Statistical Considerations for Studies Validating Imaging Biomarkers

Nancy Obuchowski, PhD

Cleveland Clinic

Disclaimers

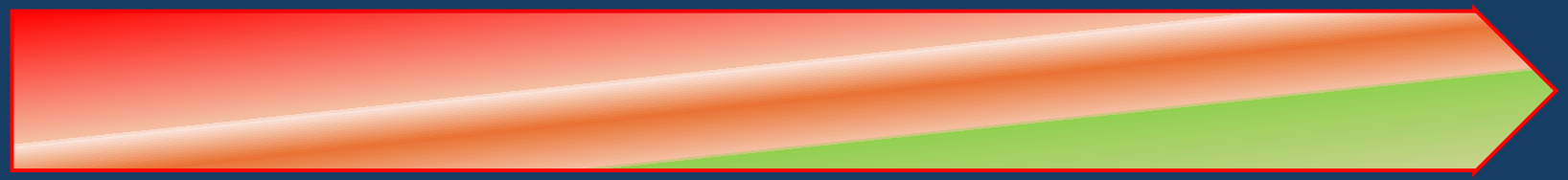
- I’m not a regulator
- FDA is a moving target (as it should be)

Imaging Biomarkers

relative to their use over the course of disease

Adapted from Ou et al, J Thorac Oncol 2022

Biomarker validation: a continuous variable



Exploratory *Probable valid* *Fit-for-purpose*

Outline: Statistical consideration for…

- Discovery
- Analytical Validation
- Clinical Validation

of quantitative imaging biomarkers (QIBs)

Outline: Statistical consideration for…

- Discovery - as introduction
- Analytical Validation – effect on clinical validation
- Clinical Validation – focus on predictive biomarker validation

of quantitative imaging biomarkers (QIBs)

Discovery

- Giant leaps with machine learning, radiomics, AI

Common Sense design for discovery:

(responsibility not to overstate our discoveries)

- Avoid convenience samples – collect from a cohort representing the target population
- *Apriori* power calculation
- *Apriori* SAP - Avoid data-driven analyses

Common Sense design for discovery:

- Avoid convenience samples – collect from a cohort representing the target population
- *Apriori* power calculation
- *Apriori* SAP - Avoid data-driven analyses
- Adjust for multiple comparisons
- Multiparametric biomarkers- combining imaging and non-imaging
  - use each biomarker as a continuous variable (rather than dichotomized) to maximize information for best performance
    - Figure out cutpoints later

Discovery…

	Design	Statistical Test
Diagnostic	Prospective/retrospective of disease presence or phenotype	Main effect test of biomarker on phenotype
Prognostic	Retrospective study of patient outcomes	Main effect test of biomarker on patient outcome
Predictive	**Secondary analysis of RCT**	**Interaction of biomarker x intervention**

Analytical Validation

– vital to clinical validation

Test-Retest Studies:
- Estimate repeatability

Phantom Studies:
-Estimate bias, assess linearity

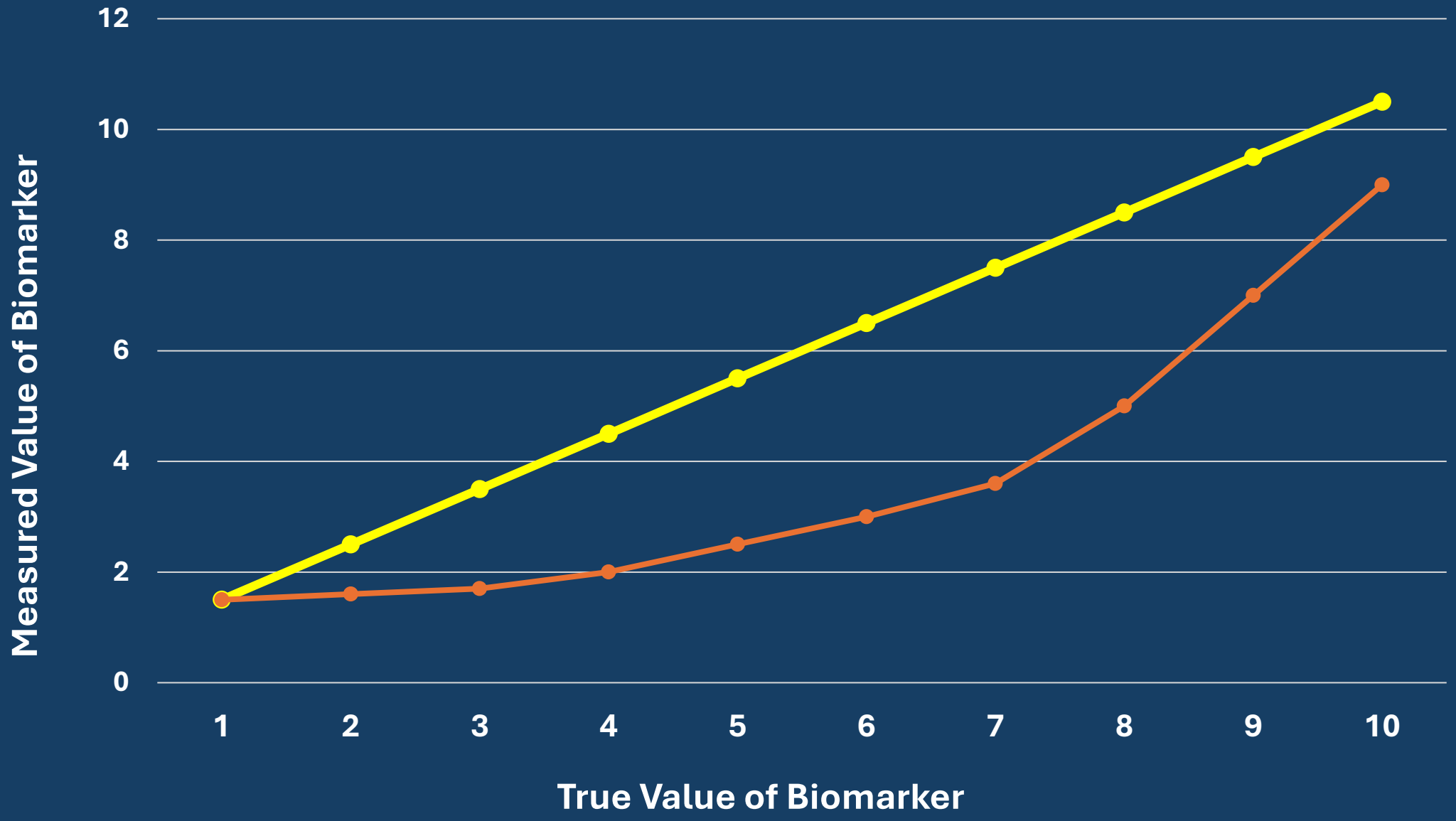
Reproducibility Studies:
- Estimate effect of imaging methods on precision and bias

