



All that Glitters Is not Gold: **Using knockoffs for type-I error controlled prognostic and predictive variable selection**

Konstantinos (Kostas) Sechidis
Associate Director of Data Science
Novartis

Basel Biometric Society (BBS) seminar
29th of August, 2024

Agenda

- Variable selection via machine learning
- Quantifying uncertainty via knockoffs
- Adapt the methods to identify predictive biomarkers
- Case study in psoriatic arthritis trials

Statistics
in Medicine

RESEARCH ARTICLE |  Full Access

Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool

Matthias Kormaksson  Luke J. Kelly, Xuan Zhu, Sibylle Haemmerle, Luminita Pricop, David Ohlssen

Statistics
in Medicine

RESEARCH ARTICLE |  Full Access

Using knockoffs for controlled predictive biomarker identification

Konstantinos Sechidis  Matthias Kormaksson, David Ohlssen

Clinical Pharmacology
& Therapeutics

original research |  Full Access

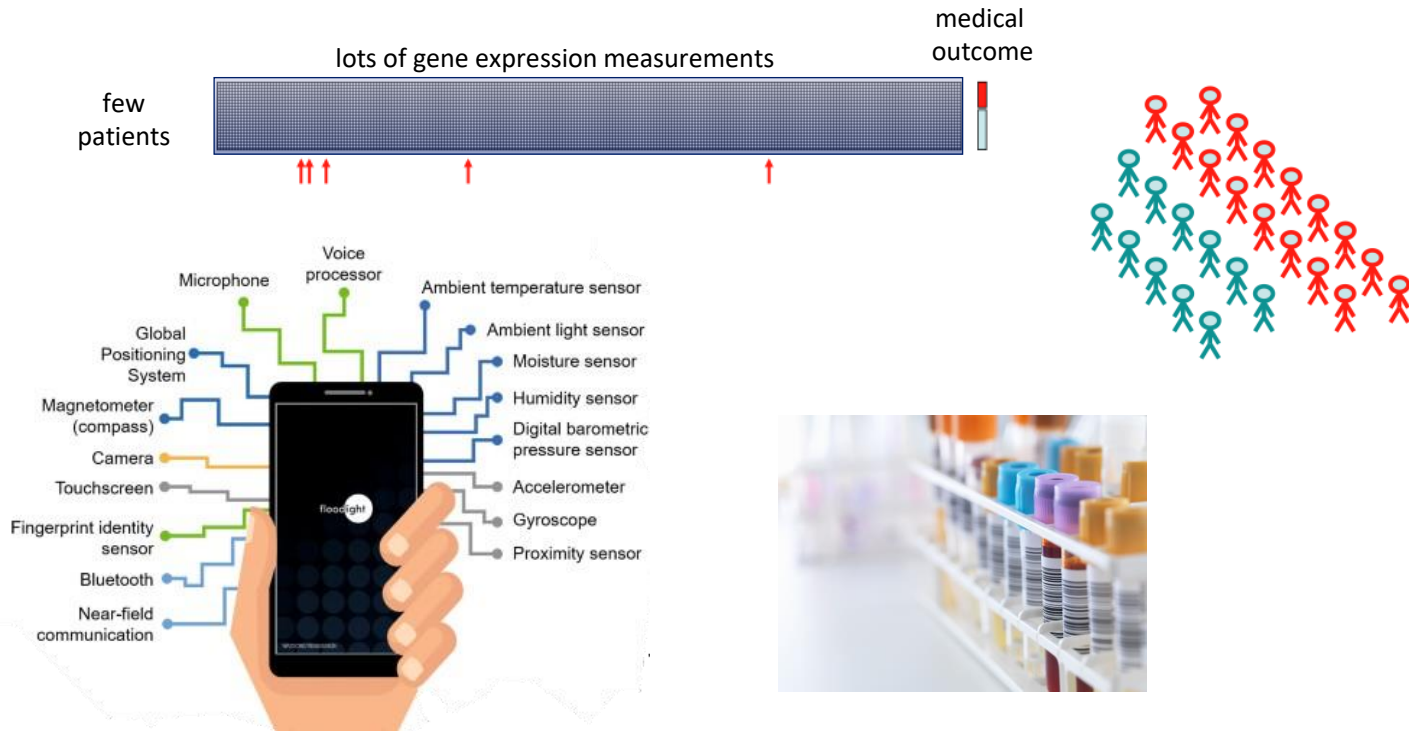
All that Glitters Is not Gold: Type-I Error Controlled Variable Selection from Clinical Trial Data

Manuela R. Zimmermann, Mark Baillie, Matthias Kormaksson, David Ohlssen, Konstantinos Sechidis 

First published: 28 February 2024 | <https://doi.org/10.1002/cpt.3211>

Variable/Feature selection

- One response Y : e.g. disease progression/status
- A large number of variables (features) X : e.g. genotype information, digital sensors ...

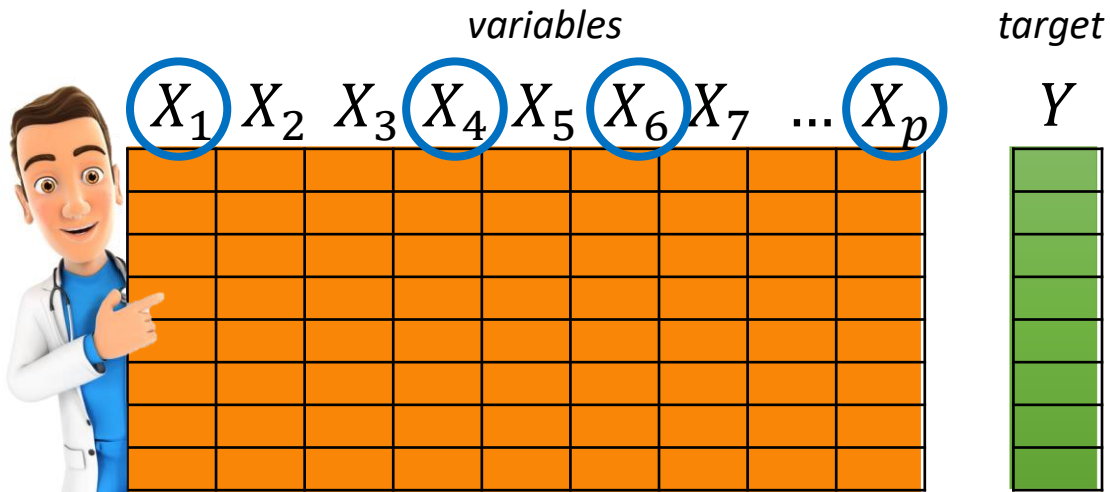


Only a subset of variables influences the outcome.

