# Risk prediction models in perinatal epidemiology

Jennifer Hutcheon, PhD
Department of Obstetrics & Gynaecology
University of British Columbia
Vancouver, Canada

Laura Schummers, SM, SD student
Department of Epidemiology
Harvard School of Public Health
Boston, MA

"If I can lose 20 pounds before becoming pregnant, how much will that reduce my risk of having pregnancy complications?"

# The woman

- BMI 37
- Age 32
- Nulliparous
- No diabetes or hypertension
- Sister had preeclampsia
- Family history of heart disease

# Risk prediction

- Many epidemiologic studies have estimated the risk of pregnancy complications among obese women *vs.* normal weight women
- Ideally, would be able to give an estimate of risk that takes her particular characteristics into account
  - Individualized risk prediction research

# Outline

- Introduction to risk prediction models
  - Differences from etiologic models
- Steps in building a risk prediction model
- Example: Pre-pregnancy weight loss counseling

# Prediction and prognosis

- Prognosis: foretelling the course of a disease
- From the Greek *prognostikos* (of knowledge beforehand).
  - pro (before)
  - gnosis (a knowing)

# Prediction and prognosis

- Achieved by creating a multivariable regression model ("prediction model") that predicts the probability of an outcome by combining information from multiple predictors
- Regression equation can then be applied to calculate the risk (predicted probability) of an outcome for a given woman

# Examples of risk prediction models

- Framingham (cardiovascular disease)
- APACHE (mortality in adult ICU)
- SNAP, SNAP-II (neonatal mortality)
- Prediction of spontaneous pregnancy in subfertile couples
- Prediction of VBAC success (vaginal birth after cesarean) in women with previous cesarean delivery

# Our interest

- Prediction of maternal and neonatal pregnancy complications based on BMI and other clinical characteristics at the time of pre-pregnancy counseling
  - At current BMI
  - After weight loss

# Prediction vs etiologic models

- Regression modeling methods taught in epidemiology tend to focus on etiologic research
- In etiologic research:
  - Goal is to understand if an exposure is causally related with an outcome (disease etiology)
  - Quantify the magnitude and direction of association between an exposure and the outcome
  - Need to control for confounding variables distorting the effect of the exposure

# Etiologic models

- Need to understand relationship between variables:
  - Confounders
  - Effect modifiers
  - Mediators
  - Common effects
- Relationships used to determine which variables should be included in regression model
- Evaluate the independent effect of the exposure by looking at the magnitude and direction of the odds ratio, relative risk, hazard ratio, etc.

# Prediction models

- Need different strategies for:

1) Variable selection

2) Variable evaluation

3) Interpretation of regression parameters

# Prediction models

## 1) Variable selection for prediction models

- <u>Causal relationship is unimportant</u>
  - Some very good predictors may be non-causal
    - Tumour markers in cancer progression
    - Skin colour in Apgar score
    - Past obstetrical history

# Prediction models

## 1) Variable selection for prediction models

- Temporality of variables is critical
  - Must consider what information will be available when outcome will be predicted
  - Proximal variables may improve model's predictive ability, but should not be included if not known at the time of prediction
    - Birth weight may predict VBAC success, but is not known before delivery, thus not useful

# Prediction models

## 2) Variable evaluation

- Variables in a prediction model should not be evaluated for their predictive ability using measures of association (e.g., odds ratio, relative risk, risk difference)

# Prediction models

- Prediction models need to be evaluated by assessing:
  - sensitivity
  - specificity
  - positive predictive value
  - negative predictive value, or
  - likelihood ratios

# Prediction models

- Predictors with identical odds ratios can have very different values of sensitivity and specificity:

| | *Outcome* | *No outcome* |
|---|---|---|
| *High risk* | 300 | 190 |
| *Low risk* | 5 | 5 |

OR= (300 x 5)/(190 x 5) =**1.6**

| | *Outcome* | *No outcome* |
|---|---|---|
| *High risk* | 5 | 5 |
| *Low risk* | 190 | 300 |

OR=(300 x 5)/(190 x 5) =**1.6**

# Prediction models

- Predictors with identical odds ratios can have very different values of sensitivity and specificity:

| | *Outcome* | *No outcome* |
|---|---|---|
| *High risk* | 300 | 190 |
| *Low risk* | 5 | 5 |

OR= (300 x 5)/(190 x 5) =**1.6**

Sensitivity= 300/305 = **98%**

| | *Outcome* | *No outcome* |
|---|---|---|
| *High risk* | 5 | 5 |
| *Low risk* | 190 | 300 |

OR=(300 x 5)/(190 x 5) =**1.6**

Sensitivity= 5/195 = **3%**

# Prediction models

- An odds ratio alone doesn’t provide adequate information to evaluate predictive ability
- Predictive value is influenced by how frequently a predictor occurs in the population
  - A predictor may have an extremely high odds ratio, but be so rare in the population that it is not a useful tool to predict adverse outcomes at the population level
  - A predictor may have a small odds ratio, but improve model’s predictive ability if it is common in the population

# Prediction models

3) **Interpretation of regression parameters**

- Not interested in interpreting any of the parameters in a prediction model
  - Rather, the model is interpreted as a whole in terms of predictive ability
  - No penalty for more complex data transformations

# Prediction vs etiologic models

**Take home messages:**

1. Model building strategies for risk prediction models differ from those for etiologic models
   - Causal relationship between predictors unimportant
   - Time at which predictors available very important
2. Odds ratios, relative risks are not sufficient to evaluate predictive ability
   - Need to use sensitivity, specificity, predictive values, and likelihood ratios

# Steps in prediction model building

1. Selection of predictors
2. Evaluation of model performance
3. Check for overfitting to the study dataset
4. Validation of model in a different population

# Our example

- Study population
  - British Columbia Perinatal Database Registry
    - Provincial population-based registry containing data abstracted from maternal and newborn medical records
  - 229,387 singleton pregnancies in 2004-2010 with available pre-pregnancy BMI
  - Restricted data to overweight/obese for prediction modeling, n= 75,225

# Our example

- Outcomes examined
  - Preeclampsia
  - Gestational diabetes
  - Macrosomia (birth weight ≥ 4500 g)
  - Shoulder dystocia
  - Cesarean delivery
  - Postpartum hemorrhage
  - Maternal mortality /severe morbidity
  - Stillbirth
  - NICU stay ≥ 48 hours
  - Alcohol/illicit drug use
  - In-hospital newborn mortality

  - Focus on **stillbirth** for this workshop

# 1. Selection of predictors

i. Create list of candidate predictors

ii. Evaluate variable quality and missingness

iii. Consider collinearity

iv. Determine final predictors

# 1. Selection of predictors

## i. Create list of candidate predictors

- This part is easy!

