PRECISION HEALTH
ANALYSIS BOOTCAMP

Biomarker discovery and evaluation techniques

Amrit Singh, PhD

Department of Anesthesiology, Pharmacology and Therapeutics, UBC

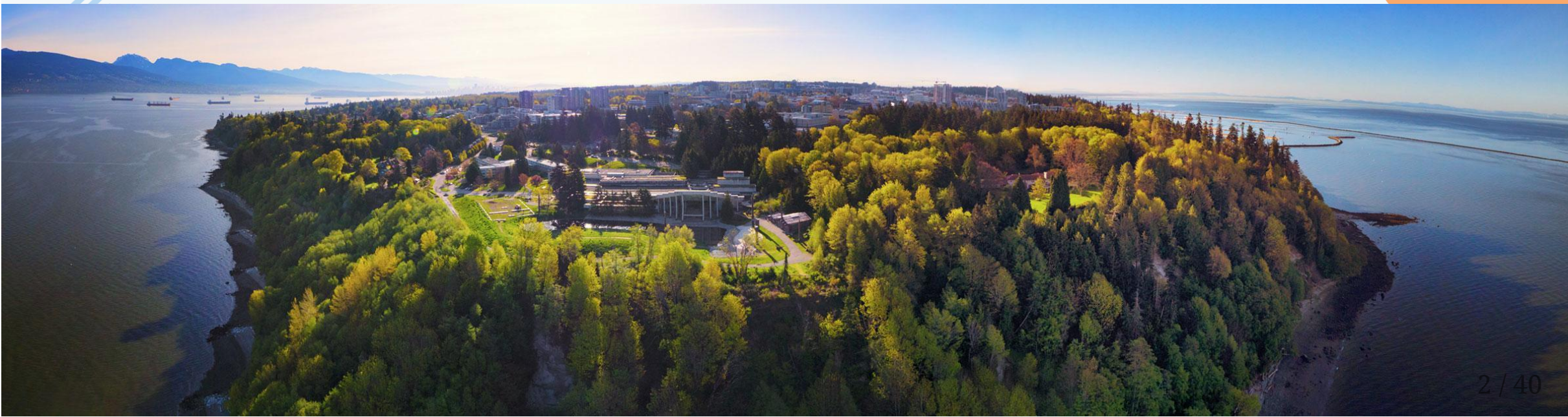
Centre for Heart Lung Innovation

August 05, 2022 | 12:00-14:00

lab
code
asingh_22g

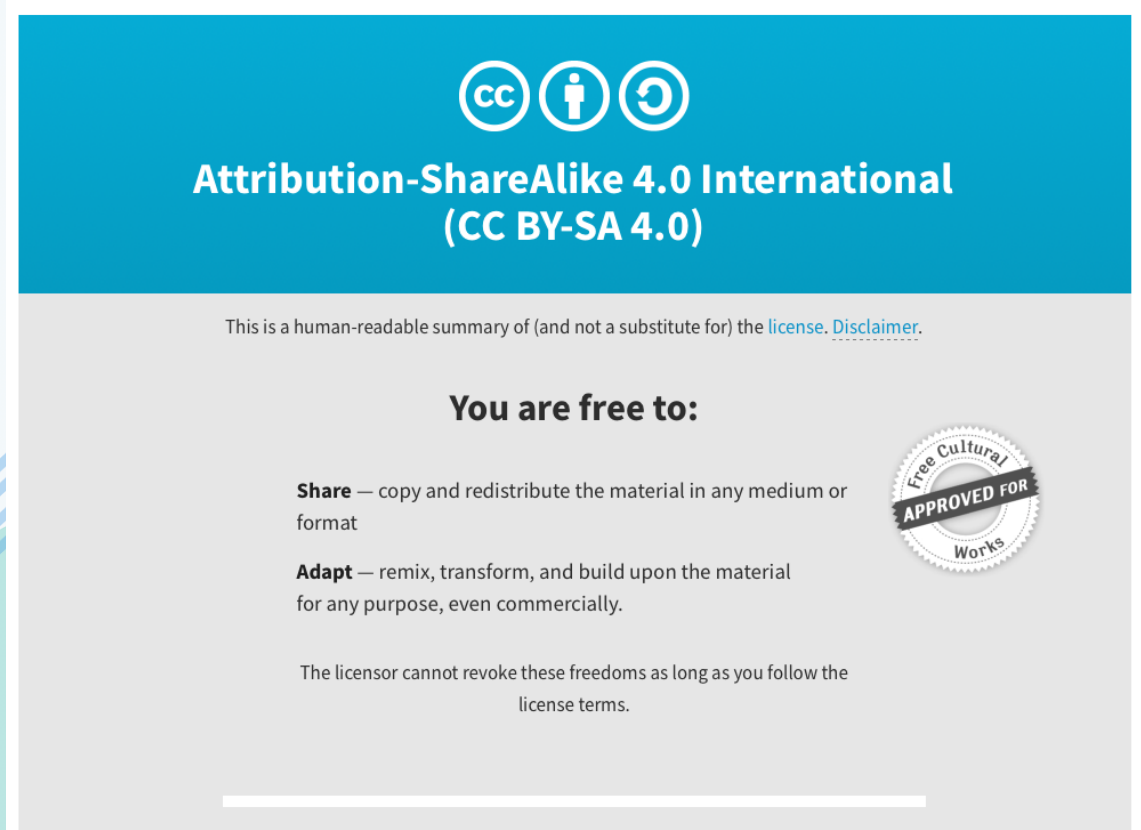
PRECISION HEALTH
ANALYSIS BOOTCAMP

We would like to begin by acknowledging that the land on which we gather is the traditional, ancestral, and unceded territory of the xwməθkwəy̓əm (Musqueam) People.



PRECISION HEALTH
ANALYSIS BOOTCAMP

Copyright Information



Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Free Cultural APPROVED FOR Works

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Read more here:
https://creativecommons.org/licenses/by-sa/4.0/

Learning outcomes

1. Describe what a biomarker is and give an example of one

2. Analyze simluated and breast cancer data to identify biomarkers

3. Discrimiminate the right and wrong to estimate the test error of a biomarker panel

4. Appreciate how easy it is to cherry-pick the data.

What is a biomarker?

