Challenges & Opportunities in Clinical Prediction Modeling

Frank E Harrell Jr

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, Tennessee

Office of Biostatistics
US FDA Center for Drug Evaluation and Research

MEMTAB 2018 Utrecht NL 2018-07-02

How Did We Get Here?

- Statistical foundations: maximum likelihood (Fisher), and Bayes
- Long tradition of methodology development in statistics and clinical epidemiology
- Thousands of methodologists
- Statistical computing platforms
- Resampling methods for model validation

Where are We?

- Flexible statistical models
  - Assume smoothness, not linearity (splines, etc.)
  - Penalized maximum likelihood estimation (shrinkage)
  - Bayesian model, penalizing through prior distributions
  - Semiparametric models for continuous ordinal Y
- Overall modeling strategies
  - Handling complexity
  - Data reduction
  - Missing data, e.g. multiple imputation

Where are We? *continued*

- Validation methods
  - Bootstrap and other resampling methods
  - Less need for external validation
  - Validation of predictive discrimination and absolute accuracy (calibration)
- Machine leaning, if black box OK
- Huge number of methods for assessing added value of biomarkers

Challenges

- Role of machine learning, and dealing with hype
- Interpreting complex models (. . . and machine learning algorithms)
- Frequentist statistical inference if using penalization
- Move more to Bayesian models
  - No point estimate of risk but a per-subject risk distribution (pointed if N large) taking all uncertainties into account
  - No overfitting, just disagreements about priors for regression coefficients
  - Handling of missing data much less ad hoc

Challenges: Interactions

- Exploratory analysis of interaction largely fails
- Interactions are frequently nonlinear and co-linear
- Curse of dimensionality and difficulty in pre-specification
- Need new approaches; focus on "interaction data reduction" and Bayes
  - Skeptical priors for interactions effects
  - Stop making dichotomous decisions
  - Interactions can be "half in" the model

Challenges, *continued*

- Methodologists keep inventing ad hoc approaches to quantifying and testing added predictive value
- Many are statistically inefficient
- Many use arbitrary categorization/binning
- Many are unnecessary
- Many indexes have problems
  - Suitable only for retrospective sampling (sensitivity, specificity, ROC curves)
  - Arbitrary and statistically insensitive
  - Improper probability accuracy scoring rules are epidemic

Challenges, *continued*

- Statisticians have forgotten the gold standards:
  - Frequentist: log-likelihood
  - Bayesian: log-likelihood + log prior
  - Explained variation
- Simpler, traditional methods handle greater complexity!
  - Interaction between a biomarker and a baseline clinical variable

Key Measures (Frequentist Versions)

- Log-likelihood; gives rise to
  - Logarithmic proper accuracy score
  - Overall LR model χ^2 (denote by LR)
  - Pseudo R^2: $1 - \exp(-\text{LR}/n)$
- Explained variation
  - Linear model: SSR / SST or $\text{var}(X\hat{\beta})$ / $\text{var}(Y)$
  - Extended by Kent and O'Quigley 1988: SST or $\text{var}(Y)$ is distribution–specific
  - Schemper 2003: excellent paper advocating for measures based on absolute rather than squared differences

Relative Explained Variation

- Base model A, added predictors B
- LR is the gold standard frequentist method for establishing evidence for some added value
- LR is an optimum, general information measure
- LR $= -n \log(1 - R^2)$ (for linear models)
  For small R^2, this is approx. nR^2
- Adequacy index (Harrell 2015): LR_A / LR_{AB}
  Proportion of explainable log likelihood that is explained by A
  Proportion of predictive information

Relative Explained Variation, *continued*

- Relative R^2:
  SSR_A / $\text{SSR}_{AB} = R^2_A$ / R^2_{AB}
  $\text{SSR}_j = \text{var}(X^j \hat{\beta}^j)$
  SSR_A / SSR_{AB} : adequacy of A
  1 - this : proportion of explainable variation explained by B
- Can use other measures than $\text{var}(X\hat{\beta})$
  - mean absolute deviation from mean $X\hat{\beta}$
  - g-index: Gini's mean difference for $X\hat{\beta}$
  - probability scale, for any of the measures

