Section III: Evaluating marker performance

- 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
- Extension: Formally incorporating treatment downsides

Describing a binary treatment rule

Given a rule $d(X)$, evaluate

- Expected outcomes given treatment recommendation:
  $E(D(0)|d(X) = 0)$, $E(D(1)|d(X) = 0)$ and
  $E(D(0)|d(X) = 1)$, $E(D(1)|d(X) = 1)$
- Average treatment effects given treatment recommendation:
  $E(\Delta(X)|d(X) = 0)$ and $E(\Delta(X)|d(X) = 1)$ where
  $\Delta(X) = E(D(0)|X) - E(D(1)|X)$

Descriptive devices for continuous markers

- 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

Risk curves plot the expected outcome as a function of a univariate marker, for each treatment: $E(D(a)|X)$ vs. $F(X)$ for $a = 0, 1$

If the X distribution in the data set is the same or similar as in the population of interest:

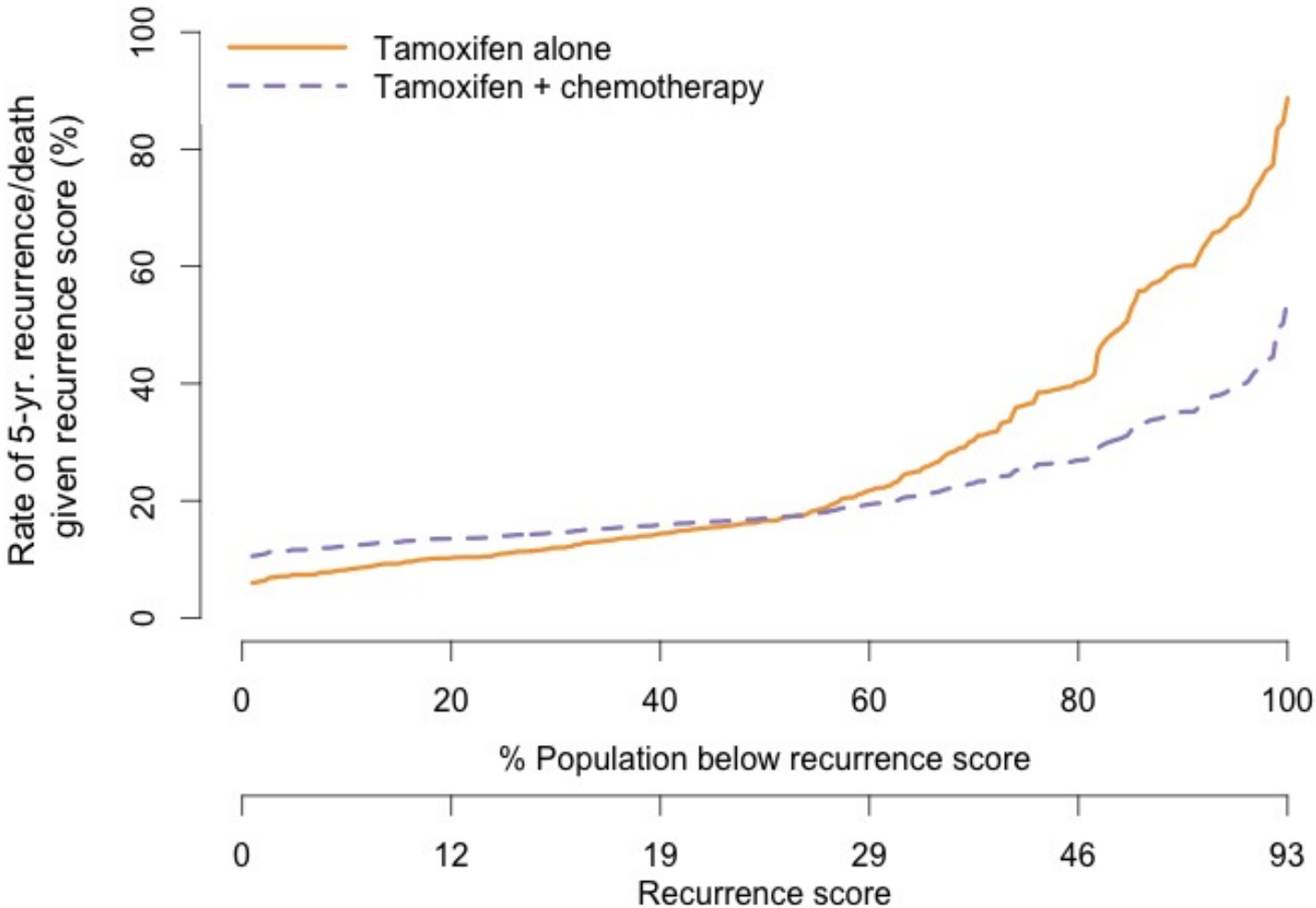
- We recommend aligning the curves for the two treatment groups with respect to marker percentile $F(X)$, rather than marker value X, i.e., plot $E(D(a)|X)$ vs. $F(X)$ for $a = 0, 1$.
- Aligning two curves with respect to marker percentile rather than marker value allows reader to see fraction of population in different regions of the curves.

Risk curves, continued

If the X distribution in the data set is not similar to the population of interest:

- We recommend either not rescaling the marker or rescaling to the marker percentile in the population of interest

