Leveraging electronic health records for predictive modeling of surgical complications

Grant Weller

ISCB 2015
Utrecht NL
August 26, 2015

Collaborators: David W. Larson, MD; Jenna Lovely, PharmD, RPh; Berton Earnshaw, PhD; Marianne Huebner, PhD

The promise of EHR in clinical research

Opportunity:

$$\begin{gathered}\text{Large amounts of rich observational data}\\ + \\ \text{modern statistical \& machine learning methods}\\ \stackrel{?}{=} \\ \text{new discovery of clinical practice improvements, decision}\\ \text{support, personalized treatments, etc.}\end{gathered}$$

Potential pitfalls:

- Messy data (missingness, censoring, mixed types, etc.)
- Variation in reporting standards (suitable for research?)
- Heterogeneity in compliance measures
- Confounding

Colorectal surgery complications

We explore opportunities for EHR in prediction and detection of complications of colorectal surgery.

- Mayo CRS: $\sim$2,000 procedures per year; 10 faculty
- Diagnoses: colorectal cancer, colitis, Crohn's, etc.

Focus on three complications:

- Surgical site infection
- Bleeding (intraop / postop)
- Ileus (partial bowel obstruction / use of NG tube)

Obtained data from years 2010–2013.

Summary points

- Significant data preparation work needed to use EHR for risk prediction
- Which method(s) to use? Exploring possibilities...
- There are opportunities to inform clinical practice

Surgical case definitions

We constructed a set of procedure-centered variables from raw data based on recorded operation start and stop times.

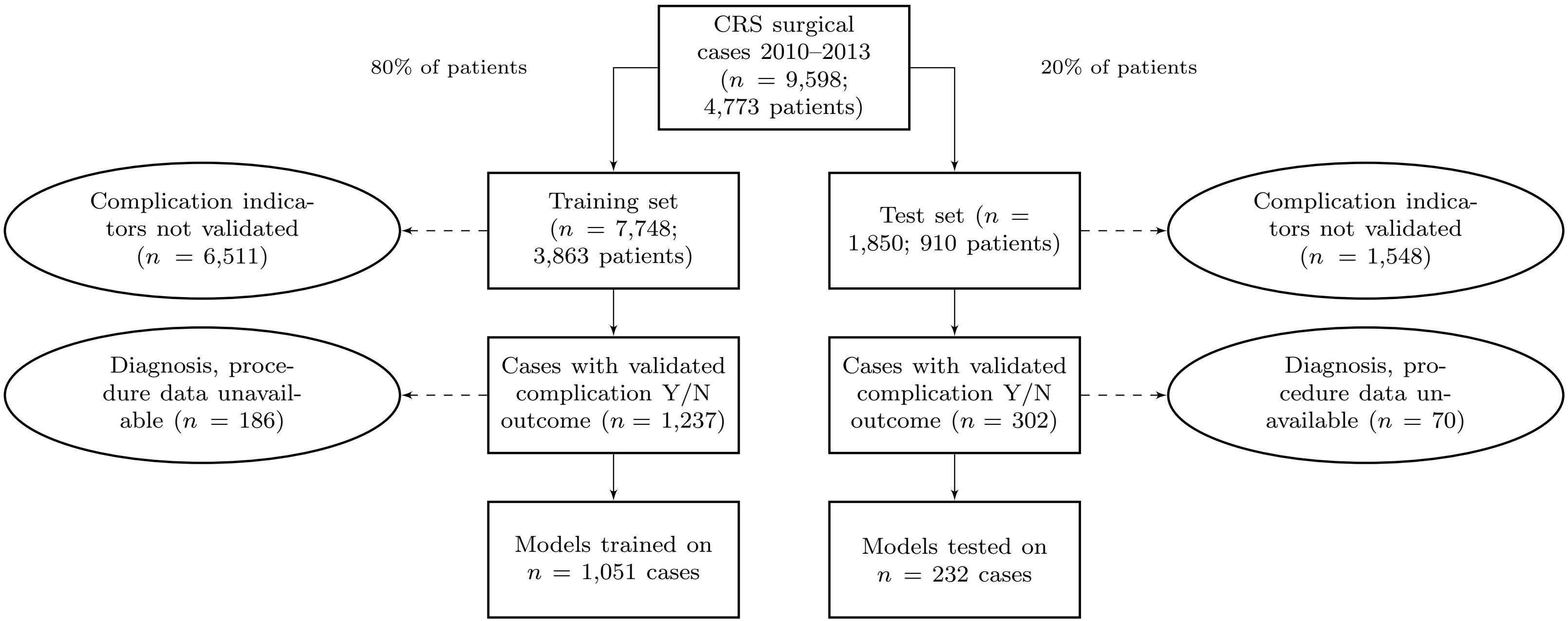
- Background / demographics
- Prior surgery history
- Labs taken within 72 hours preop
- Diagnosis and procedure information
- Post-surgical monitoring data