A defined characteristic that is measured as an indicator of normal biologial processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions.

Types

- molecular
- histologic
- radiographic
- physiologic

Categories

- susceptibility/risk biomarker
- diagnostic biomarker
- monitoring biomarker
- prognostic biomarker
- predictive biomarker
- response biomarker
- safety biomarker

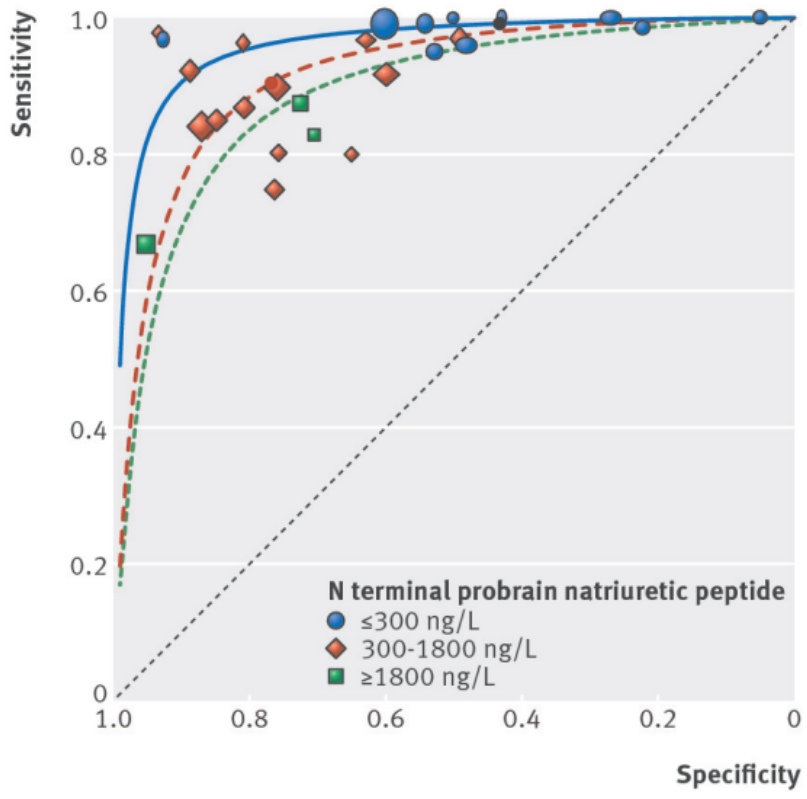
FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US); 2016-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK326791/ Co-published by National Institutes of Health (US), Bethesda (MD).]

Examples of biomarkers

Metric	Description
Diagnostic	• GFR for kidney disease • sweat chloride for Cystic Fibrosis
Monitoring	• Prostate-specific antigen (PSA) to assess cancer recurrence
Response	• Pharmacodynamic: CRP levels and tobacco use
Prognostic	• BRCA1/2 to assess the likelihood of breast cancer

FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US); 2016-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK326791/ Co-published by National Institutes of Health (US), Bethesda (MD).]

Magic bullet biomarkers: do they exist?



Roberts E et al., BMJ. 2015 Mar 4;350:h910.

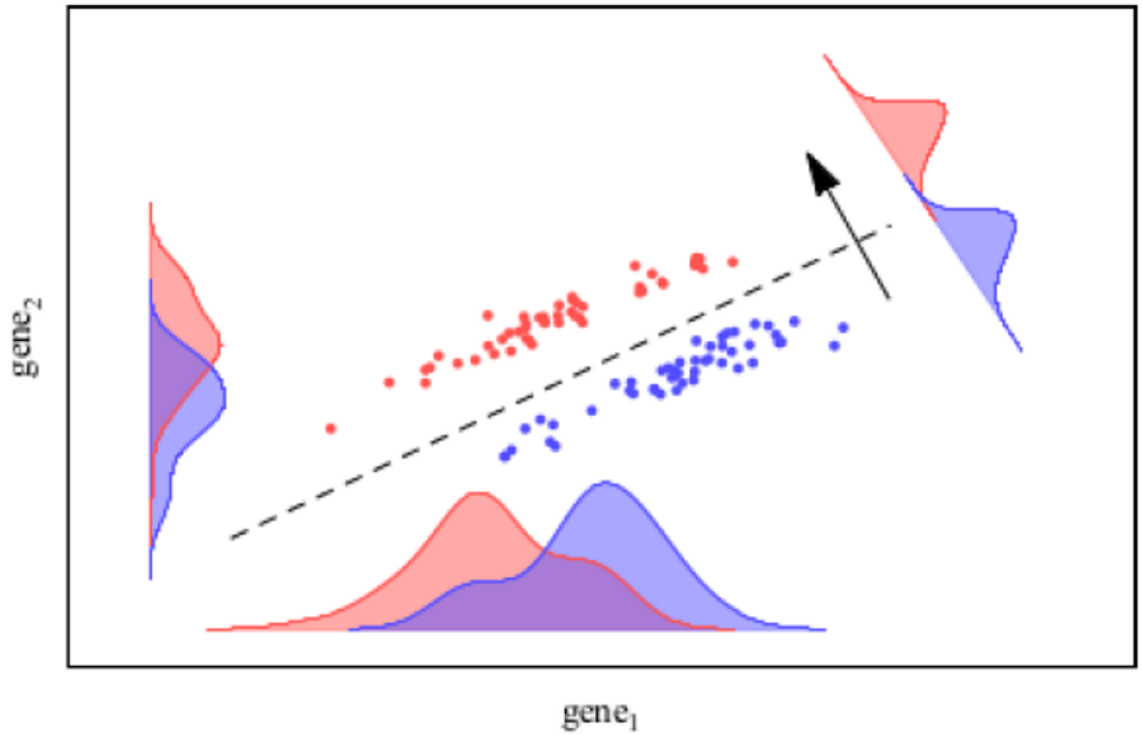
N terminal probrain natriuretic peptide (NTproBNP) is a rule-out test for acute heart failure

serum NTpBNP <= 300ng/L :

- sensitivity: 0.95 (0.93-0.96)
- negative predictive value: 0.98 (0.89-1)
- specificity: 0.43 (0.26-0.62)
- positive predictive value: 0.64 (0.57-0.73)

cardic imaging is required to confirm diagnosis if positive for heart failure.

