Evaluating the clinical value of diagnostic and prognostic markers/tests

K.G.M. Moons

Julius Center for Health Sciences and Primary Care, UMC Utrecht, www.juliuscenter.nl
Summary of various papers/consortia

- Heart: 2 papers on prediction modelling (2012)
Markers/Tests used for many purposes

- Mainly to enhance clinical predictions
  - Diagnosis $\rightarrow$ to predict probability of the result of the more invasive/costly reference (gold) standard
  - Prognosis $\rightarrow$ to predict probability of future conditions/outcomes
  - Monitoring $\rightarrow$ Predict/estimate probability of disease progress or treatment effects
• Diagnostic, Prognostic and Treatment-effect predictions are based on variables measured in a patient → obtained from:
  ◆ Patient history
  ◆ Physical examination
  ◆ Imaging tests
  ◆ Elektrofysiology (ECG, EEG)
  ◆ Blood/urine markers
  ◆ Genetic markers
  ◆ Disease characteristics
Practice

- **No diagnosis/prognosis based on single variable/test/marker**
  - doctors measure many variables $\rightarrow$ combine $\rightarrow$ estimate diagnostic + prognostic probabilities $\rightarrow$ decide upon next test/step

- **Markers/test results only part (sometimes small) of diagnostic, prognostic and treatment-effect predictions**

- **Desired knowledge/evidence for physicians:**
  - Does next test/marker has added value to what I already know from the patient (easy variables)?
  - Or simply: Does it provide added predictive value?
Biomarkers are hot

- # increases per day $\Rightarrow$ greatly vary in
  - Invasiveness / burden
  - Measurement costs
  - Predictive accuracy
Pubmed ‘Biomarkers’: > 650,000 hits

- Proteomics
- Genomics
- Metabolomics

**NycoCard CRP test**

The NycoCard CRP test is a 2-minute Point of Care test to indicate bacterial or viral cause of infection. NycoCard CRP measures C-reactive protein (CRP), an acute phase protein that increases rapidly after onset of infection.

**Test specific information**
- Sample volume: 5 μL
- Assay time: 2 minutes
- Sample material: Whole blood, serum or plasma
- Measuring range: 3 - 250 mg/L for whole blood samples and 5 - 150 mg/L for serum and plasma samples
- Stability at room temperature: 4 weeks
- Kit size: 24 and 48 tests
- NycoCard CRP Control: Positive control provided with the kit

**Clinical use of NycoCard CRP**
- Reduces unnecessary use of antibiotics
- More rapid induction of treatment
- Fewer hospital admissions
- Healthcare cost savings

**ColoCARE®**

ColoCARE is the leading throw-in-the-bowl test for detecting pre-symptomatic occult bleeding caused by gastrointestinal diseases. It is safer, easier and more pleasant to use than traditional guaiac slide tests. Simply place a ColoCARE test pad in the toilet after a bowel movement, watch for a color change, then flush the pad away. It’s clean and disposable, easy for elderly patients to see and interpret, and extremely sensitive, with no increase in the false positive rate. It is more cost-effective than guaiac slide tests because it requires no stool handling, no chemical developers, no laboratory processing, and no mailing of biohazards. Elimination of stool handling overcomes the number one patient objection to occult blood testing, resulting in wider use of the test and leading to greater success in early detection of pathological conditions.

The test pad consists of biodegradable paper chemically treated with a chromogen. The pad is floated on the water surface in the toilet bowl. If detectable blood is present, the hemoglobin reacts with the chromogen, and a blue and/or green color reaction occurs. The test pad has three reaction sites: a large test square and two smaller control squares to verify the system functions properly.
New markers/tests
(all diagnostic/prognostic devices)

- **Problem: Simply enter market**
  - Drugs rigorous phased approach
  - Not diagnostic/prognostic tests: Very liberal guidelines
    - Only safety + ‘performance’ (KEMA/DEKRA → CE approval)
    - Not: Diagnostic or prognostic accuracy → let alone added value

- **Consequences ...**
New markers/tests

1. High availability

- Only increase (‘omics’ area) and ‘point of care’ markers/tests

2. Overtesting

- Reasons: patient satisfaction; fear legal consequences; belief that new ‘toys’ always better

- Overtesting → unnecessary burden to doctors, patients, budgets

- Health care resources not used for those who need most

- Incorrect use: Swan-Ganz; ICP monitoring; preoperative ECG → Only increase in ‘omics’ area and point of care tests
Reasons/Causes

