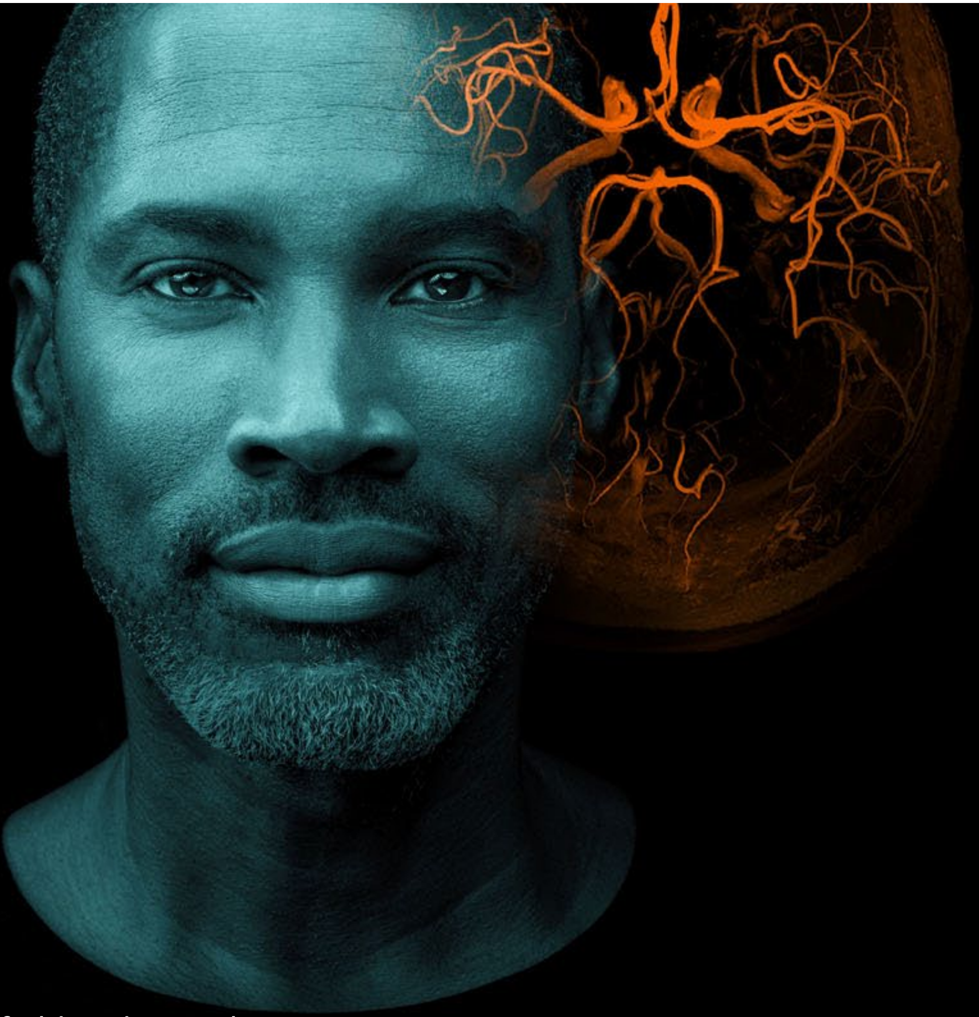
SIEMENS
Healthineers



A Structured Approach to Designing Pivotal Diagnostic Accuracy Studies for Noninvasive Tests Using Meta-Analytic Techniques

Presented By:
Ed De Vol, PhD, Principal Biostatistician, Clinical Biostatistics
Ryan Butterfield, DrPH, MBA, Senior Director, Clinical Biostatistics

Speakers

Dr. Ed De Vol

- Though born and raised in central Ohio, Ed now lives in Dallas, Texas with his wife who is originally from the Middle East. They have four grown children who are scattered across three continents – Asia, Europe and North America.
- Ed attended Oberlin College and graduated with a BA in Mathematics and then went to The University of Michigan in Ann Arbor for his MSc and PhD in Biostatistics with a thesis on martingale methods for censored data. He also has an MBA from Duke University.
- Ed took a faculty position with the Mathematics Department at Oakland University in Rochester, Michigan following his PhD. He then accepted a position at the King Faisal Specialist Hospital in Riyadh, Saudi Arabia. Following that he worked at the Baylor Healthcare System in Dallas as Vice President of Quantitative Sciences for almost ten years.
- Since 2023, Ed has worked as a Principal Biostatistician with Siemens Healthineers working primarily to support evidence generation activities for the company.

Dr. Ryan Butterfield

- Dr. Ryan Butterfield resides in St. Augustine, FL with family. On a personal note, he enjoys time with his family and golfing and reading as much as having little kids and his wife permits.
- His educational experiences resulted in a BS in Biomathematics and MPH in Biostatistics from Loma Linda University, an MBA from Jacksonville University, and a DrPH in Biostatistics from Georgia Southern University, where he was a BASS Fellow.
- Since graduating, he has worked as a Biostatistician in academia, hospital/clinical/University settings, varying levels of government (Local, State, Federal), and at several multi-international corporations at varying levels including 3M, Johnson & Johnson, and Edwards Lifesciences.
- He currently is a Senior Director of Clinical Biostatistics at Siemens Healthineers, where his team supports all stages of product development from R&D through Market entry.

Contents

- Introductions and Overview
    - Speakers
    - Learning Objectives
    - Problem and Issues in Clinical Study Design
        - in Diagnostics Research
- Background of Liver Fibrosis Diagnostics
- The How and Why of Using Meta-Analysis in Liver Fibrosis Diagnostics
    - Overview of General Meta-Analysis Approaches and Use in Regulatory Studies
    - Specialized Approaches
        - Steinhauser
- Meta-Analysis Statistical Analysis
    - Steinhauser
        - Modeling
        - Variance Extractions
        - R-Code
- Results
    - ROC Curves
    - Performance Statistics
    - etc
- Conclusions and Indications for General Statistical Application
    - Other areas where this type of approach for both specialized meta-analysis techniques and Steinhauser may be applied…beyond diagnostics?

Overview

RB

Overview

- Chronic diseases (e.g., liver fibrosis) require accurate noninvasive biomarkers
- Tests often yield continuous values -> cutoff selection critically affects accuracy
- Clinical decision-making depends on sensitivity vs specificity trade-offs
- Clinicians ask: *“At what threshold does this test perform best?”*
- Summary ROC (SROC) curves describe performance, but threshold selection remains unresolved

Learning Objectives

- **Gain perspective** for the role of advanced meta-analysis in informing *a priori* clinical study design
- Apply meta-analytic estimation to **rationally plan pivotal** regulatory studies
- **Understand challenges** in designing pivotal diagnostic accuracy studies for noninvasive tests (NITs)
- **Recognize the limitations** of conventional meta-analysis approaches when studies use multiple thresholds
- **Describe** the Steinhauser reconstruction technique and its advantages in synthesizing diagnostic performance data

Relative Problems and Issues in Clinical Study Design

General Clinical Trial Study Design

- Integration of RWE into Design
- Difficulty in estimating variability components required for power/sample size planning
- Regulatory expectations (FDA/EMA) demand robust, generalizable study designs

