10th International Conference on
Multiple Comparison Procedures,
Riverside, CA, June 20–23, 2017

Biomarker Subgroup Testing, Misclassification, and Missing Data

Gene Pennello, PhD, Team Leader,
Diagnostic Devices Branch II,
Division of Biostatistics, FDA / CDRH

Jingjing Ye, PhD,
Division of Biometrics V,
Office of Biostatistics, FDA / CDER

Abstract

Biomarker Subgroup Testing, Misclassification, and Missing Data

The objective of precision medicine has been stated as treating the "right patient with the right drug at the right time". Many predictive biomarkers facilitate precision medicine by explaining a clinically significant amount of the variation in a treatment effect. The anticipation that the treatment will only be effective in a biomarker-defined subgroup means that many proposed procedures for testing treatment effect overall and in one or more biomarker-defined subgroups are unsatisfactory. The clinical objective is not to find the largest population in whom statistical significance of the treatment effect is retained, but to determine the population (if it exists) in whom the effect is homogeneous and clinically significant. In this talk, we'll discuss frequentist and Bayesian testing procedures that have been designed to address the clinical objective of predictive biomarkers. We'll also quantify how biomarker measurement error attenuates the difference in treatment effect between biomarker defined subgroups. We'll also show that missing biomarker results (e.g., specimens unavailable or unevaluable for biomarker testing) can be addressed with Bayesian selection models even when minimal assumptions on the missing data mechanism mean that model parameters aren't fully identified.

Outline

- Biomarker Intended Uses
- Biomarker Device (Test) Evaluation
- Biomarker Subgroup Evaluation
  - Companion, Complementary Diagnostics
  - Frequentist, Bayesian
- Biomarker Misclassification
- Missing Data
- Concluding Remarks

Intended Uses for Biomarkers

- **Diagnosis,** in symptomatic patients.
- **Screening**, in asymptomatic patients.
- **Early detection**, enabling intervention at an earlier and potentially more curable stage than under usual clinical diagnostic conditions.
- **Monitoring**, e.g., of disease response during therapy, with potential for adjusting level of intervention (e.g. dose) on a dynamic and personal basis.
- **Risk assessment,** leading to preventive interventions for those at sufficient risk.
- **Prognosis**, allowing for more (less) aggressive therapy for patients with worse (better) prognosis.
- **Prediction** of safety or efficacy of a specific therapy to aid benefit/risk assessment in individual patients (e.g., predict response, predict SAE, monitor response to adjust schedule or dose or discontinue).

Last three involve prediction of a future state of health.

Test Performance Evaluation

- **Analytical performance** - does my test measure the analyte I think it does? Correctly? How reliably?

- **Clinical performance** - does my test result correlate with target condition of interest in a clinically significant way?

- **Clinical Utility** - does my test support clinical decisions for patient management such as effective treatment or preventive strategies?

Fryback-Thornbury Model

Level	Objective	Study Type*
1	Technical efficacy	*Analytical performance*
2	Diagnostic accuracy efficacy	*Clinical performance*
3	Diagnostic thinking efficacy	
4	Therapeutic efficacy	
5	Patient outcome efficacy	*Clinical outcome*
6	Societal efficacy	

Fryback DG and Thornbury JR. The Efficacy of Diagnostic Imaging. *Med Decis Making* 1991; 11(2): 88-94.

*FDA CDRH/CBER Guidance. *Design Considerations for Pivotal Clinical Investigations for Medical Devices*, 2013 (Sections 7.7, 8).

Analytical Performance Studies

- **Bias**, relative to a reference method for measuring analyte.
- **Precision**. Measurement variation in repeated testing.
  - *repeatability* of the test result taken under the same set of conditions (e.g., testing sample replicates in the same run)
  - *reproducibility* of test result taken under different conditions (e.g., testing sample replicates in different labs)
- **Limit of Detection**. Smallest analyte level detected reliably.
- **Reagent Stability**. Shelf-life, in-use, and shipment.
- **Analytical Specificity**. Measurement of a specific analyte in the presence of potential interfering substances, cross-reactivity, or cross-contamination.
- **Commutability** of different sample types, when processed samples are used in place of clinical samples.

CLSI. A Framework for Utilizing CLSI Guidelines to Evaluate Clinical Laboratory Measurement Procedures; 2nd ed, CLSI report EP19. Wayne, PA: Clinical and Laboratory Standards Institute; 2015.

Predictive, Prognostic Markers

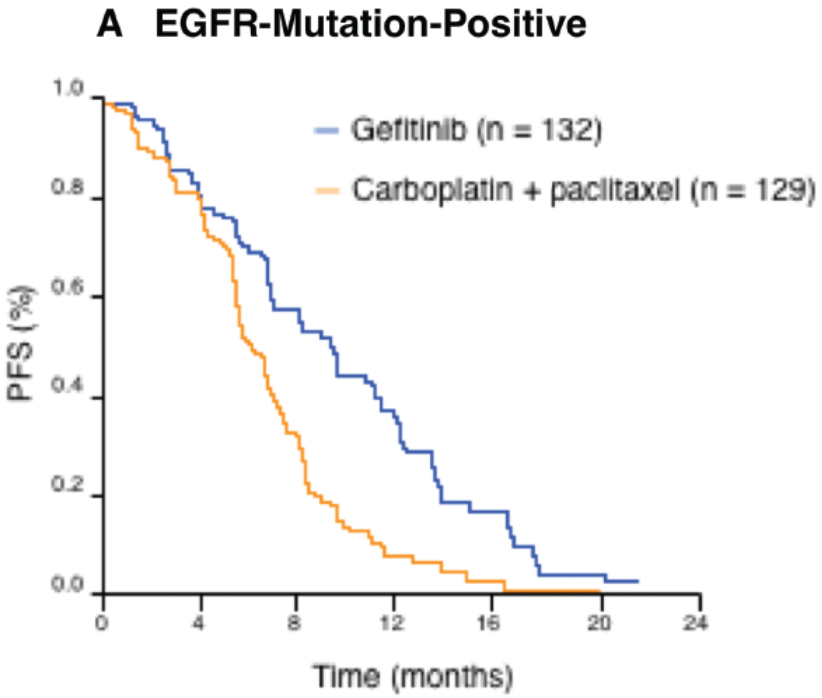
- **Predictive biomarker** informs on likely outcomes with specific treatments (e.g., relative sensitivity or resistance).
  - Other names: treatment selection biomarker, CDx
- **Prognostic biomarker** is biological characteristic indicating likelihood of disease progression in a homogeneous population of patients, either not receiving therapy (natural course) or on a standard therapy.
  - inform on outcomes independent of specific treatment (i.e. in oncology, ability of tumor to proliferate, invade, and/or spread)

Yamauchi H, Stearns V, Hayes DF. When Is a Tumor Marker Ready for Prime Time? A Case Study of c-erbB-2 as a Predictive Factor in Breast Cancer. J Clin Oncol. 2001 Apr 15;19(8):2334-56.

Intended Uses / Claims

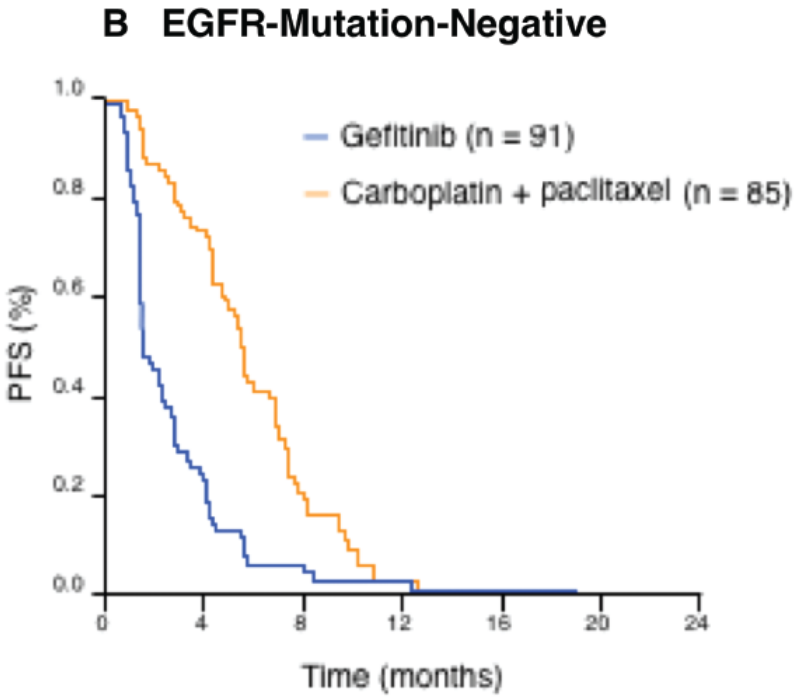
- Companion Diagnostic:
  - Provides information that is essential for the safe and effective use of a corresponding therapeutic product, allowing its benefits to exceed its risks.
  - E.g., defines the population for whom a therapeutic product is indicated.
- Complementary Diagnostic:
  - Provides clinically useful information about a therapeutic product yet is not a prerequisite for the therapeutic product's use (*not an official FDA definition*).