Analytical / Technical Performance Validation

Metrology Framework

- Technical performance characteristics:
  - Precision (closeness in agreement of replicate measurements)
  - Bias (closeness of measurements to true value)

Linearity Property

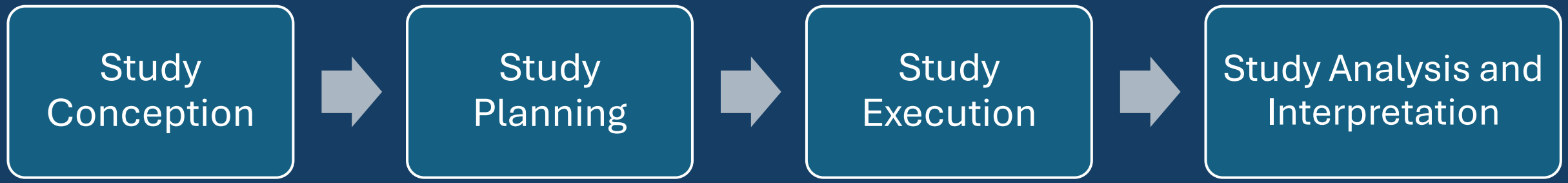


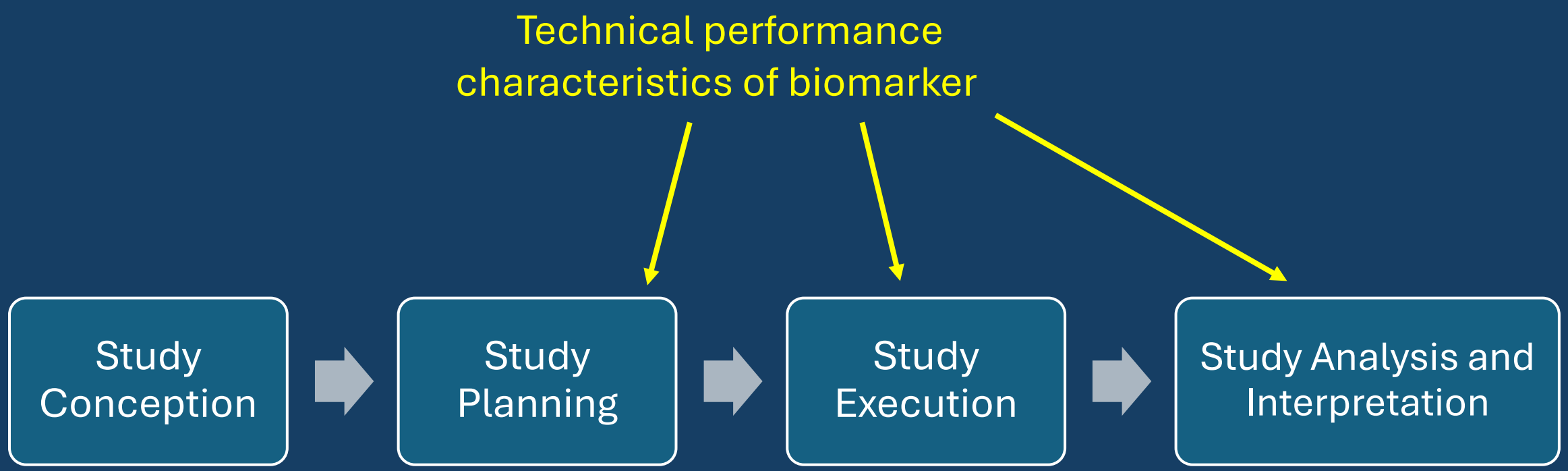
Metrology Framework

- Technical performance characteristics:
  - Precision (closeness in agreement of replicate measurements)
  - Bias (closeness of measurements to true value)
  - Property of Linearity

****Analytical validation also includes identifying cutpoints for specific intended uses and assessing accuracy at these cutpoints.**

Life Cycle of a Clinical Study




Technical performance characteristics of biomarker
Study Conception
Study Planning
Study Execution
Study Analysis and Interpretation

Three examples of how analytical validation intersects with clinical validation …

Example 1: Can ultrasound elastography discriminate subjects with liver cirrhosis (stage F4) from those without cirrhosis?

Is shear wave speed a diagnostic biomarker?

Example 1: Can ultrasound elastography discriminate subjects with liver cirrhosis (stage F4) from those without cirrhosis?

Is shear wave speed a useful diagnostic biomarker?

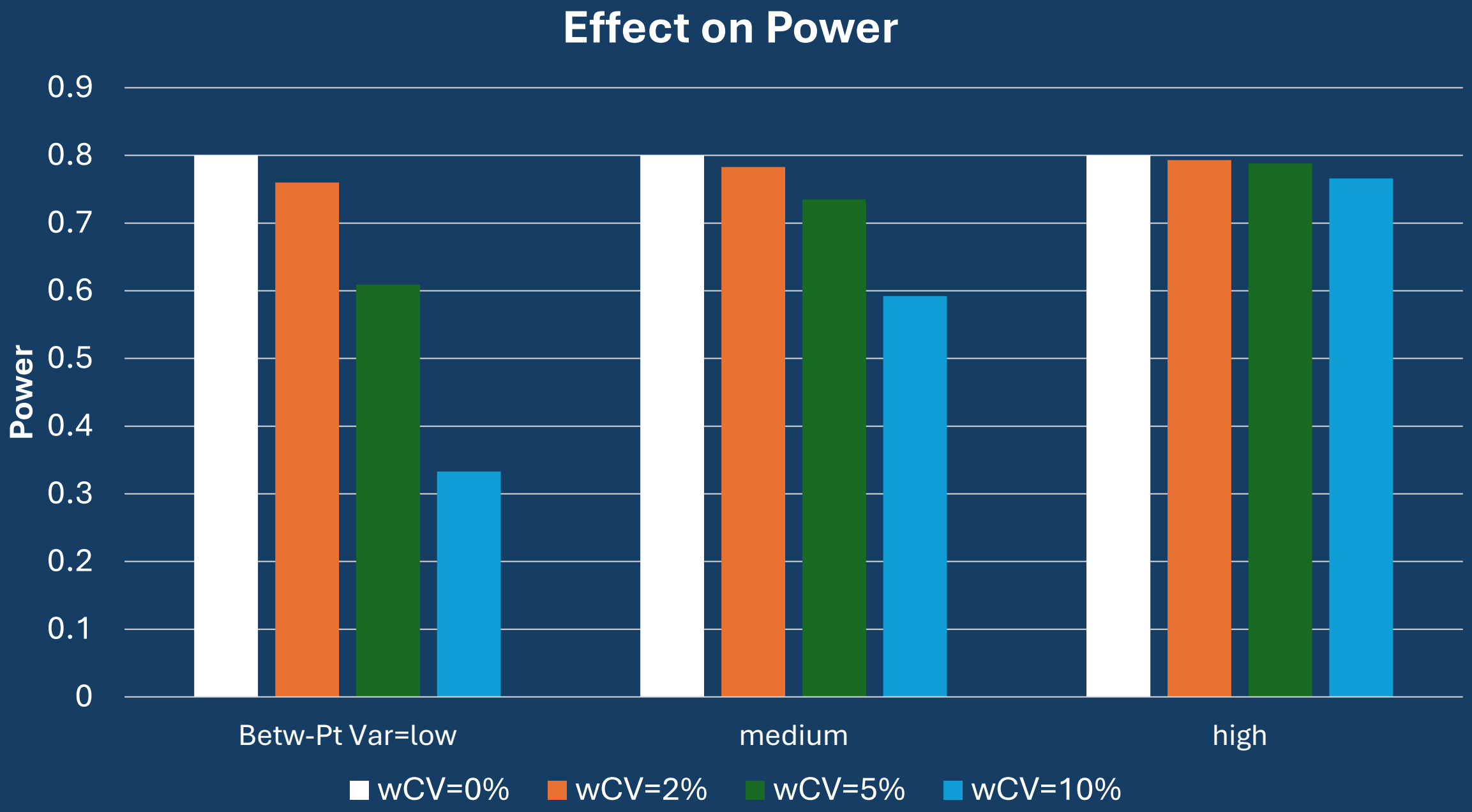
ROC analysis to assess clinical validity

