MUNI | RECETOX

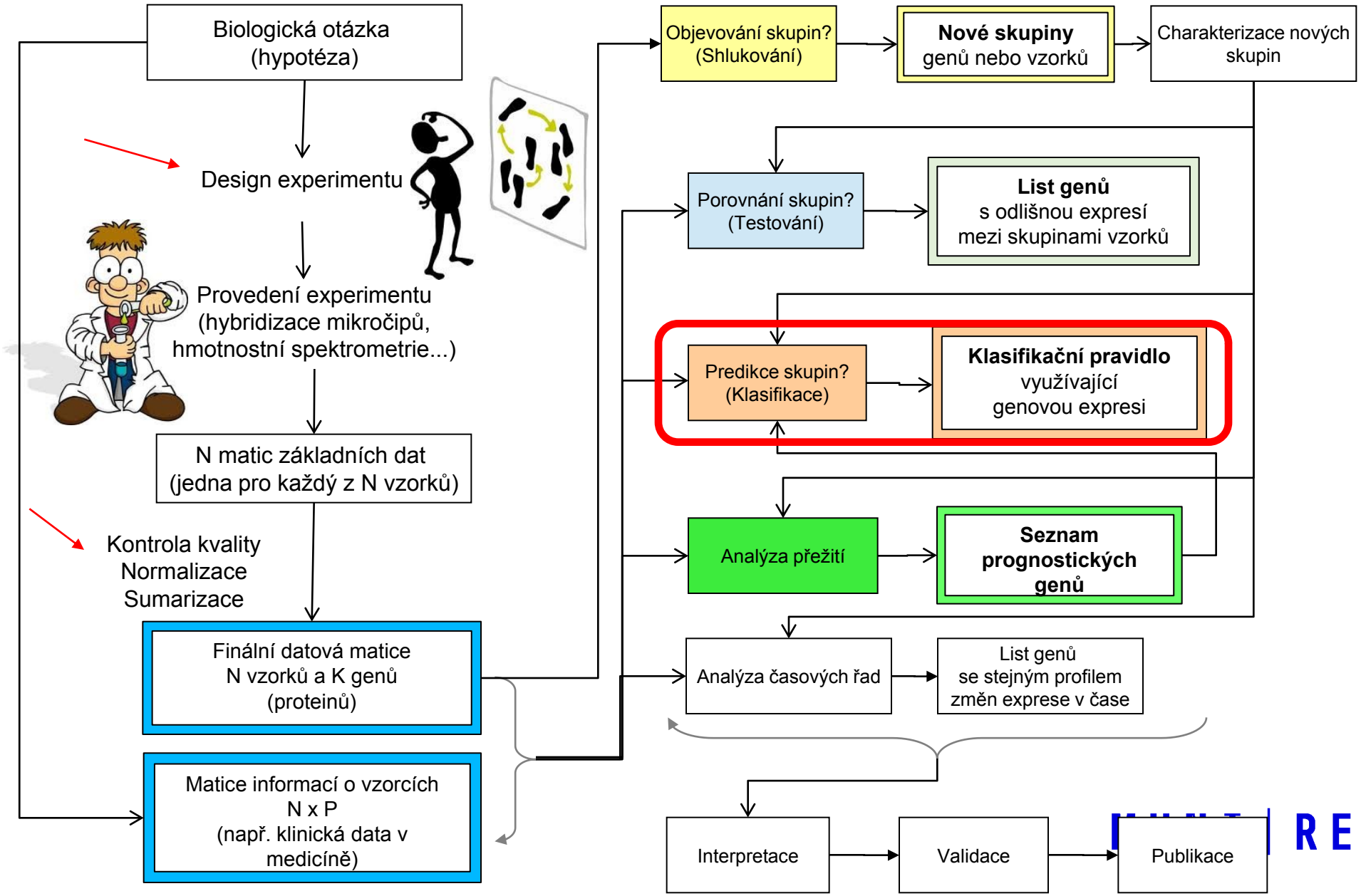
Analýza genomických a proteomických dat

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

- Jaro 2024

Klasifikace (předpovídání skupin)

Společné schéma analýzy dat



RECETOX

Tradiční schéma analýzy

- **Učení s učitelem (supervised learning)**
  - V tomto případě zobecňujeme známou strukturu dat na nové data
  - **Porovnávání skupin (class comparison)**
    - hledáme rozdíly v expresi, počtu kopií genů nebo abundanci proteinů mezi již definovanými skupinami
  - **Předpovídání skupin (class prediction)**
    - na známých skupinách se snažíme vytvořit klasifikátor, který by dokázal zařadit nového pacienta do jedné ze skupin

- **Učení bez učitele (unsupervised learning)**
  - V tomto případe struktura v datech není známá a musíme ji objevit
  - **Objevování skupin (class discovery)**
    - na základě informací o genech/proteinech hledáme nové skupiny
    - onemocnění X je velmi heterogenní a snažíme se identifikovat specifičtější podtypy, které by mohli být cílem cílené terapie

Co je to biomarker?

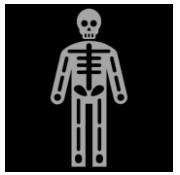
Biologický marker (biomarker):

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)

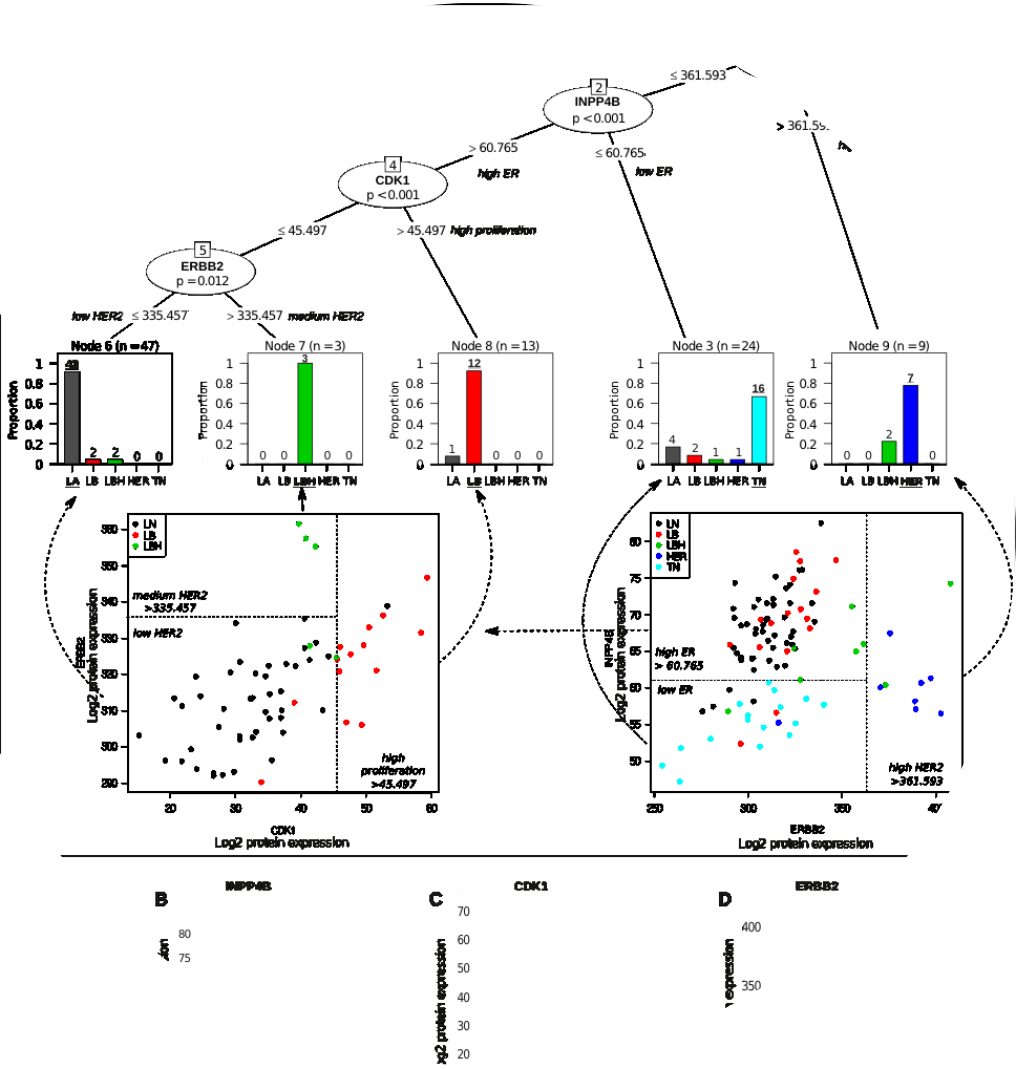
Přítomnost mikroorganizmu

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 dobrý klasifikátor 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)

… tvorba klasifikátorů z *molekulárních dat z omicsových technologií* má svá specifika…

Obsahují **množství šumu** (technická i biologická variabilita)

Nejsou skutečnými hodnotami (koncentrace, počty) sledovaných molekul

Pocházejí z komplexních technologií, které bývají **velice citlivé na vnější vlivy**

Jejich předzpracování pro statistickou analýzu je **náročné** a **vysoce specifické** pro daný typ platformy

Počet vzorků je mnohem **menší než** počet sledovaných **proměnných**.

Zkoumané **proměnné jsou často korelované** a mají mezi sebou komplexní vztahy (geny, proteiny…)

Specifika dat z omics experimentů

MUNI | RECETOX

Skandál na Duke university

Severní Karolína, USA

MUNI | RECETOX

2006 – Anil Potti, nadějný vědec z Duke University publikuje v Nature Medicine s kolegy článek o **biomarkerech rezistence na chemoterapeutika v onkologii**.

Genomické signatury byly odvozeny z analýzy exprese (mikročipy) senzitivních a rezistentních buněčných linií, výsledky validovány na pacientech.

