Section III: Evaluating performance of markers and treatment rules

- Descriptive devices
- Assessing model calibration
- Recommended measures of marker performance
    - Estimation and inference
- Critique of other marker performance measures
- Implications for comparing markers or rules

Descriptive devices

- Risk curves
- Treatment effect curves

Terminology suggests the outcome is binary, but these devices also apply to categorical and continuous outcomes.

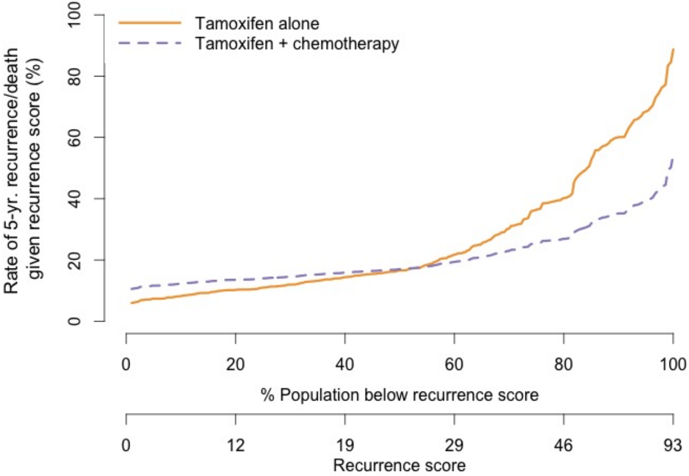
Janes et al. (*Int J Biostat* 2014)

Risk curves

For a single marker, *risk curves* plot the expected outcome as a function of the marker, for each treatment.

We recommend aligning the curves for the two treatment groups with respect to marker percentile $F(X)$, rather than marker value X, i.e., plot $E(D|A, X)$ vs. $F(X)$ for $A = 0, 1$.

Example: Oncotype DX marker in the breast cancer trial



Treatment effect curves

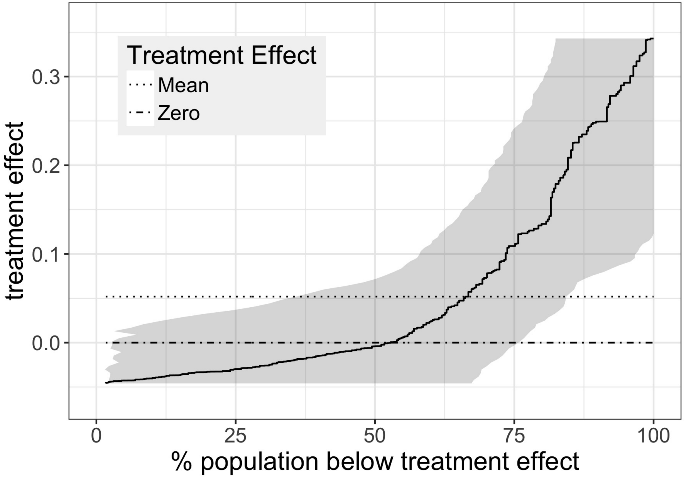
Show the distribution of the marker-specific treatment effect, $\Delta(X) = E(D|A = 0, X) - E(D|A = 1, X)$.

Different scales are possible:

