Risk-based management in obstetrics

Ewoud Schuit

The printing of this thesis is financially supported by: Department of Obstetrics and Gynecology, Academic Medical Center Amsterdam, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, and BV BMA (Mosos).

ISBN: 978-90-5335-665-4

Cover design: Nikki Vermeulen, Ridderprint BV, Ridderkerk, the Netherlands
Layout: Simone Vinke, Ridderprint BV, Ridderkerk, the Netherlands
Printing: Ridderprint BV, Ridderkerk, the Netherlands

Risk-based management in obstetrics

Risicogestuurd beleid in de verloskunde
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector manificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op
maandag 25 maart 2013 des middags te 4.15 uur

door

Ewoud Schuit

geboren op 14 april 1985
te Gorinchem

Promotoren:	Prof.dr. K.G.M. Moons
	Prof.dr. B.W.J. Mol
Co-promotoren:	Dr. A. Kwee
	Dr. R.H.H. Groenwold

CONTENTS

Chapter 1

General introduction

Risk-based management in obstetrics

Nowadays, medicine is increasingly focused on risk-based or personalized medicine. This largely means that treatment, preventive and other- e.g. diagnostic or prognostic- interventions are administered based on individual risks for having or developing a certain health outcome. In times where health care costs increase and budgets remain the same or are even cut down, there is a need to prioritize more costly and burdening care to individuals who indeed need and may benefit from such care. Identification of those individuals can be based on an individual risk factor, risk indicator or predictor (e.g. a particular patient characteristic, biomarker or test result) or on combinations of such factors. The prevailing thought is that combinations of predictors into a so-called risk prediction or decision model allows for better risk assessments and thus patient selection than when based on individual factors. In this, obstetrics are not different from other medical fields. A historical example of a prediction model and risk-based management in obstetrics is the individualized prediction of a poor neonatal outcome using the Apgar score. This score was developed in 1952 by Virginia Apgar and combines 5 clinical signs in neonates into a single and easy-to-interpret score.(1) Where obstetrics used to be a front-runner of individual risk assessment with the introduction of the Apgar score(1), other fields in medicine are now more widely applying prediction models and risk-based management. For example, the Framingham Risk Score (2) is in various countries routinely applied to determine the 10-year risk of cardiovascular disease and the EUROscore is an often used tool to predict the risk of 30-day mortality after cardiac surgery.(3)

The Apgar score, however, remains an almost paradigmatic example of how readily available predictors combined in a risk prediction model can be of help to allow for timely prognostication and consequently more effective management. Such models may serve as an alert (e.g., on the fact that a child needs immediate medical care), be used in an individual decision of further management (e.g., to refrain from intensive medical care when the chance of survival of a very preterm born child is low), or they could aid in organizational aspects (e.g. availability of doctors who can perform an instrumental delivery and availability of operating theatre and personnel when the risk of an intervention is high). Finally, in terms of risk-based patient care, prediction models allow for more individualised counselling of patients, e.g. in the decision of a pregnant woman to start labour either at home or in the hospital under supervision of an independent midwife or general practitioner.(4-7)

Obviously, besides personalized risk based management, prediction models can also be useful to identify specific subgroups of patients that benefit more from a certain preventive, therapeutic or other additional (e.g. diagnostic) intervention. A first step towards identifying such subgroups is a so-called subgroup analysis. For example, the effect of progestogen on

adverse perinatal outcome among pregnant women with a cervical length < 25 mm may be different from the effect of progestogen among women with a cervical length ≥ 25 mm.(8) Individual therapeutic trials are generally designed to assess main effects of a specific treatment (for a specific health outcome), i.e. the trials are not designed and thus underpowered to assess the therapeutic effects in smaller subgroups. In a meta-analyses the information of multiple studies is combined, thus increasing the number of study participant available for analysis. This will increase the statistical power to detect subgroup effects. Preferably, such meta-analysis with the aim to detect subgroup effects is based on individual participant data, rather than aggregated study data, since the former does not rely on published data but on the original study data. As such, for example, in individual participant data meta-analyses selective outcome reporting is of a lesser issue and one may define similar subgroups in all included studies. Other advantages of individual participant data meta-analyses are that they allow for more flexibility regarding the inclusion and exclusion of individuals, the choice of outcome, and to apply the same statistical methods (e.g. for analysis or for handling missing observations) on all data.(9-13)

The studies presented in this thesis aim to study and enhance risk-based management in obstetrics. The overall goals were threefold:

1) to develop and validate prognostic prediction models to allow for individual risk assessment;
2) to identify groups of patients that may benefit from specific treatment(s) strategies using subgroup analysis within meta-analyses of individual patient data;
3) to evaluate methods currently used in prediction modelling, and in the assessment of subgroup effects.

The outline of the thesis parallels these three goals. **Chapter 2** focuses on prognosis in obstetrics. In the first part of this thesis we describe the development of three prediction models as well as two studies in which prediction models were validated. In **Chapter 2.1** we identified risk indicators for different reasons of referral during labour from primary to secondary care in women with a singleton term pregnancy starting labour in primary care to support caregivers in primary care. In **Chapter 2.2** and **2.3** we developed and externally validated a prediction model to predict the occurrence of neonatal metabolic acidosis at birth in women pregnant of a term singleton in cephalic presentation. **Chapter 2.4** describes the development of a multinomial prediction model to assess the mode of delivery in women pregnant of a term singleton in cephalic presentation. In **Chapter 2.5** we externally validated two prediction models that predict survival for very preterm infants known to be alive at the onset of labour and very preterm infants admitted for neonatal intensive care. The models were developed in the United Kingdom and we validated them in the Dutch population.

Chapter 3 of this thesis deals with individual participant data meta-analyses aimed at quantifying treatment effects in specific patient groups. This chapter contains a protocol for an individual participant data meta-analysis and the conduct of two individual participant data meta-analyses. **Chapter 3.1** and **3.2** contain the protocol and the results of an individual participant data meta-analysis on the effectiveness of progestogen treatment in the reduction of adverse perinatal outcome in women with a twin pregnancy. In **Chapter 3.3** we describe an individual participant data meta-analysis on the effectiveness of intrapartum foetal monitoring with cardiotocography in combination with ST-analysis compared to cardiotocography alone in the reduction of neonatal metabolic acidosis in women pregnant of a term singleton in cephalic presentation.

Chapter 4 contains two methodological studies that are related to the first two chapters. **Chapter 4.1** depicts methods to address causes for unexpected predictor-outcome associations in prediction research. In **Chapter 4.2** we assess the performance of statistical analyses for meta-analysis of treatment effects, possibly treatment effects among subgroups of patients, when the analysis is conditional on a significant finding in one of the included studies that triggered the meta-analysis.

References

(1) APGAR V. A proposal for a new method of evaluation of the newborn infant. Curr Res Anesth Analg 1953 Jul;32(4):260-7.

(2) D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008 Feb 12;117(6):743-53.

(3) Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999 Jul;16(1):9-13.

(4) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

(5) Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

(6) Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012 May;98(9):691-8.

(7) Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012 May;98(9):683-90.

(8) Lim AC, Schuit E, Bloemenkamp K, Bernardus RE, Duvekot JJ, Erwich JJ, et al. 17alpha-hydroxyprogesterone caproate for the prevention of adverse neonatal outcome in multiple pregnancies: a randomized controlled trial. Obstet Gynecol 2011 Sep;118(3):513-20.

(9) Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. Stat Med 2012 Oct 15;31(23):2697-712.

(10) Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol 2007 May;60(5):431-9.

(11) Riley RD, Lambert PC, bo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.

(12) Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials 2005;2(3):209-17.

(13) Thompson SG, Higgins JP. Treating individuals 4: can meta-analysis help target interventions at individuals most likely to benefit? Lancet 2005 Jan 22;365(9456):341-6.

Chapter 2

Prognosis in obstetrics

Chapter 2.1

Risk indicators for referral during labour from primary to secondary care: a nationwide prospective cohort study

Schuit E, Hukkelhoven CWPM, van der Goes B, Overbeeke I,
Moons KGM, Mol BWJ, Groenwold RHH, Kwee A.

Submitted

Abstract

Objectives To identify risk indicators for referral during labour from primary to secondary care.

Design Cohort study using prospectively collected data from a national perinatal registry.

Setting Primary obstetric care in the Netherlands.

Participants Women with a singleton term pregnancy, starting labour in primary care between 2000 and 2007.

Main outcome measures Referral from primary to secondary care during labour due to foetal distress, meconium stained amniotic fluid, failure to progress in 1st and 2nd stage of labour, wish for pain relief, and a combination of other reasons.

Results Out of 754 727 women with a singleton term pregnancy 245 545 (33%) were referred from primary to secondary care during labour. Women were referred because of foetal distress (5%), failure to progress in 2nd stage of labour (14%), meconium stained amniotic fluid (24%), failure to progress in 1st stage of labour (17%), wish for pain relief (7%) or a combination of other less urgent reasons, e.g. malpresentation (33%). The strongest risk indicators for referral, independent of the reason, were a combination of parity and a previous instrumental vaginal delivery, gestational age, and the planned place of delivery. Additional risk indicators differed for the different reasons of referral.

Conclusions Among low risk women, complications during labour, including foetal distress, failure to progress in 1st and 2nd stage of labour, meconium stained amniotic fluid, wish for pain relief, and a combination of other reasons can be associated to readily available risk indicators. These risk indicators can contribute to the awareness of the risk of referral and they can be used during counselling of pregnant women for the intended place to start labour.

Introduction

In the Netherlands, pregnant women at low risk for complications have the choice to start labour either at home or in the hospital, supervised by independent midwives or general practitioners (primary care). Women are referred to and monitored in hospital by gynaecologists, clinical midwives or residents under the responsibility of gynaecologists (secondary care) if they are considered to be at high-risk either at the start of pregnancy (complicated by e.g. pre-existing maternal disease, adverse obstetric history), during pregnancy (e.g. intrauterine growth restriction, preeclampsia, a post date gestational age), during labour (e.g. failure to progress, need for pain relief, meconium stained amniotic fluid or non-reassuring foetal heart rate at intermittent auscultation) or after labour (e.g. complicated perineal tear or post partum haemorrhage). Annual reports show that about 55% of all pregnant women start labour in primary care and around 35% of these women is being referred to secondary care during labour.(1)

Currently, the so called List of Obstetric Indications is used in the decision of referral from primary to secondary care and contains indicators, identified by practicing evidence based medicine and by consensus amongst professionals. These indications are considered to be important in relation to adverse neonatal or maternal outcome(2;3). However, this list does not include other risk indicators that may not be related to adverse outcomes, but that may be related to a referral during labour. Knowledge of these indicators could potentially improve the quality of care by enhancing the identification of risk groups, better counselling of patients, individualised management and timely interventions. Furthermore, these risk indicators are interpreted in isolation, without combining several indicators that may influence the probability of referral.

Therefore, the aim of this study was to identify risk indicators for referral during labour from primary to secondary care which can be helpful in the counselling of patients.

Methods

Study population

We used data registered in the Netherlands perinatal registry, which is a linked database of medical registries from 3 professional organizations: the national midwives registry (primary care, LVR1), the national obstetricians' registry (secondary care, LVR2), and the national paediatricians/neonatologists registry (tertiary care, LNR).(4) The database contains 95% of all women who gave birth, resulting in high-quality data on almost all births in the Netherlands. (5) We studied women with a singleton pregnancy and gestational age between 37 and 42 weeks (term) between 1 January 2000 and 31 December 2007 that started labour in primary care (low-risk). Women who were referred because of prolonged rupture of membranes without contractions were not considered to be in labour and were therefore excluded.

Outcome

Women were assigned to one of seven distinctive outcome categories based on the main reasons of referral during labour from primary to secondary care during labour: no referral (reference category); suspected foetal distress, failure to progress in 2nd stage of labour, meconium stained amniotic fluid, failure to progress in 1st stage of labour, a wish for pain relief, either epidural anaesthesia or medication (opioids), or a combination of other reasons for referral, e.g. intra-uterine fetal death during labour, malpresentation (e.g. breech), blood loss during labour, abruptio placentae, vasa praevia, threatened preterm labour or other problems. Hence, the outcome of this study was multinomial or polytomous (i.e. more than two unordered outcome categories). Suspected foetal distress was defined as abnormal foetal heart rate during auscultation, where a normal heart rate was defined as 120-160 bpm, and no decelerations. According to the guidelines of the Royal Dutch Organization of Midwives, failure to progress in the 1st stage of labour was defined as a latent phase longer than eight hours, or an active phase with an increasing dilatation of less than 1 cm/hour for 2-4 hours. (6) Failure to progress in the 2nd stage of labour was defined as active pushing for more than 1 hour for nulliparous women and ½ hour for multiparous women. It must be noted that the definitions for a failure to progress are applied diversely. Referral after delivery (e.g. because of post partum haemorrhage, complicated perineal tear) was not considered as a transfer during labour. Women with multiple registered reasons for referral during labour were placed in the category deemed most urgent (from most to least urgent, as jointly decided by a multidisciplinary group of mid-wives and gynaecologists): suspected foetal distress > failure to progress in 2nd stage of labour > meconium stained amniotic fluid > failure to progress in 1st stage of labour > wish for pain relief > a combination of other reasons.

Potential risk indicators

We selected potential risk indicators for referral during labour based on literature and clinical reasoning. Selected candidate indicators included maternal age at childbirth, ethnicity, degree of urbanization, social economic status, a combination of parity and a history of instrumental vaginal delivery, planned place of delivery, gestational age at birth, neonatal gender, and birth weight. Maternal age and gestational age were registered as continuous characteristics but were categorized to increase clinical interpretation. Maternal age was categorized into three categories being < 25 years, 25 – 35 years (reference), or > 35 years. Gestational age at birth was categorized into three categories being 37^{+0}-37^{+6} weeks, 38^{+0} – 40^{+6} weeks (reference), ≥ 41^{+0} weeks. Ethnicity was categorized as Caucasian (reference), Creole, Hindu or other non-Western, e.g. Moroccan and Turkish. Degree of urbanization was defined as either being low (< 500 homes/km^2), average (>=500 homes/km^2 to 2 500 homes/km^2, reference) or high (>2 500 homes/km^2). Social economic status was categorized as low, average (reference) or high. Degree of urbanization and social economic status were

both based on zip code.(7) The combination of parity and a history of instrumental vaginal delivery was defined as women being multiparous without a previous instrumental vaginal delivery (reference), with a previous instrumental vaginal delivery, or being nulliparous. Instrumental vaginal delivery was defined as either vacuum or forcipal extraction or both in a previous pregnancy. Planned place of delivery was defined as a delivery under supervision of a midwife or general practitioner planned either at home (reference), in a midwife-led birth centre or in hospital. Birth weight was categorized into three categories being <p10(8), p10-4 000g (reference) and >4 000g. This categorisation was chosen because interest was in both children small for gestational age (< p10), being more vulnerable to foetal distress and children large for gestational age (> 4 000 g) giving more labour-related problems like failure to progress. Neonatal gender and birth weight are characteristics that are generally not available prior to or during labour, but since this study aimed to identify risk indicators we decided to include these characteristics.

Since the association between a risk indicator and referral from primary to secondary care during labour may change depending on another risk indicator (e.g. the risk of referral for women <25 years may change depending on parity), several interactions were investigated: interactions between maternal age and a combination of parity and a history of instrumental vaginal delivery, birth weight and a combination of parity and a history of instrumental vaginal delivery, birth weight and gestational age, birth weight and social economic status, birth weight and degree of urbanisation, birth weight and neonatal gender, birth weight and ethnicity, degree of urbanisation and social economic status, maternal age and planned place of delivery, ethnicity and planned place of delivery, degree of urbanisation and planned place of delivery, social economic status and planned place of delivery, and a three-way interaction between ethnicity, degree of urbanisation and social economic status.

Data analysis

Approximately 14% of all women had missing values for one or more of the risk indicators of interest. The proportion of missing values ranged from 175/779 682 (< 0.01%) for parity to 111 223/779 682 (14%) for planned place of delivery. Because these proportions were small and the sample size is very large, missing values were imputed using single imputation.(9)

Baseline characteristics of the study population were analysed using descriptive statistics and are presented as mean with standard deviation (SD) for continuous variables and as numbers and proportions (%) of the whole population for categorical and dichotomous variables.

To identify risk indicators we used data on women who started labour in primary care between January 1, 2000 and December 31, 2007. Maternal age at delivery, ethnicity, degree of urbanization, social economic status, gestational age at birth and birth weight were categorical, in which the largest (i.e. the most common) category was used as reference. All other variables were dichotomous.

Univariable associations between potential risk indicators and the different outcome categories were estimated with multinomial logistic regression analysis. Multinomial logistic regression allows for simultaneous estimation of the probability of the different outcomes (referral during labour from primary to secondary care due to suspected foetal distress, failure to progress in 2nd stage of labour, meconium stained amniotic fluid, failure to progress in 1st stage of labour, a wish for pain relief, either epidural anaesthesia or medication (opioids), a combination of other reasons, or no referral (reference category)).(10-14) Essentially, the multinomial logistic regression model includes several logistic regression models simultaneously, to estimate the associations between the risk indicators and each of the outcomes compared to the reference category (i.e. no referral). Hence, estimated regression coefficients for the predictors may differ per outcome.(10-14) All potential risk indicators were included in the multivariable analyses, so no preselection was made based on univariable statistics. The final risk indicators were identified by a backward stepwise selection using Akaike's Information Criterion.(15)
Statistical analyses were performed using R software, Version 2.15.2 (The R Foundation for Statistical Computing, 2009) and SAS software, Version 9.3 (SAS Institute Inc., Cary, NC, USA).

Results

Between 1 January, 2000 and 31 December, 2007, 1 443 654 women gave birth in the Netherlands. Of them, 810 298 women (56%) started labour in primary care with a singleton pregnancy of whom 746 642 women had a term pregnancy and were not referred for prolonged rupture of membranes. Approximately one-third (n=245 545, 33%) of our study population consisting of women pregnant with a term singleton starting labour in primary care was referred to secondary care during labour.
The reasons for referral to secondary care over the years 2000 till 2007 are shown in Figure 1. The total number of deliveries decreased from 187 624 in 2000 to 170 494 in 2007. The proportion of women with a singleton term pregnancy who started labour in primary care remained constant at 52%. Combined with a stable number of referral during labour of about 30 000/yr., this led to an increased incidence of referrals from 31% in 2000 to 35% in 2007. Figure 1 shows that this increase was mainly attributable to the rise in wish for pain relief from 7.0% in 2000 to 13% in 2007. Referral for foetal distress, failure to progress in 2nd stage of labour, and a combination of other reasons decreased, while referral due to failure to progress in 1st stage of labour and meconium stained amniotic fluid increased slightly.

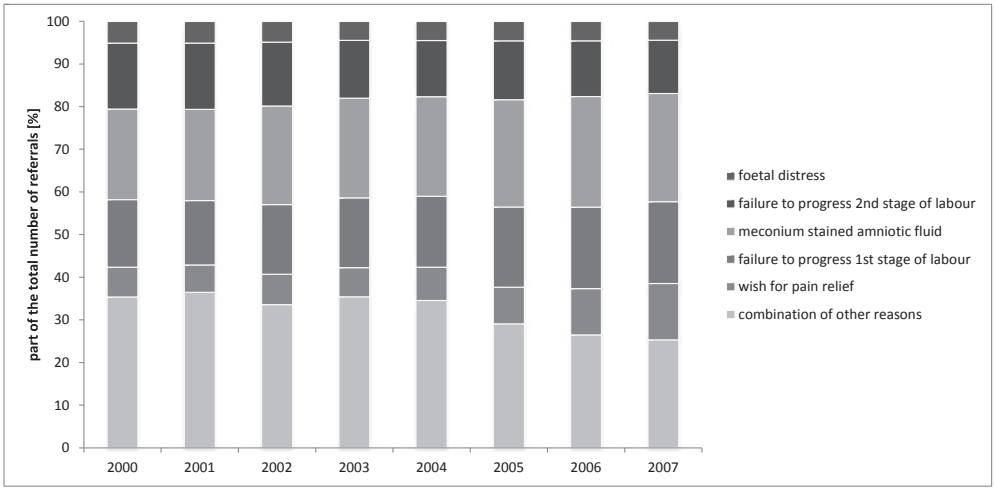


Figure 1 An overview of the different reasons for referral from primary to secondary care during labour over the years 2000 to 2007

Also the univariable associations between the potential risk indicators and the different reasons of referral are presented in Table 1. The associations between risk indicators and referral during labour were quite similar for all reasons of referral during labour, except for a referral for failure to progress in the 2nd stage of labour.

The multivariable associations for the different reasons for referral during labour from primary to secondary care are presented in Table 2, and showed to be comparable to the univariable associations. The risk indicators that showed to be significantly associated with all reasons of referral were a combination of parity and previous instrumental vaginal delivery, gestational age, and a planned hospital delivery. As compared to being multiparous without a history of an instrumental vaginal delivery, a previous instrumental vaginal delivery or being nulliparous greatly increased the risk of reasons for referral. A gestational age of 37 weeks decreased the risk of referral for all reasons but a combination of other reasons, and a gestational age >40^{+6} weeks increased the risk of referral as compared to a gestational age of 38^{+0}-40^{+6} weeks. The risk of referral was increased for all reasons when delivery was planned to be in the hospital as compared to home. None of the other risk indicators was significantly associated with all reasons of referral. In general, the risk of referral was increased for women with a low (<25) or high (>35) age, a non-Caucasian ethnicity, living in an area with an average degree of urbanization, a low social economic status, and a low (<p10) or high (>4000 grams) birth weight. Remarkably, women pregnant of a girl tended to have a lower risk of referral as compared to those pregnant with a boy. Additional interactions between different risk indicators are presented in Appendix A.

Table 1 Characteristics of women without referral during labour from primary to secondary care and from women referred due to foetal distress, failure to progress in the 2nd stage of labour, meconium stained amniotic fluid, failure to progress in the 1st stage of labour, wish for pain relief, and due to a combination of other reasons. And univariable associations between the different reasons of referral during labour and the potential risk indicators. Data are presented as mean ± standard deviation or n (%).

	Overall	No referral	Referral						Odds ratio (95% confidence interval)					
N (%)	746642	505047 (68)	241595 (32)											
Characteristic			foetal distress	Failure to progress 2nd stage of labour	meconium stained amniotic fluid	failure to progress 1st stage of labour	wish for pain relief	Combination of other reasons †	Foetal distress vs. no referral	Failure to progress 2nd stage of labour vs. no referral	meconium stained amniotic fluid vs. no referral	failure to progress 1st stage of labour vs. no referral	wish for pain relief vs. no referral	combination of other reasons vs. no referral
N (%)			**11813 (5)**	**34386 (14)**	**58132 (24)**	**40984 (17)**	**16299 (7)**	**79981 (33)**						
Maternal age, yrs.														
< 25	96060 (13)	58380 (12)	1917 (16)	3554 (10)	9078 (16)	7049 (17)	3874 (24)	12208 (15)	1.42(1.36-1.50)	0.84(0.81-0.87)	1.40(1.37-1.43)	1.54(1.49-1.58)	2.29(2.21-2.38)	1.35(1.32-1.38)
25-35	563724 (76)	383722 (76)	8844 (75)	27859 (81)	42632 (73)	30183 (74)	11120 (68)	59364 (74)	Ref	Ref	Ref	Ref	Ref	Ref
>35	86858 (12)	62945 (12)	1052 (9)	2973 (9)	6422 (11)	3752 (9)	1305 (8)	8409 (11)	0.73(0.68-0.77)	0.65(0.63-0.68)	0.92(0.89-0.94)	0.76(0.73-0.78)	0.72(0.68-0.76)	0.86(0.84-0.88)
Ethnicity														
Caucasian	613898 (82)	420706 (83)	9925 (84)	30275 (88)	44406 (76)	31668 (77)	12222 (75)	64695 (81)	Ref	Ref	Ref	Ref	Ref	Ref
Creole	15079 (2)	8755 (2)	245 (2)	261 (1)	1898 (3)	1213 (3)	324 (2)	2383 (3)	1.19(1.04-1.35)	0.41(0.37-0.47)	2.05(1.95-2.16)	1.84(1.73-1.96)	1.27(1.14-1.43)	1.77(1.69-1.85)
Hindu	7050 (1)	4307 (1)	149 (1)	265 (1)	578 (1)	570 (1)	153 (1)	1028 (1)	1.47(1.24-1.73)	0.85(0.75-0.97)	1.27(1.16-1.39)	1.76(1.61-1.92)	1.22(1.04-1.44)	1.55(1.45-1.66)
Other non-Western	110615 (15)	71279 (14)	1494 (13)	3585 (10)	11250 (19)	7532 (18)	3600 (22)	11875 (15)	0.89(0.84-0.94)	0.70(0.67-0.72)	1.50(1.46-1.53)	1.40(1.37-1.44)	1.74(1.67-1.81)	1.08(1.06-1.11)
Degree of urbanisation														
Low	148462 (20)	101198 (20)	1968 (17)	6551 (19)	9655 (17)	7134 (17)	2669 (16)	19287 (24)	0.83(0.79-0.87)	0.93(0.90-0.96)	0.84(0.82-0.86)	0.88(0.85-0.90)	0.76(0.73-0.79)	1.35(1.32-1.37)
Average	457988 (61)	312868 (62)	7335 (62)	21760 (63)	35688 (61)	25163 (61)	10876 (67)	44298 (55)	Ref	Ref	Ref	Ref	Ref	Ref
High	140192 (19)	90981 (18)	2510 (21)	6075 (18)	12789 (22)	8687 (21)	2754 (17)	16396 (20)	1.18(1.12-1.23)	0.96(0.93-0.99)	1.23(1.21-1.26)	1.19(1.16-1.22)	0.87(0.83-0.91)	1.27(1.25-1.30)
Social economic status														
Low	221039 (30)	144131 (29)	3716 (31)	9257 (27)	19952 (34)	12942 (32)	5552 (34)	25489 (32)	1.16(1.11-1.21)	0.95(0.92-0.97)	1.29(1.27-1.32)	1.15(1.12-1.17)	1.25(1.21-1.30)	1.14(1.12-1.16)
Average	345612 (46)	236539 (47)	5274 (45)	16022 (47)	25348 (44)	18514 (45)	7269 (45)	36646 (46)	Ref	Ref	Ref	Ref	Ref	Ref
High	179991 (24)	124377 (25)	2823 (24)	9107 (26)	12823 (22)	9528 (23)	3478 (21)	17846 (22)	1.02(0.97-1.07)	1.08(1.05-1.11)	0.96(0.94-0.98)	0.98(0.95-1.00)	0.91(0.87-0.95)	0.93(0.91-0.94)
parity and pIVD														
MP no pIVD	374350 (50)	313370 (62)	2255 (19)	3527 (10)	20755 (36)	7947 (19)	2165 (13)	24331 (30)	Ref	Ref	Ref	Ref	Ref	Ref
MP pIVD	39438 (5)	19713 (4)	1422 (12)	5410 (16)	4421 (8)	3129 (8)	1110 (7)	4233 (5)	10.0(9.40-10.7)	24.4(23.3-25.5)	3.39(3.27-3.51)	6.26(5.99-6.54)	8.15(7.57-8.77)	2.77(2.67-2.87)
NP	332854 (45)	171964 (34)	8136 (69)	25449 (74)	32956 (57)	29908 (73)	13024 (80)	51417 (64)	6.57(6.27-6.89)	13.2(12.7-13.6)	2.89(2.84-2.95)	6.86(6.69-7.03)	11.0(10.5-11.5)	3.85(3.79-3.91)
Planned place delivery														
At home	466662 (63)	331493 (66)	6858 (58)	21767 (63)	30867 (53)	22612 (55)	7521 (46)	45544 (57)	Ref	Ref	Ref	Ref	Ref	Ref
Midwife-led birth centre	5367 (1)	3396 (1)	94 (1)	245 (1)	474 (1)	342 (1)	146 (1)	670 (1)	1.34(1.09-1.64)	1.10(0.96-1.25)	1.50(1.36-1.65)	1.48(1.32-1.65)	1.89(1.60-2.24)	1.44(1.32-1.56)
Hospital	274613 (37)	170158 (34)	4861 (41)	12374 (36)	26791 (46)	18030 (44)	8632 (53)	33767 (42)	1.38(1.33-1.43)	1.11(1.08-1.13)	1.69(1.66-1.72)	1.55(1.52-1.59)	2.24(2.17-2.31)	1.44(1.42-1.47)
Gestational age, wks.														
37^{+0}-37^{+6}	31070 (4)	22494 (4)	418 (4)	1400 (4)	726 (1)	1148 (3)	587 (4)	4297 (5)	0.86(0.78-0.95)	0.95(0.90-1.01)	0.32(0.29-0.34)	0.70(0.66-0.75)	0.86(0.79-0.93)	1.29(1.24-1.33)
38^{+0}-40^{+6}	549026 (74)	381012 (75)	8263 (70)	24905 (72)	38960 (67)	27730 (68)	11582 (71)	56574 (71)	Ref	Ref	Ref	Ref	Ref	Ref
≥41^{+0}	166546 (22)	101541 (20)	3132 (27)	8081 (24)	18446 (32)	12106 (30)	4130 (25)	19110 (24)	1.42(1.36-1.48)	1.22(1.19-1.25)	1.78(1.74-1.81)	1.64(1.60-1.68)	1.34(1.29-1.39)	1.27(1.25-1.29)
Neonatal female gender	364997 (49)	251457 (50)	4559 (39)	15378 (45)	29031 (50)	18666 (46)	7925 (49)	37981 (47)	0.63(0.61-0.66)	0.82(0.80-0.83)	1.01(0.99-1.02)	0.84(0.83-0.86)	0.95(0.93-0.98)	0.91(0.90-0.93)
Birth weight														
<p10	59352 (8)	39980 (8)	1880 (16)	1386 (4)	4994 (9)	2767 (7)	1437 (9)	7008 (9)	2.06(1.96-2.17)	0.51(0.48-0.54)	1.15(1.11-1.18)	0.87(0.84-0.91)	1.08(1.02-1.14)	1.12(1.09-1.15)
p10-4000	572240 (77)	389242 (77)	8898 (75)	26665 (78)	42471 (73)	31063 (76)	13021 (80)	60880 (76)	Ref	Ref	Ref	Ref	Ref	Ref
>4000 g	115050 (15)	75925 (15)	1035 (9)	6335 (18)	10667 (18)	7154 (17)	1841 (11)	12093 (15)	0.60(0.56-0.64)	1.22(1.18-1.25)	1.29(1.26-1.32)	1.18(1.15-1.21)	0.72(0.69-0.76)	1.02(1.00-1.04)

† Other reasons for a referral include intra-uterine fetal death during labour, malpresentation (e.g. breech), blood loss during labour, abruptio placentae, vasa praevia, threatened preterm labour or other problems. MP = multiparous; pIVD = previous instrumental vaginal delivery; NP = nulliparous

Table 2 Multivariable associations between different reasons for referral from primary to secondary care and potential risk indicators. Interactions are described in Appendix A.

characteristic	Odds ratio (95% confidence interval) foetal distress vs. no referral	failure to progress 2nd stage of labour vs. no referral	meconium stained amniotic fluid vs. no referral	failure to progress 1st stage of labour vs. no referral	Wish for pain relief vs. no referral	Combination of other reasons vs. no referral
Maternal age, yrs.						
< 25	1.14 (0.96-1.36)	0.81 (0.68-0.97)	1.16 (1.09-1.23)	1.38 (1.26-1.50)	1.68 (1.45-1.95)	1.23 (1.16-1.30)
25-35	Ref	Ref	Ref	Ref	Ref	Ref
>35	1.17 (1.04-1.32)	1.31 (1.20-1.43)	1.05 (1.00-1.10)	1.04 (0.97-1.12)	0.87 (0.76-1.00)	1.22 (1.17-1.27)
Ethnicity						
Caucasian	Ref	Ref	Ref	Ref	Ref	Ref
Creole	1.30 (0.97-1.75)	0.63 (0.48-0.83)	2.37 (2.10-2.67)	2.50 (2.16-2.90)	1.81 (1.40-2.34)	2.35 (2.11-2.61)
Hindu	1.58 (1.06-2.36)	1.25 (0.93-1.68)	1.32 (1.05-1.66)	2.55 (2.06-3.15)	1.67 (1.10-2.53)	2.30 (1.97-2.69)
Other non-Western	0.88 (0.79-0.99)	0.86 (0.79-0.92)	1.46 (1.39-1.53)	1.60 (1.51-1.69)	1.93 (1.78-2.10)	1.25 (1.20-1.31)
Degree of urbanisation						
Low	0.84 (0.78-0.90)	0.92 (0.88-0.96)	0.88 (0.85-0.92)	0.95 (0.91-0.99)	0.90 (0.84-0.96)	1.59 (1.54-1.63)
Average	Ref	Ref	Ref	Ref	Ref	Ref
High	0.95 (0.85-1.05)	0.93 (0.87-0.99)	0.90 (0.86-0.95)	0.90 (0.85-0.96)	0.75 (0.68-0.83)	1.06 (1.01-1.11)
Social economic status						
Low	1.03 (0.96-1.11)	1.04 (1.00-1.09)	1.14 (1.10-1.18)	0.98 (0.94-1.03)	1.27 (1.20-1.35)	1.11 (1.08-1.15)
Average	Ref	Ref	Ref	Ref	Ref	Ref
High	1.02 (0.95-1.09)	1.05 (1.01-1.09)	0.94 (0.91-0.98)	0.96 (0.93-1.00)	0.86 (0.81-0.92)	1.12 (1.09-1.15)
parity and pIVD						
MP no pIVD	Ref	Ref	Ref	Ref	Ref	Ref
MP pIVD	12.4 (11.3-13.5)	34.7 (32.7-36.9)	3.70 (3.53-3.88)	7.44 (7.03-7.89)	8.86 (8.05-9.76)	3.07 (2.93-3.22)
NP	7.22 (6.77-7.69)	18.2 (17.4-19.2)	3.18 (3.10-3.26)	7.90 (7.63-8.18)	11.2 (10.5-11.9)	4.25 (4.15-4.34)
Planned place delivery						
At home	Ref	Ref	Ref	Ref	Ref	Ref
Midwife-led birth centre	1.37 (0.92-2.04)	0.93 (0.71-1.22)	1.22 (1.00-1.49)	1.02 (0.80-1.30)	1.85 (1.33-2.59)	1.58 (1.33-1.87)
Hospital	1.40 (1.31-1.50)	1.17 (1.12-1.22)	1.63 (1.58-1.69)	1.47 (1.41-1.52)	2.41 (2.27-2.55)	1.66 (1.61-1.71)
Gestational age, wks.						
37^{+0}-37^{+6}	0.68 (0.61-0.76)	0.82 (0.77-0.87)	0.28 (0.26-0.30)	0.58 (0.55-0.62)	0.69 (0.63-0.76)	1.09 (1.05-1.13)
38^{+0}-40^{+6}	Ref	Ref	Ref	Ref	Ref	Ref
≥41^{+0}	1.51 (1.43-1.58)	1.20 (1.16-1.24)	1.86 (1.82-1.91)	1.71 (1.66-1.76)	1.45 (1.39-1.52)	1.30 (1.27-1.33)
Neonatal female gender	0.59 (0.57-0.62)	0.82 (0.80-0.84)	1.00 (0.98-1.02)	0.83 (0.81-0.85)	0.93 (0.89-0.96)	0.91 (0.89-0.92)
Birth weight						
<p10	1.89 (1.65-2.16)	0.55 (0.46-0.67)	1.14 (1.07-1.22)	0.92 (0.83-1.03)	1.35 (1.15-1.59)	1.21 (1.14-1.28)
p10-4000	Ref	Ref	Ref	Ref	Ref	Ref
>4000 g	0.90 (0.79-1.03)	2.38 (2.20-2.57)	1.41 (1.35-1.48)	1.55 (1.46-1.65)	1.02 (0.90-1.16)	1.19 (1.15-1.25)

Discussion

In this study we identified risk indicators for referral from primary to secondary care during labour due to foetal distress, failure to progress in 2nd stage of labour, meconium stained amniotic fluid, failure to progress in 1st stage of labour, wish for pain relief, and a combination of other reasons. The strongest risk indicators for referral, independent of the reason, were a combination of parity and a previous instrumental vaginal delivery, gestational age, and the planned place of delivery. Additional risk indicators differed per reason of referral.

We used a well described, large, nationwide and heterogeneous cohort of pregnant women from the Netherlands. Data collection with the Perinatal Registry Netherlands was standardized with uniform forms and definitions. Furthermore, data are subjected to range and consistency checks (routine audit) and false records are sent back to the caregiver, who is given ample opportunity to correct them. This assures the data to be of high quality, which was confirmed by an earlier study in which outcome measures, e.g. perinatal mortality, in the registry were highly comparable with civil registration data.(18)

Identification of the reason of referral during labour from primary to secondary care is complicated by the fact that women can have multiple reasons for referral. By prioritization of the reasons of referral, based on urgency (i.e. a referral for foetal distress is more urgent than a wish for pain relief) and putting the health of the foetus before the mother's (i.e. referral for foetal distress is more important than failure to progress in the 2nd stage of labour), we avoid the dilution of the associations between the risk indicators and the reason for referral that would be present if women with multiple reasons of referral were analyzed for each reason separately (i.e. using multiple logistic regression models). To prevent this dilution and given only 5% of the women had multiple reasons for referral we feel that prioritization of reasons for referral was a valid decision.

There are several limitations that need further elaboration. Unfortunately, information on some potential risk indicators was not (reliably) registered, e.g. body mass index. Furthermore, care providers in primary care are not obliged to register fertility problems and smoking behaviour, consequently leading to large underrepresentation of these characteristics. Therefore, fertility problems and smoking were not considered as risk indicators in this study. Moreover, the moment at which information on planned place of delivery is registered is unknown, e.g. at the first contact with the caregiver or moments before labour. Since women can change their preference for place of delivery during their pregnancy and this change may not always be registered, the planned place of delivery may not always be correct. The quality of the information on a previous instrumental vaginal delivery is unclear. Furthermore, we were not able to correct for all possible confounders, such as differences in policy and management within the primary care setting, e.g. the increase in use of epidural anaesthesia, threshold for referral, or fusions of hospitals which increase referral times between primary and secondary care.

One of the outcome categories of interest was a combination of other reasons for referral during labour from primary to secondary care. Since this category is a combination of reasons for referral, other and less prevalent than the other reasons investigated, it is very difficult to (clinically) interpret the associations found between this outcome and the risk indicators of interest. However, since this group is part of the domain of interest, i.e. women with a term singleton pregnancy starting labour in primary care, we left these women in our study population. Given the large heterogeneity we strongly suggest not to draw any conclusions for this outcome category based on associations presented in this study.

Our study aimed to identify risk indicators for several reasons for referral during labour from primary to secondary care simultaneously using a multinomial logistic regression analysis. A potential disadvantage of this approach is that a variable is selected as a risk indicator if it is associated with one of the outcomes, i.e. the risk indicators are not necessarily related to all of the different endpoints. An advantage is that a multinomial logistic model is able to identify risk indicators of a referral for all the different reasons directly. Furthermore, in a multinomial model more women are used to fit the final model than in six separate logistic regression models where the same women are used as a reference several times.

Compared to the univariable analysis, the multivariable analysis showed opposite effects for the association between several risk indicators and referral due to any of the reasons of interest. In the univariable analysis women with a Hindu ethnicity had a lower risk of a referral due to failure to progress in the 2nd stage of labour as compared to Caucasians. In the multivariable analysis an opposite effect was found. Furthermore, living in a highly urbanized area increased the risk of referral due to foetal distress, failure to progress in the 1st stage of labour, and meconium stained amniotic fluid, while it decreased the same risks in the multivariable analysis. Apparently, the effects in the univariable analysis could be explained by other characteristics added in the multivariable analysis.

Several studies, conducted in various countries, investigated indicators associated with referral from primary to secondary care during labour.(19-23) Similar to our study, all studies found that nulliparous women were at an increased risk of referral during labour.(19-23) Being pregnant of your fourth or more child decreased the risk of referral, although it must be noted that this finding was based on small numbers.(21) Other findings that corresponded with our results were the increased risk of referral for women with a late term gestational age (20) and for nulliparous women with increasing age.(20;22) A birth weight below 2 500g or above 4 500g showed to increase the risk of referral in multiparous women, which is similar to what we found.(20) Identified risk indicators for referral in other studies that were not registered in our study included a BMI > 30 kg/m^2 (21), unavailability of the chosen midwife at the onset of labour in nulliparous women, the moment of the decision to give birth at home in multiparous women (before pregnancy decreased the risk of referral, in the last week of pregnancy increased the risk), no ultrasound, or a previous caesarean section.(20) The latter

was not considered a risk indicator in this study because in the Netherlands women with a history of a caesarean section will always start labour in secondary care.

Although many variables showed associations that were to be expected based on pathophysiologic reasoning and previous research we were surprised to find that women pregnant of a girl were less likely to be referred, not only for fetal distress, but also for non-progressive labour in the first and second stage of labour as well for pain relief, even after correction for birth weight differences. Further research into this mechanism is needed.

Although there is an increasing number of midwife-led birth centres in a growing number of Western countries that show comparable referral rates,(24) the obstetric care system in the Netherlands with its division in primary and secondary care is unique. Therefore it is debatable how generalizable and informative this study is to other Western countries. It seems however plausible that at each obstetric care unit in Western countries a distinction is made, conscious or unconscious, between women being at low or high risk of problems during labour, especially in those countries with midwife-led birth centres. This study particularly focused on low risk women and its results may help caregivers to be aware of problems that can occur during labour in these women. Except for the planned location of delivery all risk indicators are universal and are therefore generalizable to other countries.

The results of this study may aid to identify women in primary care that may encounter future problems during labour. The identified risk indicators are readily available and are different from the disease-related items that are listed in the List of Obstetric Indications. (3) The indicators may be used, e.g. combined in prognostic model, to counsel and guide individual patient management, e.g. on the preferred place of delivery, improve the quality of care by enhancing the identification of risk groups, and timely interventions. Since women can choose to start labour at home or in the hospital under supervision of primary care, one could argue why not to advise women with a high risk of referral during labour to deliver in the hospital, where the gynaecologist is stand-by. On the opposite, women with a low risk for referral could be advised to start labour in primary care. Obviously, this is still a joined decision of the pregnant woman and the midwife or primary care physician.

Future research should focus on updating the current study with new available information. From 2011 onwards, information on body mass index is being registered in the Netherlands Perinatal registry. Despite the large sample size, as with each etiologic study, the findings of the study should be verified in other countries to prove the validity of our results. In this study we focused on a referral during labour from primary to secondary care since this type of referral is associated with higher risks of maternal and neonatal complications than an antepartum referral(25) and since referral during labour is more frequent than postpartum referral.(24) Since referral during the puerperium (directly postpartum and within the first week of birth) is associated with acute adverse maternal (e.g. post partum haemorrhage > 1 000 cc and retained placenta) and neonatal (e.g low Apgar score and respiratory problems)

outcomes in which immediate secondary level care is necessary(25), future research should focus on the identification of risk indicators for this type of referral as well.

Among low risk women, complications during labour, including foetal distress, failure to progress in 2nd stage of labour, meconium stained amniotic fluid, failure to progress in 1st stage of labour, wish for pain relief, and a combination of other reasons can be associated with readily available risk indicators. These risk indicators can contribute to the awareness of the risk of referral and they can be used during counselling of pregnant women.

References

(1) Stichting Perinatale Registratie Nederland. Perinatale Zorg in Nederland 2008 (Dutch). 2011. Utrecht, Stichting Perinatale Registratie Nederland.
Ref Type: Serial (Book,Monograph)

(2) Obstetric Working Group of the National Health Insurance Board of the Netherlands. The list of obstetric indications. 1999.

(3) Commissie Verloskunde van het CVZ. Verloskundig Vademecum 2003 (Dutch). Diemen, the Netherlands: College voor Zorgverzekeringen; 2003.

(4) Meray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. J Clin Epidemiol 2007 Sep;60(9):883-91.

(5) Stichting Perinatale Registratie Nederland. Grote Lijnen 10 jaar Perinatale Registratie Nederland (Dutch). Utrecht: Stichting Perinatale Registratie Nederland; 2011.

(6) Offerhous P, de Boer J, Daemers D. Niet-vorderende ontsluiting. Aanbevelingen voor verloskundig beleid, begeleiding en preventie. KNOV-standaard. Deventer: Ovimex b.v.; 2006.

(7) Knol F. Van hoog naar laag; van laag naar hoog, de sociaal-ruimtelijke ontwikkeling van wijken in de periode 1971-1995 (Dutch). Den Haag: Elsevier Bedrijfsinformatie; 1998. Report No.: SCP-Cahier 152.

(8) Visser GH, Eilers PH, Elferink-Stinkens PM, Merkus HM, Wit JM. New Dutch reference curves for birthweight by gestational age. Early Hum Dev 2009 Dec;85(12):737-44.

(9) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-91.

(10) Begg CB, Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika 1984;71:11-8.

(11) Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KG. Polytomous logistic regression analysis could be applied more often in diagnostic research. J Clin Epidemiol 2008 Feb;61(2):125-34.

(12) Roukema J, van Loenhout RB, Steyerberg EW, Moons KG, Bleeker SE, Moll HA. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. J Clin Epidemiol 2008 Feb;61(2):135-41.

(13) Schuit E, Kwee A, Westerhuis ME, Van Dessel HJ, Graziosi GC, Van Lith JM, et al. A clinical prediction model to assess the risk of operative delivery. BJOG 2012 Jul;119(8):915-23.

(14) Wijesinha A, Begg CB, Funkenstein HH, McNeil BJ. Methodology for the differential diagnosis of a complex data set. A case study using data from routine CT scan examinations. Med Decis Making 1983;3(2):133-54.

(15) Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19(6):716-23.

(16) Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-87.

(17) Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991 Aug;10(8):1213-26.

(18) Anthony S, van der Pal-de Bruin KM, Graafmans WC, Dorrepaal CA, Borkent-Polet M, van Hemel OJ, et al. The reliability of perinatal and neonatal mortality rates: differential under-reporting in linked professional registers vs. Dutch civil registers. Paediatr Perinat Epidemiol 2001 Jul;15(3):306-14.

(19) Perinatal and maternal outcomes by planned place of birth for healthy women with low risk pregnancies: the Birthplace in England national prospective cohort study. BMJ 2011;343:d7400.

(20) Lindgren HE, Hildingsson IM, Christensson K, Radestad IJ. Transfers in planned home births related to midwife availability and continuity: a nationwide population-based study. Birth 2008 Mar;35(1):9-15.

(21) Lindgren HE, Radestad IJ, Hildingsson IM. Transfer in planned home births in Sweden--effects on the experience of birth: a nationwide population-based study. Sex Reprod Healthc 2011 Aug;2(3):101-5.

(22) Rowe R, Fitzpatrick R, Hollowell J, Kurinczuk J. Transfers of women planning birth in midwifery units: data from the Birthplace prospective cohort study. BJOG 2012 Aug;119(9):1081-90.

(23) Wiegers TA, van der ZJ, Keirse MJ. Transfer from home to hospital: what is its effect on the experience of childbirth? Birth 1998 Mar;25(1):19-24.

(24) Amelink-Verburg MP, Rijnders ME, Buitendijk SE. A trend analysis in referrals during pregnancy and labour in Dutch midwifery care 1988-2004. BJOG 2009 Jun;116(7):923-32.

(25) Amelink-Verburg MP, Verloove-Vanhorick SP, Hakkenberg RM, Veldhuijzen IM, Bennebroek GJ, Buitendijk SE. Evaluation of 280,000 cases in Dutch midwifery practices: a descriptive study. BJOG 2008 Apr;115(5):570-8.

Appendix A Interactions in multivariable associations for the different reasons of referral during labour

	Odds ratio (95% confidence interval)					
Interactions	**Foetal distress vs. no referral**	**Failure to progress 2nd stage of labour vs. no referral**	**Meconium stained amniotic fluid vs. no referral**	**Failure to progress 1st stage of labour vs. no referral**	**wish for pain relief vs. no referral**	**Combination of other reasons vs. no referral**
Maternal age* Parity and pIVD						
MP no pIVD	2.89 (2.26-3.69)	3.38 (2.74-4.18)	1.94 (1.68-2.24)	2.25 (1.92-2.63)	3.20 (2.55-4.02)	2.14 (1.86-2.47)
<25*NP	0.76 (0.64-0.91)	0.65 (0.54-0.77)	0.77 (0.72-0.82)	0.61 (0.56-0.67)	0.79 (0.68-0.91)	0.70 (0.66-0.74)
>35*MP no pIVD	0.37 (0.30-0.47)	0.35 (0.31-0.40)	0.65 (0.58-0.72)	0.52 (0.45-0.59)	0.60 (0.48-0.76)	0.68 (0.61-0.75)
>35*NP	0.90 (0.78-1.04)	0.87 (0.79-0.96)	1.17 (1.10-1.24)	1.23 (1.14-1.33)	1.46 (1.27-1.68)	1.04 (0.99-1.10)
Birth weight* Parity and pIVD						
<p10*MP no pIVD	0.87 (0.70-1.07)	0.86 (0.68-1.09)	1.07 (0.93-1.23)	0.71 (0.58-0.87)	0.72 (0.54-0.95)	0.84 (0.72-0.97)
<p10*NP	1.03 (0.90-1.18)	0.91 (0.75-1.09)	0.98 (0.92-1.05)	0.89 (0.80-0.98)	0.79 (0.67-0.92)	0.86 (0.81-0.91)
>4000g*MP pIVD	0.43 (0.35-0.52)	0.34 (0.31-0.38)	0.79 (0.73-0.86)	0.65 (0.58-0.72)	0.59 (0.48-0.72)	0.76 (0.70-0.83)
>4000g*NP	0.75 (0.65-0.87)	0.84 (0.78-0.91)	1.28 (1.22-1.35)	1.23 (1.15-1.31)	1.23 (1.09-1.40)	1.29 (1.23-1.35)
Ethnicity*Birth weight						
Creole*<p10	1.19 (0.89-1.60)	1.36 (0.89-2.07)	0.91 (0.79-1.06)	0.92 (0.75-1.12)	1.02 (0.74-1.40)	0.99 (0.87-1.13)
Creole*>4000g	2.06 (1.21-3.50)	1.32 (0.88-1.98)	1.05 (0.87-1.26)	1.04 (0.83-1.31)	1.26 (0.78-2.02)	1.27 (1.07-1.52)
Hindu*<p10	1.01 (0.66-1.54)	0.71 (0.39-1.30)	0.64 (0.46-0.89)	0.65 (0.45-0.93)	0.49 (0.25-0.95)	0.92 (0.73-1.16)
Hindu*>4000g	1.63 (0.59-4.51)	1.44 (0.81-2.55)	1.40 (0.95-2.07)	1.32 (0.86-2.02)	1.05 (0.37-2.92)	1.84 (1.33-2.55)
Other non-Western*<p10	0.90 (0.78-1.05)	1.06 (0.90-1.25)	0.89 (0.82-0.96)	0.99 (0.90-1.10)	0.95 (0.83-1.08)	1.04 (0.97-1.12)
Other non-Western*>4000g	1.47 (1.20-1.79)	1.24 (1.12-1.37)	1.10 (1.03-1.17)	1.12 (1.03-1.21)	1.18 (1.04-1.35)	1.19 (1.11-1.27)
Neonatal female gender* Birth weight						
Female*<p10	1.21 (1.09-1.34)	1.09 (0.97-1.21)	1.02 (0.95-1.08)	1.01 (0.93-1.09)	0.97 (0.86-1.08)	1.04 (0.98-1.10)
Female*>4000g	0.82 (0.70-0.95)	1.00 (0.94-1.07)	1.03 (0.99-1.08)	1.03 (0.97-1.09)	1.01 (0.91-1.12)	0.99 (0.95-1.04)
Degree of urbanisation*Social economic status*Ethnicity						
Low*High*Creole	1.87 (0.79-4.40)	0.92 (0.33-2.56)	1.09 (0.73-1.62)	0.96 (0.59-1.58)	2.82 (1.61-4.94)	2.17 (1.60-2.95)
Low*High*Hindu	0.48 (0.12-2.02)	0.87 (0.41-1.87)	1.30 (0.78-2.14)	0.88 (0.49-1.57)	3.47 (1.93-6.25)	1.40 (0.94-2.09)
Low*High*Other non-Western	1.03 (0.69-1.54)	1.19 (0.94-1.49)	0.99 (0.84-1.16)	0.90 (0.74-1.08)	1.15 (0.90-1.47)	1.20 (1.04-1.39)
Low*Low*Creole	0.52 (0.07-3.77)	2.91 (1.21-6.98)	1.16 (0.69-1.96)	2.07 (1.25-3.44)	2.14 (0.90-5.08)	0.96 (0.63-1.45)
Low*Low*Hindu	1.11 (0.33-3.68)	1.21 (0.53-2.79)	0.65 (0.28-1.53)	0.92 (0.44-1.89)	1.37 (0.48-3.93)	0.57 (0.33-1.00)
Low*Low*Other non-Western	0.79 (0.44-1.43)	1.40 (1.04-1.88)	0.97 (0.78-1.22)	1.13 (0.88-1.44)	1.59 (1.18-2.16)	0.82 (0.68-0.98)
High*High*Creole	0.68 (0.21-2.22)	0.87 (0.31-2.45)	0.50 (0.28-0.89)	0.75 (0.44-1.29)	1.17 (0.50-2.75)	1.05 (0.68-1.63)
High*High*Hindu	0.69 (0.21-2.27)	0.66 (0.26-1.68)	0.78 (0.38-1.58)	1.01 (0.55-1.84)	0.29 (0.04-2.16)	0.98 (0.56-1.71)
High*High*Other non-Western	0.67 (0.44-1.03)	1.17 (0.93-1.47)	0.88 (0.73-1.05)	0.87 (0.71-1.05)	1.61 (1.26-2.06)	1.17 (0.99-1.38)
High*Low*Creole	0.84 (0.64-1.11)	0.94 (0.72-1.23)	0.74 (0.66-0.82)	0.68 (0.59-0.78)	0.55 (0.42-0.71)	0.94 (0.85-1.04)
High*Low*Hindu	0.82 (0.57-1.16)	0.76 (0.57-1.01)	0.86 (0.72-1.04)	0.80 (0.66-0.97)	0.38 (0.24-0.59)	0.95 (0.81-1.10)
High*Low*Other non-Western	0.94 (0.83-1.08)	0.95 (0.87-1.05)	0.98 (0.93-1.04)	0.81 (0.75-0.87)	0.58 (0.51-0.65)	1.04 (0.99-1.10)

Degree of urbanisation*Social economic status						
Low*High	1.00 (0.89-1.14)	1.00 (0.94-1.08)	1.16 (1.09-1.23)	1.04 (0.97-1.12)	1.20 (1.08-1.34)	0.72 (0.68-0.75)
Low*Low	1.04 (0.90-1.20)	1.02 (0.94-1.11)	0.84 (0.78-0.90)	0.90 (0.83-0.98)	0.75 (0.66-0.85)	1.14 (1.08-1.20)
High*High	1.19 (1.01-1.39)	0.93 (0.84-1.02)	1.05 (0.97-1.14)	1.10 (1.01-1.21)	1.21 (1.03-1.41)	0.66 (0.61-0.71)
High*Low	1.19 (1.05-1.34)	0.96 (0.88-1.03)	1.09 (1.02-1.16)	1.14 (1.06-1.23)	0.94 (0.84-1.05)	0.97 (0.92-1.02)
Maternal age*Planned place delivery						
<25*midwife-led birth centre	0.74 (0.45-1.22)	0.83 (0.58-1.19)	1.21 (0.97-1.52)	0.92 (0.71-1.20)	0.73 (0.50-1.06)	0.78 (0.63-0.96)
<25*hospital	0.85 (0.76-0.94)	0.82 (0.76-0.88)	0.83 (0.79-0.88)	0.83 (0.78-0.88)	0.67 (0.62-0.73)	0.84 (0.80-0.88)
>35*midwife-led birth centre	0.53 (0.16-1.71)	2.22 (1.43-3.44)	1.17 (0.79-1.74)	0.92 (0.54-1.58)	0.81 (0.32-2.06)	1.26 (0.93-1.72)
>35*hospital	1.16 (1.02-1.32)	1.16 (1.07-1.26)	1.15 (1.09-1.22)	1.14 (1.06-1.23)	1.22 (1.08-1.37)	0.97 (0.92-1.02)
Ethnicity*Planned place delivery						
Creole*midwife-led birth centre	0.65 (0.22-1.89)	0.59 (0.21-1.67)	0.81 (0.54-1.21)	1.07 (0.68-1.69)	0.42 (0.16-1.10)	0.51 (0.35-0.76)
Creole*hospital	0.80 (0.60-1.08)	1.00 (0.76-1.34)	0.97 (0.86-1.10)	0.98 (0.84-1.14)	0.70 (0.53-0.91)	0.69 (0.62-0.77)
Hindu*midwife-led birth centre	1.82 (0.51-6.52)	0.73 (0.17-3.22)	0.85 (0.35-2.08)	0.54 (0.18-1.57)	NA	0.65 (0.32-1.31)
Hindu*hospital	0.89 (0.59-1.35)	1.04 (0.76-1.43)	1.01 (0.80-1.28)	0.80 (0.64-1.00)	0.81 (0.52-1.25)	0.58 (0.50-0.69)
Other non-Western*midwife-led birth centre	0.93 (0.54-1.59)	1.13 (0.81-1.58)	0.93 (0.74-1.17)	1.23 (0.95-1.60)	0.71 (0.48-1.04)	0.77 (0.62-0.95)
Other non-Western*hospital	1.00 (0.88-1.13)	1.10 (1.01-1.19)	0.95 (0.90-1.00)	0.98 (0.92-1.05)	0.85 (0.77-0.93)	0.78 (0.75-0.82)
Degree of urbanisation*planned place delivery						
Low*midwife-led birth centre	1.12 (0.52-2.43)	1.21 (0.76-1.93)	1.42 (0.99-2.05)	1.37 (0.89-2.10)	1.11 (0.59-2.09)	1.21 (0.93-1.58)
Low*hospital	1.06 (0.96-1.19)	1.02 (0.95-1.08)	1.01 (0.96-1.06)	0.97 (0.92-1.03)	0.92 (0.84-1.00)	0.82 (0.79-0.86)
High*midwife-led birth centre	1.00 (0.59-1.70)	0.89 (0.64-1.23)	1.13 (0.89-1.44)	0.97 (0.74-1.29)	0.65 (0.42-1.00)	0.79 (0.64-0.99)
High*hospital	0.95 (0.86-1.05)	1.03 (0.97-1.10)	0.97 (0.92-1.02)	0.94 (0.89-1.00)	0.85 (0.77-0.93)	1.06 (1.01-1.11)
Social economic status*planned place delivery						
High*midwife-led birth centre	1.16 (0.65-2.07)	1.15 (0.78-1.69)	0.62 (0.44-0.88)	0.81 (0.55-1.18)	0.47 (0.24-0.92)	0.79 (0.61-1.02)
High*hospital	0.91 (0.82-1.00)	0.97 (0.91-1.03)	0.92 (0.88-0.97)	0.96 (0.91-1.02)	0.94 (0.86-1.02)	0.85 (0.81-0.88)
Low*midwife-led birth centre	0.83 (0.48-1.42)	1.48 (1.05-2.08)	0.88 (0.69-1.13)	1.19 (0.90-1.59)	1.27 (0.86-1.88)	0.98 (0.79-1.22)
Low*hospital	0.97 (0.88-1.07)	0.97 (0.91-1.03)	0.95 (0.91-1.00)	1.01 (0.96-1.07)	0.89 (0.82-0.96)	0.90 (0.87-0.94)
Gestational age*Birth weight						
37^{+0}-37^{+6}*<p10	1.41 (1.04-1.91)	0.43 (0.26-0.72)	1.39 (1.05-1.85)	0.94 (0.68-1.28)	0.53 (0.32-0.87)	1.30 (1.13-1.50)
37^{+0}-37^{+6}*>4000g	0.67 (0.16-2.70)	1.21 (0.82-1.80)	1.64 (1.04-2.58)	1.91 (1.32-2.76)	1.40 (0.68-2.87)	1.40 (1.09-1.80)
≥41^{+0}*<p10	0.98 (0.88-1.11)	1.11 (0.98-1.26)	0.89 (0.83-0.95)	1.04 (0.95-1.14)	0.94 (0.83-1.07)	0.97 (0.91-1.04)
≥41^{+0}*<4000g	1.04 (0.91-1.19)	0.83 (0.78-0.88)	0.80 (0.76-0.84)	0.79 (0.75-0.84)	0.90 (0.81-1.00)	0.91 (0.87-0.96)

Chapter 2.2

Prediction of neonatal metabolic acidosis in women with a singleton term pregnancy in cephalic presentation.

