MUNI | RECETOX

Analýza genomických a proteomických dat

- Mgr. Eva Budinská, PhD
- Integrativní bioinformatika a biostatistika, RECETOX, PřF
- eva.budinska@recetox.muni.cz

- Jaro 2023

Meta-analýza

MUNI | RECETOX

Supertabulka omicsových dat (studie-řádky, omicsové proměnné – sloupce)

		DNA			RNA			Protein		Phenotype		Environment	
		SNP	CNV, CGH	UHTS	mRNA	miRNA	SAGE	IHC	proteomics	clinical	Imaging, metabolomics, physiology	drug, therapy	pathogen, toxin
Study design 1 human breast cancer patients, retrospective, clinical outcome, drug	Study 1				■					■			
	Study 2				■	■				■		■	
	Study 3				■					■		■	
	Study 4				■					■			
	Study 5				■								
	Study 6		■		■					■			
	...				■					■			
Study design 2 experimental, time-series, tissue culture	Study a												
	Study b												
Study design 3 cancer cell lines ...	Study x	■	■	■	■	■	■	■	■		■	■	
	Study y						■						
	Study z						■						
	...					■							

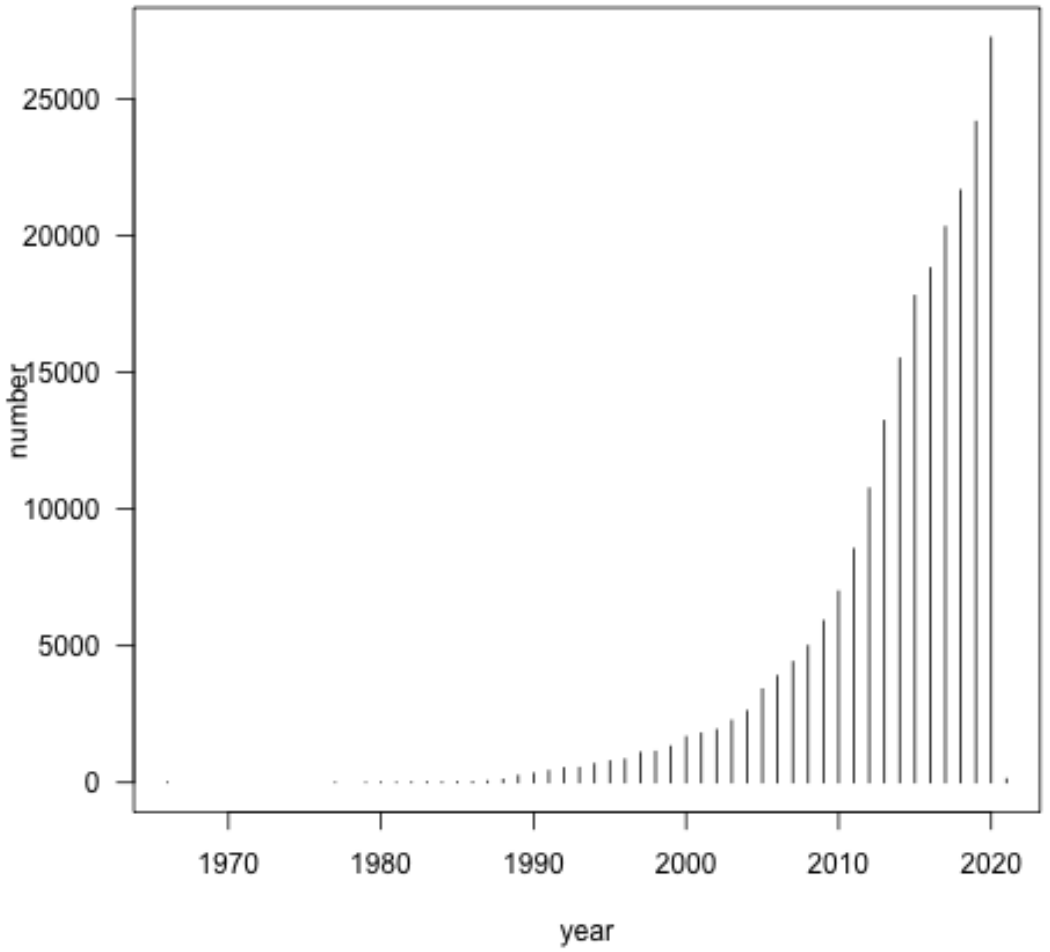
- „Horizontální“ integrace – stejné vzorky, různé typy dat (proměnné)
- „Vertikální“ integrace – stejné proměnné, různé studie (vzorky)

MUNI | RECETOX

Co je to meta-analýza?

- Statistická analýza, která kombinuje výsledky z více vědeckých studií
- Rozšířila se v 70. letech
- Stále více studií na stejné téma znamená více možností meta-analýzy

Počet nalezených publikací s klíčovým slovem “meta analysis” v databázi pubmed.