Why use a multimarker approach?

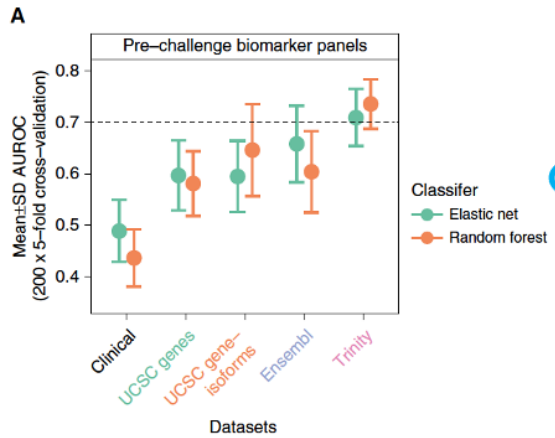


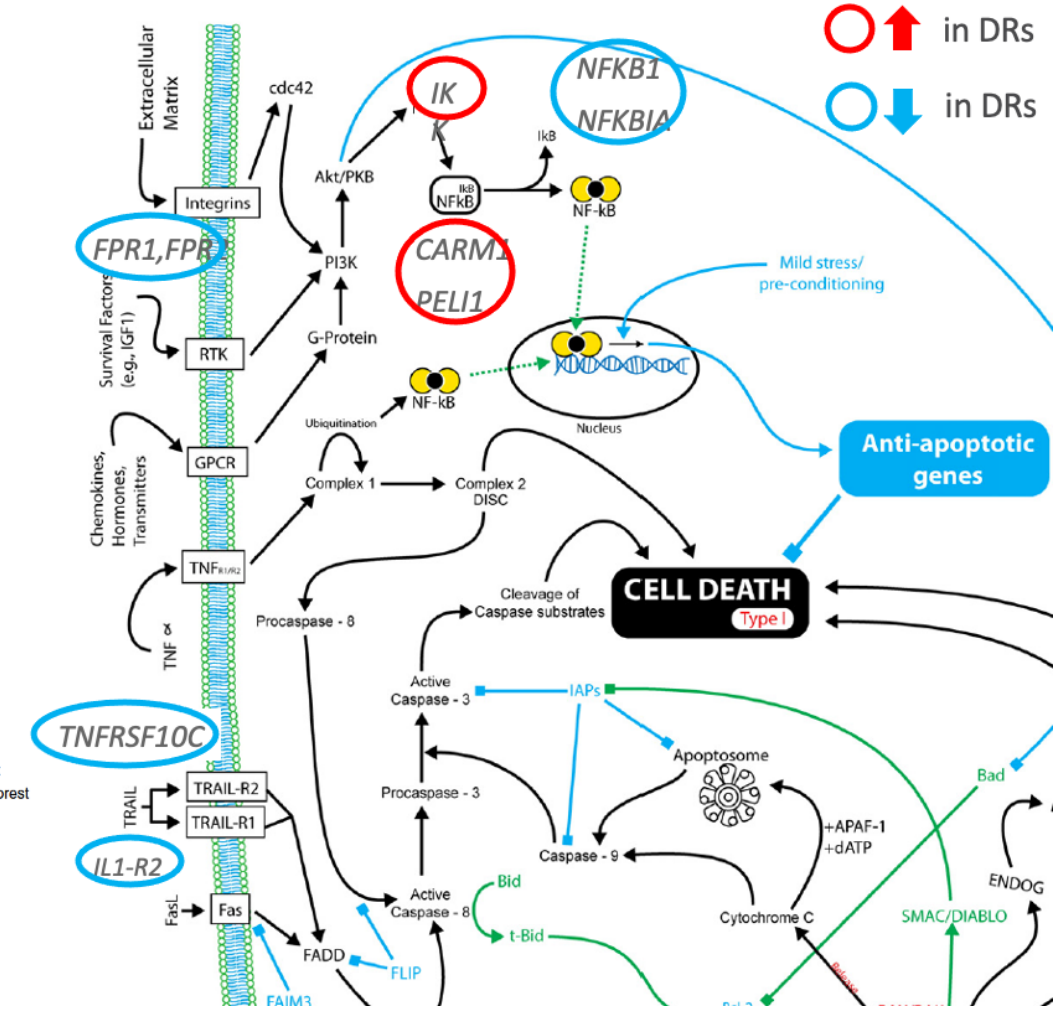
$$BS(subject) = w_1 * gene_1 + w_2 * gene_2$$

BS of subject = Biomarker Score for a given subject

BMC Bioinformatics . 2014 Mar 21;15:79. doi: 10.1186/1471-2105-15-79.

Gene expression





Singh A et al., AJRCCM 2018, 15:79-93. Portt L. et al. Biochimica et Biophysica Acta 2011, 238-259

Clinical utlity?

Single biomarker

Pros	Cons
• many real-world biomarkers • easier to develop into a clinical test • results are easier interpret	• not perfect • can be generic (CRP)

Panel of biomarkers

Pros	Cons
• increasing real-world biomarkers (PROSIGNA) • better at capturing the heterogeneity of the disease	• not perfect • results are harder interpret

Terminology review

Term	Synonym	Description
biomarker panel	biomarker signatures, multimarker panels, models, algorithms, equations	two or more biomarkers that an algorithm or equation is based on
training set	discovery cohort	dataset used to develop biomarker panels
test set	validation cohort	an independent dataset not used when developing biomarker panels
Technical replication		confirmation of findings using the same samples, (e.g. measuring the same protein using different technologies in the same set of samples)
validation		confirmation of findings using samples from the original discovery cohort

Contrast with Deep Learning (DL) common practice:

- DL pipeline use a separate cohort for train, validation and test
- Biomarker literature there is discovery and validation cohort; the discovery cohort can be divided into train and validation set however cross-validation is used instead to tune-hyperparameters.

