KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

PREDICTIVE MODELS FOR CLASSIFICATION IN GYNECOLOGY

Jury:
Prof. dr. ir. Y. Willems, voorzitter
Prof. dr. ir. S. Van Huffel, promotor
Prof. dr. ir. J. Suykens, promotor
Prof. Dr. D. Timmerman, promotor
Prof. dr. ir. J. Vandewalle
Prof. dr. J. Beirlant
Prof. dr. L. Vermeire
Prof. Dr. L. Valentin

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Lieveke AMEYE

UDC 519.237.8:618

Juli 2005

© Katholieke Universiteit Leuven – Faculteit Ingenieurswetenschappen
Arenbergkasteel, Kasteelpark Arenberg 1, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

Wettelijk depot D/2005/7515/64

ISBN 90-5682-630-1

Voorwoord

Aan het einde van mijn doctoraat, en aan het begin van dit doctoraatsproefschrift zou ik graag de mensen bedanken die het mogelijk gemaakt hebben dit werk te presenteren.

Eerst al vooral wil ik mijn drie promotoren bedanken. Prof. Sabine Van Huffel ben ik heel dankbaar om mij de kans te geven in de Biomed onderzoeksgroep mijn doctoraat voor te bereiden en mij te blijven steunen, ook in donkere dagen van wetenschappelijk onderzoek. Prof. Johan Suykens wil ik bedanken voor de inleiding tot least squares support vector machines. Tegelijk ben ik hem dankbaar voor de vrijheid om me te verdiepen in klinische toepassingen. Prof. Dirk Timmerman ben ik dankbaar om geduldig de medische achtergrond van de klinische problemen uit te leggen, om me te introduceren bij zijn collega's en het organiseren van de vele vergaderingen. Mijn drie promotoren zou ik nog eens extra willen bedanken voor het geven van de vele tips en suggesties om mijn doctoraatsmanuscript te verbeteren.

Verder zou ik ook graag de leden van mijn begeleidingscommissie, Prof. Joos Vandewalle en Prof. Jan Beirlant, willen bedanken voor de steun die ze me tijdens dit onderzoek hebben gegeven en voor het doornemen van dit proefschrift.

Bovendien zou ik ook Prof. Yves Willems, als voorzitter, Prof. Lea Vermeire en Prof. Lil Valentin willen bedanken voor hun deelname in de jury van dit doctoraatsproefschrift.

Naast de vele wetenschappelijke discussies waren mijn collega's van de Biomed groep een onuitputbare bron van administratieve weetjes en collegialiteit. Speciaal zou ik mijn verschillende bureaugenoten willen vermelden door wie er steeds in een gezellige sfeer kon gewerkt worden: Andy, Chuan, Lukas, Cynthia, Jean-Michel, Lieven, Kristiaan, Bart, Geert, John en last but not least Ben. Tevens ben ik Jos heel dankbaar voor de vele statistische discussies tijdens de treinreizen. Ida, Ilse, Lut en Bart wil ik speciaal bedanken omdat ze altijd klaar stonden om praktische vragen en problemen op te lossen.

Tevens wil ik verschillende leden bedanken van de afdelingen gynaecologie-verloskunde en neonatologie van het U.Z. Leuven waarmee op regelmatige basis is samengewerkt: Dr. Thierry Van Den Bosch, Dr. Leilani Morales, Dr. Isabelle Cadron, Dr. Caroline Van Holsbeke, Dr. Maja Konstantinovic, Dr. Hans Daniels, Prof. Bernard Spitz, Prof. Roland Devlieger, Prof. Robert Paridaens, Prof. Hugo Devlieger, Prof. Ignace Vergote, Prof. Patrick Neven, Prof. Gunnar Naulaers, én niet te vergeten de secretaresse Linda Simons.

I want to thank all the members of the IOTA group for the cooperation on this exciting topic.

Tenslotte wens ik te benadrukken dat dit proefschrift er ook gekomen is dankzij de steun van mijn familie, waarbij ik bij deze gelegenheid vooral mijn ouders en drie broers wil bedanken.

Abstract

Computer based decision support systems are considered to be important in many fields in medicine. Humans, even being experts, have limitations when diagnosing diseases, mainly because this process is subjective, and largely depends on the experiences of the assessor and his/her interpretation on the patient data. Therefore, we have built predictive models for two major decision problems in gynecology:

1. the preoperative classification of ovarian tumors, and
2. the construction of predictive survival models for premature babies of which the mother had a preterm prelabor rupture of the membranes before 26 weeks of gestation.

For the former we proceeded in two ways. Firstly, we built one model for the whole data. This constructed model beats all the previous models described in literature. We can also conclude that measurements of serum CA 125 levels do not enhance our ability to characterize an ovarian mass.

The model could offer considerable help to less experienced operators. But it did not reach the accuracy obtained by expert operators. Therefore, we decided to look for which subgroups of patients the development of new models could be worthwhile. When combining the models for the several subsets we attained the same performance of highly trained experts in order to classify ovarian tumors pre-operatively.

We might ask the question now: "Is it possible to beat those experts?" Unfortunately, we had to conclude that we could not yet build predictive models that substantially improved diagnostic accuracy in tumors that are judged to be difficult to classify by the experts.

For the second classification problem we have constructed predictive models that showed a very good capacity in predicting the survival chances in a prospective dataset.

Korte inhoud

Artsen, ook al zijn ze een expert in hun gebied, hebben beperkingen bij het stellen van de juiste diagnose. Dit komt grotendeels omdat die beoordeling subjectief is en in grote mate afhankelijk van de ervaring van de onderzoeker en zijn/haar interpretatie van de patiëntengegevens. Computergebaseerde systemen die de medische diagnostiek ondersteunen winnen aan populariteit. Daarom hebben we voorspellingsmodellen gebouwd voor twee grote problemen in de gynaecologische diagnostiek:

1. de preoperatieve classificatie van ovariale tumoren en
2. de constructie van modellen die de overlevingskansen van baby's kunnen inschatten waarvan de moeder haar vliezen prematuur gebroken heeft voor de 26ste zwangerschapsweek.

Voor het eerste probleem zijn we als volgt tewerk gegaan. We bouwden eerst een model dat toepasbaar was op alle ovariale tumoren. Dit model presteerde beter dan alle voorafgaande modellen die in de literatuur beschreven staan. We konden ook besluiten dat CA 125 metingen de performantie niet verbetert.

Het globaal model was bijzonder nuttig om de minder ervaren clinicus te ondersteunen. Maar het bereikte niet de nauwkeurigheid van de experts. Daarom besloten we om te kijken voor welke specifieke deelgroepen het nuttig kon zijn om nieuwe modellen te ontwikkelen. Wanneer we de modellen voor de verschillende deelgroepen combineerden, evenaarden we de performantie van de expert om preoperatief ovariumkanker te detecteren.

We kunnen nu de vraag stellen: "Is het mogelijk om beter te doen dan de experts?" Jammer genoeg moesten we besluiten dat we nog geen voorspellingsmodellen kunnen bouwen die een opmerkelijke verbetering van de diagnose met zich meebrengen in tumoren die door de experts als moeilijk te classifiëren bestempeld worden.

Voor het tweede classificatieprobleem hebben we voorspellingsmodellen geconstrueerd die een bijzonder goede performantie vertonen in het voorspellen van de overlevingskansen in een prospectieve dataset.

Nederlandse samenvatting

Voorspellingsmodellen voor classificatie in de gynaecologie

Hoofdstuk 1: Inleiding

Artsen, ook al zijn ze een expert in hun gebied, hebben beperkingen bij het stellen van de juiste diagnose. Dit komt grotendeels omdat die beoordeling subjectief is en in grote mate afhankelijk van de ervaring van de onderzoeker en zijn/haar interpretatie van de patiëntengegevens. Computergebaseerde systemen die de medische diagnostiek ondersteunen winnen daarom aan populariteit. Bepaalde computerprogramma's worden al frequent gebruikt om de clinicus te helpen bij het stellen van een correcte diagnose (Coeira, 1999). Vele artificiële intelligentietechnieken werden al toegepast in de medische wereld, zoals bijvoorbeeld neurale netwerken die de diagnose van borstkanker modelleren (Jain et al, 2000), of logistieke regressie, neurale netwerken en kleinste kwadraten support vector machines die preoperatief eierstokkanker voorspellen (Tailor et al, 1997; Tailor et al, 1999; Timmerman, Bourne et al, 1999; Timmerman, Verrelst et al, 1999; Lu et al, 2003; Lu et al, 2004).

We wensen te benadrukken dat geen enkel computergebaseerd systeem dat ontwikkeld is om de medische diagnostiek te ondersteunen het oordeel van een clinicus kan vervangen. Het doel van deze systemen is om de clinicus bij te staan, niet om hem of haar te vervangen.

1.1 Doel van de thesis

Het doel van de thesis is voornamelijk het ontwikkelen van voorspellingsmodellen voor classificatie waarbij we ons op twee belangrijke problemen in de medische diagnostiek concentreren:

- De preoperatieve diagnose van ovariumkanker of kanker van de eierstokken. Het doel hierbij is het ontwikkelen van modellen die de minder ervaren clinicus kunnen ondersteunen om kwaadaardige ovariale tumoren te ontdekken, zodanig dat hij de patiënte op tijd kan doorsturen naar een oncologische chirurg.
- De constructie van voorspellingsmodellen voor de overlevingskansen van premature baby's waarvan de moeder haar vliezen voortijdig gebroken heeft (voor de 26ste week zangerschap). Het doel hier is het opbouwen van een model, gebaseerd op variabelen die gekend zijn op het moment van de opname van de patiënte, dat de overlevingskansen van de baby kan voorspellen. Een dergelijk model kan een aanzienlijke hulp zijn bij het kiezen van de meest geschikte behandeling voor moeder en baby.

Beide problemen in de gynaecologische diagnostiek gaan we nu kort toelichten.

Preoperatieve voorspelling van ovariumkanker

Adnexiële tumoren zijn een veel voorkomend probleem in de gynaecologie. Veel vrouwen met gevorderde ovariumkanker ondergaan dikwijls in een lokaal ziekenhuis niet de geschikte eerste operatie (Tingulstad et al, 1996) of worden geopereerd door chirurgen zonder de nodige ervaring in oncologische chirurgie (Kehoe et al, 1994). Een ontoereikende eerste chirurgische ingreep doet de overlevingskansen van de patiënte drastisch verkleinen, aangezien de hoeveelheid van resterend kwaadaardig weefsel na de eerste operatie één van de meest belangrijke prognostische factoren is bij ovariumkanker (Vergote et al, 2001). Het tijdig doorverwijzen van patiënten met een verdachte adnexiële tumor naar een oncologische chirurg is daarom van het grootste belang om de overlevingskansen van de vrouw te laten toenemen. Dit verklaart waarom de sensitiviteit van een model, om kwaadaardige van goedaardige adnexiële massa's te onderscheiden, belangrijker is dan de specificiteit.

Nochtans is de specificiteit ook niet onbelangrijk. Met de ontwikkeling van nieuwe chirurgische technieken is het haalbaar geworden om goedaardige tumoren te behandelen met een laparoscopische ingreep – met behulp van een kijkbuisje. Deze zogenaamde "minimaal" invasieve ingreep doet de ziektecijfers en de duur van het verblijf in het ziekenhuis dalen (Nezhat et al, 1992; Lewis & Timmerman, 1993). Bij patiënten met een functionele cyste kan daarenboven elke chirurgische ingreep onnodige ziektekosten en een mogelijk verlies van vruchtbaarheid met zich meebrengen.

Daarom is het van levensbelang om nauwkeurig preoperatief goedaardige van kwaadaardige ovariale tumoren te onderscheiden. In de hoofd-

stukken 3, 4, 5 en 6 van deze thesis behandelen we de constructie van verschillende modellen om preoperatief ovariumkanker te voorspellen.

Het voorspellen van de overlevingskansen na het prematuur breken van de vliezen

Met prematuur breken van de vliezen bedoelen we het breken van de vliezen vóór het begin van de contracties. Dit gebeurt vrij vaak, namelijk in ongeveer 10% van alle zwangerschappen (Mead, 1980). Naar schatting brengt het voortijdig breken van de vliezen complicaties met zich mee in 30 tot 40% van de premature bevallingen en in 0.7% betreft het een zwangerschap van minder dan 26 weken (Gibbs & Blanco, 1982; Graham et al, 1982; Cox et al, 1988; Schuker et Mercer, 1996). We concentreren ons hier op het probleem van prematuur breken van de vliezen vóór de 26ste zwangerschapsweek wat een uitdaging is voor de medische diagnostiek. De behandeling hiervan heeft een grote impact op zowel de gezondheid van de foetus als die van de moeder.

Het doel is om modellen te ontwikkelen die zo vroeg mogelijk de overlevingskansen van de baby kunnen inschatten. Met andere woorden, een model gebaseerd op variabelen die gekend zijn op het moment van de opname van de moeder is klinisch meer bruikbaar dan een model gebaseerd op neonatale factoren. De constructie van deze modellen wordt in hoofdstuk 7 besproken.

1.2 Overzicht van eigen bijdragen

Naast deze twee grote uitdagingen in de gynaecologische diagnostiek hebben we een bijdrage geleverd bij een aantal andere onderzoeksgebieden in de neonatologie en gynaecologie. We beschrijven deze eerst kort en daarna volgt een overzicht van de thesis.

De bijdragen tot het oplossen van andere onderzoeksvragen in de neonatologie en gynaecologie zijn de volgende:

- In Ameye et al (april 2002) konden we besluiten dat kleinste kwadraten support vector machines de mentale ontwikkeling van premature baby's in belangrijke mate beter kunnen voorspellen dan logistieke regressie.
- In (Van Den Bosch, Donders et al, 2002) kwamen we tot de conclusie dat Implanon® – een contraceptief implant – niet volledig de werking van de eierstokken onderdrukt.
- In (Van Den Bosch, Van Schoubroeck et al, 2002) willen we referentiewaarden voor de dikte van het endometrium bepalen en kunnen hiertoe drie deelgroepen beschouwen, namelijk postmenopauzale vrouwen zonder hormonenbehandeling, vrouwen met sequentiële en vrouwen met een continue hormonenbehandeling.
- We stelden vast dat patiënten met een hormonenbehandeling andere referentiewaarden voor de dikte van het endometrium nodig hebben dan postmenopausale patiënten zonder hormonenbehandeling (Van Den Bosch et al, 2003).
- We concludeerden in (Van Den Bosch et al, 2004) dat het nemen van een biopsie van het endometrium d.m.v. Pipelle®, zowel de metingen van het endometrium als de echografische karakteristieken beïnvloedt.
- In (Morales, Neven, Timmerman, Christiaens et al, 2004) kwamen we tot de conclusie dat er belangrijke verschillen zijn in de effecten van tamoxifen en aromatase inhibitors op de menopausale symptomen.
- We konden besluiten in (Morales, Neven, Timmerman, Konstantinovic et al, 2004) dat behandelingen met zowel tamoxifen als niet-steroïdale aromatase inhibitors een significante toename in de frequentie en de ernst van warmteopwellingen bij postmenopausale borstkankerpatiënten tot gevolg hebben.
- We toonden aan in (Morales et al, 2005) dat na 3 maanden gebruik van tamoxifen er abnormale veranderingen in de baarmoeder geïnduceerd worden, dit in tegenstelling tot aromatase inhibitors.
- In (Okaro et al, 2003) verbeterden we de diagnostiek bij vrouwen met chronische pelvische pijn d.m.v. nieuwe echografische merkers.

De doctoraatsthesis is verder als volgt gestructureerd:

In hoofdstuk 2 beschouwen we de verschillende stappen in de constructie van een voorspellingsmodel om de medische praktijk te ondersteunen.

In hoofdstuk 3 stellen we de IOTA (International Ovarian Tumor Analysis) dataset voor. Op deze dataset ontwikkelen we de modellen die preoperatief de diagnose van ovariumkanker kunnen stellen.

Hoofdstuk 4 beschrijft de opbouw van een model dat preoperatief kwaadaardige van goedaardige tumoren kan onderscheiden. Dit model wordt verder het *globaal* model genoemd. We bespreken het belang van sensitiviteit die in deze specifieke medische context belangrijker is dan specificiteit. We eindigen dit hoofdstuk met een antwoord te bieden op de vraag of het gebruik van de serum tumor merker CA 125 de performantie in de diagnose

van ovariumkanker verder kan verbeteren. Een gedeelte van deze resultaten werd gepubliceerd in (Timmerman et al, 2004).

In hoofdstuk 5 gaan we na hoe goed het globaal model presteert op bepaalde deelgroepen van patiënten, en die performantie wordt vergeleken met de subjectieve beoordeling van experts. Voor die deelgroepen waar er een groot verschil is, bouwen we een nieuw model. We eindigen dit hoofdstuk met het combineren van de modellen voor de verschillende patiëntengroepen om te onderzoeken of we de performantie van de experts bereikt hebben.

In hoofdstuk 6 bepalen we voor welke ovariale tumoren de experts het moeilijk vinden om de juiste diagnose – goedaardig of kwaadaardig - te stellen en we onderzoeken of het gebruik van logistieke regressiemodellen de diagnosestelling bij deze moeilijke tumoren kan verbeteren (Valentin, Ameye, Jurkovic et al, 2005). We beschrijven de echografische kenmerken van de verschillende types kwaadaardige ovariale tumoren: borderline maligne, primair invasief stadium 1, primair invasief stadium 2-4, zeldzame tumoren en metastatische tumoren (Valentin, Ameye, Testa et al, 2005). Tenslotte berekenen we voor de continue verklarende variabelen de optimale cutoff waarden voor klinisch gebruik (Timmerman, Testa et al, 2005).

Hoofdstuk 7 beschrijft de opbouw van modellen om de overlevingskansen te voorspellen van premature baby's waarvan de moeder de vliezen gebroken heeft voor de 26ste zwangerschapsweek (Ameye, Van Huffel et al, 2002; Ameye et al, 2004; Ameye et al, 2005; Cadron et al, 2004).

In hoofdstuk 8 sommen we de belangrijkste conclusies van de thesis op en stellen we mogelijkheden voor verder onderzoek voor.

De appendix beschrijft kort een aantal statistische modellen, en vermeldt de bestuursleden en belangrijkste onderzoekers van de IOTA groep.

Hoofdstuk 2: Klinische data-analyse

In dit hoofdstuk wordt het algemeen kader voor klinische data-analyse geschetst.

Eerst en vooral is het belangrijk om zoveel mogelijk fouten uit de dataset te verwijderen. Aangezien het doel van de thesis het ontwikkelen van voorspellingsmodellen is om de clinicus in zijn diagnose te ondersteunen, is het belangrijk om te beschikken over betrouwbare gegevens. Een zorgvuldige univariate analyse van de gegevens is bijzonder geschikt om de fouten op te sporen. We vermelden als voorbeeld dat de dataset van ovariale tumoren – die in het volgende hoofdstuk besproken zal worden – drie keer gecorrigeerd moest worden, ook al was er een systeem voor datacollectie ontwikkeld dat

mogelijke fouten moest voorkomen. Daarnaast is het belangrijk om ons te concentreren op de ontbrekende waarden. Is er een verband tussen het ontbreken van sommige waarden en de respons of is dit volledig willekeurig? Vervolgens moeten we multicollineariteit vermijden door clusters van hoog gecorreleerde variabelen op te sporen en in elke cluster de belangrijkste variabele weerhouden op basis van bijvoorbeeld het Akaike Information Criterion (AIC - Akaike, 1981) en het Schwarz Criterion (SC - Schwarz, 1978). Tenslotte is het van belang om bij kleine datasets aandacht te hebben voor het hanteren van de juiste statistische methode bij univariate. Bij categorische data-analyse is het gebruik van Fisher exact test bijvoorbeeld meer aangewezen dan de chi-square test.

In deze thesis maken we gebruik van twee soorten modellen, nl. classificatiemodellen en overlevingsmodellen, om een antwoord te bieden op de twee eerder geformuleerde problemen in de gynaecologische diagnostiek. We gebruiken logistieke regressie, kleinste kwadraten support vector machines en veralgemeende schattingsvergelijkingen om classificatiemodellen te construeren en construeren een overlevingsmodel met behulp van het proportioneel hazards model van Cox (Harrell, 2002; Suykens, 2002; Diggle et al, 2001; Cox, 1975).

Het is belangrijk om een voorspellingsmodel te valideren, zowel intern als extern. Met interne validatie bedoelen we dat we het model toepassen in de populatie waar de gegevens opgemeten zijn. Wanneer we een dataset hebben, kunnen we deze in twee delen opsplitsen: een training set (ongeveer 70% van de data) waarop we het voorspellingsmodel ontwikkelen en een test set (de resterende 30%) om het model intern te valideren. Met externe validatie bedoelen we het valideren van het model in een andere populatie, bijvoorbeeld in andere centra. Bij deze validatie maken we gebruik van een aantal performantiematen:

- sensitiviteit: de proportie correct geclassifieerde positieve uitkomsten. Bij de ovariale tumoren is dit de proportie correct geclassifieerde kwaadaardige tumoren.
- specificiteit: de proportie correct geclassifieerde negatieve uitkomsten, met andere woorden, de proportie correct geclassifieerde goedaardige tumoren.
- de oppervlakte onder de *Receiver Operating Characteristic* (ROC) curve. De ROC-curve wordt bekomen door de sensitiviteit uit te zetten in functie van (1 min de specificiteit) voor variërende classificatiedrempelwaarden van de voorspellingsmodellen. Bij logistieke regressie kunnen we bijvoorbeeld ervoor kiezen om alle gevallen waarbij de kans op een kwaadaardige tumor groter is dan 0.5 als maligne te bestempelen, maar we kunnen ook deze cutoff kanswaarde

verlagen tot 0.15 en zo een grotere waarde voor sensitiviteit en kleinere waarde voor specificiteit bekomen. De oppervlakte onder de ROC-curve op de test set is een maat voor de performantie van het ontwikkelde voorspellingsmodel. Deze oppervlakte varieert van 0.5 tot 1 waar een waarde tussen 0.9 en 1 duidt op een uitstekend model, een waarde tussen 0.8 en 0.9 betekent een goed model, van 0.7 tot 0.8 is een redelijk model, van 0.6 tot 0.7 is slechts een bescheiden model en minder dan 0.6 op de test set duidt op een waardeloos voorspellingsmodel.

Hoofdstuk 3: Beschrijving en verkenning van de IOTA data

In dit hoofdstuk starten we met een overzicht van voorafgaand onderzoek naar statistische modellen om preoperatief ovariumkanker te voorspellen. We verkennen vervolgens de IOTA (International Ovarian Tumor Analysis) dataset die tot op dit moment de grootste internationale database van ovariumtumoren is.

Overzicht van voorafgaand onderzoek

Momenteel wordt ovariumkanker in 75% van de gevallen pas in een gevorderd stadium vastgesteld. Daardoor leiden ze tot de hoogste sterftecijfers van alle gynaecologische kankers. De vraag die we zullen onderzoeken is: 'Welke factoren en welke combinatie van deze factoren leiden tot een optimale preoperatieve classificatie van goedaardige en kwaadaardige ovariumtumoren?" Deze vraag is klinisch van groot belang, wat blijkt uit het groot aantal publicaties in dit onderzoeksgebied (Bourne et al, 1995; Granberg et al, 1989; Jacobs et al, 1990; Lerner et al, 1994; Osmer et al, 1996; Tekay et al, 1992; Tekay et al, 1996; Timmerman, Bourne et al, 1999; Timmerman, Verrelst et al, 1999). Tijdens de voorbije jaren werden er verschillende technieken geïntroduceerd om kwaadaardige van goedaardige massa's te onderscheiden, bijvoorbeeld serum tumor merkers (zoals CA 125), transvaginale echografie in combinatie met een reeks morfologische score systemen, of magnetische resonantie beeldvorming (Granberg et al, 1989; Jacobs et al, 1990; Lerner et al, 1994; Osmers et al; 1996, Sladkevicius et al, 1995; Tekay et al, 1995; Tekay et al, 1996, Valentin et al, 1994).

Serum CA 125 is de tumor merker met de hoogste sensitiviteit voor ovariumkanker (Bast et al, 1983; Jacobs & Bast, 1989; Knapp et al, 1996; Schwartz et al, 1993). Deze merker detecteert bijna 80% van de gevorderde kwaadaardige ovariumtumoren (stadium ≥ III), maar slechts 40 tot 44% van stadium I ovariumkanker (Vergote, Børmer et al, 1987; Cuckle et al, 1991;

Bourne, Campbell et al, 1994; Maggino et al, 1994). Om de betrouwbaarheid van CA 125 te verhogen heeft Jacobs de CA 125-metingen gecombineerd met morfologische kenmerken van de tumor én de menopausale toestand van de patiënte om de *Risk of Malignancy Index* (RMI) te berekenen (Jacobs et al, 1990). Daarnaast zijn er meer dan 10 morfologische score systemen in de literatuur beschreven. Deze kennen een wisselend succes bij gebruik in de klinische praktijk (Granberg et al, 1989; Granberg et al, 1990; Sassone et al, 1991; Timor-Tritsch, 1993; Lerner et al, 1994; Bourne, 1995; DePriest, 1997; Valentin, 1997; Ferrazzi, 1997; Valentin, 2000).

Ondanks al deze publicaties was de dataset waarop al deze technieken (CA 125, RMI en morfologische score systemen) uitgetest werden iedere keer te klein om tot een rationele klinische diagnose van ovariumtumoren te komen.

Sedert begin 1997 is er een nauwe samenwerking tussen D. Timmerman en I. Vergote van het UZ Leuven en ESAT-SISTA. In een pilootproject, gebaseerd op gegevens van 300 patiënten, werd er logistieke regressie toegepast (Timmerman et al, 1999). Vervolgens werden ANNs en LS-SVMs ontwikkeld om ovariumkanker te voorspellen. In een prospectieve studie haalden de niet-lineaire modellen een significant betere performantie dan de veel gebruikte 'Risk of Malignancy Index' (Jacobs et al, 1990). Maar deze niet-lineaire ANN/LS-SVM modellen waren vrij gelijkaardig aan de lineaire/logistieke regressie modellen qua performantie, m.a.w. de modelcomplexiteit van de ANNs en LS-SVMs werd niet volledig benut (Timmerman, Verrelst, Bourne et al, 1999; Lu et al, 2003). Het was mogelijk dat voor dit classificatieprobleem een niet-lineaire scheiding nodig was. Maar dan hadden we nood aan een groter aantal datapunten die moeilijk te classifiëren waren om de niet-lineaire modellen te ontwikkelen. Een grotere dataset was ook nodig omwille van de grote histologische verscheidenheid in types van ovariale tumoren, en de verschillen in klinische achtergrond en in leeftijd van de patiënten in de populatie.

Beschrijving van de IOTA dataset

Daarom werd er besloten om een multicentrische studie op te richten waaraan gerenommeerde buitenlandse universiteiten deel namen: University Hospital Malmö (Sweden), University Hospitals Leuven (Belgium), Universita del Sacro Cuore Rome (Italy), ISBM L. Sacco University of Milan (Italy), Hôpital Boucicaut Paris (France), Centre Medical des Pyramides, Maurepas (France), King's College Hospital London (UK), ISBM Ospedale, San Gerado Universita di Milano, Monza (Italy), Universita degli Studi di Napoli and Naples (Italy). Dit consortium wordt International Ovarian Tumor Analysis (IOTA) genoemd. Deze IOTA studie is in staat om de beperkingen van de voorafgaande publicaties weg te werken door gegevens

van 1066 patiënten te onderzoeken. Drie centra: Malmö, Leuven en Rome leverden 2/3 van de data, i.e. 29.5%, 24.75 en 11.8% (zie Figuur 3-1). De datacollectie gebeurde tussen juni 1999 en juni 2002. Voor elke patiënte werden de gegevens van meer dan 50 verklarende variabelen volgens een gestandaardiseerd protocol verzameld (zie sectie 3.3.2 van de thesis voor een uitgebreide beschrijving - Timmerman, Valentin et al, 2000). Deze verklarende variabelen omvatten zowel de medische voorgeschiedenis van de patiënte, echografische en morfologische kenmerken van de tumor, als ook de serum tumor merker CA 125 en indices voor de bloedcirculatie in de tumor (die opgemeten worden met kleuren Doppler beeldvorming). De responsvariabele van deze studie is de histologische diagnose, nl. "goedaardig" of "kwaadaardig". 25% van alle tumoren is kwaadaardig waarvan 5% borderline maligne, 16% primair invasief en 4% metastatisch (zie Figuur 3-2).

Er zijn opmerkelijke verschillen tussen de centra vast te stellen. Zo heeft Milano (ISBM L. Sacco University of Milan) slechts 9.2% kwaadaardige tumoren, terwijl bij Monza 37.0% van de tumoren kwaadaardig is. Daarnaast zijn er ook opmerkelijke verschillen in de verklarende variabelen. Bijvoorbeeld Rome en Milano gebruiken slechts voor 5.6% en 3.4% van hun patiënten een hormonenbehandeling, terwijl de percentages voor Malmö en Leuven 29.8% en 31.2% bedragen (zie Tabel 3-1). Verder stelt Malmö in slechts 6% van de tumoren geen bloedcirculatie vast, terwijl dit in Leuven oploopt tot 29% en in Milano zelfs 71.3% (zie Tabel 3-2). Deze variabele is dan ook eerder subjectief van aard, dit in tegenstelling tot bijvoorbeeld de leeftijd van de patiënte. Tabel 3-3 vermeldt voor elke verklarende variabele de graad van objectiviteit. Univariate analyse van de verklarende variabelen (zie Tabel 3-4, 3-5 en 3-6) toont aan dat de verschillen tussen goedaardige en kwaadaardige tumoren voor bijna alle variabelen significant zijn.

Hoofdstuk 4: Globaal IOTA model

In dit hoofdstuk ontwikkelen we een voorspellingsmodel – dat we globaal model noemen – gebruik makend van de IOTA dataset om preoperatief ovariumkanker te voorspellen.

Constructie van het globaal IOTA model

We starten met het gestratifieerd opdelen van de totale dataset in twee aparte deelsets, nl. een training set (70% van de data) en een test set (de resterende 30%). Met stratificatie bedoelen we dat de volgende percentages in beide sets dezelfde zijn:

- 75% goedaardige en 25% kwaadaardige tumoren, en

- 7.5% van de data komt van het centrum Paris Boucicaut, 5.1% van London, 24.7% van Leuven, ...

De training set wordt gebruikt om het logistiek regressiemodel te ontwikkelen en de test set om een idee te hebben van de performantie van het ontwikkelde model op een nieuwe set van gegevens.

Door middel van grafisch onderzoek kunnen we verschillende clusters van sterk gecorreleerde variabelen vaststellen (zie Figuur 4-1, 4-2, 4-3 en 4-4). In deze clusters bepalen we telkens de belangrijkste variabele door gebruik te maken van het Akaike Information Criterion (AIC - Akaike, 1981) en Schwarz Criterion (SC - Schwarz, 1978), zie tabellen 4-1 t.e.m. 4-4. Daarnaast zijn er heel wat gestructureerde ontbrekende waarden in de totale dataset. Er zijn bijvoorbeeld 513 tumoren zonder een soliede component; dit impliceert een ontbrekende waarde voor 6 variabelen, nl. *SolidD1*, *SolidD2*, *SolidD3*, *MaxSolidD123*, *VolumeSolid* en *RatioSolidLes*. Als we enkel die tumoren zouden analyseren die geen enkele ontbrekende waarde hebben, houden we slechts 97 (van de 1066) patiënten over (zie Figuur 4-5). De waarden ontbreken omdat de variabele in kwestie niet van toepassing is op de tumor. Als er bijvoorbeeld geen soliede component in de tumor aanwezig is, kunnen we de afmetingen van die soliede component niet weergeven. Daarom gaan we de ontbrekende waarden telkens opvullen met een "0". De uitzondering op deze regel zijn de bloedcirculatie-indices. Als er geen bloedcirculatie in de tumor is worden de peak systolic velocity (*PSV*), time averaged maximum velocity (*TAMXV*), pulsatility index (*PI*) en resistance index (RI) gecodeerd als volgt: respectievelijk 2.0 cm/sec, 1 cm/sec, 3.0 en 1.0.

Wanneer we nu logistieke regressie met stapsgewijze selectie van de variabelen toepassen, worden er 14 factoren weerhouden (zie Tabel 4-5). Maar voor één variabele (*loc3*) is het effect op de response in het multivariaat logistiek regressiemodel tegenovergesteld aan het effect op de response in een univariate analyse. Daarom besluiten we om deze factor uit het model te verwijderen. Wanneer we nagaan (m.b.v. generalized additive models – Hastie en Tibshirani, 1999) of het nodig is om de continue variabelen in het model te transformeren, kunnen we besluiten dat dit enkel voor één variabele van toepassing is, nl. *MaxSolidD123M*, waar we een cutoff definiëren op 50mm. Dit impliceert dat we waarden voor *MaxSolidD123M* groter dan 50 mm gelijk stellen aan de cutoff van 50 mm. Dit heeft tot gevolg dat een andere variabele niet meer significant is in het logistiek regressiemodel (*Fluid*) en uit het model weggelaten wordt. We eindigen aldus met 12 variabelen, waarvan we er één getransformeerd hebben (zie Tabel 4-11).

Als we dit model interpreteren, komen we tot de volgende conclusies: Onafhankelijke risico factoren voor de aanwezigheid van een kwaadaardig gezwel zijn de leeftijd van de patiënte, een voorgeschiedenis van ova-

riumkanker, de maximale diameter van het letsel, de maximale diameter van de soliede component (waarbij waarden boven de 50 mm geen bijkomende risico-effect met zich meebrengen), de aanwezigheid van ascites, bloedcirculatie binnen een papillaire structuur, een volledig solied letsel, onregelmatige binnenwand van de cyste en een stijging in de Doppler kleuren score.

Factoren die het risico op de aanwezigheid van ovariumkanker reduceren zijn: een huidige hormonenbehandeling, het voelen van pijn tijdens het echografisch onderzoek en de aanwezigheid van akoestische schaduwen op het echografisch beeld.

De oppervlakte onder de ROC-curve van de training set en test set bedragen respectievelijk 0.946 en 0.942, wat duidt op een uitstekend voorspellingsmodel. Een model dat goedaardige van kwaadaardige ovariale tumoren kan onderscheiden, moet een sensitiviteit van minstens 90% en een specificiteit van minstens 75% halen om van enig nut te zijn in de klinische praktijk. Daarom kiezen we voor een cutoff kanswaarde van 0.15 in de classificatietabel van de training set (zie Tabel 4-13) wat ons een sensitiviteit van 91.6% en een specificiteit van 82.8% oplevert. Als we dit model dan toepassen op de onafhankelijke test set bekomen we een sensitiviteit van 92.0% en een specificiteit van 81.0% (zie Tabel 4-14).

Bespreking van het ontwikkelde IOTA globaal model

Als we het ontwikkelde model toepassen op de centra afzonderlijk kunnen we vaststellen dat ons globaal model in ieder centrum afzonderlijk goed presteert (zie Tabel 4-15). Daarenboven presteert dit model significant beter dan alle vorige modellen die in de literatuur beschreven worden (zie Tabel 4-16).

Dit globaal model bereikt een sensitiviteit van 92% en een specificiteit van 81% op een onafhankelijke test set. Daarom kan dit model zeer hulpvol zijn om de minder ervaren clinicus te ondersteunen. Een voorafgaande studie (Timmerman, Schwärzler et al, 1999) heeft namelijk aangetoond dat deze clinici een sensitiviteit van slechts 86% bereiken en een specificiteit van 80%.

Tenslotte gaan we na of de metingen van de serum tumor merker CA 125 de performantie van het globaal model verder kunnen verhogen. Als we deze variabele toevoegen aan het globaal model is er geen significantie verbetering in de oppervlakte onder de ROC-curve, noch bij de premenopausale patiënten, noch bij de postmenopausale patiënten (zie sectie 4.11.1 en 4.11.2). Zelfs als we volledig nieuwe modellen bouwen met CA 125 erin is er geen opmerkelijke verbetering vast te stellen (Timmerman, Van Calster et al, 2005). We kunnen dus besluiten dat er geen nood is aan CA 125 metingen om kwaadaardige van goedaardige ovariale tumoren te onderscheiden

indien we over de klinische achtergrond van de patiënte en de nodige echografische kenmerken van de tumor beschikken.

Hoofdstuk 5: IOTA deelgroepen-analyse

In het vorige hoofdstuk hebben we een model ontwikkeld - om preoperatief kwaadaardige ovariale tumoren te voorspellen - dat bijzonder nuttig kan zijn voor de minder ervaren clinicus. Maar er is nog ruimte voor verbetering. In de IOTA dataset beschikken we namelijk voor elke tumor over het oordeel van de expert van ieder centrum. Elke expert heeft alle tumoren van zijn/haar centrum op basis van de echografische kenmerken beoordeeld. Voor elke tumor gaf hij zijn diagnose: "goedaardig" of "kwaadaardig". Deze heel ervaren clinici bereiken een sensitiviteit van 86.7% en een specificiteit van 93.7% op de tumoren in de test dataset. Er is op de test data geen significant verschil tussen ons globaal model en de experts in de sensitiviteit (92% versus 86.7%, p-waarde 0.25 – Tango (1998)), maar wel in de specificiteit (81% versus 93.7%, p-waarde <0.0001). Dus ons globaal model is iets meer sensitief (maar niet significant) dan de diagnose van de experts, maar heeft een veel lagere specificiteit: meer dan 10% verschil. Het doel van verdere analyses is daarom het verhogen van de specificiteit, terwijl we de goede performantie in sensitiviteit behouden. In het vorige hoofdstuk hebben we één model gebouwd voor alle tumoren. Maar een nader onderzoek van verschillende deelgroepen van tumoren kan de mogelijkheid bieden om de performantie verder te verhogen, nl. de specificiteit te doen toenemen. Daarom bespreken we in dit hoofdstuk voorspellingsmodellen voor de verschillende deelgroepen van ovariale tumoren.

Constructie van deelgroepen

In het vorige hoofdstuk hebben we opgemerkt dat er heel wat gestructureerde ontbrekende waarden in de dataset waren. Ons globaal model bevat slechts twee variabelen die oorspronkelijk gestructureerde ontbrekende waarden hadden, nl. *MaxSolidD123* en *PapFlow*. De ontbrekende waarden in *MaxSolidD123* (als er geen soliede component was) en *PapFlow* (als er geen papillaire projecties waren) zijn opgevuld geworden met nul om het globaal model te construeren. Daarom zullen we deelgroepen creëren op basis van het al of niet ontbreken van waarden in deze twee verklarende variabelen (zie Figuur 5-1). De deelgroepen zijn dan als volgt: 1) geen soliede component, 2) soliede component, maar geen papillaire projecties en 3) papillaire projecties. De deelgroep van tumoren zonder soliede component bevat 513 massa's waarvan er 313 uniloculaire cystes zijn. Van deze 313 waren er slechts 2 kwaadaardig (borderline maligne) waardoor uniloculaire tumoren bijna deterministisch goedaardig zijn. Daarom gaan we de eerste

groep nog eens in twee delen opsplitsen: uniloculair en multiloculair (zie Figuur 5-1). Als training en test set voor de vier deelgroepen nemen we telkens die tumoren uit de training en test set van het globaal model die zich in de specifieke deelgroepen bevinden.

Voorspellingsmodellen voor de 4 deelgroepen

Uniloculaire cysten hebben slechts een heel klein risico op kwaadaardigheid. Dit wordt bevestigd door een aantal studies (Goldstein et al, 1988; Obwegeser et al, 1993; Kroon & Andolf, 1995; Gerber et al, 1997; Bailey et al, 1998; Ekerhovd, 2001). Daarom stellen we een eenvoudige regel voor: als uniloculair, dan goedaardig.

Voor de multiloculaire tumoren bekomen we met slechts drie variabelen (leeftijd, aantal holten en ascites – zie Tabel 5-16) een sensitiviteit van 1/5 en een specificiteit van 50/52 wat vrij gelijkaardig is aan de performantie van het globaal model en de experten.

Een model met vier variabelen (ascites, onregelmatige binnenwand van het letsel, maximale diameter van de soliede component – waarbij waarden boven de 100mm geen bijkomend risico-effect met zich meebrengen, en akoestische schaduwen op het echografisch beeld – zie Tabel 5-26) haalt een sensitiviteit van 97.3% op de test data van de tumoren met een soliede component maar zonder papillaire projecties (cfr. globaal model: sensitiviteit 100%) en een specificiteit van 78.6%, wat een verbetering van meer dan 10% is t.o.v. het globaal model (specificiteit van 64.3%), maar toch nog altijd lager dan het oordeel van ervaren clinici (specificiteit van 92.9%).

Voor de massa's met papillaire projecties hebben we een voorspellingsmodel ontwikkeld dat gebruik maakt van vijf factoren: bloedcirculatie binnen de papillaire structuur, aantal papillaire structuren, leeftijd, maximale diameter van de soliede component met cutoff op 50mm en de resistentie-index (zie Tabel 5-36). Alhoewel we slechts 5 variabelen gebruiken, zijn we toch in staat om de specificiteit in de test data met 30% te verbeteren in vergelijking met het globaal model, i.e. 80.0% versus 51.1%, terwijl de sensitiviteit op ongeveer hetzelfde niveau wordt gehouden: 90.6% versus 93.8% voor het globaal model.

Combinatie van de 4 voorspellingsmodellen

Als we nu deze 4 modellen combineren bekomen we een sensitiviteit van 88.0% en een specificiteit van 90.3% op de totale test data. Er is geen significant verschil meer met de performantie van ervaren clinici (sensitiviteit 86.7% en specificiteit 93.7%). Dus hebben we een model gebouwd met dezelfde nauwkeurigheid als die van de experts om preoperatief ovariale tumoren te classifiëren!

Hoofdstuk 6: Vergelijking van echografische kenmerken van goedaardige en kwaadaardige tumoren

In dit hoofdstuk gaan we na of het mogelijk is om een model te bouwen dat preoperatief een onderscheid kan maken tussen goed- en kwaadaardige massa's voor die tumoren waar experts het moeilijk mee hebben om een juiste diagnose te stellen. Daarna vergelijken we de echografische kenmerken van de verschillende soorten kwaadaardige tumoren: borderline maligne, primaire invasieve tumoren in een vroeg en in een gevorderd stadium, zeldzame tumoren en metastatische massa's.

Is het mogelijk om de diagnose beter te stellen dan de experts?

Een ervaren clinicus kan op basis van een goed echografisch systeem in de meeste gevallen een juist onderscheid maken tussen goed- en kwaadaardige ovariale tumoren (Valentin, 1999; Timmerman, Schwärzler et al, 1999). In de totale IOTA dataset bereikten de experts een sensitiviteit van 88.0% en een specificiteit van 95.3%. Nochtans zal zelfs een expert in een klein aantal van de gevallen het moelijk vinden om de juiste diagnose te stellen (Valentin, 1999). In de totale dataset zijn er 90 moeilijk te classifiëren en 976 classifieerbare massa's. De sensitiviteit en de specificiteit van de experts bedraagt 56% (14/25) en 77% (50/65) in de moeilijke tumoren. De corresponderende cijfers voor de classifieerbare tumoren bedragen 91% (220/241) en 97% (712/735). Zowel de sensitiviteit als de specificiteit is significant lager bij de moeilijke massa's dan bij de classifieerbare tumoren (p-waarde <0.0001 voor beide). Typische kenmerken voor moeilijke tumoren zijn: de aanwezigheid van papillaire structuren, multilocullariteit met meer dan 10 locules en laag niveau van echogeniciteit (zie Tabel 6-2).

Als we een model ontwikkelen voor alle moeilijke tumoren vinden we geen enkele significante verklarende variabele die een onderscheid kan maken tussen de goed- en de kwaadaardige massa's. Daarom concentreren we ons op een deelgroep: de tumoren met papillaire structuren (omdat dit een typisch kenmerk is van de moeilijke tumoren). De hoogte van de papillaire structuur is dan significant, maar dit resulteert niet in een goed voorspellingsmodel: op de test data bekomen we geen betere resultaten dan de experts. Daarom kunnen we besluiten dat we nog niet in staat zijn om de diagnose beter te stellen dan de experts. Zolang we geen andere verklarende variabelen (bijvoorbeeld proteomics patterns of intraveneuze echografische contrasttoediening) gevonden hebben die in staat zijn om een correcte diagnose te stellen bij de moeilijke tumoren, zullen we moeten aanvaarden dat sommige vrouwen een onnodige operatie zullen ondergaan omwille van het

feit dat we niet in staat zijn om op een betrouwbare manier kanker uit te sluiten voor de operatie.

Echografische kenmerken van kwaadaardige ovariale tumoren.

In de klinische praktijk is het niet alleen nuttig om goedaardige van kwaadaardige ovariale tumoren te onderscheiden, maar ook de types van maligniteit: borderline tumoren, primair invasief epitheliale ovariumkanker stadium 1, primair invasief epitheliale ovariumkanker stadium 2 tot 4, zeldzame types van kwaadaardigheid en metastatische tumoren (zie Tabel 6-7 en Tabel 6-8).

Metastatische tumoren zijn in het algemeen klein, solied en goed doorbloed. Zeldzame tumoren zijn in het algemeen groot, solied en goed doorbloed. 55% (39/67) van de zeldzame en metastatische tumoren zijn solied tegenover 23% (45/199) van de borderline en primair invasief epitheliale ovariumkankertumoren ($p<0.0001$). Vocht binnen de Douglasholte en ascites zijn typische kenmerken voor metastatische en zeldzame tumoren.

Borderline ovariale tumoren en primair invasief epitheliale ovariumkanker hebben een aantal kenmerken gemeen die significant verschillen van die van stadium 2 tot 4 primair invasief epitheliale ovariumkankertumoren (p-waarden: zie Tabel 6-8): ze zijn groter; een grotere proportie bevat papillaire projecties, ze zijn vaker multiloculair en minder frequent volledig solied.

Het percentage van solied weefsel in de tumor neemt toe met de graad van kwaadaardigheid: de mediaan bedraagt 2% bij de borderline tumoren, 10% voor stadium 1 massa's, 34% voor stadium 2 tot 4 tumoren, 100% voor zeldzamé tumoren en 100% voor de metastatische tumoren (zie Tabel 6-7).

We kunnen besluiten dat de verschillende types van kwaadaardige tumoren verschillende echografische kenmerken vertonen. Het gebruik van deze echografische karakteristieken om het type van maligniteit te voorspellen moet prospectief getest worden.

Hoofdstuk 7: Voorspellingsmodellen voor de overlevingskansen van de baby na het prematuur breken van de vliezen

Met de term prematuur breken van de vliezen verwijzen we naar het breken van de vliezen vóór het begin van de contracties. Dit gebeurt in 5 tot 10% van alle zwangerschappen (Mead, 1980). In 25% van de gevallen ge-

beurt dit vóór de 37ste zwangerschapsweek, wat resulteert in 30 tot 40% van alle premature geboortes, en in 0.7% gaat het om een zwangerschap van minder dan 26 weken (Gibbs & Blanco, 1982; Graham et al, 1982; Cox et al, 1988; Taylor & Garite, 1984; Schuker & Mercer, 1996).

Hier behandelen we enkel het breken van de vliezen voor de 26ste week. Het neonataal sterftecijfer is ongeveer 40% (Taylor & Garite , 1984; Moretti & Sibai, 1988; Beydoun & Yasin, 1986; Major & Kitzmiller, 1990; Hadi et al, 1994; Fortunato et al, 1994; Morales & Talley, 1993; Rotschild et al, 1996; Kilbride et al, 1996). De behandeling van het breken van de vliezen in de niet-levensvatbare periode heeft een grote invloed op zowel de gezondheid van de foetus als die van de moeder. Het doel is om een model te construeren dat de overlevingskansen van de baby kan inschatten.

De medische gegevens van 119 opeenvolgende zwangerschappen met prematuur breken van de vliezen vóór de 26ste week – 148 baby's – werden retrospectief geanalyseerd. Deze gegevens werden opgenomen tussen 1988 en 1996. De dataset bevat 25 verklarende variabelen waaronder 5 neonatale factoren (zie Tabel 7-1) en de respons variabele is het aantal dagen dat de baby overleefde. In het geval dat de baby gestorven was voor zijn/haar geboorte, specifieert de respons variabele of dit antenataal of intranataal gebeurd is. 7% van de 148 baby's zijn antenataal gestorven, 20% intranataal, 24% op de eerste dag, 6% van dag 2 tot dag 7, 4% van dag 8 tot dag 28, 6% na de eerste maand en 35% van de baby's hebben overleefd.

In de periode van 1997 tot 200 werd een prospectieve dataset verzameld, die gegevens van 52 zwangerschappen (64 baby's) bevat. De voorspellingsmodellen worden gebouwd op de eerste dataset en getest op de prospectieve set. Voor de constructie van deze modellen maken we gebruik van de volgende methodes: het klassieke proportioneel hazards model van Cox (Cox, 1975), logistieke regressie (Hosmer and Lemeshow, 2001), veralgemeende schattingsvergelijkingen (generalized estimating equations – GEE - Diggle et al, 2001) en kleinste kwadraten support vector machines (Suykens et al, 2002). Het doel is om tot een model te komen dat zo vroeg mogelijk de overlevingskansen van de baby kan inschatten: een model gebaseerd op variabelen, die gekend zijn op het moment van de opname van de moeder is klinisch meer bruikbaar dan een model gebaseerd op neonatale factoren.

Als we het overlevingsmodel van Cox toepassen op de data, verkrijgen we een veralgemeende R^2 van minder dan 10%. Dit wil zeggen dat de performantie van het voorspellend karakter heel laag is, zodat dit model niet erg nuttig is voor gebruik in de klinische praktijk.

Als we kijken naar de respons variabele zien we in feite drie klassen voor de eerste helft van de data: antenataal, intranataal en op de eerste dag, terwijl de tweede helft eerder continu verloopt. Als we hiermee rekening houden, kunnen we veronderstellen dat we een betere performantie zullen

bekomen als we een classificatiemodel construeren – waar we de respons variabele in verschillende klassen opdelen – dan als we een continu model bouwen, waar we correcties moeten maken voor de vele samenvallende datapunten in de respons.

Vanuit klinisch standpunt zijn er twee belangrijke eindpunten in de respons variabele: 1) overleving van het geboorteproces en 2) overlevingskans op lange termijn voor de baby. Per definitie wordt de overleving van het geboorteproces bepaald als het overleven van de eerste dag. En overleven op lange termijn wordt gedefinieerd als het overleven van de eerste maand. We ontwikkelen dus classificatiemodellen om de overlevingskansen van de eerste dag en de eerste maand te voorspellen.

Nu zijn er twee manieren om naar de data te kijken. Een eerste wijze is om de data van de 148 baby's te gebruiken om tot een voorspellingsmodel te komen. We moeten dan de afhankelijkheid in de respons van baby's van dezelfde zwangerschap (tweelingen en drielingen) in rekening brengen. Anderzijds kunnen we de data ook zien als gegevens van 119 zwangerschappen. Dan hebben we te maken met volledig onafhankelijke datapunten. Voor meerlingzwangerschappen definiëren we de reponse als de respons variabele van de baby die het langst overleefde. Dus, in het geval van enkelvoudige zwangerschappen onderzoeken we of de baby de eerste dag of eerste maand overleefde, terwijl we bij meerlingzwangerschappen onderzoeken of minstens één van de baby's kans maakt om de eerste dag of eerste maand te overleven. De geconstrueerde voorspellingsmodellen staan in Tabel 7-3, 7-5, 7-7 en 7-9.

Als we deze modellen prospectief toepassen, kunnen we vaststellen dat de tweede manier om naar de data te kijken het meeste voordeel heeft: wat betreft de overlevingskansen van de eerste maand behalen we een uitstekende sensitiviteit van 97% en specificiteit van 100%. We dienen wel op te merken dat er een significant verschil is in het percentage van baby's dat overleefde in de oorspronkelijke dataset (periode 1988-1996) in vergelijking met de prospectieve set: 35% versus 52%. Het is mogelijk dat de performantie van ons model overschat blijkt te zijn, wanneer we in de toekomst het model toepassen op een populatie met een lager percentage van overleving.

Hoofdstuk 8: Conclusies en voorstellen voor verder onderzoek

We hebben voorspellingsmodellen gebouwd voor twee grote problemen in de gynaecologische diagnostiek:

- de preoperatieve classificatie van ovariale tumoren en
- de constructie van modellen die de overlevingskansen van baby's kunnen inschatten waarvan moeder haar vliezen prematuur gebroken heeft voor de 26ste zwangerschapsweek.

