Prediction (about Diagnosis & Prognosis): Accuracy vs. Consequences

Benjamin Djulbegovic
Distinguished Professor
Associate Dean for Clinical Research
Director of Clinical Translational Science Institute
University of South Florida
Moffitt Cancer Center, Tampa, FL, USA
A typical (observational) design of a diagnostic/prognostic study

<table>
<thead>
<tr>
<th>Symptom/sign/marker</th>
<th>Index Test</th>
<th>Reference Standard</th>
<th>Reference Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td></td>
<td>TP</td>
<td>TP</td>
</tr>
<tr>
<td>-</td>
<td></td>
<td>FP</td>
<td>FN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TN</td>
</tr>
</tbody>
</table>

Outcome assessment: Univariate or multivariate analysis quantifying independent association (OR) between the marker/test and outcomes

Prospective evaluation → Retrospective evaluation

Time‡
A diagnostic/prognostic study: what can go wrong?

*choice of control group & method of planned comparison

(spectrum bias - patient population and control
Substantially differ in their characteristics;
Selection bias - systematic differences in comparison

*Patients with symptom/sign

Prospective evaluation

Retrospective evaluation

Detection bias
(systematic differences in outcome assessment)

Partial verification bias
(only + test results are verified)

Reference Standard

TP

FP

FN

TN

Verification bias

Differential reference standard bias
(- test results verified by different standard)
Prototypical flow diagram of a study on diagnostic accuracy: STARD

General example

Enrolled patients

Index test

Abnormal result

Normal result

Inconclusive result

Reference standard

Reference standard

Reference standard

Reference standard

Inconclusive

Target condition present

Target condition absent

No reference standard

Target condition present

Target condition absent

Target condition present

Target condition absent

Target condition present

Target condition absent

www.consort-statement.org
Risk of Bias: Prognostic Factor Studies

- Similar to RCT RoB: Domain-based
- Reflect basic epidemiological principles
- Potential biases related to:
  1. Study participation
  2. Study attrition
  3. Prognostic factor measurement
  4. Outcome measurement
  5. Covariate measurement and account
  6. Analysis and reporting

Annals of Internal Medicine. 2006;144:427-437

Courtesy of Dr. Hayden & Moon
Phases of Evaluation of a Novel Risk Marker*

- **Proof of concept**: Do novel marker levels differ between participants with and without outcome?
- **Prospective validation**: Does the novel marker predict development of future outcomes in a prospective cohort or nested case–cohort/case–cohort study?
- **Incremental value**: Does the novel marker add predictive information to established, standard risk markers?
- **Clinical utility**: Does the novel risk marker change predicted risk sufficiently to change recommended therapy?
- **Clinical outcomes**: Does the novel risk marker improve clinical outcomes, especially when tested in a randomized clinical trial?
- **Cost-effectiveness**: Does use of the marker improve clinical outcomes sufficiently to justify the additional costs of testing and treatment?
ACCE framework

- **analytic validity** - technical test performance
- **clinical validity** - the strength of association that determines the test’s ability to accurately and reliably identify or predict the disorder of interest
- **Clinical utility** - balance of benefits and harms when the test is used to influence patient management
- **Ethical, legal and social implications** of the test/model application
Testing for Lynch syndrome in relatives of patients diagnosed with colo-rectal cancer

EGAPP Recommendations

- **Analytic validity**
  - Analytic sensitivity and specificity for preliminary and diagnostic tests (MSI, IHC, and BRAF) are high (but difficult to quantify)

- **Clinical Validity**
  - The sensitivity of MSI testing is about 89% for mutations in MLH1 and MSH2, with a lower sensitivity of about 77% for mutations in MSH6 (and PMS2). Specificity is estimated to be 90.2%, with an adequate level of evidence.
  - The sensitivity of IHC testing is 83%, regardless of the underlying MMR gene mutation. Specificity is more variable, with a central estimate of 88.8% and an adequate level of evidence.
  - Estimates of sensitivity (69%) and specificity (100%) for BRAF mutation testing among newly diagnosed CRC cases with absent IHC staining for MLH1

