MUNI | RECETOX

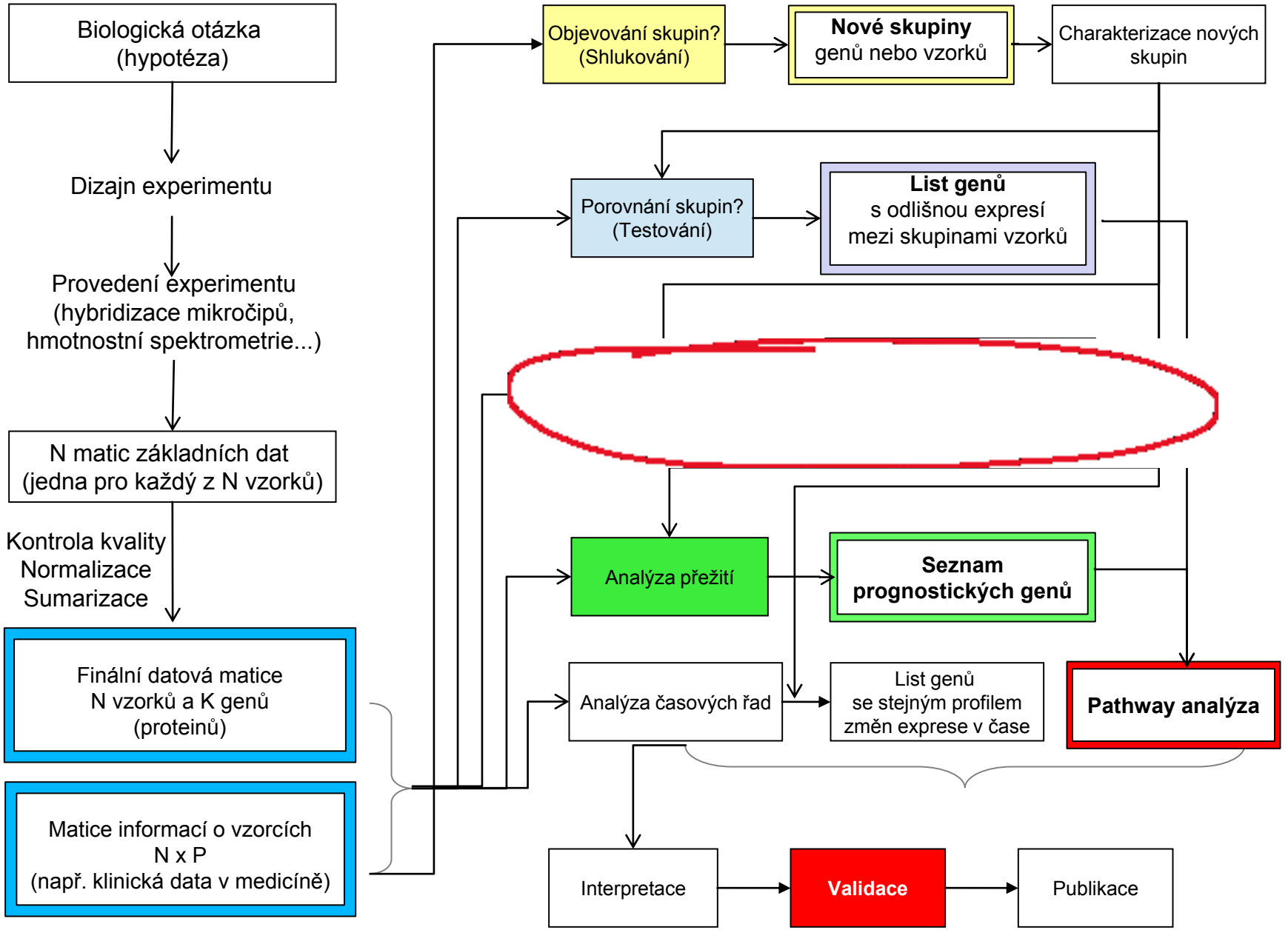
Detekce biomarkerů z omics experimentů

- Mgr. Eva Budinská, PhD
- RECETOX
- budinska@recetox.muni.cz

- Podzim 2023

Predikce skupin (klasifikace)

Jak se hledá potenciální biomarker v omics datech



Co je to biomarker?

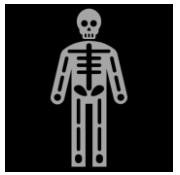
Biologický marker (biomarker):

Charakteristika, která je objektivně **měřena a hodnocena** jako **indikátor** normálních biologických procesů, patogenních procesů nebo farmakologických odpovědí na terapeutický zásah.

Biomarkers Definitions Working Group (March 2001). "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework". Clin. Pharmacol. Ther. (Review). 69 (3): 89–95.

Biomarkerem může být

Molekula a její stav (mutace DNA, hodnota exprese miRNA, zvýšená hladina proteinu…)



Aktivita buněk v konkrétních oblastech (lymfocyty v invazivním frontu nádoru)

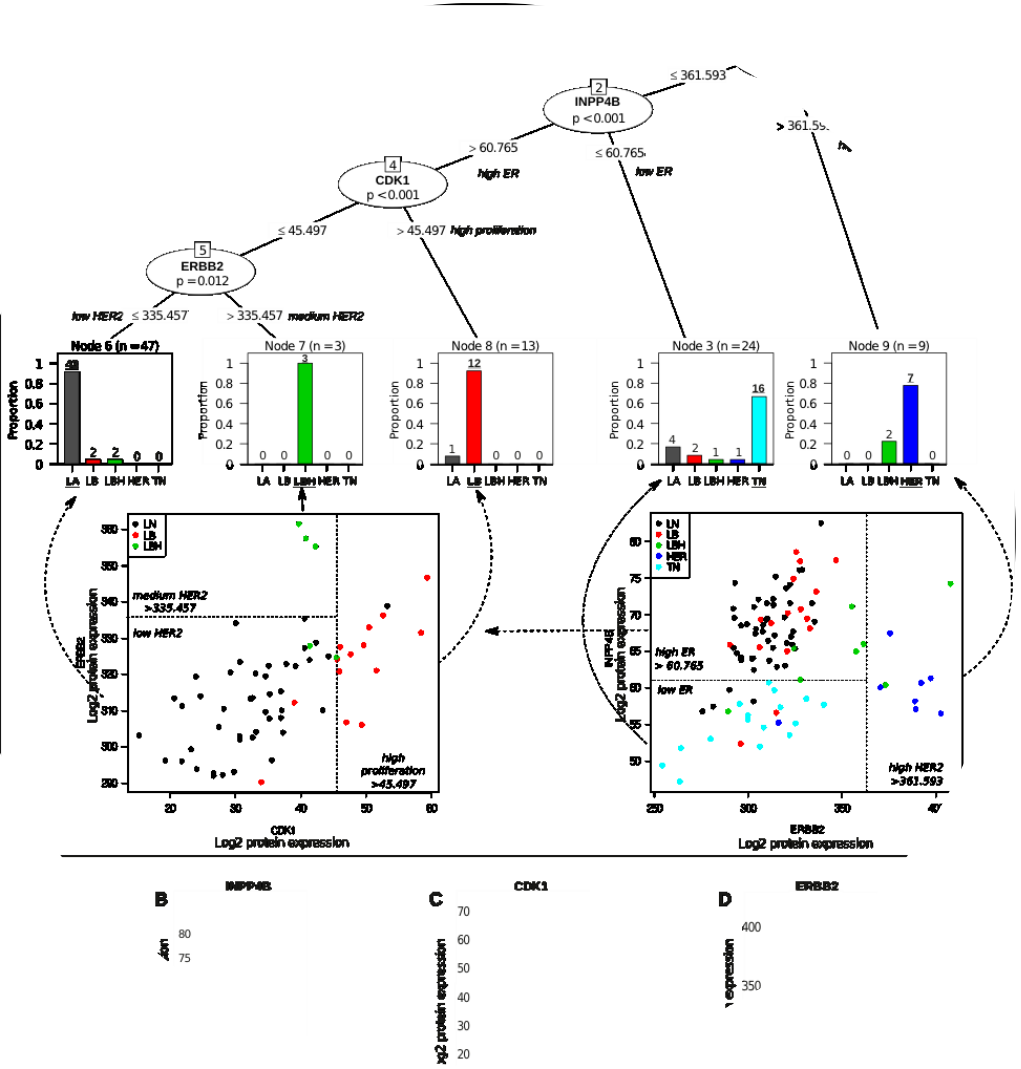
Přítomnost mikroorganizmu

Proces (zvýšená proliferace, přítomnost stromální reakce v nádoru, …)

….

Využití jednotlivých biomarkerů v rozhodovacím PRAVIDLE (modelu/testu)

MUNI | RECETOX



Biomarkery a modely

- Biomarker může být založen na **jediném analytu,** nebo na **jejich kombinaci v modelu** (klasifikátoru)

- Je to právě **kombinace více analytů** (genů, proteinů, metabolitů…), která je typická pro biomarkery z omicsových dat

Jaké je riziko onemocnění u zdravého jedince?

ká je to nemoc?

Zhoršuje se nemoc?

Jak bude nemoc probíhat?

Jaká je šance, že se nemoc vrátí?

Bude terapie úspěšná?

