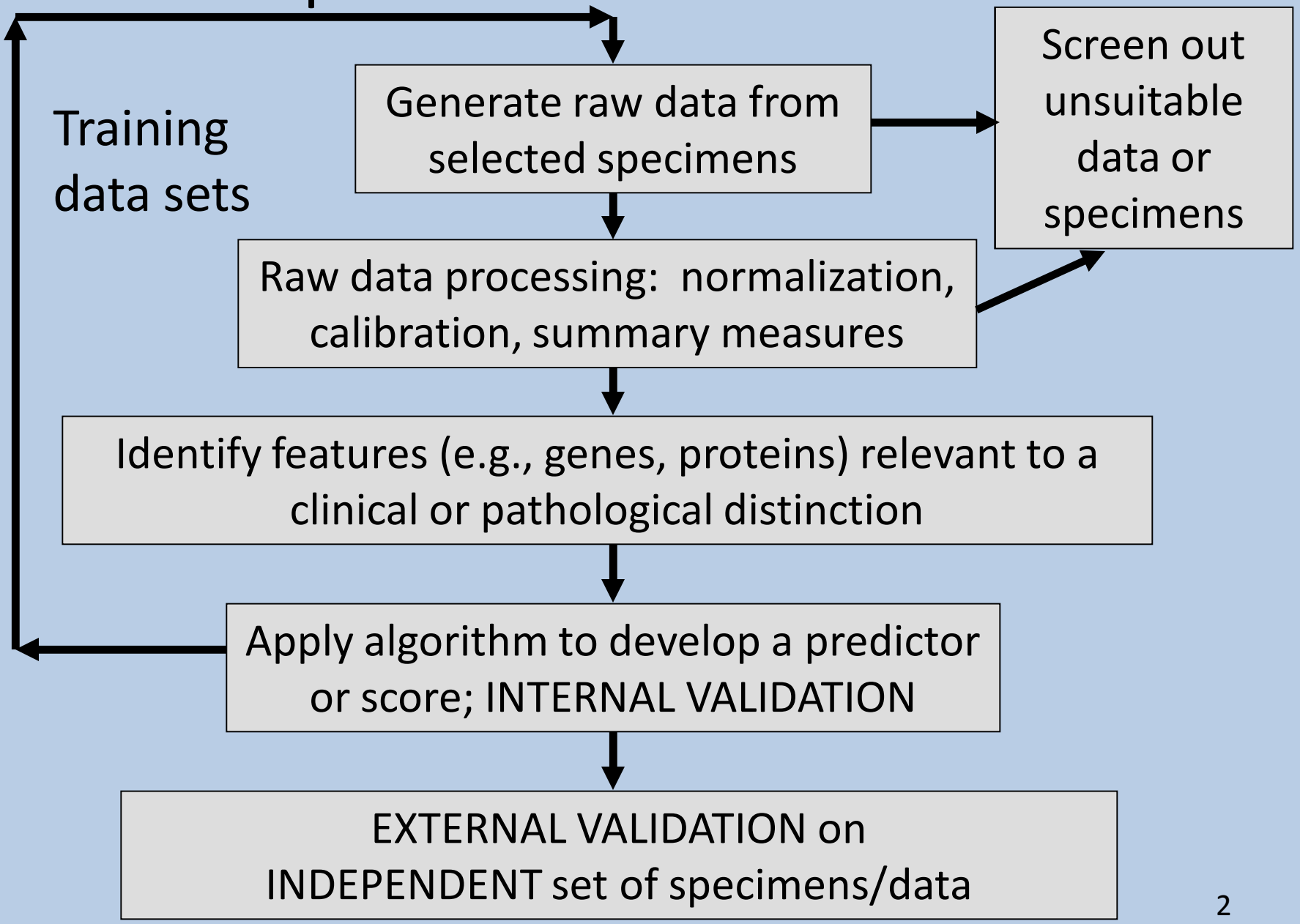
Evaluation of the Development, Validation, and Integrity of a Genomic Predictor