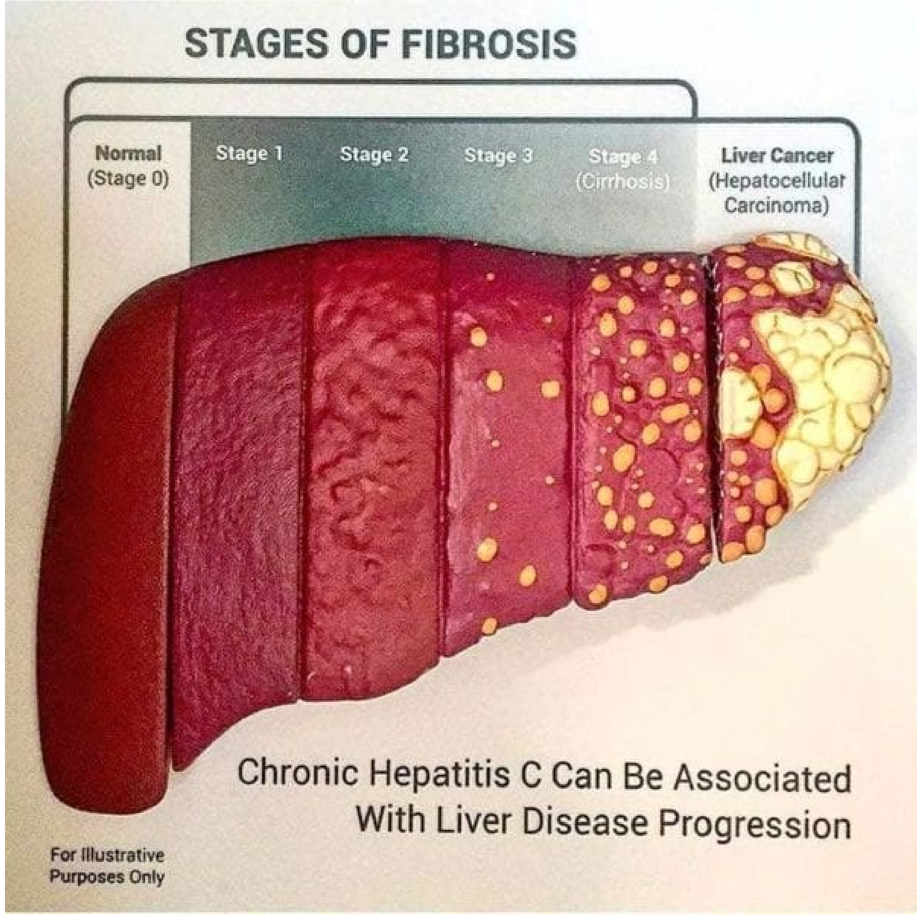
Specific to Diagnostics

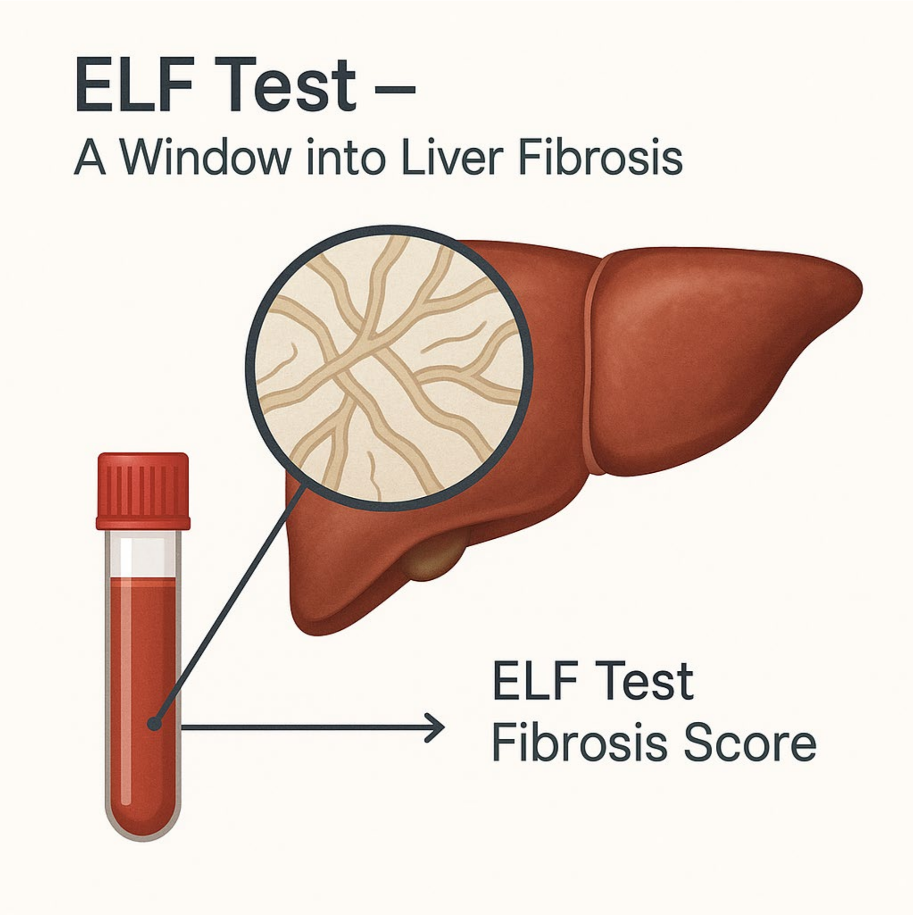
- Small or fragmented prior studies -> unstable sensitivity/specificity estimates
- Heterogeneity in cutoff thresholds across published/ unpublished studies
- Selective reporting of “optimal” thresholds -> over-optimistic accuracy estimates

Disease Background: Liver Fibrosis

EDV

Liver Fibrosis Staging and Diagnostic Testing


STAGES OF FIBROSIS
Normal (Stage 0)
Stage 1
Stage 2
Stage 3
Stage 4 (Cirrhosis)
Liver Cancer (Hepatocellular Carcinoma)
Chronic Hepatitis C Can Be Associated With Liver Disease Progression
For Illustrative Purposes Only



Liver Cancer and Fibrosis Diagnostic Testing

- In 2nd century, physician Aretaeus of Cappadocia provided the first clinically recognizable description of liver cancer
- Today, hepatitis, alcohol use, obesity, diabetes, smoking, and others are recognized as risk factors
- Worldwide - the sixth most prevalent cancer and is the third most common cause of cancer death
- **Symptoms don't tend to appear until the cancer has progressed. Survival remains poor even in high income countries.**
- **Early detection of liver fibrosis enables timely intervention – preventing progression to cirrhosis and liver cancer**

Non-invasive Tests (NITs)

- ELF

  $B0 + B1*\ln(HA) + B2*\mathit{ln}(PIINP) + B3*\mathit{ln}(TIMP\text{-}1)$

- FIB-4

$$\frac{\text{Age (years)} \times \text{AST (U/L)}}{\text{Platelet count } (10^9\text{/L}) \times \sqrt{\text{ALT (U/L)}}}$$

- VCTE (kPa)

  vibration-controlled transient elastography

- MRE (kPa)

  magnetic resonance elastography

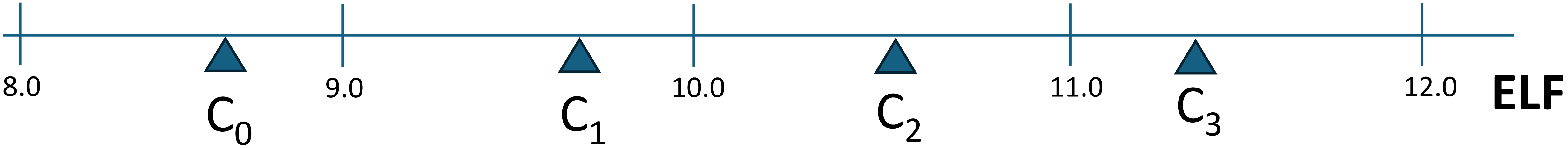
Invasive Test - Biopsy

Overview of ELF Testing

- Enhanced Liver Fibrosis (ELF) can assess the risk of disease progression in patients with advanced liver damage
- Components
  - Hyaluronic Acid – extracellular matrix
  - PIIINP – collagen synthesis
  - TIMP-1 – inhibition of matrix degradation
- Biopsy staging
  - F0 – no fibrosis
  - F1 – mild fibrosis
  - F2 – significant fibrosis
  - F3 – advanced fibrosis
  - F4 – cirrhosis

Testing ELF thresholds for Diagnostic Accuracy

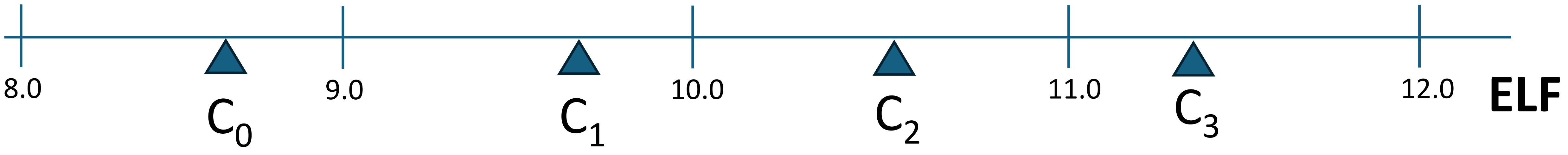
- Biopsy staging
  - F0 – no fibrosis (ELF < C_0)
  - F1 – mild fibrosis (C_0 < ELF < C_1)
  - F2 – significant fibrosis (C_1 < ELF < C_2)
  - F3 – advanced fibrosis (C_2 < ELF < C_3)
  - F4 – cirrhosis (ELF > C_3)



Diagnostic Accuracy for Cirrhosis (i.e., F4)

- Biopsy staging
  - F0 – no fibrosis (ELF < C_0)
  - F1 – mild fibrosis (C_0 ≤ ELF < C_1)
  - F2 – significant fibrosis (C_1 ≤ ELF < C_2)
  - F3 – advanced fibrosis (C_2 ≤ ELF < C_3)
  - **F4 – cirrhosis (ELF > C_3)**

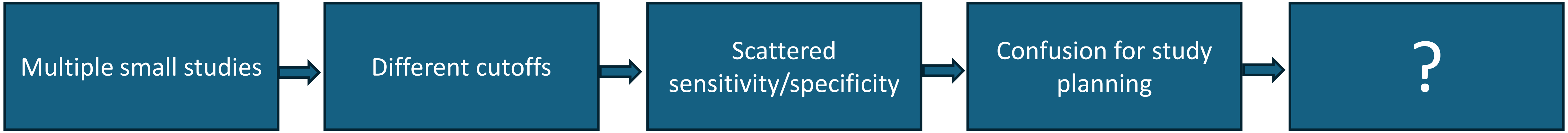
	≤ F3	> F4
ELF < C_3	**n11**	**n12**
ELF ≥ C_3	**n21**	**n22**



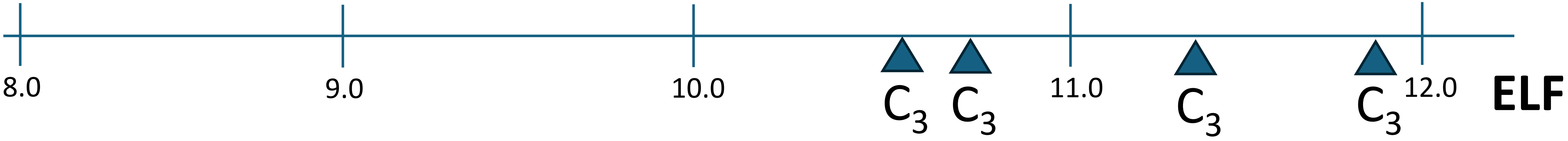
The How and Why of Using Meta-Analysis in Liver Fibrosis Diagnostics

EDV

Why Meta-Analysis Matters for NITs



Multiple small studies → Different cutoffs → Scattered sensitivity/specificity → Confusion for study planning → ?