Motivace

- Nutnost publikovat data umožňuje jejich opětovnou analýzu

- Tato reanalýza dat umožňuje:
  - Kritickou recenzi původních výsledků
  - Potvrzení / validaci výsledků z jiných studií
  - Robustnější objevy založené na větší velikosti vzorku
  - Nové objevy ve větších oblastech / kontextech

Úvod do meta-analýzy: příkladová data

- US Berkeley – výsledky přijímacích zkoušek 1973*

	Muž	Žena	Celkem
Přijat(a)	1198	557	1755
Nepřijat(a)	1493	1278	2771
Celkem	2691	1835	4526

- Byli při přijímacím řízení favorizováni muži?

Na Kalifornské univerzitě v Berkeley byla provedena observační studie zaměřená na předpojatost při přijímání absolventů. Během studijního období požádalo o přijetí ke studiu 8 442 mužů a 4 321 žen. Bylo přijato asi 44% mužů a 35% žen.

Poměr šancí nepřijetí: $\frac{1278/557}{1493/1198} = 1.84,\ 95\%\ \text{CI: } [1.62,\ 2.09\ ]$

p-hodnota: 2.2×10^{-16}

*Bickel, Hammel, O'Connell (1975) Sex bias in graduate admissions: data from berkeley. Science 187:398-403