Example: Oncotype DX marker in the breast cancer trial



Treatment effect curves

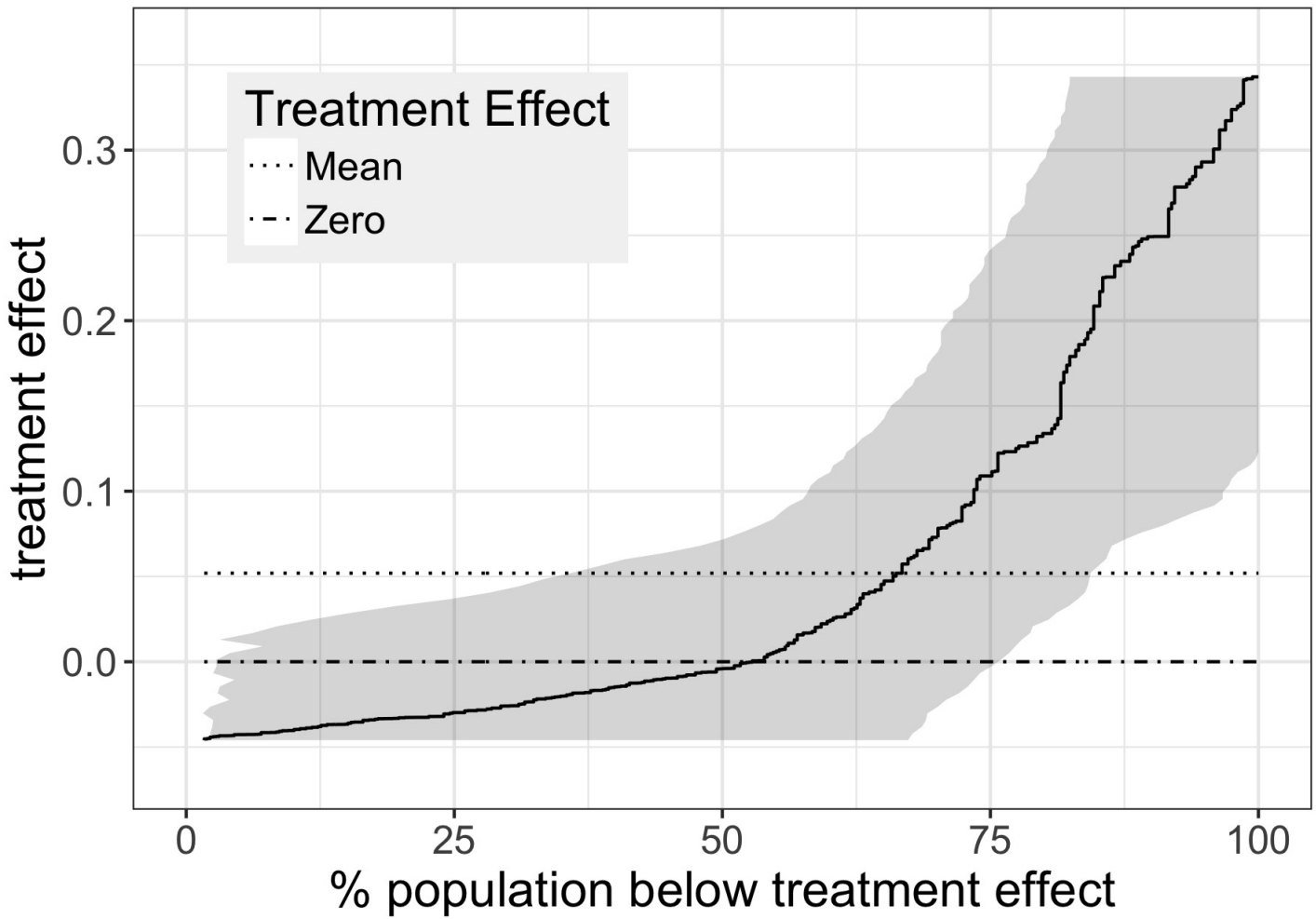
Show the distribution of the marker-specific treatment effect, $\Delta(X) = E(D(0)|X) - E(D(1)|X)$.

Different scales are possible:

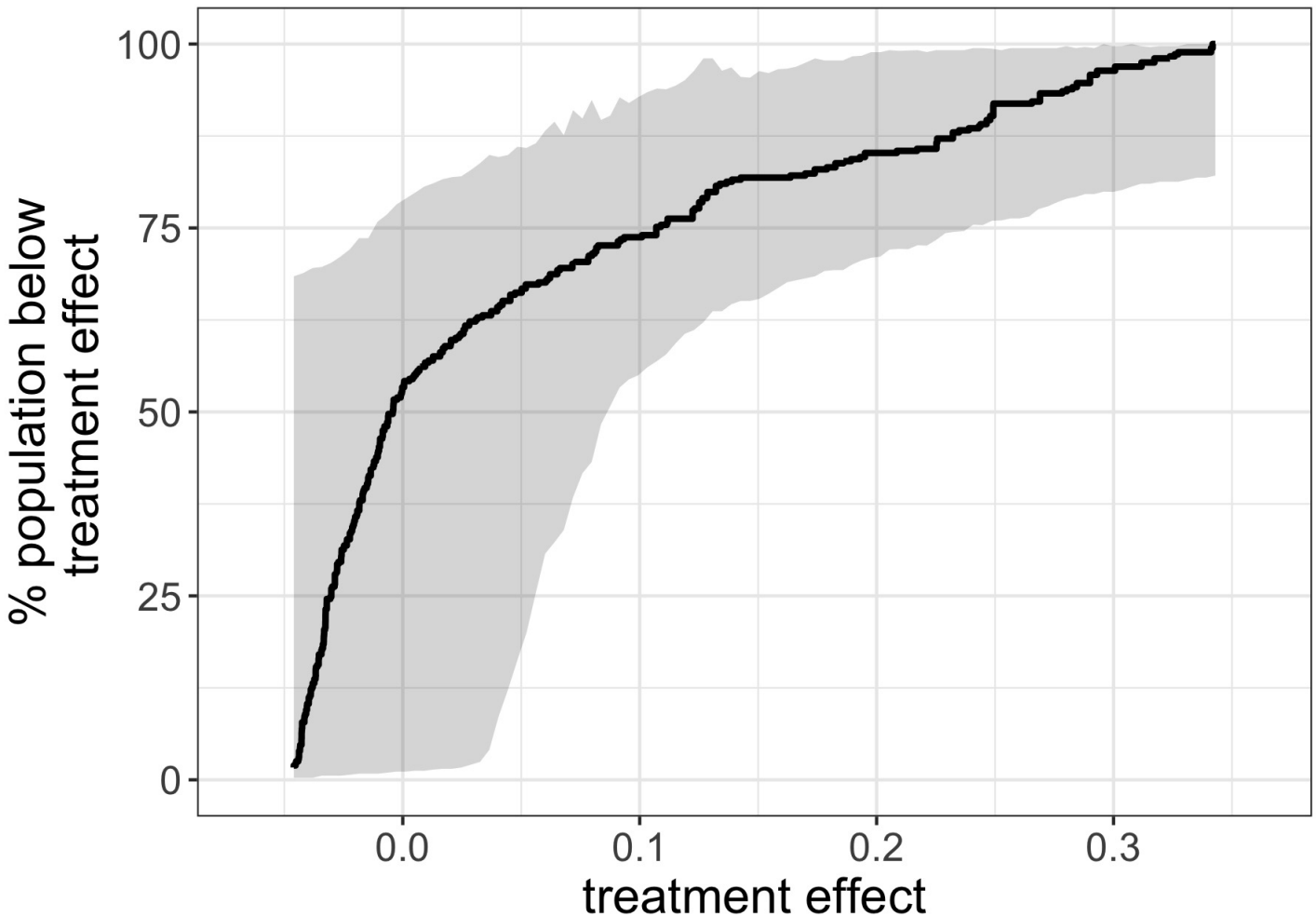
- Reverse-CDF, i.e. $\Delta(X) = \delta$ vs. $F_\Delta(\delta)$. Also called a *predictiveness curve* (Huang et al. 2007).
- Traditional CDF, i.e. $F_\Delta(\delta)$ vs. $\Delta(X) = \delta$.
- Density or histogram of $\Delta(X)$.

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



Estimation

- In RCTs we can generally assume $E(D(a)|X) = E(D|A = a, X)$
- In observational studies if we can assume X contains all of the confounders of treatment on outcome, it is plausible that $E(D(a)|X) = E(D|A = a, X)$
- A traditional/common approach is to use a parametric regression model for $E(D|A = a, X)$
- For low-dimensional X, estimating $E(D|A = a, X)$ via traditional non-parametric smoothing methods is also possible, e.g. kernel-smoothing
- For higher dimensional X, can estimate via other flexible approaches: lasso, random forests, neural networks, generalized additive models, ensemble learners...