Example 1: Can ultrasound elastography discriminate subjects with liver cirrhosis (stage F4) from those without cirrhosis?

Is shear wave speed a useful diagnostic biomarker?

Test-retest CV of 4-5% with same imaging methods;
CV of 6-7% with different imaging methods

From QIBA US profile


Effect on Power
Power
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Betw-Pt Var=low
medium
high
wCV=0%
wCV=2%
wCV=5%
wCV=10%

Correction to Sample Size:

$$\#\ \text{patients needed} = N_X(\beta_1^2\sigma_b^2 + \sigma_w^2)/\beta_1^2\sigma_b^2$$

sample size if there was
no measurement error

Obuchowski et al, JNCI *2019*

Correction to Sample Size:

$$\#\ \text{patients needed} = N_X(\beta_1^2\sigma_b^2 + \sigma_w^2)/\beta_1^2\sigma_b^2$$

slope

test-retest variability

Obuchowski et al, JNCI *2019*

Example 1

- If ignore measurement error, we would enroll N=52 subjects to estimate the ROC area to within ± 0.05.
- N=56-58 subjects are needed to correct for measurement imprecision.

Example 2: **Clinical Validation of Biomarker for Diagnostic Enrichment of RCT**

Example 2: SPECT specific binding ratio (SBR) in the posterior putamen is used as an enrichment criterion for identifying Parkinson's disease subjects likely to benefit from a new intervention.

Example 2: SPECT specific binding ratio (SBR)

SBR≤1.2 → eligible for study (likely to benefit from new trt)

SBR >1.2 → excluded from study (unlikely to benefit)

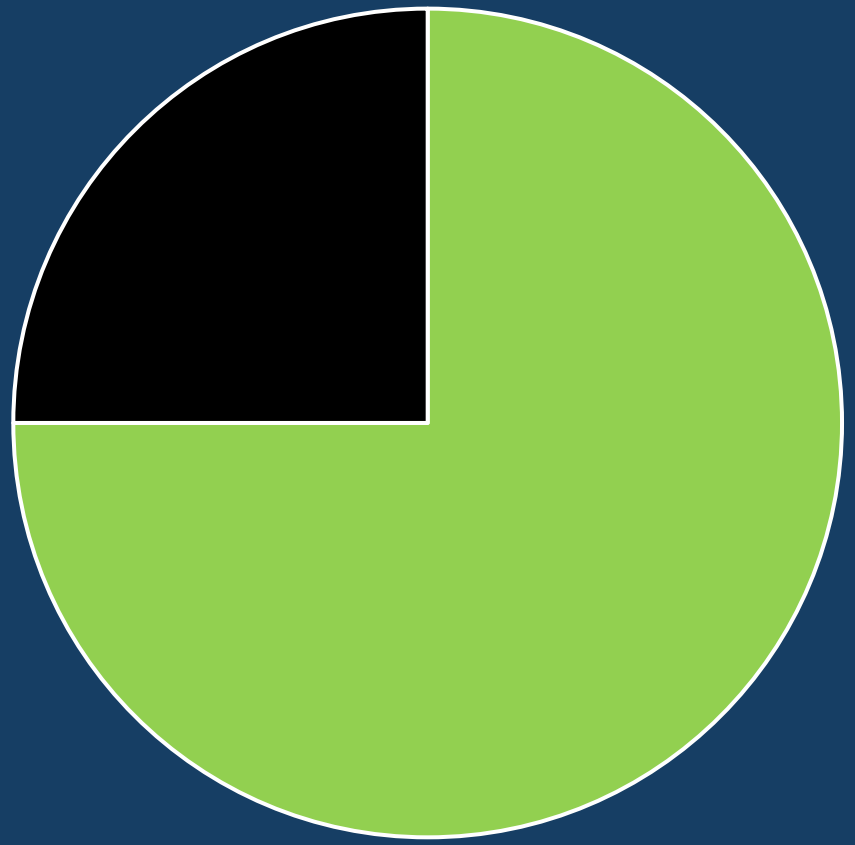
One Approach:

Ignore measurement error and enroll patients if SBR value ≤1.2

1. SPECT SBR measurements: test-retest CV=15%
2. Negligible bias

QIBA's Profile on quantifying dopamine transporters with ^{123}Iodine-labeled ioflupane

