A Comparison of Algorithms to Predict Endometriosis from Gene Expression Intensity.

ASEM BERKALIEVA & EDIE ESPEJO

Endometriosis affects 5-10% of women*.

Endometriosis is a estrogen-dependent condition that causes chronic pain and inflammation of the uterus.

Endometriosis affects 5-10% of women*.

Endometriosis is a estrogen-dependent condition that causes chronic pain and inflammation of the uterus.

Abnormal tissue growth outside of the uterus cause lesions in the abdomen and pelvic cavity.

Endometriosis affects 5-10% of women*.

Endometriosis is a estrogen-dependent condition that causes chronic pain and inflammation of the uterus.

Abnormal tissue growth outside of the uterus cause lesions in the abdomen and pelvic cavity.

Endometriosis may causes all sorts of pain "down there" and may even lead to infertility.

There is no cure.

What can be done
then, if there
is no cure?

What can be done then, if there is no cure?

1. A hysterectomy (removal of the ovaries) may halt the endometrial hormone production.

What can be done then, if there is no cure?

1. A hysterectomy (removal of the ovaries) may halt the endometrial hormone production.

2. Women can wait for menopause to recede endometrial symptoms.

What can be done then, if there is no cure?

1. A hysterectomy (removal of the ovaries) may halt the endometrial hormone production.

2. Women can wait for menopause to recede endometrial symptoms.

3. Surgery can be opted for, though for 20-50% of surgeries, endometrial growth recurs.

What can be done then, if there is no cure?

1. A hysterectomy (removal of the ovaries) may halt the endometrial hormone production.
2. Women can wait for menopause to recede endometrial symptoms.
3. Surgery can be opted for, though for 20-50% of surgeries, endometrial growth recurs.
4. Women can choose to suppress their menstruation with birth control which can help relieve pain.

We can’t stop the pain, but we can help diagnose it to route patients to treatment.

- Endometriosis is only fully diagnosed at surgery
- Use genomics data (to avoid invasive surgery) and ML to predict whether or not a patient has endometriosis
- Data are available from UCSF
    - n=148
    - Patients (aged 20-50) included
        - had pelvic pain (labeled mild to severe)
        - infertility issues
        - benign gynecological conditions
        - normal volunteers
    - Arrays were processed using Affymetrix HU133 Plus 2.0 at UCSF Genomics Core Facility
    - Data were collected for the NIH/UCSF Human Endometrial Tissue Bank

Data source.

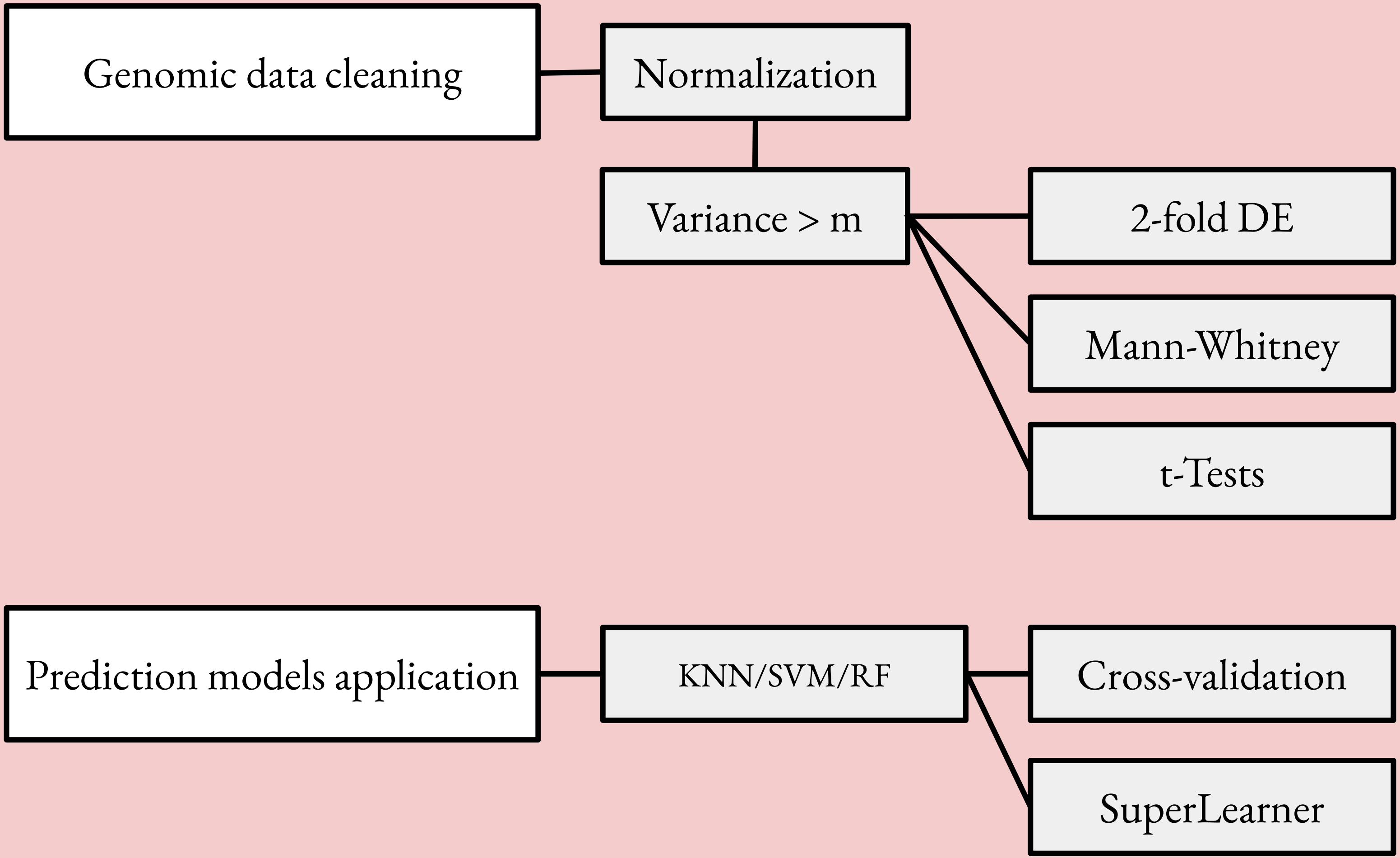
Giudice, 1999 [Link]

