Incremental value

Case study 2: Utility of IGRAs in Active TB Diagnosis in a Low Incidence Setting

John Metcalfe, MD, MPH
Assistant Professor, Division of Pulmonary and Critical Care
San Francisco General Hospital
PhD candidate, Division of Epidemiology
University of California, Berkeley

Overview

- Rationale
- Analytic Methods
- Conclusions

Limitations of Diagnostic Accuracy

- Sensitivity: probability of a TP among cases
- Specificity: probability of a TN among non-cases
  - Do not quantify or control for covariate effects
  - Often do not demonstrate worth of a test in a particular patient population
  - Often do not easily translate to clinical practice

Epidemiologic Studies
Three Broad Categories

- Association
  - "Carrying matches is associated with lung cancer."
- Causal Inference
  - "Smoking causes lung cancer."
- Prediction
  - "The 5-year survival for individual *i* with lung cancer is 40%."

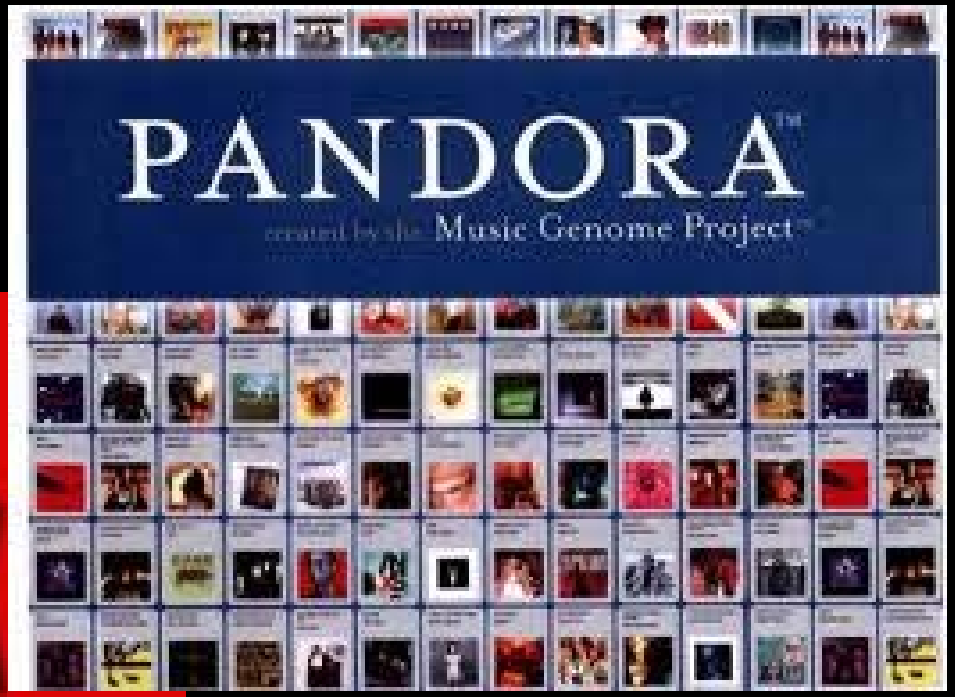
Prediction: Essential Features

- Unconcerned with causality
- Use of a statistical model $f(X)$ to estimate individual risk
- Involve some assessment of *prediction error*
- Valid insofar as your patient is similar to the patients from whom the prediction model was derived

Prediction Models: Uses

- Risk Stratification of Patients
  - Prognostication, therapeutic strategies

NETFLIX



match.com

Prediction Models: Uses

- Incremental Value
  - Extent to which a diagnostic or prognostic test improves on readily available clinical information

Overview

- Rationale
- Analytic Methods
  - Formulate Research Question
  - Model Selection
  - Model Assessment
- Conclusions

Overview

- Rationale
- Analytic Methods
  - **Formulate Research Question**
  - Model Selection
  - Model Assessment
- Conclusions

Research Question

- “Does QFT-G (as a quantitative measure) have added value beyond readily available clinical information for active TB diagnosis?”