Naïve Approach



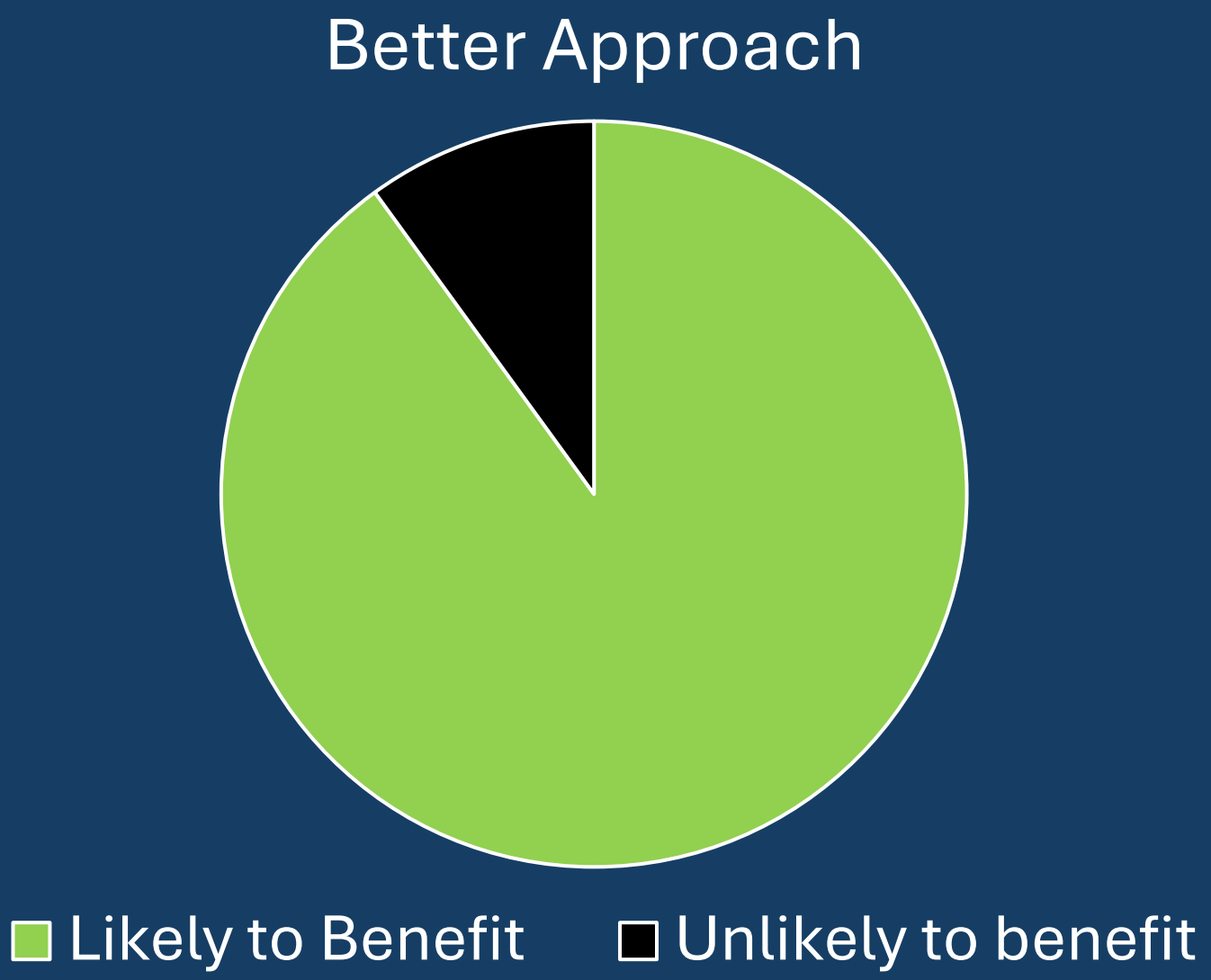
Likely to Benefit

Unlikely to benefit

Better Approach:

Construct 95% Confidence Interval (CI) for SBR
based on our knowledge of
its imprecision and bias

Enroll patients if confidence interval ≤ 1.2


Better Approach
Likely to Benefit
Unlikely to benefit

Example 3: **Estimate Treatment Effect in RCT**

Example 3: PDFF is non-invasive, early endpoint for nonalcoholic fatty liver disease

Proton-density fat fraction (PDFF) measured on MRI has been used as an endpoint in several recent RCTs:

Example 3: PDFF as non-invasive, accurate endpoint for nonalcoholic fatty liver disease

- Harrison et al [2019] conducted RCT of resmetiron vs. placebo for treatment of NASH using PDFF as primary endpoint.

  6.2% reduction with resmetiron

- Pooler et al [2019] compared PDFF values before and after bariatric surgery to monitor liver fat.

  13.2% reduction after surgery

Example 3: PDFF as non-invasive, accurate endpoint for nonalcoholic fatty liver disease

- Harrison et al [2019] conducted RCT of resmetiron vs. placebo for treatment of NASH using PDFF as primary endpoint.

  6.2% reduction with resmetiron

- Pooler et al [2019] compared PDFF values before and after bariatric surgery to monitor liver fat.

  13.2% reduction after surgery

Both studies assumed slope=1

Example 3: PDFF as non-invasive, accurate endpoint for nonalcoholic fatty liver disease

Several studies suggest slope of measured vs. true value <1.0

- Yokoo et al [2018]: 0.975
- Schneider et al [2021]: 0.940

Example 3: PDFF as non-invasive, accurate endpoint for nonalcoholic fatty liver disease

- Harrison et al [2019] conducted RCT of resmetiron vs. placebo for treatment of NASH using PDFF as primary endpoint:
  - ~~6.2%~~ reduction with resmetiron
  - 6.4 – 6.6%
- Pooler et al [2019] compared PDFF values before and after bariatric surgery to monitor liver fat.
  - ~~13.2%~~ reduction after surgery
  - 13.5 to 14.0%

Obuchowski et al, Clinical Trials, 2021

Clinical Validation

Clinical Validation

- Establish association between biomarker and endpoint of interest

AND

- Demonstrate the usefulness of the biomarker (clinical utility)

Clinical Validation design issues:

- Target population
- Intended use
- External validation is a must!
- Retrospective use of clinical trial data is ok if not same as analytical validation data
- PRoBE = prospective collection, retrospective-blinded-evaluation design
    - Appropriate for validation of diagnostic, prognostic biomarkers
    - Not ok for predictive biomarker validation

- Prospective design required for establishing clinical utility/usefulness

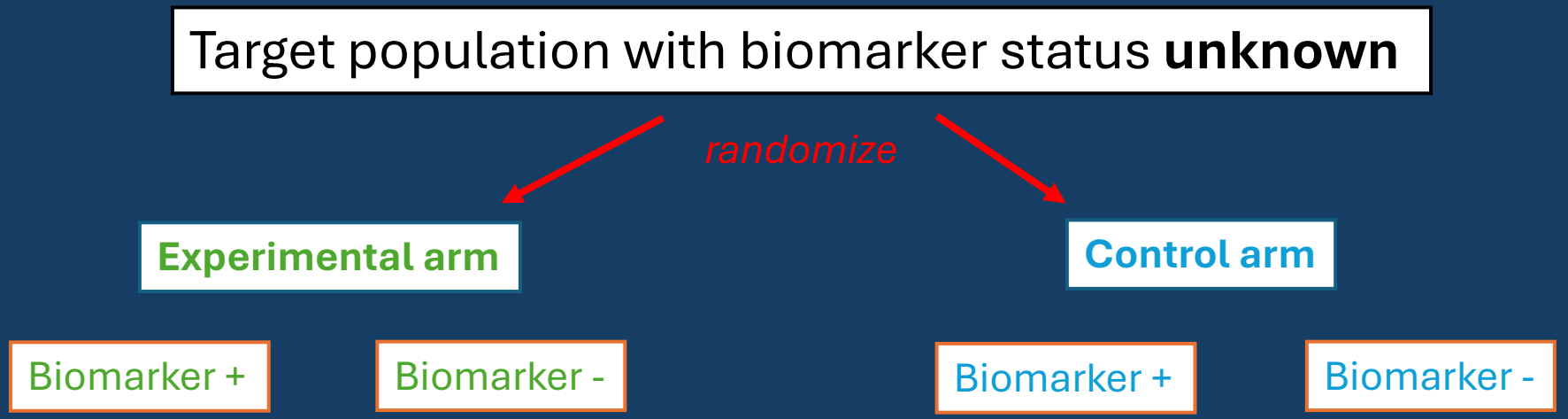
Clinical Utility Validation of Predictive Biomarkers