MUNI | RECETOX

Co musí biomarker (nebo model) splňovat

Musí být použitelný rutinně v praxi:

- **přesný** (dostatečně citlivý a dostatečně specifický)
- **robustní** (co nejméně omezen technologií měření)
- **reproducibilní** (obecně platný na cílové populaci)

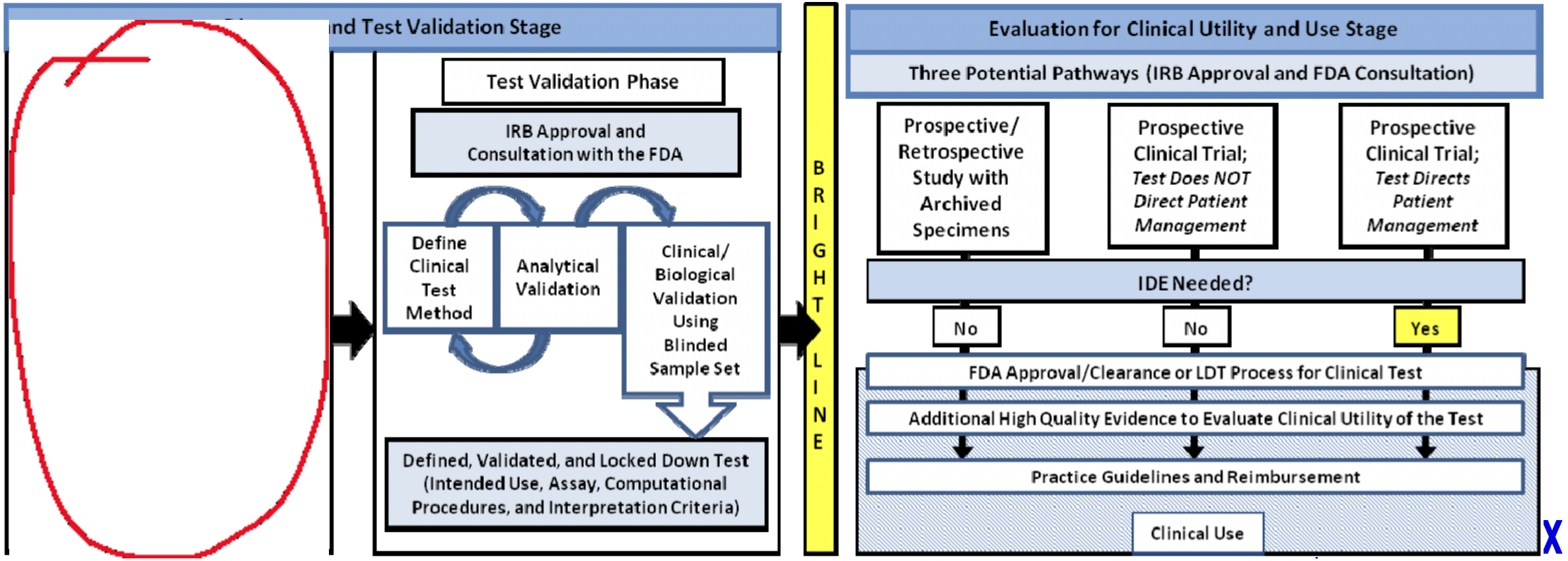
Absence jasného biologického odůvodnění testů omics biomarkerů – **proč je to problém**

Když se nedá test založený na omicsových biomarkerech biologicky odůvodnit, je o to důležitější ho správně VYTVOŘIT a poté správně VALIDOVAT, aby byla zajištěna vědecká spolehlivost!

Z důvodů vyššího rizika „přetrénování“ těchto testů je potřeba přísných kritérií, validace a odpovědnosti ještě vyšší než u samostatných testů založených na biomarkerech.

MUNI | RECETOX

Doporučení IOM komise pro vývoj testů založených na omicsových datech



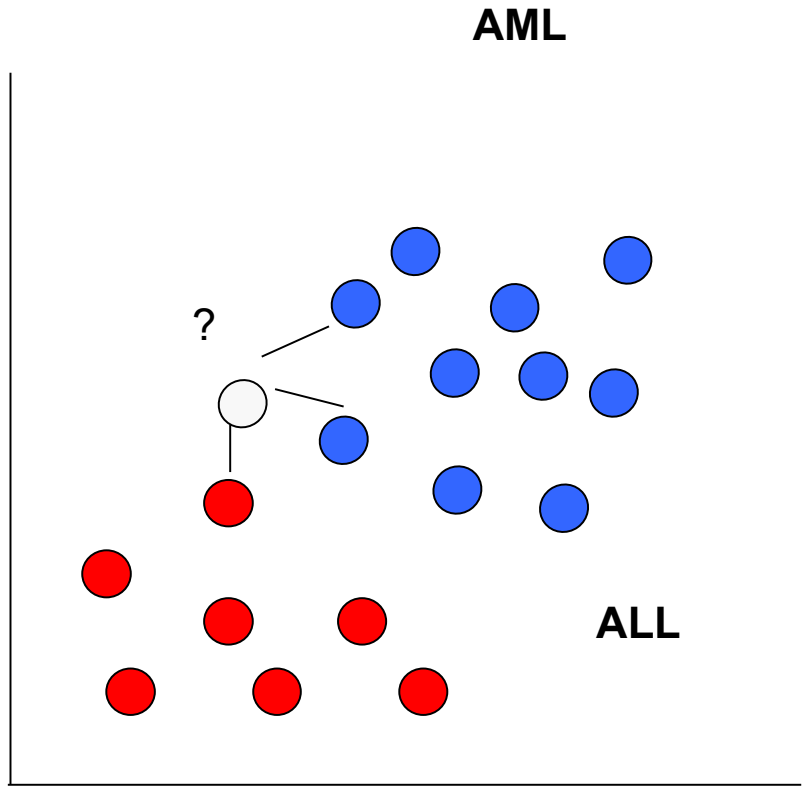
Jak (ne) predikovat téměř cokoliv

Biomarkery jako pomůcky pro diagnostiku, predikci odpovědi na léčbu nebo prognózu

- Používáme metody klasifikace!

Predikce a klasifikace

- V tomto typu analýzy se snažíme předpovědět příslušnost k jedné ze známých skupin na základě jejich molekulárního profilu
- Například určujeme:
  - diagnózu
  - odpověď na terapii
  - přežití pacienta
  - …



- Cílem je **vytvořit klasifikační pravidlo (soubor pravidel)**, které toto umožní
- Vytvoření klasifikátoru může sloužit jako **nástroj pro selekci genů**, které významně diskriminují mezi skupinami

Princip tvorby klasifikátoru

1. Výběr proměnných pro klasifikaci

- Vybíráme geny nebo proteiny, které se v klasifikátoru použijí

2. Trénování

- Na trénovacích datech vytvoříme klasifikační pravidlo (klasifikátor, model)

3. Testování

- Vytvořený klasifikátor se otestuje na testovacích datech
- K odhadnutí výkonnosti (přesnosti) klasifikátoru a optimalizaci parametrů

Výběr proměnných I.

Důvody výběru proměnných

- **Ze statistického hlediska**
  - Eliminace tisíců nerelevantních genů významně ovlivní komplexitu vybraného klasifikátoru, stane se robustnější.
- **Z biologického hlediska**
  - Výběr vhodných genů/proteinů silně korelovaných s danou skupinou pomůže pochopit mechanismus jejich působení.
- **Z praktického hlediska**
  - Čím méně genů potřebujeme pro predikci, tím snadnější je uplatnění klasifikátoru v praxi.

Výběr proměnných II.

- U omics dat je výběr proměnných trochu problematický, protože jsou velmi korelované
  - Výběr jednoho reprezentanta je víceméně náhodný
  - Malé změny v trénovacích datech, případně aplikace jiného klasifikátoru může vyústit do úplně jiné selekce genů
  - To je v pořádku, ale pozor na interpretaci!

- Při celkové interpretaci je třeba brát v potaz, že se jedná pouze o **podskupinu** genů

- Biologické závěry o podskupinách vzorků by měly být založené na studiu celé množiny významných genů

Příklad

ARTICLE

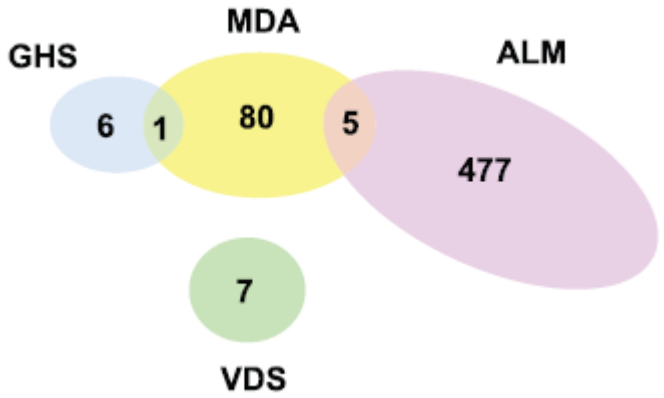
Test of Four Colon Cancer Risk-Scores in Formalin Fixed Paraffin Embedded Microarray Gene Expression Data

Antonio F. Di Narzo, Sabine Tejpar, Simona Rossi, Pu Yan, Vlad Popovici, Pratyaksha Wirapati, Eva Budinska, Tao Xie, Heather Estrella, Adam Pavlicek, Mao Mao, Eric Martin, Weinrich Scott, Fred T. Bosman, Arnaud Roth, Mauro Delorenzi

Manuscript received December 9, 2013; revised April 22, 2014; accepted July 2, 2014.