MUNI | RECETOX

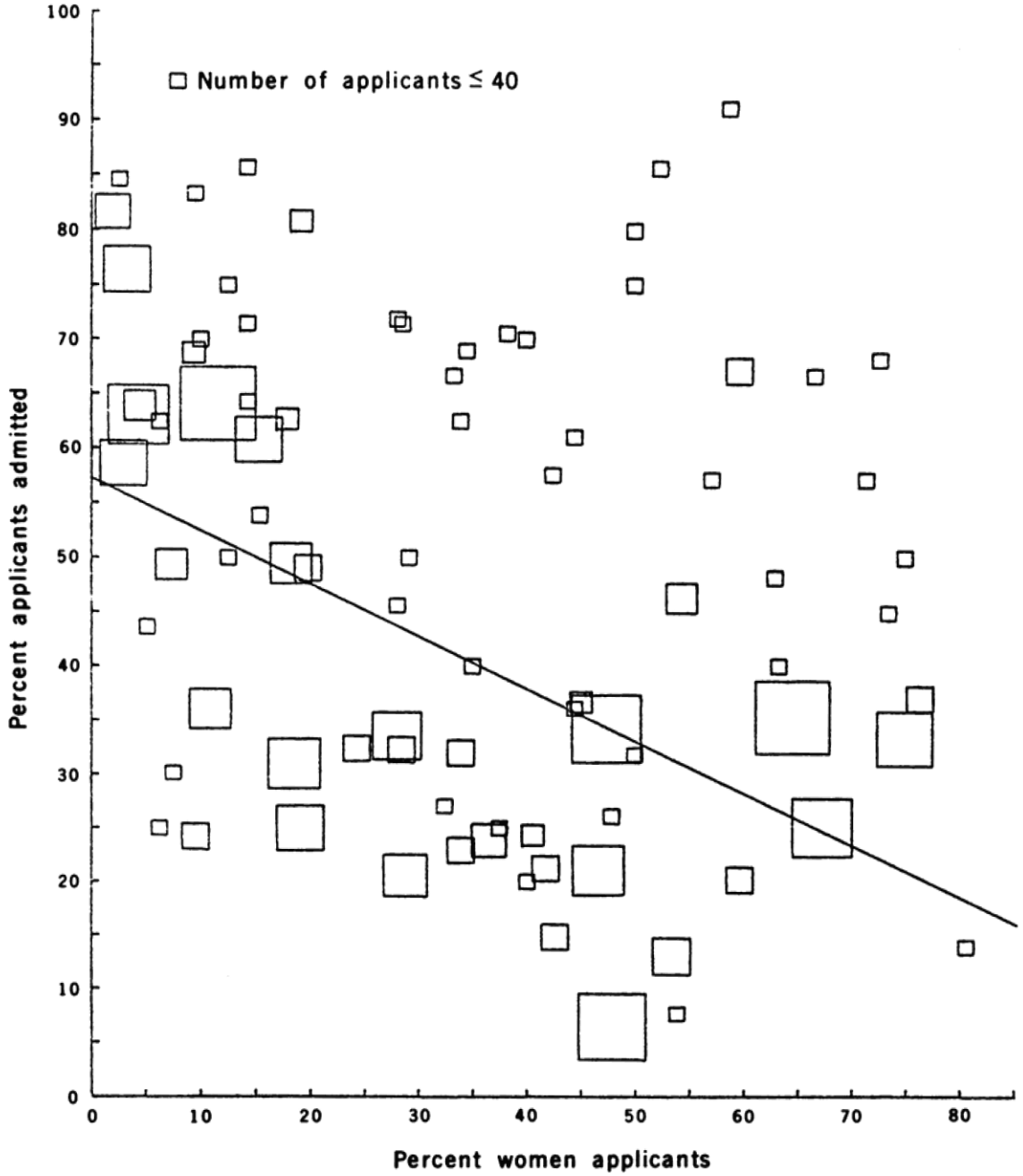


Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

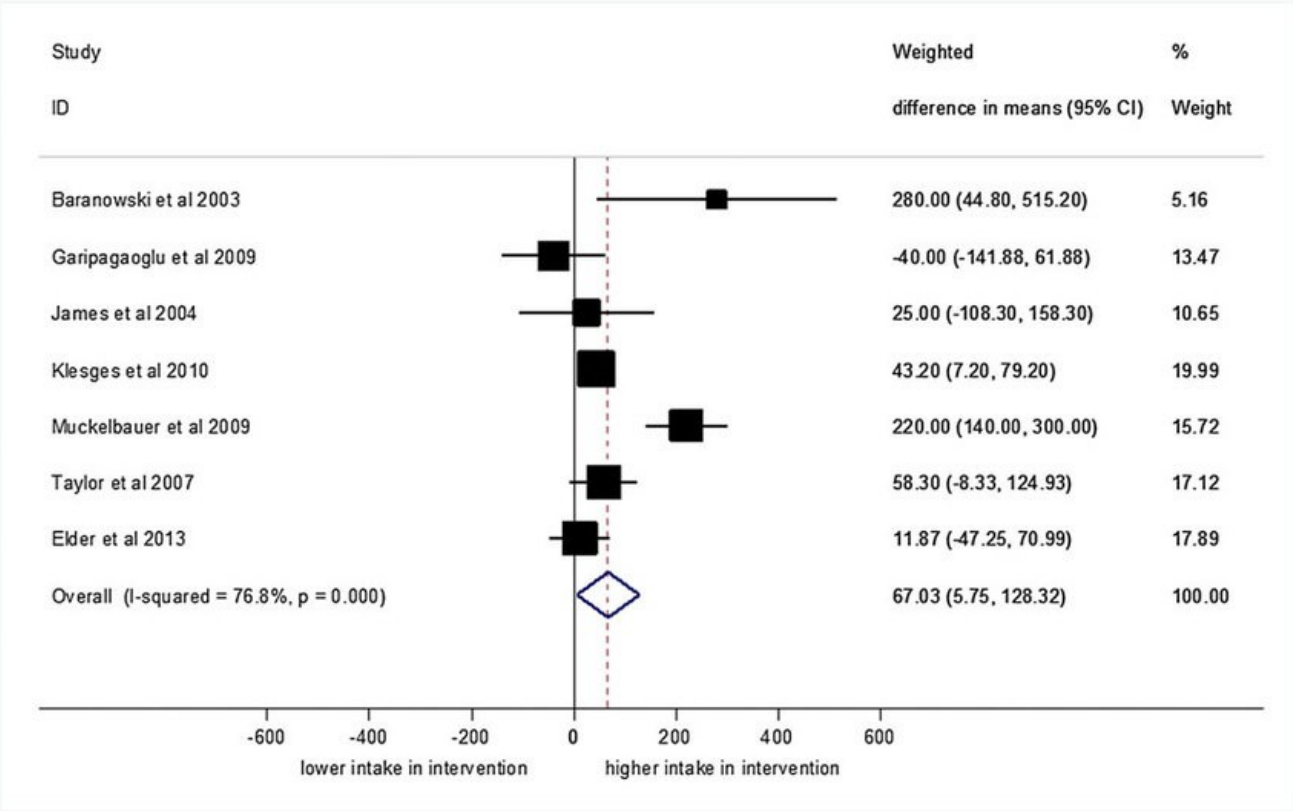
	Men		Women	
Dept	Numb. Applicants	Admitted	Numb. Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
Total	2,691	45%	1,835	30%

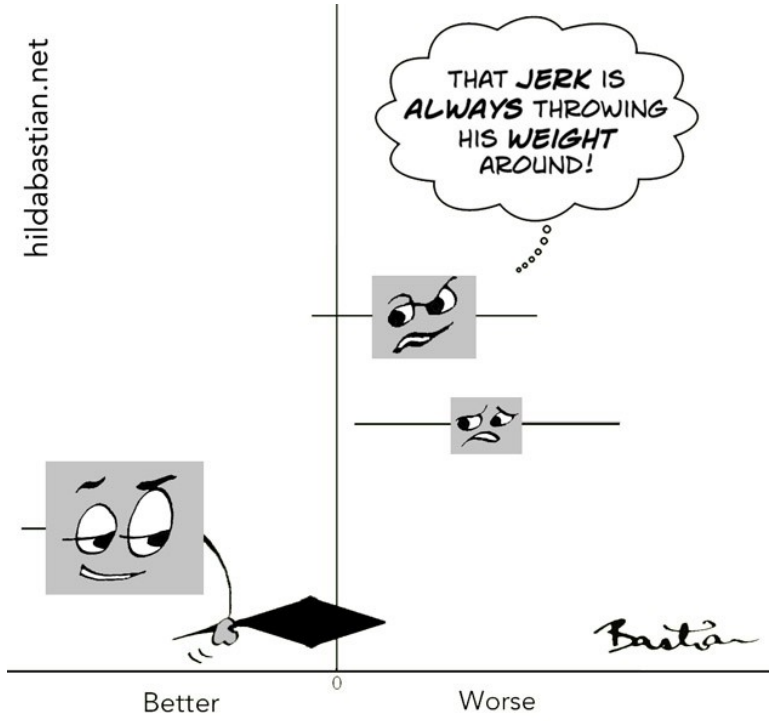
Do prvních fakult se dalo snadno dostat. Tam se hlásilo více než 50% mužů.
Do ostatních čtyř fakult bylo mnohem těžší se dostat. O tyto se zajímalo více než 90% žen.