Diagnostic Risk Modeling

Assuming (Atypical) Binary Disease Status

Y	1:diseased, 0:normal
X	vector of subject characteristics (e.g., demographics, risk factors, symptoms)
T	vector of test (biomarker, ...) outputs
α	intercept
β	vector of coefficients of X
γ	vector of coefficients of T

$$\text{pre}(X) = \text{Prob}[Y = 1|X] = \frac{1}{1+\exp[-(\alpha^* + \beta^* X)]}$$

$$\text{post}(X, T) = \text{Prob}[Y = 1|X, T] = \frac{1}{1+\exp[-(\alpha + \beta X + \gamma T)]}$$

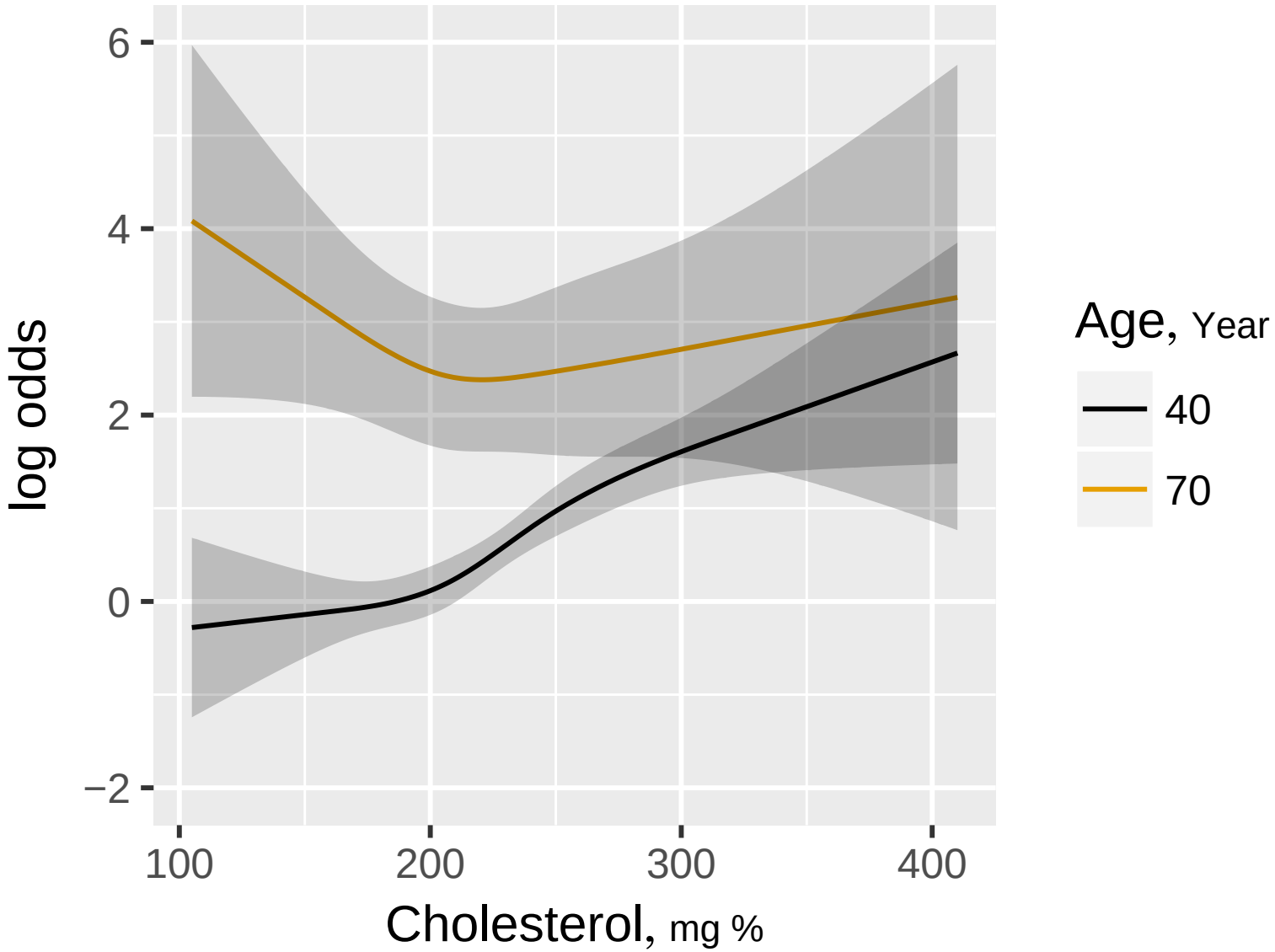
Some Summary Measures for Pre– and Post–test Probabilities