Candidate pre-pregnancy predictors of stillbirth

- Maternal BMI
- Maternal age
- Parity
- Smoking
- Pre-pregnancy diabetes
- Chronic hypertension
- History of neonatal death
- History of stillbirth
- # prior spontaneous abortions
- Alcohol/illicit drug use
- Maternal education

- All known before pregnancy (temporality)
- Not all causal (e.g. medical history)

# 1. Selection of predictors

## ii. Evaluate variable quality and missingness

- Clearly, predictors with missing observations are less desirable than those with complete data
- Missing data may identify predictors that are less useful in the real-world setting
  - 24 hour urine protein: won't wait 24 hours to deliver if other clinical signs suggest need for immediate delivery

# 1. Selection of predictors

## iii. Collinearity

- Predictors are often strongly correlated
  - Diastolic BP & Systolic BP
  - Anthropometric parameters: BMI, % body fat, abdominal circumference
- Can create difficulties in estimating regression coefficients
- Don’t collect unnecessary variables or include collinear variables in model

# 1. Selection of predictors

## iii. Collinearity

- May either combine variables or pick one based on:
  - Clinical knowledge
  - Cost
  - Logistical/feasibility issues

# 1. Selection of predictors

## iv. Determine final predictors

- No consensus on best way to select predictors
- Two general approaches
  - a) Full model approach
    - Include all predictors (after considering data quality, missingness, and collinearity)
  - b) Significance testing approach
    - Final predictors selected based on statistical significance (p-value<predetermined threshold)

# 1. Selection of predictors

**iv. a) Full model approach:**

- Theoretically, best approach in terms of minimizing bias and minimizing overfitting
- Works well if:
  - predictors known based on scientific literature
  - number of predictors is small
- In practice, often impractical
  - not feasible to define full model

# 1. Selection of predictors

## iv. b) Significance testing approach

- Common approach is to eliminate predictors based on univariable relationships with outcome
  - Keep if p-value <0.05 or <0.10
  - Stepwise regression methods
    - (e.g. forwards/backwards stepwise regression)

# 1. Selection of predictors

iv. b) **Significance testing approach cont'd**

- Selection of predictors based on p-value known to produce selection bias and overfitting
  - Non-significant variables (especially in smaller datasets) may still help predict outcome
  - Simulations have shown 'noise' variables do little to interfere with predictive ability
- Some solutions:
  - More liberal thresholds (e.g. <0.2)
  - Full model approach using penalized regression

# Our example

## i. Candidate predictors:

- Identified all available variables expected to be associated with stillbirth risk
  - Based on literature review and clinical opinion (of our clinician team member)

## ii. Data quality/missingness

- Maternal education, alcohol use, illicit drug use omitted due to missingness and data quality concerns

## iii. Collinearity

- Was not a concern with remaining predictors

# Our example

## iv. Determine final predictors

- We used a full model approach, due to:
  - Large sample size (population-based data set)
    - Relevant sample size is #events, not #pregnancies
  - Small number of candidate predictors
    - Large, population-based data sets usually less detailed

# 2. Evaluate model’s performance

# 2. Evaluate model's performance

a) **Calibration:** how well does the model's predicted values compare with actual outcomes

b) **Risk stratification capacity:** how well does the model's predictions group the population into clinically relevant risk categories

c) **Discrimination:** how well can the model's predictions separate those who have an outcome from those who don't

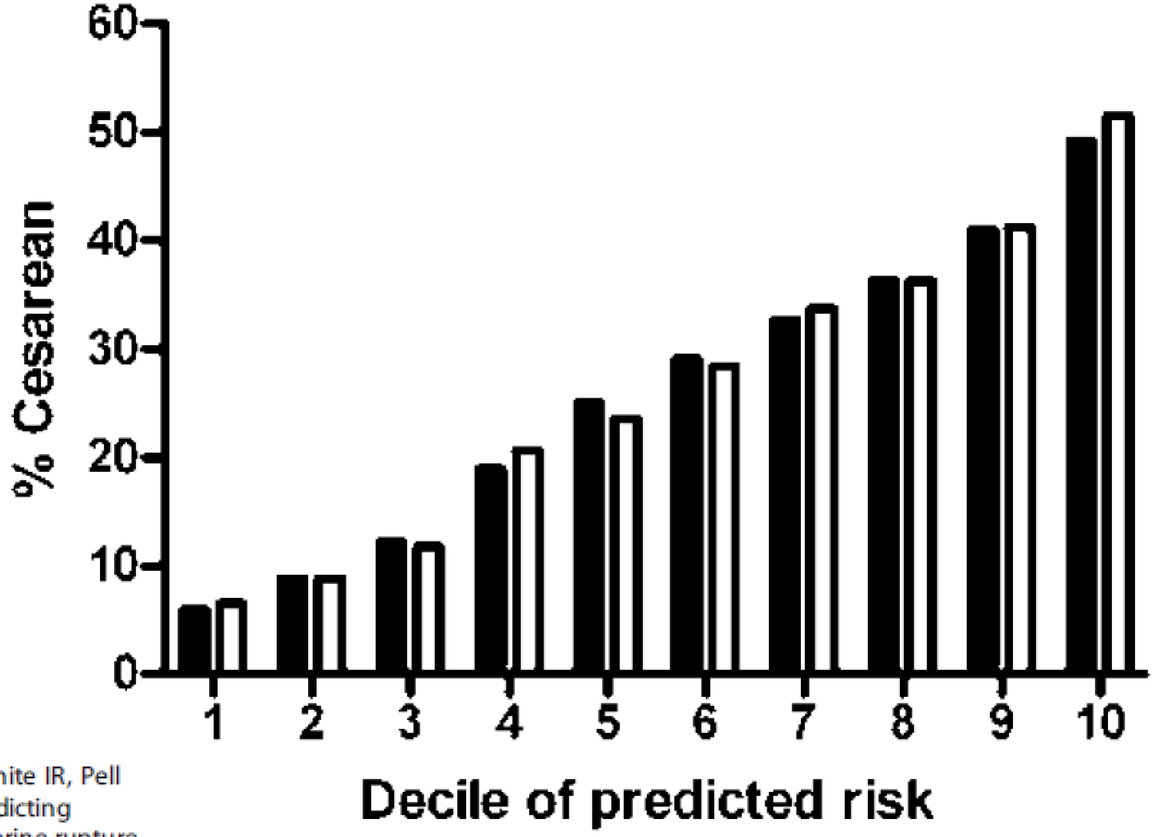
# 2. Evaluate model's performance

a) **Calibration:** how well do the model's predicted values compare with actual outcomes

- Typical approach:
  - Divide population into 10 groups (deciles) based on predicted risk (probability)
  - Calculate predicted and observed risk within each group
  - Compare predicted vs. observed visually (+/- Hosmer-Lemeshow goodness of fit test)

# 2. Evaluate model's performance

a) **Calibration:**



**Citation:** Smith GCS, White IR, Pell JP, Dobbie R (2005) Predicting cesarean section and uterine rupture among women attempting vaginal birth after prior cesarean section. PLoS Med 2(9): e252.