Lisa M McShane, PhD

Biometric Research Branch

National Cancer Institute

Development of an Omics Predictor



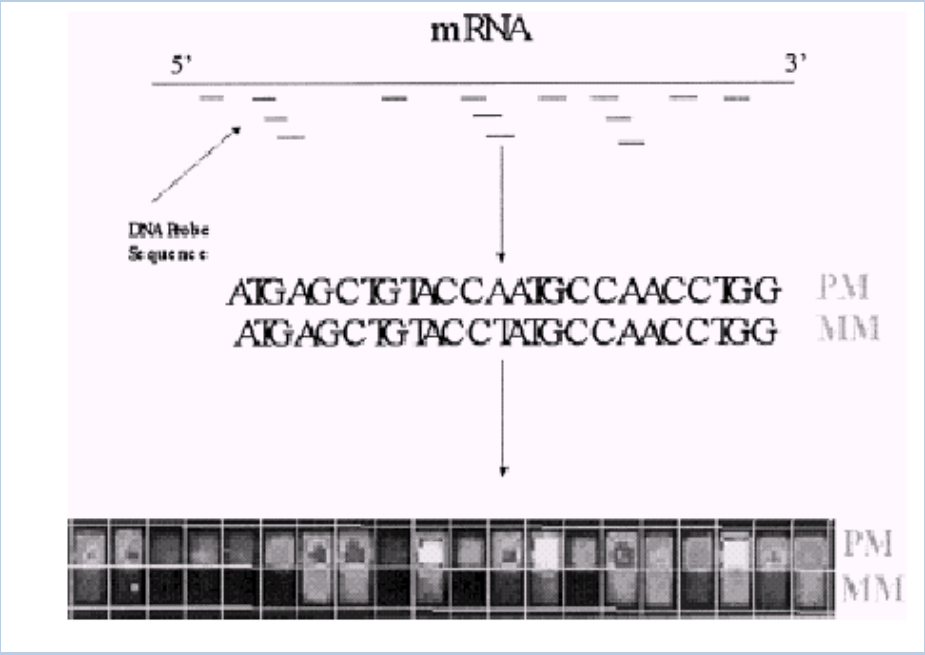
Training Set (specimens/data)

- Where did the specimens come from?
  - Was it a single source, or multiple?
  - Uniform sample collection, handling and preservation?
- Were the omics assays conducted in one or multiple labs, in one or multiple assay batches?
- Is there potential confounding of any of the above factors with the outcome that you want to predict?
  - Do patients accrued at different clinical sites have different stage distribution, or receive different treatments?
  - Are “responder” specimens obtained and/or assayed at site A but “non-responder” specimens obtained and/or assayed at site B?

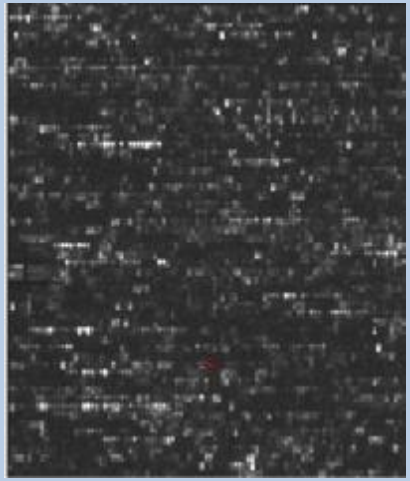
"Raw" data → "Processed" data

- Preprocessing
  - Calibration/normalization
  - Background corrections
- Summary measures
  - Example: Gene signal (probe set summaries from Affymetrix chips)
- Further normalization or standardization
  - Centering
  - Scaling
  - Centered & scaled
- All steps must be documented!