Table 1. Description of the four risk scores analyzed*

	Risk scores			
Abbreviation	**GHS**	**VDS**	**MDA**	**ALM**
Developer	Genomic Health	Veridex	MD Anderson	ALMAC diagnostics
Type of assay	Q-RT-PCR	microarray and Q-RT-PCR	microarray	microarray
Type of tissue	FFPE	fresh frozen and FFPE	fresh frozen	FFPE
Main publication	O'Connell et al. 2010.	Jiang et al. 2008.	Oh et al. 2011.	Kennedy et al. 2011.
Total number of features	7	7	114 (86 genes)	634 (482 genes)
Features used (genes)	7	6	85 (85 genes)	634 (identical platform)



MUNI | RECETOX

Je statisticky významně odlišně exprimovaný gen vhodný pro klasifikaci?

Metody klasifikace

Black-box metody

Ke klasifikaci nového vzorku používají celý trénovací soubor.

Obvykle nejsou jednoduše interpretovatelné

K-nejbližších sousedů

Support vector machines

Neuronové sítě

Metody vytvářející srozumitelná klasifikačná pravidla

Více intuitivní, jednoduše použitelné v praxi

Pouze na vybraných proměnných

Regresní modely

Diskriminační analýza

Klasifikační stromy a lesy

Top scoring pairs

AdaBoost...

Odhad výkonnosti klasifikátoru I

- **Výkonnost každého klasifikátoru musí být testována**

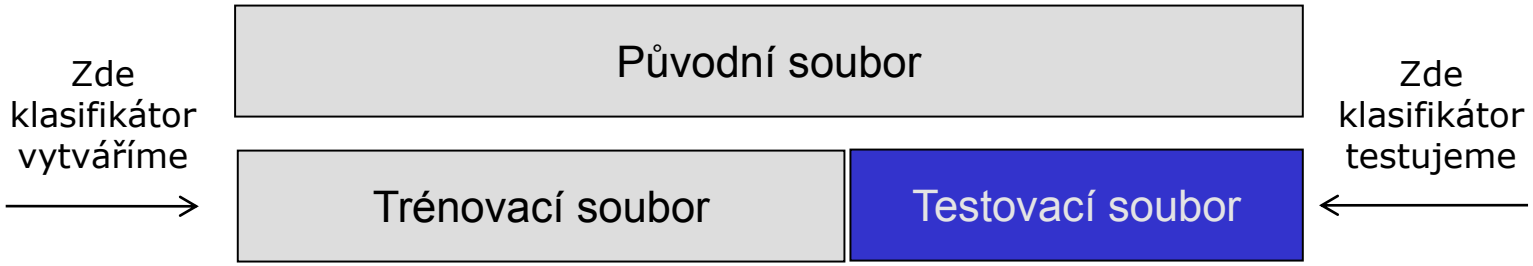
- Proč odhadovat výkonnost klasifikátoru?
  - Omezení trénovacím souborem
  - Bez předpokladu o rozložení neexistuje žádný vzorec pro výpočet velikosti vzorku
  - Často existuje jen jeden datový soubor pro trénování a testování klasifikátoru

POZOR - Odhad výkonnosti klasifikátoru na trénovacích datech je VŽDY optimisticky zkreslený proto **nutnost testovat na nezávislém souboru**

Odhad výkonnosti klasifikátoru II

Základní myšlenka:

- Převzorkováním rozdělit (opakovaně) datový soubor na trénovací a testovací, vytvořit klasifikátor na trénovacím souboru a změřit výkonnost klasifikátoru jen na datech, které neb použity pro jeho vytvoření (**křížová validace**)



- **UPOZORNĚNÍ:** Všechny kroky, které závisí na převzorkování, a které vedou k finálnímu **modelu musí být zopakované identicky na každém rozdělení na trénovací a testovací soubor**.
  - Patří sem například výběr proměnných, trénovaní klasifikátoru, optimalizace parametrů,...

Odhad výkonnosti – proč nestačí křížová validace

- Každé dva trénovací soubory vytvořené z původního datového soubor pomocí převzorkování se do jisté míry překrývají -> vytvořené klasifikáto tedy nejsou úplně nezávislé

- Variabilita je obvykle podhodnocená

- NUTNOST TESTOVAT NA JINÉM VALIDAČNÍM SOUBORU

Co získáme odhadem výkonnosti?

- Zjistíme **očekávanou výkonnost klasifikátoru** na validačním, nebo jakémkoliv jiném souboru!

- **Můžeme identifikovat nejstabilnější proměnné** (geny/proteiny) – tedy ty, které jsou vybrány nejčastěji!

- **Zjistíme**, které vzorky bývají často špatně klasifikované (pokud takové jsou, naznačuje to **odlehlé hodnoty**)

Vyhodnocení přesnosti klasifikátoru

		Klasifikace	
		Zdravý (negativní)	**Nemocný (pozitivní)**
Skutečnost	**Zdravý (negativní)**	Pravdivá negativita (PN)	Falešná pozitivita (FP) Chyba I. druhu
	Nemocný (pozitivní)	Falešná negativita (FN) Chyba II. druhu	Pravdivá pozitivita (PP)

Vyhodnocení přesnosti klasifikátoru

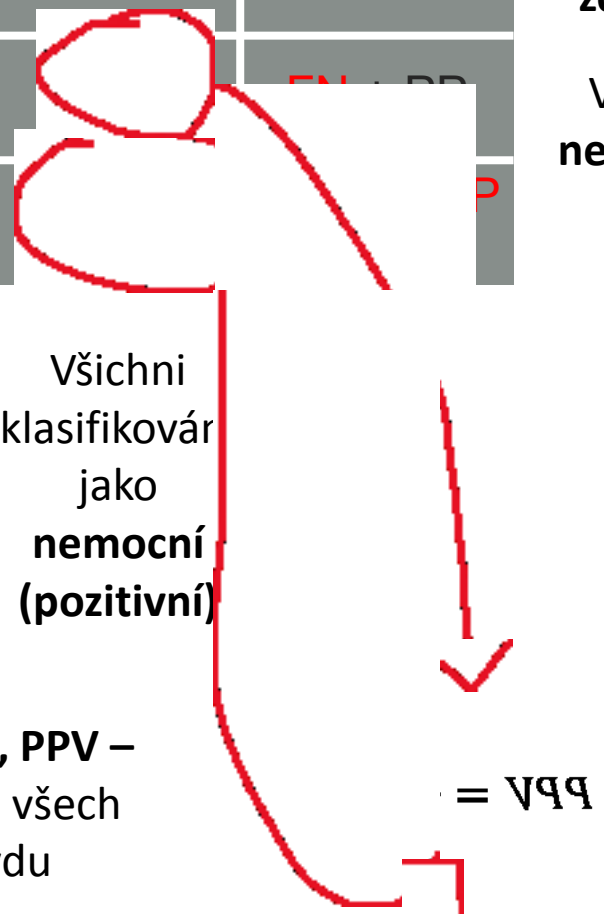
		Klasifikace		
		Zdravý (-)	Nemocný (+)	Celkem
Skutečnost	Zdravý (-)	PN	FP	PN + FP
	Nemocný (+)	FN		
	Celkem	PN + FN		

Všichni skutečně **zdraví (negativní)**

Všichni skutečně **nemocní (pozitivní)**

Všichni klasifikováni jako **zdraví (negativní)**

Všichni klasifikovár jako **nemocní (pozitivní)**



Pozitivní prediktivní hodnota (precision, PPV – positive predictive value) – jaký podíl ze všech klasifikovaných jako nemocných je opravdu nemocných?

MUNI | RECETOX

Vyhodnocení přesnosti klasifikátoru

		Klasifikace		
		Zdravý (-)	Nemocný (+)	Celkem
Skutečnost	Zdravý (-)	PN	FP	PN + FP
	Nemocný (+)	FN		
	Celkem	PN + FN	FP + P	

Všichni skutečně **zdraví (negativní)**

Všichni skutečně **tivní)**

Všichni klasifikováni jako **zdraví (negativní)**

Všichn klasifikov jako **nemoc (pozitiv**

Senzitivita / Úplnost (sensitivity/recall/TPR - true positive rate) – jaký podíl skutečně nemocných odhalíme?

$$\frac{TP}{TP + FN} =$$

MUNI | RECETOX

Vyhodnocení přesnosti klasifikátoru

		Klasifikace	
		Zdravý (-)	Nemocný (+)	Celkem
Skutečnost	Zdravý (-)		FP	
	Nemocný (+)	FN		
	Celkem	PN + FN		

Všichni skutečně **zdraví (negativní)**

Všichni klasifikováni jako **zdraví (negativní)**

Specificita (specificity) – ze všech, kteří jsou zdraví, jaký podíl byl označen za zdravých?

$$\frac{PN}{PN + FP} = x$$

MUNI | RECETOX

		Klasifikace	
		Zdravý (-)	Nemocný (+)
Skutečnost	Zdravý (-)	PN	
	Nemocný (+)	FN	PP
	Celkem	PN + FN	FP + P

Všichni skutečně **zdraví (negativní)**

Všichni klasifikováni jako **zdraví (negativní)**

Všichn klasifikov jako **nemoc (pozitivi**

Podíl falešné pozitivity (FPR) – ze všech, kteří jsou zdraví, jaký podíl byl označen za nemocných?

$$\frac{FP}{PN + FP}$$

Vyhodnocení přesnosti klasifikátoru

MUNI | RECETOX

Vyhodnocení přesnosti klasifikátoru

		Klasifikace		
		Zdravý (-)	Nemocný (+)	Celkem
Skutečnost	Zdravý (-)		FP	PN + FP
	Nemocný (+)	FN		
	Celkem	PN +		

