Analýza dat v neurologii

XLV. Grafy usnadňující studium zavádějících faktorů v asociačních studiích – III. Spojitá data

L. Dušek, T. Pavlík, J. Jarkovský, J. Koptíková
Institut biostatistiky a analýz Masarykova univerzita, Brno

✉ doc. RNDr. Ladislav Dušek, Dr.
Institut biostatistiky a analýz MU, Brno
e-mail: dusek@iba.muni.cz

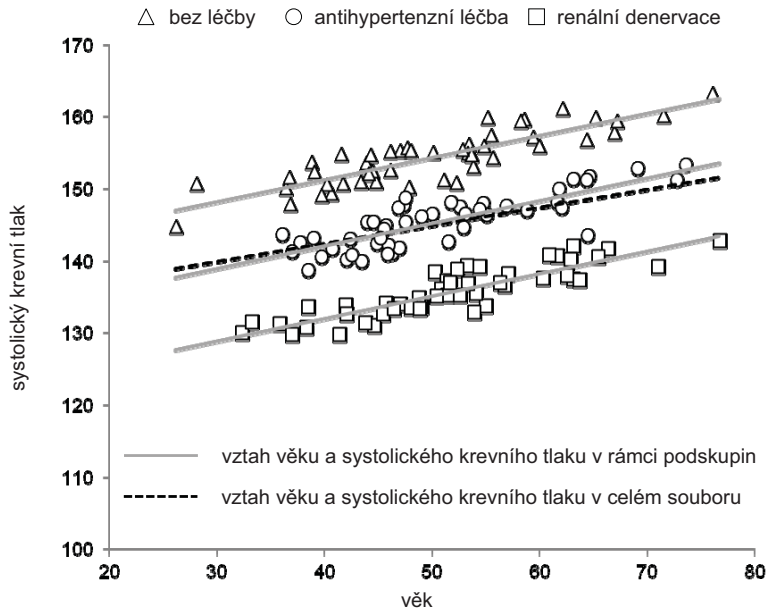
Předchozí díly seriálu se zabývaly analýzou vztahu „expozice-účinek" v asociačních studiích, přičemž všechny využité příklady pracovaly s dichotomickými (binárními) proměnnými. Tedy jak expozice, tak zkoumaný jev jako účinek expozice jsou zaznamenávány binárním způsobem „ano/ne". Tento typ dat nás ostatně provází již od výkladu kontingenčních tabulek a je základem pro analýzy vedoucí k odhadům poměru šancí nebo relativního rizika. Rozšiřme nyní výklad krátkou odbočkou pro spojité proměnné, kdy expozice (*X*) a účinek (*Y*) jsou měřeny jako kvantitativní znaky na spojité škále. Příkladem může být kvantitativní dávkování léku a výsledná hodnota krevního tlaku.

V tomto díle se pokusíme na jednoduchých příkladech ukázat, že práce se spojitými proměnnými se principiálně nijak neliší od analýzy tabulek četností. I zde kvantifikujeme vztah mezi expozicí a účinkem, resp. sledovaným jevem. U kvantitativních proměnných musíme rovněž pozorně kontrolovat potenciální zavádějící faktory a možná rizika zkreslení výsledku (viz díly XLII–XLIII seriálu).

Práce se spojitými daty je samozřejmě neméně běžná a vztah (asociace) mezi kvantitativními škálami je běžně měřen pomocí tzv. korelace. Analýzou korelace se budeme zabývat v některém z následujících dílů seriálu, zde výklad omezíme na následující otázky:

- Mohou i ve vztahu dvou spojitých proměnných působit zavádějící faktory, tak jako v tabulce četností 2 × 2?
- Můžeme ve vztahu spojitých znaků a zavádějících faktorů pozorovat slo-

V retrospektivní nemocniční studii zkoumáme vztah systolického krevního tlaku a věku pacientů s hypertenzí, přičemž možným zavádějícím faktorem je způsob léčby hypertenze. Cílem analýzy je zjistit, zda jde o faktor významně modifikující zjištěný vztah mezi věkem a systolickým krevním tlakem.