MUNI | RECETOX

Forestplot

je grafické zobrazení odhadovaných výsledků z řady vědeckých studií zabývajících se stejnou otázkou spolu s celkovými výsledky (meta-analýza)



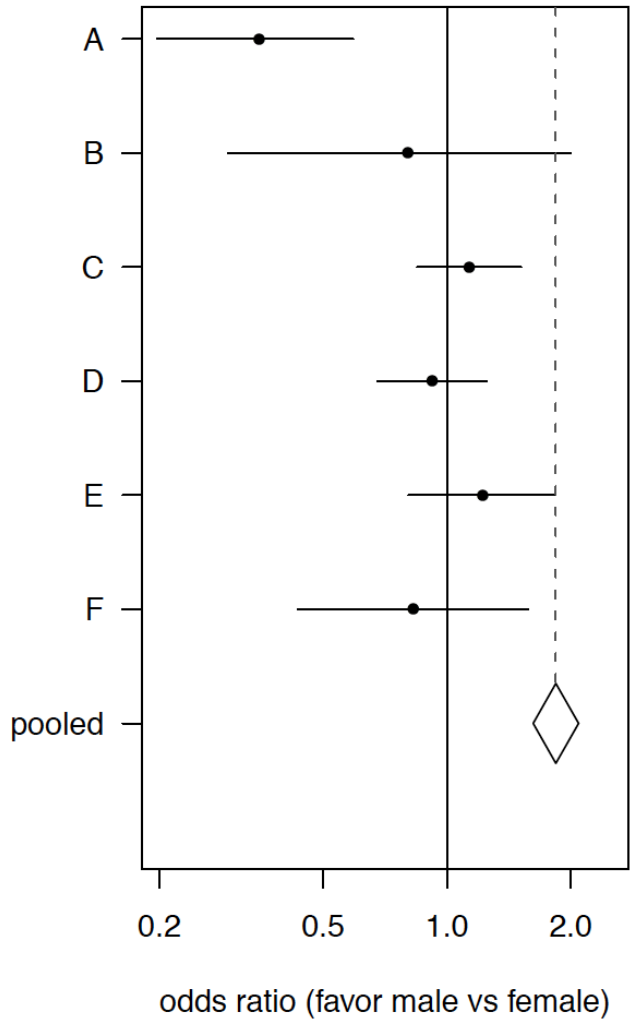


Velikost čtverce je proporční velikosti studie.

MUNI | RECETOX

Stratifikovaná analýza a forest plot

Dept.	data		odds ratio	95% C.I.	p-value
A	512	89	0.35	[0.20, 0.59]	10^{-5}
	313	19			
B	353	17	0.80	[0.30, 0.20]	0.68
	207	8			
C	129	202	1.13	[0.84, 1.52]	0.39
	205	391			
D	138	131	0.92	[0.68, 1.25]	0.60
	279	244			
E	53	94	1.22	[0.80, 1.83]	0.36
	138	299			
F	22	24	0.83	[0.43, 1.58]	0.55
	351	317			
pooled	1198	557	1.84	[1.62, 2.09]	10^{-16}
	1439	1278			