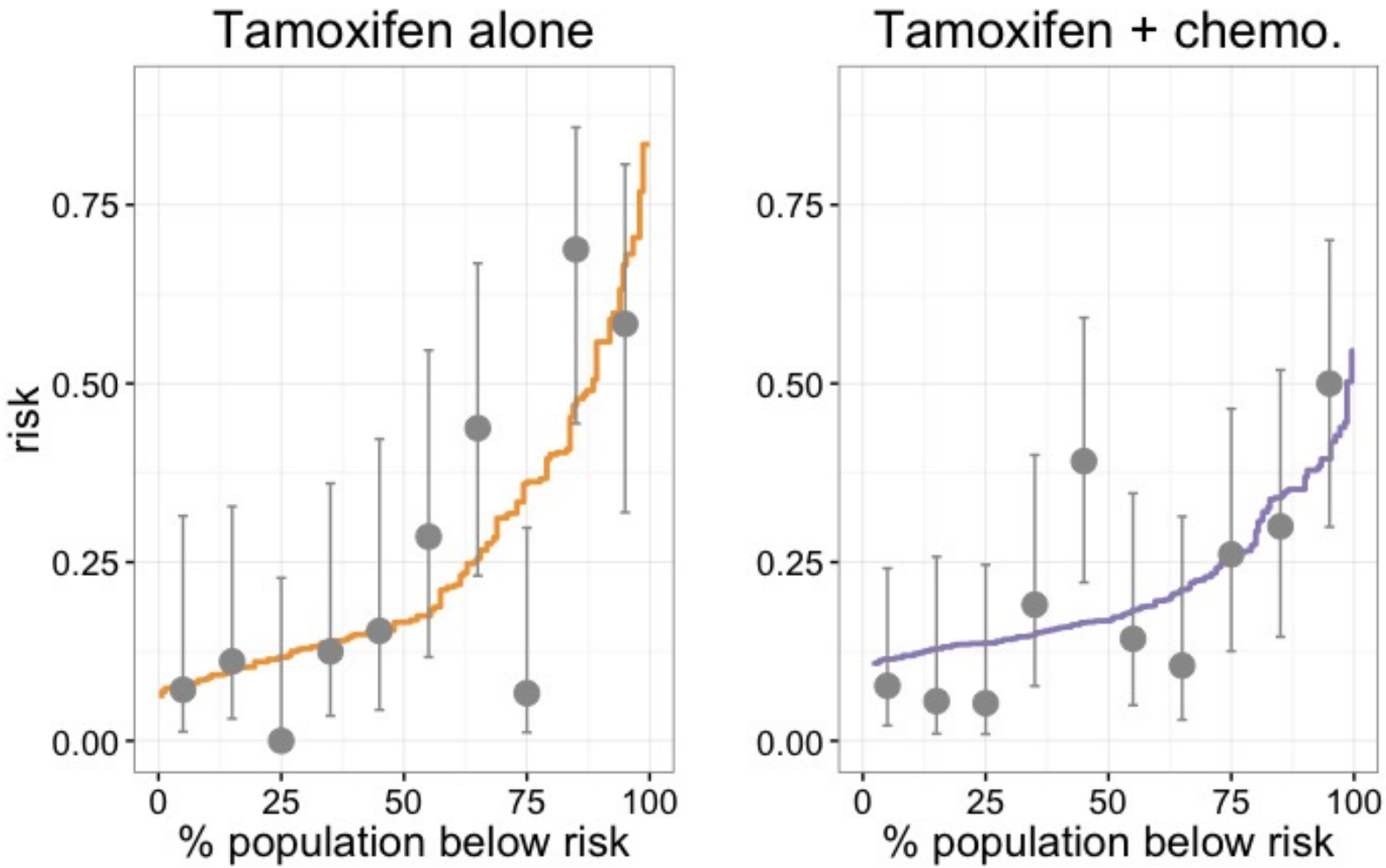
Checking model calibration

Under regression modeling approach, good calibration of $E(D|A = a, X)$ model is essential for validity.

Two approaches to assessing calibration:

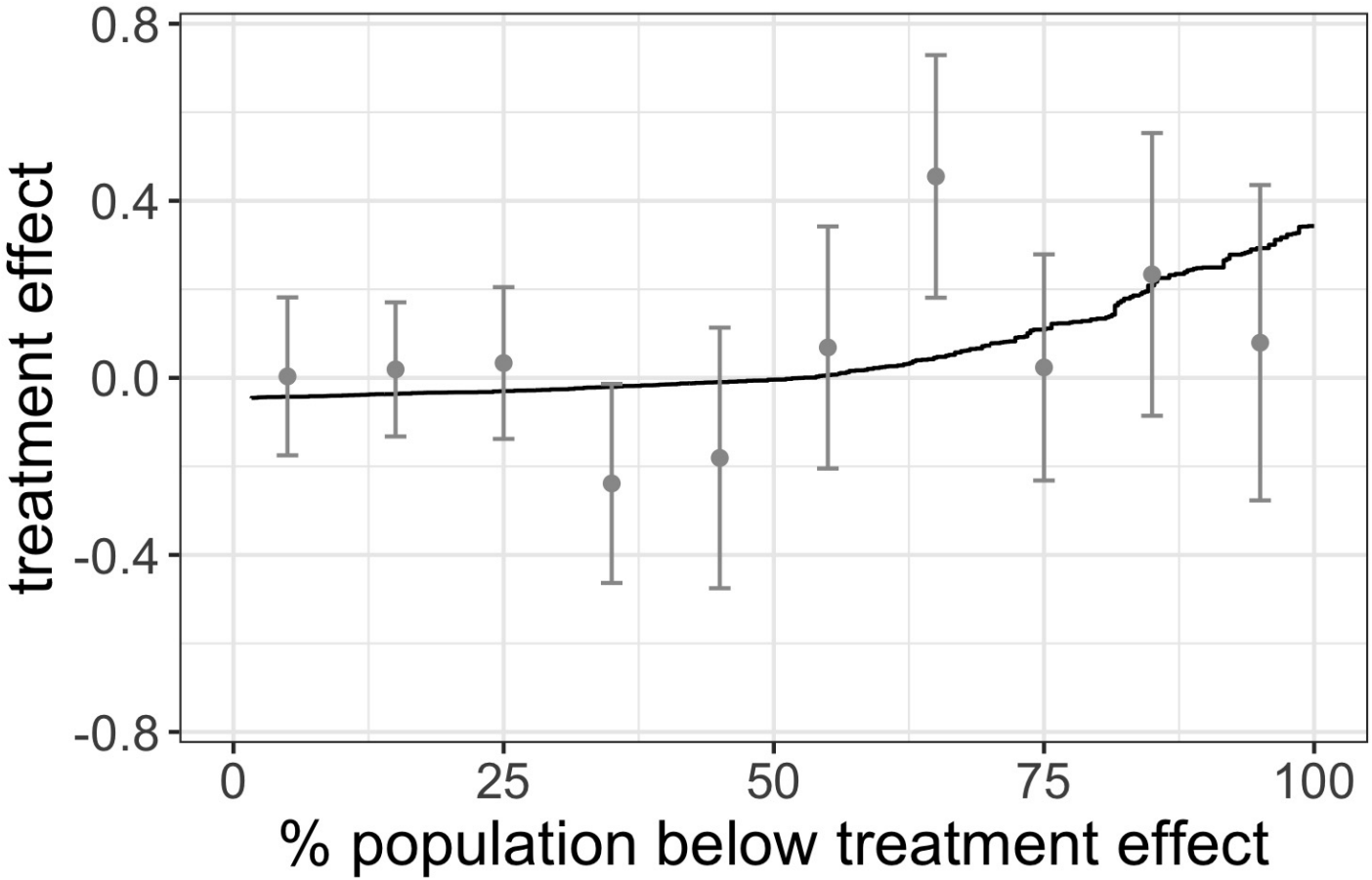
- Visual inspection: overlay observed risks and treatment effects on the plots
- Goodness of fit tests: formally compare observed vs. predicted values using Hosmer-Lemeshow goodness of fit tests
  - These goodness of fit tests are known to be valid when $E(D|A = a, X)$ estimated using a parametric model – active research area for more flexible estimation strategies