How to identify biomarkers?

A panel of biomarkers can be developed from any of the following approaches:

- differential expression
- machine learning methods
- *a prior* list of biomarkers
- combination of approaches

THATS EASY!

The hard part is trying to estimate the test error in the absence of additional data!

Case Studies

Case Study 1: Two cohorts (discovery + validation) - real data

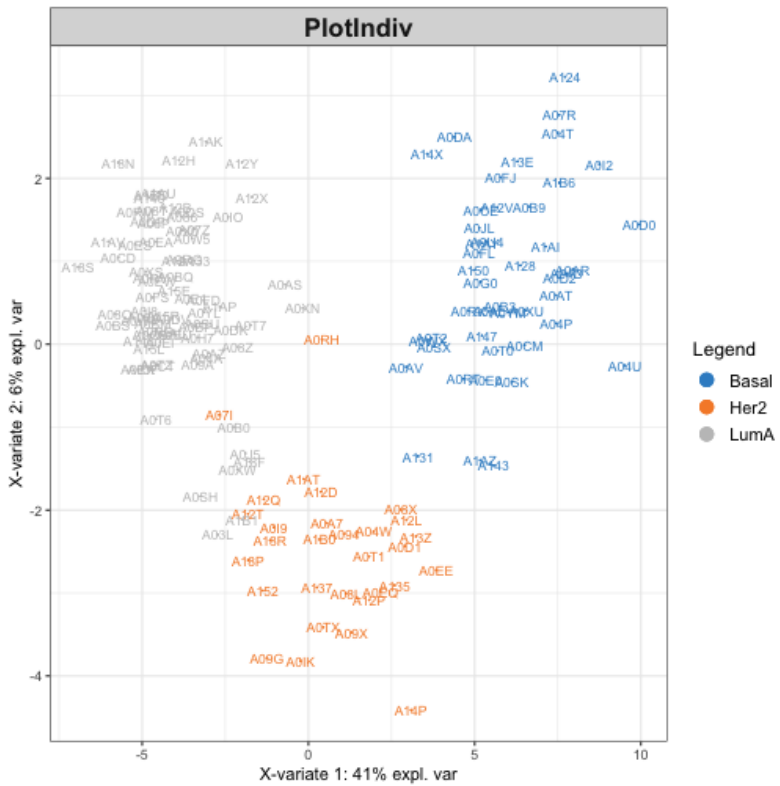
Data: Breast Cancer multi omics data from TCGA (part of mixOmics R-package)

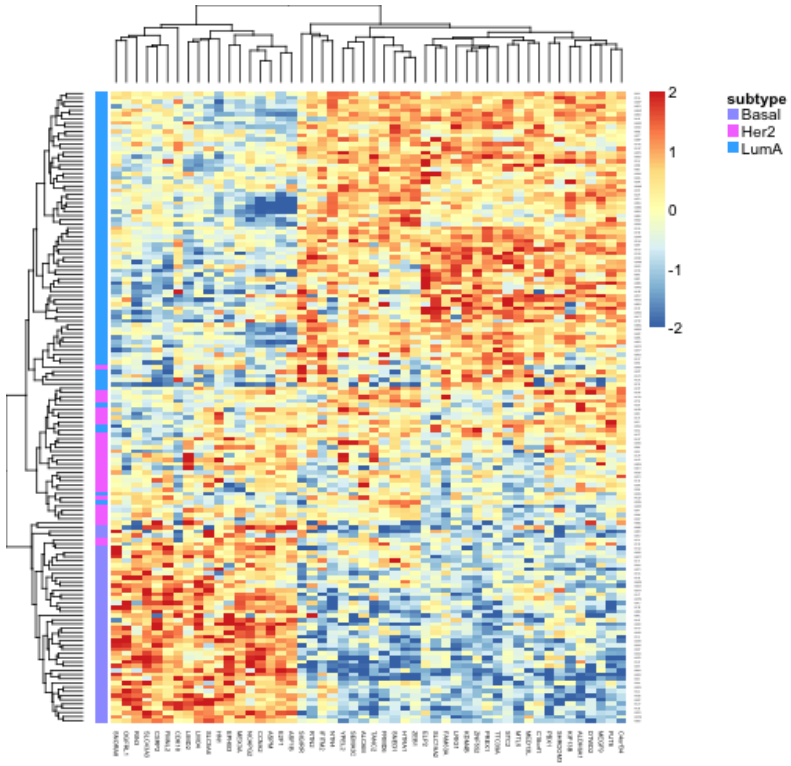
*This data set is a small subset of the full data set from The Cancer Genome Atlas that can be analysed with the DIABLO framework. It contains the expression or abundance of three matching omics data sets: mRNA, miRNA and proteomics for 150 breast cancer samples (Basal, Her2, Luminal A) in the training set, and 70 samples in the test set. The test set is missing the proteomics data set. Bioinformatics. 2019 Sep 1;35(17):3055-3062