Všichni skutečně **zdraví (negativní)**

Všichni skutečně **í (pozitivní)**

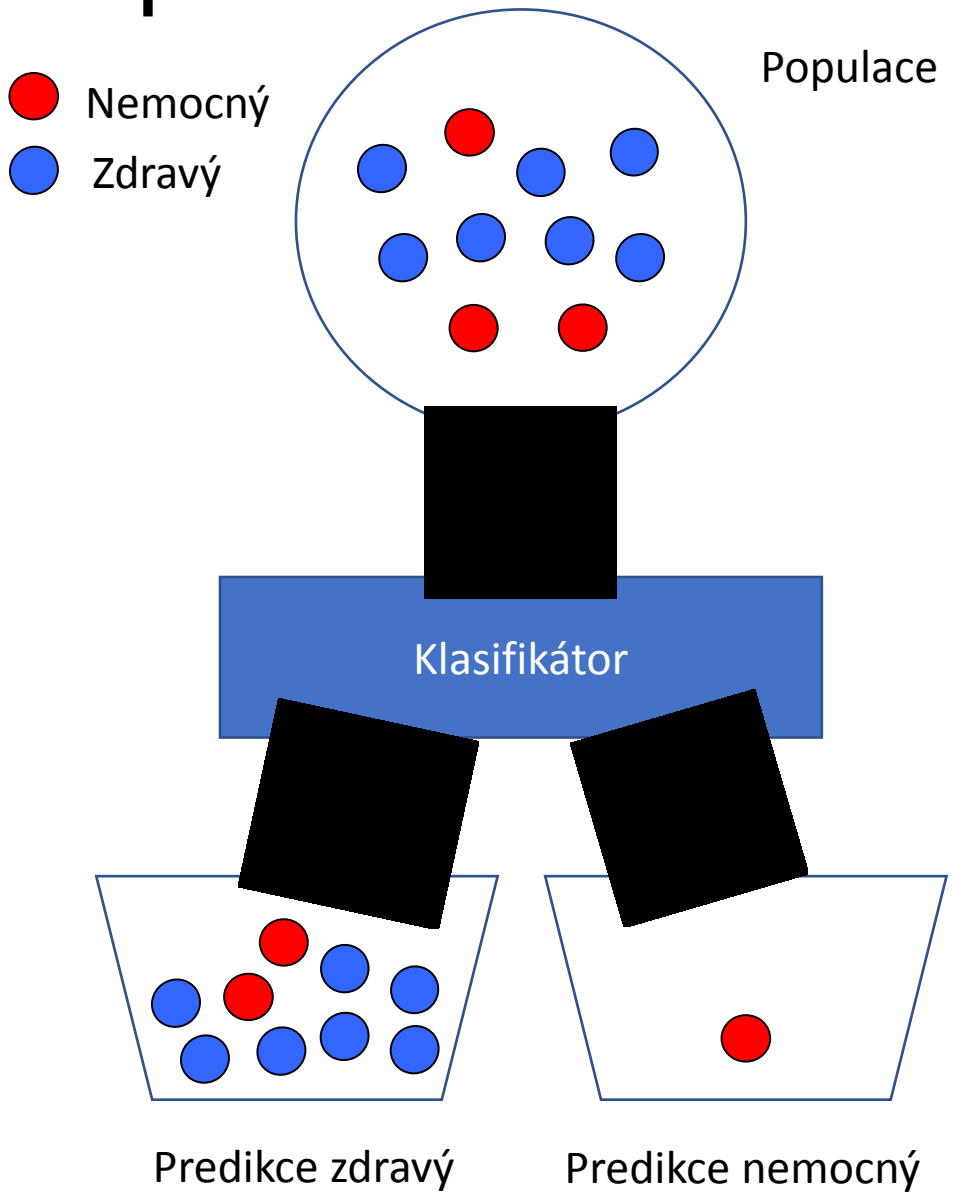
Všichni klasifikov jako **zdra (negativ**

Celková přesnost (accuracy) – jaké procento je správně klasifikováno?

$$\frac{PN + PP}{(PP + FN + PP)}$$

MUNI | RECETOX

Vyhodnocení přesnosti klasifikátoru – příklad 1



		Klasifikace		
		Zdravý	Nemocný	Celkem
Skutečnost	Zdravý	7	0	7
	Nemocný	2	1	3
	Celkem	9	1	10

$$\text{specificita} = \frac{PN}{PN + FP} = \frac{7}{7} = 100\%$$

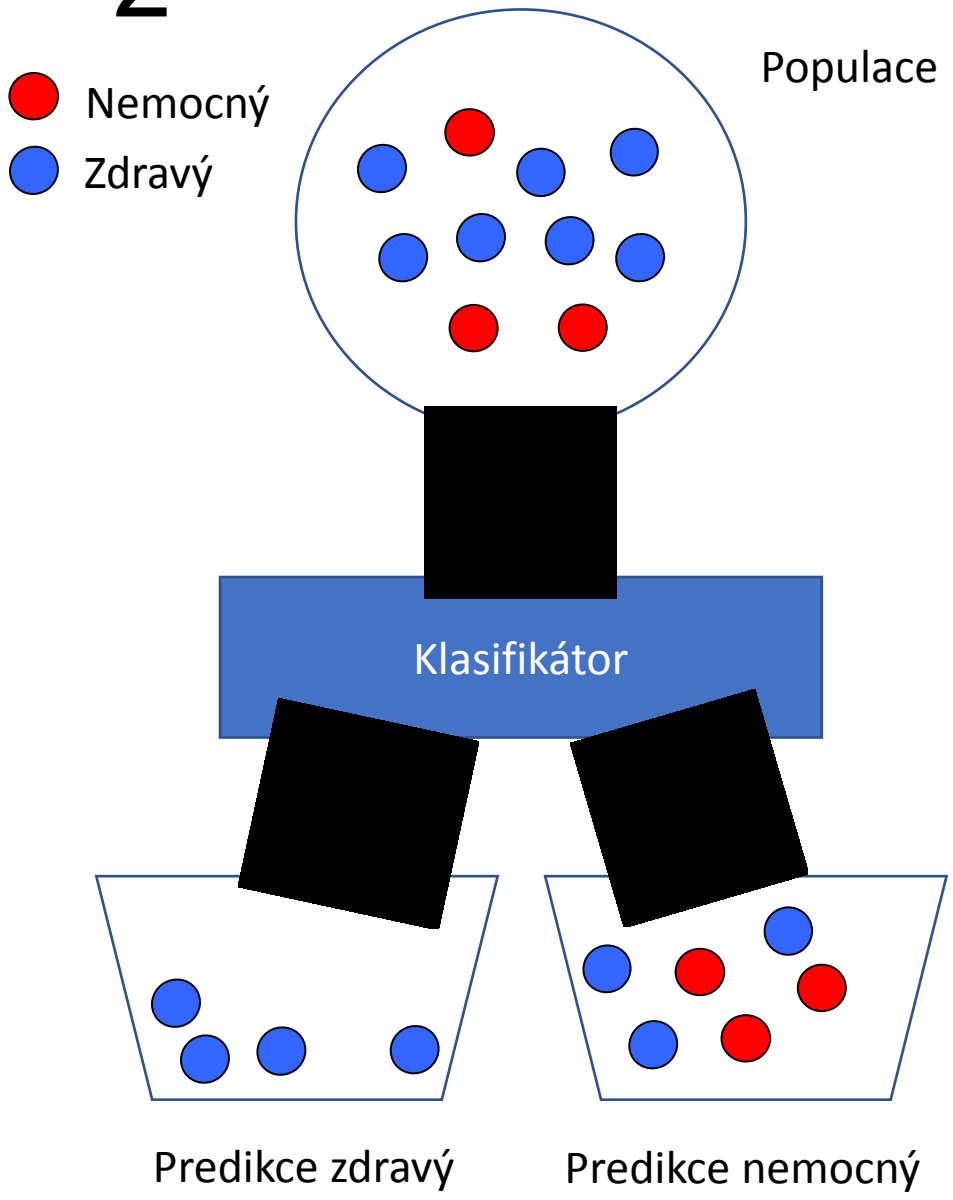
$$\text{FPR} = \frac{FP}{PN + FP} = \frac{0}{7} = 0\%$$

$$\text{senzitivita} = \frac{PP}{FN + PP} = \frac{1}{3} = 33\%$$

$$PPV = \frac{PP}{FP + PP} = \frac{1}{1} = 100\%$$

$$\text{přesnost} = \frac{PN + PP}{(PN + FP + FN + PP)} = \frac{7 + 1}{7 + 0 + 2 + 1} = \frac{8}{10} = 80\%$$