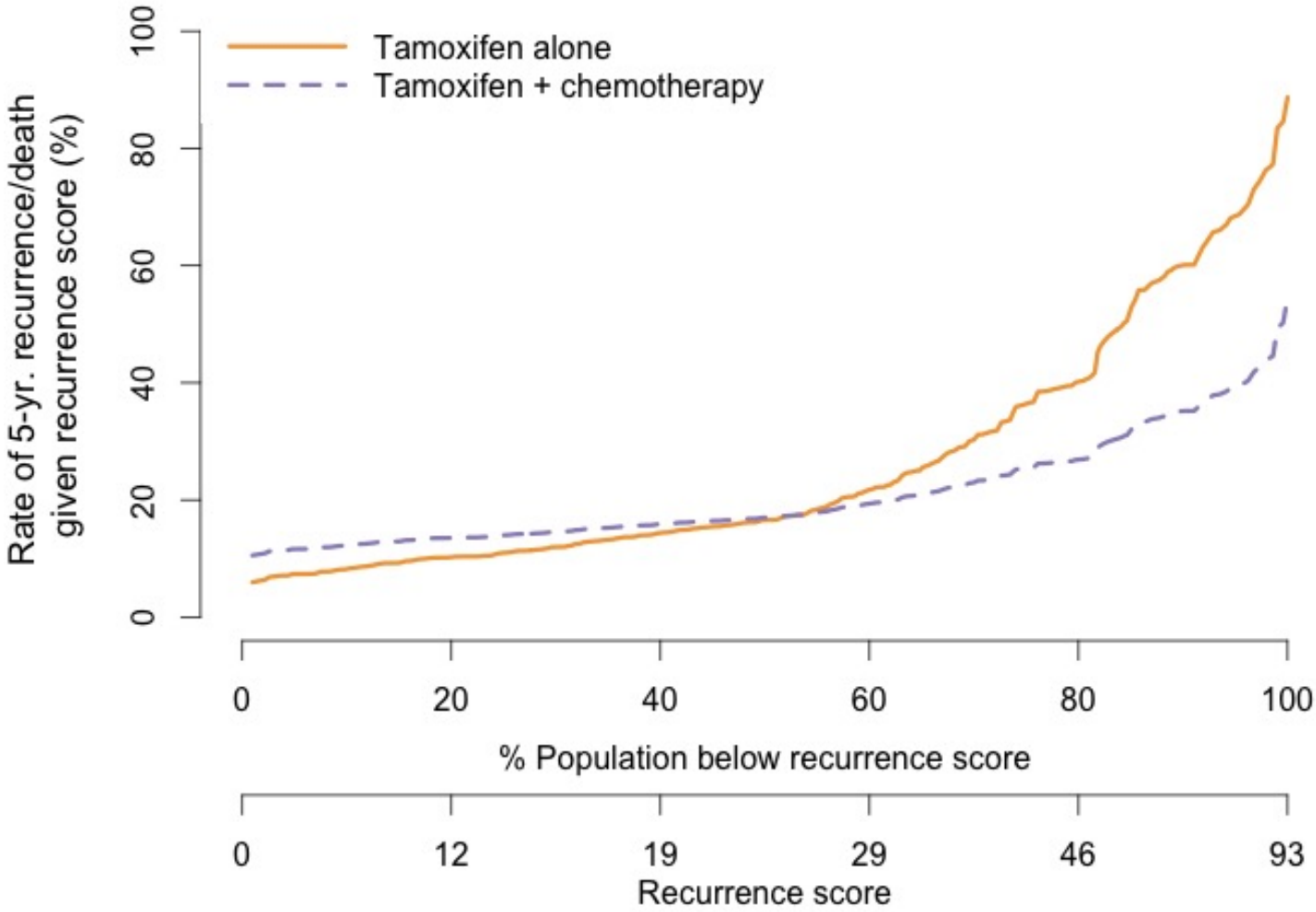
Example: Oncotype DX risk curve calibration



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 for Performance Measure Estimation

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

Mostly 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.

Expected outcome under a marker-based treatment rule

The expected outcome under the rule, $E[D(d)]$ is fundamentally important.

- But, without a baseline value to compare to, it is difficult to interpret the magnitude of $E[D(d)]$.

Evaluating contrasts in expected outcome

Natural to contrast $E[D(d)]$ against $E[D(b)]$, where b is some (possibly stochastic) default rule, i.e. to evaluate *impact*:

$$\begin{aligned}\mathcal{I}(d) &\equiv E[D(b)] - E[D(d)] \\ &= E\{E[D|A = b(X), X]\} - E\{E[D|A = d(X), X]\}.\end{aligned}$$

Reasonable choices for b include:

- Always give standard of care, i.e., $b(X) = 0$ for all X
- Always give treatment, i.e., $b(X) = 1$ for all X
- Always assign the opposite of the treatment suggested by d, i.e. $b = -d$. Leads to the largest effect, useful for testing whether the rule d is at all useful, but not clinically compelling

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

Interpretation of $\mathcal{I}(d)$

Suppose the default is standard of care, i.e. $b(X) = 0$. We have that:

$$\mathcal{I}(d) = E[D(0)] - E[D(d)] = E[\{D(0) - D(1)\}d(X)] = E[\Delta(X)d(X)]$$

Let

- $\beta(d) \equiv E[\Delta(X)|d(X) = 1]$, average treatment effect in the subgroup that are recommended treatment under d
- $\tau(d) \equiv P\{d(X) = 1\}$, proportion of subjects recommended treatment

Then

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

The parameters $\beta(d)$ and $\tau(d)$ are of interest in their own right.

If on the other hand the default is $b(X) = 1$,

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

and the two constituents are

- $\beta(d) = E[-\Delta(X)|d(X) = 0]$: average benefit of standard of care in the subgroup that are recommended $A = 0$ under d
- $\tau(d) = P\{d(X) = 0\}$: the proportion of subjects recommended standard of care

In practice we recommend reporting:

- $E[D(d)]$ and $\mathcal{I}(d)$, for a rational default rule b
- Expected outcomes under "treat all" and "treat none" policies, $\rho_1 = E[D(1)]$ and $\rho_0 = E[D(0)]$, for context
- $\tau(d)$
- $\beta(d)$

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

Empirical estimation, in RCT setting

Let $\mathbb{P}$ denote the empirical probability and $\mathbb{E}$ the empirical mean.

Estimate the proportion of individuals treated using:

$$\widehat{\tau}^e(d) = \mathbb{P}(d(X) = 1)$$

Can also estimate mean outcomes empirically:

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

Note that $\widehat{E}^e(D(d)) = \mathit{IPWE}(d)$ from Section II
Efficiency gains are possible by using augmented IPW estimators or targeted minimum loss-based estimators (TMLEs)

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 $E(D|A,X) = \mu(A,X;\beta)$:

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

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)

Estimation, in observational setting

Inverse probability of treatment weights can be incorporated into empirical estimation, to account for lack of randomization of treatment, i.e., weight observations with $A = a$ by $P(A = a)/P(A = a|X)$, estimated via a regression model.