Clinical Utility Validation of predictive biomarkers

- Sometimes the intervention is expected to work for all patients, just better in biomarker + patients

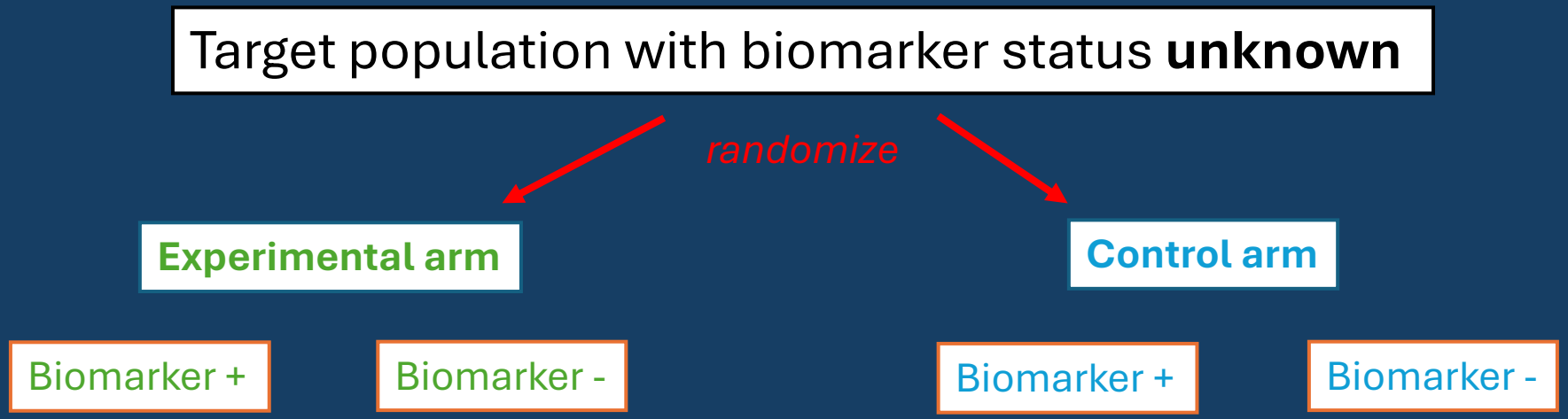
Validation of predictive biomarkers

All-comers design I



Validation of predictive biomarkers

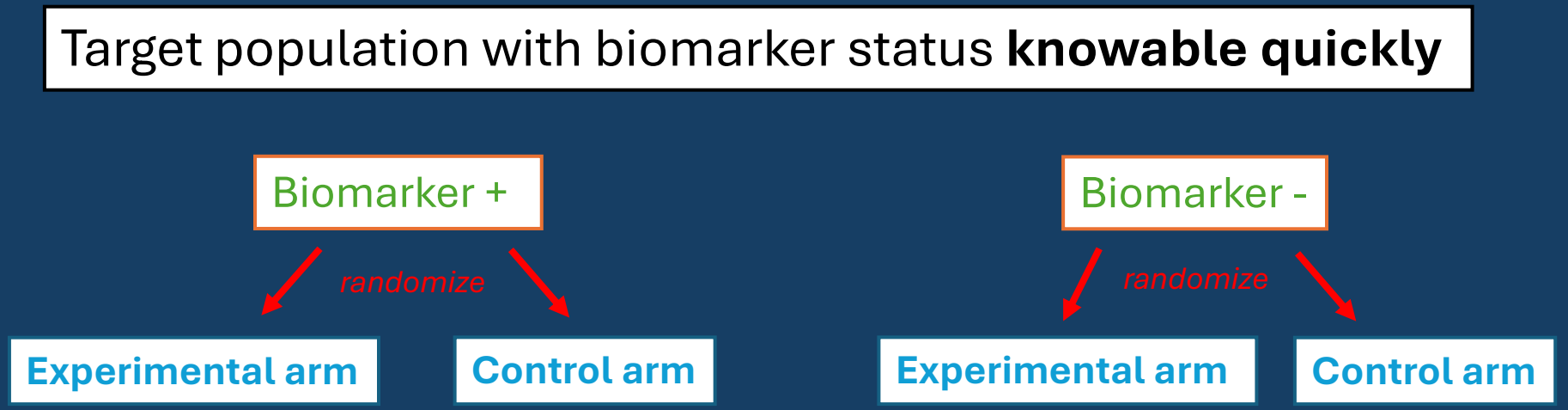
All-comers design I



- Test for interaction of trt x Biomarker
- Powerful design but prevalence of biomarker + subjects can't be low

Validation of predictive biomarkers

All-comers design II

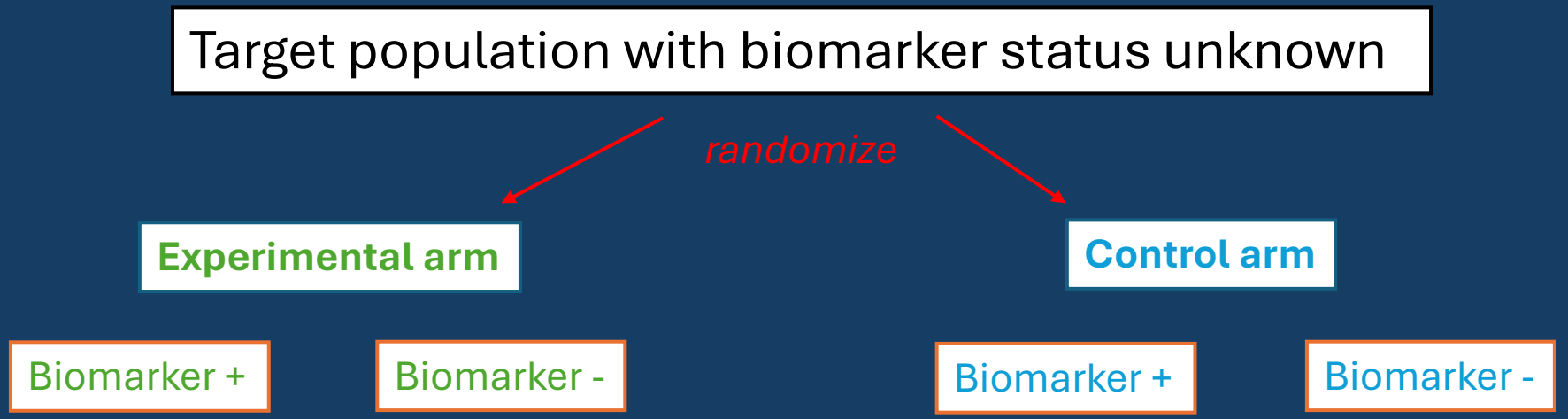


- Test for interaction of trt x Biomarker
- Powerful design, with stratified randomization

