Varimpact: exploratory variable importance integrating causal inference and machine learning

Chris Kennedy, with Alan Hubbard

Division of Epidemiology & Biostatistics,
Berkeley Institute for Data Science (BIDS),
Integrative Cancer Research Group (ICARE), and
Social Sciences Data Lab (D-Lab),
at the University of California, Berkeley

Kaiser Permanente Division of Research, Northern California

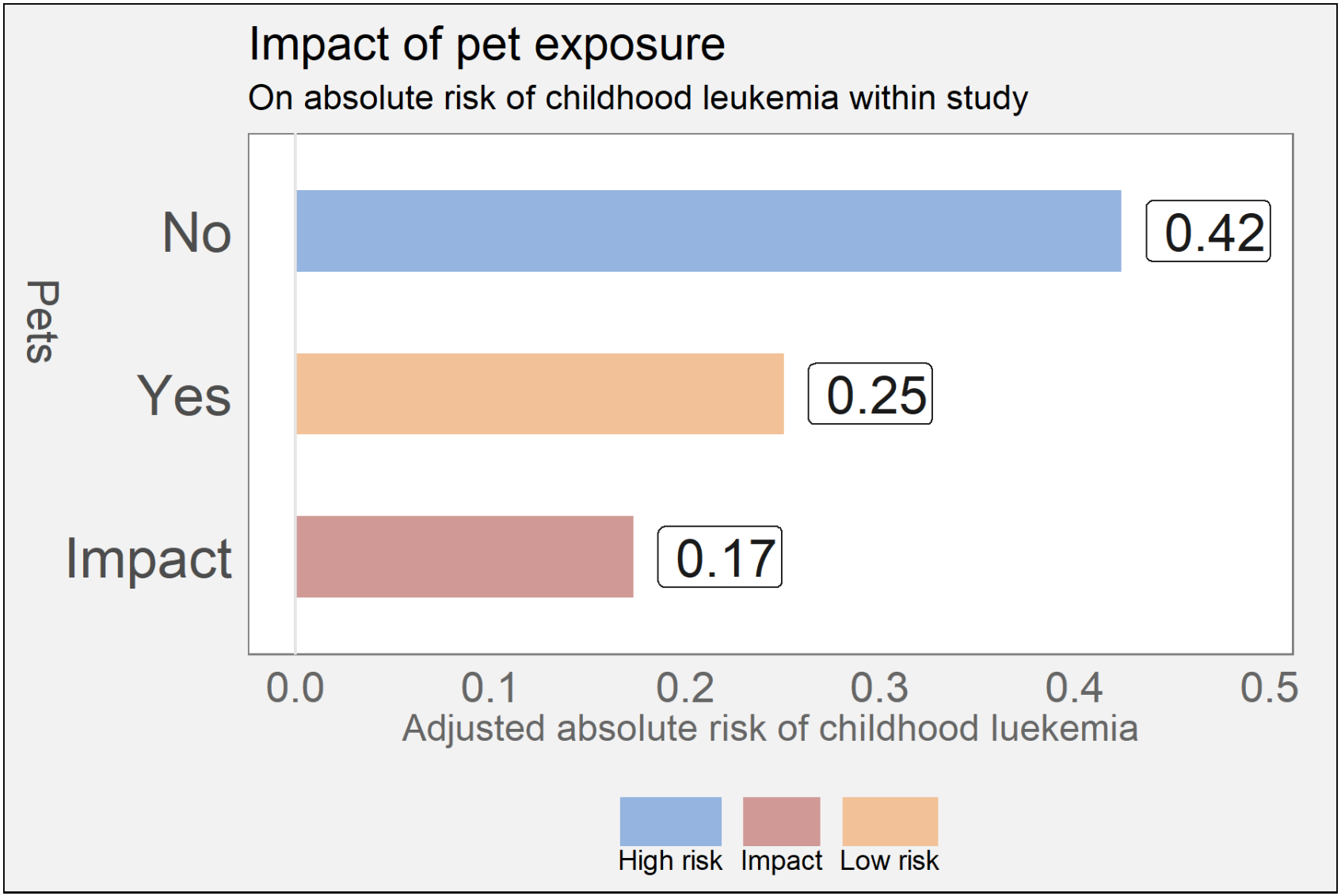
R/Medicine, Yale, Sep. 8, 2018

Scenario

- Dataset in many covariates **W** were measured at baseline
- Outcome **Y** – measured after baseline
- Question: what is your recommendation for variables that should be evaluated as treatments (**A**) in future RCTs?
  - Which variables, if hypothetically intervened upon individually, appear to most impact the outcome?

