CLINICAL PREDICTION MODELS

- Katie Colborn, PhD
- Krithika Suresh, PhD

COPRH Con

Colorado Pragmatic Research in Health Conference

ACCORDS
UNIVERSITY OF COLORADO
CHILDREN'S HOSPITAL COLORADO

Colorado Clinical and Translational Sciences Institute (CCTSI)
UNIVERSITY OF COLORADO DENVER | ANSCHUTZ MEDICAL CAMPUS

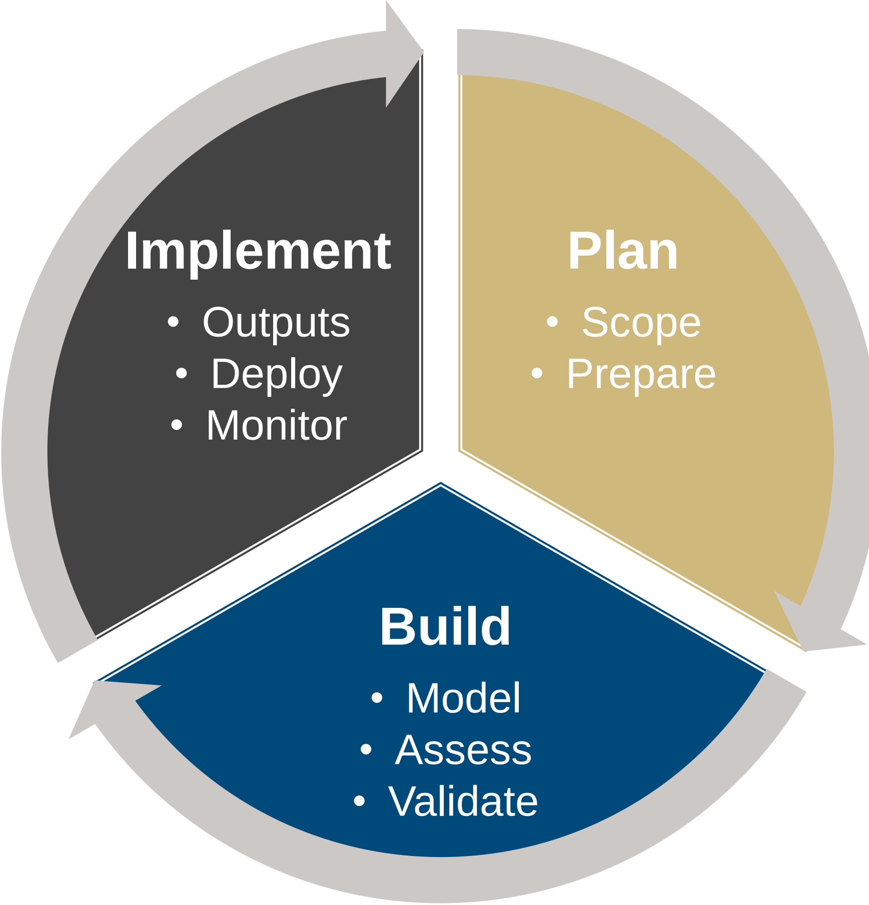
OUTLINE

- The Modelling Process
- Building Prediction Models
- Assessment and Validation

“This just isn’t doing it for me. Could we go back to using the crystal ball?”

MODELLING PROCESS

- **Plan**
  - Identify goal
  - Assemble the data
  - Handle data issues
- **Build**
  - Identify model/technique
  - Identify metrics for assessment
  - Internal validation
  - External validation
- **Implement**
  - Identify outputs
  - Make model available
  - Continued monitoring



Application

PLAN: SCOPE

- **Goal:** “Predict Prostate cancer specific mortality (PCSM) in patients with prostate cancer”
  - PCSM or Overall Survival?
- **Existing literature:**
  - AJCC criteria for prediction models
  - Alternative prediction models and methods
  - What is the gap?