- quantile regression (Koenker and Bassett 1978) curves as a function of pre
- overall mean $|\text{post} - \text{pre}|$
- quantiles of post – pre
- du_{50}: **distribution** of post when pre $= 0.5$
  diagnostic utility at maximum pre-test uncertainty
  - Choose X so that pre $= 0.5$
  - Examine distribution of post at this pre
  - Summarize with quantiles, Gini's mean difference on prob. scale
  - Special case where test is binary (atypical): compute post for T^+ and for T^-

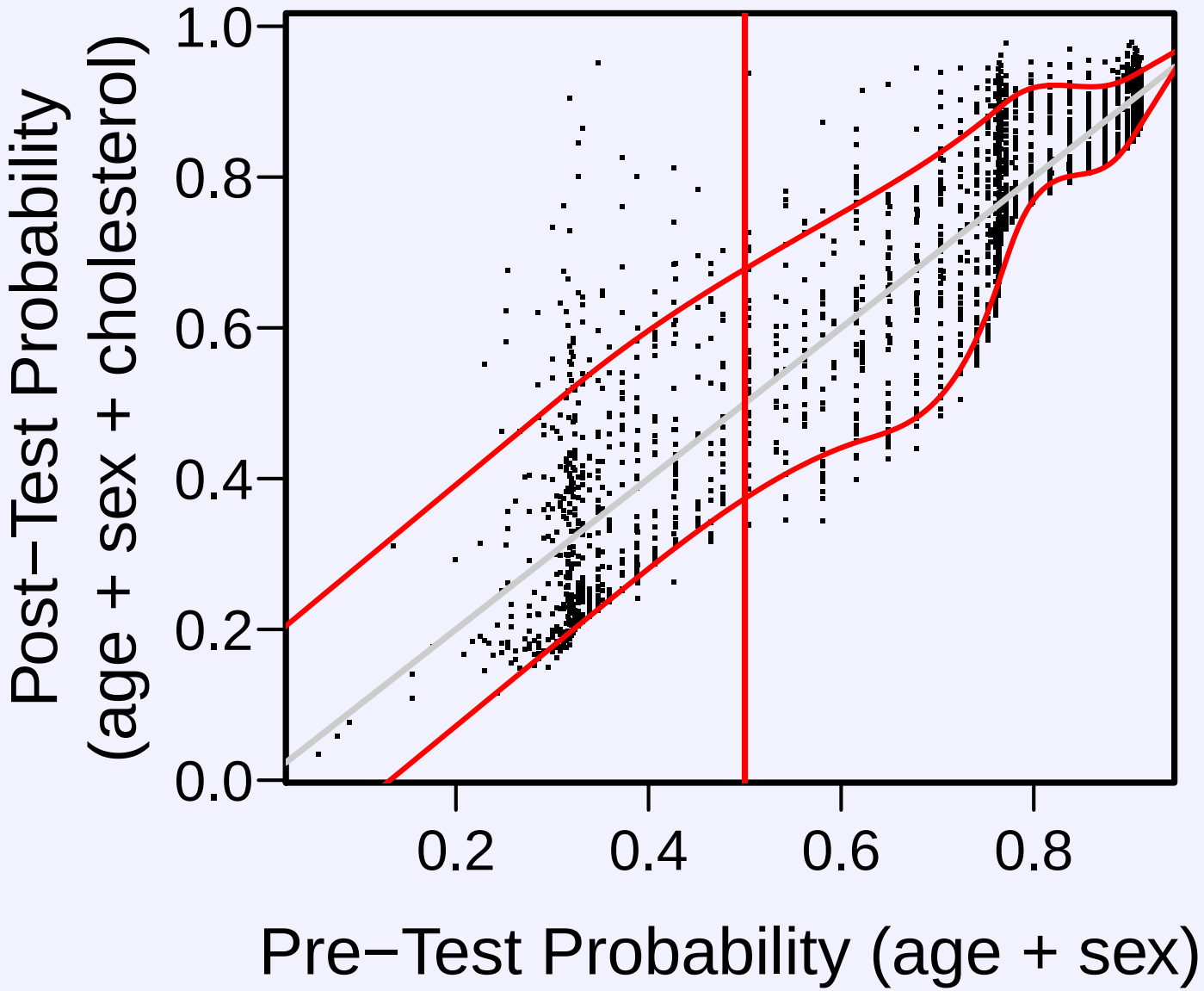
Case Study

- Patients undergoing cardiac catheterization at Duke University, for chest pain; $n = 2258$
- Diagnosis of significant coronary artery disease
- See BBR Diagnosis Chapter: `fharrell.com/links`
- Base model: age, sex; age and age $\times$ sex interactions nonlinear using splines
- New biomarker: total cholesterol
- Cholesterol interacts nonlinearly with age