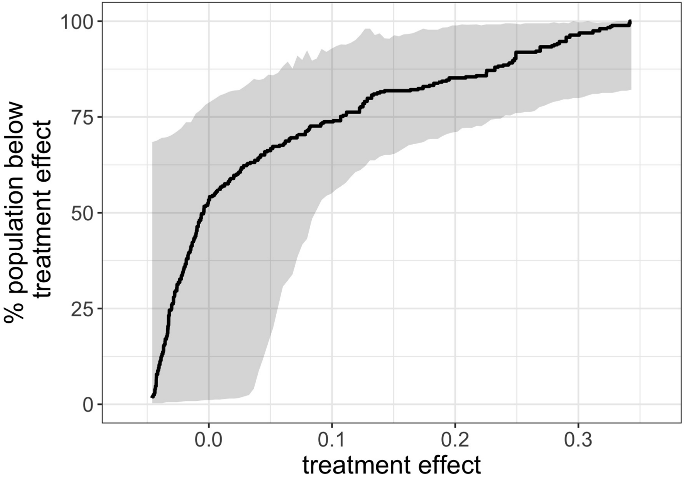
- Reverse-CDF, i.e. $\Delta(X) = \delta$ vs. $F_{\Delta}(\delta)$. Also called a *predictiveness curve* (Huang et al. 2007).
- Traditional CDF, i.e. $F_{\Delta}(\delta)$ vs. $\Delta(X) = \delta$.
- Density or histogram of $\Delta(X)$.

Unlike the risk curve plot, this device applies to multivariate X.

Treatment effect curve for the Oncotype DX marker: Reverse CDF



Treatment effect curve for the Oncotype DX marker: Traditional CDF



Checking model calibration

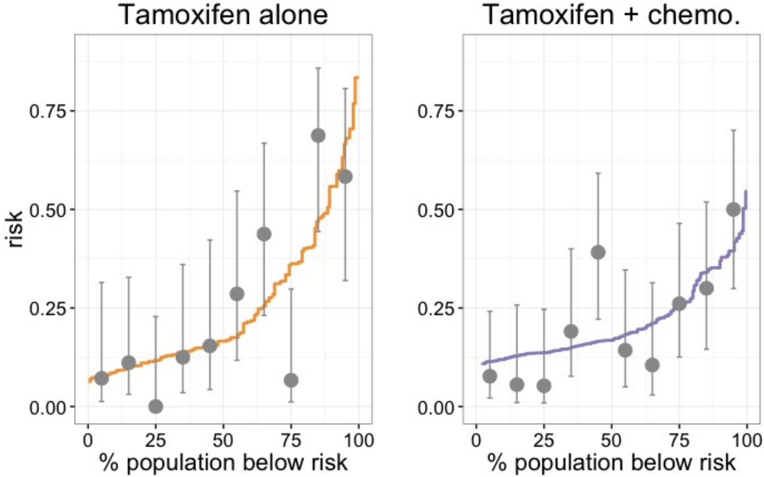
Estimating these curves requires modeling $E(D|A, X)$.

Good calibration of the $E(D|A, X)$ model is essential for validity of the risk and treatment effect curves.

Two approaches to assessing calibration:

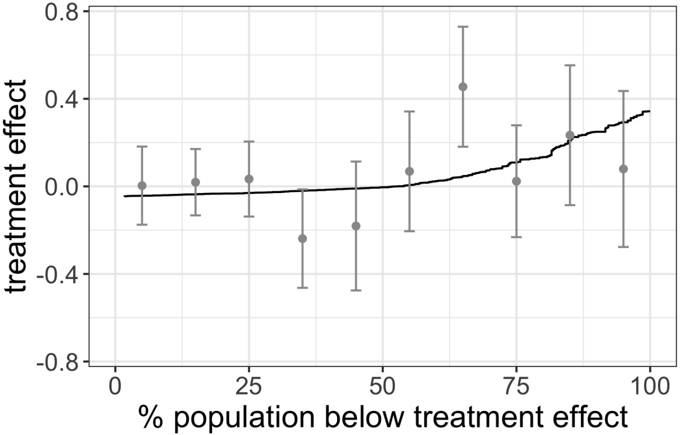
- Overlay observed risks and treatment effects on the plots
- Formally compare observed vs. predicted values using Hosmer-Lemeshow goodness of fit tests