Voor het eerste probleem zijn we als volgt tewerk gegaan. We bouwden eerst een model dat toepasbaar was op alle ovariale tumoren, waarbij we een sensitiviteit van 92% en een specificiteit van 81% bereikten op test data. Dit model, dat we het *globaal* model noemden, presteerde beter dan alle voorafgaande modellen, die in de literatuur beschreven staan. We konden ook besluiten dat CA 125 metingen de performantie niet verbetert, noch bij de premenopausale patiënten, noch bij de postmenopausale vrouwen.

Het globaal model was bijzonder nuttig om de minder ervaren clinicus te ondersteunen. Maar het bereikte niet de nauwkeurigheid van de experts: deze bereikten een specificiteit van 93.7% in vergelijking met 81% voor ons globaal model. Daarom besloten we om te kijken voor welke specifieke deelgroepen het nuttig kon zijn om nieuwe modellen te ontwikkelen. We ontdekten twee groepen waar de specificiteit van het globaal model heel laag was:

- 64.3% in de tumoren met een soliede component, maar zonder papillaire projecties, en
- slechts 51.1% in tumoren met papillaire projecties.

Daarom besloten we om voor deze deelgroepen nieuwe modellen te bouwen. Wanneer we de modellen voor de verschillende deelgroepen combineerden, bereikten we een sensitiviteit en een specificiteit van respectievelijk 88% en 90.3%. Dit was een toename van bijna 10% in de specificiteit in vergelijking met het globaal model. Er was niet langer een statistisch significant verschil met de performantie van de expert. Met andere woorden, we hebben ons doel bereikt van het bouwen van een voorspellingsmodel dat de prestatie van zeer ervaren clinici evenaart om preoperatief ovariumkanker te detecteren.

We kunnen nu de vraag stellen: "Is het mogelijk om beter te doen dan de experts?" Jammer genoeg moesten we besluiten dat we nog geen voorspellingsmodelen kunnen bouwen die een opmerkelijke verbetering van de diagnose met zich meebrengen in tumoren die door de experts als moeilijk te classifiëren bestempeld worden.

Wanneer we de echografische kenmerken van de verschillende types kwaadaardige ovariale tumoren bestudeerden, kwamen we tot de conclusie dat papillaire projecties karakteristiek zijn voor borderline tumoren en primair invasief epitheliaal ovariumkanker in het eerste stadium. Een klein percentage van solied weefsel in de tumor komt meer voor bij borderline tumo-

ren of ovariumkankertumoren in een vroeg stadium dan bij gevorderde of zeldzame tumoren.

We hebben ook optimale cutoff waarden van de echografische variabelen gedefinieerd voor gebruik in de klinische praktijk.

Voor het tweede classificatieprobleem hebben we voorspellingsmodellen geconstrueerd die de overlevingskansen van premature baby's kunnen inschatten waarvan de moeder haar vliezen gebroken heeft vóór de 26ste zwangerschapsweek. Met een sensitiviteit van 97% en een specificiteit van 100% vertoonde een logistiek regressiemodel een bijzonder goede performantie in het voorspellen van de overlevingskansen van de eerste maand in een prospectieve dataset.

Verder onderzoek

IOTA fase II is verleden maand opgestart (juni 2005). Het doel is om nieuwe gegevens van meer dan 1000 patiënten te verzamelen om prospectief onze twee modellen (globaal en deelgroepenmodel) te testen in nieuwe centra. De hoofdvraag is: "Welk model zal de beste performantie vertonen in termen van nauwkeurigheid, sensitiviteit, specificiteit, stabiliteit en gebruiksvriendelijkheid in de klinische praktijk?"

Daarnaast zal de nieuwe database de onderzoekers in staat stellen om de ovariale tumoren verder onder te verdelen (bijvoorbeeld een onderscheid maken tussen functionele en niet-functionele goedaardige tumoren, primair invasieve en metastatische tumoren) waardoor we de diagnose kunnen verfijnen.

Een tweede belangrijk punt voor verder onderzoek is de ontwikkeling van nieuwe modellen die kunnen helpen in het classifiëren van tumoren die door experts bestempeld worden als moeilijk te classifiëren tumoren. Deze nieuwe modellen kunnen nieuwe informatie bevatten, bijvoorbeeld genomische en proteomische analyses.

Het voorspellingsmodel voor de overlevingskansen van premature baby's, waarvan de moeder haar vliezen voor de 26ste zwangerschapsweek gebroken heeft, moet extern gevalideerd worden om na te gaan of de uitstekende resultaten veralgemeenbaar zijn.

Contents

VOORWOORD I

ABSTRACT III

KORTE INHOUD V

NEDERLANDSE SAMENVATTING VII

CONTENTS XXVII

ACRONYMS AND ABBREVIATIONS XXXI

CHAPTER I 1

INTRODUCTION 1

1.1 Goal of the thesis 2

1.2 Overview of own contributions 4

CHAPTER II 9

CLINICAL DATA ANALYSIS 9

2.1 Introduction 9

2.2 Preprocessing steps 9

2.3 Statistical models 12

2.4 Validation 13

2.5 Evaluation of Performance 15

2.6 Conclusions 20

CHAPTER III 21

DESCRIPTION AND EXPLORATION OF THE IOTA DATA 21

3.1 Introduction 21

3.2 Patient entry 26

3.3 Variables 27

3.4 Data exploration 35

3.5 Conclusions 42

CHAPTER IV 43

OVERALL IOTA MODEL 43

4.1 Introduction 43

4.2 Multicollinearity 45

4.3 Missing values 51

4.4 Construction of dummy variables 53

4.5 Model building steps 54

4.6 Model Interpretation 64

4.7 Model performance 66

4.8 Validation 73

4.9 Comparison with 'old' models 75

4.10 Discussion 76

4.11 Is the serum tumor marker CA 125 needed? 78

4.12 Conclusion 83

CHAPTER V 85

IOTA SUBSET ANALYSIS 85

5.1 Introduction 85

5.2 Creation of subsets 87

5.3 Subgroup 1: Unilocular cysts 91

5.4 Subgroup 2: Multilocular + unclassifiable 94

5.5 Subgroup 3 : Solid component, no papillary projections 100

5.6 Subgroup 4: Papillary projections 106

5.7 Discussion 112

5.8 Conclusions 115

CHAPTER VI 117

COMPARISON OF ULTRASOUND CHARACTERISTICS OF BENIGN AND MALIGNANT TUMORS 117

6.1 Is it possible to beat the experts? 117

6.2 Ultrasound characteristics of malignant ovarian tumors. 123

6.3 Cutoff values for the continuous ultrasound variables and blood flow indices 127

6.4 Conclusions 130

CHAPTER VII 131

PREDICTIVE MODELS FOR SURVIVAL AFTER PREMATURE RUPTURE OF MEMBRANES 131

7.1 Introduction 131

7.2 Survival analysis 137

7.3 Classification 138

7.4 Prospective Analysis 149

7.5 Discussion 152

7.6 Conclusions 154

CHAPTER VIII 157

CONCLUSIONS AND FUTURE RESEARCH 157

8.1 General conclusions 157

8.2 Future research 159

APPENDIX 161

A Logistic Regression 161

B Least Squares Support Vector Machines 164

C Generalized Estimating Equations 166

D Cox's Proportional Hazards Model 167

E International Ovarian Tumor Analysis group 169

BIBLIOGRAPHY 171

PUBLICATION LIST 185

CURRICULUM VITAE 191

Acronyms and abbreviations

AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AUC	Area under the ROC curve
CDI	Color Doppler Imaging
CT	Computerized Tomography
GEE	Generalized Estimating Equation
IM	Informative Missing
IOTA	International Ovarian Tumor Analysis
LR	Logistic Regression
LS-SVM	Least Squares Support Vector Machine
MAR	Missing at Random
MCAR	Missing Completely at Random
MRI	Magnetic Resonance Imaging
NPV	Negative Predictive Value
PI	Pulsatility Index
PPROM	Preterm Prelabor Rupture of the Membranes
PPV	Postive Predictive Value
PROM	Preterm Rupture of the Membranes
PSV	Peak Systolic Velocity
ROC	Receiver Operating Characteristic
RI	Resistance Index
RMI	Risk of Malignancy Index
SC	Schwarz Criterion
SVM	Support Vector Machines
TAMXV	Time Averaged Maximum Velocity

Chapter I

Introduction

Recent advances in technologies and computations have facilitated the collection of biomedical data that can be used for medical decision support. The collected data come from various sources and systems, ranging from clinical patient data, diagnostic images, biological data, genetic profile and population genetic information. This world of information leads to the development of computer-based tools that are able to handle and analyze various types of medical data in order to assist the clinical decision making process.

Computer based decision support systems are considered to be important in many fields in medicine. Humans, even being experts, have limitations when diagnosing diseases, mainly because this process is subjective, and largely depends on the experiences of the assessor and his/her interpretation on the patient data.

Since 1970s, the application of computer systems based upon technologies of artificial intelligence in the medical domain has been gaining popularity (Miller, 1994; Coiera, 2003; Mani et al, 1999; Van Bemmel et al, 1998). Some computer programs, for assisting clinicians in the process of diagnosis, are now in routine use in many clinical situations (Coiera, 1999). The use of microarrays in human genome analysis has urged the use of computer techniques and reframes many classes of clinical decision (Morel et al, 2004). Such computerized intelligent systems using explicit knowledge to generate patient specific advice or interpretation are typically described as clinical decision support systems. The most common types of clinical decision support systems in routine clinical use are expert systems, also known as knowledge-based systems. They contain clinical knowledge typically represented in sets of rules, usually about a very specifically defined task, and are able to reason with data from individual patients to come up with reasoned conclusions. Machine learning is on the other hand focusing on development of systems that can learn from experience and discover new knowledge. Learning techniques include e.g. logistic regression, neural networks

and support vector machines. For example, some systems are able to learn decision trees from examples taken from data, which can be used to help with diagnosis in future patients.

Healthcare has formed a rich test-bed for machine learning experiments in the past, allowing scientists to develop complex and powerful learning systems. There are many situations in which models can make a significant contribution. One direction is modeling some of the human actions or thinking processes and recognizing diseases from a variety of input sources, e.g. cardiograms, ultrasound scans, MRI, etc. Alternatively, quantitative predictive models can be built for diagnosis based upon various types of clinical information and different tests performed upon the patient, e.g. neural networks have been applied to breast cancer diagnosis (Jain et al, 2000). Logistic regression, neural networks and least squares support vector machines models have been constructed in order to predict pre-operatively ovarian cancer (Tailor et al, 1997; Tailor et al, 1999; Timmerman, Bourne et al, 1999; Timmerman, Verrelst et al, 1999, Lu et al, 2003; Lu et al, 2004) .

We should emphasize that no decision support system could or should replace medical judgment, as it is unlikely that any system could cope with the complexity of patient interactions with medical care. *The purpose of these systems is to assist clinicians, not to replace them.* The challenges for implementing a competent decision support system are definitely not easy to overcome. However, with further advancement in technology and human knowledge, an intelligent clinical decision support tool may not be too far away.

1.1 Goal of the thesis

The goal of the thesis is basically to develop predictive models for classification in which we focus on two major medical decision problems:

- The pre-operative discrimination between benign and malignant adnexal masses. The purpose is to develop models that can assist less experienced clinicians in order to detect malignant ovarian masses, so that they can refer the patient in time to an oncologic surgeon.
- The construction of predictive models for survival of premature babies of which the mother had a preterm prelabor rupture of the membranes. The purpose is to build a model, based on variables known at the moment of intake of the mother, that can predict the survival chances of the baby. Such a model can offer considerable help in counseling and choosing the best treatment modalities.

Both gynecological decision problems will be described shortly.

Pre-operative discrimination between benign and malignant ovarian tumors

Adnexal masses are a very common problem in gynecology. Since certainty about the nature of an adnexal mass can only be gained after histopathological examination, the traditional strategy for establishing the final diagnosis has been to perform an exploratory laparotomy. However, many surgeons are unexpectedly faced with ovarian cancer at the time of exploratory laparotomy. Therefore, women with advanced ovarian cancer often undergo insufficient primary surgery at local hospitals (Tingulstad et al, 1996) or in the hands of surgeons without proper skills and experience in debulking surgery (Kehoe et al, 1994). Insufficient primary surgery leads to understaging of patients with ovarian cancer and to poorer survival, since the amount of residual malignant tissue after primary surgery is one of the most important prognostic factors in ovarian cancer (Vergote et al, 2001). The timely referral of patients with a suspicious adnexal mass to an oncologic surgeon for the appropriate pre-, per-, and postoperative management is therefore of utmost importance in order to decrease mortality from ovarian cancer, which has the highest mortality figures of all pelvic cancers in females. This explains why sensitivity figures for malignancy are more important than specificity figures for tests to discriminate between benign and malignant adnexal masses.

However, specificity must not be ignored either. With the development of new surgical techniques it has become technically feasible to treat or to remove most benign adnexal tumors with laparoscopy surgery. This so-called minimally invasive surgery decreases morbidity and hospitalization length (Nezhat et al, 1992; Lewis & Timmerman, 1993). Moreover, in patients presenting with functional cysts any surgical intervention could create unnecessary morbidity, costs, and a possible loss of fertility.

Therefore, it is vital to develop accurate tests for pre-operative discrimination between benign and malignant adnexal masses. In Chapter III, IV, V and VI of this thesis, we consider the construction of different models in order to predict pre-operatively ovarian cancer.

Predictive models for survival after premature rupture of membranes

Premature rupture of membranes (PROM) refers to membrane rupture before the onset of uterine contractions. This is a fairly common event,

occurring in approximately 10% of all pregnancies (Mead, 1980). However, it is estimated that PROM complicates 30 to 40% of the preterm deliveries and in 0.7 % it concerns a pregnancy of less than 26 weeks of gestation (Gibbs & Blanco, 1982; Graham et al, 1982; Cox et al, 1988; Taylor & Garite, 1984; Schuker & Mercer, 1996).

We focus on the problem of preterm prelabor rupture of the membranes (PPROM) before 26 weeks of gestation. The management of PPROM in the previable period is among the most controversial issues in perinatal medicine and has a major impact both on fetal and maternal health and challenges the process of medical decision making.

The purpose of this study is to build a predictive model for survival of the baby as early as possible during the pregnancy, i.e. a model based on variables known at the intake of the mother is clinically more useful than a model based on neonatal factors, in order to assist the clinician choosing the best treatment modalities. The different predictive models for survival are discussed in chapter VII.

1.2 Overview of own contributions

In addition to these two major classification problems in gynecology, we have also contributed to a number of other topics in gynecology. These are shortly described below, and we conclude with an outline of the thesis.

1.2.1 Overview of contributions to other medical studies

In Ameye et al (April 2002) we constructed a model for predicting the probability of a normal mental development for premature babies using Logistic Regression (LR) and Least Squares Support Vector Machines (LS-SVM) based on a dataset of 170 observations and 26 explanatory neonatal variables. We could conclude that LS-SVMs significantly outperform LR models in order to predict the mental development at birth.

We assessed the ultrasound features of the endometrium and ovaries in an etonogestrel implant (Van Den Bosch, Donders et al, 2002) using a data set of 188 consecutive women. We could conclude that the use of an etonogestrel implant leads to incomplete ovarian inhibition and estrogen stimulation of the endometrium.

In (Van Den Bosch, Van Schoubroeck et al, 2002) we evaluated the thickness of the endometrial lining, as well as its sonographic features, in postmenopausal women in different hormonal replacement therapy schemes.

The data contained 484 vaginal ultrasound examinations of 239 consecutive postmenopausal women on hormonal replacement therapy. Of those 239 women, 106 had at least one follow-up vaginal scan. One of the conclusions was that in defining reference values for endometrium thickness, three subgroups may be considered: the women without, those on sequential and those on continuous hormonal replacement therapy.

We evaluated the thickness and the sonographic features of the endometrium in postmenopausal women using hormonal replacement therapy and related them with pathology based on 484 vaginal ultrasound examinations in 239 consecutive patients in (Van Den Bosch et al, 2003). We could conclude that patients on hormonal replacement therapy need different reference values for endometrial thickness compared to postmenopausal patients without hormonal replacement therapy.

We evaluated if – and to what extent- Pipelle® endometrial sampling influences the ultrasound features of the endometrium in (Van Den Bosch et al, 2004). The data contained three consecutive endometrial thickness measurements and records of endometrial sonographic features before and after Pipelle® sampling in 99 consecutive women. We could conclude that endometrial thickness measurements are affected by Pipelle® sampling and the quality of ultrasound features tends to be altered by biopsy.

In (Morales, Neven, Timmerman, Christiaens et al, 2004) we compared differences in menopausal symptoms and tolerance between tamoxifen and third generation aromatase inhibitors and identified factors influencing these symptoms using a questionnaire that was filled-in at baseline, after one and after three months, by 181 consecutive postmenopausal breast cancer patients. Significant differences between the early effects of tamoxifen and aromatase inhibitors on menopausal symptoms could be detected.

The aim of (Morales, Neven, Timmerman, Konstantinovic et al, 2004) was to compare the early effects of tamoxifen, and steroidal and non-steroidal aromatase inhibitors on the occurrence and severity of menopausal symptoms in postmenopausal breast cancer patients. Therefore, we used a validated menopause symptom questionnaire that was completed by 181 consecutive breast cancer patients prior to the treatment, and again after one and three months of therapy, and we used transvaginal ultrasound records before and after three months of therapy that were measured in 77 of those 181 patients. We could conclude that treatments with either tamoxifen or non-steroidal aromatase inhibitors resulted in a significant increase in the occurrence and the severity of hot flashes. Musculosketelal pain increased under non-steroidal aromatase inhibitors, while no important change occurred in patients receiving firstline tamoxifen and in those crossing-over from tamoxifen to a steroidal or non-steroidal aromatase inhibitor.

We compared in (Morales et al, 2005) early uterine changes occurring in postmenopausal breast cancer patients treated in first line with tamoxifen or third generation aromatase inhibitors and assessed the effect of those aromatase inhibitors on tamoxifen-induced uterine changes. The data contained transvaginal ultrasound records of 77 consecutive postmenopausal breast cancer patients before and after 3 months of therapy. The study demonstrated that tamoxifen induces uterine abnormalities from 3 months of therapy, in contrast to aromatase inhibitors.

In (Okaro et al, 2003), we assessed the accuracy of new ultrasound based markers and compared them to conventional ultrasound methods used in the detection of common pelvic pathology in women with chronic pelvic pain based on a data set of 120 patients. Our approach improved the detection and exclusion of significant pathology in women with chronic pelvic pain.

We conclude this introduction with an outline of the thesis.

1.2.2 Chapter-by-chapter overview of own contributions

In Chapter II, we consider the different issues of building a predictive model for clinical practice. We discuss the issues of errors in the data, missing values, multicollinearity and small datasets. We shortly consider logistic regression, Least Squares Support Vector Machines, Generalized Estimating Equations and Cox s Proportional Hazards model and compare them with the alternatives in machine learning. Then we explain two aspects of model performance: calibration and discrimination, where we focus on the importance of highly sensitive and specific predictive models. And finally, we discuss the issue of validation of predictive models, i.e. internal validation and external validation.

Chapter III introduces the IOTA (International Ovarian Tumor Analysis) trial. This trial provides the largest data set on ovarian tumors so far: more than one thousand patients. In this chapter we give an overview of state-of-the-art research in the area of pre-operative prediction of the nature of adnexal masses. We describe in full detail the outcome and explanatory variables and explore the data.

Chapter IV presents the model building steps of the construction of a predictive model - which we call *overall* model - for the pre-operative discrimination between benign and malignant adnexal masses using the IOTA trial. Once we obtain a final model, we discuss the clinical importance of sensitivity, which is in this clinical context more significant than specificity. Finally, we assess whether the serum tumor marker CA 125 can improve the

performance of the model. Some of the results have been published in (Timmerman et al, 2004).

In Chapter V we investigate how well the IOTA *overall* model performs on different subgroups of patients, and we compare this performance with the subjective assessment of experts. We examine whether there are subgroups with a big difference between the performance of the developed *overall* model and the expert s opinion. For these subgroups where improvements can be made, we develop a new model. Finally, we combine the different submodels in order to assess whether we reach the level of performance of highly trained operators.

In Chapter VI we investigate which extrauterine pelvic masses are difficult to correctly classify as benign or malignant on the basis of ultrasound findings, and investigate whether the use of logistic regression models for calculating the individual risk of malignancy can improve diagnostic accuracy of experts in difficult tumors (Valentin, Ameye, Jurkovic et al, 2005). We discuss the ultrasound characteristics of different types of extrauterine pelvic malignancies, i.e., borderline ovarian tumors, primary invasive ovarian epithelial cancer stage 1, primary invasive ovarian epithelial cancer stage 2 4, rare types of malignancy, and metastatic tumors (Valentin, Ameye, Testa et al, 2005). Finally, we calculate optimal cut-off points of the continuous explanatory variables for use in clinical practice (Timmerman, Testa et al, 2005).

Chapter VII describes the construction of predictive models (using Cox's proportional hazards model, logistic regression, Least Squares Support Vector Machines and generalized estimating equations) for survival of premature babies of which the mother had a preterm prelabor rupture of the membranes (Ameye, Van Huffel et al, 2002; Ameye et al, 2004; Ameye et al, 2005; Cadron et al, 2004) .

Chapter VIII summarizes the thesis and suggests topics for further research.

The appendix shortly describes different statistical models: logistic regression, least squares support vector machines, generalized estimating equations and Cox s proportional hazards model en states the steering comittee and principal investigators of the IOTA group.

Chapter II

Clinical Data Analysis

In this chapter, the different issues of building a predictive model for clinical practice are considered. We start with describing the preliminary steps before constructing a model, i.e. the issues of errors in the data, missing values, multicollinearity and small datasets. We shortly consider different statistical models for classification and survival. We discuss the importance of validation of predictive models, i.e. internal validation and external validation. And finally, we explain the two aspects of model performance: calibration and discrimination, where we focus on the importance of highly sensitive and specific predictive models.

2.1 Introduction

Predictive models are considered to be important in many fields in medicine because the experts have limitations when diagnosing diseases, mainly because this process depends on the experiences of the investigator. Therefore, it is important to carefully construct those predictive models. We should pay attention to the fact that medical data can contain errors and choose an appropriate statistical method in order to model the data. All predictive models should at least be internally validated on a test set and preferably also externally on a prospective set.

2.2 Preprocessing steps

Removing errors in the data

The first step in clinical data analysis is the removal of all errors in the data. This can not be considered as a statistical or medical discipline but may nevertheless not be omitted. Since the goal is to develop predictive

models to assist clinicians in their decision making, we should at least possess reliable data to construct those models.

Even when a data set is collected by using a website where the data has to fulfill a number of conditions, for example the values of certain variables should not exceed a prespecified limit or missings are not allowed, errors can not be excluded. Medical doctors, being humans, can always commit typing errors.

Careful univariate data analysis can be an excellent tool to detect those errors. They can appear as outliers in pairwise scatter plots or can be detected by the use of contingency tables.

As an example, we mention that the data set for the ovarian cancer study, which will be described in the next chapter, had to be corrected three times, although a data collection system was developed to prevent possible errors.

Missing values

Missing data appear in the majority of datasets. There are different types of missing values:

- Missing completely at random (MCAR) means that the data are missing for reasons that are unrelated to any explanatory variable or outcome for the patients. For example, a study that ran out of funds before some patients could return for follow-up visits.

- Missing at random (MAR) means that the data elements are not missing at random, but the probability that a value is missing depends on values of explanatory variables. As an example, consider a survey in which females are less likely to provide information about the physical condition (but the likelihood of responding is independent of her condition). If we know the sex of every subject and have answers for the condition for some of the females, unbiased sex-specific condition estimates can be made. That is because the values for the condition we do have for some of the women, are a random sample of all physical conditions.

- Informative missing (IM) means that data elements are more likely to be missing if their true values of the variable in question are systematically higher or lower. An example is when subjects with a bad physical condition or a very good one, are less likely to provide answers to the survey.

- Structural missing means that values are missing because a variable is not applicable. For example, when we use the number of years after

the menopause as explanatory variable, we can not apply this variable to premenopausal patients.

No matter whether we delete incomplete cases or carefully impute estimates for the missing data, it is beneficial to characterize patterns of missingness using exploratory data analysis techniques. For example in a multicenter study, we can investigating which centers are responsible for the majority of missings, or assessing whether there is a relation between the outcome and the missingness of certain explanatory variables.

Multicollinearity

If a linear combination of the explanatory variables (predictors) exists, which is almost zero, there is multicollinearity among the predictors. Due to the presence of multicollinearity, a predictor can become statistically non significant when the other variables are already in the model and the variance explained by the model will be inflated due to redundant information. A variable may seem to have little effect on the model simply because it overlaps (correlated) considerably with other predictor variables in the model, which contain much of the information of this variable.

One way to detect the presence of multicollinearity is to examine the correlation among predictors. If there is strong correlation among them, multicollinearity is assumed. A correlation matrix among the predictor variables or graphical exploration such as scatter plot matrices or biplots can be used to detect correlation between predictor variables (Gabriel, 1971).

Suppose we find a cluster of variables which are highly correlated. Then we may decide to choose only one variable in that cluster in order to avoid multicollinearity problems. For each predictor in that cluster, we can build a model containing only that variable and compare all those models based on model fit statistics, e.g. the Akaike Information Criterion (AIC Akaike, 1981; Bozdogan, 1987) and the Schwarz Criterion (SC Schwarz, 1978). Lower values of these statistics indicate a more desirable model. Or in other words, we should choose that variable in the cluster with the smallest AIC and SC.

Small datasets

Small datasets make the data analysis more difficult. We should use exact methods instead of the ordinary tests for data exploration, e.g. use Fisher s exact test instead of the chi-square test (which is a large-sample approximation) in order to check for relations between binary variables (Agresti, 1996; Fisher, 1970). When we start with developing a classical statistical model (e.g. logistic regression) we should have at least 5 observations for each explanatory variable. So, in a study with 50 patient records, only 10

variables can be modeled. An insufficient amount of data makes it very difficult to construct predictive models which can be helpful for a rational clinical disease management.

2.3 Statistical models

An important goal of statistical modeling in medicine is to develop good predictive models. In this thesis the aim is the construction of predictive models for two major medical decision problems:

- The preoperative prediction of ovarian cancer. We need to build a classification model that can discriminate between benign and malignant ovarian tumors.
- The construction of predictive models for survival of premature babies of which the mother had a preterm prelabor rupture of the membranes. We need models that can deal with survival data and can correct for the dependence in the outcome of babies of twins or triplets.

So, we need two types of models: classification models and survival models.

Logistic regression (Harrell, 2002; Hosmer and Lemeshow, 2001 - Appendix A) is a standard linear classification model of which the use has exploded during the past decade. From its original acceptance in epidemiological research, the method is now commonly employed in many fields including biomedical research, business and finance, criminology, ecology, engineering, health policy and linguistics because the parameters of this model can be easily interpreted. A drawback of this model is that any nonlinear relations between the explanatory variables and the outcome we want to introduce need to be specified explicitly into the model, e.g. adding a quadratic term in the model, or by including interaction terms between the variables.

Nonlinear black box classifiers (Bishop, 1995; Vapnik, 1995, Hastie, 2001; Schölkopf et al, 2002) have the advantage that the nonlinearity between the explanatory variables and the outcome does not need to be described explicitly. Many successful applications of artificial neural networks have been reported (Bishop, 1995). However, a problem is that the solution is characterized by many local minima. Support vector machines (SVMs) were introduced by Vapnik (1995). A typical property of SVMs is that the solution is obtained by solving a convex optimization problem, more specifically a quadratic programming problem. The solution to this convex problem is global and unique related to the fact that one employs a positive definite kernel. The support vectors are those data points that lie closest to the decision boundary and are therefore the most difficult to classify. Recently, a

modified version of SVMs has been introduced, the so-called least squares SVMs (LS-SVMs Suykens et al, 1999 - 2002 2003; Pelckmans et al, 2002; Van Gestel, 2004). In LS-SVMs the solution is given by solving a linear system instead of a quadratic programming problem which makes the implementation of this model much easier. The LS-SVM classifier model can be interpreted as a kernel version of Fisher discriminant analysis (Suykens et al, 2002). This classification model is briefly described in Appendix B.

In order to model records of babies where we should take into account the dependency between babies of twins and triplets generalized estimating equations can be used (GEE Diggle et al, 2001). We can get robust variance estimates for the parameters (related to the explanatory variables) by specifying the working correlation matrix (see Appendix C).

The analysis of survival data requires special techniques because the data are almost always incomplete, and parametric assumptions may be unjustifiable. Investigators follow subjects until they reach a prespecified endpoint, e.g. death. However, subjects sometimes withdraw from a study, or the study is completed before all subjects reached the endpoint. In these cases, the survival times (also known as failure times) are *censored*. Subjects survived up to a certain time instant beyond which their status is unknown. Methods for survival analysis must account for both censored and uncensored data.

There are many types of models that have been used for survival data. Two of the more popular types of models are the accelerated failure time model (Kalbfleisch and Prentice, 1980) and the Cox proportional hazards model (Cox, 1972). Each has its own assumptions on the underlying distribution of the survival times. Two closely related functions often used to describe the distribution of survival times are the survivor function and the hazard function (see Appendix D for definitions). The accelerated failure time model assumes a parametric form for the effects of the explanatory variables and for the underlying survivor function. Cox s proportional hazards model also assumes a parametric form for the effects of the explanatory variables, but it allows an unspecified form for the underlying survivor function. Therefore, Cox s model is most widely used in clinical data analysis. A short description of this model can be found in Appendix D.

2.4 Validation

Validation is the process of determining the validity of a model in new patients, and is an important aspect of model development. It makes no sense to develop a model that nicely describes the patterns in the data set

under study, but fails when tested prospectively. Validity may be internal or external (Justice et al, 1999; Altman and Royston, 2000).

2.4.1 Internal validation

Internal validation refers to the validation of the model in the population where the sample originated.

In principle, all aspects of model development should be validated. This includes:

- the coding and selection of predictors,
- the estimation of regression coefficients, and
- the evaluation of model performance.

When the whole process of model development is considered, the validity of the final model and the quality of its predictions can be honestly assessed.

A straightforward and fair approach is to randomly split the data set in two parts: one to develop the model (the so-called training set or development set), and another one to measure its performance (test set). With the split-sample approach, model performance is determined on similar, but independent, data. Common splits are 50:50 or 70:30. A characteristic of internal validation is that the parts of the data that are kept out of the model development phase (to test models) are kept out *at random.*

2.4.2 External validation

External validation (or generalization) refers to the validation of the model in other populations, e.g. patients from other centers, or treated in a more recent time (Justice et al., 1999).

Studying external validation can be done when a data set is available where non-random splits can be made, e.g. according to place of year of treatment. Also, the predictive model can first be developed and published, while external validity is studied after some years, either in the same center (temporal validation) or another center (spatial validation).

Note that external validation is more relevant than internal validation when a previously developed prediction model is applied in another clinical setting in a more recent time. Estimates of internal validation might be interpreted as what can maximally be expected of external validation. The aim of our thesis is to develop predictive models that are at least internally valid; external validity can only be a secondary aim.

Furthermore, we distinguish clinical and statistical validation in model development (Altman and Royston, 2000).

	Clinical validation	Statistical validation
refers to:	• whether the performance is good enough for the problem addressed.	• whether model assumptions are met ("goodness-of-fit") • whether the predictions are well-calibrated and discriminate low risk from high risk patients
determined by:	• intrinsic prognostic information, i.e. whether strong predictors are available	• continuous predictors adequately modeled • interaction terms not missed • outcome adequately modeled.

In the next section, we discuss in more detail the second point of statistical validation: the model performance, where we focus on the importance of highly sensitive and specific predictive models.

2.5 Evaluation of Performance

Model performance refers to the quality of the predictions from the model. One usually concentrates on two aspects of model performance:

- calibration, and
- discrimination.

2.5.1 Calibration

Calibration refers to whether the predicted probabilities agree with the observed probabilities. First, we may study "calibration in the large", i.e. whether predictions agree *on average* with observed probabilities. This is assessed by the Hosmer and Lemeshow goodness-of-fit test. Second, we may study calibration in more detail by looking at the correspondence between predictions and observations over the whole range of predictions.

The following graph illustrates several combinations of calibration in the large and calibration in more detail.

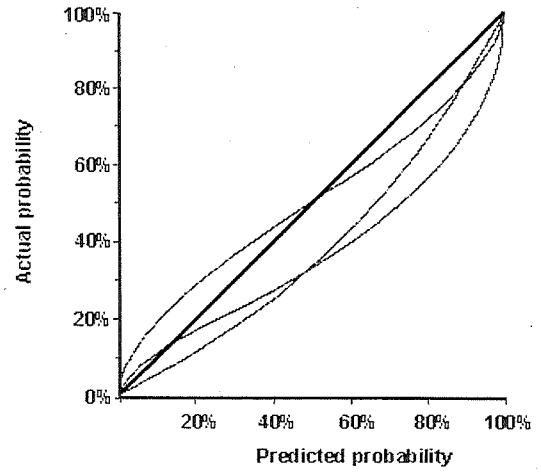


Figure 2-1 Calibration plot, i.e. plot of actual versus predicted probability.

The black line (see Figure 2-1) illustrates ideal calibration: the actual probabilities correspond exactly to the predicted probabilities. The red line illustrates the situation that predictions on average correspond to observations ("calibration-in-the-large" is OK), but that low predictions are systematically too low and high predictions are systematically too high. The other two lines represent situations where calibration in the large is not OK. An example of severe misclassification is indicated by the green line, where predictions are too high especially for higher predictions. For example, when the predicted probability was around 80%, the actual probability was around 50%.

The calibration plot may also provide useful information on the distribution of predicted values, e.g. by plotting symbols for groups of patients. The spread of predictions brings us to the other aspect of model performance: discrimination.

2.5.2 Discrimination

Discrimination refers to the ability to distinguish high risk subjects from low risk subjects, and is commonly measured by the area under the Receiver Operating Characteristic (ROC) curve (Hanley and McNeil, 1982).

We will first consider the terms sensitivity and specificity and then come to ROC curves.

Contingency Table

An intuitively appealing way to summarize the results of a fitted logistic regression model is via a contingency table. This table is the result of cross-classifying the outcome variable, y, with a dichotomous variable whose values are derived from the estimated logistic probabilities or from the LS-SVM output.

To obtain the derived dichotomous variable we have to define a cut-off point, c, and compare each estimated probability to c. If the estimated probability exceeds c then we set the derived variable equal to 1; otherwise it is equal to 0.

Classification is sensitive to the relative sizes of the two component groups and always favors classification into the larger group. This fact is independent of the fit of the model.

Table 2-1 Contingency table of predicted versus observed outcome

		Observed Outcome + Malignant	Observed Outcome - Benign	Total
Predicted Outcome	+ Malignant	a True positives	b False positives	$a+b$
	- Benign	c False negatives	d True negatives	$c+d$
	Total	$a+c$	$b+d$	N

a,b,c and d are the numbers of individuals in each cell of the table.

The two most common indices of the performance of a test are the sensitivity and the specificity. The ***sensitivity*** is the proportion of the true positives that are correctly identified by the test, given by $a / (a + c)$ in Table 2-1. The ***specificity*** is the proportion of true negatives that are correctly identified by the test, or $d / (b + d)$.

In order to know if there is a significant difference between two models in terms of sensitivity or specificity the test of Tango (1998) will be applied.

In clinical practice the test result is all that is known, so we want to know how good the test is at predicting e.g. a malignant tumor. In other words, what proportion of patients which are predicted to have a malignant tumor do really have a malignant tumor? The sensitivity and specificity do not give us this information. Instead we must approach the data from the direction of the test results. These two proportions are defined as follows:

- ***Positive predictive value*** (PPV) is the proportion of patients with positive test results which are correctly diagnosed.
- ***Negative predictive value*** (NPV) is the proportion of patients with negative test results which are correctly diagnosed.

In the notation of Table 2-1, PPV = $a / (a + b)$ and NPV = $d / (c + d)$.

The predictive values of a test depend upon the prevalence of the abnormality in the patients being tested, which may not be known. The values of PPV and NPV assume that the prevalence of e.g. a malignant tumor among the population of patients likely to be tested is the same as in the sample where we have calculated them. In a different clinical setting, the prevalence of a malignant tumor may differ considerably.

Predictive values observed in one study do not apply universally. The rarer the true abnormality is, the surer we can be that a negative test result indicates no abnormality, and the less sure that a positive result really indicates an abnormal patient.

Receiver Operating Characteristic Curves

Sensitivity and specificity rely on a single cutoff point to classify a test result as positive. A more complete description of classification accuracy is given by the area under the Receiver Operating Characteristic (ROC) curve. It is a graphical representation of the trade-off between the sensitivity and the specificity of a model (an increase in sensitivity will be accompanied by a decrease in specificity).

The ROC curve is a plot of the sensitivity against 1 minus the specificity for different probability cutoff points. Points near to the top right region of the curve correspond to lower probability cutoff points, whereas points near to the bottom left corner of the curve correspond to higher probability cutoff points.

The following graph (see Figure 2-2) shows three ROC curves representing models with excellent, good and worthless predictive accuracy.

The predictive accuracy of a model depends on how well the model classifies observations.

A model that perfectly predicts the response has an ROC curve that passes through the upper left corner, where both the sensitivity and the specificity equal the value 1.

If a model has no discriminatory power whatsoever, the ROC curve will be a straight line.

For a model with good predictive accuracy, a high level of sensitivity can be achieved at the expense of only a small decrease in specificity.

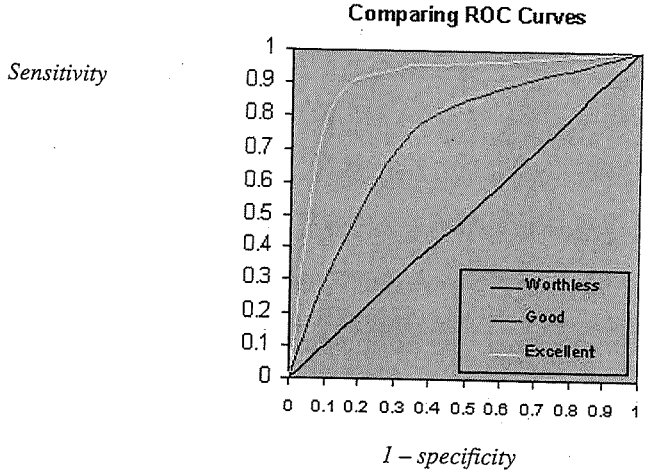


Figure 2-2 Receiver Operating Characteristic curves

Thus, the closer the ROC curve is to the upper left corner, the higher the predictive accuracy of the model. In other words, for a logistic regression model or LS-SVM model with high predictive accuracy, the ROC curve on the test set rises quickly, that is, both sensitivity and specificity are high for higher probability cutoff points. Also, the area under the curve is large for a model with high predictive accuracy. Conversely, the ROC curve rises slowly and has a smaller area under the curve on the test set for logistic regression models with low predictive accuracy.

The area under the ROC curve (AUC) is a measure of the performance of the model. It is a measure of the predictive accuracy of the model and is defined as follows :

$$AUC = (n_c + 0.5\ (t - n_c - n_d)) / t \quad ,$$

where t is the total number of pairs with different response values, n_c is the number of concordant pairs, and n_d is the number of discordant pairs. Thus, AUC is a ratio of the number of concordant pairs and half of the ties over the total number of pairs with different response values. AUC, beings the area under the ROC curve, can be computed using the trapezoidal rule.

If a model has no discriminatory power, the area under the ROC curve will be equal to 0.5, in which case the ROC curve is a straight line. When there is a perfect separation of the two groups, the area will equal 1. In this case, the ROC curve reaches the upper left corner of the plot. So, the area under the ROC curve ranges from 0.5 up to 1 for a perfect performance. The closer AUC is to 1 on the test set, the better the performance of the model. A rough guide for classifying the accuracy of a model on the test set is the following:

- 0.90 – 1 = excellent
- 0.80 – 0.90 = good
- 0.70 – 0.80 = fair
- 0.60 – 0.70 = poor
- 0.50 – 0.60 = fail

A comparison of the area under the ROC curves of two models on a same set of patients can be done using a parametric method (Hanley and McNeil, 1983) or a nonparametric approach based on the Mann-Whitney statistic (De Long, 1988).

2.6 Conclusions

Clinical data are most frequently not error free. When performing clinical data analysis, this is the first problem where we should pay attention to. Since the goal of the thesis is to develop predictive models for two major decision problems in gynecology, we should carefully build those models. We should consider the issues of missing values, multicollinearity and small datasets before starting with the construction of a model. The performance of the developed models will be assessed on an independent test set – by the use of ROC curves - in order to perform internal validation.

Chapter III

Description and exploration of the IOTA data

In this chapter we give an overview of preceding research in the area of pre-operatively predicting adnexal masses. We introduce the IOTA (International Ovarian Tumor Analysis) trial, which provides the largest data set on ovarian tumors so far. We describe in full detail the outcome and explanatory variables, and explore the data.

3.1 Introduction

Cancer is a term for diseases in which abnormal cells divide without control. A mass of extra tissue forms, which is called a tumor. Tumors can be benign or malignant. A malignant tumor consists of cancer cells, which have the ability to spread beyond the original site. They have a tendency to invade and destroy nearby tissue.

Adnexal masses are a very common problem in gynecology, and represent a number of different benign and malignant lesions or masses. Benign lesions can be either functional or non-functional. Malignant tumors are cancerous, including borderline, primary invasive and metastatic lesions.

Presently, malignant ovarian tumors are diagnosed at an advanced stage in 75 % of the cases and they result in the highest mortality figures of all gynecological cancers. It is often difficult to determine pre-operatively the nature (benign or malignant) of ovarian tumors. However, this knowledge is essential for obtaining a reliable diagnosis and for an appropriate management (both of which will influence the outcome for the patient and the medical costs). Such a classification of ovarian tumors can also improve our understanding of the biological differences between benign and malignant ovarian tumors.

The main question examined is: 'Which factors and which mathematical combination of these factors will result in an optimal classification of benign and malignant ovarian tumors?'

This question is clinically important, as reflected in the large number of publications in this field (Bourne et al, 1995; Granberg et al, 1989; Jacobs et al, 1990; Lerner et al, 1994; Osmer et al, 1996; Tekay et al, 1992; Tekay et al, 1996; Timmerman, Bourne et al, 1999; Timmerman, Verrelst et al, 1999). Since certainty about the nature of an adnexal mass can only be acquired after histopathological examination, the traditional strategy for establishing the final diagnosis has been to perform an exploratory laparotomy. However, since the management of malignant tumors is completely different from the management of benign tumors, it is of vital importance to obtain a reliable pre-operative diagnosis. In case of benign functional cysts a laparoscopy or any other surgical intervention should be avoided in order not to create unnecessary morbidity and impaired fertility. On the contrary, most benign non-functional tumors can be treated laparoscopically or with a low transverse abdominal incision.

Patients with a suspect adnexal mass require a series of expensive (and unpleasant) staging examinations and in most cases also an exploratory laparotomy through a median incision is indicated. This procedure should only be performed by a surgeon with proper skills and experience in debulking surgery, since the amount of residual malignant tissue after primary surgery is one of the most important prognostic factors in ovarian cancer. Furthermore, appropriate pre-, per-, and postoperative measures should be taken.

During recent years several techniques have been introduced to differentiate between benign and malignant ovarian lesions. These techniques include serum tumor markers (e.g. ***CA125***) , transvaginal ultrasonography with a variety of morphological scoring systems, or sometimes combined with color Doppler sonography, computerized tomography (CT scan), and magnetic resonance imaging (MRI) (Granberg et al, 1989; Jacobs et al, 1990; Lerner et al, 1994; Osmers et al; 1996, Sladkevicius et al, 1995; Tekay et al, 1995; Tekay et al, 1996, Valentin et al, 1994).

Serum CA 125 levels

The CA 125 antigen is a glycoprotein with a high molecular weight that is expressed by most epithelial ovarian cancers and is recognized by a monoclonal antibody (OC 125). Serum CA 125 is the tumor marker with the highest sensitivity for ovarian cancer (Bast et al, 1983; Jacobs & Bast, 1989; Knapp et al, 1996; Schwartz et al, 1993). This tumor marker will detect nearly 80 percent of advanced (stage ≥ III) ovarian cancers, but only 40 to 44

percent of patients with stage I disease (Vergote, Børmer et al, 1987; Cuckle et al, 1991; Bourne, Campbell et al, 1994; Maggino et al, 1994). Some authors use different cut-off levels for pre- and postmenopausal patients. In premenopausal patients the specificity is low at a cut-off level of 35 U/ml, since false positive results are frequently encountered in menstruating or pregnant women and in a wide variety of benign conditions, such as benign ovarian tumors (10%), endometriosis (20-30%), liver cirrhosis (60-70%), pancreatitis (30%), pelvic inflammatory disease, uterine fibroids and Meigs' syndrome (100%) (Bast et al, 1983; Jacobs & Bast, 1989; Schwartz, 1993; Timmerman et al, 1995). Cancer of the endometrium (30%), lung (30%), stomach (30%), pancreas (50%) and liver (50%) also frequently result in elevated (>35 U/ml) serum CA 125 levels (Schwartz, 1993). However, using a cut-off level of 65 U/ml results in a lower sensitivity for ovarian cancer.

Extensive abdominal surgery also interferes with the specificity of CA 125 as a tumor marker, since postoperatively the serum levels will rise in more than 80% of the patients, to reach a peak during the second week after the operation, and to normalize 8 weeks after surgery (Van der Zee, 1990). Finally, intermittent increases of CA 125 levels due to fluid retention by docetaxel treatment have been described in ovarian cancer patients without tumor progression (Schrijvers et al, 1996). Thus, it is recommended that CA 125 should not be measured for at least 2 months after abdominal surgery and at least 3 weeks after primary chemotherapy.

Risk of malignancy index (RMI)

In order to increase the reliability of serum CA 125 levels to differentiate pre-operatively between benign and malignant pelvic masses, Jacobs and colleagues combined the measurement of CA 125 values with morphological findings and menopausal status of the patient in calculating a risk of malignancy index (Jacobs et al, 1990). In a retrospective study using transabdominal ultrasonography the following features suggestive of malignancy have been assessed: multiloculated cysts, evidence of solid areas, evidence of metastases, presence of ascites, and bilateral lesions. An ultrasound score (U) of 0 was given when none of these features was recorded, a score of 1 was given when one feature was present and a score of 3 was given when two or more of the above features were noted. A score (M) of 1 or 3 was given to pre- and postmenopausal patients, respectively.

The following equation was proposed:

$$RMI = U \times M \times \text{serum CA 125}$$

where U and M are the ultrasound and menopausal scores, as defined above.

The main advantage of this simple and rather reliable test is that it can be applied easily in less specialized departments, since it does neither require a set of different tumor markers nor experience with color Doppler sonography.

Morphological scoring systems

More than 10 different morphological scoring systems have been reported and used with varying success (Granberg et al, 1989; Granberg et al, 1990; Sassone et al, 1991; Timor-Tritsch, 1993; Lerner et al, 1994; Bourne, 1995; DePriest, 1997; Valentin, 1997; Ferrazzi, 1997; Valentin, 2000).

Despite numerous publications on each of these techniques (CA125, RMI, morphological scoring systems), an insufficient amount of data was available to provide for a rational clinical management.

Since the beginning of 1997, there is a close cooperation between D. Timmerman and I. Vergote of the University Hospitals of Leuven and ESAT-SISTA. In a pilot-project concerning ovarian tumors, based on the data of 300 patients (Timmerman et al, 1999) stepwise multivariate logistic regression analysis was performed in order to select relevant parameters from the preliminary available data. In a next step, various logistic regression models were developed and tested. Logistic regression analysis, based on simple parameters (such as menopausal status, serum level of the tumor marker CA 125 II, a score based on the intensity of blood supply in the tumor and ultrasonographically detected presence of papillary structures), could be used to preoperatively discriminate benign and malignant ovarian masses in a reliable way. Also ANN's and LS-SVMs (Lu et al, 2003), based on simple parameters (such as age of the patient, CA 125 serum level and some morphologic features), could be trained to reliably predict malignancy. In a prospective study, the nonlinear models performed significantly better than the widely used 'Risk of Malignancy Index' (Jacobs et al, 1990). A first statistical study of the difference in performance indicated that the developed ANN/LS-SVM models were still very close to the linear/logistic regression model, in other words that the possible model complexity of the ANN's and LS-SVMs was not fully exploited (Timmerman, Verrelst, Bourne et al, 1999; Lu et al, 2003). It was possible that the classification problem at hand needed

an explicitly non-linear decision surface. This meant that a large amount of examples, which were hard to be classified (i.e. close to the decision boundary), were needed to train nonlinear models and to significantly enhance their global performance. The need for a much larger database was also corroborated by the histologically large amount of different types of ovarian tumors, each with its own biological behavior, and by the heterogeneity of the clinical history and age of the patients in the population.

Therefore, a multi-centric cooperation with renowned universities in other countries was undertaken. Only such a multi-center approach, International Ovarian Tumor Analysis (IOTA), could provide sufficient data to reach the required critical mass for this project within a reasonable period of time. The database enabled us to make a selection of relevant parameters to classify pre-operatively the ovarian tumors (e.g. to discriminate between functional, non-functional benign, primary invasive and metastatic tumors) and to develop and test refined models.

The purpose of the IOTA study is to develop mathematical models for the pre-operative classification of malignant and benign ovarian tumors, to assess their performance in comparison with subjective assessments, and to find the cut-off for several continuous variables. The information obtained will be used to build up the optimal screening strategies in order to identify cases in early stages and to select the most appropriate treatment for subsequent patients.

We start with describing the IOTA dataset in full detail.

The IOTA (International Ovarian Tumor Analysis) study is a prospective international multicenter study including nine European ultrasound centers from five countries: University Hospital Malmö (Sweden), University Hospitals Leuven (Belgium), Universita del Sacro Cuore Rome (Italy), ISBM L. Sacco University of Milan (Italy), Hôpital Boucicaut Paris (France), Centre Medical des Pyramides, Maurepas (France), King's College Hospital London (UK), ISBM Ospedale, San Gerado Universita di Milano, Monza (Italy), Universita degli Studi di Napoli and Naples (Italy) (see Figure 3-1). Three centers Malmö, Leuven and Rome contributed 2/3 of the records in the dataset, i.e. 29.5%, 24.7% and 11.8%. Recruitment was from June 1999 to June 2002. Patients representing with at least one ovarian tumor, who underwent ultrasound examination by a principal investigator at one of the participating centers were eligible for inclusion. A history was taken from each patient following a standardized protocol (Timmerman, Valentin et al, 2000).

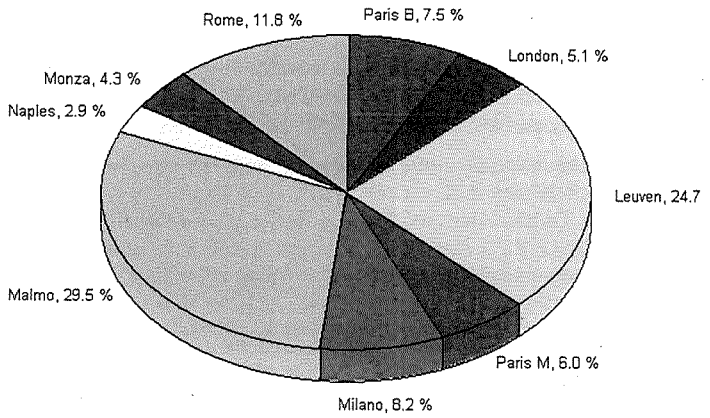


Figure 3-1 Pie chart of the center proportion in the IOTA dataset. Malmö (Sweden) and Leuven (Belgium) have the highest number of patients. 2/3 of the data consists of patients from three centers: Malmö (29.5%), Leuven (24.7%) and Rome (11.8%).

A dedicated, secure data collection system was developed for the study (Aerts et al, 2002). A unique identifier was generated automatically for each patient's record, based on the identifier for the center, the patient's birthday, the date of the ultrasound scan, and the follow-up number. Clinicians at each center could only view or update patient records from their own center. Data security was ensured by not recording the patient's name, and by encrypting all data communication between the browser-based data entry and the central IOTA database using a 56-bit SSL certificate. Data integrity was ensured by: client-side JavaScript checks, server-side XML-encoded rules based on the IOTA protocol, and manual checks by three experts.