Affymetrix GeneChip Example



Gene sequence



- One probe type per “cell”
- Typical probe = 25-mer oligo
- 11-20 PM:MM pairs per probe set
- One gene summary per probe set (MAS 5.0, RMA, etc.)
- Further normalization or standardization

Identify “Informative Features”

- Which genes are expressed at different levels between the two groups (e.g., favorable vs. unfavorable; responder vs. non-responder)?
- Potential for many false positives
  - Performing 10,000 statistical tests, each at level 0.05 will generate 500 false positives when there are truly no informative features
- Might be many different sets of equally informative features (e.g., co-regulated genes)

Predictor or Risk Score

- Link informative feature measurements to clinical outcome or characteristic
- Derive mathematical function that associates a specimen with a class or assigns a continuous score based on inputted feature measurements
- Most scores eventually subject to cut-points for clinical decision-making

Classification Methods

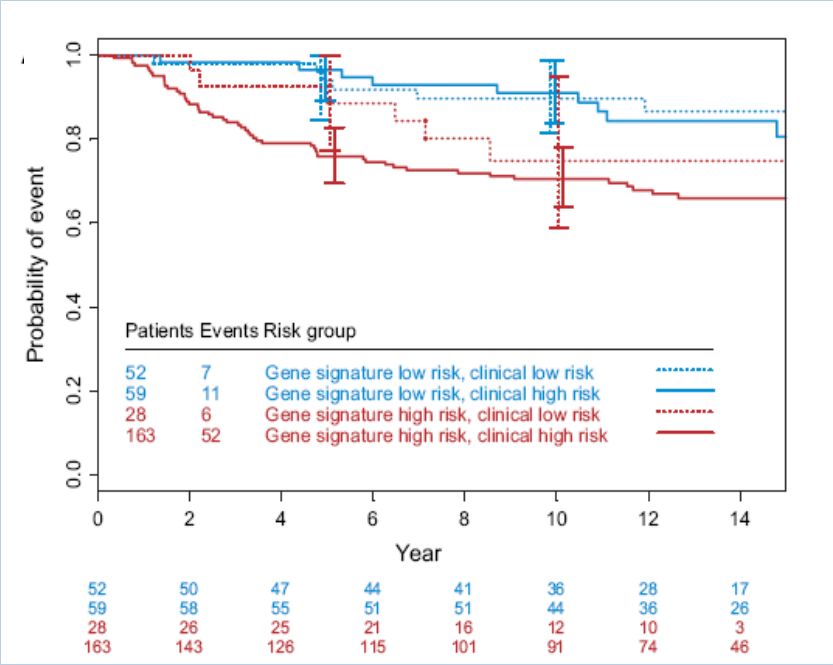
- Linear Predictor (for 2 classes)

  $L(\mathbf{x}) = w_1x_1 + w_2x_2 + \ldots + w_fx_f$

  is a weighted combination of important features to which a classification threshold is applied
  - Examples: Linear discriminant analysis, compound covariate predictor, weighted voting method, support vector machines with inner product kernel, perceptrons, naïve Bayes MVN mixture classifier
  - Features can be "metagenes"
- Distance-based
  - To which prototype pattern of informative features does the new pattern look most similar?
  - Examples: Nearest neighbor, nearest centroid
- Many more complex methods: Decision trees, random forests, completely stochastic or Bayesian model averaging

Example Clinical Predictors

MAMMAPRINT:
Outcome class predictor



Buyse et al, *JNCI,* 2006
70 genes
Prognostic/predictive?