Schuit E, Westerhuis ME, Kwee A, Zuithoff NP, Groenwold RH, Van Den Akker ES, Van Beek E, Van Dessel HJHM, Drogtrop AP, Van Geijn HP, Graziosi PGCM, Van Lith JMM, Nijhuis JG, Oei SG, Oosterbaan HP, Porath MM, Rijnders RJP, Schuitemaker NWE, Wijnberger LDE, Willekes C, Wouters MGAJ, Visser GHA, Mol BWJ, Moons KGM.

Am J Perinatol 2012 Mar;29(3):167-74.

Abstract

Objective To predict neonatal metabolic acidosis at birth using antepartum obstetric characteristics (model 1) and additional characteristics available during labor (model 2).

Study Design In 5,667 laboring women from a multicenter randomized trial that had a high-risk singleton pregnancy in cephalic presentation beyond 36 weeks of gestation, we predicted neonatal metabolic acidosis. Based on literature and clinical reasoning we selected both antepartum characteristics and characteristics that became available during labor. After univariable analyses the predictors of the multivariable models were identified by backward stepwise selection in a logistic regression analysis. Model performance was assessed by discrimination and calibration. To correct for potential overfitting we (internally) validated the models with bootstrapping techniques.

Results Of 5,667 neonates born alive, 107 (1.9%) had metabolic acidosis. Antepartum predictors of metabolic acidosis were gestational age, nulliparity, previous cesarean delivery and maternal diabetes. Additional intrapartum predictors were spontaneous onset of labor and meconium stained amniotic fluid. Calibration and discrimination were acceptable for both models (c-statistic 0.64 and 0.66, respectively).

Conclusion In women with a high-risk singleton term pregnancy in cephalic presentation, we identified antepartum and intrapartum factors that predict neonatal metabolic acidosis at birth.

Introduction

The ultimate goal of fetal surveillance during labor is the prevention of perinatal asphyxia. Perinatal asphyxia is associated with several short- and long-term complications, varying from mild hypoxic ischemic encephalopathy to cerebral palsy and death.(1-5)

In the pathway from labor and perinatal outcome to long-term consequences, there has been a strong focus on the association between adverse outcome in the perinatal period and long-term cerebral injury or death, and on univariable associations between single factors and adverse outcomes. However, in clinical practice it is important to identify women with a high(er) risk of developing adverse neonatal outcome as early as possible, e.g. at entrance in the hospital or even antenatally. Clinical practice is always multivariable: no prognosis is set by a single factor.(6) Timely prognostication may lead to more effective decision making during labor, in specific with respect to the type of fetal monitoring that is offered, as well as the interpretation of fetal heart rate patterns.

Fetal surveillance during labor is performed with cardiotocography (CTG), fetal blood sampling (FBS) or ST-analysis of the fetal electrocardiogram (ECG). However, these tests are often applied and interpreted in isolation, without taking into account other factors that may influence the probability of an adverse outcome at birth.

The aim of the present study was therefore to identify which factors measured both before and during labor could be used in a multivariable way to adequately predict neonatal metabolic acidosis at birth. For this purpose we used data from a large multicenter trial in which laboring women with a high-risk vertex singleton pregnancy beyond 36 weeks of gestation were studied.(7;8)

Methods

Setting

In a recently published randomized clinical trial in The Netherlands, laboring women with a high-risk (explained below) vertex singleton pregnancy beyond 36 weeks of gestation were randomly allocated to either intra partum monitoring by cardiotocography (CTG) plus ST-analysis of the fetal electrocardiogram (ECG; STAN) or CTG only, using a strict protocol for performance of FBS. The design and main results are presented elsewhere.(7;8)

In The Netherlands, pregnant women at low-risk are monitored by midwives or general practitioners at home or in hospital (primary care), whereas pregnant women at high-risk are monitored by gynecologists in hospital (secondary care). High-risk pregnancies are those that are complicated by pre-existing maternal disease, complicated obstetric history, hypertensive disorders, intrauterine growth restriction, ruptured membranes for more than 24 hours, a post date gestational age, failure to progress, need for pain relief, meconium stained amniotic fluid or non-reassuring fetal heart rate at intermittent auscultation by a midwife.

Outcome

The primary outcome in the original trial was metabolic acidosis based on a base deficit (BD) calculated in the extracellular fluid compartment (BDecf). However, in the present study, the primary outcome was metabolic acidosis defined as an umbilical cord-artery pH below 7.05 and a BD calculated in blood (BDblood) above 12 mmol/L. There were two reasons for using the latter outcome. First, most cord blood analyzers calculate and report BDblood instead of BDecf, which makes multivariable analysis for metabolic acidosis based on BDblood more applicable to daily clinical practice. Second, a practical reason was the higher incidence of metabolic acidosis calculated in blood, allowing for more statistical power to identify possible predictors.

Predictors under study

We selected candidate predictors for metabolic acidosis based on literature and clinical reasoning. Selected candidate predictors were divided into two categories. The first category contained only antenatal variables: maternal age, fetal gender, gestational age, parity, previous caesarean delivery, maternal diabetes mellitus, maternal hypertensive disorders and antenatal estimated fetal weight. For the latter we used a surrogate dichotomous variable using actual birth weight below tenth percentile according to birth weight reference curves of the Perinatal Registration in the Netherlands (PRN).(9;10) Maternal diabetes mellitus was defined as both pregestational type 1 and 2 as well as gestational diabetes mellitus. This first set of variables was used to develop an antenatal baseline risk model (model 1).

The second category contained variables obtained during labor i.e. onset of labor, intrapartum fever (≥ 38.5 °C), failure to progress, oxytocin augmentation, meconium stained amniotic fluid and epidural anesthesia. The cut-off for intrapartum fever was set higher than normal (38.5°C instead of 37.8°C) since women that receive epidural anesthesia are known to have elevated body temperature.(11;12) This set of characteristics was added to model 1, to determine their added predictive value and to develop a final extended model containing predictors measured both before and during labor (model 2).

Data analysis

Univariable associations between the candidate predictors and metabolic acidosis were estimated with logistic regression analysis. Pre-selection of predictors for inclusion in the multivariable logistic regression analyses was not done, since such selection based on univariable statistics often results in unstable prediction models.(13-16)

Maternal and gestational age were analyzed as continuous variables. Linearity of their association with the outcome was assessed using cubic spline analyses.(14) All other variables, including the outcome, were dichotomous. To correct for the allocated intervention followed by randomization in the original trial, we also included this intervention variable in

the multivariable analysis of model 1 and 2.

In this study the same imputed datasets were used as in the original trial(7;8), so for variables with missing values, which ranged from 0.2 to 11.9%, multiple imputation (ten times) was used to obtain complete data, without loss of power.(17-19) To account for differences between imputation sets predictor selection was performed in each imputation set separately. The predictors of the multivariable models were identified by backward stepwise selection using Akaike's Information Criterion, which is similar to selection based on a p-value of 0.157 if the predictor is modeled with one regression coefficient as for e.g. dichotomous predictors.(14;15) The predictors for the final multivariable model were selected using the majority method, meaning that predictors were selected in at least five out of ten imputed data sets.(20) Afterwards the regression coefficients of this model were estimated in each imputation set separately and combined using Rubin's rules to end up with the regression coefficients of the final multivariable model.(21)

The ability of the two models to discriminate between women with and without metabolic acidosis was studied with the area under the Receiver Operating Characteristic (ROC) curve (c-statistic). Calibration was assessed by comparing the predicted probabilities with the observed frequencies of metabolic acidosis. The agreement between the observed proportions of metabolic acidosis and the predicted risks, was studied with a calibration plot.(14;22;23) This plot also provides additional insight in the distribution of the predicted outcome incidences.

As the number of cases of metabolic acidosis was relatively low, there was a chance of finding spurious predictors and overestimated regression coefficients.(14;15;24) Such overfitted models will yield too extreme and optimistic predictions when applied in new women. To adjust for this we (internally) validated the models with bootstrapping techniques in which the predictor selection was repeated as well.(25) This yielded a shrinkage factor for the regression coefficients and for the c-statistic, to adjust both for optimism. All analyses including the bootstrapping techniques were performed in R version 2.10.0 (The R Foundation for Statistical Computing, 2009).

Results

Between January 2006 and July 2008, 5,681 women were randomised.(7;8) After randomization 14 women were excluded, because they did not meet the inclusion criteria. In the remaining 5,667, metabolic acidosis occurred in 107 neonates (incidence of 1.9%). Descriptive characteristics of the development population are presented in Table 1. The cubic spline analyses showed both maternal age and gestational age were linearly related to the outcome, i.e. metabolic acidosis.

Table 1. Characteristics of the development population (N=5,667). Data are presented as mean ± standard deviation or n (%).

Characteristic	
Antenatal	
Maternal age (y)	32.0 ± 4.8
Gestational age (wks)	40.2 ± 1.4
Multi parity	2431 (42.9)
Previous cesarean delivery	716 (12.6)
Neonatal female gender	2668 (47.1)
Birth weight below tenth percentile	519 (9.2)
Maternal diabetes mellitus	169 (3.0)
During labor	
Induced onset of labor	2341 (41.3)
Failure to progress	406 (7.2)
Oxytocin augmentation	2044 (36.1)
Meconium stained amniotic fluid	1471 (26.0)
Epidural anesthesia	2389 (42.2)
Intrapartum fever (≥ 38.5 °C)	109 (1.9)
Outcome	
Metabolic acidosis	107 (1.9)

Variables related to metabolic acidosis in univariable analysis were multi parity (odds ratio (OR) [95% confidence interval] 0.65 [0.43-0.99]), maternal diabetes (OR 2.71 [1.27-5.74]), induced onset of labor (OR 0.64 [0.42-0.97]), meconium stained amniotic fluid (OR 1.67 [1.10-2.52]) and epidural anesthesia (OR 1.52 [1.03-2.25]) (Table 2).

In model 1, four antepartum variables were identified to independently predict metabolic acidosis: gestational age (OR 1.12 [0.96-1.29] per week), multi parity (OR 0.60 [0.36-1.00]), previous caesarean delivery (OR 1.72 [0.87-3.39]) and maternal diabetes mellitus (OR 2.64 [1.21-5.72]) (Table 3). The model's c-statistic was 0.64 [0.59-0.69].

After adding all predictors obtained during labor to the variables of model 1, backward stepwise selection yielded that induced onset of labor (OR 0.74 [0.48-1.15]) and meconium stained amniotic fluid (OR 1.28 [0.83-1.96]) were independent predictors of neonatal metabolic acidosis as well (Table 3). The discriminative performance of model 2 was similar to the performance model 1 (c-statistic of model 2 0.66 [0.61-0.71]).

Table 2. Univariable associations between candidate predictors and metabolic acidosis.*Data are presented as mean ± standard deviation or n (%).

Predictor	Metabolic acidosis			
	Present N=107	Absent N=5560	Odds Ratio (95% CI)	P value§
Antenatal				
Maternal age (y)	31.7 ± 5.3	32.0 ± 4.8	0.98 (0.95-1.03)†	0.44
Gestational age (wks)	40.4 ± 1.4	40.2 ± 1.4	1.14 (0.98-1.31)‡	0.07
Multi parity	35 (32.7)	2396 (43.1)	0.65 (0.43-0.99)	0.04
Previous cesarean delivery	16 (15.0)	700 (12.6)	1.21 (0.70-2.08)	0.49
Neonatal female gender	44 (41.1)	2624 (47.2)	0.79 (0.54-1.18)	0.24
Birth weight below tenth percentile	12 (11.2)	507 (9.1)	1.21 (0.64-2.28)	0.55
Maternal diabetes mellitus	8 (7.5)	161 (2.9)	2.71 (1.27-5.74)	0.01
During labor				
Induced onset of labor	34 (31.8)	2308 (41.5)	0.64 (0.42-0.97)	0.04
Intrapartum fever	3 (2.8)	106 (1.9)	1.42 (0.42-4.83)	0.56
Failure to progress	12 (11.2)	394 (7.1)	1.70 (0.91-3.16)	0.08
Oxytocin augmentation	48 (44.9)	1996 (35.9)	1.46 (0.99-2.16)	0.05
Meconium stained amniotic fluid	39 (36.4)	1432 (25.8)	1.67 (1.10-2.52)	0.01
Epidural anesthesia	56 (52.3)	2333 (42.0)	1.52 (1.03-2.25)	0.03

* Absolute numbers (%) in this table are based on the mean of ten imputations.
† Odds ratio per year.
‡ Odds ratio per week.
§ P-values were calculated using Wald test statistic

Table 3. Multivariable logistic regression models for the prediction of metabolic acidosis.

Predictor	Model 1		Model 2	
	OR (95% CI) †	Beta†	OR (95% CI) ‡	Beta‡
Intercept		-8.14		-7.92
Randomization allocation	0.71 (0.47-1.06)	-0.35	0.72 (0.48-1.08)	-0.33
Antenatal				
Gestational age (wks) *	1.12 (0.96-1.29)	0.11	1.11 (0.95-1.30)	0.10
Multi parity	0.60 (0.36-1.00)	-0.52	0.65 (0.39-1.08)	-0.43
Previous cesarean delivery	1.72 (0.87-3.39)	0.54	1.54 (0.77-3.05)	0.43
Maternal diabetes mellitus	2.64 (1.21-5.72)	0.97	2.74 (1.25-6.01)	1.01
During labor				
Induced onset of labor	-	-	0.74 (0.48-1.15)	-0.29
Meconium stained amniotic fluid	-	-	1.28 (0.83-1.96)	0.25

* Odds ratio (OR) per week
† Shrunken with an average shrinkage factor of 0.74
‡ Shrunken with an average shrinkage factor of 0.71

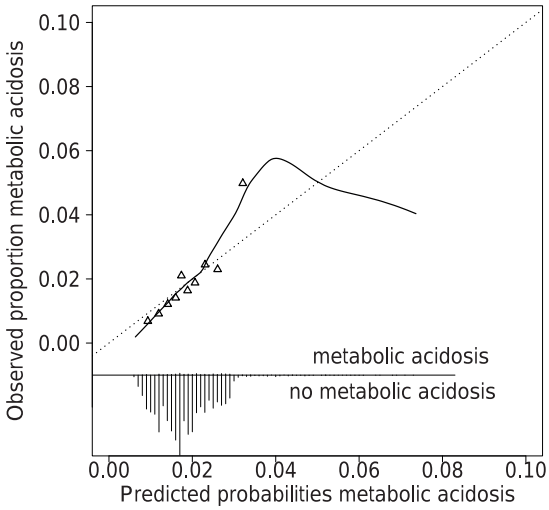


Figure 1 Calibration plot of model 1 with the observed risk of metabolic acidosis by predicted probabilities of metabolic acidosis. The triangles indicate deciles of women with similar predicted risk. Vertical bars (histogram) indicate the frequencies (number of women) across the predicted probabilities. To enhance interpretation the axes were adjusted to a scale from 0.0 to 0.10, based on the low observed and predicted outcome incidences.

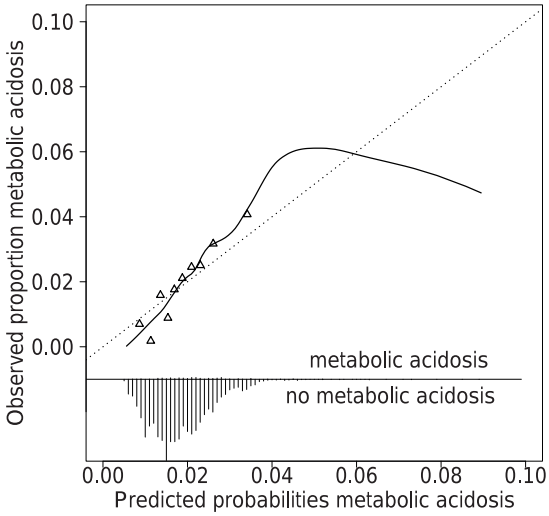


Figure 2 Calibration plot of model 2 with the observed risk of metabolic acidosis by predicted probabilities of metabolic acidosis. The triangles indicate deciles of women with similar predicted risk. Vertical bars (histogram) indicate the frequencies (number of women) across the predicted probabilities. To enhance interpretation the axes were adjusted to a scale from 0.0 to 0.10, based on the low observed and predicted outcome incidences.

Calibration of both models is shown in figures 1 and 2. At first glance these models seem to calibrate poorly, but to enhance the interpretation the axes were adjusted based on the a priori low observed (and thus predicted) outcome incidences. For model 1, the predicted probabilities ranged from 0.6 to 7.4% (Figure 1) and for model 2 from 0.6 to 8.9% (Figure 2). For model 1 calibration was good for the group with predicted probabilities below 2.5%. Some underestimation was seen for predicted risks between 2.5 and 5% and some overestimation at higher risks (Figure 1). In model 2, calibration was good until predicted probabilities of 2.5%. Underestimation was seen for predicted risks between 2.5 and 6%. At higher risks, there was some overestimation, which was largely due to the low number of cases in this group (Figure 2).

Discussion

To date, this study is the first to gain quantitative insight into the predictive capacity of combinations of obstetric characteristics obtained before and during labor for metabolic acidosis of the newborn at birth. Previous studies mainly focused on the prediction of longer-term outcomes, such as hypoxic ischemic encephalopathy, cerebral palsy and death and frequently applied a univariable approach. From a recent systematic review and meta-analysis it is known that a low cord pH at birth is highly associated with clinically important neonatal and long-term outcomes.(26) Hence, prevention of a low cord pH at birth also prevents cases with long-term adverse outcomes. Early knowledge of a woman's individual risk of developing such adverse outcome, preferably at an early stage, enhances the possibilities to anticipate the intrapartum monitoring, decision and management process.

Our results show that before the start of labor in high-risk singleton term vertex pregnancies, e.g. higher gestational age (e.g., 42 weeks) and a history of caesarean delivery increase the risk of metabolic acidosis already twofold (4.1%), as compared to the a priori risk (overall observed incidence of 1.9%). Since the discriminative ability of both models is similar one might question whether factors obtained during labor, such as induced onset and meconium stained amniotic fluid, should be included in a prediction model for neonatal metabolic acidosis, since these factors have minor added discriminative value and a model with only ante partum factors (model 1) would be more easily applicable. Despite the similar performance, meconium stained amniotic fluid still increases the risk found for model 1 even further to 4.7% risk of metabolic acidosis. Our results may allow clinicians to avoid unnecessary interventions in low-risk women and may influence decisions during labor regarding the interpretation of fetal heart rate patterns and the application of additional techniques for fetal monitoring, such as ST-analysis of the fetal ECG or FBS.

Although the aim of prognostic research is purely to predict,(6) the results of our study show that the expected direction of a relation between a predictor and the outcome does not

always match the calculated direction of the relation. In advance, one would expect women with an induced onset of labor to be at higher risk for metabolic acidosis at birth. The relation for induced onset of labor had an unexpected protective effect in our analyses. This may be explained as follows. First, despite analyzing data from a large randomized multicenter trial, model development might be influenced by mixing of two obstetric populations: one consisting of women laboring in secondary care, the other consisting of women referred to secondary care during labor from primary care. The latter group of women was referred for different reasons, i.e. wish for pain relief or failure to progress which may increase the risk of metabolic acidosis. Obviously, all of these women had a spontaneous onset of labor since they were referred during labor. Consequently, when metabolic acidosis has a higher incidence in these women with a spontaneous onset of labor as compared to women with an induced onset of labor, a protective effect of induced onset of labor on metabolic acidosis might be found. Second, it is plausible that induction of labor was conducted with far more closer supervision than a spontaneous onset of labor and that corrective action would be taken more quickly and more appropriately when deemed necessary. Since many of the women in the trial were counselled in the morning for participation in the trial and these women are often scheduled for an induced onset of labour the percentage of women in the trial with an induced onset of labour was higher than average. Combined with the closer supervision this would automatically lower the frequency of metabolic acidosis and therefore lead to a seemingly protective effect of a factor like induction of labor.

Another striking result of our study was that oxytocin augmentation did not appear to be a risk factor for metabolic acidosis. In the Netherlands, oxytocin augmentation substantially contributes to the number of admissions from primary to secondary care, in order to continuously monitor the fetus. The results of our study imply that other obstetric characteristics more importantly contribute to the prediction of metabolic acidosis.

As mentioned before, the prediction models might be influenced by mixing two obstetric populations. Referrals from primary care frequently occur in nulliparous women(27), and for reasons as i.e. fetal distress, failure to progress, a need for pain relief, or meconium stained amniotic fluid.(28) Since information on the transfer from primary to secondary care was not available, we couldn't include the reason for referral in the model. Alternatively, we chose to include parity, previous caesarean delivery, epidural analgesia and meconium stained amniotic fluid as potential predictors for metabolic acidosis. Epidural analgesia (i.e., one reason for referral) showed to be predictive in the univariable analysis, but was however not included in the multivariable model. Apparently, the predictive effect of epidural analgesia is already captured by other antenatal and during labor available variables. Likewise, other reasons for referral might actually be captured in the model and hence the model derived in the mix of two obstetric populations will likely generalize to both.

One could argue why potentially important predictors like duration of labor, performance of

fetal blood sampling, mode of delivery were not included in the model. Duration of labor, mode of delivery and – to a lesser extent – fetal blood sampling were not taken into account because these characteristics could only be determined at the end of the delivery and are therefore of less interest when one wants to perform a timely prognostication of metabolic acidosis.

To appreciate the present results, a final aspect needs to be addressed. Metabolic acidosis, which is the combined presence of both a low umbilical cord-artery pH and BD, is often used as a biochemical derivative of asphyxia. Several cut-off values for pH and BD are being used, although a consensus statement has reported pathological acidosis to occur at an umbilical cord-artery pH below 7.00 and a BD above 12 mmol/L.(29) However, in various papers often a slightly more liberal definition is used: umbilical cord-artery pH below 7.05 and a BD above 12 mmol/L.(30) In our study, the use of the latter definition, consisting of substantially more cases than for an umbilical cord-artery pH below 7.00, made the identification of predictors more powerful.

Our work shows that there is a strong variation in the antenatal baseline risk for acidosis, and this risk may even be adjusted with factors that occur during labor. At present, however, such information is hardly used in clinical practice. It should be noted that our study only aimed to determine which variables are independently predicting the occurrence of metabolic acidosis. Therefore the discriminative ability of both models (c-statistic of 0.64 and 0.65, respectively) is of less importance, especially because this is the best available model to predict neonatal metabolic acidosis.(31) Despite our internal validation procedure, further steps of model development and validation are surely needed before a model can be introduced in clinical practice.(15;16) Specific attention should be paid to applicability of the models to new laboring women, thereby remarking that the antepartum model was actually developed by using data on women that were in labor already at the moment of entering the trial.

As figures 1 and 2 show, the predicted risk of acidosis is below 1% in about 10% of women and between 1% and 2% in about 50%. Furthermore both models showed good calibration within this range with good agreement between the predicted risks of metabolic acidosis and the observed proportion of metabolic acidosis. From a clinical viewpoint it is of minor importance that at higher risks the calibration of our models is less good, because with a risk above 2% women will be categorized in the highest risk group and managed as such anyway. The variation in predicted risks is of the utmost importance, as this is guiding both in the type of fetal surveillance that is offered, as well as in its interpretation.

In summary, in women with a high-risk singleton term pregnancy in cephalic presentation, factors both before and during labor influence the a priori probability of developing metabolic acidosis of the newborn at birth. The identified predictors were internally validated by bootstrapping techniques. After external validation and proof of generalizability, this information should be used in obstetric clinical management.

References

(1) van de Riet JE, Vandenbussche FP, Le CS, Keirse MJ. Newborn assessment and long-term adverse outcome: a systematic review. Am J Obstet Gynecol 1999 Apr;180(4):1024-9.

(2) Goodwin TM, Belai I, Hernandez P, Durand M, Paul RH. Asphyxial complications in the term newborn with severe umbilical acidemia. Am J Obstet Gynecol 1992 Dec;167(6):1506-12.

(3) Williams KP, Singh A. The correlation of seizures in newborn infants with significant acidosis at birth with umbilical artery cord gas values. Obstet Gynecol 2002 Sep;100(3):557-60.

(4) van den Berg PP, Nelen WL, Jongsma HW, Nijland R, Kollee LA, Nijhuis JG, et al. Neonatal complications in newborns with an umbilical artery pH < 7.00. Am J Obstet Gynecol 1996 Nov;175(5):1152-7.

(5) Lavrijsen SW, Uiterwaal CS, Stigter RH, de Vries LS, Visser GH, Groenendaal F. Severe umbilical cord acidemia and neurological outcome in preterm and full-term neonates. Biol Neonate 2005;88(1):27-34.

(6) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

(7) Westerhuis ME, Visser GH, Moons KG, van BE, Benders MJ, Bijvoet SM, et al. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2010 Jun;115(6):1173-80.

(8) Westerhuis ME, Visser GH, Moons KG, Zuithoff N, Mol BW, Kwee A. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2011 Feb;117(2 Pt 1):406-7.

(9) Stichting Perinatale Registratie Nederland (PRN). Reference curves. Available at: http://www.perinatreg.nl/referentiecurven. Retrieved September 13. 2009.

(10) Visser GH, Eilers PH, Elferink-Stinkens PM, Merkus HM, Wit JM. New Dutch reference curves for birthweight by gestational age. Early Hum Dev 2009 Dec;85(12):737-44.

(11) Fusi L, Steer PJ, Maresh MJ, Beard RW. Maternal pyrexia associated with the use of epidural analgesia in labour. Lancet 1989 Jun 3;1(8649):1250-2.

(12) Viscomi CM, Manullang T. Maternal fever, neonatal sepsis evaluation, and epidural labor analgesia. Reg Anesth Pain Med 2000 Sep;25(5):549-53.

(13) Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996 Aug;49(8):907-16.

(14) Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-87.

(15) Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ 2009;338:b604.

(16) Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

(17) van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006 Oct;59(10):1102-9.

(18) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-91.

(19) Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995 Dec 15;142(12):1255-64.

(20) Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol 2010 Feb;63(2):205-14.

(21) Rubin DB. Multiple imputation for non response in surveys. New York: Wiley; 1987.

(22) Atkinson A. A note on the generalized information criterion for choice of a model. Biometrika 2010;67:413-8.

(23) Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991 Aug;10(8):1213-26.

(24) Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004 Aug 30;23(16):2567-86.

(25) Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003 May;56(5):441-7.

(26) Malin GL, Morris RK, Khan KS. Strength of association between umbilical cord pH and perinatal and long term outcomes: systematic review and meta-analysis. BMJ 2010;340:c1471.

(27) Evers AC, Brouwers HA, Hukkelhoven CW, Nikkels PG, Boon J, van Egmond-Linden A, et al. Perinatal mortality and severe morbidity in low and high risk term pregnancies in the Netherlands: prospective cohort study. BMJ 2010;341:c5639.

(28) Amelink-Verburg MP, Rijnders ME, Buitendijk SE. A trend analysis in referrals during pregnancy and labour in Dutch midwifery care 1988-2004. BJOG 2009 Jun;116(7):923-32.

(29) MacLennan A. A template for defining a causal relation between acute intrapartum events and cerebral palsy: international consensus statement. BMJ 1999 Oct 16;319(7216):1054-9.

(30) Low JA, Lindsay BG, Derrick EJ. Threshold of metabolic acidosis associated with newborn complications. Am J Obstet Gynecol 1997 Dec;177(6):1391-4.

(31) Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

Chapter 2.3

Prediction of neonatal metabolic acidosis in women with a singleton term pregnancy in cephalic presentation: an external validation study.

Schuit E, Amer-Wahlin I, Groenwold RH, Mol BW, Moons KG, Kwee A.

Am J Perinatol 2012 Oct;29(9):681-6.

Abstract

Objective To externally validate two previously developed prognostic models that predict the risk for developing metabolic acidosis in newborns using both antepartum (model 1) and intrapartum (combined with antepartum, model 2) risk factors: parity, previous cesarean section, maternal diabetes mellitus, gestational age, induced onset of labor, meconium stained amniotic fluid and use of ST-analysis.

Study Design The two prediction models were applied in women in active labor at more than 36 gestational weeks, with singleton fetuses in cephalic presentation, and with high-risk pregnancies (n=5049) included in a Swedish randomized trial between 1 December 1998 and 4 June 2000. The prognostic ability of the models was determined using calibration and discrimination measures.

Results Of 5049 infants in the validation population, 54 (1.1%) suffered from metabolic acidosis. After adjustment for incidence differences between the Dutch and Swedish cohorts, the prognostic models showed good calibration and moderate overall discrimination (c statistic 0.63 (95% CI 0.55-0.71) and 0.64 (95% CI 0.55-0.72), for model 1 and 2 respectively).

Conclusions External validation of the clinical prediction models for metabolic acidosis in Swedish infants showed good calibration and moderate discriminative ability. Updating of the models to enhance their predictive abilities seems indicated.

Introduction

Perinatal asphyxia is associated with several short- and long-term complications, varying from mild hypoxic ischemic encephalopathy, to cerebral palsy and death.(1-5) Perinatal asphyxia has an incidence as low as 2% in the western world but can have a major impact on newborn infants and their families.(3)

Since prevention of asphyxia is dependent on timely identification, adequate prediction of asphyxia either before or during delivery is important. Timely prognostication, for example at admission to the hospital or even antenatally, may lead to more effective decision making and possibly prevention, prior to and during labor, both in relation to delivery mode and fetal monitoring method offered, specifically in relation to the interpretation of fetal heart rate patterns.

In a previous study, factors measured both before and during labor were identified and combined in a multivariable way to adequately predict metabolic acidosis at birth.(6) Two models were developed on data from a large multicenter trial in which 5667 laboring women with a high-risk cephalic singleton pregnancy beyond 36 weeks of gestation.(7) The study showed multiparity, previous cesarean section, maternal diabetes, gestational age and use of ST-analysis for intrapartum monitoring to be antepartum predictors of metabolic acidosis (model 1).(6) Additional intrapartum predictors were spontaneous onset of labor and meconium stained amniotic fluid (model 2).(6) Both models showed good calibration and moderate discrimination with c statistics of 0.64 (95% CI 0.59-0.69) and 0.66 (95% CI 0.61-0.71), for the two models respectively. (6)

To show that prognostic models are valuable, evidence is needed that the models perform well in groups of patients other than on which they were originally developed.(8) Although developed in a Dutch population, the models may also be useful for other countries. Since local and timely variation in attitude to women at risk of giving birth to an infant with metabolic acidosis can influence the validity of the models, we studied the generalizability of the models in Swedish newborns included in a randomized trial between 1998 – 2000.(9)

Materials and Methods

Study population

We studied mothers and their newborns included in a Swedish multicenter randomized controlled trial (RCT) between December 1, 1998 and June 4, 2000.(10) Participants of the trial were women in active labor at more than 36 gestational weeks, with singleton fetuses in cephalic presentation, and with high-risk pregnancies. This includes women with suspicious or abnormal external cardiotocography, induced or oxytocin-augmented labor, meconium-stained amniotic fluid, or epidural analgesia. The women were allocated to either

intrapartum monitoring by cardiotocography plus ST-analysis of the fetal electrocardiogram or cardiotocography only to investigate the additional effect of ST-analysis on prevention of birth asphyxia. The primary outcome of the trial was metabolic acidosis at birth, which was defined as an umbilical cord-artery blood pH of less than 7.05 and a base deficit in the extracellular fluid compartment (BDecf) of more than 12.0 mmol/L. Details on the design and the results of the RCT are presented elsewhere.(9)

Prognostic models

We evaluated two models that predict metabolic acidosis(6) in a Swedish cohort. The models were developed on women from a randomized clinical trial in The Netherlands.(7) The trial included laboring women with a high-risk cephalic singleton pregnancy beyond 36 weeks of gestation that were allocated to either intrapartum fetal monitoring using cardiotocography plus ST-analysis or cardiotocography alone. High-risk pregnancies were pregnancies complicated by pre-existing maternal disease, complicated obstetric history, hypertensive disorders, intrauterine growth restriction, ruptured membranes for more than 24 hours, a post date gestational age, failure to progress, need for pain relief, meconium stained amniotic fluid or non-reassuring fetal heart rate at intermittent auscultation by a midwife.

The outcome of the prediction model was metabolic acidosis, defined as an umbilical cord-artery blood pH of less than 7.05 and a base deficit (BD) in blood (BDblood) of more than 12.0 mmol/L. This definition was different than the ones used in the Dutch and Swedish trial(7;9) where BDecf was used.(11) Most cord blood analyzers calculate and report the base deficit in blood instead of in the extracellular fluid compartment, which makes multivariable analysis for metabolic acidosis based on BDblood more applicable to daily clinical practice. Second, a practical reason was the higher incidence of metabolic acidosis found when calculating in blood, allowing for more statistical power to identify possible predictors.(6) However, since BDblood was not available for the Swedish trial the definition of metabolic acidosis was based on BDecf.

The models included multiparity, previous cesarean section, maternal diabetes mellitus and gestational age as antepartum predictors of metabolic acidosis (model 1). Additional intrapartum predictors were spontaneous onset of labor and meconium stained amniotic fluid (model 2). To take the allocated intervention of the original trial into account, the intervention variable, use of ST-analysis, was included in both models as well. Internal validation, after correction for optimism by bootstrapping, showed acceptable calibration and discrimination with c statistics of 0.64 (95% CI 0.59-0.69) and 0.66 (95% CI 0.61-0.71) for model 1 and 2, respectively.

Statistical analysis

Baseline characteristics of the development and validation population(s) were determined using descriptive statistics and presented as mean plus/minus a standard deviation for continuous variables and as numbers and percentages of the study population for categorical and dichotomous variables.

To adjust for a difference in the definition (and as a result the incidence) of the outcome (i.e. based on BDecf in the validation population rather than BDblood in the development population) the intercept of the original model was updated, which is common practice when validating prediction models in external data.(12) A correction factor was calculated by taking the logarithm of the total result of the incidence of the outcome in the validation set divided by 1 minus this incidence, divided by, the mean predicted risk in the validation set divided by 1 minus this mean. By adding this correction factor to the intercept of the original model, the model was updated.(12)

The validity of the prognostic models was assessed in terms of calibration and discrimination. Calibration refers to the agreement between the predicted risk of metabolic acidosis and the observed proportion of infants with metabolic acidosis in the validation population. Calibration was assessed graphically with a calibration plot.(13-15) Discrimination refers to the ability of the model to distinguish between neonates who died and who survived and was assessed with the concordance- or c-statistic.(14;15) For binary outcomes, the c statistic is identical to the area under the receiver operating characteristic curve.(14;15) To investigate whether the distribution of patient characteristics (case-mix) influenced the discriminative ability of the models we assessed the performance of both models by refitting them in the study population.(16) This allows for an upper bound for the discrimination of the models in the validation population.

Various subjects had missing values (ranging from 0% for fetal gender to 16.2% for metabolic acidosis). Because these are often selectively missing, which was also the case here (Table 1), it is well documented that a complete case analysis likely yields biased results.(17-19) Hence, before performing the analyses the missing values were imputed using multiple imputation (ten times). Information on one predictor in both models, i.e. previous cesarean delivery, was partly unavailable in the validation population. To overcome this problem we decided to impute information on previous caesarean delivery in multiparous women (nulliparous were all set to no previous caesarean delivery) based on the Dutch development population. (12) The performance of both models was estimated in each imputation set separately and combined using Rubin's rules to end up with the overall performance of both prediction models in the validation population.

All analyses were performed in R version 2.10.0 (The R Foundation for Statistical Computing, 2009).

Table 1 Distribution of Missing Values for Outcomes Related to Cord-Artery pH Among Baseline and Outcome Variables. Plus-minus values are means ± SD. P-values were calculated via Chi-square or Student's t-tests for dichotomous or continuous variables, respectively. Variables in bold indicate that missing data were not completely at random but related to other subject characteristics, which indicates not to perform a complete case analysis but rather apply multiple imputation first (see text).

Characteristic	Missing	Complete cases (N=4126)	Participants with at least one missing value (N=923)	P-value
Gestational age, wks	15 (0.3%)	39.7 ± 1.5	39.4 ± 2.0	< 0.01
Arterial pH	723 (14.3%)	7.22 ± 0.08	7.19 ± 0.1	< 0.01
Arterial base deficit	873 (17.3%)	5.1 ± 3.2	5.2 ± 2.9	0.81
ST-analysis	0 (0%)	2094 (51%)	471 (51%)	0.91
Multiparity	66 (1.3%)	1601 (39%)	320 (36%)	0.45
Female gender	26 (0.5%)	1937 (47%)	439 (49%)	0.29
Maternal diabetes mellitus	83 (1.6%)	88 (2%)	16 (2%)	0.77
Induced onset of labor	22 (0.4%)	709 (17%)	157 (17%)	0.90
Failure to progress	21 (0.4%)	390 (9%)	88 (10%)	0.83
Meconium stained amniotic fluid	54 (1.1%)	941 (23%)	202 (23%)	0.81
Epidural anesthesia	0 (0%)	1629 (39%)	328 (36%)	0.03
Metabolic acidosis	731 (14.5%)	45 (1%)	1 (1%)	0.69

Results

Between 1 December 1998 and 4 June 2000, 5049 women with a high-risk singleton pregnancy were randomized into the Swedish trial. Of their children, 54 (1.1%) suffered from metabolic acidosis.

Characteristics of the population

The characteristics of the Swedish validation population are shown in Table 2 next to the Dutch population in which the two prediction models were developed. In general, the characteristics of the Swedish cohort were similar to the characteristics of the cohort from the Netherlands. However, a difference was found for the number of women that had an induced onset of labor which was lower in Sweden when compared to the Netherlands (17% vs. 41%). Furthermore, as a result of the difference in outcome definition the proportion of infants that suffered from metabolic acidosis was lower in Sweden than in the Netherlands (1.1% vs. 2.1%).

Table 2 Characteristics of high-risk singleton pregnancies in the Swedish and Dutch cohorts. Data are presented as mean ± SD or n (%). Variables with an asterisk were included in model 1 and/or 2.

Characteristic	Swedish validation cohort 1998-2000 (N=5049)	Dutch development cohort 2006-2008 (N=5681)
Antenatal		
Gestational age (wks)*	39.6 ± 1.6	40.2 ± 1.4
Multi parity*	1948 (38.6)	2431 (42.9)
Previous cesarean section*	626 (12.4)	716 (12.6)
Female gender	2388 (47.3)	2668 (47.1)
Maternal diabetes mellitus*	111 (2.2)	169 (3.0)
Use of ST-analysis*	2565 (50.8)	2827 (49.9)
Intrapartum		
Induced onset of labor*	868 (17.2)	2341 (41.3)
Failure to progress	478 (9.5)	406 (7.2)
Meconium stained amniotic fluid*	1154 (22.9)	1471 (26.0)
Epidural anesthesia	1957 (38.8)	2389 (42.2)
Outcome		
Metabolic acidosis*	54 (1.1)	119 (2.1)

Performance of the prognostic models

After updating of the intercept of the original models, predicted risks of metabolic acidosis were in agreement with the observed proportions of metabolic acidosis (Figure 1 and 2). The discriminative ability of the models was moderate to poor with a c statistic lower than the one found for internal validation of 0.63 (95% CI 0.55-0.71) for the antepartum model (model 1) and 0.64 (95% CI 0.55-0.72) for the intrapartum model (model 2). Figures 1 and 2 show that the range of predicted probabilities was similar in both models with a difference of 1.1% between the means of the lowest (0.5%) and highest (1.7%) predicted perinatal risk deciles (first and last triangles, respectively, in Figure 1) for model 1 and a difference of 1.3% between the lowest (0.5%) and highest (1.8%) predicted perinatal risk deciles for model 2 (Figure 2).

Both models were refitted (i.e. the regression coefficients of the predictors of the models were estimated) in the Swedish study population to investigate whether the distribution of patient characteristics (case-mix) influenced the discriminative ability of the models.(16) This resulted in a c-statistic of 0.66 (95% CI 0.58-0.74) for model 1 and 0.66 (95% CI 0.58-0.74) for model 2. Hence, the found upper bound of the performance was higher than the performance found in the validation of the models.

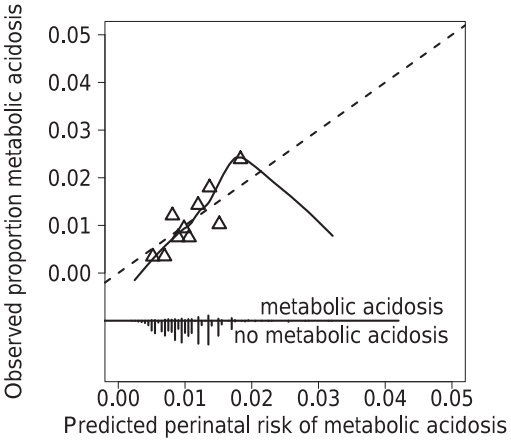


Figure 1 Observed risk of metabolic acidosis by predicted antepartum risk for term neonates in high-risk singleton pregnancies for model 1. The triangles indicate the deciles with confidence intervals of infants with similar predicted risk. Vertical bars (histogram) indicate the frequencies (number of women) across the predicted probabilities. To enhance interpretation the axes have been adjusted to a scale from 0.0 to 0.05, based on the low observed and predicted outcome incidences.

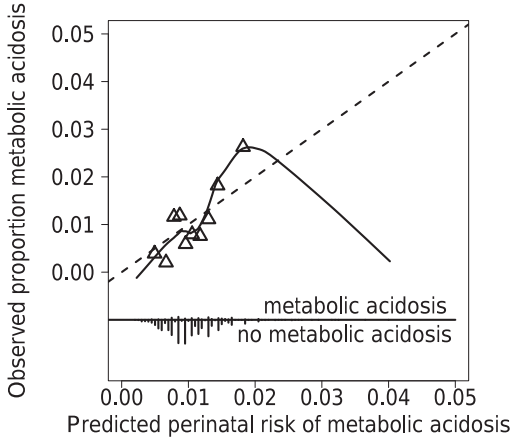


Figure 2 Observed risk of metabolic acidosis by predicted intrapartum risk for term neonates in high-risk singleton pregnancies for model 2. The triangles indicate the deciles with confidence intervals of infants with similar predicted risk. Vertical bars (histogram) indicate the frequencies (number of women) across the predicted probabilities. To enhance interpretation the axes have been adjusted to a scale from 0.0 to 0.05, based on the low observed and predicted outcome incidences.

Discussion

In this external validation study we assessed the performance of two prognostic models predicting metabolic acidosis in neonates from a high-risk singleton pregnancy. The models that were previously developed with data of Dutch mothers and their infants born between 2006 and 2008 were found to be valid in a large cohort of five thousand Swedish mothers and neonates born between 1998 and 2000.

The models showed good calibration and moderate ability to discriminate between pregnancies and neonates that did and did not develop metabolic acidosis. The c-statistics were slightly lower as found by internal validation (i.e. 0.63 (95% CI 0.55-0.71) vs. 0.64 (95% CI 0.59-0.69) for model 1 and 0.64 (95% CI 0.55-0.72) vs. 0.66 (95% CI 0.61-0.71) for model 2).(6) Given that the incidence of metabolic acidosis is low and the highest predicted/observed risk in the second model is not higher than 3.9% (Figure 2), discrimination is reasonable. Obviously, both calibration and discrimination could be better. However, since these models were the best available models to predict metabolic acidosis, it was more of interest whether the performance of the models in the Swedish validation population was similar to the performance in the Dutch development population than the actual value of the c statistic or the calibration. Refitting the models in the Swedish validation population resulted in higher c statistics than found by external validation, indicating that regression coefficients would differ in the Dutch development and Swedish validation population. Considering the low number of events in the validation population (i.e. 50) such refitting of the prediction model is much more prone to overfitting than when originally fitting the model in the derivation population (number of events 109). This is also indicated by a shrinkage factor of 0.78 and 0.76 for model 1 and 2, respectively, which indicates large optimism.

Information on the risk of metabolic acidosis in infants from a high-risk singleton pregnancy can be of great value in obstetric clinical management. For example, the models could improve the awareness of the clinician for metabolic acidosis. Furthermore, the second model might help trainees in their judgment when to call their attendant for assistance in deciding whether or not the child should be delivered.

The characteristics used in the model are known to influence the outcome of the newborn and might be familiar to caregivers. The large advantage of using the prognostic model is that the characteristics are combined in a more formal way allowing for more precise metabolic acidosis risk estimation. It is important to emphasize, however, that the prognostic models may complement clinical decision making rather than replace clinical judgment.

The prognostic models include a small number of readily available characteristics, which makes them easily applicable in daily clinical practice. To our knowledge, this is the first external validation study of a prognostic model for metabolic acidosis. Since local difference might influence the performance of prognostic models it is very important to validate these

in the population in which it is aimed to be used.
Since information on mother and child was obtained in a randomized trial data collection was standardized, resulting in high quality data. The cohort included data from an earlier (1998-2000) time period than the development cohort. It could be a limitation that the validation population was collected in a time period before the development population was collected. However, management of labor and methods used for intrapartum fetal surveillance have not changed in this time period. Therefore, we do not expect this aspect to influence the results of our study.
A second limitation was that information on previous cesarean delivery (a predictor in both models) was partly unavailable in the validation population. However, omitting such a predictor with a lot of missing values from the prediction model is a biased solution, and hence not recommended.(20) Multiple imputation is generally considered the preferred method of handling missing predictor data.(20)
A relatively large difference was found for the proportion of women that had an induced onset of labor between the validation and development populations (17% in Sweden vs. 41% in The Netherlands). The proportion of women with induced onset of labor found in the study by Westerhuis et al.(7) was actually higher than the proportion in the general population, which is about 30%.(21) Many of the women were counseled in the morning for participation in the trial. Since these women are often scheduled for an induced onset of labor the percentage of women in the trial with an induced onset of labor was higher than average.
The relatively low proportion of Swedish infants that suffered from metabolic acidosis (1% in Sweden and 2% in the Netherlands) resulted from a difference in definition. In the Swedish trial, metabolic acidosis was based on the base deficit in the extracellular fluid compartment while in the Dutch population on which the model was developed it was based on the base deficit in blood. Unfortunately, data on BDblood were not available from the Swedish population. Moreover, it is impossible to recalculate BDblood from the BDecf. It is well-known that the base deficit in umbilical cord arterial blood is influenced by the choice of fetal fluid compartment and the calculation algorithms that are used. Calculation of BD in blood will result in higher incidences of metabolic acidosis.(22) This is confirmed by several randomized controlled trials showing lower numbers of metabolic acidosis.(7;9) The intercept was adjusted to account for the difference in incidence of metabolic acidosis, resulting in good calibration and a slightly lower c statistic as compared to the internal validation. Therefore we do not expect this difference to influence the performance of the models.
The ante- and intrapartum characteristics investigated in this study are factors that are known to be related to neonatal metabolic acidosis. However, there may be other characteristics that are predictive of metabolic acidosis, such as maternal characteristics (e.g., body mass index and smoking), demographic characteristics (e.g., social economic status), pregnancy characteristics (e.g., low amniotic fluid volume) or intrapartum characteristics (e.g., the

dosage of oxytocin augmentation). None of these characteristics have been considered in the original model, since these were not available in the data set that was used to derive the prediction models. Future studies aimed at the prediction of metabolic acidosis could include these variables in order to further improve the prognostic accuracy of the models validated in this study.

Up-to-date, validation studies are scarce.(23) However, it is very important to assess the performance (i.e. generalizability) of clinical prediction model in a population different from the development population.(8) As this study showed, validation studies can be difficult to perform, since it is difficult to find a population that resembles the one on which the model is developed, for example the outcome might be defined differently or information on one of the predictors is missing. However, differences might improve the claim of generalizability in the case that a model shows good performance despite the differences between populations. So, despite the problems encountered in this study we feel confident that the results of this study are still relevant, especially since the models could add to the quality and safety in intrapartum care.

In conclusion, external validation of the clinical prediction models for metabolic acidosis in Swedish data showed good calibration and moderate to poor discrimination. Geographical and temporal generalizability therefore seems moderate. Updating of the models to enhance their predictive abilities seems indicated.

References

(1) Goodwin TM, Belai I, Hernandez P, Durand M, Paul RH. Asphyxial complications in the term newborn with severe umbilical acidemia. Am J Obstet Gynecol 1992 Dec;167(6):1506-12.

(2) Lavrijsen SW, Uiterwaal CS, Stigter RH, de Vries LS, Visser GH, Groenendaal F. Severe umbilical cord acidemia and neurological outcome in preterm and full-term neonates. Biol Neonate 2005;88(1):27-34.

(3) van de Riet JE, Vandenbussche FP, Le CS, Keirse MJ. Newborn assessment and long-term adverse outcome: a systematic review. Am J Obstet Gynecol 1999 Apr;180(4):1024-9.

(4) van den Berg PP, Nelen WL, Jongsma HW, Nijland R, Kollee LA, Nijhuis JG, et al. Neonatal complications in newborns with an umbilical artery pH < 7.00. Am J Obstet Gynecol 1996 Nov;175(5):1152-7.

(5) Williams KP, Singh A. The correlation of seizures in newborn infants with significant acidosis at birth with umbilical artery cord gas values. Obstet Gynecol 2002 Sep;100(3):557-60.

(6) Westerhuis ME, Schuit E, Kwee A, Zuithoff NP, Groenwold RH, VAN DEN Akker ES, et al. Prediction of Neonatal Metabolic Acidosis in Women with a Singleton Term Pregnancy in Cephalic Presentation. Am J Perinatol 2011 Aug 3.

(7) Westerhuis ME, Visser GH, Moons KG, van BE, Benders MJ, Bijvoet SM, et al. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2010 Jun;115(6):1173-80.

(8) Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

(9) Amer-Wahlin I, Kjellmer I, Marsal K, Olofsson P, Rosen KG. Swedish randomized controlled trial of cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram revisited: analysis of data according to standard versus modified intention-to-treat principle. Acta Obstet Gynecol Scand 2011 Sep;90(9):990-6.

(10) Amer-Wahlin I, Hellsten C, Noren H, Hagberg H, Herbst A, Kjellmer I, et al. Cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomised controlled trial. Lancet 2001;358(9281):534-8.

(11) Siggaard-Andersen O. An acid-base chart for arterial blood with normal and pathophysiological reference areas. Scand J Clin Lab Invest 1971 May;27(3):239-45.

(12) Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. Can J Anaesth 2009 Mar;56(3):194-201.

(13) Atkinson A. A note on the generalized information criterion for choice of a model. Biometrika 2010;67:413-8.

(14) Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-87.

(15) Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991 Aug;10(8):1213-26.

(16) Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010 Oct 15;172(8):971-80.

(17) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-91.

(18) Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995 Dec 15;142(12):1255-64.

(19) van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006 Oct;59(10):1102-9.

(20) Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Jr., Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem 2009 May;55(5):994-1001.

(21) Stichting Perinatale Registratie Nederland. Perinatale Zorg in Nederland 2008. 2011. Utrecht, Stichting Perinatale Registratie Nederland.
Ref Type: Serial (Book,Monograph)

(22) Wiberg N, Kallen K, Olofsson P. Base deficit estimation in umbilical cord blood is influenced by gestational age, choice of fetal fluid compartment, and algorithm for calculation. Am J Obstet Gynecol 2006 Dec;195(6):1651-6.

(23) Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006 Feb 7;144(3):201-9.

Chapter 2.4

A clinical prediction model to assess the risk of operative delivery.

Schuit E, Kwee A, Westerhuis ME, Van Dessel HJ, Graziosi GC, Van Lith JM, Nijhuis JG, Oei SG, Oosterbaan HP, Schuitemaker NWE, Wouters MGAJ, Visser GHA, Mol BWJ, Moons KGM, Groenwold RHH.

BJOG 2012 Jul;119(8):915-23.

Abstract

Objective To predict instrumental vaginal delivery (IVD) or cesarean section (CS) for suspected fetal distress (FD) or failure to progress (FTP).

Design secondary analysis of a randomised trial

Setting Three academic and six non-academic teaching hospitals in the Netherlands

Population 5,667 labouring women with a singleton term pregnancy in cephalic presentation

Methods We developed multinomial prediction models to assess the risk of operative delivery using both antepartum (model 1) and antepartum plus intrapartum characteristics (model 2). The models were validated by bootstrapping techniques and adjusted for overfitting. Predictive performance was assessed by calibration and discrimination (AUROC-area), and easy-to-use nomograms were developed.

Main outcome measures IVD or CS for FD or FTP with respect to a spontaneous vaginal delivery (reference)

Results 375 (6.6%) and 212 (3.6%) women had an IVD or CS due to FD, and 433 (7.6%) and 571 (10.1%) due to FTP, respectively. Predictors were age, parity, previous CS, diabetes, gestational age, gender, estimated birth weight (model 1) and induction of labour, oxytocin augmentation, intrapartum fever, prolonged rupture of membranes, meconium stained amniotic fluid, epidural anaesthesia, and use of ST-analysis (model 2). Both models showed excellent calibration and the ROC areas were 0.70-0.78 and 0.73-0.81 respectively.

