Biostatistika
pro matematickou biologii

Tomáš Pavlík, Ladislav Dušek

pavlik@iba.muni.cz

Jarní semestr 2012

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

Česky:

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

Anglicky:

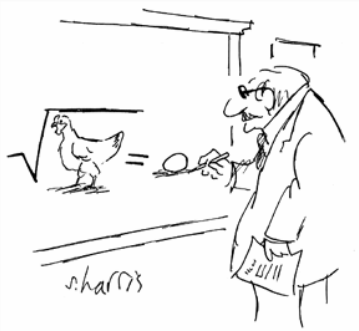
- Altman DG (1991) *Practical statistics for medical research*, Chapman&Hall/CRC, London.
- Zar JH (1999) *Biostatistical analysis*, Prentice-Hall, New Jersey.

Co znamená  ?

- Pokud bude přednáškový slide označen touto značkou, jedná se o klíčové téma, které musíte bezpodmínečně znát u zkoušky!
- Pokud někdo u zkoušky nebude adekvátně reagovat na látku z přednáškového slidu s touto značkou, výsledek zkoušky je F.

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 magnitude of reduction in CML burden is a key prognostic indicator for patients treated for CML with imatinib.

- 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 – follow-up and missing data

- In total, 658 CML patients diagnosed in years 2000–2008 with extended data record were entered into the Camelia database – **330 CML patients treated with first-line imatinib**.

- **Filter for the length of follow-up**: Patients with follow-up less than 12 months from the start of imatinib therapy were not considered in the analysis. The reason is the equal chance of achieving the CCgR or the MMR to imatinib therapy for all patients.

- **Filter for the key data missing**: Patients with missing values of key characteristics were not considered in the analysis. The key characteristics were defined as follows:
  - Date of birth, sex
  - Date of diagnosis, date of initiation of imatinib therapy
  - Sokal and Hasford prognostic scores
  - Blood count (used for definition of anemic patients)
  - Imatinib dosing at the treatment start.

Patients included in the analysis – the period of diagnosis

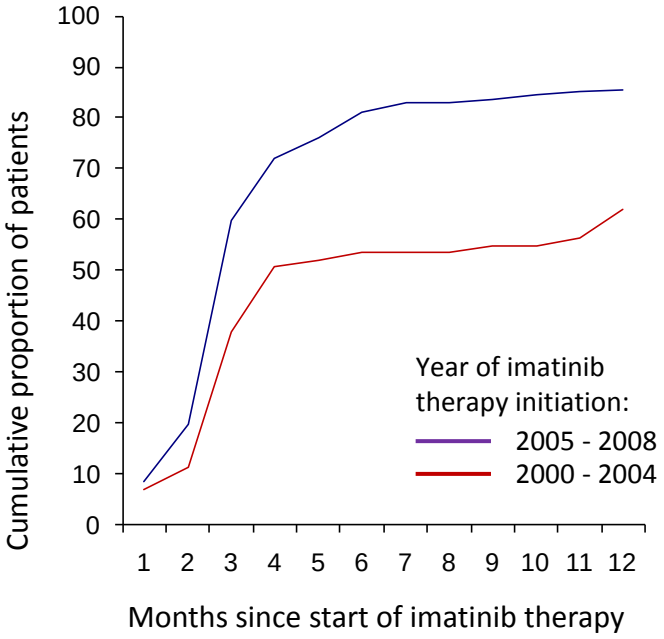
- A pilot analysis has revealed a strong association between the time to the CCgR or the MMR and the period of diagnosis represented by two time intervals 2000–2004 and 2005–2008.

- This phenomenon is related to the frequency and especially availability of follow-up examinations during the first 12 months after the start of the imatinib therapy.

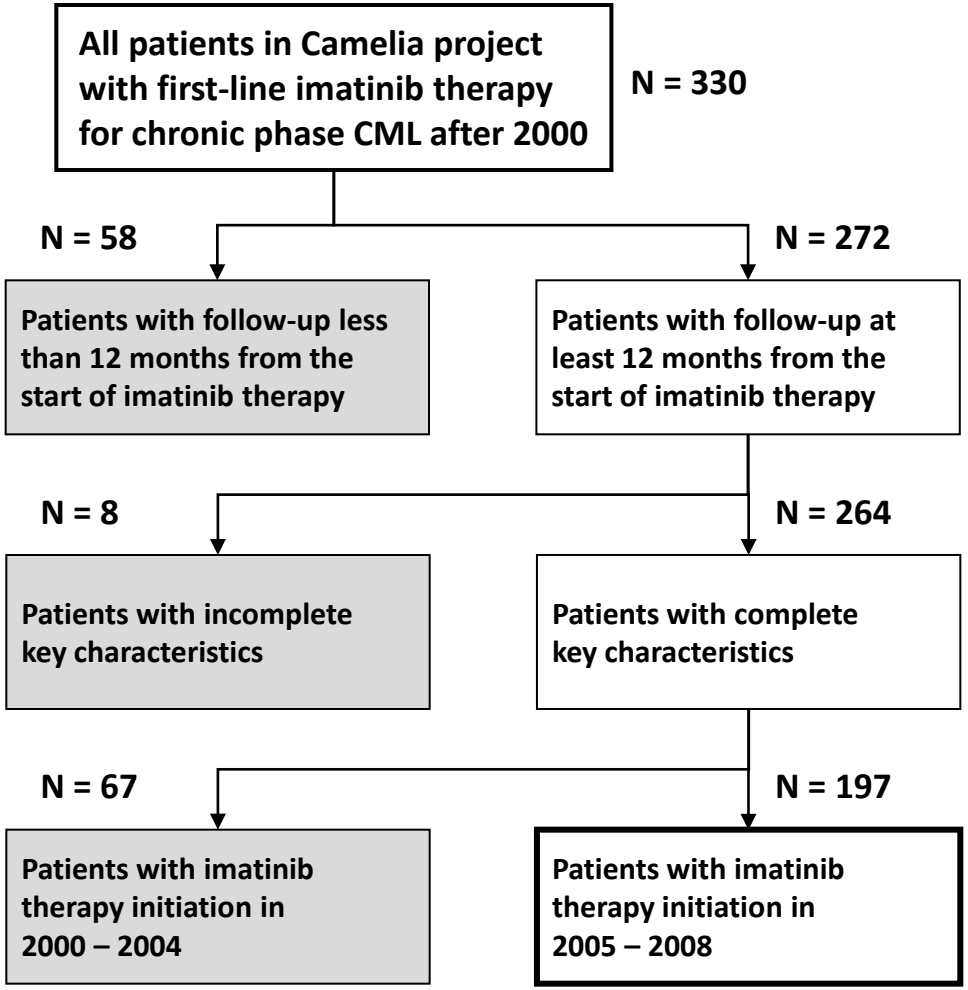
- The overall proportion of patients, who achieved the CCgR or the MMR, is similar in time periods 2000–2004 (73.2 %) and 2005–2008 (77.6 %).

- The latter time period (2005–2008) was selected for the modelling.