1. Too liberal guidelines for market access

2. Less money involved

3. Methodology to adequately study diagnostic, prognostic, monitoring markers far underdeveloped
   - Not popular → no guidelines for conduct or reporting

4. (As a consequence) → selective reporting (publication bias)
   - Kyzas et al (2005): Review cancer biomarkers: >1900 papers; → 98.5% significant
   - Doctors might think all published devices/markers are important
Hlatky et al, 2009

Criteria for Evaluation of Novel Markers of Cardiovascular Risk

Focus on prognostic cardiovascular markers
Phased approach

- From single testing → do marker levels differ between subjects with vs. without outcome?...

- … to… Quantify added value to existing predictors using so-called multivariable (clinical) prediction models

- …to… Quantify impact/clinical usefulness of such prediction models on decision making and thus patient outcomes
Central issue in current marker research

- **Key words:**
  - Added value → using multivariable analysis and prediction models (see later)
  - Clinical usefulness

- **NOT:** developing/searching new biomarker kits → many out there already
  - Review (Riley et al): 131 biomarkers for prognosis of neuroblastoma (in just few years) → can’t be all relevant
  - Challenge for new markers is to beat existing ones
Single marker/test studies

• By far prevailing in the literature
   Reviews: Kyzas et al (2003;2005); Riley et al (2003 neuroblastoma markers); Lijmer (Jama 1999):

   Aimed at
     Quantifying the marker’s sensitivity, specificity, predictive values
      – Perhaps: comparing 2 markers on difference in sens+spec

     Or even: how often marker values in patients differ from ‘normal values’ (e.g. 2 times the ULN)
For every laboratory test or diagnostic procedure there is a set of fundamental questions that should be asked. Firstly, if the disease is present, what is the probability that the test result will be positive? This leads to the notion of the sensitivity of the test. Secondly, if the disease is absent, what is the probability that the test result will be negative? This question refers to the specificity of the test.

(Campbell MI, Machin D. Medical statistics. a commonsense approach. Chichester: John Wiley & Sons, 1990)
Single marker/test studies

- Identify the sensitivity and specificity of the sign, symptom, or diagnostic test you plan to use. Many are already published and subspecialists worth their salt ought either to know them from their field or be able to track them down.


- What are ‘precautions’ with single test studies and aiming for estimating a marker’s sens and spec?
  - Design
  - Analysis
• **Question:** D-dimer level to determine presence/absence DVT (reference: leg ultrasound)
  - Can D-dimer predict result of leg ultrasound?
  - Test is intended for patients suspected of DVT.

• **Correct single test approach --> prospective**
  - select patients suspected of DVT (red, swollen leg) at office of referring physician
  - Vena punction and measure D-dimer (index test)
  - All undergo leg ultrasound (reference)
  - Quantify sens, spec and predictive values D-dimer
Pittfalls single marker/test studies

Design

• Frequent approach 1:
  1. Take from hospital files all subjects who routinely underwent ultrasound in routine care, and were positive (DVT+)
  2. Select subjects without DVT → healthy controls from general population
  3. Take D-dimer levels of DVT+ from computer; estimate D-dimer level in blood drawn from healthy controls

<table>
<thead>
<tr>
<th></th>
<th>DVT+</th>
<th>DVT-</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-dimer+</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>D-dimer-</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Pitfalls single marker/test studies

Design

<table>
<thead>
<tr>
<th></th>
<th>DVT+</th>
<th>DVT-</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-dimer+</td>
<td>95</td>
<td>5</td>
</tr>
<tr>
<td>D-dimer-</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Perfect accuracy $\rightarrow$ logic $\rightarrow$ discriminating between diseased and healthy controls (two extremes) is easy

Seems efficient design $\rightarrow$
but index test not evaluated in right persons $\rightarrow$ Healthy controls never indicated to receive D-dimer test

- Comparison with normal persons (with normal marker levels) irrelevant and biased accuracy for clinical practice
Design

- **Frequent approach 2:**
  1. Take from hospital files all subjects who routinely underwent ultrasound in routine care, and were positive (DVT+)
  2. Take from hospital files all remaining subjects who routinely underwent ultrasound in routine care, and were negative (DVT-)
  3. Take D-dimer levels of DVT+ and DVT- from computer