Important in healthcare, *i.e. identify prognostic biomarkers*

<https://www.nature.com/articles/d42473-019-00412-0>

Variable/Feature selection



A **variable** is of **relevant** if:

$$p(\text{target} | \text{variable}, \text{other_variables})$$

\neq

$$p(\text{target} | \text{other_variables})$$

The optimal set $\mathcal{S} \in \{X_1, \dots, X_p\}$:

$$Y \perp \bar{\mathcal{S}} | \mathcal{S}$$

➤ Actual set of relevant variables $\mathcal{S} = \{X_1, X_4, X_6, X_p\}$

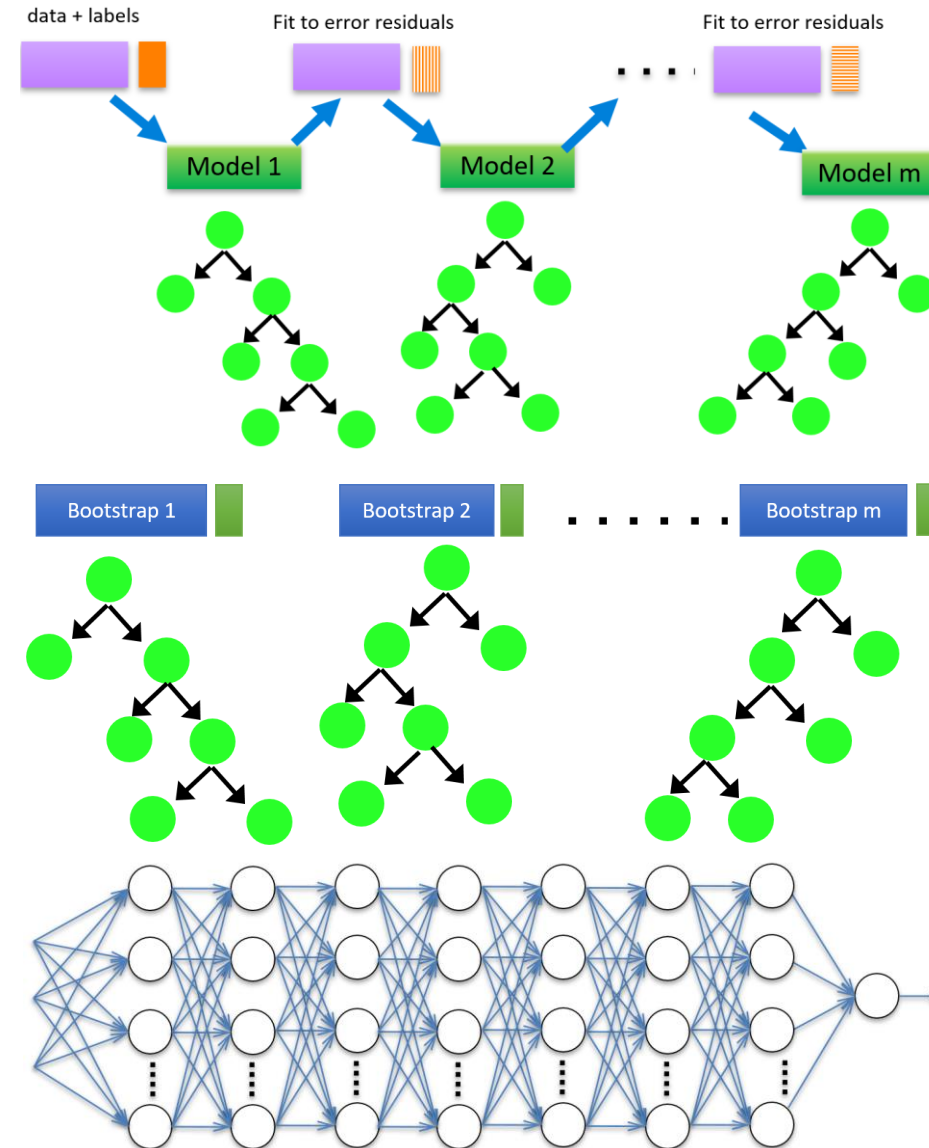
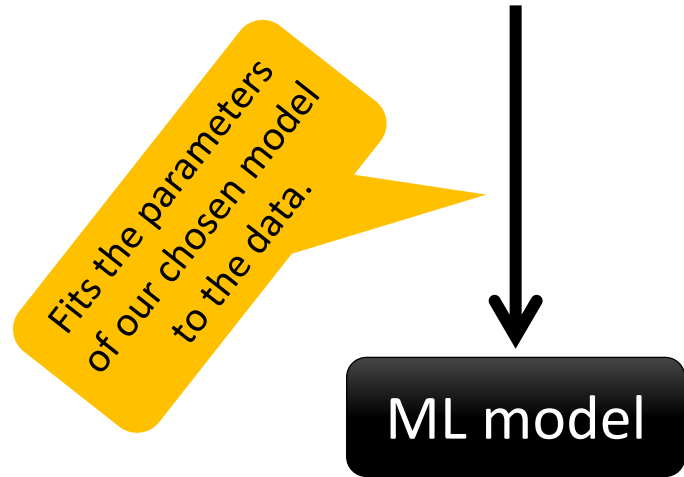
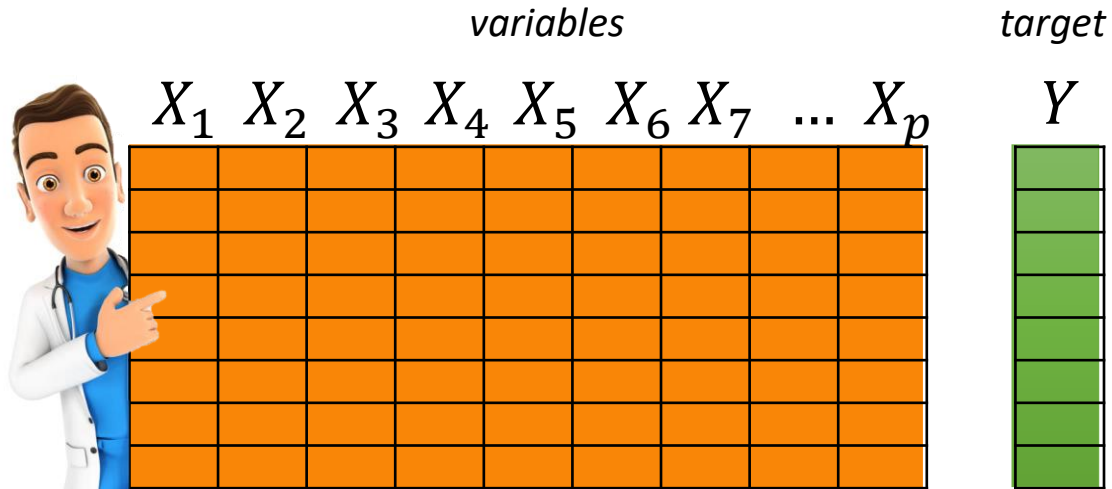
➤ Predicted set of relevant variables $\hat{\mathcal{S}} = \{X_1, X_4, X_6, X_p, X_2\}$

X_2 is a **false discovery** finding - the false discovery proportion is 1 out of 5 (20%)

Variable/Feature selection



Minimize $\sum_i (y_i - \sum_j x_{ij} \beta_j)^2$ subject to $\sum_j |\beta_j| \leq s$ LASSO