Effect of Cholesterol at Two Example Ages



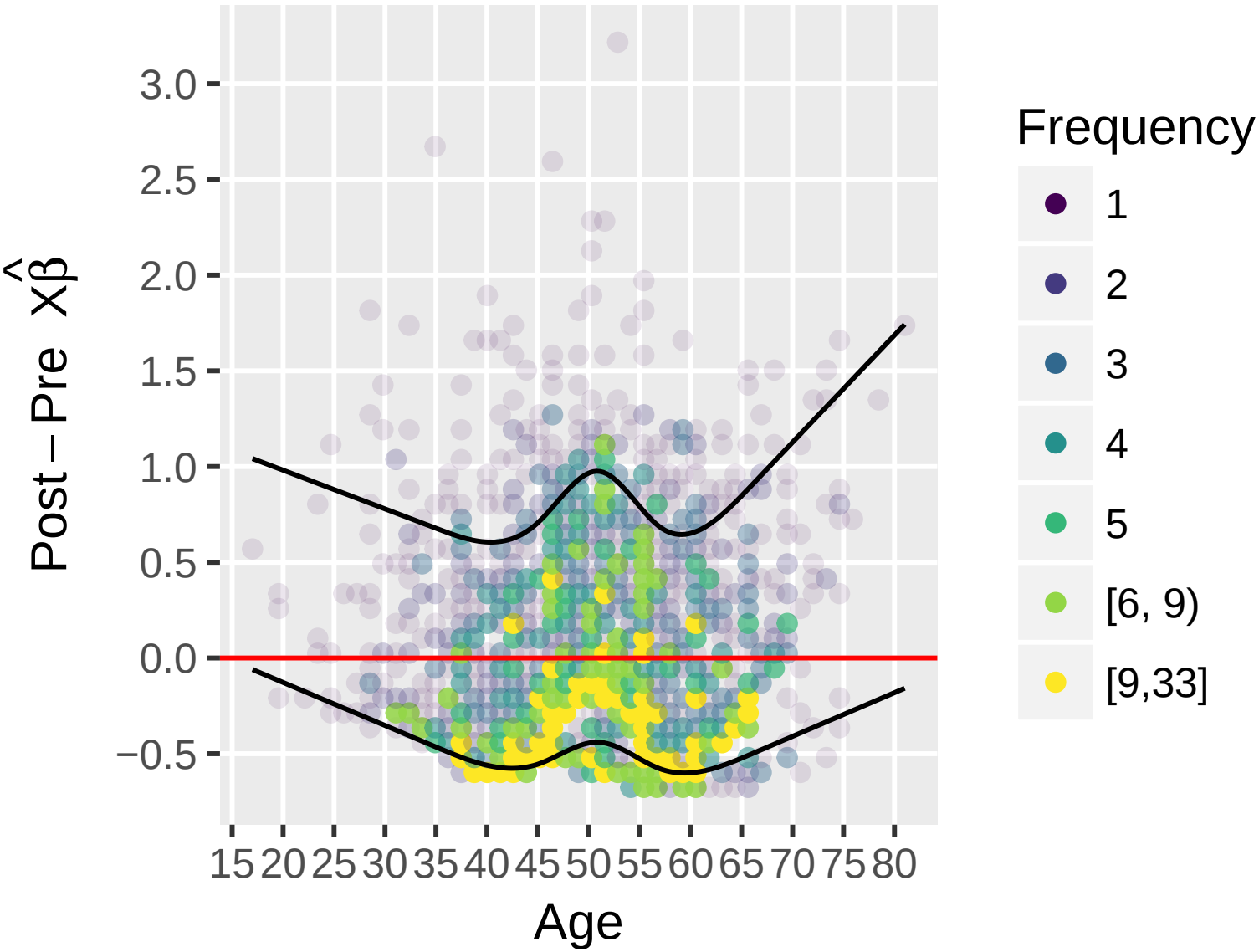
Diagnostic Utility of Cholesterol
Quantile Regression, 0.1 and 0.9 Quantiles



