Use of Classification and Regression Trees Analyses in Research on Web-based Health Behavior Change Programs

Brian G. Danaher, PhD briand@ori.org
John R. Seeley, PhD
Herbert H. Severson, PhD
H. Garth McKay, PhD

Oregon Research Institute

Presentation to the Third Meeting of the International Society for Research on Internet Interventions
October 12, 2007, University of Virginia, Charlottesville, US

Supported by grants from the National Cancer Institute

It's in the Name

- Recursive partitioning (RP)
- Classification trees
- CART (C&RT): Classification & regression trees
- CHAID: Chi-square Automatic Interaction Detector
- Signal detection analysis
- Machine learning
- Data mining

What RP does

- Uses a tree metaphor
- Recursive partitioning process
  - Predicts a defined outcome (DV)
  - Divides the sample into increasingly smaller subsamples
    - Into nodes or groups
    - Seeks perfect splits (node homogeneity)
  - Based on whether a predictor is above a cut-off point
    - Choice of predictor and cutoff value based on purity
    - Decision rules define degree of acceptable impurity

tobacco abstinence @ 3 & 6 mos. self-efficacy change and program exposure

(mean of Z-score transformation of # visits and overall duration)

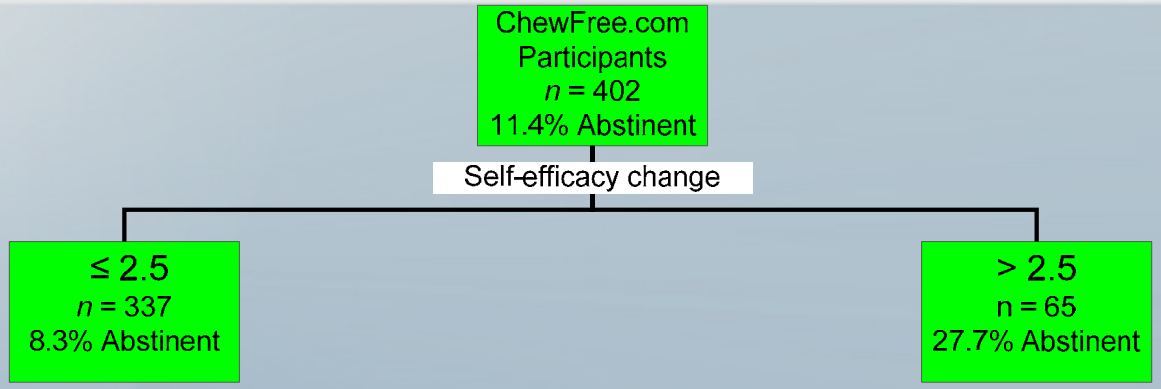
ChewFree.com Participants
n = 402
11.4% Abstinent

Self-efficacy change = "Confidence not using ST 1 year from now"
0= *not at all*, 2= *somewhat*, 4= *completely*
Baseline to 6-weeks

Program exposure = mean of Z-score transformation of # visits and overall duration