# 2. Evaluate model’s performance

a) **Calibration:**

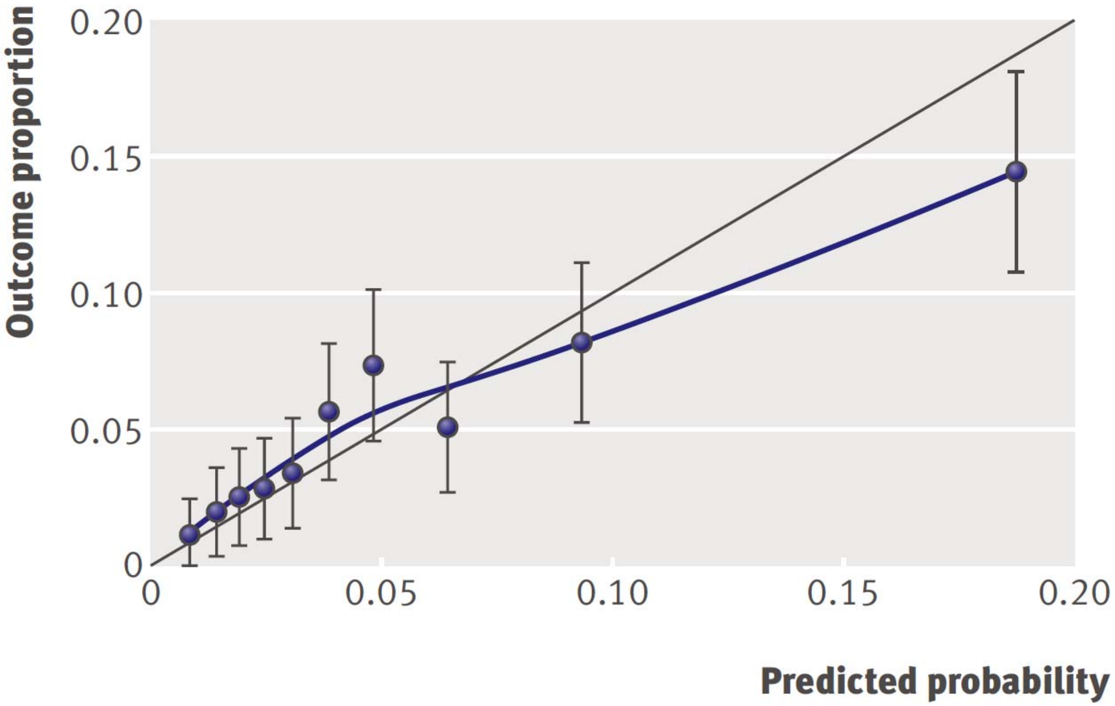


**Fig 3** | Observed rate of pre-eclampsia compared with predicted probabilities of pre-eclampsia based on clinical risk factors model

*BMJ* 2011;342:d1875

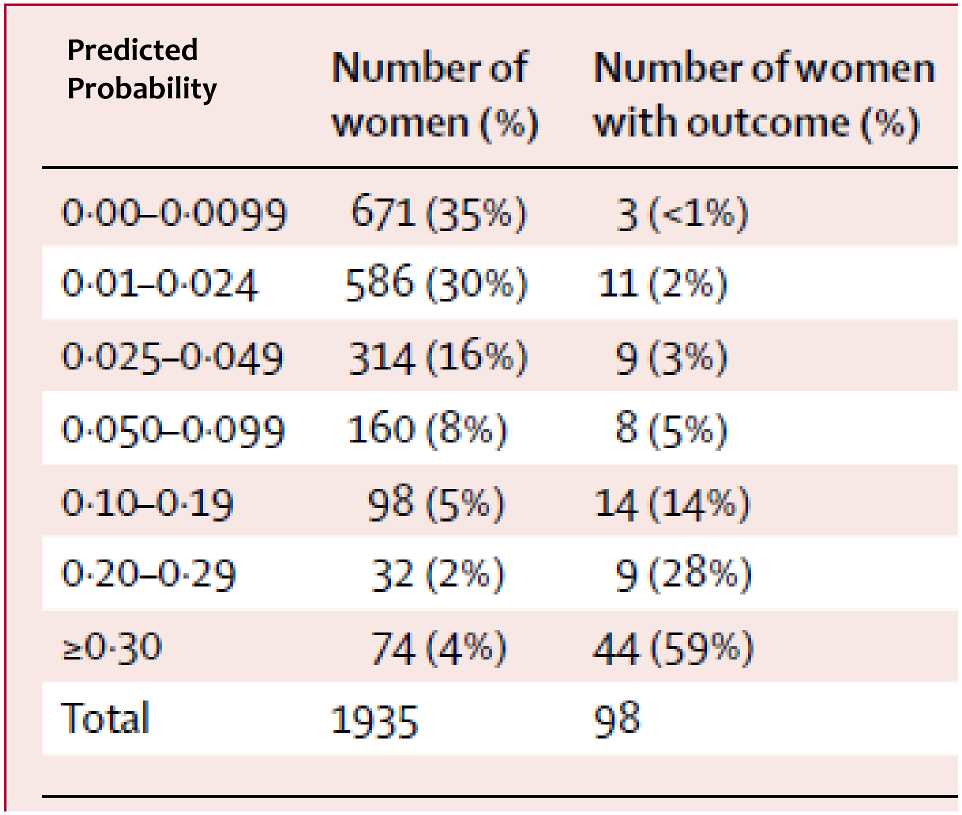
# 2. Evaluate model’s performance

b ) **Risk stratification capacity:**

- Risk stratification capacity examines how the model assigns the population into clinically distinct subgroups
- Ideal risk prediction model would divide the population into ‘minimal risk’ or ‘high risk’ groups
  - Allows surveillance and interventions to be appropriately focused

# 2. Evaluate model’s performance

## b ) **Risk stratification capacity:**



| Predicted Probability | Number of women (%) | Number of women with outcome (%) |
|---|---|---|
| 0·00–0·0099 | 671 (35%) | 3 (<1%) |
| 0·01–0·024 | 586 (30%) | 11 (2%) |
| 0·025–0·049 | 314 (16%) | 9 (3%) |
| 0·050–0·099 | 160 (8%) | 8 (5%) |
| 0·10–0·19 | 98 (5%) | 14 (14%) |
| 0·20–0·29 | 32 (2%) | 9 (28%) |
| ≥0·30 | 74 (4%) | 44 (59%) |
| Total | 1935 | 98 |

*Lancet* 2011; 377: 219–27

# 2. Evaluate model’s performance

c) **Discrimination:** how well can the model’s predictions separate those who have an outcome from those who don’t

- Sensitivity, Specificity
- Positive predictive value, negative predictive value
- Likelihood ratios
- Area under the receiver operating characteristic curve (AUC-ROC curve)

# 2. Evaluate model’s performance

c) **Discrimination:** how well can the model’s predictions separate those who have an outcome

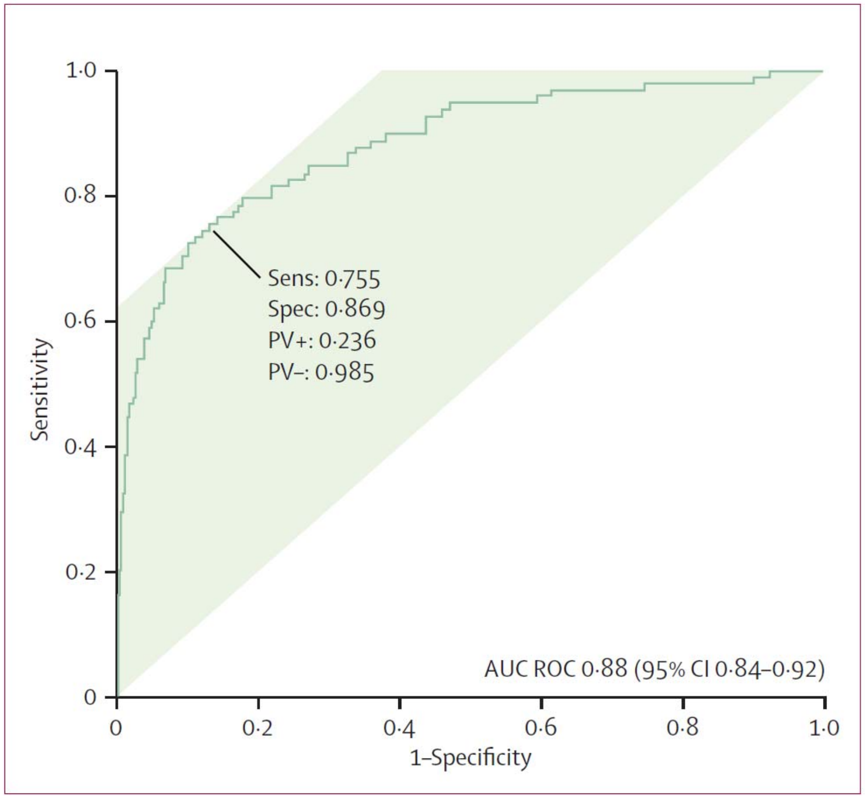
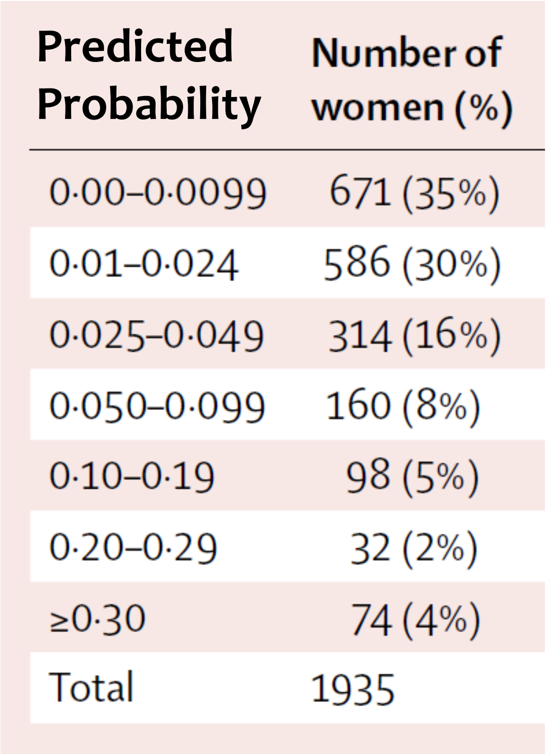