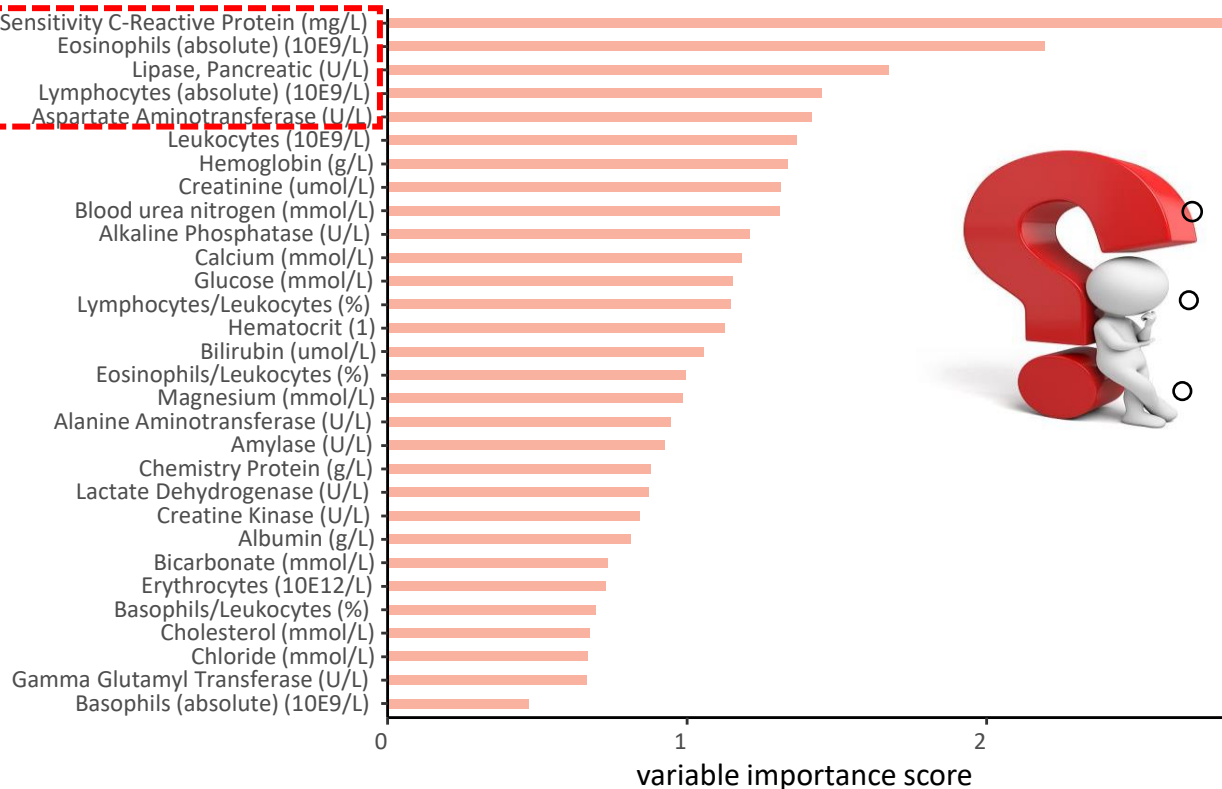
Gradient boosted trees

Random forest

(Deep) neural networks

Variable/Feature selection

ML model



Quantify uncertainty

can we control the *expected proportion of false discoveries among the discoveries?* (FDR)

can we control the *expected number of false discoveries?* (PFER)

can we control the *probability of making at least one false discovery?* (FWER)

Motivating example

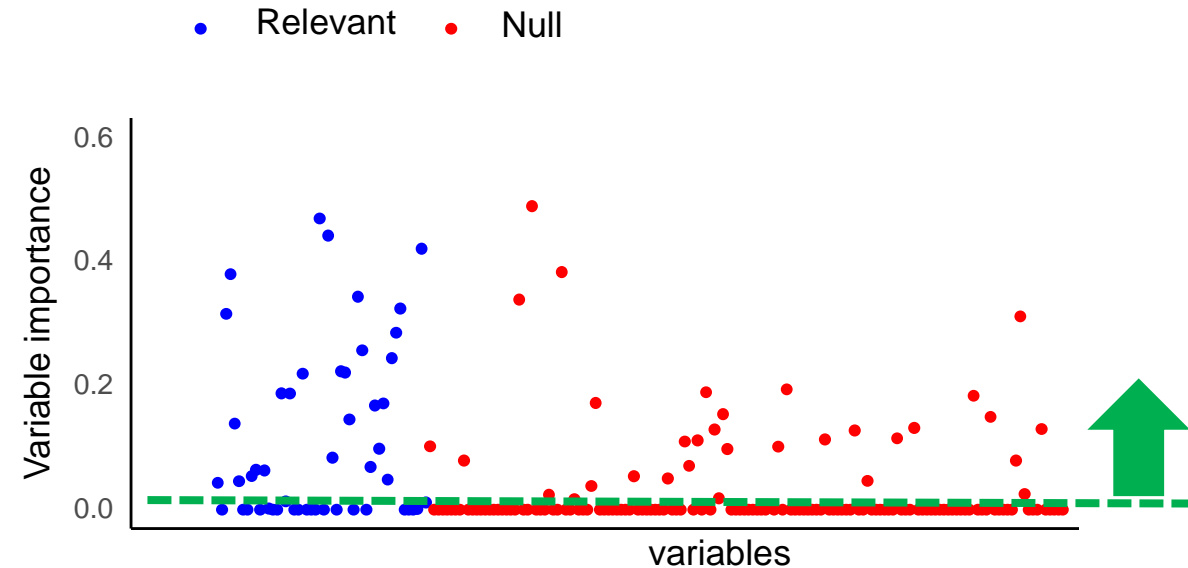
n = 500 patients

d = 200 variables (biomarkers)

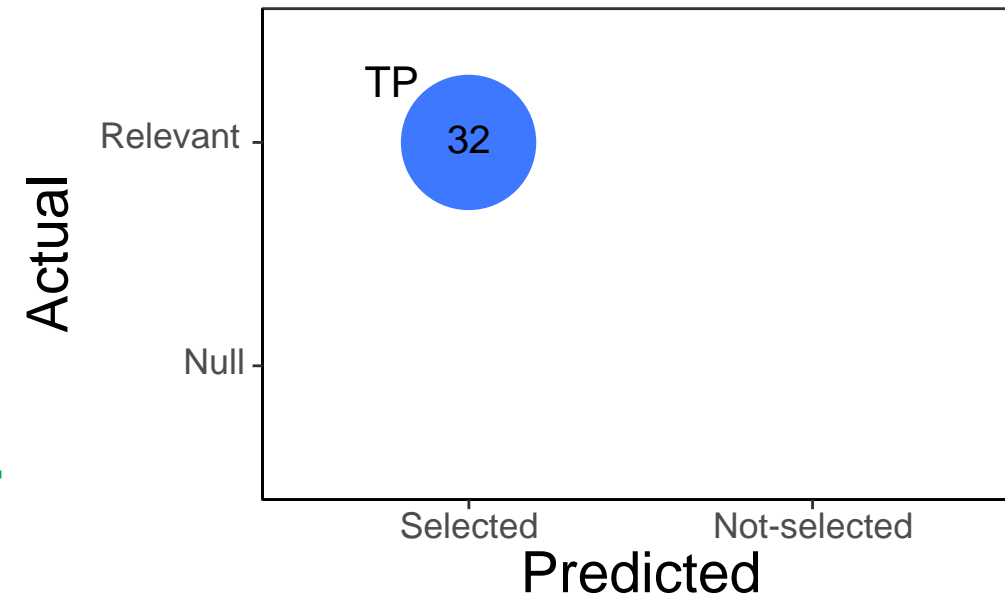
$$Y = a(X_1 + \dots + X_{50}) + \epsilon$$