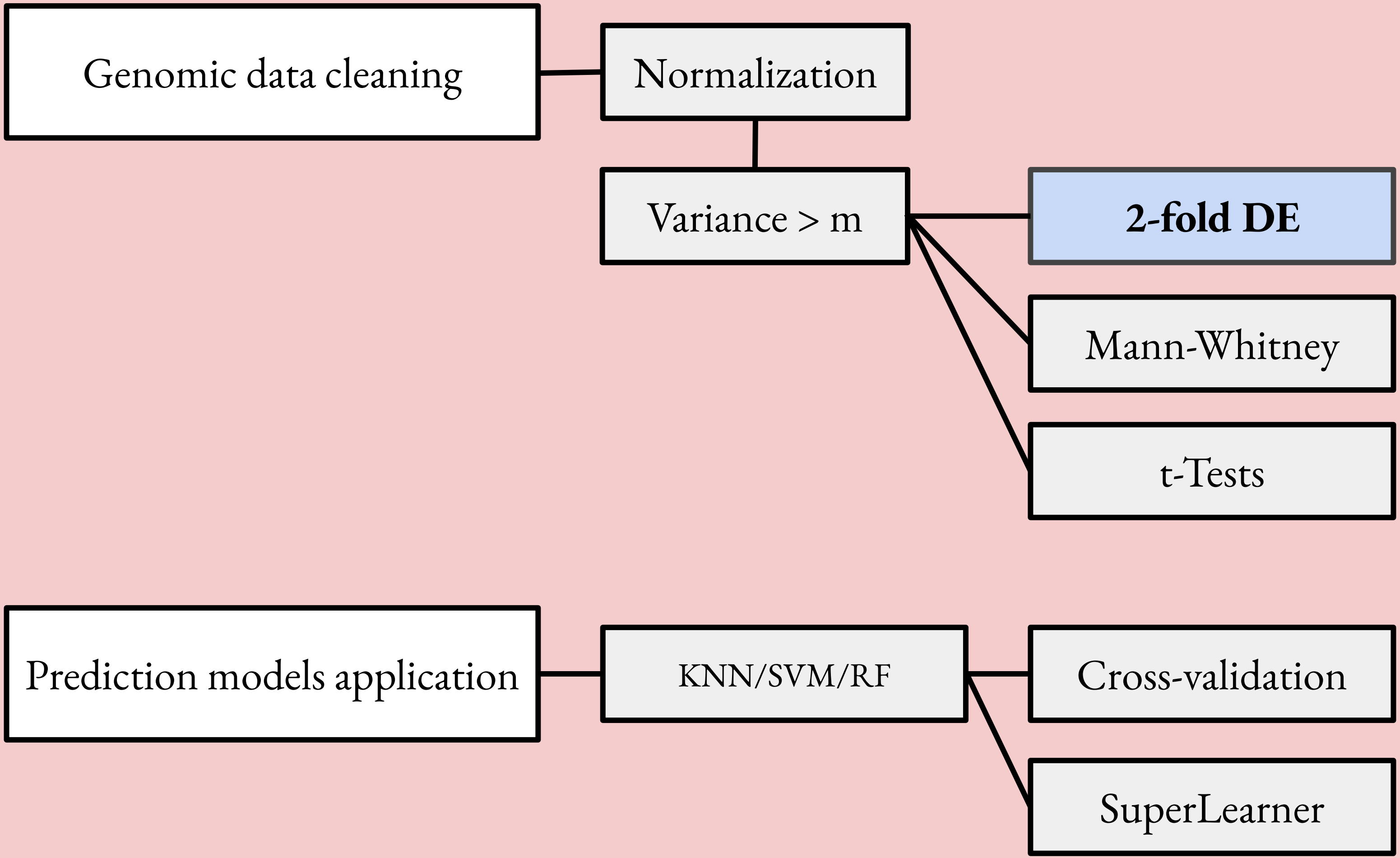
- UCSF Endometriosis Center recruits their patients (ongoing recruitment)
- NIH funded tissue bank for the research of endometriosis
  - Paid compensation
- Consultations do not require referral, but prefers it
- Center is a cross-disciplinary team
  - Gynecologists, therapists, pain management specialists
- Data are collected by specialists at the center
- Download available in CEL format online

