

# Biostatistika pro matematickou biologii

Tomáš Pavlík, Ladislav Dušek, Michal Uher

[pavlik@iba.muni.cz](mailto:pavlik@iba.muni.cz)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



# Přínos kurzu

- ➡ Orientace v principech biostatistiky, plánování a hodnocení experimentů.
- ➡ Orientace v biostatistických metodách, jejich výpočetní podstatě a jejich předpokladech.
- ➡ Schopnost aplikace metod při řešení reálných problémů z oblasti biologie a medicíny a interpretace výsledků.

**Schopnost statistického uvažování nad reálným problémem**

- ➡ Kurz slouží jako příprava pro pokročilejší přednášky statistiky a aplikované analýzy dat.
- ➡ Biostatistika v matematické biologii je předmět na pomezí základní biostatistiky a kurzu pravděpodobnosti a statistiky.

# Požadavky ke zkoušce

1. Zkouška bude vycházet z přednášek + skript
2. Zkouška bude písemná (60 bodů) + ústní (10 bodů)
3. V průběhu semestru budou 2 krátké testy (každý 15 bodů)

# Literatura

1. Přednášky
2. Skripta
3. Web

Česky:

- ➡ Zvárová J (2001) *Základy statistiky pro biomedicínské obory*, Karolinum, Praha.
- ➡ Zvára K (2006) *Biostatistika*, Karolinum, Praha.

Anglicky:

- ➡ Altman DG (1991) *Practical statistics for medical research*, Chapman&Hall/CRC, London.
- ➡ Zar JH (1999) *Biostatistical analysis*, Prentice-Hall, New Jersey.

# Přednáška I.

## Úvod do biostatistiky

- ➔ Motivační příklady
- ➔ Co je biostatistika a čím se zabývá
- ➔ Klíčové principy biostatistiky

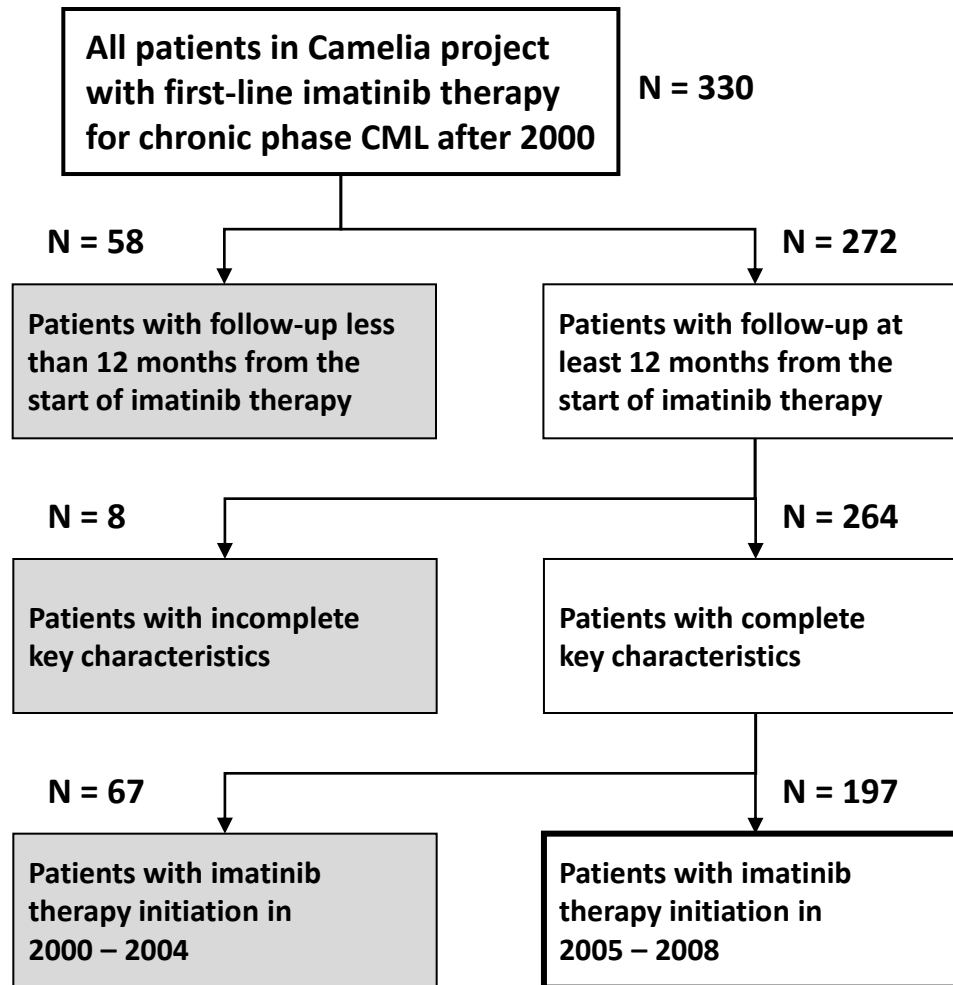


# 1. Příklady použití biostatistiky

# Př. 1 Project CAMELIA – Regression model for cytogenetic or molecular response in patients with chronic myeloid leukemia

- ➡ The aim of this work is to present a Cox regression model for the achievement of the complete cytogenetic or molecular response to a modern targeted therapy in patients in chronic phase of chronic myeloid leukemia (CML). The model is based on data coming from a population study involving approximately half of Czech and all Slovak CML patients treated since 2000.
- ➡ The primary objective of this study was to identify characteristics of CML patients associated with prolonged time to complete cytogenetic response (CCgR) or major molecular response (MMR) to imatinib therapy, which could further indicate the increased risk of disease progression.

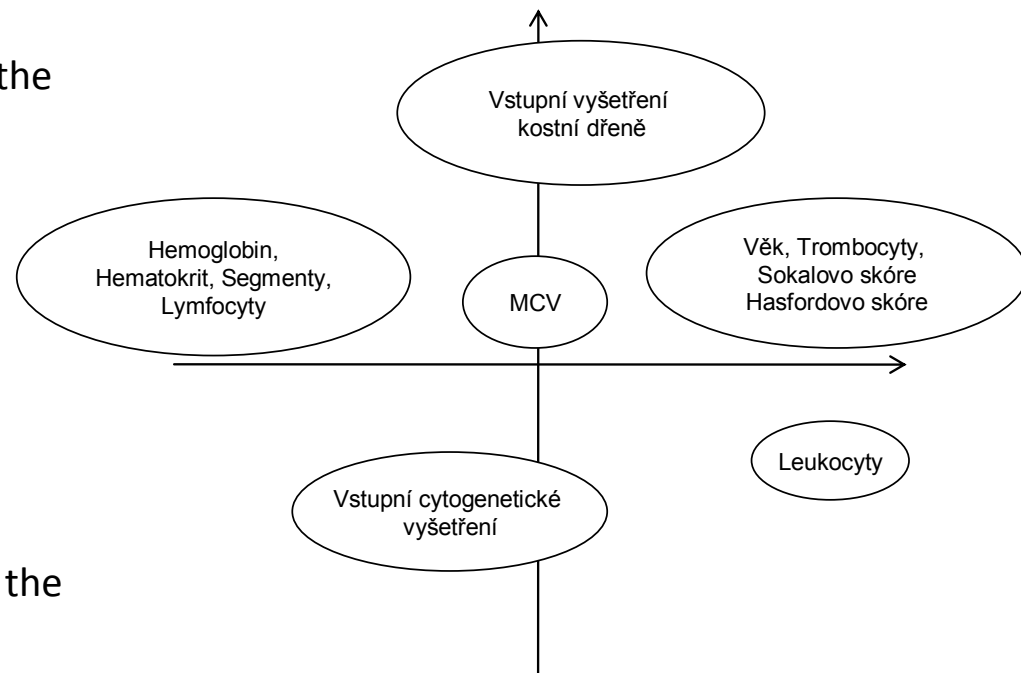
# Patients included in the analysis – summary