*Figure 1:* Performance of the fullPIERS model, developed with data from first 48 h after eligibility



| Predicted Probability | Number of women (%) | Likelihood ratio |
|---|---|---|
| 0·00–0·0099 | 671 (35%) | 0.08 |
| 0·01–0·024 | 586 (30%) | 0.36 |
| 0·025–0·049 | 314 (16%) | 0.55 |
| 0·050–0·099 | 160 (8%) | 0.98 |
| 0·10–0·19 | 98 (5%) | 3.1 |
| 0·20–0·29 | 32 (2%) | 7.3 |
| ≥0·30 | 74 (4%) | 27.5 |
| Total | 1935 | |

*Lancet* 2011; 377: 219–27

# 3. Evaluating overfitting

- A major concern when building prediction model is that the model may be "overfitted":
  - Model coefficients reflects idiosyncracies of the study dataset rather than true generalizable relationships

# 3. Evaluating overfitting

Strategies to evaluate overfitting:

a) **Data splitting:**

- Split data into two parts
- Build model using one portion (*training* sample)
- Apply model to other portion (*test* sample) and assess predictive ability

Major limitation:

- Very inefficient use of data
- Requires much larger sample sizes

# 3. Evaluating overfitting

b) **Data re-use strategies**

- Use original sample to create multiple 'simulated' samples
- Most common methods:

  i. Bootstrap validating

  ii. Cross-validating

# 3. Evaluating overfitting

b) i. **Bootstrap validating:**

- Sample with replacement from the original dataset
- Re-do all model building steps
- Repeat many times (e.g. 200)
- See how much the measures of model performance ($R^2$, AUC, etc) change between models

# 3. Evaluating overfitting

## b) ii. **Cross-validation**

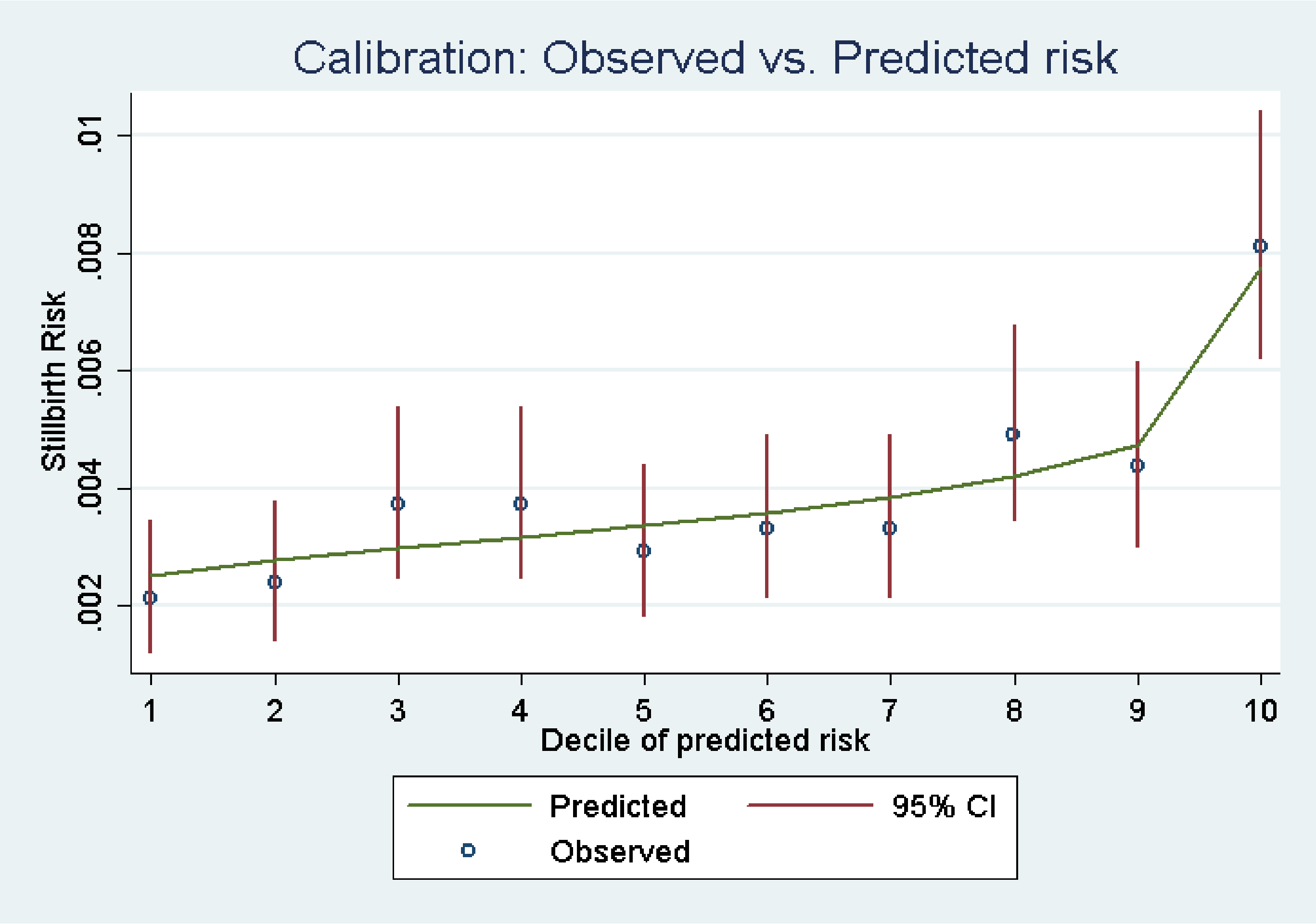
- Variation = data split into multiple groups (e.g., 10)
- Model estimated risk using all groups but one (e.g., 90% of data)
- Predict disease risk calculated in remaining group (e.g.,10%)
- Repeated holding each group out
- Predict disease risks across groups used to assess performance

# 4. External validation

- Need to establish that the model works in women other than those from whom the model was developed
- Requires data from a different cohort with similar characteristics (inclusion, exclusion criteria, etc)
  - Prospective or retrospective
- Apply prediction model equation to new population, and re-calculate calibration capacity, risk stratification, and discrimination