Conclusion In Dutch women with a singleton term pregnancy in cephalic presentation, ante- and intrapartum characteristics can assist to predict the need for an instrumental vaginal or caesarean section due to foetal distress or failure to progress.

Introduction

In the 1970s caesarean delivery rates began to rise in most Western countries(1) and continue to rise in most of these countries.(2) In 2008 the United States caesarean section rate increased for the 12th consecutive year to a total of 32.3%.(2) On the other hand instrumental vaginal deliveries remained stable over the years.(2;3) Most caesarean sections as well as instrumental vaginal deliveries are performed because of suspected foetal distress or failure to progress. Furthermore, in absolute numbers by far most of the instrumental vaginal and caesarean deliveries are performed in women with a term pregnancy with a foetus in cephalic presentation.(4) Despite these numbers, it remains difficult to predict by which mode of delivery these women will actually deliver.

The mode of delivery often depends on several maternal characteristics, e.g. maternal age(5), parity(6), body mass index(7), maternal height(8), gestational age at delivery(9), foetal head position(10), amniotic fluid volume(11;12), ultrasonic estimated foetal weight(12) and cervical length(13), but also on problems that might arise during labour (e.g., meconium stained amniotic fluid). Several prediction models using these characteristics have been developed to predict the occurrence of caesarean section.(13-18) However, since foetal distress or failure to progress followed by either a caesarean section or instrumental vaginal delivery have different consequences for the neonatal and maternal outcomes (including future pregnancies), it would be helpful to predict either of these interventions and their indications: instrumental vaginal delivery due to suspected foetal distress (IVD-FD), caesarean section due to suspected foetal distress (CS-FD), instrumental vaginal delivery due to failure to progress (IVD-FTP) and caesarean section due to failure to progress (CS-FTP).

Such a rule – based on easily and readily available characteristics – could be helpful to clinicians as it would allow for timely prognostication which may lead to more effective decision making during labour. It could be an alert (e.g., on the fact that the child is likely to be suffering from foetal distress during labour), be used in the decision for primary CS, or it could aid in organizational aspects of the delivery (e.g. availability of doctors who can perform an instrumental delivery and availability of operating theatre and personnel). Finally, it allows for more individualised counselling of the pregnant woman.

Therefore, the aim of the present study was to identify which factors that can be timely obtained, either before or early during labour, independently contribute to predicting the risk of instrumental vaginal delivery and caesarean section with a suspected foetal distress or failure to progress indication. For this purpose we used data from a large multicenter trial in which labouring women with a high-risk vertex singleton pregnancy beyond 36 weeks of gestation were studied.(19)

Methods

Setting

We used data from a recently published randomised clinical trial conducted in The Netherlands. In this trial, labouring women with high-risk vertex singleton pregnancies beyond 36 weeks of gestation were randomly allocated to either intra partum monitoring by cardiotocography (CTG) plus ST-analysis of the foetal electrocardiogram (ECG; STAN) or CTG only. The performance of foetal blood sampling (FBS) was guided by a strict protocol. Both the design as well as the main results of the study are presented elsewhere.(19)

In The Netherlands, low risk pregnant women are monitored by midwives or general practitioners at home or in hospital (primary care), whereas high risk pregnant women are monitored by gynaecologists in hospital (secondary care). High-risk pregnancies included pregnancies that were complicated by hypertensive disorders, pre-existing maternal disease, ruptured membranes for more than 24 hours, complicated obstetric history, intrauterine growth restriction, a post date gestational age (≥ 42 weeks of gestation), need for pain relief, failure to progress, meconium stained amniotic fluid or non-reassuring foetal heart rate at intermittent auscultation by a midwife.(19)

Outcome

Based on the combination of the intervention (IVD or CS) and the indication for the intervention (FD or FTP) women were assigned to one of the five distinctive outcome categories: spontaneous vaginal delivery (reference category); instrumental vaginal delivery due to suspected foetal distress (IVD-FD); caesarean section due to suspected foetal distress (CS-FD); instrumental vaginal delivery due to failure to progress (IVD-FTP); or caesarean section due to failure to progress (CS-FTP). Hence, the outcome of this study was multinomial or polytomous (i.e. more than two unordered outcome categories). Instrumental vaginal delivery was defined as either vacuum or forcipal extraction or both. Suspected foetal distress was defined as the baby having a preterminal or rapidly deteriorating abnormal CTG-pattern, a pH below 7.20 obtained by foetal blood sampling or a significant ST-event. Failure to progress in the first stage was defined as an arrest of labour of at least two hours with adequate contractions. Failure to progress in the second stage was defined as having a period of active pushing of more than sixty minutes. As such, women in the second stage of labour with a FD as well as a FTP indication were qualified as having had an intervention with an indication FD or FTP if the duration of active pushing exceeded sixty minutes (FTP) or not (FD).

Candidate Predictors

Based on literature and clinical reasoning we selected candidate predictors for the above defined outcome categories.(13;20) Candidate predictors were categorized into antepartum and intrapartum variables. The antepartum variables included maternal age, parity, gestational

age, maternal diabetes mellitus, previous caesarean delivery, foetal gender, maternal hypertensive disorder, suspected intra uterine growth restriction and antepartum estimated foetal weight (EFW). Since the latter was not registered the actual birth weight – in 100 g increments – was used as a potential predictor in the development of the models. Maternal diabetes mellitus was defined as both pregestational type 1 and 2 as well as gestational diabetes mellitus. An antepartum prediction model (model 1) was developed using this first set of variables.

The second set of candidate predictors contained variables obtained early during labour i.e. induced onset of labour, oxytocin augmentation, intrapartum fever (≥ 37.8 °C), rupture of membranes > 24 hours, epidural anaesthesia and meconium stained amniotic fluid. These intrapartum predictors were added to model 1, to determine their added predictive value (model 2).

The allocated intervention of the original trial was taken into account by inclusion of this intervention variable in the multivariable analysis of models 1 and 2. Maternal and gestational age and birth weight were analyzed as continuous variables. Restricted cubic spline analyses were used to assess linearity of their association with the outcome.(21) Furthermore, several interactions were investigated i.e. epidural anaesthesia and induced onset of labour, epidural anaesthesia and oxytocin augmentation, and epidural anaesthesia and intrapartum fever.

Data analysis

Univariable associations between candidate predictors and the different outcome categories were estimated with multinomial logistic regression analysis. Multinomial logistic regression allows for simultaneous estimation of the probability of the different outcomes (IVD-FD, CS-FD, IVD-FTP, CS-FTP, and spontaneous delivery (the reference category)).(22-25) Essentially, the multinomial logistic regression model includes several logistic regression models simultaneously, to estimate the associations between the predictors and each of the outcomes compared to the reference category. Hence, estimated regression coefficients for the predictors may differ per outcome.(22-25)

Since selection based on univariable statistics might result in unstable prediction models, we choose not to perform any preselection and to include all candidate predictors in the multivariable analyses.(21;26) In the model including antepartum predictors only (model 1), as well as the model including both antepartum and intrapartum predictors (model 2), the final predictors were identified by a backward stepwise selection in the multinomial logistic regression model using Akaike's Information Criterion.(27)

Various women had missing values for some of the potential predictors. These values were to some extent selectively missing (as published in the main trial report; Appendix 3 of Westerhuis et al(19), available online at http://links.lww.com/AOG/A178). Hence, as widely acknowledged, a complete case analysis would yield biased results.(28-30) We thus used

multiple imputation (ten imputed datasets) following the original trial analyses.(19) Since imputed data sets differ from each other, predictors were selected in each imputation set separately. For inclusion in the final prediction models, we used the majority method; i.e. predictors were included if selected in at least five out of ten imputed data sets.(31) The regression coefficients and standard errors of these final predictors were combined from the ten data sets using Rubin's rules to come to the two final prediction models.(32)

The models were (internally) validated using bootstrapping techniques. One hundred bootstrap samples of equal size as the original data (N=5667) were drawn from the original dataset with replacement, allowing for multiple sampling of the same individual. Within each bootstrap sample the entire modelling process described above was repeated. This yielded a shrinkage factor for the regression coefficients to adjust these regression coefficient and thus the final model for optimism and overfitting.(21) The area under the Receiver Operating Characteristic curve (AUC) was studied to assess the ability of the two models to discriminate between women undergoing one of the interventions versus those undergoing a spontaneous vaginal delivery. Hence, we calculated four AUCs, each time relating one outcome versus the reference category. The predicted probabilities were compared with the observed frequencies of the different outcome categories using calibration plots to assess the calibration of the two models.(21;33)

Finally, to improve clinical application, nomograms were developed to easily calculate the probability of VD-FD, CS-FD, IVD-FTP, CS-FTP, and spontaneous delivery.

All analyses were performed in R version 2.10.0 (The R Foundation for Statistical Computing, 2009).

Results

Between January 2006 and July 2008, 5,667 women met the inclusion criteria of the RCT.(19) In these, IVD-FD occurred in 375 (6.6%) women, CS-FD in 212 (3.6%), IVD-FTP in 433 (7.6%), and CS-FTP in 571 (10.1%) and spontaneous delivery in 4,077 (71.9%) women. Characteristics of these women are presented in the second column of Table 1.

Antepartum predictors related to any of the four outcomes in univariable analysis were maternal age, gestational age, nulliparity, previous caesarean delivery, pre-eclampsia or pregnancy induced hypertension, intra uterine growth restriction, neonatal female gender and birth weight (Table 1). Intra-partum predictors included induced onset of labour, oxytocin augmentation, intrapartum fever, rupture of membranes > 24 hours, meconium stained amniotic fluid, epidural anaesthesia, and the performance of ST-analysis. None of the women with CS-FTP had an intra uterine growth restriction. As a result the effect of this variable on the outcome could not be estimated reliably.(34) Therefore, although intra uterine growth restriction showed to be related to IVD-FTP and CS-FD, the variable was not considered in the multivariable analyses.

Table 1 Characteristics of the study population and the univariable associations between potential predictors and the mode of delivery and its indications.

	Overall	Delivery					Odds ratio [95% CI]			
		Spontaneous	IVD-FD	CS-FD	IVD-FTP	CS-FTP	IVD-FD vs. spontaneous	CS-FD vs. spontaneous	IVD-FTP vs. spontaneous	CS-FTP vs. spontaneous
Characteristic	**N=5667**	**N=4077**	**N=375 (7)**	**N=212 (4)**	**N=433 (8)**	**N=571 (10)**				
Antepartum										
Maternal age, y	32.0±4.8	31.9±4.8	31.8±4.7	32.6±4.8	32.1±4.2	32.5±5.0	1.00 [0.97, 1.02]	1.03 [1.00, 1.06]*	1.01 [0.99, 1.03]	1.02 [1.00, 1.04]
Gestational age, wks	40.2±1.4	40.1±1.5	40.4±1.4	40.4±1.4	40.4±1.4	40.6±1.3	1.18 [1.09, 1.28]*	1.20 [1.08, 1.34]*	1.21 [1.12, 1.30]*	1.35 [1.26, 1.45]*
Nulliparous	3236 (57)	1990 (49)	291 (78)	141 (66)	374 (87)	440 (77)	3.64 [2.84, 4.68]*	2.08 [1.53, 2.83]*	6.73 [5.07, 8.94]*	3.53 [2.87, 4.35]*
Previous caesarean delivery†	716 (13)	491 (12)	54 (14)	33 (16)	44 (10)	94 (16)	1.23 [0.91, 1.67]	1.35 [0.90, 2.03]	0.83 [0.60, 1.15]	1.44 [1.12, 1.84]*
Maternal diabetes mellitus	169 (3)	120 (3)	8 (2)	9 (4)	7 (2)	25 (4)	0.72 [0.35, 1.48]	1.51 [0.75, 3.04]	0.54 [0.25, 1.17]	1.49 [0.95, 2.33]
Pre-eclampsia or pregnancy induced hypertension	678 (12)	483 (12)	49 (13)	35 (17)	46 (11)	65 (11)	1.12 [0.82, 1.54]	1.49 [1.02, 2.17]*	0.89 [0.64, 1.22]	0.96 [0.73, 1.26]
Intra uterine growth restriction	133 (2)	106 (3)	11 (3)	15 (7)	1 (0)	0 (0)	1.13 [0.60, 2.13]	2.85 [1.62, 5.01]*	0.09 [0.01, 0.62]*	NA
Neonatal female gender	2668 (47)	1977 (48)	175 (47)	84 (40)	176 (41)	256 (45)	0.93 [0.75, 1.15]	0.70 [0.52, 0.94]*	0.73 [0.60, 0.89]*	0.86 [0.72, 1.03
Birth weight, 100 g	35.4±5.2	35.1±5.1	34.2±5.2	34.2±5.8	36.8±4.5	38.0±5.1	0.96 [0.94, 0.98)*	0.96 [0.94, 0.99]*	1.06 [1.04, 1.09]*	1.11 [1.09, 1.13]*
Intrapartum										
Induced onset of labour	2341 (41)	1714 (42)	126 (34)	105 (49)	147 (34)	249 (44)	0.70 [0.56, 0.87]*	1.35 [1.02, 1.79]*	0.71 [0.58, 0.88]*	1.07 [0.89, 1.27]
Oxytocin augmentation	2044 (36)	1294 (32)	163 (43)	78 (37)	234 (54)	275 (48)	1.65 [1.33, 2.05]*	1.26 [0.94, 1.69]	2.54 [2.08, 3.10]*	2.00 [1.67, 2.39]*
Intrapartum fever >= 37.8	470 (8)	236 (6)	45 (12)	18 (8)	78 (18)	93 (16)	2.22 [1.58, 3.12]*	1.47 [0.85, 2.55]	3.58 [2.71, 4.73]*	3.18 [2.45, 4.13]*
Rupture of membranes > 24 hours	692 (12)	502 (12)	38 (10)	15 (7)	72 (17)	65 (11)	0.80 [0.57, 1.14]	0.53 [0.29, 0.95]*	1.42 [1.09, 1.86]*	0.92 [0.70, 1.21]
Meconium stained amniotic fluid	1471 (26)	990 (24)	114 (30)	67 (32)	115 (27)	185 (32)	1.36 [1.08, 1.72]*	1.44 [1.06, 1.95]*	1.13 [0.90, 1.41]	1.49 [1.23, 1.81]*
Epidural anaesthesia	2389 (42)	1438 (35)	190 (51)	122 (57)	243 (56)	397 (69)	1.89 [1.52, 2.33]*	2.46 [1.82, 3.33]*	2.36 [1.93, 2.88]*	4.18 [3.44, 5.07]*
Use of ST-analysis	2827 (50)	2052 (50)	185 (49)	114 (54)	193 (45)	283 (50)	0.96 [0.78, 1.19]	1.15 [0.86, 1.52]	0.80 [0.65, 0.97]*	0.97 [0.81, 1.16]

Data are presented as mean ± standard deviation or n (%).

* significant

† when only multiparae were selected this resulted in an OR (95% CI) 5.91 (3.73, 9.37), OR (95% CI) 2.82 (1.72, 4.64), OR (95% CI) 10.0 (5.44, 18.4) and OR (95% CI) 8.28 (5.52, 12.4) for IVD-FD, CS-FD, IVD-FTP and CS-FTP, respectively.

In model 1, seven antepartum variables were identified to predict one of the four outcomes: maternal age, nulliparity, previous caesarean delivery, maternal diabetes mellitus, gestational age, neonatal female gender and birth weight (Table 2). The model's AUC for IVD-FD was 0.72 [95% CI 0.69, 74], CS-FD was 0.70 [95% CI 0.66, 0.73], IVD-FTP was 0.78 [95% CI 0.76, 0.80] and CS-FTP was 0.78 [95% CI 0.76, 0.80].

Addition of intrapartum characteristics including interaction terms to model 1, yielded that induced onset of labour, oxytocin augmentation, intrapartum fever, rupture of membranes, meconium stained amniotic fluid, epidural anaesthesia, use of ST-analysis and an interaction of epidural anaesthesia and oxytocin augmentation were additional predictors of one of the four outcomes (Table 3). The model's AUC was slightly higher as compared to model 1: for IVD-FD 0.73 [95% CI 0.70, 0.75], for CS-FD 0.73 [95% CI 0.70, 0.76], for IVD-FTP 0.80 [95% CI 0.78, 0.82] and for CS-FTP 0.81 [95% CI 0.80, 0.83]). The higher discriminative ability of model 2 is reflected by a larger difference between highest and lowest predicted probabilities for model 2 (Figure 1 and 2). Both models showed excellent agreement between predicted probabilities and observed proportions for all four outcomes (Figure 1 and 2).

The nomograms of both models with an illustrative example are presented in Appendix S1A and S1B, respectively.

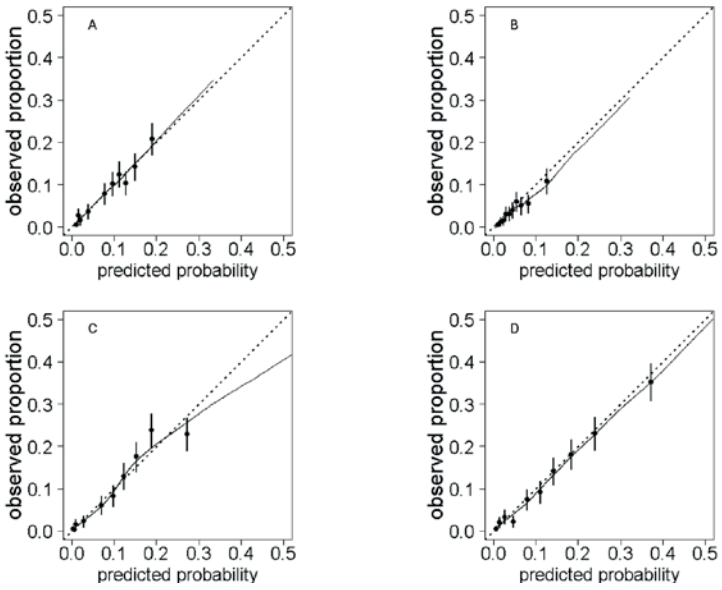


Figure 1 Calibration plots of model 1 with the observed risk of IVD-FD (A), CS-FD (B), IVD-FTP (C) and CS-FTP (D) by predicted probabilities of the IVD-FD, CS-FD, IVD-FTP and CS-FTP. The dots indicate deciles of women grouped by similar predicted risk of the different interventions and their indications. The vertical bars through the dots indicate the 95% confidence interval of the observed risks for the grouped women. To enhance interpretation, the axes were adjusted to a scale from 0.0 to 0.50, based on the low observed and predicted outcome incidences.

Table 2 Multivariable associations for model 1 (antepartum characteristics only).

	IVD-FD vs. spontaneous		CS-FD vs. spontaneous		IVD-FTP vs. spontaneous		CS-FTP vs. spontaneous	
Characteristic	**Beta***	**OR [95% CI]***	**Beta***	**OR [95% CI]***	**Beta***	**OR [95% CI]***	**Beta***	**OR [95% CI]***
Intercept	-13.1		-15.6		-11.1		-15.4	
Antepartum								
Maternal age, y	0.029	1.03 [1.01, 1.05]	0.052	1.05 [1.02, 1.09]	0.054	1.06 [1.03, 1.08]	0.056	1.06 [1.04, 1.08]
Gestational age, wks	0.26	1.29 [1.18, 1.41]	0.32	1.38 [1.22, 1.56]	0.038	1.04 [0.95, 1.13]	0.13	1.14 [1.05, 1.24]
Nulliparous	2.05	7.79 [5.26, 11.5]	1.13	3.09 [2.09, 4.55]	3.39	29.7 [17.2, 51.1]	2.65	14.1 [9.78, 20.3]
Previous caesarean delivery	1.77	5.87 [3.70, 9.32]	1.06	2.88 [1.74, 4.76]	2.39	10.9 [5.92, 20.1]	2.23	9.34 [6.17, 1.41]
Neonatal female gender	-0.19	0.83 [0.67, 1.03]	-0.50	0.61 [0.45, 0.83]	-0.25	0.78 [0.63, 0.96]	-0.013	0.99 [0.81, 1.20]
Birth weight, 100 g	-0.059	0.94 [0.92, 0.97]	-0.079	0.92 [0.89, 0.96]	0.083	1.09 [1.06, 1.11]	0.12	1.12 [1.10, 1.15]
Maternal diabetes mellitus	0.32	1.37 [0.65, 2.91]	0.99	2.69 [1.29, 5.60]	-0.24	0.79 [0.35, 1.76]	0.87	2.38 [1.44, 3.95]

* shrunken; shrinkage factor 0.99-1.02

Table 3 Multivariable associations for model 2 (ante- and intrapartum characteristics).

	IVD-FD vs. spontaneous		CS-FD vs. spontaneous		IVD-FTP vs. spontaneous		CS-FTP vs. spontaneous	
Characteristic	**Beta***	**OR [95% CI]***	**Beta***	**OR [95% CI]***	**Beta***	**OR [95% CI]***	**Beta***	**OR [95% CI]***
Intercept	-12.6		-11.6		-13. 6		-15.2	
Antepartum								
Maternal age, y	0.030	1.03 [1.01, 1.05]	0.053	1.05 [1.03, 1.08]	0.052	1.05 [1.02, 1.09]	0.057	1.06 [1.04, 1.08]
Gestational age, wks	0.25	1.28 [1.17, 1.41]	0.048	1.05 [0.96, 1.15]	0.26	1.29 [1.14, 1.46]	0.10	1.11 [1.02, 1.21]
Nulliparous	1.90	6.69 [4.47, 10.0]	3.08	21.7 [12.5, 37.8]	0.94	2.57 [1.70, 3.89]	2.28	9.78 [6.67, 14.3]
Previous caesarean delivery	1.63	5.10 [3.18, 8.17]	2.32	10.1 [5.46, 18.8]	1.03	2.81 [1.65, 4.79]	2.19	8.91 [5.78, 13.7]
Neonatal female gender	-0.19	0.83 [0.66, 1.03]	-0.26	0.77 [0.62, 0.96]	-0.48	0.62 [0.46, 0.84]	-0.025	0.98 [0.80, 1.19]
Birth weight, 100 g	-0.061	0.94 [0.92, 0.97]	0.077	1.08 [1.05, 1.11]	-0.077	0.93 [0.90, 0.96]	0.11	1.12 [1.10, 1.15]
Maternal diabetes mellitus	0.35	1.41 [0.66, 3.01]	-0.12	0.89 [0.39, 2.00]	0.80	2.23 [1.06, 4.70]	0.92	2.51 [1.49, 4.22]
Intrapartum								
Induced onset of labour	-0.16	0.85 [0.62, 1.16]	0.31	1.36 (0.95, 1.95]	0.72	2.06 [1.31, 3.25]	0.95	2.59 [1.81, 3.70]
Oxytocin augmentation	0.047	1.05 [0.72, 1.53]	1.10	3.00 [2.04, 4.41]	0.44	1.56 [0.84, 2.89]	0.95	2.58 [1.67, 3.98]
Intrapartum fever >= 37.8	0.40	1.49 [1.04, 2.14]	0.64	1.89 [1.40, 2.56]	-0.064	0.94 [0.53, 1.65]	0.33	1.39 [1.04, 1.85]
Ruptured membranes > 24h	-0.24	0.78 [0.54, 1.13]	0.23	1.25 [0.93, 1.69]	-0.72	0.49 [0.26, 0.89]	-0.084	0.92 [0.67, 1.25]
Epidural anaesthesia	0.21	1.24 [0.91, 1.70]	0.49	1.63 [1.19, 2.25]	0.70	2.01 [1.35, 2.99]	0.88	2.41 [1.81, 3.20]
Use of ST-analysis	-0.023	0.98 [0.79, 1.21]	-0.28	0.76 [0.61, 0.93]	0.15	1.16 [0.87, 1.55]	-0.039	0.96 [0.79, 1.16]
Interaction terms								
Epidural anaesthesia*Oxytocin augmentation	-0.063	0.94 [0.59, 1.50]	-0.74	0.48 [0.31, 0.73]	-0.11	0.89 [0.46, 1.74]	-0.077	0.93 [0.60, 1.42]

* shrunken; shrinkage factor 0.98-1.02

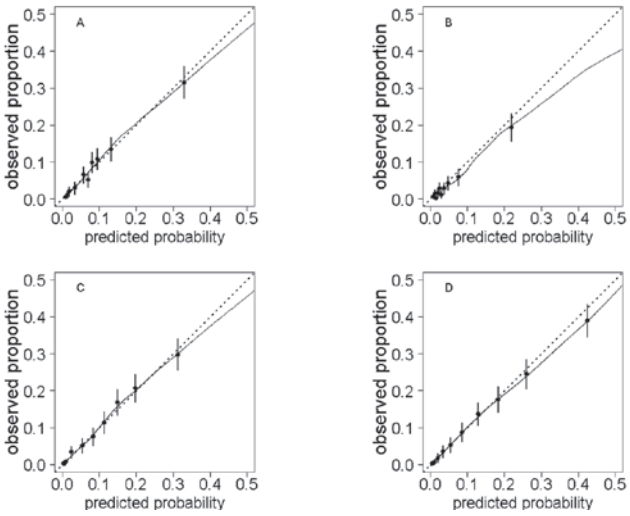


Figure 2 Calibration plots of model 2 with the observed risk of IVD-FD (A), CS-FD (B), IVD-FTP (C) and CS-FTP (D) by predicted probabilities of the IVD-FD, CS-FD, IVD-FTP and CS-FTP. The dots indicate deciles of women grouped by similar predicted risk of the different interventions and their indications. The vertical bars through the dots indicate the 95% confidence interval of the observed risks for the grouped women. To enhance interpretation, the axes were adjusted to a scale from 0.0 to 0.50, based on the low observed and predicted outcome incidences.

Discussion and Conclusion

In this study we developed models to simultaneously predict the probability of instrumental vaginal or caesarean delivery due to suspected foetal distress, or failure to progress, and spontaneous delivery based on antepartum characteristics only (model 1) and combining both ante- and intrapartum characteristics (model 2). Both models showed excellent calibration and good ability to discriminate between women undergoing the different interventions and those that had a spontaneous vaginal delivery. Model 2 showed a slightly better discriminative performance than model 1. Since this model included eight additional intrapartum characteristics compared to the antepartum model this difference was to be expected.(21)

All characteristics used in the models are readily available and easy to measure. Hence, the model will be easy and inexpensive to apply in daily clinical practice. Moreover, the developed nomograms further improve the clinical applicability of the models. The discriminative ability of the model showed only minimal improvement after the addition of intrapartum characteristics while the complexity of the model increased. The advantage of the second model is that it includes characteristics known to be associated with the different outcomes and therefore improves the face validity of the model. This is important in view of the future

application of the models in clinical practice.(35) Obviously, using the second model timely prognostication is impossible since this model includes characteristics that will only be available during labour. The first model, however, does allow for timely prognostication, since it includes antepartum characteristics only. It is important to emphasise, however, that the prognostic models may only complement clinical decision making by combining risk factors in the assessment of the chance of the outcome in an objective, more formal way rather than replace clinical judgment.

The study population contained large numbers of the outcome categories and therefore allowed for reliable estimation of the predictor effects. Consequently, the optimism found for the two models was small (Table 2 and 3). Obviously, the caesarean section rate in the Netherlands is lower than in some other developed countries.(2;3) Applying the models in these countries could result in an overall underestimation of the caesarean section proportion. Basic adjustment of the intercept could solve this problem.(36) However, when the difference in caesarean delivery rate is explained by a different attitude to a subgroup of women (e.g. women with diabetes mellitus) basic adjustment will not suffice and updating of (part of) the regression coefficients is needed. Obviously external validation will be needed to determine the performance of both models in other populations.

We used data of a well described, large and nationwide cohort of women from a randomised clinical trial in The Netherlands.(19) Since it was a randomised trial data collection was standardised resulting in high quality. The cohort included data on labouring women with a singleton term pregnancy in cephalic presentation. Since most of the labouring women in the general population are women with a singleton, term pregnancy with the child in a vertex presentation(9) we expect the models to generalise to the general population.

Our study aimed to predict both instrumental vaginal delivery and caesarean section using a single model. This is different from other studies that focused mainly on caesarean deliveries.(13-18) Furthermore, most of these studies were aimed at predicting only one outcome while a multinomial regression model has the advantage that it can estimate several outcomes simultaneously, e.g. vaginal delivery, elective caesarean section and caesarean section during labor,(17) and is therefore more informative. A variable is selected as a predictor in the model if it shows to be predictive for one of the outcomes. As a result the predictors are not necessarily predictive for all of the different endpoints, i.e. neonatal female gender in CS-FTP (Table 1-3). Another advantage is that the probability of combined outcomes can be calculated directly. For example, the probability of having an IVD-FD for a nulliparous, non-diabetic, woman of 25 years old that is in the 39th week of her pregnancy of a girl with an estimated weight of 2500 g is 13% (See Appendix S1A for a more extensive description on how to calculate this probability). For the same woman we predict a probability of 4%, 5% and 5% for CS-FD, IVD-FTP, and CS-FTP, respectively. Consequently, the probability of a spontaneous vaginal delivery is 100 – p(IVD-FD) – p(CS-FD) – p(IVD-FTP) – p(CS-FTP) = 100 –

13 – 4 – 5 – 5 = 73%, whereas the risk of a caesarean section is p(CS-FD) + *P* (CS-FTP).4 + 5 = 9%. Furthermore, in a multinomial model more women are used to fit the final model than in four separate logistic regression models where the same women are used as a reference several times.

There are some limitations. Potential predictors like body mass index(7;14;17), maternal height(13-15;18), cervical length(13), and amniotic fluid volume(11;12) were not taken into account since information on these characteristics was unknown. Although the latter are easy to measure, cervical length changes during pregnancy which makes it difficult to include in the model. Also amniotic fluid volume is often compromised by large interobserver variability. It is important to note that it is difficult to distinguish between suspected foetal distress and failure to progress in daily clinical practice since they can occur simultaneously. Nevertheless, this is of less importance since interest lies in the fact if a woman will end up having a problematic delivery or not. Consequently, emphasis may be more on the discriminative ability of the models than on the accuracy of the predictions.

The study population is a mix of two obstetric populations: women who were referred to secondary care prior to labour, and women who were referred from primary care to secondary care during labour. Since the latter group had a lower risk for an intervention prior to labour (i.e. no reason for a start of labour in secondary care) they could potentially dilute the effect of some of the variables in the antepartum model. To investigate to what extent this mixing influenced model 1 we performed an additional analysis in which we developed the model in women that had risk factors that would lead to a start of labour in secondary care only. The analysis showed results similar to the ones presented in the results section (results not shown).

The ultrasound-to-delivery interval(37) and the large intra- and interobserver variability(38) can comprise the accuracy of EFW used in the model, which might lead to an incorrect estimation of the probability of an intervention. Despite this limitation, it remains appropriate for health care providers and patients to consider past and predicted birth weights when making decisions regarding the probability of the interventions, but birth weight alone should not preclude the possibility of an intervention. Furthermore, misestimating the actual birth weight of for example 2500 g with 100 g leads to an erroneous decrease of the probability of IVD-FD of 0.9% only, indicating that a wrong estimation will have little impact on the predicted probabilities.

We are aware that caesarean section rates, instrumental vaginal delivery rates, labour induction and the use and dosage of oxytocin augmentation differ from centre to centre and from provider to provider, e.g. by age, experience and years out of training. An obvious way to include these differences would be to account for centres in a multilevel regression model. However, we are unaware of any statistical method that is able to combine multilevel and multinomial modelling. Alternatively, centre effects were investigated using a logistic

multilevel model of model 2 whereby we fitted a random intercept per centre with CD-FD as the outcome and spontaneous vaginal delivery as the reference. This analysis showed that the random intercepts of the different centres were significantly different from zero for the majority of centres. To account for these between centre differences, we've included the centres in the model as dummy variables (centre 1 is the reference). The analysis showed that, in line with the logistic multilevel analysis, several centres were significant predictors of the mode of delivery. When comparing the performance of this model (with centres) with our presented model (without centres) we found that both calibration and discrimination were slightly better for the former model with AUCs of 0.75, 0.75, 0.82 and 0.82 for IVD-FD, CD-FD, IVD-FTP and CD-FTP compared to 0.73, 0.73, 0.80 and 0.81, respectively. However, the increase in performance is only marginally improved given the increased complexity of the model and the loss of generalisability. Therefore we decided to only present the results of the models without adjustment for centres.

In summary, in women with a singleton term pregnancy in cephalic presentation, both ante- and intrapartum characteristics influence the probability of an instrumental vaginal or caesarean section due to suspected foetal distress or failure to progress. Information on the risk of an instrumental vaginal or caesarean section due to suspected foetal distress or failure to progress can be of great value in counselling women and guiding labour management. It may allow clinicians to avoid unnecessary interventions in low-risk women and may influence decisions during labour regarding the interpretation of foetal heart rate patterns and the application of additional techniques for foetal monitoring, such as ST-analysis of the foetal ECG or FBS. The nomograms will allow for fast and easy implementation in clinical practice. After external validation and proof of generalisability, these models could be used in obstetric clinical management.

References

(1) Notzon FC, Placek PJ, Taffel SM. Comparisons of national cesarean-section rates. N Engl J Med 1987 Feb 12;316(7):386-9.

(2) Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Mathews TJ, Osterman MJ. Births: final data for 2008. Natl Vital Stat Rep 2010 Dec 8;59(1):1, 3-1,71.

(3) The Netherlands Perinatal Registry. Perinatal care in the Netherlands, 2006. The Netherlands Perinatal Registry; 2008.

(4) Kwee A, Elferink-Stinkens PM, Reuwer PJ, Bruinse HW. Trends in obstetric interventions in the Dutch obstetrical care system in the period 1993-2002. Eur J Obstet Gynecol Reprod Biol 2007 May;132(1):70-5.

(5) Kozinszky Z, Orvos H, Zoboki T, Katona M, Wayda K, Pal A, et al. Risk factors for cesarean section of primiparous women aged over 35 years. Acta Obstet Gynecol Scand 2002 Apr;81(4):313-6.

(6) Rane SM, Guirgis RR, Higgins B, Nicolaides KH. The value of ultrasound in the prediction of successful induction of labor. Ultrasound Obstet Gynecol 2004 Oct;24(5):538-49.

(7) Usha Kiran TS, Hemmadi S, Bethel J, Evans J. Outcome of pregnancy in a woman with an increased body mass index. BJOG 2005 Jun;112(6):768-72.

(8) Crane JM, Delaney T, Butt KD, Bennett KA, Hutchens D, Young DC. Predictors of successful labor induction with oral or vaginal misoprostol. J Matern Fetal Neonatal Med 2004 May;15(5):319-23.

(9) Royal College of Obstetricians and Gynaecologists Clinical Effectiveness Support Unit. National Sentinel Caesarean Section Audit Report. London (UK): RCOG Press; 2001.

(10) Fitzpatrick M, McQuillan K, O'Herlihy C. Influence of persistent occiput posterior position on delivery outcome. Obstet Gynecol 2001 Dec;98(6):1027-31.

(11) Morris JM, Thompson K, Smithey J, Gaffney G, Cooke I, Chamberlain P, et al. The usefulness of ultrasound assessment of amniotic fluid in predicting adverse outcome in prolonged pregnancy: a prospective blinded observational study. BJOG 2003 Nov;110(11):989-94.

(12) Saunders N, Amis S, Marsh M. The prognostic value of fetal ultrasonography before induction of labour. Br J Obstet Gynaecol 1992 Nov;99(11):903-6.

(13) Peregrine E, O'Brien P, Omar R, Jauniaux E. Clinical and ultrasound parameters to predict the risk of cesarean delivery after induction of labor. Obstet Gynecol 2006 Feb;107(2 Pt 1):227-33.

(14) Chen G, Uryasev S, Young TK. On prediction of the cesarean delivery risk in a large private practice. Am J Obstet Gynecol 2004 Aug;191(2):616-24.

(15) Harper DM, Johnson CA, Harper WH, Liese BS. Prenatal predictors of cesarean section due to labor arrest. Arch Gynecol Obstet 1995;256(2):67-74.

(16) Hueston WJ. Development of a cesarean delivery risk score. Obstet Gynecol 1994 Dec;84(6):965-8.

(17) Roman H, Blondel B, Breart G, Goffinet F. Do risk factors for elective cesarean section differ from those of cesarean section during labor in low risk pregnancies? J Perinat Med 2008;36(4):297-305.

(18) Seshadri L, Mukherjee B. A predictive model for cesarean section in low risk pregnancies. Int J Gynaecol Obstet 2005 May;89(2):94-8.

(19) Westerhuis ME, Visser GH, Moons KG, van BE, Benders MJ, Bijvoet SM, et al. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2010 Jun;115(6):1173-80.

(20) Gopalani S, Bennett K, Critchlow C. Factors predictive of failed operative vaginal delivery. Am J Obstet Gynecol 2004 Sep;191(3):896-902.

(21) Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-87.

(22) Begg CB, Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika 1984;71:11-8.

(23) Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KG. Polytomous logistic regression analysis could be applied more often in diagnostic research. J Clin Epidemiol 2008 Feb;61(2):125-34.

(24) Roukema J, van Loenhout RB, Steyerberg EW, Moons KG, Bleeker SE, Moll HA. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. J Clin Epidemiol 2008 Feb;61(2):135-41.

(25) Wijesinha A, Begg CB, Funkenstein HH, McNeil BJ. Methodology for the differential diagnosis of a complex data set. A case study using data from routine CT scan examinations. Med Decis Making 1983;3(2):133-54.

(26) Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996 Aug;49(8):907-16.

(27) Atkinson A. A note on the generalized information criterion for choice of a model. Biometrika 2010;67:413-8.

(28) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-91.

(29) Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995 Dec 15;142(12):1255-64.

(30) van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006 Oct;59(10):1102-9.

(31) Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol 2010 Feb;63(2):205-14.

(32) Rubin DB. Multiple imputation for non response in surveys. New York: Wiley; 1987.

(33) Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991 Aug;10(8):1213-26.

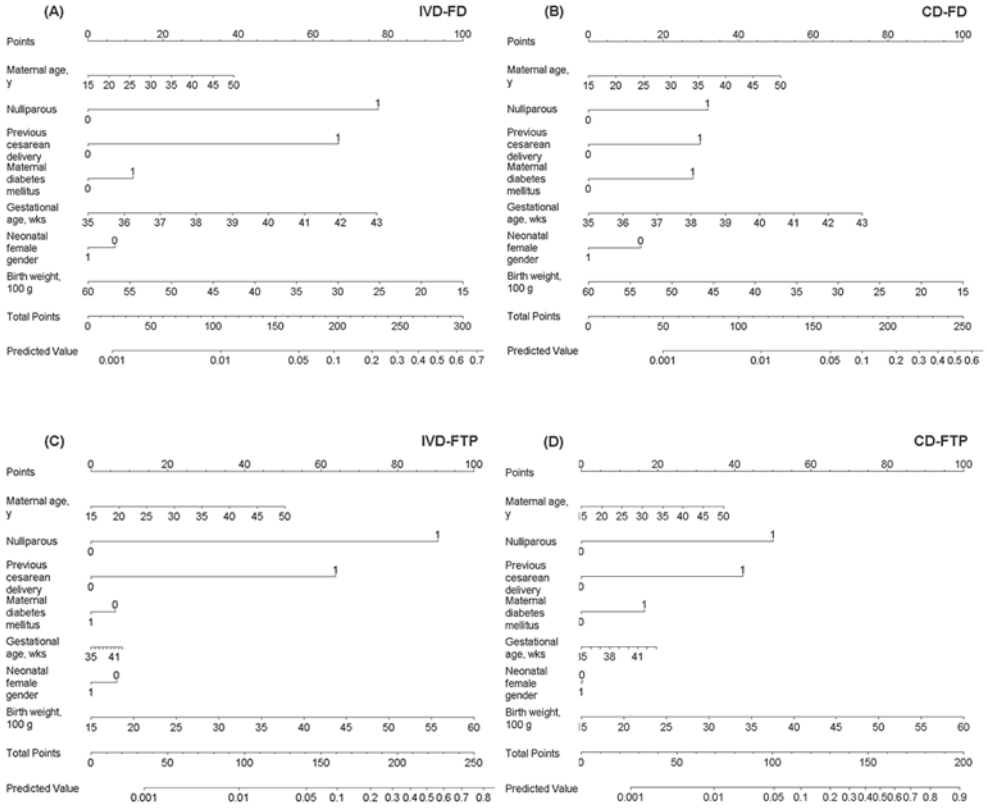
(34) Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer; 2009.

(35) Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

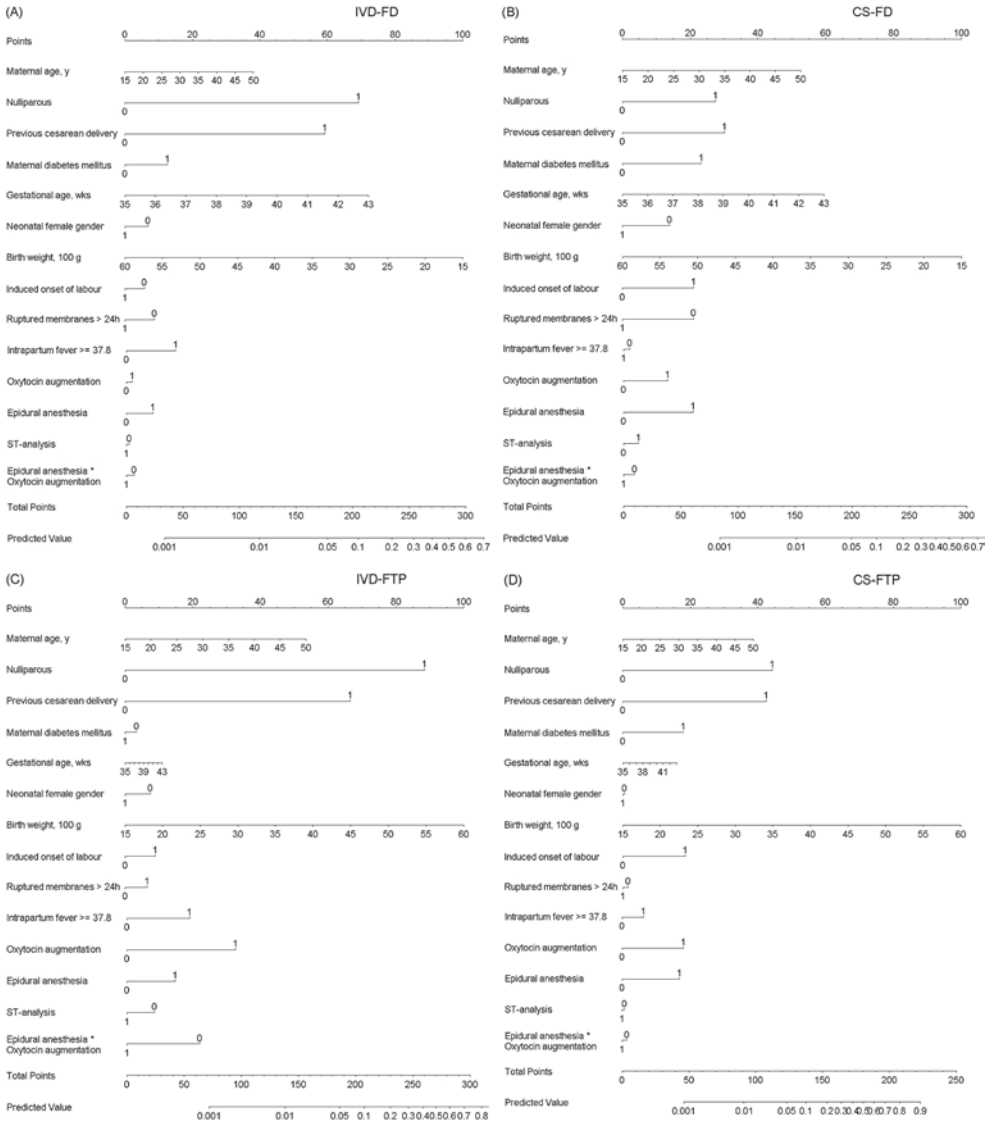
(36) Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. Can J Anaesth 2009 Mar;56(3):194-201.

(37) Cohen JM, Hutcheon JA, Kramer MS, Joseph KS, Abenhaim H, Platt RW. Influence of ultrasound-to-delivery interval and maternal-fetal characteristics on validity of estimated fetal weight. Ultrasound Obstet Gynecol 2010 Apr;35(4):434-41.

(38) Dudley NJ. A systematic review of the ultrasound estimation of fetal weight. Ultrasound Obstet Gynecol 2005 Jan;25(1):80-9.



Appendix S1A Nomograms of the multinomial model using antepartum characteristics for instrumental vaginal delivery with indication of foetal distress (A), caesarean section with indication of foetal distress (B), instrumental vaginal delivery with indication of failure to progress (C), and caesarean section with failure to progress indication (D). Continuous characteristics are presented in the nomogram on a continuous scale. Dichotomous characteristics are presented as a 0 or 1, indicating not having (0) or having (1) the characteristic. To calculate the probability of IVD-FD, CS-FD, IVD-FTP and CS-FTP one can simply fill in the characteristics of the woman at risk. For example, consider a woman of 25 years old (11 points), that is nulliparous (and consequently did not have a previous caesarean delivery, 77 points) and has no diabetes mellitus (0 points), that is in the 39th week of her pregnancy (38 points), is pregnant of a girl (0 points) with an estimated weight of 2500 g (78 points). Adding up the points of the individual predictors results in a score of 204 points, corresponding to a predicted value of 0.13, e.g. this woman has a chance of 13% of having an instrumental vaginal delivery with fetal distress indication. Using the three other nomograms we predict a probability of 0.04, 0.05 and 0.05 for CS-FD, IVD-FTP, and CS-FTP, respectively. To calculate the probability of a spontaneous vaginal delivery one can add up all the predicted probabilities for the different outcomes and subtract this from 1. Hence the probability of a spontaneous vaginal delivery is 1 – p(IVD-FD) – p(CS-FD) – p(IVD-FTP) – p(CS-FTP), whereas the risk of a caesarean section is p(CS-FD) + *P* (CS-FTP) and the risk of an intervention due to failure to progress is p(IVD-FTP) + p(CS-FTP), which are 0.73, 0.09 and 0.10, respectively.



Appendix S1B Nomograms of the multinomial model using antepartum plus intrapartum characteristics for instrumental vaginal delivery with indication of foetal distress (A), caesarean section with indication of foetal distress (B), instrumental vaginal delivery with indication of failure to progress (C), and caesarean section with failure to progress indication (D). Continuous characteristics are presented in the nomogram on a continuous scale. Dichotomous characteristics are presented as a 0 or 1, indicating not having (0) or having (1) the characteristic. An example of how to use a nomogram is described in Appendix S1A.

Chapter 2.5

Prognostic models for stillbirth and neonatal death in very preterm birth: a validation study.

Schuit E, Hukkelhoven CW, Manktelow BN, Papatsonis DN, de Kleine MJ, Draper ES, Steyerberg EW, Vergouwe Y.

Pediatrics 2012 Jan;129(1):e120-e127.

Abstract

Objective To externally validate two prognostic models for stillbirth and neonatal death in very preterm infants that are either known to be alive at the onset of labour or admitted for neonatal intensive care.

Patients and Methods All infants, with gestational age 22- 32 weeks and European ethnicity, known to be alive at the onset of labour (n = 17,582) and admitted for neonatal intensive care (n = 11,578), born in the Netherlands between 1 January 2000 and 31 December 2007. The main outcome measures were stillbirth or death within 28 days for infants known to be alive at the onset of labour and death before discharge from the neonatal intensive care unit for infants admitted for neonatal intensive care. Model performance was studied with calibration plots and c statistic.

Results 16.7% (n=2,939) of the infants known to be alive at the onset of labour died during labour or within 28 days of birth, and 7.8% (n=908) of the infants admitted for neonatal intensive care died before discharge from the intensive care. The prognostic model for infants known to be alive at the onset of labour showed good calibration and excellent discrimination (c statistic 0.92). The prognostic model for infants admitted for neonatal intensive care showed good calibration and good discrimination (c statistic 0.82).

Conclusions Our study showed good performance of two prognostic models for stillbirth and neonatal death in very preterm Dutch infants. The results imply usage in clinical practice in the Netherlands and possibly other Western countries.

Introduction

In most developed countries very preterm birth has an incidence of 1-2%(1) but accounts for approximately one third to half of all perinatal deaths.(2;3) Accurate risk assessment of perinatal death can help caregivers and parents to decide if and when to intervene in a pregnancy or to adjust postnatal care.

Two easy to apply survival graphs are used in clinical practice for discussions with parents for whom premature delivery is likely.(2;4;5) These graphs are derived from prognostic models that predict survival for very preterm infants known to be alive at the onset of labour and very preterm infants admitted for neonatal intensive care. Risk factors in the models include readily available variables such as gestational age, birth weight, sex, ethnicity and multiplicity of pregnancy.(4) The graphs were first published in 1999 based on births between 1994 and 1997 for the former Trent UK health region. An update of the graphs was published based on births between 1998 and 2001.(5) Although developed for a UK population, the models may also be useful in other Western countries, both for estimating the effect of new interventions, for developing treatment protocols and for counseling individual parents, as widely used for in the UK.(2) Since local and timely variation in attitude to the care of very preterm infants can influence the validity of the models, we studied the external validity in infants born in the first decade of the 21st century in the Netherlands.

Patients and Methods

Study population

We studied The Netherlands, which in 2010 comprised more than 16 million people and about 180,000 births a year. Similar to the study by Draper et al.(4) we included all infants alive at the onset of labour and born between 1 January 2000 and 31 December 2007 with European ethnicity and gestational age between 22 and 32 weeks to validate the model that predicts fetal death or neonatal death within 28 days. A subgroup of these infants, including only the infants that were admitted to a NICU, was used to validate the model that predicts death before discharge from the NICU. Within the Dutch study population we made a distinction between congenital malformations that were definitely not lethal, probably lethal, and definitely lethal. Similar to the study by Draper et al.(4) infants with (probably) lethal congenital malformations were excluded from the study.

Data on births were extracted from the Netherlands Perinatal Registry, which is a linked database of medical registries from three professional organizations: the national midwives registry, the national obstetricians registry and the national paediatricians/ neonatologists registry. The database contains 95% of all women who gave birth and 100% of the newborns that were admitted to a neonatal intensive care unit (NICU) after birth, resulting in high

quality data on all births.(6) We used data on the maternal age at childbirth, multiplicity of the pregnancy, the infants' sex, gestational age as calculated from ultrasound measures (93%) or from the date of last menstrual period (7%), birth weight as determined by weighing, 5-minute Apgar score as reported by the obstetrician and whether the infant was admitted for neonatal intensive care.

Outcomes

Death of infants known to be alive at the onset of labour was defined as fetal death or neonatal death within 28 days of birth. Death of infants admitted to a NICU was defined as death before discharge.

Prognostic models

We evaluated the ability of the updated logistic regression models(5) (Appendix) developed on infants from the former Trent region (development population) to predict (a) fetal death or neonatal death within 28 days of birth and (b) neonatal death before discharge from a NICU in European infants in the Dutch population (validation population). The model for infants known to be alive at the onset of labour included multiplicity of pregnancy, gestational age and birth weight. The model for infants admitted to a NICU included fetal gender as additional variable. Birth weight was included in both models as the difference between the observed birth weight and the mean birth weight of infants with the same gestational age and ethnicity (Appendix).

Statistical analysis

Missing values were imputed once with single imputation,(7) since only a very small percentage of data was missing (0.3%). Baseline characteristics of the validation and development population were analysed using descriptive statistics and presented as mean with a standard deviation for continuous variables and as numbers and percentages of the whole population for categorical and dichotomous variables.

The updated models described in the Appendix were applied to the Dutch validation population. The validity of the prognostic models was assessed in terms of calibration and discrimination. Calibration refers to agreement between the predicted risk of death and the observed proportion of infants who died in the validation population over the whole possible range of predicted risks. Calibration was assessed graphically with a calibration plot(8;9).The calibration line was described with the calibration slope (b) and with an intercept (a), given that the calibration slope is set to 1 (a|b = 1, calibration-in-the-large), as proposed by Cox. (8) Discrimination refers to the ability of the model to distinguish between infants who died and infants who survived and was assessed with the concordance (c) statistic.(10) For a binary outcome c is identical to the area under the receiver operating characteristic (ROC) curve.(11)

Results

Between 1 January 2000 and 31 December 2007, 22,827 infants with a gestational age of 22 to 32 weeks were known to be alive at the onset of labour. We excluded 1,449 infants with severe or lethal malformations and 3,796 infants without European ethnicity (i.e. Mediterranean, African, Hindu, Asian), resulting in 17,582 infants. Of those, 2,939 (16.7%) died during labour or within 28 days after birth. For the model to predict neonatal death before discharge from a NICU we excluded 6,004 (34.1%) infants that were not admitted to a NICU. The second model was validated using 11,578 very preterm infants with European ethnicity that were admitted to a NICU; 908 (7.8%) died before discharge.

Characteristics of the infants

The characteristics of the Dutch infants are shown in Tables 1 and 2 next to the infants of the former Trent region. The characteristics of the Dutch infants were generally similar to the characteristics of the infants from the former Trent region, in which the models were developed. The number of mothers that gave birth at an age below 20 was however lower in the Netherlands compared to the former Trent region (2% versus 12%). Furthermore, the proportion of preterm infants that was admitted to a NICU was lower in The Netherlands compared to the former Trent region (66% versus 95%).

No difference was found between the infants that died in both validation groups for the characteristics fetal sex, plurality of pregnancy and the age of the mother (Table 3). Differences were found for gestational age and birth weight with the neonatal deaths at a NICU being more often older (2.4 weeks) and heavier (231 grams) compared to the stillbirths and neonatal deaths in infants known to be alive at the onset of labour. The same differences were found for infants in the former Trent region although less pronounced (gestational age and birth weight differed 0.7 weeks and 60 grams, respectively, data not shown).