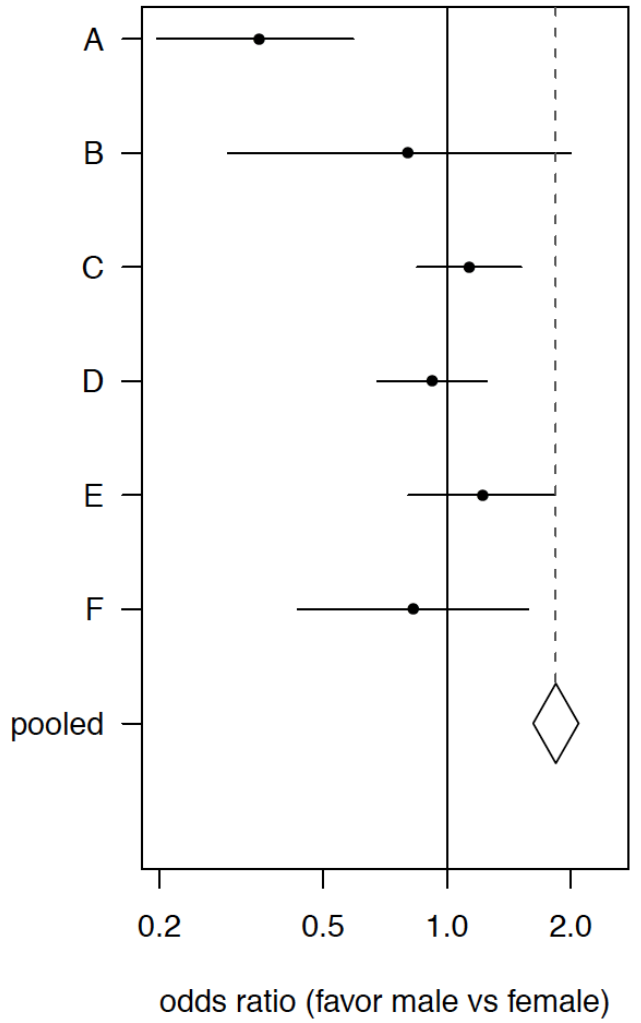


CETOX

Stratifikovaná analýza a forest plot

Dept.	data			odds ratio	95% C.I.	p-value
A	512	89		0.35	[0.20, 0.59]	10^{-5}
	313	19				
B	353	17		0.80	[0.30, 0.20]	0.68
	207	8				
C	129	202		1.13	[0.84, 1.52]	0.39
	205	391				
D	138	131		0.92	[0.68, 1.25]	0.60
	279	244				
E	53	94		1.22	[0.80, 1.83]	0.36
	138	299				
F	22	24		0.83	[0.43, 1.58]	0.55
	351	317				
pooled	1198	557		1.84	[1.62, 2.09]	10^{-16}
	1439	1278				

Efekt a interval spolehlivosti

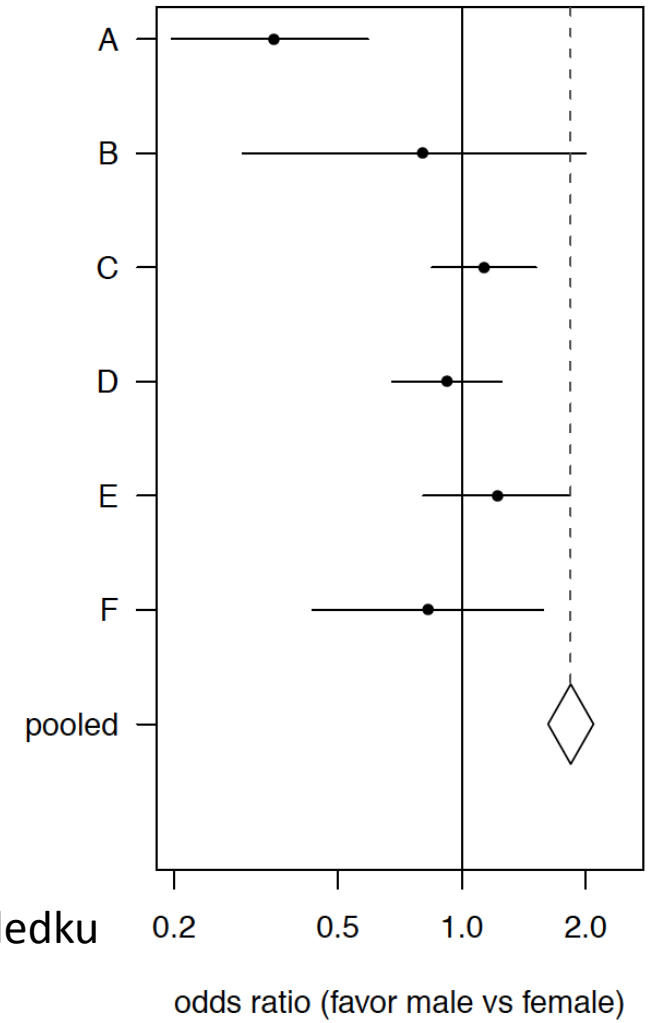


Souhrnný odhad

CETOX

Stratifikovaná analýza a tzv. forest plot

Dept.	data		odds ratio	95% C.I.	p-value
A	512	89	0.35	[0.20, 0.59]	10^{-5}
	313	19			
B	353	17	0.80	[0.30, 0.20]	0.68
	207	8			
C	129	202	1.13	[0.84, 1.52]	0.39
	205	391			
D	138	131	0.92	[0.68, 1.25]	0.60
	279	244			
E	53	94	1.22	[0.80, 1.83]	0.36
	138	299			
F	22	24	0.83	[0.43, 1.58]	0.55
	351	317			
pooled	1198	557	1.84	[1.62, 2.09]	10^{-16}
	1439	1278			