Qualitative Interaction



No. of patients at risk

Time (months)	0	4	8	12	16	20	24
Gefitinib	132	108	71	31	11	3	0
Carboplatin plus paclitaxel	129	103	37	7	2	1	0



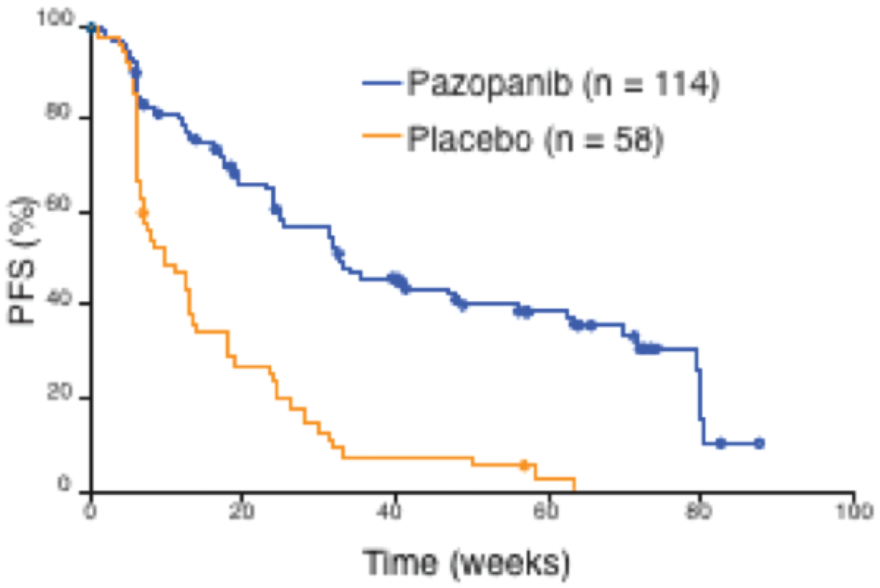
No. of patients at risk

Time (months)	0	4	8	12	16	20	24
Gefitinib	91	21	4	2	1	0	0
Carboplatin plus paclitaxel	85	58	14	1	0	0	0

Polley MC, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and Practical Considerations for Clinical Evaluation of Predictive Biomarkers, J Natl Cancer Inst;2013;105:1677–1683

Quantitative Interaction

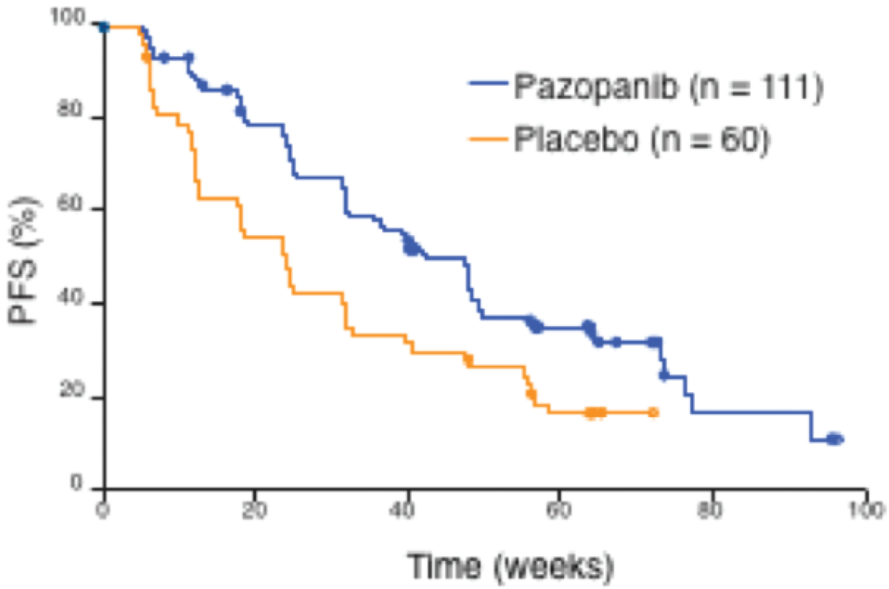
A Interleukin 6 high



No. of patients at risk

Time (weeks)	0	20	40	60	80	100
Pazopanib	114	64	42	25	4	0
Placebo	58	15	4	1	0	0

B Interleukin 6 low



No. of patients at risk

Time (weeks)	0	20	40	60	80	100
Pazopanib	111	77	53	26	4	0
Placebo	60	31	19	8	0	0

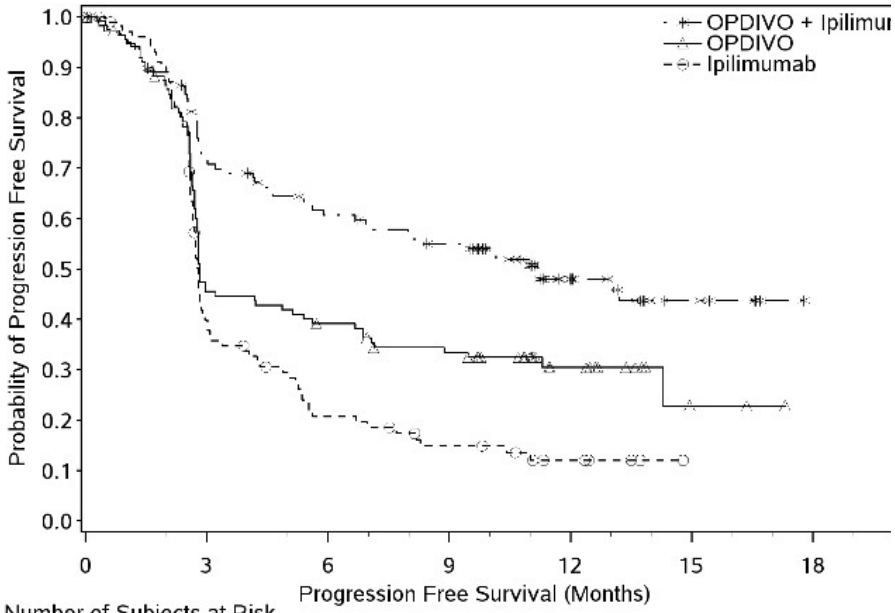
Polley MC, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and Practical Considerations for Clinical Evaluation of Predictive Biomarkers, J Natl Cancer Inst;2013;105:1677–1683

PD-L1 IHC 28-8 pharmDx

- PD-L1 expression in tumor specimens from patients with non-small cell lung cancer (NSCLC) and melanoma.
- Indications for Use
  - PD-L1 expression as detected by PD-L1 IHC 28-8 pharmDx in non-squamous NSCLC may be associated with enhanced survival from OPDIVO ® (nivolumab).
  - Positive PD-L1 status as determined by PD-L1 IHC 28-8 pharmDx in melanoma is correlated with the magnitude of the treatment effect on progression-free survival from OPDIVO®.

PD-L1 IHC 28-8 pharmDx

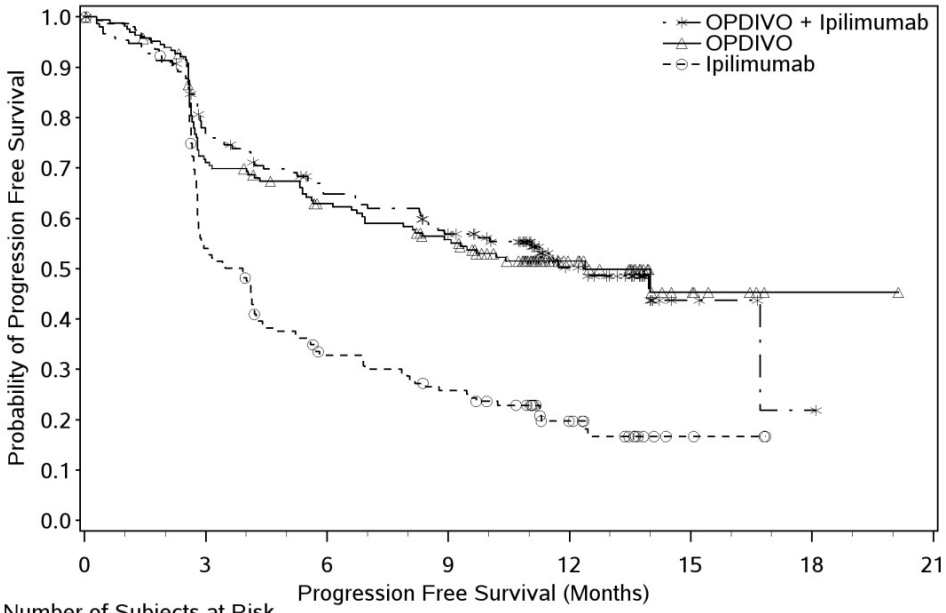
PD-L1 < 1%



Number of Subjects at Risk

OPDIVO + Ipilimumab	123	82	65	57	26	6	0
OPDIVO	117	50	42	34	13	2	0
Ipilimumab	113	39	19	12	5	0	0

PD-L1 ≥ 1%



Number of Subjects at Risk

OPDIVO + Ipilimumab	155	113	91	78	32	4	1	0
OPDIVO	171	115	97	83	34	7	1	0
Ipilimumab	164	83	47	36	16	3	0	0