Valid estimates assuming that X contains all confounders of treatment on outcome

Hernan and Robins, forthcoming; Robins, Hernan, Brumback 2000

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)
- Though still asymptotically normal when $P(\Delta(X) = 0) = 0$, the finite-sample sampling distributions are skewed.
- Confidence intervals have horrible finite sample coverage, whether derived via the bootstrap or via Wald-type methods.
  - Problem more extreme for settings with high-dimensional X, heavy marker/model selection.
- There are partial solutions.
- This is an active research area.

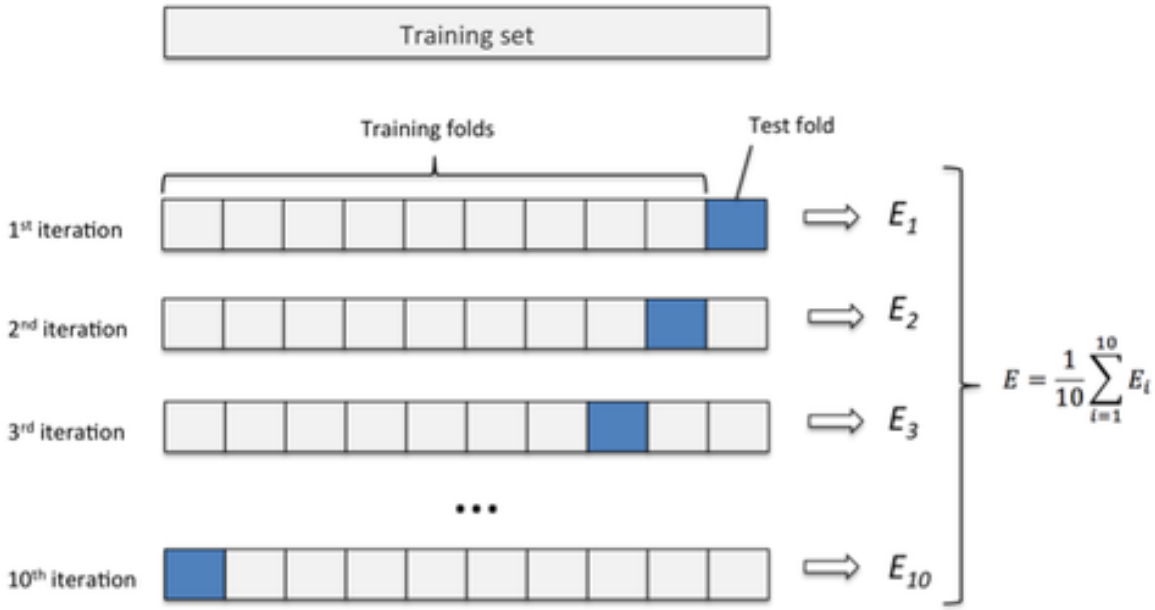
Partial solution to evaluation absent test data: bootstrap bias correction

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 this approach.

Partial solution to evaluation absent test data: cross-validation



Use training fold i to obtain $d_{n,i}$, an estimate of the optimal rule, and use "test folds" to estimate a measure of performance, E_i, e.g. the mean outcome under the rule. The final estimator is an average of these estimated measures. Wald-type normal confidence intervals that perform well in reasonably sized samples can be derived.

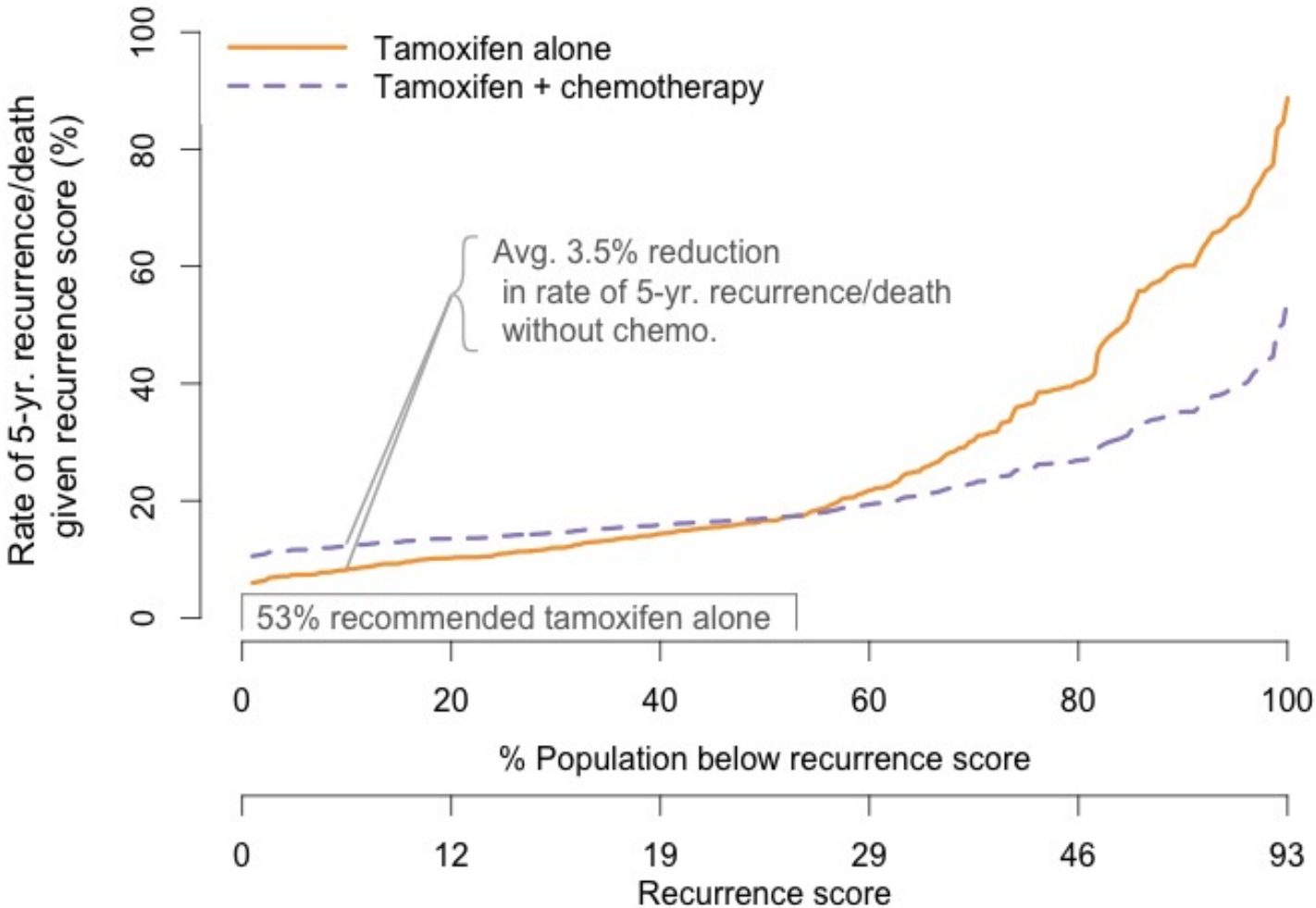
van der Laan & Luedtke *J Caus Inf*, 2015

Example: Oncotype DX marker performance

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

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% to 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.