Simpsonův paradox: "celek je v rozporu s jeho částmi"
nebezpečí shromažďování údajů spočívá v zkreslení v důsledku skrytých faktorů

CETOX

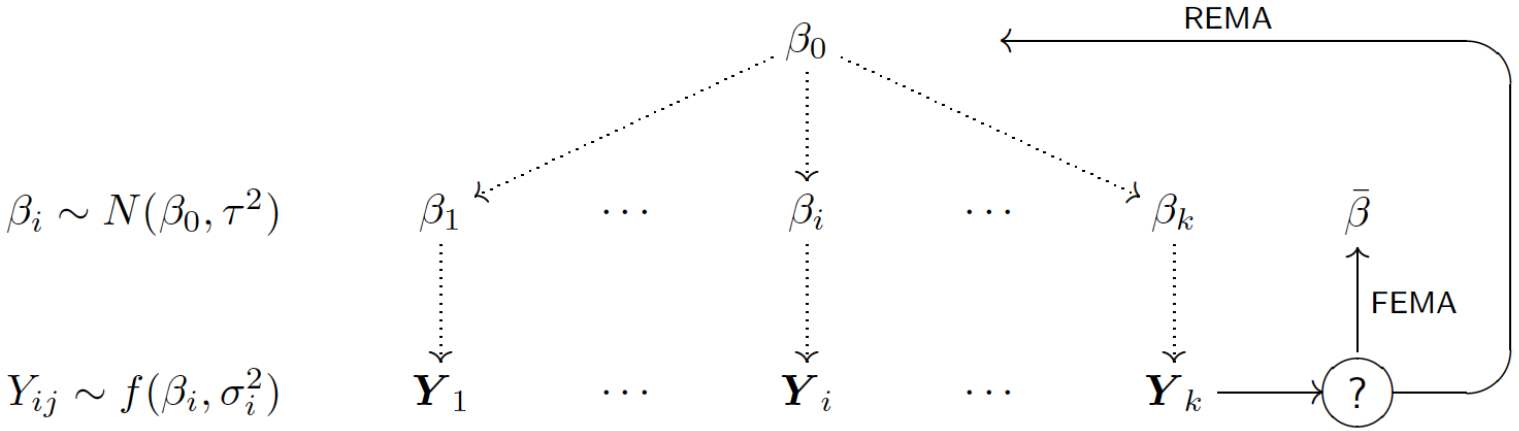
Zkreslení

„Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.“

Bickel, Hammel, O'Connell (1975) Sex bias in graduate admissions: data from berkeley. Science 187:398-403

Jak tedy udělat správně souhrn?

Meta analytické přístupy



- **Jedna studie:**
  - **Inference o β_i (β_0 + zkreslení studie: technické, designové, populace, ...)**
- **Modely s pevným efektem (FEMA):**
  - **Inference o $\bar{\beta} = \sum_i \beta_i / k$ (průměr konkrétních datových souborů k dispozici)**
  - **Interval spolehlivosti není ovlivňován variabilitou mezi studiemi (τ^2)**
- **Náhodné efekty / hierarchické modely (REMA):**
  - **Inference o β_0 (pravda; očekávání budoucích studií)**
  - **Interval spolehlivosti je úzký, je-li variabilita mezi studiemi τ^2 malá (a naopak)**

Meta-analýza

- Samostatná analýza každé studie (nebo její části)
- Průměr počítáme s použitím inverze rozptylu jako váhy:

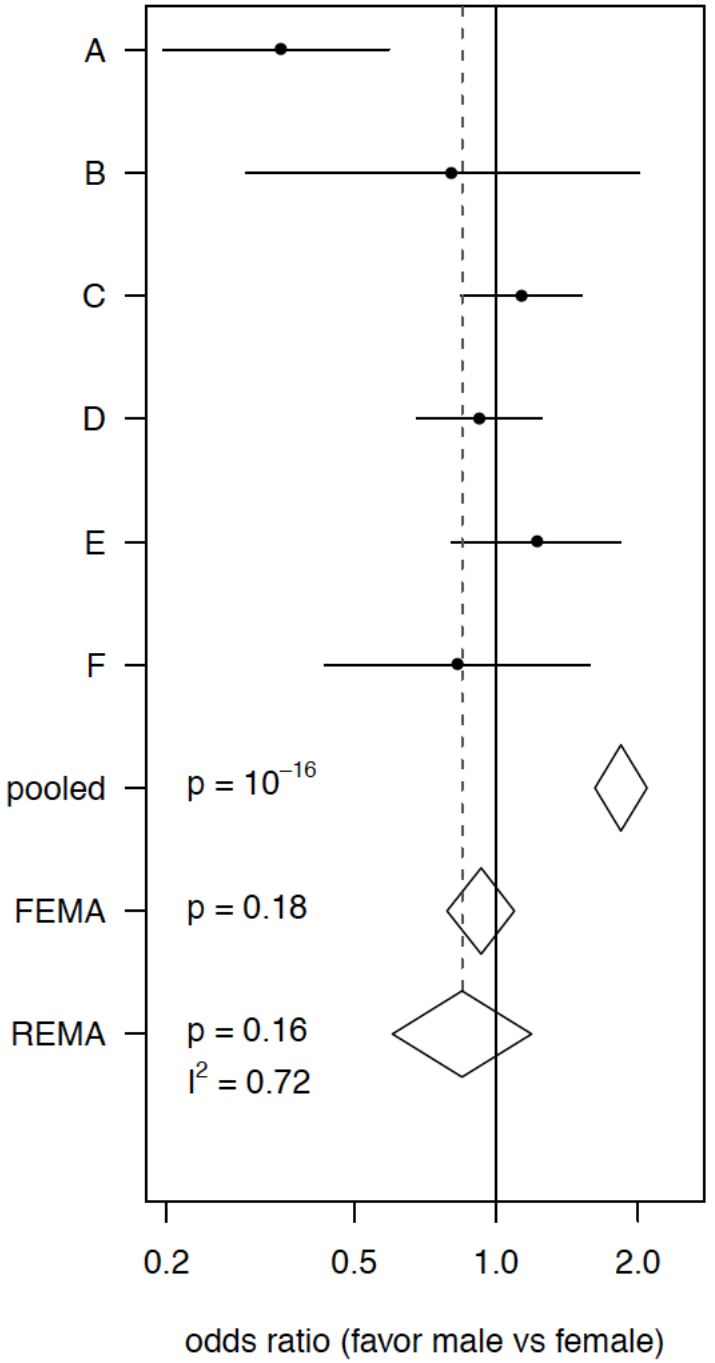
$$\hat{\beta}_0 = \frac{\sum_{i=1}^{k} \hat{\beta}_i / (\hat{\sigma}_i^2 + \hat{\tau}^2)}{\sum_{i=1}^{k} 1 / (\hat{\sigma}_i^2 + \hat{\tau}^2)}$$

β_i, β_0: effect size (per study and total)

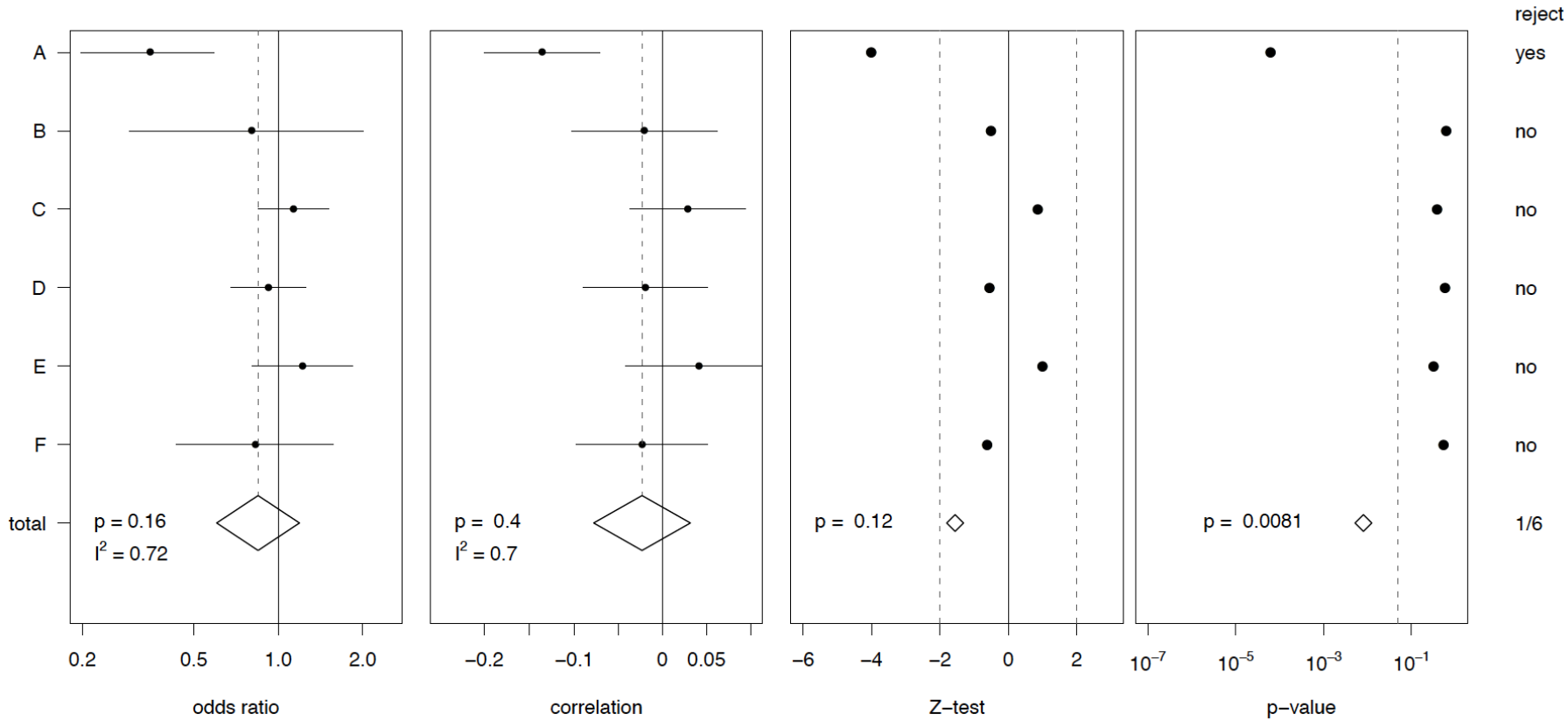
σ_i^2: within-study variance of β_i, i.e. $[\mathrm{SE}(\beta_i)]^2$