# Example: evaluating model performance

a) **Calibration** step 1 - Evaluate visually:



# Example: evaluating model performance

## a) **Calibration** step 2 - Hosmer-Lemeshow test

- Null hypothesis: The model provides an adequate fit our data;

    Hosmer-Lemeshow chi2(8) = 6.63

    Prob > chi2 = **0.58**

- Fail to reject null hypothesis, conclude fit is adequate

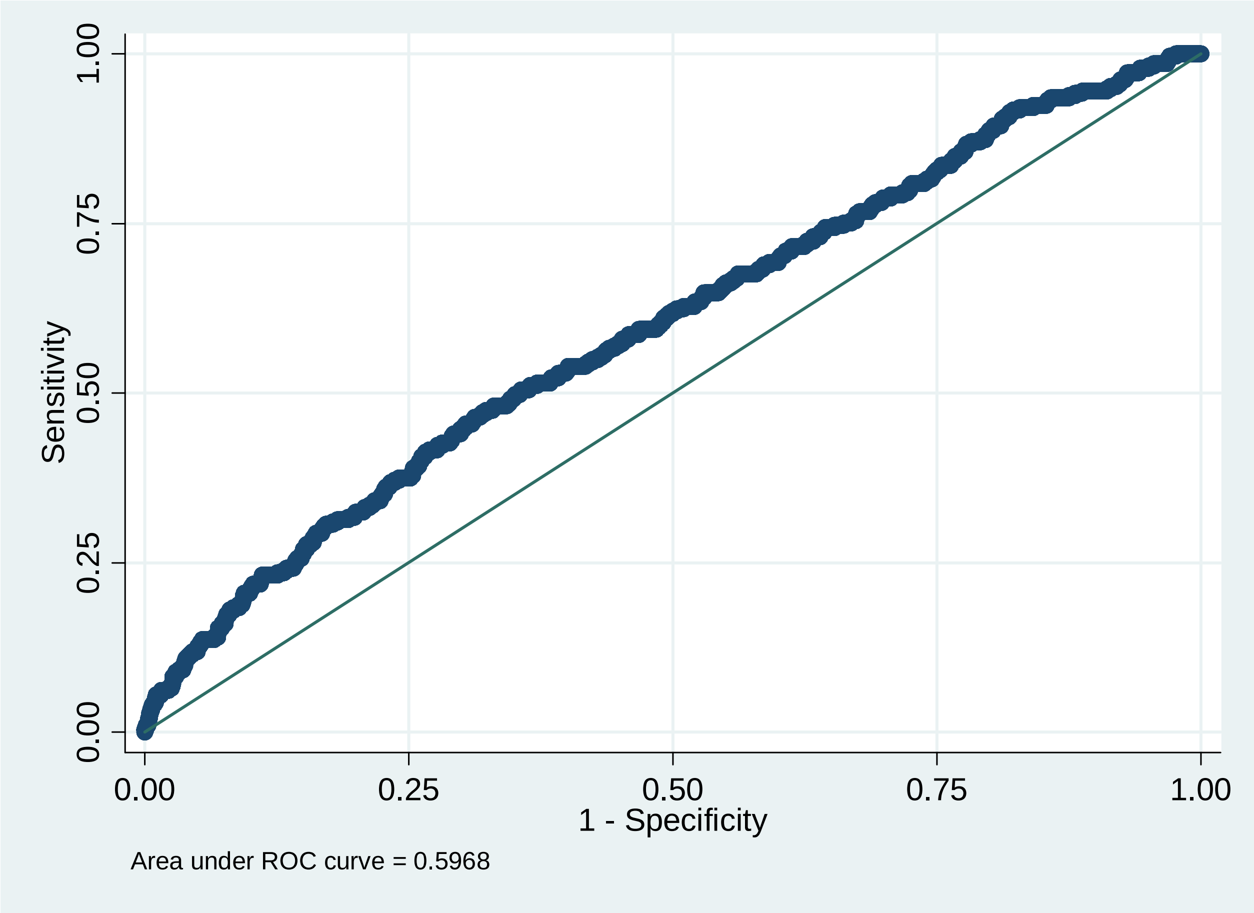
# Example: evaluating model performance

## b) **Risk Stratification capacity**

| Number of women per stratum (% of sample) | Predicted risk (%) | Observed risk n (%) |
|---|---|---|
| 16 (0.02%) | 0.0%-0.20% | 0 (0.0%) |
| 52,807 (70.2%) | 0.20%-0.4% | 164 (0.3%) |
| 18,251 (24.3%) | 0.4%-0.6% | 89 (0.5%) |
| 2,509 (3.3%) | 0.6%-0.8% | 22 (0.9%) |
| 566 (0.8%) | 0.8%-1.0% | 2 (0.4%) |
| 1,076 (1.4%) | >1.0% | 16 (1.5%) |
| 75,225 (100.0%) | 0.4% | 293 (0.4%) |

# Example: evaluating model performance

c) Discrimination: AUC = 0.6



# Example: evaluating model performance

c) **Discrimination**:

| Number of women per stratum (% of sample) | Predicted risk (%) | Observed risk n (%) | Likelihood ratio (95% CI) |
|---|---|---|---|
| 16 (0.02%) | 0.0%-0.20% | 0 (0.0%) | 0 (0.0, 0.0) |
| 52,807 (70.2%) | 0.20%-0.4% | 164 (0.3%) | 0.8 (0.7, 0.9) |
| 18,251 (24.3%) | 0.4%-0.6% | 89 (0.5%) | 1.4 (1.1, 1.7) |
| 2,509 (3.3%) | 0.6%-0.8% | 22 (0.9%) | 2.4 (1.5, 3.6) |
| 566 (0.8%) | 0.8%-1.0% | 2 (0.4%) | 0.9 (0.2, 3.6) |
| 1,076 (1.4%) | >1.0% | 16 (1.5%) | 3.9 (2.4, 6.2) |
| 75,225 (100.0%) | 0.4% | 293 (0.4%) | |

# Example: Evaluating overfitting

- Apparent AUC = 0.60
- Bootstrap validation
  - 200 samples of 75,225 drawn with replacement
  - All model-building steps repeated
  - AUC found for each bootstrap sample
    - Subtract apparent AUC (0.60) from each
    - Take mean of 200 differences (= 0.03) = average optimism
  - Subtract average optimism from apparent AUC
    - 0.60 - 0.03
  - Find optimism-corrected AUC: 0.57