3.2 Patient entry

3.2.1 Inclusion criteria

- All patients were assessed with transvaginal US by the principal investigators (or their appropriately trained medical staff) and found to have an apparently persistent extra-uterine pelvic mass, provided that

- The patient was fit for surgery
- The patient gave informed consent
- The patient had at least one remaining ovary

- In case of more than one pelvic mass, data from the apparent worse case mass were used.
- In patients with more than one examination, the record closer to the operation date was used.

3.2.2 Exclusion criteria

- Pregnant patient.
- Inability to tolerate transvaginal sonography.
- Surgery > 120 days after sonographic assessment.
- Disagreement in the classification (malignant or benign) between the original pathology report and the report of an expert reviewer.
- Incomplete submission of data.

A total of 1149 patients were recruited. Data from 83 (7%) were excluded (8 due to pregnancy, 31 because surgery was undertaken more than 120 days from the sonographic assessment, 42 due to incomplete submission of data, 2 because of disagreement over the histological diagnosis). Data from 1066 patients were available for statistical analysis and model development.

3.3 Variables

3.3.1 Outcome

The final outcome of the study were histological diagnosis, ***Outcome***, and in the cases of malignancy the surgical stage, ***Stage***.

Surgery was performed in the case of a mass being classified as persistent (still present 6 to 12 weeks after the initial scan). In cases of symptomatic adnexal masses, suspected malignancy or at the patient's request, surgery was performed more quickly – either by laparoscopy or laparotomy depending on the surgeon's clinical judgment. All surgically removed tissues

were extensively sampled for histological examination at the local center. Tumors were classified according to the criteria recommended by the International Federation of Gynecology and Obstetrics (FIGO) (Heintz et al, 2003). For malignant tumors the degree of differentiation was included. About 10% of the pathological samples were randomly selected for peer review. Two unstained slides of each prepared block of the removed adnexal masses were reviewed by a single expert gynecological pathologist (Professor Ph Moerman, Katholieke Universiteit Leuven). In cases of disagreement (malignant or benign) between the original pathology report and the reviewer the patients were excluded from the study.

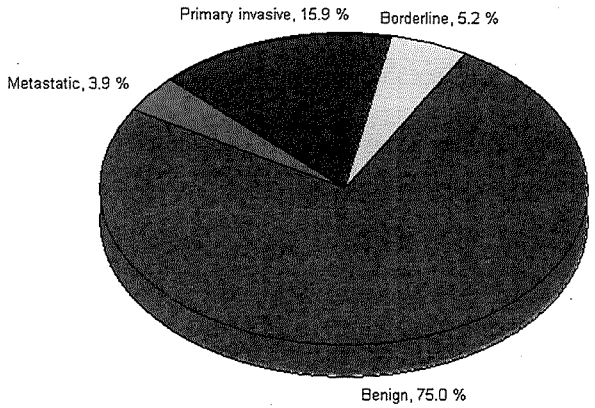


Figure 3-2 Proportion of the four outcome levels. Most ovarian tumors are benign: ¾. Of the 25% malignant masses, 5% are borderline (which are the most 'benign' malignant tumors), and 20% are 'real' malignant: 16% primary invasive and 4% metastatic.

Outcome has four categories: the benign tumor with code 0, the primary malignant tumor with code 1, the borderline tumor with code 2, and the metastatic tumor with code 3 (see Figure 3-2). 25% of all tumors is malignant of which 5% borderline, 16% primary invasive and 4% metastatic. Since our major purpose is to differentiate benign tumors from malignant ones, a binary variable ***Outcome1*** was created by grouping categories 1, 2, and 3 into one category with code 1, representing malignant tumors, while benign tumors still belong to the category with code 0.

The different levels of ***Stage*** are:

- Stage I: ovarian cancer is confined to one or both ovaries.
- Stage II: ovarian cancer has spread to pelvic organs, e.g. the uterus, but has not spread to abdominal organs.
- Stage III: ovarian cancer has spread to abdominal organs, e.g. liver and bowel.
- Stage IV: ovarian cancer has spread outside to distant sites, e.g. lung.

3.3.2 Explanatory variables

Family History

- ***FamHistOvCa***: Number of first degree relatives with ovarian cancer (0-...)
- ***FamHistBrCa***: Number of first degree relatives with breast cancer (0-...)

Personal History

- ***PersHistOvCa:*** presence of personal history of ovarian cancer (0-No; 1-Yes)
- ***PersHistBrCa:*** presence of personal history of breast cancer (0-No; 1-Yes)
- ***Age:*** age of the patient (in years)
- ***Parity:*** the number of children that patient has
- ***Hysterectomy:*** 0-No; 1-Yes
- ***Menopausal:*** 0-Premenopausal; 1-Postmenopausal
- ***PostMenoyears:*** years after menopause
- ***HormTherapy:*** presence of hormonal therapy (0-No; 1-Yes)
- ***PMB:*** presence of postmenopausal bleeding within one month prior to the ultrasonography (0-No; 1-Yes)
- ***Pain:*** presence of pelvic pain during the scan (0-No; 1-Yes)

Ultrasound variables

Ultrasonography is a study in which sound waves (called ultrasound) are bounced off tissues and the echoes are converted into a picture (sonogram). The ultrasound scanner has a microphone, which generates sound waves. The microphone is passed over the patient s body. The sound waves bounce off the organs inside the human body, and are picked up again by the microphone as they bounce back. This turns the reflected sound waves into a picture by the computer, to which the microphone is linked. From the images, it is possible to determine both the size and the nature of the structures. The interaction of ultrasound with tissue structures gives rise to information that is directly related to the acoustic (i.e. ultrasonic) properties of the tissues.

Transvaginal ultrasonography is used to look at the ovaries or part of the womb and is at present the most effective method for early diagnosis of ovarian tumors. A small ultrasound microphone is put into the vagina so that the ovaries can be seen on the scan.

- ***Bilateral:*** involvement of both ovaries (0-No; 1-Yes)
- ***LesD1***, ***LesD2*** and ***LesD3:*** measurements of the lesion in three perpendicular planes (in mm)
- ***Ascites:*** presence of fluid outside the pouch of Douglas (0-No; 1-Yes)
- ***Fluid:*** fluid in the pouch of Douglas is measured in a sagittal plane (the largest anteroposterior diameter is given, see Figure 3-3) (in mm).

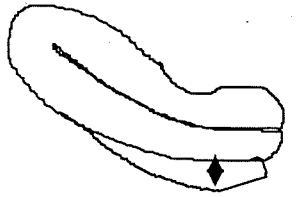


Figure 3-3: Fluid is measured in a sagittal plane.

- ***Septum:*** A septum is defined as a thin strand of tissue running across the cyst cavity from one internal surface to the contra-lateral side. The thickness of the thickest septum is measured (in mm)

where it appears to be widest (other than at its interface at the internal surface of the cyst wall).

- ***Incomplseptum:*** an incomplete septum (as seen in hydrosalpinges) is defined as a thin strand of tissue running across the cyst cavity from one internal surface to the contra-lateral side, but which is not complete in some scanning planes. Its presence is noted (0-No; 1-Yes).
- ***Papillation:*** Solid papillary projections are defined as any solid projection, which shows high echogenicity, into the cyst cavity from the cyst wall greater than or equal to 3 mm in height as shown in Figure 3-4. Its presence is noted (0=No; 1-Yes).

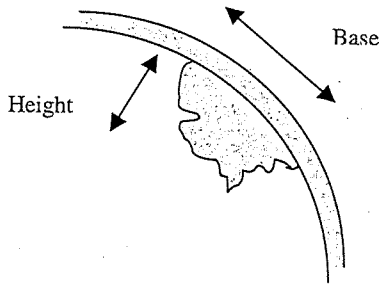


Figure 3-4: Solid papillary projection with ***PapHeight*** and ***PapBase*** indicated.

- ***PapNr: number of papillations (1,2,3,more)***
- ***PapHeight, PapBase1 and PapBase2:*** Solid papillary projections are defined as any solid projections into the cyst cavity from the cyst wall greater than or equal to 3 mm in height. The largest is measured in 3 planes: height, base1, and base2.
- ***PapFlow:*** presence of flow within some of these projections (0=No; 1-Yes).
- ***PapSmooth:*** whether the largest papillation is smooth (0), irregular (1) or uncertain (2).
- ***Locularity:*** lesions are qualified into 5 categories based on their morphology, including
    1. unilocular (a unilocular cyst without septa and without solid parts or papillary structures).

2. unilocular cyst with solid component (a unilocular cyst with a measurable solid component or at least one papillary structure).
3. multilocular (a cyst with at least one septum but no measurable solid components or papillary projections). The lesion is measured as indicated by the arrows.
4. multilocular with solid component (a multilocular cyst with a measurable solid component or at least one papillary structure)
5. solid (a tumor where the solid components comprise 80% or more of the tumor when assessed in a two-dimensional section).
6. not classifiable

- ***SolidD1, SolidD2, SolidD3:*** In cystic-solid tumors the largest solid component is measured separately (in three perpendicular planes). The solid component is noted as being smooth or irregular (e.g. cauliflower-like). In some cases a solid papillary projection is the largest solid component and thus the papillary projection is recorded both as papillary projection and as solid component.
- ***NrLocules:*** the number of locules (1, 2, 3, 4, 5-10, >10)
- ***WallRegularity:*** whether the internal wall of the lesion is smooth (0), irregular (1) or not visible (2). If there is a solid papillary projection, then the wall is irregular by definition.
- ***Shadows:*** presence of acoustic shadows (0-No, 1-Yes)
- ***Echogenicity:*** The dominant feature of the cystic contents is described as
    1. anechoic
    2. low-level echogenic
    3. ground glass appearance
    4. hemorrhagic
    5. mixed echogenic
    6. no cyst fluid.

In addition to the original variables, we calculate some supplementary variables which can be interesting to investigate their relevance in the classification task:

- *MaxLesD123* = max (*LesD1*, *LesD2*, *LesD3*) (largest diameter of lesion, in mm)
- *VolumeLes* = *LesD1*LesD2*LesD3**(π/6)/1000 (volume of lesion, in ml)
- *MaxPap123* = max(*PapHeight*PapBase1*PapBase2*) (largest diameter of papillary structure, in mm)
- *VolumePap* = *PapHeight*PapBase1*PapBase2**(π/6)/1000 (volume of papillary structure, in ml)
- *MaxSolidD123* = max(*SolidD1*SolidD2*SolidD3*) (largest diameter of solid component, in mm)
- *VolumeSolid* = *SolidD1*SolidD2*SolidD3**(π/6)/1000 (volume of solid component)
- *RatioPapLes* = *VolumePap / VolumeLes*
- *RatioSolidLes* = *VolumeSolid / VolumeLes*

Color Doppler imaging and blood flow indices

With color Doppler imaging it is possible to identify the presence of blood flow and vascular changes in new tissue growth. Such changes are imaged as fluctuating areas of color. Calculating indices derived from the Doppler flow velocity waveform gives the possibility to estimate blood flow impedance quantitatively.

This is illustrated in the following Figure 3-5 : *S* is the peak Doppler shifted frequency or peak systolic velocity (***PSV***) and *D* is the minimum Doppler shifted frequency of the end diastolic velocity. The systole is the contraction of the myocardium (or heart muscle). The diastole is the relaxation of the heart after a contraction.

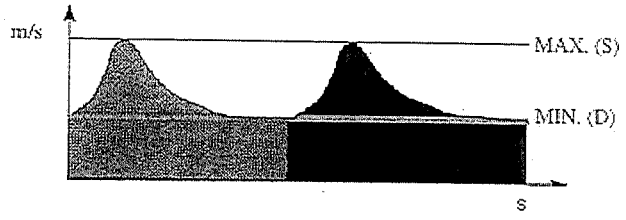


Figure 3-5 Color Doppler imaging to identify the presence of a bloodflow

At every contraction of the myocardium, the velocity of the blood flow reaches a maximum. Then it slows down and it reaches a minimum at the diastole.

Another index that can be derived from smooth curves fitted to Doppler waveforms is the mean Doppler shifted frequency *A* or time averaged maximum velocity (***TAMXV***).

From these three indices, the resistance index (***RI***) and pulsatility index (***PI***) can be calculated : ***RI*** = (*S* *D*) / *S* and ***PI*** = (*S* *D*) / *A* .

- ***ColScore:*** a subjective semi-quantitative assessment of the amount of blood flow within septa, cyst walls or solid tumor areas in Color Doppler image.
  1. no blood flow
  2. only minimal flow
  3. moderate flow
  4. highly vascular with marked blood flow
- ***Venous:*** presence of venous blood flow only (0-No; 1-Yes)
- ***PI:*** pulsatility index
- ***RI:*** resistance index
- ***PSV:*** peak systolic velocity
- ***TAMXV:*** time-averaged maximum velocity

Serum tumor marker

CA 125 antigen: a glycoprotein with a high molecular weight that is expressed by most epithelial ovarian cancers and is recognized by a monoclonal antibody (OC 125).

3.3.3 Subjective assessment by the expert

After ultrasonographic examination of the mass the investigator gives his subjective assessment of the mass:

- ***SubjAss***: 0-Benign; 1-Malignant

- ***SubjProb***: probability of malignancy (=level of certainty).
  1. benign
  2. probably benign
  3. uncertain
  4. probably malignant
  5. malignant
- ***Self_impression***: presumed histological diagnosis (e.g. dermoid, serous cystadenoma, endometrioma...)

3.4 Data exploration

It is apparent that there is a marked difference in the prevalence of malignancy between the nine participating centers (see Figure 3-6). This finding can be explained by the nature of the centers. Some are regional referral centers for gynecological malignancy, some offer tertiary scanning services, whilst others are more general units. The prevalence of malignancy is likely to be significantly higher in referral centers than in more general units and the demographic data also reflect these differences.

Figure 3-6 shows us that there are three groups of centers:

- Monza, Rome and Leuven have the highest percentage of malignant tumors, i.e. 37.0%, 35.7% and 35.4%.
- The proportion of malignant masses is rather low in Milano, Naples, Paris M (Maurepas) and Paris B (Boucicaut): 9.2%, 9.7%, 10.9% and 11.2%.
- The two remaining centers are Malmö with 21.6% and London 29.6%.

Not all centers have borderline or metastatic tumors. For example the two centers in France: Paris M and Paris B do not have metastatic cases; Naples does not have borderline or metastatic tumors in the data set.

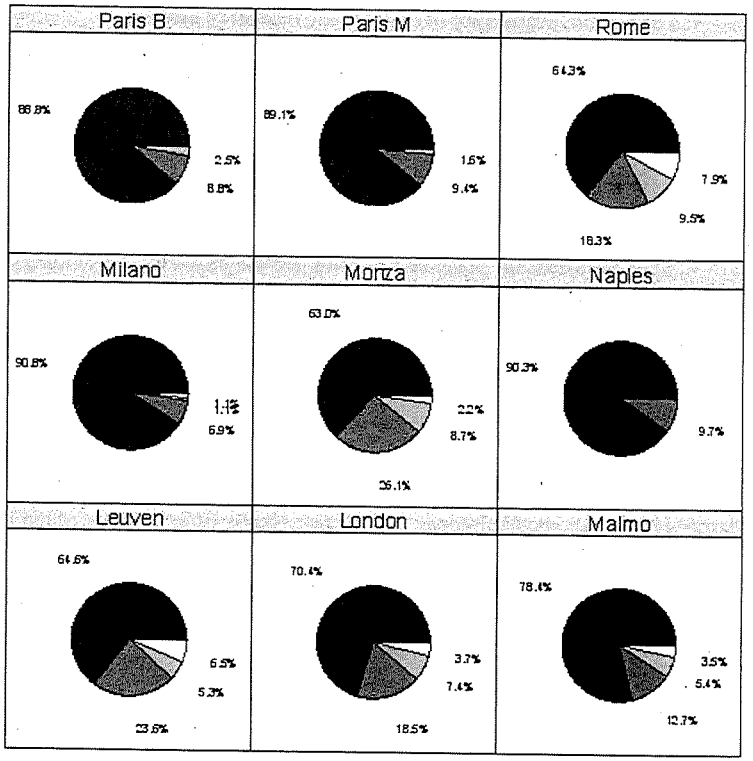


Figure 3-6 Proportion of benign, borderline, primary invasive and metastatic tumors in each center. Monza, Rome and Leuven have high percentage of malignant tumors, i.e. 37%, 36% and 35%, whereas this is rather low in Milano, Naples, Paris M (Maurepas) and Paris B (Boucicaut): it varies from 9.2% to 11.2%. There is also remarkable difference between metastatic tumors, e.g. Leuven has 6.5% metastatic masses, while Naples, Paris B and Paris M have none.
black: benign - light gray: borderline - dark gray: primary invasive - white: metastatic

There is not only a marked difference in the outcome - prevalence of malignancy - between the nine participating centers, but also in the explanatory variables. For example

- Table 3-1 shows a striking difference in the presence of hormonal therapy between Italy and the other countries. For example Rome and Milano have hormonal therapy in only 5.6% and 3.4% of the patients, while in Malmö and Leuven this is 29.8 and 31.2%.

The differences in ***ColScore*** which is a subjective assessment of the amount of blood flow in the Color Doppler image are shown in Table 3-2. E.g. Malmö gives only 6% of the cases the label "no blood flow", while in Leuven this proportion is 29%, and in Milano even 71.3%.

Table 3-1 Differences in the presence of hormonal therapy (***HormTherapy***) between the nine participating centers. The four Italian centers use the least hormonal therapy.

	HormTherapy		
Center	no n (%)	yes n (%)	Total n
Malmö	221 (70.2)	94 (29.8)	315
Leuven	181 (68.8)	82 (31.2)	263
Rome	119 (94.4)	7 (5.6)	126
Milano	84 (96.6)	3 (3.4)	87
Paris B	55 (68.8)	25 (31.2)	80
Paris M	56 (87.5)	8 (12.5)	64
London	42 (77.8)	12 (22.2)	54
Monza	45 (97.8)	1 (2.2)	46
Naples	28 (90.3)	3 (9.7)	31
Total	831	235	1066

Table 3-2 Differences in the subjective assessment of the blood flow (***ColScore***). There is a marked difference between the nine centres. For example Malmö gives to 77% of his patients the label "moderate" or "marked" blood flow, while Milano does this for only 8%.

	ColScore				
	no blood flow	minimal blood flow	moderate blood flow	marked blood flow	Total
Center	n (%)	n (%)	n (%)	n (%)	n
Malmö	18 (5.7)	55 (17.5)	176 (55.9)	66 (20.9)	315
Leuven	76 (28.9)	73 (27.8)	70 (26.6)	44 (16.7)	263
Rome	47 (37.3)	52 (41.3)	26 (20.6)	1 (0.8)	126
Milano	62 (71.3)	18 (20.7)	4 (4.6)	3 (3.4)	87
Paris B	7 (8.7)	52 (65.0)	15 (18.8)	6 (7.5)	80
Paris M	-	38 (59.4)	16 (25.0)	10 (15.6)	64
London	15 (27.8)	18 (33.3)	10 (18.5)	11 (20.4)	54
Monza	6 (13.0)	26 (56.5)	7 (15.2)	7 (15.2)	46
Naples	2 (6.5)	21 (67.7)	4 (12.9)	4 (12.9)	31
Total	233	353	328	152	1066

Note that there is a gradation in the objectivity of the explanatory variables E.g. the ***Age*** of the patient is more objective than whether the patient feels ***Pain*** during the ultrasound examination. The ultrasound variables are more objective than those related to Color Doppler imaging and blood flow indices (see Table 3-3).

Table 3-3 Gradation of objectivity (very objective, quite objective, rather subjective) specified by Prof .Timmerman (Leuven).

Variables	Gradation in the subjectivity
Family History	
FamHistOvCa	quite objective
FamHistBrCa	quite objective
Personal History	
PersHistOvCa	quite objective
PersHistBrCa	quite objective
Age	very objective
Parity	very objective
Hysterectomy	very objective
Menopausal	very objective
PostMenoyears	quite objective
HormTherapy	very objective
PMB	very objective
Pain	rather subjective
Ultrasound	
Bilateral	very objective
LesD1, LesD2, LesD3	quite objective
Ascites	quite objective
Fluid	quite objective
Septum	quite objective
IncomplSeptum	quite objective
Papillation	quite objective
PapNr	quite objective
PapHeight, PapBase1, PapBase2	quite objective
PapFlow	quite objective
PapSmooth	rather subjective
Locularity	quite objective
SolidD1, SolidD2, SolidD3	quite objective
NrLocules	quite objective
WallRegularity	quite objective
Shadows	quite objective
Echogenicity	quite objective
Color Doppler imaging and blood flow indices	
ColScore	rather subjective
Venous	rather subjective
PI	rather subjective
RI	rather subjective
PSV	rather subjective
TAMXV	rather subjective
Serum tumor marker	
CA125	very objective

Univariate analysis of the continuous variables (see Table 3-4), binary variables (see Table 3-5) and categorical variables (see Table 3-6) shows that all explanatory variables are significantly different between malignant and benign tumors, with the exception of years since the menopause (***PostMenoyears***), family history of ovarian and breast cancer (***FamHistOvCa***, ***FamHistBrCa***) and the presence of postmenopausal bleeding within one month prior to the ultrasonography (***PMB***). The median time interval between ultrasonography and surgery was 8 days in malignant tumors (maximum 90 days) and 12 days in benign tumors (maximum 120 days).

A predictable problem is that variables do not have the same number of observations:

- ***PostMenoyears*** (see Table 3-4) and ***PMB*** (see Table 3-5) have only values for women in the menopause.
- ***PapHeight***, ***PapBase1***, ***PapBase2***, ***MaxPap123***, ***VolumePap***, ***RatioPapLes***, ***PapNr*** (see Table 3-4), ***PapFlow*** and ***PapSmooth*** (see Table 3-5) have only values for the tumors with papillary projections (***Papillation***=1).
- ***SolidD1***, ***SolidD2***, ***SolidD3***, ***MaxSolidD123***, ***VolumeSolid***, ***RatioSolidLes*** (see Table 3-4) have only values for the masses with a solid component (***Locularity*** = unilocular solid , multilocular solid or solid).
- The blood flow indices ***PI***, ***RI***, ***PSV*** and ***TAMXV*** (see Table 3-4) have only values when there is blood flow in the tumor (***Colscore*** = 2, 3 or 4 see Table 3-6).

  Note 1: Not for all cases with ***ColScore*** = 2 is a measurement of ***PI***, ***RI***, ***PSV*** and ***TAMXV*** available. In the original dataset, before corrections were made, certain centers had almost no cases with ***ColScore*** = 2. After discussion with the experts about the definition of the variable ***ColScore***, some cases with ***ColScore*** 1 have been changed into ***Colscore*** 2 . For those records, no ***PI***, ***RI***, ***PSV*** or ***TAMXV*** measurements were calculated.

  Note 2: Milano performed a TAMXV measurement in only one patient, resulting in 13 supplementary missings.
- ***CA125*** is only measured in 70.9% of the benign tumors and 91.0% of the malignant cases. So, when the expert supects a benign tumor, he/she is less tended to measure ***CA125***.

Only two patients with a single unilocular tumor had a malignant tumor (see Table 3-6): one simple cyst (140x115x105 mm) with color score 2 in a 26-year-old patient with serum CA 125 of 13 U/mL proved to be a borderline malignant serous papillary cystadenoma, stage Ia; and one simple cyst

(55x38x35 mm) with color score 1 in a 42-year-old patient with serum CA 125 of 7 U/mL proved to be a borderline malignant mucinous cystadenoma, stage Ia.

Table 3-4 Univariate analysis of the continuous variables.
Fluid > 0 denotes the variable ***Fluid*** where only the values different from 0 are retained. There are 660 benign cases with no ***Fluid*** and 122 malignant cases.
Septum > 0 denotes the variable ***Septum*** where only the values different from 0 are retained. There are 457 benign cases without ***Septum*** and 123 malignant cases.

	Benign				Malignant			
	n	median	min	max	n	median	min	max
Demographic								
*Age**	800	42	17	90	266	56	17	94
PostMenoyears†	229	10	1	40	145	12	0	44
*Parity**	800	1	0	10	266	2	0	7
Morphologic								
*LesD1**	800	62	10	320	266	97.5	5	410
*LesD2**	800	48	7	242	266	75.5	8	290
*LesD3**	800	48	4	275	266	73.5	6	267
*MaxLesD123**	800	63	11	320	266	100.5	8	410
*VolumeLes**	800	73	0.2	7781	266	303	0.1	11829
Fluid > 0*	140	12	2	61	144	24	3	100
Septum > 0*	343	2.1	1	20	143	4.0	1	20
*PapHeight**	156	7	2	62	121	14	3	62
*PapBase1**	156	7.5	2	90	121	17	2	106
*PapBase2**	156	7	2	90	121	15	2	110
*MaxPap123**	156	10	3	90	121	21	4	110
*VolumePap**	156	0.2	0.006	77	121	2	0.008	226
RatioPapLes††	156	0.0003	0	0.456	121	0.006	0	0.420
*PapNr**	156	1	1	>3	121	>3	1	>3
*NrLocules**	800	1	0	>10	266	3	0	>10
*SolidD1**	309	18	2	230	244	49	4	214
*SolidD2**	309	13	2	173	244	39	2	174
*SolidD3**	309	14	2	130	244	37	2	199
*MaxSolidD123**	309	21	3	230	244	50.	4	214
*VolumeSolid**	309	1.6	0.006	1978	244	34	0.008	2291
*RatioSolidLes**	309	0.03	0	1	244	0.24	0	1
Bloodflow								
*PI**	506	0.95	0.13	5.8	246	0.74	0.25	2.26
*RI**	506	0.59	0.12	1.0	246	0.50	0.17	1.0
*PSV**	506	11.4	2.0	85.5	246	24.3	3.9	202
*TAMXV**	498	3.9	1.0	60.0	241	17.0	3.0	137
Tumor marker								
*CA125**	567	17	1.0	1409	242	167	4	31610

* p<0.0001; p=0.14; p=0.0055 (Mann-Whitney)

Table 3-5 Univariate analysis of the binary variables

	Benign		Malignant		p-value
	n	yes (%)	n	yes (%)	
FamHistOvCa	800	2.5	266	4.9	0.0559
FamHistBrCa	800	10.8	266	12.4	0.4578
PersHistOvCa	800	0.8	266	3.0	0.0096
PersHistBrCa	800	2.9	266	5.6	0.0386
Nullipara	800	41.9	266	25.2	<0.0001
Hysterectomy	800	6.3	266	10.5	0.0216
Menopausal	800	32.9	266	63.5	<0.0001
HormTherapy	800	23.5	266	17.7	0.0477
PMB	229	14.9	145	17.9	0.4291
Bilateral	800	16.6	266	31.2	<0.0001
Pain	800	28.8	266	19.6	0.0034
Ascites	800	2.9	266	42.1	<0.0001
Incomplseptum	800	9.3	266	4.1	0.0094
Papillation	800	19.5	266	45.5	<0.0001
PapFlow	156	33.3	121	84.3	<0.0001
PapSmooth (irregular)	156	49.4	121	82.6	<0.0001
WallRegularity (irregular)	800	32.8	266	81.6	<0.0001
Shadows	800	13.0	266	1.5	<0.0001
Venous	800	9.0	266	3.4	0.0040

Table 3-6 Univariate analysis of the categorical variables

Variable	Benign n (%)	Malignant n (%)
*Locularity**		
Unilocular	311 (38.9)	2 (0.8)
Unilocular - solid	88 (11.0)	44 (16.5)
Multilocular	176 (22.0)	20 (7.5)
Multilocular - solid	168 (21.0)	116 (43.6)
Solid	52 (6.5)	84 (31.6)
not classifiable	5 (0.6)	0 (-)
*Echogenicity**		
Anechogenic	303 (37.9)	107 (40.2)
Low level	149 (18.6)	60 (22.6)
Ground glass appearance	192 (24.0)	33 (12.4)
Haemorrhagic	8 (1.0)	2 (0.8)
Mixed echogenicity	114 (14.3)	18 (6.8)
No cyst fluid	34 (4.3)	46 (17.3)
*Color Score (blood flow)**		
No flow (1)	222 (27.8)	11 (4.1)
Minimal flow (2)	311 (38.9)	42 (15.8)
Moderately strong flow (3)	219 (27.4)	109 (41.0)
Very strong flow (4)	48 (6.0)	104 (39.1)

* p<0.0001

3.5 Conclusions

The pre-operative assessment of adnexal tumors is a major challenge for the gynecologist.

A retrospective assessment of all previous reports suggests that the study populations were insufficiently large for model development or evaluation due to differences in the prevalence of different types of ovarian cancers and the less common benign tumors. There were also variations in the interpretation of ultrasound derived end-points.

The aim of the IOTA (International Ovarian Tumor Analysis) study was to minimize the limitations of previous reports by studying more than 1000 patients with a persistent adnexal mass. Over 50 defined variables were recorded (using a standardized protocol) on a database for the development of a logistic regression model that could be used to achieve high sensitivity and/or specificity for malignant ovarian tumors. The multicenter approach was chosen as the most likely method to achieve a model which might be more effectively applicable to prospective studies with different clinic populations.

There is not only a marked difference in the prevalence of malignancy between the different centers: Milano 9.2% versus Monza 37.0%, but also in the explanatory variables, e.g. Milano used hormonal therapy in 3.4% of the patients while Leuven applied this to 31.2%.

Certain variables are more subjective (e.g. whether the patient feels pain during the ultrasound examination). Statistical analysis will reveal if those are useful for constructing a predictive model.

Chapter IV

Overall IOTA model

This chapter describes the model building steps of the construction of a predictive model – which we call overall model - for the pre-operative prediction of ovarian cancer using the IOTA trial. We discuss the clinical importance of sensitivity, which is in this specific context more significant than specificity and assess whether the serum tumor marker CA 125 can improve the performance of the model.

4.1 Introduction

The pre-operative assessment of adnexal tumors remains a major challenge for the gynecologist. Advances in surgery have provided more treatment options, but their potential usefulness depends upon a prior assessment of the mass using non-invasive procedures.

A test procedure with a high sensitivity for malignancy (true positive rate) could be used to indicate that surgery should be undertaken by an appropriately trained oncologist, who would also determine the progression (stage) of the disease. The false positive rate (1 - specificity) would indicate the proportion of patients who would probably have an unnecessary, or unnecessarily extensive, operation but might be reassured by the histological examination of excised tissue. Conversely, a high specificity (true negative rate) would identify those adnexal masses which could either be removed by laparoscopic surgery, or left in situ and monitored. The false negative rate (1 - sensitivity) shows the proportion of patients who would be denied appropriate surgical treatment. Appropriate surgical treatment is essential because the rupture of a stage 1 ovarian cancer during the operation may worsen the prognosis (Vergote et al, 2001).

In 1989 Granberg reported that transvaginal sonographic images of an ovarian mass could be used to predict the likelihood of malignancy (Granberg et al, 1989). This approach was later incorporated into scoring

systems to improve test performance (Sassone et al, 1991; Timor-Tritsch et al, 1993; Lerner et al, 1994). In 1989 it was also shown that an ultrasound derived index of tumoral blood flow might be an important risk factor (Bourne et al, 1989). Other workers used menopausal status and the serum CA 125 level to produce a risk of malignancy index (Jacobs et al, 1990). The logical extension of this work was to use logistic regression models (Tailor et al, 1997; Timmerman, Bourne et al, 1999) or neural networks (Timmerman, Verrelst et al, 1999; Tailor et al, 1999) to establish the most important variables and improve test performance. However, the results of prospective evaluations of these models at different centers were not very encouraging (Aslam et al, 2000; Mol et al, 2000; Valentin et al, 2001). One study reported values for sensitivity and specificity figures as low as 62% and 79%, respectively (Valentin et al, 2001), while another study reported unbalanced figures of 9 % and 91%, respectively (Aslam et al, 2000). A retrospective assessment of all reports suggests that the study populations were insufficiently large for model development or evaluation because of differences in the prevalence of different types of ovarian malignancies and rare benign tumors. There were also variations in the interpretation of ultrasound derived end-points.

The aim of the IOTA study was to minimize the limitations of previous reports by studying a total of 1000 patients with a persistent adnexal mass at a minimum of six centers working to a common protocol. Over 50 defined variables were recorded on a database for the development of a logistic regression model that could be used to achieve high sensitivity (>90%) for malignant ovarian tumors (and a specificity >75%). The multicenter approach was chosen as the most likely method to achieve a model which might be more effectively applicable to prospective studies in different clinic populations.

For the classification problem of benign versus malignant we select randomly stratified a training or development set (70% of the data) and a test (the remaining 30%) set. With stratification we mean that in each training and test set the following percentages are verified:

- 75% benign, 25% malignant, and
- 7.5% data from the center Paris Boucicaut, 5.1% data from London, 24.7% from Leuven, ...

. The training set is used to build the logistic regression model, the test set is used to have an idea of the performance of the model on a new set of records.

Two previous theses dealt with nonlinear classification models on the same dataset. Geert Fannes (2004) used *Bayesian learning with expert knowledge*. Note that he used a previous version of the data set. He included

also the non-dominant cases in the data, which resulted in 1346 tumors in total (in comparison with the 1066 tumors we have used). Chuan Lu (2005) used exactly the same training and test set as we did and developed several probabilistic machine learning approaches to this medical classification problem. She built for example Least Squares Support Vector Machines, and multilayer perceptron models on the development set and investigated the performance on the test set in order to be able to have a fair comparison with the logistic regression model developed in this chapter.

An important question in gynecology is: Do measurements of serum CA 125 help to distinguish between benign and malignant adnexal tumors ? Serum CA 125 measurements alone or as part of a multimodal strategy are widely used in clinical practice. In order to solve the question about the use of CA 125, we will first build a model on all the explanatory variables, except CA 125, and then verify if the inclusion of CA 125 improves the model performance.

We will start with addressing the problem of multicollinearity. Then we discuss the issue of missing values and how we deal with it. We consider the several model building steps of the logistic regression model and discuss the interpretation and performance of the obtained model. And finally we assess whether the serum tumor marker CA 125 can significantly improve the performance.

4.2 Multicollinearity

There are several groups in which the variables are correlated with each other:

- ***LesD1***, ***LesD2***, ***LesD3***, ***MaxLesD123*** and ***VolumeLes***
- ***SolidD1***, ***SolidD2***, ***SolidD3***, ***MaxSolidD123***, ***VolumeSolid*** and ***RatioSolidLes***
- ***PapHeight***, ***PapBase1***, ***PapBase2***, ***MaxPap123***, ***VolumePap*** and ***RatioPapLes***.
-

We start with constructing several scatter matrices to inspect possible collinearity problems. A scatter matrix is a matrix of plots with on its diagonal the histograms of the variables and in all other cells scatter plots of two variables. For example, in the left corner, we can find the histogram of the variable ***LesD1***. The cell next to it contains the scatter plot with on the Y-axis the ***LesD1***-measurements and on the X-axis the ***LesD2***-values, and so on. Note that scatter matrices can also be used to detect outliers and leverage points.

If we look at the scatter matrix of all the *Lesion* and *Solid* variables (see Figure 4-1), we can see very clearly that there is a large correlation within all five *Lesion* variables. The same conclusion applies for the five *Solid* variables. There is no remarkable correlation between the *Lesion* and the *Solid* variables. Note that we do not see a clear correlation between ***RatioSolidLes*** and any of the *Lesion* or *Solid* variables. This is because ***RatioSolidLes*** expresses the ratio of two other variables: ratio of ***VolumeSolid*** to ***VolumeLes***.

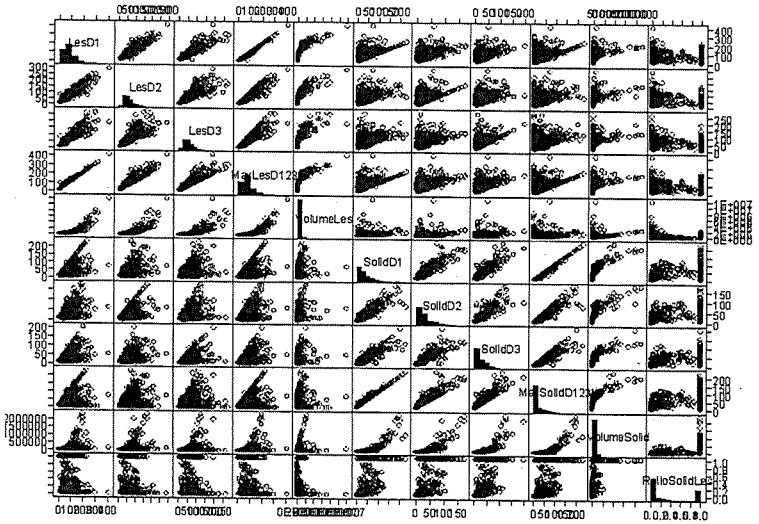


Figure 4-1 Scatter matrix of the *Lesion* and *Solid* variables. The diagonal elements of the matrix contain the histograms of the variables. The other elements contain scatter plots with on the Y-axis the variable which has his histogram in the same row, and on the X-axis the variable which has his histogram in the same column.

Figure 4-2 illustrates the high correlation between the five *Papillation* variables, the low correlation between the *Lesion* and the *Papillation* variables. There is no remarkable correlation between ***RatioPapLes*** and any of the other variables visualized on this scatter matrix, because RatioPapLes is a ratio of two variables. But, it is quite evident that if we would have a logistic regression model with already a *Lesion* variable and a *Papillation* variable in it, and add ***RatioPapLes*** to that model, we would have a problem of multicollinearity.

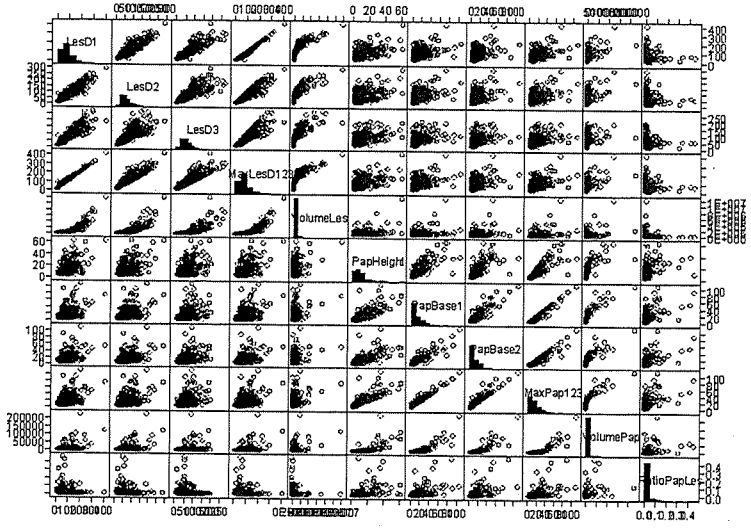


Figure 4-2 Scatter matrix of *Lesion* and *Papillation* variables. The diagonal elements of the matrix contain the histograms of the variables. The other elements contain scatter plots with on the Y-axis the variable which has his histogram in the same row, and on the X-axis the variable which has his histogram in the same column.

Figure 4-3 illustrates the low correlation between the *Solid* and the *Papillation* variables.

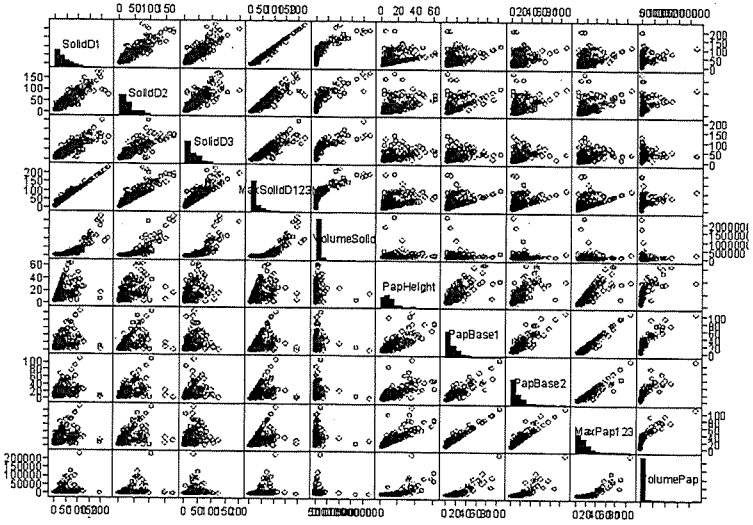


Figure 4-3 Scatter matrix of *Solid* and *Papillation* variables. The diagonal elements of the matrix contain the histograms of the variables. The other elements contain scatter plots with on the Y-axis the variable which has his histogram in the same row, and on the X-axis the variable which has his histogram in the same column.

In each of the above groups we should choose only one variable in order to avoid collinearity problems. In order to do so, we use the Akaike Information Criterion (AIC Atkinson, 1980; Van Houwelingen & Le Cessie, 1990) and the Schwarz Criterion (SC Schwarz, 1978) for comparing several models. We also report the area under the Receiver Operating Curve for each model.

We start with the group of *Lesion* variables. For each variable we build a logistic regression model containing only that variable and calculate the AIC, SC and the area under the ROC curve. For example, we calculate the model fit statistics for a model only containing ***LesD1*** and we do the same for ***LesD2***, ***LesD3***, ***MaxLesD123*** and ***VolumeLes***.

In Table 4-1, we can see that the lowest AIC and SC are reached for ***LesD2***. The highest area under the ROC curve is reached for ***LesD1***. Note that the values of AIC and SC for ***LesD1*** are very comparable to those for

LesD2. So, we should choose between *LesD1* and *LesD2*. We have to remark that, in recording the measurements of the lesion, the order of *LesD1*, *LesD2* and *LesD3* was not prespecified. This means it is possible that sometimes *LesD1* is the largest diameter of the lesion, whereas in other cases it can be the smallest diameter. There were no standard agreements on the order of the *LesD1*, *LesD2* and *LesD3* measurements. Therefore, the choice between *LesD1* and *LesD2* is quite arbitrary. If we look at the AIC and SC for *MaxLesD123*, we see that they are quite similar to those of *LesD2* and the area under the ROC curve does not differ much from the *LesD1* value. For that reason, we decide to retain the variable *MaxLesD123* for further model building.

Table 4-1 Akaike information criterion (AIC), Schwarz Criterion (SC) and area under the ROC curve for *Lesion* variables in the development set.

	AIC	SC	Area under the ROC curve
LesD1	794.824	804.075	0.696
LesD2	794.706	803.957	0.677
LesD3	806.213	815.464	0.676
MaxLesD123	**795.802**	**805.053**	**0.693**
VolumeLes	826.649	835.900	0.688

Table 4-2 Akaike information criterion (AIC), Schwarz Criterion (SC) and area under the ROC curve for *Solid* variables in the development set.

	AIC	SC	Area under the ROC curve
SolidD1	466.852	474.748	0.782
SolidD2	468.692	476.588	0.783
SolidD3	468.236	476.132	0.768
MaxSolidD123	**461.598**	**469.494**	**0.791**
VolumeSolid	517.520	525.416	0.770

Table 4-3 Akaike information criterion (AIC), Schwarz Criterion (SC) and area under the ROC curve for *Papillation* variables in the development set.

	AIC	SC	Area under the ROC curve
PapHeight	**238.243**	**244.840**	**0.776**
PapBase1	246.354	252.950	0.760
PapBase2	245.947	252.543	0.752
MaxPap123	242.041	248.638	0.771
VolumePap	276.008	266.605	0.779

In the group of *Solid* variables (see Table 4-2), ***MaxSolidD123*** has the lowest AIC and SC, and the highest area under the ROC curve. So, it is quite obvious we should retain this variable from all *Solid* measurements.

In the group of *Papillation* variables (see Table 4-3) we see that ***PapHeight*** has the lowest AIC and SC, and almost the highest area under the ROC curve. Therefore, we keep ***PapHeight***.

We can conclude that from the three clusters we retain ***MaxLesD123***, ***MaxSolidD123*** and ***PapHeight*** plus the two ratios ***RatioSolidLes*** and ***RatioPapLes***.

Still there are some other continuous factors in the dataset:

- ***PI***, ***RI***, ***PSV*** and ***TAMXV***
- ***Age*** and ***PostMenoyears***
- ***Fluid*** and ***Septum***.

In Figure 4-4 we can see correlation between ***PI*** and ***RI***, between ***PSV*** and ***TAMXV***, and between ***Age*** and ***PostMenoYears***.

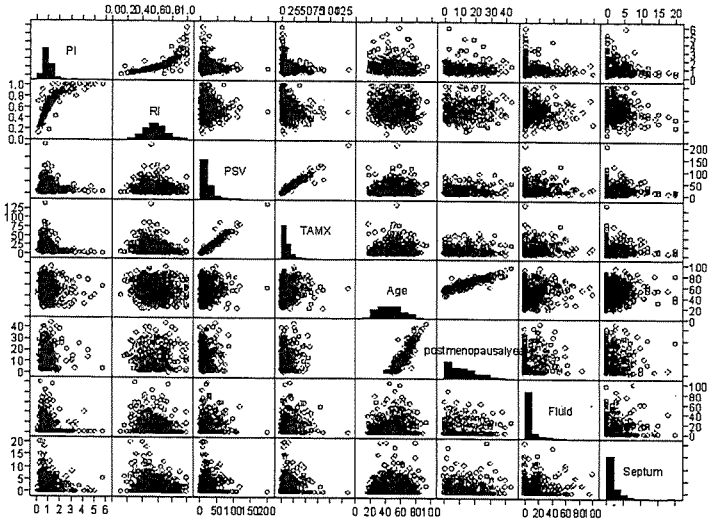


Figure 4-4 Scatter matrix of blood flow indices, Age, PostMenYears, Fluid and Septum. The diagonal elements of the matrix contain the histograms of the variables. The other elements contain scatter plots with on the Y-axis the variable which has his histogram in the same row, and on the X-axis the variable which has his histogram in the same column.

We should choose one of the variables ***PI*** and ***RI***, ***PSV*** and ***TAMXV***, and ***Age*** and ***PostMenoYears***. The values of AIC and SC for ***Age*** and ***PostMenoYears*** are not comparable because there is only a measure-

ment of ***PostMenoYears*** for postmenopausal patients. Otherwise it is structural missing. For that reason we will build models only for the different combinations of the blood flow measurements and postpone the collinearity problem between ***Age*** and ***PostMenoYears***.

Table 4-4 Akaike information criterion (AIC), Schwarz Criterion (SC) and area under the ROC curve for the blood flow indices in the development set.

	AIC	SC	Area under the ROC curve
PI + PSV	551.646	564.459	0.798
PI + TAMXV	534.899	547.655	0.801
RI + PSV	540.610	553.610	0.805
RI + TAMXV	**527.040**	**539.796**	**0.808**

The variables ***RI*** and ***TAMXV*** are included in further analyses because the model with this combination shows the smallest AIC and SC, and the largest area under the ROC curve, among the others as demonstrated in Table 4-4.

4.3 Missing values

There are four levels of structural missings in the total dataset:

- 513 cases without a solid component, i.e. missings in ***SolidD1***, ***SolidD2***, ***SolidD3***, ***MaxSolidD123***, ***VolumeSolid*** and ***RatioSolidLes***,
- 789 cases without papillary projections, i.e. missings in ***PapHeight***, ***PapBase1***, ***PapBase2***, ***MaxPap123***, ***VolumePap***, ***RatioPapLes***, ***PapNr***, ***PapSmooth*** and ***PapFlow***,
- 314 cases without measurements of the blood flow indices, i.e. missings in ***PI***, ***RI***, ***PSV*** and ***TAMXV***,
- 692 cases without values of ***PostMenoyears*** and ***PMB***. Most of them are premenopausal.
-

If we would proceed in the standard way of dropping the missing values, we would retain only 97 complete observations (see Figure 4-5). Note that the missings here are structural missings and not missing values at random. If we would drop all missing values we would not get a random sample of the total population but a very specific subsample: postmenopausal patients with papillary projections and blood flow within septa, cyst walls or solid tumor.

If we would construct several models for each possible combination, we would end up with 12 models, which we should combine afterwards:

- a model for the 160 premenopausal patients without a solid component and blood flow measurements,
- a model for the 62 postmenopausal patients without solid component and blood flow measurements,
- ...
- a model for the 97 postmenopausal patients with papillary projections and blood flow measurements.

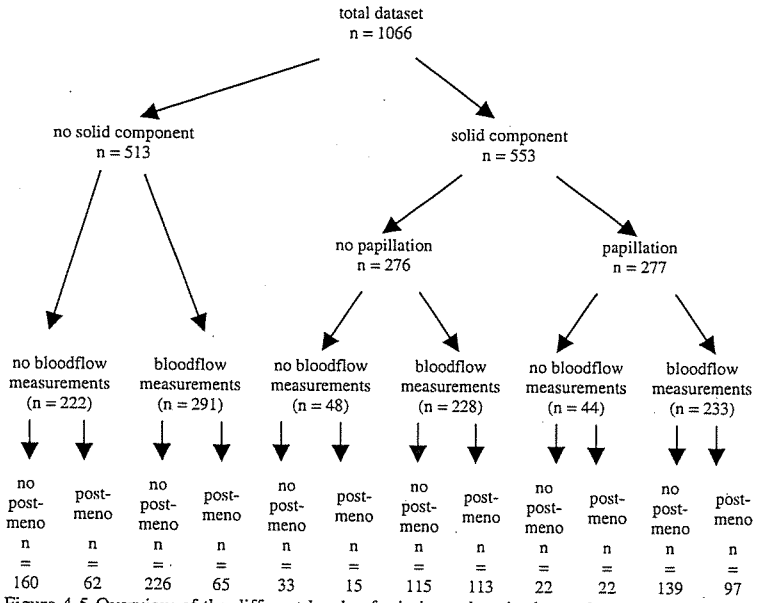


Figure 4-5 Overview of the different levels of missing values in the explanatory variables in the total dataset (development + test set).

In stead of constructing 12 models, we will impute the structural missings and build one model for the whole dataset. Most variables have structural missing observations because the variable does not apply to that patient, for example, ***PostMenoyears*** and ***PMB*** are only relevant for postmenopausal patients. In these cases, 0 is imputed. Likewise, for patients without solid papillary projections (***Papillation*** = 0), the variables ***PapNr***, ***PapHeight***, ***PapBase1***, ***PapBase2***, ***PapFlow***, and ***PapSmooth*** are missing and a "0" is imputed. The same is done for ***SolidD1***, ***SolidD2***, and ***SolidD3*** in case no solid components are present. Depending on the amount of tumor

blood flow (cf. ***ColScore***), ***PI***, ***RI***, ***PSV***, and ***TAMXV*** are missing or not. When intra-tumoral blood flow velocity waveforms were not detected, the peak systolic velocity, time averaged maximum velocity, the pulsatility index and the resistance index were coded as 2.0 cm/sec, 1 cm/sec, 3.0 and 1.0 respectively. In some cases, ***TAMXV*** is missing but the other 3 variables are not (13 cases from Milano). In these cases, ***TAMXV*** is not imputed by 1 but is regressed on ***PSV*** (simple linear regression using all available cases). This is done because there is a strong linear relationship between those two (as can be seen from Figure 4-4).

We will denote the variables with the imputed missings by the suffix –M. For example, ***PapHeightM*** and ***PapFlowM*** are the original ***PapHeight*** and ***PapFlow*** with the structural missings (cases without papillary projections) imputed by "0".

Note we can expect that the discriminatory performance will be lower for a model based on data where most structural missings are imputed with value "0" than for a model which is the combination of twelve models for the specific subgroups. But in clinical practice, it is easier to work with one overall model than with twelve specific models. Therefore, we decide to proceed in this way.

4.4 Construction of dummy variables

Before we can start with variable selection we have to create dummy variables for the categorical variables.

- For the variable ***Locularity***, we take the "unilocular" group as reference group. We define the dummy variables as follows:
  - ***loc2***=1 if ***Locularity***=2 (unilocular cyst with solid component), ***loc2***=0 otherwise
  - ***loc3***=1 if ***Locularity***=3 (multilocular cyst), ***loc3***=0 otherwise
  - ***loc4***=1 if ***Locularity***=4 (multilocular cyst with solid component), ***loc4***=0 otherwise
  - ***loc5***=1 if ***Locularity***=5 (solid tumor), ***loc5***=0 otherwise
  - ***loc6***=1 if ***Locularity***=6 (unclassifiable cyst), ***loc6***=0 otherwise.
- For the variable ***Echogenicity***, we also take the first group "anechoic" as reference group:
  - ***echo2***=1 if ***Echogenicity***=2 (low level echogenic), ***echo2***=0 otherwise

  - ***echo3***=1 if ***Echogenicity***=3 ('ground glass' appearance), ***echo3***=0 otherwise
  - ***echo4***=1 if ***Echogenicity***=4 (hemorrhagic), ***echo4***=0 otherwise
  - ***echo5***=1 if ***Echogenicity***=5 (mixed echogenic), ***echo5***=0 otherwise
  - ***echo6***=1 if ***Echogenicity***=6 (no cyst fluid), ***echo6***=0 otherwise.
- For the variable ***ColScore*** we will not define dummy variables since this is an *ordinal* categorical variable (1-no blood flow, 2-only minimal flow, 3-moderate flow and 4-marked blood flow). If we would define dummy variables we would not use anymore the order of the several categories.