Table 1 Characteristics of very preterm infants known to be alive at the onset of labour in the Dutch and Former Trent cohorts

Characteristic	Dutch validation cohort 2000-2007 (n=17,582)		Former Trent development cohort* 1998-2001 (n=3,793)	
	n	%	n	%
Sex				
Female	7,716	43.9	1,771	46.7
Male	9,833	55.9	2,014	53.1
Missing	33	0.2	8	0.2
Gestational age, weeks				
22 – 23	1,222	7.0	172	4.5
24 – 27	2,927	16.6	817	21.5
28 – 32	13,433	76.4	2,804	73.9
Mean, Sd	29.3	2.8	29.1	2.7
Birth weight, g.				
0 – 499	563	3.2	81	2.1
500 – 999	4,116	23.4	989	26.1
1,000 – 1,499	5,803	33.0	1,272	33.5
1,500 – 1,999	5,434	30.9	1,189	31.3
2,000 – 2,499	1,561	8.9	237	6.2
2,500+	68	0.4	25	0.7
Mean, Sd	1,342	489	1,319	469
Missing	37	0.2	0	0.0
NICU admission				
Yes	11,578	65.9	3,585	94.5
No	6,004	34.1	208	5.5
Plurality				
Singleton	11,814	67.2	2,811	74.1
Multiple	5,768	32.8	981	25.9
Missing	0	0.0	1	0.0
Age mother, years				
< 20	343	2.0	437	11.5
20 – 25	2,410	13.7	697	18.4
26 – 30	6,061	34.5	1,077	28.4
31 – 35	6,269	35.7	967	25.5
36+	2,498	14.2	593	15.6
Mean, Sd	30.4	4.9	28.0	6.2
Missing	1	0.0	22	0.6
Outcome				
Dead	2,939	16.7	590	15.6
Alive	14,643	83.3	3,203	84.4

* Infants with European ethnicity, only

Table 2. Characteristics of very preterm infants admitted to a neonatal intensive care unit in the validation and development cohorts

Characteristic	Dutch validation cohort 2000-2007 (n=11,578)		Former Trent development cohort* 1998-2001 (n=3,585)	
	n	%	n	%
Sex				
Female	5,027	43.5	1,673	46.8
Male	6,541	56.5	1,905	53.2
Missing	10	0.0	7	0.2
Gestational age, weeks				
22 – 23	9	0.1	53	1.5
24 – 27	2,008	17.3	748	20.9
28 – 32	9,561	82.6	2,784	77.7
Mean, Sd	29.6	2.0	29.4	2.4
Birth weight, g.				
0 – 499	40	0.3	26	0.7
500 – 999	2,503	21.6	858	23.9
1,000 – 1,499	4,850	41.9	1,256	35.0
1,500 – 1,999	3,450	29.8	1,186	33.1
2,000 – 2,499	707	6.1	235	6.6
2,500+	22	0.2	24	0.7
Mean, Sd	1,347	407	1,357	448
Missing	6	0.0	0	0.0
Apgar score after 5 minutes				
0 – 3	313	2.7	83	2.3
4 – 6	1,225	10.6	284	7.9
7 – 10	9,744	84.2	3,021	84.3
Missing	296	2.6	197	5.5
Plurality				
Singleton	7,889	68.1	2,647	73.9
Multiple	3,689	31.9	937	26.1
Missing	0	0.0	1	0.0
Age mother, years				
< 20	225	1.9	409	11.5
20 – 25	1,616	14.0	646	18.1
26 – 30	3,998	34.5	1,013	28.4
31 – 35	4,089	35.3	931	26.1
36+	1,650	14.3	564	15.8
Mean, Sd	30.4	4.9	28	6.2
Missing	0	0.0	22	0.6
Outcome				
Dead	908	7.8	382	10.7
Alive	10,670	92.2	3,203	89.3

* Infants with European ethnicity, only

Table 3. Characteristics of stillbirths and neonatal deaths in both validation populations

Characteristic	Known to be alive at onset of labour (n=2,939)		Admitted for neonatal intensive care (n=908)	
	n	%	n	%
Sex				
Female	1,237	42.1	337	37.1
Male	1,685	57.3	570	62.8
Missing	17	0.6	1	0.1
Gestational age, weeks				
22 – 23	1,216	41.4	9	1.0
24 – 27	1,189	40.5	516	56.8
28 – 32	534	18.2	383	42.2
Mean, Sd	24.9	2.7	27.3	2.1
Birth weight, g.				
0 – 499	534	18.2	19	2.1
500 – 999	1,881	64.0	543	59.8
1,000 – 1,499	359	12.2	265	29.2
1,500 – 1,999	113	3.8	64	7.0
2,000 – 2,499	29	1.0	15	1.7
2,500+	4	0.1	2	0.2
Mean, Sd	746	349	977	346
Missing	19	0.6	0	0.0
NICU admission				
Yes	782	26.6	908	100
No	2,157	73.4	0	0
Plurality				
Singleton	2,159	73.5	649	71.5
Multiple	780	26.5	259	28.5
Age mother, years				
< 20	72	2.4	21	2.3
20 – 25	399	13.6	126	13.9
26 – 30	947	32.2	303	33.4
31 – 35	1,041	35.4	322	35.5
36+	480	16.3	136	15.0
Mean, Sd	30.5	5.1	30.4	5.0

Mean birth weight per gestational age

As mentioned in the methods section, the birth weight was included in both models as the difference between the observed birth weight and the mean birth weight of infants with the same gestational age and ethnicity (Appendix). Figure 1 shows that the mean birth weight for each gestational age was similar in the Dutch infants and the infants from the former Trent region. The mean birth weights were very much the same for the infants known to be alive

at the onset of labour (Figure 1A). The mean birth weights of the Dutch infants admitted to neonatal intensive care were slightly higher than the mean birth weights of the UK infants, for infants with gestational age from 23 to 25 weeks (Figure 1B).

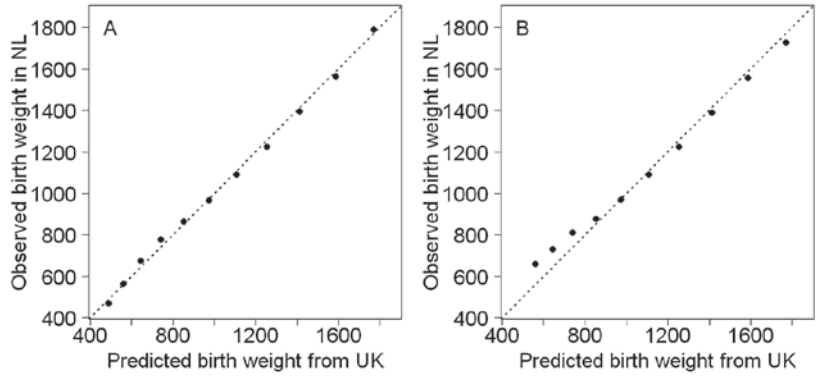


Figure 1 Calibration plot of the observed mean birth weight per gestational age by the predicted mean birth weight for the infants known to be alive at the onset of labor (A) and infants admitted to neonatal intensive care (B). The line of identity indicates perfect agreement between predicted and observed means. The gestational age increases from left to right from 22 to 32 weeks. Note that infants with a gestational age of 22 weeks were not admitted to neonatal intensive care in the Netherlands.

Performance of the prognostic models

Predicted risks of stillbirth or neonatal death were in agreement with the observed proportions of death for infants known to be alive at the onset of labour as is indicated by the line closely following the 45 degree line that indicates perfect calibration (i.e. predicted risk is equal to observed proportion) (Figure 2A). Very high predicted death risks were slightly too low (e.g. a predicted death risk of 80% corresponds to an observed proportion of deaths of 90%). Nevertheless, the mean predicted death risk and overall observed proportion of deaths was similar (16% versus 17%, respectively), as also indicated by a calibration-in-the-large (a|b = 1) of 0.09. The calibration slope b was 1.11. The discriminative ability of the model was excellent with a c statistic of 0.92 and a difference of 0.85 between the means of the highest (0.85) and lowest (0.0) predicted risk deciles.

For the infants admitted to neonatal intensive care predicted death risks showed good agreement with the observed proportions of neonatal deaths (Figure 2B). The mean predicted death risk and overall observed proportion of neonatal deaths was equal (8% versus 8%, respectively). This finding was confirmed by a value close to 0 for a|b=1 (0.11); the calibration slope was 0.86. The discrimination of the model was good (c statistic = 0.82).

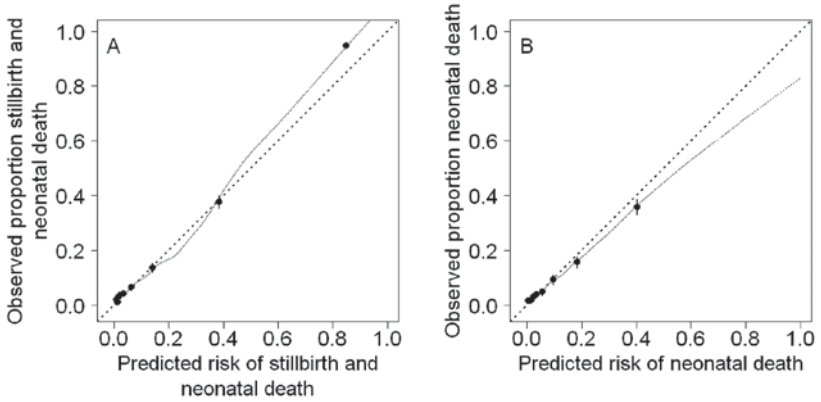


Figure 2 Observed risk of stillbirth and neonatal death by predicted risk for infants known to be alive at the onset of labor (A) and infants admitted to neonatal intensive care (B), respectively. The dots indicate deciles with confidence intervals of infants with similar predicted risk.

Discussion

Principal findings

In this external validation study we assessed the performance of two prognostic models predicting stillbirth and neonatal death for very preterm infants a) known to be alive at the onset of labour or b) admitted for neonatal intensive care. The models that were developed in the UK at the end of the 1990's were found to be valid in a large cohort of 17 thousand Dutch infants born between 2000 and 2007.

The model for very preterm infants known to be alive at the onset of labour showed remarkably good external validity with good calibration and excellent ability to discriminate between infants who died and infants who survived. The c statistic was similar as found by internal validation (0.92 versus 0.91).

The model for very preterm infants admitted to a NICU was well calibrated and the discriminative ability of the model was good. Discrimination of this model was lower than seen for the internal validation (c statistic 0.82 versus 0.89). The lower discrimination for the NICU model is reflected in the calibration plots with the highest decile of predicted risk around 40% (Figure 2B) compared to 80% for those alive at the onset of labour (Figure 2A). Very high predicted risks were rare for the NICU model. However, the overall observed proportions of deaths were similar to the mean predicted risks for both models indicating that the models estimate the risk of death for Dutch infants accurately.

The relatively low proportion of Dutch infants that was admitted for neonatal care (66% in The Netherlands versus 95% in the former Trent region) results from a difference in policy for admitting extremely preterm infants (28 weeks of gestation or less) to a NICU. The Dutch policy is conservative compared to other countries,(3;12-14) and the former Trent policy is more liberal.(15-17) During the years 2000- 2007, Dutch infants were admitted to a NICU

if their gestational age was at least 25 weeks or if they were thought to have a reasonable chance of survival. In the UK, almost all infants are admitted for neonatal intensive care. As a consequence, the extremely preterm infants admitted to a NICU in The Netherlands weigh more average given their gestational age (Figure 1) and are in a better condition. This may explain the lower unadjusted mortality in admitted infants in The Netherlands compared to former Trent (8 vs. 11%). Therefore, the influence on the model performance of the higher weight for infants with low gestational age in the Netherlands as compared to former Trent was investigated. The analyses were repeated in which the mean birth weight of the former Trent infants was substituted with the mean birth weight of the Dutch infants. The results showed similar calibration and discrimination, indicating that the difference in birth weight did not influence the performance of the models.

Implications of the study

Information on the risk of perinatal death of very preterm infants can be of great value for translating the effect of new interventions, for developing of treatment protocols and for for counseling parents and adjusting individual treatment.

The prognostic models may be useful in clinical research, to identify infants at high or low risk for an intervention under study.(18) Furthermore, risk models can be used for risk adjustment in studies carrying out between-hospitals comparisons,(2;18) for example, the CRIB-II score, which was based on the former Trent model for infants admitted for neonatal care.(19)

All the characteristics used in both models are familiar to caregivers and are known to influence the outcome of the newborn. The large advantage of using the prognostic models is that the characteristics are combined in a more formal way allowing for more accurate mortality risk estimation.

In the UK the models are widely used for counselling parents where a very preterm birth is anticipated.(2) If the models are used for counseling or guiding individual infant management, e.g. the obstetricians' decision whether or not to perform a caesarean section, to transfer a patient to a tertiary centre with neonatal intensive care facilities, or to postpone birth or the neonatologists' judgment to withdraw or intensify care, it is important to emphasise, however, that the prognostic models may only complement clinical decision making rather than replace clinical judgment.

Strengths and weaknesses

We used a well described, large, nationwide and heterogeneous cohort of infants from the Netherlands to validate the prognostic models. Data collection was standardised to ensure high quality. The Dutch cohort includes both contemporaneous data (2000-2001) and more recent data (2002-2007) than the development cohort, allowing for temporal and geographical validation.(20)

The prognostic models include a small number of readily available characteristics, which makes them easily applicable in daily clinical practice. Previously developed scoring systems

for neonates, such as the Clinical Risk Index for Babies (CRIB),(21) Score for Neonatal Acute Physiology (SNAP),(22) Neonatal Therapeutic Intervention Scoring System (NTISS),(23) and variations on these scores,(19;24) incorporate characteristics that are not consistently available (e.g. base excess) or variables, which can themselves be influenced by the caregiver (e.g. Apgar score) or actual care provision (e.g. fraction of inspired oxygen).(18;19)

Our study has some limitations. The primary outcome for all infants alive at the onset of labour was defined as stillbirth and death within 28 days of birth as opposed to death before discharge as in the Trent data. In the Netherlands perinatal registry death before discharge was only available for those infants that were admitted to a NICU. From these infants it was calculated that within the children that were admitted to a NICU fifteen percent died after 28 days of life. Since infants not admitted to a NICU are generally in a better condition than infants admitted for neonatal care and are therefore unlikely to die after 28 days of life, we expect little influence of the difference in definition.

The risk factors in the model were assessed at birth. However, the model for infants known to be alive at the onset of labour was developed for the prediction of our primary outcome before labour. Since ultrasound is very reliable in assessing multiplicity of the pregnancy before labour, any bias in the estimated model performance is expected to be low.

The ultrasound-to-delivery interval(25) and the large intra- and interobserver variability(26) can comprise the accuracy of estimated fetal weight (EFW) used in the model for infants known to be alive at the onset of labour. In low birth weight infants the actual birth weight is progressively overestimated with decreasing birth weight, although the difference is not statistically significant.(25;26) Since the interval between ultrasound measurement and delivery can vary widely and the actual birth weight is underestimated more with increasing interval length(25) and thereby overestimating the probability of stillbirth or neonatal death, we expect that model performance could be influenced.

The two models were originally developed for infants from both European and Asian origin. Asian origin was defined as mothers originating from the Indian subcontinent.(4) Since only a small part of our potential validation population had an Asian ethnicity and there was no subclassification of the Asian origin in the medical registries of the three professional organisations, we validated the models on infants of European origin only.

Conclusion

We found that two prognostic models predicting stillbirth and neonatal death in very preterm infants alive at the onset of labour and very preterm infants admitted to neonatal intensive care showed good temporal and geographical validity in a large independent cohort of infants. The prognosis at the onset of labour as well as at the admittance to a neonatal intensive care unit was accurate, implying that the models and the accompanying graphs can be used for clinical practice in the Netherlands and probably also other Western countries.

References

(1) Tucker J, McGuire W. Epidemiology of preterm birth. BMJ 2004 Sep 18;329(7467):675-8.

(2) Dorling JS, Field DJ. Value and validity of neonatal disease severity scoring systems. Arch Dis Child Fetal Neonatal Ed 2008 Mar;93(2):F80-F82.

(3) Zeitlin J, Draper ES, Kollee L, Milligan D, Boerch K, Agostino R, et al. Differences in rates and short-term outcome of live births before 32 weeks of gestation in Europe in 2003: results from the MOSAIC cohort. Pediatrics 2008 Apr;121(4):e936-e944.

(4) Draper ES, Manktelow B, Field DJ, James D. Prediction of survival for preterm births by weight and gestational age: retrospective population based study. BMJ 1999 Oct 23;319(7217):1093-7.

(5) Draper ES, Manktelow B, Field DJ, James D. Tables for predicting survival for preterm births are updated. BMJ 2003 Oct 11;327(7419):872.

(6) Stichting Perinatale Registratie Nederland. Perinatale Zorg in Nederland 2008. 2011. Utrecht, Stichting Perinatale Registratie Nederland.

(7) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-91.

(8) Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45(3-4):562-5.

(9) Harrell FE. Regression Modeling Strategies: With applications to Linear Models, Logistic Regression, and Survival Analysis. Springer; 2001.

(10) Steyerberg EW. Clinical Prediction Models: A practical approach to development, validation, and updating. New York: Springer; 2009.

(11) Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984 Apr;3(2):143-52.

(12) Draper ES, Zeitlin J, Fenton AC, Weber T, Gerrits J, Martens G, et al. Investigating the variations in survival rates for very preterm infants in 10 European regions: the MOSAIC birth cohort. Arch Dis Child Fetal Neonatal Ed 2009 May;94(3):F158-F163.

(13) Gerrits-Kuiper JA, de HR, Bouwers HA, Visser GH, den Ouden AL, Kollee LA. [At the limits of viability: Dutch referral policy for premature birth too reserved]. Ned Tijdschr Geneeskd 2008 Feb 16;152(7):383-8.

(14) Kollee LA, Cuttini M, Delmas D, Papiernik E, den Ouden AL, Agostino R, et al. Obstetric interventions for babies born before 28 weeks of gestation in Europe: results of the MOSAIC study. BJOG 2009 Oct;116(11):1481-91.

(15) Draper ES, Zeitlin J, Field DJ, Manktelow BN, Truffert P. Mortality patterns among very preterm babies: a comparative analysis of two European regions in France and England. Arch Dis Child Fetal Neonatal Ed 2007 Sep;92(5):F356-F360.

(16) Field D, Petersen S, Clarke M, Draper ES. Extreme prematurity in the UK and Denmark: population differences in viability. Arch Dis Child Fetal Neonatal Ed 2002 Nov;87(3):F172-F175.

(17) Field D, Bajuk B, Manktelow BN, Vincent T, Dorling J, Tarnow-Mordi W, et al. Geographically based investigation of the influence of very-preterm births on routine mortality statistics from the UK and Australia. Arch Dis Child Fetal Neonatal Ed 2008 May;93(3):F212-F216.

(18) Manktelow BN, Draper ES, Field DJ. Predicting neonatal mortality among very preterm infants: a comparison of three versions of the CRIB score. Arch Dis Child Fetal Neonatal Ed 2010 Jan;95(1):F9-F13.

(19) Parry G, Tucker J, Tarnow-Mordi W. CRIB II: an update of the clinical risk index for babies score. Lancet 2003 May 24;361(9371):1789-91.

(20) Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999 Mar 16;130(6):515-24.

(21) The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. The International Neonatal Network. Lancet 1993 Jul 24;342(8865):193-8.

(22) Richardson DK, Gray JE, McCormick MC, Workman K, Goldmann DA. Score for Neonatal Acute Physiology: a physiologic severity index for neonatal intensive care. Pediatrics 1993 Mar;91(3):617-23.

(23) Richardson DK, Phibbs CS, Gray JE, McCormick MC, Workman-Daniels K, Goldmann DA. Birth weight and illness severity: independent predictors of neonatal mortality. Pediatrics 1993 May;91(5):969-75.

(24) Richardson DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. J Pediatr 2001 Jan;138(1):92-100.

(25) Cohen JM, Hutcheon JA, Kramer MS, Joseph KS, Abenhaim H, Platt RW. Influence of ultrasound-to-delivery interval and maternal-fetal characteristics on validity of estimated fetal weight. Ultrasound Obstet Gynecol 2010 Apr;35(4):434-41.

(26) Dudley NJ. A systematic review of the ultrasound estimation of fetal weight. Ultrasound Obstet Gynecol 2005 Jan;25(1):80-9.

Appendix

This appendix contains the parameter estimates for the models used to construct the updated tables of Draper et al.(5)

The variables referred to below are:

eth	equal to 1 for European ethnicity of the infant.
gest	gestational age at birth in completed weeks.
mult2	equal to 1 for multiplicity of pregnancy, single equal to 0
sex	equal to 1 for female gender, male gender equal to 0.
wtgeth	difference between observed birth weight and mean birth weight for gestational age for ethnic group (model parameter estimates given below).

Linear regression model to calculate mean birth weight by ethnic group
mean birth weight = 2512 – 451*eth – 228*gest + 19*gest*eth+6.2*gest^2

Labour model: Known to be alive at the onset of labour
Logistic regression model to estimate logit of death to discharge

logit of death = -1*(logit of being alive)
= -1*(- 45.7 + 0.037*mult2 + 2.8*gest – 0.038*gest^2 + 0.019*wtgeth
– 6 .2*10-4*gest*wtgeth – 4.0*10-5*wtgeth^2
+ 1.3*10-6*gest*wtgeth^2 + 1.6*10-9*wtgeth^3)

NICU model: Admitted to neonatal care
Logistic regression model to estimate logit of death to discharge for European babies

logit of death = -1*(logit of being alive)
= -1*(- 17.0 + 0.036*mult2 + 0.43*sex + 0.68*gest + 0.020*wtgeth
– 6.4*10-4*gest*wtgeth – 3.3*10-5*wtgeth^2 + 1.0*10-6*gest*wtgeth^2
+ 1.9*10-9*wtgeth^3)

Chapter 3

Individual participant data
meta-analysis in obstetrics

Chapter 3.1

Progestogens to prevent preterm birth in twin pregnancies: an individual participant data meta-analysis of randomized trials.

Schuit E, Stock S, Groenwold RH, Maurel K, Combs CA, Garite T, Spong CY, Thom EA, Rouse DJ, Caritis SN, Saade GR, Zachary JM, Norman JE, Rode L, Klein K, Tabor A, Çetingöz E, Morrison JC, Magann EF, Briery CM, Serra V, Perales A, Meseguer J, Nassar AH, Lim AC, Moons KGM, Kwee A, Mol BWJ.

BMC Pregnancy Childbirth 2012;12:13.

Abstract

Background Preterm birth is the principal factor contributing to adverse outcomes in multiple pregnancies. Randomized controlled trials of progestogens to prevent preterm birth in twin pregnancies have shown no clear benefits. However, individual studies have not had sufficient power to evaluate potential benefits in women at particular high risk of early delivery (for example, women with a previous preterm birth or short cervix) or to determine adverse effects for rare outcomes such as intrauterine death.

Methods We propose an individual participant data meta-analysis of high quality randomized, double-blind, placebo-controlled trials of progestogen treatment in women with a twin pregnancy. The primary outcome will be adverse perinatal outcome (a composite measure of perinatal mortality and significant neonatal morbidity). Missing data will be imputed within each original study, before data of the individual studies are pooled. The effects of 17-hydroxyprogesterone caproate or vaginal progesterone treatment in women with twin pregnancies will be estimated by means of a random effects log-binomial model. Analyses will be adjusted for variables used in stratified randomization as appropriate. Pre-specified subgroup analysis will be performed to explore the effect of progestogen treatment in high-risk groups.

Discussion Combining individual patient data from different randomized trials has potential to provide valuable, clinically useful information regarding the benefits and potential harms of progestogens in women with twin pregnancy overall and in relevant subgroups.

Background

Preterm birth in twins

Preterm birth in multiple pregnancies is a major public health concern. Stillbirths and neonatal deaths in twins are 3 and 6 times higher than in singletons and a disproportionate amount of long-term morbidity is associated with multiple pregnancies.(1) Preterm birth is the principal factor contributing to these adverse outcomes, with 50% of twin pregnancies delivering before 37 weeks and 9% delivering before 32 weeks.(2) Improving outcomes in multiple pregnancies is a goal of modern obstetrics, but as yet, few interventions have been proven to be of benefit in this group. Thirty-three to 56% of preterm births in twins are due to spontaneous preterm labour, making the prevention of preterm labour an attractive strategy. (2;3)

Progestogens to prevent preterm birth

Randomized trials in singleton pregnancies have suggested that antenatal progestogens (including vaginal progesterone and synthetic progestogens such as 17-hydroyprogesterone caproate [17-OHPC]) prevents preterm delivery in women who are at high risk of preterm delivery because of a previous preterm delivery (4;5) or a short cervix.(6) These trials have led investigators to examine whether antenatal progestogens could decrease preterm birth in multiple pregnancies.

Six randomized controlled trials of progestogens to prevent preterm birth in twin pregnancies have now been published, two large trials from the USA(7;8) and one from the UK(9) and three smaller trials from Turkey(10), the UK(6) and the USA.(11) These studies randomized women to either 17-OHPC / vaginal progesterone pessary, and placebo. One study, which included 67 women with twin pregnancies, 100 mg of vaginal progesterone was found to reduce delivery before 37 weeks gestation (OR 3.48 [1.2-10.5]).(10) In all other studies (6;9),(7), treatment with 17-OHPC or vaginal progesterone did not lead to any significant reduction in preterm delivery or foetal loss. In the two largest trials, however, a non-significant increase in intrauterine death was seen in the treatment group.(8;9) Furthermore, there was found a significant difference in median gestational age favouring placebo, in the other large trial. (7) We are aware of five other trials of progestogens in multiple pregnancy that are nearing completion or publication.(12-16) In total these trials have included 3,522 women and more than 7,000 infants. Combining data from these high-quality clinical trials has potential to provide valuable information regarding the benefits and potential harms of progestogens.

Rationale for an IPD meta-analysis

Aggregated data meta-analysis involves synthesis of estimates from clinical trials. This allows for a more robust estimate of the overall treatment effect of progestogens on multiple

pregnancies as well as a more conclusive evaluation of any harmful effects. This is particularly important, as two published RCTs of progestogens in twins have shown a non-significant trend for increased intrauterine death with progestogen treatment.(8;9) A potential problem in aggregated data meta-analyses is that primary outcomes of clinical trials as well as subgroups defined in clinical trials can differ, which makes it impossible to pool the results of different studies. An Individual participant data (IPD) meta-analysis overcomes this problem as it involves synthesis of individual level data from clinical trials. This allows for the same robust estimate of the treatment effect and harmful effects as in aggregated data meta-analysis, only now more flexibility is possible regarding the choice of endpoints, subgroups and potential harms.

Performing an IPD meta-analysis as opposed to an aggregated data meta-analysis has further advantages. Firstly, IPD allows standardization of inclusion and exclusion criteria and analysis across studies, independent of bias that may arise through selective reporting.(17) Secondly, IPD allows for exploration of a differential treatment effect in relevant subgroups (i.e. treatment covariate interactions), for example, women with a monochorionic twin pregnancy and women with a short cervix.(18) Since IPD meta-analyses include more detailed data on a patient level than aggregated data meta-analyses, statistical power to carry out informative subgroup analyses is higher. Furthermore, flexibility of subgroup analyses is enhanced, thus the estimated subgroup effects may be less influenced by misclassification and bias. IPD meta-analysis therefore allows for a valid assessment of differences in treatment effects across subgroups.(19) Thirdly, IPD allows time-to-event analysis. Conventional meta-analysis only allows a pooled estimate of treatment effect at specified cut-points, i.e. delivery before 32, 34 or 37 weeks. The combined analysis of individual data however, can take account of the time between the initiation of treatment and the outcome of interest.(20) This allows time-to-delivery analysis with the construction of Kaplan-Meier analysis and the performance of Cox regression. This means associations between the timing and duration of progestogen treatment and preterm birth and intrauterine death can be explored. This is important because most published trials have reported a non-significant trend towards a shorter duration of pregnancy after the use of progestogens in women with a multiple pregnancy.(8;9;21;22)

Methods

Criteria for inclusion of studies in IPD

We propose an IPD meta-analysis of randomised controlled trials of 17- OHPC or vaginal progesterone versus placebo in women with twin pregnancies.

Participants

Inclusion criteria will be women with twin pregnancies, with chorionicity and gestation confirmed by ultrasound, who were included in a RCT comparing progestogens with placebo for the prevention of preterm birth. Women with congenital abnormalities, contraindications to progestogen treatment and cervical cerclage will be excluded.

Although studies have been performed in women with triplet pregnancies (12;21;22) these will not be included in the IPD meta-analysis. There appear to be differences in the response to progestogens in women with singleton and twin pregnancies, therefore further differences might be anticipated between women with twins and those with higher order multiples. As the number of trial participants with triplet pregnancies is comparatively small compared to the number with twin pregnancies, excluding these women should not negatively affect the power of the meta-analysis, whilst ensuring the group is as homogenous as possible. Exclusion of triplet pregnancies from the analyses will not compromise the validity of the study since randomization was stratified for twin or higher order multiple in the studies that included both twin and triplet pregnancies.(7;8;12;21;22)

Intervention

The intervention will be either weekly intramuscular injection of 17-hydroxyprogesterone caproate (17-OHPC) or daily vaginal progesterone. As these treatments may act differently and have different distribution profiles we will analyse the results of the two types of treatment separately.

Outcome measures

The primary outcome will be adverse perinatal outcome, a composite outcome of perinatal death, defined as death before discharge from the hospital, and significant neonatal morbidity at discharge, defined as one or more of respiratory distress syndrome (RDS) requiring ventilation for ≥24hr, bronchopulmonary dysplasia (BPD), intraventricular haemorrhage (IVH) grade III or IV), periventricular leucomalacia (PVL), necrotizing enterocolitis (NEC) grade II or more) culture proven sepsis, retinopathy of prematurity (ROP) requiring treatment.

Secondary outcomes will be intrauterine death <32 weeks or preterm birth <32 weeks; intrauterine death <35 weeks or preterm birth <35 weeks; intrauterine death <37 weeks or preterm birth <37 weeks; intrauterine death; foetal loss <28 weeks or early preterm birth <28 weeks gestation; time to delivery or death. If data is detailed enough, preterm birth will be analysed separately for spontaneous preterm birth and indicated preterm birth.

Subgroup analyses

Subgroup analyses will be performed for the primary outcome only, in the following groups each of ultrasonographically diagnosed monochorionic and dichorionic twins

- women who completed ≥90% of treatment
- women with a cervical length <25 mm on baseline assessment (in studies where transvaginal cervical length measurement was specified in protocol)
- women with a prior spontaneous preterm birth <37 weeks
- ethnicity
- each dose of vaginal progesterone, e.g. ≤100 mg versus ≥200 mg.

Identification of studies

We will perform an electronic search of the Cochrane Central Register of Controlled Trials (CENTRAL), PubMed, MEDLINE and ClinicalTrials.gov for published or registered randomized controlled trials including women with twin pregnancy that were randomly allocated to treatment with progestogens (including vaginal progesterone and 17-hydroxyprogesterone caproate) or placebo in the second or third trimester with the intention to prevent preterm birth. We will use the search terms “preterm birth” AND [“progesterone” OR “17 hydroxyprogesterone caproate” OR “progestogen”] AND [“pregnancy multiple” OR “pregnancy twin”] AND “randomised controlled trial” AND “human”.

Two review authors (ES and SS) will independently assess inclusion criteria and study quality and risk of bias. A third author (BWJM) will review studies in which there is any disagreement about study quality. Risks of bias will be assessed in all of the identified studies based on:(23)

- sequence generation (i.e. computer generated random number, use of random number table or other truly random process)
- allocation concealment (i.e. web-based or telephone central randomisation or consecutively numbered sealed opaque envelopes)
- blinding for participants, study personnel and outcome assessors
- Incomplete outcome data
- Selective outcome reporting
- Other sources of bias

In cases where study quality is not clear from trial protocols or publications, then the authors will be contacted for clarification.

The corresponding authors of eligible studies will be approached to take part in the IPD meta-analysis. They will be invited to take part if the study is complete and data available by 1st July 2011.

Data quality will be independently assessed by two review authors (ES and SS). A third author (BWJM) will review data in which there is any disagreement about quality. Only studies with adequate outcome data (< 10% participant attrition or exclusion, with full reporting of reasons

for withdrawals and protocol violations and no imbalance in drop-outs across groups) and adequate reporting (all of the study's pre-specified outcomes and all expected outcomes of interest made available) will be included.

Analysis

Overall effects of each treatment (17-OHPC and vaginal progesterone) in women with twin pregnancies will be estimated in the pooled IPD. Descriptive comparisons between studies will be conducted to assess between-study differences. We assume the data to be missing at random (MAR), therefore observed patient characteristics will be used to impute missing data, by means of multiple imputation.(24) Missing data will be imputed within each original study, before data of the individual studies are pooled. Treatment effects will be estimated by means of a random effects log-binomial model and, hence, the measure of association is the risk ratio. The presence of heterogeneity of outcomes across trials will be assessed using the I^2 measure.(25) Heterogeneity across studies and dependency between data originating from the same study will be taken into account by fitting a random intercept for each original study. If necessary, analyses will be adjusted for variables used in stratified randomization. Furthermore, dependency between children born from the same pregnancy will be accounted for by means of generalized estimating equations (GEE).(26) To investigate subgroup effects, the treatment effects will also be estimated within strata based on single subgrouping variables, as well as using an interaction term in the regression model.

Time-to-delivery analysis will be performed with Kaplan-Meier analysis and Cox proportional hazards regression analysis. Again, dependency between data originating from the same study will be taken into account by conducting a stratified analysis (stratified by study).(27)

We will perform subgroup analysis with pooled individual datasets of women stratified by the pre-specified criteria outlined above. Where available we will plot cervical length against gestational age of cervical length measurement. When differences in gestational age explain differences in length, we will apply standardization for gestational age. Differences in cervical length between the studies will also be explored. We will assess the absolute value of cervical length (corrected for gestational age differences) as well as the percentiles of cervical length (5th, 10th and 25th) in each dataset. We will assess interaction between the treatment effect of progestogens and cervical length, using both time to delivery and the primary and secondary endpoints. To ensure that subgroup effects are not confounded by between-trial differences, dependency between data originating from the same trial will be taken into account using a random intercept for every study in the regression model.(18)

Discussion

The proposed IPD meta-analysis is necessary to determine whether progestogen treatment in twin pregnancy is beneficial or harmful. This is the first study that combines data on the effect of progestogens in twins and the proposed IPD methodology will maximise the impact of results.

The protocol for the individual participant data meta-analysis has been designed with input from the authors of ten randomized controlled trials of progestogens in women with twin pregnancies, many of which have been published in high impact factor journals (Table 1). All authors have committed to providing data if their studies meet inclusion criteria. In total these trials have included 3,498 women and almost 7,000 infants, allowing the meta-analysis to explore effects of progestogens on rare outcomes and in high-risk subgroups. We anticipate it will provide definitive data synthesis guiding clinical practice and future research in this area.

Table 1 Overview of the studies published by authors consulted when planning the IPD meta-analysis

Study	Period	N	Intervention	Primary outcome
Briery(11)	06/04 – 06/10	30	250 mg 17-OPHC or placebo	delivery before 35 completed weeks' of gestation
Cetingoz(10)	12/04 – 02/07	67	100 mg vaginal progesterone or placebo	delivery before 37 weeks
Lim(12)	08/06 – 07/09	654	250 mg 17-OHPC in 1 mL castor oil or placebo	composite outcome (Severe RDS, BPD, IVH grade III or worse, NEC, proven sepsis or death before discharge)
Combs(7)	11/04 – 02/10	240	250 mg 17-OHPC or placebo	composite outcome (RDS, Oxygen therapy at 28d, Neonatal sepsis, Pneumonia, IVH grade III or worse, periventricular leukomalacia, NEC, retinopathy of prematurity, asphyxia)
Nassar(13)	10/06 – 10/10	290	250 mg 17-OPHC or placebo	frequency of delivery prior to completed 37 weeks of gestation (259 days)
Norman(9)	12/04– 04/08	500	Vaginal progesterone gel 90mg or placebo	delivery or intrauterine death before 34 weeks of gestation
Rode(14)	06/06 – 09/10	650	200 mg vaginal progesterone of placebo	incidence of delivery < 34 weeks
Rouse(8)	04/04 – 02/06	661	250 mg 17-OHPC in 1 mL castor oil or placebo	composite outcome (delivery or foetal death before 35 completed weeks of gestation)
Rozenberg(15)	06/06 – 06/10	160	500 mg 17-OPHC or placebo	Interval between inclusion and delivery
Serra(16)	01/06 – 05/08	246	200 mg or 400 mg vaginal progesterone or placebo	Preterm birth rate (<37 weeks)

References

(1) Scottish Perinatal and Infant Mortality and Morbidity Report, 2007. Edinburgh: Information Services Division, NHS Scotland; 2008.

(2) Schaaf JM, Mol BW, bu-Hanna A, Ravelli AC. Trends in preterm birth: singleton and multiple pregnancies in the Netherlands, 2000-2007. BJOG 2011 Sep;118(10):1196-204.

(3) Chauhan SP, Scardo JA, Hayes E, Abuhamad AZ, Berghella V. Twins: prevalence, problems, and preterm births. Am J Obstet Gynecol 2010 Oct;203(4):305-15.

(4) da Fonseca EB, Bittar RE, Carvalho MH, Zugaib M. Prophylactic administration of progesterone by vaginal suppository to reduce the incidence of spontaneous preterm birth in women at increased risk: a randomized placebo-controlled double-blind study. Am J Obstet Gynecol 2003 Feb;188(2):419-24.

(5) Meis PJ, Klebanoff M, Thom E, Dombrowski MP, Sibai B, Moawad AH, et al. Prevention of recurrent preterm delivery by 17 alpha-hydroxyprogesterone caproate. N Engl J Med 2003 Jun 12;348(24):2379-85.

(6) Fonseca EB, Celik E, Parra M, Singh M, Nicolaides KH. Progesterone and the risk of preterm birth among women with a short cervix. N Engl J Med 2007 Aug 2;357(5):462-9.

(7) Combs CA, Garite T, Maurel K, Das A, Porto M. 17-hydroxyprogesterone caproate for twin pregnancy: a double-blind, randomized clinical trial. Am J Obstet Gynecol 2011 Mar;204(3):221-8.

(8) Rouse DJ, Caritis SN, Peaceman AM, Sciscione A, Thom EA, Spong CY, et al. A trial of 17 alpha-hydroxyprogesterone caproate to prevent prematurity in twins. N Engl J Med 2007 Aug 2;357(5):454-61.

(9) Norman JE, Mackenzie F, Owen P, Mactier H, Hanretty K, Cooper S, et al. Progesterone for the prevention of preterm birth in twin pregnancy (STOPPIT): a randomised, double-blind, placebo-controlled study and meta-analysis. Lancet 2009 Jun 13;373(9680):2034-40.

(10) Cetingoz E, Cam C, Sakalli M, Karateke A, Celik C, Sancak A. Progesterone effects on preterm birth in high-risk pregnancies: a randomized placebo-controlled trial. Arch Gynecol Obstet 2011 Mar;283(3):423-9.

(11) Briery CM, Veillon EW, Klauser CK, Martin RW, Chauhan SP, Magann EF, et al. Progesterone does not prevent preterm births in women with twins. South Med J 2009 Sep;102(9):900-4.

(12) Lim AC, Schuit E, Bloemenkamp K, Bernardus RE, Duvekot JJ, Erwich JJ, et al. 17alpha-hydroxyprogesterone caproate for the prevention of adverse neonatal outcome in multiple pregnancies: a randomized controlled trial. Obstet Gynecol 2011 Sep;118(3):513-20.

(13) Nassar AH, Usta IM. Prevention of Preterm Delivery in Twin Pregnancies by 17 Alpha-hydroxyprogesterone Caproate. ClinicalTrials gov Identifier: NCT00141908 2006.

(14) Rode L, Tabor A. Does Progesterone Prevent Very Preterm Delivery in Twin Pregnancies? ClinicalTrials gov Identifier: NCT00329914 2010.

(15) Rozenberg P. Efficacy of 17 Alpha-Hydroxyprogesterones Caproate for the Prevention of Preterm Delivery. ClinicalTrials gov Identifier:NCT00331695 2006.

(16) Serra VS. Natural progesterone and preterm birth in twins. ClinicalTrials gov Identifier: NCT00480402 2007.

(17) Riley RD, Lambert PC, bo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.

(18) Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials 2005;2(3):209-17.

(19) Thompson SG, Higgins JP. Treating individuals 4: can meta-analysis help target interventions at individuals most likely to benefit? Lancet 2005 Jan 22;365(9456):341-6.

(20) Clarke MJ. Individual patient data meta-analyses. Best Pract Res Clin Obstet Gynaecol 2005 Feb;19(1):47-55.

(21) Caritis SN, Rouse DJ, Peaceman AM, Sciscione A, Momirova V, Spong CY, et al. Prevention of preterm birth in triplets using 17 alpha-hydroxyprogesterone caproate: a randomized controlled trial. Obstet Gynecol 2009 Feb;113(2 Pt 1):285-92.

(22) Combs CA, Garite T, Maurel K, Das A, Porto M. Failure of 17-hydroxyprogesterone to reduce neonatal morbidity or prolong triplet pregnancy: a double-blind, randomized clinical trial. Am J Obstet Gynecol 2010 Sep;203(3):248-9.

(23) Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration; 2011.

(24) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-91.

(25) Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-60.

(26) Carlin JB, Gurrin LC, Sterne JA, Morley R, Dwyer T. Regression models for twin studies: a critical review. Int J Epidemiol 2005 Oct;34(5):1089-99.

(27) Smith CT, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Stat Med 2005 May 15;24(9):1307-19.

Chapter 3.2

Effectiveness of progestogens to improve perinatal outcome in twin pregnancies: an individual participant data meta-analysis

Schuit E, Stock S, Rode L, Rouse DJ, Lim AC, Norman JE, Nassar AH, Serra V, Combs CA, Vayssiere C, Aboulghar MM, Wood S, Çetingöz E, Briery CM, Fonseca EB, Klein K, Tabor A, Thom EA, Caritis SN, Awwad J, Usta IM, Perales A, Meseguer J, Maurel K, Garite T, Aboulghar MA, Amin YM, Ross S, Cam C, Karateke A, Morrison JC, Magann EF, Nicolaides KH, Groenwold RHH, Moons KGM, Kwee A, Mol BWJ.

Submitted

Abstract

Background In twin pregnancies, the rates of adverse perinatal outcome and subsequent long-term morbidity are substantial and mainly due to preterm birth.

Objective To assess the effectiveness of progestogen treatment in the prevention of neonatal morbidity or preterm birth in twin pregnancies using individual participant data meta-analysis (IPDMA).

Search strategy We searched international scientific databases, trial registration websites, and references of identified articles.

Selection criteria Randomised clinical trials (RCTs) of either 17-hydroxyprogesterone caproate (17Pc) or vaginally administered natural progesterone compared to placebo or no treatment.

Data Collection and Analysis Investigators of identified RCTs were requested to share their IPD. Data was independently reviewed by two authors and extracted. The primary outcome was a composite of perinatal mortality and severe neonatal morbidity. Prespecified subgroup analyses were performed for chorionicity, cervical length, and history of spontaneous preterm birth.

Results Thirteen trials included 3764 women and their 7528 babies. Neither 17Pc nor vaginal progesterone reduced the incidence of adverse perinatal outcome (17Pc Relative Risk (RR) 1.1; 95% confidence interval (CI) 0.98-1.3, vaginal progesterone RR 0.90; 95% CI 0.74-1.1). Women treated with 17Pc had higher numbers of delivery <32 weeks' gestation (RR 1.3; 95% CI 1.04-1.7) compared with placebo. No specific subgroups were found in which progestogens were more effective.

Conclusions In unselected women with an uncomplicated twin gestation, treatment with progestogens (whether intramuscular 17Pc or vaginal natural progesterone) does not improve perinatal outcome.

Introduction

In Europe and the United States about 3% of all pregnancies are twin pregnancies (1;2). In twin pregnancies, the rates of stillbirth, neonatal death, preterm birth, (very) low birth weight, and subsequent long-term morbidity are substantially higher than in singletons (3). Preterm birth is the principal factor contributing to these adverse outcomes, with 50% of twin pregnancies delivering before 37 weeks and 9% delivering before 32 weeks (4). Six out of 10 preterm births in twins are due to spontaneous preterm labour, making the prevention of preterm labour an attractive strategy to reduce neonatal mortality and morbidity (4;5). Improving outcomes in twin pregnancies is a goal in modern obstetrics, but as yet, no interventions have been proven to be of benefit in this group.

Randomized clinical trials (RCT) have shown that antenatal progestogen therapy (vaginally administered natural progesterone and semi-synthetic progestogens such as intramuscular 17-hydroxyprogesterone caproate (17Pc)) reduces the rate of preterm delivery in women with singleton pregnancies who are at high risk because of preterm birth due to a previous pregnancy (6-9) or a sonographically short cervix in the current pregnancy (10-12).

There has been extensive international interest in determining whether the benefits of progestogens extend to twins, that is, whether these agents reduce the rate of preterm birth and thereby reduce perinatal morbidity. Two recent aggregated data meta-analyses (ADMA) have examined published trials of progestogens in twin pregnancies. One of these did not differentiate between 17Pc and vaginal progesterone (13). The other had too little information to investigate relevant subgroups (14). This is a common limitation of ADMA. Individual participant data meta-analysis (IPDMA) is a more robust design that more easily allows subgroup analysis, time-series analysis, and other advantages.(15;16) A recent IPDMA focused on women with a short cervix and found that vaginal progesterone reduced the rate of early preterm birth and the rate of composite neonatal morbidity/mortality in singleton pregnancies (12). Based on the small number of twins in that analysis, there was a trend toward reduction of early preterm birth with progesterone and a significant reduction of neonatal morbidity/mortality. That meta-analysis did not include any studies of with 17Pc. Moreover, none of the three previous meta-analyses of the effect of progestogens in twins included all published studies (12-14).

The aim of the current study was to perform an IPDMA to investigate the effects of progestogens in women with a twin pregnancy and in prespecified subgroups. Analysis was performed separately for intramuscular 17Pc and vaginally administered natural progesterone.

Methods

The reporting of the IPDMA was done according to the PRISMA guidelines (17). The study was conducted based on a previously published protocol (18).

Trial search and selection strategy

Trials were identified by searching the electronic databases Cochrane Central Register of Controlled Trials (CENTRAL), PubMed, MEDLINE, Embase, ClinicalTrials.gov and controlled-trials.com for published or registered RCTs including women with twin pregnancy that were randomly allocated to treatment with vaginally administered progesterone and intramuscular 17Pc, each versus placebo or non-intervention in the second or third trimester with the intention to prevent preterm birth.

The risk of bias was assessed by two independent reviewers using a risk of bias tool developed by the Cochrane collaboration, which contains specific items that assess adequate sequence generation, allocation concealment, incomplete outcome data, and other concerns about possible sources of bias, e.g. blinding.(19) Likewise, trials were required to have no imbalance in drop-outs across groups and adequate reporting. In cases where study quality was not clear from trial protocols/publications, or if any questions were raised, the authors were contacted for clarification. The principal investigators of all eligible RCTs were contacted to participate and were requested to provide individual participant data. Data quality was independently assessed by two review editors. A third author reviewed the data in case of disagreement about quality. Afterwards the relevant baseline characteristics and outcomes of interest, described in the following subsection, were extracted from these IPD.

Outcomes and subgroups

The primary outcome of this IPDMA was a composite of adverse perinatal outcome and was defined based on the availability of different components of perinatal outcome in the individual studies. In the 17Pc analysis, the composite outcome included perinatal death (defined as intrauterine foetal death (IUFD) or neonatal death before discharge from the hospital), or significant neonatal morbidity, defined as one or more of respiratory distress syndrome (RDS) requiring oxygen for ≥24hr, bronchopulmonary dysplasia (BPD), intraventricular haemorrhage (IVH) grade III or IV), necrotizing enterocolitis (NEC) grade II or more, or culture proven sepsis. In the vaginal progestogen analysis the composite outcome included perinatal death, RDS, IVH, and NEC. The pregnancy was considered to have the primary outcome if at least one of her children had at least one of the components of the composite outcome. Secondary outcomes included the individual neonatal morbidities listed above, IUFD or preterm birth <37, <35, <32 and <28 weeks' gestation, as well as time to delivery or death.

Secondary objectives were to assess the effect of progestogens in different prespecified

subgroups, which were based on results of included studies or previous studies in singletons:

- chorionicity as assessed by ultrasonography and defined as mono- or dichorionic. (20;21)
- cervical length at randomization ≤25 mm (yes/no).(10;12)
- cervical length before 24 weeks' gestation ≤25 mm (yes/no).(10-12)
- prior spontaneous preterm birth <37 weeks (yes/no).(6-9)

The subgroup effects were investigated for the primary outcome, adverse perinatal outcome, and time to delivery or death. Dosage of vaginal progesterone was not investigated since there is sufficient evidence that it does not affect results in twin pregnancies.(22)

Analysis

All analyses were performed on an intention-to-treat basis and were performed separately for 17Pc and vaginal progesterone as compared to control and were performed using the mother, not the fetus/newborn, as the unit of analysis.

Effects of progestogen treatment were estimated by means of a random intercept (to account for baseline differences among studies) fixed effects (effect of progestogen is assumed to be equal among studies) log-binomial model, resulting in a risk ratio (RR) with 95% confidence interval (CI). Heterogeneity across trials was assessed using the I^2 measure (23). Time-to-delivery analysis was performed with Cox proportional hazards regression analysis, resulting in a hazard ratio (HR) with 95% CI. In women who reached the 37 weeks of pregnancy, data were censored as preterm birth could not occur after term. Again, dependency between data originating from the same study was taken into account (24). Number Needed to Treat (NNT) with 95% CI was calculated when an association indicated a beneficial treatment effect that was found to be statistically significant.

Subgroup effects were investigated using an interaction term between the subgroup and the treatment in the regression model. When the interaction was found to be significant ($p < 0.05$), a stratified analysis was performed to investigate the effect of progestogen treatment in different strata of the subgroups. To allow for comparison with a previously published IPDMA (12), stratified analyses were conducted for cervical length irrespective of the significance of the interaction. Statistical analyses were performed using R software, Version 2.15.2 (The R Foundation for Statistical Computing, 2012).

Results

Included studies

Thirty-nine studies were identified through database searching (Figure 1). Thirteen studies (either published or completed but unpublished) met all inclusion criteria (Figure 1, Table 1, Appendix A) (6;10;20;21;25-33). The other studies identified from the literature search

(n=20) were either not investigating twins, described a study protocol or were not expecting to complete the study within the next year (34). Data sets containing IPD were obtained for all 13 RCTs. All studies had institutional review board approval and informed consent from all participants.

The studies randomized women to either 17Pc (21;26-29;31) or vaginal progesterone (6;10;20;25;30;32;33), each versus placebo or non-intervention. The characteristics of the 17Pc and vaginal progesterone included in this IPDMA are shown in Table 1. Eleven studies were placebo-controlled double-blind RCTs (6;10;20;21;26-30;32;33), one was an open-label trial of 17Pc versus no treatment (31), and one was a placebo-controlled trial in which participants were not blinded for treatment (25) (Figure 2).

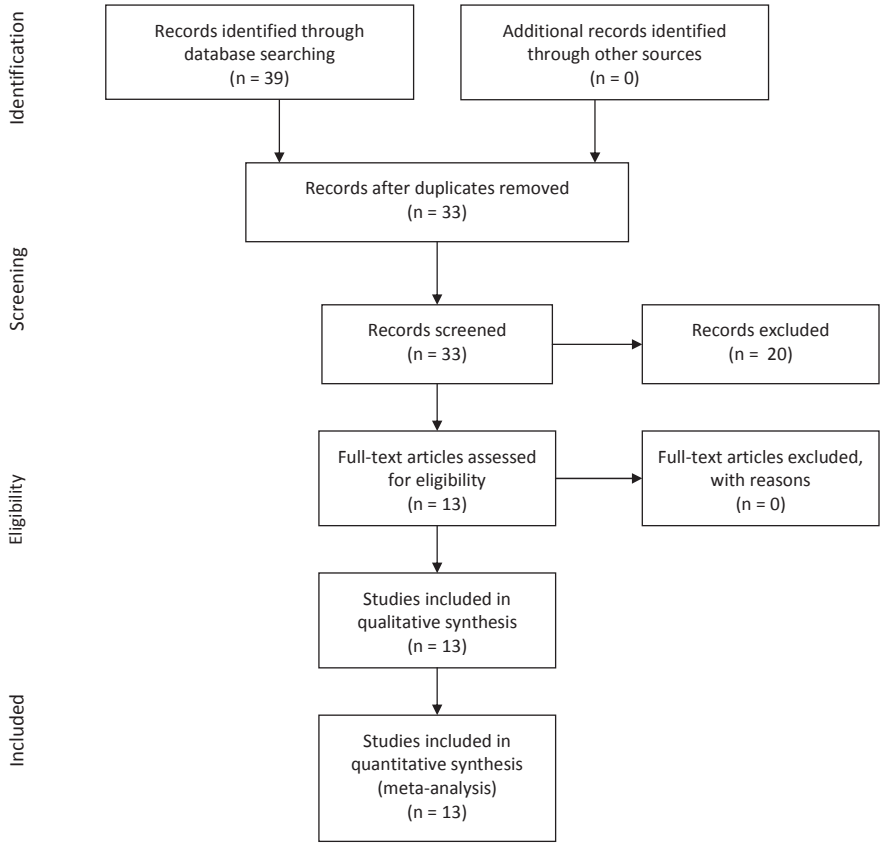


Figure 1 Flow of study identification

Table 1 Overview of the studies investigating progestogen treatment included in the IPDMA

Type of progestogen	Study	Type of study	N	Inclusion criteria	Dosage
17Pc, Weekly	**Rouse(21)**	placebo-controlled double-blind RCT, Multi-centre	661	Women carrying twins with a gestational age of 16^{+0}-20^{+3} weeks	250 mg
	Lim(28)	placebo-controlled double-blind RCT, Multi-centre	650	Women with a twin pregnancy and a gestational age between 15 and 19 weeks, with chorionicity assessed by ultrasonography.	250 mg
	Nassar(29)	placebo-controlled double-blind RCT, Single-centre	286	Viable twin pregnancy between 16-20 weeks of gestation	250 mg
	Combs(27)	placebo-controlled double-blind RCT, Multi-centre	240	Women with a dichorionic-diamniotic twin pregnancy at 15-23 weeks of gestation that an ultrasound examination showing no major foetal anomalies.	250 mg
	Senat(31)	open-label RCT, Single-centre	165	Women carrying twins with cervical length ≤ 25 mm between 20 and 32 weeks of gestation	500 mg
	Briery(26)	placebo-controlled double-blind RCT, Single-centre	30	twin pregnancy between 20–30 weeks' gestation with intact membranes	250 mg
Vaginal, Daily	**Rode(30)**	placebo-controlled double-blind RCT, Multi-centre	677	Women with a live, diamniotic twin pregnancy and chorionicity assessed by ultrasound before 16 weeks' gestation	200 mg pessaries
	Norman(20)	placebo-controlled double-blind RCT, Multi-centre	500	Women carrying twins with gestation and chorionicity established by scan before 20 weeks' gestation	90 mg gel
	Serra(32)	placebo-controlled double-blind RCT, Multi-centre	290	dichorionic diamniotic twin pregnant women	1x or 2x 200 mg pessaries
	Aboulghar(25)	Placebo-controlled RCT, Single-centre	92	Nulliparous women who conceived after IVF or ICSI with a dichorionic twin between 18-24 weeks of gestation	2x 200 mg suppositories
	Wood(33)	placebo-controlled double-blind RCT, Single-centre	81	Women with a multifoetal pregnancy between 16-20+6 weeks gestation and ultrasound confirmed minimum of 2 live foetuses	90 mg gel
	Cetingoz(6)	placebo-controlled double-blind RCT, Single-centre	67	Women carrying twins	100 mg suppositories
	Fonseca(10)	placebo-controlled double-blind RCT, Multi-centre	24	Women carrying twins who were undergoing routine ultrasonography at 20-25 weeks of gestation and had a cervical length < 15 mm	200-mg capsules

RCT = randomized clinical trial; IVF = in vitro fertilization; ICSI = intra-cytoplasmatic sperm injection

In general, all studies enrolled twin pregnancies with a gestational age of randomization between 16^{+0} and 23^{+6} weeks. Several studies had additional inclusion criteria; three studies included only women with a dichorionic twin pregnancy (25;27;32), one vaginal progesterone study included women with a cervical length below 15 mm between 20 and 25 weeks of gestation (10). One 17Pc study included women between 20 and 32 weeks of gestation and a cervical length ≤ 25 mm (31). Two trials allowed randomization of women who had a prophylactic (history-indicated) cerclage in situ (27;30). The exclusion criteria of the studies were similar with most of the studies eliminating twin pregnancies with suspected major foetal abnormalities, suspected twin-to-twin transfusion, serious maternal medical disease, cerclage in-place or planned, contraindication to progestogens, or twin gestations that were the result of intentional foetal reduction (Appendix A). Several studies used stratified randomization, (e.g. by centre (20;21;27;30), chorionicity (20;28;30) or parity (28)). One study was excluded from the analysis of the primary outcome, i.e. adverse perinatal outcome (but not the other outcomes), since only two components of the composite were registered (25).

One study found that progestogens (vaginal progesterone) significantly reduced the rate of preterm delivery before 37 weeks' gestation (6), a result that was not repeated in the other 12 trials (6;10;20;21;25-33). Five studies showed that progestogens were associated with a non-significant increase in preterm births (20;21;26-28), of which one (17Pc) found a significant decrease in median gestational age at delivery (27) and two (vaginal progesterone) showed a non-significant reduction in preterm birth in the progestogen group (10;30). In addition, two 17Pc studies showed a non-significant increase in IUFD (20;21).