Obrovský ohlas, v roce 2006 článek zařazen mezi

“The Top 6 Genetic Stories of 2006”

ARTICLES

nature medicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell

2006 – Biostatistici K. Coombes, J. Wang and K.A. Baggerly se snaží o aplikaci signatur na data výzkumníků z jejich univerzity, ovšem bez úspěchu.

Aktivně konzultují s autory článku.

Čím více se noří do dat, tím více mají pochybností o validitě závěrů a správnosti samotných dat!

2007 – Coombes a kol. publikují v Nature Medicine dopis zpochybňující Pottiho výzkum

(*Coombes, Wang, Baggerly. Microarrays: retracing steps, Nature Medicine, 2007*)

https://www.proquest.com/docview/223115891?accountid=16531&sourcetype=Scholarly%20Journals

2007 – Coombes a kol. publikují v Nature Medicine dopis zpochybňující Pottiho výzkum

(*Coombes, Wang, Baggerly. Microarrays: retracing steps, Nature Medicine, 2007*)

Reportují tyto chyby:

označení senzitivních a rezistentních buněčných linií nesedí!	tabulka se seznamem významných genů a jejich sond obsahuje systematickou chybu (posun o políčko) – geny nesedí se sondami, po korekci tabulky se podařilo reprodukovat pouze 3 ze 7 seznamů a výsledků senzitivity	Model rezistence na doxacel – podařilo se zreprodukovat pouze 31 z 50 genů publikovaných v článku, ostatních 19 bylo zřejmě přidáno ručně “aby byla validace úspěšná”	Autorský SW (algoritmus), který Potti používá, pracuje s validačními a testovacími daty společně. Po korekci této chyby jsou výsledky validace klasifikátorů špatné – na validačních datech téměř rovné náhodě.

Retracing steps - again

- https://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/HistoryOfCisPem/EmailWithNatMed/natMedLetter.pdf

Mezitím vycházejí další články:

Blood (2006), NEJM (2006), JCO (2007), Lancet Oncology (2007), JAMA (2008), PLOS (2008), PNAS (2008), Clin Can Res (2009)

V roce 2009 již **212 citací, několik klinických studií, stovky léčených pacientů**

Nature Medicine odmítá publikovat další odpovědi

Coombes a Baggerly tedy publikují své nálezy v statistickém časopise *Annals of Applied Statistics*, **2009** 3(4):1309-1334.

https://projecteuclid.org/journals/annals-of-applied-statistics/volume-3/issue-4/Deriving-chemosensitivity-from-cell-lines--Forensic-bioinformatics-and-reproducible/10.1214/09-AOAS291.full

https://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/index.html

MUNI | RECETOX

Jak skandál změnil svět omicsového výzkumu

Červenec 2010 – ředitel National Cancer Institute (NCI) Harold Varmus obdržel **dopis** od více než **30 statistiků** a bioinformatiků, ve kterém vyjádřili své obavy nad použitím několika testů založených na genové expresi, které se používali v již probíhajících klinických studiích na Duke University k predikci odpovědi na chemoterapii.

V důsledku vznikla komise Institutu medicíny (IOM), cílem které bylo sepsání doporučení pro vývoj testů z omicsových studií

MUNI | RECETOX

ANIL POTTI

CASE PROGRESSION

JULY 2010 — Potti is accused of falsifying information on his resume, and Duke launches an investigation into his work

NOV 2010 — Potti resigns

OCT 2011 — Patients in Potti's clinical trials file a lawsuit against the University

JAN 29 2015 THUR — LAWSUIT SET TO START

Více info: https://ori.hhs.gov/content/case-summary-potti-anil

Evolution of Translational Omics: Lessons Learned and the Path Forward

ISBN
978-0-309-22418-5

300 pages
6 x 9
PAPERBACK (2012)

Christine M. Micheel, Sharly J. Nass, and Gilbert S. Omenn, Editors; Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials; Board on Health Care Services; Board on Health Sciences Policy; Institute of Medicine

IOM (Institute of Medicine). 2012. *Evolution of Translational Omics: Lessons Learned and the Path Forward.* Washington, DC: The National Academies Press.

MUNI | RECETOX

PERSPECTIVE

OPEN

doi:10.1038/nature12564

Criteria for the use of omics-based predictors in clinical trials

Lisa M. McShane[1], Margaret M. Cavenagh[1], Tracy G. Lively[1], David A. Eberhard[2], William L. Bigbee[3], P. Mickey Williams[4], Jill P. Mesirov[5], Mei-Yin C. Polley[1], Kelly Y. Kim[1], James V. Tricoli[1], Jeremy M. G. Taylor[6], Deborah J. Shuman[1], Richard M. Simon[1], James H. Doroshow[1] & Barbara A. Conley[1]

The US National Canc
'omics'-based test d
omics-based tests for
mathematical model
encouraged to consid
checklist will be use
therapy.

Clinical Chemistry 60:10
1256–1257 (2014)

Perspective

Where Are All the New Omics-Based Tests?

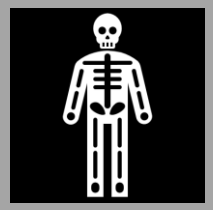
Patrick M. Bossuyt[1*]

"Why?" is the inevitable question. Why have so few biomarkers made it to everyday clinical care? Why, despite billions of dollars worldwide in omics-based research? We have been promised multiple breakthroughs, and numerous biomarker discoveries have been announced, but it is fair to say that, up to this day, clinical medicine has not

issues. A working group then developed a checklist on the basis of the key principles in the IOM report and the results of the NCI workshop *(2)*. A short version appeared in *Nature* last year, and a version with a longer explanation and elaboration was published in *BMC Medicine (3)*.

IOM komise: **Specifika testů založených na omics**

Testy na bázi omics a ve skutečnosti všechny klinické laboratorní testy podléhají **odlišnému regulačnímu rámci** než léky



Absence jasného biologického zdůvodnění na rozdíl od většiny ostatních klinických laboratorních testů založených na jediném analytu

Složitost omicsového výzkumu ztěžuje **sdílení komplexních datových souborů a výpočetních modelů**, což omezuje schopnost ostatních vědců replikovat a ověřovat zjištění a závěry těchto studií

Absence jasného biologického odůvodnění testů omics biomarkerů

Biologické zdůvodnění **testu s jedním analytem** je často zcela zřejmé: Test je užitečný, protože gen, RNA, protein nebo metabolit hraje **pochopitelnou roli** v patologii onemocnění nebo jiném vyšetřovaném biologickém procesu.

Příklady:

Testování karcinomu prsu lidským epidermálním růstovým faktorem 2 (HER2)	Měření hladiny cholesterolu lipoproteinů s nízkou hustotou (LDL) pro hodnocení srdečního rizika

MUNI | RECETOX

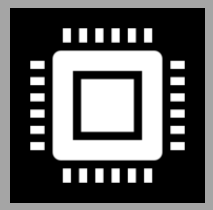
Absence jasného biologického odůvodnění testů omics biomarkerů – **proč je to problém**

Když se nedá test založený na omics 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.

Problém (ne) sdílení komplexních datových souborů a výpočetních modelů

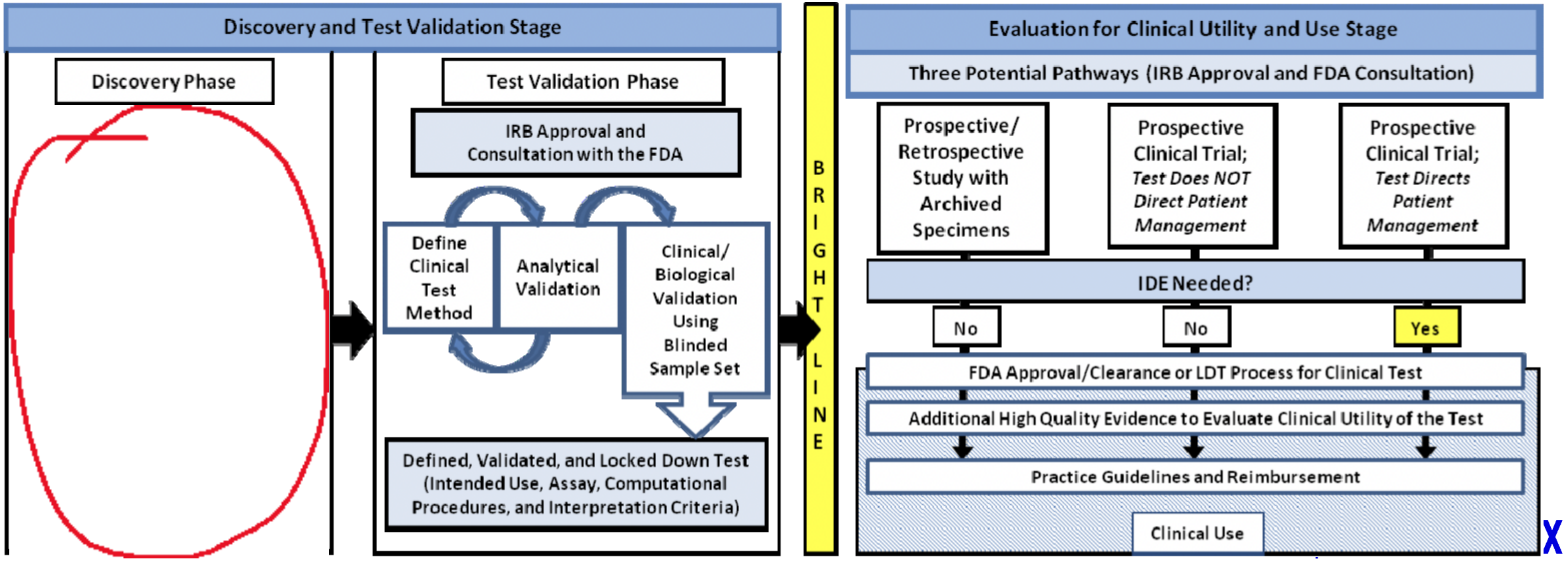
K dispozici jsou databázové úložiště pro soubory omicsových dat, ale sdílení dat není rutinní a bez přístupu k datům a přesně definovanému výpočetnímu modelu je replikace a ověření obtížnější než pro biomarkery založené na jednotlivých analytech.