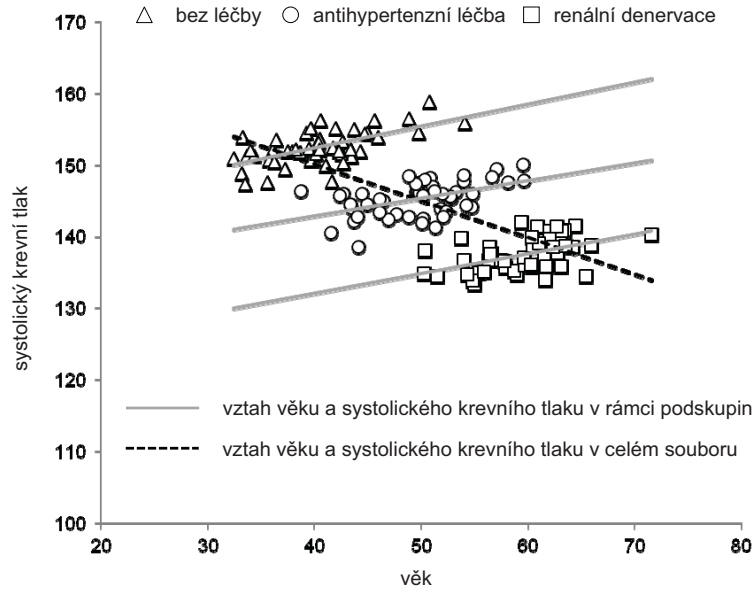


Interpretace: Graf zobrazuje vztah mezi spojitými proměnnými jednak celkově, jednak v rámci kategorií možného zavádějícího faktoru.

Vzhledem k rovnoměrné distribuci věkových kategorií v podskupinách pozorujeme konzistentní výstup dílčích analýz i analýzy celkového souboru. Hodnota krevního tlaku s věkem narůstá bez ohledu na typ aplikované léčby.

Příklad 1. Ukázka využití grafu pro analýzu vztahu spojitých proměnných s možným vlivem zavádějícího faktoru.

V retrospektivní nemocniční studii zkoumáme vztah systolického krevního tlaku a věku pacientů s hypertenzí, přičemž možným zavádějícím faktorem je nasazený způsob léčby hypertenze. Cílem analýzy je zjistit, zda jde o faktor významně modifikující zjištěný vztah mezi věkem a systolickým krevním tlakem.



Interpretace: Graf zobrazuje vztah mezi spojitými proměnnými jednak celkově, jednak v rámci kategorií možného zavádějícího faktoru.

Příklad je typickou ukázkou Simpsonova paradoxu, analýza dílčích podsouborů vede k opačnému výsledku než spojená data celého souboru. Důvodem je viditelný a významný rozdíl ve věku pacientů v různě léčených podskupinách. V důsledku této heterogenity je analýza spojených dat silně zkreslena a nevede k objektivním a reprezentativním závěrům.

Příklad 2. Ukázka využití grafu pro analýzu vztahu spojitých proměnných při významné interakci se zavádějícím faktorem.

žité interakce vedoucí až k Simpsonovu paradoxu?

Odpověď na obě otázky je kladná, což znamená, že i při interpretaci vztahů kvantitativních znaků musíme být velmi opatrní a kontrolovat rizika zkreslení. Předpokládejme, že studujeme vztah spojitých proměnných *X* a *Y* při existenci zavádějícího faktoru *Z*, který je kategoriální a stratifikuje celkový soubor dat do čtyř podsouborů. V praxi mohou nastat tři níže uvedené možnosti:

- faktor *Z* zkoumaný vztah *X* a *Y* nijak neovlivňuje a výstup analýzy celkového souboru je stejný jako u všech čtyř podsouborů, např. s rostoucí hodnotou *X* roste hodnota *Y*,
- faktor *Z* interaguje se vztahem *X* a *Y*, kdy v některých podsouborech vytvořených podle hodnoty *Z* existuje silný vztah mezi *X* a *Y*, zatímco v jiných nikoli; na celkovém souboru bez uvažování hodnot *Z* se žádný vztah projevit nemusí, neboť heterogenita dílčích podskupin anuluje statistickou významnost vztahu,
- faktor *Z* interaguje se vztahem *X* a *Y* a generuje Simpsonův paradox, tedy na celkovém souboru např. platí, že čím větší je *X*, tím menší je *Y*, ale v dílčích podsouborech pozorujeme vztah opačný.