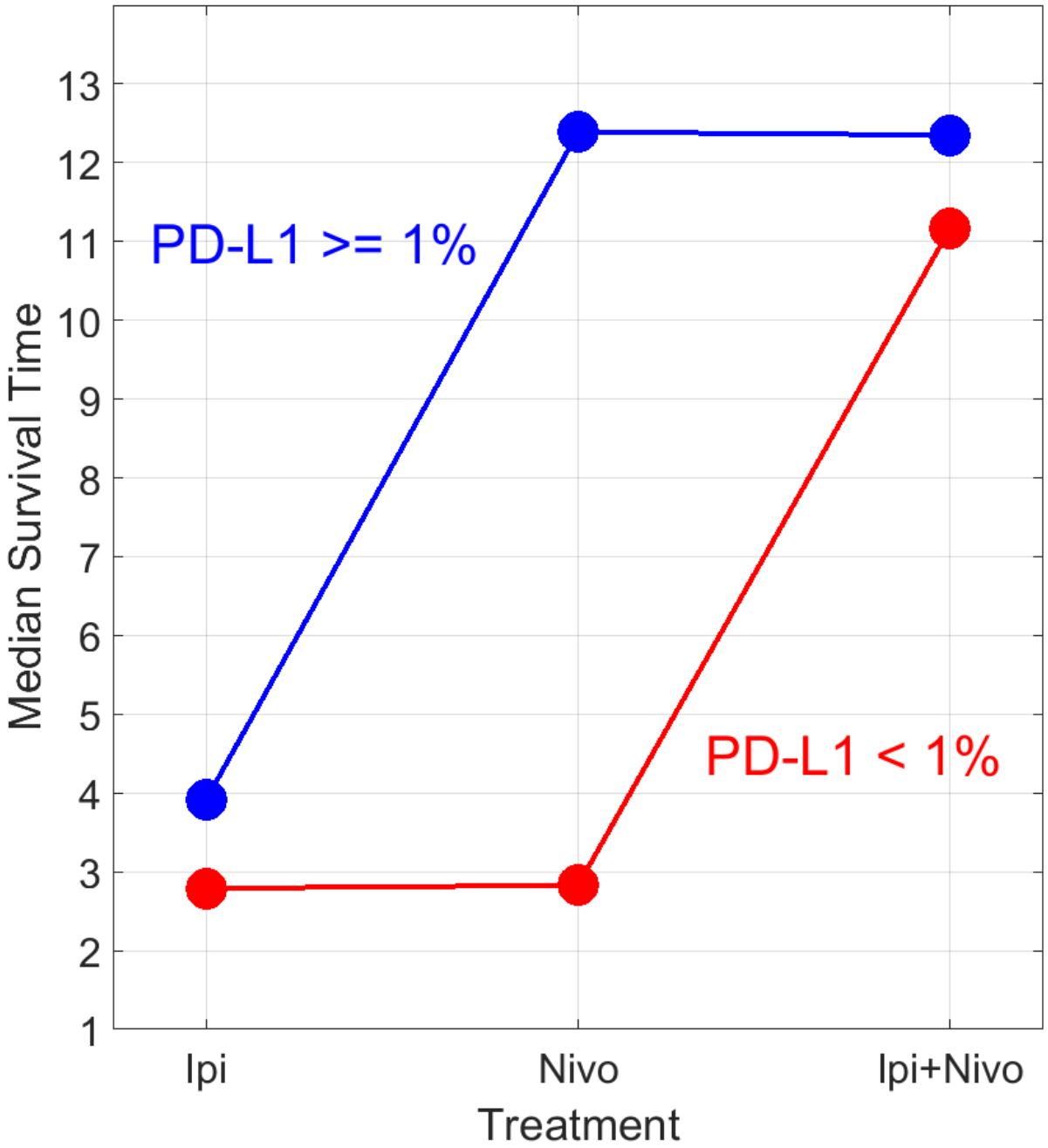
Clinical Trial CA206097, Melanoma*

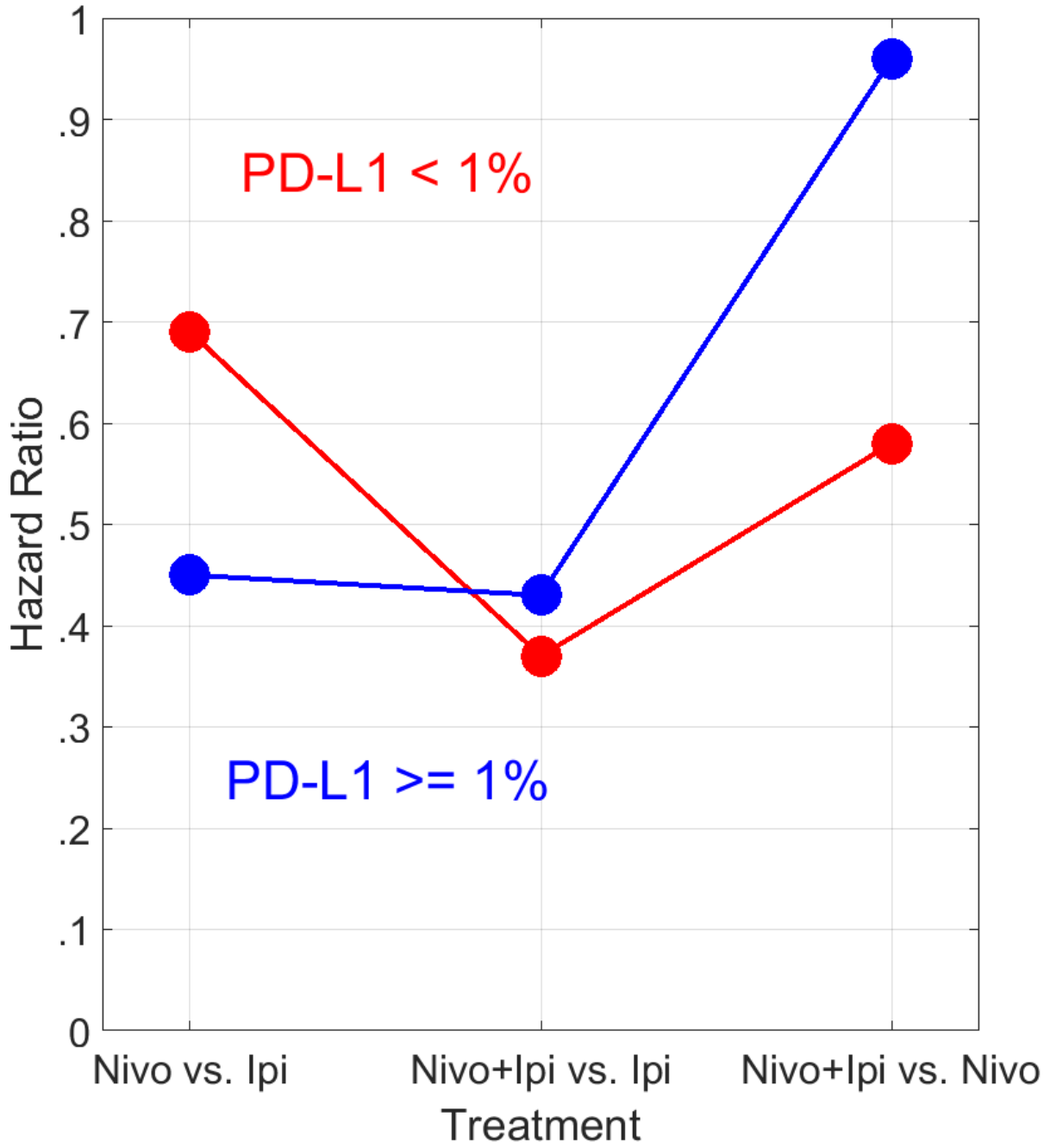
Table 8. Summary of PFS and Hazard ratios for CA209067 study by PD-L1 expression status

Median Progression Free Survival (95% CI)			
PD-L1 Expression Level	**Nivolumab monotherapy**	**Nivolumab +ipilimumab combination therapy**	**Ipilimumab monotherapy**
<1%	2.83 (2.76 , 5.13)	11.17 (6.93, NR)	2.79 (2.66, 2.96)
≥1%	12.39 (8.11, NR)	12.35 (8.51, NR)	3.91 (2.83, 4.17)
Hazard Ratios (95% CI)			
	Nivolumab vs. ipilumumab	Nivolumab + ipilimumab vs. ipilimumab	Nivolumab + ipilimumab vs. nivolumab*
<1%	0.69 (0.5, 0.93)	0.37 (0.26, 0.52)	0.58 (0.41, 0.81)
≥1%	0.45 (0.35, 0.57)	0.43 (0.32, 0.58)	0.96 (0.70, 1.33)

*Exploratory analysis

*previously untreated, unresectable or metastatic melanoma


13
12
11
10
9
8
7
6
5
4
3
2
1
Median Survival Time
PD-L1 >= 1%
PD-L1 < 1%
Ipi
Nivo
Ipi+Nivo
Treatment


PD-L1 < 1%
PD-L1 >= 1%
Hazard Ratio
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
Nivo vs. Ipi
Nivo+Ipi vs. Ipi
Nivo+Ipi vs. Nivo
Treatment

Interpretation and Causality

- Cochran: What could be done to clarify the step from association to causation?
- Fisher: **Make your theories elaborate**.
- "… when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold".
  - Cochran WG. The planning of observational studies of human populations. *JRSS A* 1965; 234-266. (p. 252).

FDA Guidance, Predictive Markers

Alosh M, Fritsch K, Huque M, Mahjoob K, Pennello G, Rothmann M, Russek-Cohen E, Smith F, Wilson S, Yue L. Statistical Considerations on Subgroup Analysis in Clinical Trials, *Statist Biopharm Res* **2015**; 7(4):286–304.

Beaver JA; Tzou A; Blumenthal GM; McKee AE; Kim G; Pazdur R; Philip R. An FDA Perspective on the Regulatory Implications of Complex Signatures to Predict Response to Targeted Therapies. *Clin Cancer Res*. **2017**, 23 (*6*), 1368-1372.

US FDA. Guidance on Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products. US FDA: Silver Spring, MD, **2012**.

US FDA. In Vitro Companion Diagnostic Devices, US FDA: Silver Spring MD, **2014**.

US FDA. Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product. US FDA: Silver Spring MD, **2016**.

US FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US); **2016**.

Biomarker Subgroup Evaluations

Purposes of Subgroup Evaluation

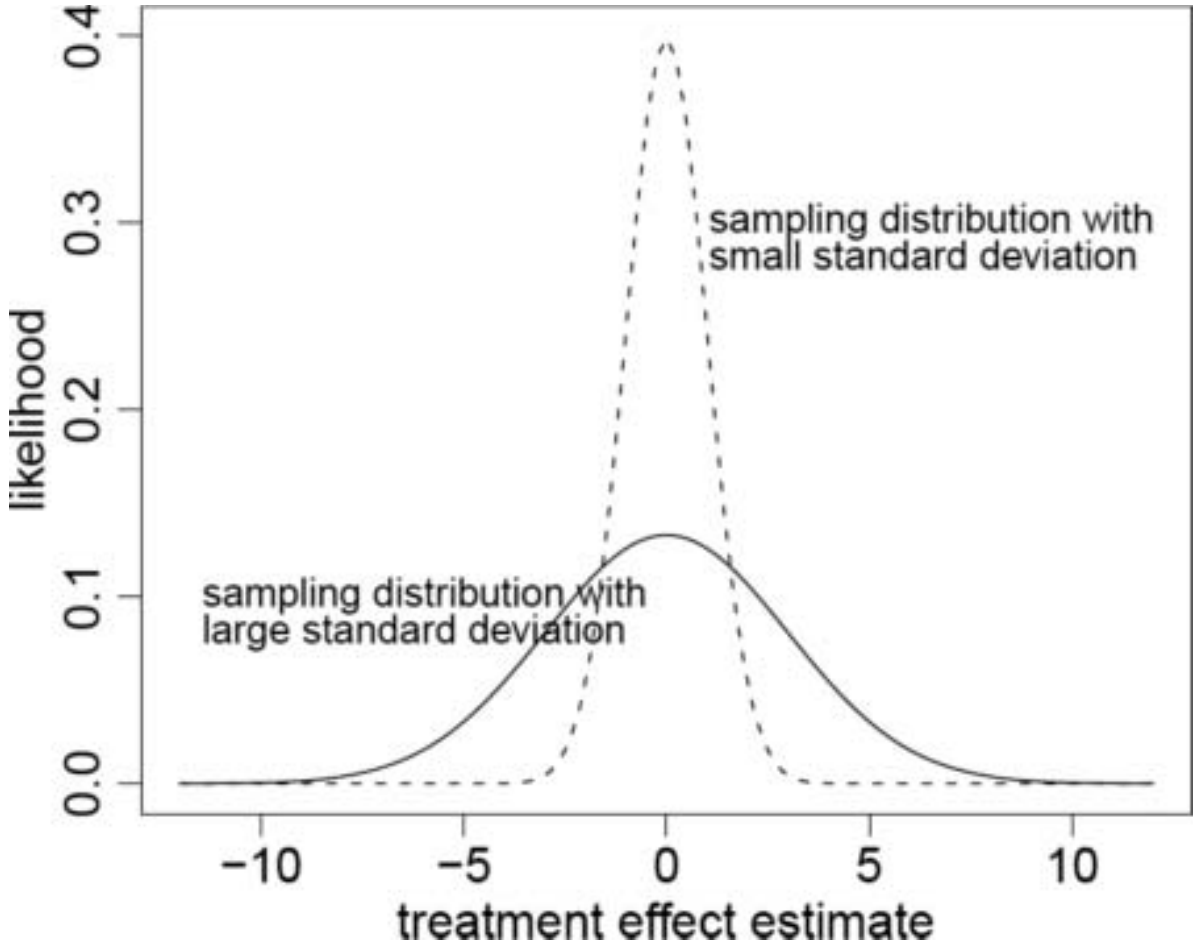
1. **Investigate the consistency of treatment effect** across subgroups of clinical importance.
2. **Explore the treatment effect** across different subgroups within an overall non-significant trial.
3. **Evaluate safety profiles** limited to one or a few subgroup(s).
4. **Establish efficacy in the targeted subgroup** when included in a confirmatory testing strategy of a single trial.

Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. BMC Med Res Methodol. 2016 Feb 18; 16:20

Check Consistency Across Subgroups

- Trial was designed to establish effectiveness in overall study population.
- Heterogeneity of treatments effects *not* anticipated to have any particular pattern *a priori* (treatment effects exchangeable).

The Subgroup Problem



Gelman, Hill, Yajima, Why We (Usually) Don't Have to Worry About Multiple Comparisons. *J Res Educ Effectiveness* 2012; 5: 189–211.

- Subgroup specific treatment effects can be falsely significant (statistically, clinically).

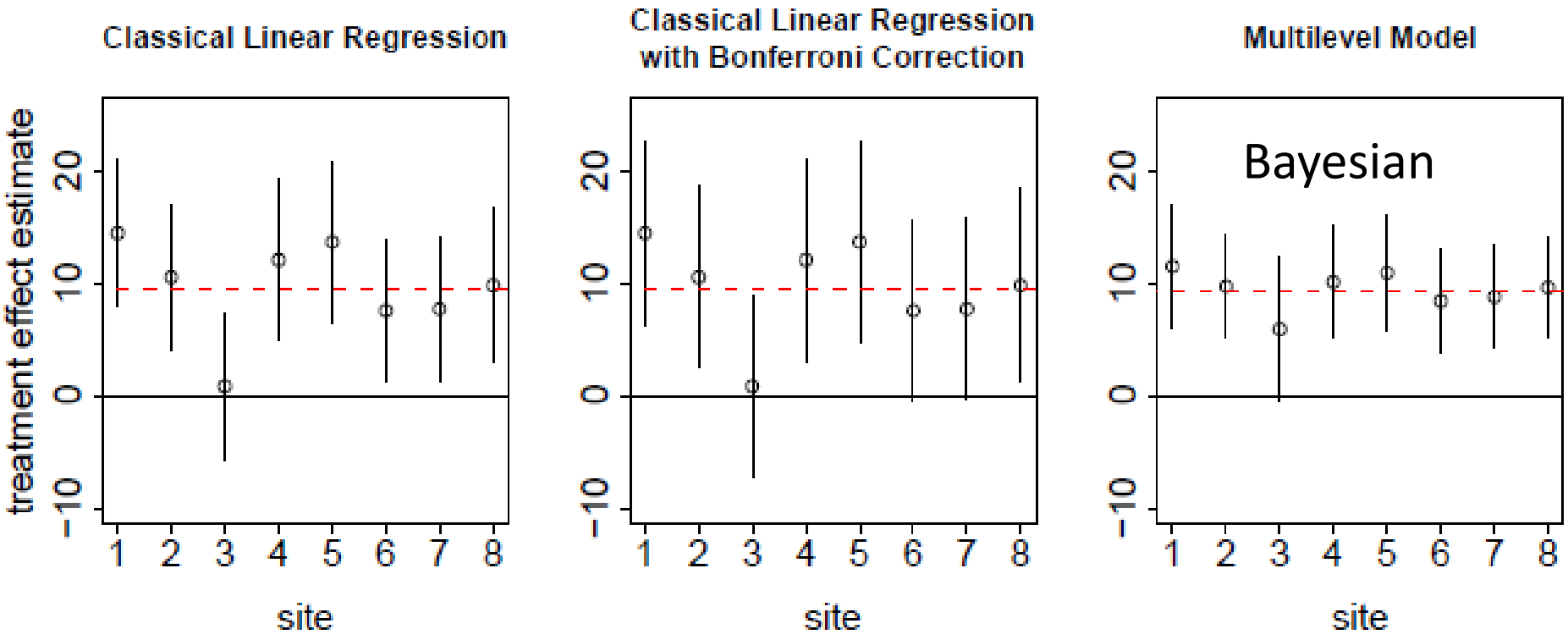


Figure 1: Treatment effect point estimates and 95% intervals across the eight Infant Health and Development Program sites. The left panel display classical estimates from a linear regression. The middle panel displays the same point estimates as in the left panel but with confidence intervals adjusted to account for a Bonferroni correction. The right panel displays posterior means and 95% intervals for each of the eight site-specific treatment effects from a fitted multilevel model.

Gelman A, Hill J, Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons, *J Res On Educational Effectiveness*. 2012; 5: 189-211.

Confirm Efficacy in Targeted Subgroup

- All-comers trial: Test for treatment effect in targeted subgroup, complement, and overall.
- Enrichment trial: Enroll targeted subgroup only.
- Heterogeneity of treatments effects *is* anticipated to have a particular pattern *a priori* (treatment effects *not* exchangeable).
- EX. A subgroup of cancer patients exhibiting the molecular target of a drug are expected to be more likely to respond to the drug than patients without the molecular target.

Companion Diagnostic (CDx) Test

H_O: treatment effect, overall

H_S: drug efficacy, test positive subset *S*

$H_{S'}$: drug efficacy, test negative subset *S'*

- Clinical Objectives: Drug claim of efficacy for
  - test positive subset, or
  - overall

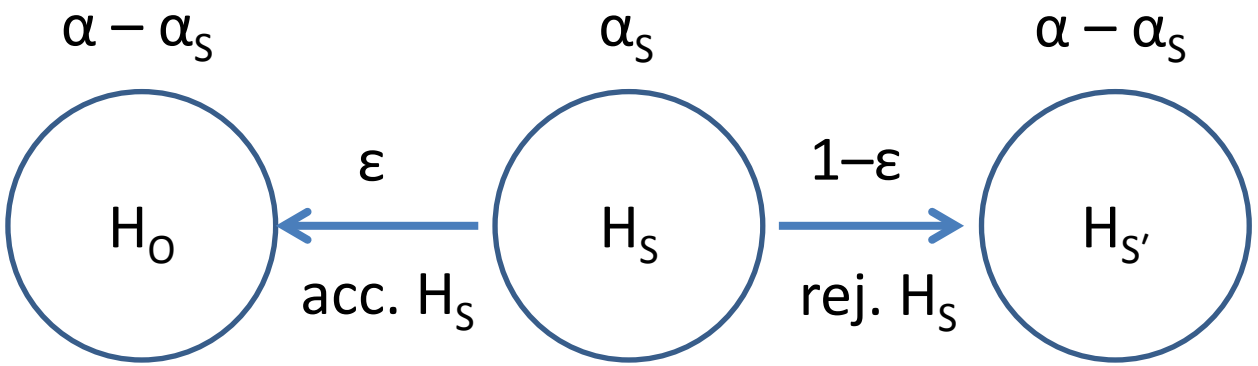
Freidlin B, Korn E, Gray R. Marker Sequential Test (MaST) design, *Clin Trials* 2014; 11: 19–27.

Millen, B.A., Dmitrienko, A., Song, G. (2014). Bayesian assessment of influence and interaction conditions in multipopulation tailoring clinical trials. *J Biopharm Stat*. 2014;24(1):94-109.

Rothmann M, Zhang, Lu, Fleming T. Testing in a Pre-specified Subgroup and the Intent-to-treat Population, *Drug Inf J.* 2012 Mar 1;46(2):175-179.

Marker Sequential Test (MaST)

$\alpha_S < \alpha$ = level of procedure



If H_S rejected at α_S,
then subgroup claim met,
if $H_{S'}$ rejected at α, overall claim met.
else if H_O rejected at $\alpha - \alpha_S$, overall claim met.

Freidlin B, Korn E, Gray R. Marker Sequential Test (MaST) design, *Clin Trials* 2014; 11: 19–27.

Bretz F, Maurer W, Brannath W, Posch M. *Stat Med.* 2009 Feb 15;28(4):586-604.

Adaptive Designs

- Adaptive sample size
- Response Adaptive (enrollment, randomization)
- Adaptive analysis (biomarker signature, threshold)
- Biomarker analysis designs
  - Blur usual recommendation to perform development and validation phases on separate data sets.
  - Conceivable for pivotal validation of a CDx if analytical accuracy and reproducibility is exceptional and highly robust .
  - Cross-validated performance may not generalize out-of-sample because classifier may have been fit to patient characteristics, specimen (or imaging) characteristics, and measurement errors (e.g., batch effects) that are peculiar to the training data set.

Baker SG, Kramer, BS, Sargent DJ, Bonetti M. Biomarkers, subgroup evaluation, and clinical trial design. *Discov Med*. 2012, 13 (70), 187–192.