The overall quality of the included studies varied from good to fair. One study is currently not peer-reviewed which made risk of bias assessment impossible. A summary of the risk of bias assessment can be found in Figure 2 & 3.Selective outcome reporting was assessed but not considered an issue since IPDMA rely on IPD rather than reported outcomes.

	random sequence allocation bias (selection bias)	allocation concealment (selection bias)	incomplete outcome data (attrition bias)	selective reporting (reporting bias)	other bias	blinding of participants and personnel (performance bias)	blinding of outcome assessment (detection bias)	remark
Rouse 2007	+	+	+	+	+	+	+	
Lim 2011	+	+	+	+	+	+	+	
Nassar 2013	?	?	?	?	?	?	?	1
Combs 2011	+	+	+	+	+	+	+	
Senat 2013	+	+	+	+	+	-	?	1,2
Briery 2009	+	+	?	+	+	+	?	3,4
Rode 2011	+	+	+	+	+	+	+	
Norman 2009	+	+	+	+	+	+	?	4
Serra 2012	+	+	+	+	+	+	?	4
Aboulghar 2012	+	+	+	-	?	-	?	2,4,5,6
Wood 2012	+	+	+	+	+	+	+	
Cetingoz 2011	+	+	+	+	?	+	+	6
Fonseca 2007	?	+	+	+	?	+	+	6,7

Figure 2 Risk of bias for each included study

[1] study not yet published

[2] participants not blinded for allocation

[3] unclear incomplete outcome data since flow diagram is missing

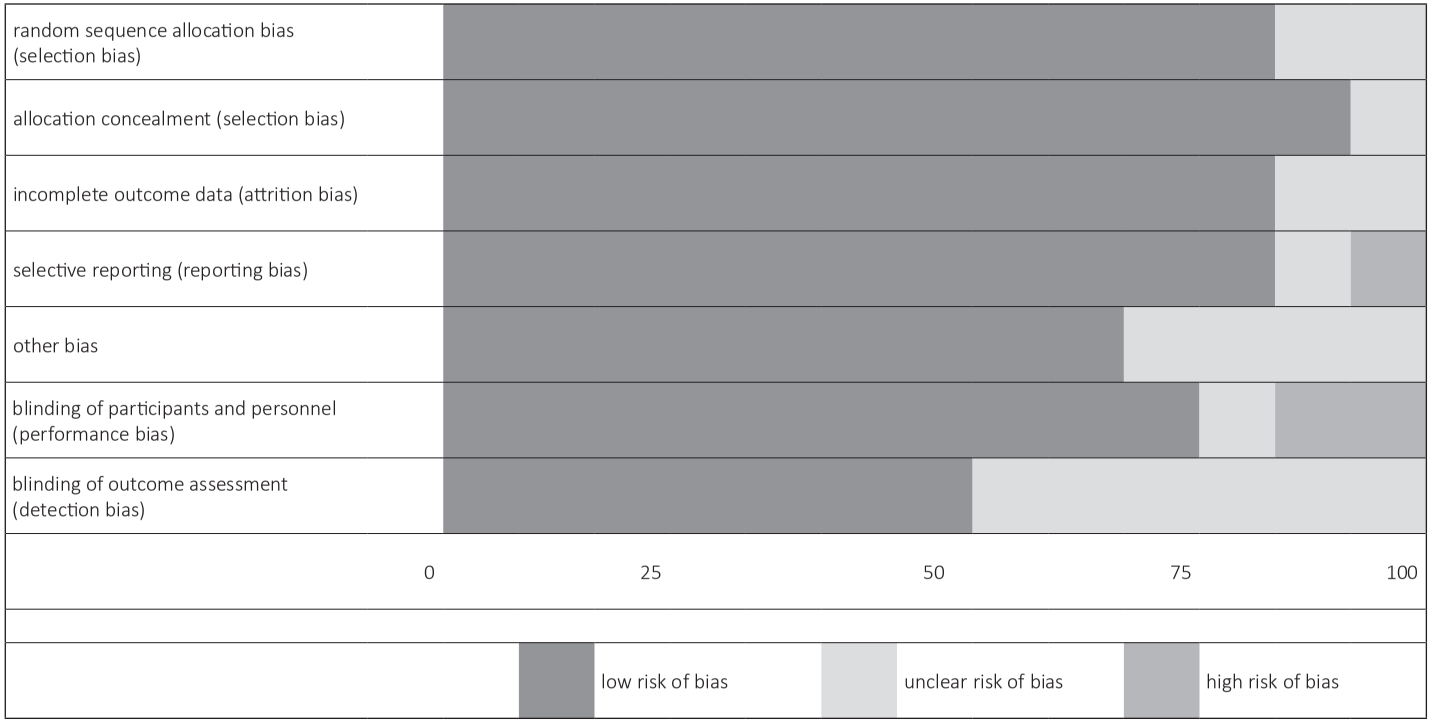
[4] unclear blinding of outcome assessment

[5] neonatal morbidity not presented

[6] unclear whether randomization was successful in twin pregnancies

[7] unclear how randomization was performed

Figure 3 Risk of bias items presented as percentage across included studies



Individual data from 2004 participants allocated to progestogen treatment, 1089 to 17Pc and 915 to vaginal progesterone, and 1760 allocated to control were included in this IPD meta-analysis. The baseline characteristics of the women administered 17Pc or placebo and vaginal progesterone or placebo are presented separately in Table 2. Baseline characteristics for the progestogen treatment group and the control group were comparable, although assisted conception had been performed more in the women in the 17Pc group, and in the vaginal progesterone studies women in the control group drank more alcohol.

Table 2 Baseline characteristics of the participants of the IPD meta-analysis according to type of progestogen and allocation. Data are presented as mean ± standard deviation or n (%). All numbers are based on the data as shared by the individual research groups

	17Pc		Vaginal progesterone	
Characteristics	**progestogen N=1089**	**Control N=944**	**progestogen N=915**	**Control N=816**
Maternal age - yrs. (±SD)	31.6 (5.6)	31.4 (5.8)	32.2 (5.0)	32.2 (5.0)
Body Mass Index - kg/m2 (±SD)	25.8 (6.0)	25.7 (6.3)	24.6 (4.8)	24.6 (5.3)
Gestational age at randomization - wks. (±SD)	19.0 (3.0)	19.0 (2.9)	20.1 (3.1)	20.6 (2.7)
Cervical length at randomization - cm (±SD)*	2.5 (1.6)	2.5 (1.5)	3.8 (0.9)	3.7 (0.9)
Cervical length before 24 weeks gestation - cm (±SD)	4.2 (1.2)	4.3 (0.9)	3.9 (0.8)	3.8 (0.9)
Nulliparous - n (%)	572 (53)	497 (53)	548 (60)	482 (60)
Race				
Black	106 (10)	103 (11)	3 (1)	3 (1)
White	773 (78)	665 (78)	278 (90)	183 (90)
Asian	44 (4)	20 (2)	4 (1)	7 (3)
Other	69 (7)	61 (7)	23 (7)	10 (5)
Smoking - n (%)	95 (9)	80 (9)	117 (14)	88 (12)
Alcohol - n (%)	36 (4)	25 (3)	179 (37)	177 (48)
Assisted conception - n (%)	493 (50)	355 (42)	366 (62)	310 (59)
Monochorionic twin - n (%)	135 (14)	117 (14)	97 (11)	111 (15)
Previous preterm delivery - n (%)	85 (8)	89 (10)	24 (3)	21 (3)

* Cervical length measurement is on the same day or before the moment of randomization. Number of cervical length measurements at or before randomization in studies of Lim, Senat, and Briery were 79, 165, and 3, respectively.

17Pc

Overall effects of 17Pc treatment

The overall effects of treatment with 17Pc are depicted in Table 3. The primary outcome, i.e. adverse perinatal outcome, occurred in 268 (25%) women treated with 17Pc and 199 (22%) in controls (RR 1.2; 95% confidence interval (CI) 0.99-1.3). The rates of the individual components of the composite outcome were comparable between the two groups.

The proportion of IUFD or deliveries before 32 weeks was increased in those treated with 17Pc compared to control (RR 1.3; 95% CI 1.02-1.6). This effect was not found for the single

outcome of IUFD (before 32 weeks) (RR 1.0; 95% CI 0.50-2.0), but was due to a higher proportion of deliveries of all babies (alive and dead) before 32 weeks in the 17Pc group compared to control (RR 1.3; 95% CI 1.04-1.7). For the other secondary outcomes no substantial differences were found between 17Pc treatment and control. All results were similar when corrected for stratified randomization.
We performed three sensitivity analyses. There was one study in which participants and clinicians were not blinded to the 17Pc treatment (31). Exclusion from the IPDMA of all the participants from this particular trial did not lead to different results than the ones presented in this paper (data not shown). A second sensitivity analysis focused on analysis of IPD from studies that included 100+ participants (21;27-29;31) and showed similar results to the main IPDMA (data not shown). The third sensitivity analysis included an analysis of IPD from studies that were published at the time of preparation of this paper. Again, this did not alter the overall results presented in this paper.

Subgroup analyses in 17Pc treatment
17Pc did not have any significant effect on adverse perinatal outcome, or time-to-delivery or death when subgroup analysis was performed according to chorionicity, cervical length at randomization ≤25 mm, cervical length before 24 weeks' gestation ≤25 mm, or incidence of prior spontaneous preterm birth <37 weeks. Exploratory subgroup analyses for adverse perinatal outcome according to cervical length indicated no benefit of 17Pc over control in women with a cervical length below 25 mm, either at randomization (38/83 vs. 23/84; RR 1.6; 95% CI 0.94-2.6; p-value for interaction 0.35) or before 24 weeks (12/21 vs. 2/9; RR 2.7; 95% CI 0.76-9.6, p-value for interaction 0.31).

Vaginal progesterone

Overall effects of vaginal progesterone treatment
The overall effects of vaginal progesterone treatment are shown in Table 3. Adverse perinatal outcome occurred in 141 (16%) women treated with vaginal progesterone and 133 (17%) in controls (RR 0.94; 95% CI 0.75-1.2). The rates of the individual components of the composite outcome were comparable between both groups. For the secondary outcomes no substantial differences were found between vaginal progesterone treatment and control. All results were similar when corrected for stratified randomization.
Two sensitivity analyses were performed. There was one study in which participants were not blinded to the vaginal progesterone (25). Exclusion from the IPDMA of these participants did not lead to different results than the ones presented in this paper (data not shown). An analysis on IPD from studies that included 100+ participants only (20;30;32) did not alter the presented results (data not shown).

Table 3 effects of progestogens on primary and secondary outcomes. Data are presented as mean ± standard deviation or n (%). All numbers are based on the data as shared by the individual research groups

	17Pc				Vaginal progesterone			
Outcome	**17Pc N=1089**	**Control N=944**	**RR (95% CI)***	**P value**	**Vaginal progesterone N=915**	**Control N=816**	**RR (95% CI)***	**P value**
Adverse neonatal outcome - n (%) ¶	268 (25)	199 (22)	1.2 (0.99-1.4)	0. 062	141 (16)	133 (17)	0.94 (0.75-1.2)	0.57
Perinatal death – n (%)	56 (5)	51 (5)	0.97 (0.67-1.4)	0.86	32 (4)	27 (3)	1.1 (0.63-1.7)	0.86
RDS – n (%)	213 (20)	152 (17)	1.2 (>0.99-1.5)	0.05	123 (14)	110 (14)	0.96 (0.75-1.2)	0.72
BPD – n (%)	32 (3)	23 (3)	1.2 (0.70-2.0)	0.51	12 (1)	10 (1)	1.1 (0.47-2.5)	0.86
IVH – n (%)	20 (2)	11 (1)	1.6 (0.76-3.4)	0.21	16 (2)	13 (2)	1.1 (0.54-2.4)	0.74
NEC – n (%)	15 (1)	11 (1)	1.2 (0.54-2.5)	0.69	8 (1)	7 (1)	1.0 (0.37-2.9)	0.95
Sepsis – n (%)	38 (4)	28 (3)	1.1 (0.70-1.8)	0.61	37 (5)	30 (4)	1.1 (0.69-1.8)	0.67
IUFD or delivery < 37 weeks - n (%)	725 (67)	578 (62)	1.1 (0.97-1.2)	0.18	484 (53)	446 (55)	0.97 (0.85-1.1)	0.61
IUFD or delivery < 35 weeks - n (%)	395 (36)	305 (32)	1.1 (0.99-1.3)	0.057	231 (26)	225 (28)	0.93 (0.79-1.1)	0.34
IUFD or delivery < 32 weeks - n (%)	162 (15)	111 (12)	1.3 (1.03-1.6)	0.026	81 (9)	80 (10)	0.92 (0.68-1.2)	0.58
IUFD or delivery < 28 weeks - n (%)	60 (6)	49 (5)	1.1 (0.76-1.6)	0.61	31 (3)	21 (3)	1.3 (0.77-2.3)	0.30
Gestational age at delivery/death - wks. (±SD)	35.0 (3.5)	35.2 (3.6)	1.1 (>0.99-1.2)[#]	0.06	35.8 (3.4)	35.7 (3.1)	0.95 (0.83-1.1)[#]	0.41

* estimated using a random effects model in which a random intercept was fitted for each individual study

¶ Composite outcome was different for 17Pc and vaginal progesterone:

17Pc: a combination of perinatal death, defined as death before discharge from the hospital, and significant neonatal morbidity at discharge, defined as one or more of respiratory distress syndrome (RDS) requiring ventilation for ≥24hr, bronchopulmonary dysplasia (BPD), intraventricular haemorrhage (IVH) grade III or IV), necrotizing enterocolitis (NEC) grade II or more) or culture proven sepsis.

Vaginal progesterone: a combination of perinatal death, and significant neonatal morbidity at discharge, defined as one or more of RDS, IVH, or NEC.

[#] Hazard ratio with 95% confidence interval

Subgroup analyses in vaginal progesterone treatment

Vaginal progesterone did not have any significant effect on adverse perinatal outcome, and time-to-delivery or death when subgroup analysis was performed according to chorionicity, cervical length at randomization ≤25 mm, cervical length before 24 weeks' gestation ≤25 mm, and prior spontaneous preterm birth <37 weeks. Exploratory subgroup analyses for adverse perinatal outcome according to cervical length indicated no benefit of vaginal progesterone over control in women with a cervical length below 25 mm, either at randomization (8/27 vs. 11/25; RR 0.63; 95% CI 0.25-1.6; p-value for interaction 0.29) or before 24 weeks (8/21 vs. 10/20; RR 0.74; 95% CI 0.37-1.5, p-value for interaction 0.28).

Discussion

Main findings

This meta-analysis based on individual participant data from 13 randomized clinical trials of progestogen treatment for the prevention of preterm birth in unselected women with an uncomplicated twin pregnancy shows that progestogen treatment, regardless of type, did not reduce the risk of adverse perinatal outcome compared to control. Women treated with 17Pc had an increased risk of delivery before 32 weeks compared to control.

Interpretation

A recent study investigating the relation between 17Pc concentration and gestational age at delivery showed that 17Pc may reduce gestational age at delivery (35). We similarly found that 17Pc increased the proportion of babies delivering before 32 weeks gestation. Given the increased risk of very early birth, we conclude that 17Pc is contradicted in twin pregnancies.
Two recent meta-analyses suggested that women with a twin pregnancy and short cervical length might benefit from treatment with vaginal progesterone (13;14). In our IPDMA we did not find a significant interaction between short cervical length and progestogen treatment. However, when stratified analyses were performed for adverse perinatal outcome we found a similar suggestion of a trend toward a beneficial effect in women with a short cervix treated with vaginal progesterone (RR 0.63; 95% CI 0.25-1.5). The effect size was similar to that previously found in singletons (RR 0.57; 95% CI 0.40-0.81) and twins (RR 0.56; 95% CI 0.30-0.97) in an IPDMA that did not include all studies currently published (12). Although the potential effectiveness of vaginal progesterone in short cervix twins may have a biologic plausibility, as short cervix might be a first sign of the onset of labour, the trends did not reach statistical significance so we cannot currently recommend routine treatment with vaginal progesterone for women with twin pregnancy and a short cervix. However, these findings should stimulate further research on progesterone in twin pregnancies with a short cervix.
Two individual studies previously reported a non-significant trend toward benefit from vaginal

progesterone in monochorionic twin pregnancies (20;30). One study showed that vaginal progesterone reduced the number of IUFD or deliveries before 34 weeks in monochorionic twins by 48%, while it was increased by 73% in dichorionic twins, both compared to the placebo group (20). We did not find a significant interaction between vaginal progesterone treatment and chorionicity. We conclude that progestogens are not effective in a subgroup of women based on chorionicity.

In the 17Pc studies, the results in this IPDMA are consistent with three of the published studies (21;27;28) of the six included studies but differ from those of one small study (26) which reported a beneficial effect of 17Pc on delivery before 35 weeks. In these six studies the baseline characteristics were comparable and there was a moderate degree of heterogeneity between the primary outcomes (I^2 = 38% and Tau^2 = 0.03).

The results in this IPDMA are consistent with six (6;10;25;30;32;33) of the seven vaginal progesterone studies included, but one study indicated a harmful effect of vaginal progesterone on delivery before 34 weeks (20). It is unclear what caused this difference, because the inclusion criteria of all studies were similar. The baseline characteristics in Table 2 do show that the participants in one study (20) had a less healthy lifestyle than the participants of the other studies, with a high proportion being smokers and alcohol users. Despite these differences all other baseline characteristics were comparable and there was no heterogeneity between the primary outcomes in the six studies (I^2 = 0% and Tau^2 = 0).

Strengths and Limitations

An IPDMA has several distinct advantages over ADMA. IPDMA involves synthesis of individual level data from the individual trials and therefore allows for verification of published results. Since IPD are available, an IPDMA allows for more flexibility regarding the inclusion and exclusion of individuals (15), and choice of endpoints and subgroups (16;36) compared to ADMA. Furthermore, an IPDMA allows for more options to perform subgroup analyses and time-to-event analysis, since it can take account of the time between the initiation of treatment and the outcome of interest (37). This is important because most published trials have reported a non-significant trend toward a shorter duration of pregnancy after the use of progestogens in women with twin pregnancies (20;21). A final advantage of IPDMA is that IPD of unpublished studies can be included in the analysis. It is however important to contact the primary investigators to assess the risk of bias, which is normally assessed using the published article. A sensitivity analysis indicated that our results were not altered by the exclusion of IPD of unpublished studies. Although the results found in this study are similar to the two previously ADMAs (13;14), we can be more confident in our finding of lack of effect of progestogens given our rigorous approach using IPD.

Other strengths of the study are that this is the largest meta-analysis so far conducted on the effects of progestogens in twins and that it includes all currently published studies on

this topic, in contrast to previously published meta-analyses (12-14). Furthermore, our study is strengthened by the low influence of publication bias. Since studies were also identified through trial registries we were able to identify on-going as well as terminated studies. Moreover, most of the trials included in this IPDMA failed to show a benefit of progestogens, but were included in a formal trials register and published according to good practice advice. All analyses were performed on the maternal level. Alternatively, the adverse perinatal outcome could have been calculated on the level of the fetus/newborn, either with or without adjustment for clustering of the children within one mother (38). Re-analysis on the level of the fetus/newborn, taking into account that those from the same mother are dependent, led to similar results (data not shown). Sensitivity analyses were performed in blinded studies, in studies that included 100+ participants, and in published studies. Given that these analyses did not lead to different results than the ones presented in this paper shows the robustness of the results of our study.

The number of women with missing data was very low Consequently, it was not likely that these missing values would influence the final conclusions. Therefore a complete case analysis was considered appropriate.

There are several deviations from our published protocol (18) that need to be discussed. First, the cut-off of cervical length was changed from < 25 mm to ≤ 25 mm to allow better comparison with another meta-analysis (12). Second, periventricular leucomalacia and retinopathy of prematurity were excluded from the composite outcome since this information was not registered in the majority of studies.

The number of women with a twin pregnancy and a short cervix in this IPDMA was small. Consequently, the effect of vaginal progesterone on adverse perinatal outcome in this specific subgroup did not reach statistical significance. Given this low precision it is not possible to draw conclusions on the effectiveness of vaginal progesterone in women with a short cervix twin pregnancy.

Conclusion

In conclusion, this IPDMA has shown that in unselected women with uncomplicated twin pregnancies, treatment with progestogens does not prolong pregnancy or improve perinatal outcome.

References

(1) EURO-PERISTAT. Better statistics for better health for pregnant women and their babies [Internet]. European perinatal health report. [www.europeristat.com]. Last accessed 29 March 2010. 2008.

(2) Martin JA, Hamilton BE, Ventura SJ, Osterman MJ, Kirmeyer S, Mathews TJ, et al. Births: final data for 2009. Natl Vital Stat Rep 2011 Nov 3;60(1):1-70.

(3) Scottish Perinatal and Infant Mortality and Morbidity Report, 2007. Edinburgh: Information Services Division, NHS Scotland; 2008.

(4) Schaaf JM, Mol BW, bu-Hanna A, Ravelli AC. Trends in preterm birth: singleton and multiple pregnancies in the Netherlands, 2000-2007. BJOG 2011 Sep;118(10):1196-204.

(5) Chauhan SP, Scardo JA, Hayes E, Abuhamad AZ, Berghella V. Twins: prevalence, problems, and preterm births. Am J Obstet Gynecol 2010 Oct;203(4):305-15.

(6) Cetingoz E, Cam C, Sakalli M, Karateke A, Celik C, Sancak A. Progesterone effects on preterm birth in high-risk pregnancies: a randomized placebo-controlled trial. Arch Gynecol Obstet 2011 Mar;283(3):423-9.

(7) da Fonseca EB, Bittar RE, Carvalho MH, Zugaib M. Prophylactic administration of progesterone by vaginal suppository to reduce the incidence of spontaneous preterm birth in women at increased risk: a randomized placebo-controlled double-blind study. Am J Obstet Gynecol 2003 Feb;188(2):419-24.

(8) Meis PJ, Klebanoff M, Thom E, Dombrowski MP, Sibai B, Moawad AH, et al. Prevention of recurrent preterm delivery by 17 alpha-hydroxyprogesterone caproate. N Engl J Med 2003 Jun 12;348(24):2379-85.

(9) O'Brien JM, Adair CD, Lewis DF, Hall DR, Defranco EA, Fusey S, et al. Progesterone vaginal gel for the reduction of recurrent preterm birth: primary results from a randomized, double-blind, placebo-controlled trial. Ultrasound Obstet Gynecol 2007 Oct;30(5):687-96.

(10) Fonseca EB, Celik E, Parra M, Singh M, Nicolaides KH. Progesterone and the risk of preterm birth among women with a short cervix. N Engl J Med 2007 Aug 2;357(5):462-9.

(11) Hassan SS, Romero R, Vidyadhari D, Fusey S, Baxter JK, Khandelwal M, et al. Vaginal progesterone reduces the rate of preterm birth in women with a sonographic short cervix: a multicenter, randomized, double-blind, placebo-controlled trial. Ultrasound Obstet Gynecol 2011 Jul;38(1):18-31.

(12) Romero R, Nicolaides K, Conde-Agudelo A, Tabor A, O'Brien JM, Cetingoz E, et al. Vaginal progesterone in women with an asymptomatic sonographic short cervix in the midtrimester decreases preterm delivery and neonatal morbidity: a systematic review and metaanalysis of individual patient data. Am J Obstet Gynecol 2012 Feb;206(2):124-19.

(13) Likis FE, Edwards DR, Andrews JC, Woodworth AL, Jerome RN, Fonnesbeck CJ, et al. Progestogens for preterm birth prevention: a systematic review and meta-analysis. Obstet Gynecol 2012 Oct;120(4):897-907.

(14) Sotiriadis A, Papatheodorou S, Makrydimas G. Perinatal outcome in women treated with progesterone for the prevention of preterm birth: a meta-analysis. Ultrasound Obstet Gynecol 2012 Sep;40(3):257-66.

(15) Riley RD, Lambert PC, bo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.

(16) Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials 2005;2(3):209-17.

(17) Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009 Jul 21;6(7):e1000100.

(18) Schuit E, Stock S, Groenwold RH, Maurel K, Combs CA, Garite T, et al. Progestogens to prevent preterm birth in twin pregnancies: an individual participant data meta-analysis of randomized trials. BMC Pregnancy Childbirth 2012;12:13.

(19) Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration; 2011.

(20) Norman JE, Mackenzie F, Owen P, Mactier H, Hanretty K, Cooper S, et al. Progesterone for the prevention of preterm birth in twin pregnancy (STOPPIT): a randomised, double-blind, placebo-controlled study and meta-analysis. Lancet 2009 Jun 13;373(9680):2034-40.

(21) Rouse DJ, Caritis SN, Peaceman AM, Sciscione A, Thom EA, Spong CY, et al. A trial of 17 alpha-hydroxyprogesterone caproate to prevent prematurity in twins. N Engl J Med 2007 Aug 2;357(5):454-61.

(22) Romero R. Progesterone to prevent preterm birth in twin gestations: what is the next step forward? BJOG 2013 Jan;120(1):1-4.

(23) Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-60.

(24) Smith CT, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Stat Med 2005 May 15;24(9):1307-19.

(25) Aboulghar MM, Aboulghar MA, Amin YM, Al-Inany HG, Mansour RT, Serour GI. The use of vaginal natural progesterone for prevention of preterm birth in IVF/ICSI pregnancies. Reprod Biomed Online 2012 Aug;25(2):133-8.

(26) Briery CM, Veillon EW, Klauser CK, Martin RW, Chauhan SP, Magann EF, et al. Progesterone does not prevent preterm births in women with twins. South Med J 2009 Sep;102(9):900-4.

(27) Combs CA, Garite T, Maurel K, Das A, Porto M. 17-hydroxyprogesterone caproate for twin pregnancy: a double-blind, randomized clinical trial. Am J Obstet Gynecol 2011 Mar;204(3):221-8.

(28) Lim AC, Schuit E, Bloemenkamp K, Bernardus RE, Duvekot JJ, Erwich JJ, et al. 17alpha-hydroxyprogesterone caproate for the prevention of adverse neonatal outcome in multiple pregnancies: a randomized controlled trial. Obstet Gynecol 2011 Sep;118(3):513-20.

(29) Nassar AH, Usta IM. Prevention of Preterm Delivery in Twin Pregnancies by 17 Alpha-hydroxyprogesterone Caproate. ClinicalTrials gov Identifier: NCT00141908 2006.

(30) Rode L, Klein K, Nicolaides KH, Krampl-Bettelheim E, Tabor A. Prevention of preterm delivery in twin gestations (PREDICT): a multicenter, randomized, placebo-controlled trial on the effect of vaginal micronized progesterone. Ultrasound Obstet Gynecol 2011 Sep;38(3):272-80.

(31) Senat M, Porcher R, Winer N, Vayssière C, Deruelle P, Capelle M, et al. Prevention of preterm delivery by 17 alpha-hydroxyprogesterone caproate in asymptomatic twin pregnancies with a short cervix: a randomized controlled trial. Am J Obstet Gynecol. In press 2013.

(32) Serra V, Perales A, Meseguer J, Parrilla J, Lara C, Bellver J, et al. Increased doses of vaginal progesterone for the prevention of preterm birth in twin pregnancies: a randomised controlled double-blind multicentre trial. BJOG 2013;120(1):50-7.

(33) Wood S, Ross S, Tang S, Miller L, Sauve R, Brant R. Vaginal Progesterone to prevent preterm birth in multiple pregnancy: a randomized controlled trial. J Perinat Med. In press 2012.

(34) Brizot M, Liao A. Prophylactic Administration of Natural Progesterone in the Prevention of Preterm Delivery in Twin Pregnancies. ClinicalTrials gov Identifier: NCT01031017 2012.

(35) Caritis SN, Simhan HN, Zhao Y, Rouse DJ, Peaceman AM, Sciscione A, et al. Relationship between 17-hydroxyprogesterone caproate concentrations and gestational age at delivery in twin gestation. Am J Obstet Gynecol 2012 Aug 6;Aug 16. [Epub ahead of print].

(36) Thompson SG, Higgins JP. Treating individuals 4: can meta-analysis help target interventions at individuals most likely to benefit? Lancet 2005 Jan 22;365(9456):341-6.

(37) Clarke MJ. Individual patient data meta-analyses. Best Pract Res Clin Obstet Gynaecol 2005 Feb;19(1):47-55.

(38) Gates S, Brocklehurst P. How should randomised trials including multiple pregnancies be analysed? BJOG 2004 Mar;111(3):213-9.

Appendix A Exclusion criteria included studies

Study	Exclusion criteria
Rode(30)	age <18 years, known allergy to progesterone or peanuts (as the active treatment contained peanut oil); history of hormone associated thromboembolic disorders; rupture of membranes; treatment for or signs of twin-to-twin transfusion syndrome (TTTS); intentional foetal reduction; known major structural or chromosomal foetal abnormality; known or suspected malignancy in genitals or breasts; known liver disease; women with higher-order multiple pregnancies
Lim(28)	Women with a previous spontaneous preterm birth before 34 weeks, serious congenital defects or death of one or more foetuses, early signs of twin-to-twin transfusion syndrome, or primary cerclage
Rouse(21)	serious foetal anomalies, spontaneous death of a foetus after 12 weeks, presumed monoamnionic placenta, suspected twin-to-twin transfusion syndrome, marked ultrasonographic growth discordance (a difference of at least 3 weeks of estimated gestational age between foetuses), planned non study progesterone therapy after 16 weeks, in-place or planned cerclage, major uterine anomaly (e.g., bicornuate uterus), treatment with 10,000 or more units of unfractionated heparin per day, treatment with low-molecular-weight heparin at any dose, and major chronic medical diseases (e.g., insulin-requiring diabetes mellitus or pharmacologically treated hypertension). Twin gestations that were the result of intentional foetal reduction were also excluded.
Norman(20)	Pregnancy complicated by a recognized structural or chromosomal foetal abnormality at the time of recruitment, or if they had contraindications to progesterone, planned cervical suture, planned elective delivery before 34 weeks' gestation, or planned intervention for twin-to-twin transfusion before 22 weeks' gestation.
Serra(32)	Single pregnancy or monochorionic twin pregnancy or triplets; chronic hepatic pathology, previous alterations in hepatic analysis during anti-contraceptive treatment; or gestational cholestasis in previous pregnancies; basal alterations in hepatic analysis profile; prophylactic cerclage until week 14; basal alterations in renal analysis profile; local allergy to micronized natural progesterone; genital pathology not allowing for correct absorption of medication; anomaly diagnosed after sonography week 12 and/or 20; smokers of more than 10 cigarettes/day; consumers of illegal substances
Nassar(29)	Known foetal anomaly in either twin; current or planned cervical cerclage; hypertension requiring medication; diabetes mellitus; asthma; history of deep vein thrombosis; pre-existing cardiac or renal disease; a seizure disorder; previous or existing liver tumours; history of herpes gestationis of pregnancy
Combs(27)	Women were excluded if they were <18 years old; had taken any progestins >15 weeks of gestation; or had symptomatic uterine contractions, rupture of the foetal membranes, any contraindication to prolonging the pregnancy, any pre-existing condition that might be worsened by progesterone, or a pre-existing medical condition carrying a high risk of preterm delivery.
Senat(31)	cervical dilatation > 3 cm; chorioamnionitis; foetal distress; placenta praevia; abruptio placenta; preterm premature rupture of membranes; polyhydramnios; Twin-twin transfusion syndrome; IUGR; preeclampsia or hypertension; other pathology justifying a preterm delivery; epilepsy drugs
Aboulghar(25)	Previous pregnancy, serious foetal anomalies for which termination may be considered, such as a major heart anomaly or major CNS anomaly e.g. anencephaly, spina bifida, intrauterine growth restriction, defined as estimated foetal weight < 10th percentile, mono-chorionic and mono-amniotic twins, uterine anomalies, such as unicornuate, septate and bicornuate uterus, triplet pregnancies and cervical cerclage
Wood(33)	Placenta praevia; pre-existing hypertension; major foetal anomaly; monoamniotic, monozygotic multiples; maternal seizure disorder; history of, or active, thromboembolic disease; maternal live disease; breast malignancy or pathology; progesterone dependent neoplasia; sensitivity to progesterone
Cetingoz(6)	abortions and deliveries between 20 and 24 weeks of gestation, prophylactic cervical cerclage
Briery(26)	severe medical disorders such as sickle cell disease, insulin-dependent diabetes mellitus, chronic hypertension, cervical dilatation ≥ 1 cm, intrauterine growth restriction (<10th percentile), growth discordancy between twins (≥20%), cerclage, uterine abnormalities or unwillingness to participate in the study protocol as described.
Fonseca(10)	Major foetal abnormalities, painful regular uterine contractions, a history of ruptured membranes, or a cervical cerclage.

Chapter 3.3

Effectiveness of cardiotocography with additional ST-analysis in intrapartum fetal monitoring of vertex singleton pregnancies beyond 36 weeks of gestation: an individual participant data meta-analysis.

Schuit E, Amer-Wahlin I, Ojala K, Vayssière C, Westerhuis MEMH, Maršál K, Tekay A, Saade GR, Visser GHA, Groenwold RHH, Moons KGM, Mol BWJ, Kwee A.

Am J Obstet Gynecol 2013 Jan 17.

Abstract

Objective To assess the effectiveness of intrapartum fetal monitoring with cardiotocography (CTG) alone as well as with additional ST-analysis (CTG+ST) in laboring women with a singleton term pregnancy in cephalic presentation in the prevention of metabolic acidosis applying individual patient data meta-analysis (IPDMA).

Study Design We conducted an IPDMA using data from four randomized trials, which enabled us to account for missing data and investigate relevant subgroups. The primary outcome was metabolic acidosis, defined as an umbilical cord-artery pH < 7.05 and a base deficit calculated in the extra cellular fluid compartment > 12 mmol/L. We performed eight explanatory subgroup analyses for eight different endpoints.

Results We analyzed data from 12 987 women and their newborns. Metabolic acidosis was present in 57 (0.9%) women in the CTG+ST group and 73 (1.1%) in the CTG alone group (relative risk (RR) 0.76; 95% confidence interval (CI) 0.53-1.10). Compared to CTG alone, use of CTG+ST resulted in a reduction in the frequency of instrumental vaginal deliveries (RR 0.90; 95% CI 0.83-0.99) and fetal blood samples (RR 0.49; 95% CI 0.44-0.55). Caesarean section rates were comparable between both groups (RR 0.99; 95% CI 0.91-1.09). Subgroup analyses showed that CTG+ST resulted in fewer admissions to a neonatal intensive care unit (NICU) in women with a duration of pregnancy beyond 41 weeks (RR 0.61; 95% CI 0.39-0.95).

Conclusion Intrapartum fetal monitoring with CTG+ST does not reduce the risk of metabolic acidosis, but does reduce the need for instrumental vaginal deliveries and fetal blood sampling.

Introduction

Perinatal asphyxia is associated with several short- and long-term complications, varying from mild hypoxic ischemic encephalopathy to cerebral palsy and death.(1-3) Fetal monitoring during delivery helps identify fetuses at risk of asphyxia. A relatively new method for continuous fetal monitoring is the STAN® methodology (Neoventa Medical, Gothenburg, Sweden) in which (classification of) the cardiotocography (CTG) is combined with ST analysis of the fetal electrocardiogram (ECG). Similar to the post-partum ECG, information can be evaluated about the amplitude of the T-wave in relation to the QRS-complex (T/QRS ratio) and the conduction in the ST-segment. Changes in the fetal ECG in combination with cardiotocography (CTG) abnormalities could be an indication of fetal hypoxia, as shown in previous animal studies.(4;5) Westgate et al. were the first to conduct a randomized controlled trial (RCT) on the effect of intrapartum fetal ECG monitoring.(6) Four subsequent RCTs focused on automatically detected T/QRS changes rather than absolute values of T/QRS.(7-10) All five RCTs were inconclusive with four studies showing no statistically significant effect.(6;8-10) In one study ST-analysis significantly reduced the incidence of neonatal metabolic acidosis. (7)
In order to study the effect of ST-analysis in addition to CTG compared to CTG alone, meta-analyses were performed using aggregated data (ADMA).(11-13) These meta-analyses showed a non-significant reduction of metabolic acidosis when using intrapartum ST-analysis. These meta-analyses relied on published data. Since not all RCTs reported all endpoints of interest, some endpoints were excluded from the meta-analyses.(11-13) Another limitation of these meta-analyses is that they did not investigate subgroups. Obviously, more information on relevant endpoints and subgroups was collected in the individual studies than was reported. A meta-analysis using individual participant data (IPDMA) allows for a more thorough investigation of endpoints and relevant subgroups by taking all this information into account. Furthermore, in IPDMA, it is possible to account for missing data.
In view of the shortcomings of conventional meta-analyses using aggregated data, we performed an IPDMA using data from RCTs to investigate the additional effect of ST-analysis in intrapartum fetal monitoring.

Materials and Methods

This study was conducted based on a previously written, but unpublished protocol. The reporting of the IPDMA was carried out according to the PRISMA guidelines.(14)

Objective
The main objective of this study was to assess the effectiveness of the combination of ST-analysis of the fetal ECG and CTG compared to CTG alone in laboring women with a term singleton pregnancy in cephalic presentation in the prevention of metabolic acidosis by means of an IPDMA.

Search strategy and selection criteria

Trials were identified by searching the following electronic databases for phase III trials of CTG + ST-analysis compared to CTG alone in laboring women with a term singleton pregnancy in cephalic presentation: Cochrane Central Register of Controlled Trials (CENTRAL), PubMed, MEDLINE, Embase, ClinicalTrials.gov and controlled-trials.com, following the search strategy of Becker et al.(11) Two review authors (ES and AK) independently assessed inclusion criteria, study quality, and risk of bias. Discrepancies were resolved by third author (RHHG). The risk of bias was assessed by two independent reviewers (ES and AK) using a modified version of the risk of bias tool developed by the Cochrane collaboration, which contains specific items that assess adequate sequence generation (i.e. computer generated random number, use of random number table or other truly random process), allocation concealment (i.e. web-based or telephone central randomization), incomplete outcome data, and other possible sources of bias.(15) Selective outcome reporting was not considered an issue since IPDMAs rely on IPD rather than reported outcomes. Studies were included if they had a low risk of bias, were focused on T/QRS changes of the fetal ECG, were completed before December 1, 2011, and the principal investigators provided the IPD relating CTG + ST-analysis versus CTG alone. The relevant baseline characteristics and outcomes of interest, described in the following subsection, were extracted by ES. Data quality, e.g. discrepancies between published and shared data, was independently assessed by two review authors (ES and AK) and a third author (RHHG) resolved discrepancies.

Outcomes

The primary outcome was metabolic acidosis$_{BDecf}$, defined as an umbilical cord-artery pH below 7.05 and a base deficit calculated in the extra cellular fluid compartment (BDecf) above 12 mmol/L, calculated with the Sigaard-Andersen algorithm.(16) Secondary outcomes included metabolic acidosis$_{BDblood}$, defined as an umbilical cord-artery pH below 7.05 and a base deficit calculated in blood (BDblood) above 12 mmol/L. Additional secondary outcomes were cord-artery pH < 7.15, cord-artery pH < 7.05, cord-artery pH < 7.00, BDecf > 12 mmol/L, BDblood > 12 mmol/L, 5 minutes Apgar score < 7, admission to a neonatal intensive care unit (NICU), hypoxic–ischemic encephalopathy, intubation, seizures, perinatal death, frequency of fetal blood samples, caesarean section, vaginal instrumental delivery, and the total frequency of operative deliveries. To increase comparability with a currently ongoing RCT conducted by the National Institute of Health Child Development (NIHCD) in the United States, we also used their primary outcome as one of our secondary outcomes. This outcome is a composite of intrapartum fetal death, neonatal death, Apgar score of 3 or lower at 5 minutes, seizure(s), cord artery pH ≤ 7.05 or lower and BDecf ≥ 12 mmol/L or less, intubation for ventilation at delivery, or presence of neonatal encephalopathy(17)

Subgroups

Secondary objectives were to assess the additional effect of ST-analysis in different subgroups differentiated by the following:

- gestational age defined as < 37 weeks, 37-40 weeks, 40-41 weeks or > 41 weeks
- parity defined as nulli- or multiparous
- previous caesarean section (yes/no)
- maternal diabetes mellitus (yes/no)
- induced onset of labor (yes/no)
- meconium stained amniotic fluid (yes/no)
- epidural anesthesia (yes/no)
- and birth weight below the tenth percentile (yes/no).

The subgroup effects were investigated for the primary outcome, metabolic acidosis$_{BDecf}$, as well as for the following secondary outcomes: composite neonatal outcome, caesarean section, need for intubation, NICU admission, hypoxic-ischemic encephalopathy, instrumental vaginal delivery and fetal blood sampling.

Analysis

All analyses were performed on all randomized women in labor of a term singleton in cephalic presentation with an indication for internal CTG monitoring. The analyses were conducted on an intention-to-treat basis, i.e. according to the treatment assigned by randomization regardless of treatment actually received.

Descriptive comparisons between studies were conducted to assess between-study differences. Treatment effects on the primary and secondary outcomes were estimated by means of a random effects log-binomial model. The measure of association was the risk ratio (RR), with RR < 1 indicating treatment benefit. Both heterogeneity across studies and dependency between data originating from the same study were taken into account by fitting a random intercept for each original study by means of a random effects model. The presence of heterogeneity in outcomes across trials was assessed using the I^2 measure and the values were interpreted as follows: 0% indicates no observed heterogeneity, while 25%, 50% and 75% indicate low, moderate, and high heterogeneity, respectively.(18) If necessary, analyses were adjusted for variables used in stratified randomization (e.g. center and/or parity) by including them as covariates in the regression model. Additionally, we calculated the Number Needed to Test (NNT) with 95% CI when an association was found to be statistically significant. The Number Needed to Test is comparable to the Numbers Needed to Treat, but refers to the number of tests, in this case number of laboring women that need to be monitored with CTG with additional ST-analysis, to prevent one case of metabolic acidosis$_{BDecf}$.

To investigate subgroup effects, the treatment effects were investigated using an interaction term between the allocation and the subgroup in the regression model defined above.

When a significant interaction was present, the treatment effect was then estimated within strata based on that subgrouping variable. For the primary outcome, metabolic acidosis$_{BDecf}$ a stratified analysis across the predefined subgroups was performed despite of the presence of a significant interaction in the regression model to investigate the direction of the additional effect of ST-analysis in different strata of the subgroups.
The four RCTs had different proportions of missing values for the primary outcome ranging from 2.4% to 14.5% (Table 1). Because these missing values are often selectively missing, which was also the case in these RCTs (Appendix A, B, C and Appendix 3 of Westerhuis et al(10)), a complete case analysis is likely to yield biased results.(19) To avoid this bias, we used observed patient characteristics to impute missing data by means of multiple imputation. Missing data were imputed (10 times) using a logistic regression model that included the following variables: centre, allocation, parity, neonatal gender, Apgar at 1 minute, Apgar at 5 minutes, arterial pH, arterial BDblood, arterial BDecf, arterial pCO2, venous pH, venous pCO2, birth weight, and indication for the intervention. The primary outcome was included in the imputation model to improve imputations for missing data on other variables of interest. Missing data were imputed within each individual study before pooling the studies.(20) Analyses were performed individually on each of the 10 imputed data sets and results were pooled using standard methods (Rubin's rule).(21)
Statistical analyses and multiple imputation were performed using R software, Version 2.15.0 (The R Foundation for Statistical Computing, 2012).

Results

Included studies
Six studies on ST-analysis in laboring women with a term singleton pregnancy in cephalic presentation were identified, of which four met the inclusion criteria (Figure 1, Table 1).(7-10) The study of Strachan et al. was excluded because it studied the PR waveform of the fetal ECG rather than the ST segment.(22) Even though the study of Westgate et al. focused on T/QRS changes, the study was excluded because the ST-analysis method used was different than methods used in more recent studies.(6) In the study by Westgate et al., the STAN 8801 recorder was used, while the other studies used the STAN S21 and/or S31. Although investigating T/QRS changes, the threshold for performing an intervention was based on the absolute T/QRS ratio and not a change in T/QRS ratio. Furthermore, biphasic ST-changes were not incorporated in the guideline. Another important difference was that ST-changes were identified by visual analysis. The STAN S21 and S31 monitors provide an automatic assessment of the ST-changes and give an automatic warning in case of significant changes. Datasets containing IPD were obtained for four RCTs, which we will refer to hereafter by first author: Amer-Wahlin et al.(7), Ojala et al.(8), Vayssière et al.(9) and Westerhuis et al.(10)

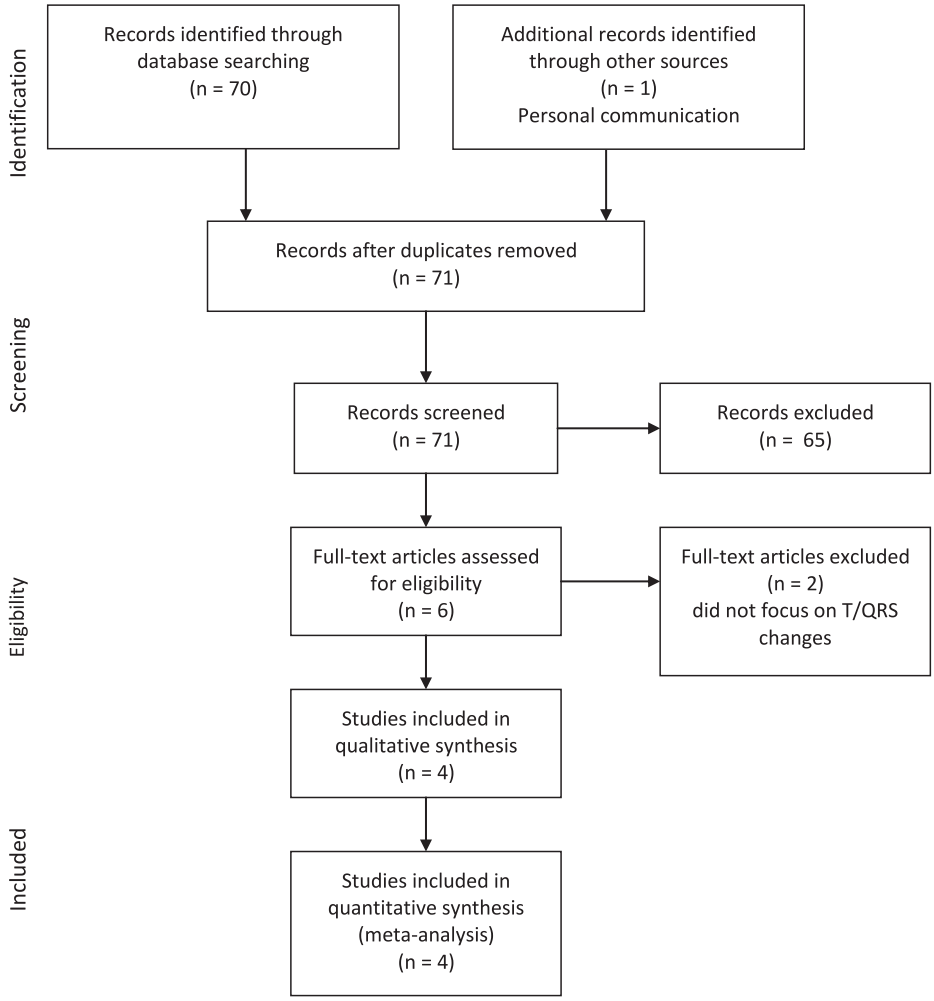


Figure 1 Flow diagram of study selection. The figure shows the process from the identification till the inclusion of studies in this IPDMA.

The characteristics of the included studies are shown in Table 1. In general, all studies had similar in- and exclusion criteria. The only exception is the study of Vayssière et al., which only included women who had an abnormal CTG or thick meconium stained amniotic fluid during labor. Because all studies also used similar interventions and controls, these studies can be considered to have a high degree of homogeneity. The study of Westerhuis et al. stratified the randomization of participants to CTG + ST-analysis or CTG alone by centre and parity (nulli- vs. multiparous).(10)

Table 1 Characteristics of studies on intrapartum monitoring comparing ST-analysis in addition to CTG compared to CTG alone

Characteristic	Amer-Wahlin 2001(7;33) Sweden	Ojala 2006(8) Finland	Vayssiere 2007(9) France	Westerhuis 2010(10;34) Netherlands
Type of study	RCT, N=5049 Multi-centre	RCT, N=1472 Single-centre	RCT, N=799 Multi-centre	RCT, N=5667 Multi-centre
Inclusion criteria	Laboring women > 36 weeks of gestation Singleton fetus Cephalic position Continuous internal CTG needed	Laboring women > 36 weeks of gestation Singleton fetus Cephalic position Amniotomy was decided	Laboring women > 36 weeks of gestation Singleton fetus Cephalic position Abnormal CTG or Thick meconium stained amniotic fluid (7%) during labor.	Laboring women > 36 weeks of gestation Singleton fetus Cephalic position Age > 18 years Indication for internal electronic fetal monitoring
Exclusion criteria	None mentioned in paper	Contra-indication scalp electrode Admittance during second stage of labor	Contra-indication scalp electrode Cardiac malformation	None mentioned in paper
Index test	ST-waveform plus cardiotocography (S21 fetal heart monitor, Neoventa Medical, Gothenburg, Sweden)(35)	ST-waveform plus cardiotocography (S21 fetal heart monitor, Neoventa Medical, Gothenburg, Sweden)(35)	ST-waveform plus cardiotocography (S21 fetal heart monitor, Neoventa Medical, Gothenburg, Sweden)(35)	ST-waveform plus cardiotocography (S21 or S31 fetal heart monitor, Neoventa Medical, Gothenburg, Sweden)(35)
Controls	CTG	CTG	CTG	CTG
Allocation concealment	Yes; adequate	Yes; adequate	Yes; adequate	Yes; adequate
Sequence generation	Yes; adequate	Yes; adequate	Yes; adequate	Yes; adequate
Blinding of participants and medical professionals	No; not possible	No; not possible	No; not possible	No; not possible
Blinding of outcome assessors	Yes; adequate	Yes; adequate	Yes; adequate	Yes; adequate
Number of participant with incomplete primary outcome data	731 (14.5%)	36 (2.4%)	34 (4.3%)	549 (9.7%)

RCT = randomized clinical trial; CTG = cardiotocography;

All trials used adequate methods to generate allocation sequences and as well as adequate methods for allocation concealment (Table 1). Because of the nature of the intervention, blinding of participants and medical professionals was not possible. Blinding the assessors to the outcome was adequate in all trials. The number of women with incomplete primary outcome data differed per study, but could be accounted for using multiple imputation. No other problems were found that could lead to bias.
Individual data from 6524 participants allocated to CTG plus ST-analysis of the fetal ECG and 6463 participants allocated to CTG alone were included in this IPDMA. The baseline characteristics of combined participants by treatment groups were similar (Table 2).

Overall effects of ST-analysis of the fetal ECG
Table 3 shows the effect of ST-analysis in addition to CTG compared to CTG alone for the primary and secondary outcomes. The primary outcome, metabolic acidosis$_{BDecf}$, was present in 57 (0.9%) women in the CTG with additional ST-analysis group and 73 (1.1%) in the CTG alone group (RR 0.76; 95% CI 0.53-1.10). Using a two-step approach (i.e. analysis like an ADMA), we found a moderate amount of heterogeneity for the primary outcome between the studies (I^2 = 42%; 95% CI 0%-81%, τ^2 = 0.09).
The frequency of fetal blood samplings (RR 0.49; 95% CI 0.44-0.55, NNT 13; 95% CI 12-16) and of instrumental vaginal deliveries (RR 0.90; 95% CI 0.83-0.99, NNT 69; 95% CI 38-357) were significantly reduced by CTG in combination with ST-analysis. ST-analysis in addition to CTG did not reduce the incidence of any other secondary outcome. The results were similar even after correction for stratified randomization.

Subgroup analyses
CTG with additional ST-analysis did not show a significant effect for metabolic acidosis$_{BDecf}$ (Table 4), composite neonatal outcome, instrumental vaginal delivery, caesarean section, need for intubation, and hypoxic-ischemic encephalopathy according to gestational age, parity, previous caesarean section, maternal diabetes mellitus, induced onset of labor, meconium stained amniotic fluid, epidural anesthesia, or birth weight below the tenth percentile. It must be noted that information regarding previous caesarean section was not available from Amer-Wahlin(7) and Ojala(8).
Significant subgroup effects were found for two secondary outcomes: fetal blood sampling and NICU admission according to gestational age at delivery and epidural anesthesia (Table 5). CTG with additional ST-analysis reduced the fetal blood sampling more in women with epidural anesthesia than in women without anesthesia (RR 0.46; 95% CI 0.40-0.52 versus RR 0.61; 95% CI 0.49-0.75, p-value of interaction: 0.03). This is rather uninformative, however, as both benefit from additional ST-analysis. Furthermore, CTG with additional ST-analysis reduced the frequency of NICU admissions in women with a gestational age at delivery beyond 41 weeks (RR 0.61; 95% CI 0.39-0.95).

Table 2 Baseline characteristics of participants in each trial and in the overall treatment group.
Data are presented as mean ± standard deviation or n (%). All numbers are based on the data as shared by the individual research groups

Characteristic	Study Publication year Country Sample size	Amer-Wahlin 2001(7;33) Sweden N=5049	Ojala 2006(8) Finland N=1472	Vayssiere 2007(9) France N=799	Westerhuis 2010(10;34) Netherlands N=5667	Combined treatment groups ST-analysis + CTG N=6524	 CTG alone N=6463
Mean maternal age - yrs (±SD)		NA	28.0 (±5.5)	30.0 (±5.7)	32.0 (±4.8)	31.0 (±5.3)	31.0 (±5.3)
Nulliparous - n (%)		3105 (61)	757 (51)	575 (72)	3236 (57)	3851 (59)	3823 (59)
Previous CS - n (%)		NA	NA	49 (6)	716 (13)	370 (11)*	395 (12)*
Diabetes mellitus - n (%)		104 (2)	115 (8)	41 (5)	169 (3)	261 (4)	168 (3)
Female gender of the newborn - n (%)		2388 (47)	733 (50)	NA	2668 (47)	2860 (47)†	2929 (48) †
Gestational age - wks (±SD)		39.6 (±1.6)	40.1 (±1.3)	40.0 (±1.9)	40.2 (±1.4)	39.9 (±1.5)	40.0 (±1.6)
Induced onset of labor - n (%)		866 (17)	277 (19)	257 (36)	2341 (41)	1879 (29)	1862 (29)
Meconium stained amniotic fluid - n (%)		1143 (23)	260 (18)	121 (15)	1471 (26)	1476 (23)	1519 (24)
Epidural anesthesia - n (%)		1957 (39)	793 (54)	725 (91)	2389 (42)	2898 (44)	2966 (46)
Birth weight - g (±SD)		3567 (±531)	3605 (±503)	3243 (±500)	3544 (±518)	3546 (±527)	3536 (±525)

NA = not available
* % based on studies of Vayssiere et al. and Westerhuis et al.
† % based on studies of Amer-Wahlin et al., Ojala et al. and Westerhuis et al.