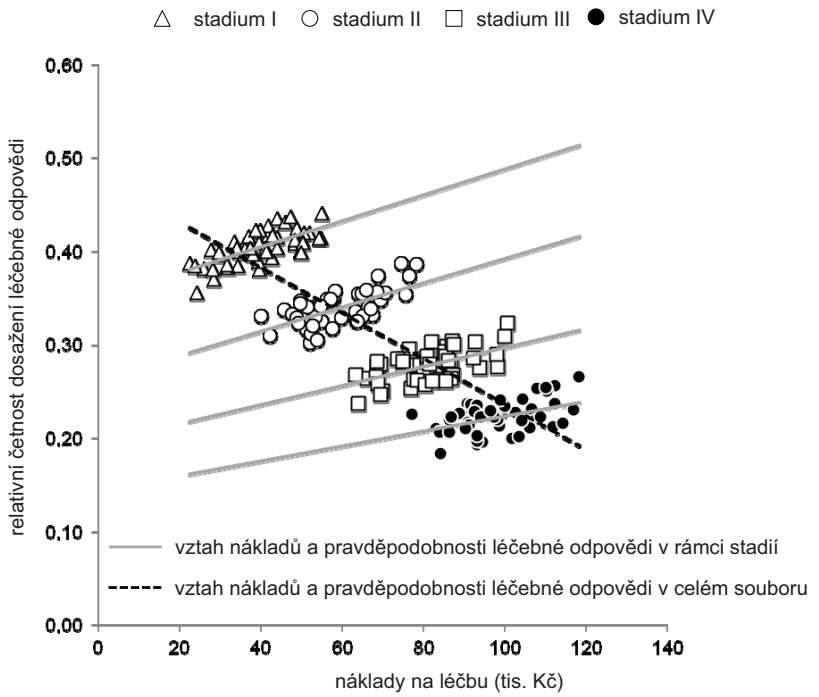
Je zřejmé, že zejména poslední případ by vedl k závažným zkreslením, pokud bychom vliv *Z* v analýze neuvažovali, neboť faktor *Z* je zdrojem velmi výrazné heterogenity hodnot v souboru. V této pozici velmi často vystupují zásadní prognostické nebo prediktivní markery, jejichž vypuštění z analýzy je samozřejmě nepřípustné. Všechny výše popsané varianty výstupů lze doložit pomocí relativně jednoduchých grafů, aniž použijeme matematické vztahy a výpočty. Jednoduchý bodový graf (*scatterplot*) vykreslující vztah spojitých proměnných *X* a *Y* nalezneme ve všech typech tabulkových nebo grafických softwarových nástrojů. Uvažujeme-li vliv faktoru *Z*, pak vykreslíme tzv. stratifikovaný bodový graf (*stratified scatterplot, grouped scatterplot*), který odlišuje jednotlivé úrovně hodnot *Z*. Ukázkou takového zobrazení je příklad 1.

Příklad 1 zobrazuje vzájemný vztah věku pacientů s hypertenzí a hodnot systolického krevního tlaku (TKs), přičemž soubor je stratifikován podle typu léčby hypertenze. Způsob léčby by mohl pozorovaný vztah ovlivnit, avšak naměřená data ukazují opak. Hodnota TKs roste s věkem pacientů jak na celkovém souboru, tak v dílčích podsouborech. V grafu také vidíme, že všechny relevantní věkové kategorie pacientů jsou přibližně rovnoměrně zastoupeny ve všech třech různě léčených skupinách pacientů, takže nevzniká prostor pro zkreslení v důsledku nerovnoměrné distribuce kategorií věku.