<table>
<thead>
<tr>
<th>D-dimer+</th>
<th>DVT+</th>
<th>DVT-</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-dimer+</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>D-dimer-</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Pitfalls single marker/test studies

Design

<table>
<thead>
<tr>
<th></th>
<th>DVT+</th>
<th>DVT-</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-dimer+</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>D-dimer-</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

• Again: efficient design → but bias due to routine care data use:
  ✤ In routine care: reason for referral to next invasive/costly tests (ultrasound) always based on previous test results (D-dimer)
  ✤ Those who underwent ultrasound in practice → notably those with positive D-dimer → more diseased (more often referred)
  ✤ = good clinical practice → bad science!
  ✤ Work-up bias / verification bias / referral bias
Pitfalls single marker/test studies

Design

- Solutions
  - Not simply take marker values from computer and compare between those with positive versus negative reference test result
  - Approach also frequent in radiology/nuclear medicine
  - Collaborate with referring specialists
    - Select cohort of patients intended for using the test
    - All undergo index test and reference standard
Pitfalls single marker/test studies

Analyses

‘Limitations’ sensitivity and specificity

1. Conceptual
   - Reverse probabilities (!= conform practice)

2. Require dichotomisation of test results

3. Assumption they are constant
   - Characteristics of a test --> THE sens and spec of a test
   - Predictive values desired parameters for practice \(\rightarrow\) sens + spec most popular
     - Reason: PVs vary across prevalences and thus populations
     - Sens and Spec not \(\rightarrow\) use Bayes theorem to obtain PVs
Pittfalls single marker/test studies

Analyses

• PVs across populations

<table>
<thead>
<tr>
<th>Setting</th>
<th>Prevalence (Cases/100,000)</th>
<th>Positive Predictive Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General population</td>
<td>35</td>
<td>0.4</td>
</tr>
<tr>
<td>Men, age 75 or greater</td>
<td>500</td>
<td>5.6</td>
</tr>
<tr>
<td>Clinically suspicious prostatic nodule</td>
<td>50,000</td>
<td>93.0</td>
</tr>
</tbody>
</table>


• More diseased --> higher prevalence --> higher PV+

⚠ screening --> clinical population

⚠ note: prevalence determined by patient characteristics
Pitfalls single marker/test studies

Analyses

- Sens and Spec across populations
  - Fletcher Ann Int Med 1986
  - CEA for colon carcinoma

- Severity disease determines sens+spec

- Vary across populations → not constant

- Constant within certain population?
Pittfalls single marker/test studies

Analyses

- 295 patients with chest pain during exercise --> suspected CAD
  - all underwent exercise stress test (index test)
  - Coronary angiography (reference)

Limitations of Sensitivity, Specificity, Likelihood Ratio, and Bayes’ Theorem in Assessing Diagnostic Probabilities: A Clinical Example

Karel G. M. Moons,1–3 Gerrit-Anne van Es,4 Jaap W. Deckers,5 J. Dik F. Habbema,2,3 and Diederick E. Grobbee1,3,6

(Epidemiology 1997;8:12–17)
# Pittfalls single marker/test studies

## Analyses

<table>
<thead>
<tr>
<th>Exercise test versus angiography</th>
<th>ECG</th>
<th>CAD + (n=207)</th>
<th>CAD - (n=88)</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>119 (58%)</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>negative</td>
<td>88</td>
<td>80 (92%)</td>
<td></td>
</tr>
</tbody>
</table>
Pitfalls single marker/test studies
Analyses

<table>
<thead>
<tr>
<th>characteristic</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>64</td>
<td>89</td>
</tr>
<tr>
<td>Female</td>
<td>30</td>
<td>97</td>
</tr>
<tr>
<td>Cholesterol (mmol/l)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.0-6.0</td>
<td>52</td>
<td>88</td>
</tr>
<tr>
<td>6.1-12.0</td>
<td>71</td>
<td>94</td>
</tr>
<tr>
<td>Baseline SBP (mm Hg)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100-140</td>
<td>65</td>
<td>96</td>
</tr>
<tr>
<td>141-240</td>
<td>50</td>
<td>86</td>
</tr>
</tbody>
</table>
Pittfalls single marker/test studies

Analyses

- **Sens + Spec not constant → vary over other test results**
  - Simply because all tests measure same underlying disease → mutually dependent
  - Single level for all patient subgroups does not exist → should not be sought