ONCOTYPE DX:
Risk score with cut-points

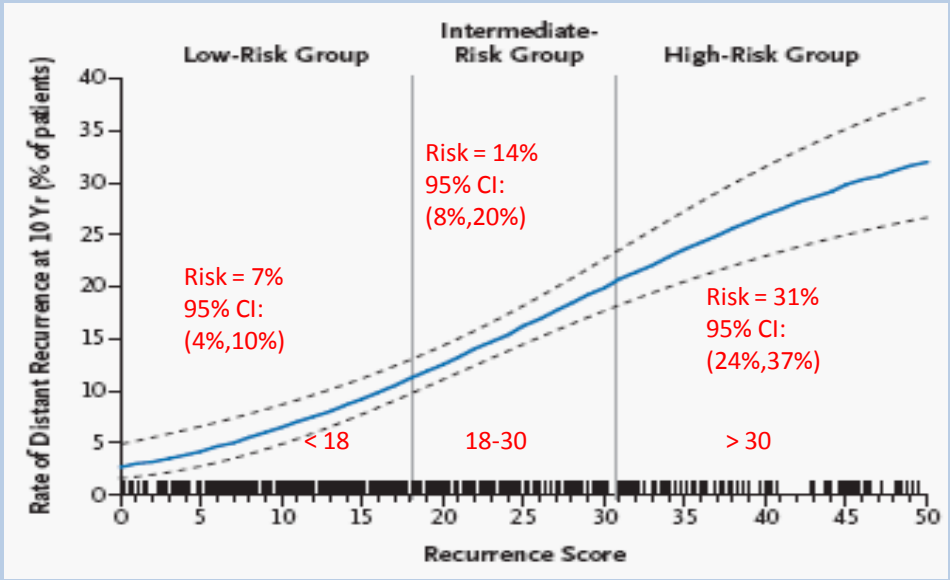


Figure 4 from Paik et al,
N Engl J Med, 2004
21 genes
Prognostic/predictive?

Classification: Avoiding Pitfalls

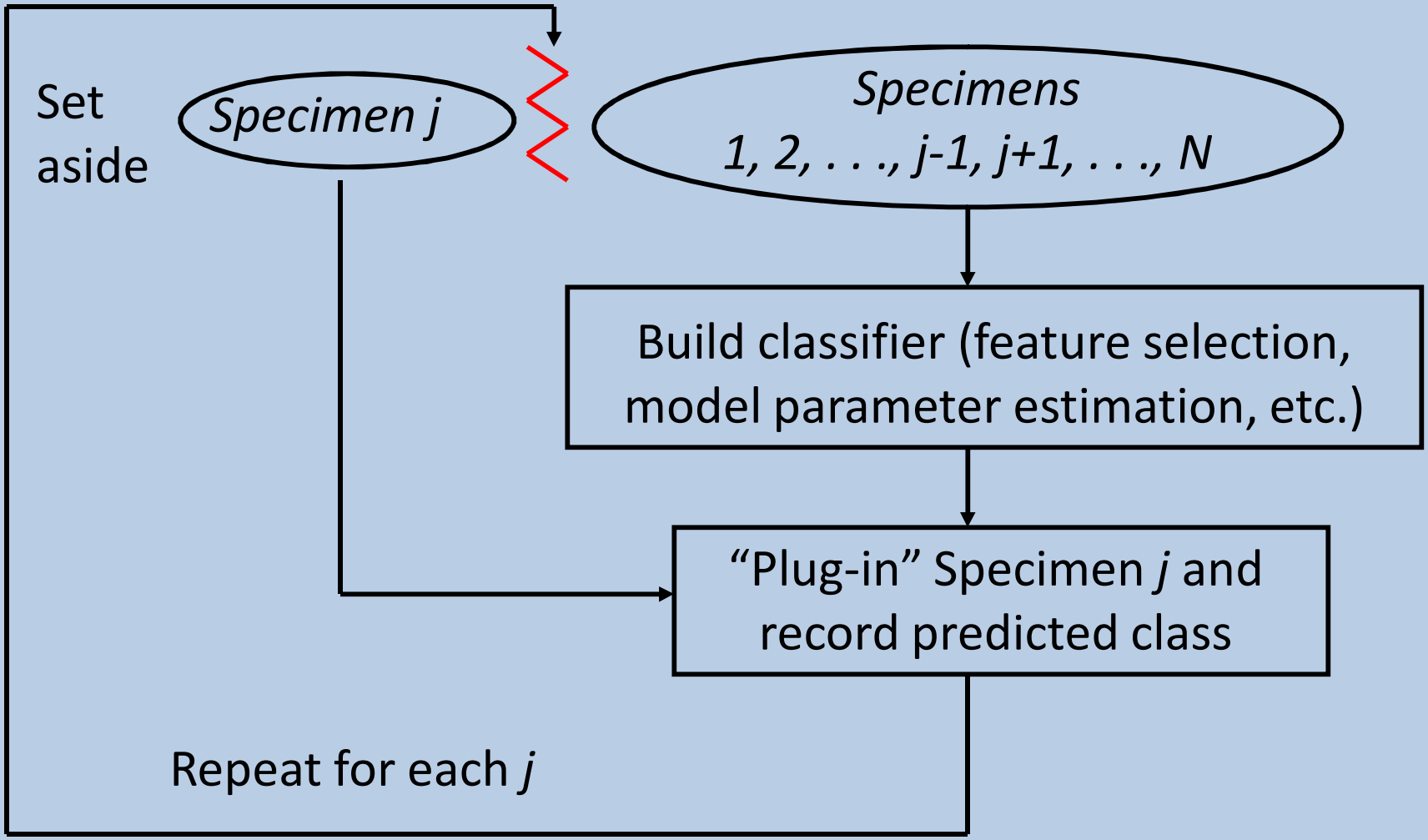
- When number of potential features is much larger than the number of cases, can always fit a classifier to have 100% prediction accuracy on data set used to build it
  - Can always perfectly fit a straight line (*two*-dimensional) between *two* points
- Estimating accuracy by "plugging in" data used to build a classifier results in highly biased estimates of prediction accuracy (re-substitution estimate)
- Internal and external validation of predictor are essential