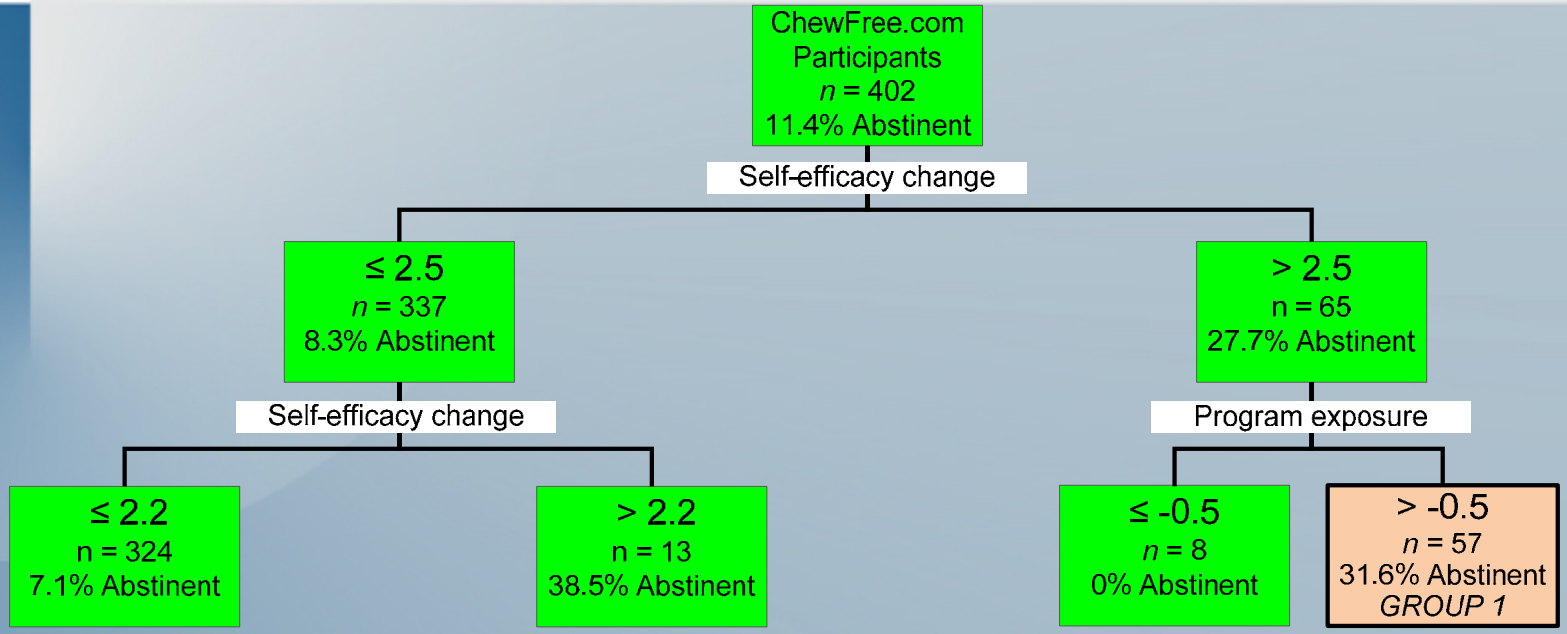
tobacco abstinence @ 3 & 6 mos. self-efficacy change and program exposure

ISRII



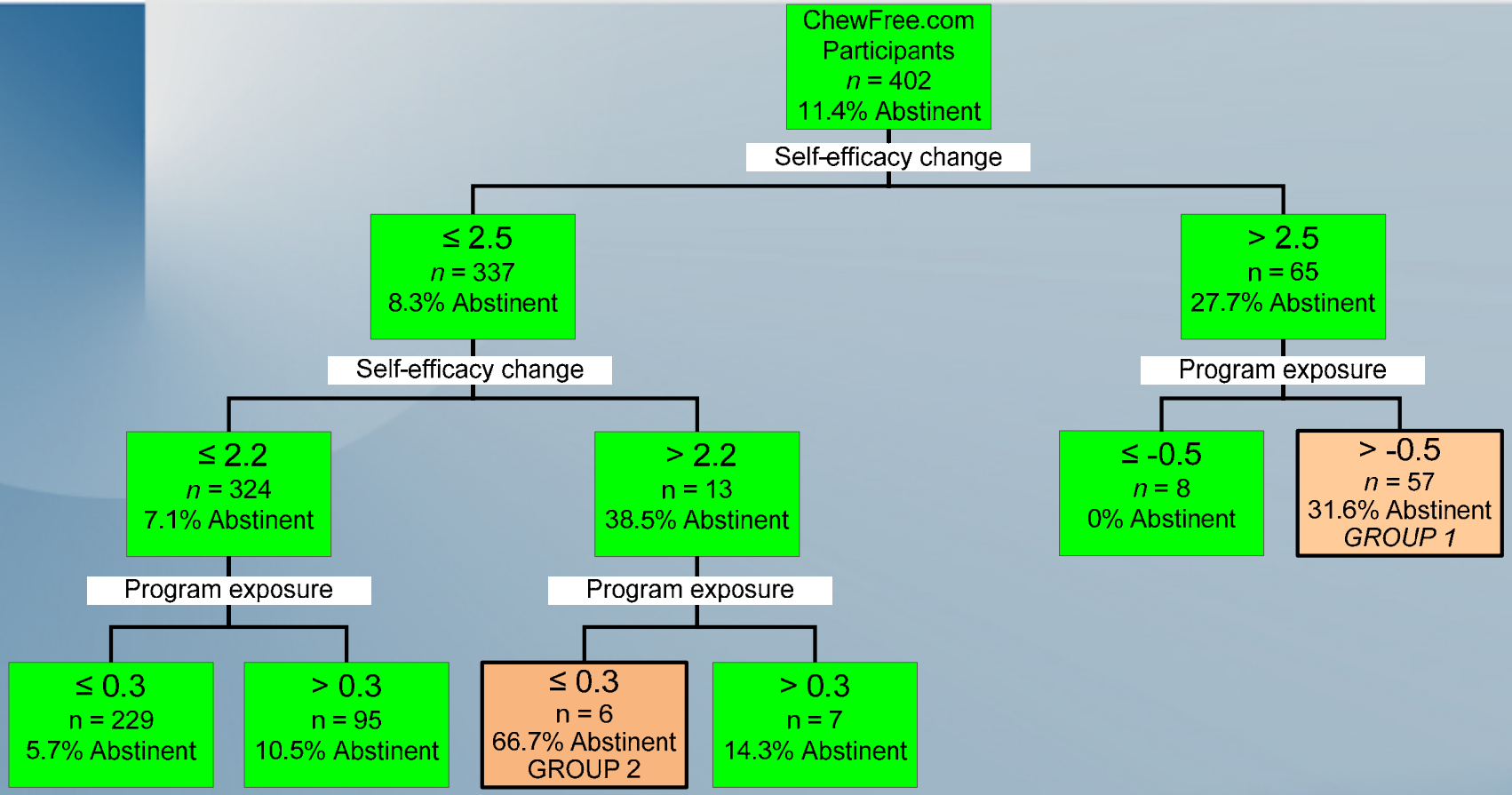
tobacco abstinence @ 3 & 6 mos. self-efficacy change and program exposure