I když nezávislé validační studie jsou drahé, potřeba replikace v omicsových studiích je nutná vzhledem ke složitosti dat, které mohou vést k chybám (od jednoduchých chyb správy dat až po nesprávně navržené výpočetní modely).

Tato úroveň složitosti neexistuje pro výzkum, vývoj a validaci testů s jedním biomarkerem.

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

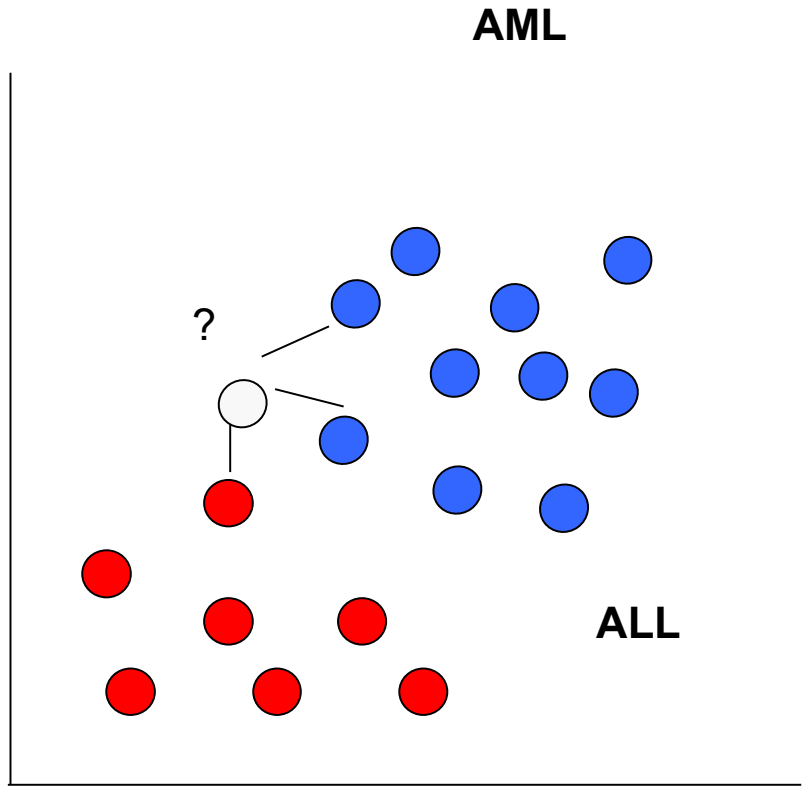
• Například určujeme:

  • 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 uplatnení 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 interpretaci je třeba brát na zřetel, že se jedná pouze o podskupinu genů
- Biologické závěry o molekulárních změnách mezi podskupinami vzorků by měly být založené na studiu celé množiny významných genů

Příklad

ARTICLE

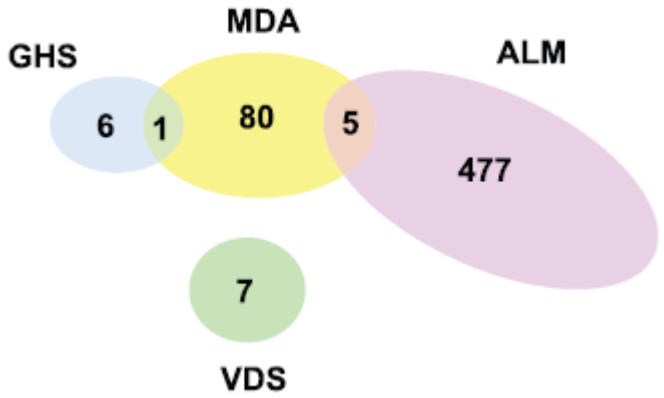
Test of Four Colon Cancer Risk-Scores in Formalin Fixed Paraffin Embedded Microarray Gene Expression Data

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

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**

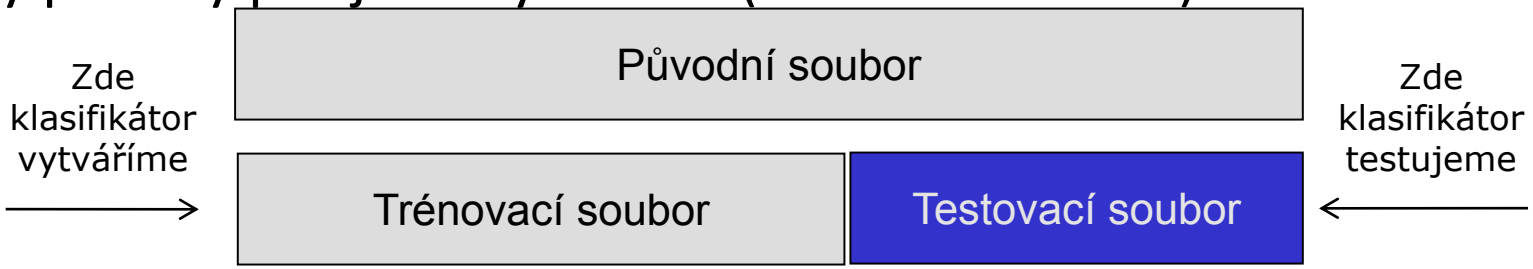
- Proč odhadovat výkonnost klasifikátoru?
  - Omezení trénovacím souborem
  - Bez předpokladu o rozložení neexistuje žádný vzorec pro výpočet velikosti vzorku
  - Často existuje jen jeden datový soubor pro trénování a testování klasifikátoru

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

Odhad výkonnosti klasifikátoru II – křížová validace

Základní myšlenka:

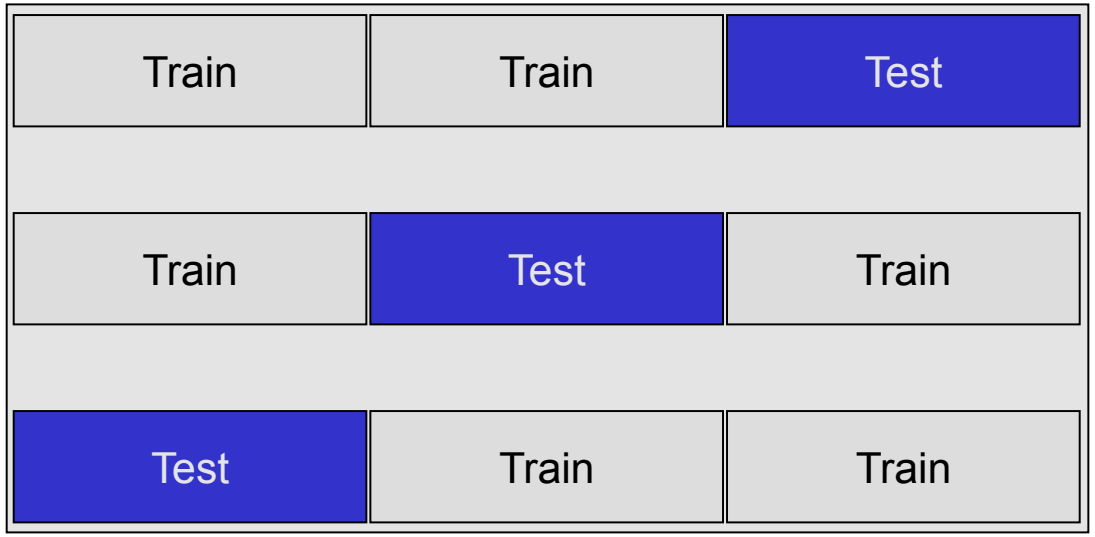
- Převzorkováním rozdělit (opakovaně) datový soubor na trénovací a testovací, vytv
klasifikátor na trénovacím souboru a změřit výkonnost klasifikátoru jen na datech, které nebyly použity pro jeho vytvoření (**křížová validace**)



Odhad výkonnosti klasifikátoru II – křížová validace

- **k-fold cross validation**

k=3



Alternativně: LODO (leave one dataset out), monte-carlo CV…

Odhad výkonnosti klasifikátoru - bootstrapping

- Vytvoření nového souboru vzorkováním s opakováním (rozdíl od křížové validace, kde je vzorek vždy jen jendou)
- Trénování klasifikátoru probíhá na nových datech a testuje se na vynechaných vzorcích
- Opakuje se B-krát
- Odhad chyby pomocí 0.632 pravidla

$$\bar{E} = 0.368E_0 + 0.632\frac{1}{B}\sum_{b=1}^{B}\hat{E}_b$$

Kde E0 je chyba na celém (původním) trénovacím souboru

Odhad výkonnosti klasifikátoru III – důležité!!!