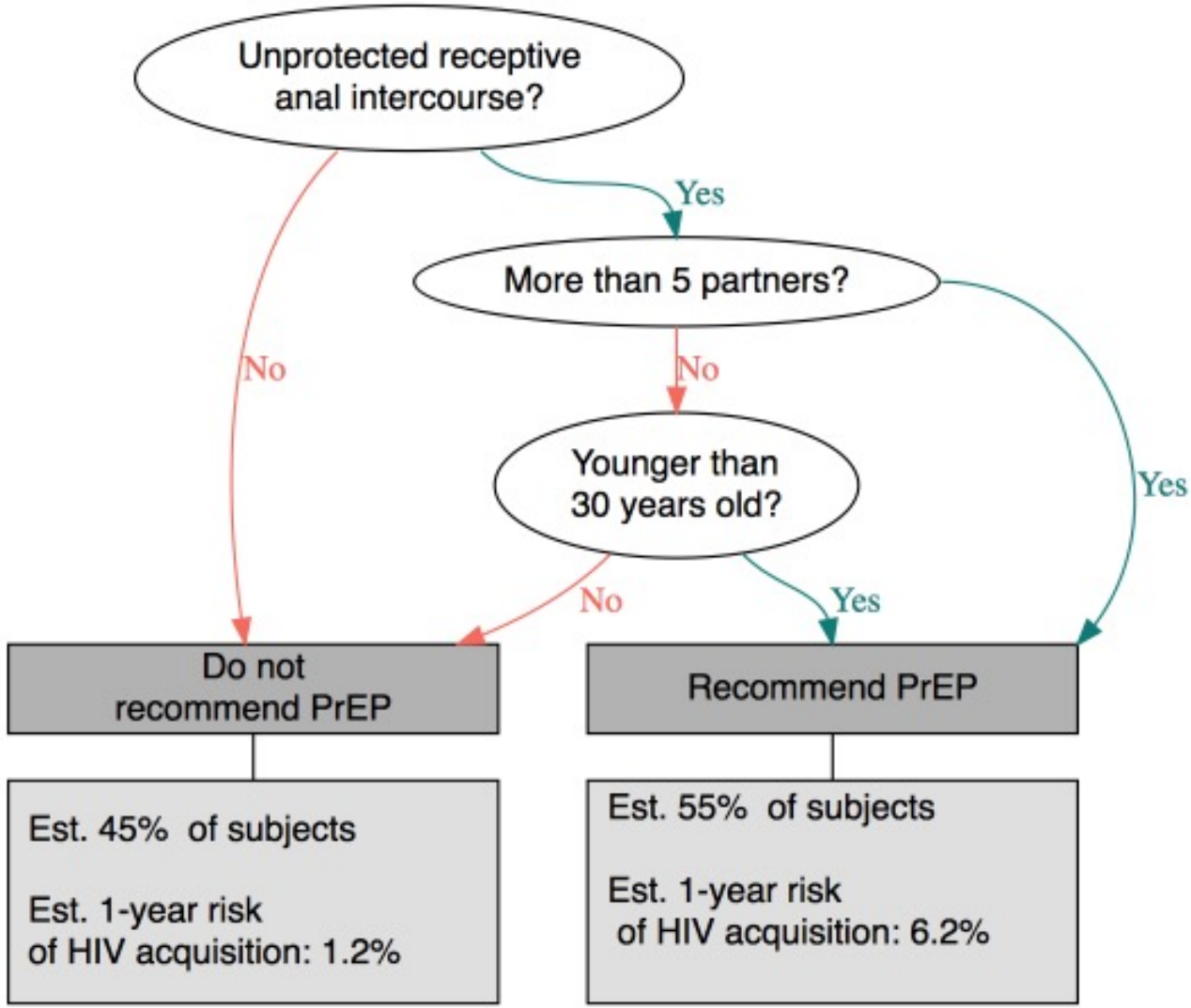
Cox proportional hazards logic regression (Ruczinski et al. 2003) used to model 1-yr. HIV risk without PrEP, $E(D|A=0,X)$, using placebo arm data.

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

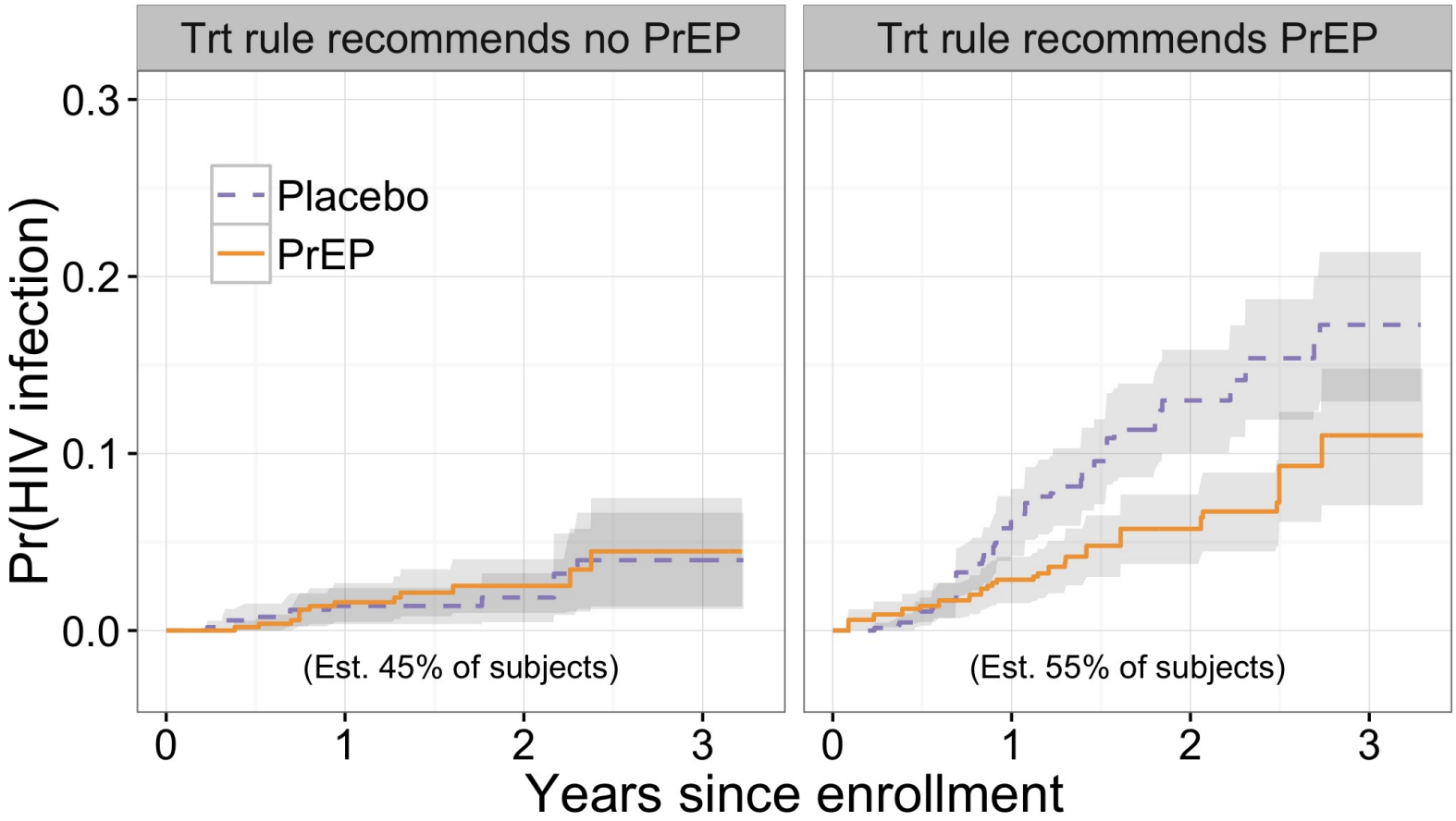
$$d_n(X) = I\left\{\widehat{E}(D|A=0,X) > 0.03\right\}$$

estimated empirically. Bootstrap bias correction used for inference.