Validation of predictive biomarkers

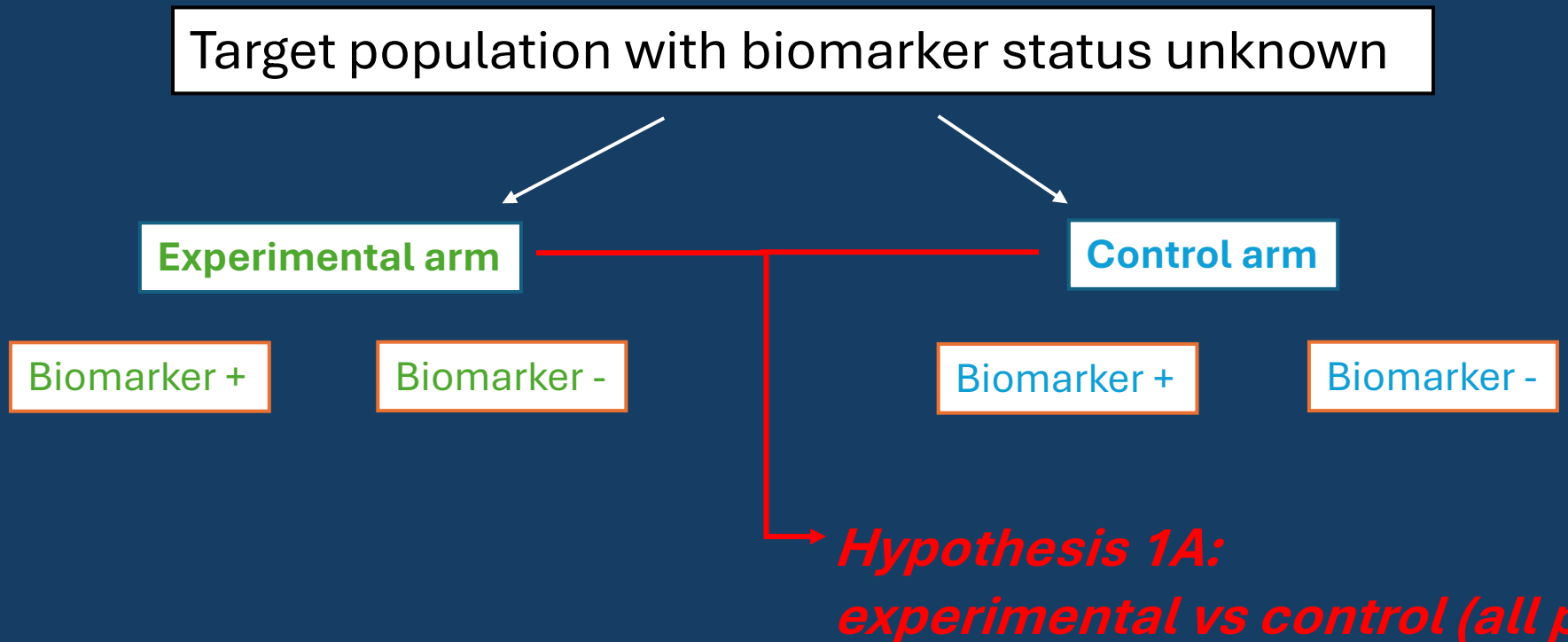
Subgroup design

*Sometimes the focus is more about the subgroup of Biomarker + subjects



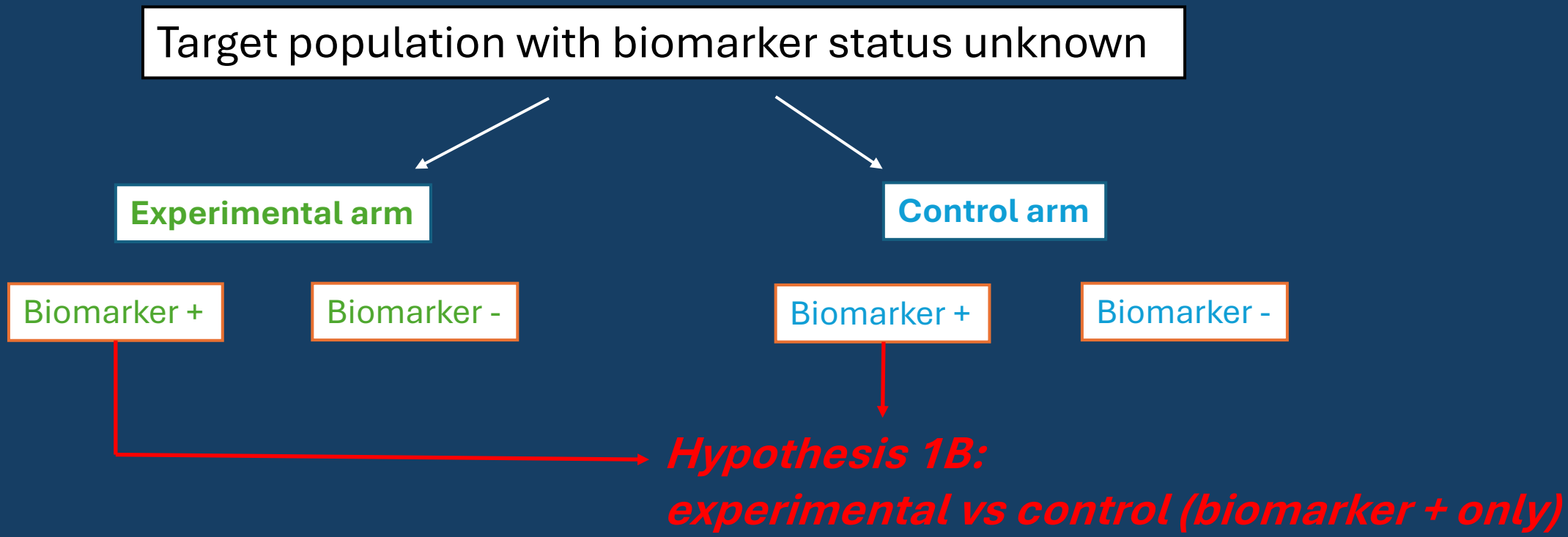
Validation of predictive biomarkers

- Multiple-hypothesis design



Validation of predictive biomarkers

- Multiple-hypothesis design



Validation of predictive biomarkers using multiple-hypotheses design

- Trt effect in 1B >> Trt effect in 1A
- Need to adjust for multiplicity
- Bonferroni adjustment most popular but too conservative
- Ideally, consider correlation between two test statistics when adjusting for multiplicity (See Spiessens and Debois, 2010)
- Need larger N

Other statistical options:

Biomarker tested as **secondary objective**:

- Need to power for it, but often don't power secondary objectives
- What is conclusion if primary objective not met? Can the biomarker still be validated?