- This is one formulation of a **causal variable importance** problem.

Preview: what we'll get out of this



Problems with traditional variable importance

- Unrealistic functional form, biased, and overfit - e.g. OLS (overfit relative to penalized regression)
- Overly coarse binary results with problematic inference - Lasso variable inclusion
- Often byproduct of procedures with wrong bias-variance trade-off - focused on predicting outcome
- Not targeted to variable importance estimate at each variable
- **Masking**: correlated variables will be artificially low in importance

Variable importance as maximal contrast

- Common scientific question: which variables have the greatest influence on the outcome?
- "If I could shift a subject's covariate value from its worst level to its best level, how much would that impact the outcome?"
- We call this a "**maximal contrast**" intervention, where the levels are chosen to yield the greatest treatment effect for that variable
- Hypothetical intervention on a variable may suggest future real interventions in RCTs

Our proposed method: varimpact

1. Conduct a separate observational study on each variable, using its levels as treatments
2. Use training sample to identify 2 levels of treatment variable to use for contrast
3. Estimate “treatment effect” on test sample at certain levels, adjusting for other vars
4. Leverage CV-TMLE so bias reduction step (fluctuation) can include full sample
5. Adjust for multiple comparisons via Benjamini-Hochberg (1995)
6. Automate data cleaning steps (missing values, etc.)

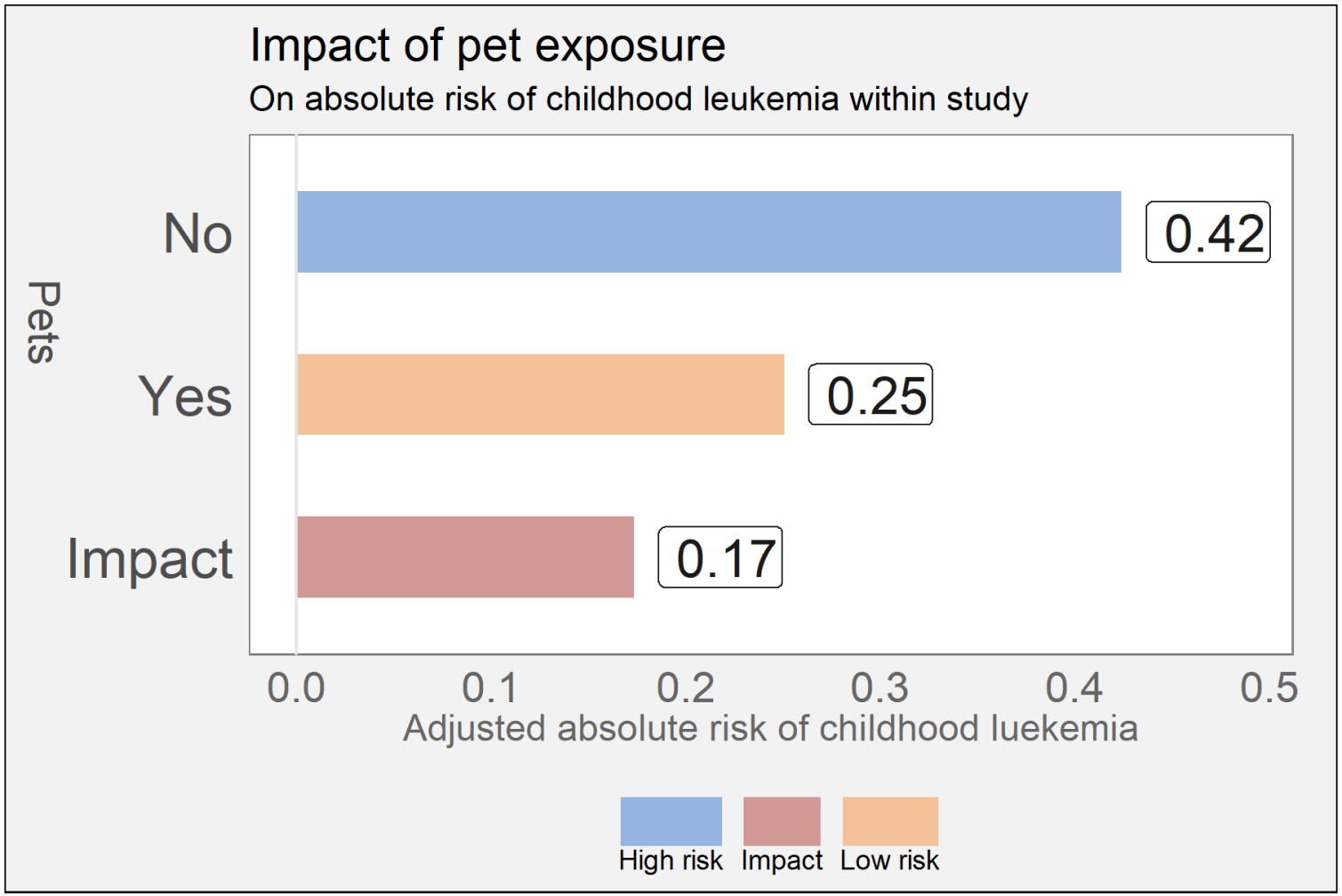
Applications:

- Childhood leukemia in Costa Rica
- Cardiovascular disease - Framingham Heart Study
- Traumatic brain injury in an urgent care setting

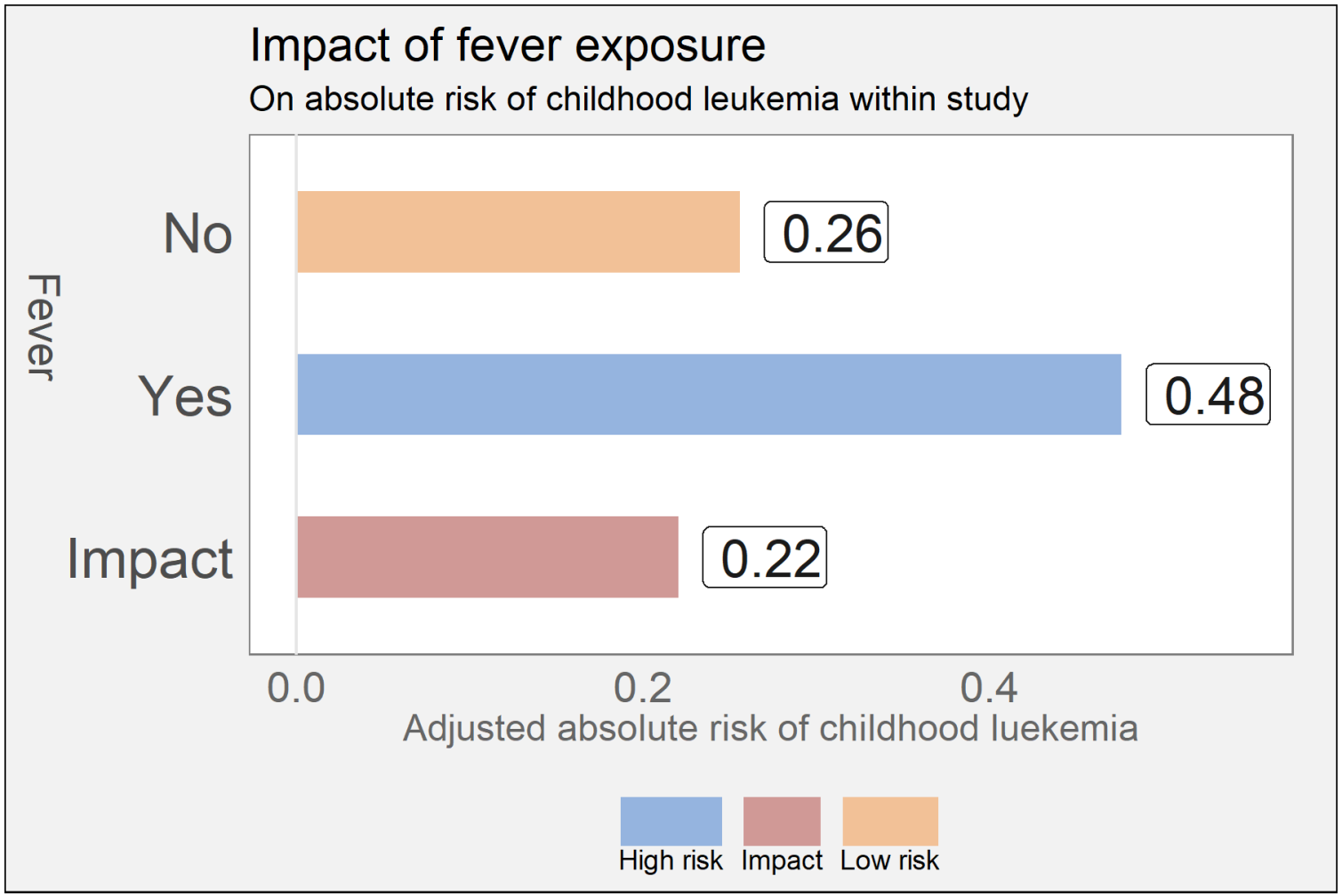
Applications: Childhood leukemia in Costa Rica