Finally, we discuss one special type of variables: variables where the zero value occurs with a frequency much greater than expected for a fully continuous distribution. This is the case for the ***Fluid***, ***Septum*** and all the imputed variables with value "0". The correct way to include such a variable is to include two terms, one that is dichotomous recording zero versus non-zero and one for the actual recorded value. For example, for the variable ***Fluid***, we will use the original variable ***Fluid*** and create a dichotomous one ***Fluidcat*** (***Fluidcat*** equals 1 if Fluid>0, and 0 otherwise).

4.5 Model building steps

As stated in Chapter II, the goal of any method is to select those variables that result in a "best" model within the scientific context of the problem. In order to achieve this goal we must have:

- a basic plan for selecting the variables for the model and
- a set of methods for assessing the adequacy of the model both in terms of its individual and its overall prediction.

We will start with the selection of the variables.

4.5.1 Variable selection

The traditional approach to statistical model building involves *seeking the most parsimonious model that still explains the data*. The rationale

for minimizing the number of variables in the model is that the resulting model is more likely to be numerically stable, and is more easily generalized. The more variables are included in a model, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data.

We will apply the stepwise method in which variables are selected either for inclusion or exclusion from the model in a sequential fashion based solely on statistical criteria: forward selection with a test for backward elimination. The stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm that checks for the importance of variables, and either includes or excludes them based on a fixed decision rule. The importance of a variable is defined in terms of a measure of the statistical significance of the coefficient for the variable, which is assessed via the likelihood ratio chi-square test. Thus, at any step in the procedure the most important variable, in statistical terms, is the one that produces the greatest change in the log-likelihood relative to a model not containing that variable.

The stepwise approach is useful and intuitively appealing in that it builds models in a sequential fashion and it provides a fast and effective means to screen a large number of variables.

The obtained variables after stepwise selection are the following fourteen (see Table 4-5): ***PersHistOvCa***, ***HormTherapy***, ***Age***, ***MaxLesD123***, ***Pain***, ***Ascites***, ***PapFlowM***, ***loc5***, ***MaxSolidD123M***, ***WallRegularityM***, ***Shadows***, ***ColScore***, ***loc3***, and ***Fluid***. No other variable met the 0.05 significance level for entry into the model.

Table 4-5 Summary of stepwise selection in the development set. In the first step MaxSolidD123M entered the model. Then WallRegularityM was added to the model, and so on. At each step, after a new variable is added, it is verified if one of the previous included variables should be removed from the model. Here, no variables were removed.

Step	variable entered	variable removed	df	# variables in model	Score Chi-Square	p-value
1	*MaxSolidD123M*		1	1	231.8244	<.0001
2	*WallRegularityM*		1	2	79.3722	<.0001
3	*ColScore*		1	3	41.6065	<.0001
4	*Age*		1	4	29.2191	<.0001
5	*Ascites*		1	5	21.8036	<.0001
6	*PapFlowM*		1	6	11.0327	0.0009
7	*Shadows*		1	7	10.1319	0.0015
8	*HormTherapy*		1	8	7.8978	0.0049
9	*Pain*		1	9	7.9187	0.0049
10	*loc5*		1	10	5.8147	0.0159
11	*MaxLesD123*		1	11	7.6030	0.0058
12	*PersHistOvCa*		1	12	4.7673	0.0290
13	*loc3*		1	13	5.1278	0.0235
14	*Fluid*		1	14	4.7302	0.0296

Following the estimate of the multivariable model, the importance of each variable included in the model should be verified. This includes a comparison of each estimated coefficient (see Table 4-6) with the coefficient from the model containing only that variable (see Table 4-7).

Table 4-6 Logistic regression parameters of the full model, obtained by stepwise selection in the training set.

	Analysis of Maximum Likelihood Estimates				
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	p-value
Intercept	1	-7.3369	0.7938	85.4279	<.0001
PersHistOvCa	1	1.9087	0.8212	5.4017	0.0201
Age	1	0.0347	0.00934	13.7584	0.0002
HormTherapy	1	-0.9877	0.3455	8.1721	0.0043
MaxLesD123	1	0.00752	0.00287	6.8470	0.0089
Pain	1	-0.9921	0.3442	8.3050	0.0040
Ascites	1	1.0321	0.4722	4.7767	0.0288
Fluid	1	0.0335	0.0152	4.8355	0.0279
PapFlowM	1	1.4231	0.3528	16.2701	<.0001
loc3	**1**	**1.0475**	**0.4328**	**5.8593**	**0.0155**
loc5	1	1.4965	0.4709	10.1008	0.0015
MaxSolidD123M	1	0.0207	0.00568	13.2169	0.0003
WallRegularityM	1	1.4794	0.3324	19.8064	<.0001
Shadows	1	-2.2558	0.8372	7.2596	0.0071
ColScore	1	0.6500	0.1702	14.5791	0.0001

Table 4-7 Logistic regression parameters of fourteen single models, i.e. for each variable in the full model, a model containing only that variable is fitted on the training set.

Parameter	Estimate
PersHistOvCa	1.2863
Age	0.0489
HormTherapy	-0.3541
MaxLesD123	0.0135
Pain	-0.7337
Ascites	3.0397
Fluid	0.0809
PapFlowM	2.2548
loc3	**-1.1980**
loc5	1.8491
MaxSolidD123M	0.0427
WallRegularityM	2.2857
Shadows	-2.6446
ColScore	1.2599

We can see that for loc3, there is a marked difference in the estimated coefficient of the full model (1.0475) compared to the estimated coefficient of a model containing only the variable loc3 (-1.1980). If we have only loc3 in our model, we can say that patients with multilocular cysts have a smaller odds of having a malignant tumor than otherwise (odds ratio 0.302). However, in the full model the conclusion is just the contrary (odds

ratio 2.851). The latter statement is clinically not valid. Therefore we decide to drop loc3 from our model.

There is no problem with the other variables in our model.

Now we come to the point of checking the assumption of linearity in the logit for continuous variables. One approach is to fit a generalized additive model (Hastie and Tibshirani, 1999).

We have several continuous variables in our model: ***Age***, ***MaxLesD123***, ***MaxSolidD123*** and ***Fluid***. ***ColScore*** is an ordinal variable (categories from 1 to 4). We will also check if the relationship is linear.

We can assess the validity of the linear assumption by fitting an additive model with relationships estimated by for example a spline smoother (Harrell et al, 1988), and then comparing it to the linear model. The tests correspond to an approximate partial test for the importance of the smoothness for each term in the model. For a single-variable model, this is equivalent to testing for a difference between a linear fit and a smooth fit that includes both linear and smooth terms. The approximate nature of the partial tests is discussed in detail by Hastie and Tibshirani (1999).

Table 4-8 Test for nonlinear effect of each continuous or ordinal term in the model.

	p-value
Age	0.3680648
MaxLesD123	0.0061144
Fluid	0.0806748
MaxSolidD123M	0.0000077
ColScore	0.0471013

From Table 4-8, we can see that a transformation is needed for ***MaxLesD123*** and ***MaxSolidD123***. The nonlinearity in ***ColScore*** is rather borderline significant and for ***Fluid*** and ***Age***, there is no transformation needed.

Figures 4-6A, 4-6B, 4-6C, 4-6D and 4-6E show for each variable the estimated spline function with its 95% confidence limits. The spline fit for ***MaxLesD123*** decreases a little from 0 to 50 to 60 mm, then it increases till 150 mm and remains rather constant for higher values of ***MaxLesD123***. The fit for ***MaxSolidD123M*** increases from 0 till 50 mm, stays stable from 50 till 100 mm, and then decreases again, but note that there are only few cases with values above 100 mm.

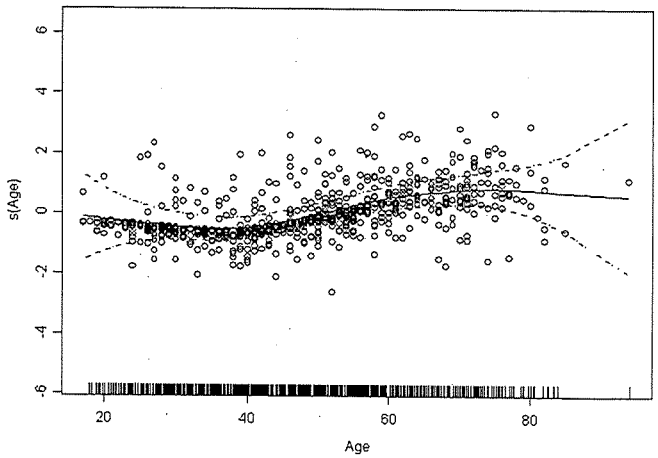


Figure 4-6A Plot of the spline smoother for the variable ***Age*** in the generalized additive model.

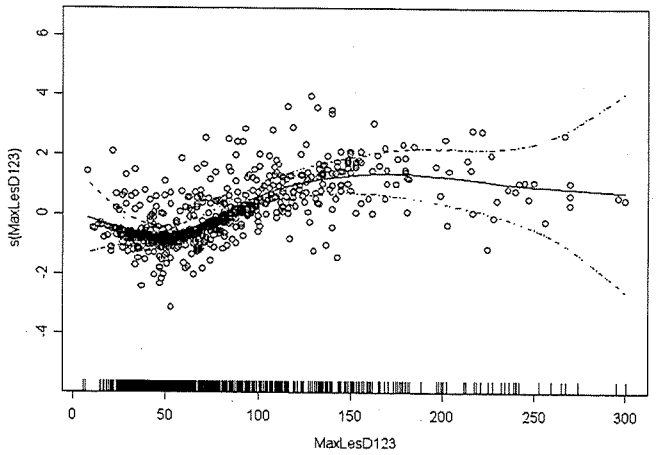


Figure 4-6B Plot of the spline smoother for the variable ***MaxlesD123*** in the generalized additive model.

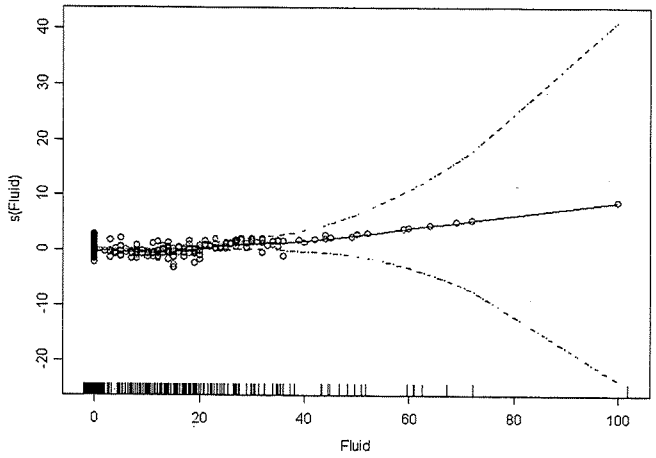


Figure 4-6C Plot of the spline smoother for the variable ***Fluid*** in the generalized additive model.

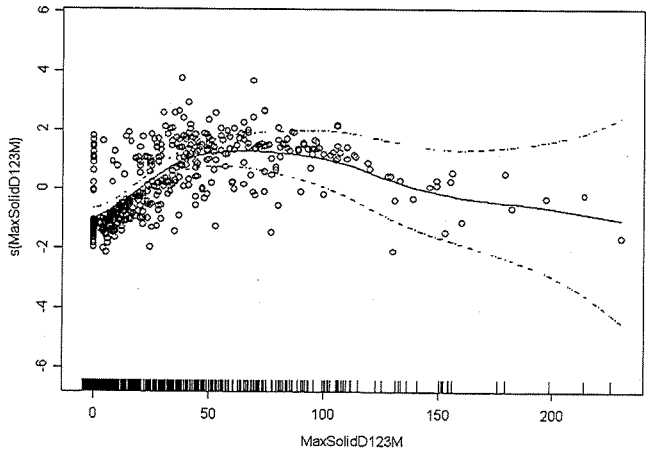


Figure 4-6D Plot of the spline smoother for the variable ***MaxSolidD123M*** in the generalized additive model.

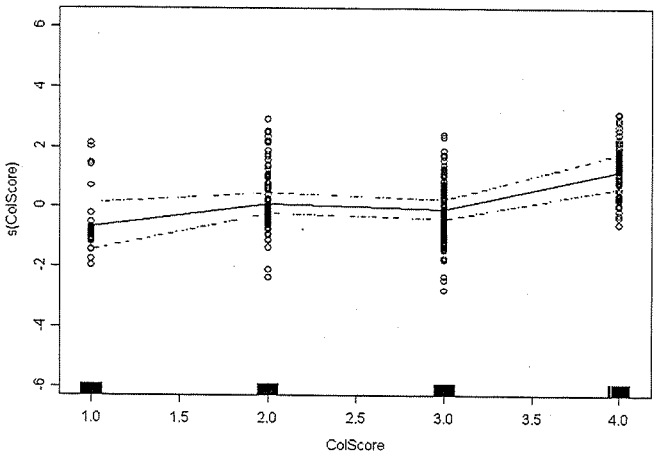


Figure 4-6E Plot of the spline smoother for the variable ***ColScore*** in the generalized additive model.

We decide to create a breakpoint in the variable ***MaxSolidD123M***: values larger than 50 mm are replaced by the value of 50. This means that we will have an increasing function till the breakpoint of 50 mm and a flat function afterwards (the transformed MaxSolidD123M with a cutoff at 50 mm is denoted by "***m***"). We do not transform ***Age*** or ***Fluid***. We can first have a look what the impact is of our transformation of ***MaxSolidD123M*** on the other variables. It may be that no further transformation is required for e.g. ***ColScore***.

We fit our selected 13 variables where ***MaxSolidD213M*** is transformed (variable ***m***). Table 4-9 shows us that ***Fluid*** becomes nonsignificant (p-value 0.1253) after the transformation of ***MaxSolidD123M***. So, we omit ***Fluid*** from our model.

Now we can check if there are still transformations needed for ***MaxLesD123*** and ***ColScore***. ***ColScore*** should not be transformed anymore (p-value 0.0969). The p-value for ***MaxLesD123*** has increased a little bit (from 0.0061 to 0.0146) but remains smaller than 0.05 (see Table 4-10). Figure 4-7 still shows that ***MaxLesD123*** decreases slightly from 0 mm to 50 mm and then starts to increase. Because there is no real clinically valid reason for the decrease we decide not to transform the variable ***MaxLesD123***.

Table 4-9 Analysis of maximum likelihood estimates of the model where the dummy variable loc3 has been omitted and with MaxSolidD123M transformed (cutoff at 50mm).

Parameter	DF	Estimate	Standard Error	Wald Chi-square	p-value
Intercept	1	-6.7991	0.7508	82.0156	<.0001
PersHistOvCa	1	1.6867	0.8078	4.3594	0.0368
HormTherapy	1	-1.0349	0.3576	8.3773	0.0038
Age	1	0.0323	0.00943	11.7181	0.0006
MaxLesD123	1	0.00880	0.00276	10.1394	0.0015
Pain	1	-0.8610	0.3418	6.3461	0.0118
Ascites	1	1.1494	0.4760	5.8303	0.0158
Fluid	**1**	**0.0240**	**0.0156**	**2.3501**	**0.1253**
PapFlowM	1	1.1649	0.3520	10.9515	0.0009
loc5	1	0.9428	0.4484	4.4218	0.0355
m	1	0.0478	0.00909	27.6503	<.0001
WallRegularityM	1	1.1379	0.3324	11.7171	0.0006
Shadows	1	-2.3768	0.8024	8.7732	0.0031
ColScore	1	0.4913	0.1717	8.1849	0.0042

Table 4-10 Test for nonlinear effect in the final model.

	p-value
MaxLesD123	0.0146
ColScore	0.0969

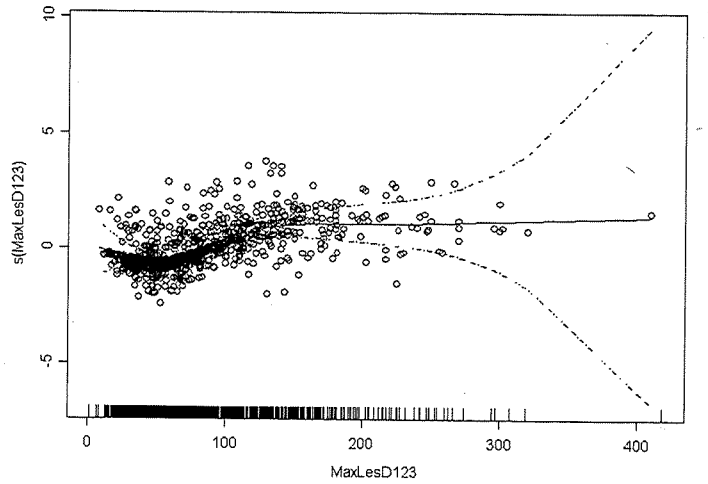


Figure 4-7 Plot of the spline smoother for MaxLesD123 in the final model

At this point we can check if we can add any other variable to our model. This is not the case. Only ***Bilateral*** seems to be borderline significant: (p-value 0.0631 when we include it). So, our final model contains the following variables: ***PersHistOvCa***, ***HormTherapy***, ***Age***, ***MaxLesD123***, ***Pain***, ***Ascites***, ***PapFlowM***, ***loc5***, ***m*** (***MaxSolidD123M*** with cutoff 50 mm), ***WallRegularityM***, ***Shadows*** and ***ColScore***.

Once we have a *final* model, we check for interactions (Harrell, 2002) among the variables in the model. In any model, an interaction between two variables implies that the effect of one variable is not constant over the levels of the other. Of course, the final decision, as to whether an interaction term should be included in a model should be based on statistical as well as practical considerations. Any interaction term in the model must make sense from a clinical perspective. Here, in our model, no interaction terms are needed.

4.5.2 Assessing the Fit of the Model

We would like to know how effectively the obtained model describes the outcome variable. This is referred to as its goodness-of-fit.

A complete assessment of the fitted model involves both the calculation of summary measures of the distance between the observed and fitted outcome values, and an examination of the individual components of these measures.

a) Summary goodness-of-fit: Hosmer-Lemeshow Test

Hosmer and Lemeshow developed a goodness-of-fit test for logistic regression models with binary responses. This test with a p-value of 0.7617 provides no evidence of lack of fit of our model.

b) Logistic Regression Diagnostics

Regression diagnostics tell us how each observation influences the fit of the logistic regression model. We use the Pearson chi-square statistic (DIFCHISQ) which detects ill-fitted observations, i.e. observations that contribute heavily to the disagreement between the data and the predicted values of the fitted model. DIFCHISQ is the change in the Pearson chi-square statistic due to deleting an individual observation.

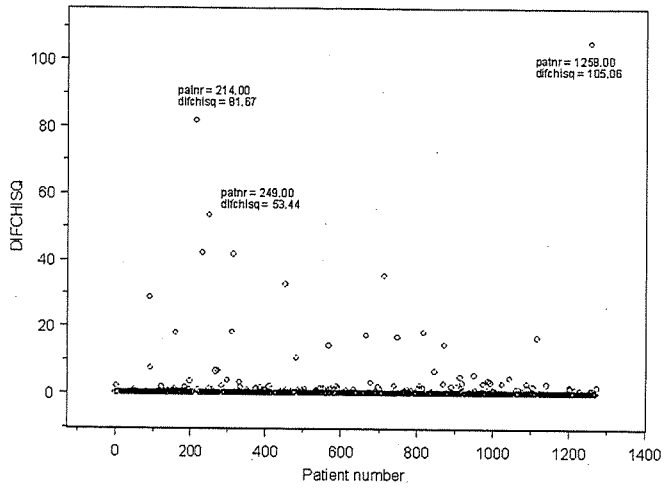


Figure 4-8 Diagnostic plot of the records using the final model. 10 cases have rather high values of the change in the Pearson chi-square statistic (DIFCHISQ).

Figure 4-8 shows a few influential observations. However, they were biologically possible. Moreover, they were not typing errors and thus they will not be discarded from the dataset for further analyses.

4.6 Model Interpretation

Table 4-11 shows the parameter estimates of our final logistic regression model, which contains twelve variables. Note that two rather subjective variables, i.e. ***Pain*** and ***ColScore*** (see Table 3-3), are retained in this model. So, although the differences in ***ColScore*** between the centers, e.g. Malmö gave only Malmö only 6% of the cases the label "no blood flow", while in Leuven this was 29%, and in Milano even 71.3%, it still seems to be an important variable.

Table 4-11 Analysis of maximum likelihood estimates of the final model developed on the training data. All the variables with a postive estimate are risk factor for ovarium cancer, while those with a negative estimate reduce the risk of malignancy. All the explanatory variables are significant at the 0.05 probability level.

Parameter	DF	Estimate	Standard error	Wald chi-square	p-value
Intercept	1	-6.7468	0.7419	82.6909	<.0001
PersHistOvCa	1	1.5985	0.8032	3.9604	0.0466
HormTherapy	1	-0.9983	0.3539	7.9553	0.0048
Age	1	0.0326	0.00941	11.9758	0.0005
MaxLesD123	1	0.00841	0.00275	9.3845	0.0022
Pain	1	-0.8577	0.3383	6.4283	0.0112
Ascites	1	1.5513	0.4030	14.8175	0.0001
PapFlowM	1	1.1737	0.3499	11.2532	0.0008
loc5	1	0.9281	0.4453	4.3437	0.0371
m	1	0.0496	0.00902	30.3141	<.0001
WallRegularityM	1	1.1421	0.3317	11.8591	0.0006
Shadows	1	-2.3550	0.8090	8.4738	0.0036
ColScore	1	0.4916	0.1699	8.3737	0.0038

In the Odds Ratio Estimates table (see Table 4-12), the odds ratio (Point Estimate) is computed by exponentiating the parameter estimate for each explanatory variable. This means that the odds of a malignant tumor are estimated to increase multiplicatively by this value for every one-unit increase of the corresponding explanatory variable or in case of the choice of categorical variable being 1. Remember that the odds of the presence of a malignant ovarian tumor were defined as the ratio of the probability of the presence of a malignant ovarian tumor over the probability of the presence of a benign ovarian tumor. Also, 95% confidence limits for odds ratios are computed by exponentiating the confidence limits for the logistic regression parameters.

Table 4-12 Odds ratio estimates of the final model developed on the training data. A personal history of ovarian cancer (PersHistOvCa), having fluid outside the Pouch of Douglas (Ascites) and the presence of flow within papillary projections (PapFlow) have a large effect on the probability of having a malignant tumor. The presence of acoustic shadows (Shadows), however, reduce the risk of malignancy.

Effect	Point estimate	95% Wald confidence limits	
PersHistOvCa	4.946	1.024	23.874
HormTherapy	0.369	0.184	0.737
Age	1.033	1.014	1.052
MaxLesD123	1.008	1.003	1.014
Pain	0.424	0.219	0.823
Ascites	4.717	2.141	10.393
PapFlowM	3.234	1.629	6.420
loc5	2.530	1.057	6.055
m	1.051	1.032	1.070
WallRegularityM	3.133	1.636	6.002
Shadows	0.095	0.019	0.463
ColScore	1.635	1.172	2.281

If there is a presence of personal history of ovarian cancer (***PersHistOvCa***), the odds of the presence of a malignant ovarian tumor are estimated to be 4.95 times larger than in patients without a personal history of ovarian cancer.

Concerning the variable ***HormTherapy***, when a patient has hormone therapy (***HormTherapy***=1), the odds of the event, the presence of a malignant ovarian tumor, are estimated to be 1/0.369 = 2.71 times smaller than in patients without hormone therapy.

The odds of the presence of a malignant ovarian tumor increase by 3.3% for each year increase in ***Age*** of the patient.

The odds of the presence of a malignant ovarian tumor increase by 0.8% for each one-unit (mm) increase in the value for ***MaxLesD123***.

If the patient had ***Pain*** during US examination, the odds of the presence of a malignant ovarian tumor are estimated to be 1/0.424=2.36 times smaller than that in patients without pain.

If the patient has ***Ascites*** (fluid outside Pouch of Douglas), the odds of the presence of a malignant ovarian tumor are estimated to be 4.72 times larger than that in patients without the presence of ***Ascites***.

If there is presence of flow within the papillary structure (***PapFlowM***), the odds of the presence of a malignant tumor are estimated to be 3.23 times larger than in the absence of flow.

If the patient has a purely solid lesion (***loc5***=1), the odds of the presence of a malignant ovarian tumor are estimated to be 2.53 times larger than otherwise.

The odds of the presence of a malignant ovarian tumor increase by 5% for each one-unit (mm) increase in the value of ***MaxSolidD123***. Values of ***MaxSolidD123*** larger than 50mm have the same effect as a value of 50mm (variable denoted by ***m***).

If the lesion has an irregular internal wall (***WallRegularityM***), the odds of the presence of a malignant ovarian tumor are estimated to be 3.13 times larger than that in a lesion with a smooth internal wall.

If acoustic ***Shadows*** are present on ultrasound examination, the odds of the presence of a malignant ovarian tumor are estimated to be 1/0.095=10.53 times smaller than that in case of the absence of acoustic shadows.

If there is a one-unit increase in the value for ***ColScore***, then the odds of the presence of a malignant ovarian tumor are estimated to be 1.64 times larger.

We can summarize our conclusions in the following way:

Independent risk factors for the presence of malignancy include: the patient's age (a 3.3% increase of odds for malignancy with each additional year of age), a personal history of ovarian cancer (odds 4.95), the maximum diameter of the lesion (0.8% for every mm increment), the maximum diameter of the solid component (5% for every mm increase with an upper limit of 50 mm; no increase above 50 mm), the presence of ascites (odds 4.72), the presence of blood flow within a solid papillary projection (odds 3.23), the presence of a purely solid lesion (odds 2.53), the presence of irregular internal cyst walls (odds 3.13), increase in color score (odds 1.64 for every one unit increase).

Factors that reduce the risk of malignancy include: the current use of hormonal therapy (odds 0.369), the presence of pain during the ultrasound examination (odds 0.424), and the presence of acoustic shadows (odds 0.095).

4.7 Model performance

4.7.1 Calibration

Calibration refers to whether the predicted probabilities agree with the observed probabilities. Since the p-value of the Hosmer and Lemeshow goodness-of-fit test equals 0.7617 (on the development set) we can conclude that the predictions agree *on average* with the observed probabilities. We can also study calibration in more detail by looking at the correspondence

between predictions and observations over the whole range of predictions by making use of a calibration plot.

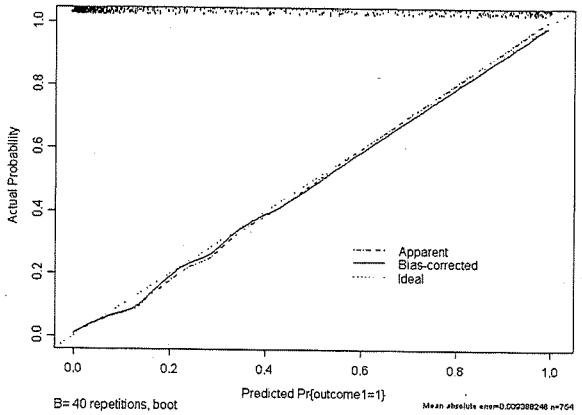


Figure 4-9A Calibration plot of the observed versus the predicted probabilities on the development set. The bias-corrected calibrated values are obtained by bootstrap resampling. Error refers to the difference between the predicted values and the corresponding bias-corrected calibrated values.

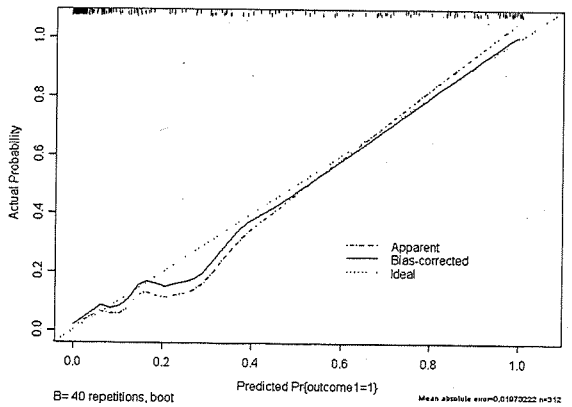


Figure 4-9B Calibration plot of the observed versus the predicted probabilities on the test set.

In the training or development set there is clearly no problem with the calibration (see Figure 4-9A), since the predicted probabilities almost perfectly agree with the observed probabilities. But in order to have an idea how the model will perform on new data, we should assess the calibration in the test set. On the test set, the predictions are too high for actual probabilities between 0.2 and 0.4, but for the rest this calibration plot is ok.

4.7.2 Discrimination

Discrimination refers to the ability to distinguish high risk subjects from low risk subjects, and is commonly measured by the area under the Receiver Operating Characteristic (ROC) curve (Hanley and McNeil, 1982).

We will first discuss the classification table and then come to ROC curves. This table is the result of cross-classifying the outcome variable, with a dichotomous variable whose values are derived from the estimated logistic probabilities.

To obtain the derived dichotomous variable we must define a cutoff point, and compare each estimated probability to that cutoff point. If the estimated probability exceeds the value of the cutoff point then we let the derived variable to be equal to 1; otherwise it is equal to 0. For example in Table 4-13, we take as cutoff point 0.01, 0.05, 0.10, 0.15 and so on. For each cutoff point, we calculate the

- Sensitivity: proportion of malignant tumors that are classified as malignant
- Specificity: proportion of benign tumors that are classified as benign.

Sensitivity figures are more important than specificity figures. Insufficient primary surgery leads to under-staging of patients with ovarian cancer and to poorer survival, since the amount of residual malignant tissue after primary surgery is one of the most important prognostic factors in ovarian cancer (Vergote et al, 2001). The timely referral of patients with a suspicious adnexal mass to an oncologic surgeon is very important in order to decrease mortality from ovarian cancer, which has the highest mortality figures of all pelvic cancers in females.

However, specificity must not be ignored. With the development of new surgical techniques it has become technically feasible to treat or to remove most benign adnexal tumors with laparoscopy surgery. This decreased morbidity and hospitalization length. Moreover, in patients presenting with functional cysts any surgical intervention could create unnecessary morbidity, costs and a possible loss of fertility.

A tumor classification model useful for clinical practice should obtain a sensitivity of at least 90% and a specificity of at least 75%.

Table 4-13 shows the classification table of the final model constructed on the training data set. To reach the above goals of sensitivity and specificity, we select a probability value of 0.15, which gives us a sensitivity of 91.6% and a specificity of 82.8%. Note that in (Timmerman et al, 2004) a biased-correction is applied to the classification table for the training set and a different probability value is choosen, i.e. 0.10.

The classification table from applying the model to the test data set is shown in Table 4-14. The selected probability cutoff value of 0.15 gives a sensitivity of 92.0% and a specificity of 81.0%.

For the clinician the predictive value of a test is more important than sensitivity and specificity. In practice the test result is all that is known, so if we want to know how good the test is at predicting a malignant tumor, we should have a look at the positive and negative predictive value of the test. Positive predictive value (PPV) means what proportion of patients which are predicted to have a malignant tumor do really have a malignant tumor. Negative predictive value (NPV) is the proportion of patients with negative test results who are correctly diagnosed.

If we immediately look at the test data set (since this gives us more reliable information how the model will perform on new data than the development set), we obtain a PPV of 60.5% and a NPV of 97.0% at the cutoff probability level of 0.15. So, 97.0% of the patients that are predicted to have a benign tumor really have a benign mass. While almost 40% of the patients that are predicted to have a malignant mass, have in reality a benign tumor.

So, the advantage of our model is the high sensitivity of 92.0%, but the related disadvantage is that in almost 40% of the cases where the model predicts a malignant tumor, this turns out to be incorrectly.

Note that predictive values observed in one study do not apply universally. We can only expect a PPV of 60.5% and a NPV of 97.0% in centers that have the same prevalence of malignancy as in our multicenter study, i.e. 25% malignancy. In centers with lower prevalence of malignant tumors, a negative test result will more often be associated with a normal outcome, while a positive result will less often indicate a truly abnormal outcome. On the contrary, sensitivity and specificity can be applied to other centers. This means that if we apply our model prospectively we can expect a sensitivity of 92.0% and a specificity of 81.0%, whatever the prevalence of malignant ovarian tumors in those centers is.

Table 4-13 Classification of malignant and benign tumours by probability level following development of a logistic regression model from the development data set (n=754).

probability level	correctly classified		incorrectly classified		Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
	malignant	benign	malignant as benign	benign as malignant					
0.01	190	149	1	414	99.5	26.5	50.0	31.5	99.3
0.05	186	365	5	198	97.4	64.8	73.1	48.4	98.7
0.10	177	433	14	130	92.7	76.9	80.9	57.7	96.9
0.15	**175**	**466**	**16**	**97**	**91.6**	**82.8**	**85.0**	**64.3**	**96.7**
0.20	173	484	18	79	90.6	86.0	87.1	68.7	96.4
0.25	169	497	22	66	88.5	88.3	88.3	71.9	95.8
0.30	166	507	25	56	86.9	90.1	89.3	74.8	95.3
0.35	163	512	28	51	85.3	90.9	89.5	76.2	94.8
0.40	160	517	31	46	83.8	91.8	89.8	77.7	94.3
0.45	155	526	36	37	81.2	93.4	90.3	80.7	93.6
0.50	153	533	38	30	80.1	94.7	91.0	83.6	93.4
0.55	143	535	48	28	74.9	95.0	89.9	83.6	91.8
0.60	135	540	56	23	70.7	95.9	89.5	85.4	90.6
0.65	127	548	64	15	66.5	97.3	89.5	89.4	89.5
0.70	123	552	68	11	64.4	98.1	89.5	91.8	89.0
0.75	115	555	76	8	60.2	98.6	88.9	93.5	88.0
0.80	100	557	91	6	52.4	98.9	87.1	94.3	86.0
0.85	92	558	99	5	48.2	99.1	86.2	94.9	84.9
0.90	72	560	119	3	37.7	99.5	83.9	96.0	82.5

Table 4-14 Classification of malignant and benign tumours by probability level following of the developed logistic regression model applied to the test data set (n=312).

probability level	correctly classified		incorrectly classified		Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
	malignant	benign	malignant as benign	benign as malignant					
0.01	74	55	1	182	98.7	23.2	41.3	28.9	98.2
0.05	72	152	3	85	96.0	64.1	71.8	45.9	98.1
0.10	70	179	5	58	93.3	75.5	79.8	54.7	97.3
0.15	**69**	**192**	**6**	**45**	**92.0**	**81.0**	**83.7**	**60.5**	**97.0**
0.20	66	200	9	37	88.0	84.4	85.3	64.1	95.7
0.25	62	204	13	33	82.7	86.1	85.3	65.3	94.0
0.30	62	208	13	29	82.7	87.8	86.5	68.1	94.1
0.35	60	212	15	25	80.0	89.5	87.2	70.6	93.4
0.40	60	219	15	18	80.0	92.4	89.4	76.9	93.6
0.45	58	223	17	14	77.3	94.1	90.1	80.6	92.9
0.50	58	226	17	11	77.3	95.4	91.0	84.1	93.0
0.55	57	231	18	6	76.0	97.5	92.3	90.5	92.8
0.60	55	234	20	3	73.3	98.7	92.6	94.8	92.1
0.65	51	235	24	2	68.0	99.2	91.7	96.2	90.7
0.70	48	236	27	1	64.0	99.6	91.0	98.0	89.7
0.75	41	236	34	1	54.7	99.6	88.8	97.6	87.4
0.80	36	237	39	0	48.0	100	87.5	100	85.9
0.85	31	237	44	0	41.3	100	85.9	100	84.3
0.90	23	237	52	0	30.7	100	83.3	100	82.0

Sensitivity and specificity rely on a single cutoff point to classify a test result as positive. A more complete description of classification accuracy is given by the area under the Receiver Operating Characteristic (ROC) curve, which is a plot of the sensitivity against 1 minus the specificity for different probability cutoff points. Points near to the top right region of the curve correspond to lower probability cutoff points, whereas points near to the bottom left corner of the curve correspond to higher probability cutoff points.

The area under the ROC curve on the training data reaches 0.946 which is an excellent performance. In the test set the area hardly decreases: 0.942 (see Figure 4-10).

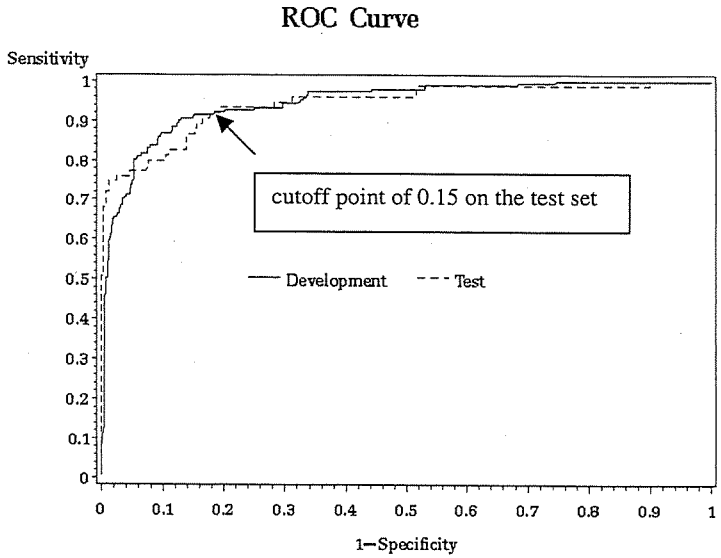


Figure 4-10 The receiver operating characteristic curves (ROC) of the logistic regression model constructed on the development data set (n=754) and applied to the test set (n=312). The 0.15 probability cutoff point is indicated on the ROC curve of the test set.

When the model is tested on the data from each individual center the overall test performance does not deteriorate (see Table 4-15). Note that we can not calculate the value for the area under the ROC-curve for the test set of Naples, since all the observations are benign. If we merge the 8 cases from Naples with the 13 cases from Monza, we find an area under the ROC-curve of 0.913 (The area under the ROC-curve of test set of Monza alone was 0.900). In the three centers that provide 2/3 of the data, i.e. Malmö, Leuven and Rome the model performance stays high: area under the ROC-curve 0.96, 0.96 and 0.93. The worst performance is in London with an area under the ROC curve of 0.76, but this value is calculated on a small number of patients: there are only 15 records in the test set of London.

Table 4-15 Performance in the test set of each center. The test sets of Naples and Monza have been combined, because otherwise no area under the ROC curve could be calculated for Naples, since all 8 patients in the test of that center had a benign tumor.

Center	n	Area under the ROC-curve	Prevalence of malignancy in the whole dataset	Prevalence of malignancy in the test set
Malmö	94	0.955	21.6%	21.3%
Leuven	78	0.964	35.4%	34.6%
Rome	37	0.926	35.7%	35.1%
Milano	25	0.957	9.2%	8.0%
Paris B	23	1	11.3%	8.7%
Paris M	19	0.838	10.9%	10.5%
London	15	0.761	29.6%	26.7%
Naples	8	0.913	9.7%	0%
Monza	13		37.0%	38.5%

4.8 Validation

Internal validity refers to the validation of the model in the population where the sample originated. This has been investigated in the previous section. There was only little difference in the performance from development to test set:

- area under the ROC curve: 0.946 versus 0.942
- sensitivity at cutoff 0.15: 91.6% versus 92.0%
- specificity at cutoff 0.15: 82.8% versus 81.0%.

External validation (or generalization) refers to the validation of the model in other populations, e.g. patients from other centers. This will be investigated in the future by setting up the second phase of the IOTA trial,

with data collecting from centers different from the ones that provided our first IOTA dataset.

There are two ways to apply our model to new patient data:

1) with a formula (see Table 4-11):

$$P(\text{malignant}) = \frac{exp(z)}{1+exp(z)}\text{, where}$$

$$z = -6.7 + 1.6\text{PersHistOvCa} - 1\text{HormTherapy} + 0.0326\text{Age} + 0.00841\text{MaxLesD123} - 0.9\text{Pain} + 1.6\text{Ascites} + 1.2\text{PapFlowM} + 0.9\text{loc5} + 0.0496\text{m} + 1.1\text{WallRegularityM} - 2.4\text{Shadows} + 0.5\text{ColScore}.$$

If the calculated probability exceeds the cutoff probability value of 0.15 then the tumor can be classified as being malignant.

2) graphically via a nomogram (see Figure 4-11). Nomograms (Banks, 1985; Lubsen et al, 1978) provide excellent graphical depictions of all the variables in the model, in addition to enabling the user to obtain predicted probabilities manually. The first row of this monogram is a score from 0 to 100. Row 2 to row 13 present the 12 variables in the model.

This graphical presentation is completely equivalent to the above formula. Binary variables with higher weights in the formula are presented with longer axes (which implies higher scores on the first row) in the monogram. E.g. ***Ascites*** has a weight of 1.6 in the formula, while ***WallRegularityM*** has a weight of 1.1. This is visualized in the monogram by the longer axis of ***Ascites*** in comparison to ***WallRegularityM***.

By using this plot, probabilities of having a malignant ovarian tumor can graphically be determined. For example, suppose we have a patient with a personal history of ovarian cancer (corresponds to ± 55 points on the first row), without ***HormTherapy*** (± 35 points), of 90 years old (± 90 points), ***MaxLesD123***=80mm (20points), with ***Pain*** (0 points), with ***Ascites*** (± 50 points), with ***PapFlowM*** (± 40 points), with a completely solid tumor (± 30 points), ***MaxSolidD123M***=30mm (± 50 points), irregular wall (± 40 points), no shadows (± 80 points) and ***Colscore***=2 (± 20 points). The total score adds up to 55+35+90+20+0+50+40+30+50+40+80 = 490 points. If we point this value on row 14 (Total Points), we see that the corresponding probability of malignancy exceeds 0.95. So, this patient would we classified as having a malignant tumor.

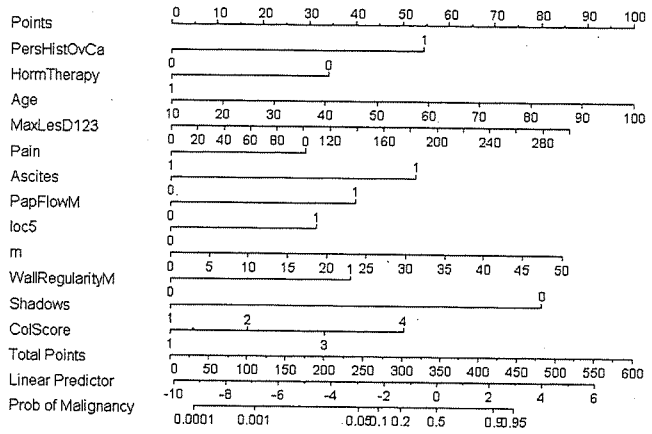


Figure 4-11 Nomogram of the overall model constructed on the development set. This plot can be used to classify new tumors. By determining the corresponding score (see first row) for the value of each of the 12 explanatory variables (row 2 – 13), and adding these scores, we can obtain the predicted outcome by assessing the corresponding probability of malignancy (row 16) for the total score of the tumor (row 14). The cutoff value of 0.15 corresponds to a total score of ± 280 points. So, tumors which have a total score larger than 280 points will be classified as malignant.

4.9 Comparison with 'old' models

In 1989 Granberg reported that transvaginal sonographic images of an ovarian mass could be used to predict the likelihood of malignancy (Granberg, 1989). This approach was later incorporated into scoring systems to improve test performance, e.g. the scoring systems of Lerner (1994). Other researchers have used menopausal status and the serum CA 125 level to produce a risk of malignancy index (Jacobs et al, 1990). Later, investigators focused on the use of logistic regression models and neural networks (Tailor et al, 1997; Tailor, 1999; Timmerman, Verrelst et al, 1999; Timmerman, Bourne et al, 1999).

Table 4-16 shows the performance of our model in comparison to three 'old' models, i.e. the risk of malignancy index (RMI-Jacobs et al, 1990), the logistic regression model of Tailor (Tailor et al, 1997) and a logistic regression model of Timmerman that was developed in the pilot study

(Timmerman, Bourne et al, 1999). If we compare the area under the ROC curve of our model (area under the ROC curve of 0.936) with the best old model (Timmerman, Bourne et al, 1999 – AUC 0.903), the difference between those two is significant (p-value 0.0187, De Long). The comparison with the performance of other models is as follows:

- our model versus RMI (Jacobs et al, 1990) : p-value 0.0038, and
- our model versus Tailor (Tailor et al, 1997) : p-value 0.0006 (for this comparison, we have applied the overall model to the 220 cases with ***TAMXV*** in the test set. Our model reaches on those 220 cases an area under the ROC curve of 0.950 (standard error 0.014).)

If we compare our model to Tailor in terms of specificity, we detect a 8% decrease. But our sensitivity is almost 30% higher, i.e. 90.3% versus 63.2%. Note that sensitivity figures are more important than specificity figures: the sensitivity for malignancy indicates the proportion of patients where surgery should be undertaken by an appropriately trained oncologist. The specificity identifies those adnexal masses which could either be removed by laparoscopic surgery, or left in situ and monitored.

In terms of sensitivity our constructed model reaches the highest values.

Table 4-16 Comparison of the performance of our model and three well known 'old' models

	Area under ROC curve	standard error	cutoff	Sensitivity (%)	Specificity (%)
'final' model*	0.936	0.020	0.15	90.3%	80.2%
RMI*	0.870	0.028	100	78.3	79.6
Tailor**	0.869	0.025	0.25	63.2	88.2
Timmerman*	0.903	0.023	0.25	79.7	80.8

* applied on the cases in the test set with CA125 (236 cases)
** applied on the cases in the test set with TAMX (220 cases)

4.10 Discussion

There was a marked difference in the prevalence of malignancy between the nine participating centers. This finding relates to different referral patterns in the participating centers. Some are regional referral centers for gynecological malignancy, some offer tertiary scanning services, whilst others are more general units. The prevalence of malignancy is likely to be significantly higher in referral centers than in more general units and the demographic data also reflect these differences. Accordingly, we have concentrated on the examination of a relatively large number of benign and malignant masses for logistic regression analysis.

We have identified several defined indices that can be used to discriminate between malignant and benign masses. The developed model is based on a large number and variety of malignant and benign ovarian masses. The size and scope of the dataset enabled the retention of a large number of independently significant variables. We are optimistic that this development should provide a more robust model than previous models. Furthermore the multicenter aspect of this study implies that any equation derived from the study is more likely to be generally applicable to other populations. In this context it is reassuring that when our model was tested on the data from each individual center the overall test performance did not deteriorate.

The results of the internal validation from this study also compare favorably with the results of previous attempts on smaller datasets to validate mathematical models prospectively (Aslam et al, 2000; Mol et al, 2001, Valentin et al, 2001; Timmerman et al, 2001). Moreover, our model has significantly better test results on the test set cases than previously published models which had been developed on smaller datasets (Jacobs et al, 1990; Tailor et al, 1997; Timmerman, Bourne et al, 1999).

Whilst the area under the ROC curve reflects the overall test performance, the optimal probability level must be chosen to suit each individual clinical setting. There is always a compromise between the sensitivity (true positive rate) and the number of false positive test results. For example by selecting a probability level of 0.15, the sensitivity for malignancy was 92% and the specificity 81% on the test set.

In previous studies we showed that highly experienced operators using subjective impression as the basis to define malignancy gave a sensitivity and specificity of 96% and 90%, respectively (Timmerman, Schwärzler et al, 1999) or 88% and 96%, respectively (Valentin, 1999). For less experienced operators the corresponding values were 86% and 80% (Timmerman, Schwärzler et al, 1999). (see Figure 4.12) So, our model which reaches a sensitivity of 92.0% and a specificity of 81.0% can be considerably helpful for less experienced operators.

This IOTA data set also provides for each mass the subjective assessment by the expert of each center. These highly experienced operators reached a sensitivity of 86.7% and a specificity of 93.7% on the cases in the test data set. There is no significant difference in sensitivity between our model and the experts (p-value 0.25), but there is one in specificity (p-value<0.0001). The masses examined in the study represent a heterogeneous grouping of benign pathology, and a range of surgical stages for the malignancies. In the analysis we have looked at the data as a whole, but an examination of different subgroups of pathology may enable us to enhance test performance further, i.e. to improve specificity.

We have shown that with sufficient numbers of patients and variables a multimodal model can be built that will classify most tumors accurately prior to surgery.

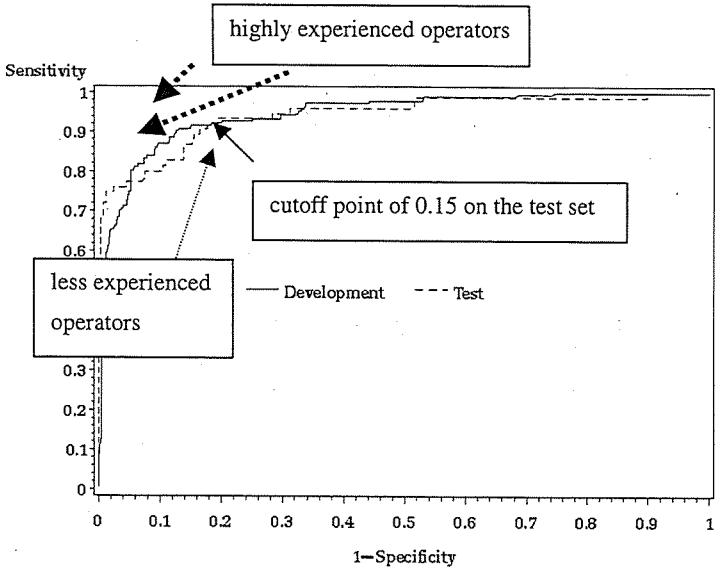


Figure 4-12 Our model reaches a sensitivity of 92% and a specificity of 81% at the probability cutoff point of 0.15 on the test set, which is still not as good as highly experienced operators: previous studies reported a sensitivity and specificity of 96% and 90%, respectively or 88% and 96% for experts. But the model can be of considerable help for less experienced operators, who reached a sensitivity and specificity of only 86% and 80% in a previous study.

4.11 Is the serum tumor marker CA 125 needed?

Now, we give an answer to this important question in gynecology: "Do measurements of serum CA 125 help to distinguish between benign and malignant adnexal tumors"?

Measurements of serum levels of CA 125 have been used as part of the preoperative evaluation of ovarian masses. In 1990 Jacobs et al assessed prior risk of a mass being malignant using menopausal status and serum CA 125 levels to produce a risk of malignancy index or RMI (Jacobs et al, 1990). In an ovarian cancer screening program the detection rate for early cancer has been shown to be significantly lower using fixed cutoff levels of

CA 125 in comparison to ultrasonography (Bourne et al, 1994). These limitations of CA 125 in relation to the detection of early stage cancer suggest that its role may be most compromised in the tumors that we find most difficult to either initially detect or subsequently classify.

In order to solve the question about the use of CA 125, we have first built a model on all the explanatory variables, except CA 125, which resulted in our 'final' overall model. Now, we can verify whether the inclusion of CA 125 improves the model performance, in other words, whether the inclusion of CA 125 to our model significantly improves the predictive performance on the test data.

All centers were encouraged to measure the level of serum CA 125 in peripheral blood from all patients, but the availability of this biochemical end-point was not an essential requirement for recruitment into the study. The immunoradiometric assay CA 125 II (Centocor, Malvern, PA or Cis-Bio, Gif-sur-Yvette, Cedex, France or Abbott Axsym system, REF 3B41-22, Abbott Laboratories Diagnostic Division, Abbott Park, IL 60064 USA) was used and the results expressed in U/ml.

The marker CA 125 has 258 'real' missings (258/1066 = 24%). The percentage of nonmissing values of the serum tumor marker is related to the centra (see Table 4-17). Certain centers, e.g. Milano, London and Monza have CA 125 measurements for all patients, whereas Paris Maurepas has CA 125 measurements in only 4 records (6.3%). The proportion of CA 125 values in the total dataset is 76% (76% in the development set and 76% in the test set).

Table 4-17 Proportion of CA 125 measurements in the different centers.