Cumulative proportion of patients with follow-up examination in first 12 months after imatinib therapy initiation

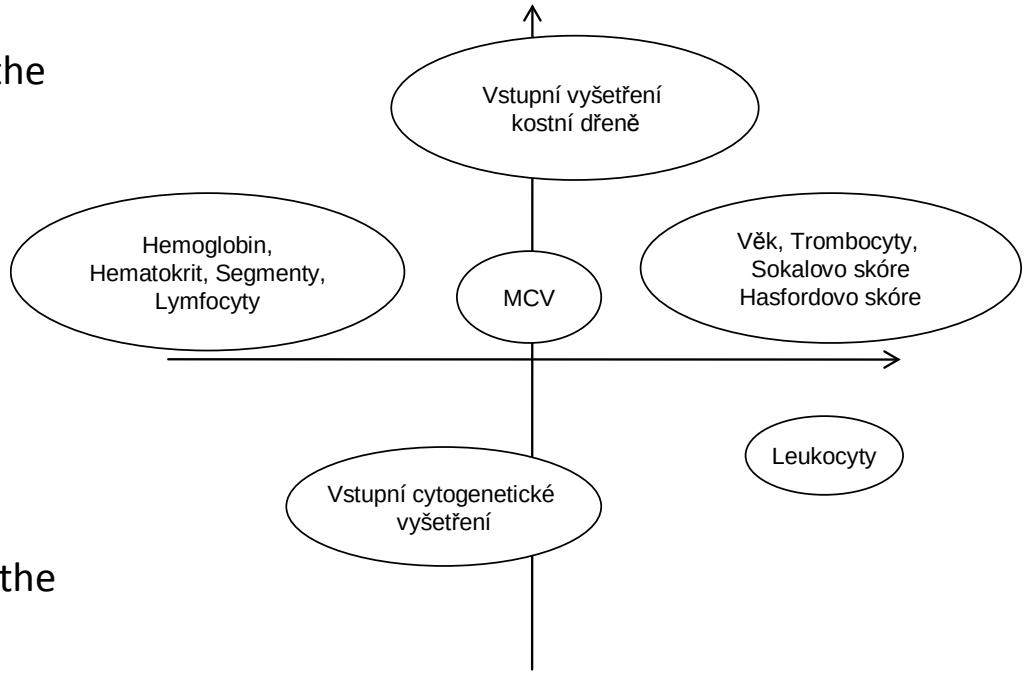


Patients included in the analysis – summary



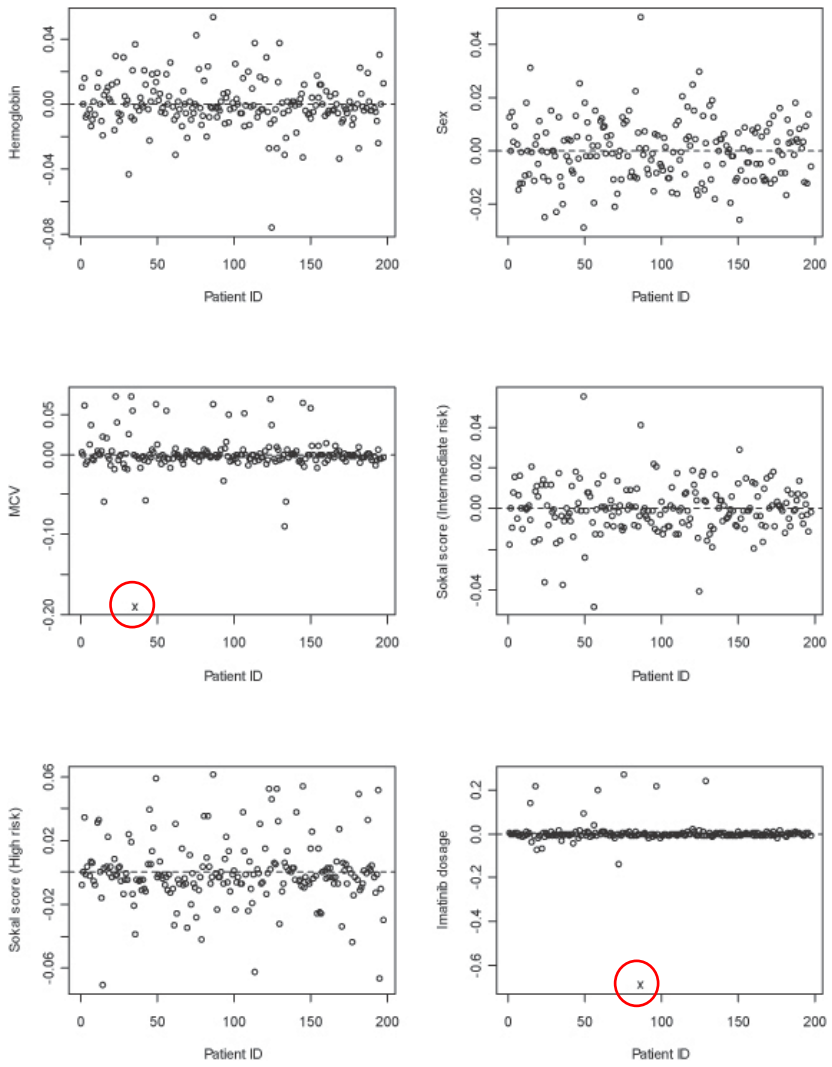
Modelling the endpoint – primary variable selection

- 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

Discussion

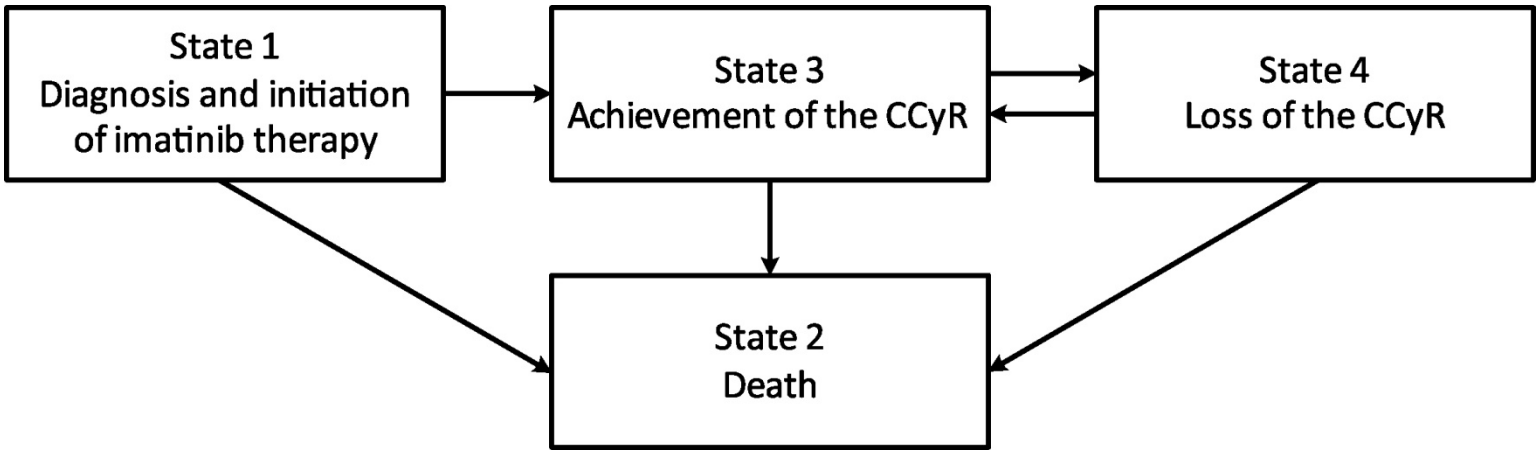
- The insignificant difference in hazard profiles between low risk and intermediate risk patients indicates that there is a space for a new prognostic score to be developed for the era of the imatinib therapy.

- The lack of statistical significance of clonal chromosome abnormalities can be explained with insufficient statistical information for there is only a small number of patients who developed Ph+ or Ph- clonal chromosome alterations.

- The same can be also true for the imatinib dosage as more observations with reduced imatinib dose would be needed to show that lower than standard dose of imatinib should be associated with limited response.

- We can conclude that the model has a potential in identification of patients, who are more likely to have problems with proper treatment response to imatinib therapy.