Previous work.

Tamaresis, et al. 2014 [Link]

- n=148 (Of subjects, 34 had no endometriosis)
- Covariates
  - Menstrual cycle phase
  - Amount of pelvic pain
  - Genomic data
- Prediction methods
  - Initial 80/20 (holdout n=28)
  - K-fold cross validation
    - Done on remaining n=120
    - k=5-10
    - Decision tree classification
- Reported 90-100% accuracy


Genomic data cleaning
Normalization
Variance > m
2-fold DE
Mann-Whitney
t-Tests
Prediction models application
KNN/SVM/RF
Cross-validation
SuperLearner

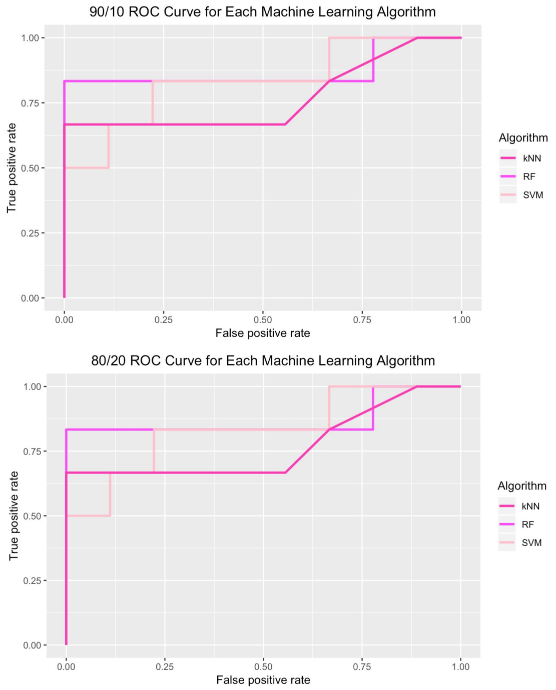

Genomic data cleaning
Normalization
Variance > m
2-fold DE
Mann-Whitney
t-Tests
Prediction models application
KNN/SVM/RF
Cross-validation
SuperLearner

Cross-validated KNN, SVM, and RF.

Table 1. Test accuracies for three separate k-fold machine learning approaches

	Test Accuracy		**AUC**	
	90/10 holdout validation	80/20 holdout validation	90/10 holdout validation	80/20 holdout validation
KNN	0.667	0.867	0.61	0.77
SVM	0.761	0.867	0.82	0.83
RF	0.733	0.867	0.83	0.87

Figure 1. ROC curve for each algorithm



Improving with SuperLearner.

Why SL?

- Outperforms individual algorithms
- Even when none of the algorithms in our SL library represents the true relationship between our predictors and outcome, SL will still asymptotically approximate the truth
- SL will only perform as well as the best weighted combination of candidate algorithms
- Avoids overfitting through cross-validation (CV.SL)

Performance of SL.

- Discrete SL vs Weighted SL
    - Both perform asymptotically as well as the oracle selected estimator
- The ratio of the dissimilarity of CV-selected estimator and truth and the dissimilarity of the oracle selected estimator and truth converges to 1

Applying SuperLearner.

Cross validation.

- 7-fold cross-validation on 90% of our full data
- 19 observations in each fold

Applying SuperLearner.

Cross validation.

- 7-fold cross-validation on 90% of our full data
- 19 observations in each fold

Loss functions.

- Non-negative least squares loss
- Non-negative log likelihood
- Area under the ROC curve (AUC)

Applying SuperLearner.

Cross validation.

- 7-fold cross-validation on 90% of our full data
- 19 observations in each fold

Loss functions.

- Non-negative least squares loss
- Non-negative log likelihood
- Area under the ROC curve (AUC)

Algorithms in SL library.

- SL.mean
- SL.bayesglm
- SL.svm
- SL.lda
- SL.glmnet
- SL.randomForest
- SL.nnet

Loss methods.

AUC loss.
Optimizes based on `cvAUC` predicted AUCs for each fold.

