



Global Drug Development (GDD)
Advanced Methodology and Data Science

Using knockoffs for controlled predictive biomarker identification

Kostas Sechidis

Associate Director Data Science, Advanced Exploratory Analytics

PSI Subgroup Analysis SIG Webinar
17th of November 2021

 **NOVARTIS** | Reimagining Medicine

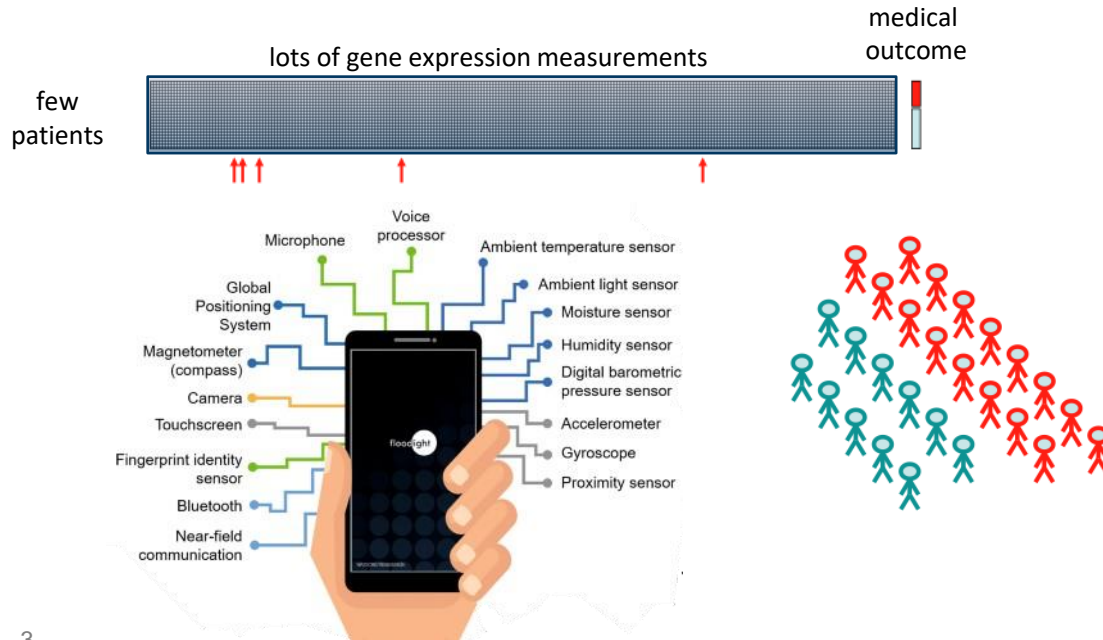
Agenda

1. Feature selection via machine learning methods
2. Quantifying uncertainty via knockoffs
3. Adapt the methods to identify predictive biomarkers
4. Case study in psoriatic arthritis trials



Feature selection

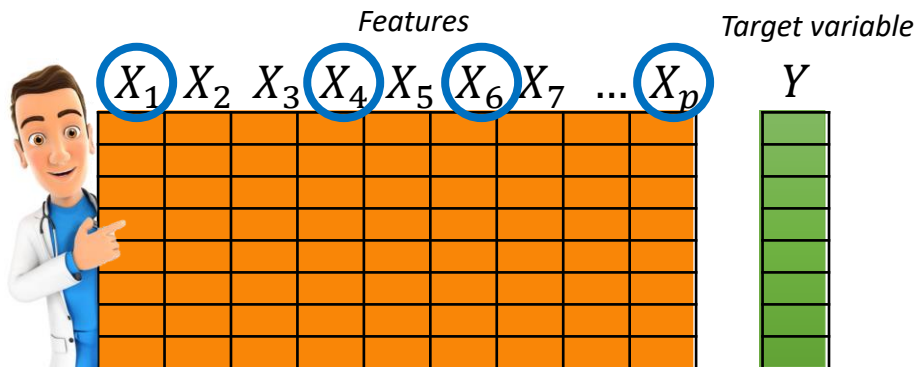
- One response Y : e.g. disease progression/status
- Thousands of variables X : e.g. genotype information, digital sensors ...