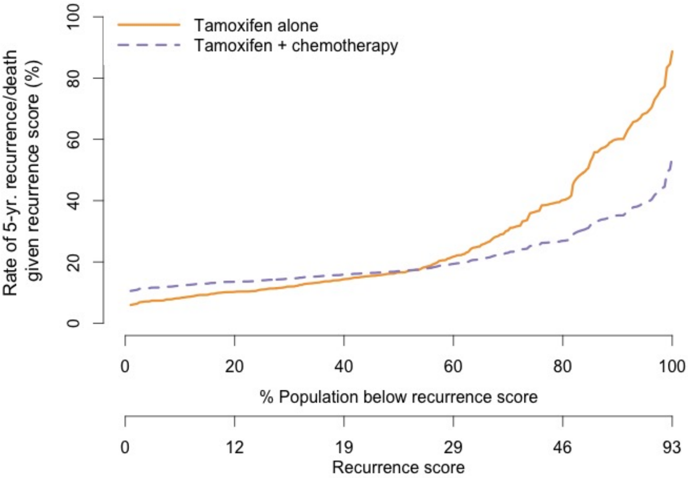
Example: Oncotype DX risk curve calibration



No significant difference between observed and predicted risks in either treatment group (p = 0.078 and 0.096, Hosmer-Lemeshow test).

Example: Oncotype DX treatment effect curve calibration





Is the marker good enough to incorporate into clinical practice?

Performance Measures

Context

Goal is to evaluate the performance of marker-based treatment rule $d(X)$.

- Could be a rule estimated from the data, $d_n(X)$, or a pre-specified rule

Focus on the setting where $A = 0$ is the default treatment choice absent X.

- X is used to identify a subgroup likely to benefit from treatment

The opposite scenario where $A = 1$ is the default and X is used to identify a subgroup not likely to benefit from treatment is handled analogously.

Evaluating a marker-based treatment rule

Suppose that $A = 0$ is the default choice absent X.

The expected outcome under the rule, $E(D(d))$, is a fundamental parameter.

It is helpful to contrast $E(D(d))$ with the expected outcome under the default approach. The *clinical impact* of rule $d(X)$ is

$$\mathcal{I}(d) = E(D \mid A = 0) - E(D(d))$$

Song and Pepe 2004; Gunter et al. 2011; Janes et al. 2011; Zhang et al. 2012

Note that

$$\begin{aligned}
\mathcal{I}(d) = {} & [E(D \mid A = 0, d(X) = 0) \cdot P(d(X) = 0) \\
& + E(D \mid A = 0, d(X) = 1) \cdot P(d(X) = 1)] \\
& - [E(D \mid A = 0, d(X) = 0) \cdot P(d(X) = 0) \\
& + E(D \mid A = 1, d(X) = 1) \cdot P(d(X) = 1)] \\
= {} & E(\Delta(X) \mid d(X) = 1) \cdot P(d(X) = 1) \\
\equiv {} & \beta(d) \cdot \tau(d)
\end{aligned}$$

The two constituents of $\mathcal{I}(d)$,

- $\tau(d)$ = the proportion of subjects impacted by X measurement, who are recommended treatment
- $\beta(d)$ = average treatment efficacy in this subgroup

are important measures in their own right.

In practice we recommend reporting $(E(D(d)), \mathcal{I}(d), \tau(d), \beta(d))$, along with the expected outcomes under "treat all" and "treat none" policies, $\rho_0 = E(D|A = 0)$ and $\rho_1 = E(D|A = 1)$.

If $A = 1$ is the default choice absent X,

$$\begin{aligned}\mathcal{I}(d) &= E(D \mid A = 1) - E(D \mid \text{treat using rule } d) \\ &= E(-\Delta(X) \mid d(X) = 0) \cdot P(d(X) = 0)\end{aligned}$$

and the two constituents are

- $\tau(d) = P(d(X) = 0) =$ the proportion impacted by X measurement, who are recommended no treatment
- $\beta(d) = E(-\Delta(X) \mid d(X) = 0) =$ average benefit of no treatment in this subgroup