Subgroup Misclassification

Subgroup Misclassification

Response $R = 0,1$ (to treatment)

Biomarker $B = 0,1$ (reference result)

Measurement $B^* = 0,1$ (test result)

- Assume misclassification error of B by B^* is *non-differential* to outcome, that is

$$B^*|B, R = B^*|B$$

i.e., $$R|B^*, B = R|B$$

Subgroup Misclassification

Consider

$$D = \Pr(R = 1|B = 1) - \Pr(R = 1|B = 0)$$

$$D^* = \Pr(R = 1|B^* = 1) - \Pr(R = 1|B^* = 0)$$

Then $$D^* \overset{NDME}{=} D \times (PPV + NPV - 1)$$

where $$PPV = \Pr(B^* = 1|B = 1)$$

$$NPV = \Pr(B^* = 0|B = 0)$$

Kuha, Skinner, Palmgren, 2005, "Misclassification Error" in *Encyc Biostat*

Notation

- $\theta_{ab} = E_{ab}(Y) =$ expectation of Y for treatment $A = a$, biomarker status $B = b$ $(A, B = 0,1)$.
  - objective response (0,1), event-free survival time
- $\theta^*_{at} = E_{at}(Y) =$ expectation of Y for treatment $A = a$, biomarker test result $T = t$ $(A, T = 0,1)$

$$\theta^*_{at} \overset{NDME}{=} \Sigma_{b=0}^{1} \theta_{ab} \Pr(B = b | T = t)$$

$$= \theta_{a0}(1 - p_t) + \theta_{a1} p_t,$$

$$p_t = \Pr(B = 1 | T = t)$$

Notation

- $\delta_b = \theta_{1b} - \theta_{0b} =$ treatment effect (mean difference) between treatment arms $a = 0,1$ given biomarker status $B = b\ (= 0,1)$
- $\Delta_{A.B} = \delta_1 - \delta_0 =$ predictive biomarker *capacity*.

- $\delta_t^* = \theta_{1t}^* - \theta_{0t}^* =$ treatment effect (mean difference) between treatment arms $a = 0,1$ given test result $T = t\ (= 0,1)$

Biomarker Stratified Design

$$\delta_t^* = \theta_{1t}^* - \theta_{0t}^*$$

$$\delta_b = \theta_{1b} - \theta_{0b}$$

Estimand $\Delta_{A.T}^* = \delta_1^* - \delta_0^*$

$$= (p_1 - p_0)(\delta_1 - \delta_0)$$

$$= (PPV + NPV - 1)\Delta_{A.B},$$

= treatment arm by biomarker interaction

$\Delta_{A.B}$ attenuated by the factor $PPV + NPV - 1$.

NDME Attenuation Result

$$\delta_{0.} - \delta_{1.} = (\delta_{.0} - \delta_{.1})(\pi_1 - \pi_0)$$

- That is,

$$E(\log h \mid T = 0) - E(\log h \mid T = 1)$$
$$= \left[E(\log h \mid S = 0) - E(\log h \mid S = 1)\right] \times \left(NPV + PPV - 1\right)$$

- The difference in log hazard ratio between groups defined by test *T* is attenuated relative to the corresponding difference for test *S*[†].

[†]Provided *PPV ≥ 1 – NPV* , i.e., *T* not negatively informative for *S*, so that *0 ≤ PPV + NPV – 1 ≤ 1. (*Pennello, *Clin Trials* 2013 Oct;10(5): 666-76 *)*

Cox Proportional Hazards Model

- For treatments *X = 0, 1*, Cox hazard for *S, X* is

$$\lambda(y \mid S, X) = \lambda_0(y) e^{\beta_S X}$$

$$\beta_s = \log \text{ hazard ratio for } S = s$$

$$\lambda(y \mid T, X) = E\left\{\lambda_0(y \mid T, S, X) \mid T, X, Y \geq y\right\}$$

NDME

$$= E\left\{\lambda(y \mid S, X) \mid T, X, Y \geq y\right\}$$

Rare Event

$$\cong \lambda_0(y) E\left\{e^{\beta_S X} \mid T, X\right\}$$

$$= \lambda_0(y)\left[e^{\beta_0 X} + \pi_T\left(e^{\beta_1 X} - e^{\beta_0 X}\right)\right]$$

Cox Proportional Hazards Model

$$\lambda(y \mid T, X) \cong \lambda_0(y)\left[e^{\beta_0 X} + \pi_T\left(e^{\beta_1 X} - e^{\beta_0 X}\right)\right]$$

$$\lambda(y \mid T=1, X=1) - \lambda(y \mid T=0, X=1)$$

$$= \lambda_0(y)\left(e^{\beta_1} - e^{\beta_0}\right)\left(\pi_1 - \pi_0\right)$$

$$= \left[\lambda(y \mid S=1, X=1) - \lambda(y \mid S=0, X=1)\right]\left(\pi_1 - \pi_0\right)$$

- Same approximate attenuation holds for log hazard ratio difference.
- Approximation is OK for rare enough outcome. More investigation is needed.

Pepe MS, Self SG, Prentice, RL. *Statist Med* 1989;, 8, 1167-1178.
Prentice RL. *Biometrika* 1982; 69, 331-342.
Lin DY, Psaty BM, Kronmal RA. *Biometrics* 1998; 54(3): 948-963.

References

Pennello GA. Analytical and clinical evaluation of biomarkers assays: when are biomarkers ready for prime time? Clin Trials. **2013**, 10 (*5*), 666–676.

Pennello GA, Ye J. Companion Diagnostics. In *Encyclopedia of Biopharmaceutical Statistics 4th Ed.,* Ed. Shein-Chung Chow. **2017**, CRC Press, to appear.

Sharma A, Zhang G, Aslam S, Yu K, Chee M, Palma JF. Novel Approach for Clinical Validation of the cobas KRAS Mutation Test in Advanced Colorectal Cancer. Mol. Diagn. Ther. **2016**, 20 (*3*), 231–240.

Missing Data in Biomarker Evaluation Studies

Diagnostic Test Evaluation

- ***Y* = Reference Standard Result**
  for a present or future state of health
  - presence or absence of disease
  - time to onset of disease, progression, death, etc.
  - true level of measurand in a sample
- ***X* = Test Result**
  - Quantitative (concentration of analyte)
  - Continuous (e.g., ratio)
  - Semi-Quantitative (ordinal)
  - Qualitative (binary)
- ***Z* = Covariates** (including comparator tests)

Missing Data in Diagnostic Studies

- Missing Reference *Y.* (*verification bias*)
  - State of health was not verified by the reference.
- Missing Test Result *X*. (*unsatisfactory test bias*)
  - Sample is unavailable or unevaluable.
  - Test result was invalid.
  - Lack of consent to use sample.
- Imperfect Reference *Y*. (*misclassification bias*)
  - *Y* is subject to error
- Imprecision in *X* or *Y*. (*measurement error bias*)
  - Result varies over repeated measurement

Test Result MAR

- Test result *X* missingness indicator *M* = 0,1.
- If *X* is MAR, then

$$M \mid X,Y,Z = M \mid Y,Z$$

$$\text{i.e., } X \mid M,Y,Z = X \mid Y,Z \quad \text{(MAR)}$$

Se, Sp unbiased in complete data.

Get *PPV, NPV* by Bayes Theorem or IPW (if *Y* = 0,1 sampling fractions are known).

MAR underlies validity of case-control studies.

Reference Result MAR

- Reference *Y* missingness indicator *V* = 0,1.
- If *Y* is MAR, then

$$V \mid X, Y, Z = V \mid X, Z$$

$$\text{i.e., } Y \mid V, X, Z = Y \mid X, Z \quad \text{(MAR)}$$

PPV, NPV are unbiased among complete data.

Get *Se, Sp* by Bayes Theorem or IPW (provided sampling fractions for *X* = 0,1 are known.

Intention to Treat (ITT)

- Includes every subject who is randomized according to their treatment assignment, regardless of non-compliance with treatment, missing outcomes, protocol deviations, withdrawal, or anything else that happened after randomization.
- The *ITT* analysis avoids overoptimistic estimates of treatment efficacy due to the exclusion of subjects on the basis of post-randomization variables.

Intention to Diagnose (ITD)

- Include every subject, regardless of whether the subject is
  - missing the test result,
  - the clinical reference result, or
  - comparator test results.
- When appropriate (meaningful for analysis), impute missing data when evaluating the test for diagnostic performance.

Bu Y, Zhou XH. *J Biopharm Stat.* 2016, 26 (6), 1118–1124
Denne JS, Pennello G, Zhao L, Chang SC, Althouse S. *Stat Biopharm Res.* 2014;6 (1), 78–88.
Li, M. *J. Biopharm Stat.* 2015, 25 (3), 397–407.
Lunceford, J.K. *Pharm Stat.* 2015, 14 (3), 233–241.

Intention to Diagnose (ITD)

- **Missing Test Results**
  - If subject is retested, include retest result (if retesting is consistent with intended use).
  - Report number and proportion of subjects without the test result by the reason it is missing (lack of consent, sample un-available, sample unevaluable, test result invalid)
  - If proportion of subjects with an invalid test result is large, the test may have a design problem.
  - Impute missing test results, if appropriate.

Begg, Greenes, Iglewicz. The influence of uninterpretability on the assessment of diagnostic tests. *J Chron Dis* 1986; 39(8): 575-584.

Missing *Y*

Bayesian MNAR Model

Data

	Test Result	
Reference Diagnosis	***T = 0***	***T=1***
D = 0	x_{00}	x_{01}
NA	$w_{\bullet 0}$	$w_{\bullet 1}$
D = 1	x_{10}	x_{11}

Bayesian MNAR Model

$$\tau = \Pr(T = 1)$$

$$p_t = \Pr(D = 1 \mid T = t),\ t = 0,1$$

$$\rho_{dt} = \Pr(V = 1 \mid D = d, T = t),\ d, t = 0,1$$

- Prior $\tau, p_t, \rho_{dt} \sim Beta(\alpha, \beta)$
- 7 parameters, 5 dofs in data

Pennello (2011). Bayesian Analysis of Diagnostic Test Accuracy When Disease State is Unverified for Some Subjects, *J Biopharm Stat*, 21: 954-970.

Missing Disease Verification Models

- Full MNAR model ρ_{dt}
  - Probability of verification depends on missing disease state, test result (*parameters under-identified*).
- Reduced MNAR model $\rho_{dt} \equiv \rho_{d}$
  - Probability of verification depends on missing disease state, not test result (*identified*).
- MAR model $\rho_{dt} \equiv \rho_{t}$
  - Probability of verification depends on test result, not missing disease state (*identified*)

MNAR Bayesian Models

- In some MNAR models, parameters are well-defined, but not fully identified.
- In some instances Bayesian MNAR models can still obtain useful inferences.
  - Neath, A., Samaniego, F. On the efficacy of Bayesian inference for non-identifiable models. *American Statistician* 1997; 51: 225-232.
  - Gustafson, P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, Chapman & Hall / CRC, 2003.
  - Gustafson, P. *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data.* Chapman & Hall / CRC, 2015.