Validation Approaches

- Internal: within-sample validation
  - Cross-validation

    (leave-one-out, split-sample, k-fold, etc.)
  - Bootstrap and other resampling methods
  - See Molinaro et al (*Bioinformatics* 2005) for comparison of methods
- External: independent-sample validation

Leave-one-out cross-validation (LOOCV)

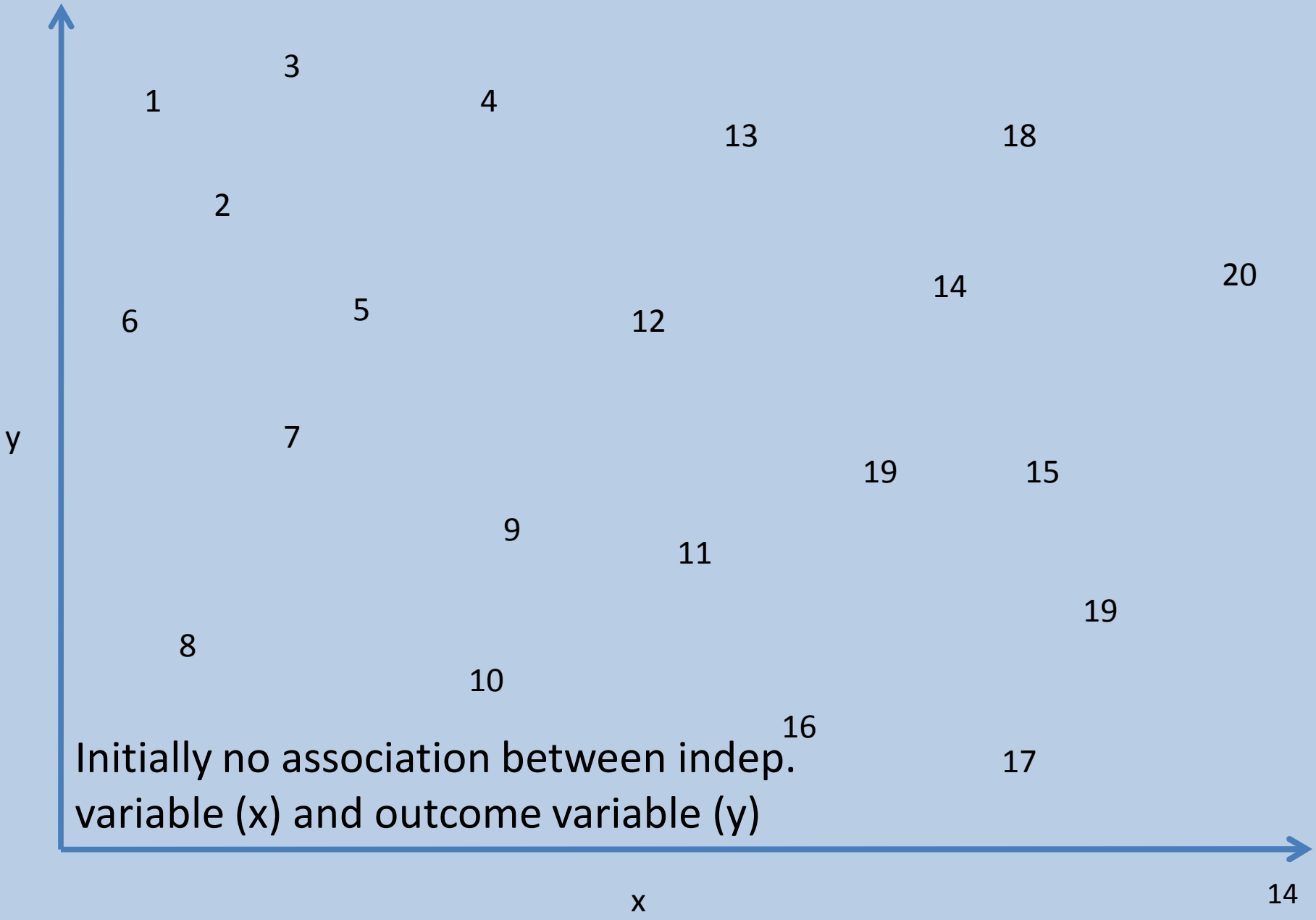


ALL steps, *including feature selection*, must be included in the cross-validation loop

Limitations of Within-Sample Validation