---

**Sensitivity and Specificity Should Be De-emphasized in Diagnostic Accuracy Studies**

*Acad Radiol* 2003; 10:670–672

Karel G. M. Moons, PhD, Frank E. Harrell, PhD
Pittfalls single marker/test studies

Analyses

- Moreover, sens + spec do not provide knowledge about added or independent value
  - Many examples single test studies showing promising results → not in multivariable analysis (accounting for mutual dependencies)
  - *Helicobacter Pylori test for peptic ulcer in patients with dyspepsia* → by itself good sens and spec → particular subgroups no added value (*Weijnen et al*; *BMJ* 2001)
  - *73 natriuretic peptides for heart failure: ANP, N-terminal ANP, BNP* → all highly significant by themselves → multivariable analysis: only BNP (*Cowie et al*; *Lancet* 1997).
  - *Same for CRP and Heart failure* → not significant anymore when combined with Interleukin-6 and TNF-alpha

*(J Am Coll Cardiol 2010;55:2129-3*
Pitfalls single marker/test studies

Analyses

• As no diagnosis/prognosis is set by one single index test anyway
  ➢ Always combination of multiple test results

  ➢ Always perform multivariable analysis and quantify independent/added value of biomarker to current predictors

  ➢ Compare etiology/causal research → impossible to publish a study on association between risk factor and outcome, without adjustment of other risk factors
Pitfalls single marker/test studies

Applications

Redundancy of Single Diagnostic Test Evaluation
Karel G.M. Moons,1,2,3 Gerri-Anne van Es,4 Bowine C. Michel,5 Harry R. Büller,6 J. Dik F. Habbema,3 (Epidemiology 1999;10:276–281)

- Perhaps two situations of single test approach (e.g. comparing disease versus healthy controls):
  - Early development phase of new marker/test
    - If can’t discriminate → Forget it
  - Markers/tests used in screening of pre-clinical diseases
Can dogs smell bladder cancer?
Knowing added predictive value is desired.
Quantifying independent/added value of markers requires multivariable (clinical prediction) modeling approach

Multivariable clinical prediction models
### Table 9-1. Apgar scoring.

<table>
<thead>
<tr>
<th>Signs</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heartbeat per minute</td>
<td>Absent</td>
<td>Slow (&lt;100)</td>
<td>Over 100</td>
</tr>
<tr>
<td>Respiratory effort</td>
<td>Absent</td>
<td>Slow, irregular</td>
<td>Good, crying</td>
</tr>
<tr>
<td>Muscle tone</td>
<td>Limp</td>
<td>Some flexion of extremities</td>
<td>Active motion</td>
</tr>
<tr>
<td>Reflex irritability</td>
<td>No response</td>
<td>Grimace</td>
<td>Cry or cough</td>
</tr>
<tr>
<td>Color</td>
<td>Blue or pale</td>
<td>Body pink, extremities blue</td>
<td>Completely pink</td>
</tr>
</tbody>
</table>

\[ \Sigma = \text{Apgar score (0-10)} \]
Clinical prediction models

- Convert predictor values of subject to an absolute probability…
  - …of having (!) a particular disease → diagnosis
  - … of developing (!) particular health state → prognosis
  - … within a certain time (hours, days, weeks, years)
  - Dying, complication, disease progression, hospitalised, quality of life, pain, therapy response
Clinical prediction models

• Predictors (for both aims) are:
  - history taking
  - physical examination
  - tests (imaging, ECG, blood markers, genetic ‘markers’)
  - disease severity
  - received therapies
Clinical prediction model

- Presented as:
  - Mathematical formula requiring computer
  - Simple scoring rules (Apgar)
  - Score charts / Nomograms (SCORE / Framingham)
Why using prediction models?

• Diseases have multiple causes, presentations and courses (McShane 2005; Riley 2003; Moons BMJ 2009)

  ▫ A patient’s diagnosis and prognosis rarely based on single predictor

  ▫ Impossible to disentangle and weigh all contributing factors by heart, and to adjust for their mutual influence
Why using prediction models?

- ... Not meant to replace physician, but to complement their clinical intuition

- Assumption:
  - Accurately/objectively estimated probabilities...
  - ...improve physicians’ behaviour / decision making ...
  - ... and thus patient outcome
Prediction models are hot

(Steyerberg 2009)
1000’s examples

- Apgar score
- Framingham risk score
- SCORE
- Euroscore (cardiac surgery)
- Goldman risk index (chest pain)
- Over 60 models for cancer prognosis (e.g. Gail model)
- APACHE score, SAPS score (IC models)
- Ottawa ankle and knee rules
- Reynolds risk score
Welcome to Your Disease Risk, the source on prevention. Here, you can find out your risk of developing five of the most important diseases in the United States and get personalized tips for preventing them.

