Analýza dat v neurologii

XLII. Simpsonův paradox a faktory modifikující účinek v analýze asociačních studií

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@cba.muni.cz

V předchozím díle seriálu jsme popsali velmi zajímavý jev, který dovede překvapit nejednoho analytika dat. Tzv. Simpsonův paradox označuje situaci, kdy výsledky analýz dílčích či menších datových souborů naznačují určitý efekt, ale po jejich spojení do většího souboru dat dostaneme výsledek opačný. Simpsonův paradox tak můžeme například pozorovat v situacích, kdy za nějakým účelem spojujeme dílčí soubory dat. Vždy jde o situaci vyžadující další rozbor a nelze ji přejít bez zvýšené pozornosti. Pokud totiž v konkrétním případě pozorujeme Simpsonův paradox, ukazuje to buď na vážné problémy s designem studie, nebo na vliv silného zavádějícího faktoru, který ovlivňuje výsledky dílčích sledování, a při spojení dat je jeho vliv maskován. Takové výsledky je nutné interpretovat s maximální opatrností, neboť existuje velké riziko jejich zkreslení.

Připomeňme si výklad předchozího dílu na dvou odlišných příkladech uvedených v tab. 1. Z jejich výsledku je bohužel patrné, že u Simpsonova paradoxu nelze paušálně prohlásit, že správný výsledek poskytují vždy dílčí pozorování anebo spojená data. Záleží totiž na typu řešeného problému a na příčinách zkreslení. Pokud jde o spojení dílčích souborů o malém počtu pozorování, pak relevantní výsledek poskytnou spíše spojená data. Avšak nerespektuje-li spojení dat důležité klasifikační (stratifikační) faktory (např. členění dle pohlaví, věku, stadia nemoci apod.), pak spojená data nemohou reprezentovat a relevantně odrážet skutečné vztahy mezi sledovanými veličinami.

Nyní přistoupíme k vysvětlení významu interakce zavádějících a expozičních faktorů v asociačních studiích. I zde se vliv zavádějícího faktoru projevuje jako rozpor mezi dílčími pozorováními a výsledkem získaným na celkovém souboru. Při-

Tab. 1. Příklady studií, ve kterých srovnání na dílčích podskupinách vycházejí protichůdně od celkového souboru.

Příklad 1.

Ve studii zaměřené na bezpečnost léčby sledujeme výskyt komplikací během hospitalizace u dvou léčebných postupů, chirurgického (**CH**) a konzervativního (**K**). Jde o vzácnou chorobu a studie je podle standardizovaného protokolu provedena ve dvou plně srovnatelných centrech.

Centrum 1	**CH:** 10 komplikací z 30 subjektů = 33,3 % **K:** 2 komplikace z 8 subjektů = 25,0 %	Protichůdný výsledek dílčích srovnání a celkového souboru je dán malým počtem pozorování v dílčích podsouborech – spojená data dílčí výsledky převáží. Spolehlivější výsledek zde poskytuje spojený soubor dat, dílčí sledování nejsou reprezentativní. Méně komplikací je tedy spojeno s chirurgickým řešením problému.
Centrum 2	**CH:** 7 komplikací z 9 subjektů = 77,8 % **K:** 15 komplikací z 24 subjektů = 62,2 %	
Celkem	**CH:** 17 komplikací z 39 subjektů = 43,6 % **K:** 17 komplikací z 32 subjektů = 53,1 %	

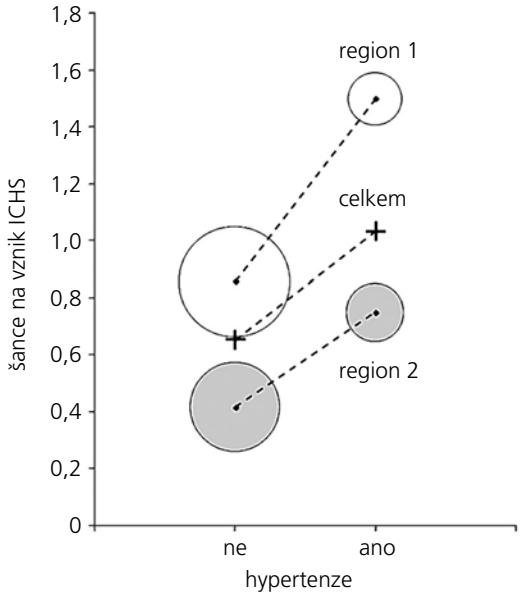
Příklad 2.