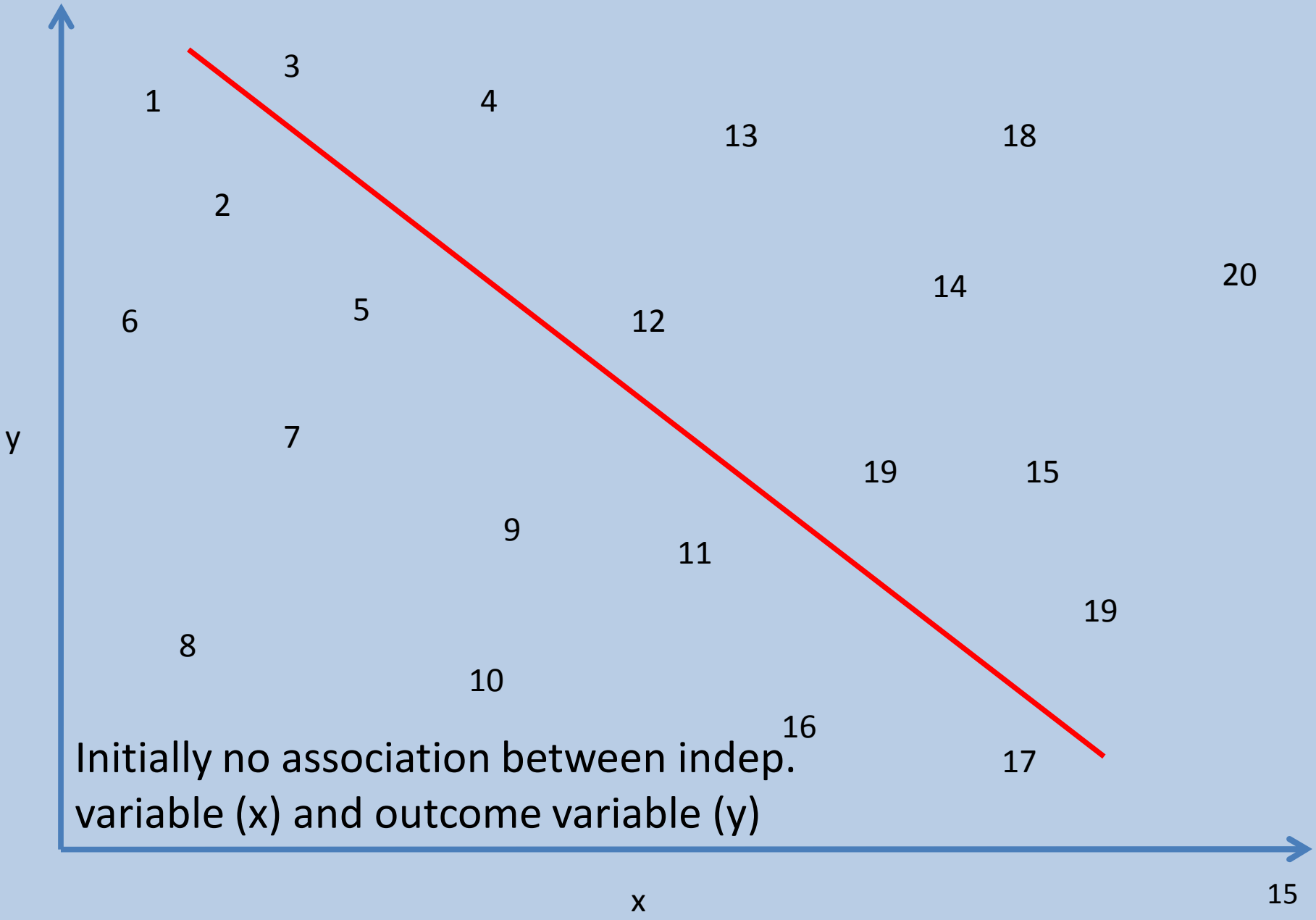
- Frequently performed incorrectly
  - Improper cross-validation (e.g., not including feature selection)
  - Special statistical inference procedures required (Lusa et al, *Statistics in Medicine* 2007; Jiang et al, *Stat Appl Genetics and Mol Biol* 2008)
- Large variance in estimated accuracy and effect sizes
- Doesn't protect against biases due to selective inclusion/exclusion of samples
- Built-in biases? (e.g., lab batch, specimen handling, etc.)