Center	# of patients in the total dataset	# of patients with a value for CA 125	proportion
Malmö	315	256	81.3 %
Leuven	263	205	77.9 %
Rome	126	68	54.0 %
Milano	87	87	100 %
Paris B	80	61	76.3 %
Paris M	64	4	6.3 %
London	54	54	100 %
Monza	46	46	100 %
Naples	31	28	90.3 %
Total	1066	809	76%

There is also a relation between the measurement of CA 125 values and the outcome (see Table 4-18). There are 20% more measurements of the serum tumor marker in malignant tumors (91.0%) than in benign masses (70.9%). So, when the expert suspects a benign tumor he/she is less tended to do the measurement.

Table 4-18 Proportion of patients with a value for CA 125 in the outcome. CA 125 is most frequently measured in borderline tumors.

Outcome	# of patients	# of patients with a value for CA 125	proportion
Benign	800	567	70.9%
Borderline	55	52	94.5%
Primary invasive	169	152	89.9%
Metastatic	42	38	90.5%
Total	1066	809	76%

Now, we can check if the inclusion of CA 125 to our model significantly improves the predictive performance on the test data or not.

4.11.1 Comparison of models with and without CA 125 on all 809 patients

If we add CA 125 to our overall model, and apply backward selection on the development set, we have to omit the variables ***loc5*** (p-value 0.1558) and ***Pain*** (p-value 0.0837). The other ten variables remain significant if CA 125 is included. The tumor marker itself has a p-value of 0.0320. The area under the ROC curve of this model reaches a value of 0.953 on the development set, but drops to 0.937 on the test set.

The application of our 'final' model (without CA 125) to the 237 cases with CA 125 in the test data results in an area under the ROC curve of 0.936.

The area under the ROC curves of the former model with CA 125 (i.e. 0.937) and our overall model without CA 125 (i.e. 0.936) are not statistically significant from each other (p-value 0.88, DeLong). So, we can conclude that the inclusion of the tumor marker CA 125 to our model does not improve the predictive performance on the test data.

In clinical practice, different cutoff levels are used for pre- and postmenopausal patients: a cutoff of 35 U/ml in premenopausal patients and 65 U/ml in postmenopausal patients (Timmerman, 2004). Therefore we will stratify our analysis by the menopausal status in order to maximize the potential for CA 125 to improve the diagnosis.

4.11.2 The impact of menopausal status

Premenopausal women

If we add CA 125 to the overall model for all premenopausal cases in the development set, and apply backward selection on the development

set, we have to omit the variables ***WallRegularityM***, ***Age***, ***Shadows***, and ***loc5***. In this reduced model CA 125 has a p-value of 0.2904.

Conclusion: CA 125 does not add any significant information in the classification of adnexal masses in premenopausal patients.

Postmenopausal women

If we add CA 125 to the overall model for all postmenopausal patients in the development set, we have to drop the variables ***PersHistOvCa***, ***Shadows***, ***Pain***, ***Ascites***, ***Age*** and ***m*** (***MaxSolidD123*** with cutoff at 50 mm) after applying backward selection of the variables. The resulting p-value for CA 125 is 0.0013. If we apply this model to all postmenopausal patients in the test set, we obtain an area under the ROC curve of 0.955.

The application of our overall model, which was developed on the whole training set, to all postmenopausal cases in the test set results in an area under the ROC curve of 0.949.

The performance of both models, i.e. 0.955 versus 0.949, is not statistically different from each other (p-value, 0.7269).

4.11.3 Constructing new models with CA125

Instead of adding CA 125 to our constructed model, we can also build completely new models with CA125. It can be that some variables that were not included in our 'final' model become important in the presence of the serum tumor marker CA125 (Timmerman, Van Calster, 2005).

If we do so on the whole development set, we obtain just the same model as in section 4.9.1.

If we stratify by menopausal status, and search for a model where CA125 is included, we obtain the following results. In the premenopausal patients in the development set, eleven variables are retained. But this model reaches an area under the ROC curve of only 0.894 on the premenopausal cases in the test set, while our 'final' model reaches an area of 0.906. So, there is no improvement of adding CA 125 to a diagnostic test for premenopausal women. On the other hand, in postmenopausal patients, we can build a model with an area under the ROC curve of 0.967 on the test set. But this is not significantly better than the application of our 'final' model (p-value 0.0574, DeLong).

4.11.4 Discussion

A wealth of publications in the literature concludes that measurements of serum CA 125 levels are a necessary part of the assessment of an adnexal mass. Therefore the findings of this large multicenter study are remarkable. We have tried to analyze the data in order to maximize the potential for CA 125 to improve the diagnosis. According to our data CA 125 does not enhance our ability to characterize an ovarian mass in premenopausal patients, and even in postmenopausal patients no significant benefit could be demonstrated.

We believe that there are several reasons why we have found CA 125 to be of little value in our study population. Previous models have always used a relatively small number of input variables. For example in Tailor et al (1997) there were 10 variables tested and only 4 retained, while Timmerman and colleagues tested 20 variables and retained only 5 (Timmerman et al, 1999). In this IOTA study 55 variables are tested and 12 significant variables are retained in the 'final' model. We were able to do this because of the large number of masses in the study. To our knowledge this is the first time where clinical variables such as a personal history of ovarian cancer, current hormone therapy, age and the presence of pain during the ultrasound scan were included in the model building. If these clinical variables are removed from the model, CA 125 becomes an independently significant variable. This demonstrates that this clinical information is more informative than CA 125 when evaluating the prior risk of an adnexal mass being benign or malignant.

Several reports in the literature confirm the significant number of false positive test results associated with the use of CA 125. Causes of this include the common pelvic pathologies of endometriosis, fibroids and infection as well as physiological events such as menstruation and pregnancy. This is important as a false positive result in these circumstances may lead to unnecessarily aggressive surgery and increased treatment costs. We also know that serum CA 125 is raised in less than 50% of stage I and borderline ovarian tumors (Vergote, Børmer et al, 1987; Cuckle et al, 1991; Bourne, Campbell et al, 1994; Maggino et al, 1994). False negative results may lead to unexpected findings of ovarian cancers in laparoscopically removed adnexal tumors, which may worsen the prognosis (Vergote et al, 2001). This is particularly problematic as some of these masses are amongst the most difficult to characterize accurately with ultrasound, and where increased diagnostic information would be helpful. In the future using combinations of tumor markers or serum proteomic pattern analysis may be of value. We also acknowledge that the variety of CA 125 kits used in our study might introduce more variation than if all CA 125 levels had been measured using the same kit.

Our analyses suggest that in the presence of a competent ultrasound unit and with knowledge of certain clinical information about the patient, there is no indication to measure serum CA 125 as part of the characterization of an adnexal mass.

4.12 Conclusion

The IOTA (International Ovarian Tumor Analysis) group minimizes the limitations of previous studies on the pre-operative classification of ovarian masses. The data sets used in those previous reports were not large enough for model development or evaluation because of differences in the prevalence of different types of ovarian malignancies and rare benign tumors. There were also variations in the interpretation of ultrasound derived end-points. The IOTA data set contains more than 1000 patients with a persistent adnexal mass at nine different centers working with a common protocol. The multicenter approach, where there is a difference in the prevalence of malignancy between the centers, was chosen as the most likely method to achieve a model which generalizes the best in prospective studies with different clinic populations. Over 50 defined variables were recorded on a database for the development of a logistic regression model.

A model with 12 variables reaches a high sensitivity and specificity for the pre-operative classification of ovarian tumors, i.e. 92% and 81%. When the model was tested on the data from each individual center the overall test performance did not deteriorate, which reassures us for the prospective testing of the model in new centers.

Independent risk factors for the presence of malignancy are: the patient's age, a personal history of ovarian cancer, the maximum diameter of the lesion, the maximum diameter of the solid component, the presence of ascites, the presence of blood flow within a solid papillary projection, the presence of a purely solid lesion, the presence of irregular internal cyst walls and an increase in color score. Factors that reduce the risk of malignancy include: the current use of hormonal therapy, the presence of pain during the ultrasound examination, and the presence of acoustic shadows.

In the past several studies have been performed to build models for the preoperative classification of ovarian tumors (Alcazar et al, 2004; Jacobs et al, 1990; Lerner et al, 1994), but none was able to reach a sensitivity of more than 90% and a specificity of more than 80% on an independent test set.

Previous studies show that highly experienced operators reached a sensitivity and specificity of 96% and 90%, respectively (Timmerman, Schwärzler et al, 1999) or 88% and 96%, respectively (Valentin, 1999). For

less experienced operators the corresponding values are 86% and 80% (Timmerman, Schwärzler et al, 1999). So, our model which reaches a sensitivity of 92.0% and a specificity of 81.0% could be of considerable help to less experienced operators. But it is not possible to beat an expert operator, since there is a significant difference in specificity.

One of the conclusions of previous studies – which suffered from small patient groups, and a small number of explanatory variables - was that measurements of serum CA 125 levels are a necessary part of the assessment of an adnexal mass. But from our large multicenter study, we can conclude that CA 125 does not enhance our ability to characterize an ovarian mass in premenopausal patients, not even in postmenopausal patients.

A previous thesis (Chuan Lu, 2005) dealt with the same dataset. Chuan used exactly the same training and test set as we did and came to the conclusion that no significant differences could be found between nonlinear (e.g. LS-SVM) models and our logistic regression model.

In the analysis we have looked at the data as a whole, but an examination of different subgroups of pathology may enable us to further enhance test performance, i.e. to improve specificity.

Chapter V

IOTA Subset analysis

In this chapter we examine whether there are subgroups with a big difference between the performance of the developed overall model (constructed in the previous chapter) and the subjective assessment of highly experienced operators in order to predict ovarian cancer. We develop new models for those subgroups where improvements can be made and combine all the different submodels in order to investigate whether we reach the high level of performance of the experts.

5.1 Introduction

In the previous chapter we developed a model for the pre-operative classification of ovarian tumors based on 12 variables. This overall model reached a high sensitivity of 92% and a specificity of 81% on an independent test set. This can be of considerable help to less experienced clinicians: a previous study (Timmerman, Schwärzler et al, 1999) showed that those operators obtained a sensitivity of only 86% and a specificity of 80%.

But improvements can be made. The IOTA data set provides for each mass the subjective assessment by the expert of each center. Each expert judged pre-operatively all the masses of his center based on ultrasonographic variables. For each tumor, he expressed pre-operatively his opinion on the benignness or malignancy of the mass, and gave a level of certainty to his diagnosis: "benign", "probably benign", "uncertain", "probably malignant", "malignant". These highly experienced operators reached a sensitivity of 86.7% and a specificity of 93.7% on the cases in the test data set. There is no significant difference in sensitivity between our overall model and the experts (92% versus 86.7%, p-value 0.25), but there is one in specificity (81% versus 93.7%, p-value<0.0001).

Table 5-1 shows the comparison - stratified by each center – between the expert's diagnosis and the overall model in terms of sensitivity and specificity for the development set of 754 patients. In all centers except one (Rome), the sensitivity of the overall model is higher or at least the same to the sensitivity of the subjective assessment of the expert from the center. But the drawback of this very good result is that in all centers the specificity of the overall model is lower than the specificity of the expert's diagnosis (exception: Milano, where the same specificity is reached). On the development set there is almost a significant difference in sensitivity between the overall model and the highly experienced operators, i.e. 91.6% versus 88.5% results in a p-value of 0.0611 (Tango, 1998). The subjective assessment of the experts yields a specificity which is more than 10% higher than the specificity of our constructed model: 95.9% versus 82.8%, p-value <0.0001. This large difference in specificity causes the difference in accuracy: subjective assessment 94.0% versus overall model 85.0%. This difference is significant (p-value <0.0001).

Table 5-1 Comparison between the subjective assessment of the expert and the overall model in terms of sensitivity and specificity on the development set.

		Subjective assessment by the expert		Overall model cutoff level 0.15	
	n	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Malmö	221	39/48 = 81	166/173 = 96	44/48 = 92	132/173 = 76
Leuven	185	63/66 = 95	115/119 = 97	65/66 = 98	100/119 = 84
Rome	89	27/32 = 84	52/57 = 91	25/32 = 78	51/57 = 89
Milano	62	6/6 = 100	55/56 = 98	6/6 = 100	55/56 = 98
Paris B	57	5/7 = 71	47/50 = 94	6/7 = 86	41/50 = 82
Paris M	45	4/5 = 80	37/40 = 93	4/5 = 80	30/40 = 75
London	39	11/12 = 92	27/27 = 100	11/12 = 92	24/27 = 89
Monza	33	11/12 = 92	21/21 = 100	11/12 = 92	18/21 = 86
Naples	23	3/3 = 100	20/20 = 100	3/3 = 100	15/20 = 75
Total	754	169/191 = 88.5	540/563 = 95.9	175/191 = 91.6	466/563 = 82.8

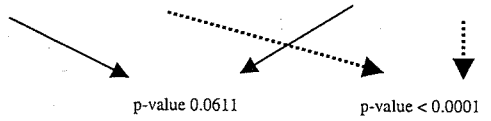


Table 5-2 shows the comparison between the subjective assessment of the expert and the overall model in terms of sensitivity and specificity on the test set, also stratified by each center. In all centers the sensitivity of the overall model is higher than the expert's diagnosis. The exception to this rule is Milano, but the values of sensitivity in that center are based on only two cases. For the specificity the conclusion is just the opposite: the specificity of the overall model is in most centers smaller than the specificity of the experts. The difference between the overall model and the subjective assessment of the doctors is not significant in sensitivity (92.0% versus 86.7%, p-

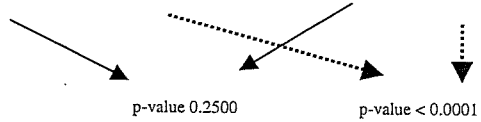
value 0.25), but is in specificity (81.0% versus 93.7%, p-value <0.0001). The difference in accuracy is also significant. The expert's diagnosis reaches an accuracy of 92.0%, while this is only 83.7% for the overall model.

So, our overall model is slightly more sensitive (although not significantly) than the opinion of the expert, but has a remarkable lower specificity: more than 10% difference on both training and test set. Therefore, the aim of further analyses will be to increase the specificity, while keeping the good sensitivity we already obtained with our model.

In the previous chapter we have looked at the data as a whole, but an examination of different subgroups may enable us to enhance test set performance further, i.e. to improve the specificity. The next step can be to develop models for specific subsets, for example exclude masses that can be easily classified subjectively. New models can be constructed that concentrate on the characterization of more complex or difficult pathology.

Table 5-2 Comparison between the subjective assessment of the expert and the overall model in terms of sensitivity and specificity on the test set. Note that no sensitivity could be calculated in Naples, since there were from Naples only benign masses in the test set.

		Subjective assessment by the expert		Overall model cutoff level 0.15	
	n	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Malmö	94	17/20 = 85	69/74 = 93	19/20 = 95	61/74 = 82
Leuven	78	22/27 = 81	49/51 = 96	26/27 = 96	44/51 = 86
Rome	37	12/13 = 92	23/24 = 96	12/13 = 92	18/24 = 75
Milano	25	2/2 = 100	22/23 = 96	1/2 = 50	22/23 = 96
Paris B	23	2/2 = 100	21/21 = 100	2/2 = 100	17/21 = 81
Paris M	19	2/2 = 100	15/17 = 88	2/2 = 100	11/17 = 65
London	15	3/4 = 75	8/11 = 73	3/4 = 75	9/11 = 82
Monza	13	5/5 = 100	7/8 = 88	4/5 = 80	5/8 = 63
Naples	8	-	8/8 = 100	-	5/8 = 63
Total	312	65/75 = 86.7	222/237 = 93.7	69/75 = 92.0	192/237 = 81.0



5.2 Creation of subsets

The issue is now the construction of appropriate subsets in order to increase the specificity.

To construct these subsets, we will return to the issue of missing values which we have discussed in section 4.3. We have noted there that there were four levels of structural missings:

- 513 cases without a solid component, i.e. missings in ***SolidD1***, ***SolidD2***, ***SolidD3***, ***MaxSolidD123***, ***VolumeSolid*** and ***RatioSolidLes***,
- 789 cases without papillary projections, i.e. missings in ***PapHeight***, ***PapBase1***, ***PapBase2***, ***MaxPap123***, ***VolumePap***, ***RatioPapLes***, ***PapNr***, ***PapSmooth*** and ***PapFlow***,
- 314 cases without measurements of the blood flow indices, i.e. missings in ***PI***, ***RI***, ***PSV*** and ***TAMXV***,
- 692 cases without values of ***PostMenoyears*** and ***PMB***. Most of them are premenopausal.

We noted in section 4.3 that we could expect the discriminatory performance for an overall model - where most structural missings are imputed with value "0" - to be lower than for a model which is the combination of all possible specific subgroups. There were 12 possible subgroups in total: premenopausal patients without a solid component and blood flow measurements, postmenopausal patients without solid component and blood flow measurements, and so on. Instead of constructing a model for each of those 12 specific subgroups, we will select some important subgroups of patients based on those variables which turned out to be important for the preoperative classification problem by the overall model.

Our overall model only used variables from the first two levels, i.e. ***MaxSolidD123*** with a cutoff at 50mm and ***PapFlow***. The structural missings in ***MaxSolidD123*** (if there was no solid component) and ***PapFlow*** (if there were no papillary projections) have been replaced by zero.

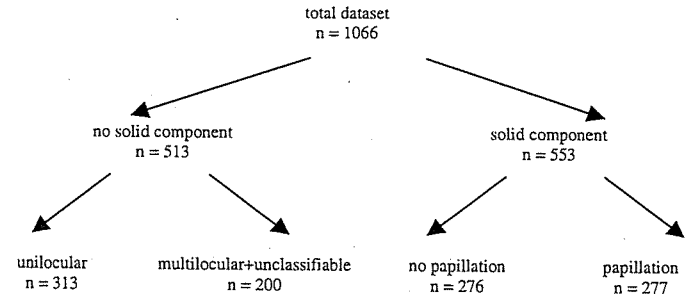


Figure 5-1 Construction of four subsets based on the absence of missing values in MaxSolidD213 and PapFlow, which are two important variables of the overall model. Note 1 unclassifiable case with solid component

Therefore, we will construct subsets based on missingness of those two variables (see Figure 5-1):

- no solid component
- solid component, but no papillary projections
- solid component and papillary projections.

The subset of no solid component contains 513 cases of which 313 are unilocular cysts. From those 313 only two tumors proved to be malignant, which makes the characteristic of unilocularity almost deterministic for being a benign tumor. Therefore, we will subdivide the group of no solid component in two groups: unilocular or multilocular (+unclassifiable), see Figure 5-1.

The purpose is to develop models on those four subgroups with comparable sensitivity as the overall model, but with an increased specificity, since this is significantly different between the overall model and the subjective assessment.

Ideally we should select randomly stratified (on outcome and center) a training set and a test set on those four subgroups, so that we can build models on the training sets and use the respective test sets to have an idea of the prospective performance on new sets of records. But in doing so, we would not be able anymore to compare the four models with the overall model applied on those subgroups of patients. For example, if we would randomly stratified select a development (70%) and a test set (the remaining 30%) on the 277 patients with papillary projections, we have the problem that some of the cases in this new test set have been used to develop the overall model, and some of the cases in the new training set are in the test set of the overall model. If we randomly stratified select training and test sets on the four subgroups there will be overlap between the new training and test sets and the original training and test set of the overall model. Now, there are two ways to continue:

- We can select new randomly stratified training and test sets (with the same proportion of malignancy) and do not compare with the overall model, or
- we can use the training and test set of the overall model stratified by the four subgroups that we consider. This means, for example, that in the original training and test set we select those cases with papillary projections to have a new training and test set on the subgroup of papillary projections. In this way, we can still compare newly developed models with the performance of the overall model on those subgroups. There are a few drawbacks of this method:

  - The training and test set do not always satisfy the proportions 70% and 30% of the data (see Table 5-3). But in fact, this is only a minor drawback, since there is no real golden rule for it. In literature, sometimes the proportions 50% - 50% or 2/3 – 1/3 or used as well.

Table 5-3 Proportion of training and test set for each subgroup

	Training	Test	Total
Unilocular	228 (72.8%)	85 (27.2%)	313
Multilocular + unclassifiable	143 (71.5%)	57 (28.5%)	200
Solid component, no papillary projections	183 (66.3%)	93 (33.7%)	276
Papillary projections	200 (72.2%)	77 (27.8%)	277
Total	754 (70.7%)	312 (29.3%)	1066

  - Another drawback of the second method is the nonequal proportion of malignancy in the training and test set for each subgroup (see Figure 5-2). However, this is more similar to the reality of external validation. This is especially the case in the group of records with a solid component but no papillation. The proportion of malignancy in the training set is 8% higher than in the test set (proportion of 47.3% versus 39.8%). Note that for three subgroups the proportion of malignancy is each time lower in the test set than in the training set:
    i. multilocular + unclassifiable: 10.4% vs. 8.8%,
    ii. solid component, no papillary projections: 47.3% vs. 39.8%
    iii. papillary projections: 44.5% vs. 41.6%.

    The reason why the proportion of malignancy in the development and test set of the overall model are nevertheless the same is because the test set contains relatively fewer unilocular cysts, which are nearly always benign. In the training set 228/754 = 30% is unilocular, while in the test set this is 85/312 = 27%.

Since it is important to be able to compare sensitivity and specificity of newly developed models with the overall model we will follow the second method of selecting training and test sets based on those of the overall model.

For each of the four subsets, we will try to keep the good sensitivity of the overall model and want to increase the specificity to ±90% on the whole dataset. In this way, the level of specificity becomes comparable to that of the subjective assessment of the expert. For each subset, we will cal-

culate the sensitivity and specificity of the assessment of the expert and compare this to the performance of the overall model.

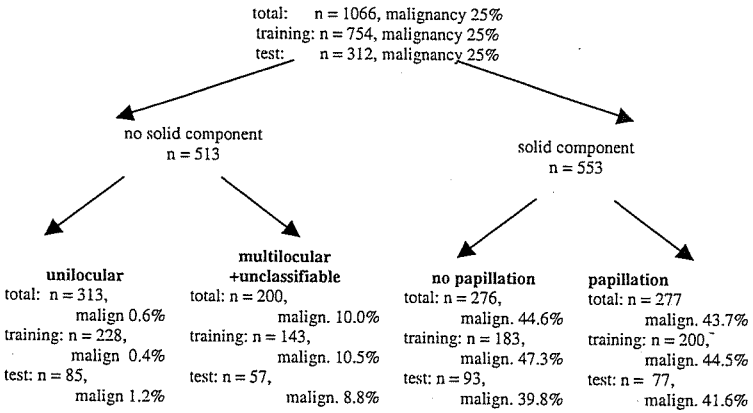


Figure 5-2 Proportion of malignancy in each subset

We will calculate the performance of the subjective assessment and the overall model on both the training and test set for the specific subgroups. Note that for the subjective assessment of the expert, the expert has not been "trained" on the training set. He or she already has his/her experience in classifying tumors from the past. But it can be interesting to see if for example the test set contains more easy cases than the training set. This is possible because the proportion of malignancy is not the same for the respective training and test sets in the four subgroups (see Figure 5-2). The overall model is also not trained on the training set of the specific subgroup, but on the training set of the total data set. We will apply each time the overall model on the training set of each subgroup, in order to be able to compare the overall model with a newly developed model for a specific subset of patients.

5.3 Subgroup 1: Unilocular cysts

The first subgroup of unilocular cysts is usually easily classified subjectively. They are almost always benign. Previous studies showed that simple unilocular cysts have only a very small risk of being malignant (Goldstein et al, 1988; Obwegeser et al, 1993; Kroon & Andolf, 1995; Gerber et al, 1997; Bailey et al, 1998; Ekerhovd, 2001).

5.3.1 Subjective assessment

In the development set two benign masses are labeled "uncertain" by the expert and the only malignant tumor (a borderline tumor) was identified as "probably benign" (see Table 5-4). The sensitivity of the expert on the development set is 0/1=0% and specificity 226/227=99.6% (see Table 5-5).

In the test set also two benign masses are labeled "uncertain", and one borderline tumor is noted as benign (see Table 5-6). The sensitivity equals 0/1=0% and specificity 84/84=100% (see Table 5-7).

Table 5-4 Subjective probability of malignancy in the development set of unilocular cysts

		subjective probability of malignancy					
		benign	probably benign	uncertain	probably malignant	malignant	Total
outcome	benign	200	25	2	-	-	227
	borderline	-	1	-	-	-	1
	prim.inv.	-	-	-	-	-	-
	metastatic	-	-	-	-	-	-
	Total	200	26	2	-	-	228

Table 5-5 Subjective assessment of malignancy in the development set of unilocular cysts

		subjective assessment of malignancy		
		benign	malignant	Total
outcome	benign	226	1	227
	borderline	1	-	1
	prim.inv.	-	-	-
	metastatic	-	-	-
	Total	227	1	228

Table 5-6 Subjective probability of malignancy in the test set of unilocular cysts

		subjective probability of malignancy					
		benign	probably benign	uncertain	probably malignant	malignant	Total
outcome	benign	76	6	2	-	-	84
	borderline	1	-	-	-	-	1
	prim. inv.	-	-	-	-	-	-
	metastatic	-	-	-	-	-	-
	Total	77	6	2	-	-	85

Table 5-7 Subjective assessment of malignancy in the test set of unilocular cysts

		subjective assessment of malignancy		
		benign	malignant	Total
outcome	benign	84	-	84
	borderline	1	-	1
	prim.inv.	-	-	-
	metastatic	-	-	-
	Total	85	-	85

5.3.2 Overall model

We can calculate the sensitivity and specificity of the overall model at the probability cutoff level of 0.15 that was chosen in the previous chapter. The values for sensitivity and specificity in both development and test set are just the same as to those of the expert, i.e. 0/1=0% and 226/227=99.6% on the development set, and 0/1=0% and 84/84=100% on the test set (see Table 5-8 and Table 5-9).

Table 5-8 Performance of the overall model in the development set of unilocular cysts

		overall model		
		benign	malignant	Total
outcome	benign	226	1	227
	borderline	1	-	1
	prim.inv.	-	-	-
	metastatic	-	-	-
	Total	227	1	228

Table 5-9 Performance of the overall model in the test set of unilocular cysts

		overall model		
		benign	malignant	Total
outcome	benign	84	-	84
	borderline	1	-	1
	prim.inv.	-	-	-
	metastatic	-	-	-
	Total	85	-	85

5.3.3 Subset model

The subjective assessment of the expert and the overall model both reach an excellent specificity but are not able to detect one rare malignant unilocular tumor. The overall model has just the same performance of the expert, but uses twelve variables. However, a very easy rule could be: if unilocular, then benign. With this rule, we would have just the same sensitivity, and obtain a specificity of 100% on both training and test set.

5.3.4 Conclusions

Unilocular cysts have only a very small risk of being malignant. This is confirmed by a number of previous studies (Goldstein et al, 1988; Obwegeser et al, 1993; Kroon & Andolf, 1995; Gerber et al, 1997; Bailey et al, 1998; Ekerhovd, 2001).

Therefore we propose the simple rule: if unilocular, then benign.

5.4 Subgroup 2: Multilocular + unclassifiable

The second subgroup contains all multilocular tumors and four of the five tumors with unclassifiable locularity. The reason why not all five cases with unclassifiable locularity are in this subset is that there is one tumor which has a solid component (but no papillation). Therefore, this case is in the next subgroup (solid component, no papillary projections).

The development set contains 10.5% malignant tumors and the test set 8.8%.

5.4.1 Subjective assessment

In the development set five benign masses are labeled "uncertain" by the expert and one "probably malignant". Out of the 15 malignant tumors only 5 have a label of "malignant" or "probably malignant": two borderline and three primary invasive cases (see Table 5-10). The sensitivity of the expert is 6/15=40%, while the specificity is 126/128=98.4% (see Table 5-11).

In the test set a primary invasive tumor has the label "probably benign" and a borderline tumor was denoted as "uncertain" (see Table 5-12). Both were classified as benign by the expert. The sensivity is only 3/5=60%, while the specificity remains very high 51/52=98% (see Table 5-13).

Table 5-10 Subjective probability of malignancy in the development set of multilocular cysts

		subjective probability of malignancy					
		benign	probably benign	uncertain	probably malignant	malignant	Total
outcome	benign	91	31	5	1	-	128
	borderline	2	3	2	2	-	9
	prim. inv.	-	2	1	-	3	6
	metastatic	-	-	-	-	-	-
	Total	93	36	8	3	3	143

Table 5-11 Subjective assessment of malignancy in the development set of multilocular cysts

		subjective assessment of malignancy		
		benign	malignant	Total
outcome	benign	126	2	128
	borderline	7	2	9
	prim. inv.	2	4	6
	metastatic	-	-	-
	Total	135	8	143

Table 5-12 Subjective probability of malignancy in the test set of multilocular cysts

		subjective probability of malignancy					
		benign	probably benign	uncertain	probably malignant	malignant	Total
outcome	benign	37	11	3	1	-	52
	borderline	-	-	1	-	-	1
	prim. inv.	-	1	-	1	1	3
	metastatic	-	-	-	1	-	1
	Total	38	12	4	3	1	57

Table 5-13 Subjective assessment of malignancy in the test set of multilocular cysts

		subjective assessment of malignancy		
		benign	malignant	Total
outcome	benign	51	1	52
	borderline	1	0	1
	prim. inv.	1	2	3
	metastatic	0	1	1
	Total	53	4	57

5.4.2 Overall model

The sensitivity of the overall model on the development set is 6/15=40%, which is just the same low value as for the expert. On the other hand, the specificity is much higher: 123/129=95.3%. The values for sensitivity and specificity on the test set are 2/5=40% and 49/52=94.2% on the test set (see Table 5-14 and Table 5-15).

Table 5-14 Performance of the overall model in the development set of multilocular cysts

		overall model		
		benign	malignant	Total
outcome	benign	123	5	128
	borderline	6	3	9
	prim. inv.	3	3	6
	metastatic	-	-	-
	Total	132	11	143

Table 5-15 Performance of the overall model in the test set of multilocular cysts

		overall model		
		benign	malignant	Total
outcome	benign	49	3	52
	borderline	1	-	1
	prim. inv.	2	1	3
	metastatic	-	1	1
	Total	52	5	57

5.4.3 Subset model

The overall model performs quite similar to the diagnosis of the expert: sensitivity 2/5 versus 3/5 and specificity 49/52 versus 51/52 on the test set. We will investigate whether it is possible to develop a new, more simple (less variables) model with similar performance.

Variable selection on the development set results in only three factors: ***Age***, ***NrLocules*** and ***Ascites***. If we construct a generalized additive model with a spline smoother for ***Age*** and ***NrLocules***, and test the importance of the smooth for those two variables, the nonlinear effect is not significant: p-value of 0.6254 for ***Age*** and 0.3308 for ***NrLocules***. There is no problem with the overall goodness-of-fit on the development set (p-value 0.6136 for Hosmer and Lemeshow test).

Table 5-16 Analysis of Maximum Likelihood Estimates of the subset model for multilocular cysts

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	p-value
Intercept	1	-7.8208	1.5554	25.2822	<.0001
Age	1	0.0542	0.0198	7.4682	0.0063
NrLoculesM	1	0.6502	0.2254	8.3232	0.0039
Ascites	1	2.2131	1.0384	4.5419	0.0331

Table 5-17 Odds Ratio Estimates of the subset model for multilocular cysts

Effect	Point estimate	95% Wald confidence limits	
Age	1.056	1.015	1.098
NrLoculesM	1.916	1.232	2.980
Ascites	9.144	1.195	69.992

Table 5-16 shows us the parameter estimates of the logistic regression model developed on the training data, and Table 5-17 the odds ratios which are computed by exponentiating the parameter estimate for each explanatory variable. All three variables are risk factors for the presence of a malignant ovarian tumor. The interpretation of the subset model is as follows:

- The odds of the presence of a malignant ovarian tumor increase by 5.6% for each year increase in ***Age*** of the patient.
- If the patient has ***Ascites*** (fluid outside Pouch of Douglas), the odds of the presence of a malignant ovarian tumor are estimated to be 9 times larger than that in patients without the presence of ***Ascites***.
- For each one-unit increase in ***NrLocules***, the odds of malignancy double.

Table 5-18 Classification table for the development set of the subset model for multilocular cysts

Probability level	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.20	53.3	89.8	86.0
0.25	53.3	90.6	86.7
0.30	53.3	92.2	88.1
0.35	**53.3**	**94.5**	**90.2**
0.40	40.0	96.1	90.2

Table 5-19 Classification table for the test set of the subset model for multilocular cysts

Probability level	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.20	40.0	88.5	84.2
0.25	40.0	92.3	87.7
0.30	20.0	94.2	87.7
0.35	**20.0**	**96.2**	**89.5**
0.40	20.0	96.2	89.5

The subset model reaches an area under the ROC curve of 0.863 on the development set. Table 5-18 shows the classification table for this development set. The sensitivity remains for different cutoff probability values very low. But remember the sensitivity of the experts was also only 40%. Taking the cutoff point at 0.35 seems to be a good compromise: our sensitivity is larger than the value of the experts and we have a high value of 94.5% for the specificity. (Taking the cutoff lower in order to obtain a higher sensitivity is not a good idea, since the corresponding specificity drops quickly.)

On the test set, the area under the ROC curve measures 0.777; and the selected probability value of 0.35 gives a sensitivity of 20% and a specificity of 96.2% (see Table 5-19).

We can compare this subset model with the subjective assessment and the overall model in terms of performance:

- sensitivity
  - expert's opinion: 3/5
  - overall model: 2/5
  - subset model: 1/5
- specificity
  - expert's opinion: 51/52
  - overall model: 49/52
  - subset model: 50/52.

Note that the differences in sensitivity are not statistically significant, due to the very small number of malignant tumors in the test set, i.e. only 5 cases. Also the differences between the three (i.e. expert's opinion, overall model and subset model) in terms of specificity are not significant.

External validation or generalization of the subset model in patients from other centers can be applied in two ways:

- with a formula (see Table 5-16):

$$P(malignant) = \frac{\exp(-7.8208 + 0.0542Age + 0.6502NrLocules + 2.2131Ascites)}{1 + \exp(-7.8208 + 0.0542Age + 0.6502NrLocules + 2.2131Ascites)}$$

- graphically via a nomogram (see Figure 5-3). Masses with a total score larger than ±160 points (this corresponds to the 0.35 cutoff level) are denoted as malignant. For example a patient of 80 years old (±90 points), where the mass has two locules (±40 points), and without ascites (0 points) has a total score of 90+40 = 130, which lies below the 0.35 cutoff level. So, the model predicts this tumor as benign. Similarly, a tumor of an 80 years old woman with two locules and with ascites (± 55 points) has a total score of 90+40+55=185 points > 160 points, and is identified as malignant.

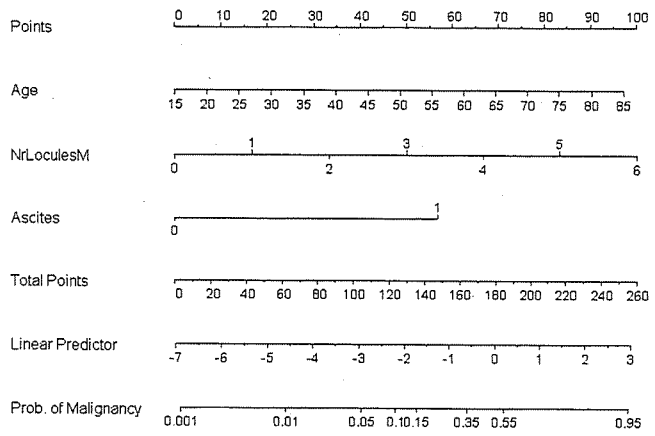


Figure 5-3 Nomogram of the subset model for multilocular cysts.

5.4.4 Conclusions

Our model with only three variables (*Age*, *NrLocules* and *Ascites*) obtains a sensitivity of 1/5 and a specificity of 50/52 on the test set which are quite similar to the overall model and the subjective assessment of the expert.

The advantage of applying this model to the multilocular cysts instead of the overall model is the small number of variables. The overall model contains 12 variables, while this model obtains the same performance with only three variables.

5.5 Subgroup 3 : Solid component, no papillary projections

In this subset, the proportion of malignant tumors is quite large, i.e. 47.3% in the development set and 39.8% in the test set. Note that the proportion of malignancy is lower in the test set.

5.5.1 Subjective Assessment

In the development set two benign masses are labeled "malignant" by the expert, while three borderline cases are denoted as "benign" and "probably benign" by the expert (see Table 5-20). The sensitivity of the subjective assessment of the expert is 82/86=95.3%, while the specificity is 92/97=94.8% (see Table 5-21). Note that the experts have misclassified one metastatic tumor as benign.

In the test set three benign tumors are identified as "probably malignant" by the expert. Three malignant tumors get the label of "benign" or "probably benign"(see Table 5-22). The sensivity of the expert's opinion is 33/37=89.2%, the specificity 52/56=92.9% (see Table 5-23). In the test set two primary invasive tumors and one metastatic tumor are misclassified as benign.

Note that misclassifying borderline tumors as benign, is clinically less relevant than misclassifying advanced ovarian cancer. While insufficient primary surgery of primary invasive and metastatic masses leads to poorer survival, since the amount of residual malignant tissue after primary surgery is one of the most important prognostic factors in ovarian cancer (Vergote et al, 2001), this is not the case for borderline tumors. Borderline tumors can not be considered as strictly malignant tumors: their clinical behavior, invasiveness and the subsequent treatment is actually very different from those of primary invasive and metastatic tumors, especially in younger patients who desire childbearing. Some types of borderline tumors can be considered more benign than malignant tumors.

Table 5-20 Subjective probability of malignancy in the development set of tumors with a solid component but no papillary projections

		subjective probability of malignancy					
		benign	probably benign	uncertain	probably malignant	malignant	Total
outcome	benign	61	24	10	-	2	97
	borderline	1	2	2	1	1	7
	prim. inv.	-	-	1	15	41	57
	metastatic	-	-	1	7	14	22
	Total	62	26	14	23	58	183

Table 5-21 Subjective assessment of malignancy in the development set of tumors with a solid component but no papillary projections

		subjective assessment of malignancy		
		benign	malignant	Total
outcome	benign	92	5	97
	borderline	3	4	7
	prim. inv.	-	57	57
	metastatic	1	21	22
	Total	96	87	183

Table 5-22 Subjective probability of malignancy in the test set of tumors with a solid component but no papillary projections

		subjective probability of malignancy					
		benign	probably benign	uncertain	probably malignant	malignant	Total
outcome	benign	34	15	4	3	-	56
	borderline	-	1	-	-	-	1
	prim. inv.	1	1	1	6	17	26
	metastatic	-	-	1	1	8	10
	Total	35	17	6	10	25	93

Table 5-23 Subjective assessment of malignancy in the test set of tumors with a solid component but no papillary projections

		subjective assessment of malignancy		
		benign	malignant	Total
outcome	benign	52	4	56
	borderline	1	-	1
	prim. inv.	2	24	26
	metastatic	1	9	10
	Total	56	37	93

5.5.2 Overall model

The sensitivity and specificity of the overall model, with cutoff probability level on 0.15, are 83/86=96.5% and 61/97=62.9% on the development set and 37/37=100% and 36/56=64.3% on the test set (see Table 5-24 and Table 5-25).

Note the perfect sensitivity of 100% in the test set. This means that the overall model has detected all malignant masses. This performance is better than the sensitivity of the expert, i.e. sensitivity of 89.2%. But the large disadvantage of this very high sensitivity is the much lower corresponding specificity of 64.3%, compared to the 92.9% specificity of the highly experienced doctors.

Table 5-24 Performance of the overall model in the development set of tumors with a solid component but no papillary projections

		overall model		
		benign	malignant	Total
outcome	benign	61	36	97
	borderline	2	5	7
	prim. inv.	-	57	57
	metastatic	1	21	22
	Total	64	119	183

Table 5-25 Performance of the overall model in the test set of tumors with a solid component but no papillary projections

		overall model		
		benign	malignant	Total
outcome	benign	36	20	56
	borderline	-	1	1
	prim. inv.	-	26	26
	metastatic	-	10	10
	Total	36	57	93

5.5.3 Subset model

We will investigate whether it is possible to build a model with a high sensitivity (such as the overall model has) and a good specificity of at least 75%.

Variable selection on the development set results in only four factors: ***Ascites***, ***WallRegularityM***, ***MaxSolidD123*** and ***Shadows***.

The generalized additive model suggests the need of a transformation for the factor ***MaxSolidD123*** (p-value 0.0137). Figure 5-4 shows us that the probability of having a malignant tumor increases when ***MaxSolidD123M*** increases until ±100mm and then decreases due to a few points. We propose to have a cutoff value at 100mm. This means that values above 100mm are replaced by 100mm. Thus the probability of malignancy can increase until ***MaxSolidD123*** reaches 100mm. After that point, the probability stays constant. We will denote the transformed variable by ***MaxSolidcutoff100***. Note that in the overall model we also have transformed the variable MaxSolidD123M, but there we decided to take the cutoff point at 50mm. It is clear from Figure 5-4 that this is not the best way to do if we are only looking at the subset of tumors with a solid component, but no papillary projections.

There is no problem with the overall goodness-of-fit (p-value 0.2626, Hosmer and Lemeshow test).

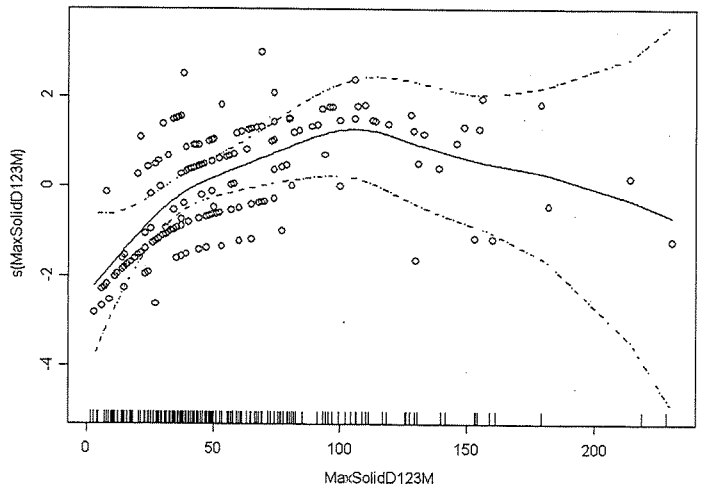


Figure 5-4 Spline smoother for the variable *MaxSolidD123M* in the generalized additive model for tumors with a solid component, but no papillary projections.

Table 5-26 Analysis of Maximum Likelihood Estimates of the subset logistic regression model for tumors with a solid component but no papillary projections

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.5339	0.4834	27.4730	<.0001
Ascites	1	1.7887	0.5325	11.2836	0.0008
WallRegularityM	1	1.8267	0.4092	19.9263	<.0001
maxsolidcutoff100	1	0.0267	0.00758	12.3561	0.0004
Shadows	1	-2.2584	0.8305	7.3940	0.0065

Table 5-27 Odds Ratio Estimates of the subset logistic regression model for tumors with a solid component but no papillary projections

Effect	Point estimate	95% Wald confidence limits	
Ascites	5.981	2.106	16.985
WallRegularityM	6.214	2.786	13.857
maxsolidcutoff100	1.027	1.012	1.042
Shadows	0.105	0.021	0.532

Table 5-26 shows us the parameter estimates of the logistic regression model developed on the training data, and Table 5-27 the odds ratios which are computed by exponentiating the parameter estimate for each explanatory variable. ***Ascites***, ***WallRegularityM*** and ***MaxSolidcutoff100*** are risk factors for the presence of a malignant ovarian tumor, while the presence of ***Shadows*** reduces the risk. The interpretation of the subset model is as follows:

- If the patient has ***Ascites*** (fluid outside Pouch of Douglas); the odds of the presence of a malignant ovarian tumor are estimated to be 6 times larger than those in patients without the presence of ***Ascites***.
- The odds of the presence of a malignant ovarian tumor increase by 2.7% for each one-unit (mm) increase in the value of ***MaxSolidD123***. Values of ***MaxSolidD123*** larger than 100mm have the same effect as a value of 100mm (variable denoted by ***maxsolidcutoff100***).
- If the lesion has an irregular internal wall (***WallRegularityM***), the odds of the presence of a malignant ovarian tumor are estimated to be 6 times larger than that in a lesion with a smooth internal wall.
- If acoustic ***Shadows*** are present on ultrasound, the odds of the presence of a malignant ovarian tumor are estimated to be 1/0.105=9.5 times smaller than that in case of the absence of acoustic shadows.

Table 5-28 Classification table for the development set of the subset model for tumors with a solid component but no papillary projections

Probability level	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.30	89.5	70.1	79.2
0.35	**89.5**	**76.3**	**82.5**
0.40	87.2	79.4	83.1

Table 5-29 Classification table for the test set of the subset model for tumors with a solid component but no papillary projections

Probability level	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.30	97.3	74.4	81.7
0.35	**97.3**	**78.6**	**86.0**
0.40	94.6	80.4	86.0

The area under the ROC curve is 0.895 on the development set. If we choose a cutoff probability value on the development set, we should fulfill the goal of having a model with ±90% sensitivity and ±75% specificity. Therefore, we choose the probability value of 0.35, which gives a sensitivity of 89.5% and a specificity of 76.3% (see Table 5-28). On the test set, the area under the ROC curve measures 0.943; and the selected probability value of 0.35 gives a sensitivity of 97.3% and a specificity of 78.6% (see Table 5-29). So, the sensitivity is higher in the test set than in the training set. This

was also the case for the overall model in contrast to the subjective assessment of the experts.

In order to externally validate the model in new patient data, we present two ways to calculate the predicted outcome:

- with a formula (see Table 5-26):

$$P(malignant) = \frac{\exp(-2.53 + 1.79Ascites + 1.83WallRegularityM + 0.0267MaxSolidcutoff100 - 2.26Shadows)}{1 + \exp(-2.53 + 1.79Ascites + 1.83WallRegularityM + 0.0267MaxSolidcutoff100 - 2.26Shadows)}$$

- or graphically via a nomogram (see Figure 5-5). The probability cut-off point of 0.35 corresponds to a total score of ±160 points. So, patients of which the score adds up to a value higher than ±160 will be denoted as malignant, the others as benign. For example, a patient without ***Ascites*** (0 points), where the lesion has an irregular internal wall (70 points), where ***MaxSolidD123***=120mm (100 points) and without ***Shadows*** (85 points) has a total score of 255 points and will thus be predicted as having a malignant tumor.

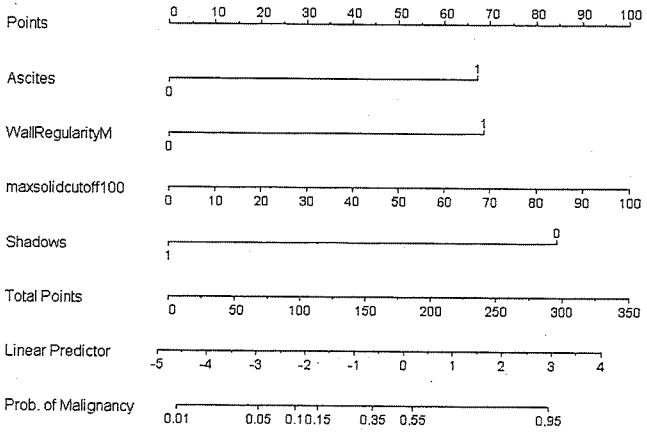


Figure 5-5 Nomogram of the subset model for tumors with a solid component but no papillary projections

5.5.4 Conclusions

The above subset model with only four variables (***Ascites***, ***WallRegularityM***, ***MaxSolidcutoff100*** and ***Shadows***) obtains a sensitivity of 36/37 = 97.3%, which is similar to the overall model (sensitivity of 100%), and a specificity of 78.6%, which is an improvement to the specificity of the overall model (specificity of 64.3%), but still lower than that of the subjective assessment (specificity of 92.9%).

5.6 Subgroup 4: Papillary projections

In this last subgroup the proportion of malignant tumors is the highest of all subgroups, i.e. 44.5% in the development set and 41.6% in the test set. Papillary projections are considered to be a strong sign of malignancy (Granberg et al, 1989), but they are also found in many benign pelvic masses, which is here the case, e.g. 55.5% of the masses with papillary projections are benign in the development set, 58.4% in the test set. Therefore, masses containing papillary projections are particularly difficult to classify as benign or malignant.

5.6.1 Subjective Assessment

In the development set five borderline tumors are classified as "benign" and "probably benign" by the expert. Ten benign masses are labeled "probably malignant" and "malignant". The sensitivity equals 81/89=91% and specificity 96/111 = 86.5% (see Table 5-30 and Table 5-31).

In the test set no malignant tumors are denoted as benign, nine have got the label of "uncertain" by the expert. In this test set the sensitivity of the expert reaches 29/32=90.6% and specificity 35/45=77.8% (see Table 5-32 and Table 5-33).

Table 5-30 Subjective probability of malignancy in the development set of tumors with papillary projections

		subjective probability of malignancy					
		benign	probably benign	uncertain	probably malignant	malignant	Total
outcome	benign	36	37	28	8	2	111
	borderline	1	4	3	4	11	23
	prim. inv.	-	-	1	18	39	58
	metastatic	-	-	2	1	5	8
	Total	37	41	34	31	57	200

Table 5-31 Subjective assessment of malignancy in the development set of tumors with papillary projections

		subjective assessment of malignancy		
		benign	malignant	Total
outcome	benign	96	15	111
	borderline	7	16	23
	prim. inv.	-	58	58
	metastatic	1	7	8
	Total	104	96	200

Table 5-32 Subjective probability of malignancy in the test set of tumors with papillary projections

		subjective probability of malignancy					
		benign	probably benign	uncertain	probably malignant	malignant	Total
outcome	benign	10	15	11	5	4	45
	borderline	-	-	5	6	1	12
	prim. inv.	-	-	4	3	12	19
	metastatic	-	-	-	1	-	1
	Total	10	15	20	15	17	77

Table 5-33 Subjective assessment of malignancy in the test set of tumors with papillary projections

		subjective assessment of malignancy		
		benign	malignant	Total
outcome	benign	35	10	45
	borderline	1	11	12
	prim. inv.	2	17	19
	metastatic	-	1	1
	Total	38	39	77

5.6.2 Overall model

The sensitivity and specificity of the overall model (with cutoff 0.15) are 86/89=96.6% and 56/111=50.5% in the development set and 30/32=93.8% and 23/45=51.1% in the test set (see Table 5-34 and Table 5-35).

Note that the sensitivity of our model is each time higher than the expert's assessment: 96.6% versus 91% on the development set, and 93.8% versus 91% on the test set. But on both sets, the specificity of the overall model is more than 25% smaller.

Table 5-34 Performance of the overall model in the development set of tumors with papillary projections

		overall model		
		benign	malignant	Total
outcome	benign	56	55	111
	borderline	3	20	23
	prim. inv.	-	58	58
	metastatic	-	8	8
	Total	59	141	200

Table 5-35 Performance of the overall model in the test set of tumors with papillary projections

		overall model		
		benign	malignant	Total
outcome	benign	23	22	45
	borderline	2	10	12
	prim. inv.	-	19	19
	metastatic	-	1	1
	Total	25	52	77

5.6.3 Subset model

Since the sensitivity of the subjective assessment and the overall model are quite comparable 90.6% versus 93.8%, but there is more than 25% difference in specificity 77.8% versus 51.1%, our goal will be to construct a model with sensitivity of ±90% and specificity of ±80%.

Variable selection retains the following five variables: ***PapFlow***, ***PapNr***, ***Age***, ***MaxSolidD123*** and ***RIM***. Note that ***RIM*** denotes the variable ***RI*** where the structural missings are imputed by "1".

Transformations are advisable for ***MaxSolidD123*** (p-value 0.0077) and ***RIM*** (p-value 0.0081). No transformations are needed for ***Age*** (p-value 0.2123) or ***PapNr*** (p-value 0.6530).