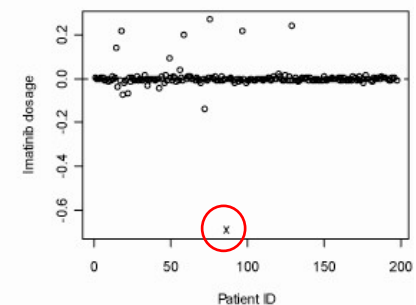
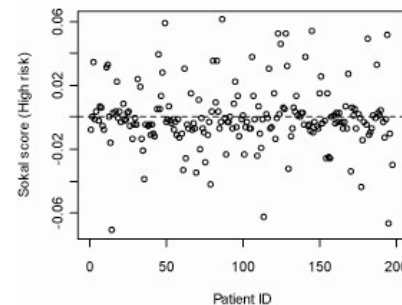
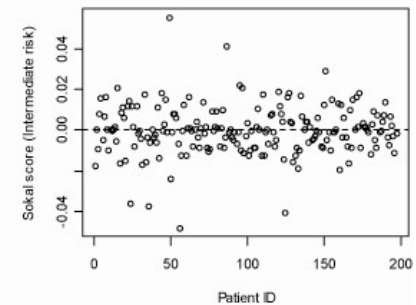
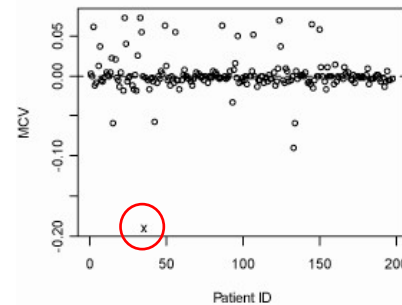
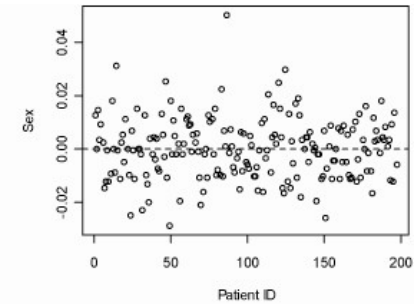
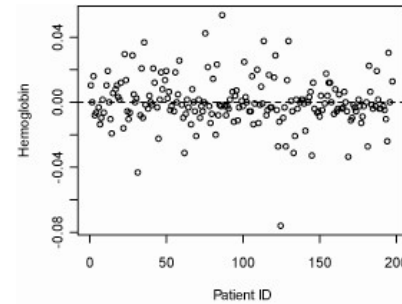
# Modelling the endpoint – primary variable selection

- ➡ The continuous explanatory variables were analysed using a cluster analysis and principal component analysis to identify highly correlated prognostic factors.
- ➡ Four distinct clusters and two separate clinical variables identified with the multivariate techniques – only one member from each of the identified groups of prognostic factors was used as a covariate.
- ➡ In addition, following categorical variables were also considered for the modelling:
  - ➡ Patient's sex
  - ➡ Imatinib dosage
  - ➡ Clonal chromosomal abnormalities in the Ph+ cells
  - ➡ Clonal chromosomal abnormalities in the Ph- cells
- ➡ Clinical centre was incorporated to the model as a random effect.



# Modelling the endpoint – regression diagnostics

- ➡ Regression diagnostic was performed to find out whether the model adequately describes the data.
- ➡ Highly influential observations (outliers) were subsequently filtered out.
- ➡ Finally, N=5 outliers were filtered out with N=192 considered in the final model.

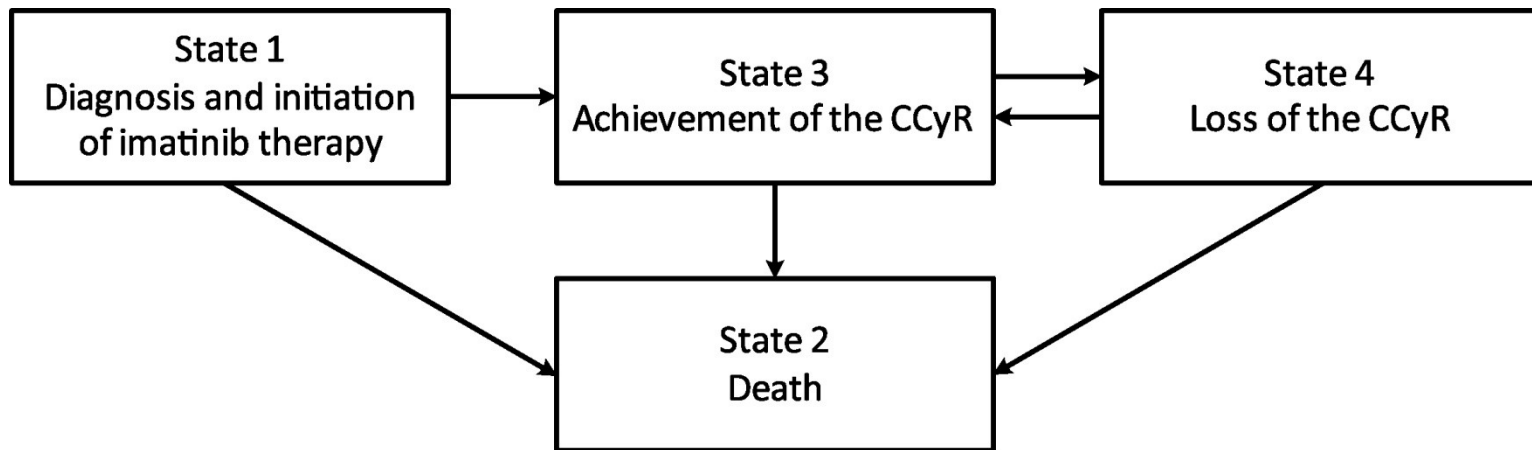


# Modelling the endpoint – the final model

- ➡ Hazard ratios identified with the final model according to the achievement of cytogenetic or molecular response to imatinib therapy in chronic CML patients treated with imatinib in first-line after 2004
- ➡ N=192

Risk factor	Risk category / Basal category	Hazard ratio	95% CI	p-value
Sex	Male / Female	1.88	1.33–2.66	<0.001
Hemoglobin	Hb < 110 g/l / Hb 110 g/l and more	1.89	1.23–2.87	0.004
Sokal score	Medium risk / Low risk	1.34	0.93–1.93	0.120
Sokal score	High risk / Low risk	2.43	1.45–4.08	<0.001
Clinical centre*	-	-	-	<0.001

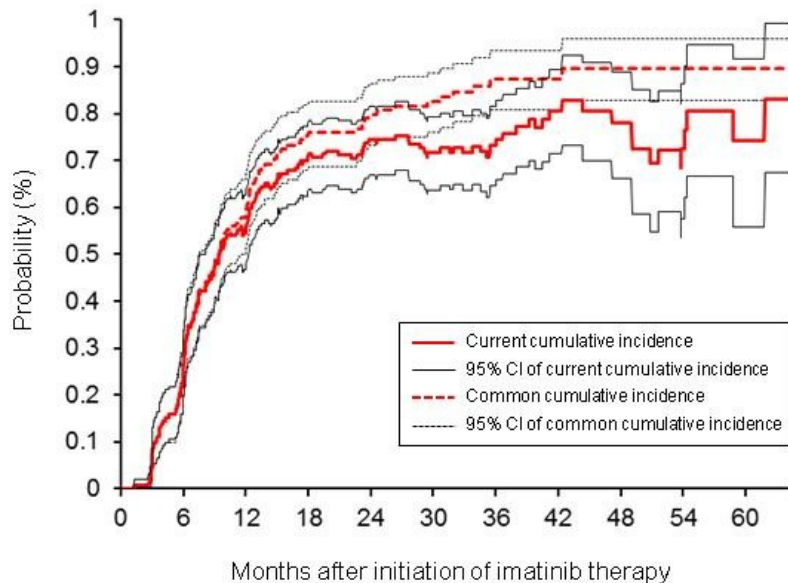
## Př. 2 Multi-state model of CML therapy



- Obviously, all living patients can move from the CCyR (state 3) to the cytogenetic relapse (state 4) and vice versa repeatedly.

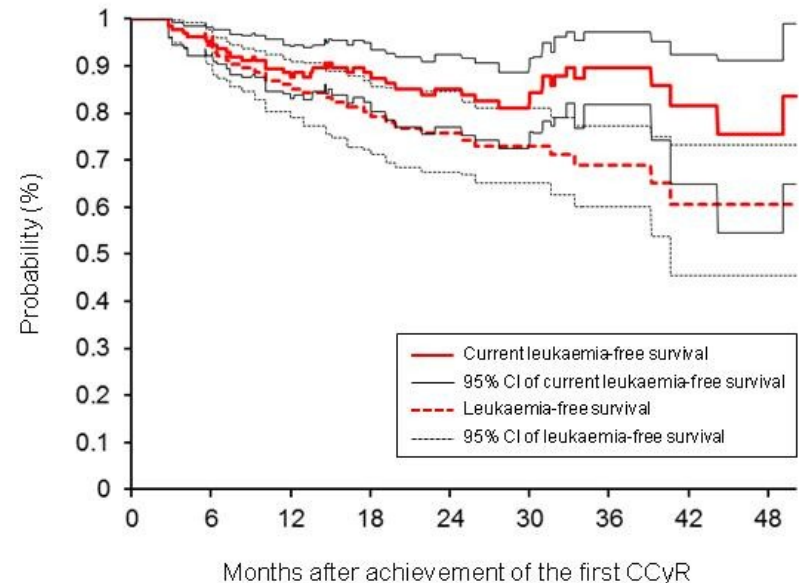
# Current cumulative incidence (CCI) and current leukaemia-free survival (CLFS)

$CCI(t)$



The common cumulative incidence curve **overestimates** the probability of being alive and in remission after the initiation of the imatinib therapy.

$CLFS(t)$



The common leukaemia-free survival **underestimates** the probability of being alive and in remission after the achievement of first remission on the imatinib therapy.