Data Distributions

$$x_{\bullet 1} + w_{\bullet 1} \sim Bin(x_{\bullet\bullet} + w_{\bullet\bullet}, \tau),$$

$$(x_{0t}, x_{1t}, w_{\bullet t}) \,|\, x_{\bullet t} + w_{\bullet t} \sim Mult_3(x_{\bullet t} + w_{\bullet t},$$

$$(\rho_{0t}(1-p_t),\ \rho_{1t}p_t,\ (1-\rho_{0t})(1-p_t)+(1-\rho_{1t})p_t)\ )$$

Gibbs Sampler with Data Augmentation

$$\tau^{(i)} \mid \underline{x}, \underline{w}^{(i)} \sim Beta(\alpha + x_{\bullet 1} + w_{\bullet 1}^{(i)}, \beta + x_{\bullet 0} + w_{\bullet 0}^{(i)})$$

$$p_t^{(i)} \mid \underline{x}, \underline{w}^{(i)} \sim Beta(\alpha + x_{1t} + w_{1t}^{(i)}, \beta + x_{0t} + w_{0t}^{(i)})$$

$$\rho_{dt}^{(i)} \mid \underline{x}, \underline{w}^{(i)} \sim Beta(\alpha + x_{dt}, \beta + w_{dt}^{(i)})$$

$$w_{1t}^{(i+1)} \mid w_{\bullet t}, \underline{p}^{(i)}, \underline{\rho}^{(i)} \sim Bin\left(w_{\bullet t}, \frac{(1-\rho_{1t}^{(i)})\, p_t^{(i)}}{(1-\rho_{1t}^{(i)})\, p_t^{(i)} + (1-\rho_{0t}^{(i)})(1-p_t^{(i)})} \right),$$

$$w_{0t}^{(i+1)} = w_{\bullet t} - w_{1t}^{(i+1)}, \; t = 0,1$$

Hepatic Scintigraphy

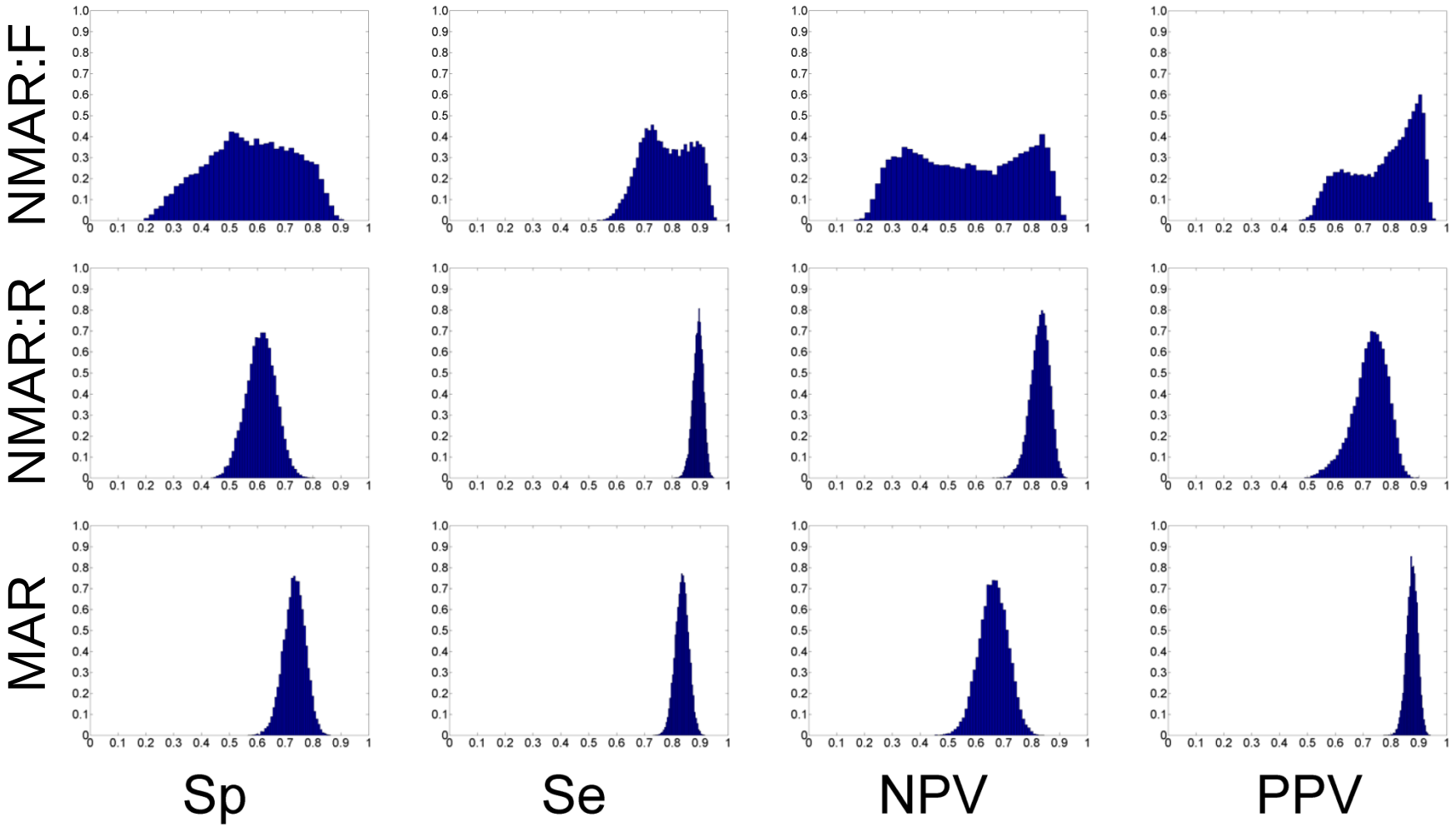
Scintigraphy	Liver Disease		
	D=0	D=1	NA
T = 0	54	27	140
T = 1	32	231	166
Total	86	258	306

$$\widehat{Sp}_{CC} = \frac{54}{86} = 62.8\% \qquad \widehat{NPV}_{CC} = \frac{54}{81} = 66.7\%$$

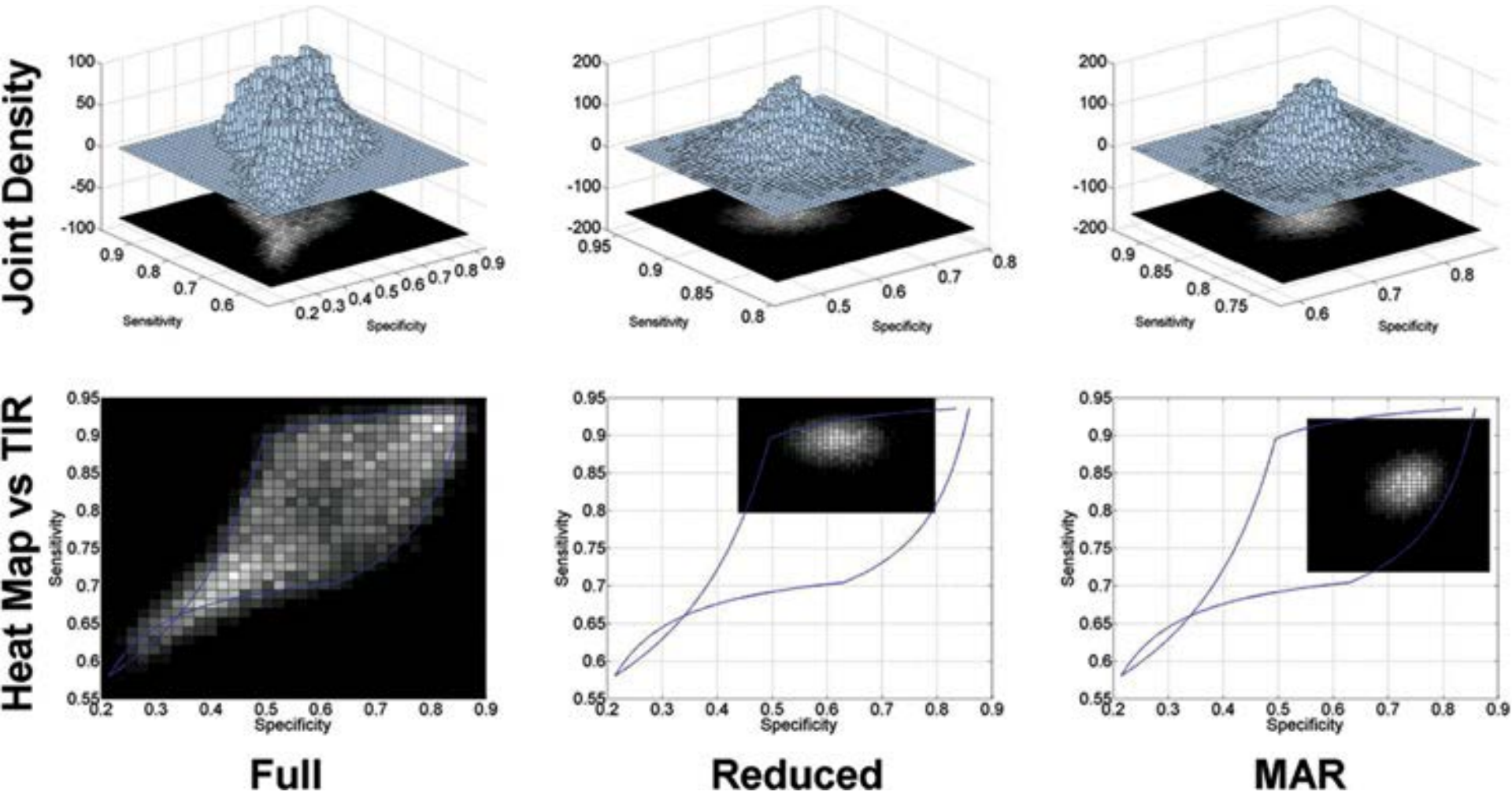
$$\widehat{Se}_{CC} = \frac{231}{258} = 89.5\% \qquad \widehat{PPV}_{CC} = \frac{231}{263} = 87.8\%$$

Drum D.E., Christacopoulos, J.S. (1969). Hepatic scintigraphy in clinical decision making, *J. Nucl. Med.*, 13, 908-915.

Posterior Distributions



Se/Sp Posterior Distribution



Solid Line is the test ignorance region.