Case Study 2: One cohort - simulated data

Case Study 1: Method 1

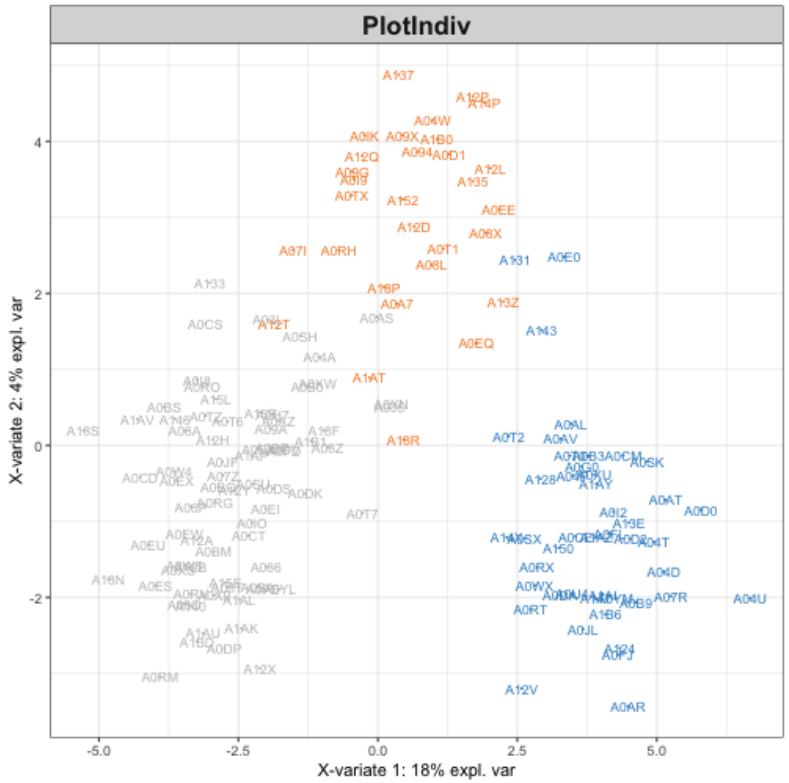
- Step 1: Differential expression (50 gene panel)
- Step 2: train machine learning model

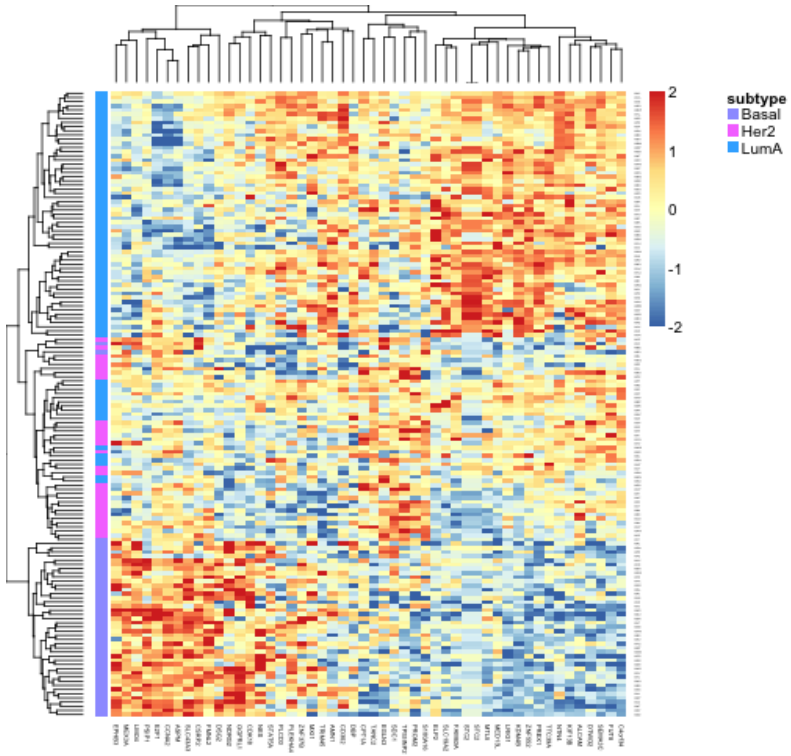




Case Study 1: Method 2

- using a machine learning model with variable selection (select 50 genes): sPLS-DA





> ZNF552, KDM4B, PREX1, LRIG1, CCNA2, TTC39A, C4orf34, FUT8, ASPM, MEX3A, SLC43A3, SEMA3C, STC2, LMO4, E2F1, MED13L, FMNL2, DTWD2, SLC19A2, KIF13B, FAM63A, CSRP2, NTN4, MTL5,

Overlap between biomarkers between the two approaches

```
## Loading required package: grid
```

How to determine the performance of biomarkers?

Metric	Description
Area under the receiver operating characteristic curve (AUROC)	• binary classification • true vs. false positive rate
Mean Square Error (MSE)	• regression • difference between actual vs. observed squared
Accuracy	• classification • proportion of correctly classified
F1 Score	• classification • combines precision and recall

See formulas here

J Natl Cancer Inst. 2015 Jun 24;107(8):djv153.

Case Study 1: Apply biomarker panels to test cohorts

Method 1: Diff_plsda_biomarker_panel

Train (model built using this data)

```
##
##         Basal Her2 LumA
##   Basal    43    0    0
##   Her2      2   29    3
##   LumA      0    1   72
```

Accuracy: 0.96 and Balanced Accuracy: 0.96

Test (apply model to independent cohort)

```
##
##         Basal Her2 LumA
##   Basal    18    0    0
##   Her2      3   13    1
##   LumA      0    1   34
```

Accuracy: 0.93 and Balanced Accuracy: 0.92

Method 2: splsda_biomarker_panel

Train (model built using this data)

```
##
##         Basal Her2 LumA
##   Basal    43    0    0
##   Her2      2   29    3
##   LumA      0    1   72
```

Accuracy: 0.96 and Balanced Accuracy: 0.96

Test (apply model to independent cohort)

```
##
##         Basal Her2 LumA
##   Basal    20    0    0
##   Her2      1   14    1
##   LumA      0    0   34
```

Accuracy: 0.97 and Balanced Accuracy: 0.97

Q: What if you don't have a test cohort?

Anwers: split data and pretend you do have test cohorts:

Resampling strategies

Type	Description
K-fold cross validation (CV)	• split data into K folds of equal size (can also be stratified across disease groups)
repeated K-fold cross-validation	• repeat K-fold CV multiple times until metric values convergences
nested cross-validation if hyperparameter tuning is required	• repeat K-fold CV with each fold and select the best model to apply to test data

- bias: how are far is the model away from the true model (K=N has the lowest biased but high variance whereas lower values of K have higher bias but lower variance); K of 5 or 10 offer a good compromise.