Risk-based PrEP recommendation



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

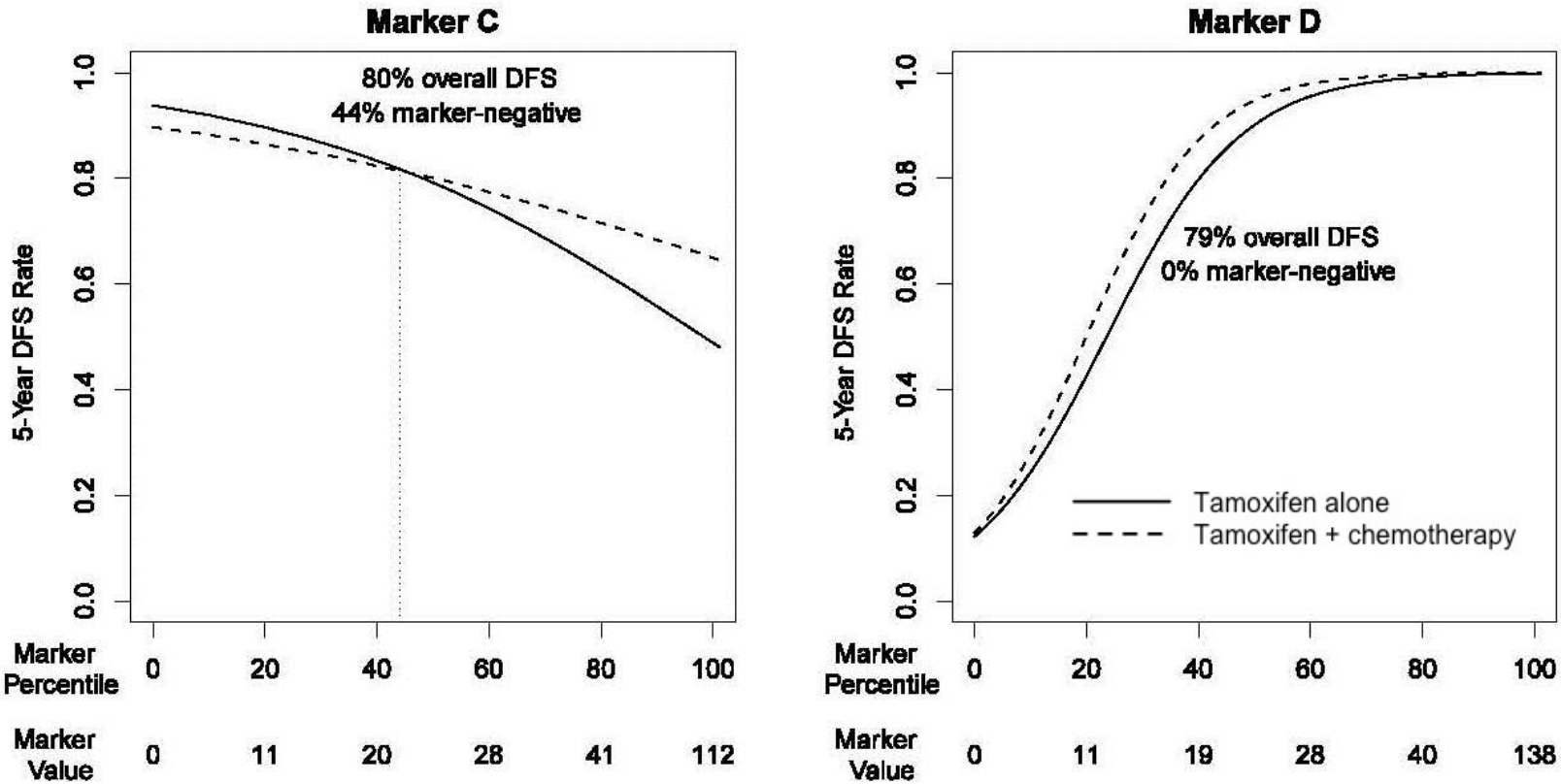
Testing for a marker-by-treatment interaction is a useful first step.

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

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

- 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.

Some biomarker development guidance documents advocate reporting accuracy measures

- without clearly distinguishing between 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 under standard of care

$D(1)$ = potential outcome with treatment

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

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 even in an RCT- so we do not know whether a subject benefits from treatment.

Therefore, in general the accuracy measures are not estimable from 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 measures of marker performance: $E[D(d)]$, $\mathcal{I}(d)$, $\tau(d)$, $\beta(d)$.