Developed over the past ten years by world-renowned experts, Your Disease Risk collects the latest scientific evidence on disease risk factors into one easy-to-use tool.

To get started, choose one of the diseases below.

What is your risk?

- Cancer: There’s much more to it than just smoking and lung cancer.
- Diabetes: Over 18 million in the U.S. suffer from it. Take steps now to lower your risk.
- Heart disease: The #1 killer in the U.S. is also one of the most preventable.
- Osteoporosis: Calcium isn’t the only way (or even the best way) to protect yourself.
- Stroke: Most cases of this feared disease can be avoided by lifestyle changes.
Life Expectancy Calculator

Start Here

Your life expectancy is influenced by a number of factors, from your family history to your personal lifestyle. Please begin by entering some basic information about yourself, then select "Family History" to the left.

- Male
- Female

Current age: 

Weight: 
Height: 

Frame size: 
- Small
- Medium
- Large

Education completed:
- High school only
- Some college
- College graduate

How would a friend describe you?
- Easy-going and relaxed
- Aggressive, intense and quick to anger

Get a free life insurance quote.
No one has ever claimed that the results were not valid.

To try this model yourself go to [Business Bankruptcy Predictor](#).
What evidence needed to apply prediction models in practice?

Steps in prediction modeling


- 1. Developing the prediction model
- 2. Validate the model in other subjects
- 3. Update existing models to local situations
- 4. Quantify impact of using a model on doctor’s decision making and patient outcome (cost-effectiveness)
1. Development studies

- Many reviews (G Collins 2010/2011; S Mallet 2010; W Bouwmeester 2012) show that majority of prediction models still poorly developed → in all disciplines

- In fact: no real challenges anymore → Much literature:
  

  - Analysis including quantifying added value of new test (Royston, BMJ 2009; Books by Harrell 2001; Steyerberg 2008; Royston & Sauerbrei 2009)
What evidence needed to apply models in practice?

Steps in prediction modeling

1. Developing the prediction model

2. Validate the model in other subjects

3. Update existing models to local situation

4. Quantify model’s impact on doctor’s decision making and on patient outcome (cost-effectiveness)
Phase 2. Validation studies
Unfortunately scarce
In contrast to development studies: sexy
Phase 2. Validation study characteristics


• Aim: to demonstrate accuracy/performance of original model in subjects not used to develop model
  - Calibration, discrimination (c-index), classification

• Validating a model is not …
  - …Repeat analysis in new data and check if you come up with same predictors, regr. coeffs, predictive performance
  - …Fit the previously found predictors/model and estimate its predictive performance
Phase 2. Validation study characteristics


• Use original developed model \(\rightarrow\) apply (!) to new data \(\rightarrow\) Compare predicted with observed outcomes
  - Discrimination, calibration and classification

• Validation studies thus require that original, developed prediction models properly reported
  - Original beta’s – plus intercept / baseline hazard
  - Not just simplified score (too often still done)
  - Clear definition and measurement method of predictors + outcome (so future researchers can repeat/use them)
  - Reporting guideline underway: TRIPOD (end 2013)
Phase 2. Types of Validation studies


- **4 (increasingly stringent) types:**
  1. Internal validation (in fact part of development phase)
  2. Temporal validation
  3. Geographical validation
  4. Other setting / domain (type of patients)
Types of Validation Studies

1. Internal validation (split sample, bootstrapping)
   - Not random split sample $\rightarrow$ no difference
   - Best = Bootstrapping
   - Note: not new data (Bleeker SE et al, JCE 2002)

2. Temporal validation
   - Same setting, measurements and investigators (often), but later in time
     - Many similarities $\rightarrow$ ‘high’ chance of good performance
   - Split sample: if large database -- split over time
Types of Validation Studies

3. Geographic
   - Other centers + often other investigators
   - Also often other protocols
   - May be – if very large database or combination of data sets (= IPD meta analysis) -- split sample by country

4. Setting/domain/subgroup
   - Secondary → primary care
   - Adults → children
   - Men → women
Types of Validation Studies