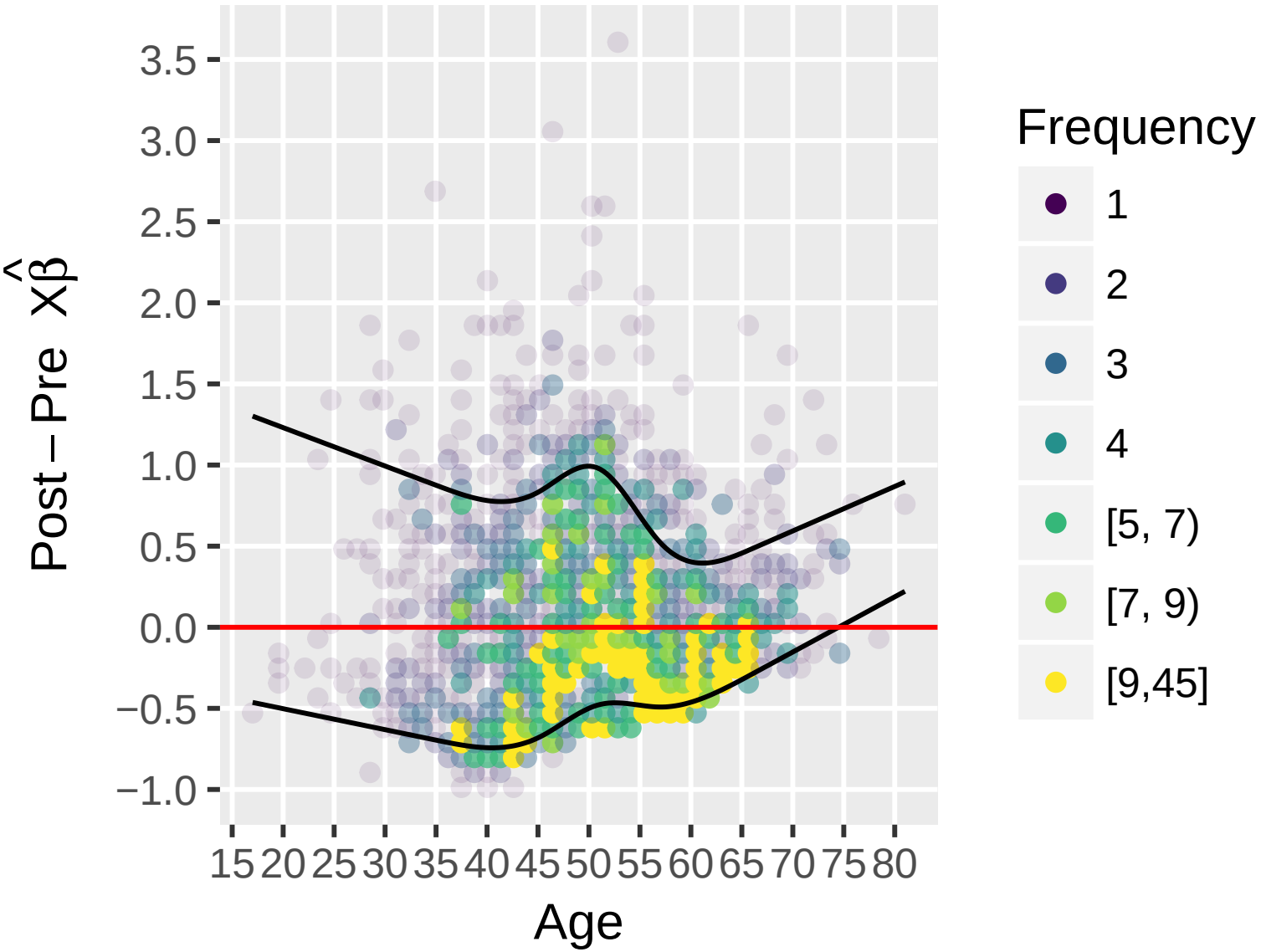
Quantifying Explained Variation

	Base	Base+Chol
LR χ^2	496.85	596.99
c	0.77	0.79
R^2	0.27	0.32
Brier	0.18	0.17
g_p	0.24	0.27
Adequacy	0.83	1.00
var$(X\hat{\beta})$	1.18	1.51
Relative $R^2(X\hat{\beta})$	0.78	1.00
var$(\hat{P})$	0.05	0.06
Relative $R^2(\hat{P})$	0.84	1.00