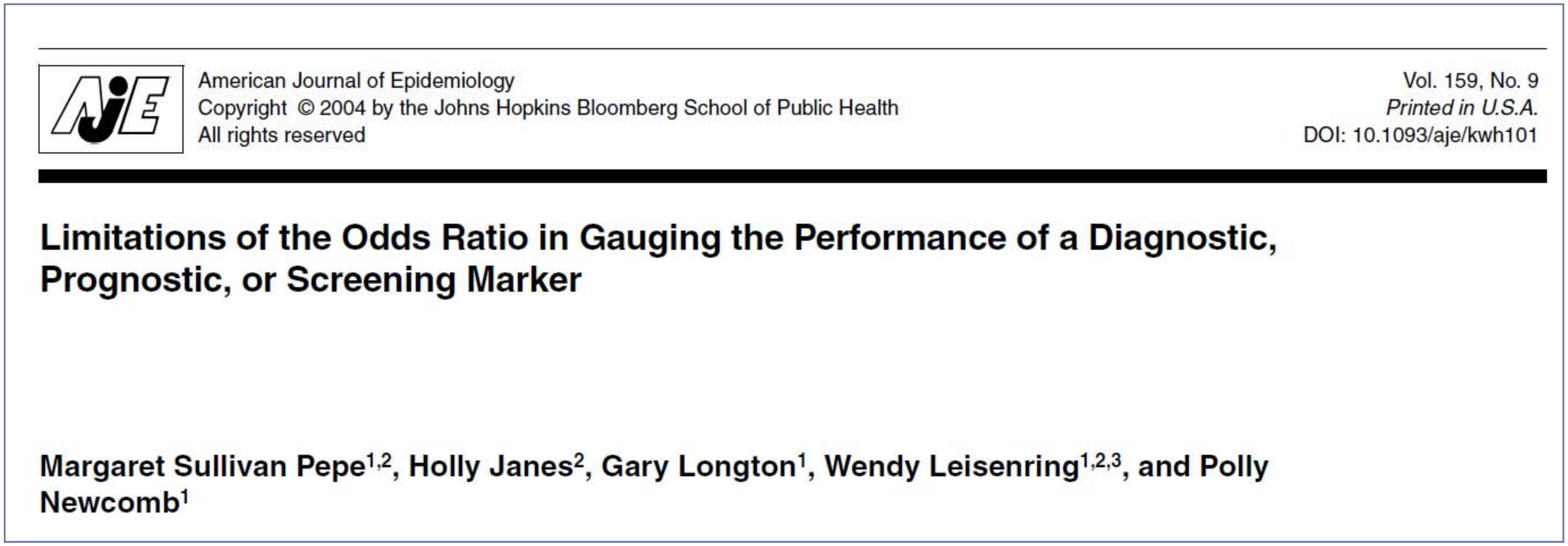
# Example: Conclusions

- Decent calibration
- Inadequate discrimination and risk stratification
- Minimal overfitting, but irrelevant given poor performance
- Did not evaluate external validity
  - Because of poor internal performance
- Conclusion: cannot predict stillbirth on individual level
  - With the available variables in our data

# Most prediction models don’t work

- Individual-level prediction is much harder than finding differences between groups
- Inadequate performance is common
  - Poor discrimination (AUC<0.7)
  - Poor calibration

# Most prediction models don't work



American Journal of Epidemiology
Copyright © 2004 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved

Vol. 159, No. 9
*Printed in U.S.A.*
DOI: 10.1093/aje/kwh101

**Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker**

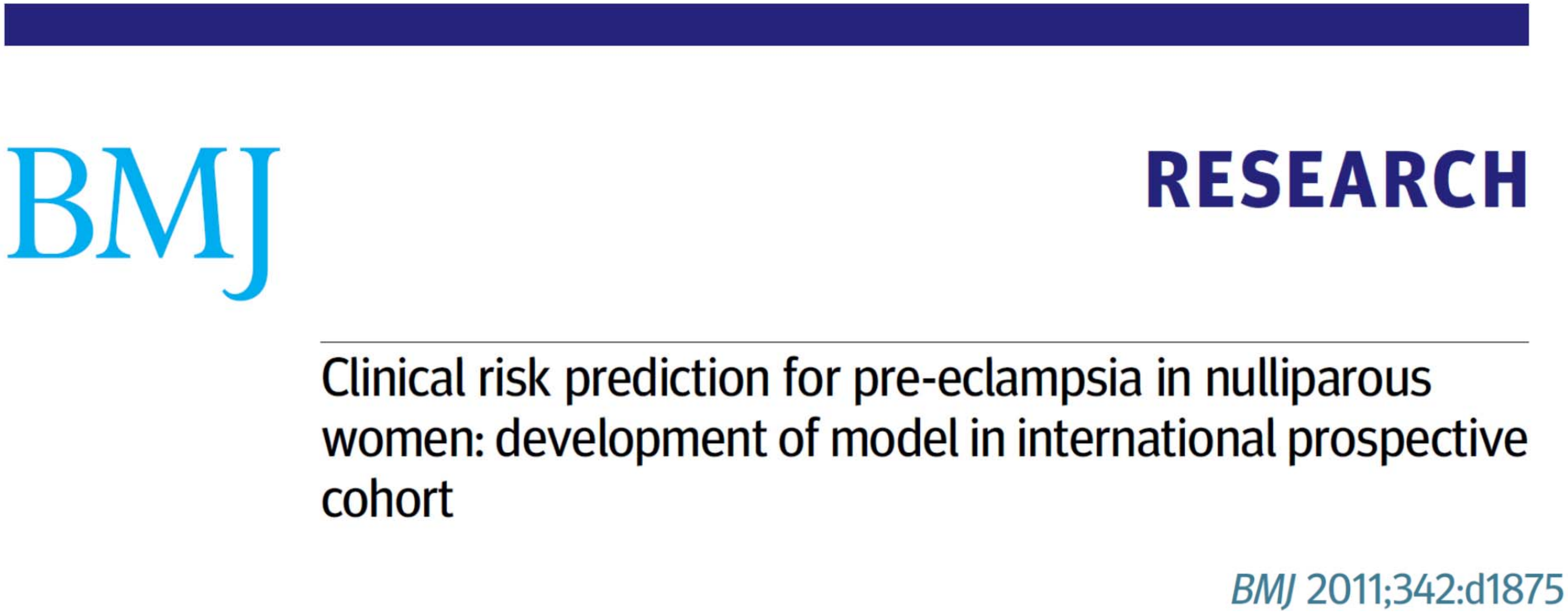
**Margaret Sullivan Pepe[1,2], Holly Janes[2], Gary Longton[1], Wendy Leisenring[1,2,3], and Polly Newcomb[1]**

- Used simulation to evaluate relationship between odds ratio and classification accuracy

# Most prediction models don’t work

- “A marker strongly associated with outcome (or disease) is often assumed to be effective for classifying persons according to their current or future outcome. However, for this assumption to be true, the associated OR must be of a magnitude rarely seen in epidemiologic studies.”
  - Marker with an OR of as high as 3 is in fact a very poor classification tool
  - Found that OR of at least 16 may be needed