# Př. 3 Je použití inzulínového analoga u diabetiků bezpečné?

Diabetologia

DOI 10.1007/s00125-009-1418-4

ARTICLE

## Risk of malignancies in patients with diabetes treated with human insulin or insulin analogues: a cohort study

L. G. Hemkens • U. Grouven • R. Bender • C. Günster •  
S. Gutschmidt • G. W. Selke • P. T. Sawicki

Received: 29 August 2008 / Accepted: 26 May 2009

© The Author(s) 2009. This article is published with open access at Springerlink.com

- ➡ Hemkens a kol. (2009) publikovali vyšší riziko vzniku zhoubného nádoru při užívání inzulínu glargin při srovnání s adekvátní dávkou humánního inzulínu.

# Co tato studie znamená ze statistického hlediska?

Jedná se o

1. observační studii
2. studii s „pokusem“ o adjustaci na dávkování inzulínu
3. studii s velmi krátkou délkou sledování pacientů ve skupině s vysokou dávkou inzulínu glargin (v průměru 7,3 měsíců)
4. studii s vyloučením pacientů s kombinovanou terapií

# 1. Observační studie má své výhody...

Randomizovanou studii někdy nelze v klinické praxi provést.

Hlavními důvody mohou být

- ➡ etické hledisko
- ➡ randomizaci nelze použít
- ➡ raritní výskyt sledovaného onemocnění

V těchto případech má observační studie své opodstatnění,

**ALE!**



# ... a jednu velkou nevýhodu!

Observační studie nemůže zaručit stejné zastoupení rizikových faktorů v jednotlivých sledovaných skupinách!

I při použití adjustačních metod mohou být výsledky ovlivněny nenáhodným rozdělením pacientů do jednotlivých skupin.

Použití výsledků observačních studií pro vytváření klinických doporučení tak může být nekorektní, ...

... což je i případ studie Hemkense a kol.

## 2. Adjustace na dávkování inzulínu

Adjustace na dávkování použitá v německé studii neodpovídá statistickým standardům.

- ➡ Je nepřijatelné adjustovat statistický model na informaci, která je získána až v průběhu sledování.
- ➡ Adjustace na dávkování musí být provedena s pomocí časově proměnného faktoru, ne s použitím průměrné hodnoty.

**Coxův model nebyl v německé studii použit správně!**

### 3. Krátká délka sledování pacientů

Může být vůbec u pacientů sledovaných necelý rok označeno použití inzulínu jako příčina vývoje nádorového onemocnění?

**Vždy je třeba důkladně rozlišit příčinu a důsledek!**

# 4. Vyloučení pacientů s kombinovanou terapií

Vyloučení pacientů s kombinovanou terapií je ze statistického hlediska umělý krok, který může vést ke zkreslení výsledků.

- ➡ Nelze úplně vyloučit pacienty ze studie na základě informace, kterou opět získáme až v průběhu sledování.
- ➡ Doba sledování pacientů s kombinovanou léčbou měla být zahrnuta do analýzy.

**Autoři se dopustili umělé a nekorektní selekce pacientů!**

# Závěr

Studie Hemkens a kol. (2009) je ze statistického hlediska nekorektní a její výsledky jsou neinterpretovatelné.

Lze jednoznačně souhlasit s tvrzením:

*“There is no evidence of an overall increase in the rate of cancer development in patients on insulin glargine”.*

# Další příklady použití biostatistiky

- ➡ Modelování demografické struktury obyvatelstva
- ➡ Hodnocení úspěšnosti screeningových programů v onkologii
- ➡ Identifikace vlivu genetických a vnějších rizikových faktorů na vznik různých onemocnění – astma, diabetes, hypertenze
- ➡ Identifikace podskupin pacientů s leukémií na základě genetických dat
  
- ➡ Prostorové modelování koncentrací PAH, PCB, DDX a HCB v půdě
- ➡ Prediktivní modelování potencionálního rozšíření biologických společenstev
- ➡ Definice indikačních taxonů a jejich vztah k parametrům prostředí
- ➡ Analýza vztahu dávka - odpověď mezi koncentrací toxické látky, např. pesticidu a reakcí biologických receptorů

## 2. O čem ta biostatistika vlastně je?

*„Statistics is the art and science of making decisions in the face of uncertainty.*

*Biostatistics is statistics as applied to the life and health sciences.“*

Abdelmonem A. Afifi

# Biostatistika

- ➡ Biostatistika je **aplikace statistických metod** v řešení biologických a klinických problémů.
- ➡ Snahou je **získat z pozorovaných dat užitečnou informaci**. V popředí zájmu je pozorovaná variabilita mezi studovanými subjekty, kterou chceme vysvětlit.
- ➡ Je **orientována na konkrétní problém**, ne na teoretické aspekty. To však neznamená, že lze statistické metody používat bezhlavě.

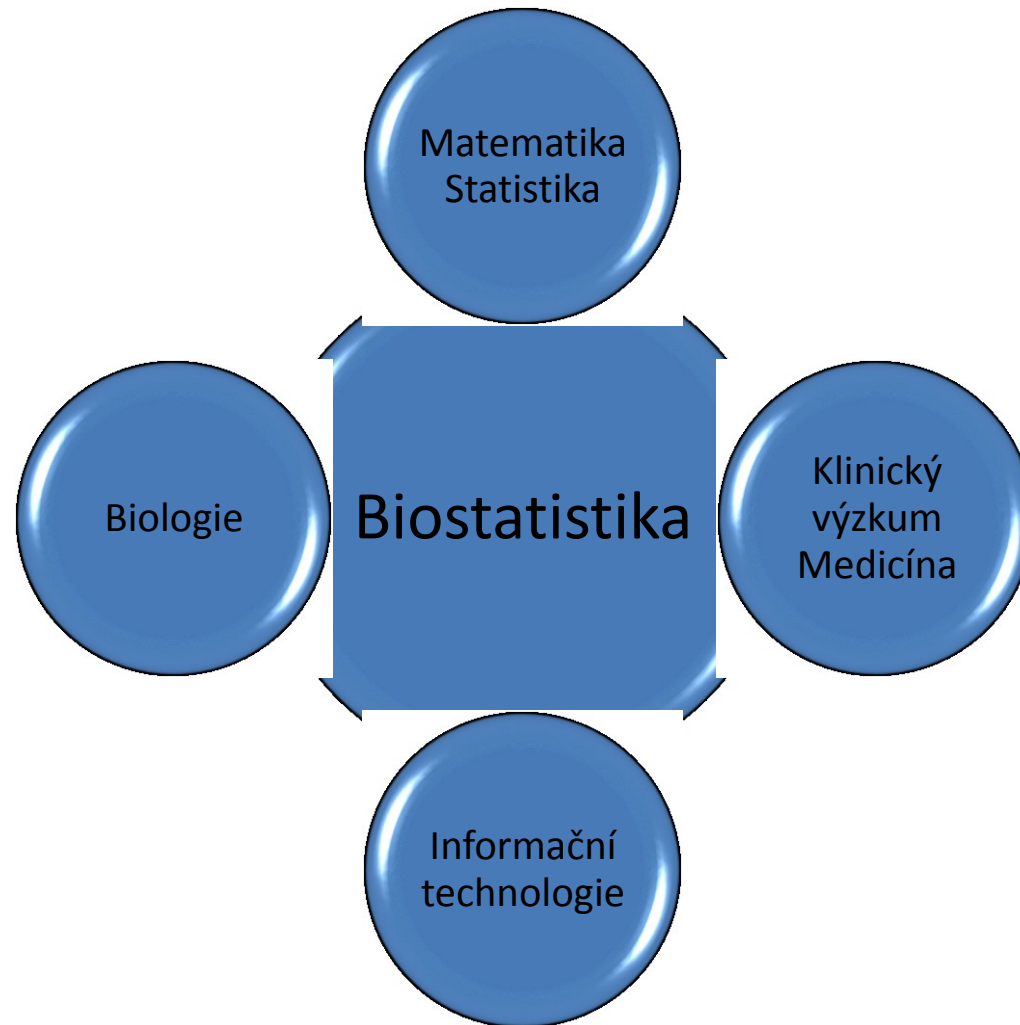


# Význam biostatistiky

11 nejdůležitějších událostí medicíny v minulém tisíciletí (NEJM, 2001):

- ➡ Elucidation of human anatomy and physiology
- ➡ Discovery of cells and their substructures
- ➡ Elucidation of the chemistry of life
- ➡ **Application of statistics to medicine**
- ➡ Development of anesthesia
- ➡ Discovery of the relation of microbes to disease
- ➡ Elucidation of inheritance and genetics
- ➡ Knowledge of the immune system
- ➡ Development of body imaging
- ➡ Discovery of antimicrobial agents
- ➡ Development of molecular pharmacotherapy