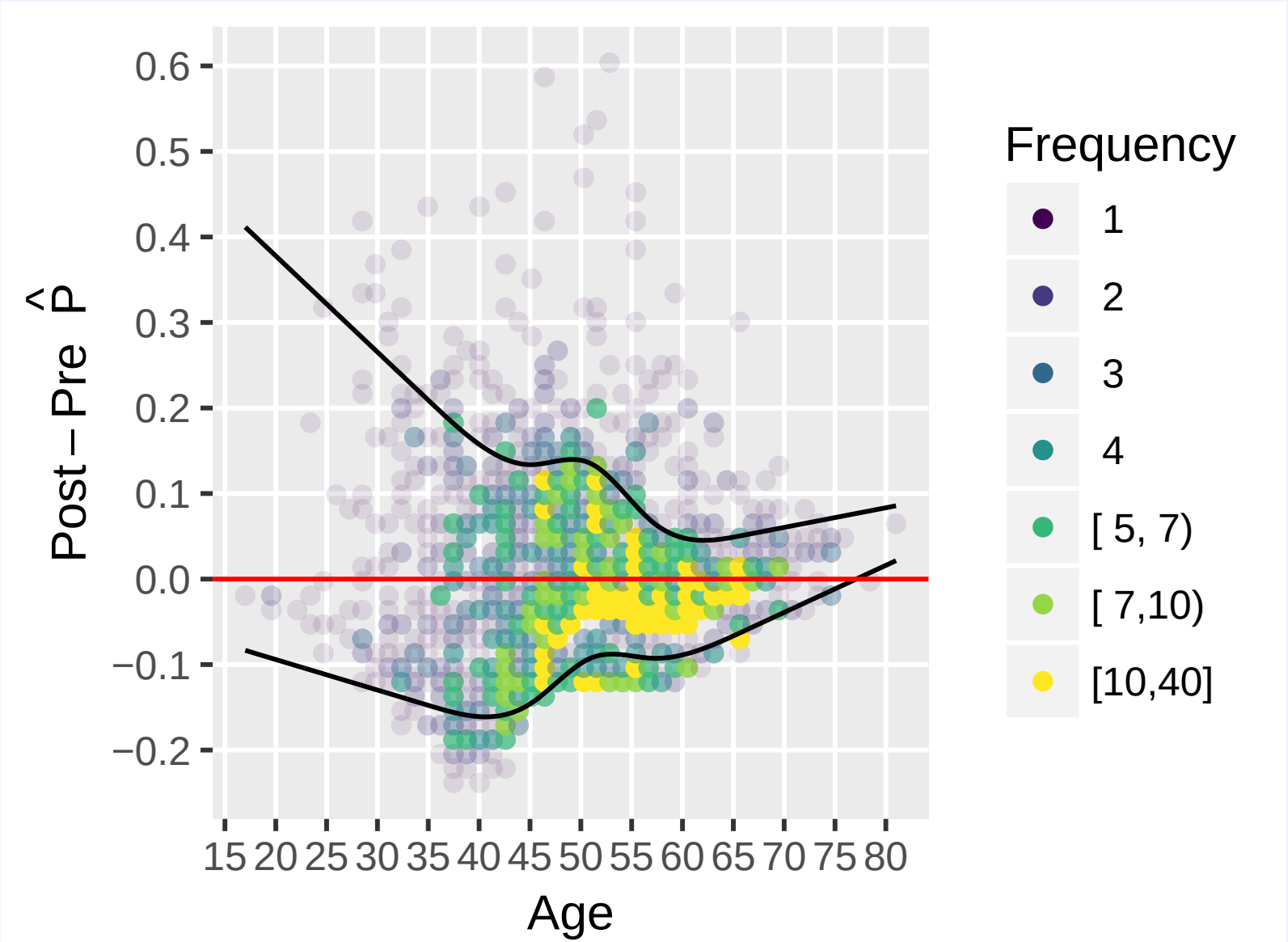
Diagnostic Utility of Cholesterol vs. Age, Logit Scale; No Cholesterol × Age Interaction