Vyhodnocení přesnosti klasifikátoru – příklad 2



		Klasifikace		
		Zdravý	Nemocný	Celkem
Skutečnost	Zdravý	4	3	7
	Nemocný	0	3	3
	Celkem	4	6	10

$$\text{přesnost} = \frac{PN + PP}{(PN + FP + FN + PP)} = \frac{4 + 3}{4 + 3 + 0 + 3} = \frac{7}{10} = 70\%$$

$$PPV = \frac{PP}{FP + PP} = \frac{3}{6} = 50\%$$

$$\text{specifičnita} = \frac{PN}{PN + FP} = \frac{4}{7} = 57\%$$

$$\text{senzitivita} = \frac{PP}{FN + PP} = \frac{3}{3} = 100\%$$

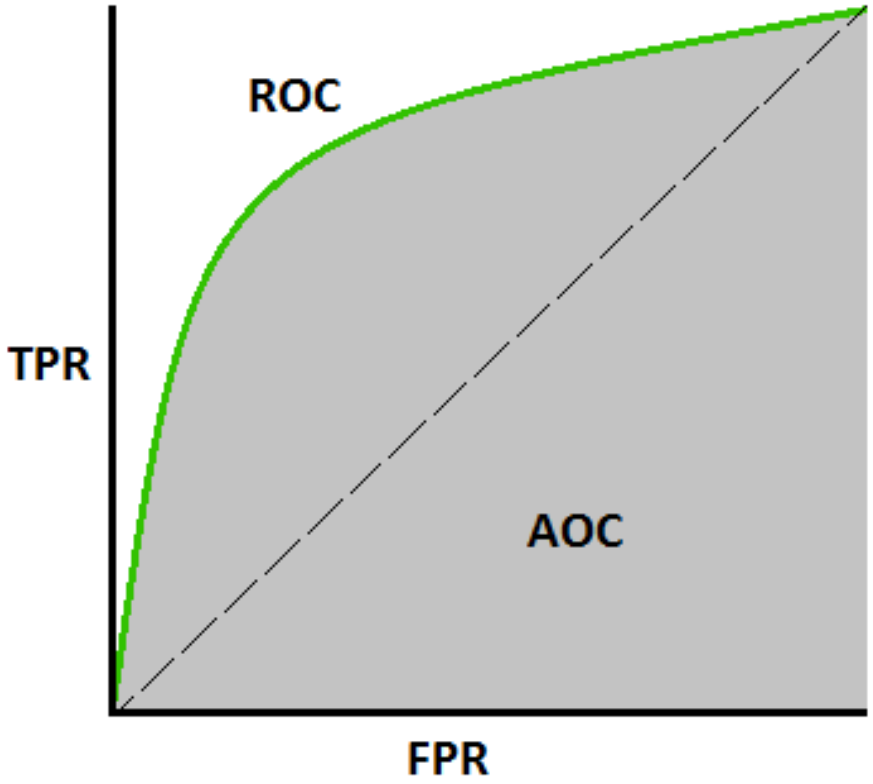
$$FPR = \frac{FP}{PN + FP} = \frac{3}{7} = 43\%$$

ROC křivka

- Receiver operator characteristics (ROC)
- Mějme binární klasifikátor který má být založený na nějaké proměnné (například na velikosti exprese genu)
- Musíme zvolit hranici exprese genu, která bude rozdělovat vzorky na pozitivní a negativní
- ROC křivka ukazuje, **jak dobrý klasifikátor jsme schopni na základě této proměnné sestavit** z pohledu senzitivity a specificity

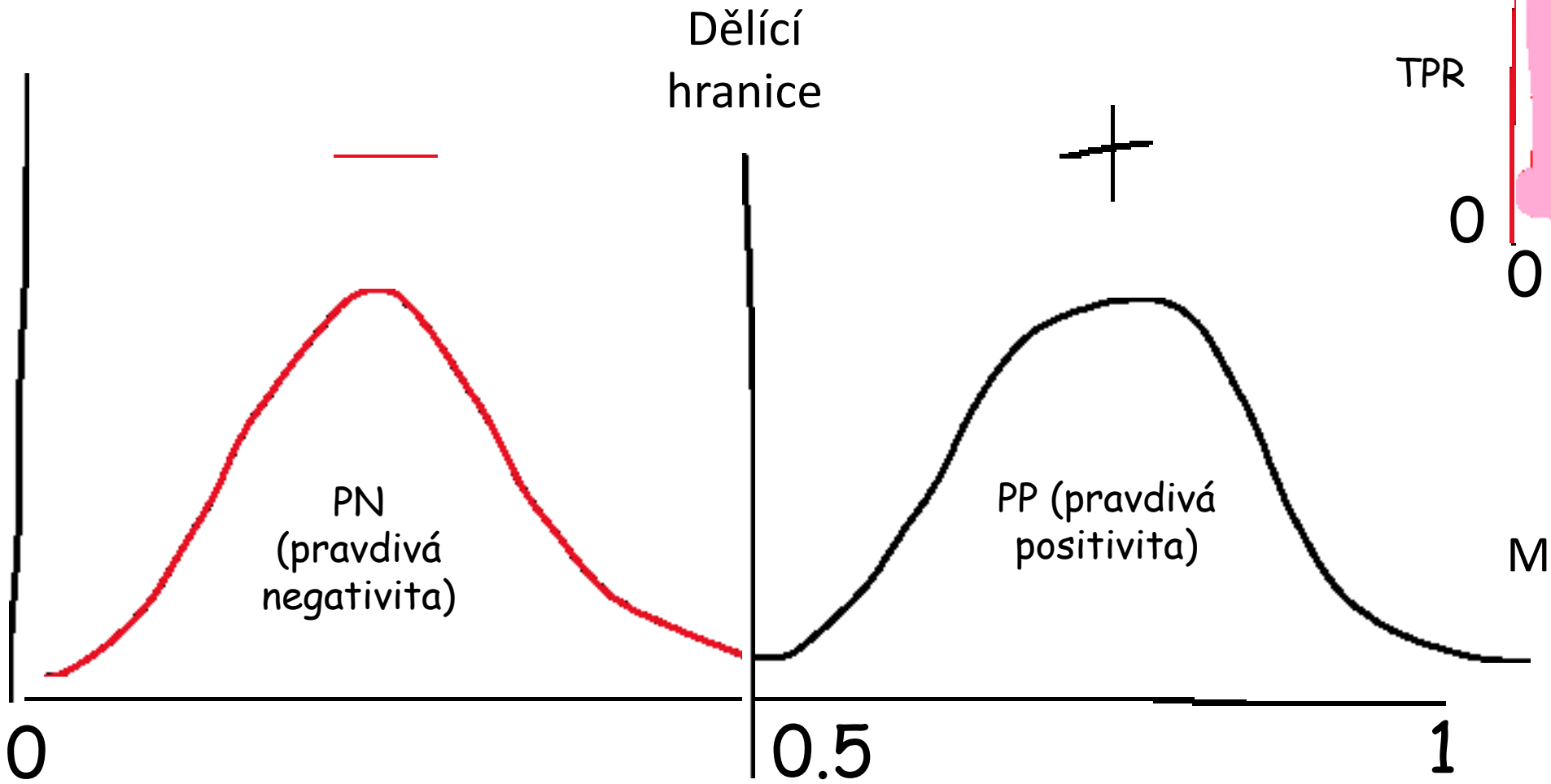
ROC křivka

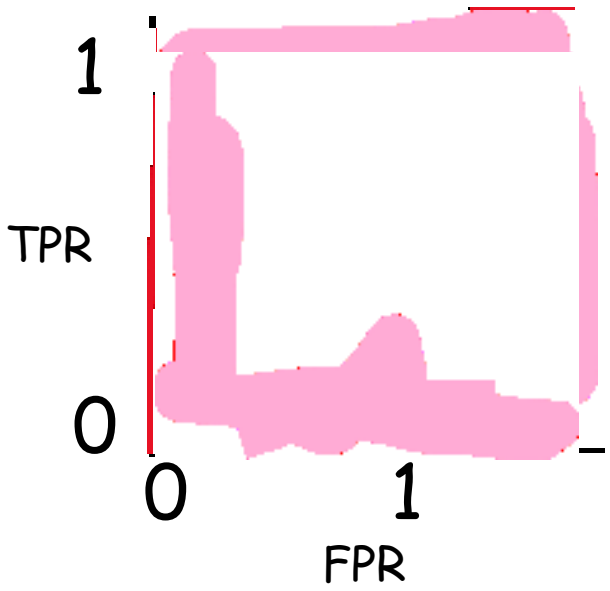
- Receiver operator characteristics (ROC)



- ROC křivka zobrazuje vztah mezi FPR a TPR
- AUC – area under curve (plocha pod křivkou) - míra přesnosti testu, vyjadřuje šanci, že model bude schopen rozlišit naše skupiny