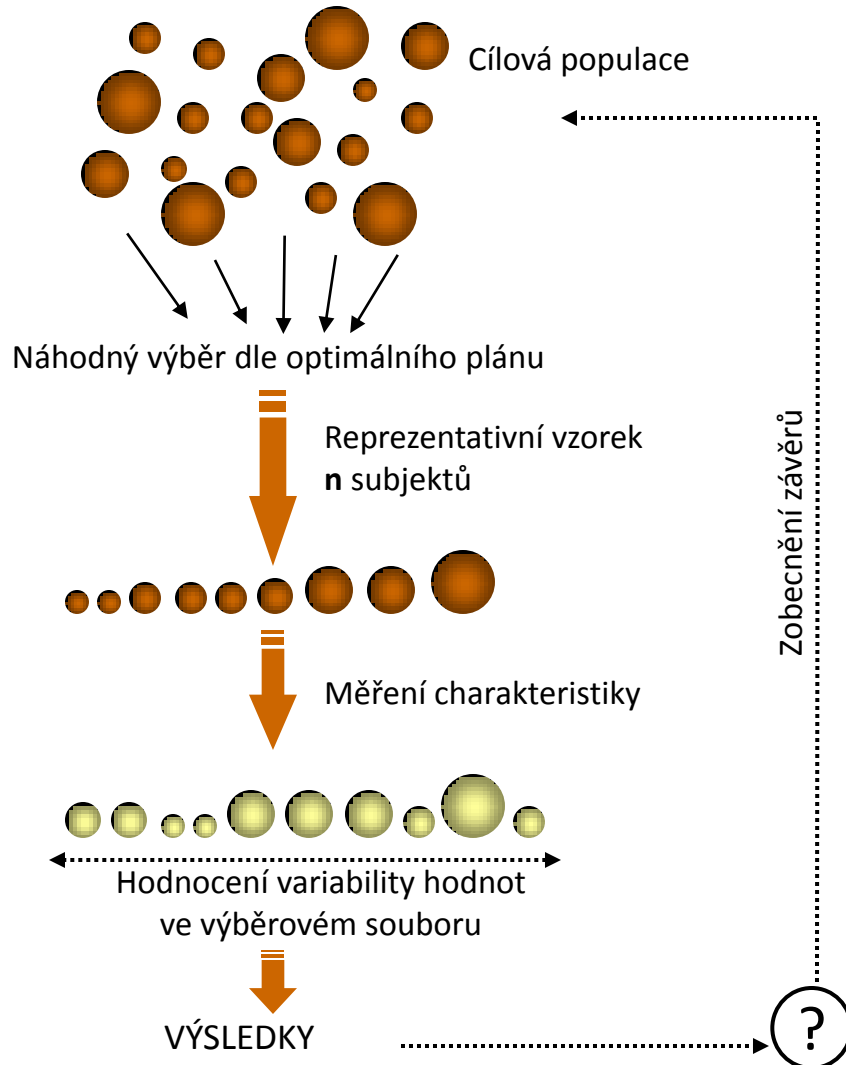
# Biostatistika souvisí s dalšími vědami



# Jaké úlohy můžeme řešit?

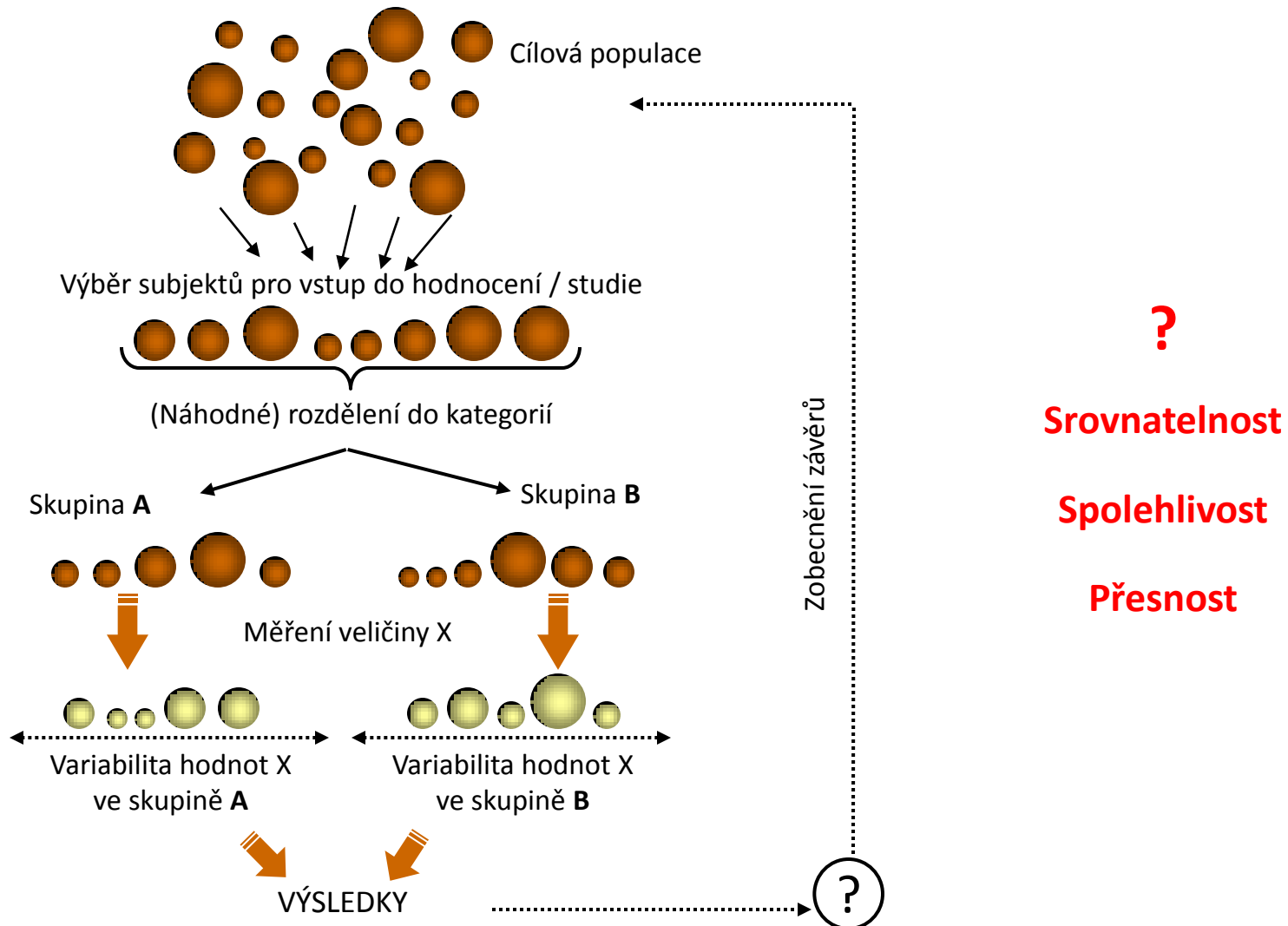
- ➔ **Popis cílové populace** – odhady charakteristik cílové populace
- ➔ **Srovnání skupin** – testování hypotéz
- ➔ **Regresní analýza** – stochastické modelování pro vysvětlení variability
- ➔ **Predikce a klasifikace** – stochastické modelování a klasifikační algoritmy pro předpovídání neznámých hodnot