Only a subset of features actually influences the outcome.

Important in healthcare,
i.e. identify prognostic biomarkers

Feature selection



A **feature** is of **interest (relevant)** if:
 $p(\text{target}|\text{feature}, \text{other_features}) \neq p(\text{target}|\text{other_features})$

The optimal set $\mathcal{S} \in \{X_1, \dots, X_p\}$:
 $Y \perp \bar{\mathcal{S}} | \mathcal{S}$

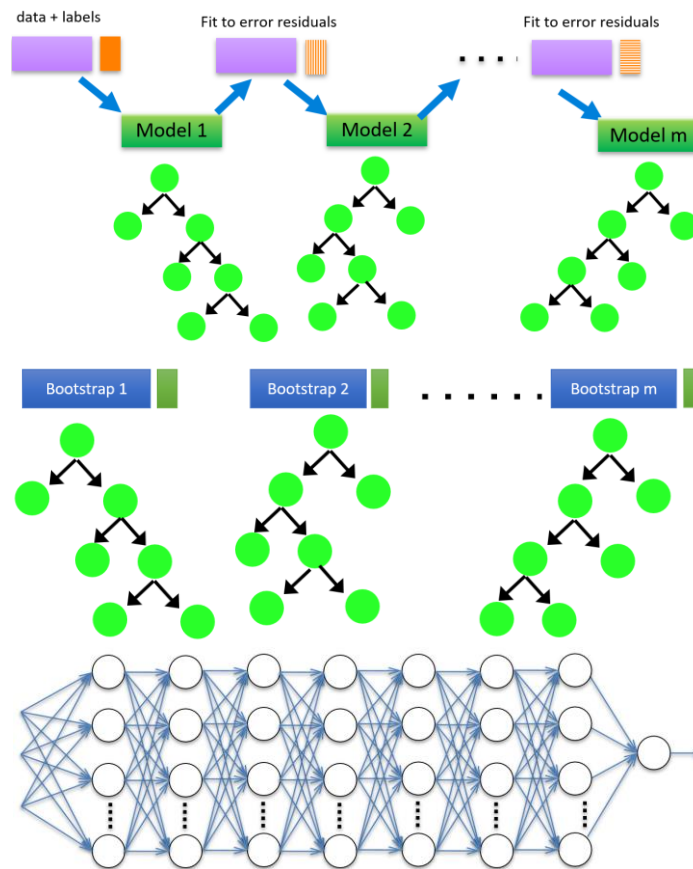
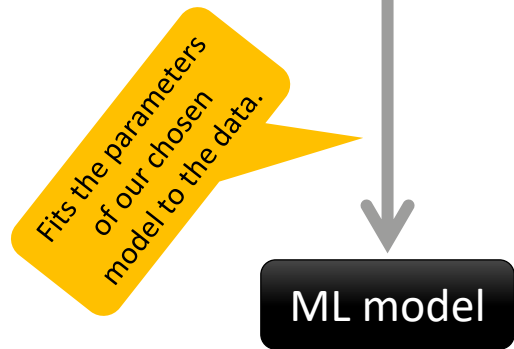
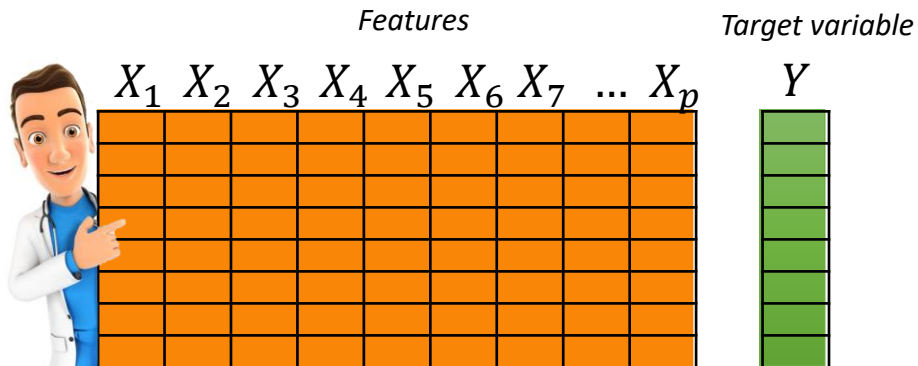
- Actual set of relevant features $\mathcal{S} = \{X_1, X_4, X_6, X_p\}$
- Predicted set of relevant features $\hat{\mathcal{S}} = \{X_1, X_4, X_6, X_p, X_2\}$