Constructed about 200 total input features for modeling.

- Outliers removed
- Missing data imputed via Bayesian regression model

Preparing data for modeling

Most complication outcomes undocumented in our data.



For comparison, we also obtain clinical "rule engine" data

- Deterministic SQL-based rule set
- Generates yes/no/unknown

Summary statistics

Complication rates:

Complication	Training	Test
SSI	9.7%	6.0%
Bleeding	13.6%	13.9%
Ileus	11.5%	10.3%

Related work (see poster of M. Huebner):

- Complication rates vary by diagnosis / procedure
- Co-occurrence rates vary as well

Summary points

- Significant data preparation work needed to use EHR for risk prediction
- Which method(s) to use? Exploring possibilities...
- There are opportunities to inform clinical practice

Exploration of prediction methods

We explore the predictive performance of several statistical and machine learning methods:

- Regularized (LASSO) logistic regression
- Random forests
- Naive Bayes
- Support vector machines
- Boosted classifiers

We construct four models for each method:

- Pre-surgery
- Post-op days 0, 1, 2

See Hastie et al. (2009) for a good overview.

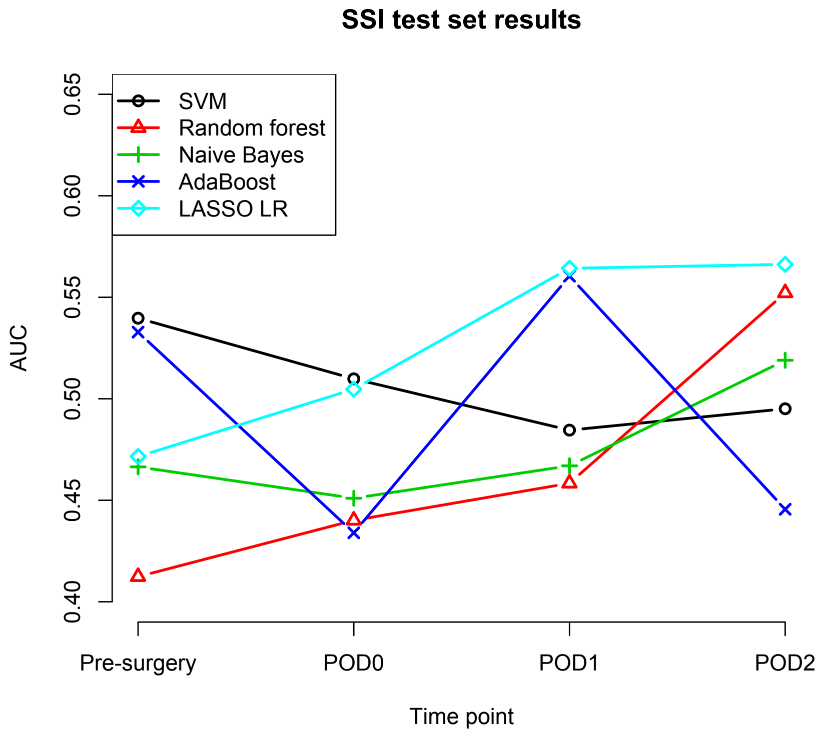
Evaluation

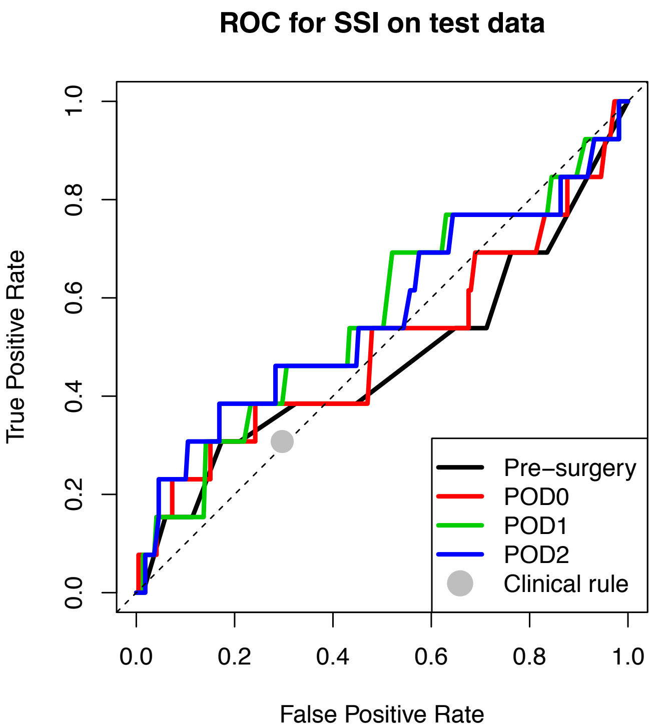
We investigate several aspects of prediction methods:

- Area under receiver operating characteristic (ROC) curve
- Discovery of most relevant features for prediction
- Comparison with deterministic clinical rule
- Evaluation at different data collection points (pre-op, POD 0–2)
- Comparison of response surfaces

Test set results: SSI

AUC at different timepoints (left); LassoLR results (right)