# Prediction of pre-eclampsia in nulliparous women



BMJ

RESEARCH

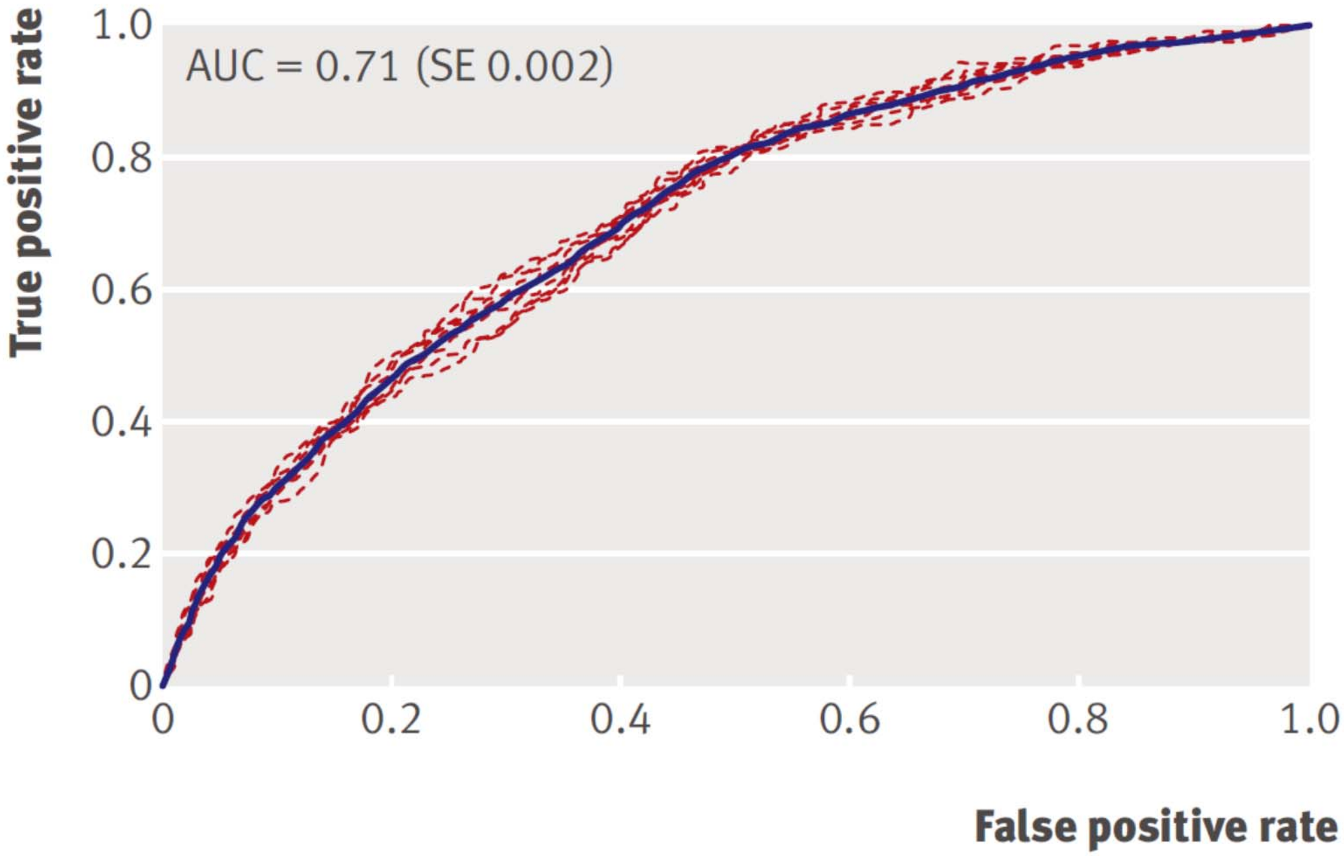
Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort

*BMJ* 2011;342:d1875

# Prediction of pre-eclampsia in nulliparous women

- Tested 39 candidate predictors including most known and potential risk factors for pre-eclampsia:
  - Education, family income, living situation, immigration status
  - Personal and family obstetrical and medical history
  - Diet history and supplement use
  - Lifestyle work, exercise, and sleep
  - Stress and domestic violence
  - Blood glucose and serum lipids
  - Doppler ultrasound

# Prediction of pre-eclampsia in nulliparous women



*BMJ* 2011;342:d1875

# Most prediction models don’t work

Reviews

## New Insights on Vaginal Birth After Cesarean

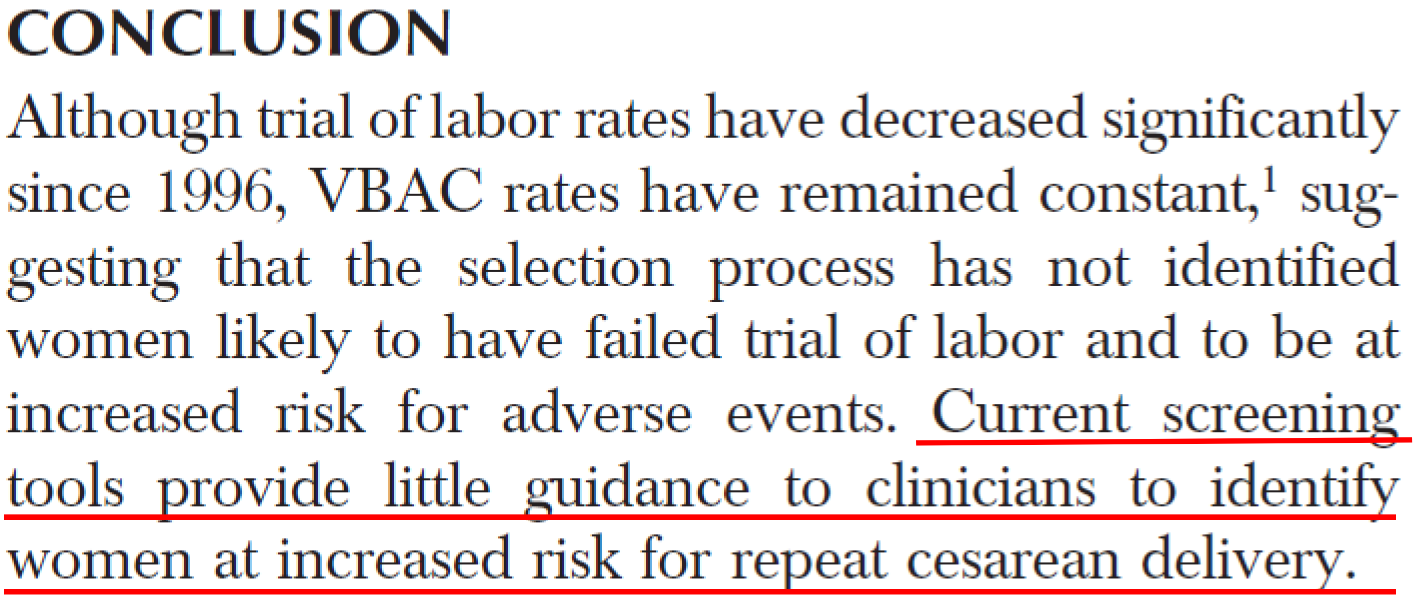
*Can It Be Predicted?*

*Karen B. Eden,* PhD, *Marian McDonagh,* PharmD, *Mary Anna Denman,* MD, *Nicole Marshall,* MD, *Cathy Emeis,* PhD, CNM, *Rongwei Fu,* PhD, *Rosalind Janik,* BA, *Miranda Walker,* MA, *and Jeanne-Marie Guise,* MD, MPH

- Review of 16 published models predicting VBAC (or failed trial of labour)

*(Obstet Gynecol 2010;116:967–81)*

# Most prediction models don’t work



**CONCLUSION**

Although trial of labor rates have decreased significantly since 1996, VBAC rates have remained constant,[1] suggesting that the selection process has not identified women likely to have failed trial of labor and to be at increased risk for adverse events. Current screening tools provide little guidance to clinicians to identify women at increased risk for repeat cesarean delivery.

*(Obstet Gynecol 2010;116:967–81)*

# Most prediction models don’t work

- Many are published (and used) despite lack of adequate evaluation!
    - Dismal external validation

# Most prediction models don’t work

- Challenges in creating good prediction models doesn’t mean that attempt isn’t worthwhile
- Demonstrating that estimates can’t (and shouldn’t) be tailored to each woman is an important message
    - E.g., shouldn’t treat women whose sister had/did not have pre-eclampsia differently

# Most prediction models don’t work

- Aside from the individual-level interpretation, results from prediction models are appealing because they present the *absolute risk* of the outcome of interest
- With cohort data (common in perinatal epidemiology!), we should present absolute measures more
  - We can (and should) use logistic regression to estimate adjusted probabilities!