- Three year interval
- 660 consecutive TB suspects
- 65 culture-confirmed TB cases

Evaluation of Quantitative IFN-γ Response for Risk Stratification of Active Tuberculosis Suspects

John Z. Metcalfe[1], Adithya Cattamanchi[1], Eric Vittinghoff[2], Christine Ho[3,4], Jennifer Grinsdale[3], Philip C. Hopewell[1,3], L. Masae Kawamura[3], and Payam Nahid[1,3]

[1]Division of Pulmonary and Critical Care Medicine, San Francisco General Hospital, and [2]Department of Epidemiology and Biostatistics, University of California, San Francisco; [3]Tuberculosis Control Section, Department of Public Health; and [4]Centers for Disease Control and Prevention, San Francisco, California

***Rationale*: The contribution of interferon-γ release assays (IGRAs) to appropriate risk stratification of active tuberculosis suspects has not been studied.**

***Objectives*: To determine whether the addition of quantitative IGRA results to a prediction model incorporating clinical criteria improves risk stratification of smear-negative–tuberculosis suspects.**

***Methods*: Clinical data from tuberculosis suspects evaluated by the San Francisco Department of Public Health Tuberculosis Control Clinic from March 2005 to February 2008 were reviewed. We excluded tuberculosis suspects who were acid fast–bacilli smear–positive, HIV-infected, or under 10 years of age. We developed a clinical prediction model for culture-positive disease and examined the benefit of adding quantitative interferon (IFN)-γ results measured by QuantiFERON-TB Gold (Cellestis, Carnegie, Australia).**

***Measurements and Main Results*: Of 660 patients meeting eligibility criteria, 65 (10%) had culture-proven tuberculosis. The odds of active tuberculosis increased by 7% (95% confidence interval [CI], 3–11%) for each doubling of IFN-γ level. The addition of quantitative IFN-γ results to objective clinical data significantly improved model performance (c-statistic 0.71 vs. 0.78; $P < 0.001$) and correctly reclassified 32% of tuberculosis suspects (95% CI,11–52%; $P < 0.001$) into higher-risk or lower-risk categories. However, quantitative IFN-γ results did not significantly improve appropriate risk reclassification beyond that provided by clinician assessment of risk (4%; 95% CI, −7 to +22%; $P = 0.14$).**

***Conclusions*: Higher quantitative IFN-γ results were associated with active tuberculosis, and added clinical value to a prediction model incorporating conventional risk factors. Although this benefit may be attenuated within highly experienced centers, the predictive accuracy of quantitative IFN-γ levels should be evaluated in other settings.**

Keywords: interferon-gamma release assay; Quantiferon; risk prediction; risk reclassification

AT A GLANCE COMMENTARY

Scientific Knowledge on the Subject

The role of interferon-γ release assays (IGRAs) in the evaluation of active tuberculosis suspects is controversial. To date, whether IGRAs improve classification of smear negative tuberculosis suspects into clinically relevant risk categories has not been examined.

What This Study Adds to the Field

Quantitative interferon-γ levels measured by QuantiFERON-TB Gold improves risk stratification of smear-negative active tuberculosis suspects when added to objective clinical and demographic risk factors. However, this benefit is attenuated when the judgment of experienced clinicians is also considered.

Interferon-γ release assays (IGRAs) are *in vitro* immunodiagnostic tests that measure effector T cell mediated interferon (IFN)-γ response to *Mycobacterium tuberculosis*–specific antigens. IGRAs are as sensitive and more specific than the tuberculin skin test for detecting latent tuberculosis infection (LTBI) (1, 2) and have better correlation with gradient of *M. tuberculosis* exposure (3–8). In 2005, the Centers for Disease Control and Prevention recommended that QuantiFERON TB-Gold (QFT-G; Cellestis, Carnegie, Australia), the first FDA-approved, commercially available IGRA to experience widespread use, could be used for targeted screening of LTBI in all circumstances in which the tuberculin skin test (TST) is used (9).

Although the advantages of IGRAs in diagnosing LTBI are well established, their role in evaluating active tuberculosis suspects remains unclear. IGRAs have variable, although often suboptimal, sensitivity and specificity for diagnosing active tuberculosis (1, 2, 10–16). To date, with the exception of studies examining these assays in parallel with the TST (11, 17), IGRAs have not been considered in light of conventional risk factors for active disease. In addition, whether IGRAs improve prediction of individual patients' risk for active tuberculosis has not been examined.