Příklad 2 je ukázkou opačné situace. Zatímco v rámci různě léčených skupin pacientů roste krevní tlak s věkem, na celkovém souboru bez uvažování typu léčby bychom dostali výstup opačný, tedy čím vyšší věk, tím nižší krevní tlak. Jde o ukázku Simpsonova paradoxu na spojitých datech, přičemž jednoduchý bodový graf zároveň identifikuje i příčinu této extrémní situace. Analýza spojených dat je nutně zkreslená, neboť různě léčené skupiny pacientů nejsou vzájemně srovnatelné věkem. Nerovnoměrná distribuce věku v podskupinách nejen zkresluje srovnání, ale také brání spojení podskupin.

Je zřejmé, že identifikace vlivu zavádějícího faktoru v korelačních grafech není nijak složitá a jistě ji zvládne i laik v oblasti statistiky. Přesto jsme v praxi velmi často svědky zcela chybných závěrů způsobe-

Ve farmakoekonomické analýze zkoumáme vztah mezi náklady na léčbu a pravděpodobností dosažení léčebné odpovědi, přičemž možným zavádějícím faktorem je klinické stadium onemocnění (pokročilost nemoci). Cílem analýzy je zjistit, zda pokročilost nemoci neovlivňuje výslednou analýzu modifikací vztahu mezi náklady a četností dosažené léčebné odpovědi.



Interpretace: Graf zobrazuje vztah mezi spojitými proměnnými jednak celkově, jednak v rámci kategorií možného zavádějícího faktoru.

Je zřejmé, že data analyzovaná v rámci klinických stadií onemocnění konzistentně indikují pozitivní vztah mezi náklady investovanými do léčby a relativní četností dosažení léčebné odpovědi. Avšak analyzujeme-li soubor jako celek bez ohledu na stav choroby, dostáváme vztah opačný, kdy vyšší ekonomická investice indikuje horší výsledky léčby.

Příklad je typickou ukázkou Simpsonova paradoxu, kterému se v této situaci téměř nelze vyhnout. Klinická stadia nemoci jsou totiž objektivně významně rozdílná nejen v ceně léčby, ale i v dosažitelném léčebném úspěchu. Pokročilost nemoci je významný prognostický faktor. Jelikož nemůžeme dosáhnout vzájemné srovnatelnosti klinických stadií v úspěšnosti léčby a zároveň v ceně (nejde o rozdíly v důsledku např. špatného výběru vzorku), nelze tento typ analýzy provést bez znalosti klinického stadia a jeho vlivu. Analýza celkového souboru nemá smysl.

Příklad 3. Simpsonův paradox na příkladu korelace spojitých proměnných.

ných ignorováním zavádějících faktorů v korelačních analýzách. Ukázku takového výstupu dokumentuje příklad 3, který sice pracuje s vymyšlenými daty, ale je inspirován skutečnou událostí z České republiky. Farmakoekonomická analýza zde usiluje o prokázání vztahu mezi náklady na léčbu určitého onemocnění a výsledkem léčby, tedy její účinností. Analýza celkového souboru vede k překvapivému zjištění, že čím větší jsou náklady, tím horší je dosažený výsledek, což může v konečném důsledku vést ke kritice daného segmentu léčby nebo dokonce k jeho omezování. Při detailnějším rozboru dat ale zjišťujeme, že náklady jsou v silném vztahu s pokročilostí choroby; čím později je nemoc zachycena, tím je léčba dražší a tím menší má pacient šanci na dobrý výsledek. Pokud tedy zkoumáme vztah nákladů a výsledku léčby v rámci jednotlivých stadií choroby, zjišťujeme ve všech kategoriích silně pozitivní vztah, tudíž vyšší investované náklady vedou k lepšímu výsledku. To je ale patrné pouze v rámci stadií choroby, které se vzájemně v dosažitelném výsledku léčby velmi významně liší. Pokročilost (stadium) choroby zde vystupuje jako významný zavádějící faktor a výstup popsaný v příkladu 3 je typická ukázka Simpsonova paradoxu. V rámci dílčích podsouborů platí, že vyšší náklady znamenají lepší výsledek, avšak spojený soubor vede k opačnému vztahu v důsledku silného prognostického významu stadia choroby. Výsledek zjištěný na spojených datech je tak nutně zavádějící, bez stratifikace souboru na stadia choroby nelze podobnou analýzu odpovědně provádět.