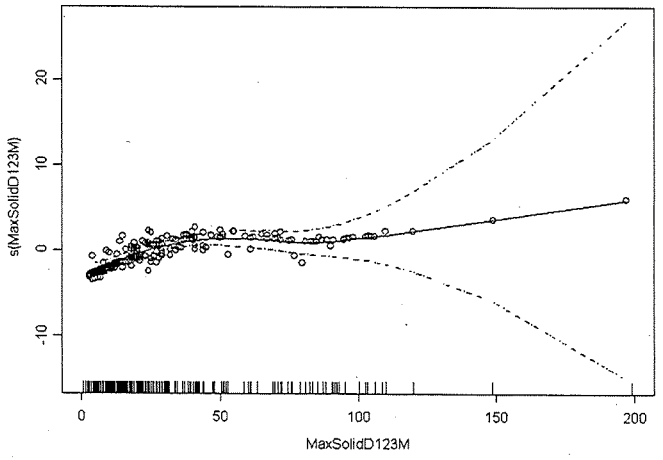


Figure 5-6A Plot of the spline smoother for MaxSolidD123 in the generalized additive model for tumors with papillary projections

The probability of having a malignant tumor increases till ***MaxSolidD123*** reaches ±50 mm (see Figure 5-6A) and stays almost constant afterwards. We will transform this variable by defining a cutoff point at 50mm: no increase possible after 50mm. For the variable ***RIM*** (see Figure 5-6B) the probability decreases if ***RIM*** increases but near to 1 the probability of having a malignant tumor seems to increase. Note that we have imputed the missing values of ***RI*** all by "1". We can verify whether this imputation causes the nonlinearity in the smoother. If we omit the missings in ***RI*** and check whether the variable ***RI*** should be transformed, we obtain a p-value of 0.3431. So, the nonlinearity in ***RIM*** is only caused by our imputation. There is no nonlinear relation between ***RI*** and the outcome. Therefore, we decide not to transform ***RIM***.

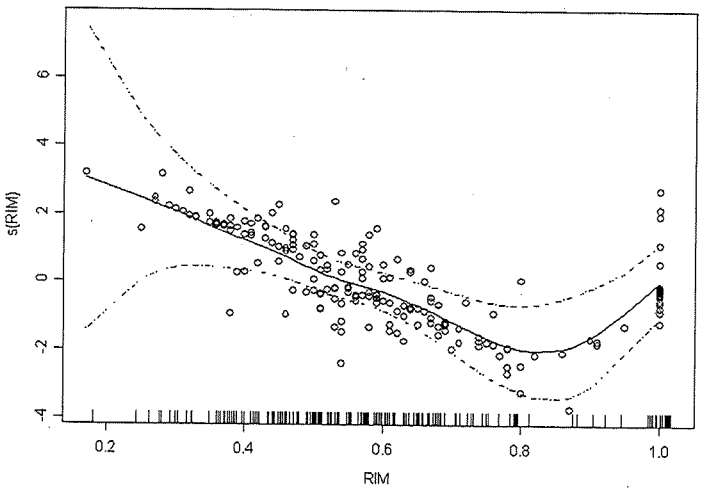


Figure 5-6B Plot of the spline smoother for RIM in the generalized additive model for tumors with papillary projections

There is no problem with the overall goodness-of-fit (p-value 0.8366, Hosmer and Lemeshow test).

Table 5-36 shows us the parameter estimates of the logistic regression model developed on the training data, and Table 5-37 the odds ratios which are computed by exponentiating the parameter estimate for each ex-

planatory variable. ***PapFlowM***, ***PapNrM***, ***Age*** and ***m*** (***MaxSoldidD123*** with cutoff at 50mm) are risk factors for the presence of a malignant ovarian tumor, while higher values of ***RIM*** reduce the risk.

Table 5-36 Analysis of Maximum Likelihood Estimates of the logistic regression model for tumors with papillary projections

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.3606	1.3795	9.9918	0.0016
PapFlowM	1	1.2969	0.4903	6.9949	0.0082
PapNrM	1	0.5926	0.1720	11.8695	0.0006
Age	1	0.0384	0.0163	5.5327	0.0187
m	1	0.0784	0.0154	25.8830	<.0001
RIM	1	-3.4151	1.3911	6.0269	0.0141

Table 5-37 Odds Ratio Estimates of the logistic regression model for tumors with papillary projections

Effect	Point estimate	95% Wald confidence limits	
PapFlowM	3.658	1.399	9.563
PapNrM	1.809	1.291	2.534
Age	1.039	1.006	1.073
m	1.082	1.049	1.115
RIM	0.033	0.002	0.502

The interpretation of the subset model is as follows:

- If there is presence of flow within the papillary structure (***PapFlowM***), the odds of presence of a malignant tumor are estimated to be 3.7 times larger than otherwise.
- For each one unit increase in ***PapNr***, the odds of the presence of malignant tumor is multiplied with a factor 1.8.
- The odds of the presence of a malignant ovarian tumor increase by 3.9% for each year increase in ***Age*** of the patient.
- The odds of the presence of a malignant ovarian tumor increase by 8.2% for each one-unit (mm) increase in the value of ***MaxSolidD123***. Values of ***MaxSolidD123*** larger than 50mm have the same effect as a value of 50mm (variable denoted by ***m***).
- If the ***RIM*** decreases with 10%, the odds of presence of a malignant tumor is multiplied by a factor of 1.4 (=exp((-3.4151)*(-0.1))).

Table 5-38 Classification table in development set of tumors with papillary projections

Probability level	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.25	95.5	75.7	84.5
0.30	**94.4**	**80.2**	**86.5**
0.35	91.0	82.0	86.0
0.40	87.6	84.7	87.5

The area under the ROC curve is 0.930 on the development set. A chosen probability value of 0.30 gives a sensitivity of 94.4% and a specificity of 80.2% (see Table 5-38). On the test set, the area under the ROC curve measures 0.902; and the selected probability value of 0.30 gives a sensitivity of 90.6% and a specificity of 80.0% (see Table 5-39).

Table 5-39 Classification table in test set of tumors with papillary projections

Probability level	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.25	90.6	73.3	80.5
0.30	**90.6**	**80.0**	**84.4**
0.35	81.3	80.0	80.5
0.40	81.3	82.2	81.8

There are two ways to apply this model to new data:

- with a formula (see Table 5-34):

$$P(\text{malignant}) = \frac{\exp(-4.4 + 1.3PapFlow + 0.6PapNr + 0.0384Age + 0.0784m - 3.41RIM)}{1 + \exp(-4.4 + 1.3PapFlow + 0.6PapNr + 0.0384Age + 0.0784m - 3.41RIM)}$$

- graphically via a nomogram (see Figure 5-3). Masses with a total score larger than ±150 points (this corresponds to the cutoff value of 0.30) are classified as malignant. For example a 45 years old patient (±30 points) with no ***PapFlow***, with three papillations (±30 points), with a ***MaxSolidD123*** of more than 50 mm (100 points) and a ***RI*** of 0.3 (± 60 points) has a total score of 220 points and is by consequence classified as having a malignant tumor.

5.6.4 Conclusions

The five variables (***PapFlow***, ***PapNr***, ***Age***, ***m*** and ***RIM***)model for masses with papillary projections reaches a sensitivity of 90.6% and a specificity of 80% which is very similar to the subjective assessment of the expert.

Note that although this model only uses five variables whereas the overall model uses twelve, we are able to improve the specificity with 30%, i.e. 80.2% versus 50.5% on the development set and 80.0% versus 51.1% on the test set, while keeping the sensitivity at approximately the same level: 94.4% versus 96.6% in the development set, and 90.6% versus 93.8% in the test set.

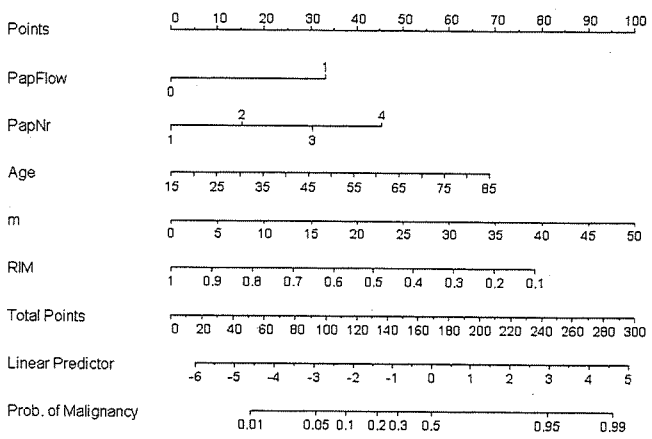


Figure 5-5 Nomogram of the logistic regression model for tumors with papillary projections.

5.7 Discussion

Now we have four subset models. The combination of those four models is presented in Figure 5-6. If we combine the models for the different subsets, we obtain in the test set

- a sensitivity of $\frac{0+1+36+29}{1+5+37+32} = \frac{66}{75} = 88\%$ and
- a specificity of $\frac{84+50+44+36}{84+52+56+45} = \frac{214}{237} = 90.3\%$.

The combination of the four models misclassifies only 3 malignant tumors more than the overall model – sensitivity decreases from 69/75 (92%) to 66/75 (88%) – but there is a remarkable gain in specificity compared to the overall model: from 81.0% to 90.3%. Of those 3 misclassified malignant tumors, there are 2 borderline tumors and 1 primary invasive mass. But note that borderline tumors can not be considered as strictly malignant tumors. Misclassifying such tumors, is clinically less relevant than misclassifying advanced ovarian cancer tumors (primary invasive and metastatic tumors). Borderline tumors can be considered more benign than malig-

nant tumors. So, in comparison to the overall model, we missed only 1 important malignant mass, i.e. a primary invasive case, while we gained almost 10% in specificity.

Another advantage of this combination of subset models is the smaller number of variables. The overall model uses 12 variables, while our combination model uses 10 factors:

- ***Locularity***: to determine in which subset the tumor is included,
- ***Age***,
- ***NrLocules***,
- ***Ascites***,
- ***WallRegularityM***,
- ***MaxSolidD123M***
  - using a cutoff at 100mm (MaxSolidD123cutoff100) for tumors with a solid component, but no papillation and
  - using a cutoff at 50mm (m) for tumors with papillary projections,
- ***Shadows***,
- ***PapFlow***
- ***PapNr***
- ***RIM*** (variable RI, with structural missings imputed by "1".

Finally, we compare our combination model to the subjective assessment of the experts on the test set. We have the following values:

- sensitivity: 88% (66/75) versus 86.7% (65/75), the difference is not statistically significant: p-value of 0.8576,
- specificity: 90.3% (214/237) versus 93.7% (222/237), the difference is not significant: p-value of 0.1167!

This means that we are successfully succeeded in constructing a model that is able to reach the excellent performance of highly trained experts!

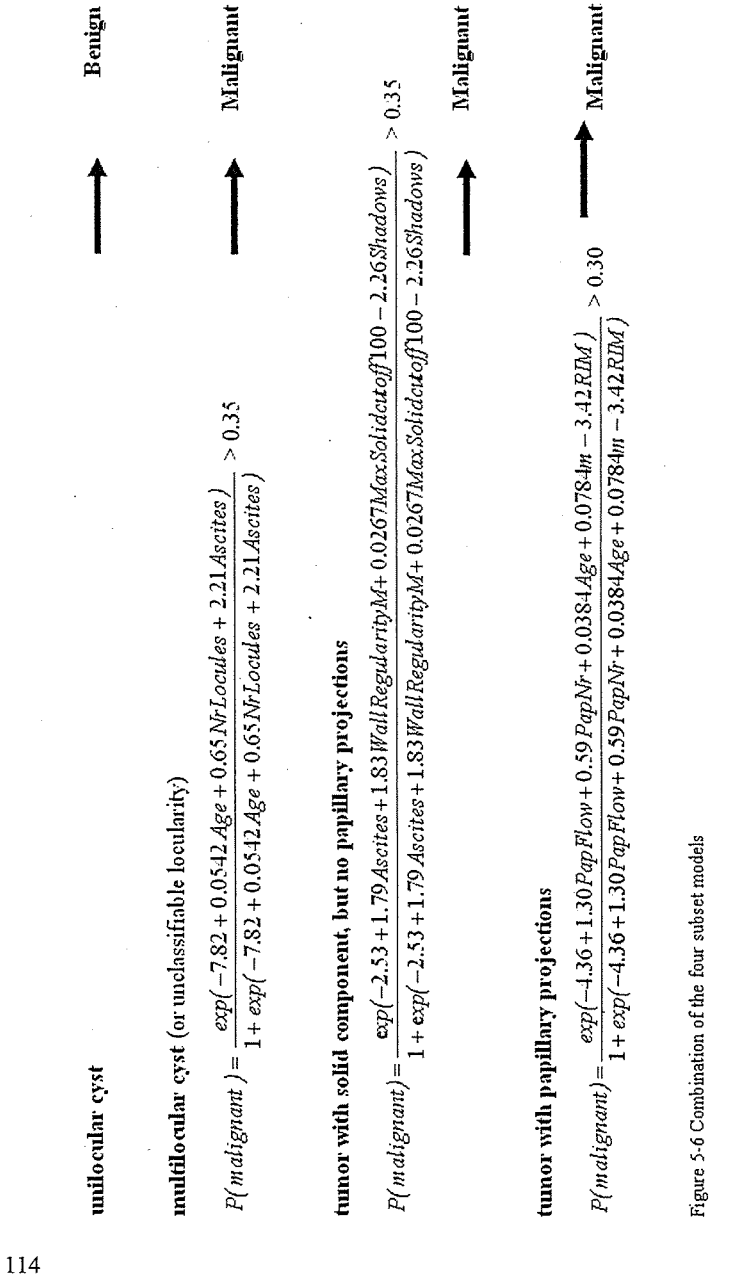


Figure 5-6 Combination of the four subset models

5.8 Conclusions

In the previous chapter we constructed the overall model (sensitivity 92%, specificity 81%) that could be of considerable help to less experienced clinicians. But this model does not reach the excellent performances of the experts (sensitivity of 86.7%, specificity of 93.7%). There is no significant difference in sensitivity, but the more than 10% difference in specificity is significant.

So, improvements could be made. Therefore, we decided to construct different subsets of patients and to develop a model for each subset. The subsets are:

- unilocular cysts,
- multilocular cysts,
- tumors with a solid component, but no papillary projections and
- tumors with papillary projections.

The combination model of those four subset models reaches an excellent performance, i.e. sensitivity of 88% and a specificity of 90.3%. Now, there is no longer a statistically significant difference with the expert's opinion. This means that we have reached our goal of building a model that achieves the performance of highly trained experts in order to pre-operatively classify ovarian tumors!

Chapter VI

Comparison of ultrasound characteristics of benign and malignant tumors

In this chapter we investigate if it is possible to construct a model that pre-operatively discriminates between benign and malignant tumors in those masses that are judged by the experts to be difficult to classify. We compare the ultrasound characteristics of those difficult masses with those that experts are able to classify. Then we focus on the different types of malignant ovarian tumors, i.e. borderline tumors, early and advanced ovarian cancer, rare tumors and metastases and examine their ultrasound characteristics.

6.1 Is it possible to beat the experts?

Using subjective evaluation of gray scale and Doppler ultrasound findings, an experienced ultrasound examiner using a good ultrasound system can correctly classify ovarian tumors as benign or malignant in most cases (Valentin, 1999; Timmerman, Schwärzler et al, 1999). In our dataset the experts reach a sensitivity of 88.0% and a specificity of 95.3% on the total dataset. However, in a small proportion of cases even a very experienced ultrasound examiner will find it difficult to discriminate between benignity and malignancy (Valentin, 1999). The experts gave those difficult cases the label "uncertain" in our dataset.

The aim is to determine which ovarian masses are judged to be difficult to correctly classify by the experts, and to investigate if the use of a logistic regression model for calculation of individual risk of malignancy can improve diagnostic accuracy in those difficult masses.

From now on, the difficult cases – which the experts labeled as "uncertain" – are called the unclassifiable tumors. The other ones - which the

experts labeled as "benign", "probably benign", "probably malignant" and "malignant" – are denoted as classifiable masses.

6.1.1 Characteristics of unclassifiable masses

In the entire dataset, there are 90 unclassifiable masses and 976 classifiable tumors. In the unclassifiable masses the sensitivity and specificity of the subjective assessment by the experts are 56% (14/25) and 77% (50/65). For the classifiable masses the corresponding figures are 91% (220/241) and 97% (712/735). Both the sensitivity and specificity are significantly lower for the difficult cases than for the classifiable masses (p-value <0.0001 for both).

Borderline tumors are over-represented among the unclassifiable masses, being three times more common among the unclassifiable masses than among the other ones (see Table 6-1). Of those borderline tumors, 47% (26/55) are correctly classified (i.e., classified as malignant) by the expert, 29% (16/55) are incorrectly classified (i.e., classified as benign), and 24% (13/55) are unclassifiable. Of the non-borderline tumors, i.e. benign, primary invasive and metastatic tumors, 90% (906/1011) are correctly classified, 3% (28/1011) are incorrectly classified, and 8% (77/1011) are unclassifiable. This difference between borderline and non-borderline tumors is statistically significant (P <0.0001).

Table 6-1 Histological diagnoses of unclassifiable and classifiable masses. The borderline tumors are significantly more frequently judged as difficult to classify by the experts.

	All n=1066	unclassifiable mass n=90	classifiable mass n=976	p-value
Benign	800 (75%)	65 (72%)	735 (75%)	0.5175
Borderline	55 (5%)	13 (14%)	42 (4%)	<0.0001
Primary invasive	169 (16%)	8 (9%)	161 (16%)	0.0691
Metastatic	42 (4%)	4 (4%)	38 (4%)	0.7751

Ultrasound findings of unclassifiable versus classifiable tumors are shown in Table 6-2. Papillary projections, >10 locules in a cyst without solid components, low level echogenicity of cyst fluid, and color score 3 (indicating moderate vascularization) are characteristic for those tumors that experts judge as difficult to classify. Logistic regression shows that the above four variables are the only ultrasound variables which are independently characteristic for unclassifiable masses (the other significant variables are related to the above four).

- If there is papillation, the odds of the presence of an unclassifiable mass are estimated to be 5.1 times larger than in cases without papillary projections (p-value <0.0001).
- The odds of the presence of an unclassifiable mass are estimated to be 3.7 times larger in records with more than 10 locules in a cyst without solid components (p-value 0.0175).
- The odds of the presence of un unclassifiable mass are estimated to be 2.5 times larger in cases with low level echogenicity of cyst fluid than in cases with another type of echogenicity (p-value 0.0002).
- If color score = 3, the odds of the presence of an unclassifiable mass are estimated to be 1.8 times larger than for cases with color score = 1, 2 or 4 (p-value 0.0109).

In the IOTA dataset, 64% (35/55) of the borderline tumors contain papillary projections vs. 24% (242/1011) of the non-borderline tumors (p-value < 0.0001), 11% (6/55) vs. 2% (19/1011) were multilocular cysts with >10 locules (p-value 0.012), and 33% (18/55) vs.19% (191/1011) contained cyst fluid with low-level echogenicity (p-value 0.023).

Table 6-2 Ultrasound characteristics of the classifiable and unclassifiable masses. The median and range is presented for the continuous variables, percentages are presented for binary variables.

	All n=1066	unclassifiable mass n=90	classifiable mass n=976	p-value
Volume lesion, ml	93 (0.1 – 11829)	149 (0.1 – 7781)	91 (0.2 – 11829)	0.0199
Bilateral	216 (20%)	18 (20%)	198 (20%)	0.9486
Ascites	135 (13%)	11 (12%)	124 (13%)	0.8952
Type of mass				<0.0001
Unilocular	313 (29%)	4 (4%)	309 (32%)	<0.0001
Unilocular solid	132 (12%)	18 (20%)	114 (12%)	0.0219
Multilocular	196 (18%)	12 (13%)	184 (19%)	0.1959
Multilocular with >10 locules	25 (3%)	5 (6%)	20 (2%)	0.0526
Multilocular solid	284 (27%)	41 (46%)	243 (25%)	<0.0001
Solid	136 (13%)	15 (17%)	121 (12%)	0.2454
Not classifiable	5 (0.5%)		5 (0.5%)	0.9999
Number of locules	1 (0 – 12)	2.5 (0 – >10)	1 (0 – >10)	0.0092
Pain	282 (26%)	19 (21%)	263 (27%)	0.2619
Echogenicity				<0.0001
Anechoic	410 (38%)	21 (23%)	389 (40%)	0.0021
Low level	209 (20%)	36 (40%)	173 (18%)	<0.0001
Ground glass	225 (21%)	14 (16%)	211 (22%)	0.1774
Haemorrhagic	10 (1%)	-	10 (1%)	0.9999
Mixed	132 (12%)	9 (10%)	123 (13%)	0.4732
No cyst fluid	80 (8%)	10 (11%)	70 (7%)	0.1747

Table 6-2 Continued. Ultrasound characteristics of the classifiable and unclassifiable masses. The median and range is presented for the continuous variables, percentages are presented for binary variables.

	All n=1066	unclassifiable mass n=90	classifiable mass n=976	p-value
Papillation	277 (26%)	54 (60%)	223 (23%)	<0.0001
† Irregular papillation	177 (64%)	33 (61%)	144 (65%)	0.6347
† Flow in papillation	154 (56%)	26 (48%)	128 (57%)	0.2210
† Number of papillations	2 (1 - 4)	1.5 (1 - 4)	2 (1 - 4)	0.1888
† Volume of largest papillation, ml	0.5 (0.006 – 226)	0.3 (0.008 – 87)	0.5 (0.006 – 226)	0.1576
† Ratio papillation volume/lesion volume;	0.004 (0.000–0.456)	0.004 (0.000–0.105)	0.004 (0.000 – 0.456)	0.1342
† Height of papillation	10 (2 – 62)	8.5 (3 – 36)	10 (2 – 62)	0.1553
Mass with solid components	552 (52%)	74 (82%)	478 (49%)	<0.0001
††Volume of largest solid component, ml	8 (0.006 – 2291)	4 (0.008 – 2291)	10 (0.006 – 2091)	0.0039
††Ratio volume largest solid component / volume of lesion	0.084 (0.000 – 1)	0.014 (0.000 – 1)	0.106 (0.000 – 1)	0.0031
Thickness of thickest septum, mm	0 (0 - 20)	1.95 (0 – 8)	0 (0 – 20)	0.0060
Incomplete septum	85 (8%)	5 (6%)	80 (8%)	0.3794
Irregular walls	479 (45%)	61 (68%)	418 (43%)	< 0.0001
Shadows	108 (10%)	9 (10%)	99 (10%)	0.9658
Doppler findings				
Color score 1	233 (22%)	8 (9%)	225 (23%)	0.0019
Color score 2	353 (33%)	34 (38%)	319 (33%)	0.3259
Color score 3	328 (31%)	42 (47%)	286 (29%)	0.0006
Color score 4	152 (14%)	6 (7%)	146 (15%)	0.0313
Venous flow only	81 (8%)	4 (4%)	77 (8%)	0.3010
† † †Pulsatility index	0.89 (0.13 – 5.80)	0.81 (0.34 – 5.80)	0.89 (0.13 – 5.09)	0.7624
† † †Resistance index	0.57 (0.12 – 1.00)	0.55 (0.25 – 1.00)	0.57 (0.12 – 1.00)	0.5609
† † †Peak systolic velocity, cm/s	14.0 (2 – 202)	16.4 (3 – 77)	14.0 (2 – 202)	0.4053
† † †Time averaged maximum velocity, cm/s;	8.9 (1 – 137)	9.4 (1.4 – 60)	8.7 (1 – 137)	0.6767

† percentages and p-values calculated in the subgroup of masses with papillary projections
†† percentages and p-values calculated in the subgroup of masses with solid components
† † † Pulsatility index, resistance index and peak systolic velocity were measured in 752 patients, in 78 with an unclassifiable mass and in 674 with a classifiable mass; time averaged maximum velocity was measured in 739 patients, in 78 with an unclassifiable mass and in 661 with an classifiable mass.

Now we are ready to construct a predictive model to discriminate between benign and malignant tumors that are judged by the highly experienced ultrasound examiners as difficult to classify.

6.1.2 Is it possible to predict ovarian cancer in the difficult tumors?

We start with selecting a training and a test set. We will follow the same procedure as we applied for selecting training and test sets for the subset models. In the original training and test set of the overall model we selected those cases with ***SubjProb*** = 3 (labeled "uncertain" by the experts) in order to have a new training and test set on the subgroup of unclassifiable masses. The training set contains 58 cases and the test set 32 masses. The prevalence of malignancy is larger in the test set, i.e. 38% (12/32) versus only 22% (13/58) in the training set. Note that this difference in proportion of malignancy can cause difficulties in comparing the performance of a model on training and test set. Nevertheless, we will continue.

If we start with constructing an overall model on the training set (where structural missings are replaced by zero) no explanatory variable is selected by logistic regression, not even the serum tumor marker CA 125.

We try to select an important subset in the unclassifiable masses, where it might be possible to build a model. In the previous section, we could conclude that papillary projections are typical for unclassifiable masses. Therefore, we decide to construct a model for the 54 unclassifiable masses with papillary projections. Consequently, the training set is reduced to 34 cases and the test set to 20 cases. The prevalence of malignancy is now only 18% (6/34) in the training set versus 45% (9/20) in the test set. There are approximately three times more malignant tumors in the test set than in the training set.

Variable selection on the training set results in only one factor: ***PapHeight***. And this variable is just borderline significant (p-value 0.0517 – see Table 6-3). Also here, the serum tumor marker CA 125 does not add any significant information to the model.

Table 6-3 Analysis of Maximum Likelihood Estimated of the logistic regression model for 'difficult' tumors with papillary projections

	Estimate	Standard error	p-value
Intercept	-3.6595	1.2386	0.0031
Papillary height (mm)	0.2011	0.1034	0.0517

Table 6-4 Odds ratio estimates of the logistic regression model for 'difficult' tumors with papillary projections. The height of the papillation and the number of papillations increase the risk on a malignant tumor.

Effect	Point estimate	95% Wald confidence limits	
Papillary height (mm)	1.223	0.999	1.497

For each one-mm increase in the height of the largest papillary projection the odds of malignancy increase 1.22 times (see Table 6-4).

The generalized additive model suggests no transformation is needed. There is also no problem with the overall goodness-of-fit (p-value, 0.2289, Hosmer and Lemeshow test).

The area under the ROC curve is only 0.72 on the development set, and reduces even to 0.586 on the test set. So, we could conclude that we are not able to build a reliable predictive model using logistic regression in order to classify the difficult cases. But note the unequal proportion of malignancy in training and test set. To verify, if this bad performance is not due to the large difference of prevalence of malignancy in training and test set, we decide to choose new training and test set, which are randomly stratified by the outcome, and to repeat the whole analysis.

The randomly selected training and test set contains 65 and 25 difficult cases. In both sets the proportion of 28% malignancy is verified. Again, no explanatory variable is selected when replacing all the structural missings by a zero. If we focus on the tumors with papillation, the training set is reduced to 39 cases and the test set to 15 cases. The former has 28% and the latter 27% malignant masses.

Table 6-5 shows the parameter estimates of the logistic regression model developed on the 34 cases of the training set. Only three explanatory variables are retained. Also here, the marker CA 125 did not add any significant information to the model. Table 6-3 shows the parameter estimates of the logistic regression model developed on the 34 cases of the training set. Only three explanatory variables are retained. Also here, the marker CA 125 did not add any significant information to the model.

Table 6-5 Analysis of Maximum Likelihood Estimated of the logistic regression model for 'difficult' tumors with papillary projections

	Estimate	Standard error	p-value
Intercept	-3.6294	1.3886	0.0090
Papillary height (mm)	0.2052	0.0908	0.0239
Number of pappilations	0.9826	0.4852	0.0429
Thickness of septum (mm)	-0.6111	0.2504	0.0147

Table 6-6 Odds ratio estimates of the logistic regression model for 'difficult' tumors with papillary projections. The height of the papillation and the number of papillations increase the risk on a malignant tumor.

Effect	Point estimate	95% Wald confidence limits	
Papillary height (mm)	1.228	1.027	1.467
Number of pappilations	2.671	1.032	6.915
Thickness of septum (mm)	0.543	0.332	0.887

The interpretation of the model is as follows:

- for each one-unit increase in height of the largest papillary projection the odds of malignancy increase 1.23 times,
- for each one-unit increase in the number of papillary projections the odds increase 2.67 times, and
- for each one-unit increase in thickness of the thickest septum the odds decrease 0.54 times.

In the training set the area under the receiver operator characteristic (ROC) curve for the logistic regression model is 0.88. A probability cutoff value of 0.10 on the training set yields a sensitivity with regard to malignancy of 82% (9/11) and a specificity of 50% (14/28); a cutoff of 0.50 is associated with a sensitivity of 54% (6/11) and a specificity of 86% (24/28).

If we apply this model to the test set, the area under the ROC curve reduces to 0.73. So, our model does not perform well on new data. Here, at the cutoff value of 0.10 the sensitivity is 75% (3/4) and specificity 18% (2/11). Using the cutoff probability value of 0.50, the sensitivity equals 75% (3/4) and specificity 54% (6/11).

We can compare these values to the performance of the experts on difficult masses with papillary projections. They reached a sensitivity and specificity of 64% (7/11) and 82% (23/28) on the training set, and 50% (2/4) and 82% (9/11) in the test set.

So, we can conclude that we can not yet build predictive models that substantially improve diagnostic accuracy in tumors that are judged to be difficult to classify by the experts.

6.2 Ultrasound characteristics of malignant ovarian tumors.

For optimal planning of treatment it is clinically useful to be able to distinguish not only between benignity and malignancy but also between different types of malignancy, for instance between ovarian borderline tumors, early and advanced ovarian cancer, and metastases. In this section we focus on the ultrasound characteristics of the different types of malignant ovarian tumors, i.e. borderline ovarian tumors, primary invasive epithelial ovarian cancer stage 1, primary invasive epithelial ovarian cancer stage 2 - 4, rare types of pelvic malignancy, and metastatic tumors.

Table 6-7 Ultrasound characteristics of the different types of malignant ovarian tumors. The median and range is presented for the continuous variables, percentages are presented for binary variables.

	Borderline	Epithelial cancer Stage 1	Epithelial cancer Stage 2-4	Rare	Metastatic
n	55 (21%)	42 (16%)	102 (38%)	25 (9%)	42 (16%)
Largest diameter, mm	108 (8-300)	126 (24-410)	85 (24-245)	109 (16-250)	82 (27-224)
Volume, ml	375 (0.1-7163)	695 (6-11829)	209 (6-2541)	507 (1-4020)	163 (6-2790)
Fluid in the pouch of Douglas	15 (27%)	18 (43%)	73 (72%)	16 (64%)	22 (52%)
All cases, mm	0 (0-32)	0 (0-75)	18 (0-100)	15 (0-60)	7 (0-63)
When Fluid > 0 (Yes), mm	14 (5-32)	23 (3-75)	26 (4-100)	21 (12-60)	26.5 (5-63)
Ascites	5 (9%)	13 (31%)	62 (61%)	14 (56%)	18 (43%)
Locularity					
Unilocular	2 (4%)	-	-	-	-
Unilocular solid with papillary projections	13 (24%)	12 (29%)	8 (8%)	-	2 (5%)
Unilocular solid without papillary projections	1 (2%)	3 (7%)	3 (3%)	2 (8%)	-
Multilocular	10 (18%)	6 (14%)	2 (2%)	1 (4%)	1 (2%)
Multilocular solid	26 (47%)	18 (43%)	50 (49%)	8 (32%)	14 (33%)
Solid	3 (5%)	3 (7%)	39 (38%)	14 (56%)	25 (60%)
Echogenicity of cyst fluid					
Anechoic	19 (35%)	12 (29%)	53 (52%)	8 (32%)	15 (36%)
Low level	18 (33%)	16 (38%)	12 (12%)	7 (28%)	7 (17%)
Ground glass	13 (24%)	10 (24%)	4 (4%)	2 (8%)	4 (10%)
Hemorrhagic	-	-	1 (1%)	-	1 (2%)
Mixed	3 (5%)	4 (10%)	7 (7%)	2 (8%)	2 (5%)
No cyst fluid	2 (4%)	0	25 (25%)	6 (24%)	13 (31%)
Papillary projection present;	35 (64%)	28 (67%)	42 (41%)	7 (28%)	9 (21%)
Number of papillations*	>3 (1->3)	>3 (1->3)	>3 (1->3)	>3 (1->3)	>3 (1->3)
Papilla irregular*	29 (53%)	22 (52%)	34 (33%)	7 (28%)	8 (19%)
Irregular wall	40 (73%)	36 (86%)	90 (88%)	19 (76%)	32 (76%)
Shadowing	1 (2%)	-	1 (1%)	-	2 (5%)
Thickness of septum if present	3.9 (1.4-15)	4 (1-20)	5 (1-20)	8 (2.6-15)	4.4 (1.4-12)
Volume of largest papilla*, ml	2 (0.008-226)	4 (0.02-123)	2 (0.01-90)	1 (0.02-90)	2 (0.2-12)
Volume of largest solid component**;ml	4 (0.008-412)	46 (0.9-478)	33 (0.5-1202)	206 (3-2291)	57 (0.3-1702)
Proportion of tumor being solid;	0.02 (0.0001-1)	0.10 (0.001-1)	0.34 (0.001-1)	1 (0.023-1)	1 (0.006-1)
Doppler results					
Color score 1	4 (7%)	4 (10%)	1 (1%)	1 (4%)	1 (2%)
Color score 2	17 (31%)	6 (14%)	12 (12%)	2 (8%)	5 (12%)
Color score 3	21 (38%)	19 (45%)	38 (37%)	9 (36%)	22 (52%)
Color score 4	13 (24%)	13 (31%)	51 (50%)	13 (52%)	14 (33%)
No flow recorded	4 (7%)	4 (10%)	1 (1%)	1 (4%)	1 (2%)
Only venous flow recorded	3 (5%)	2 (5%)	3 (3%)	1 (4%)	-
Arterial flow recorded	48 (87%)	36 (86%)	98 (96%)	23 (92%)	41 (98%)
Flow in papillary projection*	25 (71%)	25 (89%)	37 (88%)	7 (100%)	8 (89%)
PI	0.74 (0.35-2.26)	0.63 (0.29-1.90)	0.82 (0.25-2.25)	0.64 (0.33-1.44)	0.76 (0.36-2.05)
RI	0.48 (0.30-0.87)	0.44 (0.26-0.78)	0.55 (0.17-1)	0.44 (0.27-0.73)	0.53 (0.20-0.85)
PSV, cm/s	20 (5-72)	26 (3.9-86)	25 (5-120)	30.(9-202)	23 (6-86)
TAMXV, cm/s	14 (3-52)	18 (3-87)	17 (3-81)	21 (8-137)	17 (5-56)

* calculated only for masses with papillary projections
* calculated only for masses with solid components

Table 6-8 Tests for the ultrasound characteristics between borderline, epithelial cancer stage1 and epithelial cancer stage 2-4. For all the variables we started with calculating the p-value for testing if there was any difference between the three groups. If this p-value was smaller than 0.05, we calculated the p-values for the pairwise comparisons. Using the Bonferroni correction for multiple comparisons, those p-values should be compared to 0.05/3=0.017. In case of categorical variables (Locularity, Echogenicity and Color score) the permutation method was applied in order to calculate the p-values for the different levels of the variable. Therefore, those p-values can also be compared to 0.017.

	Borderline versus Epithelial cancer Stage 1	Borderline versus Epithelial cancer Stage 2-4	Epithelial cancer Stage 1 versus Epithelial cancer Stage 2-4	Overall test
Largest diameter, mm	0.2440	0.0352	0.0007	0.0019
Volume, ml	0.1950	0.0213	0.0001	0.0004
Fluid in the pouch of Douglas	0.1084	<0.0001	0.0012	<0.0001
All cases, mm	0.0695	<0.0001	0.0006	<0.0001
When Fluid > 0 (Yes), mm	0.2696	0.0019	0.1336	0.0058
Ascites	0.0082	<0.0001	0.0017	<0.0001
Locularity	0.6617	<0.0001	<0.0001	<0.0001
Unilocular	-	0.3965	-	
Unilocular solid with papillary projections	-	0.0345	0.0058	
Unilocular solid without papillary projections	-	1	0.7964	
Multilocular	-	0.0017	0.0204	
Multilocular solid	-	0.9983	0.9283	
Solid	-	<0.0001	0.0005	
Echogenicity of cyst fluid	0.7398	<0.0001	<0.0001	<0.0001
Anechoic	-	0.1465	0.0608	
Low level	-	0.0084	0.0022	
Ground glass	-	0.0008	0.0018	
Hemorrhagic	-	1	1	
Mixed	-	1	0.9944	
No cyst fluid	-	0.0017	0.0003	
Papillary projection present	0.7566	0.0072	0.0054	0.0034
Number of papillary projections*	-	-	-	0.9420
Papilla irregular*	0.9730	0.0180	0.0331	0.0118
Irregular wall	0.1239	0.0140	0.6776	0.0165
Shadowing	-	-	-	1
Thickness of septum if present	-	-	-	0.4856
Volume of largest papilla*	-	-	-	0.2163
Volume of largest solid component**	0.0004	<0.0001	0.8940	<0.0001
Proportion of tumor being solid	0.0212	<0.0001	0.0003	<0.0001
Doppler results***	0.2995	<0.0001	0.0262	0.0010
Color score 1	-	0.1373		
Color score 2	-	0.0157		
Color score 3	-	1		
Color score 4	-	0.0045		
No flow recorded	0.7237	0.0513	0.0255	0.0192
Only venous flow recorded	-	-	-	0.6360
Arterial flow recorded	-	-	-	0.0552
Flow in papillary projection	-	-	-	0.0887
PI	0.1085	0.2442	0.0202	0.0450
RI	0.0599	0.1957	0.0078	0.0172
PSV, cm/s	-	-	-	0.1510
TAMXV, cm/s	-	-	-	0.1209

* calculated only for masses with papillary projections
* calculated only for masses with solid components

Table 6-7 shows descriptive statistics for the ultrasound characteristics of the five different types of malignant ovarian tumors. Table 6-8 shows the p-values when comparing pairwise borderline, epithelial cancer stage 1, and epithelial cancer stage 2-4 tumors.

The metastatic tumors tend to be small, solid, and well vascularized. The rare malignancies tend to be large, solid, and well vascularized. 55% (39/67) of the rare and metastatic tumors are purely solid vs. 23% (45/199) of borderline tumors and primary invasive ovarian epithelial cancers (p-value <0.0001).

Fluid in the pouch of Douglas or ascites are common features both of metastatic tumors and rare types of tumor. Borderline ovarian tumors and stage 1 primary invasive ovarian epithelial cancers have several ultrasound features in common that differ from those of stage 2 – 4 primary invasive ovarian epithelial cancers (all p-values can be found in Table 6-8):

- they are larger (p-value 0.0213, p-value 0.0001),
- a much greater proportion of them contains papillary projections (p-value 0.0072, p-value 0.0054),
- they are more often multilocular cysts (p-value 0.0017, p-value 0.0204),
- but they are less often purely solid tumors (p-value <0.0001, p-value 0.0005),
- their cyst fluid is more often characterized as having low level echogenicity (p-value 0.0084, p-value 0.0022) or ground glass appearance (p-value 0.0008, p-value 0.0018).

With increasing degree of invasiveness (from borderline epithelial ovarian tumors via stage 1 invasive epithelial ovarian tumors to stage 2 – 4 epithelial ovarian tumors) ascites becomes increasingly more common (9% vs. 31% vs. 61% of cases), and among tumors with solid components (n = 179) the proportion of tumor consisting of solid tissue increases (median 2% – 10% – 34%). However, the proportion of tumor consisting of solid tissue is highest in rare tumors (median proportion 100%) and in metastases (median proportion 100%) (see Table 6-7).

Multivariable logistic regression in the 55 borderline tumors and the 42 stage 1 ovarian epithelial cancers shows ascites (odds ratio 5 .991, 95% CI confidence interval 1.743 - 20.586), nulliparity (odds ratio 3.726, 95% CI 1.254 - 11.074), and age (odds ratio 1.056, 95% CI 1.019 - 1.095) to be the only independent predictors of invasiveness.

We can conclude that different types of malignant ovarian tumors manifest different characteristics at transvaginal gray scale and color Dop-

pler ultrasound examination. The use of these ultrasound characteristics to predict the type of malignancy needs to be tested prospectively.

6.3 Cutoff values for the continuous ultrasound variables and blood flow indices

For the practicing clinician, it is important to know at what point specific morphological markers become of importance to the clinical management of the patient. In this section, we will calculate 'optimal' cutoff values for the ultrasound and blood flow variables.

So far, we considered in the previous analyses the dominant ovarian tumors of 1066 patients. But in 165 patients, the tumors were bilateral. This means we have also data of an extra 165 'non dominant' tumors. And there were also 2 non dominant tumors of which the corresponding dominant mass was not included in the dataset. So, in total we have data of 1066 + 165 + 2 = 1233 adnexal tumors. Overall, 903 (73%) tumors are benign and 330 (27%) are malignant. Table 6-9 summarizes the differences in ultrasound morphology and blood flow indices between benign and malignant masses. All the variables show significant differences.

Table 6-9 Univariate analysis of the continuous ultrasound variables and blood flow indices in all (dominant + non dominant) masses. The statistics for *PapHeight* and *RatioPapLes* are calculated only for those masses with papillary projections. *MaxSolidD123* and *RatioSolidLes* are calculated for masses with a solid component and the blood flow indices are calculated for those tumors where a blood flow was present.

	Benign				Malignant			
	n	median	min	max	n	median	min	max
Ultrasound								
MaxLesD123*	903	59	8	320	330	85	8	410
Fluid* > 0	140	12	2	61	145	24	3	100
Septum* > 0	372	2.1	1	20	167	4	1	20
PapHeight*	161	7	2	62	139	14	3	62
***RatioPapLes**†*	161	0.003	0	0.456	139	0.006	0	0.420
MaxSolidD123*	334	21	3	230	303	46	4	214
RatioSolidLes*	334	0.035	0	1	303	0.265	0	1
Bloodflow								
PI*	556	0.95	0.13	7.30	300	0.76	0.23	3.31
RI*	556	0.59	0.12	1.0	300	0.52	0.17	1.0
PSV*	556	11.2	2.0	85.5	300	23.9	2.3	202
TAMXV*	546	6.6	1.0	60.0	293	16.0	1.7	137

* p<0.0001; † p=0.0011 (Mann-Whitney)

We calculate optimal cutoff values for the continuous variables using the minimum p-value approach of Mazumdar and Glassman (2000). Table 6-10 demonstrates those calculated cutoff points for the morphological and color Doppler variables.

Table 6-10 The optimal cutoff values for the ultrasound and blood flow variables using the minimum p-value approach. The performance measures (accuracy, sensitivity and specificity) are calculated by classifying cases with values ≥ the cutoff as malignant and cases with value < the cutoff point as benign. The exceptions to this rule are *PI* and *RI*: tumors with e.g. values for RI ≥ 0.5 are classified as benign.

	cutoff	accuracy (%)	sensitivity (%)	specificity (%)
Ultrasound				
MaxLesD123*	100 mm	73.9	43.0	85.2
Fluid* > 0	15 mm	66.9	75.0	58.6
Septum* > 0	3 mm	68.6	62.9	71.2
PapHeight*	10 mm	70.0	74.1	66.5
***RatioPapLes**†*	0.006	61.0	54.0	67.1
MaxSolidD123*	35 mm	70.8	69.6	71.9
RatioSolidLes*	0.020	64.7	85.8	45.5
Blood flow				
PI*	0.6	69.9	30.7	91.1
RI*	0.5	67.9	46.0	79.7
PSV*	15 cm/sec	68.9	74.0	66.2
TAMXV*	10 cm/sec	70.9	71.7	70.5

If we focus on the size of a lesion (*MaxLesD123*), we can see that the optimal cutoff value above which a mass is more likely to be malignant is when the maximum diameter is above 100 mm. However size is not an independent variable and irrespective of size other morphological features of a cyst must be considered. When we are only considering the variable MaxLesD123, we obtain a sensitivity of hardly 43%.

Our data show that the optimal cutoff height of a papillary projection (*PapHeight*) to define malignancy is 10 mm. However, the situation is not so simple. While we can statistically produce an optimal value for a variable to define the risk of malignancy, this may not be of practical help clinically. The real question is whether there are any indices whose presence or absence can categorically exclude or indicate the presence of carcinoma. For example, (see Figure 6-1) if we take as cutoff value 30 mm malignancy is highly likely. For tumors with *PapHeight* smaller than 30 mm, no conclusion can be drawn. So, by taking 30 mm as cutoff value we can almost exclude the presence of a benign tumor.

As another example, we look at the variable *Fluid*. With the calculated cutoff of 15 mm (see Table 6-10), we reach a sensitivity of 75%. However, clinically it can be more useful to use 50 mm as cutoff. Tumors with more than 50 mm of fluid have a high probability of being malignant (see Figure 6-2).

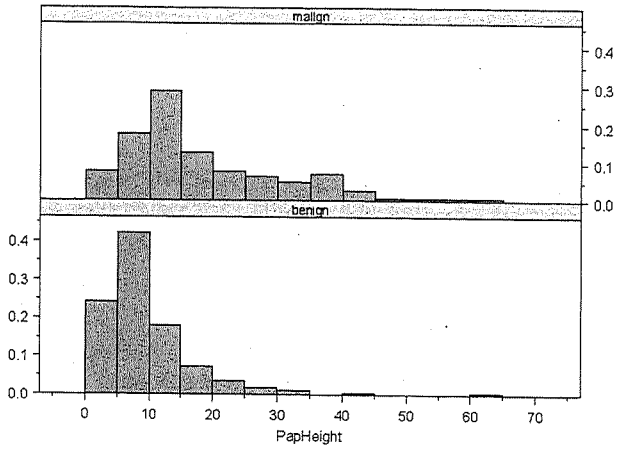


Figure 6-1 Histogram, stratified by the outcome, of the height of a papillary projection. Tumors with *PapHeight* ≥ 30 mm, are highly likely to be malignant masses.

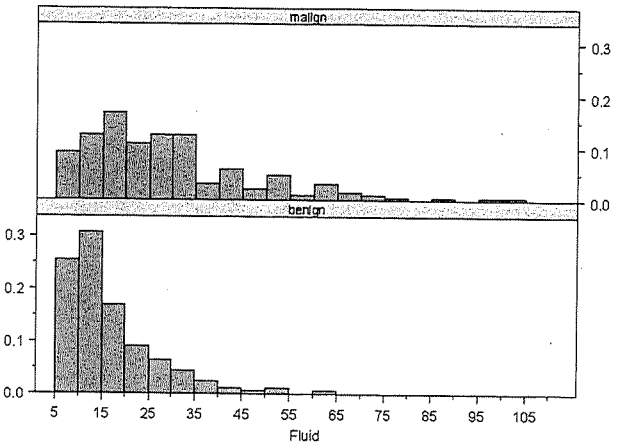


Figure 6-2Histogram, stratified by the outcome, of the diameter of any fluid seen in the pouch of Douglas. Tumors with *Fluid* ≥ 50 mm, are highly likely to be malignant masses.

So, 'optimal' cutoff points can be calculated, but the use of only one variable to classify ovarian tumors is limited. By considering cutoff points we are nevertheless able to exclude or indicate ovarian cancer in a number of cases.

6.4 Conclusions

Experts using high-end ultrasound systems find slightly less than 10% of the ovarian tumors difficult to classify as benign or malignant. Borderline tumors are difficult to classify as strictly benign or malignant. Masses with papillary projections, multilocular cysts with > 10 locules, cysts with low level echogenicity of cyst fluid and masses moderately vascularized at color Doppler ultrasound examination seem to be more difficult to classify than other types of tumors.

We found no logistic regression model that was useful to distinguish between benignity and malignancy in difficult masses. However, a logistic regression model built on unclassifiable masses with papillary projections suggested that the more papillary projections and the larger the papillary projections the greater the risk of malignancy.

It is important to be able to reliably discriminate between benign and malignant adnexal masses in order to be able to correctly evaluate the need for surgery and to choose appropriate time and mode of operation. As long as we have not found other explanatory variables capable of distinguishing between difficult benign and malignant ovarian tumors (e.g. proteomics or intravenous ultrasound contrast), we must accept that some women will need to undergo an unnecessary operation or an unnecessarily extensive operation because of our inability to reliably exclude malignancy before operation.

The possibility to discriminate preoperatively between ovarian borderline tumors, early and advanced ovarian cancer, and pelvic metastases may improve the possibility to provide optimal disease management.

Cutoff points can be calculated for the continuous ultrasound and color Doppler variables. This enables us to exclude or indicate malignancy in a number of cases.

Chapter VII

Predictive models for survival after premature rupture of membranes

We discuss the construction of different predictive models (using Cox's proportional hazards model, logistic regression, Least Squares Support Vector Machines and generalized estimating equations) for survival of premature babies of which the mother had a preterm prelabor rupture of the membranes in this chapter. For each developed model, we assess the performance in a prospective data set and discuss the clinical impact.

7.1 Introduction

Prelabor rupture of membranes (PROM) refers to membrane rupture before the onset of uterine contractions and occurs in approximately 5-10 % of all pregnancies (Mead, 1980). In 25 % of cases this occurs before 37 weeks of gestation causing 30 – 40 % of all preterm births and in 0.7 % it concerns a pregnancy of less than 26 weeks of gestation (Gibbs & Blanco, 1982; Graham et al, 1982; Cox et al, 1988; Taylor & Garite, 1984; Schuker & Mercer, 1996).

Here only preterm prelabor rupture of the membranes (PPROM) before 26 weeks of gestation is studied. The management of PPROM in the previable period is among the most controversial issues in perinatal medicine and has a major impact both on fetal and maternal health. The neonatal mortality rate after midtrimester PPROM is approximately 40 % (Taylor & Garite , 1984; Moretti & Sibai, 1988; Beydoun & Yasin, 1986; Major & Kitzmiller, 1990; Hadi et al, 1994; Fortunato et al, 1994; Morales & Talley, 1993; Rotschild et al, 1996; Kilbride et al, 1996).

Efforts to decrease the incidence of premature birth have involved numerous strategies, including maternal risk assessment, use of tocolytic agents, home uterine activity monitoring, biochemical markers and a decrease in maternal activity. The most important current therapeutic options are the administration of steroids, tocolytic agents and antibiotics.

The purpose of this study is to assess the outcome of preterm prelabor rupture of the membranes before 26 weeks of gestation and to develop a predictive model for survival.

The medical records of 119 consecutive pregnancies of PPROM at or before 26 completed weeks of gestation, including 148 babies, were retrospectively analyzed. All women were admitted with a living fetus at the university hospital Gasthuisberg Leuven between 1988 and 1996. The dataset contains 25 explanatory variables and the outcome is the number of days the baby survived. In case the baby died before he/she was born, the outcome specifies whether it was antenatal or intranatal. 7% of the 148 babies died antenatally, 20% intranatally, 24% on the first day, 6% from day 2 till day 7, 4 % from day 8 till day 28, 6% after the first month and 35% of the babies survived.

In 1999 a data analysis has already been performed on this dataset of 148 babies (Spitz et al, 1999). For that analysis the outcome was summarized into a 0 -1 variable: 0 if the baby died before the end of the first month, and 1 if the baby survived the first month. Corrections were made for the fact that members of twin or triplet pregnancies do have the same mother by using Generalized Estimating Equations (GEE) (Diggle et al, 2001). The variables of the obtained model determining the survival chances were: the birth weight of the baby, the sex of the baby, the gestational age at the time of rupture of the membranes, and a binary variable expressing whether or not steroids have been administered during the pregnancy.

In the period of 1997 until 2000 a prospective dataset has been constructed. The prospective dataset contains medical records of 64 consecutive cases of PPROM at or before 26 completed weeks of gestation. All women were admitted with a living fetus at the university hospital Gasthuisberg Leuven between 1997 and 2000.

The purpose is to develop several predictive models for survival based on the records in the period 1988 – 1996 and to test the performance of these predictive models on the prospective set (period 1997 – 2000). In addition we also compare the prospective performance of our models with the performance of the model in paper (Spitz et al, 1999). For the construction of predictive models for survival we use the following statistical methods: the classical Cox's Proportional Hazards model (Allison, 1995; Kalbfleisch et al, 1980; Marubini et al, 1997; Therneau et al, 2000), Logistic Regression (Agresti, 1996; Harrell, 2002; Hosmer et al, 2001), Generalized Estimating Equations (Diggle et al, 2001; Hosmer et al, 2001) and Least Squares – Support Vector Machines (LS-SVM) (Suykens et al, 1999-2002-2003). The first three ones are linear models, while the last one is a nonlinear

black box model. LS-SVM classifiers are used here to check the quality of Logistic Regression and Generalized Estimating Equations as linear models in relation to a nonlinear classifier.

While the model of 1999 (Spitz et al, 1999) could help at the moment of labor or birth, we now want to create a model that is based on variables known at the intake of the patient with PPROM and which is not dependent on neonatal variables. Such a model can help in counseling and choosing the best treatment modalities.