Acid fast bacilli (AFB) smear-positive tuberculosis suspects can often be triaged with relative ease. However, in suspects whose sputa or other tissue are smear-negative for AFB, clinicians use demographic and clinical risk factors, symptoms, and chest radiograph findings to classify patients into low-, intermediate-, or high-risk categories for active tuberculosis. Patients classified as high risk are typically initiated on antituberculosis therapy, whereas treatment is withheld for low-risk patients. In this study, we use novel risk reclassification methods (18) to assess whether addition of quantitative IFN-γ response measured by QFT-G (Cellestis) to routine clinical evaluation improves risk stratification of individuals suspected of having smear-negative pulmonary and extrapulmonary tuberculosis.

Some of the results of these studies have been previously reported in abstract form (19).

(*Received in original form June 30, 2009; accepted in final form October 1, 2009*)

Supported by grants T32 HL007185 (J.Z.M.), K23 HL094141 (A.C.), R01 AI034238 (P.C.H.), and K23 HL092629 (P.N.) from the National Institutes of Health.

Correspondence and requests for reprints should be addressed to John Z. Metcalfe, M.D., M.P.H., University of California, San Francisco Division of Pulmonary and Critical Care Medicine, 1001 Potrero Avenue, Rm 5K1, San Francisco, CA 94110-0111. E-mail: john.metcalfe@ucsf.edu

This article has an online supplement, which is accessible in this issue's table of contents at www.atsjournals.org

Am J Respir Crit Care Med Vol 181. pp 87–93, 2010
Originally Published in Press as DOI: 10.1164/rccm.200906-0981OC on October 1, 2009
Internet address: www.atsjournals.org

Overview

- Rationale
- Analytic Methods
  - Formulate Research Question
  - **Model Selection**
  - Model Assessment
- Conclusions

Model Building: Variable Selection

- Options:
  - Empiric selection
  - Stepwise variable selection
  - Data adaptive estimation

- A misspecified model =
  - Lost statistical efficiency
  - Uninterpretable coefficients

Logistic regression

CART
Random Forests
Elastic Net
Logic Regression
Ridge Regression
Polynomial Spline Regression
Deletion/Substitution/Addition
Neural Nets
Gradient Boosting
LOESS
Bayesian additive regression trees

- *“Just as the ability to devise simple but evocative models is the signature of the great scientist,*

  so overelaboration and over parametrization is often the mark of mediocrity.”

  -Box 1976

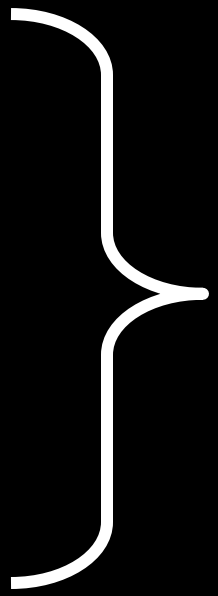
Variable Selection: Deletion/Substitution/Addition (DSA)

- Data Adaptive Estimation
  - Based on minimizing L2 loss function
- Considers non-linear terms and all possible interactions between predictors
- Avoids overfitting through *K*-fold cross validation

Sinisi Stat Appl Genet Mol Biol 2004

Variable Selection

- Age
- Gender
- Race/Ethnicity
- Foreign Birth
- Night sweats
- Weight loss
- Fever
- Hemoptysis
- CXR status
- Contact Status
- Homelessness
- Diabetes mellitus
- Previous Active TB
- Quantitative IFN-g Response