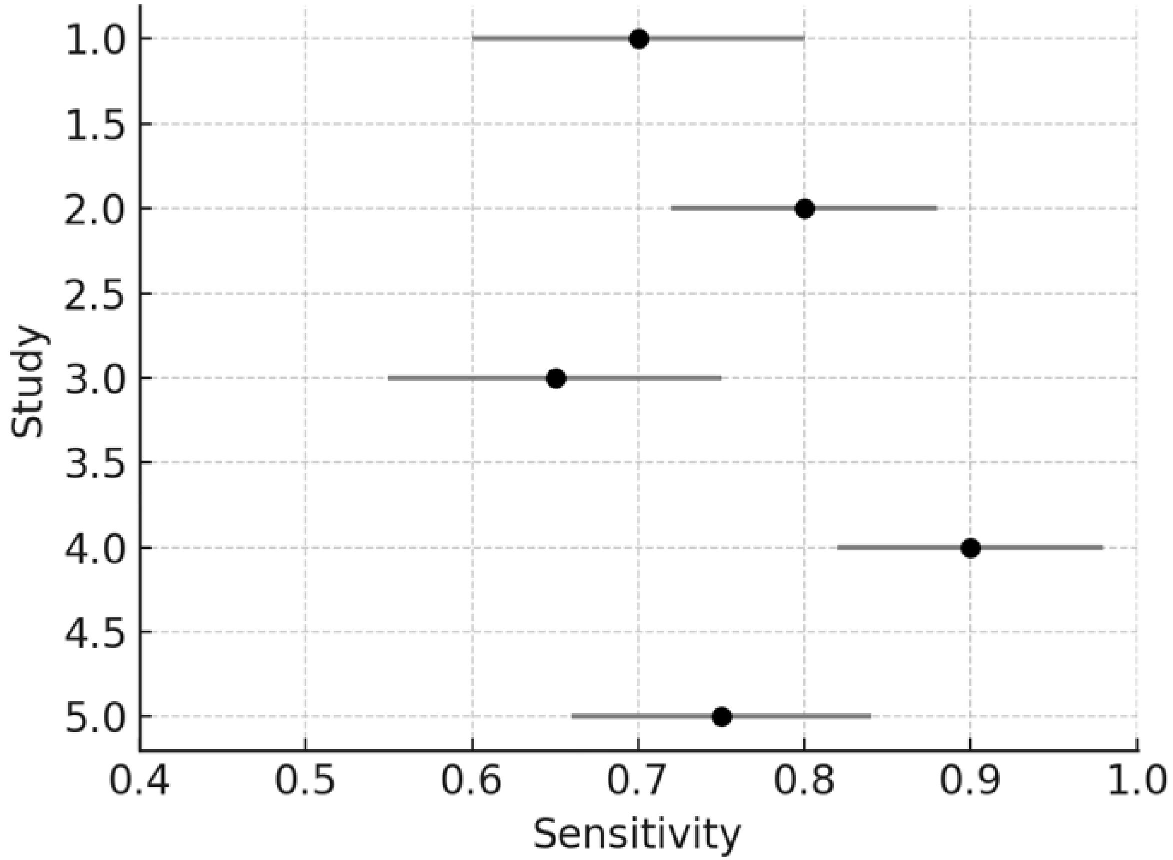
Classical Meta-Analysis

Studies S_i

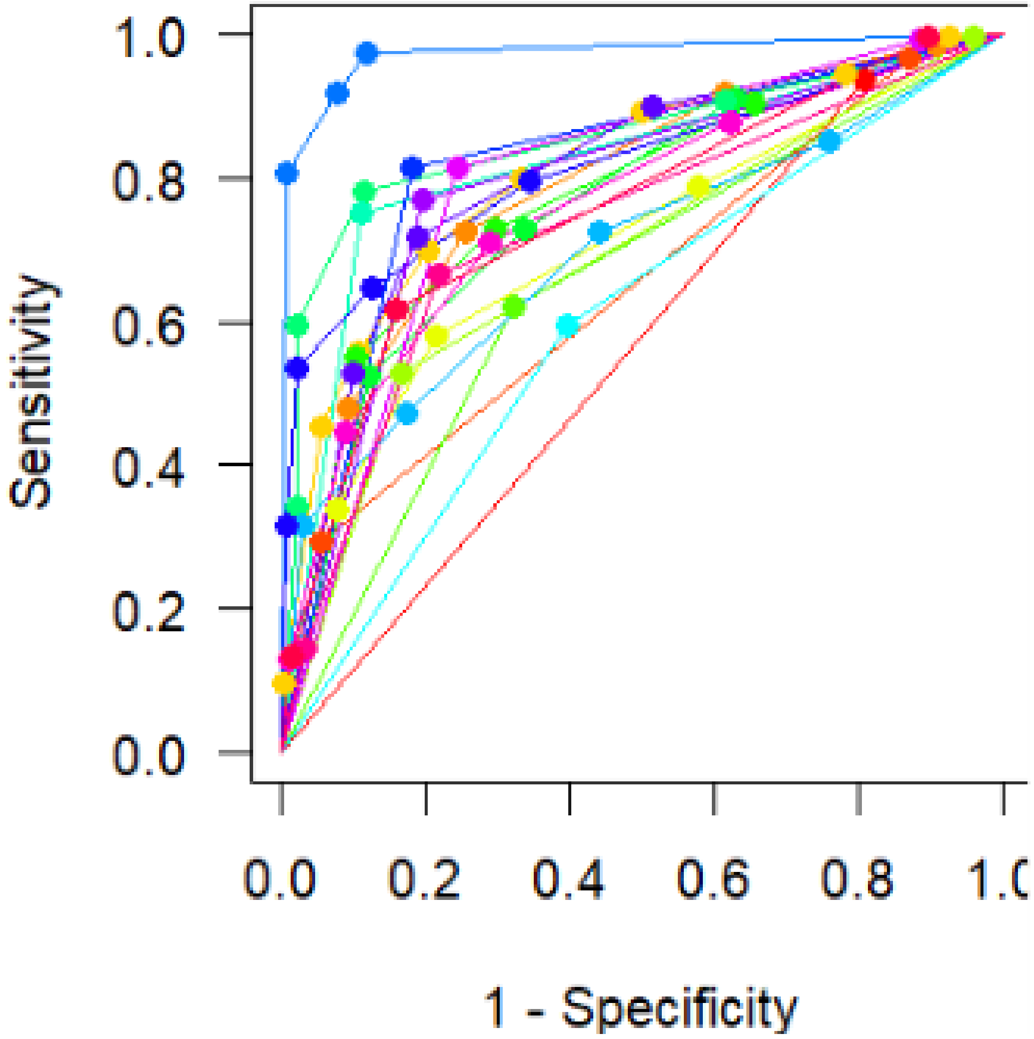
Single threshold C	Disease +	Disease -
< C	TP	FP
≥ C	FN	TN