Dangers of selective inclusion/exclusion of cases



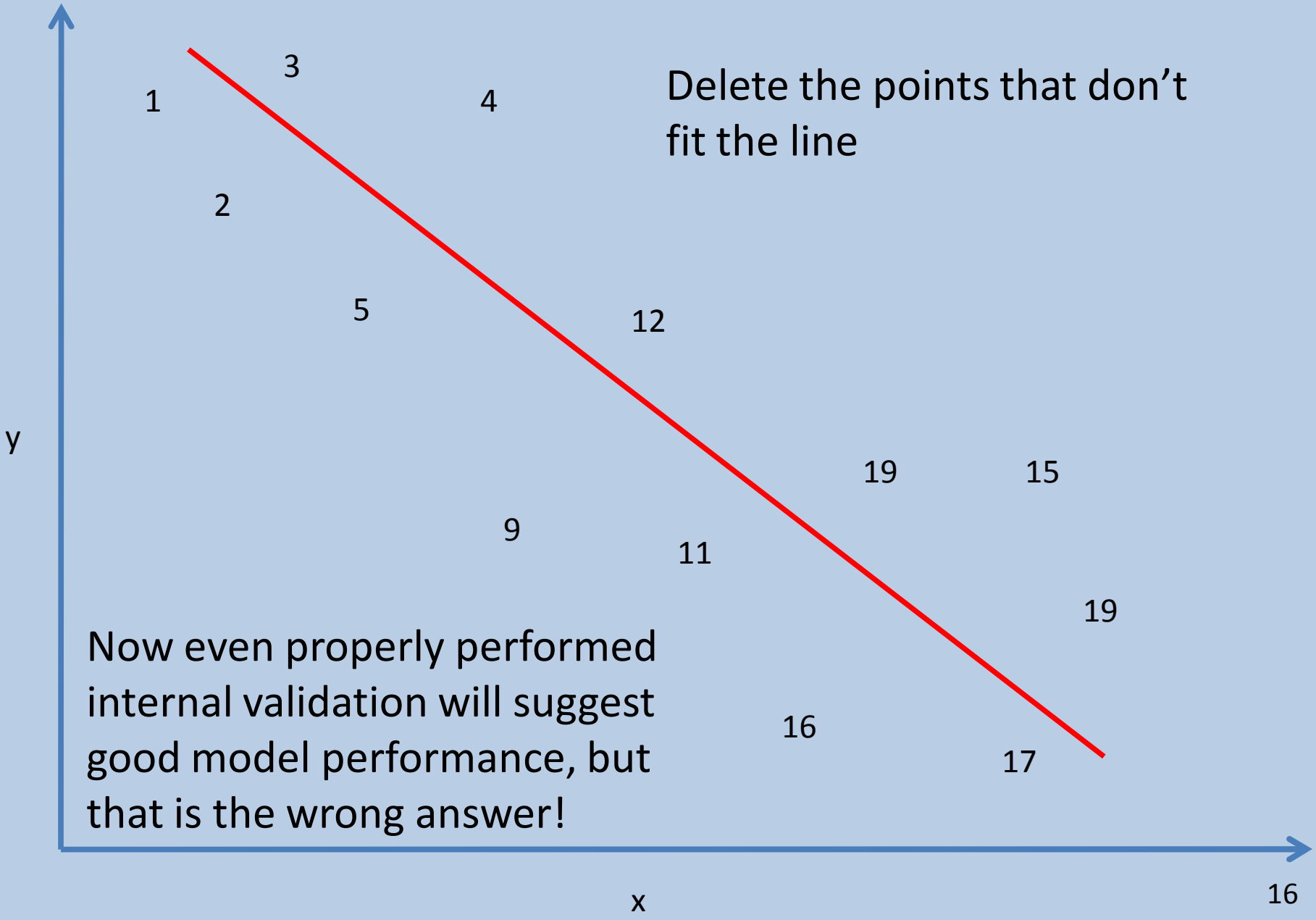
Initially no association between indep. variable (x) and outcome variable (y)