Application

PLAN: PREPARE

- **Data:** Obtained patient data from 10 centers (n=~20,000)
- **Data issues:**
  - Missing data: To impute or not?
  - Variable coding
  - Inconsistencies
- **Training and Validation:**
  - Split based on center?
  - Random split?
  - Percentage split?
- **Transparency**
  - Circulate model building plan
  - Hold validation data externally

BUILD

- Model/Technique
- Assess
- Validate

"The boss wants me to create a computer algorithm that can convert hindsight into foresight."

BUILD: MODEL/TECHNIQUE

- Regression Methods (stepwise, penalized)
- Tree-based Methods
- Random Forest
- Other methods
  - Neural nets
  - Deep learning
  - Support Vector Machines
  - Boosting

Regression Methods

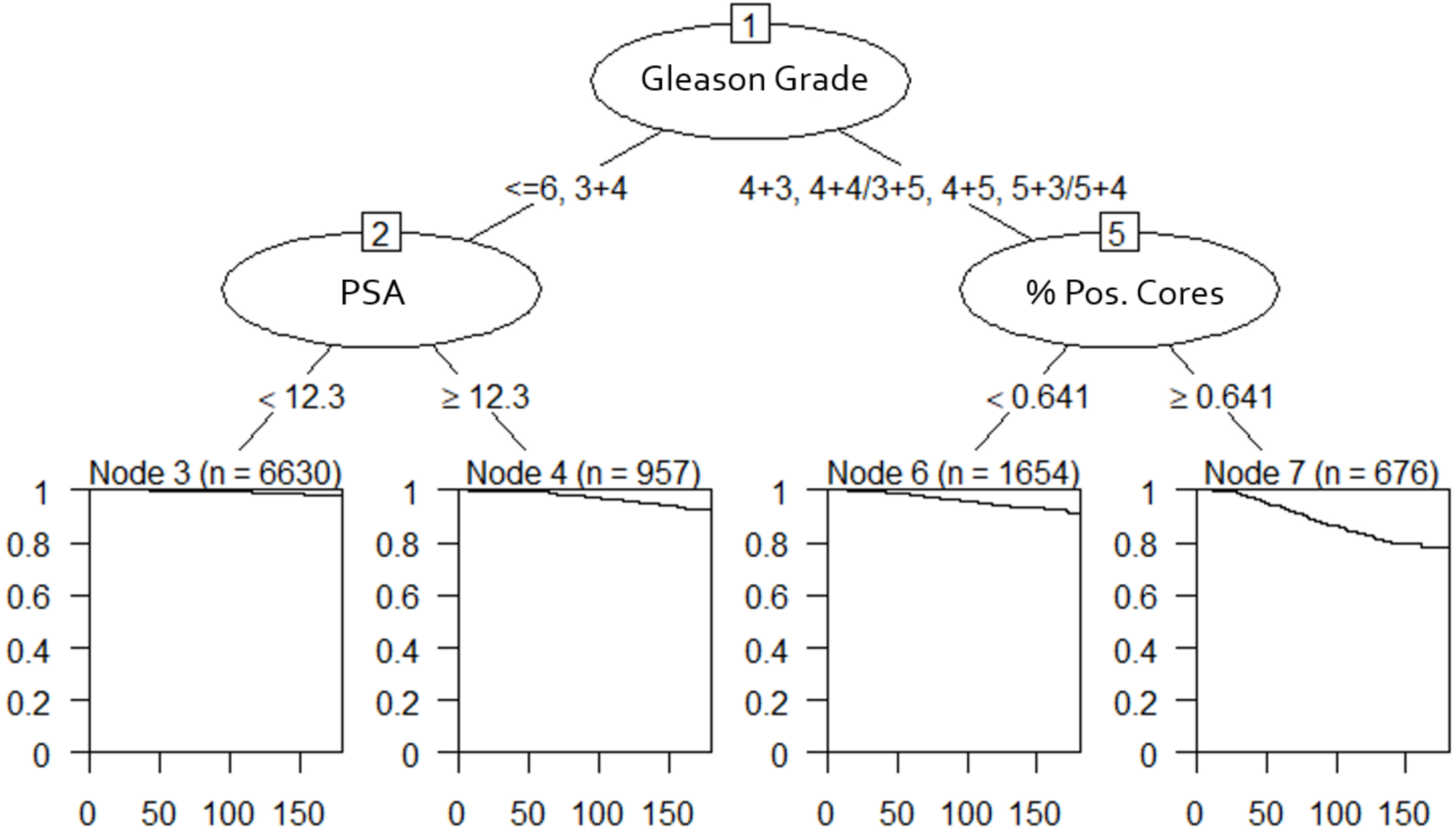
- **Stepwise**
- **Penalized regression** (Lasso, Ridge, Elastic net)
- **Advantages:**
  - Good prediction (all penalized regression methods)
  - Variable selection and prediction (Lasso and Elastic net)
- **Disadvantages**
  - No variable selection (Ridge)
  - Inference is more difficult