Srovnáváme hospitalizační mortalitu u pacientů se zhoubným nádorem slinivky břišní ve dvou centrech. Zásadním stratifikačním faktorem je zde klinické stadium onemocnění, kde uvažujeme dvě kategorie: nepokročilé (**N**) a pokročilé (**P**) onemocnění.

Centrum 1	**N:** 10 úmrtí z 60 subjektů = 16,7 % **P:** 90 úmrtí ze 190 subjektů = 47,3 %	Stratifikovaná a nestratifikovaná analýza vyšly protichůdně, což potvrzuje zásadní význam klinického stadia onemocnění jako zavádějícího faktoru. Analýza spojeného souboru neuvažující velmi rozdílnou zátěž jednotlivých center je zkreslená. Větší hospitalizační mortalitu vykazuje centrum 2, a to pro obě kategorie pokročilosti choroby.
Centrum 2	**N:** 30 úmrtí ze 170 subjektů = 17,6 % **P:** 45 úmrtí z 80 subjektů = 56,3 %	
Celkem	**Centrum 1:** 100 úmrtí z 250 subjektů = 40,0% **Centrum 2:** 75 úmrtí z 250 subjektů = 30,0 %	

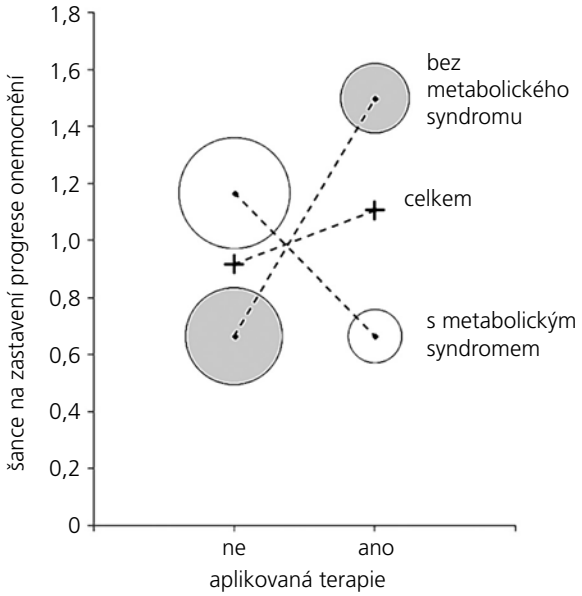
Graf pro vizualizaci vlivu interakce dvou faktorů (jeden z nich považujeme pro účely hodnocení za analyzovaný faktor, druhý za zavádějící) na hodnocený cílový parametr (jev) znázorňuje šanci na výskyt jevu v kategoriích daných kombinacemi hodnot obou faktorů. Relativní četnost výskytu těchto kombinací je popsána poměrem ploch kruhů v grafu. Vliv analyzovaného faktoru je zobrazen v jednotlivých kategoriích hodnot zavádějícího faktoru a dále celkově (šance v kategoriích zavádějícího faktoru jsou propojeny čárkovanou čarou, související relativní četnosti jsou vyznačeny stejným barevným odstínem plochy kruhu).

1a) Bez interakce se zavádějícím faktorem.



Hodnotíme vliv hypertenze na vznik ICHS a jako možný zavádějící faktor uvažujeme dva různé regiony. Z grafů je patrné, že jak jednotlivě v obou regionech, tak na sloučených datech je hypertenze konzistentně rizikovým faktorem pro vznik ICHS (šance vzniku ICHS narůstá). Mezi analyzovaným rizikovým faktorem (hypertenze), a možným zavádějícím faktorem (region) tak nebyla zjištěna interakce.

1b) S interakcí se zavádějícím faktorem.



Hodnotíme úspěšnost určité terapie při zastavení progrese onemocnění, jako možný zavádějící faktor uvažujeme přítomnost metabolického syndromu. Z grafu je patrná silná interakce mezi vlivem terapie a metabolickým syndromem; při přítomnosti zavádějícího faktoru vede terapie k opačnému efektu než u pacientů bez metabolického syndromu.

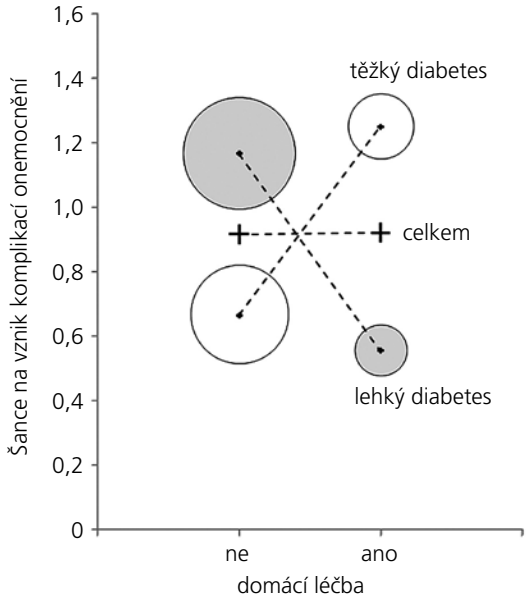
Příklad 1. Ukázka grafu – vizualizace vlivu interakce dvou faktorů na hodnocený cílový parametr.