Non-negative log-likelihood loss.
Here is log loss. Resulting coefficients are non-negative.

$$-(y\log(p) + (1-y)\log(1-p))$$

Non-negative least squares loss.
For non-negative **x**,

$$\arg\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x} - \mathbf{y}||_2$$

SuperLearner Results.

Discrete SuperLearner.
Which algorithm does discrete SuperLearner use?

	NNLS	**NNLL**	**AUC**
Algorithm	SL.bayesglm	SL.bayesglm	SL.bayesglm
Weight	0.502	0.389	0.301
Accuracy	0.867	0.867	0.867

Table 2. Algorithms chosen by discrete SL and their associated weights.
Holdout row represents accuracy on 10% of data unused for training model.

SuperLearner Results.

All methods chose Bayesian GLM with the default hyperparameters from library(arm).

	NNLS	NNLL	AUC
Algorithm	SL.bayesglm	SL.bayesglm	SL.bayesglm
Weight	0.502	0.389	0.301
Holdout	0.867	0.867	0.867

Table 2. Algorithms chosen by discrete SL and their associated weights. Holdout row represents accuracy on 10% of data unused for training model.

SuperLearner Results.

All methods misclassified the same two subjects.

	NNLS	**NNLL**	**AUC**
Algorithm	SL.bayesglm	SL.bayesglm	SL.bayesglm
Weight	0.502	0.389	0.301
Holdout	0.867	0.867	0.867

Table 2. Algorithms chosen by discrete SL and their associated weights.
Holdout row represents accuracy on 10% of data unused for training model.

SuperLearner Results (AUC Loss).

Table 3. SuperLearner Summary Output

Algorithm	**Risk**	**Coef**
SL.mean	0.65	0.12
SL.svm	0.18	0.02
SL.glmnet	0.20	0.12
SL.randomForest	0.19	0.12
SL.lda	0.18	0.12
SL.nnet	0.43	0.21
SL.bayesglm	0.14	0.30

Discrete SL under AUC chose Bayes GLM with default hyperparameters.

SuperLearner Performance (AUC Loss).

Table 4.
CV.SuperLearner Summary Output

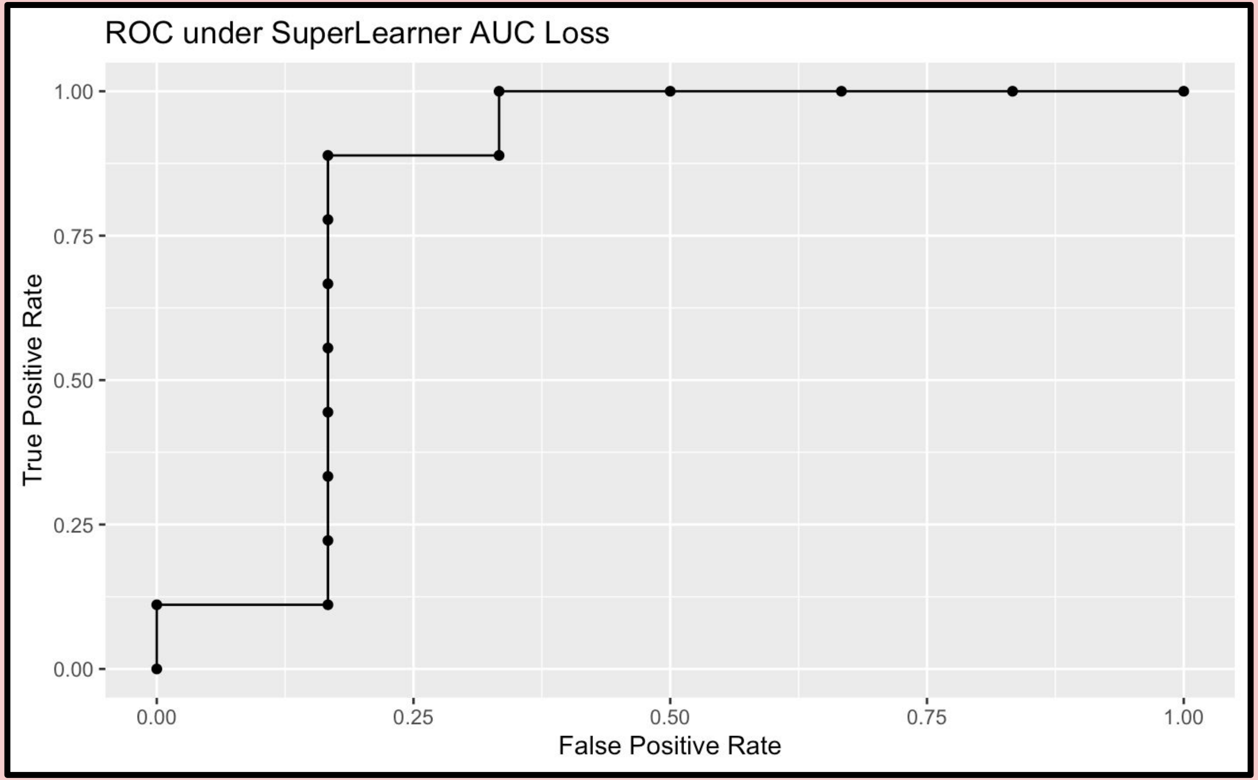
Algorithm	**Average**	**Min**	**Max**
SuperLearner	0.867	0.795	0.964
Discrete SL	0.837	0.711	0.954
SL.mean	0.500	0.500	0.500
SL.svm	0.853	0.755	0.952
SL.glmnet	0.813	0.705	0.976
SL.randomForest	0.844	0.715	0.928
SL.lda	0.845	0.711	0.952
SL.nnet	0.662	0.500	0.917
SL.bayesglm	0.874	0.761	0.998