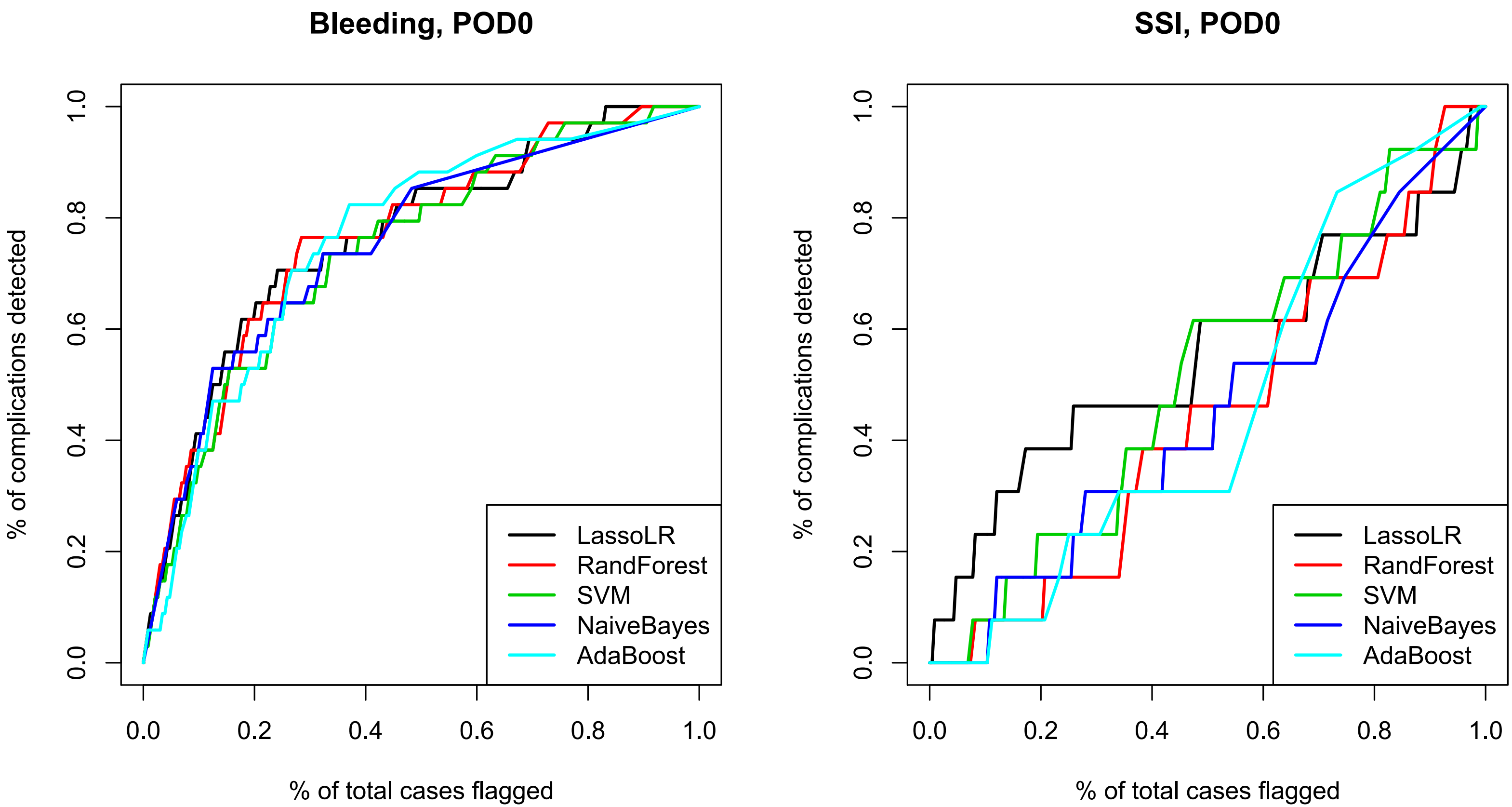
Note small number of test cases.