ROC křivka



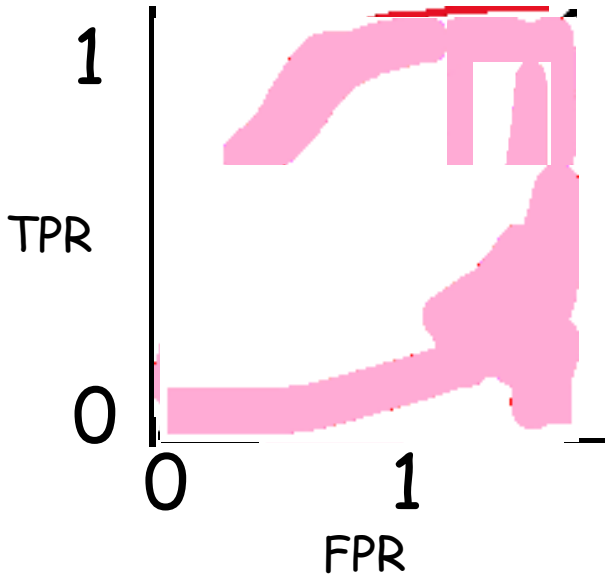


AUC = 1

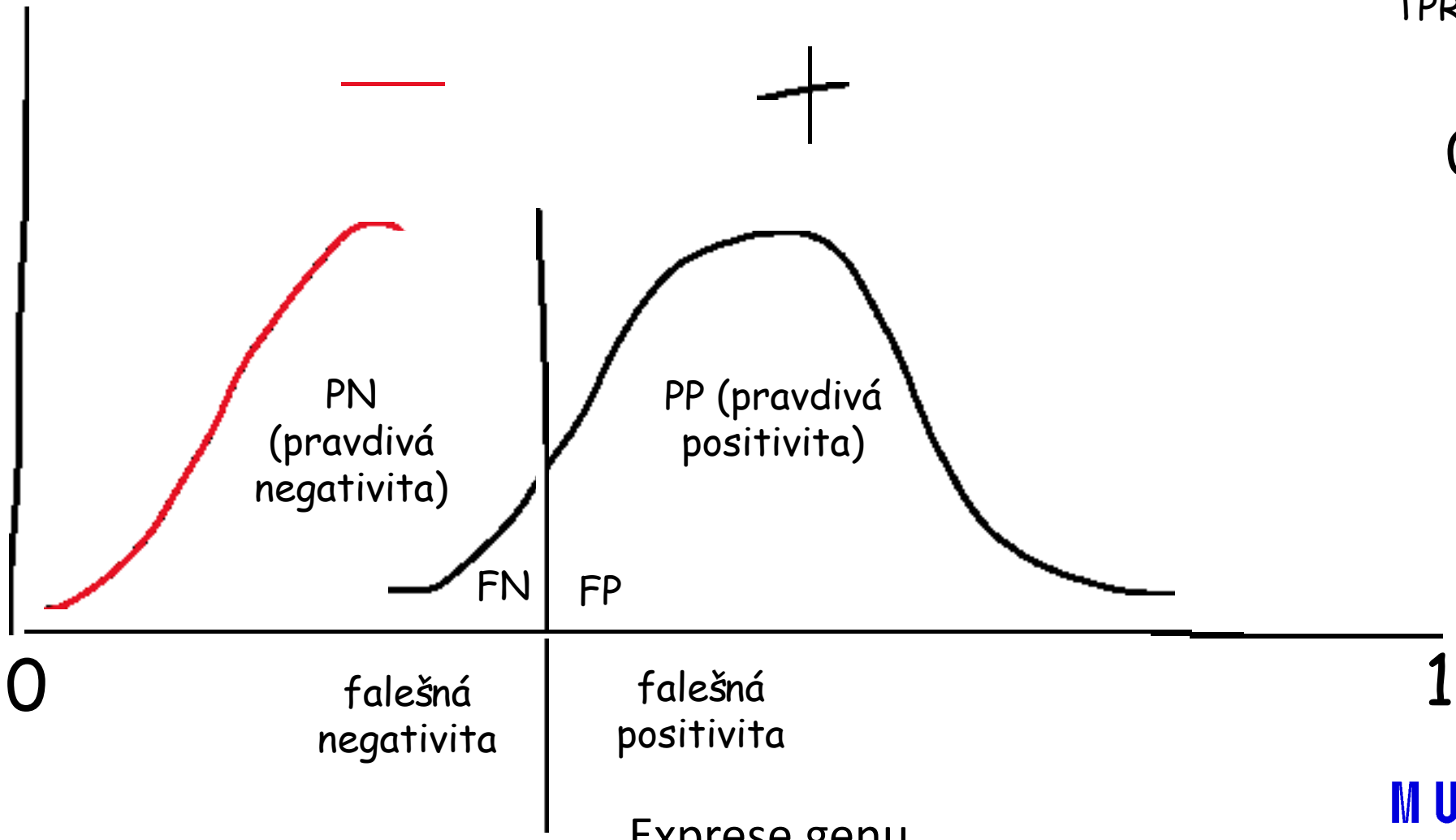
Model perfektně diskriminuje

MUNI | RECETOX

ROC křivka

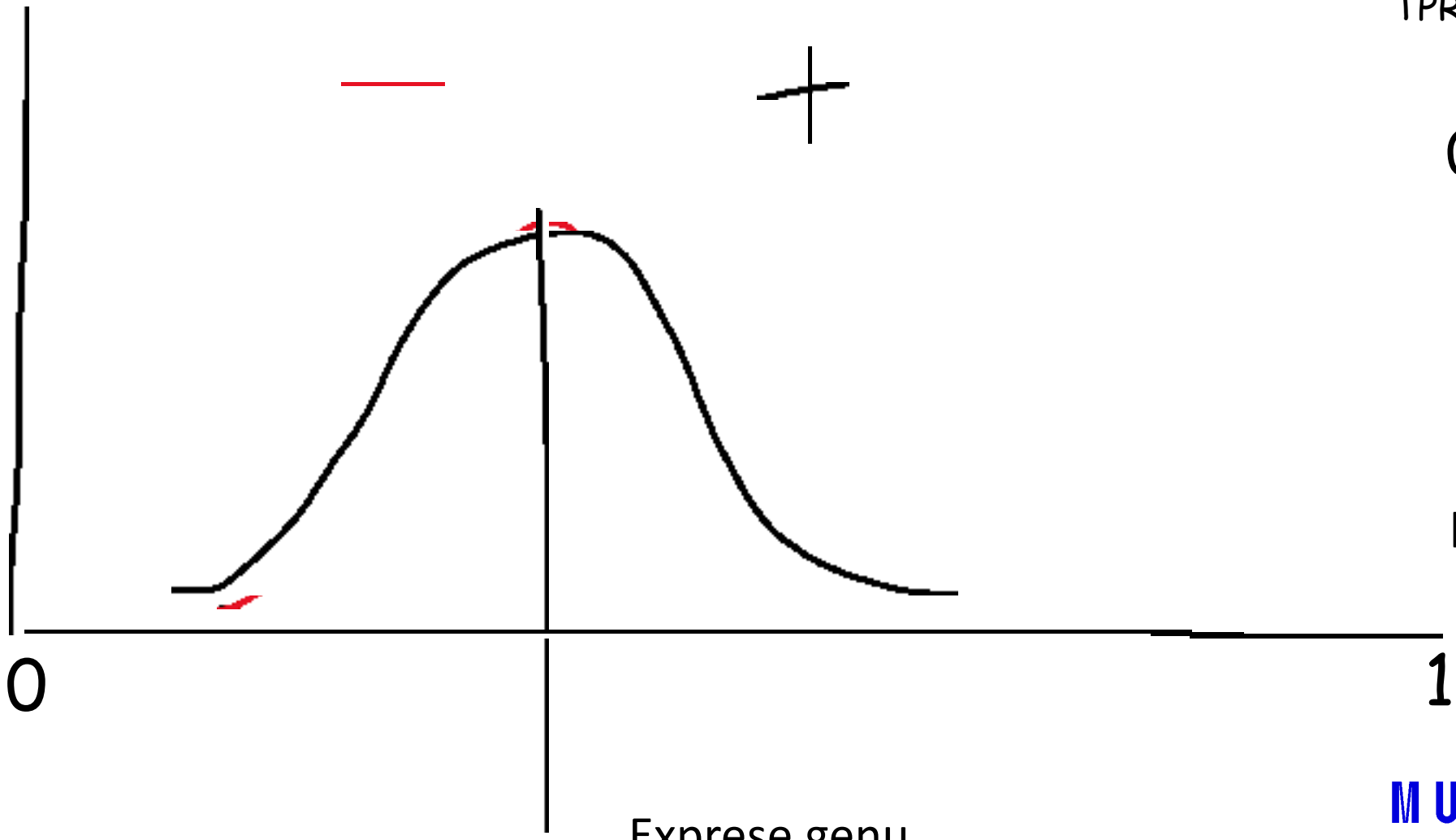


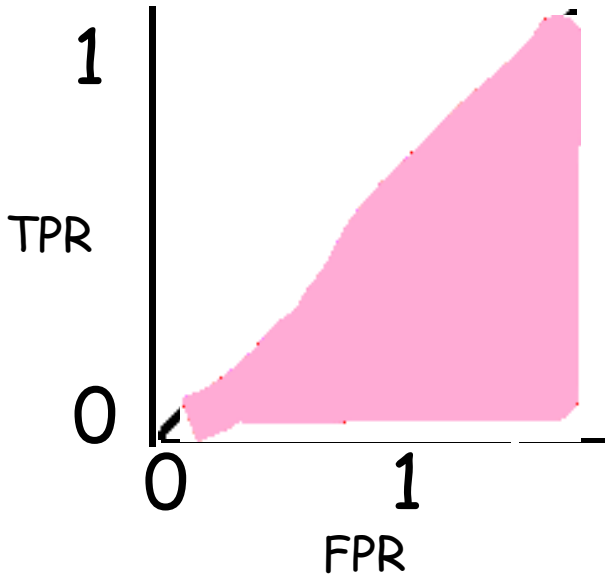
AUC = 0.8



MUNI | RECETOX

ROC křivka



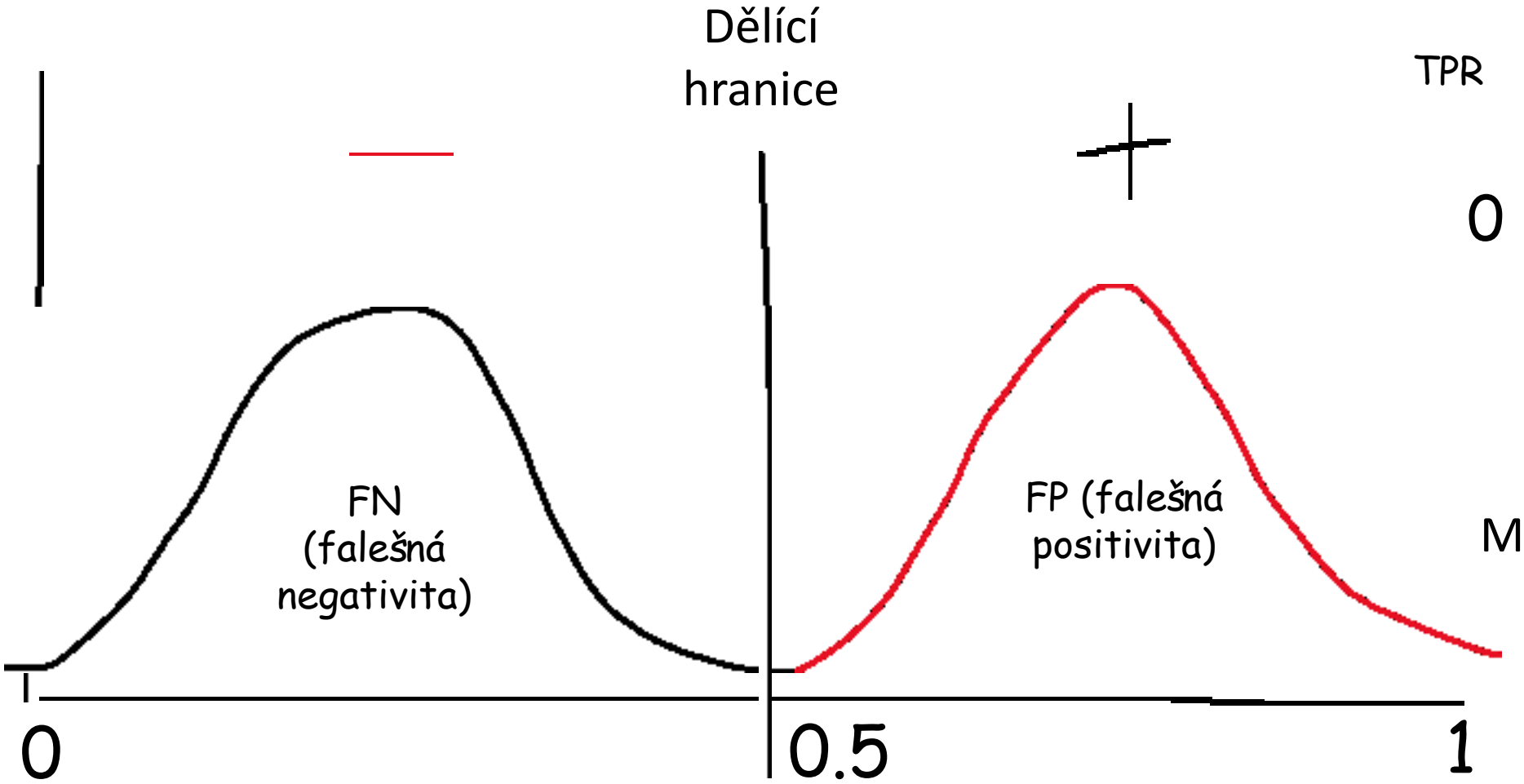


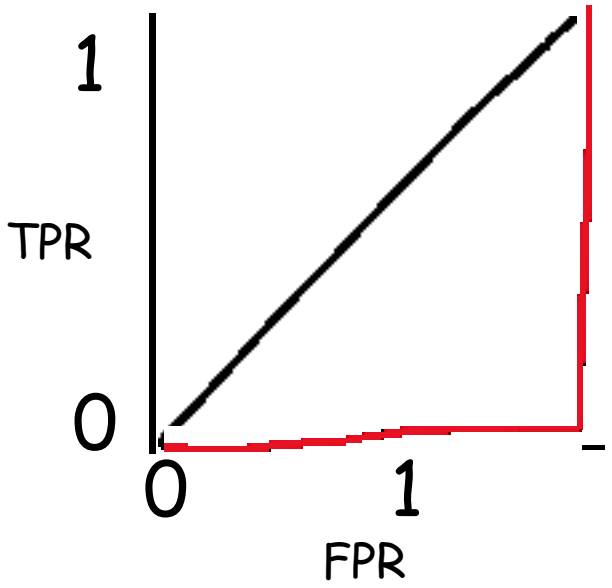
$AUC = 0.5$

Model není lepší než hod mincí (proměnná nemá žádnou diskriminační schopnost)

MUNI | RECETOX

ROC křivka





AUC = 0

Model určuje přesně naopak!

MUNI | RECETOX

ROC křivka

Animace principu (jak se křivka kreslí)

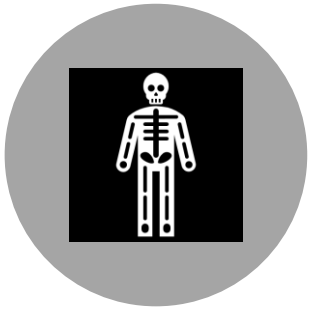
http://arogozhnikov.github.io/2015/10/05/roc-curve.html

Bez validace není publikace (?)

Není validace jako validace

Validace samotného **výběru biomarkeru** na cílové populaci!

STATISTICKÁ VALIDACE



JE PŘÍMO SOUČÁSTÍ PROCESU VÝBĚRU BIOMARKERU

JEŠTĚ PŘED JAKOUKOLIV JINOU VALIDACÍ!

MUNI | RECETOX

Validace **technologie**