≥ C FN TN

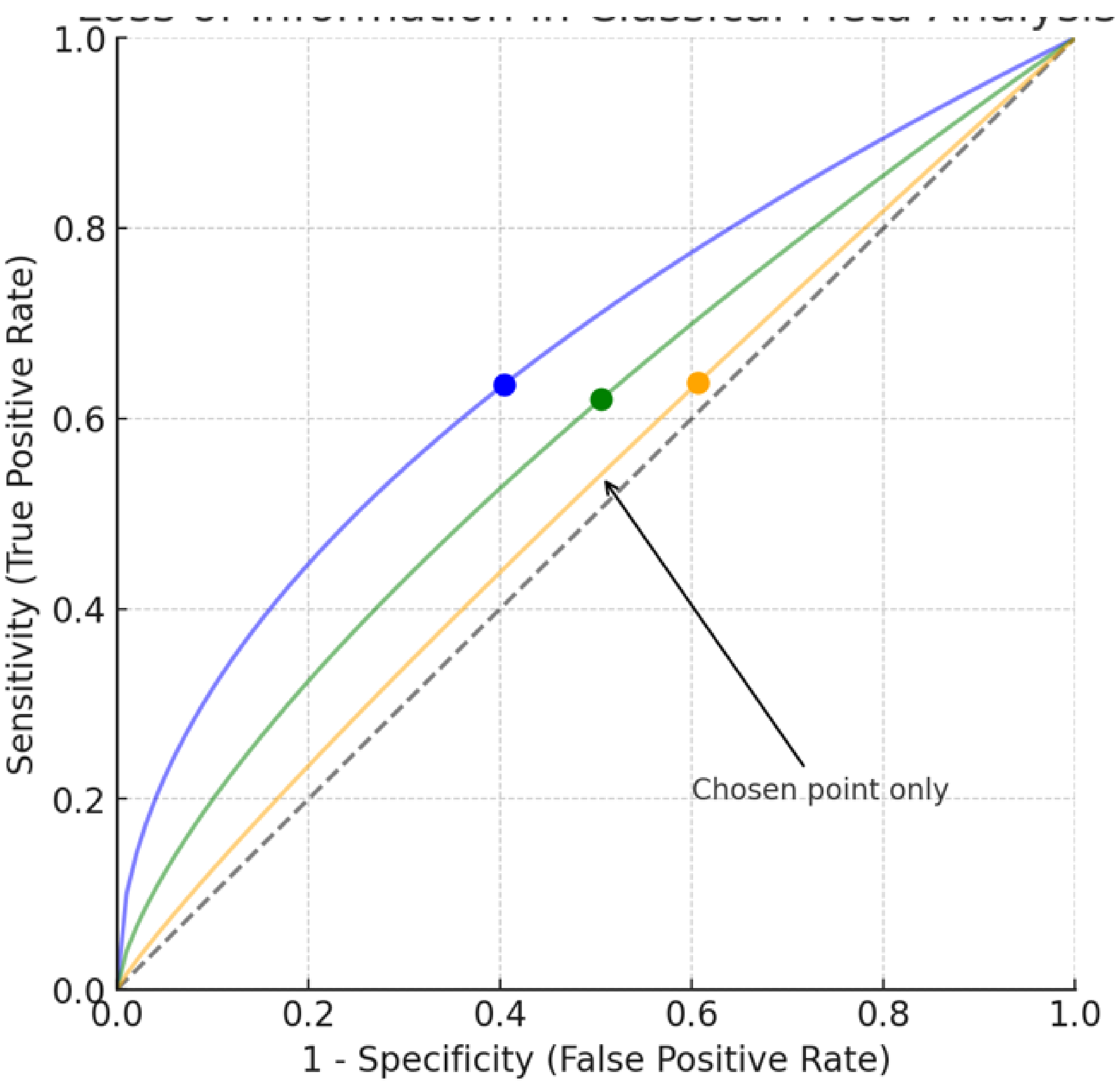
≥ C FN TN



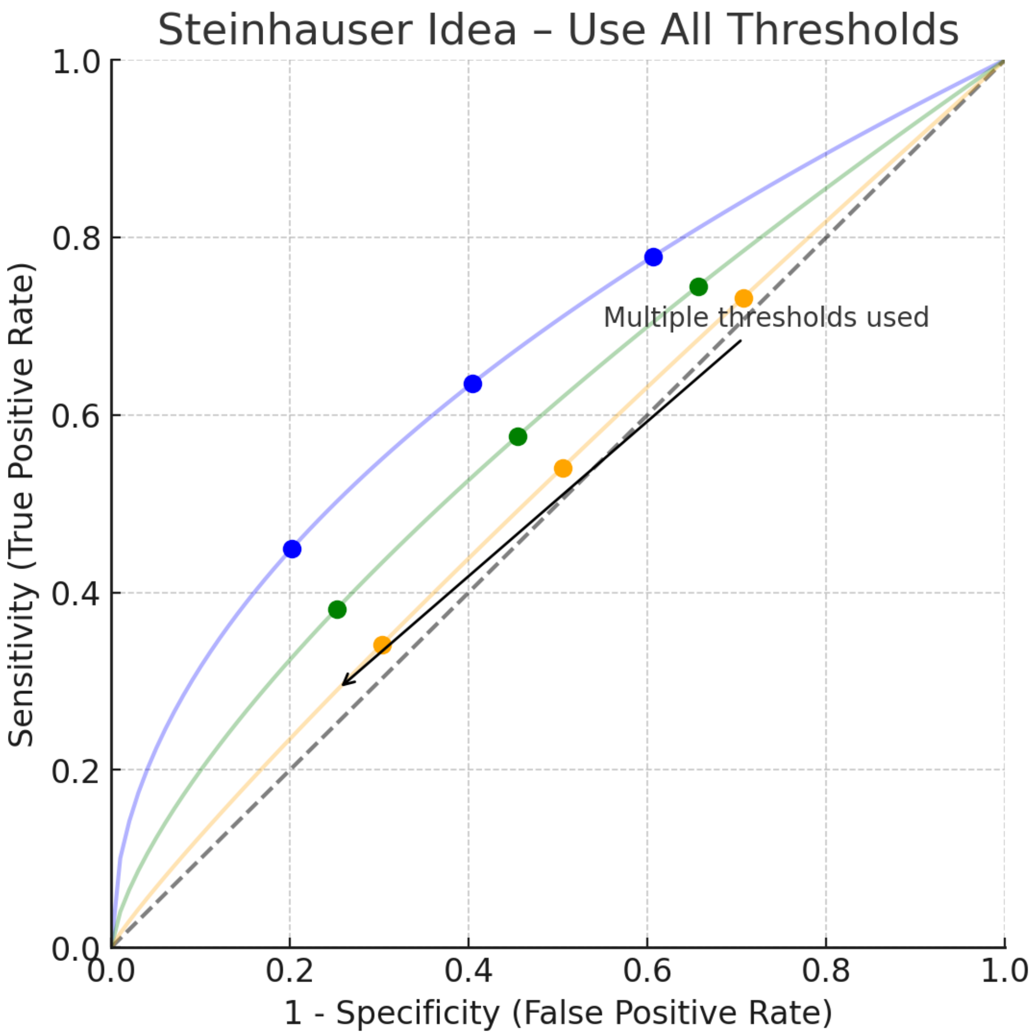
Heterogeneity of Thresholds (i.e., the data)



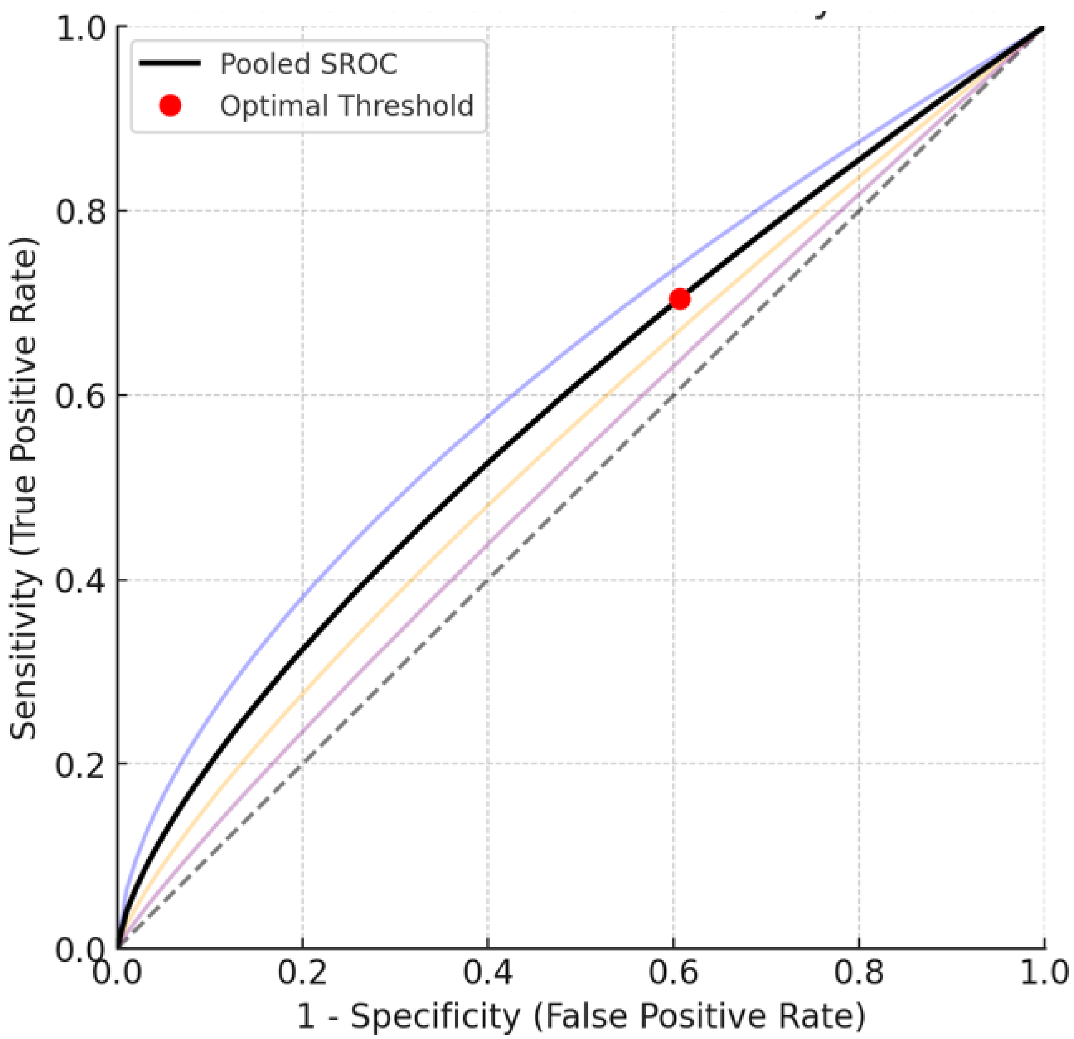
Loss of Information in Classical Meta-Analysis