# Popis cílové populace – popis pozorované variability

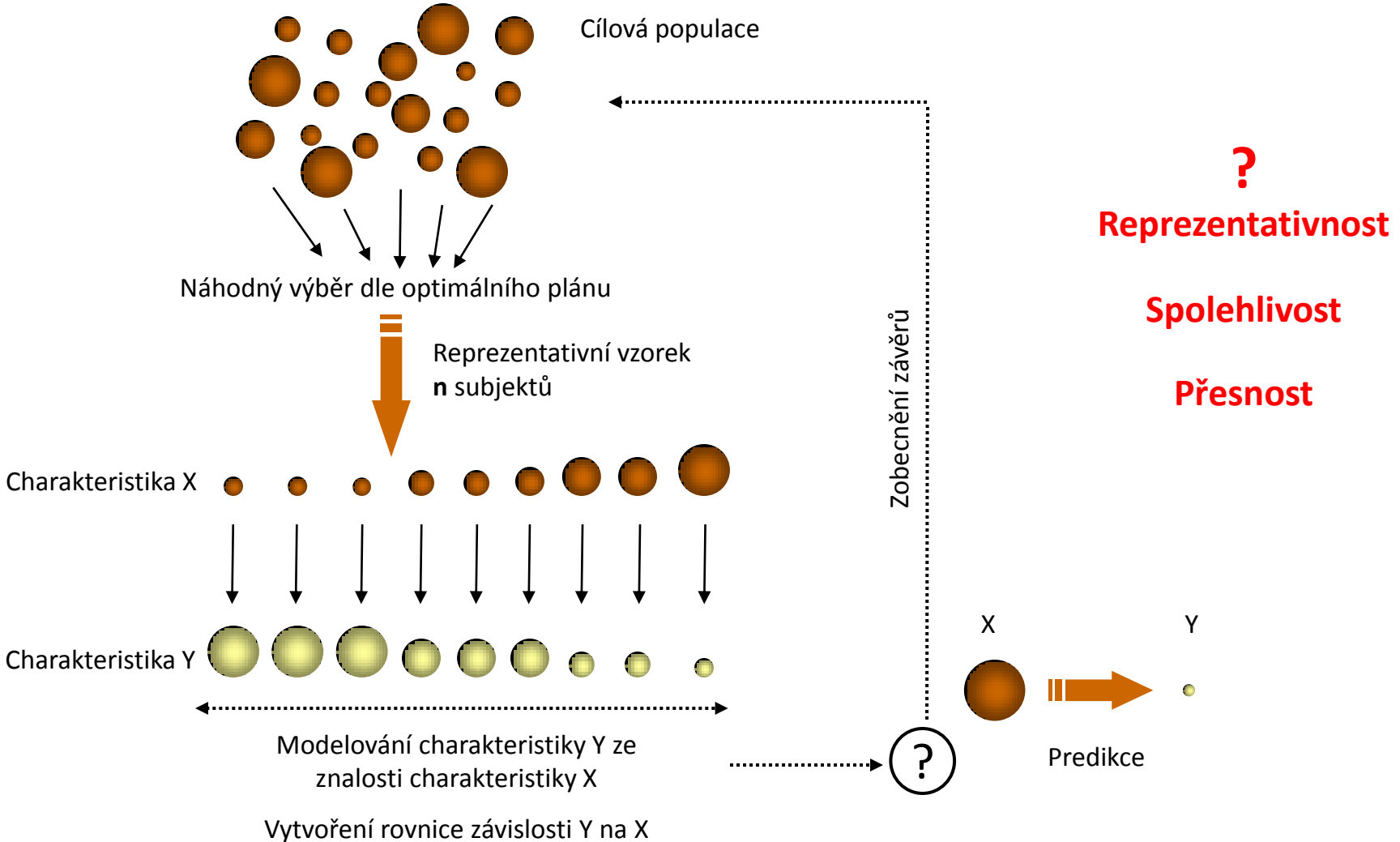


**?**  
**Reprezentativnost**  
**Spolehlivost**  
**Přesnost**

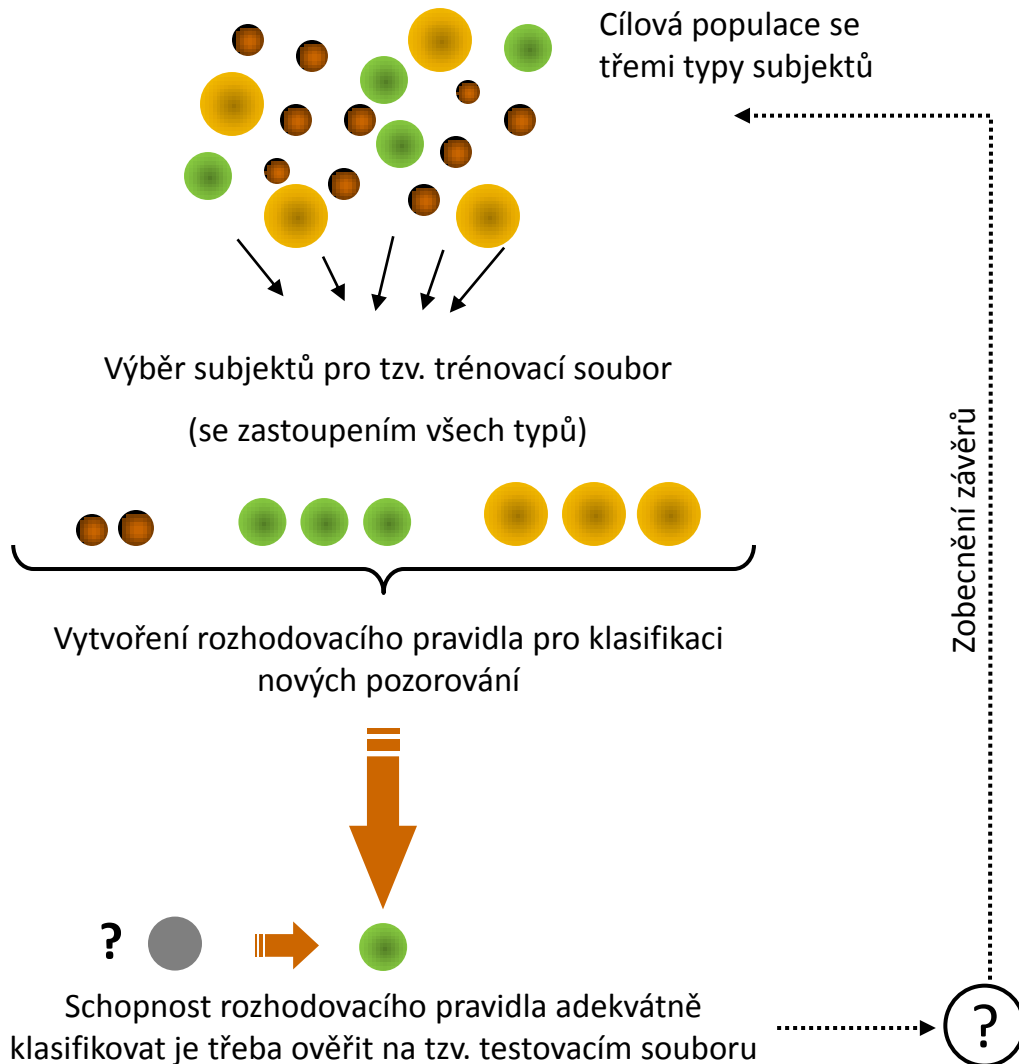
# Srovnání skupin – srovnání pozorované variability



# Predikce neznámých hodnot + stochastické modelování

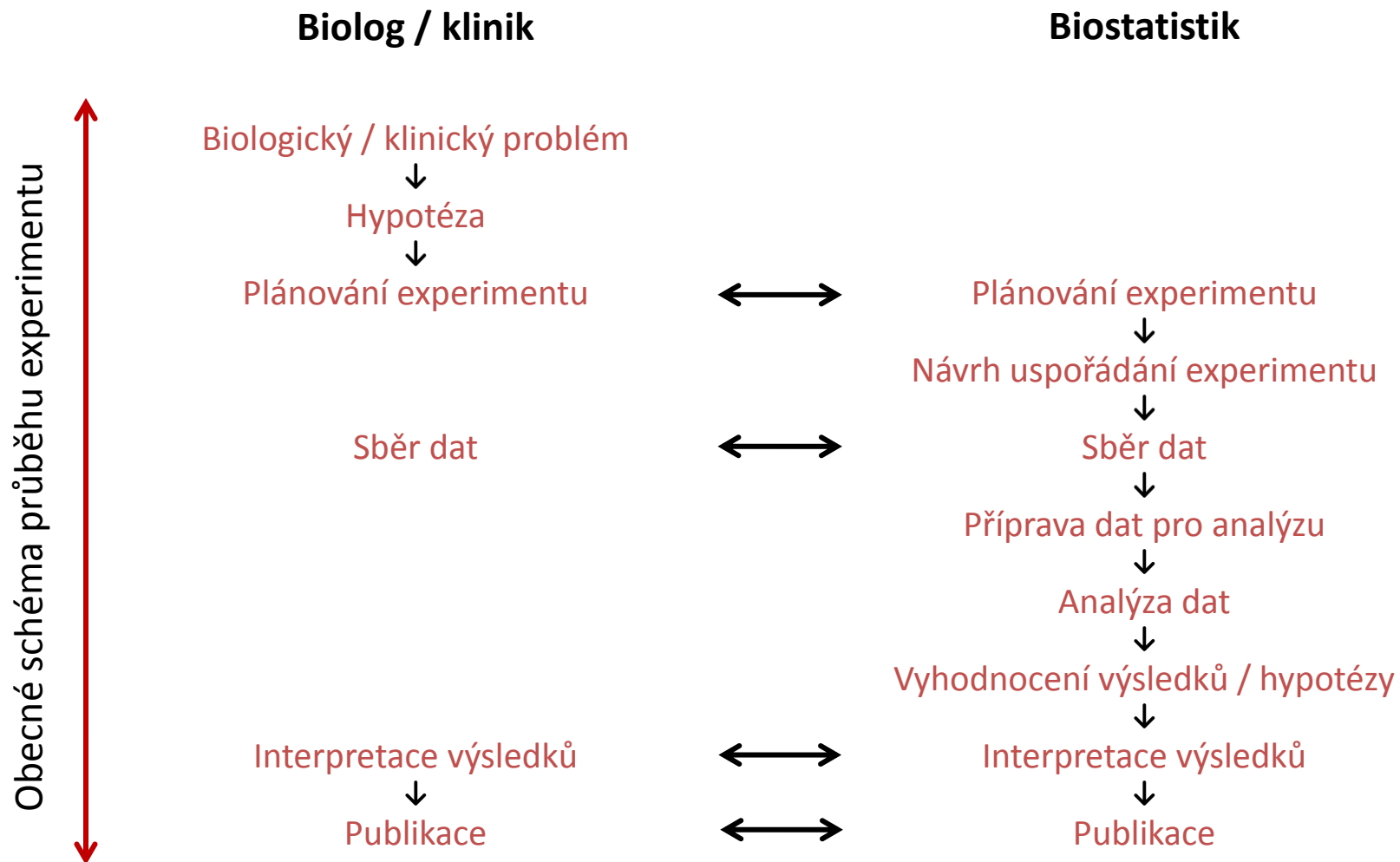


# Klasifikace nových pozorování – klasifikační algoritmy



**?**  
**Reprezentativnost**  
**Spolehlivost**  
**Přesnost**

# Biostatistiku lze najít všude...





# 3. Klíčové aspekty biostatistiky

*„Statistical analysis allows us to put limits on our uncertainty, but not to prove anything.“*