pomeňme si hlavní typy možného vlivu zavádějících faktorů při studiu vztahu (asociace) mezi expozicí rizikovým faktorem a výskytem klinické události. Pro jednoduchost uvažujme expozici pouze jedním faktorem a vliv jednoho zavádějícího faktoru; např. v retrospektivní studii zkoumáme vliv léčby v domácím prostředí (expozice) na výskyt vážných komplikací sledované choroby (událost), přičemž možným zavádějícím faktorem je u této choroby diabetes jako významná komorbidita. Výstupem studie je v nejjednodušším případě tabulka četností 2 × 2 a z ní provedený odhad poměru šancí (*OR*). Vliv diabetu budeme zkoumat ve čtyřech kategoriích (*stratech*) dle jeho tíže (diabetes žádný, mírný, těžký, velmi těžký). Získáme tak čtyři dílčí, specifické odhady *OR (strata-specific OR estimates, conditional OR estimates)* a zároveň adjustovaný odhad metodou dle Mantela-Haenszela (OR_{MH}), který podrobíme testu hypotézy $OR_{MH} = 1$ (test dle Cochrana-Mantela-Haenszela, viz díly 39–40 seriálu). Možné varianty výsledku jsou následující:

1. Dílčí odhady *OR* provedené přes jednotlivá *strata* daná tíží diabetu jsou homogenní a bez statisticky významného rozdílu (test Breslowa a Daye) a odhad OR_{MH} se neliší od hrubého odhadu *OR* kalkulovaného na celkové tabulce četností bez uvažování vlivu diabetu. Při takovém výsledku diabetes nepůsobí jako významný a vlivný zavádějící faktor a nezkresluje výsledky studie. To může být mimo jiné důsledkem faktu, že výskyt různých kategorií diabetu je mezi případy (v domácí léčbě) a kontrolami stejný. Výsledkem studie může být statisticky významná hodnota *OR* (např. při *OR* > 1 je domácí léčba rizikovým faktorem pro komplikace) nebo hodnota statisticky nevýznamná (domácí léčba výskyt komplikací neovlivňuje).
2. Hrubý (neadjustovaný) odhad *OR* vyjde statisticky významný a ukazuje na rizikový vliv domácí léčby (*OR* > 1), avšak odhad adjustovaný na vliv diabetu (OR_{MH}) významný není. Původní hodnotu v tomto případě zřejmě zkreslil rozdíl ve výskytu diabetu mezi případy a kontrolami. Podobný závěr učiníme i při opačném výsledku, kdy hrubý odhad *OR* je statisticky nevýznamný, ale adjustovaný odhad OR_{MH} významný je. Pokud v obou případech ukazují dílčí odhady pro jednotlivá *strata* na stejný „směr" vlivu domácí léčby (pro-

V retrospektivní studii zkoumáme vztah léčby v domácím prostředí a výskytu komplikací choroby, přičemž možným zavádějícím faktorem je u této choroby diabetes (DM) jako významná komorbidita. Cílem analýzy je zjistit, zda jde o faktor významně modifikující účinek domácí léčby na výskyt komplikací.

Vizualizace interakcí mezi léčbou, diabetem a výskytem komplikací onemocnění



Graf pro vizualizaci interakce dvou faktorů a jejího vlivu na hodnocený cílový parametr znázorňuje šanci na výskyt sledovaného jevu v jednotlivých kategoriích obou faktorů. Relativní četnost výskytů těchto kombinací je popsána poměrem ploch kruhů v grafu. Vliv analyzovaného faktoru je zobrazen v jednotlivých kategoriích hodnot zavádějícího faktoru a dále celkově (šance v kategoriích zavádějícího faktoru jsou propojeny čárkovanou čarou, související relativní četnosti jsou vyznačeny stejným barevným odstínem plochy kruhu).

1. Spočteme hrubé poměry šancí (crude *OR*) v jednotlivých kategoriích modifikujícího faktoru i celkově.

   $OR_{\text{lehký DM}}$ = 0,476 (0,262; 0,866); p = 0,013
   $OR_{\text{těžký DM}}$ = 1,875 (1,134; 3,100); p = 0,014
   OR_{celkem} = 1,005 (0,692; 1,459); p = 0,980

2. Otestujeme homogenitu *OR* mezi úrovněmi modifikujícího faktoru pomocí testu dle Breslowa a Daye:
   p = 0,001 → zamítáme hypotézu o homogenitě odhadů *OR* mezi kategoriemi modifikujícího faktoru.