Summary points

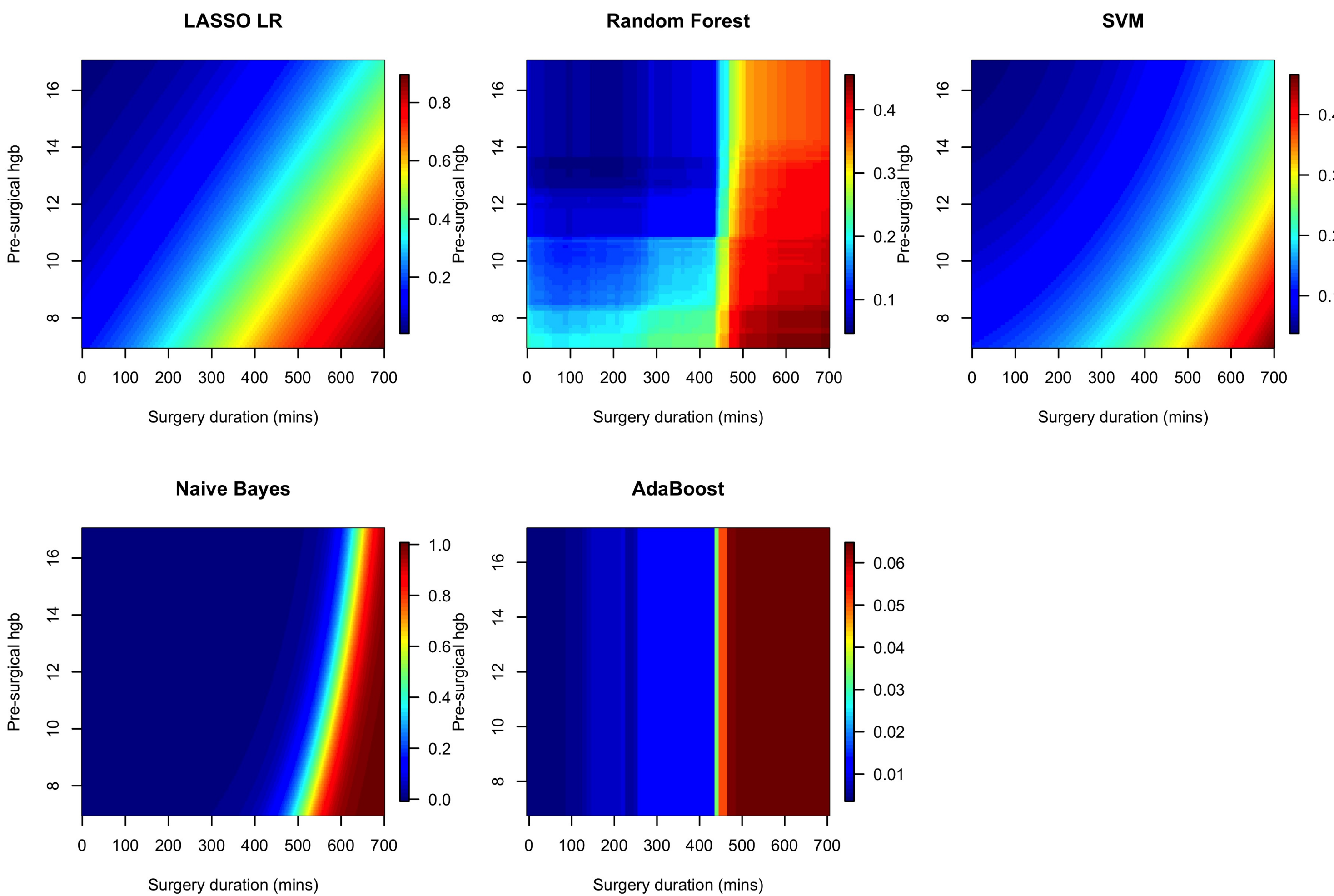
- Significant data preparation work needed to use EHR for risk prediction
- Which method(s) to use? Exploring possibilities...
- There are opportunities to inform clinical practice