Empirical estimation

Estimate the performance of rule $d(X)$ empirically using

$$\begin{aligned}
\widehat{\tau}^e(d) &= \mathbb{P}(d(X)=1) \\
\widehat{E}^e(D(d)) &= \mathbb{E}(D \mid A=0, d(X)=0)\cdot(1-\widehat{\tau}^e(d)) \\
&\quad + \mathbb{E}(D \mid A=1, d(X)=1)\cdot\widehat{\tau}^e(d) \\
\widehat{\beta}^e(d) &= \mathbb{E}(D \mid A=0, d(X)=1) \; - \; \mathbb{E}(D \mid A=1, d(X)=1) \\
\widehat{\mathcal{I}}^e(d) &= \mathbb{E}(D \mid A=0) - \widehat{E}^e(D(d))
\end{aligned}$$

where $\mathbb{P}$ is the empirical probability and $\mathbb{E}$ is the empirical mean

Note that $\widehat{E}^e(D(d)) = IPWE(d)$ from Section II

Janes et al. (*Int J Biostat* 2014)

Model-based estimation

Alternatively, performance of $d(X)$ can be estimated in a *model-based* fashion, using a model for $E(D|A,X) = \mu(A,X;\beta)$:

$$\begin{aligned}
\widehat{E}^m(D(d)) &= \mathbb{E}(\mu(0,X;\widehat{\beta})(1-d(X))) + \mathbb{E}(\mu(1,X;\widehat{\beta})d(X)) \\
\widehat{\beta}^m(d) &= \mathbb{E}([\mu(0,X;\widehat{\beta}) - \mu(1,X;\widehat{\beta})]\; d(X)) \\
\widehat{\mathcal{I}}^m(d) &= \mathbb{E}(D \mid A=0) - \widehat{E}^m(D(d))
\end{aligned}$$

Model-based estimators are more efficient than empirical estimators. However they are biased if the $E(D|A,X)$ model is mis-specified.

Janes et al. (*Int J Biostat* 2014)

Inference

When evaluating performance of a pre-specified rule $d(X)$, all estimates of performance are asymptotically normal. Quantile bootstrap confidence intervals work well.

Similarly, when training data are used to derive $d_n(X)$ and independent test data are used to estimate performance, estimators are asymptotically normal and the bootstrap can be used for inference.

One exception to the above is when $P(\Delta(X) = 0) > 0$, i.e. there exist subjects with $\Delta(X)$ identically 0. Performance estimates may not be asymptotically normal and the bootstrap may not perform well.

Inference, continued

However, when the same data are used to derive $d_n(X)$ and to estimate performance

- Estimates are biased (overoptimistic)
- Estimators are not asymptotically normal. Bootstrap-based confidence intervals may not have good coverage.
- Performance of normal-theory/bootstrap inferential methods is expected to be worse for settings with: small n, high-dimensional X, heavy marker/model selection
- There are partial solutions (next slide).
- This is an active research area.

Partial solutions to drawing inference absent test data

Cross-validation (CV)

- Sample B training/test data splits. For each, obtain $d_n^b(X)$ using training data and estimate performance using test data. Average performance estimates. Shift naive performance estimates and confidence intervals down by the estimated bias – the difference between naive and CV performance estimates.

Bootstrap bias correction: the "refined bootstrap" (Efron and Tibshirani 1994)

- Sample B bootstrap datasets. For each, obtain $d_n^b(X)$ and calculate the difference in estimated performance of this rule using the bootstrap vs. original data. The average of these differences estimates the bias. Shift naive performance estimates and confidence intervals down by the estimated bias.