Relevant

ML model **LASSO regularization**



$$\text{False Discovery Proportion} = \frac{\text{FP}}{\text{TP} + \text{FP}} = \frac{33}{32 + 33} = 0.51$$



Motivating example

n = 500 patients

d = 200 variables (biomarkers)

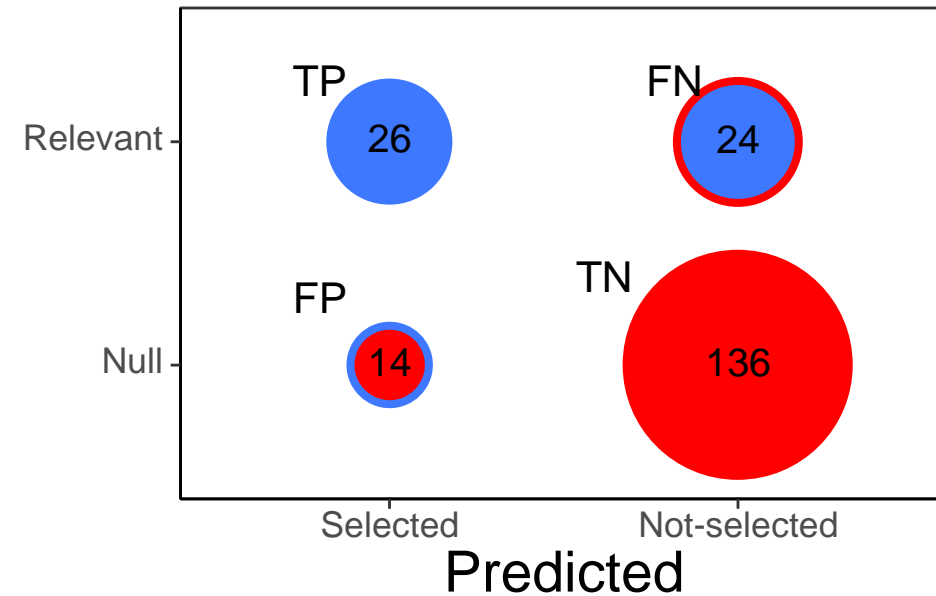
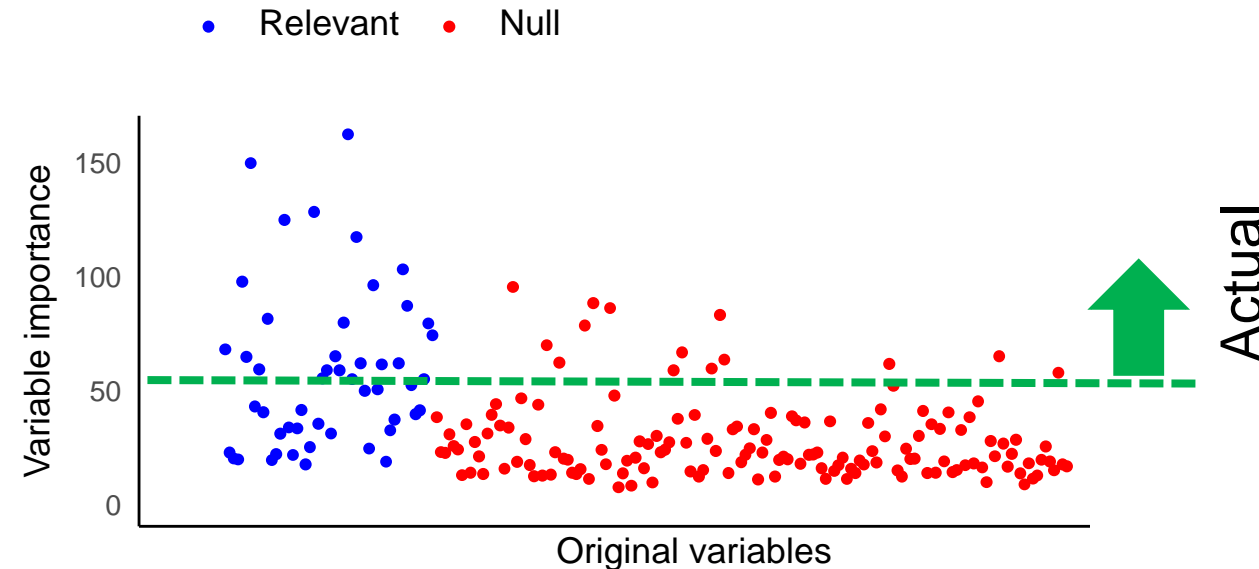
$$Y = a(X_1 + \dots + X_{50}) + \epsilon$$

Relevant

ML model

Random Forest

$$\text{False Discovery Proportion} = \frac{\text{FP}}{\text{TP} + \text{FP}} = \frac{14}{26 + 14} = 0.35$$



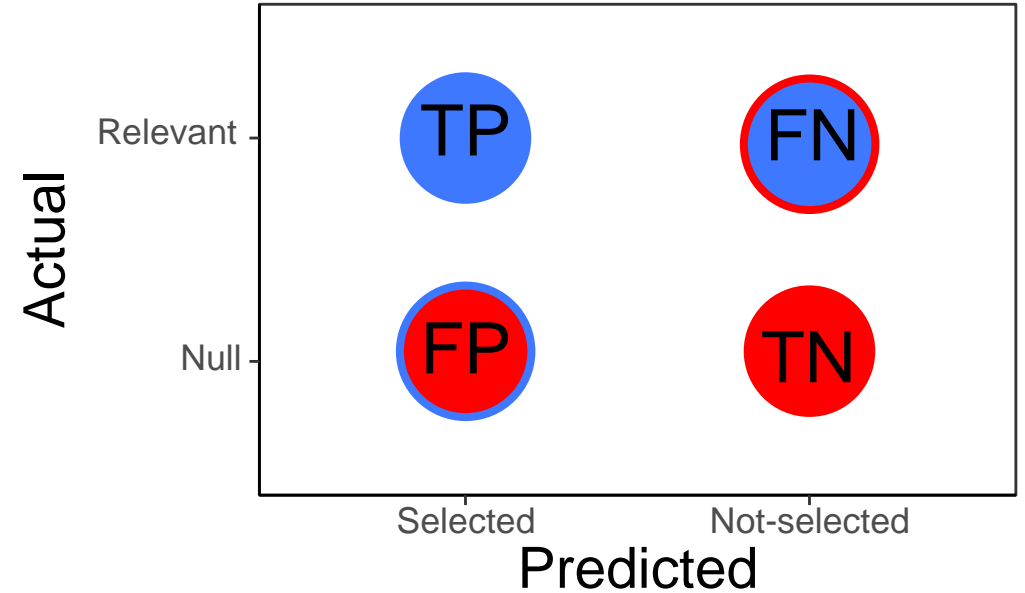
False Discovery Rate

False Discovery Proportion:

$$\text{FDP} = \frac{\text{FP}}{\text{TP} + \text{FP}}$$

False Discovery Rate:

$$\text{FDR} := \mathbb{E} [\text{FDP}]$$



*J. R. Statist. Soc. B (1995)
57, No. 1, pp. 289-300*

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel



*...for each variable, a corresponding p-value
...the tests should be independent*