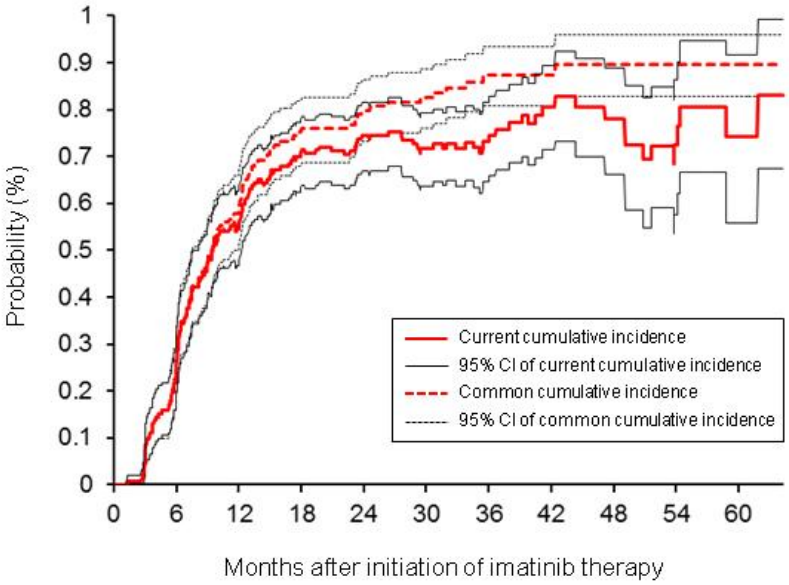
Př. 2 Multi-state model of CML therapy



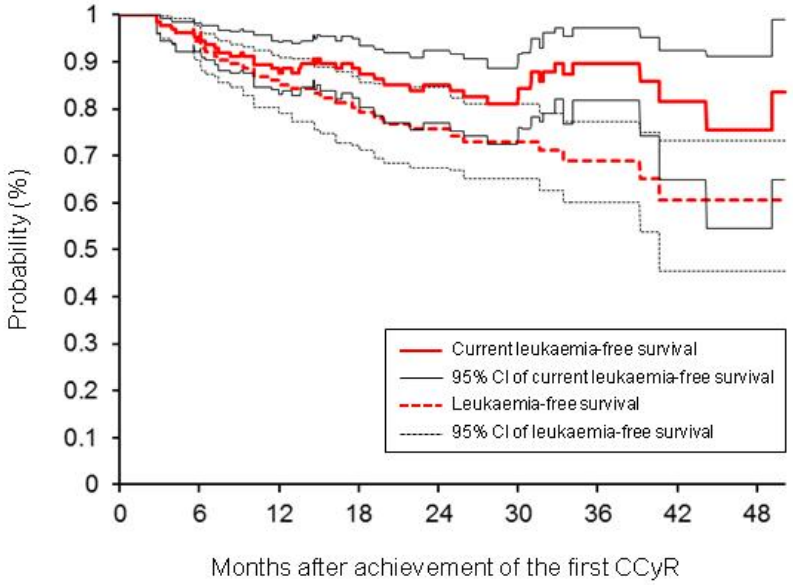
- Obviously, **all living patients can move from the CCyR** (state 3) **to the cytogenetic relapse** (state 4) **and vice versa repeatedly**.

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

CCI(*t*)



CLFS(*t*)



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

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

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

Pavlík *et al. BMC Medical Research Methodology* 2011, **11**:140
http://www.biomedcentral.com/1471-2288/11/140

RESEARCH ARTICLE Open Access

Estimation of current cumulative incidence of leukaemia-free patients and current leukaemia-free survival in chronic myeloid leukaemia in the era of modern pharmacotherapy

Tomáš Pavlík[1], Eva Janoušová[1], Zdeněk Pospíšil[2], Jan Mužík[1], Daniela Žáčková[3], Zdeněk Ráčil[3], Hana Klamová[4], Petr Cetkovský[4], Marek Trněný[4], Jiří Mayer[3] and Ladislav Dušek[1*]

Abstract

Background: The current situation in the treatment of chronic myeloid leukaemia (CML) presents a new challenge for attempts to measure the therapeutic results, as the CML patients can experience multiple leukaemia-free periods during the course of their treatment. Traditional measures of treatment efficacy such as leukaemia-free survival and cumulative incidence are unable to cope with multiple events in time, e.g. disease remissions or progressions, and as such are inappropriate for the efficacy assessment of the recent CML treatment.

Př. 3 Je použití inzulinového 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 inzulinu glargin při srovnání s adekvátní dávkou humánního inzulinu.

Co tato studie znamená ze statistického hlediska?

Jedná se o

1. observační studii
2. studii s „pokusem“ o adjustaci na dávkování inzulinu
3. studii s velmi krátkou délkou sledování pacientů ve skupině s vysokou dávkou inzulinu 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 inzulinu

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 inzulinu 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

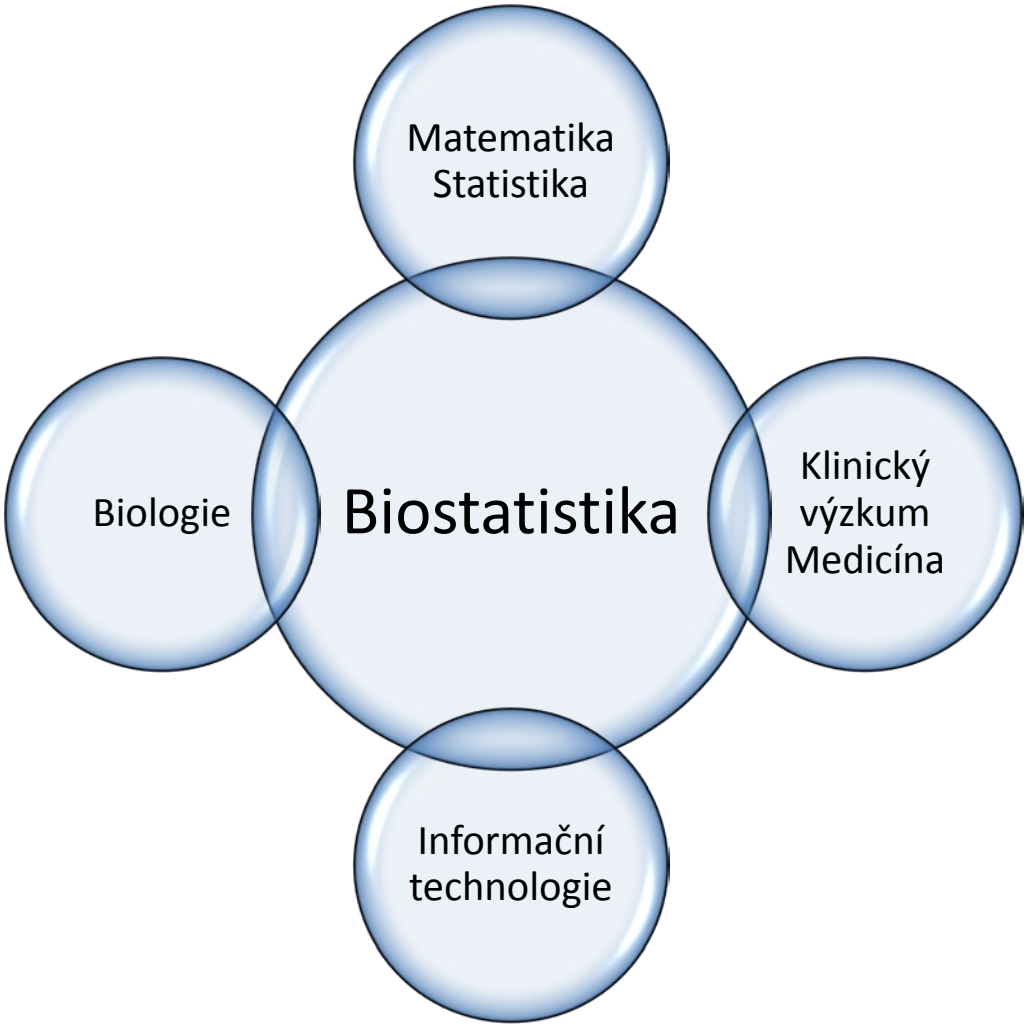
- Biostatistika je **aplikace statistických metod** v řešení biologických a klinických problémů.
- Snahou je **získat z pozorovaných dat užitečnou informaci**. V popředí zájmu je pozorovaná variabilita mezi studovanými subjekty, kterou chceme vysvětlit.
- Je **orientována na konkrétní problém**, ne na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