7.1.1 Description of the original dataset (1988 – 1996)

The medical records of 119 consecutive pregnancies of PPROM at or before 26 completed weeks of gestation, including 148 babies, were retrospectively analyzed. All women were admitted with a living fetus at the university hospital Gasthuisberg Leuven between 1988 and 1996. The diagnosis of PROM was clinically evident by the presence of a pool of amniotic fluid in the vagina, or confirmed by a positive nitrazine and ferning test. During the study period no substantial change in the approach of neonatal care occurred; every premature baby born after viability during the study period received maximal support.

We retrieved for each case 25 variables (see Table 7-1). The variables that are related to the mother, i.e. the variables ***TPROM*** until ***MODE_OF_DELIVERY*** are measured in the group of 119 pregnancies, while the variables that are related to the babies themselves, i.e. ***SEX***, ***A1***, ***A5*** and ***WEIGHT***, are measured in the group of 148 babies.

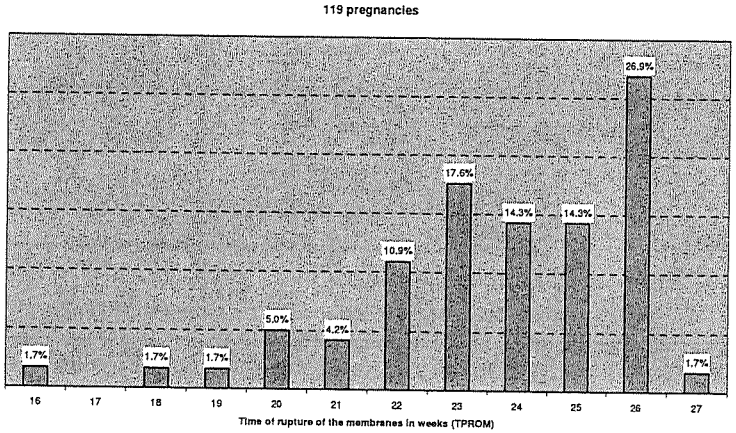


Figure 7-1 Gestational age at the time of rupture of the membranes

Table 7-1 Description and basic statistics of the explanatory variables

Variable	*Description of variable*
TPROM	*Gestational age at the time of PPROM (weeks)* mean ± SD : 23.7 ± 2.2 weeks, median 24, min 16 and max 27 weeks
PART	*Gestational age at the time of delivery (weeks)* mean ± SD : 26.2 ± 2.9 weeks, median 26, min 22 and max 36 weeks
LATENCY	*Time interval between PPROM and the delivery (days)* mean ± SD : 18.5 ± 25 days, median 8, min 1 and max 141 days
AGE	*Maternal age (years)* mean ± SD : 29 ± 5 years, median 28, min 19 and max 41 years
A	*Number of abortions* 0: 71%, 1: 22%, 2: 5%, 3: 2%
P	*Number of births* 0: 51%, 1: 31%, 2: 12%, 3: 3%, 4: 1%, 5: 2%
G	*Number of pregnancies* 1: 39%, 2: 28%, 3: 22%, 4: 7%, 5: 1%, 6: 2%, 7: 1%
SMOKE	0 = *non smoker* (80%), 1 = *smoker outside of pregnancy* (10%), 2 = *smoker during pregnancy* (10%)
CIGARETTE	*Average number of cigarettes smoked during pregnancy* (n = 12) mean ± SD : 12 ± 7 cigarettes, median 10, min 5 and max 30 cigarettes
F	*Number of fetuses* Single pregnancy: 78.2%, twins: 19.3%, triplet: 2.5%
ANTIBIOTIC	*Antibiotic treatment* 0 = none (68%), 1 = prophylactic (16%) , 2 = therapeutic (16%)
TOCOLYTIC	*Tocolytic treatment* 0 = none (31%) , 1 = short term use (9%), 2 = long term use (60%)
STEROID	*Administration of steroids* 0 = none (87%), 1 = administered (13%)
TRH	*Administration of thyreotropin releasing hormone*: 0 = none (57%), 1 = administered (43%)
INF_CULT	0 = negative culture (62%), 1 = positive culture (38%)
INF_BIOC	*Biochemical infection* 0 = no signs (72%), 1 = signs (28%)
INF_KLIN	*Clinical infection ad admission* 0 = no signs (76%), 1 = signs (24%)
INF_LAT	*Clinical infection during hospital stay* 0 = no signs (54%), 1 = signs (46%)
AMNIOTIC_FLUID	0 = anamnion (32%), 1 = oligoamnion (40%), 2 = normal amount of amniotic fluid (28%)
MODE_OF_LABOR	0 = none (20%), 1 = spontaneous (57%), 2 = induction (14%), 3 = optimization (9%)
MODE_OF_DELIVERY	1 = spontaneous (47%), 2 = vacuum (1%), 3 = forceps (0%), 4 = breech (20%), 5 = caesarean section (32%)
SEX	0 = boy (50%), 1 = girl (50%)
A1	*Apgar score after 1 minute* mean ± SD : 3 ± 3 , median 3, min 0 and max 9
A5	*Apgar score after 5 minutes* mean ± SD : 5 ± 4, median 7, min 0 and max 10
WEIGHT	*Birth weight* mean ± SD : 940g ± 409g, median 870g, min 400g and max 2500g

The smallest gestational age at the time of rupture of the membranes is 16 weeks (see Figure 7-1). For women with a gestational age of less than 22 weeks at time of PPROM the records were included only if the delivery took place after 22 weeks, because deliveries before 22 weeks are considered as miscarriages and these are not systematically registered in our unit. As a

result, the number of patients with PPROM before 22 weeks will be underestimated and the prediction models developed on this database will be too optimistic, because we can not correct the data for the babies which were born before 22 weeks and died.

The outcome is the number of days the baby survived. In case the baby died before he/she was born, the outcome specifies whether it was antenatal or intranatal. 7% of the 148 babies died antenatally, 20% intranatally, 24% on the first day, 6% from day 2 till day 7, 4 % from day 8 till day 28, 6% after the first month and 35% of the babies have survived.

Not all the variables are measured at the same time. For example the gestational age at the time of PPROM (***TPROM***) and smoking behaviour of the mother (***SMOKE***) are already known the first time the medical doctor examines the patient, while the administration of steroids (***STEROID***) and thyreotropin releasing hormone (***TRH***) are process variables measured during the time interval between PPROM and delivery. The weight (***WEIGHT***) and sex (***SEX***) of the baby are in this study only defined when the baby is born, although it is possible to estimate these factors before birth based on an ultrasound scan. The purpose of this study is to construct a predictive model for survival as early as possible during the pregnancy, i.e. a model based on process variables measured at the first examination of the medical doctor is clinically more useful than a model based on factors measured after the baby is born.

The dataset contains also records of twins and triplets, where we should take into account the influence of one baby on the other. The total number of pregnancies is 119 (148 babies). Out of the 119 pregnancies, there are 93 (78.2%) single pregnancies and 26 (21.8%) multiple pregnancies, of which 23 are twin pregnancies and 3 triplet pregnancies. This implies that 93 babies (62.8%) are from a single pregnancy, while 55 ones (37.2%) are from a multiple pregnancy. For babies of multiple pregnancies, the factors concerning the mother and the process variables measured during the time interval between PPROM and delivery are the same but the amniotic fluid (***AMNIOTIC_FLUID***) volume after PPROM, the birth weight, sex and apgar score 1 (***A1***) and 5 minutes (***A5***) after birth can differ for the babies of the same pregnancy.

The survival chances are higher in the group of single pregnancies than in the group of multiple pregnancies. This is illustrated in Figure 7-2.

If we look at Figure 7-2, we see that 44.1% of the babies of a single pregnancy have survived, while in the group of multiple pregnancies this is only 20%. In this last group significantly more babies die during the birth process.

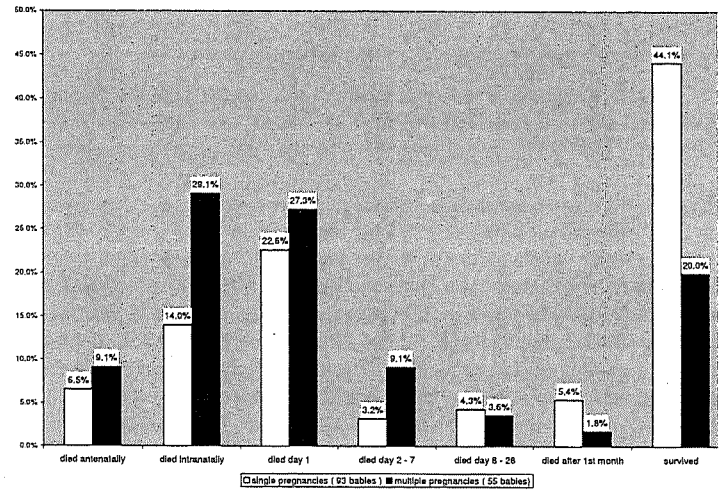


Figure 7-2 Comparison of the survival chances between babies of single pregnancies and babies of multiple pregnancies.

Figure 7-3 illustrates that babies of a twin or triplet do not have the same survival chance. For example:

- Twin 1: one baby has died on day 1 and one baby has died after the first month
- Twin 5: one baby has died on day 1 and one baby has survived
- Triplet 25: one baby has died during the first month and two babies have survived.

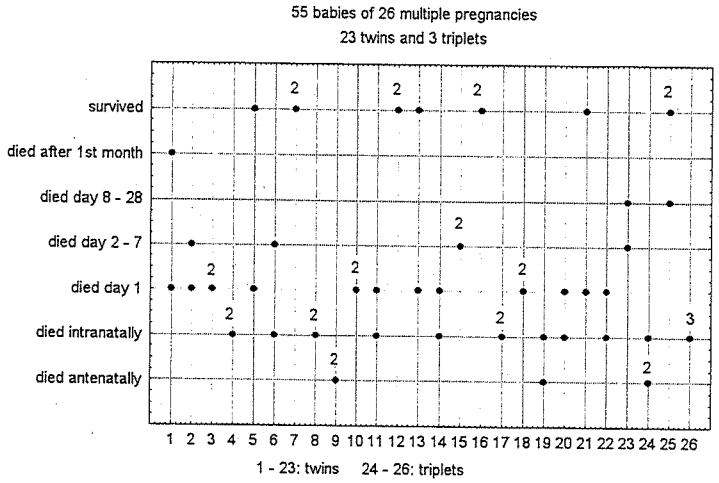


Figure 7-3 Comparison of the survival chances of babies of the same multiple pregnancy

7.2 Survival analysis

We apply Cox's proportional hazards model for survival analysis. The outcome variable is the number of days that the baby survived in case the baby was born alive. In case the baby died before he/she was born, the outcome specifies whether it was antenatal or intranatal. When constructing a survival model, a time origin has to be specified. First, we take time of birth of the baby as time origin. Therefore, we exclude the babies that died antenatally or intranatally (number of observations is 106 instead of 148). Note that the obtained parameter values will be too optimistic because we do not consider all babies who died antenatally or during delivery anymore.

Since there is a lot of tied data in our dataset (a lot of babies die on the same day, e.g. 24% die on the first day) we have to correct for it. We use the exact method (Allison, 1995; Kalbfleisch et al, 1980). In order to obtain robust variance estimates that allow for dependence between babies of twins and triplets, we use the method of Wei, Lin, and Weissfeld (Allison, 1995; Wei et al, 1989).

Using variable selection, we obtain the following model:

Table 7-2 Cox's Proportional Hazards Model

Variable	Parameter estimate	Standard error	Chi-Square	p-value
TPROM	-0.18	0.06	7.61	<0.01
WEIGHT	-0.21	0.06	13.92	<0.01
SEX	-0.95	0.32	5.57	<0.01
STEROID	-1.94	0.73	7.05	<0.01
AMNIOTIC_FLUID	-0.89	0.44	3.99	0.05

We use the generalized R^2 to get an idea of the predictive power of our survival model. The generalized R^2 is defined as follows: $R^2 = 1 - \exp(-G^2/n)$, with G^2 the likelihood-ratio chi-square statistic for testing the null hypothesis that all covariates have zero-valued coefficients, and n the sample size, which is 106 in this case. The generalized R^2 of the obtained model is smaller than 10%. So, the predictive performance is very low, which makes this model not very useful for clinical practice.

Secondly, we take another time origin. We use the time point when the mother has a rupture of the membranes as time origin. So, the outcome becomes now the number of weeks that the baby survived after rupture of the membranes. Statistical analysis using this outcome as dependent variable has shown that the predictive performance of the obtained model (the generalized R^2) has not improved.

The lack of a good predictive performance using the classical approach for survival data forces us to have another look at the data. If we look at the outcome where 7% of the 148 babies died antenatally, 20% intranatally, 24% on the first day, 10% from day 2 till day 28, 6% after the first month and 35% of the babies have survived, we see that we have three classes for the first half of the data: antenatal, intranatal, on the first day, while the other half is rather continuous. Taking this into account, we can presume that the predictive performance would be better if we construct a classification model – where we transform the outcome into several classes - rather than if we construct a continuous model where we have to make serious corrections for tied datapoints.

7.3 Classification

From a clinical point of view, there are two important end points in the outcome:

- survival of the birth process

- long term survival chance for the baby.

Normally, we would define surviving the birth process, as not dying antenatally or intranatally. But, a lot of babies that die on the first day, die only a few minutes after they were born. In clinical practice it is quite difficult to make a strict distinction between dying intranatally and dying on the first day. Therefore we take dying antenatally, dying intranatally and dying on the first day as one class. So, we define surviving the birth process as surviving the first day. The long term survival time point has been defined by medical doctors as surviving the first month (which is the same end point as in the paper (Spitz et al, 1999). Thus, we classify whether the baby survives the first day or not and whether he/she survives the first month.

We follow two ways in order to construct predictive models for this classification problem:

The most common way to construct a predictive model, based on the records of the 148 babies, where we have to take into account the dependency between some of them (babies of a twin or a triplet) is by using Generalized Estimating Equations, where we can get robust variance estimates by specifying the working correlation matrix.

Another way to look at the problem is to see the data as records of 119 pregnancies. Now, we can consider all observations as independent and just apply ordinary Logistic Regression. In case of a multiple pregnancy, we define the outcome as the outcome of the baby that has survived the longest. So, in case of single pregnancies we just examine whether the baby survives the first day or the first month, while in case of multiple pregnancies we look whether at least one of the babies makes a chance to survive the first day or first month.

Note that if we follow this way of analyzing the data, we can not use all 25 explanatory variables anymore. Since, the amount of amniotic fluid (***AMNIOTIC_FLUID***), the weight and sex of the baby, and the apgar scores after 1 and 5 minutes (***A1*** and ***A5***) can differ for the babies of a multiple pregnancy, we can not take these variables into account for model building.

We start with this second way of looking at the data. So, initially we construct a predictive model based on the records of all 119 pregnancies (93 or 78,2% single pregnancies and 26 or 21.8% multiple pregnancies), where in case of multiple pregnancies we define surviving the birth process as surviving the birth process of at least one of the babies. Secondly, we construct a classification model for all 148 babies where we have to take into account the dependence between babies of the same pregnancy (twins or triplets).

7.3.1 Total group of 119 pregnancies

Survival chances of the birth process

The data contains records of both the 93 single pregnancies and the 26 multiple pregnancies. The aim is to detect those antenatal factors that have a significant influence on the chances of the baby to survive the birth process (first day). In case of single pregnancies we examine whether the baby survives the first day, in case of multiple pregnancies we look whether at least one of the babies makes a chance to survive the birth process. In 65 of the 119 pregnancies the baby (single pregnancies) or at least one of the babies (multiple pregnancies) survived the first day.

But remember that not all the variables are measured at the same time. For example the gestational age at the time of PPROM (***TPROM***) and smoking behaviour of the mother (***SMOKE***) are already known at the first medical examination, while ***STEROID*** and ***TRH*** are process variables measured during the time interval between PPROM and delivery. The purpose of this study was to construct a predictive model for survival as early as possible during the pregnancy. Indeed, a model based on process variables measured at the first examination of the medical doctor is clinically more useful than a model based on factors measured after the baby is born.

At the first medical examination, the following variables can be measured: gestational age at the time of PPROM (***TPROM***), number of abortions (***A***), parity (***P***), number of pregnancies (***G***), smoking behaviour of the mother (***SMOKE***, ***CIGARETTE***), age of the mother (***AGE***) and the number of fetuses (***F***) (single or multiple pregnancy).

Only one of the eight factors seems to be important by variable selection: the gestational age at the time of PPROM (***TPROM***) with a p-value of 0.02. We can not perform a reliable classification with this one factor (area under the ROC-curve (De Long, 1988, Hanley et al, 1982) is 0.71 on the whole set of 119 observations). Note that the difference between a single pregnancy and a multiple pregnancy does not have a significant influence if we predict that at least one of the babies survives the first day.

If the delivery does not take place immediately after the first examination, we take more variables into account. The medical doctor can look whether there are infections (***INF_CULT***, ***INF_BIOC***, ***INF_KLIN***, ***INF_LAT***). He can give the mother an antibiotic or tocolytic treatment, or administration of steroids or thyreotropin releasing hormone. Variable selection (while including always ***TPROM*** in our model) showed that the administration of steroids (***STEROID***) or thyreotropin releasing hormone (***TRH***) is significant in order to survive the birth process. The p-values are respectively 0.01 and < 0.01. But if we construct a model with those two variables and the gestational age at time of PPROM (***TPROM***) we still can not per-

form a reliable classification (the area under the ROC-curve is 0.83 on the whole set of 119 observations).

Once the delivery begins, three other variables can be considered: ***LATENCY***, ***MODE_OF_LABOR*** and ***MODE_OF_DELIVERY***. If we include the selected variables ***TPROM***, ***STEROID*** and ***TRH*** in our model and perform variable selection on the additional variables ***LATENCY***, ***MODE_OF_LABOR*** and ***MODE_OF_DELIVERY***, the results show that only the time interval between PPROM and delivery (***LATENCY***) provides additional significant information in order to estimate the survival chances of the baby. If we take this additional factor into account we can perform a fairly good classification (area under the ROC-curve is 0.92 on the whole set, see Figure 7-4).

Table 7-3 Logistic Regression model for predicting the survival chances of the birth process in the group of 119 pregnancies

Variable	Parameter estimate	Standard error	Chi-Square	p-value
Intercept	-22.51	4.92	20.95	<0.01
TPROM	0.86	0.20	19.56	<0.01
STEROIDS	2.89	1.20	5.84	0.02
TRH	1.67	0.53	4.82	0.03
LATENCY	0.08	0.02	18.44	<0.01

Table 7-4 Classification table of the Logistic Regression model for predicting the survival chances of the birth process in the group of 119 pregnancies

Probability level	% correctly classified	% of babies that are predicted to survive the birth process in the group of babies that effectively survived the birth process (Sensitivity)	% of babies that are predicted not to survive the birth process in the group of babies that effectively died (Specificity)
0.60	79%	0.78	0.79
0.50	80%	0.82	0.77
0.45	84%	0.91	0.75
0.40	83%	0.92	0.72
0.35	**84%**	**0.94**	**0.72**
0.30	82%	0.94	0.68

If we take a look at the classification table, we see that we should take 0.35 as cut-off value, where we classify babies with P(survive the birth process) > 0.35 into the class of babies that survive the birth process, and classify babies with P(survive the birth process) < 0.35 into the 2nd class of babies which die before or during the birth process. Taking 0.35 as cut-off value, we obtain a sensitivity of 94% and a specificity of 72%.

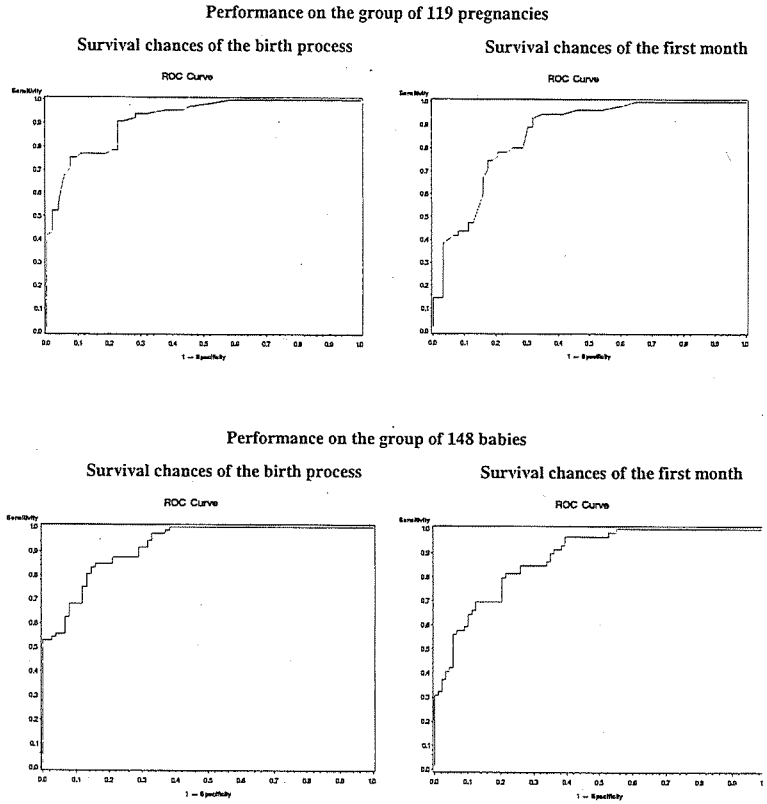


Figure 7-4 ROC curves of the predictive Logistic Regression and Generalized Estimating Equations model

To get an idea of the predictive power of this classification model on a new set of records, we randomly divide the original set of 119 pregnancies into two sets: a training set, which contains 70% of the data, and a test, containing the remaining 30%.

The obtained four factors are used in the Logistic Regression model. We use the training set to find the parameter values of the models and the test set to see how good a model is able to classify a new set of observations. We repeat this procedure 20 times. We calculate the parameter values for 20 Logistic Regression models based on 20 randomly selected training sets, and apply those models on the 20 test sets. So, we can get an idea of the mean value and the variance of the area under the ROC-curve.

Note that Logistic Regression is a linear model. In order to know if it is necessary to consider nonlinear terms in our model, we apply the nonlinear black box model LS-SVM with RBF kernel as described in a previous chapter. We use the same four factors as input variables and apply the LS-SVM model to the randomly selected training sets to train the model. Then we apply the trained models to the 20 corresponding test sets.

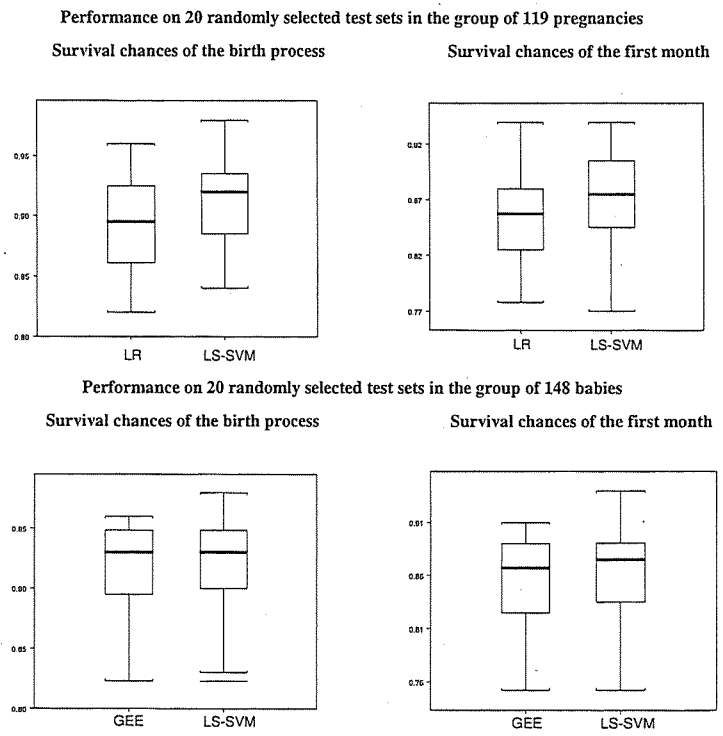


Figure 7-5 Boxplots of the area under the ROC curve for Logistic Regression (LR), Generalized Estimating Equations (GEE) and Least Squares Support Vector Machines (LS-SVM)

The mean value of the area under the ROC-curve for those 20 randomizations is 0.90 (stdev 0.04) in case of Logistic Regression. For the 20 LS-SVM models the mean is 0.91, stdev 0.04. The results of those 20 randomizations are illustrated in Figure 7-5. Figure 7-5 shows a boxplot for the area under the ROC-curves for the two applied models: LR and LS-SVMs. The difference between those two models is clearly non significant [14,15],

so there is no need to consider nonlinear terms in our Logistic Regression model.

Now, we construct a predictive model to determine the survival chances of the first month.

Survival chances of the first month

The aim is again to detect those antenatal factors that have a significant influence on the chances of a baby to survive the first month. In case of single pregnancies we look whether the baby will survive the first month, in case of multiple pregnancies we look if at least one of the babies has a chance to survive. In 55 of the 119 pregnancies the baby (single pregnancies) or at least one baby (multiple pregnancies) survived. We will follow exactly the same procedure as in the previous section:

First, we apply stepwise selection on the variables that are known at the first medical examination: ***TPROM***, ***AGE***, ***A***, ***P***, ***G***, ***SMOKE***, ***CIGARETTE*** and ***F***. Then we include the selected significant factors in our model and perform variable selection on the variables that are determined in the time period between the rupture of the membranes and the delivery: ***INF_CULT***, ***INF_BIOC***, ***INF_KLIN***, ***INF_LAT***, ***ANTIBIOTIC***, ***TOCOLYTIC***, ***STEROID*** and ***TRH***. Finally, we add the remaining variables to our model: ***LATENCY***, ***MODE_OF_LABOR*** and ***MODE_OF_DELIVERY*** and perform variable selection on those three.

This whole procedure gives us the same significant variables as in the previous section in order to construct a classification model: ***TPROM***, ***STEROID***, ***TRH*** and ***LATENCY***. The area under the ROC-curve for this model is 0.86 (Figure 7-4).

Table 7-5 Logistic Regression model for predicting the survival chances of the first month in the group of 119 pregnancies

Variable	Parameter estimate	Standard error	Chi-Square	p-value
Intercept	-17.18	4.17	16.96	<0.01
TPROM	0.64	0.16	14.98	<0.01
STEROID	0.89	0.87	4.77	0.03
TRH	1.24	0.48	6.63	0.01
LATENCY	0.05	0.01	14.77	<0.01

Table 7-6 Classification table of the Logistic Regression model for predicting the survival chances of the first month in the group of 119 pregnancies

Probability level	% correctly classified	% of babies that are predicted to survive the first month in the group of babies that effectively survived the first month (Sensitivity)	% of babies that are predicted not to survive the first month in the group of babies that effectively died (Specificity)
0.60	78%	0.73	0.83
0.50	78%	0.76	0.79
0.45	77%	0.80	0.75
0.40	75%	0.80	0.71
0.35	78%	0.89	0.68
0.30	**78%**	**0.95**	**0.63**
0.25	75%	0.95	0.59

According to the classification table, we should take 0.30 as cut-off value, where we classify babies with P(survive the first month) > 0.30 into the class of babies that survive the first month, and classify babies with P(survive the first month) < 0.30 into the class of babies which die before the end of the first month. Taking 0.30 as cut-off value, we obtain a sensitivity of 95% and a specificity of 63%.

Now, we will apply Logistic Regression and LS-SVMs to the same 20 randomizations. In case of Logistic Regression the mean value of the area under the ROC-curve is 0.85 (stdev 0.04). For the 20 LS-SVM models the mean is 0.87, stdev 0.05. The results of those 20 randomizations are illustrated in Figure 5. As in previous section the difference between those two models is clearly non significant, so there is no need to consider nonlinear terms in our Logistic Regression model.

7.3.2 Total group of 148 babies

Now, we investigate the other way to construct a predictive model, i.e. using the information of all 148 babies, where we have to take into account the dependency between some of them (babies of a twin or a triplet). Therefore we use Generalized Estimating Equations, where we can get robust variance estimates by specifying the working correlation matrix.

We will follow the same procedure as in the previous section for the selection of the significant variables, i.e. application of variable selection in three stages:

- Stage 1: stepwise selection on the variables that are known at the first medical examination: *TPROM*, *AGE*, *A*, *P*, *G*, *SMOKE*, *CIGARETTE* and *F*
- Stage 2: include the selected significant factors in the model and perform stepwise selection on the variables that are determined in the time period between the rupture of the membranes and the delivery: *INF_CULT*, *INF_BIOC*, *INF_KLIN*, *INF_LAT*, *ANTIBIOTIC*, *TOCOLYTIC*, *STEROID* and *TRH*
- Stage 3: add the remaining variables to the model: *LATENCY*, *MODE_OF_LABOR* and *MODE_OF_DELIVERY* and perform stepwise selection on those three.

Here, we can add a fourth stage:

- Stage 4 : include the selected variables from Stage 1, 2 and 3, and perform stepwise selection on the additional baby specific variables: *AMNIOTIC_FLUID*, *SEX*, *A1*, *A5* and *WEIGHT*.

Survival chances of the birth process

72 babies of the 148 (49%) have survived the birth process. The important factors according to the GEE model in order to predict the survival chances of the birth process are: the gestational age at the time of PPROM (*TPROM*), the volume of the amniotic fluid (*AMNIOTIC_FLUID*), administration of steroids (*STEROID*), the birth weight (*WEIGHT*) and the apgar score 1 minute after birth (*A1*). With these 5 variables the area under the ROC-curve reaches 0.92 (Figure 7-4).

Table 7-7 GEE model for predicting the survival chances of the birth process in the group of 148 babies

Variable	Parameter estimate	Standard error	Z	p-value
Intercept	-13.26	2.84	-4.67	<0.01
TPROM	0.36	0.10	3.44	<0.01
STEROID	3.80	0.74	5.16	<0.01
AMNIOTIC_FLUID	1.15	0.50	2.28	0.02
WEIGHT	0.0027	0.0009	3.00	<0.01
A1	0.47	0.17	2.84	<0.01

Table 7-8 Classification table of the GEE model for predicting the survival chances of the birth process in the group of 148 babies

Probability level	% correctly classified	% of babies that are predicted to survive the birth process in the group of babies that effectively survived the birth process (Sensitivity)	% of babies that are predicted not to survive the birth process in the group of babies that effectively died (Specificity)
0.60	82%	78%	87%
0.50	83%	83%	86%
0.45	**85%**	**85%**	**84%**
0.40	83%	85%	82%
0.35	82%	86%	79%
0.30	82%	88%	76%

If we take a look at the classification table, we see that we should take 0.45 as cut-off value, where we classify babies with P(survive the birth process) > 0.45 into the class of babies that survive the birth process, and classify babies with P(survive the birth process) < 0.45 into the 2nd class of babies which die before or during the birth process. Taking 0.45 as cut-off value, we obtain a sensitivity of 85% and a specificity of 84%.

Figure 5 shows that there is no significant difference between the performance of LR and LS-SVM for this dataset. In case of Logistic Regression the mean value of the area under the ROC-curve 0.92 (stdev 0.04). For the 20 LS-SVM models the mean is 0.92, stdev 0.04.

Survival chances of the first month

59 of the 148 babies (40%) have survived the first month. The important factors according to the GEE model in order to predict the survival chances of the first month are: the gestational age at the time of PPROM (***TPROM***), a binary variable indicating whether it is a single pregnancy or a multiple pregnancy (***F***), the volume of the amniotic fluid (***AMNIOTIC_FLUID***), administration of steroids (***STEROID***) and thyreotropin releasing hormone (***TRH***), the birth weight (***WEIGHT***) and the sex of the baby (***SEX***). With these 7 explanatory variables we get an area under the ROC-curve of 0.88 (Figure 4).

Table 7-9 GEE model for predicting the survival chances of the first month in the group of 148 babies

Variable	Parameter estimate	Standard error	Z	p-value
Intercept	-11.60	2.89	-4.01	<0.01
TPROM	0.24	0.11	2.17	0.03
F	1.72	0.55	3.11	<0.01
AMNIOTIC_FLUID	1.08	0.48	2.24	0.02
STEROID	2.33	0.80	2.93	<0.01
TRH	1.05	0.52	2.01	0.04
WEIGHT	0.0025	0.0007	3.37	<0.01
SEX	1.39	0.47	2.96	<0.01

Table 7-10 Classification table of the GEE model for predicting the survival chances of the first month in the group of 148 babies

Probability level	% correctly classified	% of babies that are predicted to survive the birth process in the group of babies that effectively survived the first month (Sensitivity)	% of babies that are predicted not to survive the birth process in the group of babies that effectively died (Specificity)
0.60	78%	0.58	0.92
0.50	79%	0.64	0.89
0.45	77%	0.69	0.82
0.40	77%	0.73	0.80
0.35	**79%**	**0.80**	**0.79**
0.30	78%	0.83	0.74
0.25	74%	0.86	0.65
0.20	74%	0.90	0.64

According to the classification table, we should take 0.35 as cut-off value, where we classify babies with P(survive the first month) > 0.35 into the class of babies that survive the first month, and classify babies with P(survive the first month) < 0.35 into the class of babies which die before the end of the first month. Taking 0.35 as cut-off value, we obtain a sensitivity of 80% and a specificity of 79%.

Figure 7-5 shows that there is no significant difference between the performance of LR and LS-SVM for this dataset. In case of Logistic Regression the mean value of the area under the ROC-curve is 0.85 (stdev 0.04). For the 20 LS-SVM models the mean is 0.87, stdev 0.05.

7.4 Prospective Analysis

The prospective dataset contains medical records of 64 consecutive cases (52 pregnancies) of PPROM at or before 26 completed weeks of gestation. All women were admitted with a living fetus at the university hospital Gasthuisberg Leuven between 1997 and 2000. The aim is to test prospectively the predictive models that we have developed based on records in the period 1988 – 1996 (section 6.3) on the new set of records.

Let us first have a look at a comparison between the original dataset (period 1988 – 1996) and the prospective dataset (period 1997 – 2000). While in the original dataset only 35% of the babies survived (31% in the period '88 – '91 and 41% in the period '92 – '95), 52% of the babies survived in the prospective dataset (see Figure 7-6). So, a lot more babies have survived in the prospective dataset than in the original dataset.

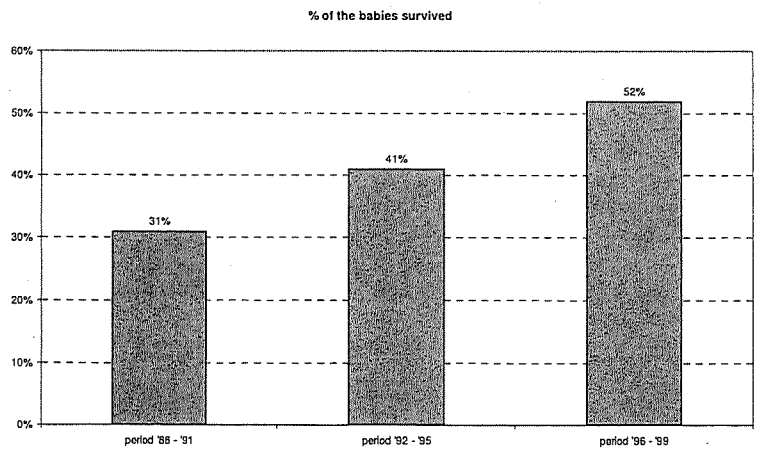


Figure 7-6 Survival chances of the babies in the original dataset (period 1988-1996) and prospective dataset (period 1997-2000)

We can also compare the explanatory variables that were selected for the several classification models.

Table 7-11 Comparison of the selected (in the several predictive models) explanatory variables between the original dataset and the prospective dataset

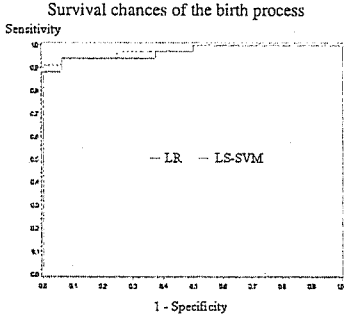
	Original dataset (period '88 – '96)	Prospective dataset (period '97 – '99)
TPROM	mean 24 weeks	mean 22 weeks
LATENCY	mean 24 days	mean 29 days
STEROID	18%	33%
TRH	56%	17%
AMNIOTIC_FLUID	normal amount : 24%	normal amount: 25%
SEX	50% girls	33% girls
WEIGHT	mean 939g	mean 926g
A1	mean 3.2	mean 3.5
A5	mean 5.1	mean 5.0

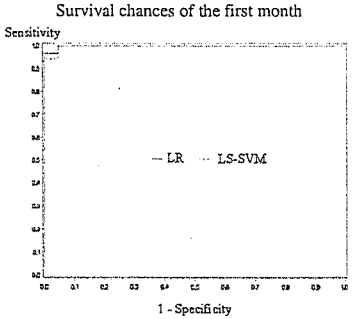
Concerning these variables, more use is made of ***Steroids*** and less of ***TRH***, and there are more boys in the prospective set than in the original dataset.

For testing the predictive models prospectively, we will follow the same order of section 6.3. The prospective dataset contains records of 64 babies (52 pregnancies). So, we will first test the logistic regression model, developed on the 119 records in the period 1988 – 1996, on the prospective set of 52 records (period 1997 – 2000). Then we apply GEE, developed on 148 observations (period 1988 – 1996), on the 64 cases of the prospective dataset. Here, we will also apply Least Squares Support Vector Machine classifiers using the same inputs as for logistic regression and GEE, to check the quality of logistic regression and GEE as linear models in relation to a nonlinear classifier.

As shown in Figure 7-7 and Table 7-12, there is no statistically significant difference in the performance between the linear LR and the nonlinear LS-SVM. Note that we obtain a specificity of 100% on the prospective set, i.e. all the babies that have effectively died have been predicted to die.

As shown in Figure 7-8 and Table 7-13, there is also no significant difference in the performance between the linear and the nonlinear model and we obtain again a specificity of 100%.

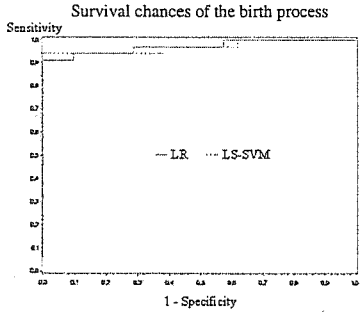


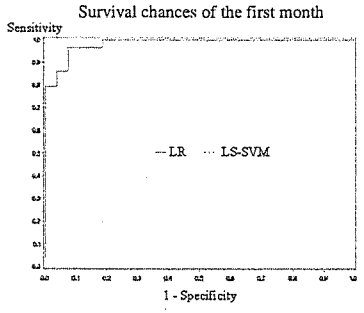


	LR	LS-SVM
Area under the ROC-curve	0.97	0.98
Correct classification	88%	88%
Sensitivity	82%	82%
Specificity	100%	100%

	LR	LS-SVM
Area under the ROC-curve	0.998	0.997
Correct classification	98%	96%
Sensitivity	97%	93%
Specificity	100%	100%

Figure 7-7: ROC curves of Logistic Regression (LR) and Least Squares – Support Vector Machines (LS-SVM), and performance measures (area under the ROC curve, accuracy, sensitivity and specificity) in the prospective set of 52 pregnancies.





	GEE	LS-SVM
Area under the ROC-curve	0.97	0.97
Correct classification	88%	91%
Sensitivity	82%	85%
Specificity	100%	100%

	GEE	LS-SVM
Area under the ROC-curve	0.98	0.98
Correct classification	89%	89%
Sensitivity	79%	79%
Specificity	100%	100%

Figure 7-8 ROC curves of Generalized Estimating Equations (GEE) and Least Squares – Support Vector Machines (LS-SVM), and performance measures (area under the ROC curve, accuracy, sensitivity and specificity) in the prospective set of 64 babies.

7.5 Discussion

Now, we can compare the performance of the two models:

- Logistic Regression based on the records of 119 pregnancies
- GEE based on the records of 148 babies.

Concerning the prediction of the survival chances of the first day, we see that both models reach just the same performance: 88% correct classification, sensitivity 82%, specificity 100%. Concerning the prediction of the survival chances of the first month, the Logistic Regression model is better than the GEE model: LR reaches 98% correct classification, sensitivity 97%, specificity 100%, while GEE only reaches 89% correct classification with a sensitivity of 79% and a specificity of 100%. So, the point of view for analyzing the data as explained in section 7.3, i.e. seeing the data as records of 119 pregnancies instead of 148 babies, has its benefits.

Another advantage is that the LR model does not make use of any neonatal variables, contrary to GEE which makes us of e.g. the birth weight of the baby. The LR model is based on variables known at the intake of the patient with PPROM. Such a model can help in counseling and choosing the best treatment modalities.

Each time we calculated the intrinsically nonlinear LS-SVM model, we did not obtain a significantly better performance. Therefore, there is no need to include nonlinear terms in the Logistic Regression or GEE model.

We can also compare our GEE model for survival of the first month with the GEE model in the paper (Spitz, 1999). As mentioned in the introduction the variables of the model determining the survival chances in (Spitz et al, 1999) were: the birth weight of the baby (***WEIGHT***), the sex of the baby (***SEX***), the gestational age at the time of rupture of the membranes (***TPROM***), and whether steroids have been administered during the pregnancy (***STEROID***). Two interaction terms were also included: one between ***STEROID*** and ***SEX***, and one between ***STEROID*** and ***TPROM***.

This model reaches an area under the ROC-curve of 0.96, which is not significantly different from our GEE model for predicting the survival chances of the first month. But, the proportion correctly classified data is higher using our model: we obtain 89% correct classification, while in the paper (Spitz et al, 1999) the GEE model reaches only 81% (sensitivity: 100% and specificity 66%) (Table 7-12).

Note that the performance in the prospective set is higher than the performance in the original set where the different models were developed on, e.g., concerning the prediction of the survival chances of the first day, Logistic Regression obtains an area under the ROC-curve of 0.92 on the original set and 0.97 on the prospective set of observations. As stated in section 6.4 there is a significant difference in the survival rate between the original and the prospective dataset. While in the original dataset (period 1988 – 1996) only 35% of the babies survived, 52% of the babies survived in the period 1997 – 2000. It might be that the predictive power of >95% area under the ROC-curve is an overestimation when in the future the survival rate is not as high as the survival rate in our prospective set.

Table 7-12 Classification table of the GEE model in paper (Spitz et al, 1999)

Probability level	% correctly classified	% of babies that are predicted to survive the first month in the group of babies that effectively survived the first month (Sensitivity)	% of babies that are predicted to not survive the first month in the group of babies that effectively died (Specificity)
0.6	81%	100%	66%
0.5	**81%**	**100%**	**66%**
0.45	79%	100%	63%
0.4	79%	100%	63%
0.35	79%	100%	63%
0.30	79%	100%	63%

To get an idea of the predictive power on a randomly selected test set in the total dataset, i.e. the records of the period 1988 – 2000, we will randomly divide the total set into two sets: a training set, containing 70% of the data, and a test set, containing the remaining 30% and repeat this procedure 20 times. Now, we can get an idea of the mean value and the variance of the area under the ROC-curve on a randomly selected test set in the total dataset.

We can conclude that on a randomly selected test set in the total dataset we obtain an area under the ROC-curve of almost 90%. So, for the future, we can expect a predictive power of at least 87% as area under the ROC-curve, and > 95% if the survival rate would stay as high as in our prospective dataset (period 1997 – 2000).

Table 7-13 Comparison in performance between the additive Logistic Regression (LR) and Generalized Estimating Equations (GEE) models and the intrinsically nonlinear Least Squares – Support Vector Machines (LS-SVM)

Performance on 20 randomly selected test sets in the total group of 119 + 52 = 171 pregnancies

Survival chances of the birth process

	GEE	LS-SVM
Area under the ROC-curve	0.89 (SE:0.04)	0.89 (SE:0.04)
Correct classification	83%	83%
Sensitivity	90%	87%
Specificity	73%	78%

Survival chances of the first month

	GEE	LS-SVM
Area under the ROC-curve	0.88 (SE:0.04)	0.88 (SE:0.04)
Correct classification	79%	80%
Sensitivity	84%	86%
Specificity	73%	75%

Performance on 20 randomly selected test sets in the total group of 148 + 64 = 212 babies

Survival chances of the birth process

	LR	LS-SVM
Area under the ROC-curve	0.88 (SE:0.04)	0.88 (SE:0.04)
Correct classification	78%	79%
Sensitivity	81%	77%
Specificity	76%	80%

Survival chances of the first month

	LR	LS-SVM
Area under the ROC-curve	0.87 (SE:0.04)	0.87 (SE:0.04)
Correct classification	78%	79%
Sensitivity	72%	75%
Specificity	82%	81%

7.6 Conclusions

The development of a prediction model for survival can allow targeting interventions toward those at highest risk. Traditional methods for predicting women predestined to deliver preterm relied upon obstetrical history, demographic factors or subjective symptoms that were neither very sensitive nor specific.

In our models we found that the gestational age at the time of PPROM, administration of steroids or thyreotropin releasing hormone and the time interval between time of PPROM and delivery were the most important variables determining the chance of survival.

Comparing the two datasets (period 1988-1996 and 1997-2000) we saw that in the retrospective dataset only 35 % of the babies survived while in the prospective set 52 % survived. By looking at the different variables for an explanation we saw that steroids were more frequently administered (33% vs. 18%) in the prospective set but that on the other hand there were more boys than girls (67% vs. 50%). Despite these differences it is clear that the logistic regression model derived from the retrospective dataset performs well in the new prospective dataset.

Of all the variables determining survival, administration of steroids is the most important factor that can be influenced at the time of diagnosis. Our model is in agreement with the data in the literature concerning the sur-

plus value of steroids. All the other significant variables cannot be influenced by external factors.

With an accuracy of 98 %, a sensitivity of 97 % and a specificity of 100 % the logistic model showed a very good predictive capacity in predicting the survival chances of the first month in our prospective dataset. However, we have to await external validation in order to test whether these results can be generalized. Incorporation of statistical models finds its way into medicine and can help us in daily dilemma's concerning prognosis and survival. Still treatment of early PPROM remains a challenging task for those who take care of obstetric patients.

Chapter VIII

Conclusions and future research

8.1 General conclusions

We have built predictive models for two major decision problems in gynecology:

3. the preoperative classification of ovarian tumors, and
4. the construction of predictive survival models for premature babies of which the mother had a preterm prelabor rupture of the membranes before 26 weeks of gestation.

For the former we proceeded in two ways. Firstly, we built one model for the whole data where we reached a sensitivity of 92% and a specificity of 81% on a test set that was not used at all to construct the model. This model, which we called *overall* model, beats all the previous models described in literature. We can also conclude that measurements of serum CA 125 levels are not a necessary part of the assessment of an adnexal mass. They do not enhance our ability to characterize an ovarian mass in premenopausal patients, not even in postmenopausal patients.

The overall model could offer considerable help to less experienced operators. But it did not reach the accuracy obtained by expert operators, since there was still a significant difference in specificity: experts reached a specificity of 93.7%, compared to only 81% for our overall model. Therefore, we decided to look at different subgroups of patients in order to detect for which specific group the development of new models could be worthwhile. We detected two groups where the specificity of the overall model was very poor:

- 64.3% in the tumors with a solid component, but without papillary projections, and
- only 51.1% in tumors with papillary projections.

Therefore we decided to build new models for those subsets. When combining the models for the several subsets, we obtained a sensitivity and a specificity of 88% and 90.3%, respectively. This was almost a 10% increase in specificity compared to the overall model. There was no longer a statistically significant difference with the performance of the expert's subjective assessment. This means that we have reached our goal of building a model that equals the performance of highly trained experts in order to classify ovarian tumors pre-operatively.

We might ask the question now: "Is it possible to beat those experts?" Unfortunately, we had to conclude that we could not yet build predictive models that substantially improved diagnostic accuracy in tumors that are judged to be difficult to classify by the experts.

When comparing the ultrasound characteristics of different types of ovarian malignant tumors we came to the conclusion that papillary projections are characteristic of borderline tumors and stage 1 primary invasive epithelial ovarian cancer. A small proportion of solid tissue in a malignant mass at ultrasound examination makes it more likely to be a borderline tumor or a stage I epithelial ovarian tumor than an advanced tumor or a rare type of tumor.

We also defined optimal cutoff values of ultrasound variables for use in clinical practice.

For the second classification problem we constructed predictive models (using Cox's proportional hazards model, logistic regression, Least Squares Support Vector Machines and generalized estimating equations) for survival of premature babies of which the mother had a preterm prelabor rupture of the membranes. With an accuracy of 98 %, a sensitivity of 97 % and a specificity of 100 % a logistic model showed a very good predictive capacity in predicting the survival chances of the first month in a prospective dataset.

8.2 Future research

IOTA phase II has started last month (June 2005). The aim is to collect data of 1000 patients to test prospectively our two models (overall and subset model) for the pre-operative discrimination between benign and malignant ovarian tumors in new centers. The main question is: "When tested prospectively which one will show the best performance in terms of sensitivity, accuracy, specificity, stability and ease-of-use in clinical practice?"

In addition, this additional database will enable researchers to further subclassify ovarian tumors (e.g. to discriminate between functional, non-functional benign, primary invasive and metastatic tumors), thereby refining the diagnosis.

A second important issue for further research is the development of new models to help in classifying tumors that are difficult to classify on the basis of clinical variables and ultrasound examination even by experienced examiners. These new models could contain new information, such as genomic and/or proteomic pattern analyses.

The predictive model for survival of premature babies of which the mother had rupture of the membranes before 26 weeks of gestation has to be externally validated in order to test whether the excellent performance results can be generalized. Incorporation of statistical models finds its way into medicine and can help us in daily dilemma's concerning prognosis and survival. But still treatment of early PPROM remains a challenging task for those who take care of obstetric patients.

Appendix

A Logistic Regression

Suppose the response variable Y can only take two possible values, denoted by 0 and 1. For example, $Y = 1$ if a tumor is malignant and $Y = 0$ otherwise. In that case, logistic regression can be used to explore the relation between the response variable Y and several explanatory variables $X_1, X_2, \ldots, X_k$. Let $p = p(X_1, X_2, \ldots, X_k)$ denote the probability that the disease is present, given the explanatory variables $X_1, X_2, \ldots, X_k$. The logistic regression model can then be written as follows :

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \ ,$$

where $\frac{p}{1-p}$ is called the *odds* that the disease is present given $X_1, X_2, \ldots, X_k$. This equation models the log of the odds or the *logit* as a linear function of the explanatory variables; β_0 is the intercept parameter and $\beta_1, \ldots, \beta_k$ are the slope parameters. The regression parameters $\beta_0, \beta_1, \ldots, \beta_k$ are estimated by the method of maximum likelihood. The logistic regression model can also be rewritten as follows:

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)} \ .$$

Note that the logit can range from minus to plus infinity, while p falls between 0 to 1.

The odds ratio $\exp(\beta_i)$ is the ratio of odds for a one-unit increase in X_i. This means that the odds increase multiplicatively $\exp(\beta_i)$ for every one-unit increase in X_i.

Since logistic regression does not make any assumptions about the distribution of the explanatory variables, it can be used when the explanatory variables are a mixture of categorical and continuous variables, which will be the case for the problems studied in this thesis.

Variable selection

The traditional approach to statistical model building involves *seeking the most parsimonious model that still explains the data* (Hosmer and Lemeshow, 2001; Harrell, 2002). The rationale for minimizing the number of variables in the model is that the resulting model is more likely to be numerically stable, and is more easily generalized. The more variables included in a model, the larger the estimated standard errors become, and the more dependent the model becomes on the observed data.

The selection process should begin with a careful univariate analysis of each variable. For binary, nominal or ordinal variables, this can be done by using the Pearson chi-square test. For continuous variables, the most desirable univariate analysis involves fitting a univariate logistic regression.

Upon completion of the univariate analyses, we select variables for the multivariate analysis. One approach to variable selection is to use a stepwise method in which variables are selected either for inclusion or exclusion from the model in a sequential fashion based solely on statistical criteria. There are two main versions of the stepwise procedure: a) forward selection with a test for backward elimination and b) backward elimination followed by a test for forward selection. The stepwise approach is useful and intuitively appealing in that it builds models in a sequential fashion and it allows for the examination of a collection of models which might not otherwise have been examined. "Best subset selection" is another selection method that has not been used extensively in logistic regression. With this procedure a number of models containing one, two, three variables, and so on, are examined to determine which ones are considered the "best" according to some specified criteria.

Stepwise, best subsets, and other mechanical selection procedures have been criticized because they can yield a medical implausible model and can select irrelevant variables. However, employing e.g. a stepwise selection procedure can provide a fast and effective means to screen a large number of variables. It is only when the analyst understands the strengths, and especially the limitations, of the methods that these methods can serve as useful tools in the model building process.