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

Odhad výkonnosti – proč převzorkování nestačí

- Každé dva trénovací soubory vytvořené z původního datového souboru s pomocí převzorkování se do jisté míry překrývají -> vytvořené klasifikátory 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

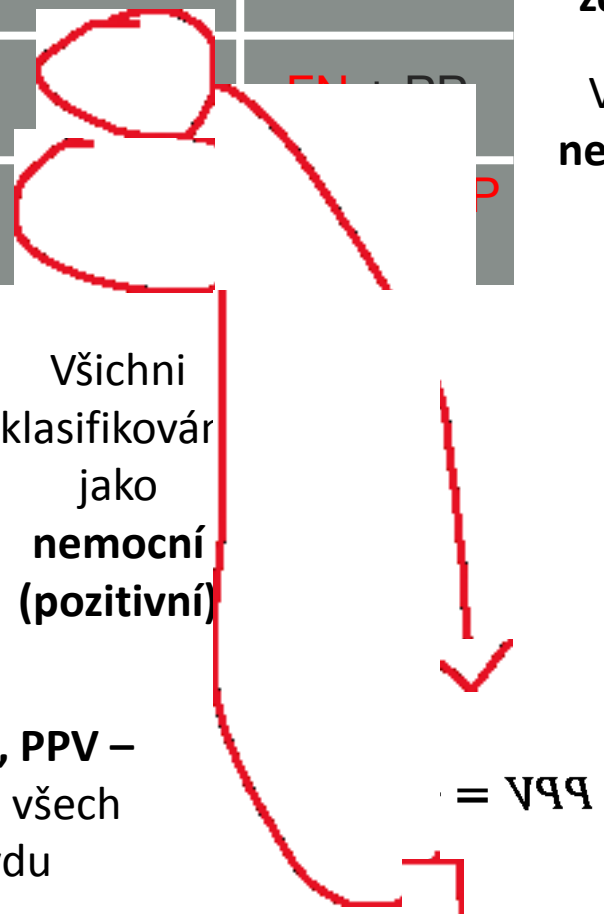
		Klasifikace		
		Zdravý (-)	Nemocný (+)	Celkem
Skutečnost	Zdravý (-)	PN	FP	PN + FP
	Nemocný (+)	FN		
	Celkem	PN + FN		

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)**

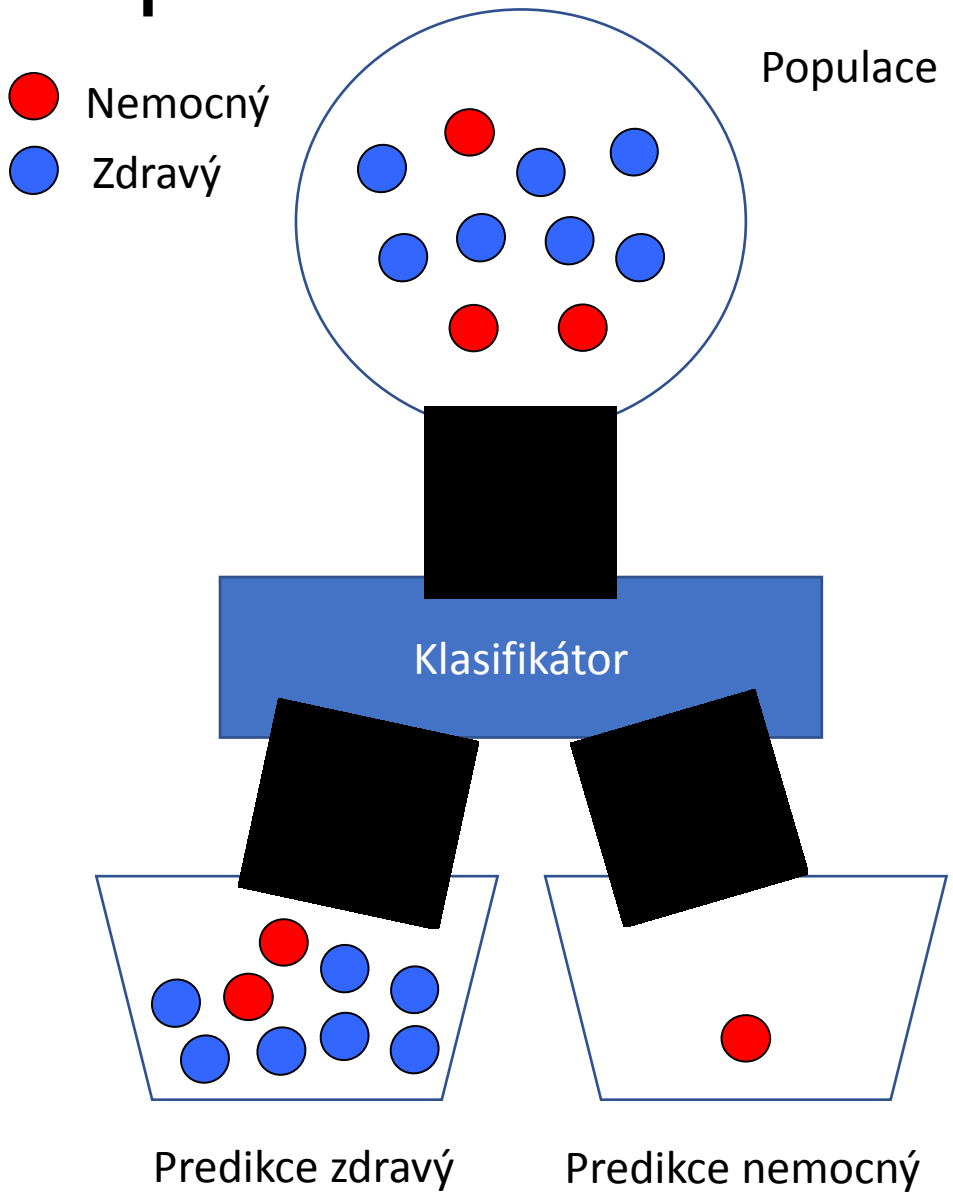
Všichni klasifikov jako **zdra (negativ**

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

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

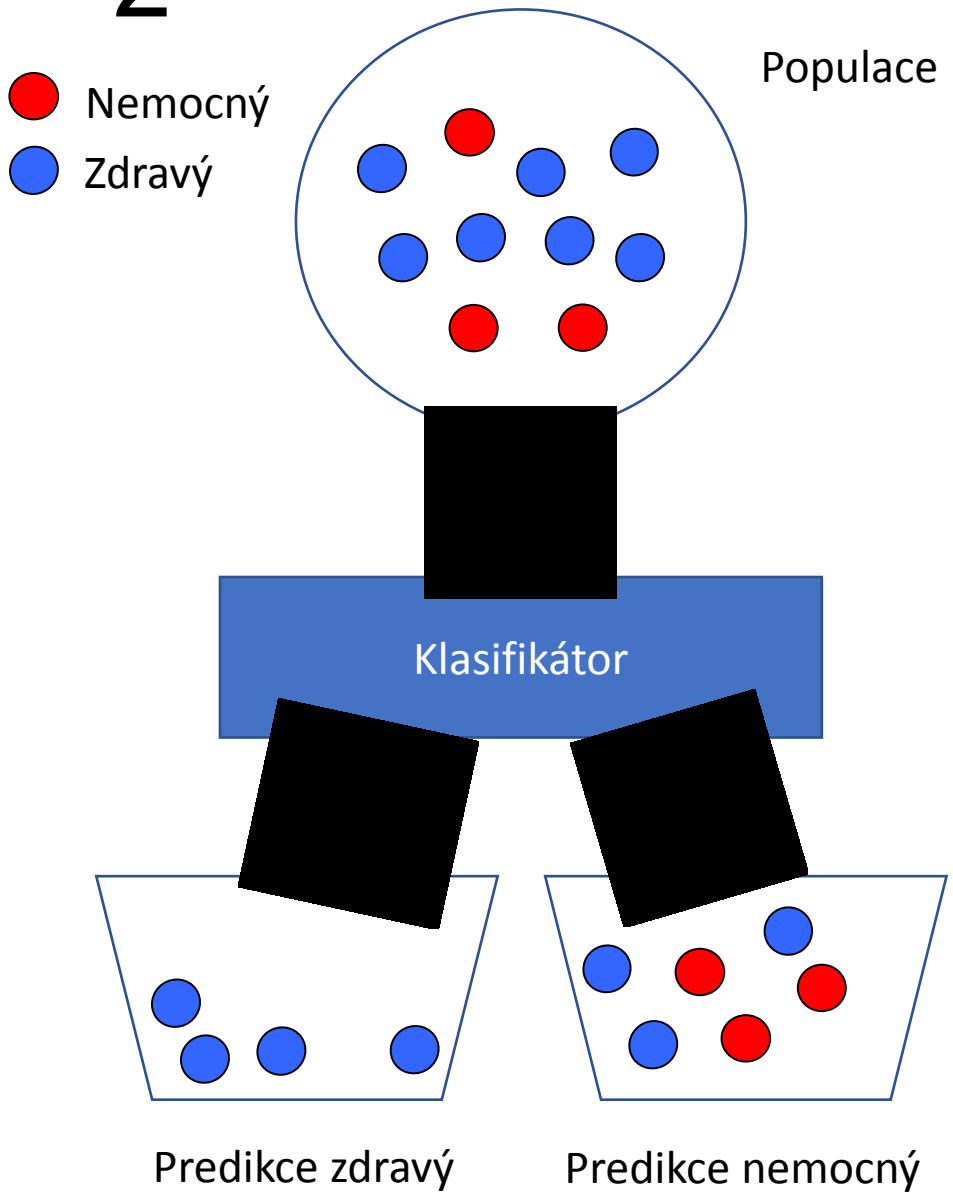
$$\text{FPR} = \frac{FP}{PN + FP} = \frac{0}{7} = 0\%$$