• Note temporal, geographic and domain/setting validation can be done:
  ➢ Prospectively
  ➢ Retrospectively using large existing data sets
  ➢ Often called ‘external’ validation

• YES: usually researchers find poor accuracy when validating existing model in their data
  ➢ Key message: suppress your reflexes
  ➢ Do not immediately fit (yet) a new model
Typical Result

- Systematically too high predictions
  - Higher outcome prevalence/incidence in development set
  - Intercept too large for new subjects
Typical Result

Slope plot $< 1.0$

- Low prob too low
- High prob too high
- Typical overfitted model in development set
- Too extreme regression coefficients (OR/HR)
Logical: reasons poor validation


• 1. Different outcome occurrence
• 2. Different patients
3. Different interpretation of predictors or (incorrect) proxies of predictors are used

4. Changes in care over time
   - Improvement in measurements: e.g. imaging tests
     - Previous CTs less accurate than spiral CT for pulmonary embolism detection

5. Original model could have missed important predictor
Reasons poor validation

• **BUT: No matter what reason of poor validation:**
  - Reflex: one develops ‘own new’ model from their validation study data
  - >100 models for brain trauma; >60 models for breast cancer; >100 CVD risk in general population; > 100 diabetes models

• **Understandable:**
  - We finally learned the ‘tricks’ to develop models (in standard software)
  - ‘Own’ model makes you famous (Apgar; Goldman; Gail; Wells)
  - Validation is only to support (citation index of) others
Reasons poor validation


- **Unfortunate habit**
  - Previous knowledge neglected
  - Prediction research becomes completely particularistic
    - Every country, setting, hospital, subgroup, etc.
  - Validation data sets often smaller → even less generalisable models
  - Perhaps new model needed: but likely not!
What evidence needed to apply models in practice?

Steps in prediction modeling

• 1. Developing the prediction model
• 2. Validate the model in other subjects

• 3. Update existing models to local situation
• 4. Quantify model’s impact on doctor’s decision making and on patient outcome (cost-effectiveness)
Phase 3. Updating prediction models

- Recent insights: update/adjust existing model with new data → rather than fitting (‘our’) new model
  - Certainly if validation set is relatively small(er)

- Updating is particularly important when new predictors/markers are found → to be added to existing models: e.g.
  - CRP to Framingham risk model
  - Frequently heard: search for new blood markers
Phase 3. Updating prediction models


- After validation existing model → unsatisfactory accuracy → update → ranges from:
  - Simple adjustment of base line risk (intercept / hazard)
  - Adjusting the regression coefficients of the predictors in model
    - All together in same way (if overfitted model)
    - Different adjustments
  - Adding previously missed or new predictors/markers
Phase 3. Updating prediction models


- Adjust for difference in overall prevalence/incidence (intercept adjustment) is often sufficient

- If also slope different \(\rightarrow\) adjust predictor weights
- Or search for adding/new predictors
Phase 3. Updating prediction models


- Final notes

  - Updating done after (!) model (external) validation → if unsatisfactory accuracy in new subjects
    - Not recommend updating without first validating
  - Aim of validation studies is not to find similar predictive accuracy as in development set
    - But to find satisfactory accuracy in validation set
    - Depends on preferences/consequences of false predictions in validation situation
      - AUC of 0.60 is not per se bad
Steps in prediction modeling

- 1. Developing the prediction model
- 2. Validate the model in other subjects
- 3. Update existing models to local situation
- 4. Quantify impact of using model/test/marker/test strategy on doctor’s decision making and patient outcomes
Recall assumption of prediction rules:

- accurately estimated probabilities…
- …improve physicians’ decision making/behaviour…
- … and thus patient outcome

… studied in so-called Impact studies

(Phase 4. Impact studies

(Campbell BMJ 2000; Reilly and Evans Ann Int M. 2006; Moons BMJ 2009 + Heart 2012)
Phase 4. Impact studies

(Aim) Whether actual use (!) of prediction model/test/marker truly improves …

- … Physicians behaviour (treatment indications) …
- … Patient outcome or Health care costs …

… as compared to not using such model/marker/test

Impacts studies are thus intervention studies

- Intervention = use and subsequent treatment actions based on the model predictions
Phase 4. Impact studies

(Campbell BMJ 2000; Reilly and Evans Ann Int M. 2006; Moons BMJ 2009 + Heart 2012)