Implications for practitioners

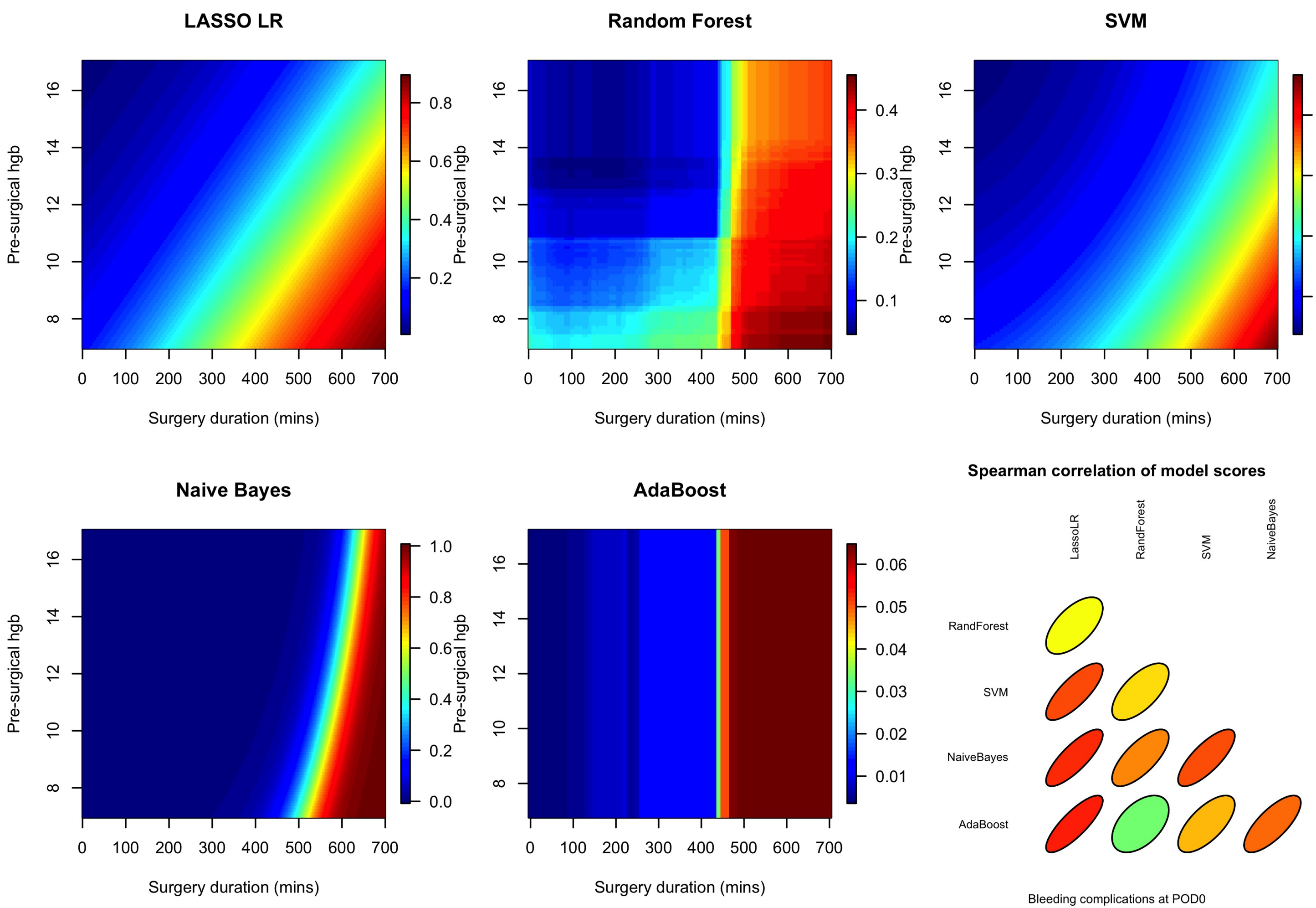
Examining prioritization and resource trade-offs.



Example of prediction: bleeding, POD0



Example of prediction: bleeding, POD0



Model comparisons

The differences in models' predictive ability were small considering the small size of the training set.

- Linear methods with regularization perform as well as more complex approaches here
- Value as a data mining tool to identify 'movable' predictors

Clinically relevant findings:

- Duration of surgery is strongest predictor of complications
- Wound type, existing conditions also risk indicators
- Probabilistic approach comparable to clinical rule

Future possiblities

Further development of probabilistic risk calculators from observational data is a promising area of research.

- Time to event modeling (Wolfson et al., 2015)
- Predictions with dynamic covariates
- Implementation in decision support tools

All the above require standardized, well-documented EHR data, including follow-up.

Development needed in both data collection / architecture and in inferential methodology.

References and contact

Email: GrantWeller@savvysherpa.com
Web: http://grantbweller.com

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, volume 2. Springer.

Wolfson, J., Bandyopadhyay, S., Elidrisi, M., Vazquez-Benitez, G., Musgrove, D., Adomavicius, G., Johnson, P., and O'Connor, P. (2015). A naive Bayes machine learning approach to risk prediction using censored, time-to-event data. *Statistics in Medicine*, 34(21):2941–2957.