SuperLearner Performance (AUC Loss).

Table 5. Discrete SL Selection Per Fold

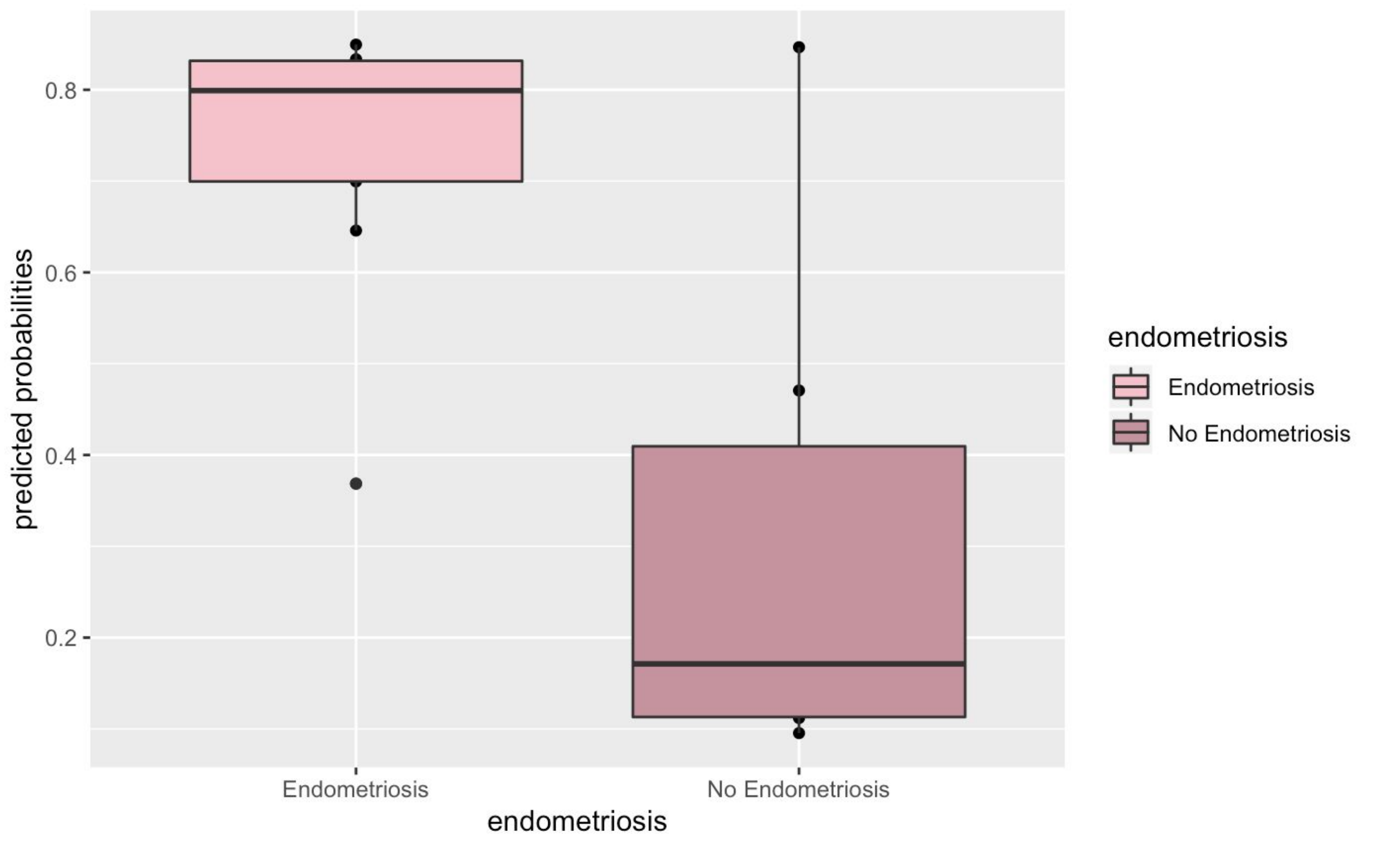
Fold	Discrete SL	Notes
1	-	Even weights on all algorithms. LDA had lowest risk.
2	LDA (0.26)	Close weights (0.13/0.12) on the rest.
3	LDA (0.43)	Bayes GLM and GLM next most weighted.
4	LDA (0.18)	Rest weighted closely.
5	Bayes GLM (0.54)	GLM next most weighted (0.2). NN is weighted 0.
6	Bayes GLM (0.46)	RF next weighted (0.22)
7	LDA (0.31)	GLM has 0 weight. Rest are evenly weighted.