X

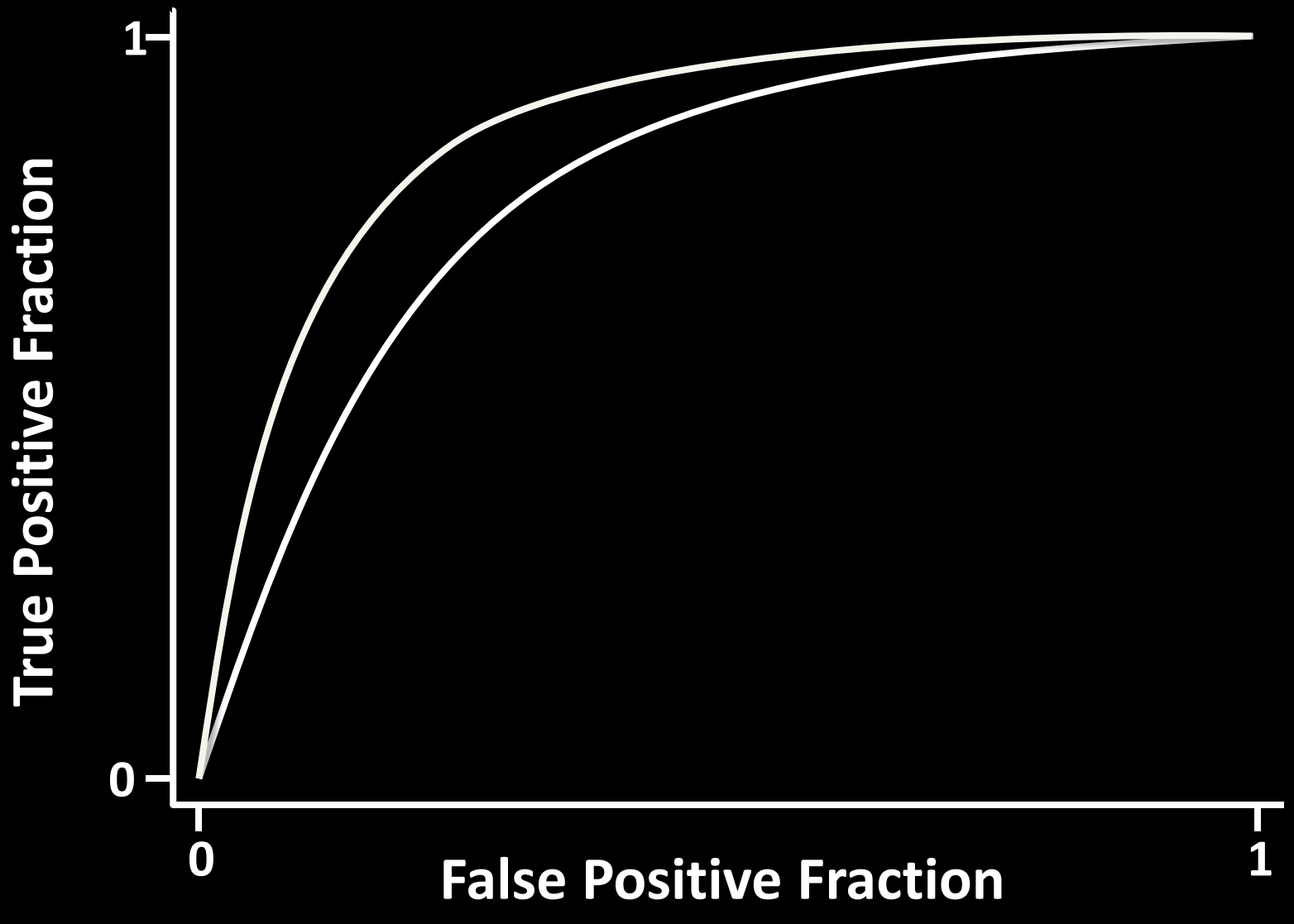
Y

Model Performance

- Calibration: How well do predicted probabilities agree with actual observed risk?

- Discrimination: How well does the model separate out those who do and do not have the disease of interest?

Receiver Operating Characteristic (ROC)



Receiver Operating Characteristic (ROC): What Is It?