Table 3 Primary and secondary outcomes per study and the overall effect of ST-analysis in addition to CTG compared to CTG alone
Data are presented as n (%). All numbers are based on the data as shared by the individual research groups

Outcome	Study Sample size	Amer-Wahlin(7;33) N=5049 n (%)	Ojala(8) N=1472 n (%)	Vayssiere(9) N=799 n (%)	Westerhuis(10;34) N=5667 n (%)	Combined treatment groups ST-analysis + CTG N=6524 n (%)	 CTG alone N=6463 n (%)	Relative Risk (95% CI)*	P value	Number needed to test (95% CI)	I^2 (95% CI) [%]
Primary outcome											
Metabolic acidosis$_{BDecf}$ (pH<7.05 & BDecf >12 mmol/L)		54 (1)	10 (1)	19 (2)	46 (1)	57 (0.9)	73 (1.1)	0.76 (0.53-1.10)	0.13	NC	42 (0-81)
Secondary outcomes											
Metabolic acidosis$_{BDblood}$ (pH<7.05 & BDblood>12 mmol/L)		NA	23 (2)	NA	107 (2)	58 (1.6)†	72 (2.0)†	0.82 (0.58-1.16)	0.25	NC	83 (56-93)
Arterial pH <7.15		997 (20)	218 (15)	159 (20)	861 (15)	1118 (17)	1117 (17)	0.99 (0.91-1.08)	0.79	NC	0 (0-84)
Arterial pH <7.05		178 (4)	28 (2)	29 (2)	117 (2)	165 (2.5)	187 (2.9)	0.87 (0.70-1.09)	0.20	NC	68 (6-89)
Arterial pH <7.00		67 (1)	7 (0)	14 (2)	50 (1)	65 (1.0)	72 (1.1)	0.89 (0.62-1.26)	0.48	NC	57 (0-86)
BDecf > 12 mmol/L		150 (3)	34 (2)	123 (15)	204 (4)	266 (4)	246 (4)	1.07 (0.90-1.29)	0.42	NC	0 (0-27)
BDblood > 12 mmol/L		NA	83 (6)	NA	413 (7)	244 (7)†	251 (7)†	0.98 (0.82-1.16)	0.80	NC	0 (0-44)
Apgar at 5 minutes < 7		61 (1)	17 (1)	13 (2)	76 (1)	89 (1.4)	78 (1.2)	1.14 (0.84-1.54)	0.41	NC	0 (0-0)
Admitted to a NICU		387 (8)	49 (3)	10 (1)	86 (2)	258 (4)	274 (4)	0.92 (0.78-1.09)	0.32	NC	0 (0-0)
Hypoxic-ischemic encephalopathy		7 (0)	1 (0)	NA	2 (0)	3 (0.1)	7 (0.2)	0.42 (0.11-1.64)	0.21	NC	0 (0-68)
Need for intubation		NA	12 (1)	8 (1)	22 (3)	16 (1.1)	26 (1.7)	0.64 (0.35-1.20)	0.16	NC	18 (0-92)
Seizures		NA	2 (0)	2 (0)	9 (1)	4 (0.3)	9 (0.6)	0.46 (0.14-1.51)	0.20	NC	0 (0-65)
Perinatal death		3 (0)	0 (0)	1 (0)	5 (0.1)	5 (0.1)	4 (0.1)	1.24 (0.33-4.61)	0.75	NC	0 (0-75)
Composite perinatal outcome‡		75 (1)	28 (2)	33 (4)	91 (2)	101 (1.6)	125 (1.9)	0.80 (0.62-1.05)	0.10	NC	0 (0-82)
Fetal blood sampling		NA	166 (11)	356 (45)	879 (16)	460 (12)	941 (24)	0.49 (0.44-0.55)	<0.0001	13 (12-16)	9 (0-91)
Caesarean section		447 (9)	82 (6)	209 (26)	796 (14)	768 (12)	766 (12)	0.99 (0.91-1.09)	0.91	NC	18 (0-87)
CS fetal distress		194 (4)	30 (2)	119 (15)	164 (3)	253 (4)	254 (4)	0.99 (0.83-1.17)	0.87	NC	42 (0-81)

Table 3 (continued)

Outcome	Study Sample size	Amer-Wahlin(7;33) N=5049 n (%)	Ojala(8) N=1472 n (%)	Vayssiere(9) N=799 n (%)	Westerhuis(10;34) N=5667 n (%)	Combined treatment groups: ST-analysis + CTG N=6524 n (%)	Combined treatment groups: CTG alone N=6463 n (%)	Relative Risk (95% CI)*	P value	Number needed to test (95% CI)	I^2 (95% CI) [%]
Instrumental vaginal delivery		542 (11)	149 (10)	226 (28)	815 (14)	823 (13)	909 (14)	0.90 (0.83-0.99)	0.02	69 (38-357)	0 (0-77)
IVD fetal distress		239 (5)	84 (6)	161 (20)	337 (6)	393 (6)	428 (7)	0.91 (0.80-1.05)	0.19	NC	1 (0-85)
IVD failure to progress		261 (5)	30 (2)	NA	361 (6)	305 (5)	347 (6)	0.87 (0.75-1.01)	0.07	NC	0 (0-87)
Operative delivery		989 (20)	231 (16)	435 (55)	1611 (28)	1591 (24)	1675 (26)	0.94 (0.88-1.01)	0.10	NC	0 (0-79)
Operative delivery due to fetal distress		433 (9)	114 (8)	280 (35)	501 (9)	646 (10)	682 (11)	0.94 (0.84-1.05)	0.26	NC	50 (0-84)
Operative delivery due to failure to progress		478 (9)	67 (5)	NA	870 (15)	694 (11)	722 (12)	0.95 (0.86-1.05)	0.31	NC	65 (0-90)

NA = not available, NC = not calculated

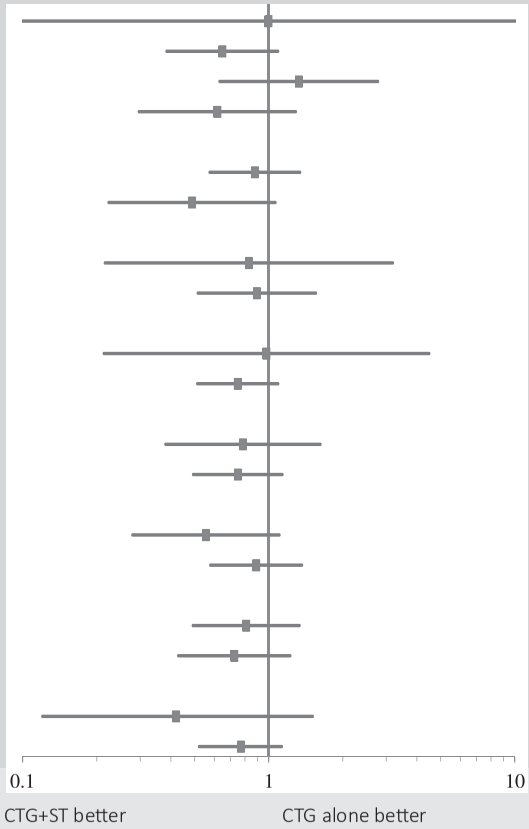
* estimated using a random effects model in which a random intercept was fitted for each individual study

† % based on studies of Ojala et al. and Westerhuis et al.

‡ composite of intrapartum fetal death, neonatal death, Apgar score of 3 or lower at 5 minutes, seizure(s), cord artery pH of 7.05 or lower and base deficit of 12 mmol/L or less, intubation for ventilation at delivery, or presence of neonatal encephalopathy.

Table 4 Risk of neonatal metabolic acidosis$_{BDecf}$ calculated in the extra cellular fluid compartment in relevant subgroups

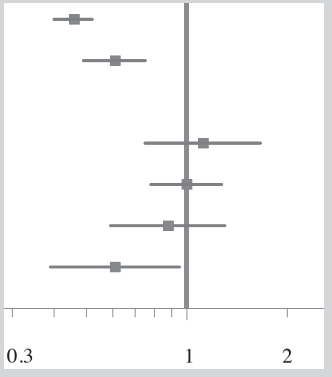
	ST-analysis + CTG (N=6524) Event/n (%)	CTG alone (N=6463) Event/n (%)	Relative risk (95% CI)	Relative risk (95% CI)*	p-value interaction
Gestational age					
≤37 wks	1/326 (0.3)	1/325 (0.3)		1.00 (0.07-14.8)	0.84
37-40 wks	25/3262 (0.8)	38/3167 (1.2)		0.65 (0.38-1.09)	0.36
40-41 wks	18/1575 (1.1)	13/1508 (0.9)		1.32 (0.63-2.78)	0.08
> 41 wks	12/1356 (0.9)	21/1454 (1.4)		0.62 (0.30-1.29)	0.48
Parity					0.17
Nulliparous	46/3850 (1.2)	52/3823 (1.4)		0.88 (0.57-1.34)	
Multiparous	10/2674 (0.4)	21/2640 (0.8)		0.49 (0.22-1.07)	
Previous caesarean section†					0.91
Yes	4/370 (1.1)	5/395 (1.3)		0.83 (0.22-3.17)	
No	26/2853 (0.9)	29/2844 (1.0)		0.89 (0.51-1.56)	
Maternal diabetes mellitus					0.81
Yes	4/261 (1.5)	3/168 (1.8)		0.98 (0.21-4.49)	
No	52/6215 (0.8)	70/6253 (1.1)		0.75 (0.51-1.09)	
Induced onset of labor					0.90
Yes	14/1879 (0.7)	17/1862 (0.9)		0.78 (0.38-1.62)	
No	41/4580 (0.9)	55/4534 (1.2)		0.75 (0.49-1·14)	
Meconium stained amniotic fluid					0.23
Yes	13/1476 (0.9)	25/1519 (1.6)		0.56 (0.28-1·11)	
No	43/5005 (0.9)	47/4902 (1.0)		0.89 (0.58-1·37)	
Epidural anesthesia					0.75
Yes	30/2898 (1.0)	38/2966 (1.3)		0.81 (0.49-1·33)	
No	26/3626 (0.7)	35/3497 (1.0)		0.72 (0.43-1·23)	
Birth weight					0.35
< tenth percentile	4/667 (0.6)	9/641 (1.4)		0.29 (0.03-2.37)	
≥ tenth percentile	48/5728 (0.8)	62/5691 (1.1)		0.78 (0.41-1.47)	



* estimated using a random effects model in which a random intercept was fitted for each individual study
† % based on studies of Vayssiere et al. and Westerhuis et al.

Table 5 Risk of an instrumental vaginal delivery in relevant subgroups

Outcome	Subgroup	ST-analysis + CTG (N=6524) Event/n (%)	CTG alone (N=6463) Event/n (%)	Relative risk (95% CI)	Relative risk (95% CI)*	Number needed to test (95% CI)	p-value interac-tion
Fetal blood sampling	**Epidural anesthesia**						0.03
	Yes	323/2898 (11)	715/2966 (24)		0.46 (0.40-0.52)	8 (7-9)	
	No	137/3626 (4)	226/3497 (6)		0.61 (0.49-0.75)	37 (27-60)	
Admitted to NICU	**Gestational age**						
	≤37 wks	52/326 (16)	46/325 (14)		1.12 (0.75-1.66)	NC	Ref
	37-40 wks	129/3262 (4)	124/3167 (4)		1.00 (0.78-1.27)	NC	0.62
	40-41 wks	48/1575 (3)	53/1508 (4)		0.88 (0.59-1.30)	NC	0.36
	> 41 wks	29/1356 (2)	51/1454 (4)		0.61 (0.39-0.95)	73 (39-669)	0.038



NC = not calculated

* estimated using a random effects model in which a random intercept was fitted for each individual study

Comment

This meta-analysis based on individual participant data from four randomized clinical trials of ST-analysis showed that CTG with additional ST-analysis of the fetal ECG does not reduce metabolic acidosis$_{BDecf}$ and caesarean section rate, but reduces the frequency of fetal blood sampling and instrumental vaginal deliveries as compared to CTG alone. Subgroup analyses showed an additional advantage of additional ST-analysis for women with a gestational age beyond 41 weeks in the reduced frequency of NICU admissions.

The results of this study are in line with the three previously published ADMAs.(11-13) The slight differences in the point estimates for the primary outcome between the meta-analyses and this IPDMA is explained by the inclusion of the trial of Westgate et al.(6) in the ADMAs. First of all, Westgate et al. showed that additional ST-analysis significantly decreases the incidence of metabolic acidosis$_{BDecf}$. Therefore, exclusion of this study from the IPDMA gives a conservative result on the added value of ST-analysis. Secondly, the IPDMA accounted for missing data. Imputation of missing values resulted in 7 (15%) additional cases of metabolic acidosis$_{BDecf}$ in the CTG with additional ST-analysis group and 6 (9%) in the CTG only group. Due to the increase in numerator (more cases of metabolic acidosis$_{BDecf}$) as well as the denominator (no participants with missing outcome values, so all participants included in the analyses), the estimated effect of ST-analysis was expected to be different from the ADMAs.

The point estimate of perinatal metabolic acidosis$_{BDecf}$ in this study is consistent only with two of the four included studies and was opposite to the effects found by Ojala et al. and Vayssiere et al. It is unclear what might have caused this difference. Vayssière et al. used slightly different inclusion criteria (Table 1), which may have led to the inclusion of women at higher risk (Table 2), potentially influencing the effect of ST-analysis on metabolic acidosis$_{BDecf}$. However, the inclusion criteria of Ojala et al. were very similar to the studies of Amer-Wahlin et al. and Westerhuis et al. It must be noted that the studies by Ojala et al. and Vayssière et al. were not powered to find a difference in metabolic acidosis$_{BDecf}$. Despite these differences, the baseline characteristics were comparable. However, there was a moderate degree of heterogeneity between the primary outcomes in the four studies. This heterogeneity is likely a result of different directions of the effect of ST-analysis in the individual studies, with the studies of Ojala et al. and Vayssière et al. that showed RRs above 1 (harmful) and those of Amer-Wahlin et al. and Westerhuis et al. that found RRs below 1 (beneficial). Given the amount of heterogeneity, a fixed effect assumption was considered unrealistic for the outcomes and we therefore used, similar to the ADMAs, random effects models to account for this heterogeneity.

Monitoring with ST-analysis in addition to CTG leads to a reduction in the frequency of operative vaginal deliveries and fetal blood samples. Although the incidence of metabolic acidosis$_{BDecf}$ was reduced by 25%, this reduction was not statistically significant, which might

be due to its low incidence of 1.0%. This incidence was much lower than anticipated by the sample size calculations used by Amer-Wahlin et al.(1.3%) and Westerhuis et al. (3.5%).(7;10) It is important to note that while adverse neonatal outcome might be a more clinically relevant outcome, additional ST-analysis did not lead to a significant reduction. Therefore, results of the ongoing RCT being carried out by the NIHCD in the United States in which the primary outcome is adverse neonatal outcome will be of crucial importance in guiding future management.(17)

As mentioned by Neilson et al., little information about the long-term development of the infants participating in the studies exists.(23) A short follow-up in the Swedish RCT showed a decreased number of neonates with moderate or severe neonatal encephalopathy in the ST-analysis arm.(24) A long-term follow-up study of the Dutch RCT is currently ongoing. Follow-up studies should provide more insight into long-term behavioral and neurological outcomes which will in-turn allow for the investigation of long-term cost-effectiveness. However, as there already seems to be a clear association between neonatal acidosis and long term poor neurological outcome(2), we feel that neonatal asphyxia, defined as metabolic acidosis in the umbilical artery, is the best available surrogate marker.

This IPDMA did not show a reduction of metabolic acidosis for those monitored with ST-analysis compared to CTG alone, but did show a reduced frequency of operative deliveries and fetal blood sampling. Consequently, such discrepancies call for formal cost-effectiveness assessment of ST-analysis compared to CTG alone. Two up-to-date formal cost-effectiveness studies have been performed.(25;26) One study, a long-term cost-effectiveness study based on a probabilistic decision model using Swedish maternity ward data, showed that ST-analysis is cost-effective in comparison to CTG alone.(25) The other study, a cost-effectiveness study based on the study of Westerhuis et al., concluded that the additional costs of monitoring by ST-analysis are very limited when compared to monitoring with CTG only and very low compared with the total costs of delivery.(26) Since the results in this IPDMA are comparable to the results of the study on which the latter cost-effectiveness analysis was based, we feel that the results of this cost-effectiveness analysis can be translated to this IPDMA. To reach a final conclusion on the cost-effectiveness of ST-analysis it may be necessary to perform a formal and extensive cost-effectiveness study using all available individual participant data collected in this meta-analysis. However, this fell outside the scope of the current IPDMA study.

To appreciate the present results, a few limitations need to be addressed. Fetal blood sampling is part of standard obstetric care in all countries in which the included studies were performed. All studies included in the IPDMA that registered information on fetal blood sampling showed a decisive effect of ST-analysis on the use of fetal blood sampling.(8-10) This reduction could be explained by lower and more explicit decision thresholds for initiating the blood sampling, making it less useful. As fetal blood sampling is more common in the

countries included in the study, it might be difficult to generalize these results to countries where this is less common. Comparison with the results on the outcome fetal blood sampling from the ongoing NIHCD-trial in the USA will be particularly interesting.

To investigate the effect of the algorithm used to calculate the BDblood and to be able to utilize umbilical cord gas data from the studies of Amer-Wahlin et al. and Vayssiere et al., which did not record BDblood, a sensitivity analysis was performed. In this sensitivity analysis, the BDblood was calculated using the Corning and the Roche algorithms, which both use a fixed value for the hemoglobin concentration (9.3 mmol/L).(27) After calculation of metabolic acidosis$_{BDblood}$ using these algorithms, the effect of additional ST-analysis was investigated again and was found to be similar to the results in Table 3 (data not shown).

Ojala et al. excluded 11 and Westerhuis et al. 14 women after randomization because it was discovered later that they did not fulfill the inclusion criteria. Technically, this means that analyses in the studies, and therefore also in this IPDMA, were not performed according to the intention-to-treat principle. However, since the excluded women did not fulfill the inclusion criteria, it was justified to leave them out of the analyses.

Implementation of the STAN-methodology should be performed carefully, taking into account the learning curve . The favorable effect of ST-analysis in the study by Amer-Wahlin et al. was mainly observed in the second part of the trial. The interim analysis showed several avoidable protocol violations in participants giving birth to babies with metabolic acidosis$_{BDecf}$. These cases resulted in structured feedback and renewed training.(28) Recent observational studies investigating the effects of long term use of ST-analysis have shown a decrease in the incidence of metabolic acidosis$_{BDecf}$ over time.(28-30) Furthermore, the cases of adverse neonatal outcome described in literature are mainly due to problems with the interpretation of the CTG or violation of guidelines,(31;32) which further supports the hypothesis that the real impact of ST-analysis is still unknown.

In conclusion, this large IPDMA adds to the literature that the addition of ST-analysis to cardiotocography in fetal monitoring does not reduce the incidence of metabolic acidosis$_{BDecf}$, but does reduce the frequency of instrumental vaginal deliveries and the need for fetal blood sampling.

References

(1) Lavrijsen SW, Uiterwaal CS, Stigter RH, de Vries LS, Visser GH, Groenendaal F. Severe umbilical cord acidemia and neurological outcome in preterm and full-term neonates. Biol Neonate 2005;88(1):27-34.

(2) Malin GL, Morris RK, Khan KS. Strength of association between umbilical cord pH and perinatal and long term outcomes: systematic review and meta-analysis. BMJ 2010;340:c1471.

(3) van de Riet JE, Vandenbussche FP, Le CS, Keirse MJ. Newborn assessment and long-term adverse outcome: a systematic review. Am J Obstet Gynecol 1999 Apr;180(4):1024-9.

(4) Westgate JA, Bennet L, Brabyn C, Williams CE, Gunn AJ. ST waveform changes during repeated umbilical cord occlusions in near-term fetal sheep. Am J Obstet Gynecol 2001 Mar;184(4):743-51.

(5) Widmark C, Lindecrantz K, Murray H, Rosen KG. Changes in the PR, RR intervals and ST waveform of the fetal lamb electrocardiogram with acute hypoxemia. J Dev Physiol 1992 Sep;18(3):99-103.

(6) Westgate J, Harris M, Curnow JS, Greene KR. Plymouth randomized trial of cardiotocogram only versus ST waveform plus cardiotocogram for intrapartum monitoring in 2400 cases. Am J Obstet Gynecol 1993 Nov;169(5):1151-60.

(7) Amer-Wahlin I, Hellsten C, Noren H, Hagberg H, Herbst A, Kjellmer I, et al. Cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomised controlled trial. Lancet 2001 Aug 18;358(9281):534-8.

(8) Ojala K, Vaarasmaki M, Makikallio K, Valkama M, Tekay A. A comparison of intrapartum automated fetal electrocardiography and conventional cardiotocography--a randomised controlled study. BJOG 2006 Apr;113(4):419-23.

(9) Vayssiere C, David E, Meyer N, Haberstich R, Sebahoun V, Roth E, et al. A French randomized controlled trial of ST-segment analysis in a population with abnormal cardiotocograms during labor. Am J Obstet Gynecol 2007 Sep;197(3):299-6.

(10) Westerhuis ME, Visser GH, Moons KG, van BE, Benders MJ, Bijvoet SM, et al. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2010 Jun;115(6):1173-80.

(11) Becker JH, Bax L, mer-Wahlin I, Ojala K, Vayssiere C, Westerhuis ME, et al. ST Analysis of the Fetal Electrocardiogram in Intrapartum Fetal Monitoring: A Meta-Analysis. Obstet Gynecol 2012 Jan;119(1):145-54.

(12) Neilson JP. Fetal electrocardiogram (ECG) for fetal monitoring during labour. Cochrane Database Syst Rev 2006;3:CD000116.

(13) Potti S, Berghella V. ST waveform analysis versus cardiotocography alone for intrapartum fetal monitoring: a meta-analysis of randomized trials. Am J Perinatol 2012 Sep;29(8):657-64.

(14) Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009 Jul 21;6(7):e1000100.

(15) Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

(16) Siggaard-Andersen O. An acid-base chart for arterial blood with normal and pathophysiological reference areas. Scand J Clin Lab Invest 1971 May;27(3):239-45.

(17) Belfort MA, Saade GR. ST segment analysis (STAN) as an adjunct to electronic fetal monitoring, Part II: clinical studies and future directions. Clin Perinatol 2011 Mar;38(1):159-67, vii.

(18) Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-60.

(19) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct;59(10):1087-91.

(20) Koopman L, van der Heijden GJ, Grobbee DE, Rovers MM. Comparison of methods of handling missing data in individual patient data meta-analyses: an empirical example on antibiotics in children with acute otitis media. Am J Epidemiol 2008 Mar 1;167(5):540-5.

(21) Rubin DB. Multiple imputation for non response in surveys. New York: Wiley; 1987.

(22) Strachan BK, van Wijngaarden WJ, Sahota D, Chang A, James DK. Cardiotocography only versus cardiotocography plus PR-interval analysis in intrapartum surveillance: a randomised, multicentre trial. FECG Study Group. Lancet 2000 Feb 5;355(9202):456-9.

(23) Neilson JP. Fetal electrocardiogram (ECG) for fetal monitoring during labour. Cochrane Database Syst Rev 2012;4:CD000116.

(24) Noren H, mer-Wahlin I, Hagberg H, Herbst A, Kjellmer I, Marsal K, et al. Fetal electrocardiography in labor and neonatal outcome: data from the Swedish randomized controlled trial on intrapartum fetal monitoring. Am J Obstet Gynecol 2003 Jan;188(1):183-92.

(25) Heintz E, Brodtkorb TH, Nelson N, Levin LA. The long-term cost-effectiveness of fetal monitoring during labour: a comparison of cardiotocography complemented with ST analysis versus cardiotocography alone. BJOG 2008 Dec;115(13):1676-87.

(26) Vijgen SM, Westerhuis ME, Opmeer BC, Visser GH, Moons KG, Porath MM, et al. Cost-effectiveness of cardiotocography plus ST analysis of the fetal electrocardiogram compared with cardiotocography only. Acta Obstet Gynecol Scand 2011 Jul;90(7):772-8.

(27) Mokarami P, Wiberg N, Olofsson P. An overlooked aspect on metabolic acidosis at birth: Blood gas analyzers calculate base deficit differently. Acta Obstet Gynecol Scand 2012 May;91(5):574-9.

(28) Amer-Wahlin I, Kallen K, Herbst A, Rydhstroem H, Sundstrom AK, Marsal K. Implementation of new medical techniques: experience from the Swedish randomized controlled trial on fetal ECG during labor. J Matern Fetal Neonatal Med 2005 Aug;18(2):93-100.

(29) Noren H, Blad S, Carlsson A, Flisberg A, Gustavsson A, Lilja H, et al. STAN in clinical practice--the outcome of 2 years of regular use in the city of Gothenburg. Am J Obstet Gynecol 2006 Jul;195(1):7-15.

(30) Noren H, Carlsson A. Reduced prevalence of metabolic acidosis at birth: an analysis of established STAN usage in the total population of deliveries in a Swedish district hospital. Am J Obstet Gynecol 2010 Jun;202(6):546-7.

(31) Westerhuis ME, Kwee A, van Ginkel AA, Drogtrop AP, Gyselaers WJ, Visser GH. Limitations of ST analysis in clinical practice: three cases of intrapartum metabolic acidosis. BJOG 2007 Oct;114(10):1194-201.

(32) Westerhuis ME, Porath MM, Becker JH, Van Den Akker ES, van BE, van Dessel HJ, et al. Identification of cases with adverse neonatal outcome monitored by cardiotocography versus ST analysis: secondary analysis of a randomized trial. Acta Obstet Gynecol Scand 2012 Jul;91(7):830-7.

(33) Amer-Wahlin I, Kjellmer I, Marsal K, Olofsson P, Rosen KG. Swedish randomized controlled trial of cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram revisited: analysis of data according to standard versus modified intention-to-treat principle. Acta Obstet Gynecol Scand 2011 Sep;90(9):990-6.

(34) Westerhuis ME, Visser GH, Moons KG, Zuithoff N, Mol BW, Kwee A. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2011 Feb;117(2 Pt 1):406-7.

(35) Rosen KG, Lynch P. Intrapartum fetal monitoring: its basis and current developments. Prenatal and Neonatal Medicine 2000;5:155-68.

Chapter 4

Methodological studies

Chapter 4.1

Unexpected predictor-outcome associations in clinical prediction research: causes and solutions.

Schuit E, Groenwold RHH, Harrell FE Jr, de Kort WLAM, Kwee A,
Mol BWJ, Riley RD, Moons KGM.

CMAJ 2013 Jan 21.

Introduction

Clinical prediction models aim to predict individual clinical outcomes using multiple predictor variables. Nowadays, prediction models are abundant in the medical literature and their number is still increasing.(1-3) Established causal risk factors are often good predictors. For example, the Framingham risk score, which predicts the 10-year risk of cardiovascular disease, includes the variables blood pressure, and smoking status, which are well-established risk factors.(1) However, predictors are not necessarily causally related to the clinical outcome, e.g. tumor markers in cancer progression or recurrence.

Even though causality is not necessary, clinical researchers often anticipate a certain direction in the relation between predictor and outcome. For example, higher blood pressure is expected to increase, and not decrease the risk of cardiovascular disease.(1) Thus, when observing a negative association between blood pressure and cardiovascular disease (suggesting a protective effect), this is unexpected and unlikely to be found in another population and could, consequently, hamper generalizability. Moreover, unexpected findings may suggest that the prediction model is actually invalid, thus lowering the face validity of the model, such that readers and potential users will not trust the model to guide their practice.(2)

Our aim is to describe causes for unexpected findings in prediction research and to provide possible solutions. In the first section we describe three clinical examples that will be used for illustrative purposes throughout the paper. The subsequent section outlines causes for unexpected findings and their potential solutions. The last section provides a general discussion and conclusion.

Clinical examples

We use data from three prognostic studies in which an unexpected predictor-outcome relation was found during the development of the prognostic model. These studies are used to illustrate causes and solutions for unexpected findings in clinical prediction research. For illustration purposes we use selective samples of the original data; the validity of the original models is not questioned in any way.

Example 1. Metabolic acidosis in neonates.

Metabolic acidosis in neonates is associated with several short- and long-term complications, including death. Westerhuis et al. developed a prediction model to identify women at risk of giving birth to a child with metabolic acidosis as early as possible.(6) Metabolic acidosis can be the result of a lack of oxygen in the fetus. Since an elevated maternal body temperature leads to more oxygen consumption, it was therefore unexpected that a higher maternal body temperature actually reduced the risk of neonatal metabolic acidosis.(6)

Example 2. Diagnosing deep vein thrombosis.
Deep vein thrombosis is a serious condition with potentially lethal complications such as pulmonary embolism. The gold standard to diagnose deep vein thrombosis is ultrasound, which requires referral to a radiology department. Oudega et al. developed a prediction models to diagnose deep vein thrombosis.(3) In general, deep vein thrombosis is less common among men than women.(8;9) Therefore, the observation that male gender increased the probability of a diagnosis of deep vein thrombosis was unexpected.(3)

Example 3. Anemia in whole blood donors.
To protect donors from being critically anemized after blood donation and to guaranty high donor blood quality, the iron status of blood donors is assessed prior to donation by measuring hemoglobin levels. Deferrals are demoralizing for donors and the risk of subsequent donor lapse increases. Two sex-specific models to predict low hemoglobin levels in whole blood donors were developed by Baart et al.(4) The total number of whole blood donations in the past 2 years was expected to increase the chance of low hemoglobin levels, but (unexpectedly) actually lowered the probability of low hemoglobin levels.

Causes and solutions for unexpected findings in prediction research
Causes for unexpected findings in prediction research include chance, misclassification of the predictor, selection bias, mixing of effects (confounding), intervention effects and heterogeneity (Box 1). One rigorous possible solution that appeals to all possible causes for unexpected findings is to delete the predictor with the unexpected finding from the model. However, this is undesirable as this likely reduces the predictive ability of the clinical prediction model,(5) and reduces face validity of the model. Hereafter, we will discuss and illustrate the causes and solutions depicted in Box 1 in more detail and illustrate these using the aforementioned clinical examples.

Chance
The direction of the found (estimated) predictor-outcome relation may be opposite from the anticipated direction merely by chance. For example, the observed relation between gestational age and neonatal metabolic acidosis (example 1) was observed to be an odds ratio (OR) of 1.17 (95% confidence interval (CI) 1.09-1.27), which is in line with clinical experience (i.e. expected), since the chance of metabolic acidosis increases with increasing gestational age. However, if we take random samples from this dataset, by chance we may observe an opposite relation. For example, we took 1,000 random samples of size 50, 100, 250, 500 or 1000 subjects from this same data source. Among these samples, the proportion of unexpected findings (OR < 1) decreased with increasing sample size: 37.3%, 33.5%, 25.4%, 18.3% and 11.0% respectively. Hence, the probability of finding an unexpected finding of the predictor-outcome relation by chance strongly depends on sample size.

Box 1 Causes of and solutions for unexpected findings in prediction research.

Cause and descriptionof an unexpected finding	Solution How to deal with unexpected findings
Chance Due to chance, the direction of the predictor-outcome relation can be unexpected, especially when sample sizes are small.	*Chance* • Advanced predictor selection strategies • Proper design of prediction model studies with suitable sample size calculation • Delete predictor from the model
Misclassification Unexpected findings due to misclassification can occur when a predictor is measured or coded with error, the predictor-outcome relation is modelled incorrectly, or two or more variables are included even though they are collinear.	*Misclassification* • Redo the measurement/reclassify • Delete erroneous value (if known) and impute • Model predictor on its continuous scale and consider non-linear trends • Include strongest predictor of two collinear variables • Combine two collinear variables into one predictor • Delete predictor from the model
Selection when selection is related to both the predictor and the outcome, either at inclusion, during follow-up, or during the outcome assessment (Figure 1).	*Selection* • Apply weighting to undo selection process • Add subjects to undo the selection process • Clearly define the domain in which the model is applicable • Delete predictor from the model
Mixing of effects (confounding) When two causes of the outcome are mutually related, the observed effect of one can be mixed up with the effect of the other, potentially resulting in an unexpected finding (Figure 1).	*Mixing of effects (confounding)* • Add extra variable to the model • Delete predictor from the model
Intervention effects A predictor value can trigger a medical intervention, which subsequently lowers the probability of the outcome, thereby attenuating the observed relation between the predictor and the outcome (Figure 1).	*Intervention effects* • Add intervention to the model • Delete predictor from the model
Heterogeneity Predictor effects may differ across subgroups (i.e., interaction or heterogeneity of predictor effects). If the distribution the subgrouping factor in the study population differs from the distribution of this factor in the 'typical' patient population, this may lead to an unexpected finding.	*Heterogeneity* • Add interaction term to the model • Delete predictor from the model

Misclassification

The status of a predictor may be measured or coded with error (e.g. coding women with intrapartum fever as women without and vice versa) and hence may lead to incorrect classification (i.e. predictor misclassification). Furthermore, the predictor-outcome relation may be modeled incorrectly, for example when an incorrect transformation is used for a continuous predictor (e.g. linear instead of non-linear), or when categorizing a continuous variable.(3;12) Finally, collinearity of variables may result in apparent misspecification.

Collinearity arises when two or more predictors are highly correlated and so explain similar components of the variability in patient outcome. For example body mass index and weight are by definition strongly correlated. Including these predictors together can lead to poor estimation of the individual predictor estimates, in particular an inflated standard error and low power, which may increase the possibility of unexpecting findings. However, in terms of predictive accuracy of the overall model, collinearity will usually not affect performance as long as the collinearities in future data are similar to those identified in the data used to develop the model.

Possible solutions for misclassification of a predictor include redoing the measurement (if possible), and modeling the continuous predictor appropriately (e.g., by splines or fractional polynomials).(6) Alternatively, if it is known which values are measured with significant error, these could be deleted and imputed.(7) In the case of collinearity options include omitting some of the affected predictor variables from the model, or combining them into a single variable (e.g. mean arterial pressure instead of systolic and diastolic blood pressure) by adding and/or summing them. However, it may be entirely sensible to include collinear predictors together in the model in order to improve the overall predictive accuracy. In this situation an important recommendation is to interpret any collinear predictors in combination, rather than separately. For example, in a model where the highly collinear variables age and age2 are both included, one should discuss the quadratic relationship due to age and not focus on the individual estimates for age or age2. A more extreme solution is to adopt a different regression technique, such as ridge regression, but this itself may lead to biased predictor effect estimates and make the model hard to interpret.

Misclassification was a potential cause of the unexpected finding in the example of predicting risk of metabolic acidosis in neonates (example 1). After reclassification of intrapartum fever from 37.8 ºC to 38.5ºC the initial unexpected finding in the predictor outcome relation (OR 0.86; 95%CI 0.68-1.08) disappeared (OR 1.43; 95%CI 0.99-2.08).(6) The choice of temperature (threshold value) to define fever is thus influential in the direction of the predictor effect, and a better approach may be to rather analyse temperature as a continuous predictor here.

Selection bias

If the study population is a selective sample from the total patient population (domain), this may result in biased estimates, for example when the selection is related to both the predictor and the outcome. Selection can occur at different phases during a study, e.g. at inclusion (e.g. index event bias), during follow-up (e.g. selective dropout), or during the measurement of the outcome at interest (e.g. when not all patients undergo the same reference test, referred to as differential-verification).(8)

The mechanism resulting in selection bias is schematically shown in Figure 1. If both the predictor and the outcome (possibly through a symptom of the disease of interest) affect

the probability of selecting a subject for a study, this may induce a bias (i.e., selection or collider stratification bias). A possible solution for selection bias is to apply weighting, in which a subgroup can be given extra weight to compensate possible underrepresentation. (9) The extent of the underrepresentation, however, is typically unknown, and the weights will therefore depend on unverifiable assumptions.(9) Another solution is to clearly define the domain in which the model is applicable, e.g. only in patients suspected of deep vein thrombosis in secondary care.

In the diagnostic study on deep vein thrombosis (example 2) males unexpectedly had a higher probability of a diagnosis of deep vein thrombosis than women (OR = 1.84; 95% CI 1.41-2.40). This may be due to an over-representation of women without deep vein thrombosis in the study population. If female gender is a risk factor for deep vein thrombosis, primary care physicians may suspect deep vein thrombosis more often in women than in men and consequently more women without deep vein thrombosis are referred to secondary care.. If we assume that women were twice as likely to be included in the study compared to men, the unexpected finding disappeared by weighting these overrepresented women without deep vein thrombosis with ½ (i.e. one divided by the likelihood of inclusion in the study) in the multivariable regression model: OR 0.92; 95%CI 0.62-1.36.

The problem of selection bias can be extended to meta-analyses, which may suffer from publication bias and selective reporting biases. A nice example is the meta-analysis of Tandon et al.(10), which found that the presence of mutant p53 tumor suppressor gene is prognostic for disease-free survival (hazard ratio (HR) = 0.45; 95% CI 0.27-0.74) but not for overall survival (HR = 1.09; 95% CI 0.60-2.81) in patients presenting with squamous cell carcinoma arising from the oropharynx cavity. Of the total 6 studies included in the meta-analysis, all reported on the prognostic effect of p53 for overall survival, but only 3 for disease-free survival. Results on disease-free survival were only reported when deemed prognostic, thus there appears to be a selective availability of data here, leading to unexpected findings.(11) A possible solution to this problem is to perform a bivariate meta-analysis, which synthesizes both outcomes jointly and accounts for their correlation,(12) to reduce the impact of missing disease-free survival results in 3 studies by 'borrowing strength' from the available overall survival results. A bivariate meta-analysis gave similar overall survival conclusions but gave an updated summary hazard ratio for disease-free survival of 0.76 (95% CI: 0.40 to 1.42), indicating no significant evidence that p53 is prognostic for disease-free survival anymore. (11)

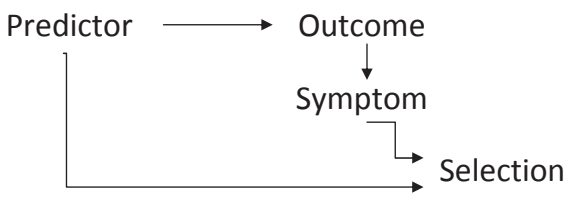
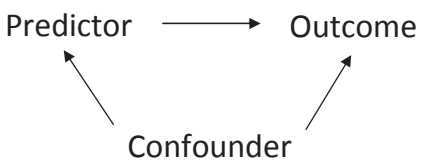
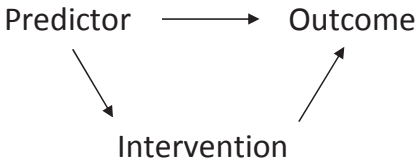
Selection bias 	If both the Predictor and the Outcome (possibly through a Symptom of the disease of interest) affect the probability of selecting a subject for a study, this may induce a selection bias
Mixing of effects (confounding) 	When a variable is related to both the Predictor and the Outcome, is not in the causal pathway and not included in the model, the observed predictor-outcome relation is the combined effect of the predictor and the omitted variable
Intervention effects 	The observed predictor-outcome relation becomes a combination of the direct effect of the Predictor on the Outcome and the indirect effect through the Intervention.

Figure 1 Directed acyclic graphs of scenarios that may result in unexpected findings in prediction research. In the figure the mechanisms resulting in selection bias, mixing of effects and intervention effects are schematically shown in a directed acyclic graph together with a short explanation of the mechanism.

Mixing of effects (confounding)

When two causes of a disease are mutually related, the observed effect of one can be mixed up with the effect of the other. In causal research, this phenomenon is referred to as confounding. Similarly, if two factors are mutually related (e.g. smoking and alcohol consumption), and one is causal for outcome but the other is not, then exclusion of the causal factor would lead to the non-causal factor having a strong predictor effect unexpectedly. For example, omitting smoking from a prediction model for lung cancer, would lead to alcohol consumption unexpectantly predicting lung cancer risk simply because it is confounded by smoking (those who smoke more tend to drink more). Since in prognostic research the interest lies in the joint predictive accuracy of multiple predictors, confounding is usually not deemed relevant.(13) However, the mechanism is the same in both descriptive and causal research: when omitting from the model a variable that is related to both an included predictor and the outcome, the observed predictor-outcome relation is the combined effect of the included predictor and the omitted variable (Figure 1). Consequently, an unexpected

finding of the predictor-outcome relation can be observed. The potential for mixing of effects is design and population specific. Hence, mixing of effects likely affects generalization of the model to other populations. An obvious solution would be to include the variable that was initially omitted from the prediction model. Clearly, this is impossible when the variable is not observed.

Mixing of effects was observed in the example of predicting anemia in whole blood donors (example 3). A lower risk of low hemoglobin levels was found when the number of whole blood donations in the past two years increased (OR = 0.92; 95% CI 0.90-0.93), also known as the 'healthy donor effect', which was corrected by including the recent history of hemoglobin level to the model (OR = 1.00; 95% CI 0.98-1.02). The most recent historic value of hemoglobin level is related to both the current hemoglobin level and the current risk of anemia, thus being a confounder.

Intervention effect

Predictor values may guide the decision to initiate a medical intervention. If effective, this intervention then lowers the probability of the outcome, thus attenuating the observed predictor-outcome relation. Similar to the mixing of effects the overall observed relation is a combination of the direct effect of the predictor on the outcome and the indirect effect through the intervention (Figure 1). However, expectations of the direction of the predictor-outcome relation apply to the direct effect. Theoretically, the overall observed predictor-outcome relation could even be the reverse of the direct effect between predictor and outcome (in case of an extremely effective intervention), thereby leading to an unexpected finding. Without further consideration, it seems unlikely that an intervention reduces the risk of the outcome to a level that is even lower than observed in the reference group of women who didn't have the indication and therefore didn't receive the intervention.

The solution to deal with an unexpected finding due to an intervention effect would be to include the intervention in the prediction model.(14) Obviously, this is not possible if everyone in the study has the same intervention. In that case it is likely that the unexpected finding actually has another cause than an intervention effect. Note that if an intervention is equally effective in all patients, modeling the intervention effect doesn't require an interaction between predictor and intervention in the model. If the intervention is more effective in, for example, those having the predictor, then an interaction between intervention and predictor is required (see below for discussion on heterogeneity).

In the prediction of metabolic acidosis in neonates (example 1) there could be an intervention effect present due to caesarean section. An unexpected finding was observed for the relation between intrapartum fever and metabolic acidosis (OR = 0.86; 95% CI 0.68-1.08). Upon inclusion of caesarean section in the model, intrapartum fever was positively related to metabolic acidosis: OR 1.08; 95%CI 0.86-1.34, which was in line with expectations.

Heterogeneity

The effect of a predictor may differ across subgroups of patients. This is referred to as a differential predictor effect, interaction, effect modification or heterogeneity of the predictor. When heterogeneity is not accounted for in the prediction model, the observed predictor effect is a (weighted) average of predictor effects within the different subgroups. If the predictor-outcome relations across subgroups are opposite, the direction of the observed relation depends on the proportional contributions of the subgroups. Expectations are likely based on the majority subgroup in a typical patient population, which is not necessarily the majority in the study population. Hence, heterogeneity of a predictor effect can lead to unexpected findings, if not accounted for. This differs from selection bias, in that the relative size of the subgroups is not related to both predictor and outcome. The principle solution to deal with heterogeneity is to include an interaction term in the model.

Heterogeneity is actually an unlikely cause for an unexpected finding in prediction models. Firstly, heterogeneity that results in genuinely opposite direction of effects is rare in epidemiology. Secondly, it seems unrealistic to assume that the group of patients who are typically the majority represent only the minority of patients in a specific study population.

In the prognostic model of metabolic acidosis in neonates (example 1), the effect of intrapartum fever on metabolic acidosis (OR 0.86; 95%CI 0.68-1.08) was unexpected. Alongside the impact of misclassifying how fever is defined (see discussion above), this unexpected finding could also have been the result of an interaction between intrapartum fever and epidural analgesia: OR 0.47 (95%CI 0.35-0.64) for women who received epidural analgesia vs. OR 3.16 (95% CI 2.16-4.64) for women without epidural analgesia.

Discussion

A first step in evaluating the validity of a clinical prediction model is to check whether the direction of the predictors-outcome relation are as expected. We identified six causes for unexpected findings: chance, misclassification, selection, confounding, intervention effects and heterogeneity. Obviously the aforementioned causes for unexpected findings can occur simultaneously. In that case finding the reasons for the unexpected finding will become more complicated, yet the solutions described still hold and can be applied simultaneously.

The major problem of an unexpected finding in prediction research is that it may hamper the generalizability of a prediction model. Even though the performance of the model may be good in the population in which the model was developed, it will probably be weaker when applied to a different setting/population, indicating poor generalizability. Hence, despite high methodological standards used in the development of the model it will not (i.e. not without further adjustments(3;22)) be applicable outside the population in which it was developed. It is therefore of utmost importance to signal unexpected findings. Obviously,

when the direction of a predictor-outcome relation is well-established in both literature and clinical experience, it is easy to identify unexpected (or incorrect) findings. Things become complicated when there is no pre-existing knowledge and it is unknown what direction is to be expected. Then, one has to make assumptions on the relation and therefore it is called an unexpected finding rather than an incorrect finding. A more subtle unexpected finding occurs when the direction of a predictor-outcome association is as expected but the magnitude of the effect is larger or smaller than expected. Still, the proposed solutions of this study could be used to solve this problem.

The examples also show that unexpected findings in prediction research are not only theoretically challenging, but are a phenomenon that can actually occur in any field of prediction research. When confronted with an unexpected finding, one should evaluate the different reasons for an unexpected finding (chance, misclassification, selection bias, mixing of effects, heterogeneity effects, intervention effects or a combination of these). Directed acyclic graphs like Figure 1 may help to identify possible causes. As mentioned before, heterogeneity or an intervention effect rarely results in unexpected findings, but the clinical examples illustrate that it can occasionally happen. When an unexpected finding is observed, it is more likely that it results from chance, misclassification, selection bias, or mixing of effects.

The potential for unexpected findings may differ between study designs. For example, incorrectly conducted case-control studies may be more prone to selection bias than cohort studies. Furthermore, mixing of effects becomes more likely when using (retrospective) routinely collected health care registry data, in which the number and detail of observed patient characteristics is typically limited.

It must be noted that in multivariable prediction models the problem of unexpected findings is likely to be smaller than in the univariable examples shown in this paper, because mixing of effects and intervention effects are accounted for by adding the appropriate covariates to the model.

In conclusion, unexpected findings of the predictor-outcome relation can occur in any kind of prediction research, and likely hamper generalizability and potential uptake of the model for clinical use. Researchers are encouraged to give explanations for possible unexpected findings in their prediction model, including the causes as well as the attempts undertaken to solve the problem, using the proposed framework for causes and solutions for unexpected findings in prediction research.

References

(1) D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008 Feb 12;117(6):743-53.

(2) Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

(3) Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005 Jul;94(1):200-5.

(4) Baart AM, de Kort WL, Atsma F, Moons KG, Vergouwe Y. Development and validation of a prediction model for low hemoglobin deferral in a large cohort of whole blood donors. Transfusion 2012 Apr 23.

(5) Janssen KJ, Donders AR, Harrell FE, Jr., Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010 Jul;63(7):721-7.

(6) Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society Series A-Statistics in Society 1999;162:71-94.

(7) Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol 2006 Aug;35(4):1074-81.

(8) de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for differential-verification bias in diagnostic-accuracy studies: a Bayesian approach. Epidemiology 2011 Mar;22(2):234-41.

(9) Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004 Sep;15(5):615-25.

(10) Tandon S, Tudur-Smith C, Riley RD, Boyd MT, Jones TM. A systematic review of p53 as a prognostic factor of survival in squamous cell carcinoma of the four main anatomical subsites of the head and neck. Cancer Epidemiol Biomarkers Prev 2010 Feb;19(2):574-87.

(11) Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. Stat Med 2011 Jan 26.

(12) Riley R. Multivariate meta-analysis: the effect of ignoring within-study correlation. JRSS Series A 2009;172(4):789-811.

(13) Brotman DJ, Walker E, Lauer MS, O'Brien RG. In search of fewer independent risk factors. Arch Intern Med 2005 Jan 24;165(2):138-45.

(14) Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? Anesth Analg 2005 May;100(5):1525-32, table.

Chapter 4.2

Conduct of subgroup analyses in meta-analysis triggered by previous (false-)significant findings: problems and solutions

Schuit E, Roes KCB, Mol BWJ, Kwee A, Moons KGM, Groenwold RHH.

Submitted

Abstract

Background: The final step to assess effectiveness of a treatment, overall or subgroups effects, is to perform a meta-analysis of multiple studies. Meta-analyses of subgroup effects are typically triggered by a significant finding in a previously conducted study; a previous finding that is potentially false-significant. We studied statistical analysis for meta-analysis investigating subgroup effects when studies that triggered the meta-analysis are included in the meta-analysis and investigated potential solutions.

Methods: We analytically assessed the bias of the estimates obtained by a meta-analysis, conditional on the number of included studies, the number of included false-significant studies, and the sample size of the included studies. The type I error rate and power of the meta-analysis were assessed using simulations. Bias-correction was performed by subtracting the analytically derived bias from the meta-analysis effect estimate.

Results: Bias and type I error rate depended on the number of false-significant studies included for meta-analysis. When 20% of the included studies were false-significant, the bias of the meta-analysis was 0.08 (z-score, instead of 0) and type I error was 20% (instead of 5%). Bias-correction yielded a type I error rate of 5%. In case a treatment was effective, yet no significant studies were included in a meta-analysis of ten studies, the power of the meta-analysis was close to 50%.

Conclusion: False-significant subgroup effects may trigger a meta-analysis, yet inclusion of these false-significant studies yields biased effect estimate and inflated type-I error rates. This bias can be adjusted for using a simple correction factor.

Introduction

Meta-analysis, either using aggregate or individual participant data, involves the synthesis of results or data from several studies and is considered the best design to determine the effects of particular exposures or interventions.(1;2) Irrespective of whether aggregate or individual participant data are used for meta-analysis, meta-analyses do not stand alone; i.e., they are often the final analysis of a period of accumulating scientific evidence, either investigating overall or subgroup effects.

It obviously makes sense to combine the information obtained from multiple studies, by conducting a meta-analysis, since the statistical power of a meta-analysis will be higher than those of individual studies. The choice to initiate a meta-analysis may be based on an observed effect or trend towards an effect in a single study. We distinguish three relevant reasons to conduct a meta-analysis of medical treatments, namely an interest in main effects, an interest in an interaction between treatment and a certain patient characteristic, and an interest in safety.

Randomized studies are designed to show an effect of the treatment on the main or primary outcome. Individual studies are not designed to show interactions or effects on safety endpoints. Consequently, an observed effect found for a secondary outcome, a safety outcome, or a subgroup can be due to chance, yet it may trigger the conduct of a meta-analysis. This may have important consequences for the meta-analysis when the trigger study is part of the meta-analysis.

In this paper, we discuss the implication of the practice of conducting a meta-analysis that was triggered by results of one or more individual studies that are also included in that meta-analysis. First, we will derive analytically what the impact is of the inclusion of studies with false-significant results in a meta-analysis on the bias in the effect estimate from the meta-analysis. Then, using simulations, we will assess the impact of this conduct on the type I error rate (i.e. rate of false-significant studies), and the power (i.e. rate of true-significant studies) of meta-analysis. Moreover, we suggest a simple correction method to adjust for potentially inflated type I error rates and bias, using the analytical bias under a null-effect. In this study we focus on meta-analyses that investigate treatment effects for which the individual studies were not designed, more specifically for subgroups, i.e. meta-analyses based on randomized clinical trials.

Bias in meta-analysis

Meta-analysis

In meta-analysis, effect estimates from individual studies are pooled; these estimates can relate to main effects, subgroup effects, or effects within subgroups. In this manuscript we

focus on subgroup effects, which are differential treatment effects between subgroups of study participants. Alternatively, one can think of pooling effect estimates obtained within different subgroups (i.e., without looking at the difference in estimates between subgroups). Throughout the following, we will use the term effect, to indicate a certain treatment effect, which can be either of these effects. We stress, however, that a trigger for a meta-analysis will often be a significant finding of a subgroup effect or an effect in a subgroup in one of the included studies, rather than the main effect.

Suppose that five trials have been conducted in which subgroup effects of a certain treatment were assessed (i.e., a differential treatment effect in the subgroups, expressed for example by an interaction term in a regression model). Then, under the assumption of a null-effect for the subgroup, any non-significant study included in the meta-analysis is correctly non-significant, whereas significant ones are false-significant. By chance, one of the five trials may show a false-significant association between treatment and outcome for the subgroup of interest. Since the results of the five trials are inconsistent, this may lead to a meta-analysis, in order to pool all available information to obtain a final answer to the question about the effects of the treatment for the subgroups.

The overall estimate obtained in a meta-analysis usually is a weighted average of the effect estimates of the individual studies, both significant and non-significant studies included in the meta-analysis. An overall effect estimate can be calculated by:(3)

$$effect = \frac{\sum_i w_i \hat{E}_i}{\sum_i w_i} \tag{1}$$

With w_i the weight that represents the contribution of study *i*, which is equal to the inverse of the within study variance (for a fixed effects model), or equal to the inverse of the within and between study variance (for a random effects $\hat{E}_i$ model), and the effect estimate in study *i*.

If subgroups of studies included in a meta-analysis have approximately the same size and the treatment has no effect, the combined effect of the non-significant studies will be relatively close to zero, since those studies did not show a significant effect. In contrast, the combined effect of the false-significant (i.e. under the assumption of a null-effect) studies will be relatively far from zero. Suppose, a treatment under study has no effect on the outcome for a certain subgroup, yet a meta-analysis was triggered by a false-significant result in one study while there were five studies included in total. In that case, the overall estimate in the meta-analysis is a weighted average of an effect that is far from zero (one false-significant study) and four effects that are close to zero (four non-significant studies): the estimate in the meta-analysis then likely differs from zero. Obviously, the deviation from zero depends on both the effect size and the sample sizes (i.e. of the subgroup of interest) of the significant and non-significant studies. Since the true effect of the treatment is a zero effect, the estimate of the

meta-analysis is a biased estimate of the true treatment effect. On average, in case of endless repetition of meta-analyses (some including significant studies, others not), one still expects to observe a subgroup effect of zero, i.e., no bias.

Analytical derivation of bias

We consider a continuous outcome and the subgroup treatment effect is defined as the difference between the treatment effects between the subgroups, i.e. the difference in the difference of the mean outcome value of the two treatment groups (A and B). In the following, we assume that the true subgroup treatment effect equals zero.

The amount of bias in the effect estimate in the meta-analysis can be derived by analysis, and is defined as the average difference between the estimated effect in the meta-analysis (equation 1) and the true treatment effect. In case of no subgroup effect, the mean effect of a meta-analysis is expected to be zero, and the test-statistic of the effect follows a z-distribution. The effect estimate obtained from a meta-analysis is a weighted average of the effects from the included studies, which include both significant and non-significant studies. In the significant studies, the subgroup effect should exceed a certain value, in order to achieve significance. Similarly, in the non-significant studies, the subgroup effect will not exceed this value. Based on a known boundary of significance (e.g. 0.05 significance level), and under the assumption of a null-effect, the expected means of the significant and non-significant studies can be derived from a truncated normal distribution (Figure 1).(4) In case of a standard normal distribution and a one-sided test at a 0.05 significance level the significant effects will lie in the rejection region of a normal distribution that is truncated at a z-score of 1.65, whereas the non-significant effects will lie in the non-rejection region of that truncated normal distribution. The mean effects of the non-significant studies ($\bar{z}_{ns}$) and the significant studies ($\bar{z}_{s}$) can be calculated as described by Barr and Sherrill(4) and are under a null effect equal to-0.108 and 2.07, respectively, for a one-sided significance level of 0.05. Suppose that under a null effect a meta-analysis was based on the significant results in one study (i.e. false-significant) while there are five studies included in total, then the overall effect estimate obtained by meta-analysis is expected to be (4 * (-0.108) + 1 * (2.07)) / 5 = 0.33. Notice that this overall expectation of the subgroup effect estimate differs from zero and is thus biased since the true subgroup effect was zero. In case of a two-sided statistical test, two significant studies may have opposite effects which may cancel out. Therefore the above approach is only applicable to a one-sided test. We notice that most meta-analysis apply two-sided statistical tests. We come back to this issue in the discussion of this paper.

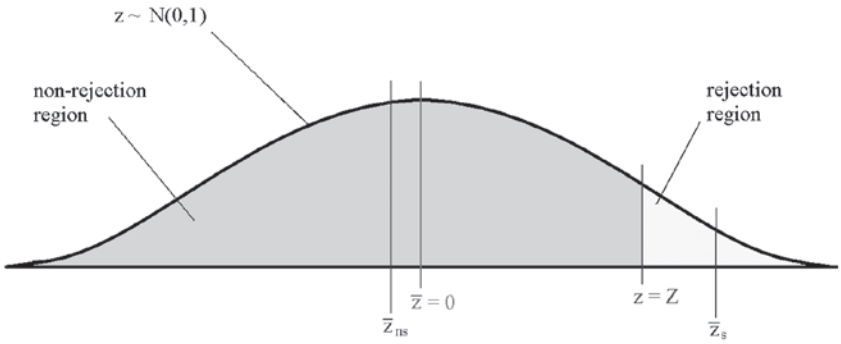


Figure 1 Truncated standard normal distribution. The overall mean of z, $\bar{z}$ is zero. Truncation at z=Z splits up the distribution into a rejection region and non-rejection region. These regions have a mean of $\bar{z}_s$ and $\bar{z}_{ns}$, respectively.(4) For Z = 1.65, $\bar{z}_{ns}$ = 0.108 and $\bar{z}_s$ = 2.07.

Apart from the mean subgroup effects of the significant and non-significant studies, the size of this bias depends on the number of small and large studies, which are significant or non-significant. Under a null-effect the effect estimate is equal to the bias and can under a fixed effects approach be derived from equation 1 (Appendix A):

$$effect = \frac{\sqrt{2n_s}\left(N_{s_{ns}}\,\bar{z}_{ns} + N_{s_s}\,\bar{z}_s\right) + \sqrt{2n_l}\left(N_{l_{ns}}\,\bar{z}_{ns} + N_{l_s}\,\bar{z}_s\right)}{N_s n_s + N_l n_l} \tag{2}$$

With $N_{s_{ns}}$ the number of non-significant small studies, $N_{l_{ns}}$ the number of non-significant large studies, N_{s_s} the number of significant small studies, N_{l_s} the number of significant large studies, n_s the sample size (i.e. of the subgroup of interest) of the small studies, n_l the sample size (i.e. of the subgroup of interest) of the large studies, $\bar{z}_{ns}$ the mean difference between treatment A and B in the non-significant studies (i.e. -0.108, for a one-sided 0.05 significance level), $\bar{z}_s$ the mean difference between treatment A and B in the significant studies (i.e., 2.07 for a one-sided 0.05 significance level).

Application of equation 2 (i.e., under the assumption of no difference in the mean outcome value between treatment A and B) showed that the bias in the overall subgroup effect obtained in the meta-analysis increased with an increasing number of false-significant studies included in the meta-analysis (Figure 2). When none of the total 10 studies showed a false-significant result the bias was negative, while the bias in the overall estimate was positive when one or more studies had a false-significant result. The upper bound of the bias was found when all of the significant studies were large (i.e. none of the small studies showed a false-significant effect), while the lower bound of the bias was found when all false-significant studies were small (i.e. none of the small studies showed a false-significant effect). For other combinations, e.g. one large false-significant study and an increasing number of false-significant small studies, the bias of the subgroup effect estimate from the meta-analysis was found to be in-between the extremes (data not shown).