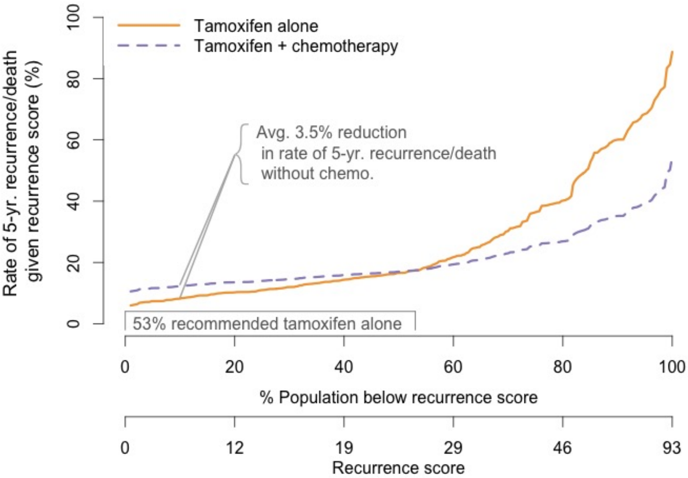
There are variations on each of these approaches.

Example: Oncotype DX marker performance

Utilize logistic regression model for $E(D|A, X)$ to estimate risk and treatment effect curves and to estimate the rule

$$d_n(X) = I\left(\widehat{E}(D|A = 0, X) > \widehat{E}(D|A = 1, X)\right)$$

Performance of $d_n(X)$ is estimated empirically. Bootstrap bias correction is used for inference.

Tamoxifen alone
Tamoxifen + chemotherapy
Avg. 3.5% reduction in rate of 5-yr. recurrence/death without chemo.
53% recommended tamoxifen alone
Rate of 5-yr. recurrence/death given recurrence score (%)
0
20
40
60
80
100
% Population below recurrence score
0
20
40
60
80
100
Recurrence score
0
12
19
29
46
93

Absent X, chemotherapy is the default.

Given X,

- $\widehat{\tau} = 53.0\%$ avoid chemo, and associated toxicity and cost (0.2 to 80.1)
- $\widehat{\beta} = 3.5\%$ lower risk of 5-yr. recurrence/death in subset avoiding chemo. (-12.9 to 10.8)
- Estimated clinical impact is $\widehat{\mathcal{I}} = 1.5\%$ lower 5-yr. recurrence/death rate (-3.6 5.7)
  - 21% event rate under default "chemo. for all" policy is reduced to 19.5% with use of X.
  - 25% event rate under "chemo for none"

Said another way,

- The overall efficacy of chemo. is a 3.9% absolute reduction in the 5-yr. recurrence/death rate.
- The efficacy of X-based chemo. is a $3.9 + 1.5 =$ 5.4% reduction in the 5-yr. recurrence/death rate.

Example: HIV prevention trial

RCT of PrEP vs. placebo for prevention of HIV infection in MSM.

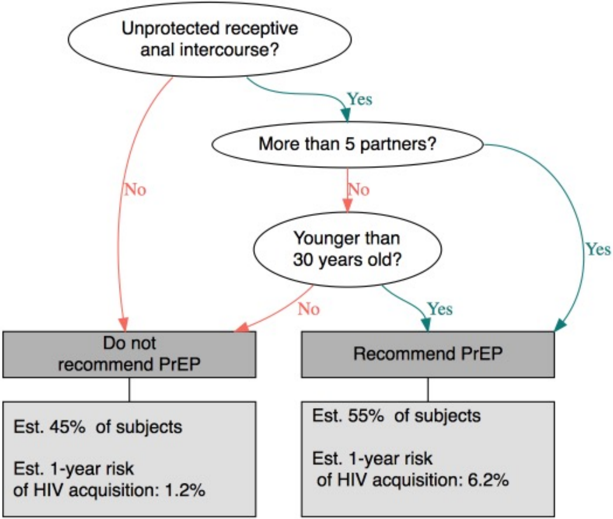
Cox proportional hazards logic regression (Ruczinski et al. 2003) used to model 1-yr. HIV risk without PrEP, $E(D|A=0,X)$, using placebo arm data.