Application

Tree-based Methods

Algorithm:

- Start will all patients in top node
- For every variable, evaluate every possible binary split
- Choose the best variable/threshold combination
- Repeat for all terminal nodes until no split is possible
- Stop when terminal node size is too small



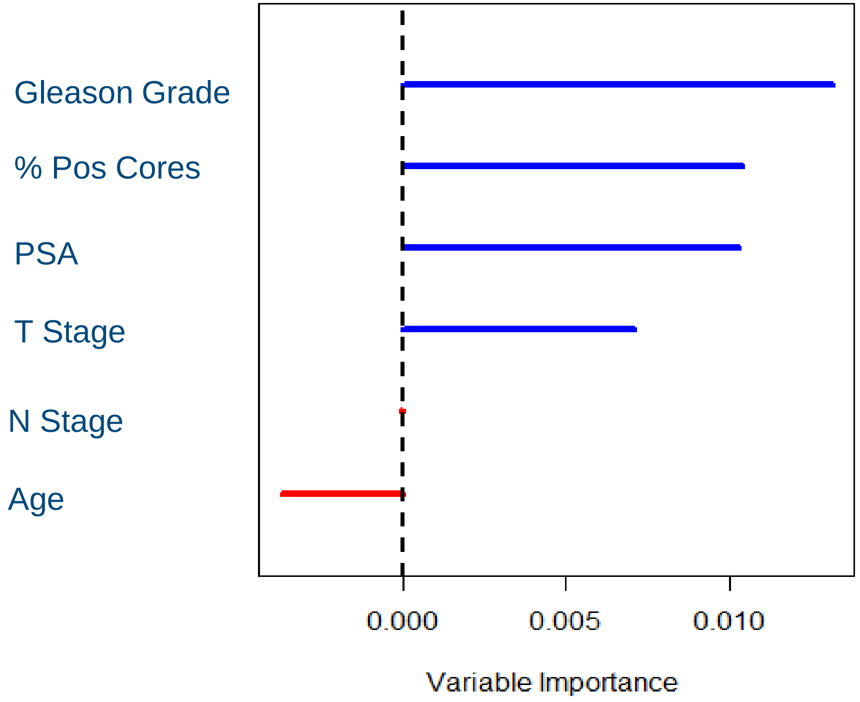
Tree-based Methods

- **Advantages:**
  - Simple, easy to understand and use
  - Naturally identify thresholds (when they exist, which is not always)
  - Naturally identify interactions
- **Disadvantages:**
  - Poor prediction
  - Unstable (small changes to the data can result in large changes to the tree)

Application

Random Forest

- A forest is comprised of many trees
- “Ensemble” method
- **Advantages:**
  - Can be used for both regression and classification tasks
  - Easy to view relative importance of input variables
  - Won’t overfit (with enough trees)
- **Disadvantages:**
  - Very slow with large number of trees
  - Ineffective for real-time predictions
  - Not a descriptive tool



Which Algorithm to Use?

1. The size, quality, nature of your data
2. What you want to do with your data
3. The available computation time

- Match the method to your goal
- **Goal:** Parsimonious model
  - **Method:** Lasso (NOT Ridge or Random Forest)
- **Goal:** Interpretable model
  - **Method:** Elastic net, Survival Tree, Regression
- Choice of model can be less important than getting the basics right (confounding, censoring, etc.)