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

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

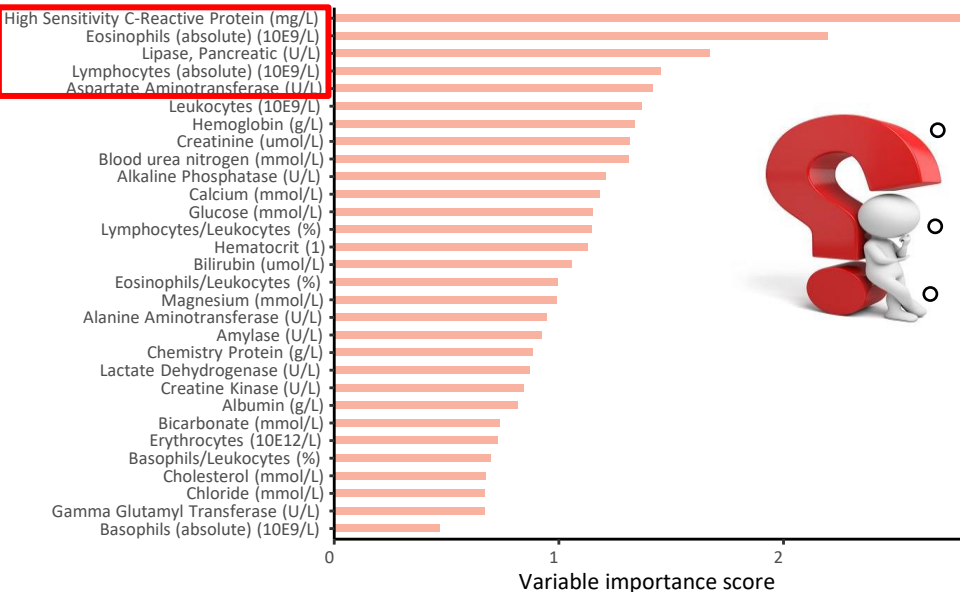
Random forest

(Deep) neural networks

Feature selection

Quantify uncertainty

ML model



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

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

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

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	\tilde{X}_1	\tilde{X}_2	...	\tilde{X}_p
1.128	-0.300	0.416	...	-0.328	-0.120	-0.868	...	-1.396
-0.725	-0.310	-0.568	...	-0.396	0.132	-0.213	...	0.822
-0.107	-0.876	-1.689	...	-2.554	0.351	-1.441	...	0.218
0.791	0.308	0.804	...	-0.515	-0.756	-1.289	...	-1.554
0.233	-0.038	0.425	...	-1.015	-0.330	0.216	...	-0.228
-0.350	0.931	-1.041	...	0.818	-1.293	0.172	...	-0.108
-0.849	-1.402	0.472	...	-0.208	-0.032	0.422	...	-0.015
-0.386	0.215	-0.513	...	1.822	0.381	-1.104	...	0.218
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-0.350	0.931	-1.041	...	0.818	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



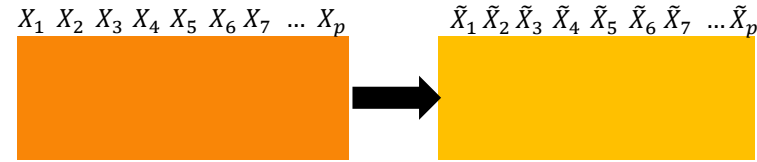
Theory and Methods
Derandomizing Knockoffs
 Zhimei Ren ✉, Yuting Wei & Emmanuel Candès
 Received 28 Dec 2020, Accepted 24 Jul 2021, Accepted author version posted online: 04 Aug 2021, Published online: 14 Sep 2021
 Download citation | <https://doi.org/10.1080/01621459.2021.1962720> | Check for updates