$\hat{\tau}^2$: between-study variance

- Je-li τ^2 nula (nemusí být realistické!) - jedná se o meta analýzu pevných efektů (FEMA)
- V opačném případě je τ^2 odhadnutá z dat, a jedná se o analýzu náhodných efektů (REMA) – předpokládáme, že mezi studiemi je náhodná variabilita
- I^2: podíl variability způsobený heterogenitou mezi studiemi



Které metriky porovnávat?



- Odds ratio (poměr šancí): regresní koeficient (jeho průměr při použití REMA)
- Korelace: míra závislosti nebo vzájemné informace (její průměr u REMA)
- Z-statistika: významnost (se znaménkem) => Stouerova metoda *akumulace*: $\sum Z/sqrt(k)$
- p-hodnoty: významnost (bez znaménka) => Fisherova metoda *akumulace*: $-2\sum \log p$
- Metoda součtu hlasů: počet zamítnutých hypotéz

MUNI | RECETOX

Jak kombinovat analýzy (data)

1. Kombinovat **nezpracovaná data**

(+) snadno použitelné (-) potenciální zkreslení, žádné posouzení heterogenity

2. Kombinovat **koeficienty** (změna násobku, riziko a poměry šancí, ...)

(+) fyzická interpretovatelnost (-) ovlivněná jednotkou měření

3. Kombinovat **korelaci** / závislost (R2, $\tanh^{-1}$ (r), ...).

(+) bez jednotky (-) ovlivněná vzorkováním / dizajnem

4. Kombinovat **významná měření** (t-test, Z-test, p-hodnota atd.)

(-) silný efekt + nízká síla = slabý efekt + vysoká síla

5. Kombinovat **rozhodnutí** (odmítnutí / přijetí hypotézy, seznamy genů)

(+) snadno použitelné (-) postrádá sílu

Jak na to?
1. krok: Správa datových souborů

- Připravte si přehled relevantních dostupných datových souborů
  - Prohledávejte literaturu, veřejné databáze a web
- Zajistěte jejich neprůsečnost
  - Reorganizujte soubory dat tak, aby neobsahovaly redundantní vzorky
- Nejednotné názvy a reprezentace proměnných
  - Přejmenujte a překódujte proměnné tak aby byly stejné napříč datovými soubory
- Zajistěte shodnou anotaci molekul
  - Přemapujte sondy (sady sond) napříč platformami, zarovnat k referenční sekvenci; redukovat na jednu sondu na gen (mikročipy)
  - Ujistěte se, že jste použili anotaci na stejný referenční genom
- Zkontrolujte kvalitu kvantitativních proměnných (např. genová exprese)
  - Zajistěte stejnou jednotku / transformaci; v případě potřeby přejmenujte a změňte měřítko

Jak na to?
2. krok: Analýza

- Jak provést kombinovanou analýzu heterogenních datových souborů?
  - Rozdíly v dizajnech studií, populacích a kritériích pro výběr vzorků
  - Nesrovnatelné kvantitativní údaje; systematické chyby měření
- Jak vytvořit celkové výsledky na základě všech datových sad?
- Jak posoudit a začlenit heterogenitu?
- Jak vizualizovat a prezentovat výsledky analýzy?
- Jak analýzu přizpůsobit omics datům?
- Jak přistoupit ke komplexní analýze, jako je například hierarchické shlukování a predikce?

Na co si dávat při meta-analýze pozor

Jsou studie opravdu porovnatelné?

- Reprezentují stejné populace?
- Jsou provedená měření porovnatelná?
- …

Závěrem

- Více omicsových datových souborů lze společně analyzovat v rámci „standardních“ statistických metod (např. zobecněné lineární modely, metaanalýzy, hierarchické vzorkovací modely).
- Rozšíření na komplexní analýzu (např. predikce, shluková analýza) je možné začleněním REMA pro kombinování sumárních statistik ve vhodné fázi analýzy.