- Outcome: childhood leukemia case-control status (binary)
- Sample size: 818 observations, 39% positive
- Covariates
  - Exposures (11): breastfeeding, birth order, allergies, etc.
  - Demographic covariates (3): sex, age, income

Childhood leukemia results: **Pets**



Childhood leukemia results: **Fever**



Childhood leukemia results: **Farm Animals**

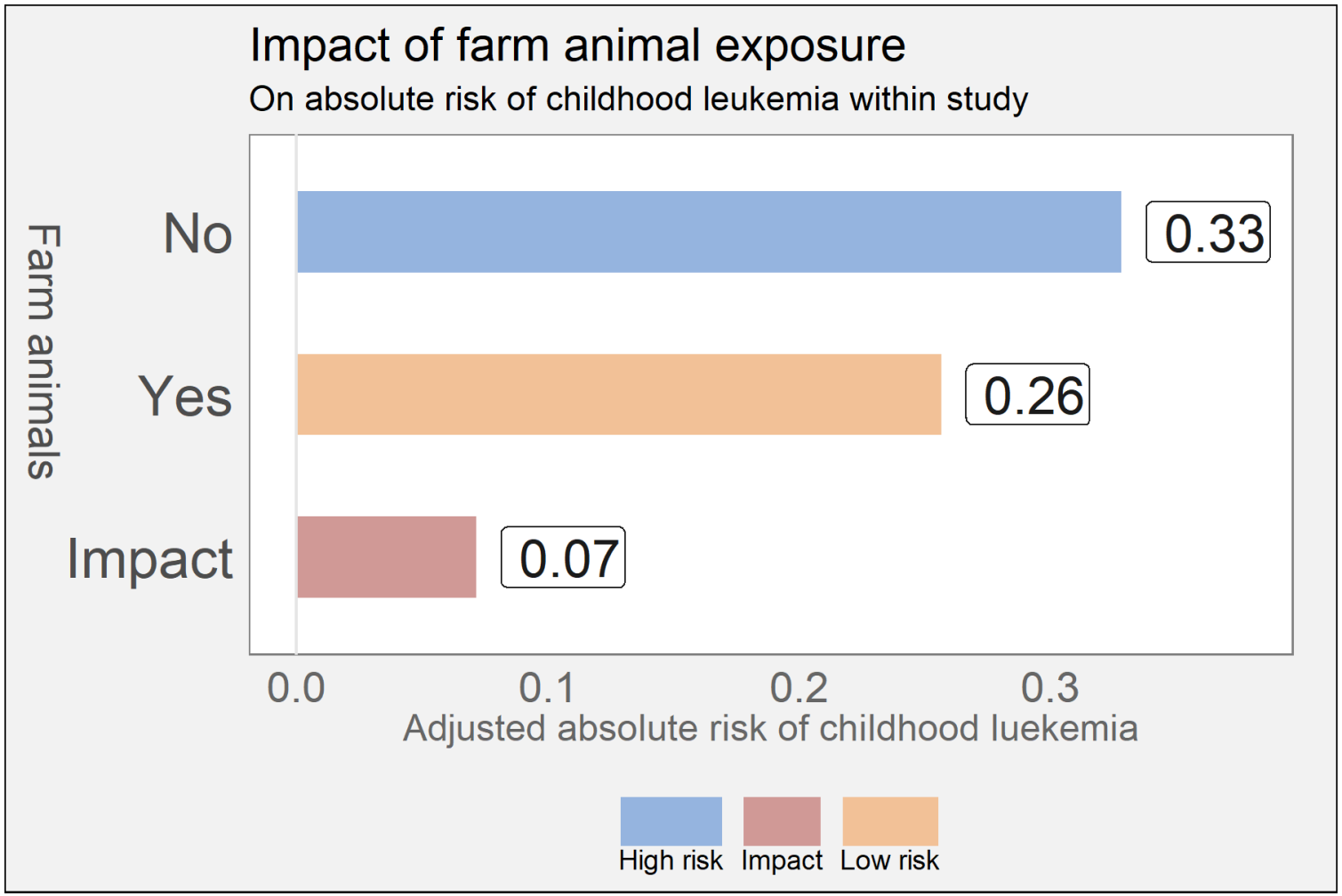


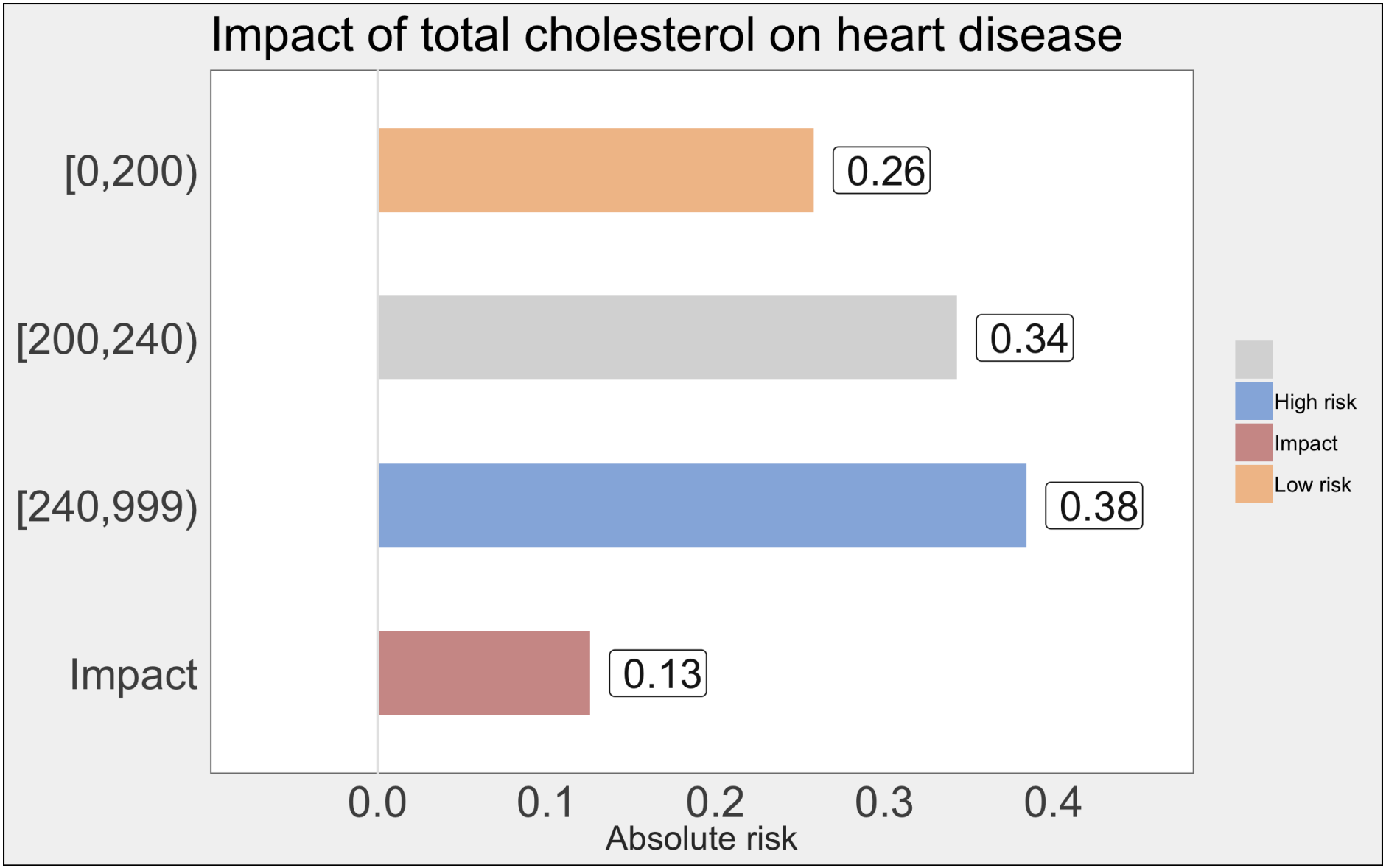
Table 1: Exposures significantly impacting childhood leukemia

Rank	Variable	Estimate	P-value	Adj. P-value	CI 95
1	fever	0.2384	0.0000	0.0000	(0.135 - 0.342)
2	pets	0.1775	0.0000	0.0000	(0.101 - 0.254)
3	farmanim	0.0826	0.0049	0.0180	(0.0199 - 0.145)