... extensions to FWER, PFER

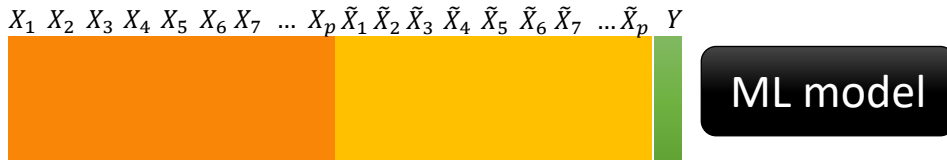
Knockoff filters

- 1st step: construct knockoff variables

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



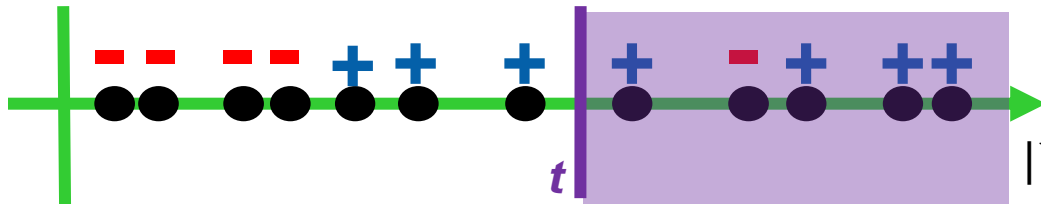
- 2nd step: calculate a knockoff statistic



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**

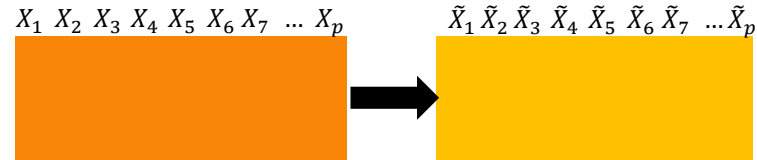


$$\widehat{\text{FDP}}(t) = \frac{1 + |\{j: W_j \leq -t\}|}{|\{j: W_j \geq t\}|} = 0.50$$

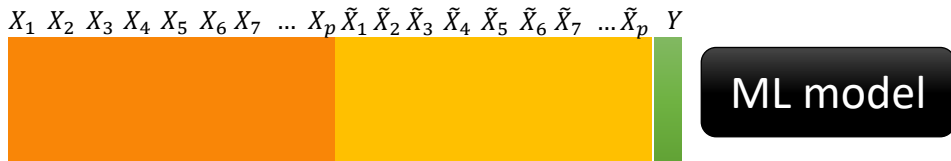
Knockoff filters

- 1st step: construct knockoff variables

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



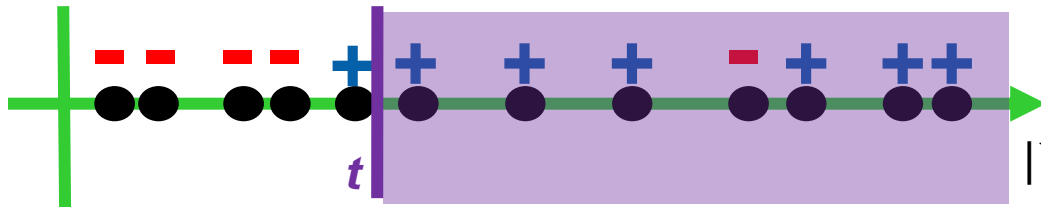
- 2nd step: calculate a knockoff statistic



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



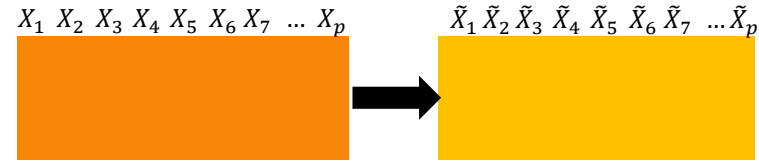
$$\widehat{\text{FDP}}(t) = \frac{1 + |\{j: W_j \leq -t\}|}{|\{j: W_j \geq t\}|} = 0.33$$