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

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

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

Vyhodnocení přesnosti klasifikátoru – příklad 2



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

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

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

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

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

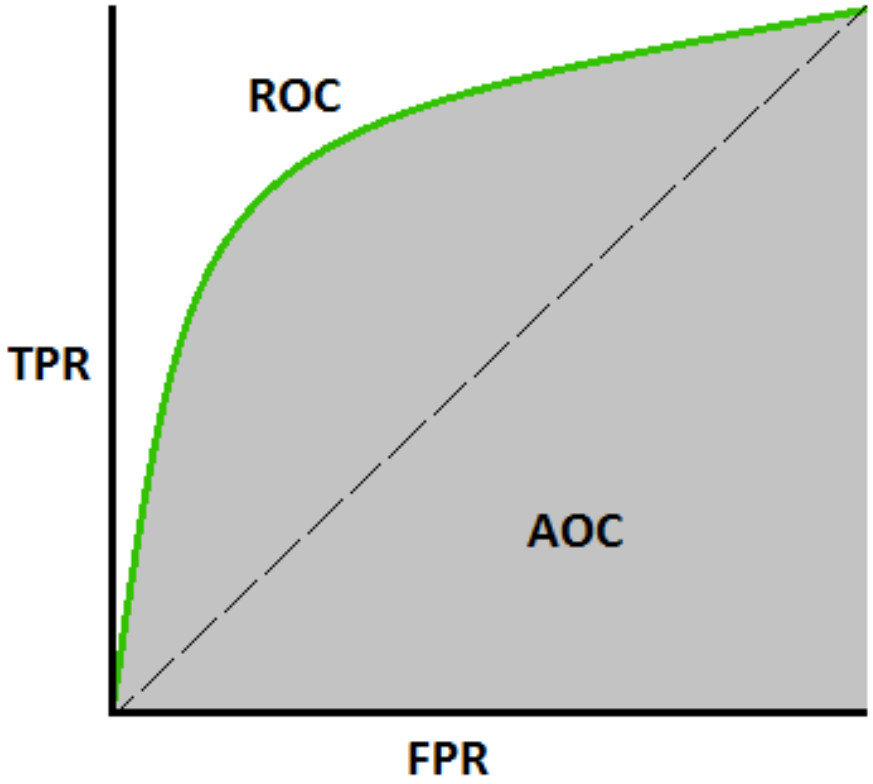
$$FPR = \frac{FP}{PN + FP} = \frac{3}{7} = 43\%$$

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

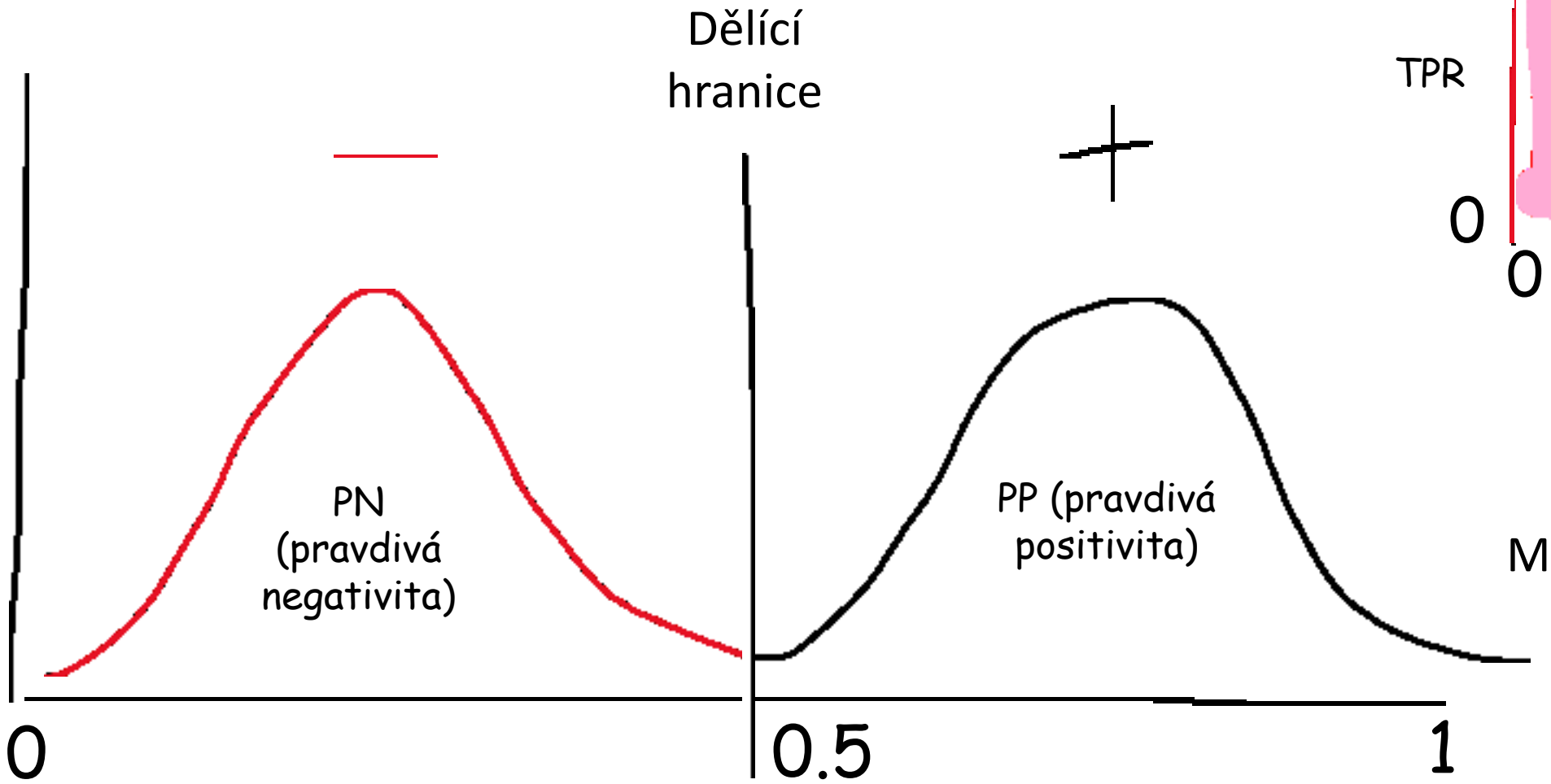
ROC křivka

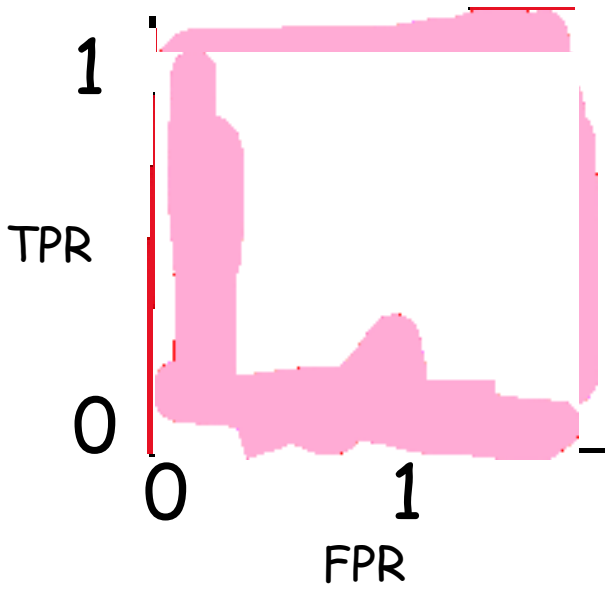
- Receiver operator characteristics (ROC)