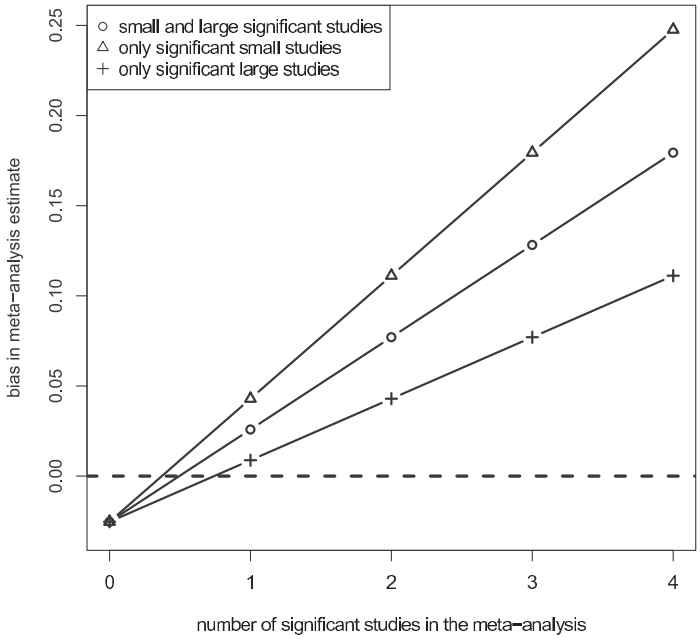


Figure 2 Bias in effect estimates from meta-analysis due to inclusion of false-positive studies. The dashed horizontal line indicates the bias over all simulated meta-analyses, which is equal to zero. The triangles indicate the bias in relation to the number of false-significant studies (we assumed no effect) in the meta-analysis when only large studies showed a false-significant subgroup effect. The crosses show the bias when all false-significant studies were small.

Simulation studies of statistical inference in meta-analysis

Simulation set-up

We set up a simulation study to assess the performance of statistical analysis in meta-analysis, i.e. the type I error rate (i.e. incorrect rejection of the null-hypothesis) and power (rate of correct rejection of the null-hypothesis). We considered a continuous outcome, measured in individuals allocated to treatment A and treatment B. To evaluate type I error rates, we simulated a meta-analysis with 10 studies investigating subgroup effects, in which a continuous outcome of individuals allocated to treatment A or B were drawn from a standard normal distribution. First, subgroup treatment effect (i.e. the mean difference in the difference of the mean outcome value between treatment A and B) and its significance (based on a one-sided two-sample t-test) were estimated in each individual study. Next, data on all participants in the 10 studies were combined and the mean subgroup effect estimate and its significance were estimated on the meta-analysis level. The overall mean effect in the meta-analysis was estimated by a fixed as well as a random effects model. Effect estimates

from individual studies were weighted by the inverse of the variance within studies in a fixed effects approach, and by the inverse of the variance within and between (τ^2) studies in the random effects approach.(3) In a separate scenario, we investigated the statistical power. Here, instead of a true effect of zero, we simulated a 0.2 difference in the continuous outcome between treatment A and B, by sampling outcomes from a standard normal distribution for treatment A (i.e., mean 0, variance 1) and from a normal distribution with a mean 0.2 and a variance 1 for treatment B. Again, both a fixed and random effects model were applied.
Under the assumption of a difference in effect of 0.2 between treatment A and B, a variance of 1, a desired power of 80% and a type I error of 5% we needed approximately 330 individuals in each arm of the meta-analysis. The individuals were divided over 10 studies, of which 5 were considered large (N=52 per treatment arm) and 5 considered small (N=13 per treatment arm). Since the sample size (i.e. of the subgroup of interest) was based on the meta-analysis all individual studies had a power of less than 80%, which is common when investigating subgroup effects, i.e. individual studies are generally designed to investigate a treatment effect on the primary outcome in the whole study population, not in a specific subgroup. Also in the above mentioned scenario in which data were simulated under a true treatment effect of zero, these sample sizes were used. It must be noted that in general power is not described for a meta-analysis since it is not used for hypothesis testing, rather than to estimate effects. However, the use of power in this specific study allows us to assess the precision of the overall meta-analysis.
The first scenario was simulated 10 000 000 times. This resulted in a standard error of the type I error rate of $6.9e^{-5}$ in case of four false-significant studies in a meta-analysis of 10 studies, i.e., averaged type I error rates have a 95% probability of being in the range of 4.86-5.14%. The standard error of the power with 10 000 000 repetitions was $1.3e^{-4}$, i.e. averaged power has a 95% probability of being in the range of 79.7-80.3%. The type I error rate was estimated by the mean number of false-significant meta-analyses. Since a true subgroup effect was simulated, the power was estimated as the mean number of true-significant meta-analyses. The proportion of small studies in the meta-analysis was varied from 10% to 80%. Furthermore, the ratio of the sample sizes (i.e. of the subgroup of interest) in the small and large studies was varied and ranged from larger studies that had a sample size that was equal, or 2, 4, or 8 times larger than in the small studies.

Bias-correction

A possible solution to correct for a potentially inflated type I error rate and inadequate power is to apply a bias-correction based on the analytical bias, which we derived above. To evaluate this approach we subtracted the analytical bias (calculated using equation 2) from the simulated effect estimate in the meta-analyses. Next, we determined the type I error rate and power again using this corrected estimate. All statistical analyses were conducted in R for Windows, version 2.15.2.(5)

Results of simulations

The results of the simulation study with no difference in subgroup treatment effect between treatment A and B and the use of a fixed effects model showed that the type I error rate of the overall meta-analysis increased when the number of studies with a false-significant subgroup effect included in the meta-analysis increased (Figure 3). The most extreme type I error rates were found when only large studies were false-significant (highest) or when only small studies included in the meta-analysis showed a false-significant effect (lowest). Other combinations of false-significant small and large studies resulted in a type I error rate of the overall meta-analysis that was in-between these boundaries (data not shown). After the bias-correction, the type I error rate of the meta-analysis was below 5%, independent of the number of studies with a false-significant subgroup effect (Figure 3). As expected, when using a random effects model, the type I error rate was lower than seen for a fixed effects model, but was still substantial (data not shown). When only the small studies were false-significant the type I error rate for a meta-analysis using a random effects model showed to be similar to the fixed effects model, as expected.

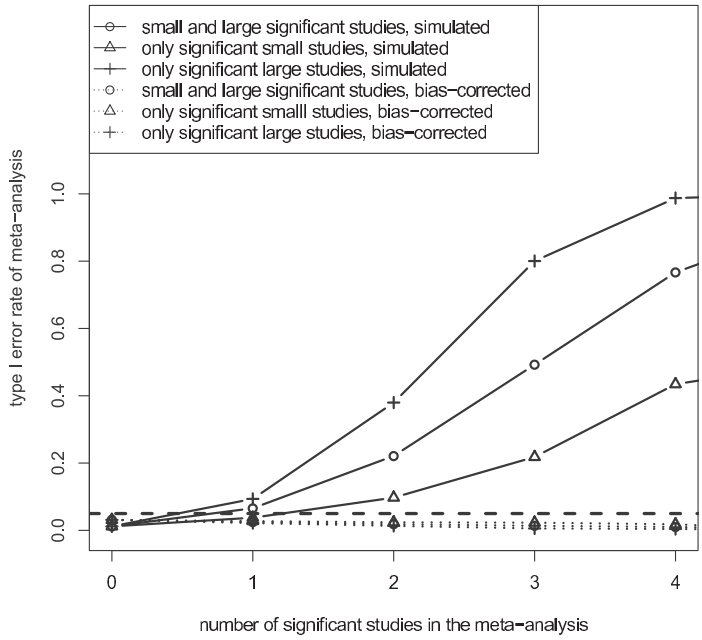


Figure 3 Type I error rate in meta-analysis in relation to the number of false-significant studies (we assumed no subgroup effect) included in the meta-analysis. The dashed horizontal line indicates the type I error of all simulated meta-analyses together, which is equal to 5%. The solid lines show the type I error rate obtained by simulation and the dotted lines indicate the type I error rate after bias-correction. The triangles indicate the type I error rate in relation to the number of false-significant studies in the meta-analysis when all false-significant studies were large. The crosses show the type I error rate when all false-significant studies were small.

Under a simulated treatment effect of a mean difference of 0.2 between treatment A and B and using a fixed effects model, the power of the meta-analysis as expected increased when the number of studies with a true-significant subgroup effect in the meta-analysis increased (Figure 4). When there were no true-significant studies included in the meta-analysis the power of the meta-analysis was approximately 50%. This increased to about 70% with inclusion of one true-significant study, and approached 90% when two or more true-significant studies were included. The extremes of the power of the meta-analysis were found when only large studies were true-significant (highest) or when only small studies included in the meta-analysis showed a true-significant effect (lowest). The bias-correction (i.e., subtracting the amount of bias under the null hypothesis from the effect estimate in the meta-analysis) did not affect the power of the meta-analysis if a non-zero treatment effect was present, since the simulated effect (0.2) was substantially higher than the effect what was corrected for (0). As expected, the random effects model resulted in less power of the meta-analysis as compared to the fixed effects model for all possible combinations of true-significant and false-negative small and large studies (data not shown).

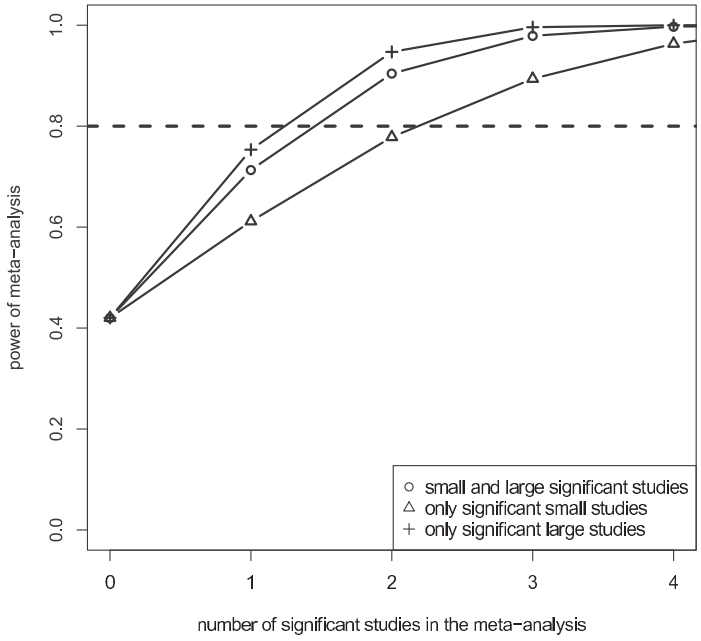


Figure 4 Power in the meta-analysis in relation to the number of true-significant studies (we assumed a subgroup effect) included in the meta-analysis. The dashed horizontal line indicates the power of all simulated meta-analyses together, which is equal to 80%. The triangles indicate the power in relation to the number of true-significant studies in the meta-analysis when all true-significant studies were large. The crosses show the power when all true-significant studies were small.

The power of the overall meta-analysis was only marginally influenced by the proportion of small studies (varied from 10-80% of the total number of included studies) and the ratio of the sample size (i.e. of the subgroup of interest) between the small and large studies (sample size in large studies was 2, 4 or 8 times larger than in the small studies) in the meta-analysis (i.e. similar to solid line with circles in Figure 4).

Discussion

We showed that inclusion of studies with false-significant subgroup effects in a meta-analysis may bias the subgroup effect estimates from that meta-analysis and consequently the type I error rates will not meet the prespecified nominal level, particularly when the number of false-significant studies included in a meta-analysis increases. Using a simple correction based on the number of included studies, the number of included significant studies, and the sample sizes of the subgroups of interest of the included studies, the nominal type I error rate can be controlled without losing power of the meta-analysis in the event of a non-zero treatment effect.

There are several findings from our study that need further elaboration. First, a negative bias was found for the situation in which none of the included studies showed a false-significant subgroup effect (i.e. under the assumption of no treatment effect) (Figure 2). This is a result of applying a one-sided test; the average effect size of the meta-analysis may be negative as a result of the fact that the majority of the non-significant studies included in the meta-analysis showed a non-significant negative effect.

Second, if there were no studies with a true-significant subgroup effect included in the meta-analysis (i.e., under the assumption of a non-zero treatment effect), the power of the meta-analysis was close to 50% (Figure 4), which is substantially lower than 80%. This is remarkable because researchers often combine several non-significant studies in order find a significant effect in meta-analysis. When none of the individual studies included in the meta-analysis shows a significant effect, even though the true (simulated) effect is e.g. 0.2 (as in our example), the pooled effect in the meta-analysis will consequently be lower than 0.2. As a result, the statistical power to detect an effect will go down. The distribution of standardized effect sizes (z-scores) of adequately sized meta-analyses (type I error rate of 5% and power of 80%) will approximately correspond to a normal distribution of mean 2.83 and standard deviation 1. However, the distribution of standardized effect sizes in the selected group of meta-analyses that only include non-significant trials is shifted: mean 1.61 and standard deviation 1. Since the mean of the latter distribution is close to a test statistic of 1.65, approximately half of all meta-analyses without any significant studies will have a significant result, i.e. the power of such a meta-analysis is close to 50%.

Third, the power of the meta-analysis was only marginally influenced by the bias-correction. The bias-correction was calculated based on the null-hypothesis that there was no overall effect, however, to simulate the power we used an effect estimate of 0.2. Consequently, the bias used for the correction of the effect size in the meta-analysis is low compared to the actual simulated effect size. Hence, the power is not influenced by the bias-correction. Ideally, one would use a bias-correction based on the 'true' effect. Since this true treatment effect is never known, we used a bias-correction based on the null hypothesis.

Fourth, the type I error rates and power were lower when applying a random effects model than using a fixed effects model. This was expected since a random effects model gives, under the simulation of fixed effects, less efficient estimates. In a random effects model the weights of the studies with a higher precision (i.e. the studies with larger subgroup sample sizes) are lower compared to a fixed effects model. Consequently, the influence of the larger studies on the effect estimates in the meta-analysis is smaller and therefore the effect estimate of the difference between treatment A and B in the meta-analysis will be lower and less often significant. This results in a lower type I error rates as well as a lower power.

Our simulations show that a bias-correction is an effective way to maintain the nominal type I error rate. We determined this bias-correction analytically assuming a z-distribution of the ratio of the effect estimate and its standard error of the studies included in the meta-analysis. Consequently, this approach for analytical bias may be representative for other statistics that follow a z-distribution or a chi-square distribution, such as the log(odds ratio). Hence, the results of this study are generally applicable, but not automatically generalisable to a dichotomous situation.

To appreciate the present results, a few potential limitations need to be addressed. As with every simulation study assumptions have to be made. In this study we focused on two scenarios of which one simulated a treatment difference of 0.2 on a continuous scale between both treatment arms, in which sample sizes were equal in both arms and in which the meta-analysis included 10 studies. Obviously, results may be different for large (dichotomous) treatment effects, unbalanced treatment arms, and substantially less or more studies included in the meta-analysis; these topics were beyond the scope of this study.

Findings are often interpreted dichotomous, i.e. there either is an effect or not. Consequently, the smallest finding may lead to the conduct of a meta-analysis. We may distinguish three relevant triggers to start a meta-analysis of medical treatments. The first are intended effects in which one or more consecutive studies found a significant treatment effect and other studies found an opposite effect, or when studies found a trend towards a treatment effect. Based on these positive studies or to increase the power (e.g. when none of the individual studies found an effect) it is then decided to conduct a meta-analysis. The second trigger is an interaction effect between the treatment and a certain subgroup, e.g. effect of statins on cholesterol levels in males and females. If in a single study no overall treatment effect was present,

the chance of a false-significant treatment-subgroup interaction increases. Consequently, individual studies show more false-significant treatment effects among subgroups and more meta-analyses are performed, which in turn lead to false-significant results in the overall meta-analyses. This problem can be avoided if the interaction is investigated only when the study showed an overall effect. The third domain of interest is safety. Opposed to main effects, several endpoints could be defined that are relevant for safety. Since multiple safety endpoints are investigated- and correction for multiple testing is often absent- the chance of finding a treatment effect in one of the studies increases. Consequently, a meta-analysis is set up. As mentioned before, for outcomes or analyses for which the study was initially not designed the risk of chance findings is particularly high due to multiple testing. This includes secondary outcomes, subgroup effects, and safety outcomes. Therefore, we believe that our findings will be generalisable to low incidence secondary outcomes and safety outcomes.

It seems unlikely that different meta-analyses will be conducted on the same topic using completely different data, which is distinct from individual studies that are often performed independently by different research groups. Moreover, a meta-analysis is always conditional on information from previous studies. Consequently, a frequentist approach is ambiguous. When investigating subgroup effects in a meta-analysis the most obvious method to correct for this would be to correct for multiplicity. A general solution that we showed in this study is to apply a bias-correction. The alternative is performing a validation study. The latter simply means that the treatment effect is investigated in an independent population to check whether the positive finding can be reproduced. However, in a meta-analysis this seems rather difficult since generally all available information will be used in the meta-analysis.

We conclude that in the situation in which the results from one or more studies investigating subgroups triggered the initiation of a meta-analysis, inclusion of false-significant studies in the meta-analysis will result in a bias and inflation of the type-I error rate. This bias may be substantial, but can be adjusted for using a bias-correction.

References

(1) Riley RD, Lambert PC, bo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.

(2) Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials 2005;2(3):209-17.

(3) DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-88.

(4) Barr DR, Sherrill ET. Mean and variance of truncated normal distributions. The American Statistician 1999;53(4):357-61.

(5) R: A Language and Environment for Statistical Computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2012.

Appendix A

The overall effect observed in a meta-analysis is a weighted average of the effects in the included studies:

$$effect = \frac{\sum_i w_i \hat{E}_i}{\sum_i w_i} \quad (1)$$

With w_i the weights that represent the contribution of each study and are equal to the inverse of the within study variance (fixed effects model) or equal to the inverse of the inverse of the within and between study variance (random effects model), and $\hat{E}_i$ the effect estimate in study *i*.

We consider a continuous outcome under a fixed effects model and given that we can have studies with significant and non-significant study results, equation 1 can be rewritten as:

$$effect = \frac{\frac{1}{\text{var}(\bar{x}_s)}\bar{x}_s + \frac{1}{\text{var}(\bar{x}_{ns})}\bar{x}_{ns}}{\frac{1}{\text{var}(\bar{x}_s)} + \frac{1}{\text{var}(\bar{x}_{ns})}} \quad (2)$$

With $\bar{x}_{ns}$ the mean effect in the non-significant studies which is defined as $\bar{x}_{ns} = \bar{z}_{ns}\, sd(x_{ns})$, $\bar{x}_s$ the mean effect in the significant studies defined as $\bar{x}_s = \bar{z}_s\, sd(x_s)$. $\bar{z}_s$ and $\bar{z}_{ns}$ are the expected means of the significant and non-significant studies in case of a truncated standard normal distribution and a one-sided test a 5% significance level (Figure 1). $\text{var}(x_s)=2/N_s$ and $\text{var}(x_{ns})=2/N_{ns}$, with N_s and N_{ns} as the sample size (i.e. of the subgroup of interest) of the significant and non-significant studies. Using these definitions the equation can be further simplified to:

$$effect = \frac{\bar{z}_s\sqrt{2n_s} + \bar{z}_{ns}\sqrt{2n_{ns}}}{n_s + n_{ns}} \quad (3)$$

This equation can be further specified by dividing the significant and non-significant studies into small and large studies:

$$effect = \frac{\sqrt{2n_s}\left(N_{s_{ns}}\bar{z}_{ns} + N_{s_s}\bar{z}_s\right) + \sqrt{2n_l}\left(N_{l_{ns}}\bar{z}_{ns} + N_{l_s}\bar{z}_s\right)}{N_s n_s + N_l n_l} \quad (4)$$

With $N_{s_{ns}}$ the number of non-significant small studies, $N_{l_{ns}}$ the number of non-significant large studies, N_{s_s} the number of significant small studies, N_{l_s} the number of significant large studies, n_s the sample size (i.e. of the subgroup of interest) of the small studies, N_l the sample size (i.e. of the subgroup of interest) of the large studies.

Chapter 5

Summary and general discussion

Summary and general discussion

The goal of this thesis was threefold: 1.) to develop and validate prognostic prediction models to allow for individual risk assessment; 2.) to investigate treatment strategies and identify groups of patients who may benefit from specific treatment(s) strategies using subgroup analysis within meta-analyses of individual patient data; and 3.) to evaluate methods currently used in prediction modelling, and in the assessment of subgroup effects. With the studies presented in this thesis we aimed to enhance risk-based management in obstetrics. Below is given a summary of the studies included in this thesis.

Development and validation of prognostic prediction models for individual risk assessment in obstetrics

In **Chapter 2** of this thesis we focused on prognosis in obstetrics, specifically on the identification of risk indicators for a referral during labour and the prediction of neonatal metabolic acidosis, operative delivery, and stillbirth and neonatal death of very preterm births.

In the Netherlands, pregnant women at low risk for complications have the choice to start labour either at home or in the hospital (primary care), supervised by independent midwives or general practitioners. Women are referred to and monitored by gynaecologists in the hospital (secondary care) when they are considered to have a high-risk pregnancy either at the start of or during pregnancy, during labour or after labour. Currently, risk indicators for referral during labour are based on complicating factors that arise during labour and indicators which are associated with adverse neonatal or maternal outcome(1;2), but do not include single or combined risk indicators that may be solely related to a referral during labour. Therefore, in **Chapter 2.1**, we aimed to identify risk indicators for referral during labour from primary to secondary care due to foetal distress, failure to progress in the 2nd stage of labour, meconium stained amniotic fluid, failure to progress in 1st stage of labour, wish for pain relief, or other reasons. These risk indicators were identified in Dutch women with a singleton term pregnancy, starting labour in primary care between 2000 and 2007, who were registered in a national perinatal registry. Out of 746,642 women with a singleton term pregnancy 241,595 (32%) were referred from primary to secondary care during labour. The strongest risk indicators for referral, independent of the reason, were nulliparity and a previous instrumental vaginal delivery. Additional risk indicators were maternal age, ethnicity, degree of urbanization, social economic status, planned home delivery, gestational age, neonatal gender, and birth weight. The identified risk indicators can support caregivers in primary care.

Perinatal asphyxia is associated with several short- and long-term complications, varying from mild hypoxic ischemic encephalopathy to cerebral palsy and death.(3-7) Therefore, it is important to identify women with a high(er) risk of developing adverse neonatal outcome

as early as possible, e.g. at entrance in the hospital or even antenatally. Clinical practice is always multivariable: no prognosis is set by a single factor.(8) However, foetal surveillance during labour with cardiotocography (CTG), foetal blood sampling (FBS) or ST-analysis of the foetal electrocardiogram (ECG) is often applied and interpreted in isolation, without taking into account other factors that may influence the probability of an adverse outcome at birth. Therefore, the aim of **Chapter 2.2** was to predict neonatal metabolic acidosis at birth using the best combination of antepartum (model 1) and additional intrapartum characteristics (model 2). The study population consisted of participants of a large multicentre trial in which labouring women with a high-risk vertex singleton pregnancy beyond 36 weeks of gestation were studied.(9;10) Of 5,667 singletons, 107 (1.9%) had metabolic acidosis. Antepartum predictors of metabolic acidosis were gestational age, nulliparity, previous caesarean section and maternal diabetes. Additional intrapartum predictors were spontaneous onset of labour and meconium stained amniotic fluid. Calibration and discrimination were acceptable for both models. In conclusion, in women with a high-risk singleton term pregnancy in cephalic presentation, we identified antepartum and intrapartum factors that predict neonatal metabolic acidosis at birth.

To show that prognostic models are valuable and generalizable, evidence is needed that the models perform well in groups of patients other than on which they were originally developed. (11) Clinicians may not trust the model and will not do so until the model is proven to be generalizable to a different population than on which the model was developed. Since local and timely variation in attitude to women at risk of giving birth to an infant with metabolic acidosis can influence the validity of the models developed in Chapter 3, in **Chapter 2.3** we applied the models in Swedish new-borns included in a randomized trial between 1998 – 2000.(12) The models were applied in 5,049 women in which 54 (1.1%) suffered from metabolic acidosis. After adjustment for differences in the incidence of metabolic acidosis between the Dutch and Swedish cohorts, the prognostic models showed good calibration and moderate overall discrimination. Updating of the models to enhance their predictive abilities seems indicated.

In the 1970s caesarean section rates began to rise in most Western countries(13) and continue to rise in most of these countries.(14) Most caesarean sections as well as instrumental vaginal deliveries are performed because of suspected foetal distress or failure to progress. Furthermore, in absolute numbers by far most interventions are performed in women with a singleton term pregnancy cephalic presentation.(15) Despite these numbers, it remains difficult to predict by which mode of delivery these women will actually deliver. Therefore, the aim of the study presented in **Chapter 2.4** was to identify which combination of factors that can be timely obtained, both antepartum (model 1) and antepartum plus intrapartum before or early during labour, best predicts the risk of either instrumental vaginal delivery or caesarean section with a suspected foetal distress or failure to progress indication. Data

were used from a large multicentre trial in which 5,667 labouring women with a high-risk vertex singleton pregnancy beyond 36 weeks of gestation were studied.(9) Based on the combination of the intervention and the indication for the intervention, women were assigned to one of the five distinctive outcome categories: spontaneous vaginal delivery (reference category); instrumental vaginal delivery due to suspected foetal distress; caesarean section due to suspected foetal distress; instrumental vaginal delivery due to failure to progress; or caesarean section due to failure to progress. 375 (6.6%) and 212 (3.6%) women had an instrumental vaginal delivery or caesarean section due to foetal distress, and 433 (7.6%) and 571 (10.1%) due to failure to progress, respectively. Predictors were age, parity, previous CS, diabetes, gestational age, gender, estimated birth weight (model 1) and induction of labour, oxytocin augmentation, intrapartum fever, prolonged rupture of membranes, meconium stained amniotic fluid, epidural anaesthesia, and use of ST-analysis (model 2). Both multinomial models showed excellent calibration and good discrimination.

In most developed countries very preterm birth has an incidence of 1-2%(16) but accounts for approximately one third to half of all perinatal deaths.(17;18) In the United Kingdom, two easy to apply survival graphs are used in clinical practice for discussions with parents for whom premature delivery is likely.(17;19;20) These graphs are derived from prognostic models that predict survival for very preterm infants known to be alive at the onset of labour and very preterm infants admitted for neonatal intensive care using readily available variables such as gestational age, birth weight, sex, ethnicity and multiplicity of pregnancy.(19) Although developed for a UK population, the models may also be useful in other Western countries. The goal of **Chapter 2.5** was to study the external validity of these models. We applied the models in all infants, with gestational age 22- 32 weeks and European ethnicity, known to be alive at the onset of labour (n = 17,582) and admitted for neonatal intensive care (n = 11,578), born in the Netherlands between 1 January 2000 and 31 December 2007. 16.7% (n=2,939) of the infants known to be alive at the onset of labour died during labour or within 28 days of birth, and 7.8% (n=908) of the infants admitted for neonatal intensive care died before discharge from the intensive care. The prognostic model for infants known to be alive at the onset of labour showed good calibration and excellent discrimination. The prognostic model for infants admitted for neonatal intensive care showed good calibration and good discrimination. We concluded that the results imply usage in clinical practice in the Netherlands and possibly other Western countries.

Identifying subgroups of patients in whom treatments are (more) effective

Chapter 3 of this thesis focused on individual participant data meta-analyses in obstetrics aimed at quantifying treatment effects of progestogen in twin pregnancies (**Chapters 3.1** and **3.2**) and ST-analysis in term singleton pregnancies in cephalic presentation (**Chapter 3.3**), in specific patient groups.

Chapters 3.1 and **3.2** are respectively the protocol and the conduct of an individual participant data meta-analysis on the effectiveness of progestogen treatment in the reduction of adverse perinatal outcome in twin pregnancies. These two studies are summarized in parallel hereafter. In twin pregnancies, the rates of stillbirth, neonatal death, preterm birth, (very) low birth weight, and subsequent long-term morbidity are substantially higher than in singletons (21), which is mainly due to preterm birth(22). Up to now no interventions to reduce preterm birth have been proven to be of benefit in twin pregnancies. Randomized clinical trials in singleton pregnancies have shown that antenatal progestogen therapy reduces the rate of preterm delivery in women who are at high risk of preterm birth due to a previous preterm delivery (23-26) or a short cervix (27-29). As a result, the effectiveness of progestogens in the reduction preterm birth has been extensively investigated in twin pregnancies. Recently, three meta-analyses have been published of which one used a small number of twins and did not investigate 17-OHPC(29), another had too few information to investigate relevant subgroups(30), and the third did not differentiate between 17-OHPC and vaginal progesterone(31). Furthermore, none of the three meta-analyses included all studies on progestogens in twins (29-31). Therefore, the aim of the study described in Chapters 7 and 8 was to perform an individual participant data meta-analysis to investigate the effects of progestogens, separately for intramuscular 17-OHPC and vaginally administered natural progesterone, in women with a twin pregnancy and in pre-specified subgroups based on chorionicity (mono- or di-), cervical length (< 25 mm or ≥ 25 mm), and a history of spontaneous preterm birth. Thirteen trials included 3,764 women and their 7,528 babies. Neither 17-OHPC nor vaginal progesterone reduced the incidence of adverse perinatal outcome. Women treated with 17-OHPC had higher numbers of delivery <32 weeks' gestation than those in the control group. No specific subgroups were found in which progestogens were more effective than placebo. We concluded that in unselected women with an uncomplicated twin gestation, treatment with progestogens, regardless of the type, does not improve perinatal outcome.

As described above, perinatal asphyxia is associated with several short- and long-term complications.(3;7;32) Foetal monitoring during delivery aims to identify foetuses at risk of asphyxia. A relatively new method for continuous foetal monitoring is the STAN® methodology (Neoventa Medical, Gothenburg, Sweden) in which (classification of) the cardiotocography is combined with ST analysis of the foetal electrocardiogram. Changes in the foetal electrocardiogram in combination with cardiotocography abnormalities could be an indication of foetal hypoxia, as shown in previous animal studies.(33;34) Five randomized controlled trials that investigated the effect of intrapartum foetal electrocardiogram monitoring on the reduction of neonatal metabolic acidosis were inconclusive with four studies showing no statistically significant effect.(9;35-37) In one study ST-analysis significantly reduced the incidence of neonatal metabolic acidosis.(38) Based on these studies, three meta-analyses

were performed using aggregated data.(39-41) The studies included in the aggregated data meta-analysis, however, did not report on all endpoints of interest and did not investigate subgroups. In view of the shortcomings of conventional meta-analyses using aggregated data, we presented in **Chapter 3.3** an individual participant data meta-analysis to investigate the effect of ST-analysis additional to cardiotocography on neonatal metabolic acidosis in intrapartum foetal monitoring. Four randomized trials included 12,987 labouring women with a term singleton in cephalic presentation. Additional ST-analysis did not result in a statistically significant reduction of the rate of metabolic acidosis as compared to the cardiotocography alone group, but resulted in a reduction in the frequency of instrumental vaginal deliveries and foetal blood samples. Subgroup analyses showed that cardiotocography with additional ST-analysis resulted in fewer admissions to a neonatal intensive care unit in women with a duration of pregnancy beyond 41 weeks. We concluded that intrapartum foetal monitoring with CTG+ST does not reduce the risk of metabolic acidosis, but does reduce the need for instrumental vaginal deliveries and foetal blood sampling.

Methods to assess differential treatment effects

Chapter 4 of this thesis focused on methodological issues related to the first two chapters of the thesis, i.e. development and validation of prognostic prediction models for individual risk assessment and identification of subgroups of patients in whom treatments are (more) effective.

Clinical prediction models aim to predict individual clinical outcomes using multiple predictor variables. Even though causality is irrelevant, clinical researchers often anticipate a certain direction (i.e., either positive of negative sign) of the relation between predictor and outcome. Thus, when observing a negative association (suggesting a protective effect) when expecting a positive association, this is unexpected and unlikely to be found in another population and could, consequently, hamper generalizability. Moreover, unexpected findings may suggest that the prediction model is actually invalid, thus lowering the face validity of the model.(8) The aim of **Chapter 4.1** was to discuss causes for unexpected findings in prediction research and to provide possible solutions. Possible causes for unexpected findings in prediction research include chance, misclassification, selection bias, mixing of effects (confounding), intervention effects, and heterogeneity. The type of design or analytical method to address an unexpected finding depends on the cause of the unexpected finding. Unexpected findings in prediction models should always be reported by researchers to improve the potential uptake and impact of the model, including the potential causes of the unexpected findings as well as the attempts undertaken to solve them.

Meta-analysis involves the synthesis of results or data from several studies and is considered the best design to determine the effects of particular exposures or interventions on

subsequent outcomes. The choice to initiate a meta-analysis may be based on an observed effect or trend towards an effect in a single study. We distinguish three relevant reasons to conduct a meta-analysis of medical treatments, namely an interest in main effects, an interest in an interaction between treatment and a certain patient characteristic, and an interest in safety. Randomized studies are designed to show an effect of the treatment on the main or primary outcome and are not designed to show interactions or effects on safety endpoints. Consequently, an observed effect found for a secondary outcome, a safety outcome, or a subgroup can be due to chance, yet it may trigger the conduct of a meta-analysis. This may have important consequences for the meta-analysis when the trigger study is part of the meta-analysis. In **Chapter 4.2** we investigated this potential problem, specifically for subgroup effects in a meta-analysis of randomized trials. We analytically assessed the bias of the estimates obtained by a meta-analysis, conditional on the number of included studies, the number of included false-significant studies, and the sample size of the included studies. The type I error rate and power of the meta-analysis were assessed using simulations. Bias-correction was performed by subtracting the analytically derived bias from the meta-analysis effect estimate. Bias and type I error rate depended on the number of false-significant studies included for meta-analysis. When 20% of the included studies were false-significant, the bias of the meta-analysis was 0.08 (z-score, instead of 0) and type I error was 20% (instead of 5%). Bias-correction yielded a type I error rate of 5%. In case a treatment was effective, yet no significant studies were included in a meta-analysis of ten studies, the power of the meta-analysis was close to 50%. We concluded that false-significant subgroup effects may trigger a meta-analysis, yet inclusion of these false-significant studies yields biased effect estimates and inflated type-I error rates. This bias can be adjusted for using a simple correction factor.

Discussion and implications for future research and clinical care in obstetrics

The studies presented in this thesis focused on predicting individual risks for developing particular outcomes in obstetrics (**Chapter 2**) and to identify specific groups of patients who may benefit from a certain treatment (**Chapter 3**). Furthermore, we addressed various methodological problems encountered in clinical prediction research and meta-analysis (**Chapter 4**). Below, we discuss the implications of the main findings in this thesis for future research and clinical care in obstetrics. This discussion follows the outline of this thesis based on the three different chapters of this thesis.

Development and validation of prognostic prediction models for individual risk assessment in obstetrics

The prediction models developed and validated in **Chapters 2.2 till 2.5** were all based on easily and readily available characteristics, and seem to be of help to clinicians as they allow

for timely prognostication, which may lead to more effective decision making and treatment. The models may serve as an alert (e.g., on the fact that the child has an above average chance of developing metabolic acidosis), be used in an individual decision of treatment (e.g., neonatologists' judgment to withdraw or intensify care), or could aid in organizational aspects (e.g. availability of doctors who can perform an instrumental delivery and availability of operating theatre and personnel). Finally, the here developed and validated prediction models in **Chapters 2.2 till 2.5** allow for more individualised counselling of patients, e.g. in the decision of a pregnant woman to start labour either at home or in the hospital under supervision of an independent midwife or general practitioner.

In general, irrespective of whether current methodological high-level standards are used in the development prediction models, they are seldom applied in obstetrics. The reasons for this may be twofold: prognostic models are often too complex for daily use in clinical settings without computer support (although the introduction of computerised patient records will clearly facilitate their application in routine care.(42)), and because many prognostic models have not been validated in other populations, clinicians may (and perhaps should) not trust (i.e. face validity) probabilities provided by these models.(43) Additionally, despite that models may have been validated and proven to be generalizable to other settings the performance may be questioned after a change in policy or guidelines. For example, in the validated models on stillbirth and neonatal death in preterm births (**Chapter 2.5**). Recently the perinatal guidelines of extremely preterm infants have been published.(44) The main message was that there are medical-scientific arguments to actively treat infants born at 24 weeks of pregnancy, which was formerly not the case. This active treatment leads to an increased survival in this group of infants. Since the data on which the models were validated do not include infants born under this new policy, the performance of the models in this specific group of children born at 24^{+0} to 24^{+6} weeks of pregnancy is unclear and should ideally be revalidated using data collected after the guideline change.

Consequently, the first step towards clinical application of the models presented in this thesis would be to validate the models in different but related populations.(11) Since the models to predict neonatal metabolic acidosis and still birth and neonatal death in very preterm births were already validated, this mainly applies to the model to predict the mode of delivery. If needed the model can be updated using methods that vary from simple recalibration to more extensive methods referred to as model revision. Recalibration includes adjustment of the intercept (i.e. correction for a difference in incidence of the outcome) of the model and overall adjustment of the associations (relative weights) of the predictors with the outcome. Model revision includes adjustment of individual predictor-outcome associations and addition of new predictors. Mostly, simple recalibration methods are sufficient.(11)

The next step towards the application of prediction models in clinical practice would be to investigate the impact on clinical practice of the models described in this thesis. The underlying assumption of prediction models is that accurately estimated probabilities improve doctors' decision making and consequently patient outcome.(45) The effect of a previously developed, validated, and (if needed) updated prognostic model on behaviour and patient outcomes should be studied separately in so called impact studies. In such an impact study it is possible to quantify the effect of using a prognostic model on doctors' behaviour, patient outcome, or cost effectiveness of care compared with not using such model. Consequently, a control group is needed. The optimal, though costly, design of such an impact study will be a randomized trial in which in one arm care as usual is performed, while in the other arm the clinician has the possibility to use the model to assist in decision making.(45) An alternative design is a before-after study with the same doctors or centres. In such a study the outcomes of interest are registered before and after application of a prediction model and compared afterwards. A disadvantage of this design is the sensitivity to temporal changes in therapeutic approaches. In medicine, examples of impact studies are scarce, but a few good examples exist..(46-48) These impact studies may use an assistive approach—simply providing the model's predicted probabilities of an outcome between 0% and 100% based on which to the provider can make a treatment decision—or a decisive approach that explicitly suggests treatment decisions for each estimated probability category. The assistive approach clearly leaves room for intuition and judgment, but a decisive approach may have greater effect.(45;48)

Obstetrics is an interesting field for the development of prediction models, since the predicted outcomes may have major impact on both the lives of the mother and her child. An intervention has impact on two lives at once, which potentially introduces a conflict of interests between mother and the child. Often interests intertwine, e.g. in the conduct of an operative intervention. An intervention will either be performed due to suspected foetal distress, which will potentially decrease the risk of an adverse neonatal outcome but will increase the chance of complications in the mother, while the opposite will be true for an intervention due to failure to progress. A method to solve this conflict is to combine maternal and neonatal outcomes into one model, e.g. using a multinomial regression model as we described in **Chapter 2.4** in the prediction of an operative delivery. Such an approach is relatively new as well as that we distinguished two distinct time points at which prediction is possible. Firstly, one can make a prediction of the outcome, e.g. mode of delivery, before labour using antepartum characteristics, i.e. maternal age, parity, etc., only. Secondly, the outcome can be predicted during labour using a combination of antepartum characteristics and information that becomes available during labour such as epidural anaesthesia and meconium stained amniotic fluid. We believe this approach should be applied more in prediction research within obstetrics, as it will create better awareness of potential health risks and it will allow for adequate anticipation on expected (i.e. predicted) problems.

Investigating treatment strategies and identifying subgroups of patients in whom treatments are (more) effective

With the conduct of the two individual participant data meta-analyses, one on the effectiveness of progestogen treatment in the reduction of adverse perinatal outcome in women with a twin pregnancy and one on the effectiveness of ST-analysis in combination with cardiotocography in reduction of neonatal metabolic acidosis in intrapartum foetal monitoring of women pregnant with a term singleton in cephalic presentation, we aimed to get better informed on the effectiveness of these strategies. However, as with many studies, meta-analysis gave answers to the questions asked but also resulted in new questions.

The fact that progestogens are not effective in the reduction of preterm birth in women with a multiple pregnancy as a group leaves the need to find a new treatment. We currently investigate within the Dutch Obstetrics Consortium whether preventive cervical pessary placement is effective in the reduction of adverse perinatal outcome, through the reduction of preterm birth, in women with multifoetal pregnancies.(49)

A pre-specified subgroup analysis in our progestogen meta-analysis studied women with a short cervix. In this subgroup of women vaginally administered progestogen showed a trend towards being effective, a result that deserves future scientific exploration. Besides in singleton and twins, progestogens have also been investigated in triplets.(50-52) Although these studies all showed no benefit of progestogen treatment, a new individual participant data meta-analysis to investigate subgroup effects in these women will also be informative.

The results of the individual participant data meta-analysis on intrapartum foetal monitoring with additional ST-analysis did not show statistically significant results, as the primary outcome neonatal metabolic acidosis had a low incidence of approximately 1%. Currently, a randomised trial on the effectiveness of ST-analysis is conducted by the National Institutes of Child Health and Human Development, which uses a composite perinatal outcome as a primary outcome that consists of intrapartum foetal and neonatal death, Apgar score below or equal to 3 at 5 minutes after birth, neonatal seizure, metabolic acidosis, intubation for ventilation at delivery, and presence of neonatal encephalopathy.(53) Apart from the results of their study it would be interested to combine these study data with the data used in our meta-analysis to come to a more final answer on the effectiveness of additional use of ST-analysis in intrapartum foetal monitoring.

Individual participant data meta-analysis is an excellent study design to investigate treatment effects, but is applied limited, specifically within obstetrics. Therefore, the conduct of more individual participant data meta-analysis should be a goal for obstetrics research and in the future, especially since it allows to adequately answer questions that individual studies cannot, e.g. subgroup effects.(54) Recently a group of international investigators that perform

clinical trials and observational studies in maternal foetal medicine and obstetrics founded the Global Obstetrics NETwork (GONET). The group brings together investigators and groups in maternal foetal medicine and obstetrics with the idea that it might be helpful to have a forum to describe the different types of collaborations and networks, to learn from each other as well as to potentially open the avenues for collaboration.{2012 971 /id} This new network potentially has two advantages. First, it may facilitate the collection of data needed to perform meta-analyses on the individual participant level. Secondly, through international collaboration, e.g. through GONET, it may be possible to run randomized trials in multiple countries simultaneously. Currently, studies are often conducted separately in several countries at the same time because researchers work with different local protocols (e.g. USA vs. Germany) or simply because they are unaware of other studies being conducted on the same topic. This approach runs the danger of studies that are unnecessarily performed, which means that participants are unnecessarily exposed to a treatment and that precious funding money is wasted. Conducting randomized trial in multiple countries simultaneously will not only lower the frequency of unnecessary studies but will also increase the efficiency of the research since the study will include participants faster, which will lower the overall costs and maybe even more importantly, the research question will be answered more quickly. In conclusion, an international collaboration like GONET has the potential to conduct randomized trials more efficiently and to increase the frequency of individual participant data meta-analysis.

Although the identification of subgroups is already a step towards more individualized treatment a subgroup is still a rather general population within which some individuals might still benefit more from treatment than others and in which some might still suffer from adverse effects. So, subgroup analyses do not fully incorporate all available patient characteristics that are needed for personalized medicine. Prediction models may aid in the further individualization of management, since making treatment decisions on the basis of a predicted treatment effect for individual patients may in some situations result in more net benefit on a group level than treating all patients (in the case of a positive trial result) or treating no one (in the case of a negative trial result).(56) Although this approach is occasionally used in the research of cancer(57-59) and cardiovascular disease,(60-62) the full potential has yet to be recognised by both researchers and clinicians, also in obstetrics. It is interesting to investigate such an approach for important clinical outcomes in obstetrics.

Methods to assess differential treatment effects

The two methodological studies presented in the **Chapter 4** of this thesis addressed various methodological problems encountered in clinical prediction research and meta-analysis.

Chapter 4.1 focussed on the reasons for unexpected predictor-outcome associations in prediction research and provides solutions for unexpected findings in prediction research

and we strongly encourage their use if encountered. Although it is widely known that researchers should assess the bias in individual studies being part of a meta-analysis, the exact consequences are less clear. The study presented in **Chapter 4.2** showed the impact of the inclusion of false-significant studies in a meta-analysis and provides a simple correction method for an inflated bias and type I error rate. In the conduct of a meta-analysis, we suggest to consider the possibility of unreliable effect estimates and test results, depending on the data used within the meta-analysis, i.e. number of (potentially false-significant) studies included and size of the studies.

Finally, future methodological research in prediction research follows from our study that describes a multinomial model to predict the mode of delivery. In regression models the reliability of the effect estimates depends on the combination of the number of participants in the smallest outcome category and the number of covariates, also known as the number of events per variable (EPV).(63) In logistic regression it is generally accepted that the estimates are reliable when there are for each covariate estimated in the model at least ten individuals with the outcome in the smallest outcome category (63;64), or even a bit lower than ten when bootstrapping techniques are used.(65) If this rule is not satisfied, models are likely to be overfitted on the data and therefore resulting in unreliable predictor-outcome associations. In multinomial regression models, however, the EPV needed in order to achieve reliable predictor-outcome associations is unknown. Therefore, the EPV in multinomial regression is of interest for future research.

Concluding remarks

With this thesis we aimed to study and enhance risk-based management in obstetrics. This goal was achieved in three ways: by development and validation of prognostic prediction models to allow for individual risk assessment, by identification of groups of patients who may benefit from specific treatment(s) strategies using subgroup analysis within meta-analyses of individual patient data, and by the evaluation of methods currently used in prediction modelling and in the assessment of subgroup effects. Obstetrics would benefit from the conduct and implementation of proper prognostic research and subgroup analysis to redeem its position at the top of the risk-based management ladder, where it once started after the introduction of the Apgar score.(66)

References

(1) Obstetric Working Group of the National Health Insurance Board of the Netherlands. The list of obstetric indications. 1999.

(2) Commissie Verloskunde van het CVZ. Verloskundig Vademecum 2003 (Dutch). Diemen, the Netherlands: College voor Zorgverzekeringen; 2003.

(3) van de Riet JE, Vandenbussche FP, Le CS, Keirse MJ. Newborn assessment and long-term adverse outcome: a systematic review. Am J Obstet Gynecol 1999 Apr;180(4):1024-9.

(4) Goodwin TM, Belai I, Hernandez P, Durand M, Paul RH. Asphyxial complications in the term newborn with severe umbilical acidemia. Am J Obstet Gynecol 1992 Dec;167(6):1506-12.

(5) Williams KP, Singh A. The correlation of seizures in newborn infants with significant acidosis at birth with umbilical artery cord gas values. Obstet Gynecol 2002 Sep;100(3):557-60.

(6) van den Berg PP, Nelen WL, Jongsma HW, Nijland R, Kollee LA, Nijhuis JG, et al. Neonatal complications in newborns with an umbilical artery pH < 7.00. Am J Obstet Gynecol 1996 Nov;175(5):1152-7.

(7) Lavrijsen SW, Uiterwaal CS, Stigter RH, de Vries LS, Visser GH, Groenendaal F. Severe umbilical cord acidemia and neurological outcome in preterm and full-term neonates. Biol Neonate 2005;88(1):27-34.

(8) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

(9) Westerhuis ME, Visser GH, Moons KG, van BE, Benders MJ, Bijvoet SM, et al. Cardiotocography plus ST analysis of foetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2010 Jun;115(6):1173-80.

(10) Westerhuis ME, Visser GH, Moons KG, Zuithoff N, Mol BW, Kwee A. Cardiotocography plus ST analysis of foetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2011 Feb;117(2 Pt 1):406-7.

(11) Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

(12) Amer-Wahlin I, Kjellmer I, Marsal K, Olofsson P, Rosen KG. Swedish randomized controlled trial of cardiotocography only versus cardiotocography plus ST analysis of foetal electrocardiogram revisited: analysis of data according to standard versus modified intention-to-treat principle. Acta Obstet Gynecol Scand 2011 Sep;90(9):990-6.

(13) Notzon FC, Placek PJ, Taffel SM. Comparisons of national cesarean-section rates. N Engl J Med 1987 Feb 12;316(7):386-9.

(14) Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Mathews TJ, Osterman MJ. Births: final data for 2008. Natl Vital Stat Rep 2010 Dec 8;59(1):1, 3-1,71.

(15) Kwee A, Elferink-Stinkens PM, Reuwer PJ, Bruinse HW. Trends in obstetric interventions in the Dutch obstetrical care system in the period 1993-2002. Eur J Obstet Gynecol Reprod Biol 2007 May;132(1):70-5.

(16) Tucker J, McGuire W. Epidemiology of preterm birth. BMJ 2004 Sep 18;329(7467):675-8.

(17) Dorling JS, Field DJ. Value and validity of neonatal disease severity scoring systems. Arch Dis Child Fetal Neonatal Ed 2008 Mar;93(2):F80-F82.

(18) Zeitlin J, Draper ES, Kollee L, Milligan D, Boerch K, Agostino R, et al. Differences in rates and short-term outcome of live births before 32 weeks of gestation in Europe in 2003: results from the MOSAIC cohort. Pediatrics 2008 Apr;121(4):e936-e944.

(19) Draper ES, Manktelow B, Field DJ, James D. Prediction of survival for preterm births by weight and gestational age: retrospective population based study. BMJ 1999 Oct 23;319(7217):1093-7.

(20) Draper ES, Manktelow B, Field DJ, James D. Tables for predicting survival for preterm births are updated. BMJ 2003 Oct 11;327(7419):872.

(21) Scottish Perinatal and Infant Mortality and Morbidity Report, 2007. Edinburgh: Information Services Division, NHS Scotland; 2008.

(22) Schaaf JM, Mol BW, bu-Hanna A, Ravelli AC. Trends in preterm birth: singleton and multiple pregnancies in the Netherlands, 2000-2007. BJOG 2011 Sep;118(10):1196-204.

(23) Cetingoz E, Cam C, Sakalli M, Karateke A, Celik C, Sancak A. Progesterone effects on preterm birth in high-risk pregnancies: a randomized placebo-controlled trial. Arch Gynecol Obstet 2011 Mar;283(3):423-9.

(24) da Fonseca EB, Bittar RE, Carvalho MH, Zugaib M. Prophylactic administration of progesterone by vaginal suppository to reduce the incidence of spontaneous preterm birth in women at increased risk: a randomized placebo-controlled double-blind study. Am J Obstet Gynecol 2003 Feb;188(2):419-24.

(25) Meis PJ, Klebanoff M, Thom E, Dombrowski MP, Sibai B, Moawad AH, et al. Prevention of recurrent preterm delivery by 17 alpha-hydroxyprogesterone caproate. N Engl J Med 2003 Jun 12;348(24):2379-85.

(26) O'Brien JM, Adair CD, Lewis DF, Hall DR, Defranco EA, Fusey S, et al. Progesterone vaginal gel for the reduction of recurrent preterm birth: primary results from a randomized, double-blind, placebo-controlled trial. Ultrasound Obstet Gynecol 2007 Oct;30(5):687-96.

(27) Fonseca EB, Celik E, Parra M, Singh M, Nicolaides KH. Progesterone and the risk of preterm birth among women with a short cervix. N Engl J Med 2007 Aug 2;357(5):462-9.

(28) Hassan SS, Romero R, Vidyadhari D, Fusey S, Baxter JK, Khandelwal M, et al. Vaginal progesterone reduces the rate of preterm birth in women with a sonographic short cervix: a multicentre, randomized, double-blind, placebo-controlled trial. Ultrasound Obstet Gynecol 2011 Jul;38(1):18-31.

(29) Romero R, Nicolaides K, Conde-Agudelo A, Tabor A, O'Brien JM, Cetingoz E, et al. Vaginal progesterone in women with an asymptomatic sonographic short cervix in the midtrimester decreases preterm delivery and neonatal morbidity: a systematic review and metaanalysis of individual patient data. Am J Obstet Gynecol 2012 Feb;206(2):124-19.

(30) Sotiriadis A, Papatheodorou S, Makrydimas G. Perinatal outcome in women treated with progesterone for the prevention of preterm birth: a meta-analysis. Ultrasound Obstet Gynecol 2012 Sep;40(3):257-66.

(31) Likis FE, Edwards DR, Andrews JC, Woodworth AL, Jerome RN, Fonnesbeck CJ, et al. Progestogens for preterm birth prevention: a systematic review and meta-analysis. Obstet Gynecol 2012 Oct;120(4):897-907.

(32) Malin GL, Morris RK, Khan KS. Strength of association between umbilical cord pH and perinatal and long term outcomes: systematic review and meta-analysis. BMJ 2010;340:c1471.

(33) Westgate JA, Bennet L, Brabyn C, Williams CE, Gunn AJ. ST waveform changes during repeated umbilical cord occlusions in near-term foetal sheep. Am J Obstet Gynecol 2001 Mar;184(4):743-51.

(34) Widmark C, Lindecrantz K, Murray H, Rosen KG. Changes in the PR, RR intervals and ST waveform of the foetal lamb electrocardiogram with acute hypoxemia. J Dev Physiol 1992 Sep;18(3):99-103.

(35) Ojala K, Vaarasmaki M, Makikallio K, Valkama M, Tekay A. A comparison of intrapartum automated foetal electrocardiography and conventional cardiotocography--a randomised controlled study. BJOG 2006 Apr;113(4):419-23.

(36) Vayssiere C, David E, Meyer N, Haberstich R, Sebahoun V, Roth E, et al. A French randomized controlled trial of ST-segment analysis in a population with abnormal cardiotocograms during labor. Am J Obstet Gynecol 2007 Sep;197(3):299-6.

(37) Westgate J, Harris M, Curnow JS, Greene KR. Plymouth randomized trial of cardiotocogram only versus ST waveform plus cardiotocogram for intrapartum monitoring in 2400 cases. Am J Obstet Gynecol 1993 Nov;169(5):1151-60.

(38) Amer-Wahlin I, Hellsten C, Noren H, Hagberg H, Herbst A, Kjellmer I, et al. Cardiotocography only versus cardiotocography plus ST analysis of foetal electrocardiogram for intrapartum foetal monitoring: a Swedish randomised controlled trial. Lancet 2001 Aug 18;358(9281):534-8.

(39) Becker JH, Bax L, mer-Wahlin I, Ojala K, Vayssiere C, Westerhuis ME, et al. ST Analysis of the Fetal Electrocardiogram in Intrapartum Fetal Monitoring: A Meta-Analysis. Obstet Gynecol 2012 Jan;119(1):145-54.

(40) Neilson JP. Fetal electrocardiogram (ECG) for foetal monitoring during labour. Cochrane Database Syst Rev 2012;4:CD000116.

(41) Potti S, Berghella V. ST waveform analysis versus cardiotocography alone for intrapartum foetal monitoring: a meta-analysis of randomized trials. Am J Perinatol 2012 Sep;29(8):657-64.

(42) James BC. Making it easy to do it right. N Engl J Med 2001 Sep 27;345(13):991-3.

(43) Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006 Feb 7;144(3):201-9.

(44) de Laat MW, Wiegerinck MM, Walther FJ, Boluyt N, Mol BW, van der Post JA, et al. [Practice guideline 'Perinatal management of extremely preterm delivery']. Ned Tijdschr Geneeskd 2010;154:A2701.

(45) Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

(46) Foy R, Penney GC, Grimshaw JM, Ramsay CR, Walker AE, MacLennan G, et al. A randomised controlled trial of a tailored multifaceted strategy to promote implementation of a clinical guideline on induced abortion care. BJOG 2004 Jul;111(7):726-33.

(47) Marrie TJ, Lau CY, Wheeler SL, Wong CJ, Vandervoort MK, Feagan BG. A controlled trial of a critical pathway for treatment of community-acquired pneumonia. CAPITAL Study Investigators. Community-Acquired Pneumonia Intervention Trial Assessing Levofloxacin. JAMA 2000 Feb 9;283(6):749-55.

(48) Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012 May;98(9):691-8.

(49) Hegeman MA, Bekedam DJ, Bloemenkamp KW, Kwee A, Papatsonis DN, van der Post JA, et al. Pessaries in multiple pregnancy as a prevention of preterm birth: the ProTwin Trial. BMC Pregnancy Childbirth 2009;9:44.

(50) Caritis SN, Rouse DJ, Peaceman AM, Sciscione A, Momirova V, Spong CY, et al. Prevention of preterm birth in triplets using 17 alpha-hydroxyprogesterone caproate: a randomized controlled trial. Obstet Gynecol 2009 Feb;113(2 Pt 1):285-92.

(51) Combs CA, Garite T, Maurel K, Das A, Porto M. Failure of 17-hydroxyprogesterone to reduce neonatal morbidity or prolong triplet pregnancy: a double-blind, randomized clinical trial. Am J Obstet Gynecol 2010 Sep;203(3):248-9.