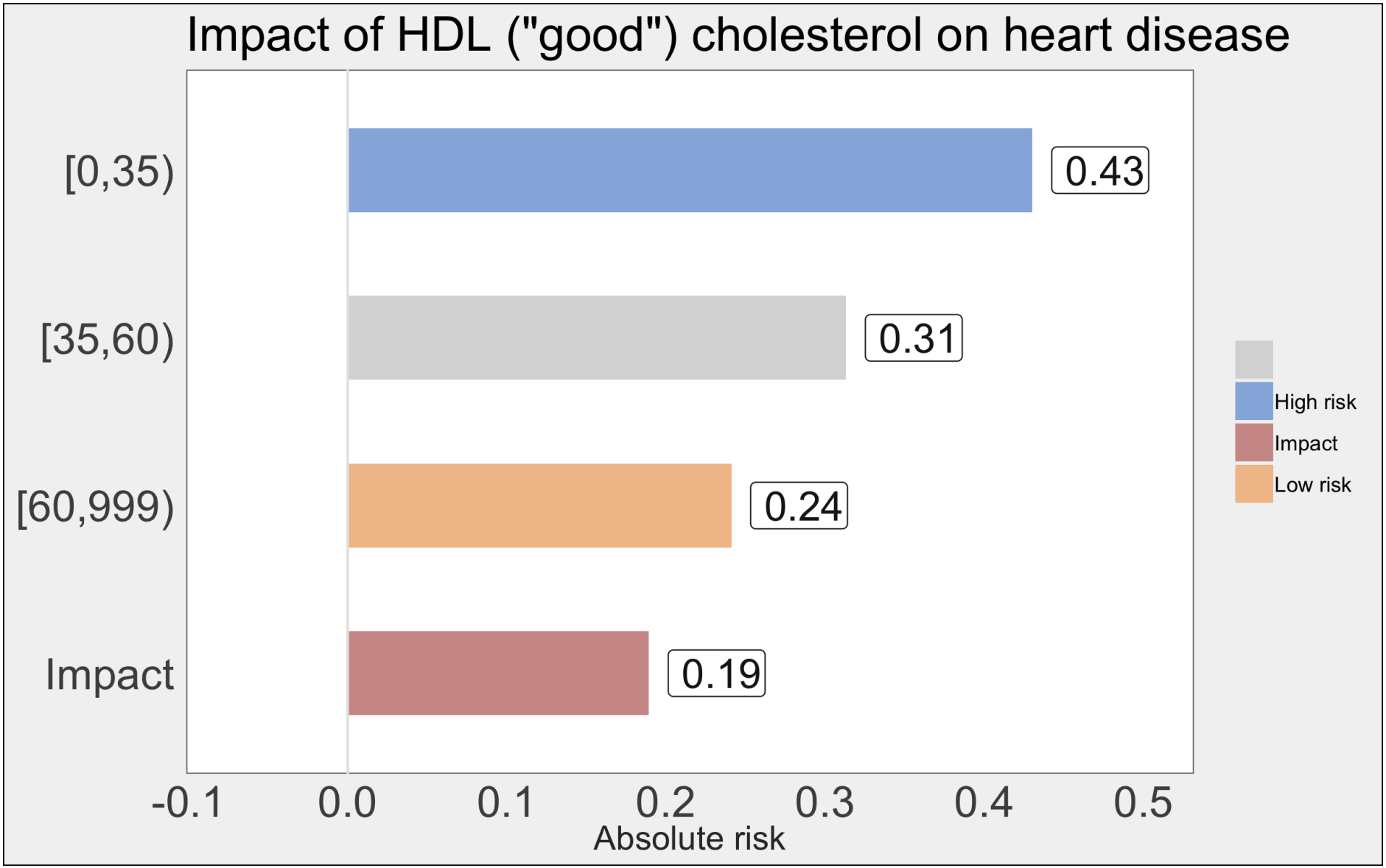
Applications: Framingham Heart Study

- Outcome: heart disease (CHD)
- Sample size: 3,263 observations, 26% positive
- Covariates (6): age, blood pressure, smoking status, diabetes status, total cholesterol, HDL cholesterol

Framingham results: **total cholesterol**



Framingham results: **HDL cholesterol**



Applications: Traumatic Brain Injury in an urgent care setting

- Outcome: traumatic brain injury (binary)
- Sample size: 784 observations, 43% positive
- Baseline covariates: 130