zopakujeme-li experiment, dostaneme stejné výsledky? (**technické replikáty** stejných vzorků)

potvrdí výsledek na tom samém vzorku i jiná (standardní) technologie? **mikročip** vs. **qPCR**

Validace **biologická**

Dává to celé smysl?

Na jakých úrovních se biomarker projevuje – genová exprese? Koreluje s proteinovou expresí?...

Validace **aplikovatelnosti**

Lze nalezené biomarkery uvést do klinické praxe?

Jsou nalezené geny přepsány do proteinů?

Lze pro nalezené proteiny nalézt protilátku na imunohistochemické barvení?

MUNI | RECETOX

nature biotechnology

The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models

MAQC Consortium*

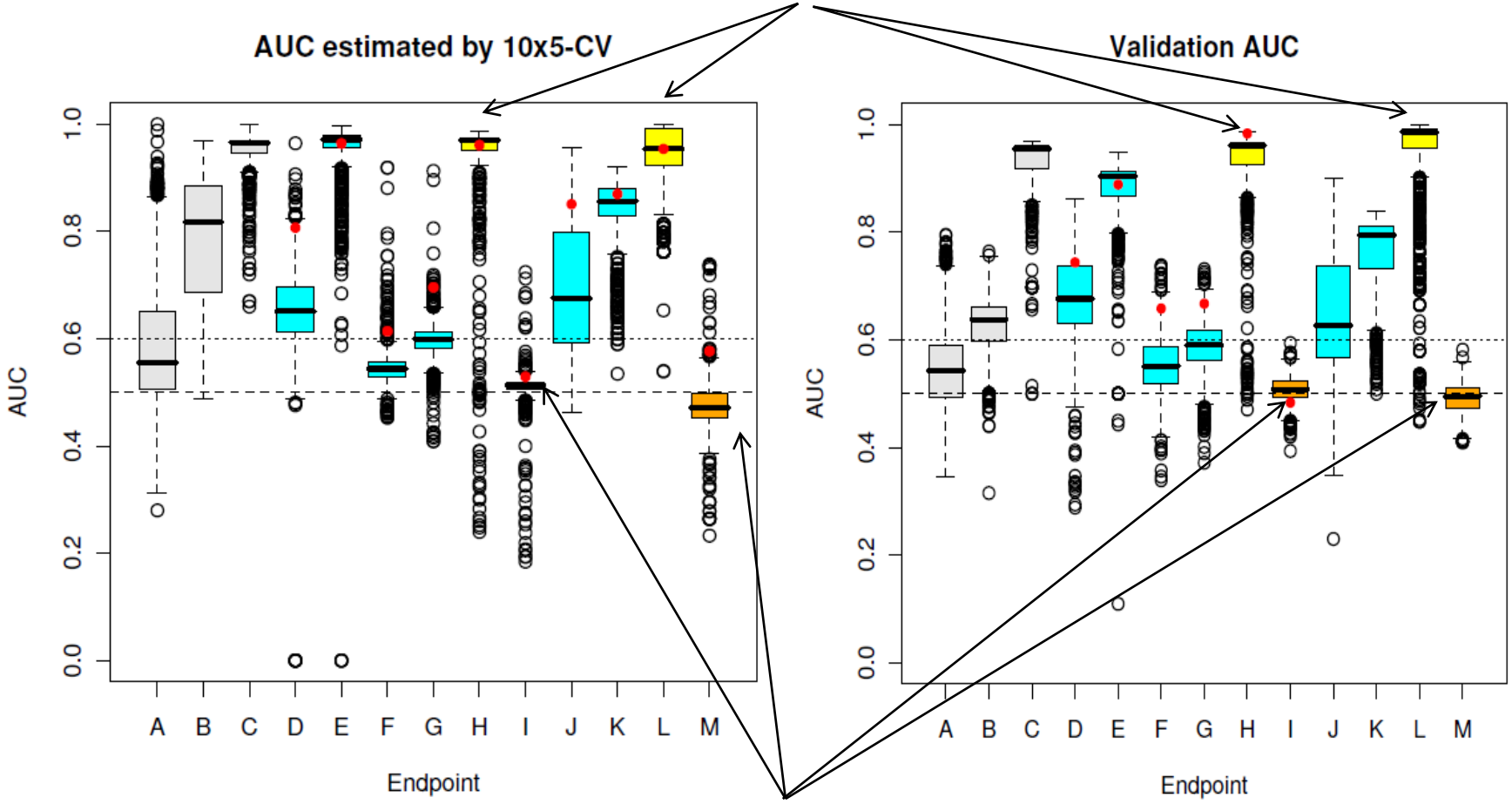
- **36** nezávislých tímů analytiků z celého světa analyzovalo **6** mikročipových studií a vytvořilo klasifikátory pro predikci **13** endpointů (ER+ vs ER-, ...)
- Každý tým navrhl plán tvorby a validace klasifikátoru
- Tyto plány byly předem posouzeny odbornými statistiky a ohodnoceny dle jejich názoru na škále od 1 do 10