- A summary of Se and Sp across a range of cut points for a predictor or model
  - Plot of sensitivity vs. 1-specificity (Range 0.5 to 1)
- Area Under the Curve (AUC) = *the probability that the predicted risk is higher for a case than for a non-case*
  - ***not*** *the probability that individuals are classified correctly (i.e.., not a function of actual predicted probabilities)*

Receiver Operating Characteristic (ROC): Limitations

- Insensitive in model comparison
  - A new variable can influence more accurate risk classification, despite little change in the c statistic
- Clinical Relevance?
  - Scale ≠ absolute risk
  - Patients seldom present in pairs

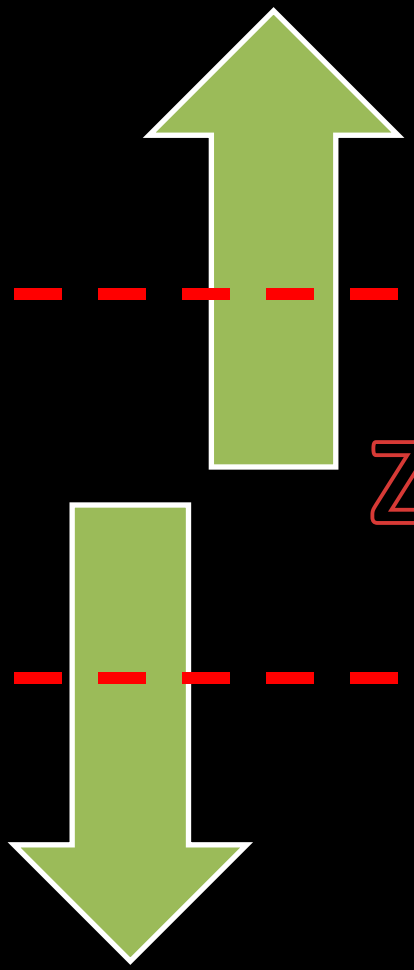
Cook 2007 Circ
Janes 2008 Ann Int Med

Risk Reclassification: To the Rescue?

Essential Question:

Can a new model more accurately stratify individuals into higher or lower risk categories of clinical importance?