K-fold cross validation (CV)

K-fold cross validation (4-fold CV)

K-fold cross validation (4-fold CV)

K-fold cross validation (4-fold CV)

- Note: many approaches can be used to create the final model

K-fold cross validation (4-fold CV)

- Note: many approaches can be used to create the final model

K-fold cross validation (4-fold CV)

- Note: many approaches can be used to create the final model

K-fold cross validation (4-fold CV)

- Note: many approaches can be used to create the final model

K-fold cross validation (4-fold CV)

- Note: many approaches can be used to create the final model

K-fold cross validation (4-fold CV)

- Note: many approaches can be used to create the final model

K-fold cross validation (4-fold CV)

- Note: many approaches can be used to create the final model

K-fold cross validation (4-fold CV)

- Note: many approaches can be used to create the final model

Case Study 2: Simulation scenerio (true error rate = 50%)

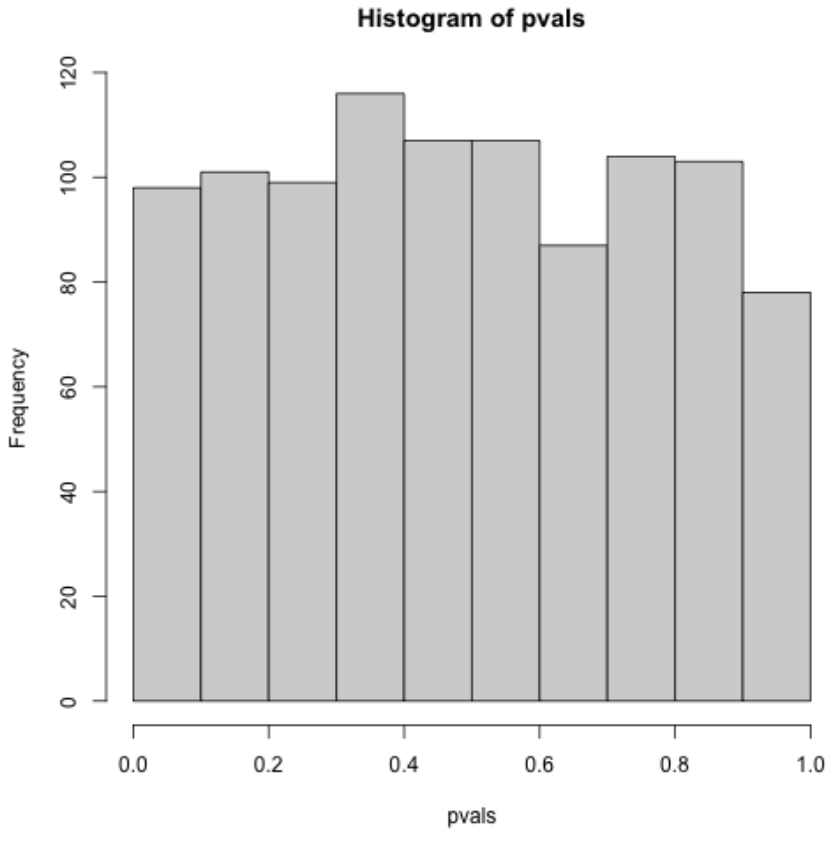
```
n <- 100; p <- 1000
group <- rep(c("group1", "group2"), each = n/2)
eset <- matrix(rnorm(n * p), nr = n)
colnames(eset) <- paste0("gene", 1:p)

pvals <- apply(eset, 2, function(i){
  t.test(i~group)$p.value
})

biomarkers <- pvals[order(pvals)][1:10]
as_tibble(data.frame(names = names(biomarkers),
           pvalues = biomarkers))
```

```
## # A tibble: 10 × 2
##    names    pvalues
##    <chr>      <dbl>
##  1 gene646 0.000225
##  2 gene345 0.000709
##  3 gene649 0.00121
##  4 gene407 0.00146
##  5 gene128 0.00254
##  6 gene308 0.00885
##  7 gene291 0.00931
##  8 gene140 0.0105
##  9 gene873 0.0110
## 10 gene920 0.0118
```

```
hist(pvals)
```



How to cross-validate?

Wrong way to do cross-validation

- subset dataset to selected biomarkers than apply cross-validation

```
wrong_way_model <- plsda(X = eset[, names(biomarkers)]

cv <- perf(wrong_way_model, validation = "Mfold", fold
cv$auc
```

```
## $comp1
## AUC.mean    AUC.sd
##   0.8888        NA
```

Right-way to do cross-validation

- DO NOT subset dataset and repeat the same model building process during each cross-validation train/test fold.

```
right_way_model <- splsda(X = eset, keepX = 10, Y = gr

cv <- perf(right_way_model, validation = "Mfold", fold
cv$auc
```

```
## $comp1
## AUC.mean    AUC.sd
##   0.5456        NA
```

I want to include differential expression in my approach

- within each cross-validation fold, start with **ALL** the data and perform differential expression, then use the top features to build the model and apply to the held out/test set.

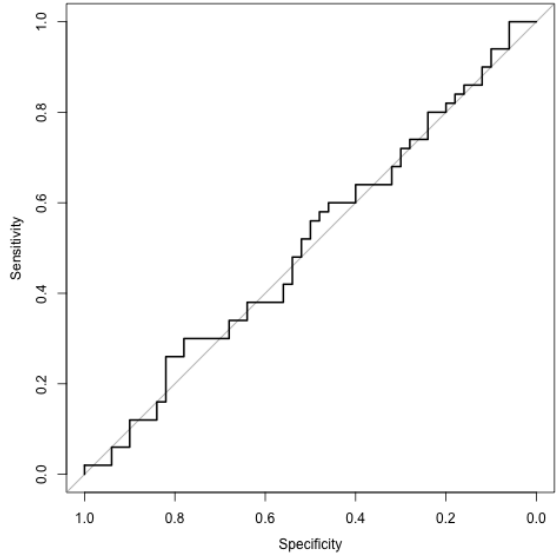
```
cvIndex <- caret::createFolds(factor(group), k = 5)  #

predictions <- lapply(cvIndex, function(fold){
  ## Create train and test dataests
  Xtrain <- eset[-fold, ]
  Xtest <- eset[fold, ]
  ytrain <- group[-fold]

  ## differential expression
  pvals <- apply(Xtrain, 2, function(i){
    t.test(i~ytrain)$p.value
  })
  biomarkers <- pvals[order(pvals)][1:10]

  ## plsda model
  model <- mixOmics::plsda(X = Xtrain[, names(biomarke
  as.numeric(predict(model, Xtest[, names(biomarkers)]
})
```