Current WHO guideline recommends PrEP for subjects estimated to be at or above 3% 1-yr. risk without PrEP. Performance of rule

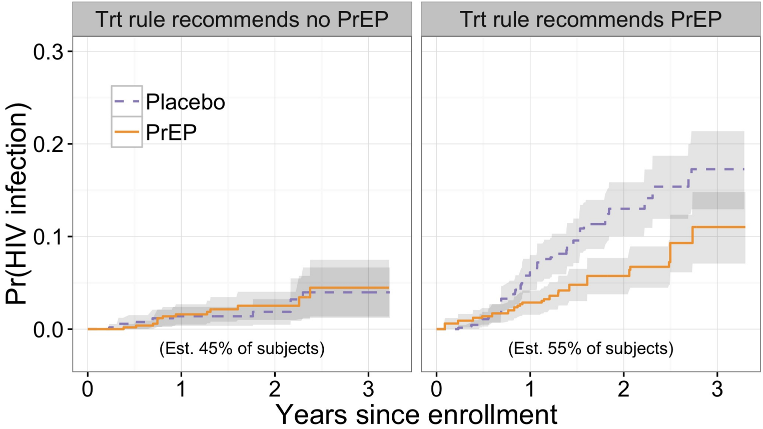
$$d_n(X) = I\left(\widehat{E}(D|A=0,X) > 0.03\right)$$

estimated empirically. Bootstrap bias correction used for inference.

Risk-based PrEP recommendation


Unprotected receptive anal intercourse?
Yes
No
More than 5 partners?
No
Yes
Younger than 30 years old?
No
Yes
Do not recommend PrEP
Recommend PrEP
Est. 45% of subjects
Est. 1-year risk of HIV acquisition: 1.2%
Est. 55% of subjects
Est. 1-year risk of HIV acquisition: 6.2%

Based on Cox logic regression model fit using placebo-arm data. Recommend PrEP if 1-yr. risk is 3% or higher.



Without PrEP, est. 1-yr HIV incidence is 4.0% (2.9 - 5.2%)
PrEP for all yields est. incidence 2.3% (1.4 - 3.2%)
PrEP for high risk subjects yields est. incidence 2.4% (1.8 - 2.9%)

In contrast to a PrEP for all policy:

- PrEP for high risk subjects is estimated to increase 1-yr. HIV incidence by 0.1%
- But requires treating only 55.2% of the population

Other Performance Measures

A marker-by-treatment interaction is insufficient

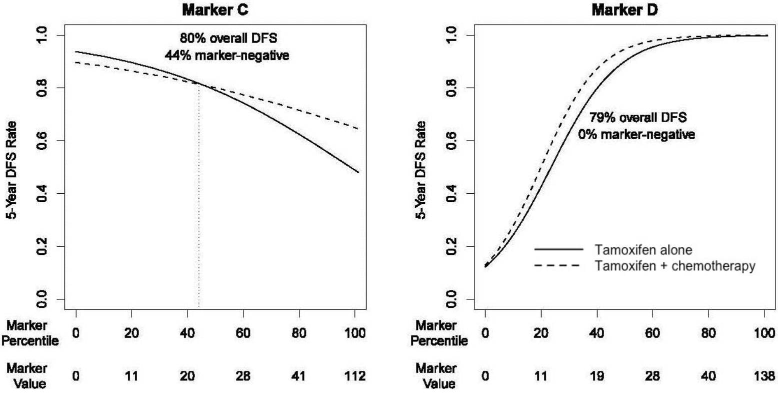
Testing for a marker-by-treatment interaction is a useful first step.

- An interaction is necessary, but not sufficient, for the marker to have value

However, the interaction coefficient does not *quantify* marker performance.

- Interpretation depends on the scale of the $E(D|A, X)$ model, the other variables in the model, and the scale of the marker
- Easy to construct examples of markers with the same interaction coefficient, but different clinical impact

Example



Two markers with the same marker-by-treatment interaction, but very different performance.

Janes et al. (*Ann. Int. Med.* 2011)

What about biomarker accuracy?

Sensitivity, specificity, PPV, and NPV are classic performance measures for diagnostic, screening, and prognostic markers.