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

Other performance measures

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

Variance in treatment effect,

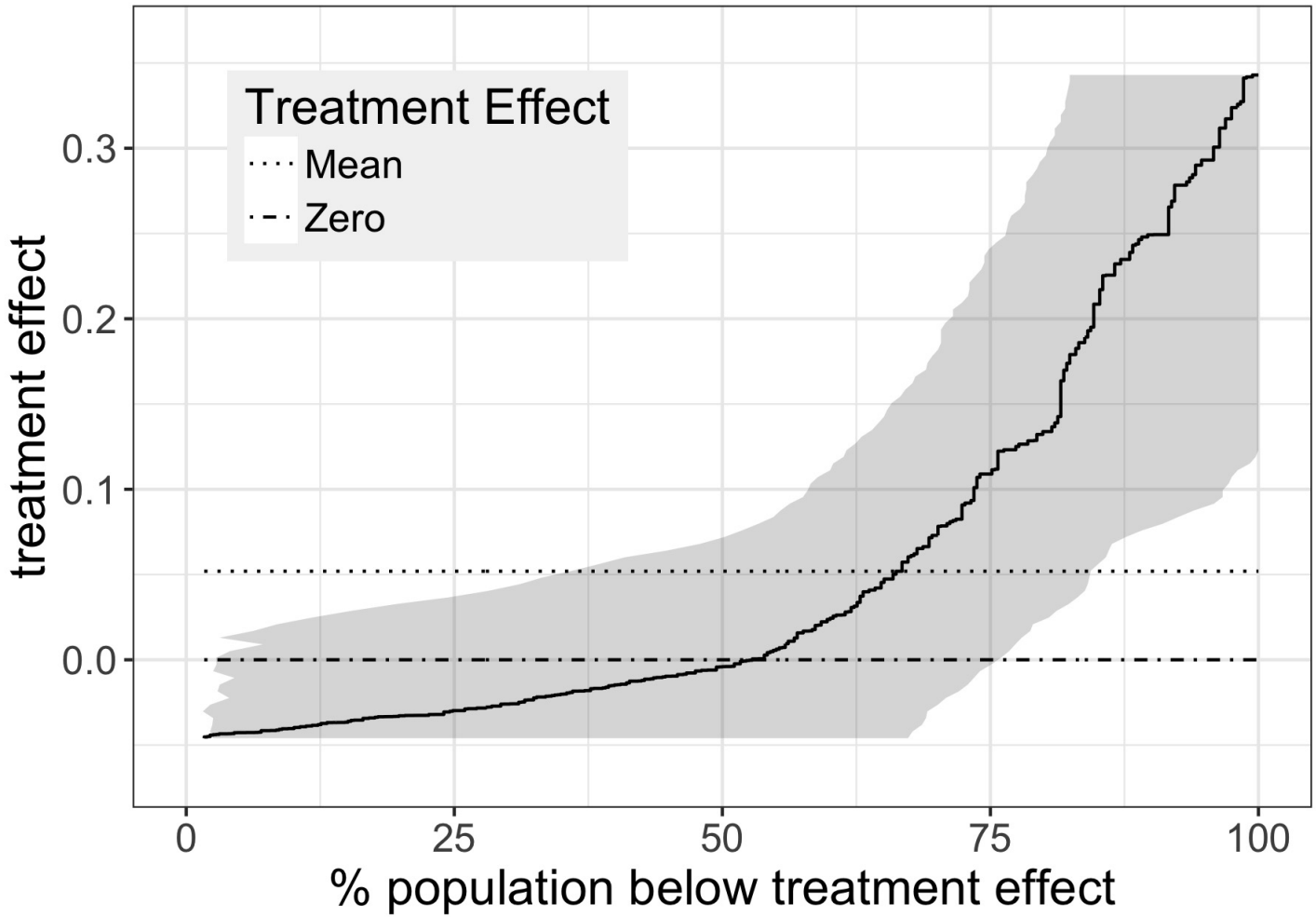
$$V_\Delta \equiv \int (\Delta(X) - (\rho_0 - \rho_1))^2 \, \partial F_\Delta$$

Total gain,

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

- 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

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

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

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

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

Formally incorporating treatment downsides

Incorporating treatment downsides into the treatment rule

The rule

$$d(X) = I\left(E(D(0)|X) - E(D(1)|X) > 0\right)$$

is optimal if the goal is to minimize $E[D(d)]$.

If, however, there are additional downsides of treatment not captured in D, it is compelling to consider rules of the form

$$d^{\delta}(X) = I\left\{E(D(0)|X) - E(D(1)|X) > \delta\right\},$$

for $\delta > 0$.

Decision theory justification

Let C_D be the cost/disutility of one unit of D and C_A be the cost/disutility of treatment.

The net benefit of rule d_δ captures both its impact on outcomes and its impact on treatment. Compared to a default rule $b(X) = 0$ for all X,

$$\begin{aligned}
\text{NB}(d^\delta) &\equiv (\text{Exp. cost under } b) - (\text{Exp. cost under } d^\delta) \\
&= [E(D(0)) - E(D(d^\delta))]C_D - P(d^\delta(X) = 1)C_A \\
&= \mathcal{I}(d^\delta)C_D - \tau(d^\delta)C_A
\end{aligned}$$

Net benefit is maximized under rule

$$d^\delta(X) = I\left\{E(D(0)|X) - E(D(1)|X) > \delta\right\}.$$

Vickers et al. *Trials* 2007; Janes et al. *MDM* 2014

Composite outcome justification

Alternatively, define the composite outcome

$$D^* = DC_D + AC_A.$$

The same rule,

$$d^{\delta}(X) = I\left\{E(D(0)|X) - E(D(1)|X) > \delta\right\}$$

minimizes the expected composite outcome, $E[D^*(d)]$.

Choice of treatment effect threshold, δ

Decision theory suggests that the optimal δ corresponds to the cost/disutility of treatment A relative to the cost/disutility of one unit of the outcome D

- E.g. if treatment-associated toxicity is 1/10 the cost of the binary clinical outcome, the optimal $\delta = 0.10$

"Cost" is used broadly here; units may be dollars or quality-adjusted life years (QALYs) or probabilities of downstream events.

Vickers et al. *Trials* 2007; Janes et al. *MDM* 2014

The performance of treatment rule $d^\delta(X)$ can be evaluated using the aforementioned metrics:

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

Evaluating performance using net benefit

Furthermore, if the optimal rule for maximizing net benefit is used ($\delta = \frac{C_A}{C_D}$), the net benefit in C_D units is

$$\text{NB}(d^\delta) = \mathcal{I}(d^\delta) - \tau(d^\delta)\delta$$

- Appealing that this NB formulation depends only on δ, and not on C_D or C_A

Thus, $\text{NB}(d^\delta)$ can be interpreted as the discounted reduction in the expected outcome under marker-based treatment.

Special case: if $\delta = 0$, $\text{NB}(d^\delta) = \mathcal{I}(d^\delta)$.

If $b(X) = 1$ is the default rule,

$$\begin{aligned}
\text{NB}(d^\delta) &\equiv (\text{Exp. cost under } b) \; - \; (\text{Exp. cost under } d^\delta) \\
&= [E(D(1)) - E(D(d^\delta))]C_D \\
&\quad + [1 - P(d^\delta(X) = 1)]C_A \\
&= \mathcal{I}(d^\delta)C_D \; + \; \tau(d^\delta)C_A \\
&= \mathcal{I}(d^\delta) \; + \; \tau(d^\delta)\delta,
\end{aligned}$$

where the last line holds if $\delta = C_A/C_D$ and NB is in C_D units.

Thus, $\text{NB}(d^\delta)$ can be interpreted as the augmented reduction in the expected outcome under marker-based treatment.

Example: Oncotype DX marker performance

Suppose it is determined that $\Delta(X) > 0.01$ is large enough to warrant a chemotherapy recommendation; women with $\Delta(X) < 0.01$ should be recommended no chemo.

We estimate that using this rule to recommend no chemo. would:

- Allow 56.3% of women to avoid chemo.
- Reduce the 5-yr. recurrence/death rate in this subgroup by 1.7%
- Reduce the population 5-yr. recurrence/death rate by 0.5%
- Yield a NB of 0.011. Thus, 1.1% is the augmented reduction in the 5-yr. recurrence/death rate.

Optimal treatment rule under resource constraints

A related concern is finite resources that may place constraints on the size of the treated population.

In this setting, the optimal rule is again

$$d^{\delta}(X) = I\{E(D(0)|X) - E(D(1)|X) > \delta\},$$

where δ is selected to ensure $P\left[E(D(0)|X) - E(D(1)|X) > \delta\right] = \tau$.

Above evaluation metrics $E[D(d^{\delta})]$, $\mathcal{I}(d^{\delta})$, and $\beta(d^{\delta})$ apply ($\tau(d^{\delta}) = \tau$ by definition).

Summary

- Descriptive devices are useful for visualizing and interpreting data
- Impact of rule on expected outcome, and its constituents, are recommended performance measures
- Contrasts in these measures are recommended for comparing markers or rules
- Inference is challenging in the absence of independent test data
- Extensions allow treatment downsides to be incorporated into the treatment rule and its evaluation