Quantifying uncertainty via knockoffs



Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection

Emmanuel Candès, Yingying Fan, Lucas Janson ✉, Jinchi Lv

First published: 08 January 2018 | <https://doi.org/10.1111/rssb.12265> |

Y	X_1	X_2	...	X_p
1.128	-0.300	0.416	...	-0.328
-0.725	-0.310	-0.568	...	-0.396
-0.107	-0.876	-1.689	...	-2.554
0.791	0.308	0.804	...	-0.515
0.233	-0.038	0.425	...	-1.015
-0.350	0.931	-1.041	...	0.818
-0.849	-1.402	0.472	...	-0.208
-0.386	0.215	-0.513	...	1.822
⋮	⋮	⋮	⋮	⋮
-0.350	0.931	-1.041	...	0.818

\tilde{X}_1	\tilde{X}_2	...	\tilde{X}_p
-0.120	-0.868	...	-1.396
0.132	-0.213	...	0.822
0.351	-1.441	...	0.218
-0.756	-1.289	...	-1.554
-0.330	0.216	...	-0.228
-1.293	0.172	...	-0.108
-0.032	0.422	...	-0.015
0.381	-1.104	...	0.218
⋮	⋮	⋮	⋮
0.808	0.048	...	-1.515

- 1st step: Construct knockoffs (fake variables)
- 2nd step: Calculate a knockoff statistic
- 3rd step: Calculate a threshold to control FDR

... extensions to FWER, PFER

Journal of the American Statistical Association

Volume 116, Number 5, October 2021

Theory and Methods

Derandomizing Knockoffs

Zhimei Ren ✉ , Yuting Wei & Emmanuel Candès

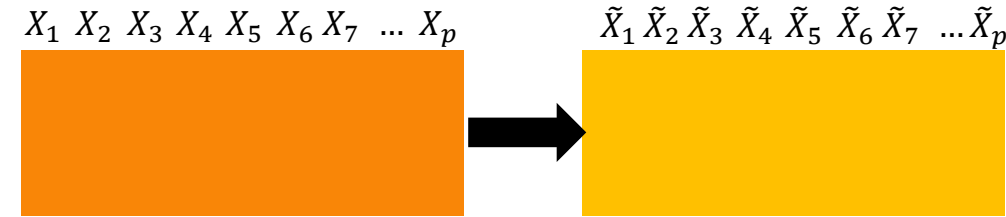
Received 28 Dec 2020, Accepted 24 Jul 2021, Accepted author version posted online: 04 Aug 2021, Published online: 14 Sep 2021

Download citation
 <https://doi.org/10.1080/01621459.2021.1962720>
 Check for updates

Knockoff filters

➤ 1st step: construct knockoff variables

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \stackrel{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$



➤ 2nd step: calculate a knockoff statistic

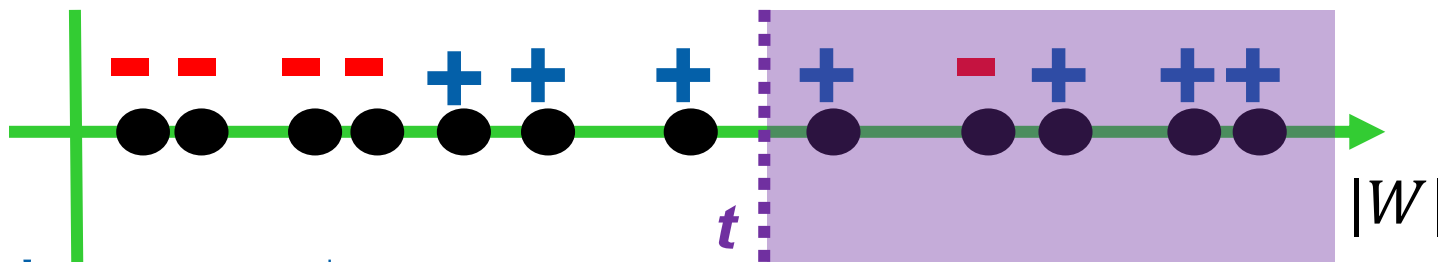
$X_1 X_2 X_3 X_4 X_5 X_6 X_7 \dots X_p \tilde{X}_1 \tilde{X}_2 \tilde{X}_3 \tilde{X}_4 \tilde{X}_5 \tilde{X}_6 \tilde{X}_7 \dots \tilde{X}_p Y$



Random forests $W_j^{\text{RF}} = |Z_{X_j}| - |Z_{\tilde{X}_j}|$

LASSO $W_j^{\text{LASSO}} = |\widehat{b}_{X_j}(\lambda)| - |\widehat{b}_{\tilde{X}_j}(\lambda)|$

➤ 3rd step: Calculate a threshold to control FDR, eg FDR = 0.30

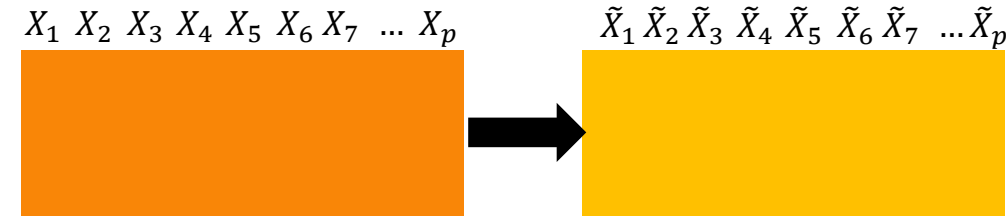


$$\widehat{\text{FDP}}(t) = \frac{1 + |\{j: W_j \leq -t\}|}{|\{j: W_j \geq t\}|} = 0.50$$

Knockoff filters

➤ 1st step: construct knockoff variables

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \stackrel{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$$