Knockoff filters

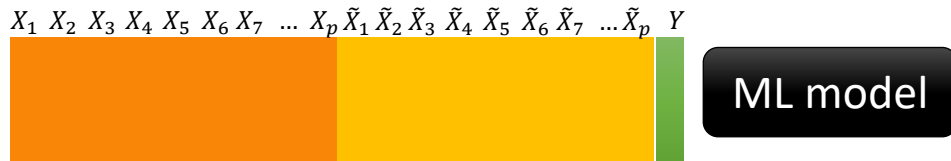
- 1st step: construct knockoff variables

$$\begin{pmatrix} X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3 \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3 \end{pmatrix}$$

$$\begin{pmatrix} X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3 \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3 \end{pmatrix}$$



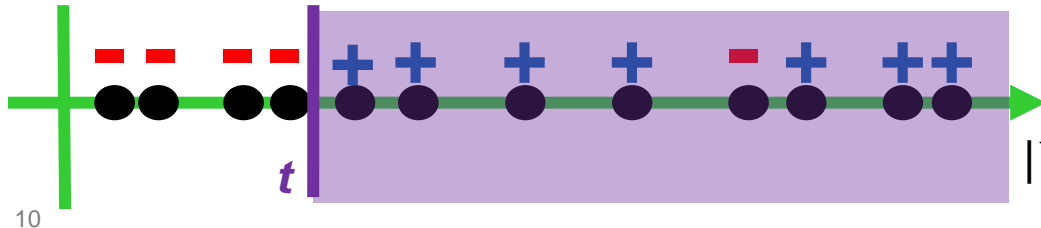
- 2nd step: calculate a knockoff statistic



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**



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



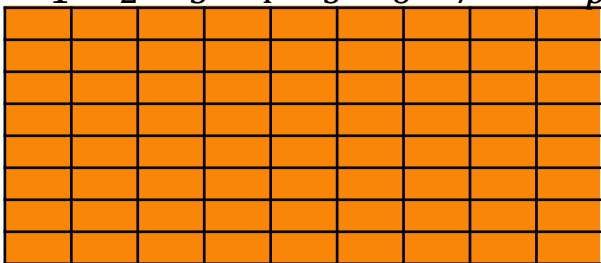
Using knockoffs in clinical trial datasets

Features

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 markers



Statistics in Medicine

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

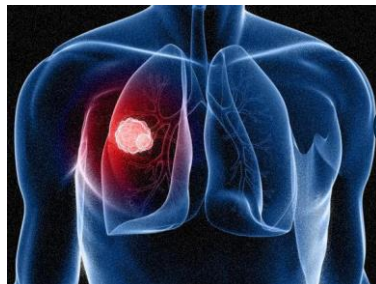
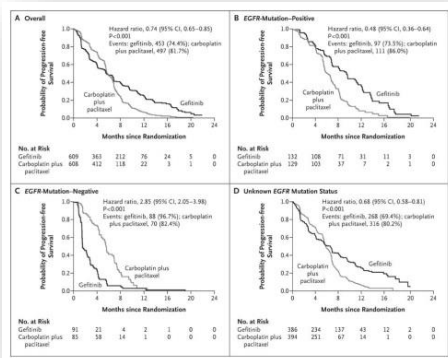
Matthias Kormaksson ✉, Luke J. Kelly, Xuan Zhu, Sibylle Haemmerle, Luminita Pricop, David Ohlssen

First published: 25 April 2021 | <https://doi.org/10.1002/sim.8955> | Citations: 1

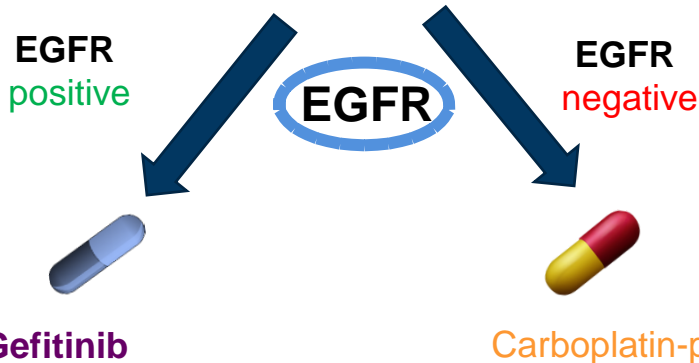
From FS to predictive biomarker discovery



Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma



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





EGFR 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	?