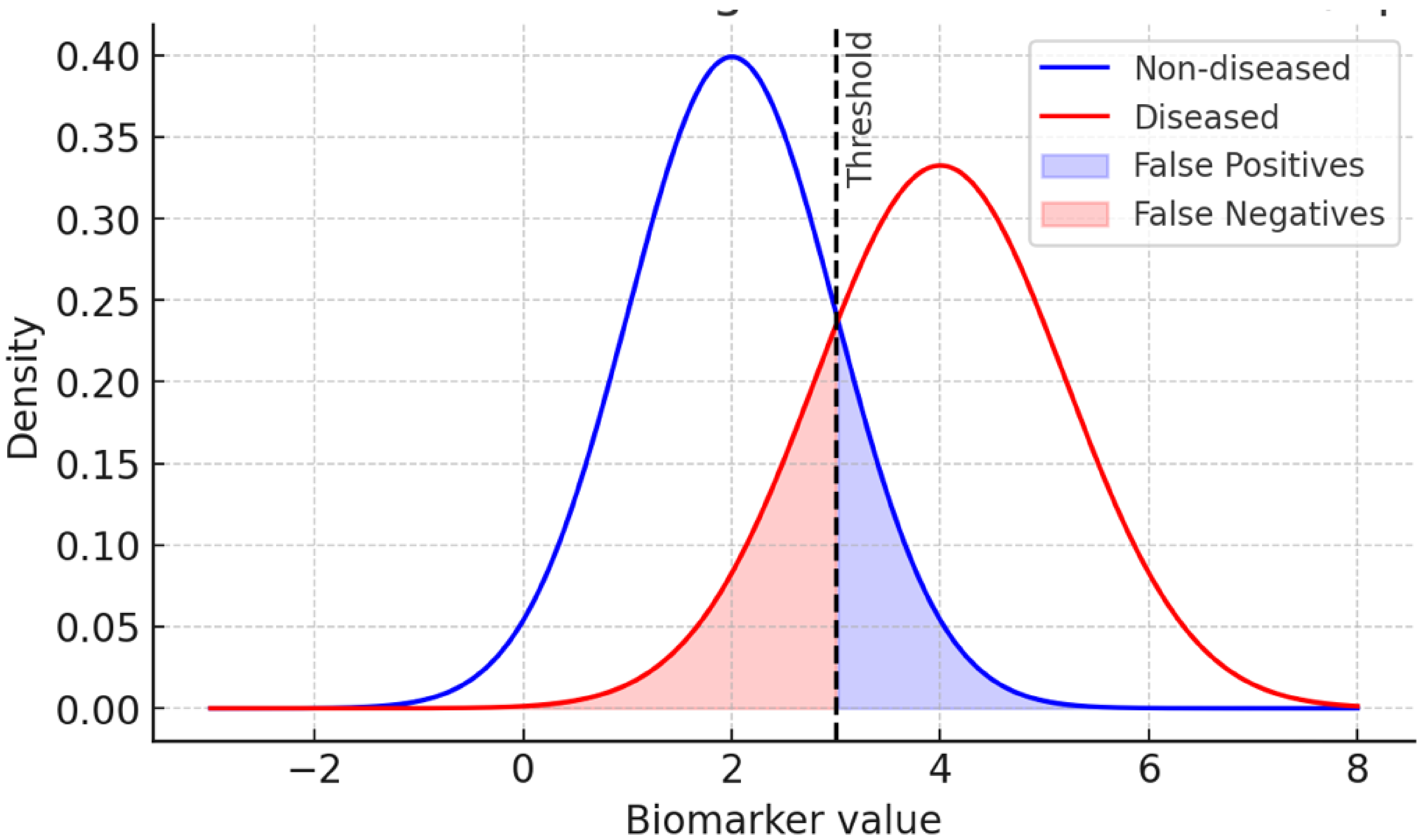
Steinhauser idea – Use All Thresholds as Input



Pooled SROC Curve With Study Curves



Distributional Thinking: Threshold Defines Se/Sp



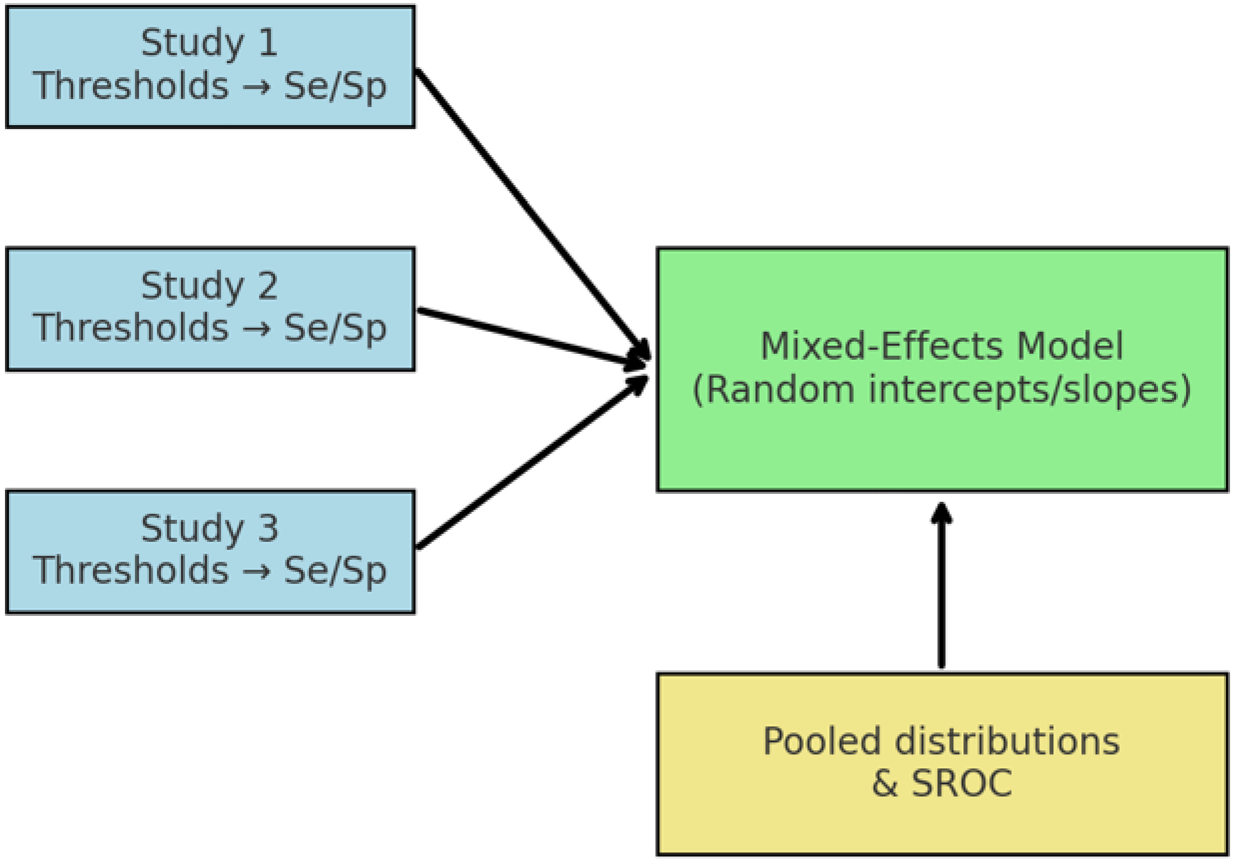
Specialized Statistical Framework

- Key idea – estimate the cumulative distribution functions (CDFs) of the biomarker for diseased and non-diseased individuals
- Assumptions – Normal or logistic distribution for biomarker values
- Transformations - use a logit or probit transformation to linearize sensitivity and specificity

$$h\left(\mathrm{Sp}(x)\right) = \frac{x - \mu_0}{\sigma_0},$$

$$h\left(1 - \mathrm{Se}(x)\right) = \frac{x - \mu_1}{\sigma_1},$$

Hierarchical Mixed-Effects Model



Study 1
Thresholds → Se/Sp

Study 2
Thresholds → Se/Sp

Study 3
Thresholds → Se/Sp

Mixed-Effects Model
(Random intercepts/slopes)

Pooled distributions
& SROC

Linear mixed effects model for Specificity (h(Sp(x))

- **Response**
  - the transformed proportion of true negatives
- **Fixed effects**
  - intercept – fixed and representing the overall average transformed specificity across all studies
  - slope – fixed and representing the overall average rate of change in the transformed specificity with respect to the threshold x
- **Random effects**
  - intercept – random for study s_i, capturing study-specific deviations from the overall average intercept
  - slope – random for study s_i, capturing study-specific deviations from the overall average slope

Justification for specialized Meta-Analysis Methods

- With regards to diagnostics:
  - Standard bivariate/meta-analytic models use only one threshold per study -> discards valuable data
  - Steinhauser technique uses *all reported thresholds* to reconstruct biomarker distributions
  - Provides pooled sensitivity/specificity estimates across cutoffs
  - Enables determination of an **optimal threshold** across studies
  - Extracts variance components necessary for *a priori* power/sample size calculations
  - Improves planning for pivotal clinical studies intended for regulatory submission

Statistical Analysis

RB

Ten-Step Workflow

1. Define Objective – Clarify test purpose, clinical question, regulatory target
2. Decompose Task – Break into components: review, extraction, modeling, design
3. Data identification – Systematically gather studies and threshold data (TP/FP/TN/FN)
4. Harmonize Cutoffs – Standardize thresholds and outcome definitions across studies
5. Reconstruction – Apply Steinhauser algorithm to transform Se/Sp into distributions
6. Model Components – Fit linear mixed-effects models across thresholds and studies
7. Extract Variance – Derive within- and between-study variance of Se/Sp estimates
8. Simulate Scenarios – Explore power/sample size under alternative cutoffs
9. Select Designs – Identify candidate designs meeting sensitivity/specificity goals
10. Optimize & Refine – Balance regulatory, clinical, and practical constraints

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Clearly state the clinical question, e.g., “Estimate diagnostic accuracy of ELF for $\geq$ F3 fibrosis.”
- Specify intended use (rule-in, rule-out, balanced) and regulatory context (FDA/EMA)
- Determine required outputs: sensitivity, specificity, AUC, optimal threshold, variance estimates

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Protocol development and registration
- Break down the study design problem into smaller components:
  - Literature retrieval
  - Data extraction (TP, FP, TN, FN at thresholds)
  - Modeling steps
  - Variance estimation -> study design implications
- Assign roles and analytic tasks (systematic review team vs. statistical modeling team)

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Systematic review of published diagnostic accuracy studies
- Extract all reported thresholds per study (not just “best” cutoffs)
- Records TP, FP, TN, FN for each threshold
- Note disease spectrum, reference standards, study-level covariates