FDA and IOM biomarker development guidance documents advocate reporting accuracy measures

- without properly distinguishing between approaches for diagnostic and prognostic and predictive/treatment selection markers

Accuracy measures have been proposed for treatment selection markers, for the setting of a binary outcome D

Huang et al. (*Biometrics* 2012), Zhang et al. (*Ann Appl Stat* 2014), Sitlani and Heagerty (*Stat Med* 2014), Simon (*JNCI* 2015)

Accuracy measures rely on potential outcomes

$D(0)$ = potential outcome without treatment

$D(1)$ = potential outcome with treatment

$\textit{Trt. benefit} \equiv D(0) = 1, D(1) = 0$

$\textit{No trt. benefit} \equiv D(0) = D(1)$ or $D(0) = 0, D(1) = 1$

The accuracy of rule $d(X)$ is then measured by:

Sensitivity $= P(d(X) = 1 \mid \textit{Trt. benefit})$

Specificity $= P(d(X) = 0 \mid \textit{No trt. benefit})$

PPV $= P(\textit{Trt. benefit} \mid d(X) = 1)$

NPV $= P(\textit{No trt. benefit} \mid d(X) = 0)$

Fundamental problem

Almost never can both potential outcomes be observed and so we do not know whether a subject benefits from treatment.

Therefore, in general the accuracy measures are not estimable from data– even RCT data.

Illustration: Two binary markers in an RCT (n = 2000)

Unobservable data: Marker-positivity by potential outcome

		Benefit from trt. (n = 400)	Bad outcome regardless of trt. (n = 600)	Good outcome regardless of trt. (n = 600)	Harmed by trt. (n = 400)
Marker 1	Negative	200	250	400	250
	Positive	200	350	200	150
Marker 2	Negative	100	350	500	150
	Positive	300	250	100	250

Observable data: Marker-positivity by observed outcome

	Treatment arm	No trt. (n = 1000)		Trt. (n = 1000)	
	Outcome	Good	Bad	Good	Bad
Marker 1	Negative	325	225	300	250
	Positive	175	275	200	250
Marker 2	Negative	325	225	300	250
	Positive	175	275	200	250

Janes et al. (*JNCI* 2015)

The biomarkers have very different accuracy, but the same observed data:

Marker 1	Sensitivity = 50%	Specificity = 56%
	PPV = 22%	NPV = 82%
	Prop. marker-positive = 56%	
Marker 2	Sensitivity = 75%	Specificity = 63%
	PPV = 33%	NPV = 91%
	Prop. marker-positive = 56%	

"Pragmatic" accuracy measures have been proposed which assume $D(0) \perp D(1)$ given X.

- This assumption is unlikely to hold in any clinical context
- This example illustrates the fallacy of these pragmatic measures

Pragmatic Accuracy (Both Markers)	$\text{Sensitivity}_i = 61\%$ $\text{PPV}_i = 27\%$	$\text{Specificity}_i = 46\%$ $\text{NPV}_i = 78\%$
Marker 1 Truth	$\text{Sensitivity} = 50\%$ $\text{PPV} = 22\%$	$\text{Specificity} = 56\%$ $\text{NPV} = 82\%$
Marker 2 Truth	$\text{Sensitivity} = 75\%$ $\text{PPV} = 33\%$	$\text{Specificity} = 63\%$ $\text{NPV} = 91\%$

Our recommendation

In general, accuracy estimates depend on unverifiable assumptions about the joint distribution of potential outcomes.

We recommend instead focusing on identifiable marker performance measures: $E(D(d))$, $\mathcal{I}$, τ, β.

These measures do not depend on the joint distribution of potential outcomes.

Other performance measures

Recall $\rho_0 = E(D|A=0)$ and $\rho_1 = E(D|A=1)$. Note that $E(\Delta(X)) = \rho_0 - \rho_1$.