Představme si nyní, že by někdo použil vztah mezi náklady na léčbu a dosaženým výsledkem jako indikátor kvality zdravotnických zařízení a přitom by neuvažoval o spektru jimi léčených pacientů („*case mix*"). Rozdíly mezi jednotlivými zařízeními v zastoupení léčených stadií choroby by tak „zezadu" analýzu zcela zkreslily a její výstupy by byly nebezpečně zavádějící. Podobně jako stadium choroby ovšem může výsledný vztah ovlivňovat řada jiných faktorů, např. věk pacientů, jejich anamnéza, způsob příjmu do nemocnice, komorbidity apod. Odpovědný analytik musí prověřit všechny potenciální zavádějící faktory, o kterých ví a které má v rámci studie k dispozici. Zejména je nutné na tomto rozboru trvat, má-li být měřená asociace využita k závažným rozhodnutím, jako je třeba skórování poskytovatelů péče, změna léčebných postupů apod.

V tomto a v předchozích dílech seriálu jsme detailně probrali problematiku identifikace těchto faktorů, určení a kvantifikace jejich vlivu a zejména interpretaci výsledků. Doufejme, že jsme čtenáře přesvědčili o nutnosti detailního popisu klinických dat, a to u všech typů klinických studií. Pouze komplexní panel demografických, sociálních a klinických deskriptorů umožní objektivně posoudit potenciální vliv těchto faktorů (v literatuře někdy na-

zývaných *„baseline variables“*) a zabránit zkreslení zkoumaných vztahů. Jako zavádějící faktor mohou působit všechny typy znaků od nominálních, přes kategoriální, ordinální až po znaky spojité. Zavádějících faktorů ovlivňujících vztah „expozice-účinek“ („expozice-následek“) může být současně více a ve svých účincích se mohou vzájemně stimulovat nebo inhibovat. Vzájemné interakce mohou být velmi komplikované a jejich rozkrytí často vede i k novým poznatkům. V následujícím přehledu shrnujeme hlavní pojmy s touto problematikou spojené.

Typický zavádějící faktor v asociační studii je v přímém nebo nepřímém vztahu k sledovanému účinku (následku expozice) a může být rovněž ve vztahu se studovanou expozicí (např. léčbou). Jako pozitivní zavádějící faktor označujeme znak, jehož vazba k expozici a následku má stejný „směr“, je tedy buď na obě strany pozitivní, nebo negativní. Takové zavádějící faktory vedou typicky k nadhodnocení rizika plynoucího z expozice. Negativní zavádějící faktor má vůči expozici opačný vztah než k jejímu následku; v důsledku toho může v asociační analýze riziko spjaté s expozicí podhodnocovat a maskovat. Podle vztahu ke zkoumanému riziku označujeme zavádějící faktory jako protektivní nebo rizikové.

Typický zavádějící faktor nazýváme též zkreslující (rušící) faktor, neboť zkresluje vztah „expozice-účinek“, zejména pokud je jeho výskyt (distribuce) odlišný ve skupinách vytvořených podle typu expozice. Pozorujeme tak zavádějící efekt, který chceme identifikovat a závěry studie od něj očistit. Typický zavádějící faktor tedy závěry studie zkresluje, ale z definice nevstupuje do přímé interakce s vlivem expozice; nestojí tudíž mezi expozicí a účinkem, ani jinak neovlivňuje účinek expozice. Vliv zkreslujícího faktoru se snažíme eliminovat pomocí adjustace, stratifikace souboru podle úrovní faktoru nebo zajištěním vyváženého výskytu faktoru ve srovnávaných skupinách (*„matching“*).