MAQC II – endpointy

štúdia	endpoint	model
A	Lung tumorigen vs non tumorigen	mouse
B	Non genotoxic liver carcinogens vs non-carcinogens	rat
C	Liver toxicants vs non-toxicants based on overall necrosis score	rat
D	Breast cancer - Pre-operative treatment response (pCR, pathologic complete response)	human
E	Breast cancer – Estrogen receptor status	human
F	Multiple myeloma – overall survival milestone outcome	human
G	Multiple myeloma – event-free survival milestone outcome	human
H	Clinical parameter S1 – positive control, gender	human
I	Clinical parameter S1 – random assignment, negative control	human
J	Neuroblastoma – overall survival milestone outcome	human
K	Neuroblastoma – event-free survival milestone outcome	human
L	Newly established parameter – positive control, gender	human
M	Newly established parameter – negative control, random	human
		human

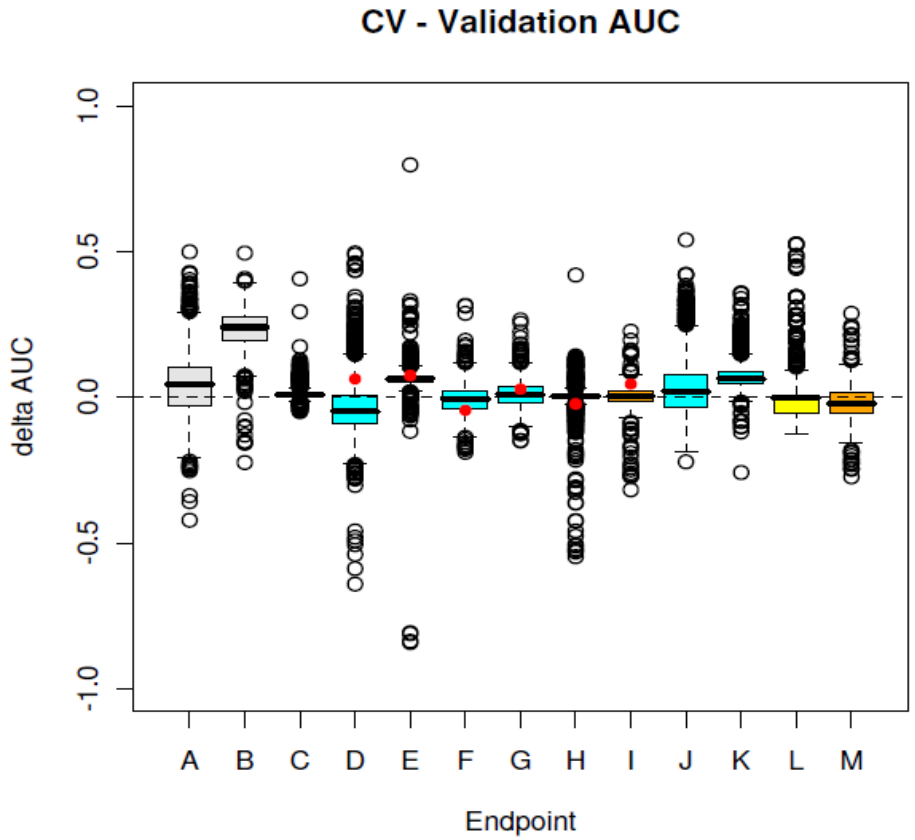
Výkonnost klasifikátorů dle experimentu

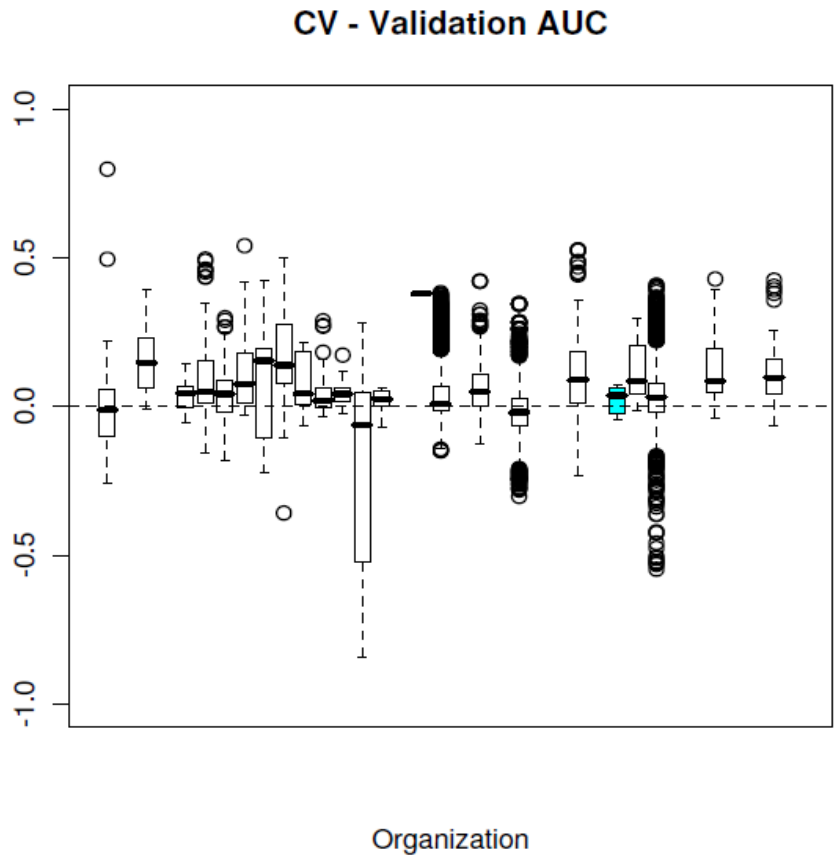
Úspěšnost odhadu pohlaví, pozitivní kontrola



Úspěšnost predikce náhodného zařazení, negativní kontrola

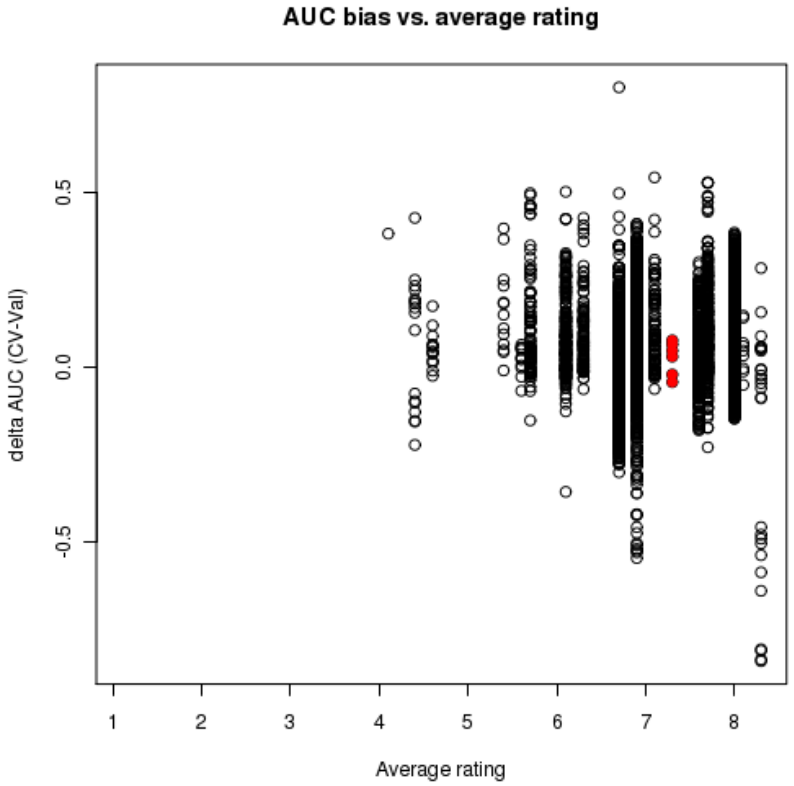
Rozdíl výkonnosti odhadnuté na základě krosvalidace (CV) a na validačním souboru (Validation)





Rozdíl v AUC (plocha pod ROC křivkou) mezi odhadem výkonu krosvalidací a výkonu na validačním souboru by měl být 0

Aby to nebylo jednoduché...



To, že se algoritmus zdál hodnotitelům správný neznamená, že opravdu byl...

Rozdíl v AUC (plocha pod ROC křivkou) mezi odhadem výkonu krosvalidací a výkonu na validačním souboru jako funkce **průměrného hodnocení externími hodnotiteli navržených algoritmů**

Bez validace není (dobrá) publikace

Doporučené předměty

- PřF:Bi7490 Pokročilé neparametrické metody
- PřF:Bi0034 Analýza a klasif. dat - Informace o předmětu
- PřF: ENV003 Environmentální informace a modelování – specifika u chemických dat