Význam biostatistiky

11 nejvýznamnějších 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

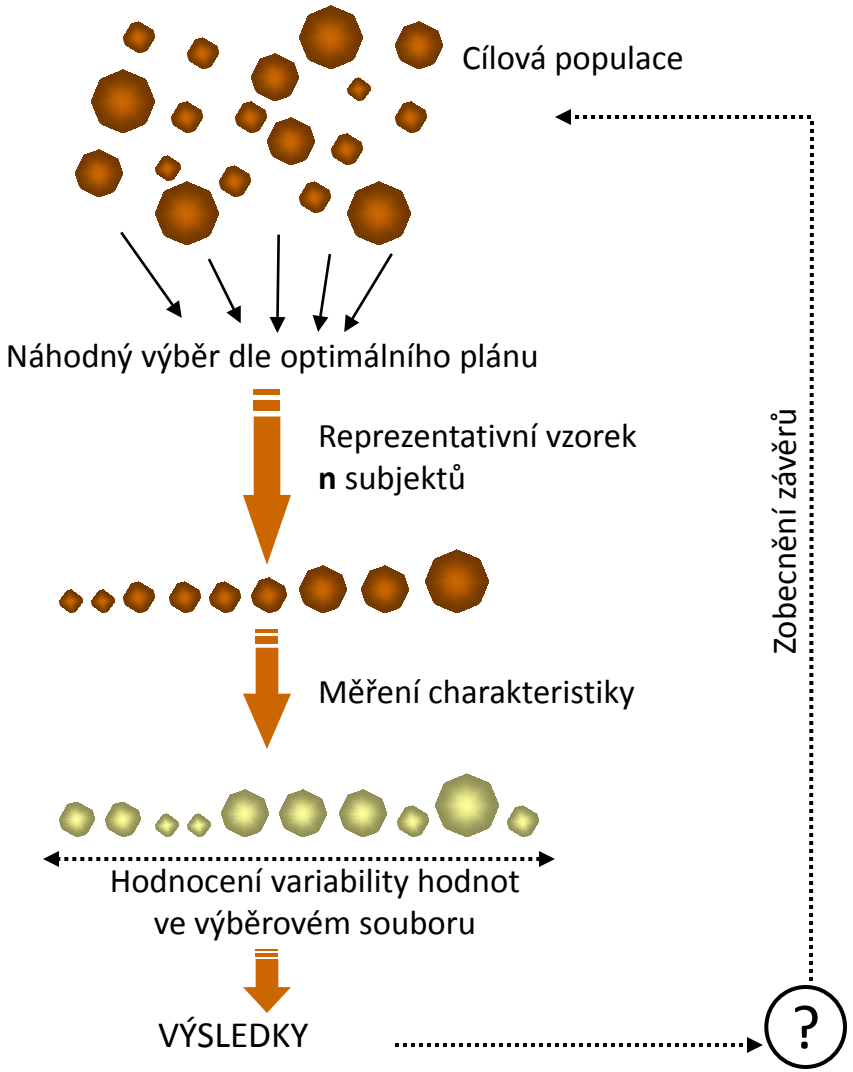
Biostatistika souvisí s dalšími vědami



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



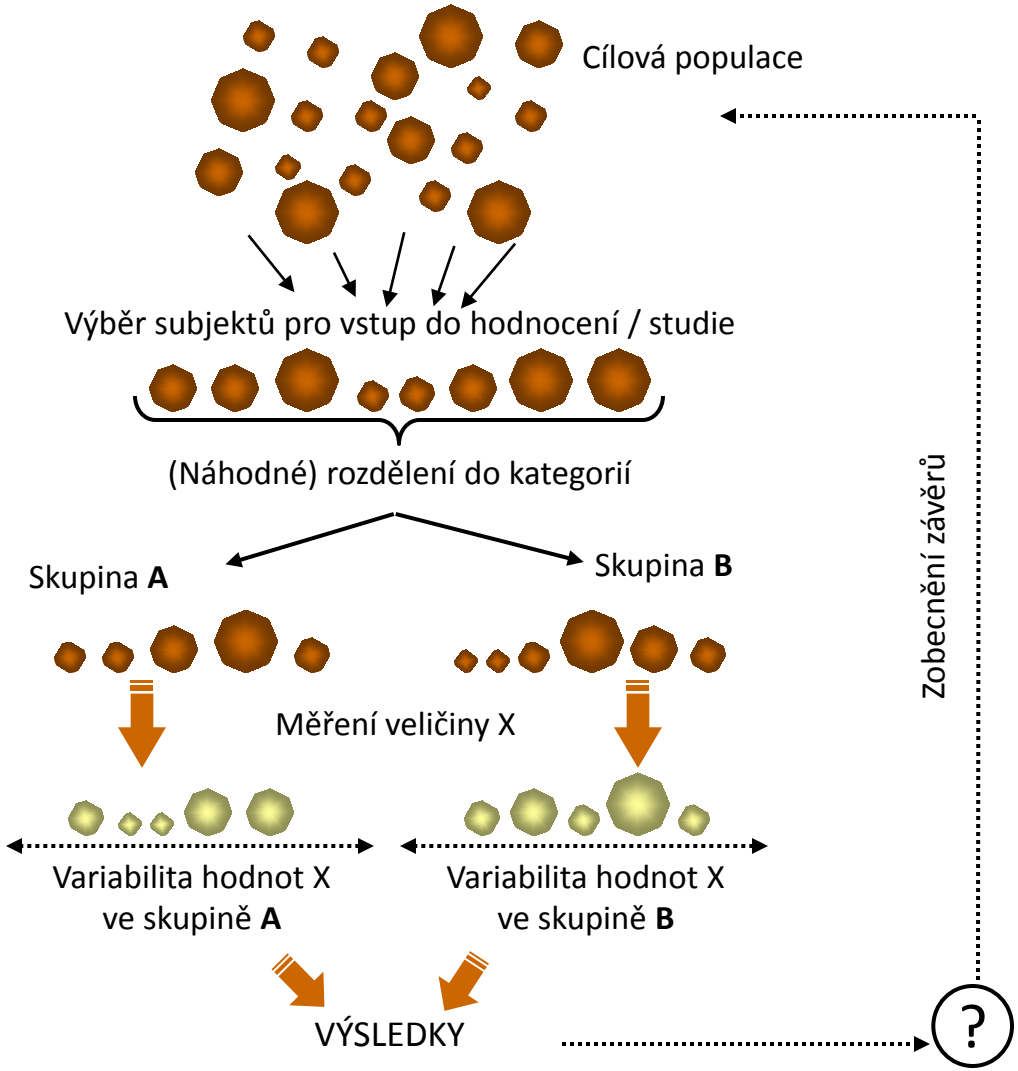
?