Douglas G. Altman

# Klíčové aspekty biostatistiky



# Klíčové aspekty – zkreslení

- ➡ V jakémkoliv hodnocení se snažíme vyhnout **zkreslení výsledků** („biased results“) – tedy zkreslení výsledků jinými faktory než těmi, které jsou cíli studie.
- ➡ Statistické srovnání není nikdy 100% spolehlivé, existuje náhoda a tedy i pravděpodobnost chybného úsudku – to nelze ovlivnit.
- ➡ Chceme použít adekvátní metody pro odstranění vlivů, které by zkreslily výsledky a nebyly přítom náhodné (např. zastoupení pohlaví).

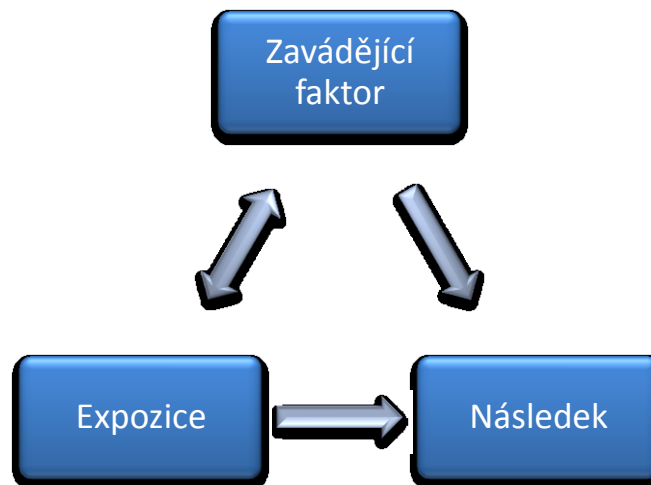


# Klíčové aspekty – zkreslení

➡ Pojem **zavádějící faktor**

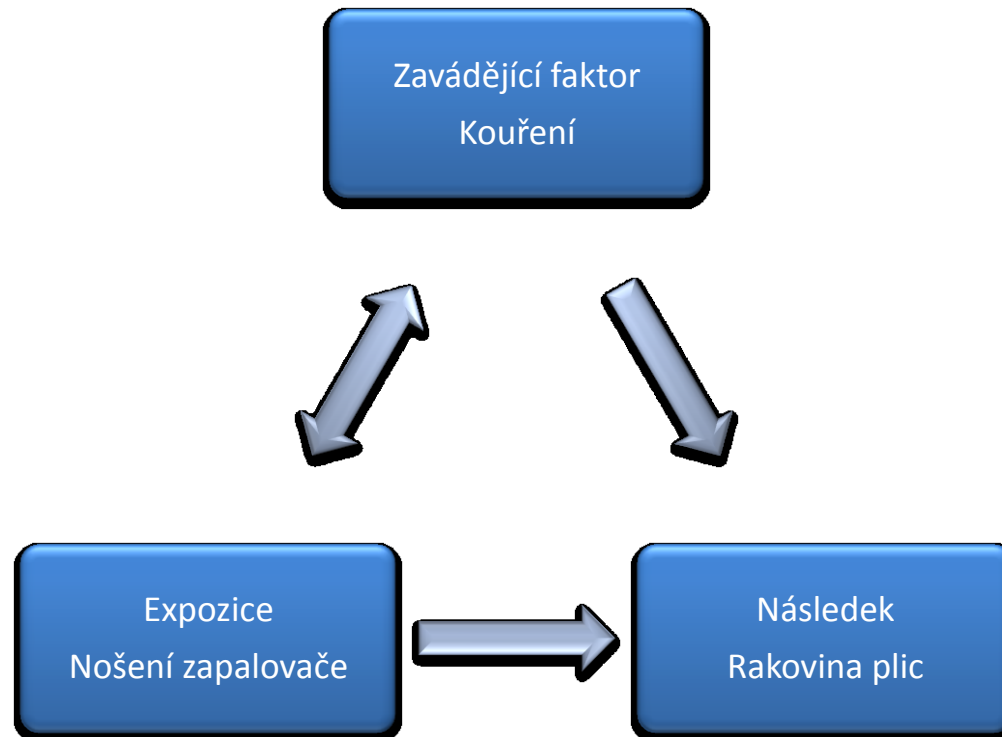
➡ Pro zavádějící faktor současně platí, že

- ➡ přímo nebo nepřímo ovlivňuje sledovaný následek,
- ➡ je ve vztahu se studovanou expozicí ,
- ➡ není mezikrokem mezi expozicí a následkem.

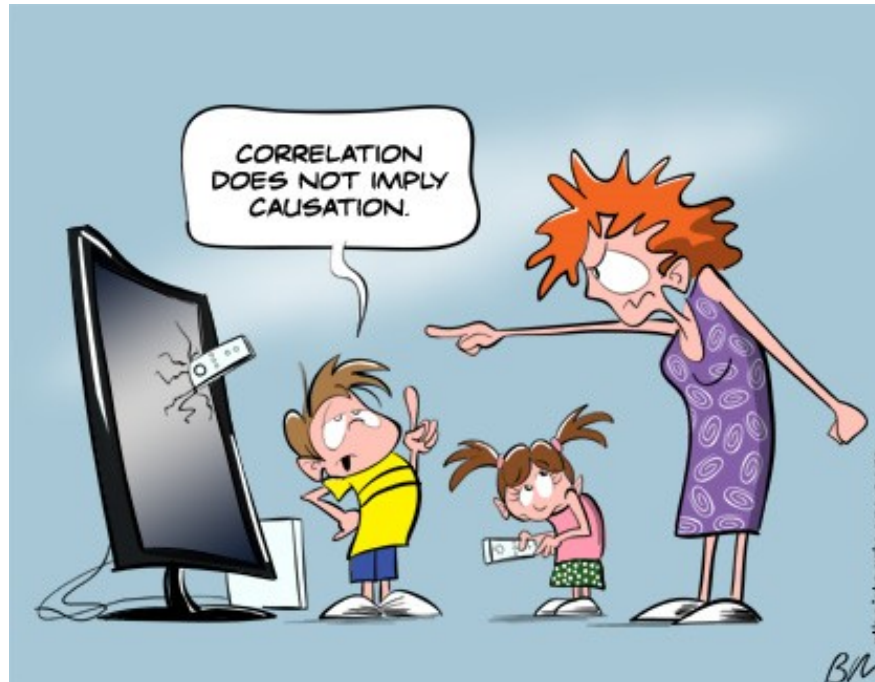


# Klíčové aspekty – zkreslení

- ➡ Příklad na **zavádějící faktor** (přímo nebo nepřímo ovlivňuje sledovaný následek, je ve vztahu se studovanou expozicí, není mezikrokem mezi expozicí a následkem).