(52) Lim AC, Schuit E, Bloemenkamp K, Bernardus RE, Duvekot JJ, Erwich JJ, et al. 17alpha-hydroxyprogesterone caproate for the prevention of adverse neonatal outcome in multiple pregnancies: a randomized controlled trial. Obstet Gynecol 2011 Sep;118(3):513-20.

(53) Belfort MA, Saade GR. ST segment analysis (STAN) as an adjunct to electronic foetal monitoring, Part II: clinical studies and future directions. Clin Perinatol 2011 Mar;38(1):159-67, vii.

(54) Riley RD, Lambert PC, bo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.

(55) Global Obstetrics Network. 2012.

(56) Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ 2011;343:d5888.

(57) Gill S, Loprinzi CL, Sargent DJ, Thome SD, Alberts SR, Haller DG, et al. Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer: who benefits and by how much? J Clin Oncol 2004 May 15;22(10):1797-806.

(58) Loprinzi CL, Thome SD. Understanding the utility of adjuvant systemic therapy for primary breast cancer. J Clin Oncol 2001 Feb 15;19(4):972-9.

(59) Steyerberg EW, Homs MY, Stokvis A, Essink-Bot ML, Siersema PD. Stent placement or brachytherapy for palliation of dysphagia from esophageal cancer: a prognostic model to guide treatment selection. Gastrointest Endosc 2005 Sep;62(3):333-40.

(60) Califf RM, Woodlief LH, Harrell FE, Jr., Lee KL, White HD, Guerci A, et al. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J 1997 Jun;133(6):630-9.

(61) Kent DM, Hayward RA, Griffith JL, Vijan S, Beshansky JR, Califf RM, et al. An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase. Am J Med 2002 Aug 1;113(2):104-11.

(62) Rothwell PM, Warlow CP. Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European Carotid Surgery Trialists' Collaborative Group. Lancet 1999 Jun 19;353(9170):2105-10.

(63) Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995 Dec;48(12):1503-10.

(64) Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984 Apr;3(2):143-52.

(65) Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer; 2009.

(66) APGAR V. A proposal for a new method of evaluation of the newborn infant. Curr Res Anesth Analg 1953 Jul;32(4):260-7.

Nederlandse samenvatting

Tegenwoordig is de gezondheidszorg steeds meer risicogestuurd en gericht op het individu. Dit betekent dat behandeling wordt toegepast op basis van individuele risico's voor het hebben of het ontwikkelen van een bepaalde ziekte of ziektebeeld. De algehele gedachte is dat het combineren van risicofactoren in zogenoemde predictiemodellen het mogelijk maakt om het risico van een individu op een bepaalde ziekte beter in te schatten en om daarmee specifieke groepen individuen te identificeren die (meer) baat zouden kunnen hebben bij een bepaalde behandeling.
De onderzoeken binnen dit manuscript hebben tot doel om het risicogestuurd beleid binnen de verloskunde te bestuderen en verbeteren. Dit doen we op drie manieren: 1.) het ontwikkelen en valideren van predictiemodellen om risico-inschatting op individueel niveau mogelijk te maken; 2.) het onderzoeken van behandelstrategieën en het identificeren van specifieke patiëntgroepen die (meer) baat zouden kunnen hebben bij een bepaalde behandeling met behulp van subgroepanalyses binnen meta-analyses van individuele patiënt data; en 3.) het evalueren van methoden binnen predictie-onderzoek, en methoden voor het bepalen van subgroepeffecten.

Ontwikkeling en validatie van predictiemodellen om individuele risico-inschatting mogelijk te maken binnen de verloskunde
In **Hoofdstuk 2** van dit proefschrift hebben we ons gericht op prognostiek binnen de verloskunde, specifiek op het identificeren van risico-indicatoren voor een overdracht tijdens de bevalling van de eerste naar de tweede lijn, het voorspellen van neonatale metabole acidose (verzuring van het kind), een vaginale kunstverlossing of keizersnede en doodgeboorte en neonatale sterfte in preterm geboren kinderen.

In Nederland hebben zwangeren met een laag risico op complicaties de keuze om de bevalling thuis of in het ziekenhuis te starten onder begeleiding van een verloskundige of huisarts (eerste lijn). Vrouwen worden doorverwezen naar een gynaecoloog in het ziekenhuis (tweede lijn) bij een verhoogd risico op complicaties bij de start of tijdens de zwangerschap, tijdens de bevalling of erna. Op dit moment zijn risico-indicatoren voor een verwijzing tijdens de bevalling gebaseerd op complicerende factoren die zich voordoen tijdens de bevalling en indicatoren die geassocieerd zijn met een slechte uitkomst van moeder of kind (1;2), maar bevatten geen indicatoren die, individueel of in combinatie, alleen aan een verwijzing tijdens de bevalling gerelateerd zijn. Daarom was het doel van **Hoofdstuk 2.1** om risico-indicatoren te identificeren voor een verwijzing tijdens de bevalling van de eerste naar de tweede lijn vanwege foetale nood, niet vorderende uitdrijving, meconium-houdend vruchtwater, niet vorderende ontsluiting, verzoek om pijnstilling, of andere redenen. Deze risico-indicatoren werden geïdentificeerd in Nederlandse vrouwen met een aterme eenlingzwangerschap die de bevalling startten in de eerste lijn tussen 2000 en 2007. Van de 746642 vrouwen werden

er 241595 (32%) tijdens de bevalling verwezen naar de tweede lijn. De sterkste risico-indicatoren voor verwijzing waren nulliparitеit en een eerdere kunstverlossing. Additionele indicatoren waren de leeftijd van de moeder, etniciteit, stedelijkheid van woonomgeving, sociaal economische status, geplande bevalling in het ziekenhuis, zwangerschapsduur, geslacht van het kind, en geboortegewicht. De geïdentificeerde risico-indicatoren kunnen zorgverleners in de eerste lijn ondersteunen in de counseling van de zwangere in de beslissing waar de bevalling te starten.

Perinatale asfyxie is gerelateerd aan verschillende korte en lange termijn complicaties, variërend van milde hypoxisch-ischemische encefalopathie tot cerebrale parese en sterfte.(3-7) Daarom is het belangrijk om zwangeren die een hoog/hoger risico hebben op het ontwikkelen van een slechte neonatale uitkomst zo snel mogelijk te kunnen identificeren, dus bij binnenkomst in het ziekenhuis of zelfs al ver voor de bevalling. Deze identificatie zal nooit kunnen gebeuren op basis van één factor.(8) In de dagelijkse praktijk worden foetussen van vrouwen met een verhoogd risico tijdens de bevalling bewaakt met cardiotocografie, microbloedonderzoek of ST-analyse van het foetale elektrocardiogram. De resultaten hiervan worden vaak in isolatie toegepast en geïnterpreteerd, zonder rekening te houden met andere factoren die mogelijk invloed hebben op de kans op een slechte uitkomst bij geboorte. Daarom was het doel van **Hoofdstuk 2.2** om neonatale metabole acidose, een gevolg van asfyxie, bij geboorte te voorspellen op basis van de beste combinatie van informatie die voor de bevalling bekend is (model 1) en in combinatie met informatie die tijdens de bevalling bekend wordt (model 2). De onderzoekspopulatie bestond uit barende vrouwen uit de tweede lijn met een aterme eenlingzwangerschap in hoofdligging die deelnamen aan een groot gerandomiseerd onderzoek.(9;10) Van de 5667 kinderen hadden er 107 (1.9%) metabole acidose. Predictoren voor metabole acidose die al voor de bevalling bekend zijn waren zwangerschapsduur, nulliparitеit, een eerdere keizersnede en diabetes. Additionele predictoren waren een spontaan begin van de bevalling en meconium-houdend vruchtwater. De overeenkomst tussen voorspelde en geobserveerde kansen (calibratie) was redelijk, net als de mogelijkheid van de modellen om onderscheid te maken tussen vrouwen met en zonder een kind met metabole acidose (discriminatie). In conclusie, in vrouwen met een aterme hoog risico eenlingzwangerschap in hoofdligging konden we metabole acidose bij geboorte voorspellen door middel van factoren die voorafgaand aan en tijdens de bevalling bekend waren.

Om aan te tonen dat predictiemodellen waardevol en generaliseerbaar zijn is er bewijs nodig dat de modellen ook goed presteren in patiënten anders dan die op welke de originele modellen zijn ontwikkeld.(11) Doordat clinici op voorhand weinig vertrouwen kunnen hebben in een model en dat al helemaal niet zullen hebben tot aangetoond is dat het model te generaliseren is naar andere populaties worden modellen in kliniek maar mondjesmaat

toegepast. Variatie, zowel in locatie als in de tijd, kan invloed hebben op de manier waarop met vrouwen met een verhoogd risico op een acidotisch kind wordt omgegaan en dus kan deze variatie de betrouwbaarheid van de modellen ontwikkeld in Hoofdstuk 2.2 beïnvloeden. Daarom hebben we de modellen in **Hoofdstuk 2.3** toegepast in Zweedse neonaten die deelnamen aan een gerandomiseerd onderzoek tussen 1998 en 2000.(12) De modellen werden toegepast in 5049 vrouwen van wie 54 (1.1%) kinderen metabole acidose hadden. Na een correctie voor verschillen in de incidentie van metabole acidose tussen de Nederlandse en Zweedse populatie, hadden de modellen een goede calibratie en een redelijk discriminerend vermogen. Om het voorspellend vermogen van de modellen verder te verhogen bevelen we aan om de modellen te updaten.

In de jaren 70 begon het aantal keizersneden in de meeste Westerse landen toe te nemen(13) en in de meeste van deze landen groeit dit aantal nog steeds.(14) De meeste keizersneden en kunstverlossingen worden gedaan vanwege problemen bij het kind (foetale nood) of bij de moeder (niet vorderende bevalling). Verder worden in absolute aantallen veruit de meeste interventies gedaan in vrouwen met een aterme eenlingzwangerschap in hoofdligging. (15) Ondanks deze aantallen blijft het lastig om te voorspellen op welke manier vrouwen uiteindelijk zullen bevallen. Daarom was het doel van **Hoofdstuk 2.4** om te onderzoeken welke combinatie van gemakkelijk te bepalen factoren, bekend zowel voor (model 1) als vroeg tijdens de bevalling (model 2), het best het risico voorspellen op een kunstverlossing of keizersnede met de indicatie foetale nood of niet vorderende bevalling. Dezelfde data als in Hoofdstuk 2.2 werd gebruikt, namelijk data van een groot gerandomiseerd onderzoek waarin 5667 barende vrouwen met een aterm hoog risico eenlingzwangerschap in hoofdligging werden bestudeerd.(9) Gebaseerd op de combinatie van de soort interventie (vaginale kunstverlossing of keizersnede) en indicatie voor de interventie werden vrouwen in één van de volgende vijf categorieën ingedeeld: spontane vaginale bevalling (referentie categorie), kunstverlossing vanwege foetale nood; keizersnede vanwege foetale nood; kunstverlossing vanwege niet vorderende bevalling; of een keizersnede vanwege niet vorderende bevalling. 375 (6.6%) en 212 (3.6%) vrouwen hadden respectievelijk een kunstverlossing of keizersnede vanwege foetale nood, en 433 (7.6%) en 571 (10.1%) vanwege een niet vorderende bevalling. Predictoren waren leeftijd van de moeder, pariteit, een eerdere keizersnede, diabetes, zwangerschapsduur, geslacht van het kind, geschat geboortegewicht (model 1) en een geïnduceerde bevalling, gebruik van oxytocine, koorts tijdens de bevalling, langdurig gebroken vliezen, meconium-houdend vruchtwater, epiduraal anesthesie, en het gebruik van ST-analyse (model 2). Beide modellen hadden een uitstekende calibratie en een goed discriminerend vermogen. Als uit validatie blijkt dat de modellen ook goed presteren in andere onderzoekspopulaties zouden de modellen toegepast kunnen worden en kunnen helpen in de counseling en in de organisatie van de zorg.

In ontwikkelde landen heeft extreme vroeggeboorte (≤ 32 weken) een incidentie van ongeveer 1-2%(16) maar is het verantwoordelijk voor bijna 1/3-1/2 van alle perinatale sterfte.(17;18) In het Verenigd Koninkrijk worden twee eenvoudige scorekaarten gebruikt in de discussie met ouders met een dreigende vroeggeboorte.(17;19;20) Deze grafieken zijn afgeleid van predictiemodellen die met de hulp van informatie die direct voor handen is, zoals zwangerschapsduur, geboortegewicht, geslacht, etniciteit en de omvang van de zwangerschap, de overlevingskans voorspellen voor extreem vroeggeboren kinderen die levend waren bij de start van de bevalling en voor extreem vroeggeboren kinderen die opgenomen waren op een neonatale intensive care unit.(19) Ondanks dat de modellen ontwikkeld zijn voor het Verenigd Koninkrijk, kunnen ze ook van nut zijn in andere Westerse landen. Het doel van **Hoofdstuk 2.5** was om de generaliseerbaarheid van deze modellen te onderzoeken. De modellen werden gevalideerd op alle kinderen die in Nederland zijn geboren tussen 1 januari 2000 en 31 december 2007 met een zwangerschapsduur tussen de 22 en 32 weken, met een Europese etniciteit, die leefden bij aanvang van de bevalling (n=17852) en opgenomen waren op de intensive care (n=11578). 16.7% (n=2939) van de kinderen levend bij aanvang van de bevalling overleed tijdens de bevalling of binnen 28 dagen na geboorte, en 7.8% (n=908) van de kinderen opgenomen op de intensive care overleed voor ontslag van de intensive care. Het predictiemodel voor kinderen levend bij aanvang van de bevalling had een goede calibratie en uitstekend discriminerend vermogen. Het predictiemodel voor kinderen opgenomen op de neonatale intensive care had een goede calibratie en een goed discriminerend vermogen. We concludeerden dat de resultaten van het onderzoek reden geven om de modellen in Nederland toe te passen en waarschijnlijk ook in andere Westerse landen.

Identificatie van subgroepen van patiënten in wie behandelingen (meer) effect hebben
Hoofdstuk 3 van dit proefschrift richt zich op individuele patiënt data meta-analyses binnen de verloskunde gericht op het kwantificeren van behandeleffecten van progesteron in tweelingzwangerschappen (**Hoofdstuk 3.1** en **3.2**) en van ST-analyse in aterme eenlingzwangerschappen in hoofdligging (**Hoofdstuk 3.3**), overall en in specifieke patiëntgroepen.

Hoofdstuk 3.1 en **3.2** beschrijven respectievelijk het protocol en de resultaten van een individuele patiënt data meta-analyse, een onderzoeksopzet waarin data van individuele onderzoeken gecombineerd wordt, waarin de effectiviteit van progesteronbehandeling in de reductie van een slechte perinatale uitkomst in tweelingzwangerschappen werd onderzocht. Beide onderzoeken worden hierna gezamenlijk beschreven. In tweelingzwangerschappen komt doodgeboorte, neonatale sterfte, vroeggeboorte, (erg) laag geboortegewicht, en opeenvolgende lange termijn morbiditeit substantieel vaker voor dan in eenlingen (21); iets

wat voornamelijk toe te schrijven is aan het hogere aantal vroeggeboorten (22). Tot op heden is er geen interventie gevonden om vroeggeboorte bij tweelingzwangerschappen te reduceren. Gerandomiseerde onderzoeken in eenlingen laten zien dat progesteronbehandeling het aantal vroeggeboortes reduceert in vrouwen die een verhoogd risico hebben op vroeggeboorte vanwege een eerdere vroeggeboorte (23-26) of een korte baarmoedermond (cervix) (27-29). Als gevolg hiervan is de effectiviteit van progesteron, zowel intramusculair toegediend 17-hydroxyprogesteron caproate (17Pc) als vaginaal toegediend progesteron, ook in tweelingen onderzocht. Recent zijn er drie meta-analyses gepubliceerd (op basis van gepubliceerde studieresultaten) waarvan er één slechts een klein aantal tweelingen onderzocht en zich alleen richtte op vaginaal progesteron (29), er één te weinig informatie had om specifieke subgroepen te onderzoeken (30), en de derde geen onderscheid maakte tussen 17Pc en vaginaal progesteron terwijl de effecten van beiden typen progesteron kunnen verschillen (31). Geen van deze drie meta-analyses beschreef alle onderzoeken over progesteron in tweelingen (29-31). Daarom was het doel van onze studie, beschreven in **Hoofdstuk 3.1** en **3.2**, om een individuele patiënt data meta-analyse uit te voeren om de effecten van progesteron te onderzoeken, apart voor 17Pc en vaginaal progesteron, in vrouwen met een tweelingzwangerschap en specifieke patiëntgroepen gebaseerd op chorioniciteit, cervix lengte, en een eerdere spontane vroeggeboorte. In totaal includeerden de dertien onderzoeken in onze meta-analyse samen 3764 vrouwen en hun 7528 baby's. Zowel 17Pc als vaginaal progesteron leidde niet tot een reductie in slechte perinatale uitkomst. In vrouwen behandeld met 17Pc kwam een bevalling < 32 weken vaker voor dan in de controle groep. Er werden geen specifieke groepen geïdentificeerd die baat hadden bij progesteronbehandeling. We concludeerden dat progesteronbehandeling, onafhankelijk van het type, in vrouwen met een ongecompliceerde tweelingzwangerschap niet leidt tot een verbetering in perinatale uitkomst.

Zoals eerder beschreven is perinatale asfyxie geassocieerd met verschillende korte en lange termijn complicaties.(3;7;32) Foetale bewaking tijdens de bevalling heeft tot doel om baby's te identificeren die een risico hebben op dergelijke asfyxie. Een relatief nieuwe methode voor continue foetale bewaking is de STAN® methode (Neoventa Medical, Gothenburg, Sweden) waarin (classificatie van) cardiotocografie wordt gecombineerd met ST-analyse van het foetaal elektrocardiogram. Zoals in onderzoek in dieren eerder is aangetoond kunnen veranderingen in het foetaal elektrocardiogram in combinatie met abnormaliteiten op de cardiotocografie een indicatie zijn van foetale hypoxie.(33;34) Vijf gerandomiseerde onderzoeken die het effect van additionele ST-analyse onderzochten op de reductie van metabole acidose lieten wisselende effecten zien, met vier onderzoeken die geen statistisch significant effect lieten zien,(9;35-37) en één studie die een significante reductie in metabole acidose liet zien in de ST-analyse groep.(38) Gebaseerd op deze onderzoeken werden er drie meta-analyses

gedaan gebruikmakend van geaggregeerde data (dus op basis van gepubliceerde resultaten). (39-41) Deze onderzoeken konden vanwege een tekort aan informatie niet rapporteren over alle relevant uitkomsten en onderzochten daarom ook geen specifieke subgroepen. Vanwege deze tekortkomingen presenteren we in **Hoofdstuk 3.3** een individuele patiënt data meta-analyse waarin we het effect te onderzochten van foetale bewaking tijdens de bevalling met ST-analyse in combinatie met cardiotocografie op de reductie van neonatale metabole acidose. De vier gerandomiseerde onderzoeken in de meta-analyse includeerden 12987 barende vrouwen met een aterm eenlingzwangerschap in hoofdligging. Additionele ST-analyse resulteerde niet in een statistisch significante daling van metabole acidose in vergelijking met cardiotocografie alleen, maar zorgde wel voor een reductie in het aantal kunstverlossingen en microbloedonderzoeken. Subgroepanalyse liet zien dat additionele ST-analyse zorgde voor minder opnames op een neonatale intensive care unit in vrouwen met een zwangerschapsduur voorbij de 41 weken. We concludeerden dat foetale bewaking tijdens de bevalling met cardiotocografie met ST-analyse niet zorgt voor een reductie in het risico op metabole acidose, maar wel voor een reductie in het aantal kunstverlossingen en microbloedonderzoeken.

Methoden binnen predictie-onderzoek en in het bepalen van subgroepeffecten

Hoofdstuk 4 van dit proefschrift was gericht op mogelijke methodologische problemen gerelateerd aan Hoofdstuk 2 en 3, namelijk de ontwikkeling en het valideren van predictiemodellen en de identificatie van groepen patiënten voor wie een behandeling (meer) effectief is.

Klinische predictiemodellen hebben als doel binnen individuen te voorspellen wat de kans is op een bepaald uitkomst, gebruikmakend van meerdere voorspellende variabelen (predictoren). Ondanks dat de relatie tussen een variabele en de uitkomst niet per se causaal hoeft te zijn (variabele leidt niet per se tot uitkomst) verwachten clinici toch vaak een bepaalde richting (positief of negatief) van de relatie. Dus, als een negatieve relatie wordt gevonden wanneer een positieve relatie wordt verwacht, dan is deze bevinding onverwacht. Hoogstwaarschijnlijk zal deze relatie niet in een andere populatie gevonden worden en kan dit daardoor tot een slechte generaliseerbaarheid van het model leiden. Verder kunnen onverwachte relaties suggereren dat het predictiemodel onjuist is, waardoor de geloofwaardigheid van het model omlaag gaat.(8) Het doel van **Hoofdstuk 4.1** was om redenen voor onverwachte bevindingen in predictie-onderzoek te beschrijven en mogelijke oplossingen te bieden. Mogelijke oorzaken voor onverwachte relaties zijn toeval, misclassificatie, selectiebias, verwarren van effecten (confounding), interventie-effecten, en heterogeniteit. De onderzoeksopzet of de analytische methode om met een onverwachte bevinding om te gaan hangt af van de oorzaak van de onverwachte bevinding. We concludeerden dat om de acceptatie en impact van een model

te vergroten onverwachte relaties in predictie-onderzoek door onderzoekers altijd moeten worden gerapporteerd, evenals de potentiële oorzaken van de onverwachte bevindingen als ook de pogingen die gedaan zijn om ze ongedaan te maken.

In een meta-analyse worden de resultaten of data van verschillende onderzoeken gecombineerd. Dit wordt gezien als het best mogelijke design om effecten van bepaalde interventies te onderzoeken. De keuze om een meta-analyse te starten is vaak gebaseerd op een geobserveerd effect of een trend in de richting van een effect die gevonden wordt in een individuele studie. Gerandomiseerde onderzoeken zijn vaak zodanig opgezet dat ze een effect aan kunnen tonen op de primaire uitkomst van de studie, maar dus niet voor het aantonen van interacties of effecten op veiligheidseindpunten. Daardoor kan een geobserveerd effect voor een secundaire uitkomst, veiligheidseindpunt of subgroep een foutpositieve bevinding zijn, maar kan dit wel zorgen voor het initiëren van een meta-analyse. Wanneer deze foutpositieve studie op basis waarvan de meta-analyse gedaan wordt deel uit maakt van de meta-analyse kan dit voor de (resultaten van de) meta-analyse belangrijke consequenties hebben. Daarom onderzochten we in **Hoofdstuk 4.2** dit potentiële probleem. Eerst hebben we analytisch afgeleid wat de impact is van de inclusie van foutpositieve onderzoeken in een meta-analyse op de bias in de effectschatting van de meta-analyse. Vervolgens hebben we, door middel van simulaties, onderzocht wat de invloed van deze aanpak is op de type I fout en de power van de meta-analyse. Verder introduceerden we nog een simpele correctiemethode om te corrigeren voor potentieel verhoogde type I fouten en de bias, door middel van de analytisch afgeleide bias. We richtten ons specifiek op subgroepeffecten binnen meta-analyses van gerandomiseerde onderzoeken. De bias en type I fout hingen af van het aantal foutpositieve onderzoeken in de meta-analyse. Als 20% van de onderzoeken foutpositief was dan was de bias in de effectschatting van de meta-analyse 0.08 (z-score, in plaats van 0) en de type I fout 20% (in plaats van 5%). Biascorrectie verlaagde de type I fout tot 5%. Bij de aanname van een behandeleffect, maar geen significante onderzoeken in de meta-analyse was de power van de meta-analyse ongeveer 50%.We concludeerden dat foutpositieve subgroepeffecten kunnen leiden tot het opstarten van een meta-analyse, maar dat inclusie van deze foutpositieve onderzoeken in de meta-analyse kan zorgen voor bias in de effectschatting en vergrootte type I fouten. Deze bias en type I fout kunnen gecorrigeerd worden door een simpele correctiemethode.

References

(1) Obstetric Working Group of the National Health Insurance Board of the Netherlands. The list of obstetric indications. 1999.

(2) Commissie Verloskunde van het CVZ. Verloskundig Vademecum 2003 (Dutch). Diemen, the Netherlands: College voor Zorgverzekeringen; 2003.

(3) van de Riet JE, Vandenbussche FP, Le CS, Keirse MJ. Newborn assessment and long-term adverse outcome: a systematic review. Am J Obstet Gynecol 1999 Apr;180(4):1024-9.

(4) Goodwin TM, Belai I, Hernandez P, Durand M, Paul RH. Asphyxial complications in the term newborn with severe umbilical acidemia. Am J Obstet Gynecol 1992 Dec;167(6):1506-12.

(5) Williams KP, Singh A. The correlation of seizures in newborn infants with significant acidosis at birth with umbilical artery cord gas values. Obstet Gynecol 2002 Sep;100(3):557-60.

(6) van den Berg PP, Nelen WL, Jongsma HW, Nijland R, Kollee LA, Nijhuis JG, et al. Neonatal complications in newborns with an umbilical artery pH < 7.00. Am J Obstet Gynecol 1996 Nov;175(5):1152-7.

(7) Lavrijsen SW, Uiterwaal CS, Stigter RH, de Vries LS, Visser GH, Groenendaal F. Severe umbilical cord acidemia and neurological outcome in preterm and full-term neonates. Biol Neonate 2005;88(1):27-34.

(8) Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

(9) Westerhuis ME, Visser GH, Moons KG, van BE, Benders MJ, Bijvoet SM, et al. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2010 Jun;115(6):1173-80.

(10) Westerhuis ME, Visser GH, Moons KG, Zuithoff N, Mol BW, Kwee A. Cardiotocography plus ST analysis of fetal electrocardiogram compared with cardiotocography only for intrapartum monitoring: a randomized controlled trial. Obstet Gynecol 2011 Feb;117(2 Pt 1):406-7.

(11) Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

(12) Amer-Wahlin I, Kjellmer I, Marsal K, Olofsson P, Rosen KG. Swedish randomized controlled trial of cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram revisited: analysis of data according to standard versus modified intention-to-treat principle. Acta Obstet Gynecol Scand 2011 Sep;90(9):990-6.

(13) Notzon FC, Placek PJ, Taffel SM. Comparisons of national cesarean-section rates. N Engl J Med 1987 Feb 12;316(7):386-9.

(14) Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Mathews TJ, Osterman MJ. Births: final data for 2008. Natl Vital Stat Rep 2010 Dec 8;59(1):1, 3-1,71.

(15) Kwee A, Elferink-Stinkens PM, Reuwer PJ, Bruinse HW. Trends in obstetric interventions in the Dutch obstetrical care system in the period 1993-2002. Eur J Obstet Gynecol Reprod Biol 2007 May;132(1):70-5.

(16) Tucker J, McGuire W. Epidemiology of preterm birth. BMJ 2004 Sep 18;329(7467):675-8.

(17) Dorling JS, Field DJ. Value and validity of neonatal disease severity scoring systems. Arch Dis Child Fetal Neonatal Ed 2008 Mar;93(2):F80-F82.

(18) Zeitlin J, Draper ES, Kollee L, Milligan D, Boerch K, Agostino R, et al. Differences in rates and short-term outcome of live births before 32 weeks of gestation in Europe in 2003: results from the MOSAIC cohort. Pediatrics 2008 Apr;121(4):e936-e944.

(19) Draper ES, Manktelow B, Field DJ, James D. Prediction of survival for preterm births by weight and gestational age: retrospective population based study. BMJ 1999 Oct 23;319(7217):1093-7.

(20) Draper ES, Manktelow B, Field DJ, James D. Tables for predicting survival for preterm births are updated. BMJ 2003 Oct 11;327(7419):872.

(21) Scottish Perinatal and Infant Mortality and Morbidity Report, 2007. Edinburgh: Information Services Division, NHS Scotland; 2008.

(22) Schaaf JM, Mol BW, bu-Hanna A, Ravelli AC. Trends in preterm birth: singleton and multiple pregnancies in the Netherlands, 2000-2007. BJOG 2011 Sep;118(10):1196-204.

(23) Cetingoz E, Cam C, Sakalli M, Karateke A, Celik C, Sancak A. Progesterone effects on preterm birth in high-risk pregnancies: a randomized placebo-controlled trial. Arch Gynecol Obstet 2011 Mar;283(3):423-9.

(24) da Fonseca EB, Bittar RE, Carvalho MH, Zugaib M. Prophylactic administration of progesterone by vaginal suppository to reduce the incidence of spontaneous preterm birth in women at increased risk: a randomized placebo-controlled double-blind study. Am J Obstet Gynecol 2003 Feb;188(2):419-24.

(25) Meis PJ, Klebanoff M, Thom E, Dombrowski MP, Sibai B, Moawad AH, et al. Prevention of recurrent preterm delivery by 17 alpha-hydroxyprogesterone caproate. N Engl J Med 2003 Jun 12;348(24):2379-85.

(26) O'Brien JM, Adair CD, Lewis DF, Hall DR, Defranco EA, Fusey S, et al. Progesterone vaginal gel for the reduction of recurrent preterm birth: primary results from a randomized, double-blind, placebo-controlled trial. Ultrasound Obstet Gynecol 2007 Oct;30(5):687-96.

(27) Fonseca EB, Celik E, Parra M, Singh M, Nicolaides KH. Progesterone and the risk of preterm birth among women with a short cervix. N Engl J Med 2007 Aug 2;357(5):462-9.

(28) Hassan SS, Romero R, Vidyadhari D, Fusey S, Baxter JK, Khandelwal M, et al. Vaginal progesterone reduces the rate of preterm birth in women with a sonographic short cervix: a multicenter, randomized, double-blind, placebo-controlled trial. Ultrasound Obstet Gynecol 2011 Jul;38(1):18-31.

(29) Romero R, Nicolaides K, Conde-Agudelo A, Tabor A, O'Brien JM, Cetingoz E, et al. Vaginal progesterone in women with an asymptomatic sonographic short cervix in the midtrimester decreases preterm delivery and neonatal morbidity: a systematic review and metaanalysis of individual patient data. Am J Obstet Gynecol 2012 Feb;206(2):124-19.

(30) Sotiriadis A, Papatheodorou S, Makrydimas G. Perinatal outcome in women treated with progesterone for the prevention of preterm birth: a meta-analysis. Ultrasound Obstet Gynecol 2012 Sep;40(3):257-66.

(31) Likis FE, Edwards DR, Andrews JC, Woodworth AL, Jerome RN, Fonnesbeck CJ, et al. Progestogens for preterm birth prevention: a systematic review and meta-analysis. Obstet Gynecol 2012 Oct;120(4):897-907.

(32) Malin GL, Morris RK, Khan KS. Strength of association between umbilical cord pH and perinatal and long term outcomes: systematic review and meta-analysis. BMJ 2010;340:c1471.

(33) Westgate JA, Bennet L, Brabyn C, Williams CE, Gunn AJ. ST waveform changes during repeated umbilical cord occlusions in near-term fetal sheep. Am J Obstet Gynecol 2001 Mar;184(4):743-51.

(34) Widmark C, Lindecrantz K, Murray H, Rosen KG. Changes in the PR, RR intervals and ST waveform of the fetal lamb electrocardiogram with acute hypoxemia. J Dev Physiol 1992 Sep;18(3):99-103.

(35) Ojala K, Vaarasmaki M, Makikallio K, Valkama M, Tekay A. A comparison of intrapartum automated fetal electrocardiography and conventional cardiotocography--a randomised controlled study. BJOG 2006 Apr;113(4):419-23.

(36) Vayssiere C, David E, Meyer N, Haberstich R, Sebahoun V, Roth E, et al. A French randomized controlled trial of ST-segment analysis in a population with abnormal cardiotocograms during labor. Am J Obstet Gynecol 2007 Sep;197(3):299-6.

(37) Westgate J, Harris M, Curnow JS, Greene KR. Plymouth randomized trial of cardiotocogram only versus ST waveform plus cardiotocogram for intrapartum monitoring in 2400 cases. Am J Obstet Gynecol 1993 Nov;169(5):1151-60.

(38) Amer-Wahlin I, Hellsten C, Noren H, Hagberg H, Herbst A, Kjellmer I, et al. Cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomised controlled trial. Lancet 2001 Aug 18;358(9281):534-8.

(39) Becker JH, Bax L, mer-Wahlin I, Ojala K, Vayssiere C, Westerhuis ME, et al. ST Analysis of the Fetal Electrocardiogram in Intrapartum Fetal Monitoring: A Meta-Analysis. Obstet Gynecol 2012 Jan;119(1):145-54.

(40) Neilson JP. Fetal electrocardiogram (ECG) for fetal monitoring during labour. Cochrane Database Syst Rev 2012;4:CD000116.

(41) Potti S, Berghella V. ST waveform analysis versus cardiotocography alone for intrapartum fetal monitoring: a meta-analysis of randomized trials. Am J Perinatol 2012 Sep;29(8):657-64.

List of publications

(1) **Schuit E**, Groenwold RH, Harrell FE, de Kort WL, Kwee A, Mol BW, et al. Unexpected predictor-outcome associations in clinical prediction research: causes and solutions. CMAJ 2013 Jan 21.

(2) **Schuit E**, Amer-Wahlin I, Ojala K, Vayssiere C, Westerhuis ME, Marsal K, et al. Effectiveness of electronic fetal monitoring with additional ST-analysis in vertex singleton pregnancies beyond 36 weeks of gestation: an individual participant data meta-analysis. Am J Obstet Gynecol 2013 Jan 17.

(3) Roos C, Spaanderman ME, **Schuit E**, Bloemenkamp KW, Bolte AC, Cornette J, et al. Effect of maintenance tocolysis with nifedipine in threatened preterm labor on perinatal outcomes: a randomized controlled trial. JAMA 2013 Jan 2;309(1):41-7.

(4) Lim AC, **Schuit E**, Papatsonis D, van EJ, Porath MM, van Oirschot CM, et al. Effect of 17-alpha hydroxyprogesterone caproate on cervical length in twin pregnancies. Ultrasound Obstet Gynecol 2012 Oct;40(4):426-30.

(5) **Schuit E**, Hukkelhoven CW, Manktelow BN, Papatsonis DN, de Kleine MJ, Draper ES, et al. [Predicting death in very preterm infants]. Ned Tijdschr Geneeskd 2012;156(38):A4846.

(6) **Schuit E**, Moons K, Groenwold R, Kwee A, Mol B. A clinical prediction model to assess the risk of operative delivery. BJOG 2012 Oct;119(11):1418-9.

(7) **Schuit E**, Amer-Wahlin I, Groenwold RH, Mol BW, Moons KG, Kwee A. Prediction of neonatal metabolic acidosis in women with a singleton term pregnancy in cephalic presentation: an external validation study. Am J Perinatol 2012 Oct;29(9):681-6.

(8) **Schuit E**, Kwee A, Westerhuis ME, Van Dessel HJ, Graziosi GC, Van Lith JM, et al. A clinical prediction model to assess the risk of operative delivery. BJOG 2012 Jul;119(8):915-23.

(9) **Schuit E**, Stock S, Groenwold RH, Maurel K, Combs CA, Garite T, et al. Progestogens to prevent preterm birth in twin pregnancies: an individual participant data meta-analysis of randomized trials. BMC Pregnancy Childbirth 2012;12:13.

(10) **Schuit E**, Hukkelhoven CW, Manktelow BN, Papatsonis DN, de Kleine MJ, Draper ES, et al. Prognostic models for stillbirth and neonatal death in very preterm birth: a validation study. Pediatrics 2012 Jan;129(1):e120-e127.

(11) Becker JH, Kuipers LJ, **Schuit E**, Visser GH, Van Den Akker ES, Van BE, et al. Predictive value of the baseline T-QRS ratio of the fetal electrocardiogram in intrapartum fetal monitoring: a prospective cohort study. Acta Obstet Gynecol Scand 2012 Feb;91(2):189-97.

(12) Lim AC, **Schuit E**, Bloemenkamp K, Bernardus RE, Duvekot JJ, Erwich JJ, et al. 17alpha-hydroxyprogesterone caproate for the prevention of adverse neonatal outcome in multiple pregnancies: a randomized controlled trial. Obstet Gynecol 2011 Sep;118(3):513-20.

(13) Westerhuis ME, **Schuit E**, Kwee A, Zuithoff NP, Groenwold RH, Van Den Akker ES, et al. Prediction of neonatal metabolic acidosis in women with a singleton term pregnancy in cephalic presentation. Am J Perinatol 2012 Mar;29(3):167-74.

List of co-authors and their affiliations

Prof. Dr. M.A. Aboulghar, The Egyptian IVF Centre, Maadi, Cairo, Egypt

Dr. M.M. Aboulghar, The Egyptian IVF Centre, Maadi, Cairo, Egypt

Dr. Y.M. Amin, The Egyptian IVF Centre, Maadi, Cairo, Egypt

Prof. Dr. J. Awwad, Department of Obstetrics and Gynecology, American University of Beirut Medical Centre, Beirut, Lebanon

Dr. E.S.A. Van den Akker, Department of Obstetrics and Gynecology, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands

Dr. Isis Amer-Wahlin, Department of Women and Child Health, Karolinska Institute Solna, Stockholm, Sweden

Dr. E. van Beek, Department of Obstetrics and Gynecology, St Antonius Hospital, Nieuwegein, the Netherlands

Dr. C.M. Briery, Department of Obstetrics and Gynecology, Willis Knighton Medical Centre, Shreveport, Louisiana, USA

Dr. C. Cam, Department of Obstetrics and Gynecology, Zeynep Kamil Women and Children Diseases Education and Research Hospital, Uskudar, Istanbul, Turkey

Prof. Dr. S.N. Caritis, Department of Obstetrics and Gynecology, Magee-Womens Hospital, University of Pittsburgh, Pittsburgh, USA

Dr. E. Çetingöz, Department of Obstetrics and Gynecology, Zeynep Kamil Women and Children Diseases Education and Research Hospital, Uskudar, Istanbul, Turkey

Dr. C.A. Combs, Obstetrix Medical Group, San Jose, California, USA

Dr. H.J.H.M. van Dessel, Department of Obstetrics and Gynecology, TweeSteden Hospital, Tilburg, the Netherlands

Prof. Dr. E.S. Draper, Dept of Health Sciences, University of Leicester, Leicester, United Kingdom

Dr. A.P. Drogtrop, Department of Obstetrics and Gynecology, TweeSteden Hospital, Tilburg, the Netherlands

Dr. E.B. Fonseca, Departamento de Obstetrícia e Ginecologia, Hospital do Servidor Publico Estadual, "Francisco Morato de Oliveira," and School of Medicine, University of São Paulo, São Paulo, Brazil

Prof. Dr. T. Garite, The Centre for Research, Education & Quality (CREQ), Obstetrix Medical Group, Steamboat Springs, Colorado, USA

Dr. H.P. van Geijn, Department of Obstetrics and Gynecology, Free University Medical Centre, Amsterdam, the Netherlands

MSc. B.Y. van der Goes, Department of Obstetrics and Gynaecology, Academic Medical Centre, Amsterdam, the Netherlands

Dr. G.C.M. Graziosi, Department of Obstetrics and Gynecology, St Antonius Hospital, Nieuwegein, the Netherlands

Dr. R.H.H. Groenwold, Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, the Netherlands

Prof. Dr. F.E. Harrell Jr., Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

Dr. C.W.P.M. Hukkelhoven, The Netherlands Perinatal Registry, Utrecht, the Netherlands

Prof. Dr. A. Karateke, Department of Obstetrics and Gynecology, Zeynep Kamil Women and Children Diseases Education and Research Hospital, Uskudar, Istanbul, Turkey

Dr. K. Klein, Department of Obstetrics and Feto maternal Medicine, Medical University of Vienna, Vienna, Austria

Dr. M.J.K. de Kleine, Department of Neonatology, Máxima Medical Centre, Eindhoven, The Netherlands

Dr. W.L.A.M. de Kort, Sanquin Blood Bank, Southeast Region, Nijmegen, the Netherlands

Dr. A. Kwee, Department of Obstetrics and Gynecology, University Medical Centre Utrecht, Utrecht, the Netherlands

Dr. A.C. Lim, Department of Obstetrics and Gynecology, Academic Medical Centre Amsterdam, Amsterdam, the Netherlands

Prof. Dr. J.M.M. van Lith, Department of Obstetrics and Gynecology, Leiden University Medical Centre, Leiden, the Netherlands

Dr. E.F. Magann, Department of Obstetrics and Gynecology, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

Dr. B.N. Manktelow, Department of Health Sciences, University of Leicester, Leicester, United Kingdom

Prof. Dr. K. Maršál, Department of Obstetrics and Gynecology, Clinical Sciences, Lund University, Lund, Sweden

MSN. K. Maurel, The Centre for Research, Education & Quality (CREQ), Obstetrix Medical Group, Fountain Valley, California, USA

Dr. J. Meseguer, Departamento de Obstetricia y Ginecología. Hospital Universitario Virgen de la Arrixaca, Murcia, Spain

Prof. Dr. B.W.J. Mol, Department of Obstetrics and Gynecology, Academic Medical Centre Amsterdam, Amsterdam, the Netherlands

Prof. Dr. K.G.M. Moons, Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, the Netherlands

Dr. J.C. Morrison, Department of Obstetrics and Gynecology, University of Mississippi Medical Centre, Jackson, USA

Prof. Dr. A.H. Nassar, Department of Obstetrics and Gynecology, American University of Beirut Medical Centre, Beirut, Lebanon

Prof. Dr. K.H. Nicolaides, Department of Obstetrics and Gynecology, King's College Hospital, London, United Kingdom

Prof. Dr. J.G. Nijhuis, Department of Obstetrics and Gynecology, Maastricht University Medical Centre, Maastricht, the Netherlands

Prof. Dr. J.E. Norman, MRC Centre for Reproductive Health, University of Edinburgh, Edinburgh, Scotland

Dr. S.G. Oei, Department of Obstetrics and Gynecology, Máxima Medical Centre, Veldhoven, the Netherlands

Dr. K. Ojala, Department of Obstetrics and Gynecology, Oulu University Hospital, Oulu, Finland

Dr. H.P. Oosterbaan, Department of Obstetrics and Gynecology, Jeroen Bosch Hospital, 's Hertogenbosch, the Netherlands

MSc. I. Overbeeke-van Os, Department of Obstetrics and Gynaecology, University Medical Centre Utrecht, Utrecht, the Netherlands

Dr. D.N.M. Papatsonis, Department of Obstetrics and Gynaecology, Amphia Hospital, Breda, The Netherlands

Prof. Dr. A. Perales, Department of Obstetrics and Gynecology, University Hospital La Fe, Valencia, Spain

Dr. M.M. Porath, Department of Obstetrics and Gynecology, Máxima Medical Centre, Veldhoven, the Netherlands

Dr. R.J.P. Rijnders, Department of Obstetrics and Gynecology, Jeroen Bosch Hospital, 's Hertogenbosch, the Netherlands

Dr. R.D. Riley, School of Health and Population Sciences, University of Birmingham, Birmingham, UK

Dr. L. Rode, Department of Foetal Medicine 4002, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

Prof. Dr. K.C.B. Roes, Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, the Netherlands

Prof. Dr. S. Ross, Department of Obstetrics and Gynecology, University of Calgary, Canada

Prof. Dr. D.J. Rouse, Department of Obstetrics and Gynecology, Brown University, Providence, Rhode Island and Providence Plantations, USA

Prof. Dr. G.R. Saade, Department of Obstetrics and Gynecology, University of Texas Medical Branch, Galveston, USA

Dr. N.W.E. Schuitemaker, Department of Obstetrics and Gynecology, Diakonessenhuis, Utrecht, the Netherlands

Prof. Dr. Vicente Serra, Maternal-Foetal Medicine Unit, Instituto Valenciano de Infertilidad, University of Valencia, Valencia, Spain

Prof. Dr. C.Y. Spong, National Institute of Child Health and Human Development, Bethesda, Maryland, USA

Prof. Dr. E.W. Steyerberg, Centre for Medical Decision Sciences, Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

Drs. S. Stock, MRC Centre for Reproductive Health, University of Edinburgh, Edinburgh, Scotland

Prof. Dr. A. Tabor, Department of Foetal Medicine 4002, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

Dr. A. Tekay, Department of Obstetrics and Gynecology, Oulu University Hospital, Oulu, Finland

Prof. Dr. E.A. Thom, The Biostatistics Centre, George Washington University, Rockville, Maryland, USA

Prof. Dr. I.M. Usta, Department of Obstetrics and Gynecology, American University of Beirut Medical Centre, Beirut, Lebanon

Prof. Dr. C. Vayssière, Department of Obstetrics and Gynecology, CHU Toulouse and Inserm UMR1027, University Toulouse III, Toulouse, France

Dr. Y. Vergouwe, Centre for Medical Decision Sciences, Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

Prof. Dr. G.H.A. Visser, Department of Obstetrics and Gynecology, University Medical Centre Utrecht, Utrecht, the Netherlands

Dr. M.E.M.H. Westerhuis, Department of Obstetrics and Gynecology, University Medical Centre Utrecht, Utrecht, the Netherlands

Dr. L.D.E. Wijnberger, Department of Obstetrics and Gynecology, Leiden University Medical Centre, Leiden, the Netherlands

Dr. C. Willekes, Department of Obstetrics and Gynecology, Maastricht University Medical Centre, Maastricht, the Netherlands

Dr. S. Wood, Department of Obstetrics and Gynecology, University of Calgary, Canada

Dr. M.G.A.J. Wouters, Department of Obstetrics and Gynecology, Free University Medical Centre, Amsterdam, the Netherlands

Dr. J.M. Zachary, The Biostatistics Centre, George Washington University, Rockville, MD, USA

Dr. N.P.A. Zuithoff, Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, the Netherlands

Dankwoord

"Bedankt, en we zullen het nooit, nooit vergeten!"

Het leven en dus ook onderzoek zijn onmogelijk alleen te doen. Voor beide heb je mensen nodig die je steunen, aanvullen en waar nodig verbeteren. Via dit dankwoord wil ik iedereen die een rol heeft gespeeld in mijn promotieonderzoek, op onderzoeksgebied maar zeker ook daarbuiten, persoonlijk bedanken.

Beste Prof. Dr. B.W.J. Mol, beste BW, het mag geen geheim zijn dat jouw motto luidt: "een dag niet gerandomiseerd, is een dag niet geleefd". Ondanks dat ik zelf nooit hoofdonderzoeker van een studie binnen het Consortium ben geweest heb ik veel profijt gehad van jouw motto. Data van vele studies heb je me de afgelopen drie jaar toevertrouwd en dat blijf je ook de komende jaren doen. Ik heb veel bewondering voor jouw inzet en kennis, het vermogen om mensen te motiveren en je oneindige stroom aan ideeën. Daarom is het ook ontzettend jammer dat we je de komende jaren, in ieder geval in Nederland, niet meer gaan zien. Ik wil je wederom veel plezier en succes wensen met je nieuwe Australische avontuur. Ik, maar ook het Consortium, zal je zeker gaan missen. Heel erg bedankt voor de kansen die je me geboden hebt en wellicht kunnen we in de toekomst, aan de andere kant van de wereld, weer samenwerken?!

Beste Prof. Dr. K.G.M. Moons, beste Carl, het is me een waar genoegen geweest om de afgelopen jaren met je samen te werken. Net als BW heb je ontzettend brede kennis en versta jij de kunst om mensen te motiveren en van hun successen te laten genieten. Dit typeert jou en ook jouw manier van onderzoek doen. Wat ik knap vind is dat jij als geen ander in staat bent om naar 'the bigger picture' te kijken en om zo nu en dan even een stapje terug te doen. Verder heb ik bewondering voor de manier waarop je ingewikkeld lijkende onderwerpen voor de leek erg begrijpelijk op kunt schrijven. Ik heb veel van je geleerd en ben blij dat ik dat ook de komende jaren mag blijven doen!

Beste Dr. A. Kwee, beste Anneke, jij was degene die mij 3 jaar geleden opbelde om te vertellen dat ik mijn promotieonderzoek mocht beginnen! Na een stage bij PRN kreeg ik nu de kans om direct met clinici samen te werken en daardoor dichter bij de praktijk te staan. Jij bent mijn link naar de praktijk. Ik heb bewondering voor je vakinhoudelijke kennis, je eerlijkheid en je onaflatende persoonlijke belangstelling. Zelfs in tijden dat je het erg druk had met je dagelijkse werkzaamheden, interviews, misleidende krantenkoppen en een onvergetelijk optreden bij Nieuwsuur had je altijd tijd voor mij. Ik vond het erg prettig om met je samen te werken en dus hoop ik dat, ondanks dat de officiële verbinding na mijn promotie weg zal vallen, we ook in de toekomst nog mooie projecten samen kunnen doen.

Beste Dr. R.H.H. Groenwold, beste Rolf, wat ben ik blij dat Carl besloot om jou aan mijn begeleidingsteam toe te voegen. Net als Anneke ben je een bekend gezicht, maar dan vooral voor de mensen die dagelijks de trap nemen in de D-vleugel van het AZU. De afgelopen jaren

heb ik je regelmatig gestalked, zowel in persoon als via e-mail, en nooit stond ik voor een dichte deur. Altijd nam je de tijd om even met mij mee te denken en je wist op iedere vraag het antwoord door je schijnbaar oneindige kennis. Door jouw snelle en bruikbare reacties was stagnatie onmogelijk. Ik ben je dankbaar voor je onuitputtelijke feedback en de prettige manier van samenwerken. Daarom hoop ik dat we de komende jaren nog vaker samen projecten mogen ondernemen.

Beste Prof. E.W. Steyerberg, beste Ewout, jij bent degene die mijn switch van de biomedische technologie naar de klinische epidemiologie mogelijk heeft gemaakt. Je gaf me de kans, in de vorm van een stage, om eens te ruiken aan epidemiologisch en dan in het bijzonder predictieonderzoek. Door de gezamenlijke begeleiding van jou, Yvonne en Chantal is dit project een succes geworden en heeft dit er mede voor gezorgd dat ik later mijn promotieonderzoek aan het Julius Centrum heb kunnen starten. Hier wil ik jullie alle drie voor bedanken.

Beste collega's van de methodologieclub van het Julius Centrum, ik heb altijd met belangstelling geluisterd naar jullie onderzoeken en ben jullie dankbaar voor jullie input tijdens de overleggen en paperbesprekingen.

Beste Paulien, Gerdien, Liselotte, Thomas, Stan, Sanne, Carla, Frederiek, Evelien, Lennie, Charles, Patricia, Judith, Chantal, Irene, Corine, Nienke, Mirjam, Mart en anderen die ik nu vast vergeet, ik heb genoten van alle lunches, de promovenski, de borrels en etentjes.

Lieve Irene en Nienke, de afgelopen drie jaar hebben we samen op één kamer gezeten en in die tijd is er een hoop gebeurd. Irene is zelfs moeder geworden. Ik vond het fijn om met jullie op de kamer te zitten (ook vanwege de hoeveelheid snoep!). We hebben een hoop met elkaar gedeeld, mooie maar ook minder mooie momenten. Ik hoop jullie beiden nog regelmatig tegen te komen in het Julius. Veel succes en geluk met jullie promotieonderzoek en verdere ondernemingen!

Uiteraard heb ik ook veel te danken aan het Verloskundig Consortium. Beste Maya, Zelda en Cynthia, ik vind het knap hoe jullie ervoor zorgen dat het trialbureau zo vloeiend blijft lopen met al dat werk wat jullie in te weinig tijd moeten doen. Ik wil alle researchmedewerkers bedanken voor hun inzet voor het consortium en het counselen van deelnemers en verzamelen van data. Data die ik, nadat het eerder als hoofdartikel gepubliceerd werd, weer heb kunnen gebruiken om mijn eigen onderzoeken te kunnen doen. En dan wil ik natuurlijk alle onderzoekers bedanken: Gert-Jan, Katrien, Brenda, Sophie, Carolien, Joepe, Arianne, Michelle, Anneloes, Lidewij, Jelle, Hannah, Raïssa, Parvin, Sander, Emily, Floortje, Jolande, Babette, Joost, Jantien, Thomas, Karst, Marta, David, Stijn, Kim, Nico en anderen. Ik heb het met jullie ontzettend naar mijn zin gehad binnen het Consortium. De kasteelcursussen en buitenlandse congressen waren onvergetelijk! Arianne, Carolien, Sophie en Joepe ik heb met alle plezier met jullie

samengewerkt om te komen tot een mooie publicatie van jullie hoofdartikel. Het is toch mooi dat we elkaar zo met onze eigen ervaring en expertise hebben kunnen helpen!

I would like to acknowledge all collaborators of the individual participant data meta-analyses, on progestogens in twins and ST-analysis in women with a singleton term pregnancy in cephalic presentation, and the collaborators on Chapters 2.5 and 4.1. I am very grateful that I had the opportunity to collaborate with some of the leading researchers within their field of expertise. Your input has been incredible and greatly improved the quality of the studies.

Beste Jeroen, ondanks dat we elkaar weinig zien vind ik het toch erg mooi dat jij samen met Nienke mijn paranimf bent. We hebben het beiden ontzettend druk, maar ik hoop dat we elkaar in de toekomst vaker kunnen zien.

En dan mijn vrienden uit 't Fèr: Lieve Simon, Jolien, Eric, Camilla, Junior, Minna, Robin, Charlene, Remco, Merel, Boy, Jorn, Sanne, Erik-Jan, Winnie, Gaby, Frank, Mike, Carlijn, Jeffrey, Mariska en Tommy! Vaak hoor ik dat het toch wel uniek is wat wij als groep hebben; we gaan met 21 vrienden een weekend naar de Ardennen of met 18 man 1,5 week naar Tenerife. Ik wil jullie via deze weg bedanken voor alle ontspannen momenten die ik tijdens mijn promotieonderzoek goed kon gebruiken. Dat er nog maar vele gezellige momenten mogen volgen!

Lieve Jos, Jeannette en Jasper, of zal ik zeggen schoonouders en zwager?! Want na meer dan tien jaar samen met Stephanie kunnen we daar toch wel van spreken. Ik wil jullie bij deze bedanken voor jullie steun en vertrouwen in mij, niet alleen als vriend van jullie dochter/zus, maar ook als persoon. En natuurlijk de pokeravondjes. Ik realiseer me dat ik me gelukkig mag prijzen met een dergelijk liefdevol tweede thuis.

Lieve Pa, Ma, Martin en Lizzy, ik wil jullie bedanken voor jullie eindeloze vertrouwen, steun en stimulatie, niet alleen in mijn promotieonderzoek maar ook de vele jaren daarvoor. Ik begon mijn promotieonderzoek na terugkomst uit Eindhoven nog vanuit jullie huis, maar na een jaar was het eigen huis daar. Ondanks dat Stephanie en ik onze eigen stek hebben is het nog altijd heel fijn om thuis te komen en ik hoop dit nog heel lang te kunnen doen, net als de uitstapjes met zijn zessen. Bedankt voor alles!

Lieve Stephanie, mijn allerliefste. Het leven van een promovendus en diens partner gaat niet altijd over rozen. Hier en daar een avond of weekend door werken (zeker tegen het einde) heeft ons toch de nodige tijd samen gekost. Ook dat ik soms met mijn hoofd ergens anders zat heb je maar moeten slikken. Desondanks heb je me altijd gesteund en daar ben ik je ontzettend dankbaar voor. Ondertussen hebben we ook een verhuizing en bijbehorende verbouwing tot een goed einde gebracht. Samen kunnen wij alles aan, ik hou van je! Everyday for the rest of our lives!

Curriculum Vitae

Ewoud Schuit was born on the 14th of April 1985 in Gorinchem. In 2003 he completed pre-university education at the Sint-Oelbertgymnasium in Oosterhout. The same year he started as a student at the department of Biomedical Engineering at the Eindhoven University of Technology. He obtained his Bachelor of Science degree in 2007 and his Master of Science degree in 2009. After graduation, Ewoud started an internship at the Center of Medical Decision Sciences at the Erasmus Medical Center under supervision of Prof. Ewout Steyerberg and Dr. Yvonne Vergouwe in collaboration with the Netherlands Perinatal Registry by person of Dr. Chantal Hukkelhoven. In this research project he validated two prognostic models, developed in the United Kingdom to predict the risk of stillbirth and neonatal death in very preterm births, in the Dutch population. After completion of this internship he started his PhD studies at the Julius Center for Health Sciences and Primary Care under supervision of Prof. Carl Moons and Dr. Rolf Groenwold in close collaboration with the nationwide consortium for women's health research by person of Prof. Ben Willem Mol and Dr. Anneke Kwee. The results of this work are described in this thesis. In 2012 he obtained his Master of Science degree after completion of the Postgraduate Master of Clinical Epidemiology at the Utrecht University.