➤ 2nd step: calculate a knockoff statistic

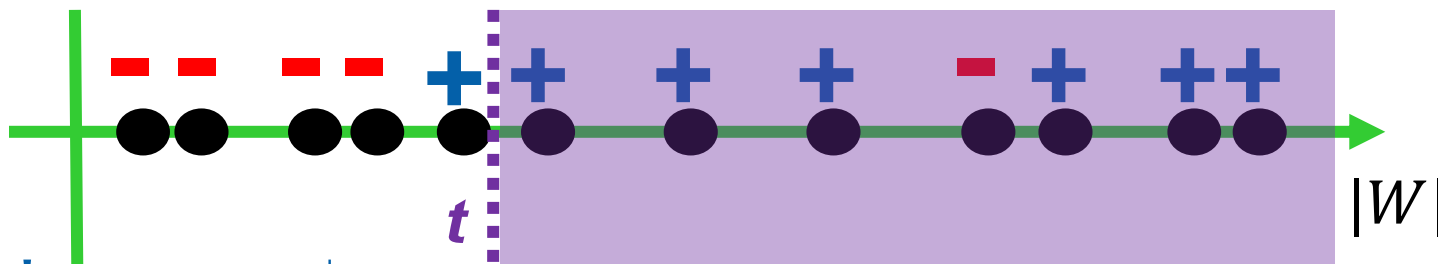
$X_1 X_2 X_3 X_4 X_5 X_6 X_7 \dots X_p \tilde{X}_1 \tilde{X}_2 \tilde{X}_3 \tilde{X}_4 \tilde{X}_5 \tilde{X}_6 \tilde{X}_7 \dots \tilde{X}_p Y$



Random forests $W_j^{\text{RF}} = |Z_{X_j}| - |Z_{\tilde{X}_j}|$

LASSO $W_j^{\text{LASSO}} = |\widehat{b}_{X_j}(\lambda)| - |\widehat{b}_{\tilde{X}_j}(\lambda)|$

➤ 3rd step: Calculate a threshold to control FDR, eg FDR = 0.30

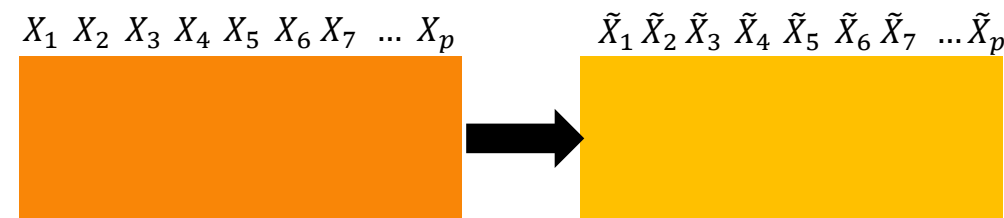


$$\widehat{\text{FDP}}(t) = \frac{1 + |\{j: W_j \leq -t\}|}{|\{j: W_j \geq t\}|} = 0.33$$

Knockoff filters

➤ 1st step: construct knockoff variables

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \stackrel{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$$



➤ 2nd step: calculate a knockoff statistic

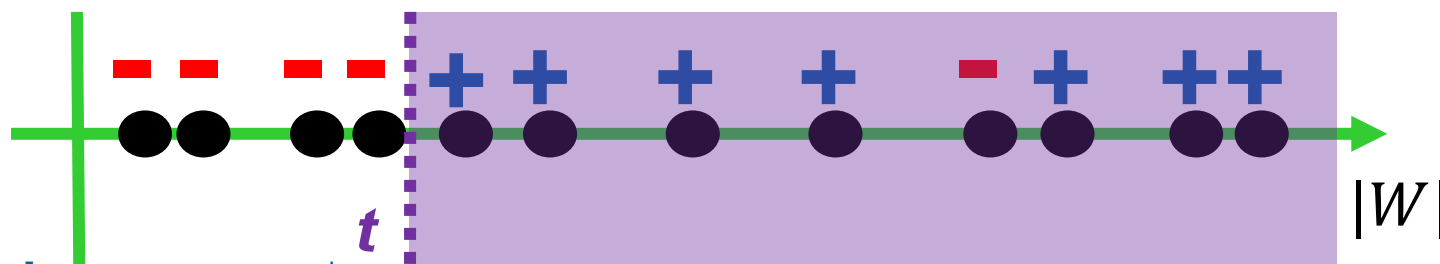
$X_1 X_2 X_3 X_4 X_5 X_6 X_7 \dots X_p \tilde{X}_1 \tilde{X}_2 \tilde{X}_3 \tilde{X}_4 \tilde{X}_5 \tilde{X}_6 \tilde{X}_7 \dots \tilde{X}_p Y$



Random forests $W_j^{\text{RF}} = |Z_{X_j}| - |Z_{\tilde{X}_j}|$

LASSO $W_j^{\text{LASSO}} = |\widehat{b}_{X_j}(\lambda)| - |\widehat{b}_{\tilde{X}_j}(\lambda)|$

➤ 3rd step: Calculate a threshold to control FDR, eg FDR = 0.30



$$\widehat{\text{FDP}}(t) = \frac{1 + |\{j: W_j \leq -t\}|}{|\{j: W_j \geq t\}|} = 0.28$$



Using knockoffs in clinical trial data

variables

Target variable

X_1 X_2 X_3 X_4 X_5 X_6 X_7 ... X_p

Y


- 1st step: Construct knockoffs (fake variables)
- 2nd step: Calculate a knockoff statistic
- 3rd step: Calculate a threshold to control FDR

prognostic biomarkers

Statistics in Medicine

RESEARCH ARTICLE | [Full Access](#)

Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool

Matthias Kormaksson  Luke J. Kelly, Xuan Zhu, Sibylle Haemmerle, Luminita Pricop, David Ohlssen

predictive biomarkers

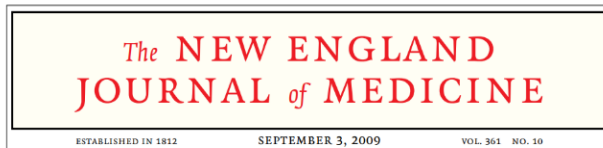
Statistics in Medicine

RESEARCH ARTICLE | [Full Access](#)

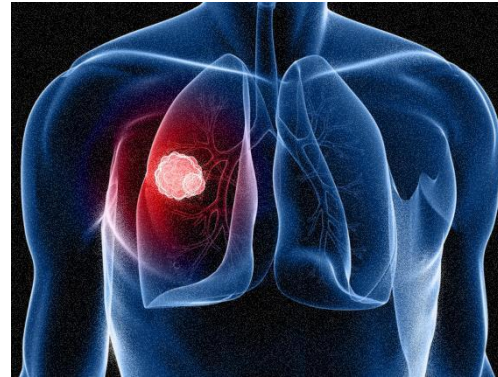
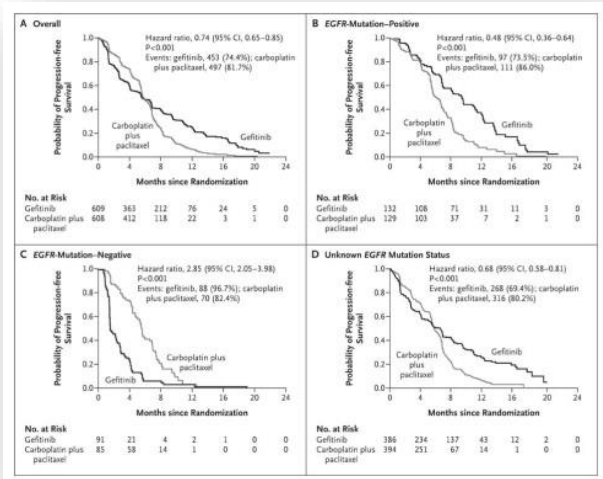
Using knockoffs for controlled predictive biomarker identification

Konstantinos Sechidis  Matthias Kormaksson, David Ohlssen

From FS to predictive biomarker discovery



Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma



A framework for discovering predictive biomarkers (eg EGFR), by controlling FDR

EGFR
positive

EGFR

EGFR
negative



Gefitinib




Carboplatin-paclitaxel


EGFR mutation is predictive ...