Sequential design:

- Don't have to adjust for multiplicity
- If the primary endpoint is not met, then won't be able to test biomarker + subpopulation

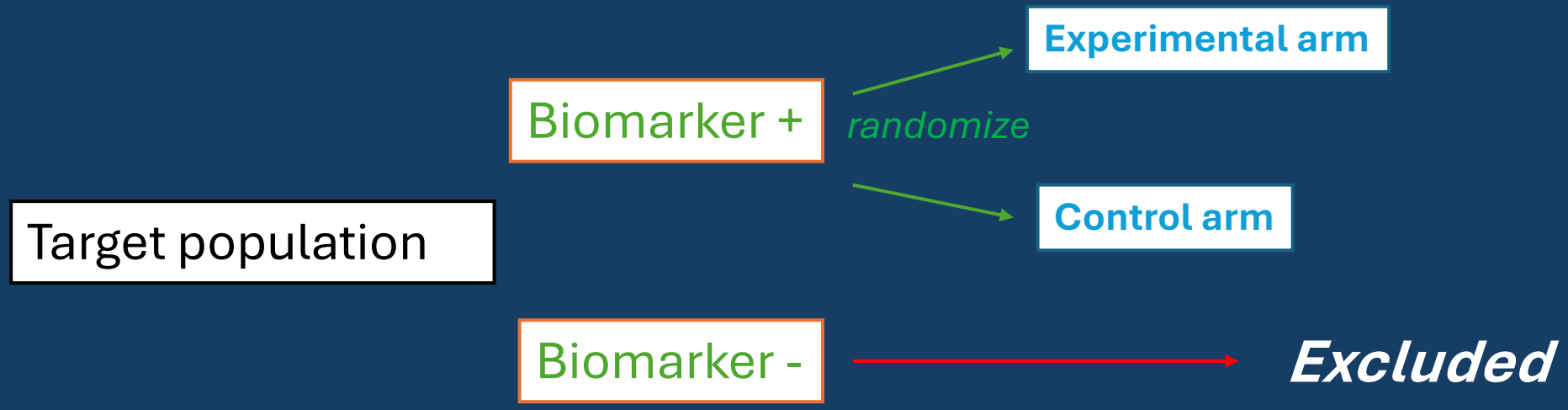
Clinical Utility validation of predictive biomarkers

Sometimes, though, the RCT must be restricted to patients who are biomarker +

Validation of predictive biomarkers

Enrichment design

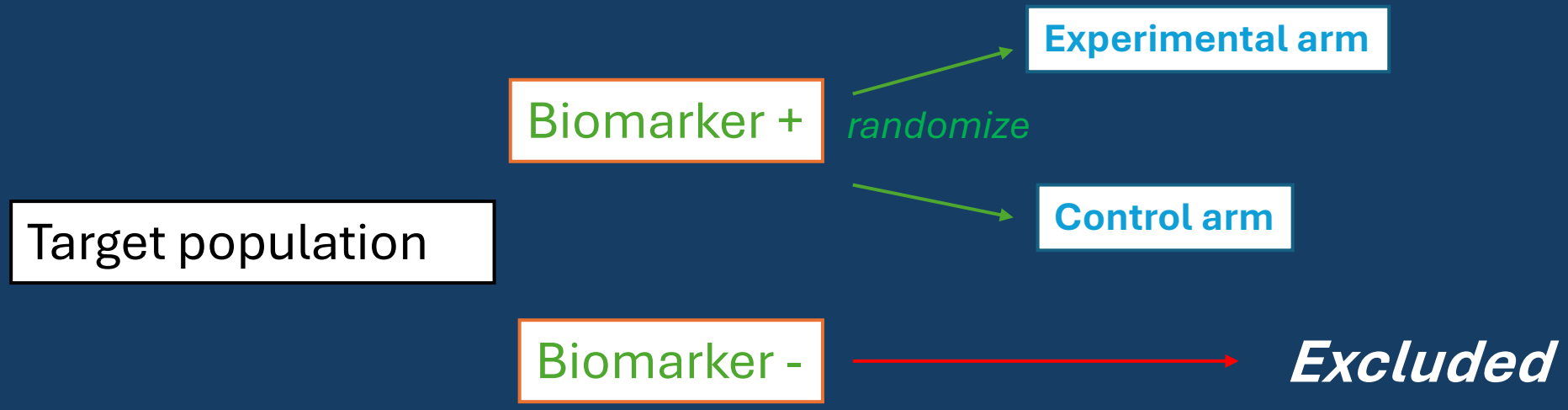
- Restrict trial to Biomarker + patients only



Validation of predictive biomarkers

Enrichment design

- Restrict trial to Biomarker + patients only

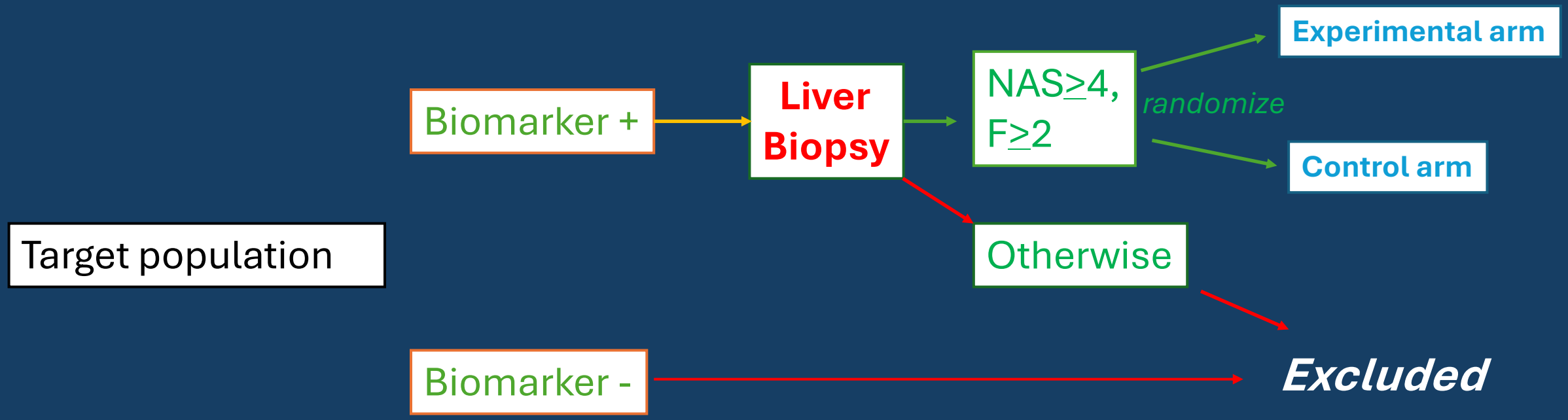


- Test trt effect >> hypothesized value

Validation of predictive biomarkers

Enrichment design

- For biomarkers used to identify patients for inclusion in NAFLD RCTs



Statistical Considerations

Analytical Validation

	Biopsy Result	
Biomarker Finding:	**Eligible**	**Ineligible**
Positive	True positive	Screen failure (FP)
Negative	Screen failure (FN)	True negative