Pokud faktor ovlivňuje expozici v jejím účinku, tedy pokud je efekt expozice různý pro různé úrovně faktoru, hovoříme o faktoru modifikujícím účinek. Identifikace tohoto jevu většinou zásadně mění výstup studie, neboť asociaci „expozice-účinek“ nejde uspokojivě vysvětlit bez uvažování vlivu modifikujícího faktoru. Zavádějící neboli doprovodný faktor se tak posunul do pozice faktoru vysvětlujícího, který může mít s účinkem expozice i příčinnou souvislost. Takový příčinný vztah často odráží i určitý biologický mechanizmus účinku, např. vztah mezi nadměrnou konzumací alkoholu a onemocněním jater. Vysoká konzumace alkoholu pak může potencovat efekt další expozice nebo ovlivňovat účinek léčby.

V klinických studiích rovněž často pracujeme s řadou obecných deskriptorů onemocnění nebo pacienta, které příčinný vztah k účinku expozice nemají, ale s vysvětlujícími faktory souvisí. Např. nadměrný alkoholizmus je často asociován s určitými sociálními charakteristikami jedince (nízká úroveň vzdělání, nezaměstnanost, kriminalita…). I tyto proměnné potom mohou vykazovat silný vztah k následku expozice, resp. k studovanému onemocnění, ačkoli nejsou nijak spojeny s biologickým mechanizmem účinku. Použijeme-li je např. jako nepřímé prediktory onemocnění, hovoříme o tzv. zástupných faktorech (*„surrogate factor“*, *„proxy factor“*). Zástupné faktory mohou být úspěšně využity jako „náhrada“, pokud kauzální prediktory nemoci či jiných rizikových jevů nejsou z objektivních důvodů k dispozici.

Je evidentní, že analýzy asociačních studií, identifikace a kvantifikace vlivu zavádějících faktorů navozují řadu situací, kdy není možné data interpretovat bez podrobné znalosti podstaty problému. Účast odborníka orientovaného ve studované problematice je nezbytná již při plánovaní studie, neboť prvním krokem k úspěchu je nezanedbat žádný s deskriptorů, který by mohl výsledky následných analýz zkreslit. Dále velmi záleží na postavení, které jednotlivé faktory dostanou ve struktuře dat a ve kterém bude analyzován jejich vliv. V tomto a předchozích dílech seriálu jsme probrali hlavní metodické postupy řešení těchto problémů, i když jsme téma samozřejmě zcela nevyčerpali.

Význam této problematiky dokládá i to, že je po více než 60 letech od vydání původní Simpsonovy práce v literatuře stále živá a inspiruje výzkum, který metodiku odhalování vlivu zavádějících faktorů neustále rozvíjí a obohacuje. Čtenářům doporučujeme práci kolektivu Hernán et al z roku 2011, v níž jsou velmi srozumitelnou formou probírány různé pozice zavádějících faktorů od zkreslení výsledků studie, přes kolizi s účinkem expozice až po prognostický význam. Obdobně Heydtmann (2002) a Bandyoapdhyay et al (2011) uvádějí instruktivní a praktické příklady zkreslení typu Simpsonova paradoxu. Nejnověji Baker (2013) rozvíjí grafické možnosti identifikace různých vlivů zavádějících faktorů, včetně faktorů, které nejsou v designu studie zahrnuty a pozorovány. K této poslední práci se v některém z dalších dílů seriálu vrátíme při rozboru zdrojů zkreslení v metaanalýzách.

Literatura

Baker SG. Causal inference, probability theory and graphical insights. Statist Med 2013; 32: 4319–4330. doi: 10.1002/sim.5828.

Bandyoapdhyay PS, Nelson D, Greenwood M, Brittan G, Berwald J. The logic of Simpson's paradox. Synthese 2011; 181(2): 185–208.

Hernán MA, Clayton D, Keiding N. The Simpson's paradox unraveled. Int J Epidem 2011; 40(3): 780–785. doi: 10.1093/ije/dyr041.

Heydtmann M. The nature of truth: Simpson's paradox and the limits of statistical data. Q J Med 2002; 95(4): 247–249.

Simpson EH. The interpretation of interaction in contingency tables. J Roy Stat Soc B 1951; 13: 238–241.