EGFR: Epidermal Growth Factor Receptor

From FS to predictive biomarker discovery

X_1	X_2	...	X_p	T	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
-0.300	0.416	...	-0.328	1	1.128	1.128	?	?
-0.310	-0.568	...	-0.396	0	-0.725	?	-0.725	?
-0.876	-1.689	...	-2.554	1	-0.107	-0.107	?	?
0.308	0.804	...	-0.515	0	0.791	?	0.791	?
-0.038	0.425	...	-1.015	1	0.233	0.233	?	?
0.931	-1.041	...	0.818	0	-0.350	?	-0.350	?
-1.402	0.472	...	-0.208	1	-0.849	-0.849	?	?
0.215	-0.513	...	1.822	0	-0.386	?	-0.386	?
0.425	-0.208	⋮	-0.513	1	-1.324	-1.324	?	?
0.931	-1.041	...	0.818	0	-0.350	?	-0.350	?

$T = 1$ 

$T = 1$ 

$T = 0$ 

$T = 0$ 

Knockoffs for predictive biomarker discovery

1st step: Construct knockoffs – **SAME AS BEFORE**

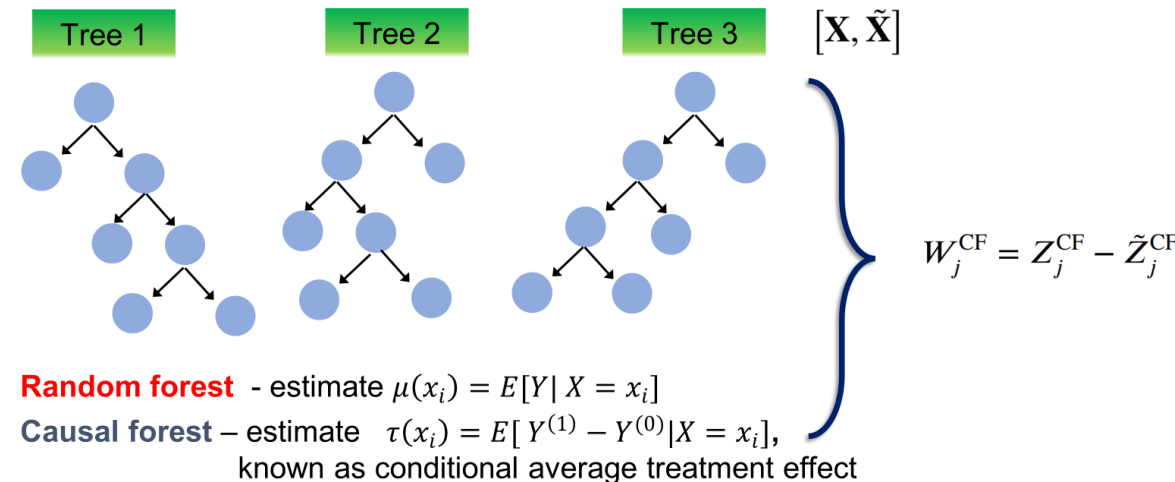
2nd step: Calculate a knockoff statistic – **NOVEL METHODS**

3rd step: Calculate a threshold to control FDR – **SAME AS BEFORE**

Filter 1: Using LASSO regression coefficients of the interaction terms

$$\begin{aligned}
 & \mathbb{E}(Y|X = \mathbf{x}, T = t) = \alpha t + \beta \mathbf{x} + \gamma t \mathbf{x} \\
 & [\mathbf{t}, \mathbf{X}, \tilde{\mathbf{X}}, \mathbf{t} : \mathbf{X}, \mathbf{t} : \tilde{\mathbf{X}}] \\
 & \hat{\mathbf{b}}(\lambda) = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \mathbf{y} - [\mathbf{t}, \mathbf{X}, \tilde{\mathbf{X}}, \mathbf{t} : \mathbf{X}, \mathbf{t} : \tilde{\mathbf{X}}] \mathbf{b} \right\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\} \\
 & \mathbf{b} = [\alpha, \beta, \tilde{\beta}, \gamma, \tilde{\gamma}]
 \end{aligned}
 \left. \vphantom{\begin{aligned} \mathbb{E}(Y|X = \mathbf{x}, T = t) = \alpha t + \beta \mathbf{x} + \gamma t \mathbf{x} \\ [\mathbf{t}, \mathbf{X}, \tilde{\mathbf{X}}, \mathbf{t} : \mathbf{X}, \mathbf{t} : \tilde{\mathbf{X}}] \\ \hat{\mathbf{b}}(\lambda) = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \mathbf{y} - [\mathbf{t}, \mathbf{X}, \tilde{\mathbf{X}}, \mathbf{t} : \mathbf{X}, \mathbf{t} : \tilde{\mathbf{X}}] \mathbf{b} \right\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\} \\ \mathbf{b} = [\alpha, \beta, \tilde{\beta}, \gamma, \tilde{\gamma}] \end{aligned}} \right\} W_j^{\text{INT-LCD}} = |\hat{\gamma}_j(\lambda)| - |\tilde{\hat{\gamma}}_j(\lambda)|$$

Filter 2: Using importance scores derived from causal forest



Novartis case study: Psoriatic arthritis (PsA)

- PsA is an inflammatory disease that affects many areas of the body.
- Cosentyx is indicated for the treatment of adult patients with active psoriatic arthritis.
- Four Phase III trials were analysed: FUTURE 2-5

Trial/ Dose	Placebo	75 mg	150 mg NL	150 mg	300 mg	Total
FUTURE2 (NCT01752634)	98	99	0	100	100	397
FUTURE3 (NCT01989468)	137	0	0	138	139	414
FUTURE4 (NCT02294227)	114	0	113	114	0	341
FUTURE5 (NCT02404350)	332	0	222	220	222	996
Total	681	99	335	572	461	2148



<https://doi.org/10.1007/s40267-021-00814-5>

- Primary endpoint is a binary composite score ACR50 in week 16.
 - **Y=1 responder** 😊
 - **Y=0 non responder** 😞
- 57 variables (baseline variables)

Predictive markers by controlling FDR = 20%

$$\Pr(Y = 1|T = 1, X = \mathbf{x}) - \Pr(Y = 1|T = 0, X = \mathbf{x})$$



Predictive markers

C-reactive protein

Age

Fatigue score

Sex

Body Surface Area

Psoriasis Nail Subset

Asymmetric Peripheral

Polyarticular Arthritis

Overall population

Subgroups defined by Age

- [19,44]
- {44,55}
- (55,84]