ISRII



tobacco abstinence @ 3 & 6 mos. self-efficacy change and program exposure

ISRII



Self-efficacy change ≤ 2.5 AND
Self-efficacy change > 2.2 AND
Program exposure ≤ 0.3

Self-efficacy change > 2.5 AND
Program exposure > -0.5

tobacco abstinence @ 3 & 6 mos. self-efficacy change and program exposure

		Actual				
		Yes	No	Total		
Predicted	Yes	22	41	63	0.35	positive predictive value
	No	24	315	339	0.93	negative predictive value
	Total	46	356	402	0.84	predictive acccuracy
		0.48 sensitivity	0.88 specificity			

Considerations

- Node purity
- Stopping rules
  - Degree of improvement required
  - Minimal *N* for each parent/child
- Misclassification costs (weights)
  - Specificity and sensitivity
- Prior probability
- Competing splits
- Validation

More Considerations

- Content area understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

Exploration & Prediction

- Exploration can use entire sample
- Prediction requires validation methods
  - Split-sample (train, test, validate)
  - *k*-fold (bootstrapping)
    - We used 10-fold validation in our example

Logistic Regression

- LR: Ss homogeneous in outcome but heterogeneous in risk predictors
- RP: Ss homogeneous within a node both in outcome and risk predictors
  - More useful for designing tailored interventions for subgroups of high-risk individuals