Reprezentativnost

Spolehlivost

Přesnost

Srovnání skupin – srovnání pozorované variability



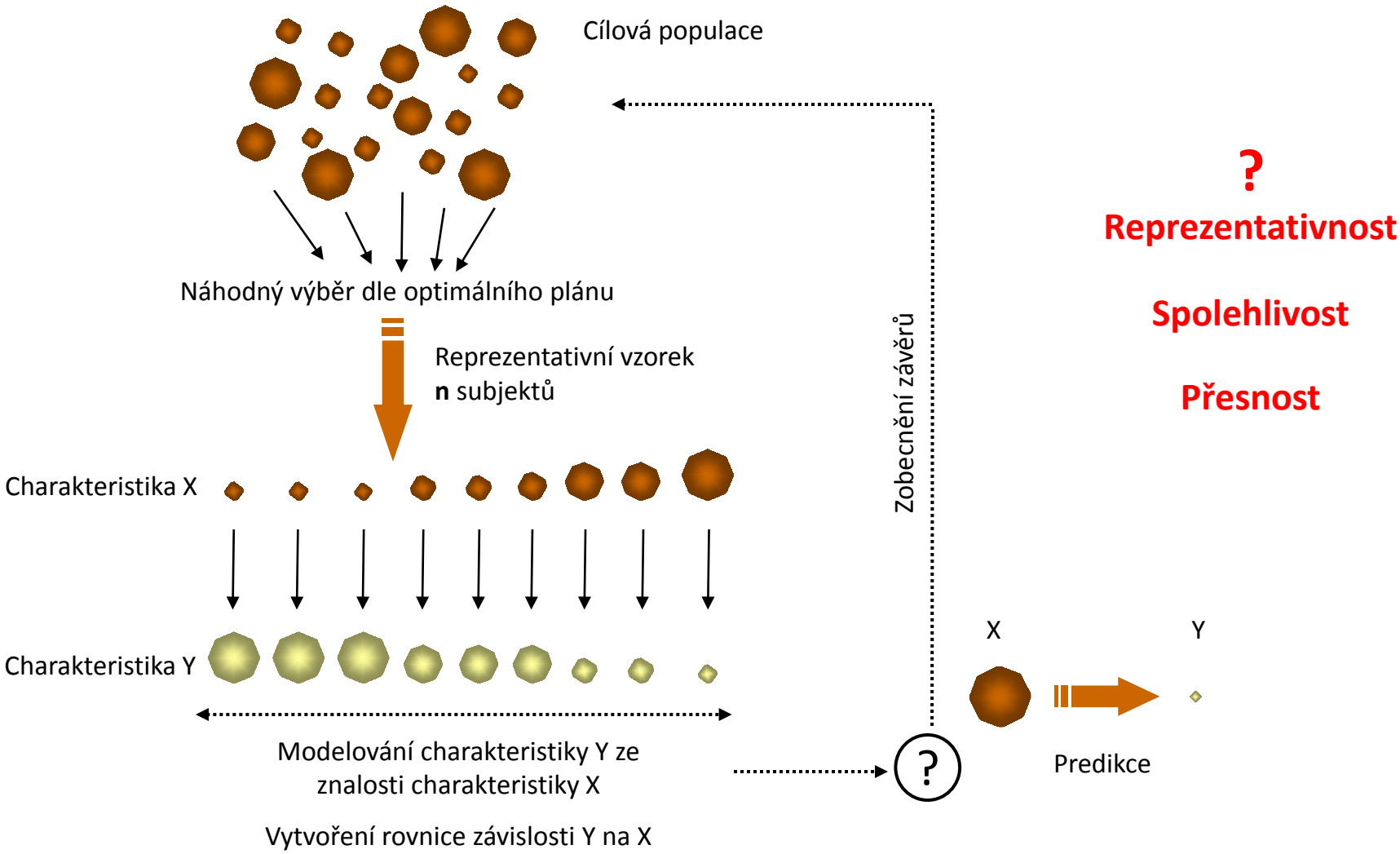
?