Test Ignorance Region

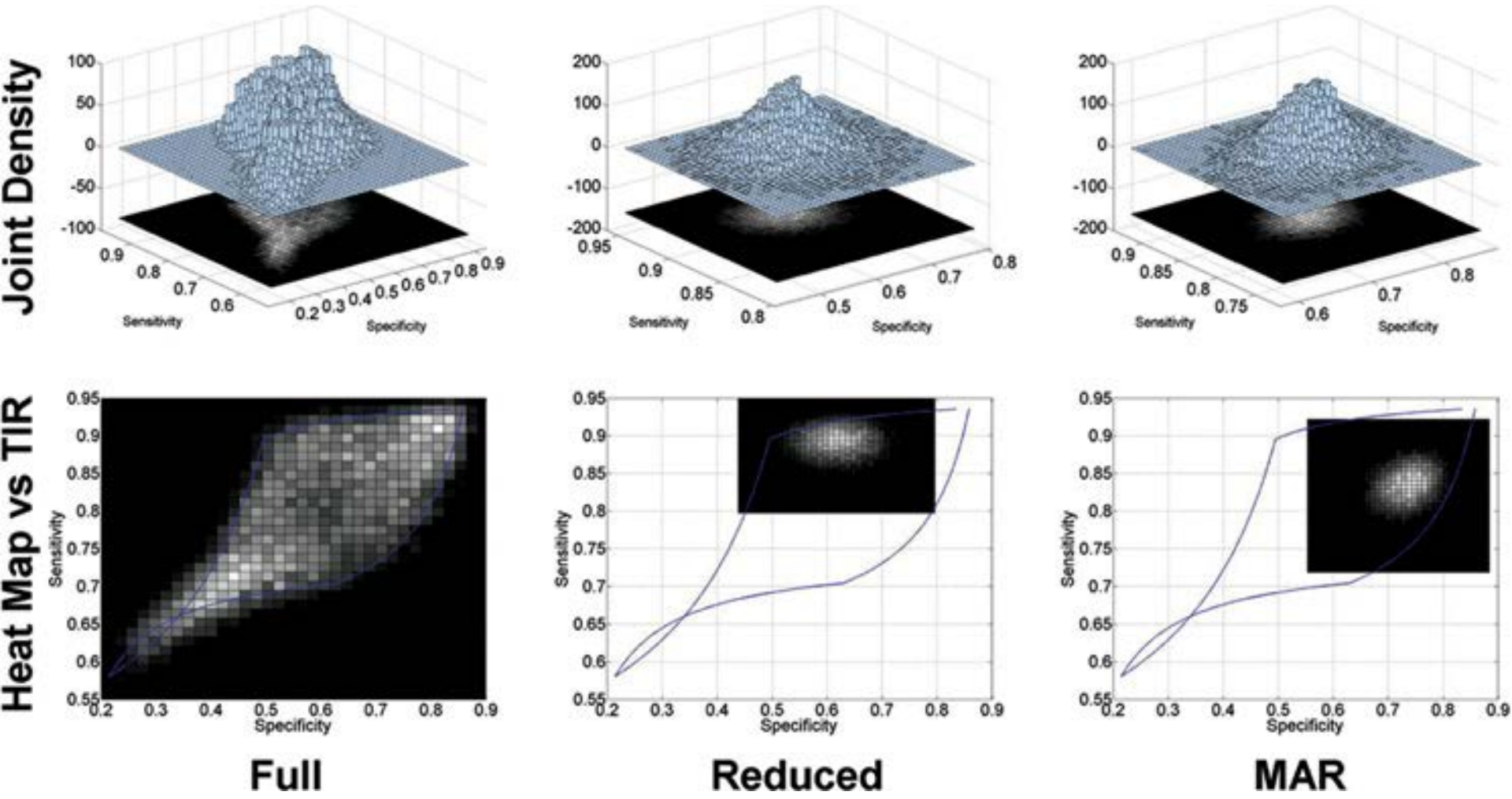
TIR delineates the possible estimates of *Se, Sp* over all possible disease states of unverified subjects:

$$g_1(\widehat{Se}) = \begin{cases} 1 - \dfrac{x_{01} + w_{\cdot 1}}{x_{\cdot\cdot} + w_{\cdot\cdot} - x_{11} / \widehat{Se}}, & \dfrac{x_{11}}{x_{1\cdot} + w_{\cdot 0}} \le \widehat{Se} \le \dfrac{x_{11}}{x_{1\cdot}} \\ \dfrac{x_{00} + w_{\cdot 0}}{x_{\cdot\cdot} + w_{\cdot\cdot} - x_{10} / (1 - \widehat{Se})}, & \dfrac{x_{11}}{x_{1\cdot}} \le \widehat{Se} \le \dfrac{x_{11} + w_{\cdot 1}}{x_{1\cdot} + w_{\cdot 1}} \end{cases}$$

$$g_2(\widehat{Se}) = \begin{cases} \dfrac{x_{00}}{x_{\cdot\cdot} + w_{\cdot\cdot} - (x_{10} + w_{\cdot 0}) / (1 - \widehat{Se})}, & \dfrac{x_{11}}{x_{1\cdot} + w_{\cdot 0}} \le \widehat{Se} \le \dfrac{x_{11} + w_{\cdot 1}}{x_{1\cdot} + w_{\cdot\cdot}} \\ 1 - \dfrac{x_{01}}{x_{\cdot\cdot} + w_{\cdot\cdot} - (x_{11} + w_{\cdot 1}) / \widehat{Se}}, & \dfrac{x_{11} + w_{\cdot 1}}{x_{1\cdot} + w_{\cdot\cdot}} \le \widehat{Se} \le \dfrac{x_{11} + w_{\cdot 1}}{x_{1\cdot} + w_{\cdot 1}} \end{cases}$$

Kosinski, A.S. and Barnhart, H.X. (2003). *Statist. Med.*, 22:2711-2721.

Se/Sp Posterior Distribution



Solid Line is the test ignorance region.

Hepatic Scintigraphy, 1/7th of Missing Data

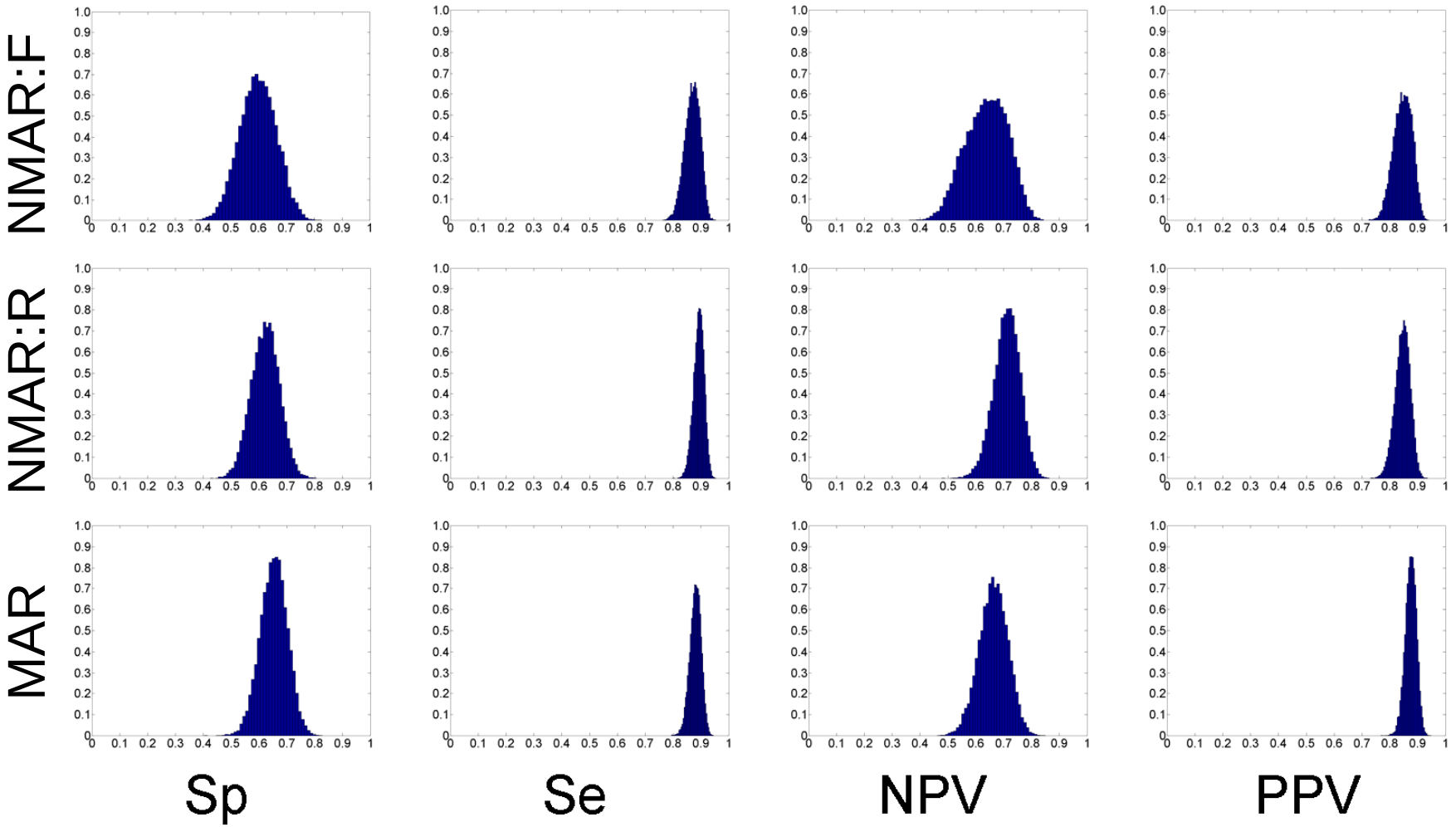
Scintigraphy	Liver Disease		
	D=0	D=1	NA
T = 0	54	27	20
T = 1	32	231	24
Total	86	258	44

$$\widehat{Sp}_{CC} = \frac{54}{86} = 62.8\% \qquad \widehat{NPV}_{CC} = \frac{54}{81} = 66.7\%$$