- 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



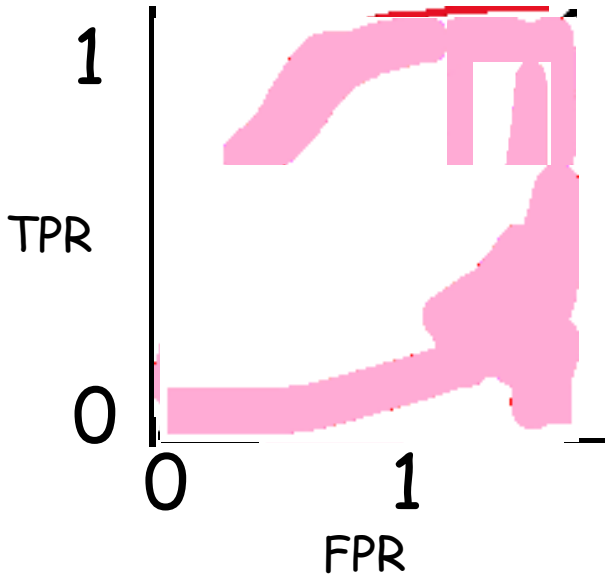


AUC = 1

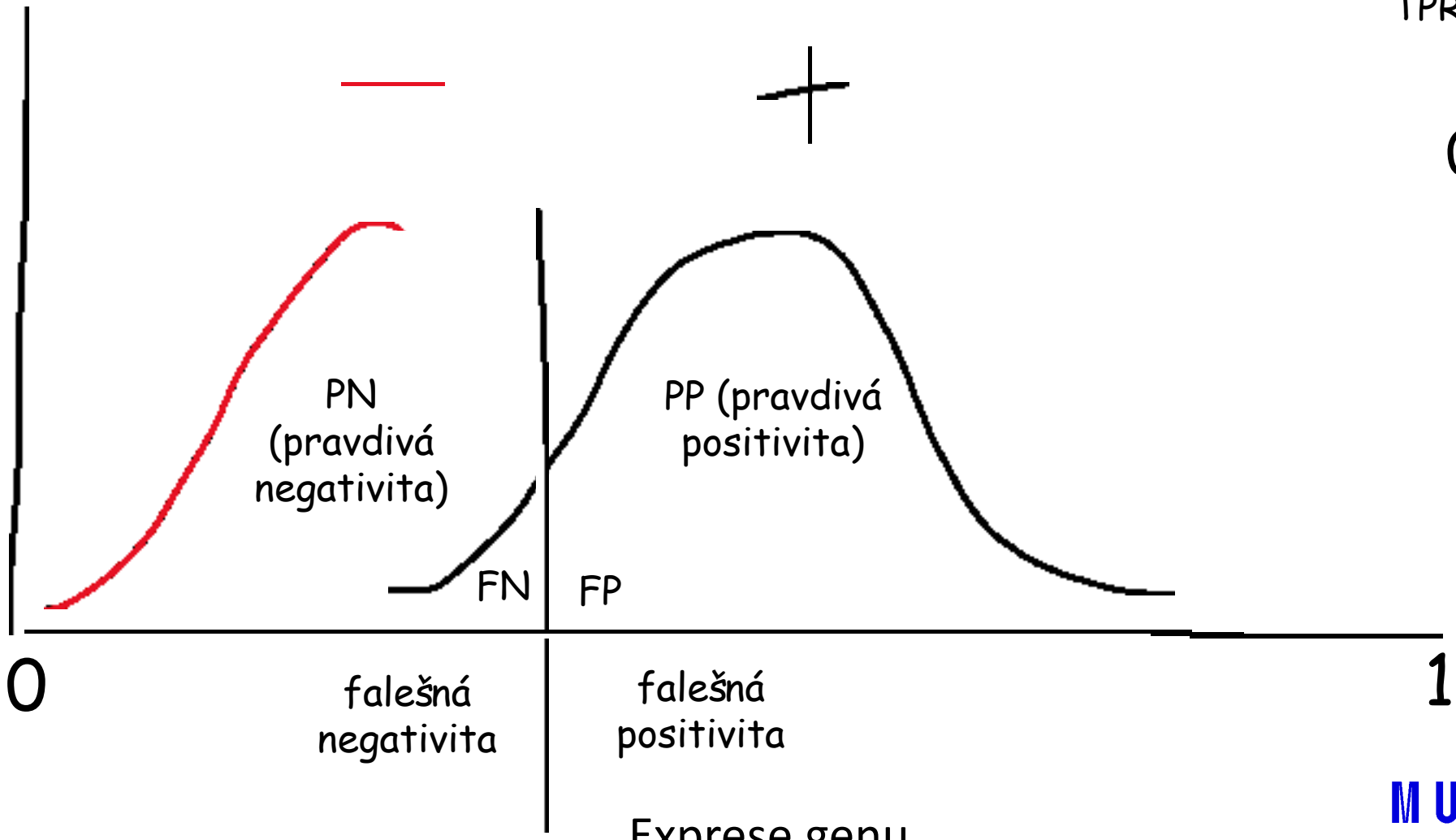
Model perfektně diskriminuje

MUNI | RECETOX

ROC křivka

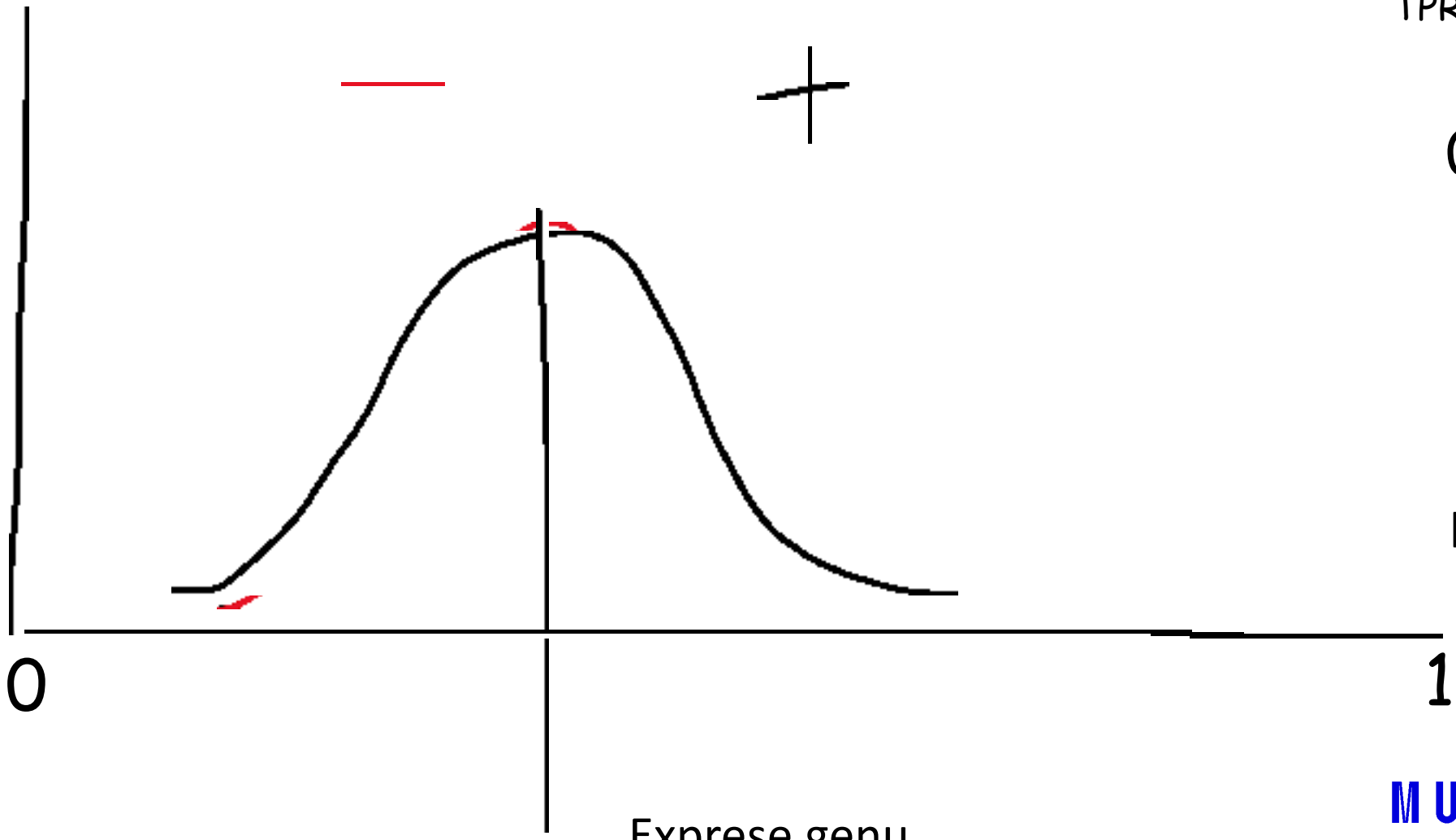


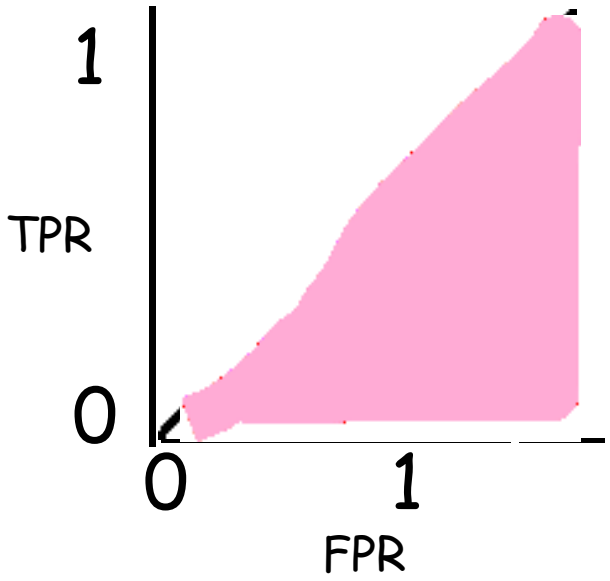
AUC = 0.8



MUNI | RECETOX

ROC křivka



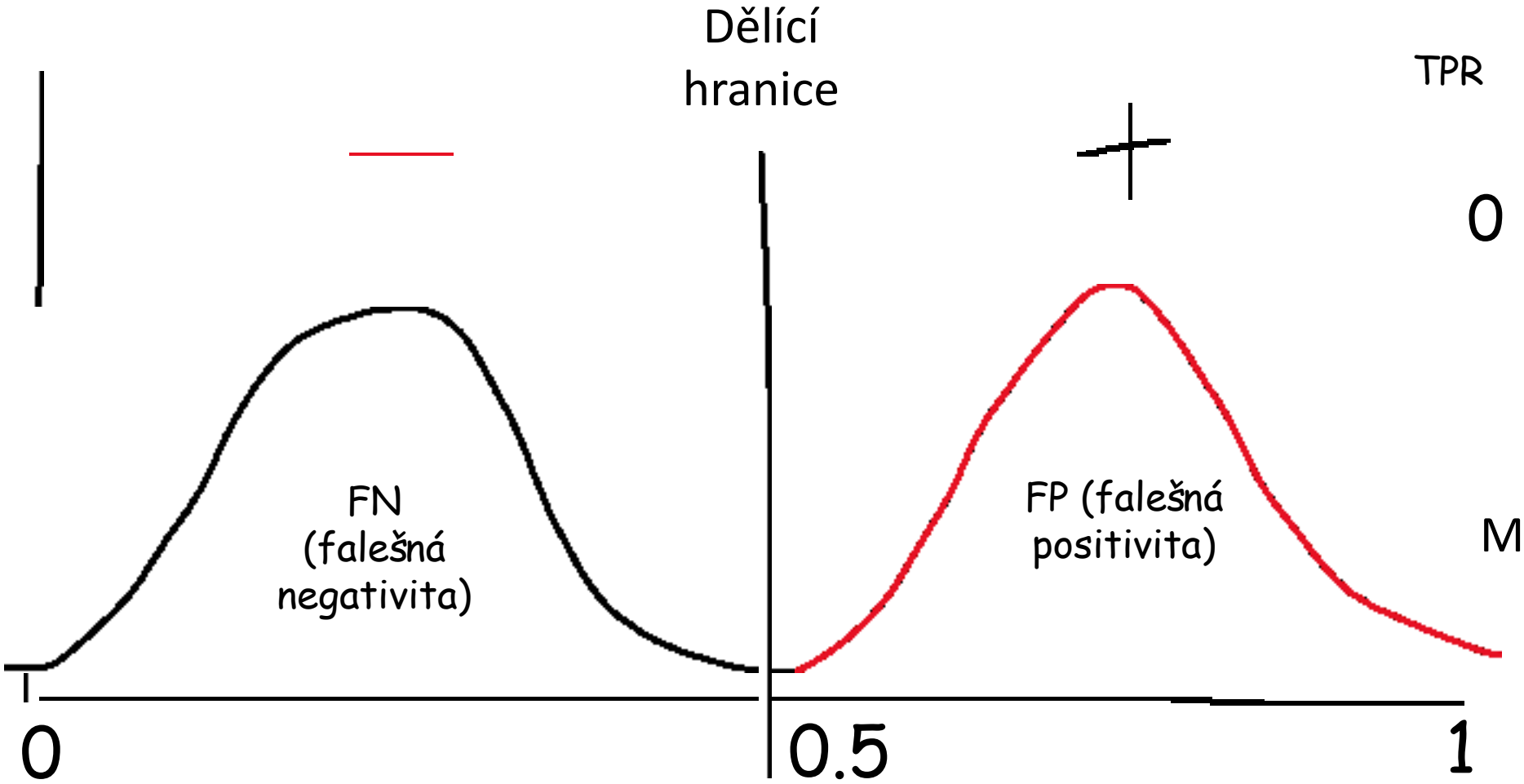


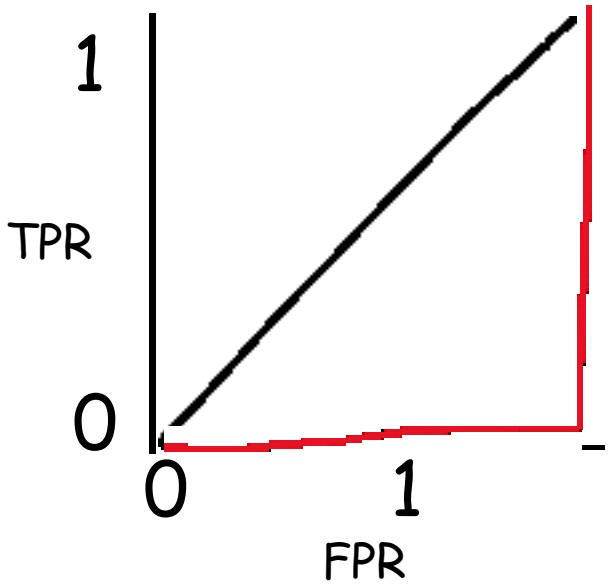
$AUC = 0.5$

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

Standardy pro mikročipy

nature biotechnology

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