Variance in treatment effect,

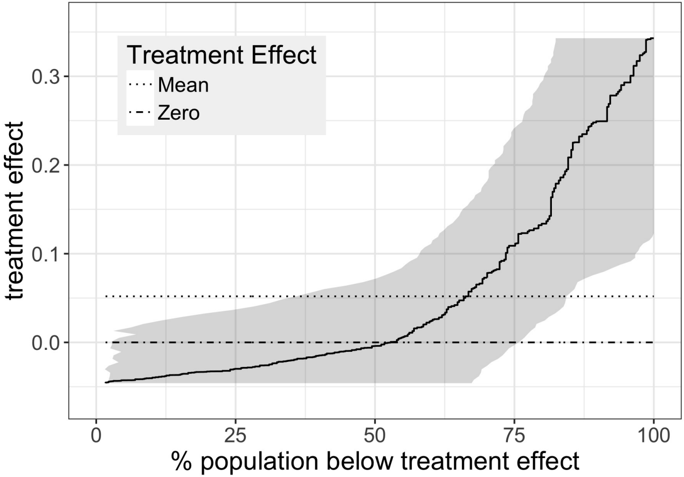
$$V_\Delta \equiv \int (\Delta(X) - (\rho_0 - \rho_1))^2 \, \partial F_\Delta$$

Total gain,

$$\mathrm{TG} \equiv \int |\Delta(X) - (\rho_0 - \rho_1)| \, \partial F_\Delta$$

- Two “global” performance measures– do not require specifying a treatment rule
- They lack a clinically relevant interpretation

Janes et al. (*Int J Biostat* 2014)


Treatment Effect
Mean
Zero
treatment effect
0.3
0.2
0.1
0.0
0
25
50
75
100
% population below treatment effect

Any one performance measure is insufficient

We advocate reporting

$$\begin{gathered} E(D(d)) \\ \mathcal{I}(d) = \beta(d) \cdot \tau(d) \\ \tau(d) = P(d(X) = 1) \\ \beta(d) = E(\Delta(X) \mid d(X) = 1) \end{gathered}$$

No single measure says it all.

- E.g., a large β may not be compelling if τ is small
- E.g., if treatment has downsides not captured in D, $\mathcal{I}$ is insufficient and we need τ to capture treatment "cost"

Implications for comparing markers or treatment rules

Estimate contrasts in the above performance measures.

Again, our recommendation is to contrast

$$E(D(d))$$

$$\mathcal{I}(d) = \beta(d) \cdot \tau(d)$$

$$\tau(d) = P(d(X) = 1)$$

$$\beta(d) = E(\Delta(X) \mid d(X) = 1)$$

Performance measures can be compared using Wald-type hypothesis tests.

Janes et al. (*Int J Biostat* 2014)

Example: Two simulated markers in the breast cancer context

	Marker X_1 Estimate (95% CI)	Marker X_2 Estimate (95% CI)	X_1 vs. X_2 Estimated Diff. (95% CI)	 P-value
$\widehat{\tau}^e$	0.461 (0.000,0.700)	0.377 (0.304,0.470)	0.084 (-0.358,0.236)	0.768
$\widehat{\beta}^e$	0.029 (-0.106,0.082)	0.238 (0.170,0.309)	-0.209 (-0.342,-0.129)	< 0.002
$\widehat{\beta}^m$	0.023 (0.000,0.057)	0.262 (0.209,0.310)	-0.239 (-0.294,-0.178)	< 0.002
$\widehat{\mathcal{I}}^e$	0.013 (-0.010,0.044)	0.090 (0.060,0.122)	-0.076 (-0.111,-0.042)	< 0.002
$\widehat{\mathcal{I}}^m$	0.010 (0.000,0.037)	0.099 (0.071,0.129)	-0.088 (-0.115,-0.061)	< 0.002

Summary

- Descriptive devices are useful for visualizing data
- Clinical impact and its constituents are recommended performance measures
- Contrasts in these measures are recommended for comparing markers or rules