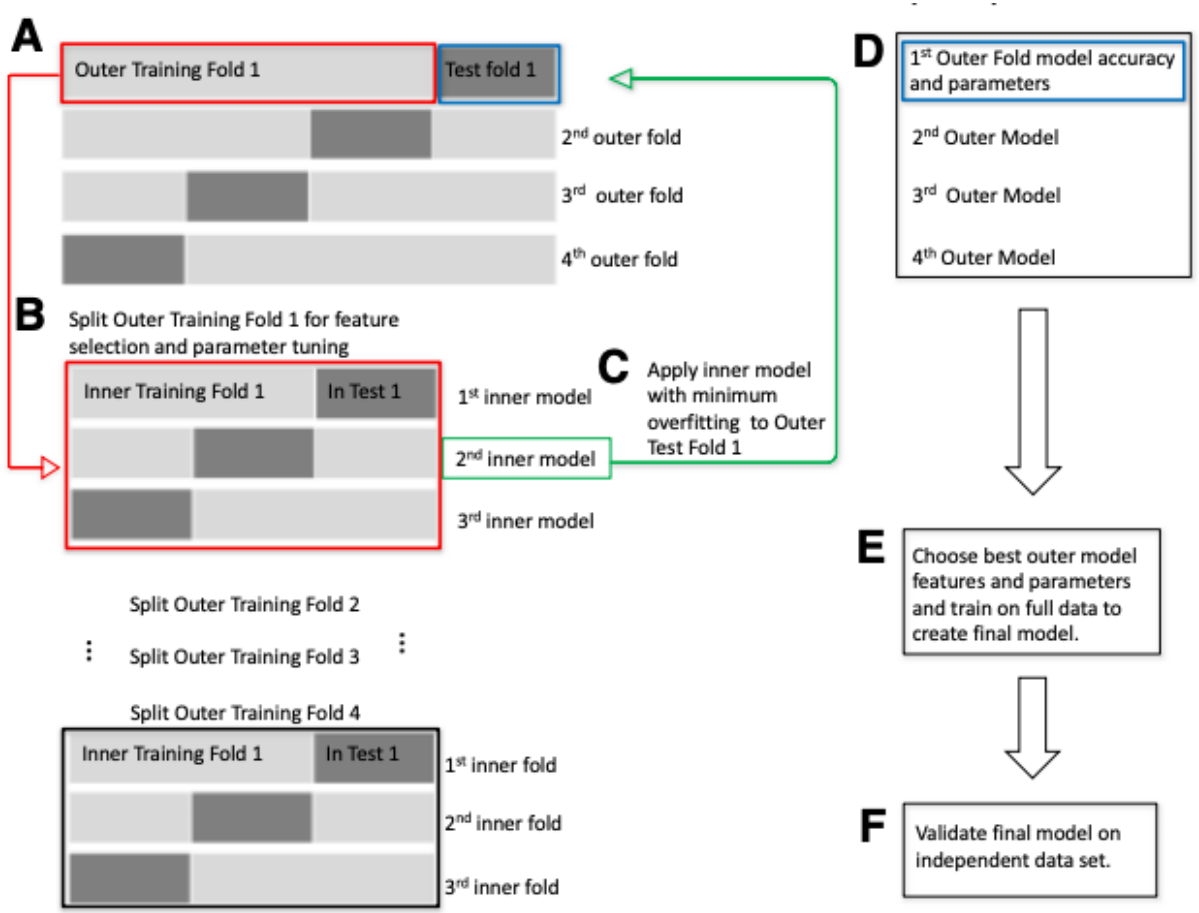
```
pROC::auc(group[unlist(cvIndex)], unlist(predictions),
```




```
## Area under the curve: 0.5096
```

Nested cross-validation

- apply cross-validation within each cross-validation fold (*e.g.* hyperparamter tuning)



Bioinformatics . 2020 May 1;36(10):3093-3098

Hyperparameter tuning (tune the number of variables to select)

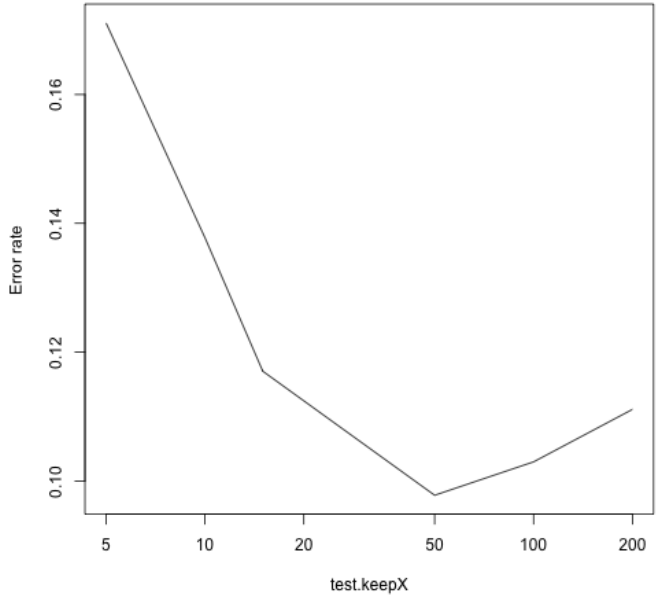
use a grid of values to find the biomarker panel with the least error rate

```
set.seed(1)
test.keepX = c(5, 10, 15, 50, 100, 200)
tune <- tune(method = "splsda",
             X = breast.TCGA$data.train$mrna,
             Y = breast.TCGA$data.train$subtype,
             ncomp=1,
             test.keepX = test.keepX,
             folds=5, dist="centroids.dist", progressBar
```

```
## Calling 'tune.splsda'
```

```
final_model <- splsda(X = breast.TCGA$data.train$mrna,
             Y = breast.TCGA$data.train$subtype,
             ncomp=1,
             keepX = test.keepX[which.min(tune$error.rat
# selectVar(final_model, ncomp=1)$name
```

```
plot(tune$error.rate ~ test.keepX, type="l", log="x",
```