$T = 1$ 

$T = 0$ 

$T = 1$ 

$T = 0$ 

Knockoffs for predictive biomarker discovery

$\mathcal{S}^{\text{Pred.}}$: the actual set of predictive biomarkers

$\mathcal{H}_0^{\text{Pred.}}$: the actual of non-predictive

$$\text{FDR}_{\text{Pred.}} = \mathbb{E} \left[\frac{|\hat{\mathcal{S}}^{\text{Pred.}} \cap \mathcal{H}_0^{\text{Pred.}}|}{|\hat{\mathcal{S}}^{\text{Pred.}}|} \right]$$

$\hat{\mathcal{S}}^{\text{Pred.}}$: the set of biomarkers selected as predictive

- **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 treatment interaction terms

$$\mathbb{E}(Y|X = \mathbf{x}, T = t) = \alpha t + \beta \mathbf{x} + \gamma t \mathbf{x}$$

$$[t, \mathbf{X}, \widetilde{\mathbf{X}}, t : \mathbf{X}, t : \widetilde{\mathbf{X}}]$$

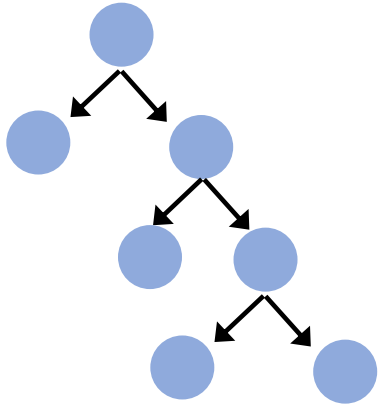
$$\hat{\mathbf{b}}(\lambda) = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \mathbf{y} - [t, \mathbf{X}, \widetilde{\mathbf{X}}, t : \mathbf{X}, t : \widetilde{\mathbf{X}}] \mathbf{b} \right\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}$$

$$\mathbf{b} = [\alpha, \beta, \widetilde{\beta}, \gamma, \widetilde{\gamma}]$$

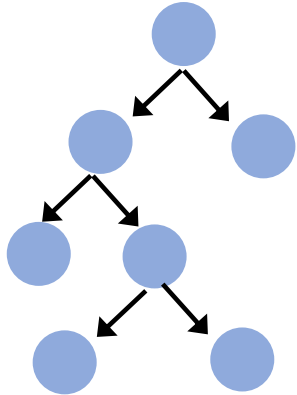
$$W_j^{\text{INT-LCD}} = |\hat{\gamma}_j(\lambda)| - |\widetilde{\hat{\gamma}}_j(\lambda)|$$

Filter 2: Using importance scores derived from causal forest

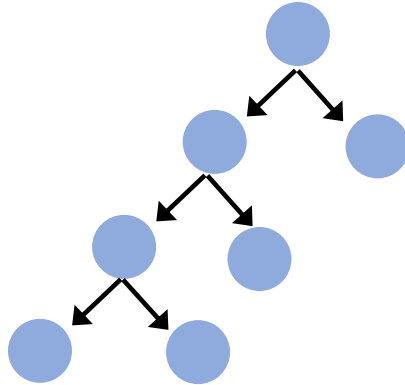
Tree 1



Tree 2



Tree 3



$[X, \tilde{X}]$

$$W_j^{\text{CF}} = Z_j^{\text{CF}} - \tilde{Z}_j^{\text{CF}}$$

Random forest - estimate $\mu(x_i) = E[Y | X = x_i]$

Causal forest – estimate $\tau(x_i) = E[Y^{(1)} - Y^{(0)} | X = x_i]$,
known as conditional average treatment effect