Traumatic brain injury results

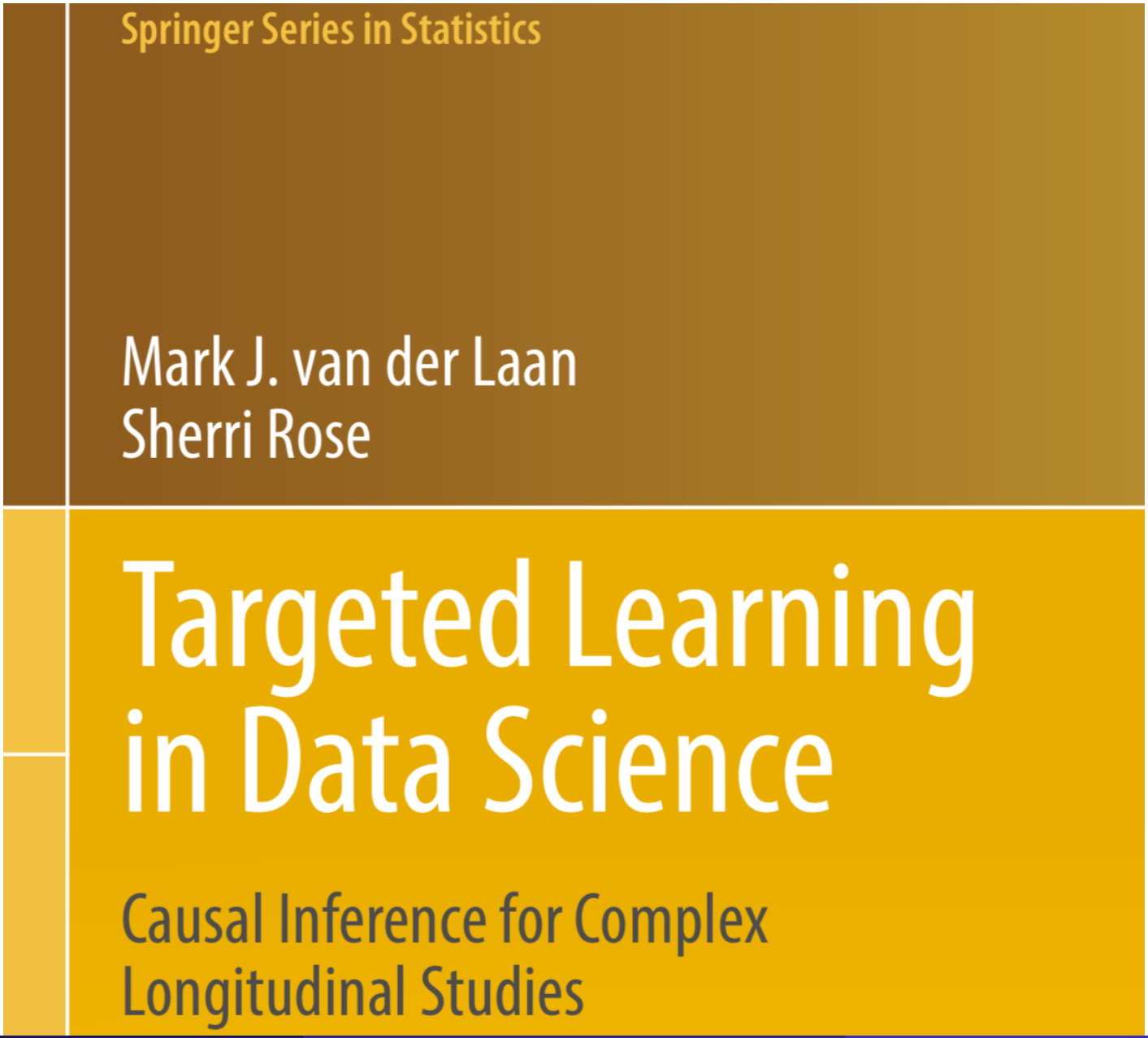
Table 2: Consistent variable importance for TBI

Variable	Type	Estimate	P-value
D-Dimer	ordered	0.4142	0.0000
Mechanism	factor	0.5170	0.0000
Blunt injury	factor	0.4629	0.0000
INR	ordered	0.1992	0.0000
Drug use	ordered	0.1422	0.0000
Alcohol use	factor	0.1505	0.0000
Factor II	ordered	0.0877	0.0000
Race	factor	0.2611	0.0001
PTT	ordered	0.1128	0.0001
Temperature	ordered	0.1466	0.0002
Height	ordered	0.1178	0.0004
Latino	factor	0.1032	0.0021

Example code

```
# Define library of estimators.
estimators = c("SL.glmnet", "SL.ranger",
               "SL.xgboost", "SL.mean")
# Use all cores on computer.
future::plan("multiprocess")
# Estimate variable impacts.
result = varimpact(outcome, data,
                   Q.library = estimators,
                   g.library = estimators)
# Plot impact of a variable.
plot_var("cholesterol", result)
```

More in latest Targeted Learning book (ch. 9)



Thanks

Questions, comments?

ck37@berkeley.edu

R package: **`github.com/ck37/varimpact`**

Twitter: @c3k