Srovnatelnost

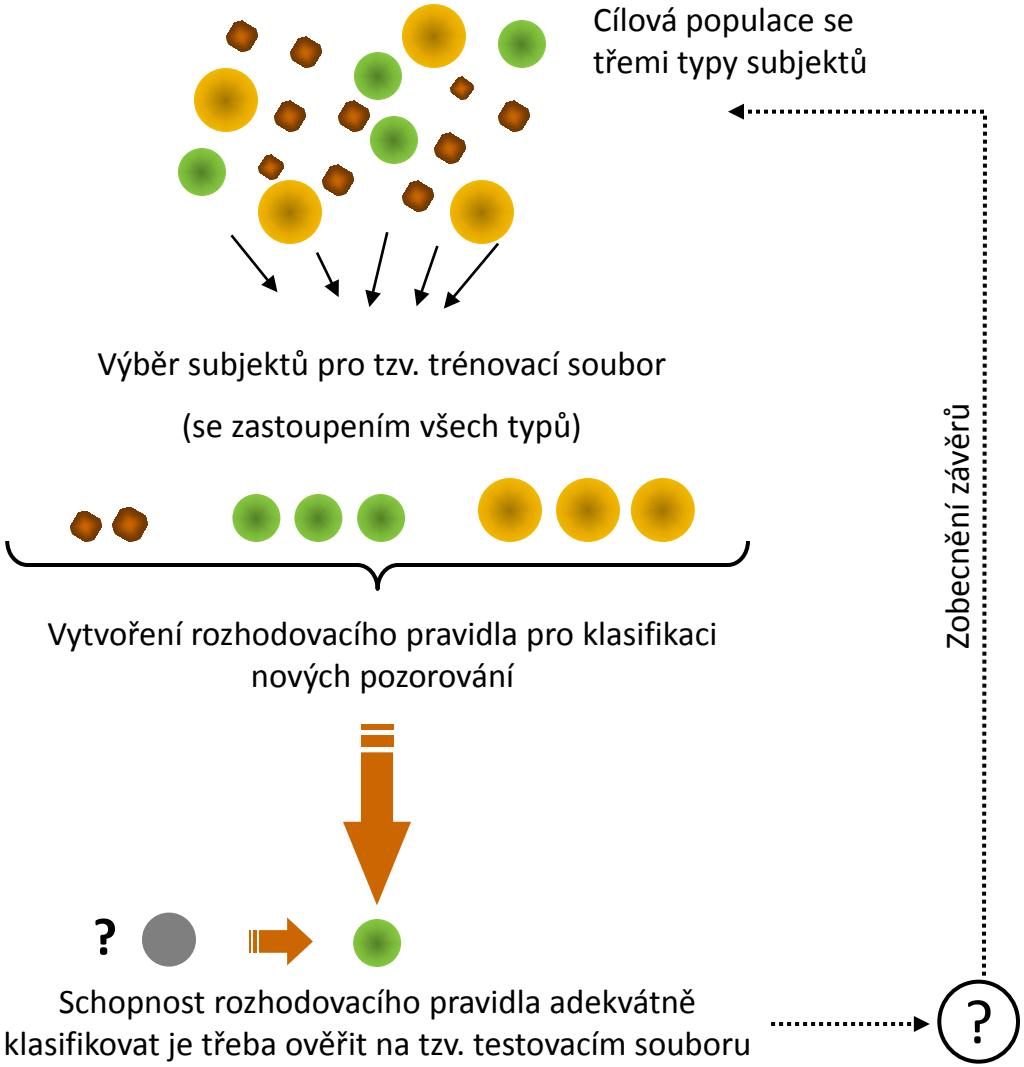
Spolehlivost

Přesnost

Predikce neznámých hodnot – stochastické modelování



Klasifikace nových pozorování – klasifikační algoritmy



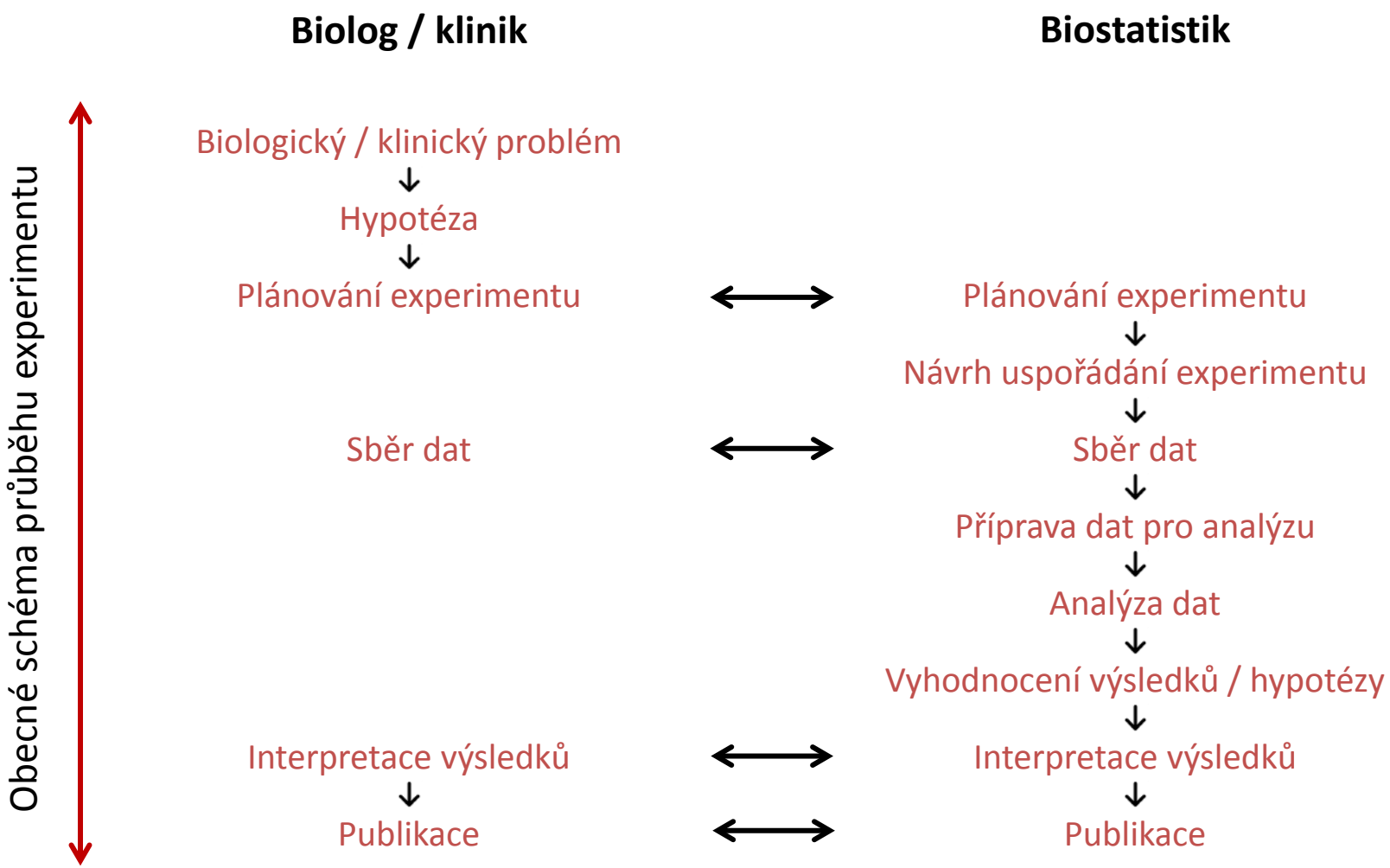
?

Reprezentativnost

Spolehlivost

Přesnost

Biostatistiku lze najít všude…



Biostatistiku lze najít všude…

A jak je to ve skutečnosti?

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

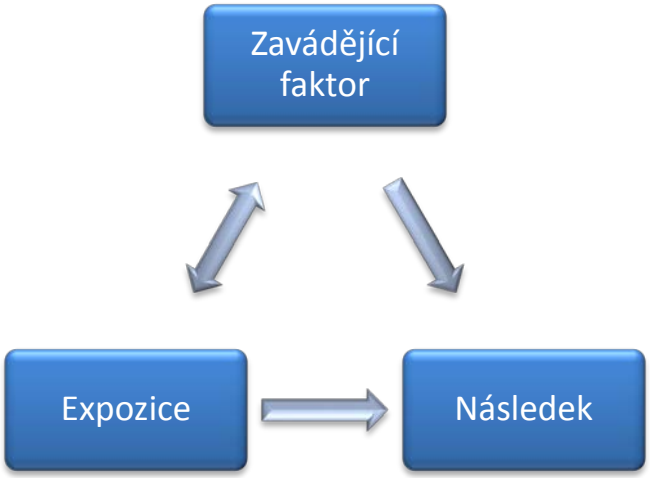
- V jakémkoliv hodnocení se snažíme vyhnout **zkreslení výsledků** („biased results“) – tedy zkreslení výsledků jinými faktory než těmi, které jsou cíli studie.

- Statistické srovnání není nikdy 100% spolehlivé, existuje náhoda a tedy i pravděpodobnost chybného úsudku – to nelze ovlivnit.

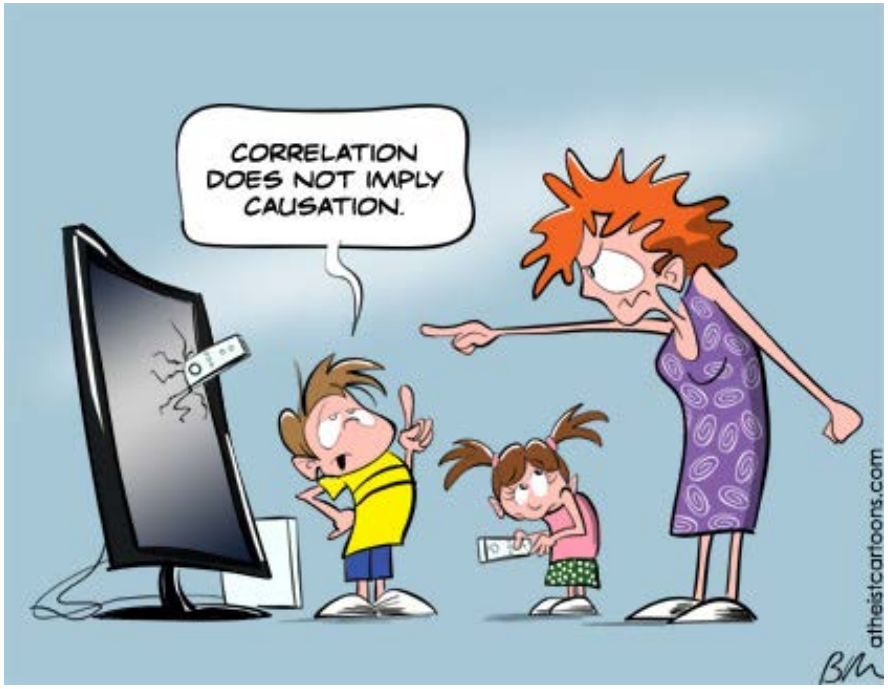
- Chceme použít adekvátní metody pro odstranění vlivů, které by zkreslily výsledky a nebyly přitom náhodné (např. zastoupení pohlaví).

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.

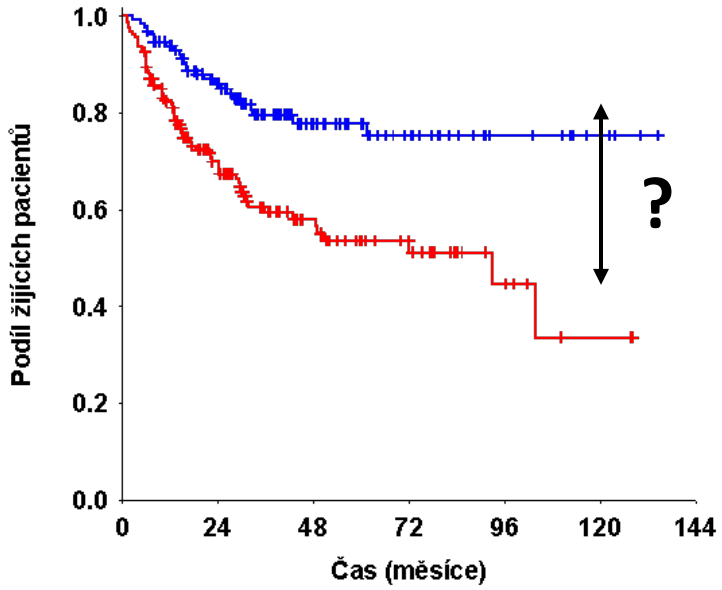


Klíčové aspekty – zkreslení



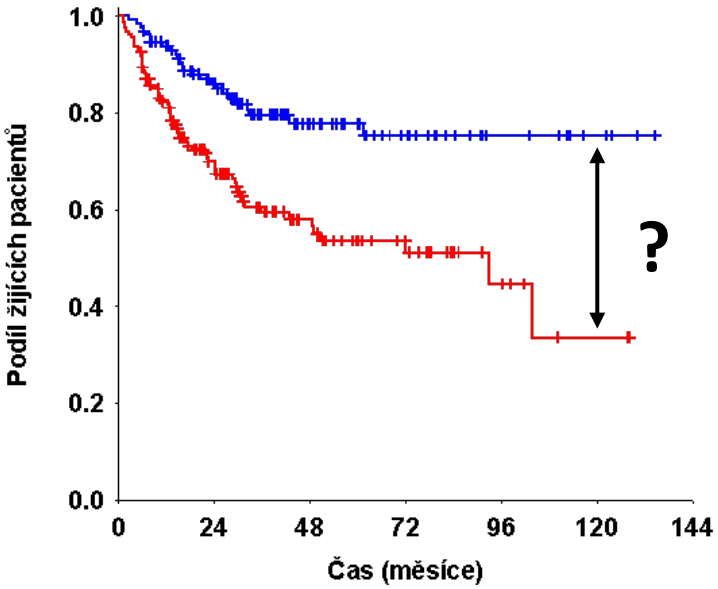
Klíčové aspekty – zkreslení

- 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?