BUILD: ASSESS

- Identify the metrics to assess predictive performance
- **Discrimination**
  - Area under the ROC curve (AUC), Concordance-index (C-index)
- **Calibration**
  - Calibration plots
- **Overall measures of prediction error**
  - Brier Score
- Provide comparison
  - To null model (with no covariates)
  - To alternate models

BUILD: VALIDATE

- **Apparent:** Performance of model on data used to develop the model
  - Will get optimistic estimates of performance
- **Internal:** Performance on population underlying the sample (“reproducibility”)
  - Test/training set, cross-validation, bootstrap
- **External:** Performance on related but slightly different population
  - Different centers, years, therapies, variable definitions

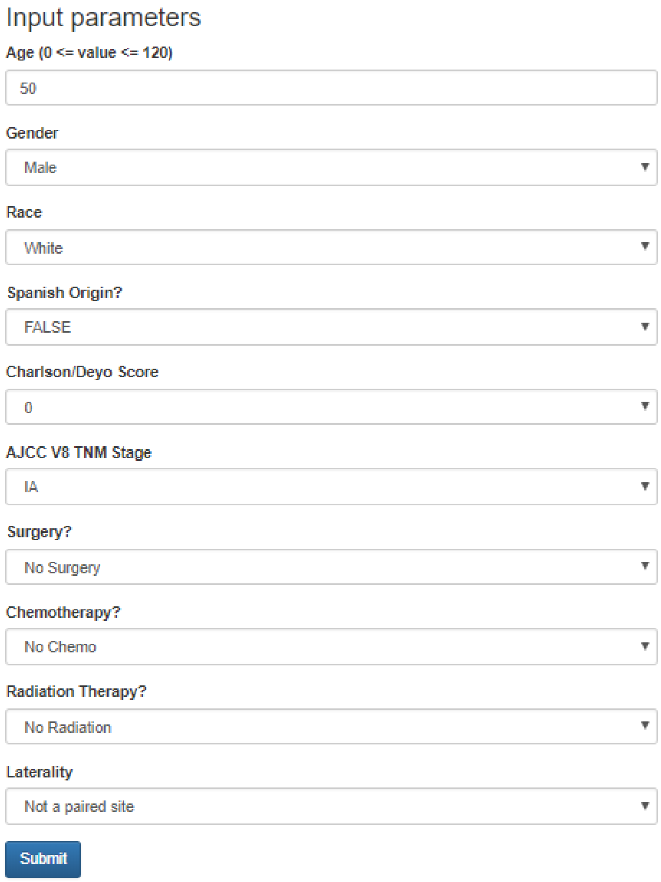
IMPLEMENT: OUTPUTS

- Nomograms
- Point estimates
- Tree-based methods
- Score charts
- Web-based applications (R Shiny apps)

IMPLEMENT: DEPLOY & MONITOR

Application

R Shiny App (Small Cell Lung Cancer)



Input parameters

Age (0 <= value <= 120)

50

Gender

Male ▾

Race

White ▾

Spanish Origin?

FALSE ▾

Charlson/Deyo Score

0 ▾

AJCC V8 TNM Stage

IA ▾

Surgery?

No Surgery ▾

Chemotherapy?