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Standardize definitions of disease outcomes (e.g., fibrosis stages ≥ F2, ≥ F3, F4)
- Normalize thresholds to consistent units (e.g., ELF score ranges)
- Address incomplete reporting: continuity corrections, imputations when needed
- Ensure comparability across studies before pooling

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. **Steinhauser Reconstruction**
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Transform sensitivity/specificity values into linearized scale (e.g., logit or probit)
- Model these as points on the biomarker’s CDF in diseased vs. non-diseased groups
- Fit linear mixed-effects regression with:
  - Fixed effects -> means (m0, m1) and variances (s0, s1) of biomarker distributions
  - Random effects -> study-level intercepts/slopes capturing heterogeneity

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Fit alternative random-effects structures (CIDS, DICS, CICS, etc.)
- Compare models using REML, AIC, or cAIC criteria
- Select the most stable/parsimonious model
- Back-transform estimates to obtain sensitivity/specificity across thresholds

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Quantify within-study and between-study variability in Se/Sp estimates
- Derive standard errors and covariance of Se/Sp at clinically-relevant cutoffs
- These become the variance inputs for power and sample size calculations in new studies

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Use estimated distributions to simulate “future” study performance
- Explore trade-offs: high sensitivity design vs high specificity design vs balanced design
- Evaluate performance under varying prevalence rates or disease severity distributions

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- From simulations, identify feasible study designs that meet regulatory performance targets
- For example:
  - Rule-out study: prioritize sensitivity $\geq$ 90%
  - Rule-in study: prioritize specificity $\geq$ 90%
  - Balanced study: maximize Youden index
- Propose sample size ranges and cutoff strategies

Ten-Step Workflow

1. Define Objective
2. Decompose Task
3. Data Identification & Acquisition
4. Harmonize Cutoffs/Outcomes
5. Steinhauser Reconstruction
6. Model Across Components
7. Extract Variance Components
8. Simulate Scenarios
9. Select Candidate Designs
10. Optimize/Refine

- Refine design recommendations considering:
  - Practical feasibility (sample size, patient recruitment)
  - Regulatory guidance (FDA/EMA minimum performance criteria)
  - Clinical utility (trade-offs clinicians actually need)
- Finalize study design plan with variance-backed justification

CASE STUDY

EDV

ELF Meta-Analysis (submitted to *APT*)

Project team: Arun Sanyal (MD), Matt Gee (Med Science), Ed De Vol (Biostats), Don Chalfin (MD), Toana Kawashima (Biostats), Roma Levy (Med Writing)

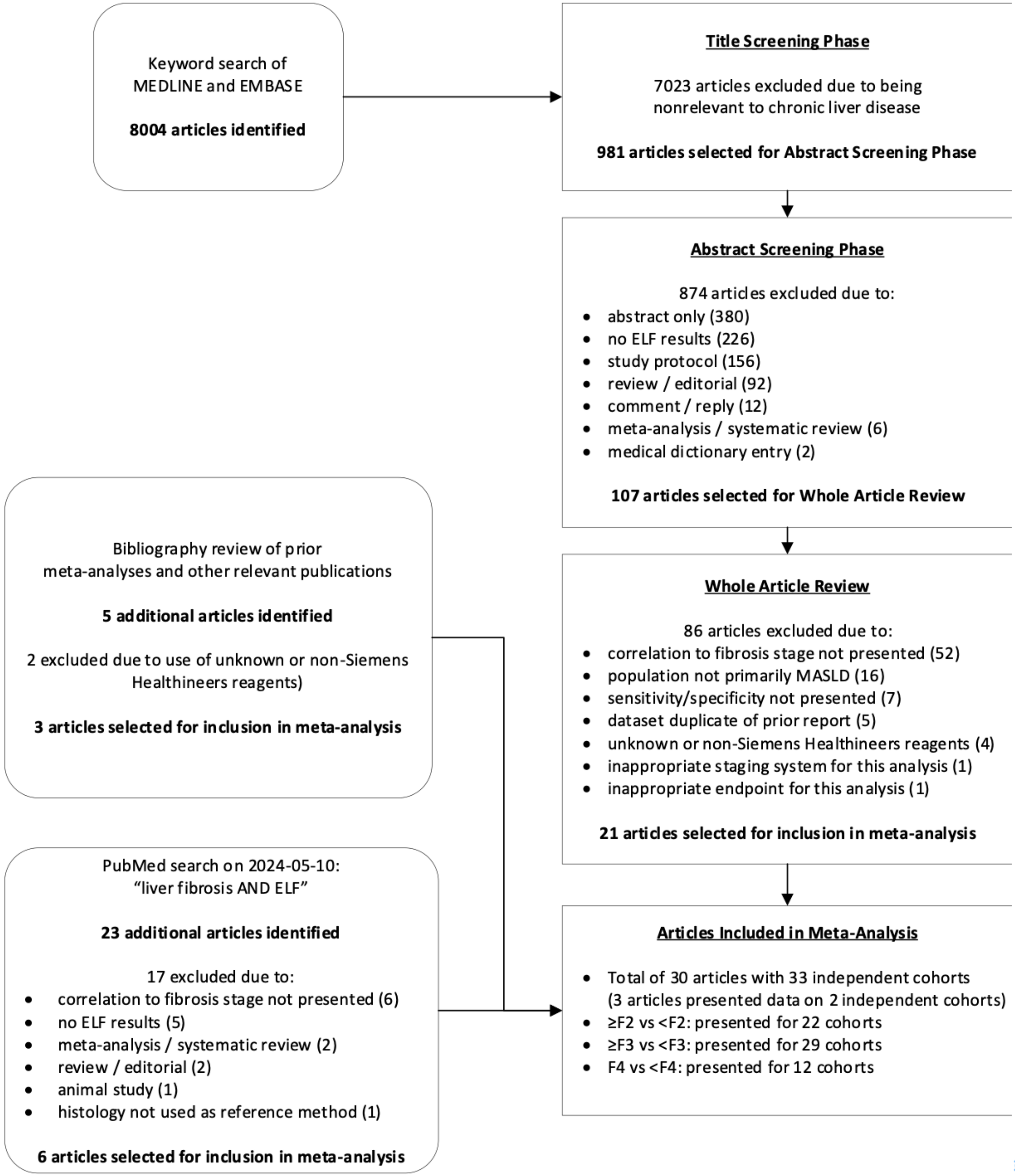
Where did this impact the company:

- Providing background and supporting evidence to support regulatory submissions
- Scientific advancement through publication submission

- **Scope and Problem:**
  - Develop an industry best practice of scientific documentation demonstrating the performance of ELF etc. through the use cutting edge statistical techniques in meta-analysis
  - Meta-Analysis of diagnostic studies difficulties: only studies using the same thresholdss could be used…up to now…using this technique we can combine studies with differing thresholds and still use that information for inclusion

Study Aim & Methods

- Aim: evaluate ELF accuracy for detecting >F2, >F3, F4 and validate optimal cutoffs
- Methods
  - Systematic review & meta-analysis of 30 studies (33 cohorts)
  - Reference: biopsy-based fibrosis staging
  - Analyzed across thresholds using advanced statistical models



Keyword search of MEDLINE and EMBASE

8004 articles identified

Title Screening Phase

7023 articles excluded due to being nonrelevant to chronic liver disease

981 articles selected for Abstract Screening Phase

Abstract Screening Phase

874 articles excluded due to:
- abstract only (380)
- no ELF results (226)
- study protocol (156)
- review / editorial (92)
- comment / reply (12)
- meta-analysis / systematic review (6)
- medical dictionary entry (2)

107 articles selected for Whole Article Review

Bibliography review of prior meta-analyses and other relevant publications

5 additional articles identified

2 excluded due to use of unknown or non-Siemens Healthineers reagents)

3 articles selected for inclusion in meta-analysis

Whole Article Review

86 articles excluded due to:
- correlation to fibrosis stage not presented (52)
- population not primarily MASLD (16)
- sensitivity/specificity not presented (7)
- dataset duplicate of prior report (5)
- unknown or non-Siemens Healthineers reagents (4)
- inappropriate staging system for this analysis (1)
- inappropriate endpoint for this analysis (1)

21 articles selected for inclusion in meta-analysis

PubMed search on 2024-05-10: "liver fibrosis AND ELF"