Subgroups defined by Fatigue Score

- [0,24] (high fatigue)
- {24,36} (medium fatigue)
- (36,52] (low fatigue)

Subgroups defined by Sex

- F
- M

Subgroups defined by Body Surface Area

- <3%
- >=3% - <10%
- >=10%

Subgroups defined by Psoriasis Nail Subset

- N
- Y

Subgroups defined by Asymmetric Peripheral Arthritis

- N
- Y

Subgroups defined by Polyarticular Arthritis

- N
- Y

CRD (90% CI)

0.26(0.23-0.29)

patients

1600

0.36 (0.31-0.4)
0.22 (0.16-0.28)
0.18 (0.13-0.23)

565
541
494

0.2 (0.15-0.25)
0.25 (0.2-0.3)
0.32 (0.26-0.38)

548
548
504

0.22 (0.17-0.25)
0.3 (0.25-0.34)

828
772

0.22 (0.18-0.26)
0.25 (0.18-0.32)
0.32 (0.25-0.37)

769
367
464

0.21 (0.16-0.26)
0.28 (0.24-0.31)

502
1098

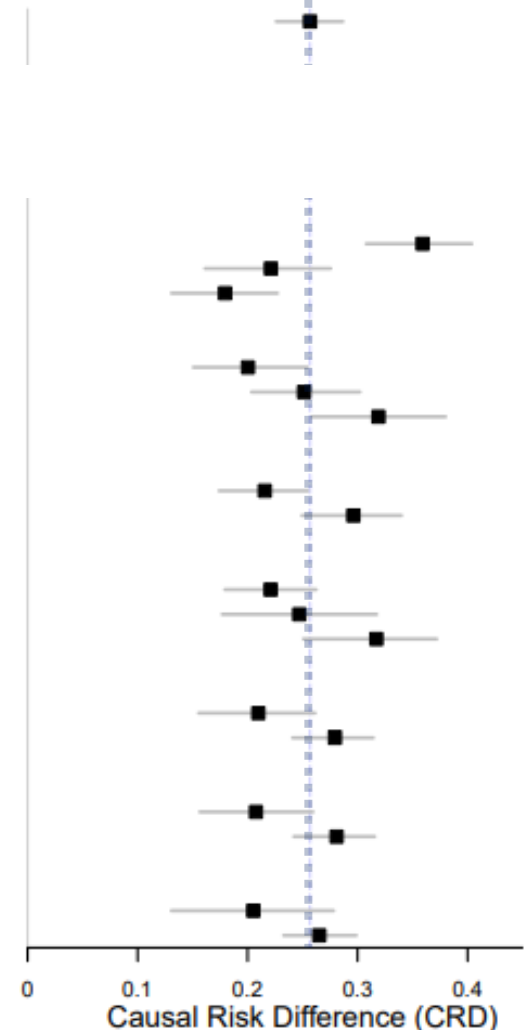
0.21 (0.16-0.26)
0.28 (0.24-0.32)

517
1083

0.21 (0.13-0.28)
0.27 (0.23-0.3)

239
1361

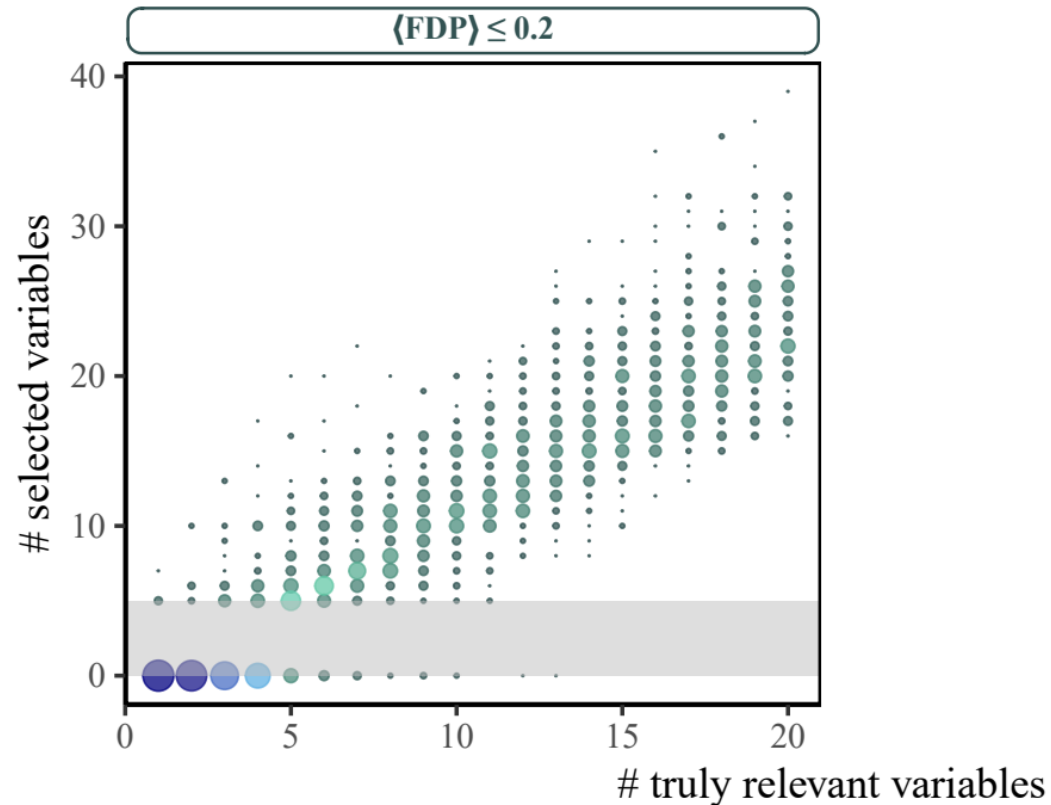
Worse than overall ← Better than overall →



Knockoff framework in practice

When we put a framework like this into practice many issues arise:

- *How to handle categorical variables?*
- *How the methods scale with sample size?*
- *How to choose the knockoff statistic?*
- *What is the computational cost?*
- *Which type-I error measure to control?*



$$\widehat{\text{FDP}}(t) = \frac{1 + |\{j: W_j \leq -t\}|}{|\{j: W_j \geq t\}|}$$

Conclusions

- ✓ Knockoffs provide a powerful framework for ML based **controlled discoveries**.
- ✓ Our work used knockoffs for controlled **predictive biomarker** identifications.
- ✓ We developed the **knockofftools**, an R package for controlled discoveries of prognostic/predictive markers in a wide variety of scenarios in terms of endpoint, error-types, filter types.



Biomarker type

- Prognostic
- Predictive

Endpoint type

- Continuous
- Binary
- Time to event

Filter type

- Regularised regression
- Random Forest
- Causal Forest

Error types

- FDR
- k-FWER
- PFER

Thank you

Kostas Sechidis

kostas.sechidis@novartis.com