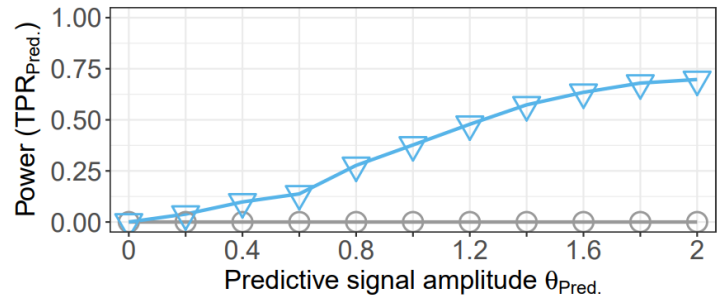
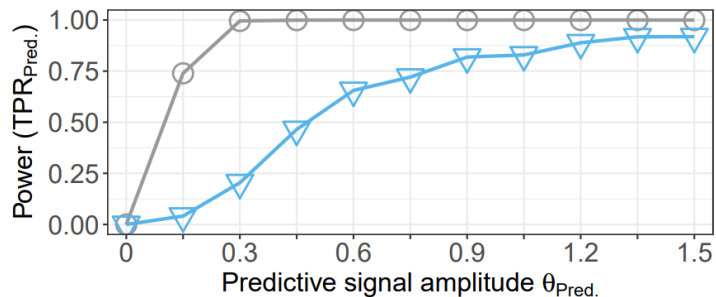
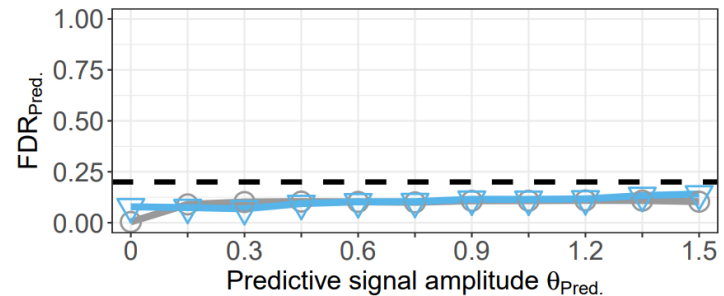
Simulation studies

(a) Knockoff filters *control FDR* to the nominal value

(b) *LASSO* filter *more powerful* when there are only *linear interactions* between features

(c) *CF* filter *more powerful* when there are *nonlinear interactions* between features

○ linear interaction KO filter ▼ CF variable importance KO filter



NVS case study: Psoriatic arthritis (PsA)

- ❑ **Psoriatic arthritis (PsA)** is an inflammatory disease that affects many areas of the body and is associated with impaired physical function and poor QoL
- ❑ **Cosentyx (secukinumab)** is indicated for the treatment of adult patients with active psoriatic arthritis and has been tested in various clinical trials.
- ❑ **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



- ❑ **Primary endpoint is a binary composite score ACR50** in week 16, which considers the number of tender and swollen joints but also includes patient/physician global assessment as well as pain and functional ability.

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

Predictive markers by controlling FDR = 20%

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



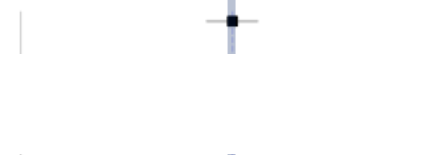
Worse than overall ← Better than overall →

Predictive markers

Overall population

CRD (90% CI)
0.26(0.23-0.29)

patients
1600



C-reactive protein

Age

Fatigue score

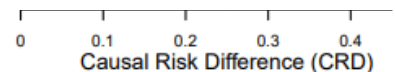
Sex

Body Surface Area

Psoriasis Nail Subset

Asymmetric Peripheral

Polyarticular Arthritis



Conclusions and future directions

- Knockoffs provide a framework for ML based controlled discoveries
- Our work used knockoffs for controlled predictive biomarker identifications
- We are currently using that methods for omics based discoveries



Thank you