23 additional articles identified

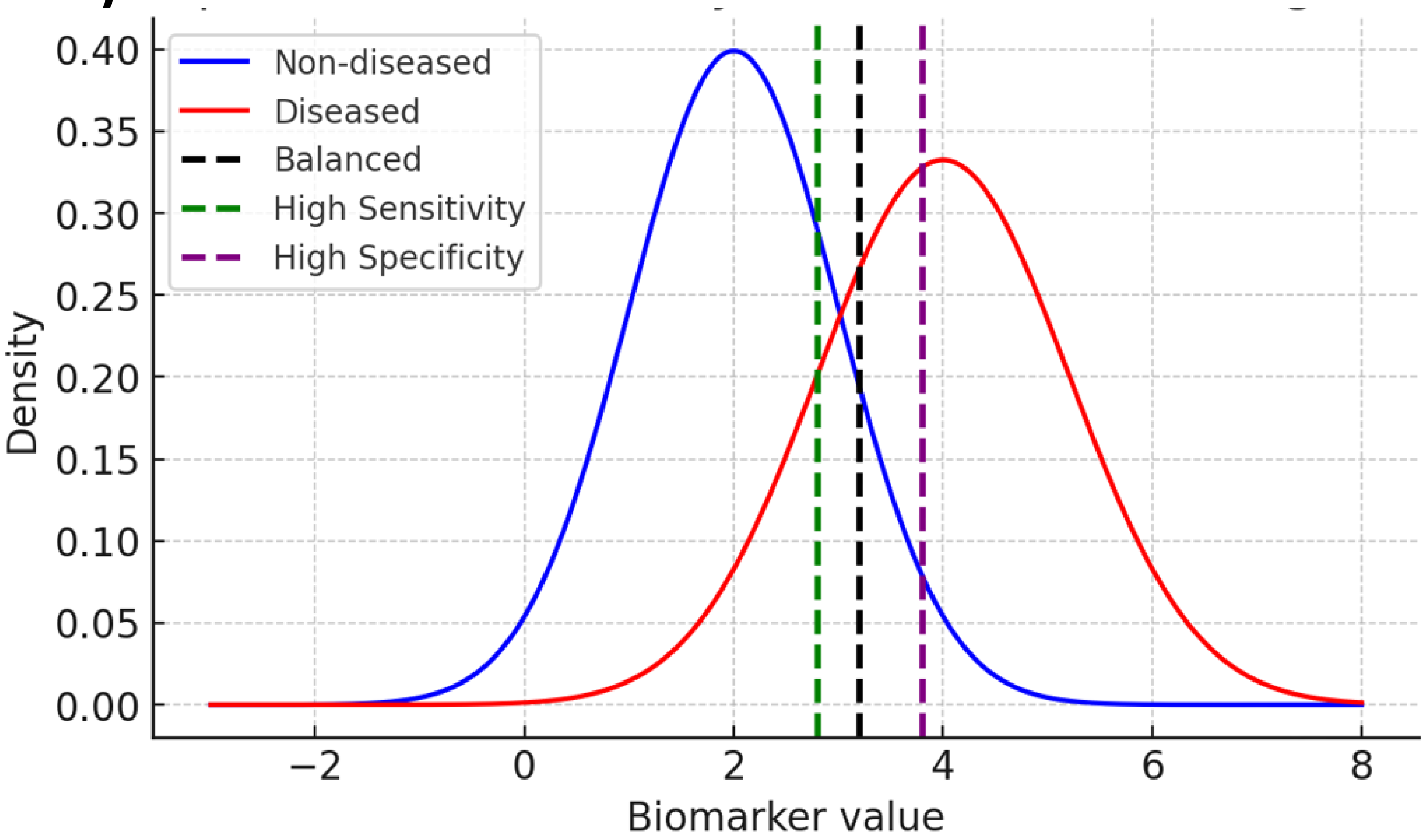
17 excluded due to:
- correlation to fibrosis stage not presented (6)
- no ELF results (5)
- meta-analysis / systematic review (2)
- review / editorial (2)
- animal study (1)
- histology not used as reference method (1)

6 articles selected for inclusion in meta-analysis

Articles Included in Meta-Analysis

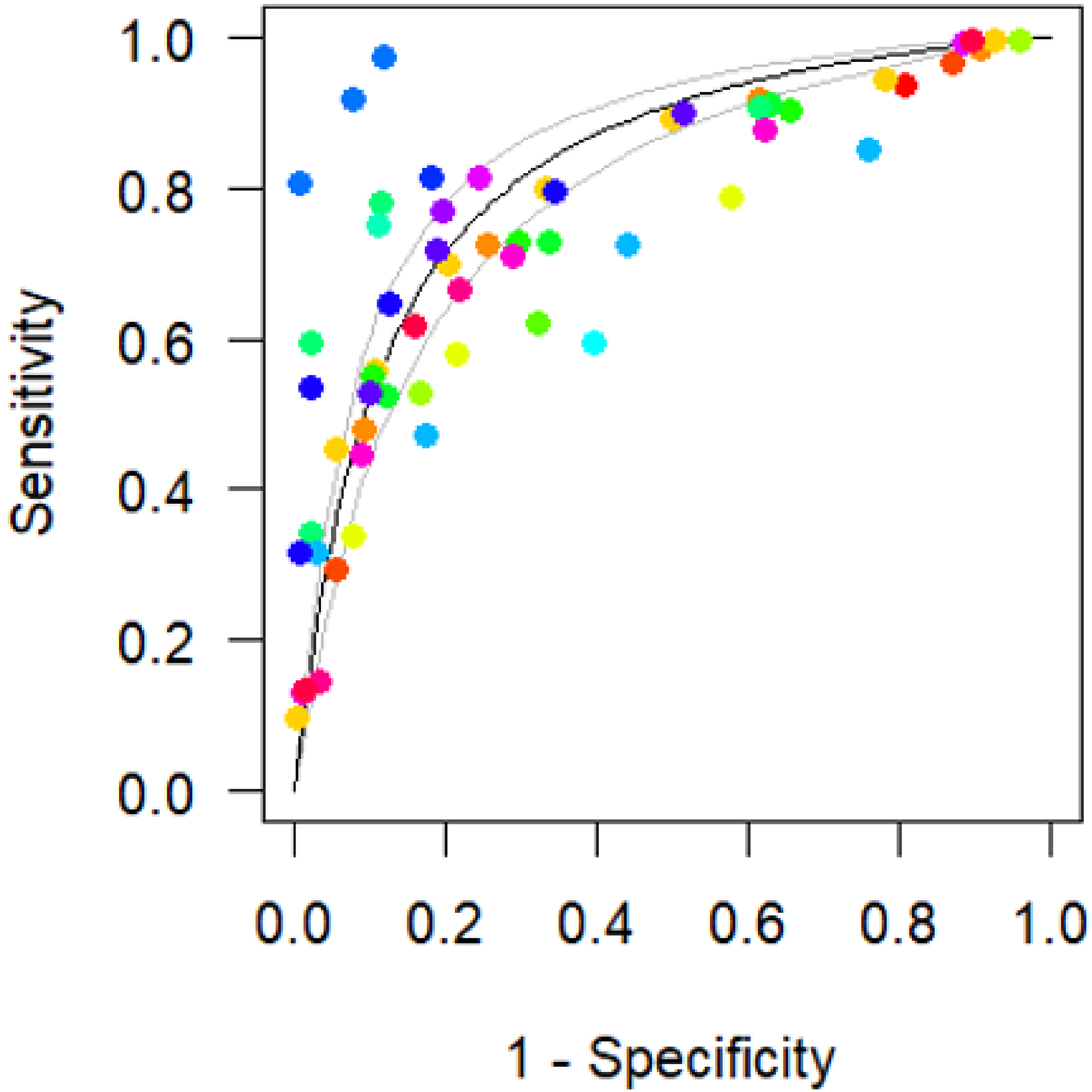
- Total of 30 articles with 33 independent cohorts (3 articles presented data on 2 independent cohorts)
- ≥F2 vs <F2: presented for 22 cohorts
- ≥F3 vs <F3: presented for 29 cohorts
- F4 vs <F4: presented for 12 cohorts

Optimal Thresholds By Youden Index (λ Weights)



Diagnostic Accuracy

- Results across all cohorts
  - >F2: AUC 0.827 (good)
  - >F3: AUC 0.829 (good)
  - F4: AUC 0.773 (fair-good)
- ELF provides strong discrimination across fibrosis stages



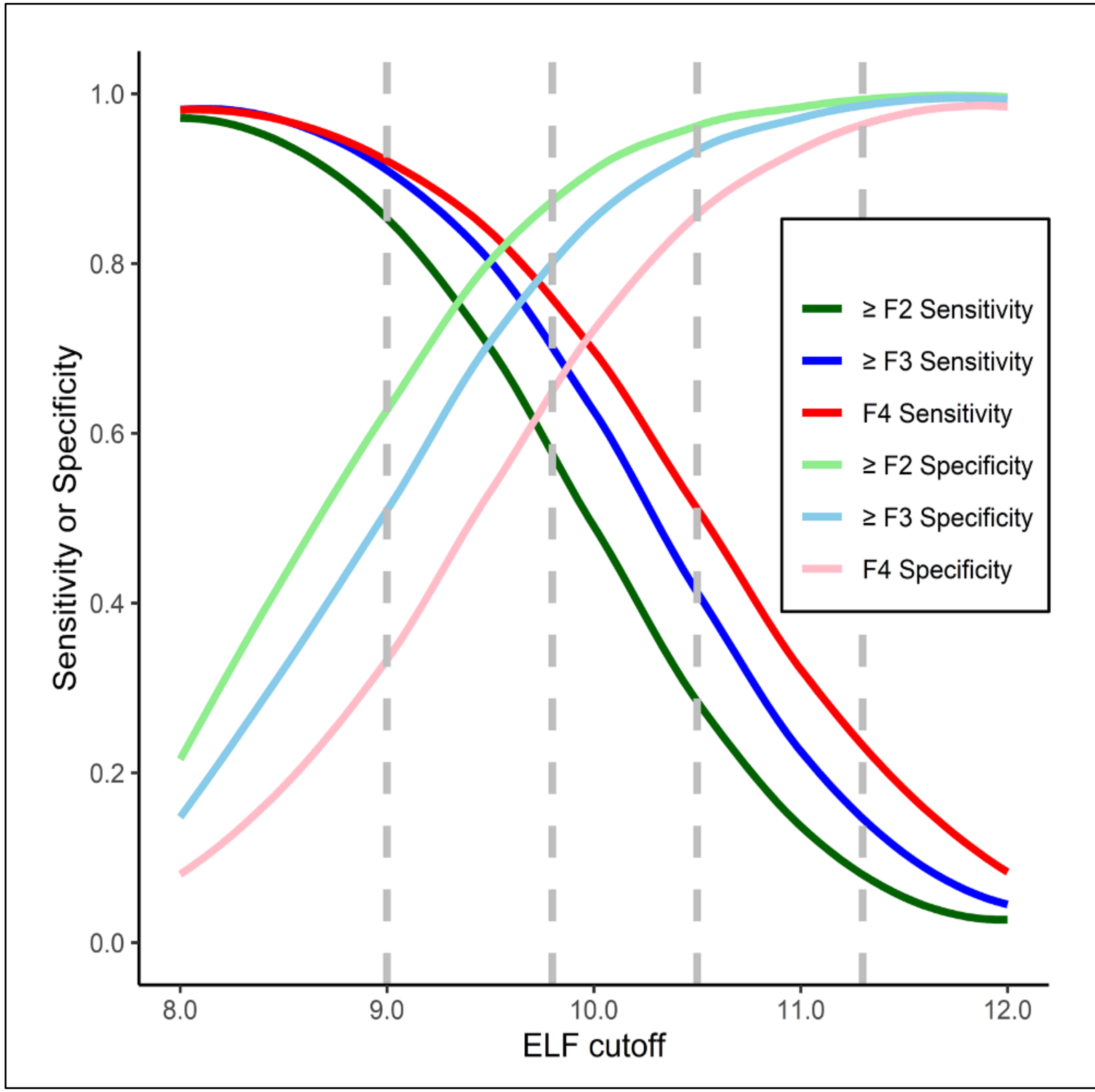
R package – *diagmeta* (Steinhauser et al. 2016)