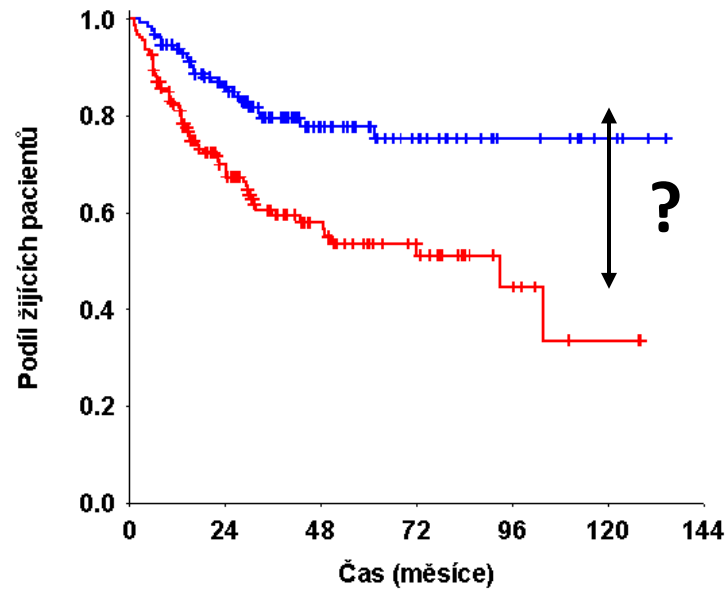


# Klíčové aspekty – zkreslení



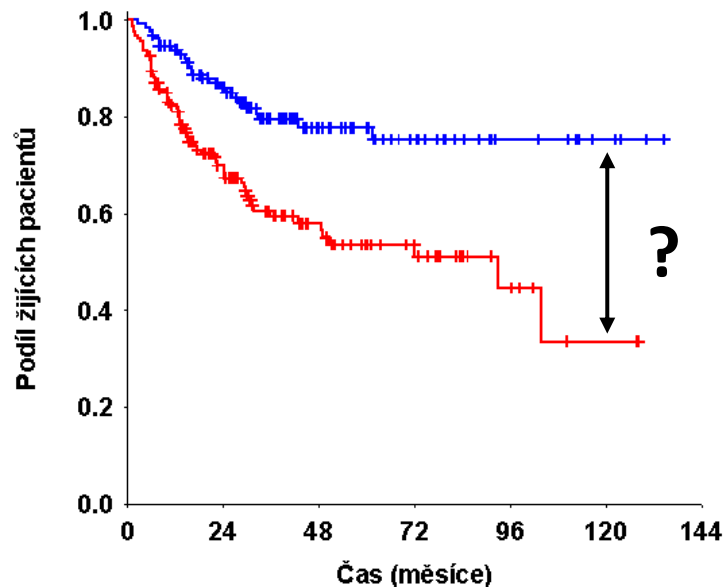
# Klíčové aspekty – zkreslení

- ➡ Čím by mohl být způsoben pozorovaný rozdíl v 10letém přežití pacientů s nádorem trávicího traktu?



# Klíčové aspekty – zkreslení

- ➡ Čím by mohl být způsoben pozorovaný rozdíl v 10letém přežití pacientů s nádorem trávicího traktu?



Léčba?

Nějaký prognostický faktor?

Stadium nemoci?

Věk?

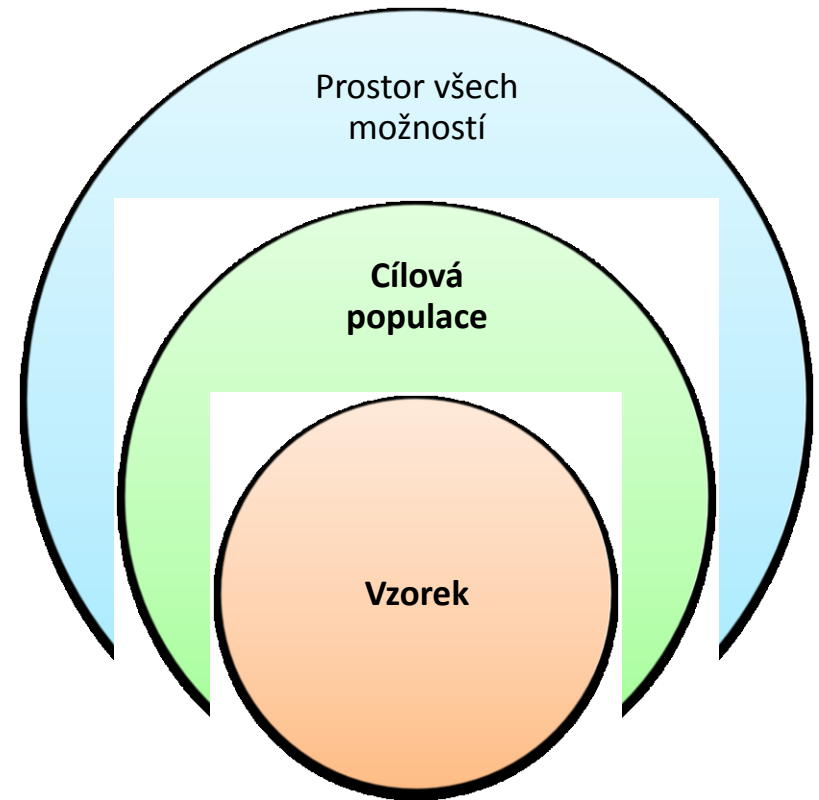


# Klíčové aspekty – zkreslení

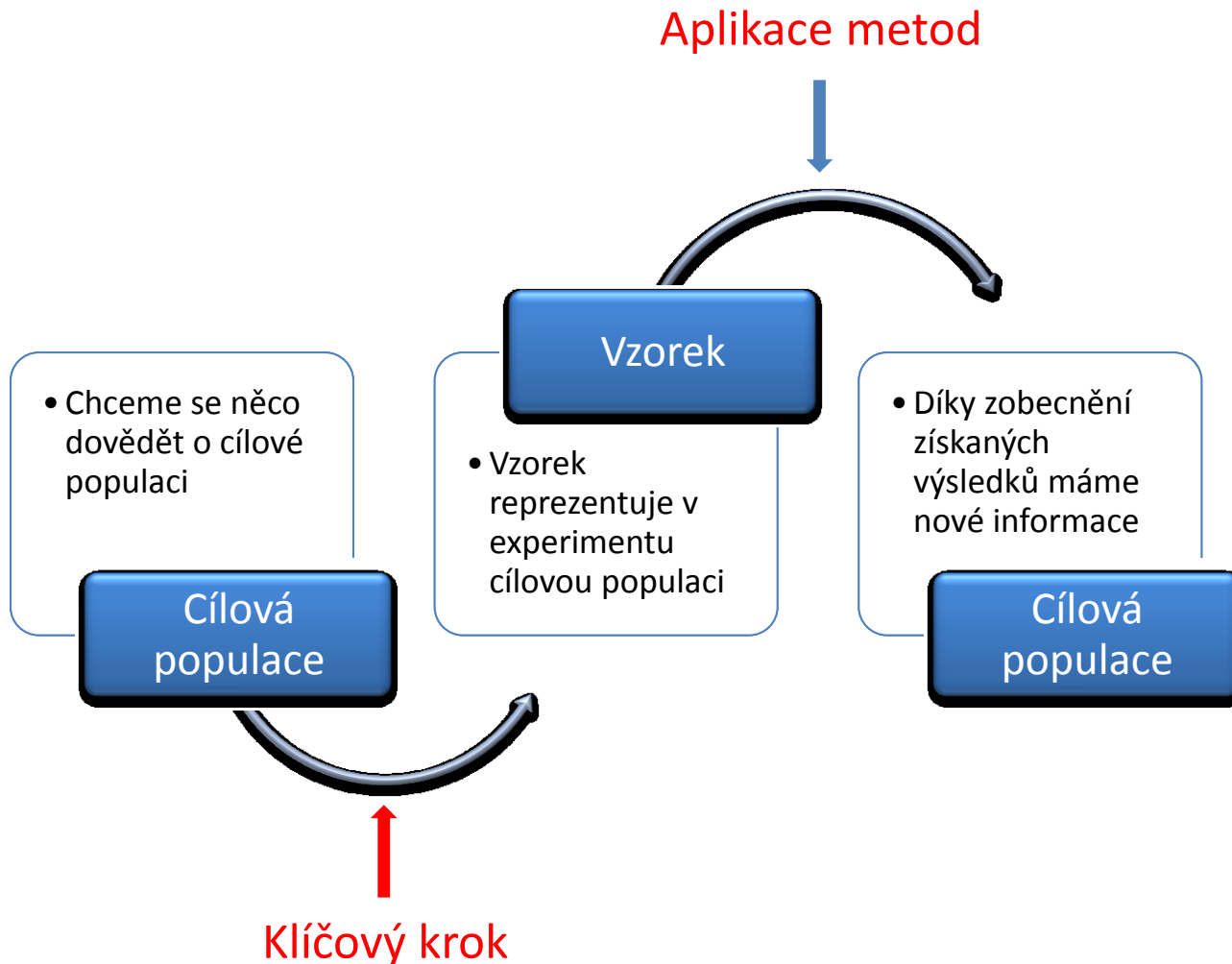
- ➡ Zkreslení výsledků souvisí s pojmem „**validita klinické studie**“.
- ➡ Medicína založená na důkazech – zajímají nás pouze „**kvalitní**“ důkazy.
- ➡ Hlavním aspektem kvality je **validita získaných výsledků**.
  
- ➡ **Interní validita studie:** odráží, jak moc lze rozdíly v účinnosti a bezpečnosti pozorované u srovnávaných skupin přisuzovat sledované intervenci. Chceme minimalizovat nenáhodnou chybu (zkreslení).
  
- ➡ **Externí validita studie:** odráží zobecnitelnost (z hlediska korektnosti) výsledků na jiné populace a experimentální podmínky.

# Klíčové aspekty – reprezentativnost

- ➔ Pojem **cílová populace** – skupina subjektů, o které chceme zjistit nějakou informaci.
- ➔ Pojem **experimentální vzorek** – podskupina cílové populace, kterou „máme k dispozici“.
  - ➔ **Musí odpovídat svými charakteristikami cílové populaci.**
  - ➔ Chceme totiž zobecnit výsledky na celou cílovou populaci.
  - ➔ Souvislost s náhodným výběrem.