- **Clinical Utility**
  - **Probands** with Lynch syndrome: Subtotal colectomy with ileorectal anastomosis is recommended as a reasonable alternative to segmental resection in these cases, but it has not been shown to be superior at follow-up.
  - **Family members with Lynch syndrome**—counseling and genetic testing; Colonoscopy is recommended every 1–2 years for both patients and their relatives with Lynch syndrome, beginning at age 20–25 years.
  - **Female probands and relatives with Lynch syndrome**—In women with Lynch syndrome, transvaginal ultrasound and endometrial biopsy every 1–2 years, beginning at age 30–35 years
Methods for prediction

Statistical approaches

- Linear regression (continuous data)
  - Optimal results for linear predictor-outcome relationships
  - Inappropriate to model non-linear relationships
  - Very sensitive to outliers and inconsistencies
  - Limited to predicting numerical output
- Logistic regression (binary outcomes)
  - Predictors can be categorical variables
  - Assume linear relationships between the predictors and the logic from of the output variable
- Cox regression (time-to-event outcomes)
  - Semi-parametric
  - May handle censored data
  - Requires proportionality assumption to be met

Artificial Intelligence

- Neural networks
  - Able to model complex predictor-outcome relationship
  - Work well with very large datasets
  - Difficult interpretation (Black box)
- Support vector machines
  - Work with data that is not linear separable
  - Highly accurate results
  - Difficult interpretation (Black box)
- Rough set methods
  - Models inconsistencies in dataset
  - Readily interpretable results in the form of decision rules.
  - Need categorical variables
  - Model is dependent on the selecting training set
How good is a test/marker?

• Accuracy in predictions
  – ROC-ers
    • Metrics that assess how close predictions (calculations) are to actually observed outcomes (classifications)

• Consequences of (miss)classifications
  – VOI-ers
    • How useful is for decision-making?
Overall performance measures

- Calculation of the distance from the predicted and actual outcome is central to quantifying overall model performance

- **Brier Score**: Quadratic scoring rule that calculates the difference between the predicted and observed patient’s outcome.
  
  Given predicted probability of outcome $p_i$ and binary outcome (dead-alive) $Y_i$ it is defined as
  
  $$\sum_i (Y_i(1 - p_i)^2 + (1 - Y_i)p_i^2)$$

  - A Brier score of 0 indicates a perfect model, while the score of 0.25 indicates an non-informative model
    - the value achieved when issuing a predicted probability of 50% to each patient.

- **Explained variation ($R^2$)**: Nagelkerke’s $R^2$
  
  - Different derivation for continuous, dichotomous and time-to-event data

  Pearson $R^2$

- **Scaled Brier score** by its maximum has interpretation similar to Pearson Correlation Coefficient.

*Steyerberg et al Epidemiology 2010*
Metrics for evaluation of PREDICTIONS

Plot of predicted vs. observed outcomes

- **Calibration (“reliability”)**
  - Agreement between predicted probabilities and observed frequencies of the event of interest
  - Perfect calibration: 45° line, with a slope 1 and intercept 0
    - The slope and intercept is calculated using regression model
  - **Hosmer-Lemeshow goodness-of-fit test** to compare observed vs. predicted probabilities (most commonly into deciles) and tests the hypothesis that the difference between observed and predicted events is zero for all groups.
  - The higher the p-value, the closer predicted agree with the observed outcomes.