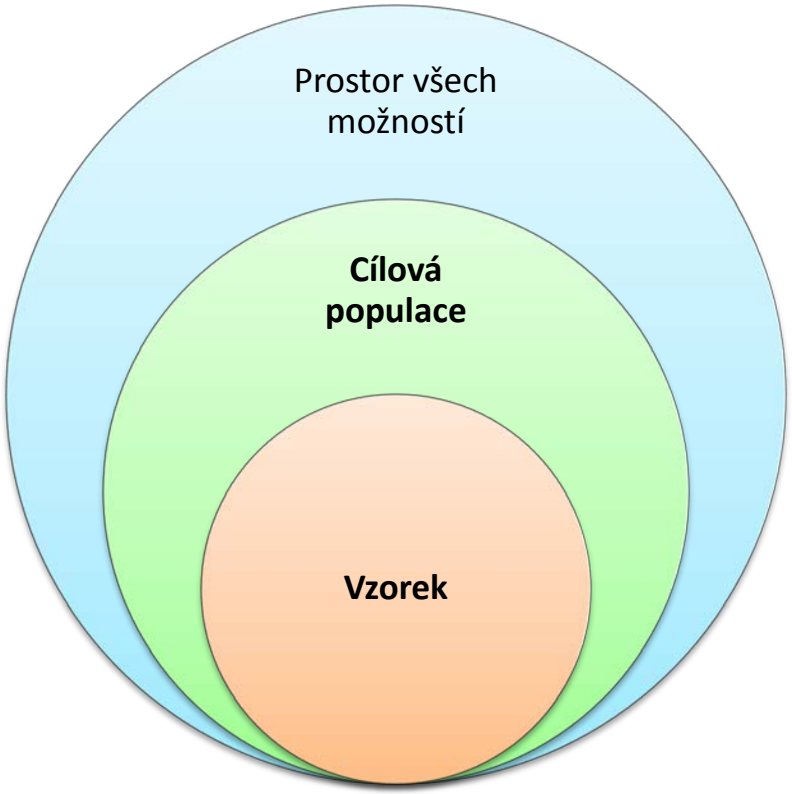
Nějaký prognostický faktor?

Stadium nemoci?

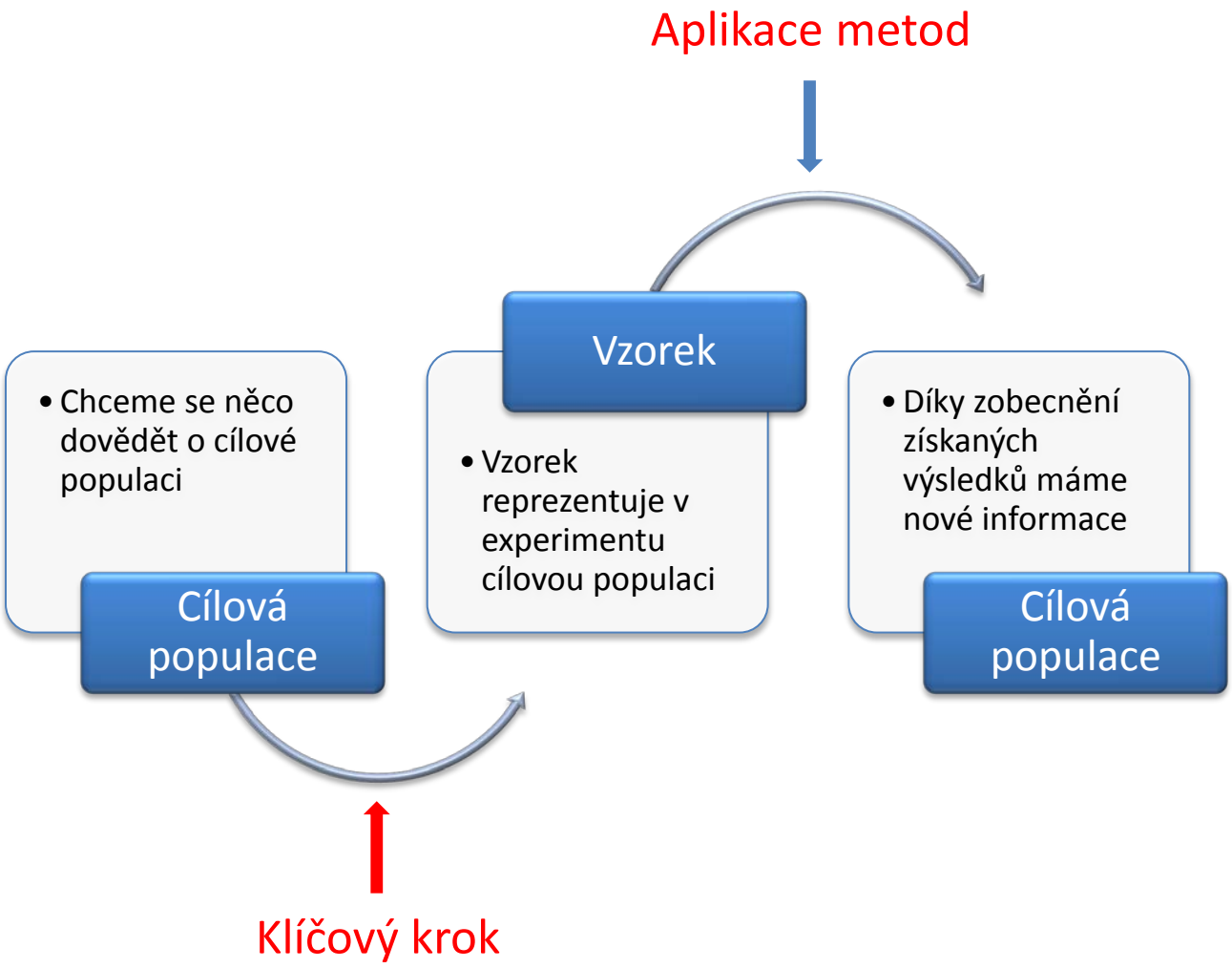
Věk?

Klíčové aspekty – 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.

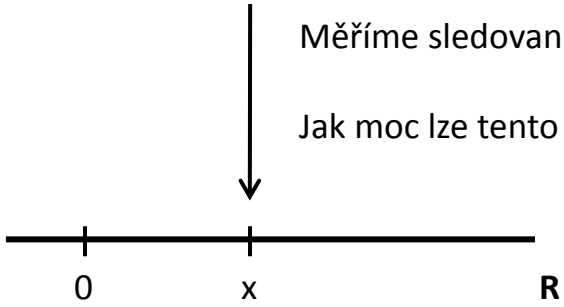
Klíčové aspekty – spolehlivost

- 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.