# Example: Population average risk

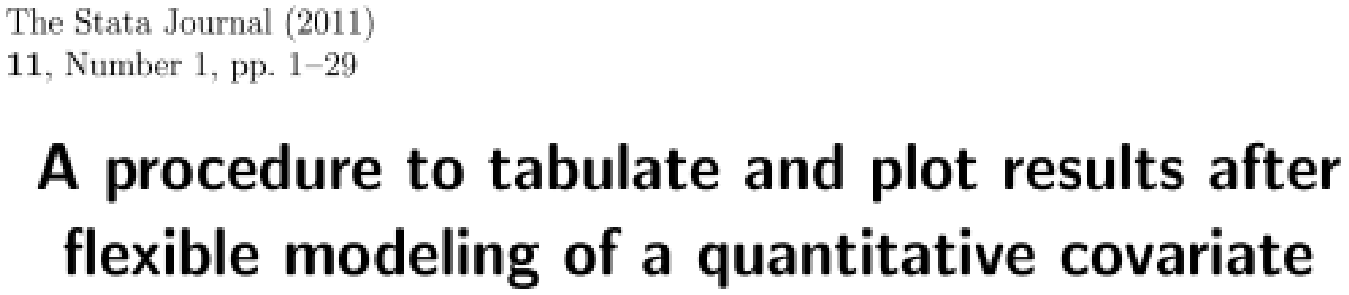
- Prediction model did not perform adequately
  - Precludes individual-level interpretation of results
- We can estimate population-level average risks
  - (Probability, cumulative incidence)
  - Still clinically useful!
  - Probabilities more clinically relevant than odds ratios
  - Smaller gradient of BMI than previously examined

# Example: Analysis

- Logistic regression
  - Adjusted for confounders (not all predictors!)
    - Maternal age, height, parity, smoking
    - NOT adjusted for: number of prior spontaneous abortions or history of stillbirth/neonatal death
- BMI modeled as a continuous variable using a restricted cubic spline
- Predicted odds at each BMI value
- Expit transformation to find predicted probability

# Example: Analysis

- Analyses in Stata 12.0
- Logistic regression
  - BMI defined by 4 variables (cubic spline)
  - XBLC command used to find predicted odds when L=0
  - Orsini & Greenland, 2011



The Stata Journal (2011)
11, Number 1, pp. 1–29

**A procedure to tabulate and plot results after flexible modeling of a quantitative covariate**

- Margins command to specify covariate values

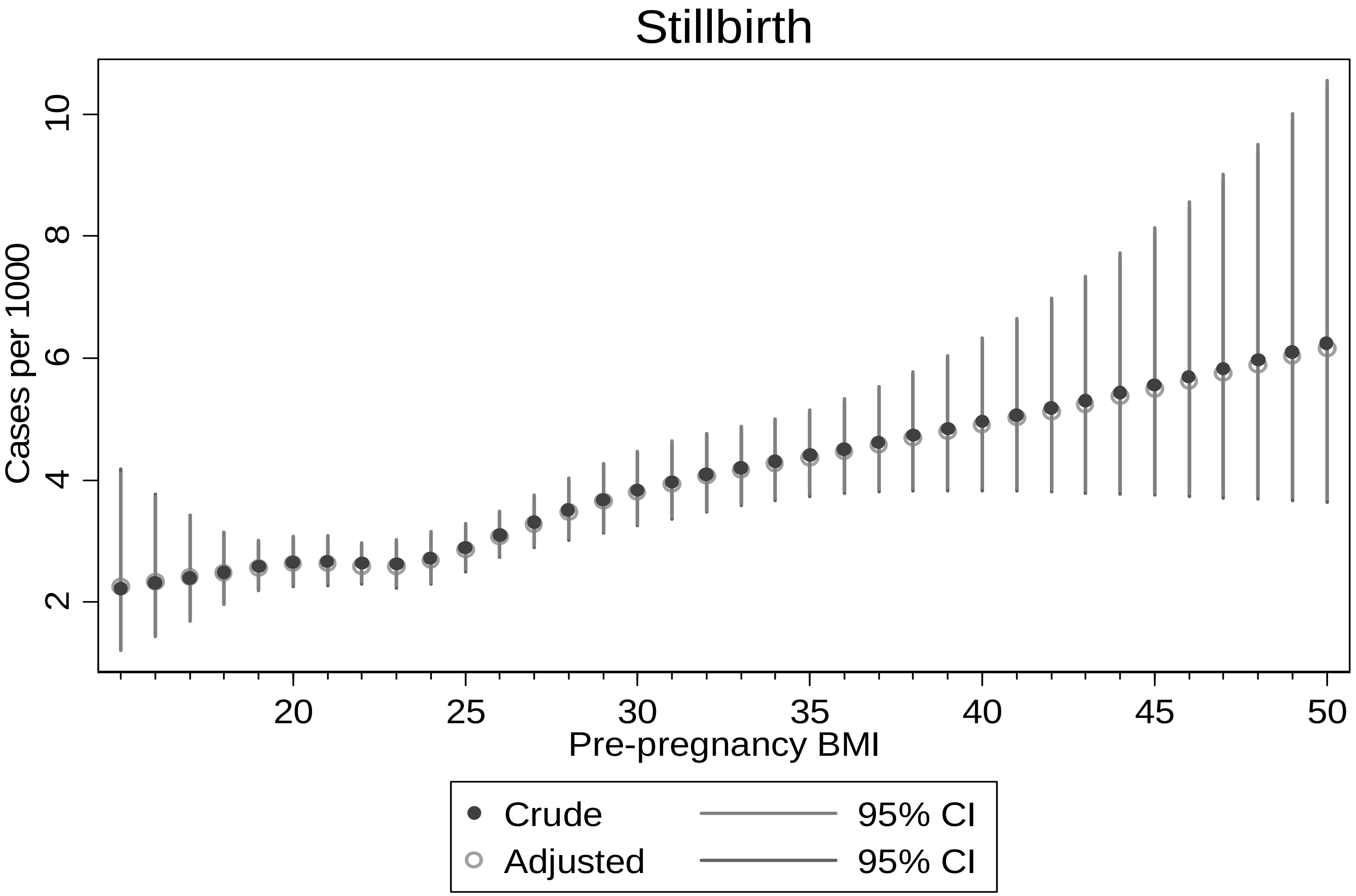
# Example: Analysis

```
*Run logistic regression model
logit stillbirth _bmi1 _bmi2 _bmi3 _bmi4  age smoke_c par_c height_m_c

*Predict odds of stillbirth at each value of BMI
xblc _b*, covname(_bmi1) at(15(1)50) pr eform  /*
*/generate(weight_bmi_sb odds_bmi_sb odds_bmi_sb_lb odds_bmi_sb_ub)

*Transform odds to probabilities
gen prob_bmi_sb =odds_bmi_sb/(1+odds_bmi_sb)
gen prob_bmi_sb_lb=odds_bmi_sb_lb/(1+ odds_bmi_sb_lb)
gen prob_bmi_sb_ub=odds_bmi_sb_ub/(1+ odds_bmi_sb_ub)
```

# Example: Population average risk



# Concluding thoughts

- Growing interest in individualized risk prediction in reproductive and perinatal health
- Need to ensure that appropriate methods used answer risk prediction questions:
  - Calibration, risk stratification, discrimination
  - Potential for overfitting
  - External validation
- Developing a good clinical prediction model is tough!

# Key references & suggested readings

- Patrick Royston et al " Prognosis and prognostic research: Developing a prognostic model" BMJ 2009;338:b604
- Karel Moons et al "Prognosis and prognostic research: what, why, and how?" BMJ 2009; 338:b375
- Ewout W Steyerberg "Clinical Prediction Models" Springer 2009

# Acknowledgements

- Study co-authors
  - Dr. Katherine Himes (PI)
  - Dr. Lisa Bodnar
  - Dr. Ellice Lieberman

CHILD
& FAMILY
RESEARCH
INSTITUTE