The importance of each variable included in the selected model should be verified. This process of deleting, re-estimating, and verifying continues until all important variables are included in the model and those excluded are clinically and/or statistically unimportant.

Once we have obtained a model that contains the essential variables, we should look more closely at the variables in the model. For continuous variables we should check the assumption of linearity in the logit. One approach is to fit a generalized additive model (Hastie and Tibshirani, 1999). See next subsection.

When we have ascertained that each of the individual variables is scaled correctly, we check for interactions among the variables in the model (Harrell, 2002). In any model, an interaction between two variables implies that the effect of one of the variables is not constant over levels of the other. E.g. suppose we have an interaction between the explanatory variables *age* and a binary variable *X*, in order to model the outcome, which denotes if a tumor is benign or malignant. This would imply that the effect of age on the outcome depends on the value of *X*, for example it can be that when X=0 the older the patient is, the more risk he or she has on a malignant tumor, while for X=1 it is just the opposite: the older the patient is, the smaller the risk. The final decision as to whether an interaction term should be included in a model, should be based on statistical as well as practical considerations. Any interaction term in the model must make sense from a clinical perspective.

Generalized additive models

The generalized additive model extends the logistic regression model by fitting nonparametric functions to estimate relationships between the outcome and the explanatory variables. The nonparametric functions are estimated from the data using smoothing operations. This generalization replaces the linear model

$$\text{logit}(p) = log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

with an additive model

$$\text{logit}(p) = \log\frac{p}{1-p} = \alpha + f_1(X_1) + ... + f_k(X_k),$$

where α is a constant intercept term and f_i corresponds to a nonparametric function describing the relationship between the logit(p) and the ith explanatory variable. E.g. cubic B-splines (De Boor, 1978) or lowess (locally weighted least-squares regression smoother – Cleveland, 1979) can be used to fit smooth relationships.

Goodness-of-fit

Hosmer and Lemeshow developed a goodness-of-fit test for logistic regression models with binary responses (Hosmer and Lemeshow, 2001). This test involves dividing the data into approximately ten groups of roughly equal

size based on the percentiles of the estimated probabilities. The observations are sorted in increasing order of their estimated probability of having an event outcome. So, the first group contains the N/10 subjects having the smallest estimated probabilities, and the last group contains the N/10 patients having the largest estimated probabilities.

The discrepancies between the observed and expected number of observations in these groups are summarized by the Pearson chi-square statistic. Using an extensive set of simulations, Hosmer and Lemeshow (1980) showed that, when the fitted logistic regression model is the correct model, the distribution of this statistic is well approximated by the chi-square distribution with f degrees of freedom, where f is the number of groups minus 2, i.e. 10-2 = 8. A small p-value suggests that the fitted model is not an adequate model.

The advantage of such a summary goodness-of-fit statistic is that it provides a single, easily interpretable value that can be used to assess the quality of the model.

B Least Squares Support Vector Machines

Least Squares Support Vector Machines (LS-SVM) classifiers are a modified version of Support Vector Machines that can be used for binary classification (Suykens et al, 1999). Given a training set $\{x_{1i}, ..., x_{ki}, y_i\}_{i=1, ..N}$ with input data $x_{1i}, ..., x_{ki}$ - where x_{ki} denotes the value of the k-th explanatory variable for the i-th datapoint - and corresponding class labels $y_i \in \{-1, +1\}$. The LS-SVM classifier takes the following (primal) form:

$$y(x) = \text{sign}\left[w^T \varphi(x) + b\right] \quad (1)$$

where the input data (that is said to belong to the input space) is mapped to a high dimensional feature space (which can be infinite dimensional) by a mapping function φ(x) (that does not have to be specified). This means that the classification is done in a high dimensional feature space where w is an element.

To determine the parameters of this model, the following optimization problem has to be solved:

$$\min_{w,b,e}\left[\frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{i=1}^{N}(e_i)^2\right], \quad (2)$$

subject to:

$$y_i\left[w^T\varphi(x_i) + b\right] = 1 - e_i,$$

for $i = 1, \ldots, N$. The first term in equation (2) is called the regularization term and is representative of the model complexity; the second term is representative for the training set error. The process of limiting model complexity is called regularization and is necessary to prevent overfitting and enhance the generalization performance of this model. Note that γ is a hyperparameter of the model, which is called the regularization parameter. Also note that γ determines the degree of balance that has to be reached between model complexity and training set error. If γ is chosen to be infinite, no regularization is performed and only the training set error is minimized.

The optimization problem is dealt with by solving the Lagrangian, which results in:

$$w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i) \tag{3}$$

and the following dual problem to be solved in α and b:

$$\begin{bmatrix} 0 & y \\ y^T & \Omega + \mathrm{I}/\gamma \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1_N \end{bmatrix}, \tag{4}$$

where

$$y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix},$$

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}, \quad 1_N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

and where the elements of the matrix Ω are given by:

$$\Omega_{ml} = y_m y_l \varphi(x_m)^T \varphi(x_l). \tag{5}$$

The mapping function $\varphi(x)$ does not have to be constructed explicitly. We only have to specify the inner product in the feature space. This is represented by a (symmetric) kernel function that satisfies the Mercer condition:

$$K(r,s) = \varphi(r)^T \varphi(s),$$

where r and s belong to the input space. In this thesis we only consider the Radial Basis Function (RBF) kernel given by:

$$K(r,s) = \exp\left(-\frac{\|r-s\|_2^2}{\sigma^2}\right), \tag{6}$$

where σ is another hyperparameter. Using a RBF kernel will result in a non-linear classifier. Applying the kernel trick to Equation (5) results in

$$\Omega_{ml} = y_m y_l K(x_m, x_l). \tag{7}$$

After applying Equation (6) to Equation (7), all the coefficients of the set of equations given in Equation (4) are known. Finally, substitution of Equation (3) in equation (1) results in the LS-SVM classifier in the dual form:

$$y(x) = \operatorname{sign}\left[\sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)^T \varphi(x) + b\right]$$

$$= \operatorname{sign}\left[\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b\right].$$

In the dual form, the model parameters are α and b that have been determined in Equation (4). This Equation is a set of N+1 linear equations in N+1 unknowns ($\alpha_1, \alpha_2, \ldots, \alpha_N$ and b). The size is therefore not determined by the dimensions of the input space but by the number of objects in the training set. Finally, in order to optimize model performance, the hyperparameters γ and σ can be determined e.g. using crossvalidation on the training set.

The LS-SVM classifier model can be interpreted as a kernel version of Fisher discriminant analysis (Suykens et al, 2002).

C Generalized Estimating Equations

Generalized estimating equations are an extension of the logistic regression model (Diggle et al, 2001; Lipsitz et al, 1991; Lipsitz et al, 1994). Where in the case of logistic regression the data points are assumed to be independent of each other, generalized estimating equations can handle clustered data. For example, suppose we have the following binary outcome: a premature baby survives the birth process – yes or no. If there are only babies of single pregnancies in the data, we can apply ordinary logistic regression. But when there are also babies of twins or triplets in the dataset, we should take into account that in a multiple pregnancy the outcome of one baby influences the outcome of the other. In other words, babies of twins or triplets do not have independent outcomes - each twin or triplet is a cluster of data. Generalized estimating equations can handle such clustered data.

The generalized estimating equation for estimating the vector of regression parameters β is given by

$$S(\beta)=\sum_{i=1}^{K}\frac{\partial\mu_i^T}{\partial\beta}V_i^{-1}(Y_i-\mu_i(\beta))=0$$

with K the number of subjects (i.e. pregnancies), Y_i the vector of measurements on the i-th subject, μ_i the corresponding vector of means, and V_i the covariance matrix of Y_i. In case of our binary responses is Y_i the vector of the outcomes on the i-th pregnancy and μ_i (which we will denote as p_i) the corresponding vector of probabilities. The objective function is then

$$log\frac{p_{ij}}{1-p_{ij}}=\beta_0+\beta_1X_{1ij}+\beta_2X_{2ij}+...+\beta_kX_{kij}\ ,$$

with p_{ij} the probability of survival of the birth process for the j-th baby of the i-th pregnancy, X_{1ij} the value of the first explanatory variable for the j-th baby of the i-th pregnancy, and X_{kij} the value for the k-th explanatory variable for the j-th baby of the i-th pregnancy.

Let $R_i(\alpha)$ be a "working" correlation matrix that is fully specified by the vector of parameters α. The covariance matrix of Y_i is then modeled as

$$V_i=A_i^{1/2}R_i(\alpha)A_i^{1/2},$$

where A_i is an diagonal matrix with $v(\mu_{ij})$ - the estimated variance of Y_{ij} - as the j-th diagonal element. The working correlation matrix is usually unknown and must be estimated. It is estimated in an iterative fitting process. The jk-th element of the working correlation matrix of the i-th pregnancy $R_i(\alpha)$ is an estimate of the correlation between baby j and baby k of that pregnancy. Some of the more popular ones are corr(Y_{ij}, Y_{ik}) = 0 (independence), corr(Y_{ij}, Y_{ik}) = α (exchangeable, i.e. baby j and baby k of the i-th twin have the same influence on each other), corr(Y_{ij}, Y_{ik}) = α_{jk} (unstructured, i.e. baby j can have a different influence on baby k than baby k has on baby j).

D Cox's Proportional Hazards Model

The Cox's proportional hazards model is the most commonly used regression model for survival data (Cox, 1972 – 1975; Kalbfleisch, 1980; Marubini and Valsecchi, 1997).

We start with defining two functions that describe the distribution of the survival times of the subjects, and then shortly explain the proportional hazards model.

Let T be a nonnegative random variable representing the failure time of an individual from a homogeneous population. The survival distribution function (also known as the survivor function) of T is defined as

$$S(t)=P(T\geq t).$$

A mathematically equivalent way of specifying the distribution of T is through its hazard function. The hazard function $\lambda(t)$ specifies the instantaneous failure rate at t. If T is a continuous random variable, $\lambda(t)$ is expressed as

$$\lambda(t)=\lim_{\Delta t\to 0^+}\frac{P(t\leq T<t+\Delta t\ |T\geq t)}{\Delta t}=\frac{f(t)}{S(t)},$$

where $f(t)$ is the probability density function of T.

Cox's semi parametric proportional hazards model is widely used in the analysis of survival data to explain the effect of explanatory variables on hazard rates. The survival time of each member of a population is assumed to follow its own hazard function, $\lambda_i(t)$, expressed as

$$\lambda_i(t)=\lambda_0(t)\exp(\beta_1X_{1i}+...+\beta_kX_{ki}),$$

where $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function, X_{1i}, ..., X_{ki} are the values of the k explanatory variables for the ith individual and β_1, ..., β_k are the regression parameters associated with the explanatory variables. Those parameters are assumed to be the same for all individuals. The survivor function of the ith individual can then be expressed as

$$S_i(t)=[S_0(t)]^{\exp(\beta_1X_{1i}+...+\beta_kX_{ki})}$$

where

$$S_0(t)=\exp\left(-\int_0^t\lambda_0(u)du\right)$$

is the baseline survivor function. To estimate the parameters β_1, ..., β_k Cox (1972, 1975) introduced the partial likelihood function, which eliminates the unknown baseline hazard $\lambda_0(t)$ and accounts for censored survival times.

The partial likelihood of Cox also allows time-dependent explanatory variables. An explanatory variable is time-dependent if its value for any given individual can change over time. Time-dependent variables have many applications in survival analysis. For example, we can use a time-dependent

variable to model the effect of subjects changing from one treatment group to the other.

E International Ovarian Tumor Analysis group

IOTA Steering Committee

Dirk Timmerman, Lil Valentin, Thomas H. Bourne, William P. Collins, Sabine Van Huffel, Ignace Vergote

IOTA principal investigators

(in alphabetical order)

Jean-Pierre Bernard, Maurepas, FRANCE.

Thomas H. Bourne, London, U.K.

Enrico Ferrazzi, Milan, ITALY.

Davor Jurkovic, London, UK

Fabrice Lécuru, Paris, FRANCE.

Andrea Lissoni, Monza, ITALY.

Ulrike Metzger, Paris, FRANCE.

Dario Paladini, Napels, ITALY.

Antonia Testa, Roma, ITALY.

Dirk Timmerman, Leuven, BELGIUM.

Lil Valentin, Malmö, SWEDEN.

Caroline Van Holsbeke, Leuven, BELGIUM.

Sabine Van Huffel, Leuven, BELGIUM.

Ignace Vergote, Leuven, BELGIUM.

Gerardo Zanetta†, Monza, ITALY.

Bibliography

Aerts S, Antal P, Timmerman D, et al. Web based data collection for ovarian cancer: a case study. *Proceedings of the 15th IEEE Symposium on Computer Based Medical Systems (CBMS)*, Maribor, Slovenia, 2002.

Agresti A. *Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics, 1996.

Akaike H. Likelihood of model and information criteria. *Journal of econometrics*, 16: 3-14, 1981.

Alcazar JL, Errasti T, Laparte C, Jurado M, Lopez-Garcia G. Assessment of a new logistic model in the preoperative evaluation of adnexal masses. *J Ultrasound Med.* 20(8):841-8, 2001.

Allison PD. *Survival Analysis using the SAS System: A Practical Guide*. SAS Institute Inc., Cary, NC, USA, 1995.

Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*, 19:453-473, 2000.

Ameye L, De Brabanter J, Suykens JAK, Cadron I, Devlieger R, Timmerman D, Spitz B, Van Huffel S. Predictive models for survival after premature rupture of membranes. Internal Report 04-116, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004.

Ameye L, De Brabanter J, Suykens JAK, Cadron I, Devlieger R, Timmerman D, Spitz B, Van Huffel S. Predictive Models for Long Term Survival after Premature Rupture of Membranes. *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBS), 1-4 September 2005, Shanghai, China, in press.

Ameye L, Lu C, Lukas L, De Brabanter J, Suykens JAK, Van Huffel S, Daniels H, Naulaers G, Devlieger H. Prediction of mental development of

preterm newborns at birth time using LS-SVM. *Proceedings of the European Symposium on Artificial Neural Networks* (ESANN'2002), Bruges, Belgium, pp. 167-172, Apr. 2002.

Ameye L, Van Huffel S, De Brabanter J, Suykens J, Spitz B, Cadron I, Devlieger R, Timmerman D. Study of rupture of membranes before 26 weeks of gestation. *Proceedings of the 2nd European Medical and Biological Engineering Conference* (EMBEC'02), Vienna, Austria, pp. 760-761, Dec. 2002.

Antal P, Fannes G, Timmerman D, De Moor B, Moreau U. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artif Intell Med*, 29:39-60, 2003.

Antal P, Fannes G, Timmerman D, Moreau Y, De Moor B. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artif Intell Med*, 30: 257-281, 2004.

Aslam N, Banerjee S, Carr JV, Savvas M, Hooper R, Jurkovic D. Prospective evaluation of logistic regression models for the diagnosis of ovarian cancer. *Obstet Gynecol,* 96(1):75-80, 2000.

Atkinson AC. A note on the generalized information criterion for choice of a model. Biometrika, 67: 413-418, 1980.

Bailey C, Ueland F, Land GL, et al: The malignant potential of small cystic ovarian tumors in women over 50 years of age. *Gynecol Oncol,* 69:3-7, 1998

Banks J. Nomograms. In Kotz S and Johnson NL, editors, *Encyclopedia of Statistical Sciences*, Volume 6. Wiley, New York, 1985.

Bast RC, Jr, Klug TL, St John E et al. A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. *N Eng J Med*, 309: 883-8, 1983.

Beydoun SN, Yasin SY. Premature rupture of the membranes before 28 weeks: conservative management. *Am J Obstet Gynecol*, 155: 471, 1986.

Bishop CM. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

Bourne T. Screening for early familial ovarian cancer and the diagnosis of endometrial and ovarian pathology using transvaginal ultrasonography with and without color Doppler. Thesis. Göteborg. Sweden. 1995.

Bourne TH, Campbell S, Reynolds K et al. The potential role of serum CA 125 in an ultrasound-based screening program for familial ovarian cancer. *Gynecol Oncol*, 52: 379-385, 1994.

Bourne T, Campbell S, Steer CV, et al: Transvaginal colour flow imaging: a possible new screening technique for ovarian cancer. *Br Med J,* 299:1367-1370, 1989.

Bourne TH, Gruböck K, Tailor A. The study of ovarian tumors. In: Bourne TH, Jauniaux E, Jurkovic D (eds). *Transvaginal colour Doppler*, Springer-Verlag, Berlin, 1995.

Bozdogan H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52: 345-370, 1987.

Cadron I, Spitz B, Devlieger R, Ameye L, Van Huffel S, De Brabanter J, Suykens J, Timmerman D. Can we predict survival in cases of Preterm prelabor rupture of the membranes (PPROM) before 26 weeks of gestation? Internal Report 04-04, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004. Submitted to *Gynecol Obstet Inves*.

Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74: 829-836, 1979.

Coiera E. *A guide to Health Informatics*. Hodder & Stoughton Educational, UK, 2nd edition, 2003.

Coiera E. Artificial intelligence systems in routine clinical use. *Online*, 1999.

Cox DR. Regression models and life tables. *J Roy Statistical Society*, Series B, 20: 187-220, 1972.

Cox DR. Partial Likelihood. *Biometrika*, 62: 269-276, 1975.

Cox S, Williams L, Leveno K. The natural history of preterm ruptured membranes: what to expect of expectant management. *Obstet Gynecol*, 71, 558-561, 1988.

Cuckle HS, Wald NJ. Screening for ovarian cancer. In Miller AB, Chamberlain U, Day NE, Hakama M, Prorok PC (eds) *Cancer Screening*. pp 228-239. Cambridge: Cambridge University Press, 1991.

De Boor. *A practical Guide to Splines*. Springer-Verlag, New York, 1978.

De Long ER, De Long DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44: 837-845, 1988.

DePriest P, Gallion H, Pavlik E et al. Transvaginal sonography as a screening method for the detection of early ovarian cancer. *Gynecol Oncol*, 65: 408-414, 1997.

Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press, 2001.

Edwards A, Clarke L, Piessens S, Graham E, Shekelton P. Acoustic streaming: a new technique for assessing adnexal cysts. *Ultrasound Obstet Gynecol*, 22(1): 74-8, 2003.

Ekerhovd E, Wienerroith H, Staudach A, Granberg S. Preoperative assessment of unilocular adnexal cysts by transvaginal sonography: A comparison between sonographic morphological imaging and histopathologic diagnosis. *Am J Obstet Gynecol*, 184: 48-54, 2001.

Fannes G. Bayesian learning with expert knowledge: *Transforming informative priors between Bayesian networks and multilayer perceptrons*. PhD dissertation, Katholieke Universiteit Leuven, Faculty of Engineering, Department of Electrical Engineering, Leuven, Belgium, 2004.

Ferrazzi E; Zanetta G; Dordoni D et al. Transvaginal ultrasonographic characterization of ovarian masses: comparison of five scoring systems in a multicenter study. *Ultrasound Obstet Gynecol*, 10: 192-197, 1997.

Finkler NJ, Benacerraf B, Lavin PT et al. Comparison of serum CA 125, clinical impression, and ultrasound in the preoperative evaluation of ovarian masses. *Obstet Gynecol*, 72: 659-664, 1988.

Fisher RA. Statistical methods for research workers, 14th edition. Edinburgh, Oliver and Boyd, Ltd, 1970.

Fortunato SJ, Welt SI, Eggleston MK Jr, Bryant EC. Active expectant management in very early gestations complicated by premature rupture of the fetal membranes. *J Reprod Med*, 39: 13, 1994.

Gabriel KR. The biplot graphical display of matrices with applications to principal component analysis. *Biometrika*, 58: 453-467, 1971.

Gerber B, Muller H, Kulz T, et al: Simple ovarian cysts in premenopausal patients. *Int J Gynaecol Obstet*, 57:49-55, 1997.

Gibbs RS, Blanco JD. Premature rupture of the membranes. *Obstet Gynecol*, 60: 671-673, 1982.

Goldstein SR, Subramanyam B, Snyder J, et al: The postmenopausal cystic adnexal mass: the potential role of ultrasound in conservative management. *Obstet Gynecol*, 73:8-10, 1988

Graham L, Gilstrap LC, Hauth JC, Kodeck-Garza S. Conservative management of patients with premature rupture of fetal membranes. *Obstet Gynecol*, 59: 607-610, 1982.

Granberg S, Ekerhovd E, Timmerman D, Bourne T. The use of ultrasound to assess the morphology of ovarian tumors. In Timmerman D, Deprest J, Bourne TH (eds) *Ultrasound and Endoscopic Surgery in Obstetrics and Gynecology*. Berlin: Springer-Verlag, 2002.

Granberg S, Norström A, Wikland M. Tumors in the Lower Pelvis as Imaged by Transvaginal Sonography. *Gynecol Oncol*, 37: 224-229, 1990.

Granberg S., Wikland M., Jansson I. Macroscopic characterization of ovarian tumors and the relation to the histological diagnosis: Criteria to be used for ultrasound evaluation. *Gynecol Oncol* 35: 139-144, 1989.

Hadi HA, Hodson CA, Strickland D. Premature rupture of membranes between 20 and 25 weeks' gestation: role of amniotic fluid volume in perinatal outcome. *Am J Obstet Gynecol*, 170: 1139, 1994.

Hanley AJ, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 143(1): 29-36, 1982.

Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: Determing relationships between predictors and response. *J Nat Ca Instit*, 80: 1198-1202, 1988.

Harrell FE Jr. *Regression Modeling Strategies*. Springer Series in Statistics, Springer – Verlag, New York, 2002.

Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Monographs on statistics and applied probability. Chapmann and Hall/CRC, Boca Raton, Florida, 1999.

Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning – data mining, inference and prediction*. Springer, New York, 2001.

Heintz APM, Odicino F, Maisonneuve P, et al: Carcinoma of the Ovary, in 25th Annual Report on the Results of Treatment in Gynecological Cancer. *Int J Gynecol Obstet*, 83: S135-S137 (suppl 1), 2003.

Hermann UJ, Locher GW, Goldhirsch A. Sonographic patterns of ovarian tumors: Prediction of malignancy. *Obstet Gynecol*, 69: 777-781, 1987.

Hillaby K, Aslam N, Salim R, Lawrence A, Raju KS, Jurkovic D. The value of detection of normal ovarian tissue (the 'ovarian crescent sign') in the differential diagnosis of adnexal ,asses. *Ultrasound Obstet Gynecol*, 23(1): 63-67, 2004.

Hosmer DW Jr., Lemeshow S. *Applied Logistic Regression*. John Wiley & Sons, Inc., 2001.

Jacobs I, Bast RC. The CA 125 tumour-associated antigen: a review of the literature. *Hum Reprod*, 4: 1-12, 1989.

Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzingskas. A risk of malignancy index incorporating CA125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br J Obstet Gynaecol*, 97:920-929, 1990.

Jain A, Jain A, Jain S, Jain L. *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*, volume 39 of *Series in Machine Perception & Artificial Intelligence*. World Scientific, 2000.

Jermy K, Luise C, Bourne T. The characterization of common ovarian cysts in premenopausal women. *Ultrasound Obstet Gynecol*, 17: 140-144, 2001.

Justice AC, Covinsky KE, Berline JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 130:515-524, 1999.

Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. John Wiley & Sons Inc., New York, 1980.

Kehoe S, Powell J, Wilson S, Woodman C. The influence of the operating surgeon's specialization on patient survival in ovarian carcinoma. *Br J Cancer*, 70: 1014-1017, 1994.

Kilbride HW, Yeast J, Thibeault DW. Defining limits of survival: Lethal pulmonary hypoplasia after midtrimester premature rupture of membranes. *Am J Obstet Gynecol*, 175: 675, 1996.

Knapp RC, Jacobs I, Schwartz PE. Clinical perspectives in using CA 125. *Contemp Obstet Gynecol*, 1-7, 1996.

Kroon E, Andolf E: Diagnosis and follow-up of simple ovarian cysts detected by ultrasound in postmenopausal women. *Obstet Gynecol*, 85:211-214, 1995.

Lerner J.P., Timor-Tritsch I.E., Federman A., Abramovich G. Transvaginal ultrasonographic characterization of ovarian masses with an improved, weighted scoring system. *Am J Obstet Gynecol*, 170: 81-85, 1994.

Lewis BV, Timmerman D. Complications de la chirurgie coelioscopique [courrier des lecteurs]. *J Gynécol Obstét Biol Reprod*, 22: 202, 1993.

Lipsitz SH, Kim K, Zhao L. Analysis of repeated categorical data using generalized estimating equations, *Stat Med*, 13: 1149-1163, 1994.

Lipsitz SH, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association, *Biometrika*, 78: 153-160, 1991.

Lubsen J, Pool J, Van Der Does E. A practical device for the application of a diagnostic or prognostic function. *Method Inf Med*, 17: 127-129, 1978.

Lu C. *Probabilistic machine learning approaches to medical classification problems*. PhD dissertation, Katholieke Universiteit Leuven, Faculty of Engineering, Department of Electrical Engineering, Leuven, Belgium, 2005.

Lu C, Suykens J, Van Huffel S, Vergote I, Timmerman D. Chapter 11: Linear and nonlinear preoperative classification of ovarian tumors. In Ichimura T, Yoshida K, editors, *Knowledge Based Intelligent System for Health Care*, volume 7 of *International Series on Advanced Intelligence*, pages 343-382, Australia, 2004.

Lu C, Van Gestel T, Suykens J, Van Huffel S, Vergote I, Timmerman D. Preoperative prediction of malignancy of ovarian tumor using least squares support vector machines. *Artif Intell Med*, 28(3):281-306, 2003.

Maggino T, Gadducci A, D'Addario V et al. Prospective multicenter study on CA 125 in postmenopausal pelvic masses. *Gynecol Oncol*, 54: 117-23, 1994.

Major CA, Kitzmiller JL. Perinatal survival with expectant management of midtrimester rupture of membranes. *Am J Obstet Gynecol*, 163: 838, 1990.

Marubini E, Valsecchi G. *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley & Sons Inc., England, 1997.

Mead PB. Management of the patient with premature rupture of the fetal membranes: a review. *Clin Perinatol*, 7: 243-260, 1980.

Miller RA. Medical diagnostic decision support systems – past, present and future: a threaded bibliography and brief commentary. *J Am Med Inform Assoc*, 1(1): 8-27, 1994.

Mol BWJ, Boll D, De Kanter M, et al.:Distinguishing the benign and malignant adnexal mass: an external validation of prognostic models. *Gynecol Oncol*, 80:162-167, 2001

Morales L, Neven P, Timmerman D, Christiaens M-R, Vergote I, Van Limbergen E, Carbonez A, Van Huffel S, Ameye L, Paridaens R. Acute effects of tamoxifen and third generation aromatase inhibitors on menopausal symptoms of breast cancer patients. *Anti-Cancer Drugs*, 15(8): 753-760, 2004.

Morales L, Neven P, Timmerman D, Konstantinovic ML, Vergote I, Christiaens MR, Van Limbergen E, Weltens C, Carbonez A, Van Huffel S, Ameye L, Paridaens R. Menopausal symptoms and uterine changes in postmenopausal breast cancer patients receiving tamoxifen or third generation aromatase inhibitors. *Eur J Cancer (Suppl)*, 2(9): 19-21, 2004.

Morales L, Timmerman D, Neven P, Konstantinovic ML, Carbonez A, Van Huffel S, Ameye L, Weltens C, Christiaens MR, Vergote I, Paridaens R. Third generation aromatase inhibitors may prevent endometrial growth and reverse tamoxifen-induced uterine changes in postmenopausal breast cancer patients. *Ann Oncol*, 16: 70-74, 2005.

Morales WJ, Talley T. Premature rupture of membranes at 25 weeks: a management dilemma. *Am J Obstet Gynecol*, 168: 503, 1993.

Morel NM, Holland JM, Van Der Greef J, Marple EW, Clish C, Loscalzo J, Naylor S. Primer on medical genomics part xiv: introduction to systems biology – a new approach to understanding disease and treatment. In *Mayo Clin Proc*, volume 79, pages 651-658. Mayo Foundation for Medical Education and Research, May 2004.

Moretti M, Sibai BM. Maternal and neonatal outcome of expected management of premature rupture of membranes in the midtrimester. *Am J Obstet Gynecol*, 159: 390, 1988.

Mazumdar M, Glassman JR. Tutorial in biostatistics. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments, *Stat Med*, 19: 113-132, 2000.

Nezhat F, Nezhat C, Welander CE, Benigno B. Four ovarian cancers diagnosed during laparoscopy management of 1011 women with adnexal masses. *Am J Obstet Gynecol*, 167: 790-796, 1992.

Okaro E, Condous G, Khalid A, Timmerman D, Ameye L, Van Huffel S, Bourne T. The use of ultrasound based "soft markers" for the prediction of pelvic pathology in women with chronic pelvic pain - Can we reduce the need for laparoscopy? Internal Report 03-195, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2003. Submitted to *Br J Obstet Gynaecol.*

Obwegeser R, Deutinger J, Bernascheck G: The risk of malignancy with an apparently simple adnexal cyst on ultrasound. *Arch Gynecol Obstet*, 253:117-120, 1993

Osmers R. Sonographic evaluation of ovarian masses and its therapeutical implications [editorial]. *Ultrasound Obstet. Gynecol.* 8: 217-111, 1996.

Ozols RF, Rubin SC, Thomas GM, Robboy SJ. Epithelial ovarian cancer. In Hoskins WJ, Perez CA, Young RC, editors, *Principles and practice of gynecologic oncology*, Lippincott Williams and Wilkins, Philadelphia, pages 981-1058, 2000.

Pelckmans K, Suykens JAK, Van Gestel T, De Brabanter J, Lukas L, Hamers B, De Moor B, Vandewalle J. LS-SVMlab: a Matlab/C toolbox for Least Squares Support Vector Machines. K.U. Leuven Belgium, ESAT-SISTA, : Internal Report 02-44 (downloadable from http://www.esat.kuleuven.ac.be/sista/lssvmlab), 2002.

Petricoin EF III, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C; Fishman DA, Kohn EC, Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. Vol 359. Feb 16, 2002.

Rotschild A, Ling EW, Puterman ML, Farquharson D. Neonatal outcome after prolonged preterm rupture of the membranes. *Am J Obstet Gynecol*, 162: 46, 1996.

SAS Institute Inc. *Logistic Regression Examples Using the Sas System*, Cary, NC: Sas Institute Inc., 1995.

Sassone AM, Timor-Tritsch IE, Artner A et al. Transvaginal sonographic characterisation of ovarian disease: evaluation of a new scoring system to predict ovarian malignancy. *Obstet Gynecol*, 78: 70-76, 1991.

Schölkopf B, Smola AJ. *Learning with kernels*. MIT Press, London, England, 2002.

Schrijvers D, van Dam P, van Oosterom A. CA 125: a misleading tumor marker? [brief communication]. *Int J Gyn Obstet*, 55: 179-80, 1996.

Schuker JL, Mercer BM. Midtrimester premature rupture of the membranes. *Semin Perinatol*, 20: 389, 1996.

Schwartz PE. The role of tumor markers in the preoperative diagnosis of ovarian cysts. *Clin Obstet Gynecol*, 36: 384-394, 1993.

Schwarz G. Estimating the dimension of a model. *Annals of Statistics*, 6: 461-464, 1978.

Sladkevicius P, Valentin L. Interobserver agreement in the results of Doppler examinations of extrauterine tumors. *Ultrasound Obstet Gynecol*, 6:91-96, 1995.

Spitz B, Vossen C, Devlieger R, Van Assche FA. Rupture of membranes before 26 weeks of gestation. *J Perinat Med,* 27: 451-457, 1999.

Suykens JAK, Vandewalle J, De Moor B. *Artificial Neural Networks for Modelling and Control of Non-Linear systems*. Kluwer Acad. Publ., Boston, 1995.

Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. *Least Squares Support Vector Machines*. World Scientific Publishing Co., Pte, Ltd. (Singapore), 2002.

Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Support vector machines: least squares approaches and extensions. In Suykens JAK, Horvath G, Basu S, Micchelli C, Vandewalle J, editors, *Advances in Learning Theory: Methods, Models and Applications,* volume 190 of *Nato Science Series III: Computer and Systems Sciences*, pages 155-177. IOS Press, 2003.

Suykens JAK, Vandewalle J. Least Squares support vector machines classifiers. *Neural Process Lett,* 9(3): 293-300, 1999.

Tailor A, Jurkovic D, Bourne TH et al. Sonographic prediction of malignancy in adnexal masses using an artificial neural network. *Br J Obstet Gynaecol*, 106: 21-30, 1999.
Tailor A, Jurkovic D, Bourne TH, Collins WP, Campbell S. Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. *Ultrasound Obstet Gynecol*, 10: 41-7, 1997.

Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Stat Med*, 17:891-908, 1998.

Taylor J, Garite TJ. Premature rupture of membranes before fetal viability. *Obstet Gynecol*, 64: 615, 1984.

Tekay A, Jouppila P. Controversies in assessment of ovarian tumors with transvaginal color Doppler ultrasound. *Acta Obstet Gynecol* Scand, 75: 316-329, 1996.

Tekay A, Jouppila P. Validity of pulsatility and resistance indices in classification of adnexal tumors with transvaginal color Doppler ultrasound. *Ultrasound Obstet Gynecol*, 2: 338-344, 1992.

Therneau TM, Grambsch PM. *Modeling Survival data*, extending the Cox Model. Springer, New York, 2000.

Timmerman D. Lack of standardization in gynecological ultrasonography (Editorial). *Ultrasound Obstet Gynecol* 16: 395-398, 2000.

Timmerman, D. The use of mathematical models to evaluate pelvic masses; can they beat an expert operator? *Best Pract. Res. Cl. Ob.,* 18: 91-104, 2004.

Timmerman D. *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology.* PhD Thesis (ISBN 906186 857 2). Leuven: Leuven University Press, 1997.

Timmerman D, Bourne TH, Tailor A, Collins WP, Verrelst H, Vandenberghe K, Vergote I. A comparison of methods for the preoperative discrimination between malignant and benign adnexal masses: The development of a new logistic regression model. *Am J Obstet Gynecol*, 181: 57-65, 1999.

Timmerman D, Konstantinovic ML, Vergote I. Diagnostic models for ovarian cancer. *British Medical Ultrasound Society Bulletin*, 11: 20-23, 2003.

Timmerman D, Moerman P, Vergote I. Meigs' syndrome with elevated serum CA 125 levels: two case reports and review of the literature. *Gynecol Oncol*, 59: 405-408; 1995.

Timmerman D, Schwärzler P, Collins WP, Claerhout F, Coenen M, Amant F, Vergote I, Bourne TH. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. *Ultrasound Obstet Gynecol*, 13:11-16, 1999.

Timmerman D, Testa A, Bourne T, Ameye L, Paladini D, Jurkovic D, Valentin L. Clinical lessons learned from the ultrasound examination of 1233 adnexal tumors in the IOTA multicenter study, 2005. In preparation.

Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L. A lo-

gistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis (IOTA) group. Technical report, 2004. Submitted to *J Clin Oncol.*

Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group. *Ultrasound Obstet Gynecol*, 16:500-505, 2000.

Timmerman D, Van Calster B, Jurkovic D, Valentin L, Testa AC, Bernard JP, Van Holsbeke C, Van Huffel S, Vergote I, Bourne T. Do measurements of serum CA125 help to distinguish between benign and malignant adnexal tumors? Outcomes from a multicentre study by the International Ovarian Tumour Analysis (IOTA) group, 2005. In preparation.

Timmerman D, Verrelst H, Bourne TH, De Moor B, Collins WP, Vergote I, Vandewalle J. Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. *Ultrasound Obstet Gynecol*, 13:17-25, 1999.

Timmerman D, Verrelst H, Collins WP et al . Distinguishing the benign and malignant adnexal mass: an external validation of prognostic models. Letter to the editor. *Gynecol Oncol*, 83: 166-168, 2001.

Timmerman D, Verrelst H, Van Holsbeke C et al. Prospective evaluation of a new risk of malignancy index. *Gynecol Oncol*, 76: 255, 2000.

Timor-Tritsch LE, Lerner JP, Monteagudo A, Santos R. Transvaginal ultrasonographic characterization of ovarian masses by means of color flow-directed Doppler measurements and a morphologic scoring system. *Am J Obstet Gynecol*, 168: 909-913, 1993.

Tingulstad S, Hagen B, Skjeldestad FE, Onsrud M, Kiserud T, Halvorsen T, Nustad K. Evaluation of a risk of malignancy index based on serum CA125, ultrasound findings and menopausal status in the pre-operative diagnosis of pelvic masses. *Br J Obstet Gynaecol*, 103:826-831, 1996.

Valentin L. Comparison of Lerner score, Doppler ultrasound examination, and their combination for discrimination between benign and malignant adnexal masses. *Ultrasound Obstet Gynecol.* 15(2):143-147, 2000.

Valentin L. Gray scale sonography, subjective evaluation of the color Doppler image and measurement of blood flow velocity for distinguishing benign and malignant tumors of suspected adnexal origin. *Eur J Obstet Gynaecol Reprod Biol*, 72:63-72, 1997.

Valentin L. Pattern recognition of pelvic masses by gray-scale ultrasound imaging: the contribution of Doppler ultrasound. *Ultrasound Obstet Gynecol*, 14: 338-347, 1999.

Valentin L. Prospective cross-validation of Doppler ultrasound examination and gray-scale ulrasound imaging for discrimination of benign and malignant pelvic masses, *Ultrasound Obstet Gynecol*, 14: 273 – 283, 1999.

Valentin L, Hagen B, Tingulstad S, Eik-Nes S. Comparison of 'pattern recognition' and logistic regression models for discrimination between benign and malignant pelvic masses: a prospective cross validation. *Ultrasound Obstet Gynecol*, 18: 357-365, 2001.

Valentin L, Sladkevicius P, Marsál K. Limited contribution of Doppler velocimetry to the differential diagnosis of extrauterine pelvic tumors. *Obstet Gynecol*, 83: 425-433, 1994.

Valentin L, Ameye L, Jurkovic D, Metzger U, Lécuru F, Van Huffel S, Timmerman D. Which extrauterine pelvic masses are difficult to correctly classify as benign or malignant on the basis of ultrasound findings, and is there a way of making a correct diagnosis?, 2005. Submitted to *Ultrasound Obstet Gynecol.*

Valentin L, Ameye L, Testa A, Lécuru F, Bernard JP, Paladini P, Van Huffel S, Timmerman D. Ultrasound characteristics of different types of adnexal malignancies, 2005. Submitted to *Ultrasound Obstet Gynecol.*

Valentin L, Ameye L, Testa A, Lécuru F, Bernard JP, Paladini D, Van Huffel S, Timmerman D. Ultrasound characteristics of benign and malignant pelvic extrauterine tumors containing papillary projections, 2005. In preparation.

Van Bemmel JH, Musen MA, editors. *Handbook of Medical Informatics. Springer*, 1998.

Van den Bosch T, Donders G, Riphagen I, Debois P, Ameye L, De Brabanter J, Van Huffel S, Van Schoubroeck D, Timmerman D. Ultrasonographic features of the endometrium and the ovaries in women on etonogestrel implant. *Ultrasound Obstet Gynecol*, 20: 377-380, 2002.

Van Den Bosch T, Van Schoubroeck D, Ameye L, De Brabanter J, Van Huffel S, Timmerman D. The influence of hormone replacement therapy on the ultrasound features of the endometrium : a prospective study. *Eur J Cancer*, 38(6): S78-S79, 2002.

Van Den Bosch T, Van Schoubroeck D, Ameye L, De Brabanter J, Van Huffel S, Timmerman D. Ultrasound assessment of endometrial thickness and endometrial polyps in women on hormonal replacement therapy. *Am J Obstet Gynecol*, 188(5): 1249-1253, 2003.

Van Den Bosch T, Van Schoubroeck D, Ameye L, Van Huffel S, Timmerman D. Ultrasound examination of the endometrium before and after Pipelle enodmetrial sampling. Internal Report 04-222a, ESAT-SISTA,

K.U.Leuven (Leuven, Belgium), 2004. Accepted for publication in *Ultrasound Obstet Gynecol.*

Van der Zee AGJ, Duk JM, Aalders JG et al. The effect of abdominal surgery on the serum concentration of the tumour-associated antigen CA 125. *Br J Obstet Gynaecol*, 97: 934-8, 1990.

Van Gestel T, Suykens JAK, Baesens B, Viaene S, Vanthienen J, Dedene G, De Moor B, Vandewalle J. Benchmarking Least Squares Support Vector Machines Classifiers. *Machine Learning*, 54(1) 5-32, 2004.

Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Stat Med*, 8: 1303-1325, 1990.

Vapnik V. *The nature of statistical learning theory*. Springer-Verlag, 1995.

Vergote I, De Brabanter J, Fyles A, Bertelsen K, Einhorn N, Sevelda P, Gore ME, Kaern J, Verrelst H, Sjovall K, Timmerman D, Vandewalle J, Van Gramberen M, Trope CG. Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. *Lancet*, 357(9251):176-182, 2001.

Vergote IB, Børmer O, Abeler VM. Evaluation of CA 125 in the monitoring of ovarian cancer. *Am J Obstet Gynecol*, 157: 88-92, 1987.

Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *J Am Stat Ass*, 84: 1065-1073, 1989.

Publication list

Published/accepted articles in international journals

Morales L, Neven P, Timmerman D, Christiaens M-R, Vergote I, Van Limbergen E, Carbonez A, Van Huffel S, **Ameye L**, Paridaens R. Acute effects of tamoxifen and third generation aromatase inhibitors on menopausal symptoms of breast cancer patients. *Anti-Cancer Drugs*, 15(8): 753-760, 2004.

Morales L, Neven P, Timmerman D, Konstantinovic ML, Vergote I, Christiaens MR, Van Limbergen E, Weltens C, Carbonez A, Van Huffel S, **Ameye L**, Paridaens R. Menopausal symptoms and uterine changes in postmenopausal breast cancer patients receiving tamoxifen or third generation aromatase inhibitors. *European Journal of Cancer Supplements*, 2(9): 19-21, 2004.

Morales L, Timmerman D, Neven P, Konstantinovic ML, Carbonez A, Van Huffel S, **Ameye L**, Weltens C, Christiaens MR, Vergote I, Paridaens R. Third generation aromatase inhibitors may prevent endometrial growth and reverse tamoxifen-induced uterine changes in postmenopausal breast cancer patients. *Ann Oncol*, 16: 70-74, 2005.

Van Den Bosch T, Donders G, Riphagen I, Debois P, **Ameye L**, De Brabanter J, Van Huffel S, Van Schoubroeck D, Timmerman D. Ultrasonographic features of the endometrium and the ovaries in women on etonogestrel implant. *Ultrasound Obstet Gynecol*, 20: 377-380, 2002.

Van Den Bosch T, Van Schoubroeck D, **Ameye L**, De Brabanter J, Van Huffel S, Timmerman D. The influence of hormone replacement therapy on the ultrasound features of the endometrium : a prospective study. *Eur J Cancer*, 38(6): S78-S79, 2002.

Van Den Bosch T, Van Schoubroeck D, **Ameye L**, De Brabanter J, Van Huffel S, Timmerman D. Ultrasound assessment of endometrial thickness and endometrial polyps in women on hormonal replacement therapy. *Am J Obstet Gynecol*, 188(5): 1249-1253, 2003.

Van den Bosch T, Van Schoubroeck D, **Ameye L**, Van Huffel S, Timmerman D. Ultrasound examination of the endometrium before and after Pipelle endometrial sampling. Internal Report 04-222a, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004. Accepted for publication in *Ultrasound Obstet Gynecol.*

Submitted articles in international journals

Cadron I, Spitz B, Devlieger R, **Ameye L**, Van Huffel S, De Brabanter J, Suykens J, Timmerman D. Can we predict survival in cases of Preterm prelabor rupture of the membranes (PPROM) before 26 weeks of gestation? Internal Report 04-04, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004. Submitted to *Gynecol Obstet Inves.*

Okaro E, Condous G, Khalid A, Timmerman D, **Ameye L**, Van Huffel S, Bourne T. The use of ultrasound based "soft markers" for the prediction of pelvic pathology in women with chronic pelvic pain - Can we reduce the need for laparoscopy? Internal Report 03-195, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2003. Submitted to *Br J Obstet Gynaecol.*

Timmerman D, Testa A, Bourne T, **Ameye L**, Paladini D, Yurkovic D, Valentin L. Clinical lessons learned from the ultrasound examination of 1233 adnexal tumors in the IOTA multicenter study, 2005. In preparation.

Timmerman D, Testa AC, Bourne T, Ferrazzi E, **Ameye L**, Konstantinovic ML, Collins WP, Vergote I, Van Huffel S, Valentin L. A logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis (IOTA) group. Submitted to the *J Clin Oncol.*

Valentin L, **Ameye L**, Jurkovic D, Metzger U, Lécuru F, Van Huffel S, Timmerman D. Which extrauterine pelvic masses are difficult to correctly classify as benign or malignant on the basis of ultrasound findings, and is there a way of making a correct diagnosis? Submitted to *Ultrasound Obstet Gynecol.*

Valentin L, **Ameye L**, Testa A, Lécuru F, Bernard JP, Paladini P, Van Huffel S, Timmerman D. Ultrasound characteristics of different types of adnexal malignancies. Submitted to the *J Clin Oncol.*

Valentin L, **Ameye L**, Testa A, Lécuru F, Bernard JP, Paladini D, Van Huffel S, Timmerman D. Ultrasound characteristics of benign and malignant pelvic extrauterine tumors containing papillary projections. In preparation.

Articles in international conference proceedings

Ameye L, Lu C, Lukas L, De Brabanter J, Suykens JAK, Van Huffel S, Daniels H, Naulaers G, Devlieger H. Prediction of mental development of preterm newborns at birth time using LS-SVM. *Proceedings of the European Symposium on Artificial Neural Networks* (ESANN'2002), Bruges, Belgium, pp. 167-172, Apr. 2002.

Ameye L, Van Huffel S, De Brabanter J, Suykens J, Spitz B, Cadron I, Devlieger R., Timmerman D. Study of rupture of membranes before 26 weeks of gestation. *Proceedings of the 2nd European Medical and Biological Engineering Conference* (EMBEC'02), Vienna, Austria, pp. 760-761, Dec. 2002.

Ameye L, De Brabanter J, Suykens JAK, Cadron I, Devlieger R, Timmerman D, Spitz B, Van Huffel S. Predictive Models for Long Term Survival after Premature Rupture of Membranes. Internal Report 05-76, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2005. *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBS), 1-4 September 2005, Shanghai, China, in press.

Abstracts in conference proceedings

Ameye L, De Brabanter J, Van Huffel S, Suykens J, Daniels H, Naulaers G, Devlieger H. Study of Mental Development Index of premature babies. Abstract at Joint Study Day of Interuniversity Attraction Pole (IUAP) IV/2, Leuven, Belgium, December 5, 2000.

Ameye L, De Brabanter J, Van Huffel S, Suykens J, Daniels H, Naulaers G, Devlieger H. Prediction of Mental Development of preterm newborns at birth time. Abstract at 9th annual meeting of the Belgian Statistical Society (BSS 2001), Oostende, Belgium, October 12-13, 2001.

Ameye L, Suykens J, De Brabanter J, Van Huffel S. Nonlinear computer intensive models for survival data. Abstract at 10th annual meeting of the Belgian Statistical Society (BSS 2002), Kerkrade, The Netherlands, October 18-19, 2002.

Morales L, Timmerman D, Neven P, Konstantinovic ML, **Ameye L**, Van Huffel S, Vergote I, Paridaens R. Prospective study of uterine sonographic changes after 3 months endocrine treatment in postmenopausal breast cancer patients: tamoxifen is estrogenic while aromatase inhibitors induce atrophy and can reverse the effect of tamoxifen. Poster presentation at the 26th Annual San Antonio Breast Cancer Symposium, USA 3-6.12.2003, Abstract in Breast Cancer Res Treat 2003; 82, suppl. 1, 105-106.

Morales L, Neven P, Timmerman D, Konstantinovic ML, Van Huffel S, **Ameye L**, Vergote I, Paridaens R. Prospective study on the effect of endocrine treatments on menopausal symptoms in breast cancer patients. Poster presentation at the 26th Annual San Antonio Breast Cancer Symposium, USA 3-6.12.2003, Abstract in Breast Cancer Res Treat 2003, 82, suppl. 1, 106.

Okaro E, Condous G, Khalid A, Timmerman D, Rao S, **Ameye L**, Van Huffel S, Bourne T. Patient history and ultrasound findings: A logistic regression model for the prediction of pelvic pathology in women with chronic pelvic pain. Invited speaker at the 13th World Congress of Ultrasound in Obstetrics and Gynecology organised by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG), Paris 31.8-4.9.2003. Abstract in Ultrasound Obstet. Gynecol. 2003; 22, suppl. 1, 18-19.

Okaro E, Condous G, Khalid A, Alkatib M, Timmerman D, **Ameye L**, Van Huffel S, Bourne T. Can transvaginal ultrasound detect and localize endometriosis in women with chronic pelvic pain? Invited speaker at the 13th World Congress of Ultrasound in Obstetrics and Gynecology organised by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG), Paris 31.8-4.9.2003. Abstract in Ultrasound Obstet. Gynecol. 2003; 22, suppl. 1, 19.

Timmerman D, Morales L, Konstantinovic ML, Moerman P, **Ameye L**, Van Huffel S, Paridaens R, Vergote, I. Evaluation of endometrial pathology in tamoxifen-treated breast cancer patients: the value of the 'pedicle artery test' using transvaginal color Doppler imaging. Invited speaker at the 13th International meeting of the European Society of Gynaecological Oncology (ESGO), Brussels, 6-10.4.2003, Abstract in J Gynecol Cancer; 13, suppl. 1, 67.

Timmerman D, Valentin L, Testa A, Ferrazzi E, **Ameye L**, Jurkovic D, Metzger U, Lécuru F, Bernard JP, Vergote I, Van Huffel S. Classifying ovarian masses and lessons from the IOTA trial. Invited lecture at the 14th World Congress of Ultrasound in Obstetrics and Gynecology, organised by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG), Stockholm, 31.8-4.9.2004. Abstract in Ultrasound Obstet. Gynecol. 2004; 24, 261.

Timmerman D, Testa AC, Bourne T, Ferrazzi E, **Ameye L**, Konstantinovic ML, Van Holsbeke C, Vergote I, Van Huffel S, Valentin L. A logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis (IOTA) Group. Abstract for the 14th International meeting of the

European Society of Gynaecological Oncology (ESGO), Istanbul, 25-29.9.2005.

Van den Bosch T, Donders G, Riphagen I, Debois P, **Ameye L**, De Brabanter J, Van Huffel S, Van Schoubroeck D, Timmerman D. Ultrasonographic features of the endometrium and the ovaries in women on etonogestrel implant. Sixièmes Journées Européennes de la Société Française de Gynécologie, Paris, 10-12.10.2002. Abstract published in Proceedings.

Van den Bosch T, Donders G, Riphagen I, Debois P, **Ameye L**, De Brabanter J, Van Huffel S, Van Schoubroeck D, Timmerman D. Ultrasonographic features of the endometrium and the ovaries in women on etonogestrel implant. Abstract at the 12th World Congress on Ultrasound in Obstetrics and Gynecology, New-York, USA, November 2002.

Internal reports

Ameye L, De Brabanter J, Suykens JAK, Cadron I, Devlieger R, Timmerman D, Spitz B, Van Huffel S. Predictive models for survival after premature rupture of membranes. Internal Report 04-116, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004.

Curriculum vitae

Personal information

Name:	Ameye Lieveke
Address:	Pereboomstraat 1, 8840 Westrozebeke
Email:	Lieveke.Ameye@esat.kuleuven.ac.be
Date of birth:	27/12/1977
Place of birth:	Roeselare
Nationality:	Belgian

Studies

Secondary education

1989-1995 Latin - Mathematics
Bisschoppellijk Lyceum der Grauwe Zusters, Roeselare

University studies - Katholieke Universiteit Leuven

1995-1997	candidature mathematics	KULAK	magna cum laude
1997-1999	licence (applied) mathematics	KUL	cum laude
1999-2000	ISPS – Master in Statistics	KUL	magna cum laude

Work experience

From 01/10/2000: PhD student of the Katholieke Universiteit Leuven at the Biomed group of ESAT-SCD(SISTA), under the supervision of Prof. dr. ir. Sabine Van Huffel, Prof. dr. ir. Johan Suykens and Prof. Dr. Dirk Timmerman