RP in Web-based Interventions

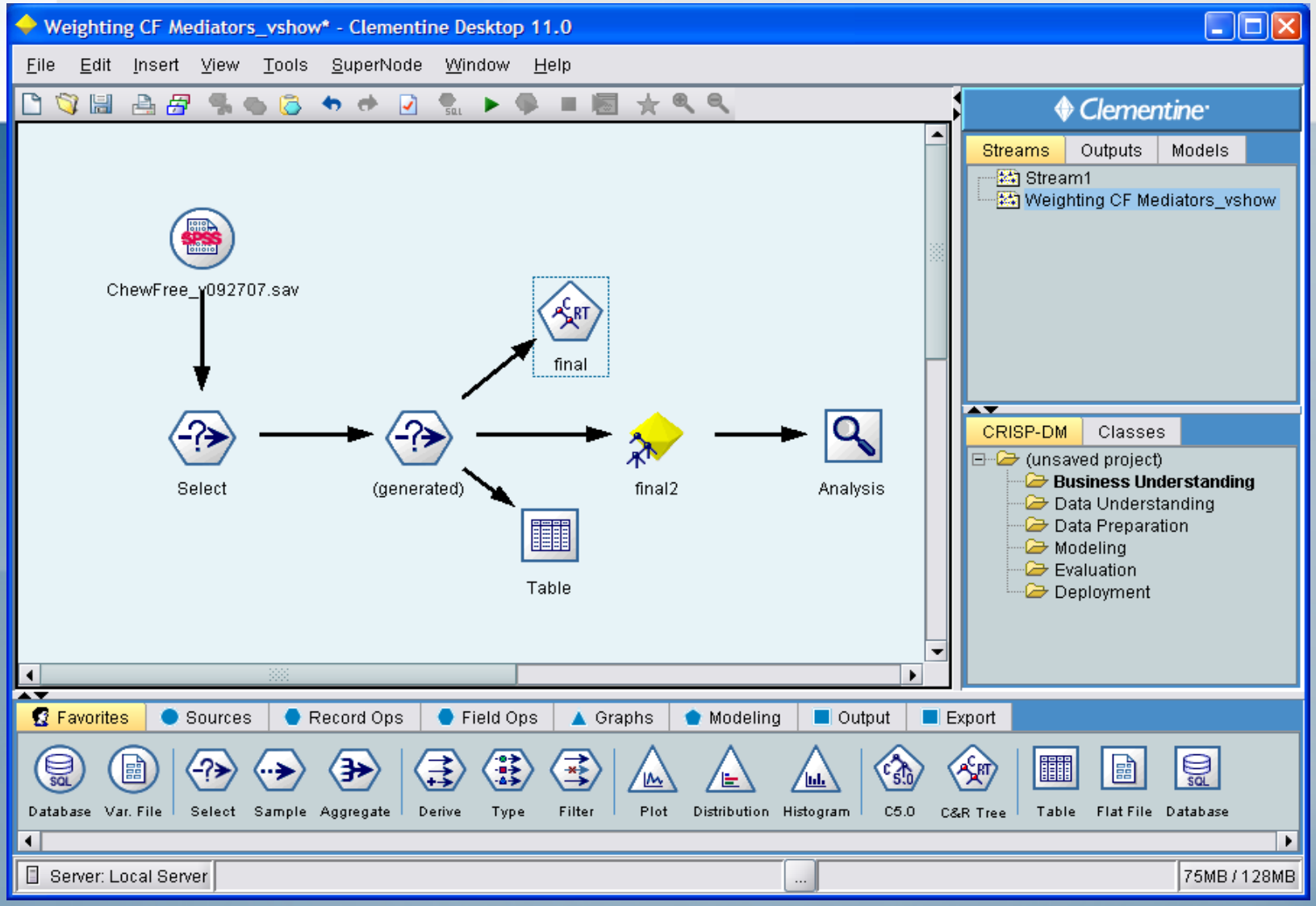
- Recruitment
- Engagement
- Attrition
- Outcome
- Underlying mechanisms
- Web forum text analytics
- Tailoring

Takeaway: RP

- Deserves an important place in analysis toolset
- Complements logistic regression, survival analysis, and other often-used tests
- Used widely but not in yet in research on Internet interventions

Takeaway RE RP

- RP tools with usable interfaces are becoming more widely available
  - SPSS classification trees module
- GUI within data mining programs
  - SPSS Clementine which focus on decision analysis
  - SAS Enterprise Miner


Weighting CF Mediators_vshow* - Clementine Desktop 11.0
File Edit Insert View Tools SuperNode Window Help
ChewFree_v092707.sav
final
Select
(generated)
final2
Analysis
Table
Clementine
Streams Outputs Models
Stream1
Weighting CF Mediators_vshow
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export
Database Var. File Select Sample Aggregate Derive Type Filter Plot Distribution Histogram C5.0 C&R Tree Table Flat File Database
Server: Local Server
75MB / 128MB

Resources

- **Lemon** et al., Classification and regression tree analysis in public health. *Annals of Behavioral Medicine*. 2003,*26*(3):172-181.
- **Zhang & Singer**, *Recursive partitioning in the health sciences*. Springer, 1999. [Yale]
- **Kraemer** et al. pubs with signal detection [Stanford]
- **Vanasse** et al., Smoking cessation within the context of family medicine: Which smokers take action? *Preventive Medicine*, 2004, *38*, 330-337.
- **Calvocoressi** et al., Applying recursive partitioning to a prospective study of factors associated with adherence to mammography screening guidelines. *American Journal of Epidemiology*, 2005, *162*(12), 1215-1224.

Brian G. Danaher, PhD
briand@ori.org

John R Seeley, PhD
H. Garth McKay, PhD
Herbert H. Severson, PhD