Dangers of selective inclusion/exclusion of cases



Initially no association between indep. variable (x) and outcome variable (y)

Dangers of selective inclusion/exclusion of cases



Delete the points that don't fit the line

Now even properly performed internal validation will suggest good model performance, but that is the wrong answer!

Corrupted Validation Data

- Suppose all model building steps are completely sound
- Still, results can be misleading if the *validation* data are corrupted
  - Test model on validation data with corrupted specimen labels (e.g., responder/nonresponder) or outcome variables (e.g., drug sensitivity measure)
  - Test model on validation data with corrupted omics (e.g., gene expression profile) data
  - Selective exclusion of validation specimens that don't fit the model developed on the training set

Information Leak from Validation Data Into Model Building Process

- Identify genes that are good predictors in the *validation* set
- Force those genes into the "informative set" of genes obtained from the training data
  - Cluster the validation data using the gene list that contains those found to be informative on the training data *plus* the forced genes from the validation data
  - Build the model with genes forced into it
- BIASED VALIDATION!

Combining training and validation data

- Build model on training set only
- Present performance results for that model on the full set of combined training and test sets?
- This is a hybrid between re-substitution method (invalid) and correct validation, and the overall result is HIGHLY BIASED!

Questions to Ask

- What data sets were the "starting points" for both the training and validation sets?
  - Inclusion/exclusion criteria?
  - Are the data accurate for both the training and validation sets (going back to *original* sources)?
  - Plugging data provided into computer code is a good start, but it does not confirm validity of data or assure reported prediction performance is free of biases

Questions to Ask

- If there was a fully specified predictor building algorithm, can the predictor be re-derived using the training data *only*?
- If there was no fixed predictor building algorithm, is there documentation of a *strict blinded* validation?
  - Split sample (internal) validation
  - Independent (external) validation

Questions to Ask

- Are results presented with appropriate separation of training and validation data?
- Are the *best* results of many attempts presented, or was a single predictor evaluated?
- Does the predictor always produce the same result given the same data?

Questions to Ask

- Is the predictor presented (and reportedly validated) *really the one being used in the trial*?