MAQC Consortium*

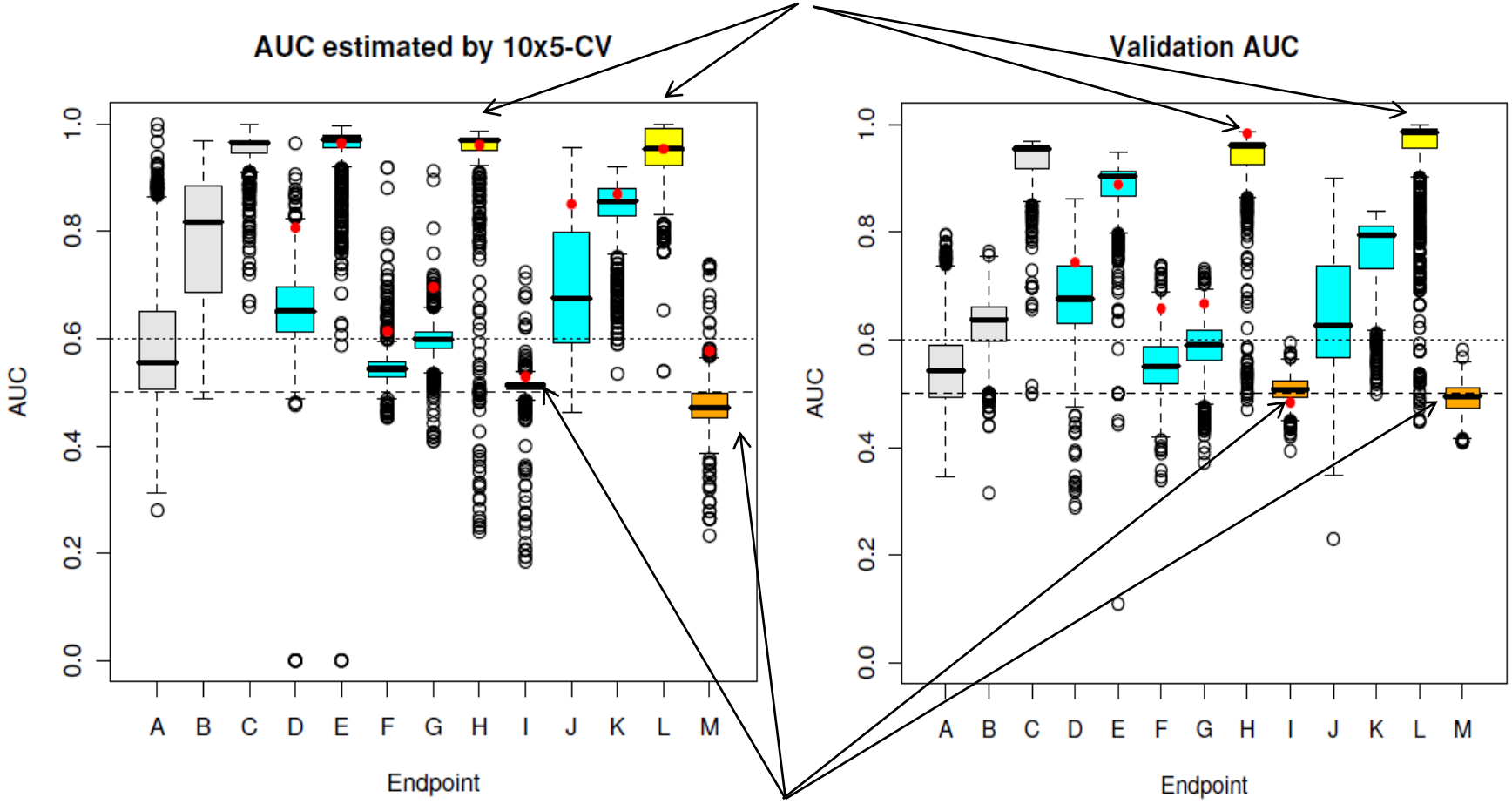
- **36** nezávislých tímů analytiků z celého světa analyzovalo **6** mikročipových studií a vytvořilo klasifikátory pro predikci **13** endpointů (ER+ vs ER-, ...)
- Každý tým navrhl plán tvorby a validace klasifikátoru
- Tyto plány byly předem posouzeny odbornými statistiky a ohodnoceny dle jejich názoru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

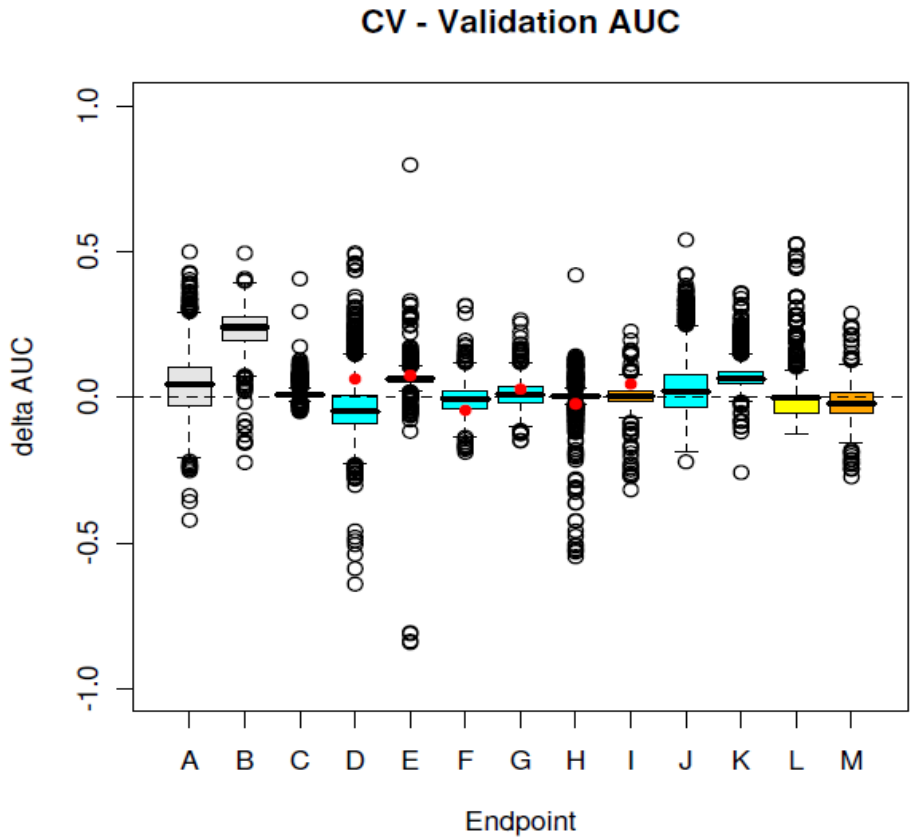
Výkonnost klasifikátorů dle experimentu

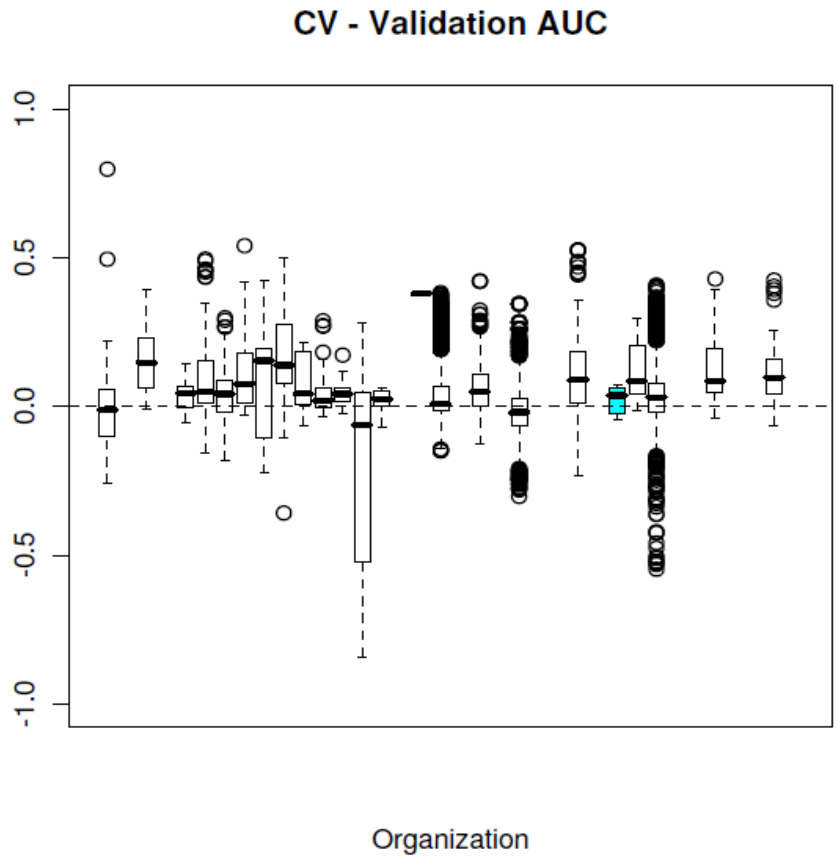
Úspěšnost odhadu pohlaví, pozitivní kontrola