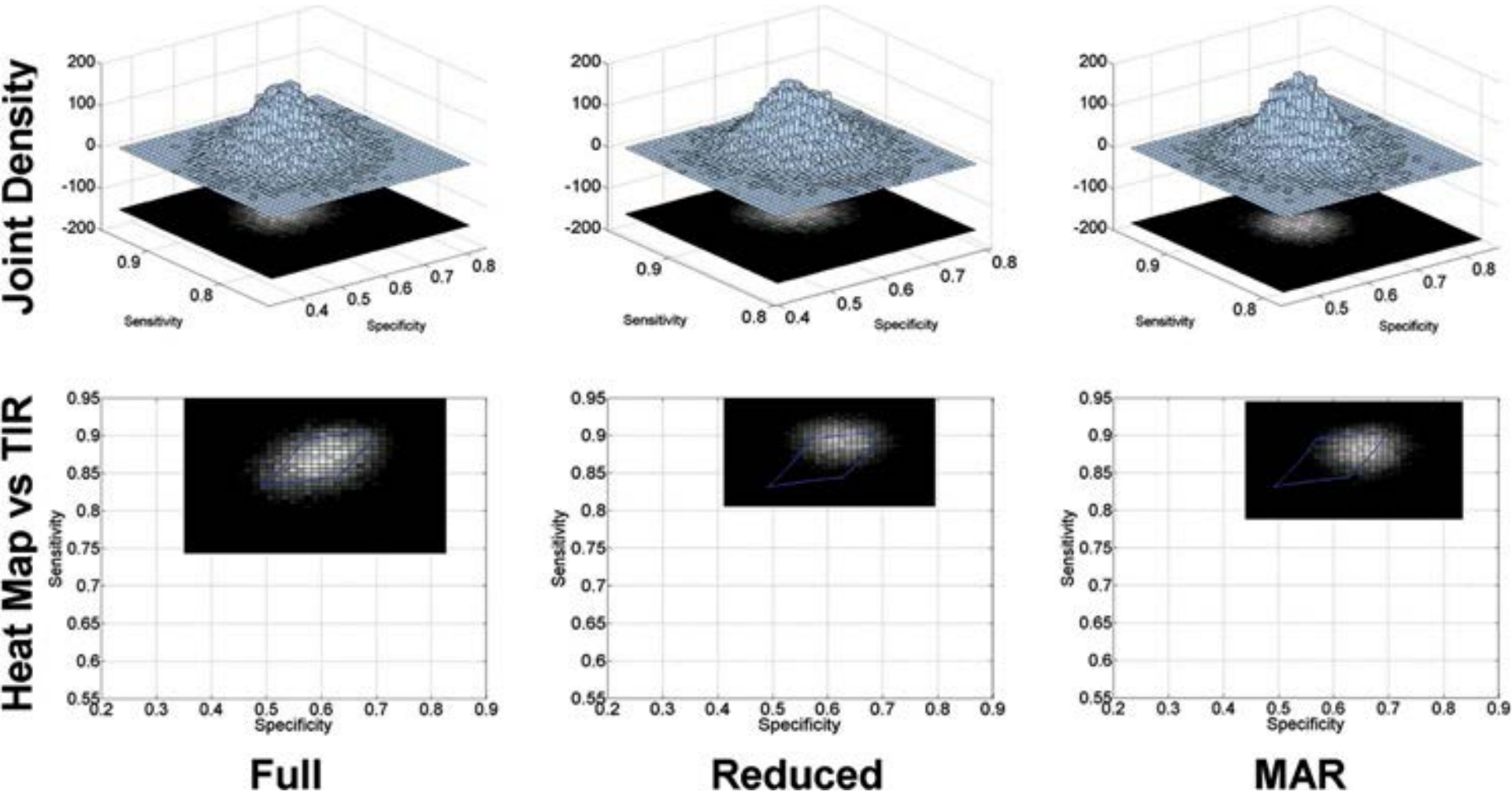
$$\widehat{Se}_{CC} = \frac{231}{258} = 89.5\% \qquad \widehat{PPV}_{CC} = \frac{231}{263} = 87.8\%$$

Drum D.E., Christacopoulos, J.S. (1969). Hepatic scintigraphy in clinical decision making, *J. Nucl. Med.*, 13, 908-915.

Posterior Distributions



Se/Sp Posterior Distribution, 1/7th Miss



Solid Line is the test ignorance region.

Concluding Remarks

- A biomarker's capacity to predict treatment response or treatment effect is only as good as the test used to measure it.
- In biomarker subgroup analysis, *a priori* biological understanding may determine
  - clinical objectives of validation study,
  - statistical analysis plan.
- Biomarker signature (classifier) development and validation has been a challenge
  - No FDA approvals to date.
  - Analytical validation can be very complex.

Concluding Remarks

- Subgroup multiplicity can be out-of-control for some diagnostic devices.
  - Smoothing, not hypothesis testing may be best option for data interpretation, labeling.

B-type Natriuretic Peptide (BNP)

- Demographics
  - age, sex, race, ethnicity, height, weight, BMI
- Physical Exam, Signs and Symptoms
  - vital signs, acute dyspnea (due to cardiac vs non-cardiac cause), chest discomfort, pulmonary rales, peripheral edema, nocturnal cough, etc.
- Medical History, Risk Factors and Comorbidities
  - diabetes, kidney disease / renal dysfunction (eGFR<60), hypertension, history of MI, cardiac ischemia, prior diagnosis of HF, atrial fibrillation, myocarditis, tachycardia, LV hypertrophy, pulmonary embolism, COPD, pneumonia, asthma, obstructive sleep apnea, sepsis, anemia, cirrhosis of liver, obesity [BMI ≥ 30], severe obesity [BMI ≥ 37.5], etc.
- Medication list
  - beta-blockers, diuretics, ACE inhibitors, angiotensin II receptor blockers [ARBs], aldosterone antagonists, statins, etc.
    - Prior to admission to ED
    - ED medications (includes date and time of administration)

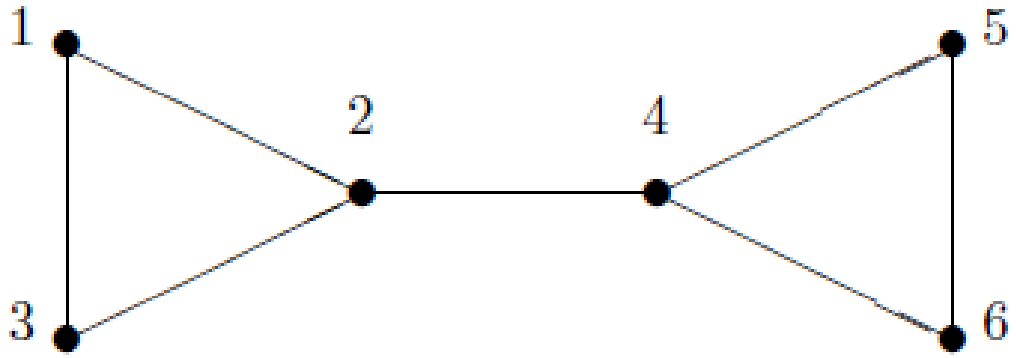
B-type Natriuretic Peptide (BNP)

- Diagnostic Procedure Results (if available per standard of care in the assessment of the dyspnea patient)
  - ECG, Echocardiogram, Radionuclide angiography , Chest radiography
- Ejection fraction (EF)
- Laboratory Results
  - electrolytes, BUN, creatinine, eGFR, thyroid function tests, liver function tests, HbA1c or fasting glucose, CBC, urinalysis, troponin
- Standard-of-care BNP or NT-proBNP value at admission
- Diagnosis: heart failure vs no heart failure
- Classification of HF if present and known:
  - HFrEF (heart failure with reduced ejection fraction) (≤ 40% EF)
  - HFpEF (heart failure with preserved ejection fraction) (> 40% EF)
- Severity of HF if present: NYHA Functional Class (I, II, III, IV)

Graph Theory

Corollary 5.3.2. Let *A, B*, and *C* denote sets that partition of factors in a graphical model such that every chain between a factor in *A* and a factor in *B* involves at least one factor in *C*; then the relationships among the factors in *A* and *C* can be examined in the marginal table obtained by summing over the factors in *B*.

Example 5.3.3. Model [123][24][456] is graphed below:



- Accurate conclusions can be drawn from the marginal tables $n_{123\cdot\cdot\cdot}$, $n_{1234\cdot\cdot}$, $n_{\cdots 456}$, and $n_{\cdot 2\cdot 456}$.

Concluding Remarks

- **Benefit-Risk Evaluation of Diagnostic Devices**
  - **Validation**
    - Baker, S. G. (2009). Putting risk into perspective: Relative utility curves. *JNCI*, 101:1538–1542
    - Baker, S. G., Van Calster, B., Steyerberg, E. W. (2012). Evaluating a new marker for risk prediction using the test tradeoff: An update. *Int J Biostat* 8(1):Article 5, 101:1538–1542.
    - Evans SR, Pennello G, Pantoja-Galicia N, et al, for the Antibacterial Resistance Leadership Group. Benefit-risk Evaluation for Diagnostics: A Framework (BED-FRAME). *Clin Infect Dis* 2016; 63(6):812-7.
    - Pennello G, Pantoja-Galicia N, Evans S. Comparing diagnostic tests on benefit-risk. *J Biopharm Stat* 2016; 26(6): 1083–1097.
    - Pepe MS, Janes H, Li Cl, Bossuyt PM, Feng Z, Hilden J. Early-phase studies of biomarkers: What target sensitivity and specificity values might confer clinical utlity? *Clin Chem* 2016;62(5):737-742.
    - Vickers, A. J., Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Med Decis Making* 26:565–574.
  - **Signature Development**
    - Gunter L, Zhu J, Murphy SA. Variable selection for qualitative interactions. *Statistical Methodology* 8 (2011) 42–55.
    - Schnell PM, Tang Q, Offen WW, Carlin BP. A Bayesian Credible Subgroups Approach to Identifying Patient Subgroups with Positive Treatment Effects. *Biometrics* 2016; 72, 1026–1036.

Take Home Messages

- **Analytical Performance**
  - Characterizing the analytical performance of a diagnostic device (its reliability in measuring the analyte) is a prerequisite to applying it to specimens in a clinical performance study.
- **Clinical Performance**
  - Clinical significance should be demonstrated.
  - Intended Use determines clinical data requirements.
  - Claims in labeling depend on studies conducted.

Take Home Messages

- **Biomarker Signature Discovery / Development**
  - Classifier development, validation been challenging
  - CDRH has a pre-submission program to meet with & provide informal feedback to device sponsors.
- **Independent validation**
  - Validate a biomarker assay on specimens independent of those used to develop the assay.
- **"Intent to Diagnose" Analysis**
  - An analysis of device performance should include all study subjects, even if the device result or the reference (gold) standard result is missing (unavailable, unevaluable, invalid, indeterminate).