No Chemo ▾

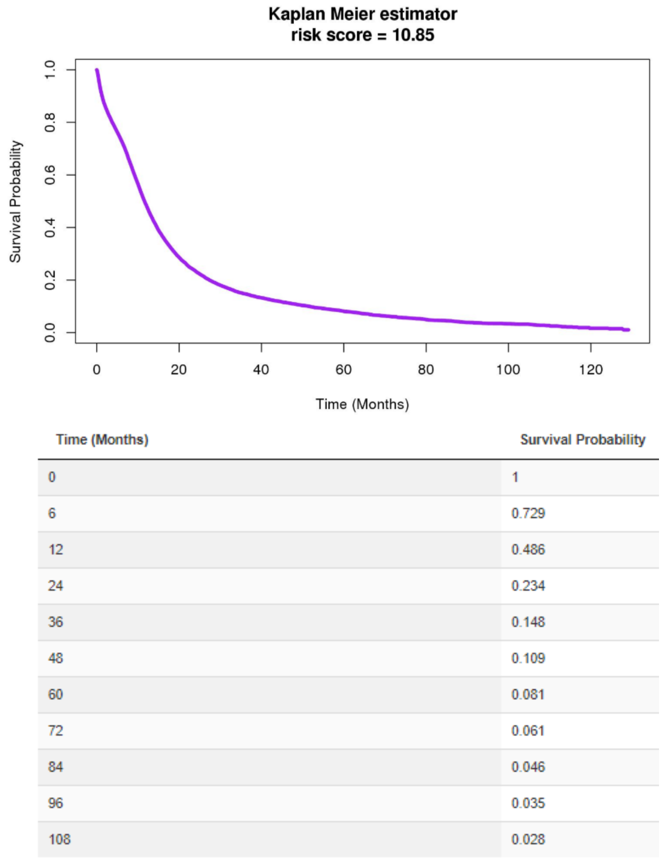
Radiation Therapy?

No Radiation ▾

Laterality

Not a paired site ▾

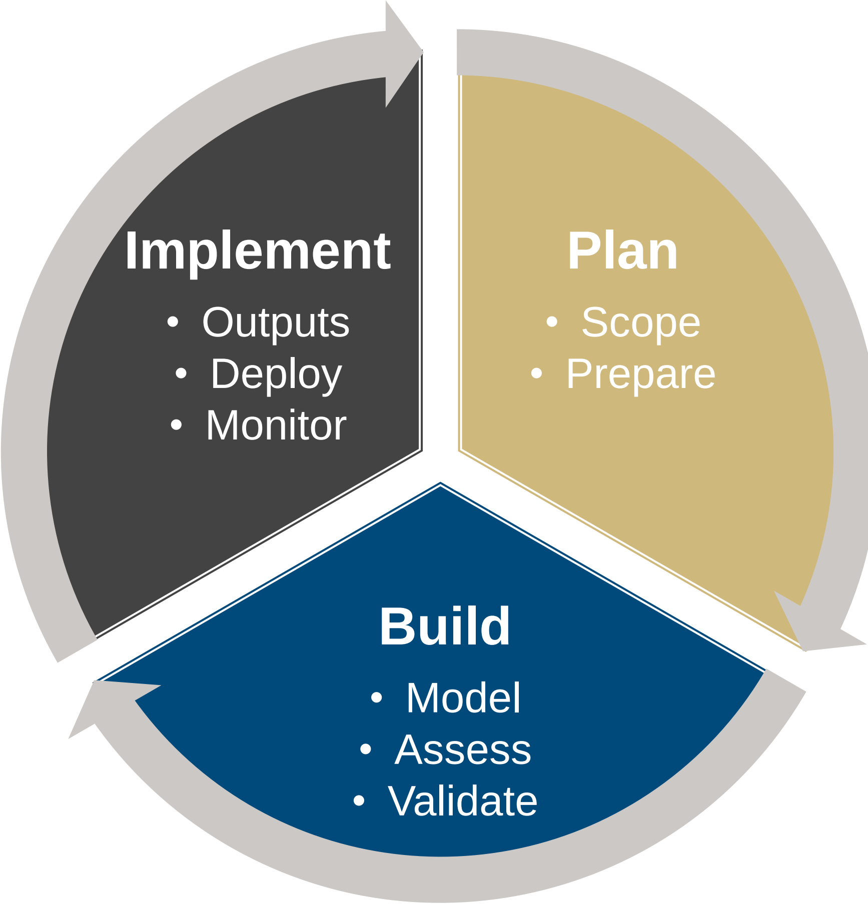
Submit



Time (Months)	Survival Probability
0	1
6	0.729
12	0.486
24	0.234
36	0.148
48	0.109
60	0.081
72	0.061
84	0.046
96	0.035
108	0.028

http://lce.biohpc.swmed.edu/lungcancer/sclc_nomogram/index.php

SUMMARY



Plan

- Scope
- Prepare

Build

- Model
- Assess
- Validate

Implement

- Outputs
- Deploy
- Monitor