Reclassification - Diagnostic Utility



Empiric Treatment

>20 % Risk

Zone of Uncertainty

No Treatment – Await Culture Results

<5 % Risk

Net Reclassification Improvement (NRI) Summary Index

- *Net* improvement in reclassification

GOOD BAD BAD GOOD

$$[P(up|D=1)-P(down|D=1)]-[P(up|D=0)-P(down|D=0)]$$

Correct classification of cases + Correct classification of non-cases

Pencina *Stat Med* 2007

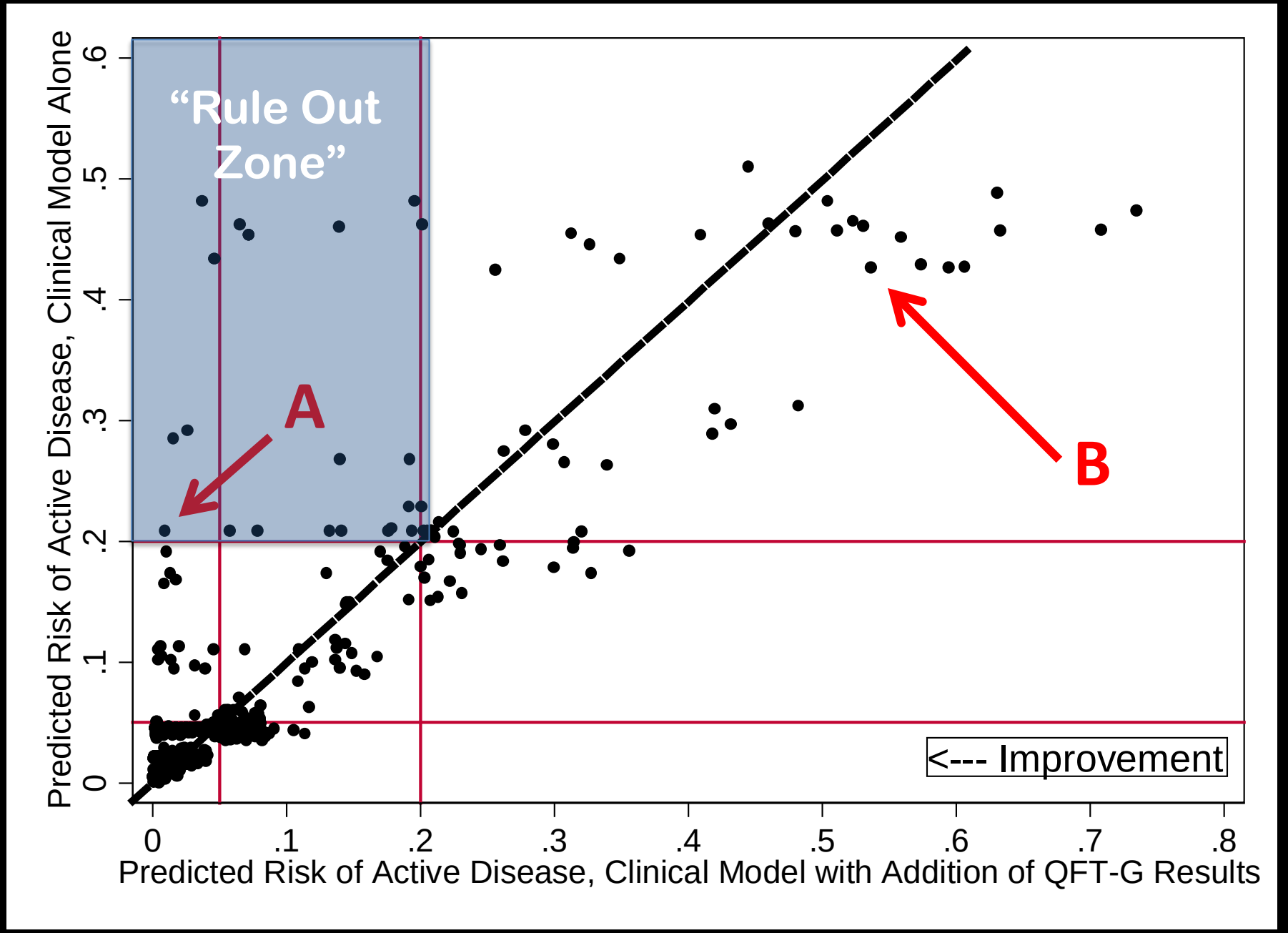
Model with Clinical Predictors Alone	Model with Clinical Predictors and Quantitative IFN-γ Results				Percent Appropriately Reclassified
	≤5% risk	5–20% risk	>20% risk	Total No.	
In 595 patients who ruled out for active tuberculosis					
≤5% risk	334	121	0	455	−27
5–20% risk	20	34	14	68	9
>20% risk	9	18	45	72	38
Total No.	363	173	59	595	

- Overall NRI= 3.7%
- Reclassification among patients who ruled out for active tuberculosis = -14.8%
- Reclassification among patients who developed culture-positive disease = 18.5%

Net Reclassification Improvement (NRI): Limitations

- Risk category thresholds should not be arbitrary
  - Theoretically, should be based on weighing cost versus benefit of incorrect versus correct designation
- Treats all reclassifications in the correct direction as equivalent
  - Rare that FP and FN have similar effect on a patients' survival or quality of life
  - Alternatives: report the components of NRI, or weight reclassifications

Examining Changes in Predicted Risk: Noncases



Overview

- Rationale
- Analytic Methods
  - Formulate Research Question
  - Model Selection
  - **Model Assessment**
- Conclusions

Model Assessment

- *'generalization* performance of a learning method relates to its prediction capability on independent test data'

- However, data often too limited to set aside separate validation and test sets…

Hastie *Elements of Stat Learning* 2009

Model Assessment: *K*-fold Cross-validation

- Data split into *K* roughly equally-sized parts
- Model is fit using the other *K*-1 parts
- Prediction error then calculated using the fitted model to predict the *k*th part

Hastie *Elements of Stat Learning* 2009

Overview

- Rationale
- Analytic Methods
  - Formulate Research Question
  - Model Selection
  - Model Assessment
- Conclusions

Conclusions

- It is possible to go beyond diagnostic accuracy studies!
- Reclassification may have more clinical applicability than AUC

- QFT (as dichotomous or quantitative) is unlikely to improve clinical diagnosis of immunocompetent smear-negative TB suspects
  - Though may have value in subgroups