Úspěšnost predikce náhodného zařazení, negativní kontrola

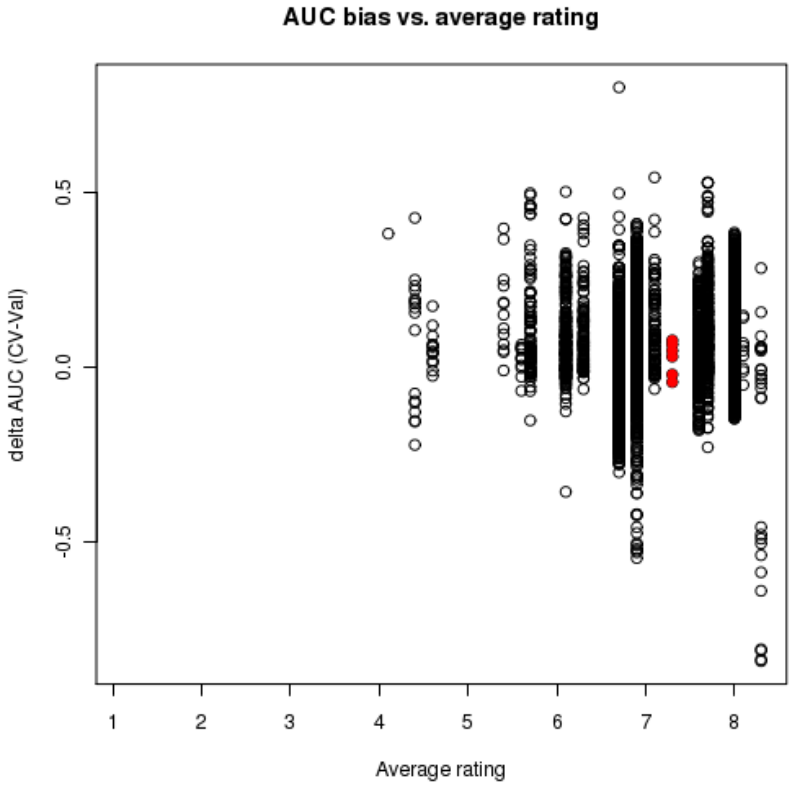
Rozdíl výkonnosti odhadnuté na základě krosvalidace (CV) a na validačním souboru (Validation)





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

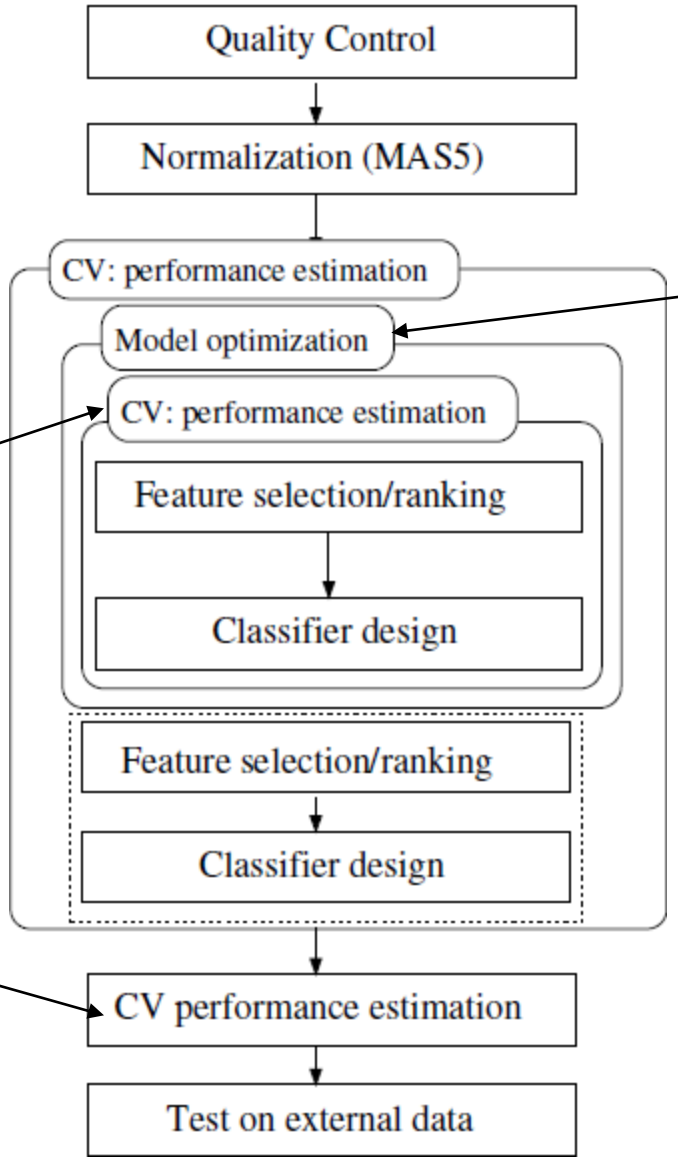
Aby to nebylo jednoduché...



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**

Jeden z navržených a úspěšných algoritmů validace



Na k-1 fold datech výběr modelu (typ nebo parametrů) – opakuje se pro každý model (sadu parametrů modelu)

Nejlepší model bude jiný v každé externí CV

Znovu se opakuje interní krosvalidace na novém rozdělení k-fold (zde je k jiné číslo než v externí CV)

Statistika performance vybraných modelů v k-fold externí CV

MUNI | RECETOX

Bez validace není (dobrá) publikace

Finální validace

- **Vždy na nezávislém datovém souboru**
- Velmi důležitá pro otestování skutečné robustnosti klasifikátoru
- Absolutně nevyhnutné v medicíně

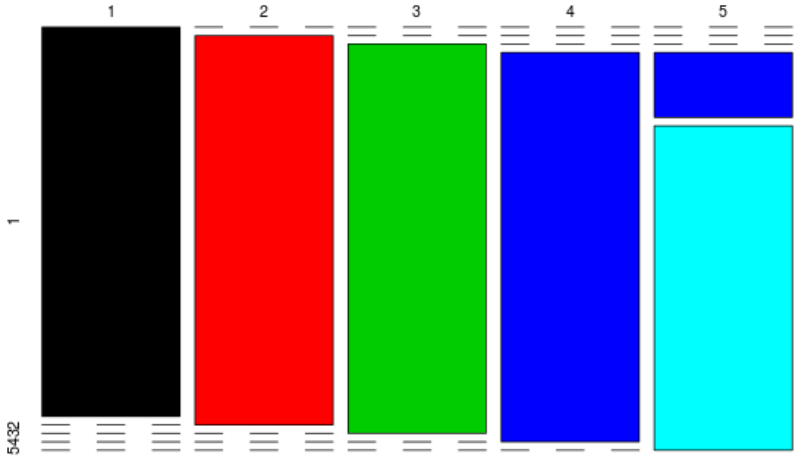
- Testovací soubor by měl splňovat následující vlastnosti:
  - Musí obsahovat parametry použité v klasifikátoru
  - Musí být známá příslušnost vzorků ke skupinám, které se klasifikátor snaží diskriminovat
  - Podobná struktura s ohledem na klinické a patologické parametry (např. stejné rozložení věku, zastoupení pohlaví apod.)

Design experimentu je důležitý!

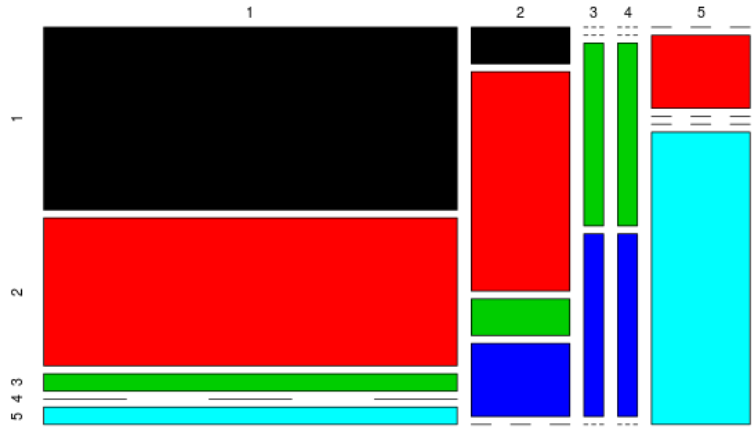
- **Myslete na dostatečně velký trénovací i testovací datový soubor!**

Příklad: 5 podtypů karcinomu prsu – 96 vzorků (N1=48, N2=16, N3=8, N4=8, N5=16)

Trénovací soubor



Testovací soubor



Málo vzorků ve skupině, nemožnost tuningu, malá variabilita ->
přetrénování => nefunguje na testovacím souboru.
Stačí jeden špatně klasifikovaný vzorek a výrazně se sníží výkonnost!

- **Datové soubory musí reprezentovat populaci, na které budete klasifikátor používat**

MUNI | RECETOX

Další 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