- **Design = like intervention studies**

  - When ‘effects of some intervention on patient outcome’ is mentioned → reflex = comparative study → good reflex!
    - In sharp (!) contrast to previous prediction modeling phases

  - Second reflex = randomized comparison

  - Indeed: best design = RCT

    - Preferably cluster randomised (e.g. stepped wedge) trial (Moons BMJ 2009 + Heart 2012)

    - Randomising practices
      - Less contamination across doctors in same practice → reduced contrast

    - Not randomising patients
      - Learning effects of doctors → reduced contrast
Disadvantages Cluster RCTs:

- Long duration → Certainly if patient outcomes occur late in time
- Large studies (costs)
- Prediction model always studied in combination with current treatments
  - If new treatment → new cluster RCT

Thousands clinical prediction models → increase per day

Simply not enough resources (budget plus patients) to study them all in a long term, expensive cluster RCT
Phase 4. Impact studies

(Campbell BMJ 2000; Reilly and Evans Ann Int M. 2006; Moons BMJ 2009 + Heart 2012)

• Before reflexing to RCTs → Alternative, cheaper/easier designs:
  - To better indicate which tests/markers/models should indeed undergo an RCT

• 1. Cross sectional randomised study with therapeutic decision (physicians or patients behavior) as outcome (no f-up)
  - Outcome never changes if physicians/patients don’t change behavior based on model predictions
  - Disadvantages
    - If changes decision making → Still need to quantify whether change in therapeutic decisions actually change patient outcomes
Phase 4. Impact studies

(Campbell BMJ 2000; Reilly and Evans Ann Int M. 2006; Moons BMJ 2009 + Heart 2012)

• 2. Modeling study

- Risk-Benefits (decision) models:

  - Linked evidence approach -- combining predictive accuracy studies and RCTs
  
  - Use predictive probabilities of validated model
  
  - Results of benefits and risks of existing therapies for that disease (e.g. obtained from RCTs)
  
  - To quantify effect of actually using the model (or test/marker) with model-directed therapies on patient outcome
Phase 4. Impact studies

(Campbell BMJ 2000; Reilly and Evans Ann Int M. 2006; Moons BMJ 2009 + Heart 2012)

=Gives indication of expected risks/benefits when introducing model/test(marker) combined with therapies
  – plus its cost-effectiveness
  – plus specific scenarios (e.g. treatment-probability thresholds) or subgroups may be tested

=Gives indication:
  – whether a real RCT is indicated or not
  – How to enrich the RCT design -- Eg excluding/focusing specific groups

From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study

Hendrik Koffijberg1, Bas van Zaane2 and Karel GM Moons1,2
Phase 4. Impact studies

(Campbell BMJ 2000; Reilly and Evans Ann Int M. 2006; Moons BMJ 2009 + Heart 2012)

• 3. Before-After study
  ▶ Compare patient outcomes in period before introducing model/test/marker with period after introducing
  ▶ E.g. Wells rule for DVT; Ottawa ankle/knee rule

• 4. External/historical control group

• Disadvantages 3+4
  ▼ Time changes (notably in therapeutic guidelines/therapies)
  ▼ Confounding by indication / case mix differences → adjustment in analysis (like non-randomized intervention studies)
Take home messages

• Number of markers increases per day
  - Simply enter market → overtesting → can’t be all relevant

• No diagnosis or prognosis estimated by single test/marker
  - Marker always form (small) part of many results

• Added/independent value of a marker test is relevant to know for physicians → and thus to quantify in research

• Many markers significant/relevant in isolation → not in combination
Take home messages

- Sensitivity and specificity of single marker:
  - ‘irrelevant’ (except in early phase of marker evaluation)
  - no information on added value
  - No constants of marker
  - Require dichotomisation of marker values (loss of information)

- Added value new marker/test is relevant to quantify
  - method: multivariable prediction modeling
Take home messages

• Phased approach of prediction modeling:
  - Development
  - Validation
  - Updating
  - Impact

• Validation is not aiming to find same predictive accuracy as in development set

• Validation requires proper reporting of original developed models, plus how predictors and outcomes defined/measured
  - not only of simplified scores
Take home messages

• Validation leads often to poor accuracy → do not panic → try an update first

• Impact studies are not per se large scale RCTs

• No developed model applied (or in guideline) without at least one external validation → preferably with impact assessment

• We need more collaborative IPDs → to develop, externally validate and improve prediction models