*Steyerberg et al. Epidemiology 2010*
Metrics for evaluation of PREDICTIONS

- **Discrimination**
  - Discriminating between those with (TP) and without outcomes (TN)
  - **c statistics**: probability of correct classification for a pair of subjects with and without the outcome
  - **c statistics** = \( \text{AUC (for binary outcome)} \)
  - **Discrimination slope**: how well subjects with and without outcomes are separated
  - calculated as the **absolute difference in average predictions** for those with and without the outcome
  - Visualize via **box-plots** or histograms
  - Better discrimination - less overlap between those with and without outcomes

- **ROC Curve**
  - Evaluation of Prognostat in predicting survival of patients referred to hospice
Improving accuracy of CLASSIFICATIONS: how well a new model reclassifies events and non-events

Net reclassification index (NRI)

- Measures how well the new model reclassifies event and non-event
- Depends on how we decide to classify observations
  - Example: risk of CVD at 10 yrs > 5.6% vs. • 5.6%
  - Re-classify in such a way that the patients experiencing an event go up in risk and patients not experiencing an event go down in risk
    - Each individual is classified by both model 1 and 2
    - 2 cross-tabulation tables (events and non-events)
**Improving accuracy of classification:**
how well a new model reclassifies events and non-events

**Net reclassification index (NRI)**

<table>
<thead>
<tr>
<th></th>
<th>Model without HDL</th>
<th>Model with HDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>No CHD (n=3081)</td>
<td>• 5.6%</td>
<td>1872</td>
</tr>
<tr>
<td></td>
<td></td>
<td>142*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;5.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>166**</td>
</tr>
<tr>
<td>CHD (n=183)</td>
<td>• 5.6%</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10**</td>
</tr>
<tr>
<td></td>
<td>&gt;5.6%</td>
<td>3*</td>
</tr>
</tbody>
</table>

*Reclassification in the wrong direction; ** Reclassification in the right direction

NRI = Net fraction of reclassifications in the right direction by making decisions based on predictions with a marker compared to decisions without the marker = 

\[
NRI = \left[ \frac{\#\text{Events moving up}}{\#\text{Events}} \right] - \left[ \frac{\#\text{Events moving down}}{\#\text{Events}} \right] + \left[ \frac{\#\text{Non-events moving down}}{\#\text{Non-events}} \right] - \left[ \frac{\#\text{Non-events moving up}}{\#\text{Non-events}} \right]
\]

\[
NRI = 2 \cdot \text{AUC}
\]

Steyerberg et al
Rev Esp Cardiol. 2011
Improving accuracy by “reclassification”: Integrated discrimination improvement (IDI)

- **IDI integrates the NRI over all possible cut-offs**
  - Measure of improved sensitivity without sacrificing specificity
  - Formula measures how much increase in ‘p’ for events, and how much decrease for the non-event
  - equivalent to difference
    - in discrimination slopes, or
    - to differences in Pearson $R^2$ measure, or
    - to differences in scaled Brier scores
  - Visualized by box plot for 2 models
    - One with and one without a marker

Steyerberg et al, *Epidemiology* 2010

taken from Kenedy K, 2009
Do IDI and NRI reflect consequences?

- That is, its weights differ from misclassification costs related to decision-analytic Pt
  - NRI = DCA Net benefits, only Pt = Prevalence

  **NO- they are “clinically absurd”**

  “The test criterion involves cost parameters that can be far beyond the scope of statistical expertise, involving matters of valuation and quality of life”...and it cannot rely on potentially “absurd implicit defaults”
Predictions vs. decisions

- C index = 0.85
- Specificity = 50% at sensitivity = 90%
- Brier score = 19%
- H-L: p = 0.1
- IDI = 0.07
- NRI = 5%

- What we should do?
Evaluating consequences: decision-analytic approach

- DCA
- Regret-DCA
- Relative Utility Curves
- Weighted NRI
Weighting the consequences of under-treatment vs. over-treatment

- Weighting the consequences of FN (failing to benefit) vs. FP (unnecessary harms)
  - Pt = 1/(1 + X)
    - where X = (value of FN / value of FP)
  - FN = FP † Pt = 0.5
  - 4 FN > FP † Pt = 0.2
  - 4 FP > FN † Pt = 0.8
If value of a true-positive result is fixed at 1‡ Decision curve analysis

1. Select a \( p_t \)
2. Positive test defined as \( \hat{p} \geq p_t \)
3. Calculate “Clinical Net Benefit” as:

\[
\frac{\text{TruePositiveCount}}{n} - \frac{\text{FalsePositiveCount}}{n} \left( \frac{p_t}{1 - p_t} \right)
\]