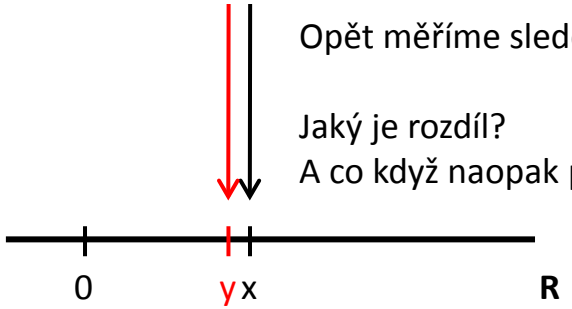
Klíčové aspekty – spolehlivost



Měříme sledovanou veličinu a následně spočítáme odhad.

Jak moc lze tento bodový odhad zobecnit na cílovou populaci?

Klíčové aspekty – spolehlivost

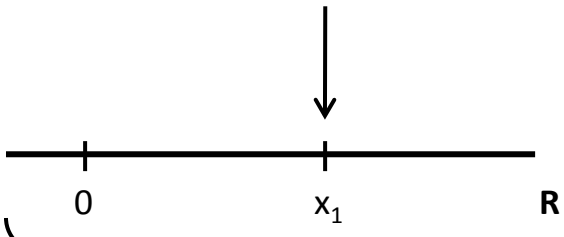


Opět měříme sledovanou veličinu.

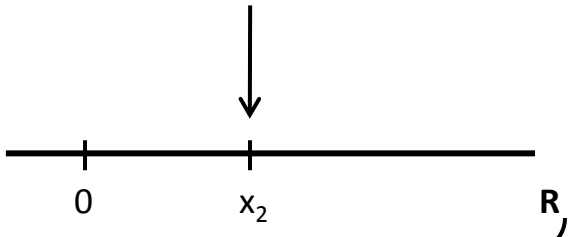
Jaký je rozdíl?
A co když naopak přidáme někoho jiného?

Klíčové aspekty – spolehlivost

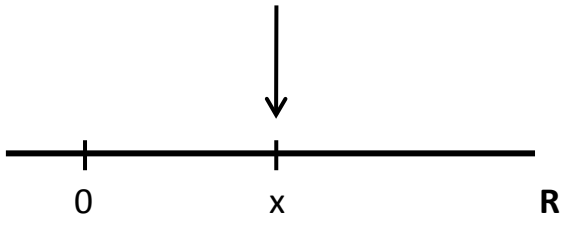
Výběr číslo 1



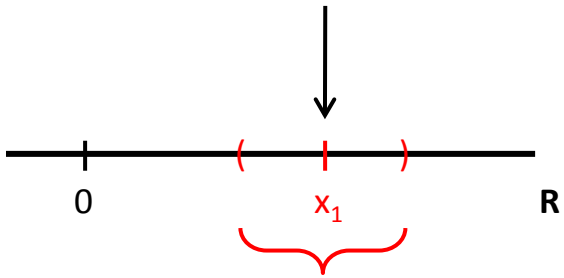
Výběr číslo 2



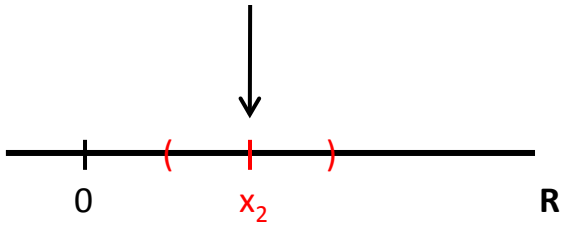
Celá cílová populace



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



Interval spolehlivosti na základě výběru číslo 1.



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

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

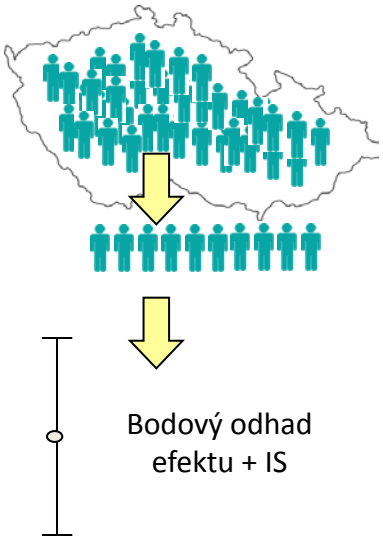
Statistická významnost	Praktická významnost: ANO	Praktická významnost: NE
ANO	OK, praktická i statistická významnost je ve shodě.	Významný výsledek je statistický artefakt, prakticky nevyužitelný.
NE	Výsledek může být pouhá náhoda, neprůkazný výsledek.	OK, praktická i statistická významnost je ve shodě.

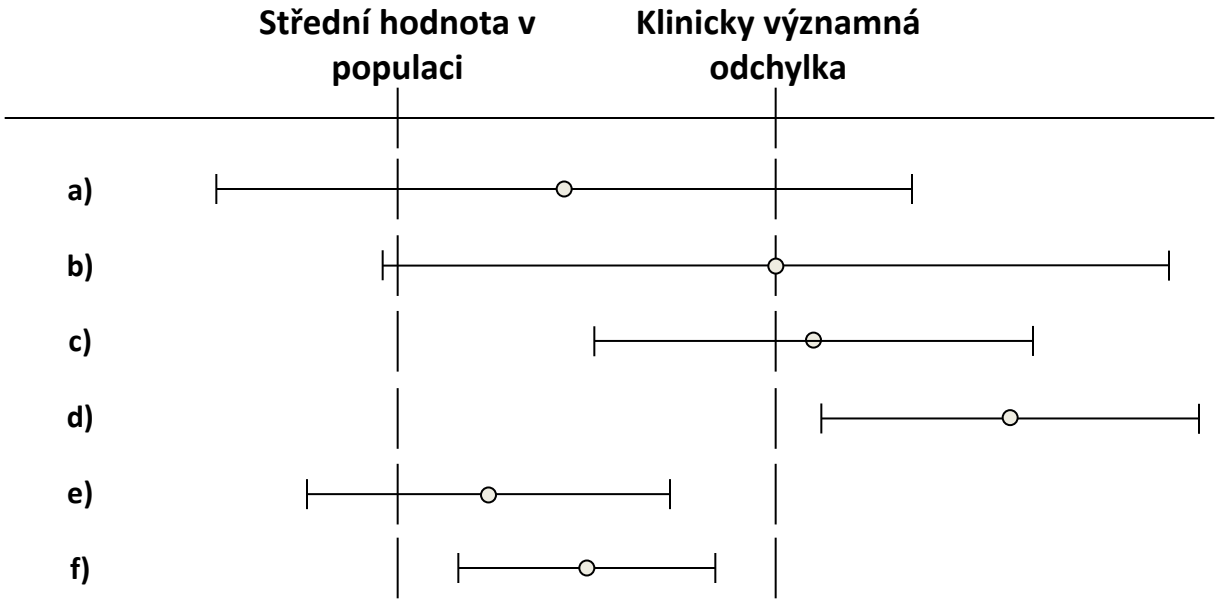
Klíčové aspekty – významnost

Statistická významnost	Praktická významnost: ANO	Praktická významnost: NE
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

Reklama na příští týden...

Statistika, biostatistika a analýza dat

Statistika	Biostatistika	Analýza dat
▪ Primárně je zaměřena na vývoj metod a algoritmů pro řešení teoretických problémů. ▪ Nicméně i statistika je vždy primárně motivována reálnými problémy. ▪ Vychází z teorie pravděpodobnosti.	▪ Propojení znalosti statistických metod a dané problematiky v řešení biologických a klinických úloh. ▪ Na prvním místě není teoretický vývoj, ale aplikace.	▪ Velmi obecná oblast bez jasné definice. ▪ Prostupuje různými odvětvími. ▪ Zahrnuje komplexní postupy hodnocení dat (čištění, kódování). ▪ Nemusí být založena na statistice.

Poděkování…

Rozvoj studijního oboru „Matematická biologie“ PřF MU Brno je finančně podporován prostředky projektu ESF č. CZ.1.07/2.2.00/07.0318 „Víceoborová inovace studia Matematické biologie“ a státním rozpočtem České republiky