Figure 2.
AUC=0.8333



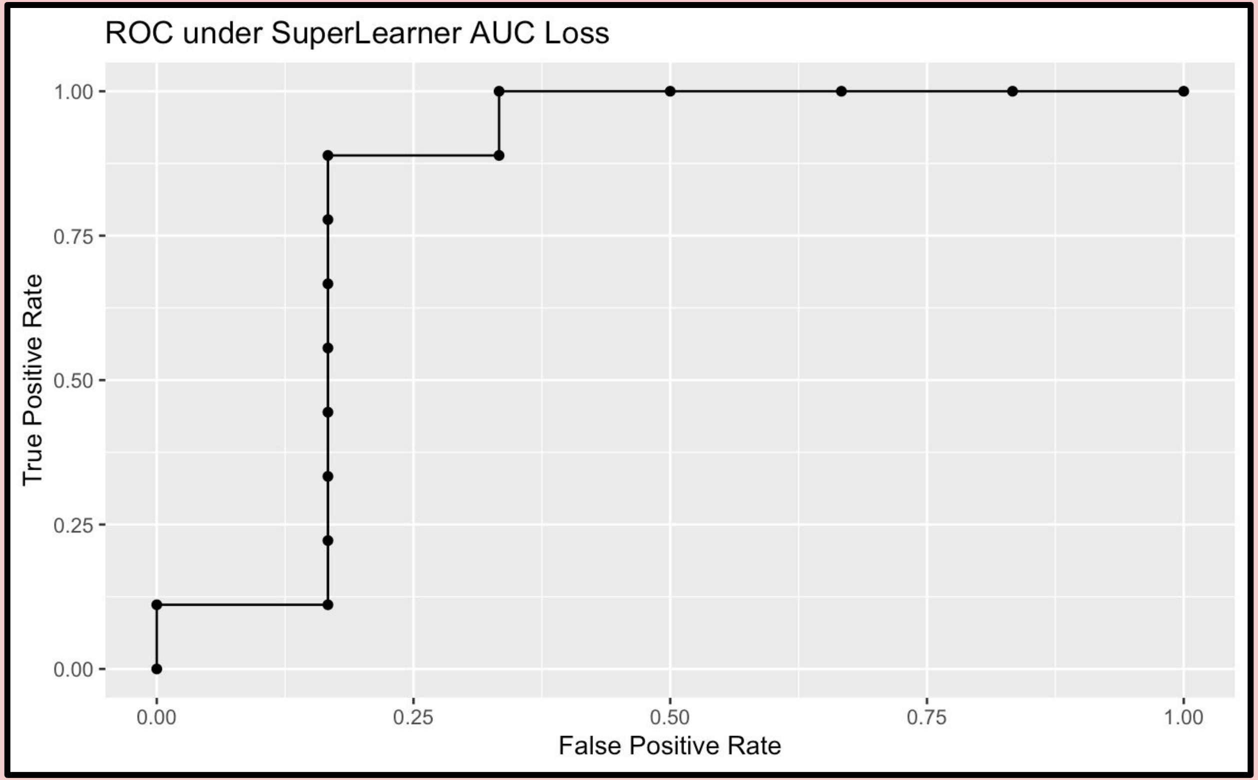
An endometriosis patient has a higher assigned probability than a randomly chosen "healthy" patient 83.33% of the time.

Figure 3.
Boxplot of predicted probabilities



These boxplots show how well the classes were separated by probability.

Figure 2.
AUC=0.8333



An endometriosis patient has a higher assigned probability than a randomly chosen "healthy" patient 83.33% of the time.