Diagnostic Utility vs. Age, Logit Scale
Cholesterol × Age Interaction Included

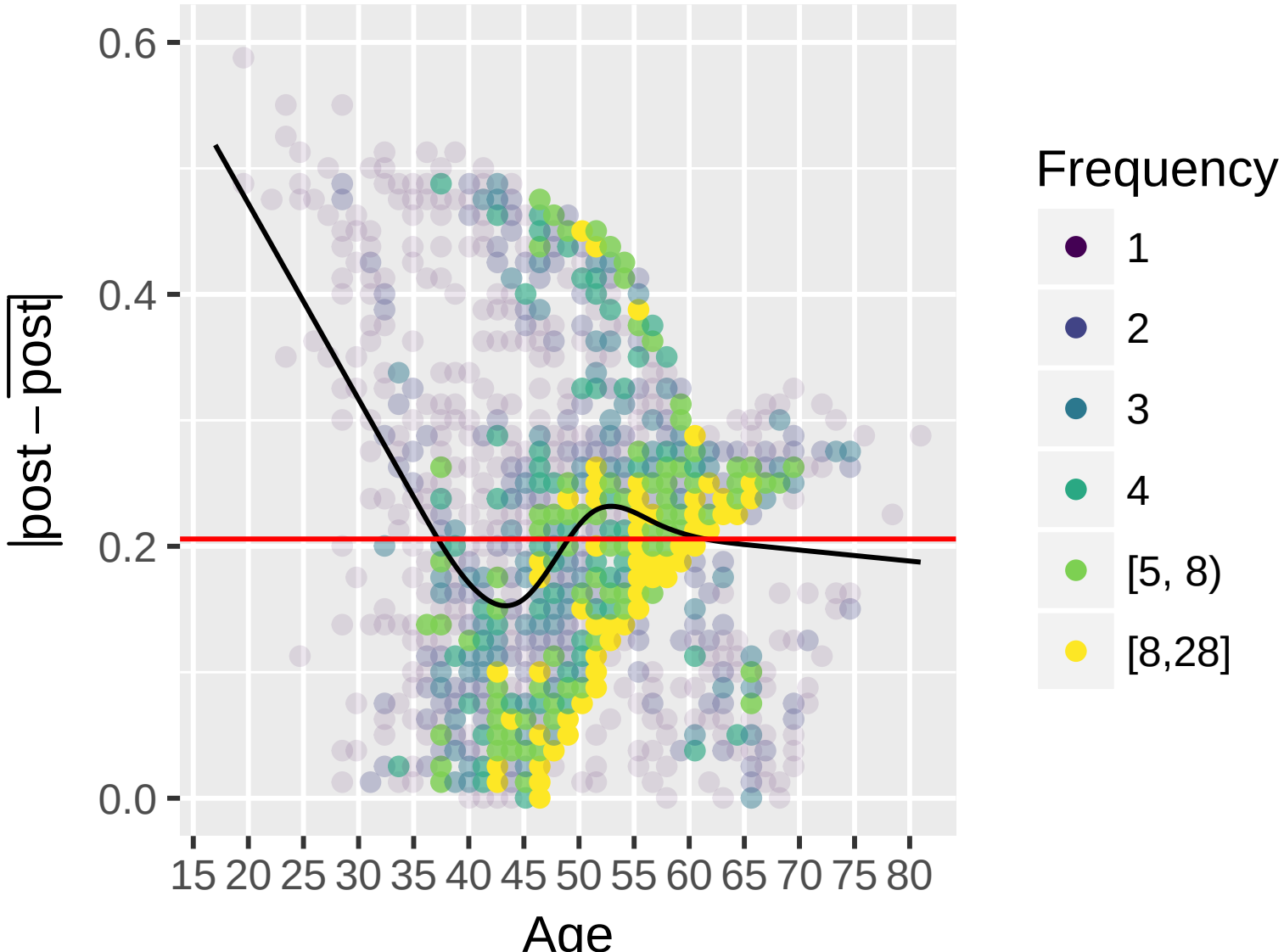


Diagnostic Utility vs. Age, Probability Scale Interaction Included

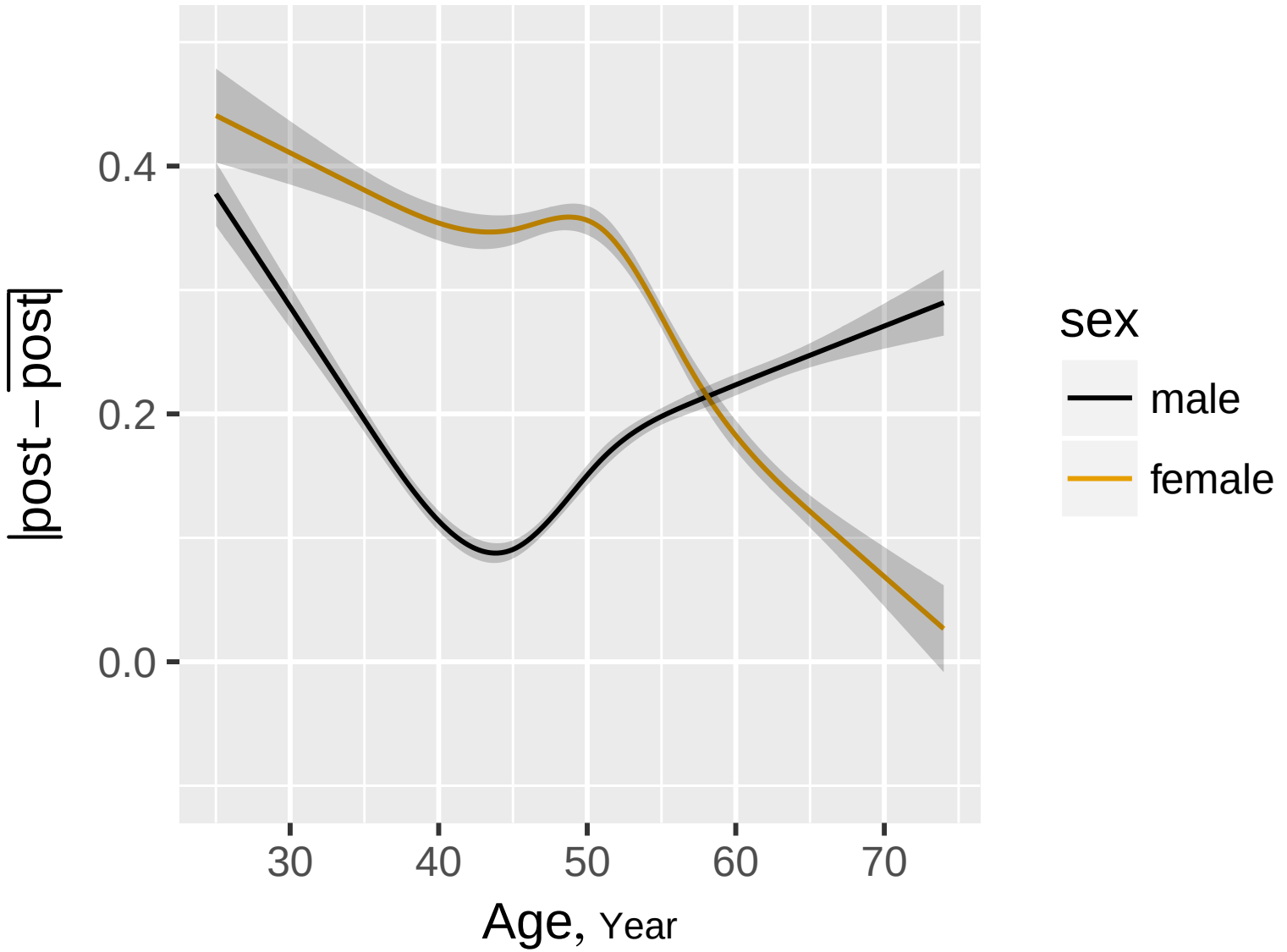


Explained Variation vs. Age, Probability Scale

Absolute difference between $\hat{P}$ and $\overline{P}$ in post-test model



Explained Variation vs. Age and Sex, Probability Scale



Summary

- There are many remaining challenges in clinical prediction model development
- Need general approaches for reliable interaction modeling for precision medicine/HTE
- Bayesian modeling opens vast possibilities
- Need to unlearn a lot of ad hoc methods for assessing added value of biomarkers
- Simple regression and likelihood approaches are
  - more powerful
  - more precise
  - less arbitrary (no binning)
  - more insightful
  - more flexible
- Need to spend effort translating likelihood and explained variation measures for clinicians

References

Harrell, F. E. (2015). *Regression Modeling Strategies, with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Second edition. New York: Springer. isbn: 978-3-319-19424-0 (cit. on p. 10).

Kent, J. T. and J. O'Quigley (1988). "Measures of dependence for censored survival data". In: *Biometrika* 75, pp. 525–534 (cit. on p. 9).

Koenker, R. and G. Bassett (1978). "Regression quantiles". In: *Econometrica* 46, pp. 33–50 (cit. on p. 13).

Schemper, M. (2003). "Predictive accuracy and explained variation". In: *Stat Med* 22, pp. 2299–2308 (cit. on p. 9).

value of R^2 with binary response data;measures of average absolute prediction errors with continuous response

.