Study	Cutoff	TP	FP	TN	FN
1	7.0				
1	8.5				
1	10.5				
2	7.5				
2	10.0				
3	9.5				
4	7.5				
4	8.0				
4	9.0				
4	11.0				
…	…	…	…	…	…

```
diag4F2 <- diagmeta(data = data01,
       TP, FP, TN, FN,
       Cutoff,
       Study,
       distr = 'logistic',
       log.cutoff = FALSE,
       method.weights = 'size',
       incr = 0.5, model = "CICS")
```

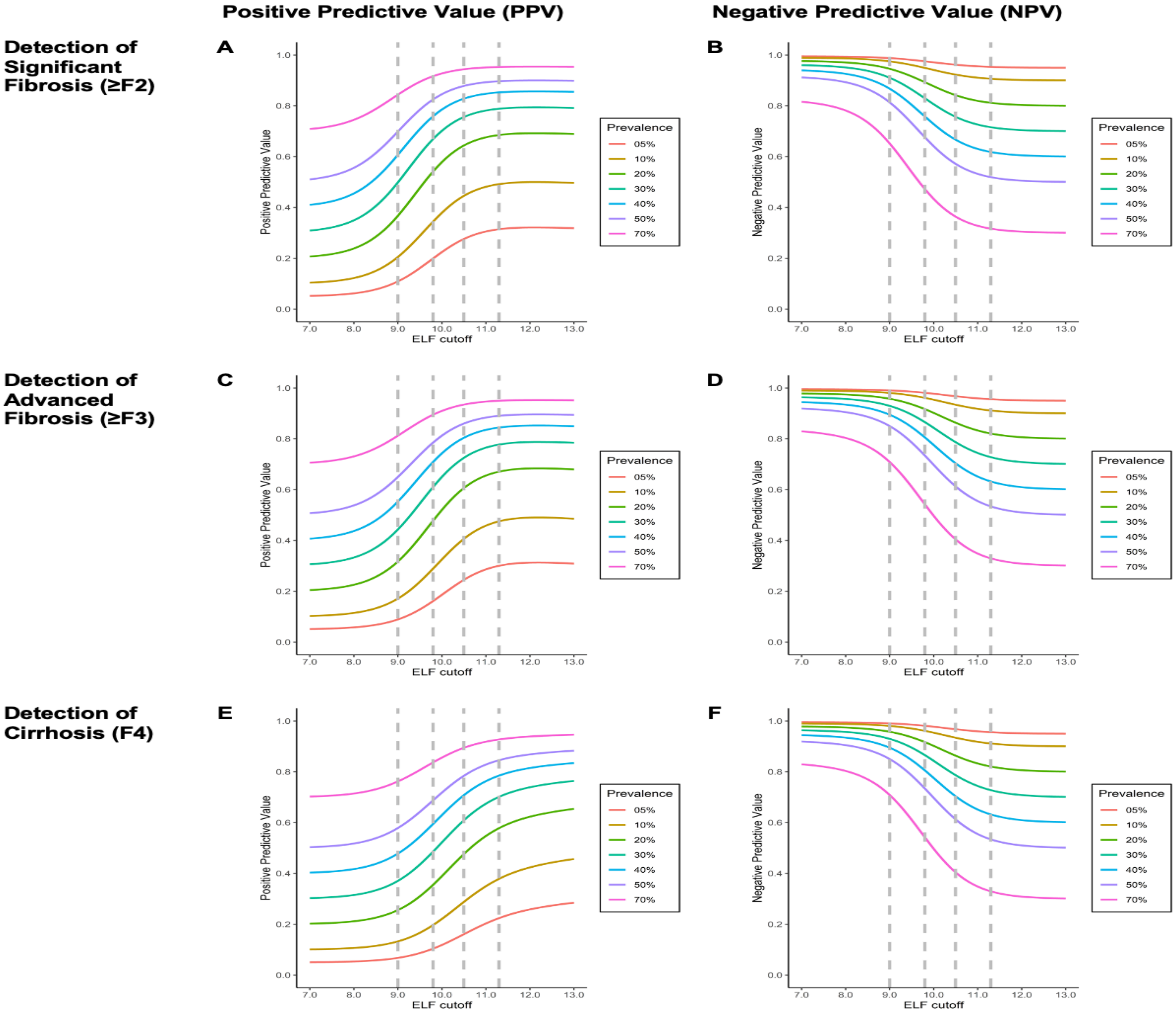
Cutoffs & Performance

- Key thresholds (dashed lines):
  - >9.0: high sensitivity, good for ruling out significant fibrosis (>F2)
  - >9.8: balanced cutoff for advanced fibrosis (71% sensitivity, 81% specificity) (>F3)
  - >10.5: More sensitive cirrhosis detection
  - >11.3: High specificity "rule-in" cutoff for cirrhosis (>F4)



Clinical Utility

- “Rule-out”:
  - ELF <9.0 -> unlikely to have significant fibrosis
- “Balanced”:
  - ELF≥9.8 -> high chance of advanced fibrosis
- “Rule-in”:
  - ELF≥11.3 -> strong chance of cirrhosis



Stage comparison	AUC (95% CI)	Cutoff	Sensitivity (95% CI)	Specificity (95% CI)	Prevalence	Positive Predictive	Negative Predictive
GE F2	0.83 (0.79, 0.86)	9.00	0.86 (0.80, 0.90)	0.63 (0.52, 0.73)	0.20	0.37	0.95
		9.34*	0.76 (0.69, 0.82)	0.76 (0.67, 0.83)	0.20	0.44	0.93
		9.80	0.58 (0.50, 0.66)	0.88 (0.82, 0.92)	0.20	0.54	0.89
		10.50	0.28 (0.22, 0.35)	0.96 (0.94, 0.97)	0.20	0.64	0.84
		11.30	0.08 (0.06, 0.11)	0.99 (0.99, 1.00)	0.20	0.69	0.81
GE F3	0.82 (,)	9.00	…	…	…	…	…
		9.63*					
		9.80					
		10.50					
		11.30					
F4	0.773 (,)	9.00	…	…	…	…	…
		9.80					
		10.00*					
		10.50					
		11.30					

Conclusions & Clinical Takeaways

- The use of specialized techniques provide greater robustness to determining practical thresholds for clinical use.
- With the larger availability of published and unpublished data, meta-analysis should be a readily available tool for both study planning and summary understanding.
- Not only are aggregate studies useable but often subject level meta-analyses are now more readily conducted.

References

- Steinhauser, S., Schumacher, M., Rucker, G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC Med Res Methodol, 16(1):97, 2016.
- Chevret S (2019). ‘diagmeta: Meta-Analysis of Diagnostic Test Accuracy Studies.’ R package version 1.5.0. https://CRAN.R-project.org/package=diagmeta.

Acknowledgements and Recognition

Submitted for publication 10 October

Special thanks to:

- Colleagues at Siemens Healthineers:
  - Chadwick Brown, Matt Gee, Toana Kawashima, Don Chalfin, Ross Molinaro, Roma Levy
- The BASS Planning Committee for accepting our talk

Abstract

- Designing pivotal diagnostic accuracy studies for noninvasive tests (NITs) presents significant challenges, particularly when preparing for regulatory approval. While sensitivity and specificity are standard metrics, estimating their variability is difficult—especially for continuously-scaled NITs where test cutoffs vary widely across studies. This heterogeneity complicates efforts to synthesize data for study planning. The Steinhauser meta-analytic technique, which reconstructs pseudo-individual data from published summary statistics, offers a promising method to estimate pooled performance metrics while accounting for variable cutoffs. This talk develops and demonstrates a ten-step workflow for applying the Steinhauser technique to support the design of pivotal studies for FDA submission. Using the Enhanced Liver Fibrosis (ELF) test as a case study, data from ~30 studies were analyzed to estimate sensitivity, specificity, and associated variance components across unified cutoffs. The process informed five candidate study designs, addressing differing regulatory performance targets. Results yielded stable variance estimates suitable for powering and confidence interval planning to assist in future study design. This structured approach provides NIT developers with a practical tool to overcome common design challenges and supports the growing acceptance of specialized meta-analytic methods by regulatory agencies. This approach will also be compared to more standard approaches such as likelihood and Bayesian methods to assess the generalizable utility of this technique.