Figure 4.
AUC=0.8518

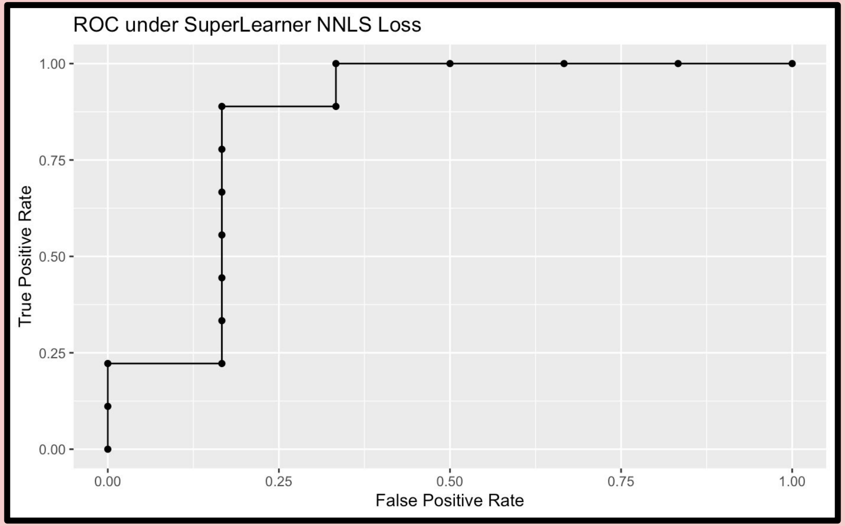
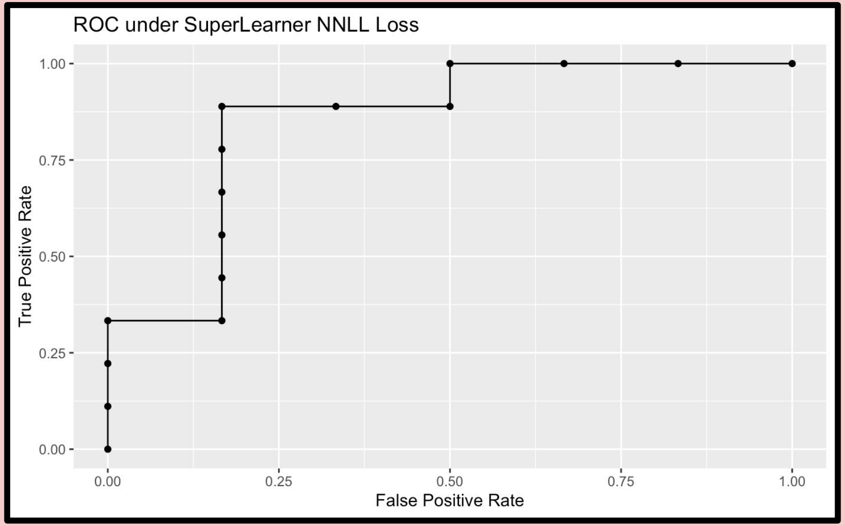
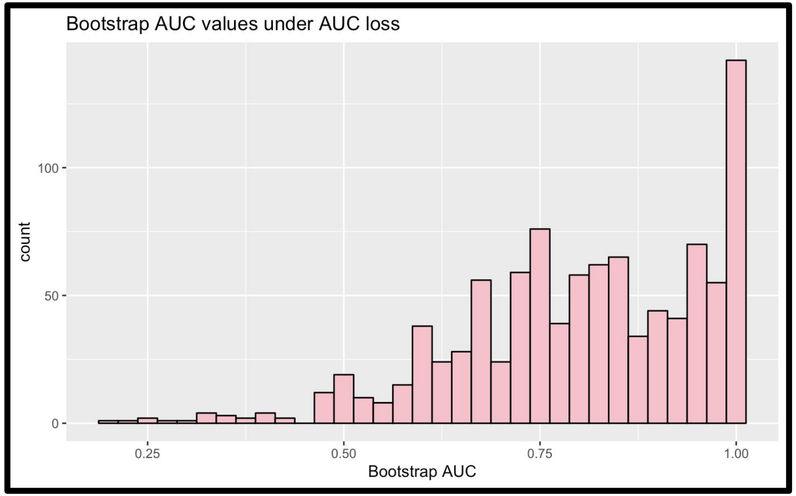