Statistical Considerations

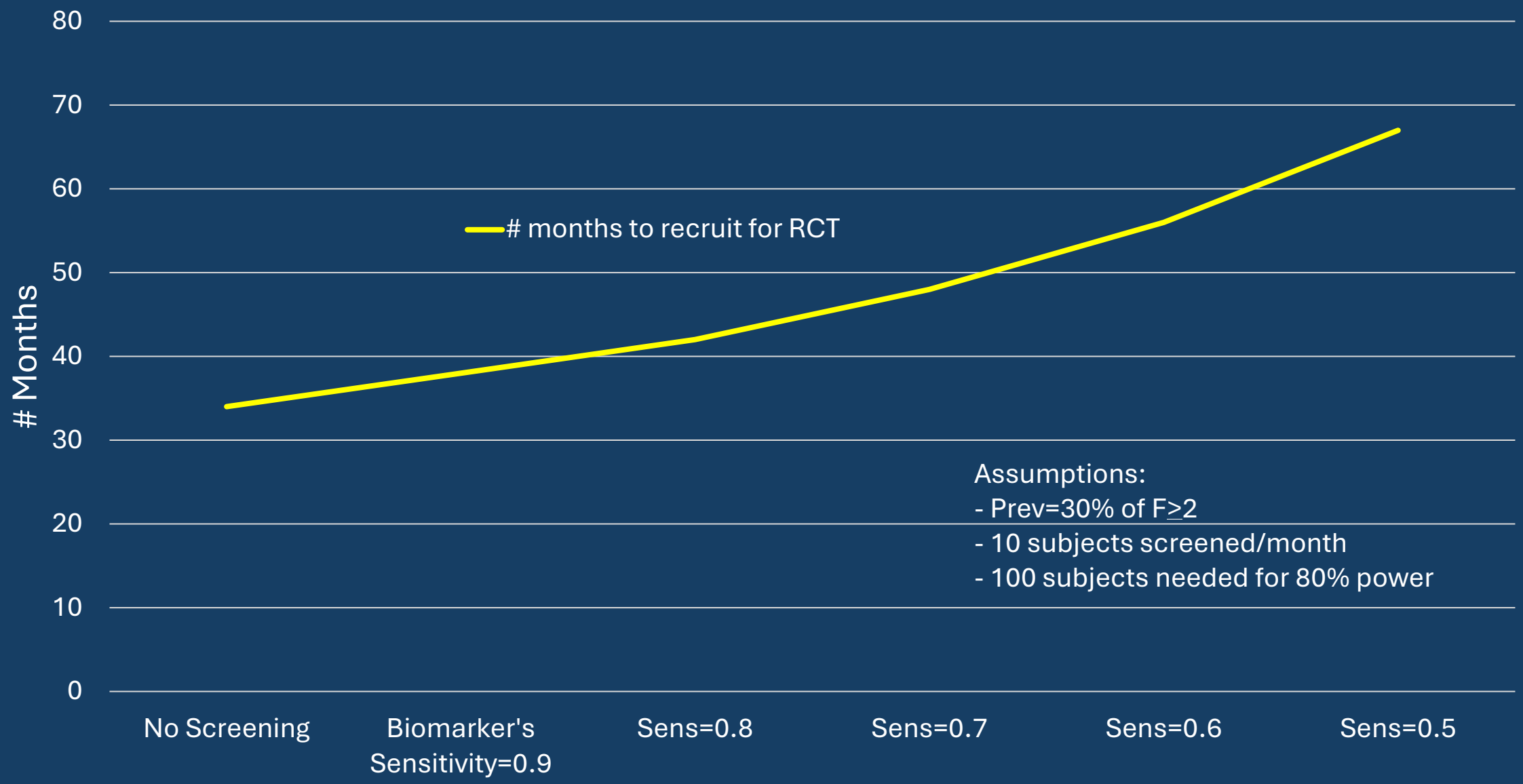
Clinical Validation

	Biopsy Result	
Biomarker Finding:	**Eligible**	**Ineligible**
Positive	**True positive**	**Screen failure (FP)**
~~**Negative**~~	~~Screen failure (FN)~~	~~True negative~~

Clinical validation: Screen + failure rate = (1-PPV) = FP / (TP + FP)

Other Statistical Considerations:

- How long will it take to enroll enough TPs?


80
70
60
50
40
30
20
10
0
Months
months to recruit for RCT
Assumptions:
- Prev=30% of F≥2
- 10 subjects screened/month
- 100 subjects needed for 80% power
No Screening
Biomarker's Sensitivity=0.9
Sens=0.8
Sens=0.7
Sens=0.6
Sens=0.5

Other Statistical Considerations:

- How many eligible subjects will be excluded from trial?
- Depending on biomarker and its mechanism of action, what is the distribution of fibrosis stage among TPs?
    - Is the relative frequency of fibrosis stages maintained after enrichment?
        - No, not if magnitude of biomarker measurement correlates with fibrosis stage
    - Is there differential effect of treatment based on fibrosis stage?
        - Is trt effect smaller when F2 subjects are excluded?
        - Are results of RCT generalizable?

Other Statistical Considerations:

- How many eligible subjects will be excluded from trial?
- Depending on biomarker and its mechanism of action, what is the distribution of fibrosis stage among TPs?
    - Is the relative frequency of fibrosis stages maintained after enrichment?
        - No, not if magnitude of biomarker measurement correlates with fibrosis stage
    - Is there differential effect of treatment based on fibrosis stage?

Choice of cutpoint must be based on more than accuracy

Conclusions:

1. Discovery biomarkers are hypothesis-generating
   - Need both analytical and clinical validation before “fit-for-purpose”
2. Target population
   - Identify early!
   - Study designs, even at discovery phase, depend on this
3. Role of imaging with non-imaging biomarkers - multiparametric
   - Imaging may have lower specificity in detection but critical for localizing, staging
   - Use biomarkers as continuous variables to maximize performance
4. Measurement error leads to low powered studies, wrong treatment effect estimates, and misinterpretation of findings
   - Methods to adjust for it are easy to use
5. Predictive biomarkers
   - Must consider cutpoints carefully, especially effect on clinical utility (Less reliance on binary results/cutpoints would benefit the field)

References

Ou et al. Biomarker discovery and validation: statistical considerations. J Thorac Oncol 2022

Ou et al. Discussion of trial designs for biomarker identification and validation through use of case studies. JCO Precis and Oncol 2019

Spiessens, Debois. Adjusted significance level for subgroup analyses in clinical trials. Contemporary Clinic Trials 2010

deSouza et al. Validated imaging biomarkers as decision-making tools in clinical trials and routine practice. Insights Imaging 2019

Obuchowski et al. Statistical considerations for planning clinical trials with QIBs. J Nat Cancer Institute 2019

Obuchowski et al. Importance of incorporating QIB technical performance characteristics when estimating treatment effects. Clinical Trials 2021