# Klíčové aspekty – reprezentativnost



# Klíčové aspekty – srovnatelnost

- ➔ Korektní výsledky při srovnávacích analýzách lze získat pouze při srovnávání srovnatelného.
- ➔ V kontrolovaných klinických studiích je srovnatelnost zajištěna randomizací.
- ➔ U studií bez randomizace je nutné se tématu srovnatelnosti skupin věnovat.
- ➔ Metody adjustace, matching, propensity scores.



# Klíčové aspekty – spolehlivost

- ➡ Ve většině studií nás zajímá kvantifikace sledovaného efektu nebo charakteristiky, obecně náhodné veličiny, ve formě jednoho čísla, bodového odhadu.
- ➡ Bodový odhad je však sám o sobě nedostatečný.
- ➡ Je nutné ho doplnit intervalovým odhadem, který odpovídá pravděpodobnostnímu chování sledované veličiny, tedy odpovídá určité spolehlivosti výsledku.

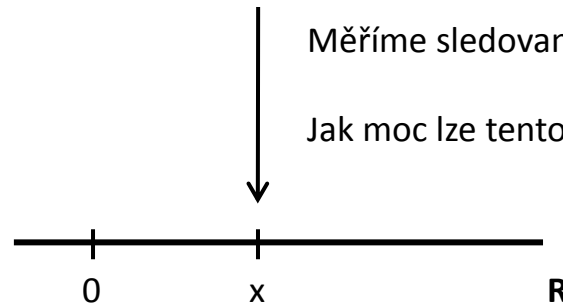


# Klíčové aspekty – spolehlivost

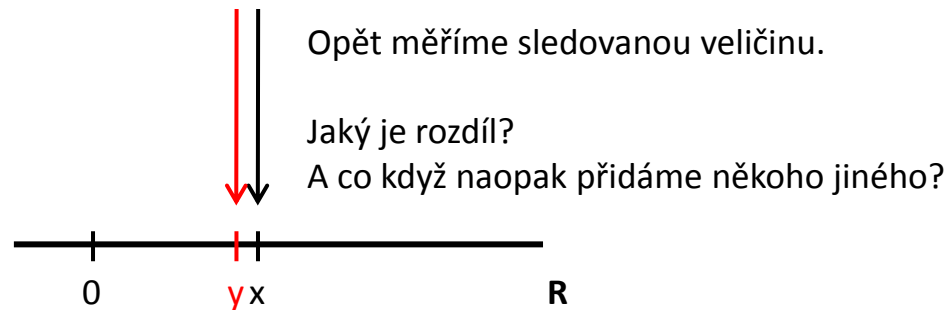


Měříme sledovanou veličinu a následně spočítáme odhad.

Jak moc lze tento bodový odhad zobecnit na cílovou populaci?

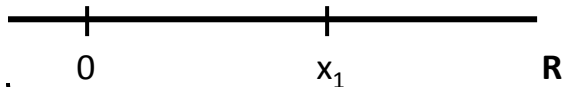


# Klíčové aspekty – spolehlivost

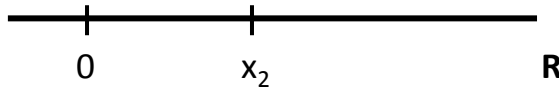


# Klíčové aspekty – spolehlivost

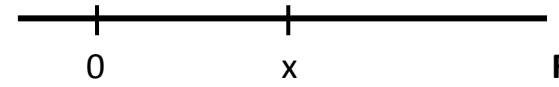
Výběr číslo 1



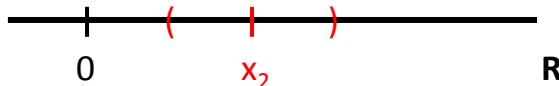
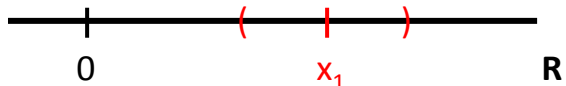
Výběr číslo 2



Celá cílová populace



Pracujeme-li s výběrem z cílové populace, je třeba na základě variability pozorovaných dat spočítat tzv. interval spolehlivosti pro bodový odhad.



Umíme-li „změřit“ celou cílovou populaci, nepotřebujeme interval spolehlivosti, protože jsme schopni odhadnout sledovaný parametr přesně – v praxi je tato situace nereálná.

Interval spolehlivosti na základě výběru číslo 1.





# Klíčové aspekty – významnost

- ➡ Analytické výsledky studie nemusí odpovídat realitě a skutečnosti.  
Statistická významnost jednoduše nemusí znamenat příčinný vztah!
- ➡ Statistická významnost pouze indikuje, že pozorovaný rozdíl není náhodný (ve smyslu stanovené hypotézy).
- ➡ Stejně důležitá je i praktická významnost, tedy významnost z hlediska lékaře nebo biologa.
- ➡ Statistickou významnost lze ovlivnit velikostí vzorku.



# Klíčové aspekty – významnost

		Praktická významnost	
		ANO	NE
Statistická významnost	ANO	OK, praktická i statistická významnost je ve shodě.	Významný výsledek je statistický artefakt, prakticky nevyužitelný.
	NE	Výsledek může být pouhá náhoda, neprůkazný výsledek.	OK, praktická i statistická významnost je ve shodě.



# Klíčové aspekty – významnost

Praktická významnost

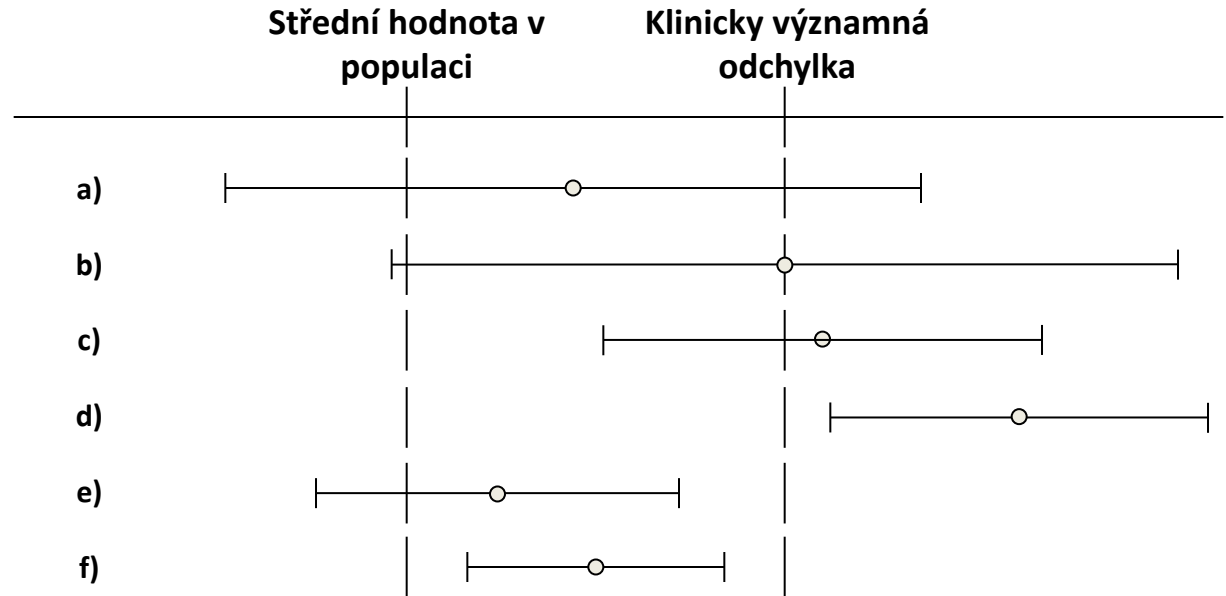
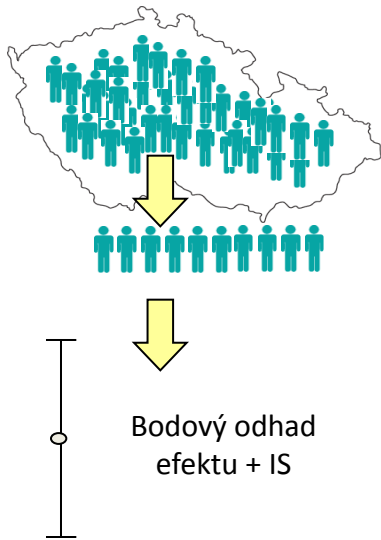
		Praktická významnost	
		ANO	NE
Statistická významnost	ANO	OK, praktická i statistická významnost jsou ve shodě.	Významný výsledek je statistický artefakt, prakticky nevyužitelný.
	NE	Výsledek může být pouhá náhoda, neprůkazný výsledek.	OK, praktická i statistická významnost jsou ve shodě.



Statisticky nevýznamný výsledek neznamená, že pozorovaný rozdíl ve skutečnosti neexistuje! Může to být způsobeno nedostatečnou informací v pozorovaných datech!



# Klíčové aspekty – významnost



Možnost	Statistická významnost	Klinická významnost
a)	ne	možná
b)	ne	možná
c)	ano	možná
d)	ano	ano
e)	ne	ne
f)	ano	ne