Figure 5.
AUC=0.8518

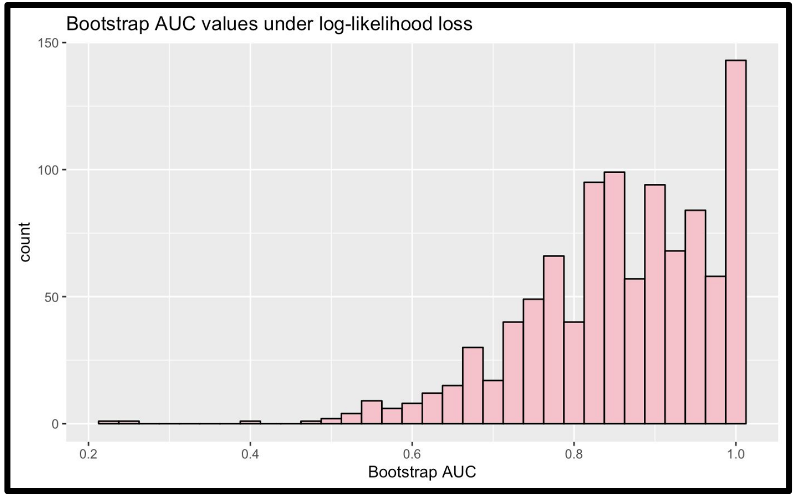


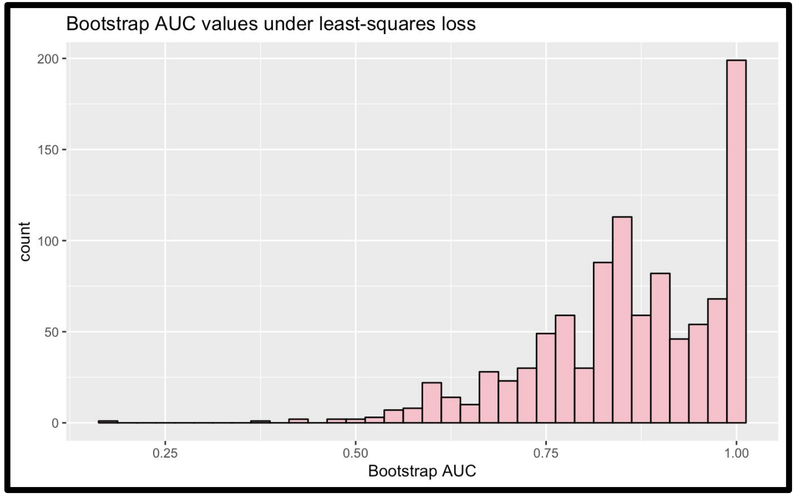
An endometriosis patient has a higher assigned probability than a randomly chosen "healthy" patient 85.18% of the time.



Figures 6-8.

- 1,000 bootstraps of our final validation set (we left out 10% to begin with for predictions, n=15) to study AUC behavior
- 95% quantile method CIs
  - AUC: (0.5, 1)
  - NNLL: (0.61, 1)
  - NNLS: (0.58, 1)





Influence Curve Based Confidence Intervals.

Running cvAUC. Because each of the loss methods misclassified the same subjects, their IC-based 95% confidence intervals were all computed to be the same.

Our below results match the significance found from the bootstrap confidence intervals, but are found using robust methods.

cvAUC	0.8611
SE	0.1730
95% CI	(0.552, 1.000)

Future work.

Our SuperLearner library has ways to grow! When you add more prediction algorithms, SL will only perform better.

We would want to fit algorithms on DE genes selected by other methods for comparison.

Adjust hyperparameters/tuning parameters in SL with background genomics knowledge.

References.

Van der Laan, Mark. Polley, Eric. Hubbard, Alan. Super Learner. [Paper]

LeDell, Van der Laan, Petersen. AUC-Maximizing Ensembles through Metalearning. [Paper]

Koidl, Kevin. Loss Functions in Classification Tasks. Trinity College Dublin. [PDF]

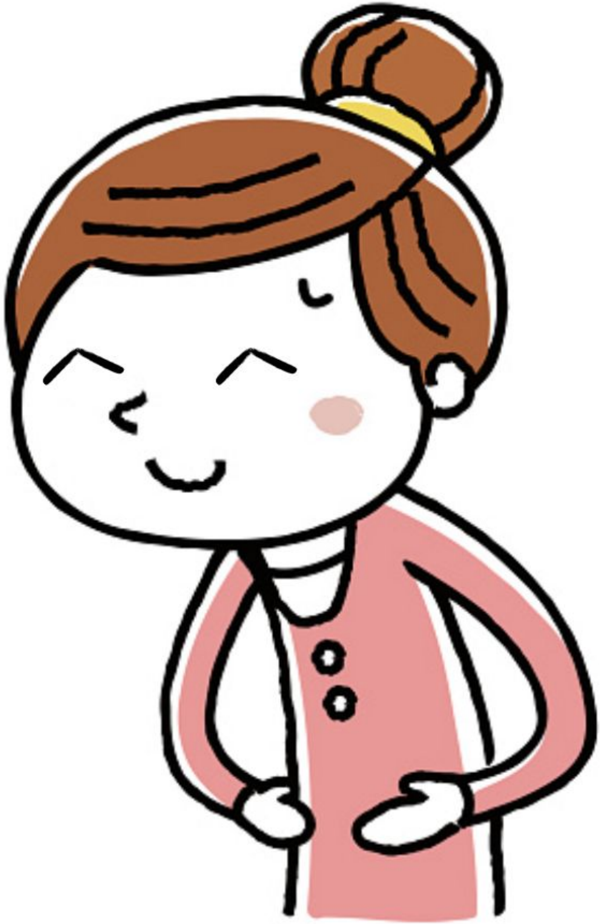
Steyerberg, Ewout. Clinical Prediction Models. Leiden University Medical Centre. 2009. [Text]

Godoy, Daniel. Understanding binary cross-entropy. [Article]

Week in Life with Endometriosis. [Video]

Endometriosis. Health Engine. [Article]

Endometriosis - Diagnosis and Treatment. Mayo Clinic. [Article]



Thanks for
your attention.