4. Vary \( p_t \) over an appropriate range
Decision Curve Analysis

• Example:
  – N=1000, dichotomous event, 10% as threshold
    (avoiding FN is more 9 times more important than FP)

<table>
<thead>
<tr>
<th>Predicted Probability</th>
<th>Outcome</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 10%</td>
<td>80</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>&lt;10%</td>
<td>20</td>
<td>700</td>
<td></td>
</tr>
</tbody>
</table>

\[
NetBenefit = \frac{80}{1000} - \left( \frac{200}{1000} \cdot \frac{.1}{1-.1} \right) = .0578
\]

5.78 net true positive results per 100 compared with treating all

Med Decis Making 2006;26:565–574
Decision Curve Analysis: use of SUPPORT model facilitate hospice referral

BMC medical informatics and decision making 2010, 10:51.
BMC Medical Informatics and Decision Making 2011, 11:77

No Rx > model if the tolerance of FP/FN $> 11.5$
Importance of anticipating consequences of our actions: regret-theory approach

- We can **always** make a mistake
  - Recommend ineffective treatments, or
  - Fail to recommend effective treatments
- Sense of loss, or **regret**
- “theory of choice that completely ignores feeling such as the pain of losses and the **regret of mistakes** is not only descriptively unrealistic but also might lead to prescriptions that do not maximize the utility of outcomes as they are actually experienced”.
  - Requires the ability to imagine other possibilities than the current state of the world
    - Avoiding wrong choice, or bad outcomes
    - An example of **counterfactual thinking**

Regret-based Decision Curve Analysis

Mathematically more parsimonious ...
Enables experiencing consequences of decisions both at the emotional (system 1) and cognitive (system 2) level
Linked to direct elicitation of preferences using DVAS method (regret of omission vs. regret of commission)

BMC medical informatics and decision making 2010, 10:51.
BMC Medical Informatics and Decision Making 2011, 11:77
Weighted NRI

• The default values for the weights $S_1$ and $S_2$ are $1$/prevalence, which reduces $wNRI$ to NRI

$$wNRI = s_1 (P(\text{event|up})P(\text{up}) - P(\text{event|down})P(\text{down})) + s_2 (P(\text{non-event|down})P(\text{down}) - P(\text{non-event|up})P(\text{up})).$$

For binary classification, this can be reduced to

$$wNRI = s_1 \frac{TP_2 - TP_1}{N} + s_2 \frac{FP_1 - FP_2}{N}.$$ 

$wNRI = \cdot \frac{NB}{p_t}$

where $NB$ is DCA net benefit, and $p_t$ is the probability of disease at which we are indifferent between Rx and No Rx

Pencina et al Stat Med 2011
Relative Utility Curves

Utility relative to perfect prediction at the optimal cut point

Values consistent with Rx (or no Rx) in the absence of prediction (prevalence > Risk Threshold)

RU (model 2 - model 1) = 0.18 - 0.15 = 0.03 (largest possible: 1 - 0.15 = 0.85)

Is model 2 worth doing?
At R = 0.013 * prevalence (0.01) = 0.0003
1 / 0.0003 = 3333 measurements of breast density to Rx one woman

Risk of disease in the absence of treatment at which a person is indifferent between Rx and NoRx
R = 1 / (1 + X), where X = consequences of FN / consequences of FP

Who would develop breast cancer in the absence of RX (chemoprevention) (at R = 0.013, avoiding FN is 85x more important than FP)

Baker JNCI 2009
Conclusions

- **Overall performance of model**
  - Brier score, Pearson $R^2$, Nagelkerke $R^2$

- **Predictions**
  - Use measures for calibration and discrimination
    - Intercept/slope; H-L statistics; AUC, c statistics

- **Evaluation of incremental value of a marker**
  - AUC; NRI; IDI

- **Decisions**
  - DCA (NB or regret-based), wNRI, relative utility curves