Estimate test error using only the TRAINING data

```
cvIndex <- caret::createFolds(factor(breast.TCGA$data.

predictions <- lapply(cvIndex, function(fold){
  ## Create train and test dataests
  Xtrain <- breast.TCGA$data.train$mrna[-fold, ]
  Xtest <- breast.TCGA$data.train$mrna[fold, ]
  ytrain <- breast.TCGA$data.train$subtype[-fold]

  ## find optimal panel
  test.keepX = c(5, 10, 15, 50, 100, 200)
  tune <- tune(method = "splsda",
               X = Xtrain,
               Y = ytrain,
               ncomp=1,
               test.keepX = test.keepX,
               folds=5, dist="centroids.dist", progressB

  print(test.keepX[which.min(tune$error.rate)])

  final_model <- splsda(X = Xtrain,
               Y = ytrain,
               ncomp=1,
               keepX = test.keepX[which.min(tune$error.r

  ## plsda model
  predict(final_model, Xtest)$class$centroids.dist
})
```

Accuracy

```
conf_matrix <- table(unlist(predictions), breast.TCGA$
sum(diag(conf_matrix))/sum(conf_matrix)
```

```
## [1] 0.9066667
```

Balanced Accuracy

```
mean(diag(conf_matrix)/colSums(conf_matrix))
```

```
## [1] 0.8888889
```

Estimate test error using only the TEST (independent) data

```
predicted_labels <- predict(final_model, breast.TCGA$data.test$mrna)$class$centroids.dist

conf_matrix <- table(predicted_labels, breast.TCGA$data.test$subtype)
```

Accuracy

```
sum(diag(conf_matrix))/sum(conf_matrix)
```

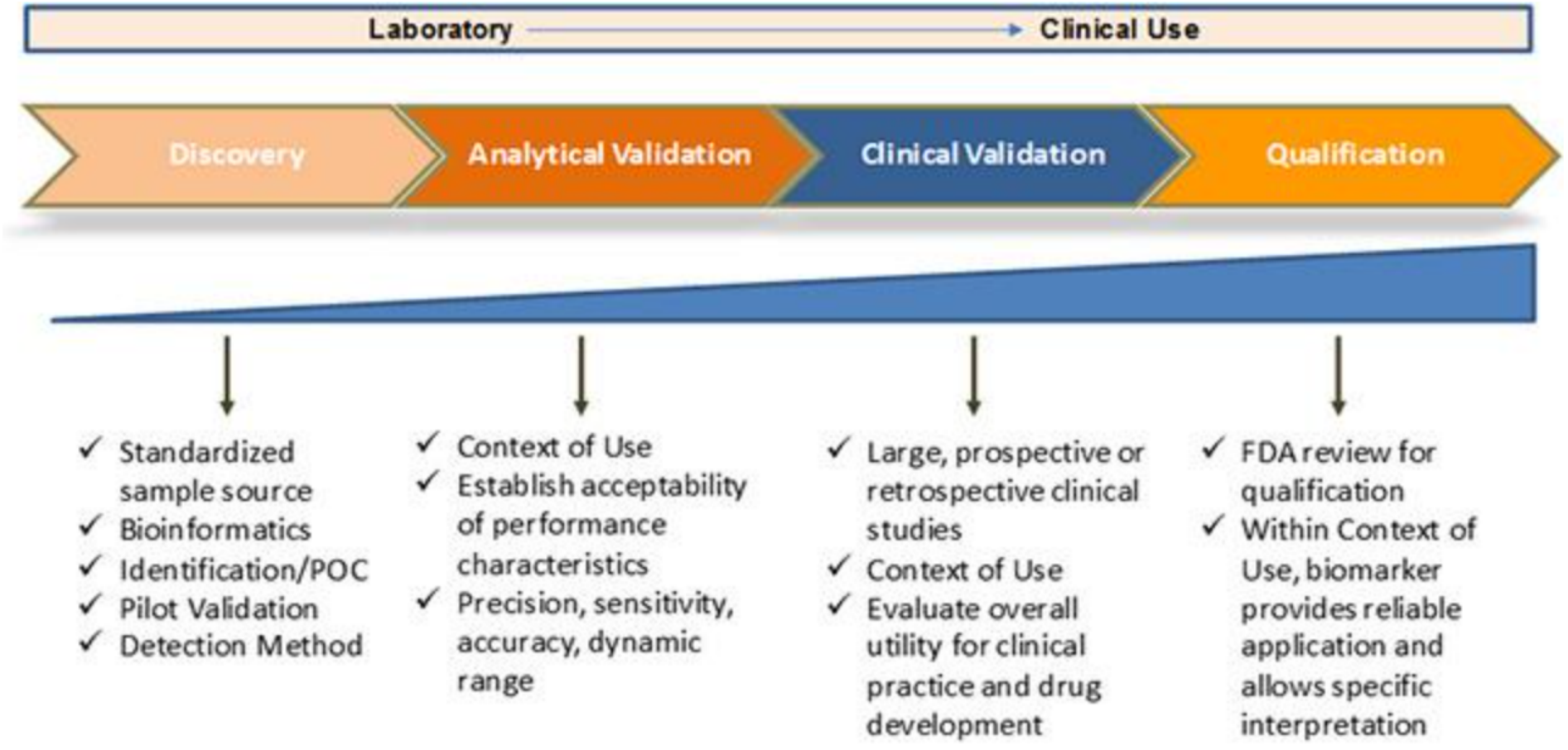
```
## [1] 0.9
```

Balanced Accuracy

```
mean(diag(conf_matrix)/colSums(conf_matrix))
```

```
## [1] 0.8857143
```

Should you biomarker? yes



Resources

Code

- Tidymodels
- mixOmics

Slides

- xaringan

Biomarkers

- Biomarker Discovery and Validation: Statistical Considerations
- BEST (Biomarkers, EndpointS, and other Tools) Resource

Cross-validation

- Resampling techniques

PRECISION HEALTH
ANALYSIS BOOTCAMP

THANK YOU!

August 05, 2022 | 12:00-14:00

lab
code
asingh_22g