3. Výsledky popsané v bodech 1 a 2 ukazují na silný vliv modifikujícího faktoru, který v podstatě činí celkové hodnocení poměru šancí (tedy bez ohledu na diabetes) nesmyslným. Za těchto okolností poskytne výpočet adjustovaného odhadu poměru šancí (*adjusted OR*) napříč úrovněmi modifikujícího faktoru pomocí Mantelovy-Haenszelovy metody statisticky nevýznamný výsledek: OR_{MH} = 1,048 (0,722; 1,522); p = 0,877.

Shrnutí: V analýze byla prokázána významná heterogenita vlivu domácí léčby na výskyt komplikací u pacientů s lehkým a těžkým diabetem. Vliv domácí léčby je nutné hodnotit odděleně pro obě kategorie tíže diabetu, neboť diabetes zde vystupuje jako faktor modifikující účinek této léčby. Oba dílčí poměry šancí kalkulované v rámci kategorií diabetu jsou statisticky významné, ale ukazují na opačný směr účinku léčby. Z výsledků analýzy vyplývá silná interakce mezi oběma analyzovanými faktory, což dokládá i grafické znázornění.

Příklad 2. Příklad vlivu binárního faktoru modifikujícího účinek na výskyt komplikací léčby.

tektivní anebo rizikový), můžeme prezentovat adjustovaný odhad OR_{MH}, neboť vliv diabetu je významný a nesmíme ho pominout.

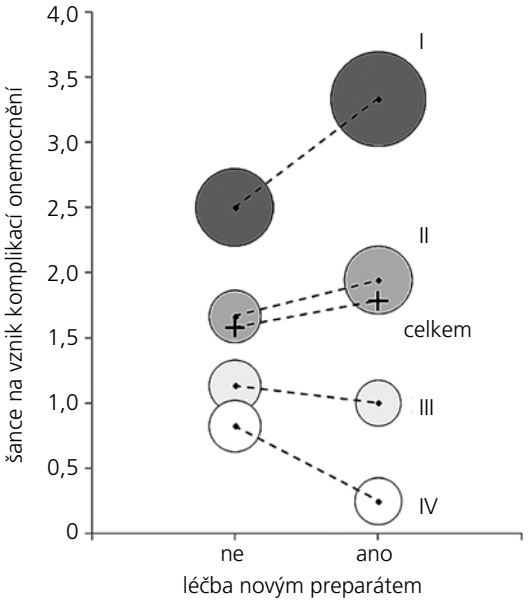
3. Odhad OR_{MH} adjustovaný na vliv diabetu vyjde blízký hodnotě 1 a je statisticky nevýznamný v důsledku protichůdných výsledků v jednotlivých *stratech* daných diabetem. Např. u pacientů bez diabetu a s lehkým diabetem může domácí léčba působit protektivně (dílčí odhady *OR* < 1), zatímco u pacientů se silným diabetem se může domácí léčba projevit jako rizikový faktor komplikací (dílčí *OR* > 1). Takový výsledek by sám o sobě stál za studium, samozřejmě shrnutí studie v jediný odhad *OR* postrádá smysl a je zavádějící. V tomto případě jsme svědky významné interakce faktorů ovlivňujících výskyt komplikací, tedy diabetu a sledované léčby.

Čtenář jistě tuší, že zejména výše popsané varianty 2 a 3 mají k projevu Simpsonova paradoxu nejblíže. V těchto variantách pozorujeme výsledek, kdy celková sumarizace výsledků je v rozporu s dílčími pozorováními. Nejsložitější je ovšem varianta 3, neboť protichůdný vliv zavádějícího faktoru v rámci různých kategorií celkového souboru je vůbec obtížné objevit, a pokud se tak stane, pak se zásadně mění interpretace celé analýzy (aplikace adjustovaného odhadu OR_{MH} není smysluplná). Hovoříme o nehomogenitě ukazatelů asociace zkoumaných jevů v dílčích tabulkách, která je důsledkem interakce daného stratifikačního faktoru (v příkladu výše jde o tíži diabetu) a expozice, jíž je v našem příkladu domácí léčba. Takto působící stratifikační (zavádějící) faktor se nazývá **faktor modifikující účinek** *(effect modulating factor).*

Identifikace a kvantifikace vlivu faktorů, které modifikují účinek zkoumané experimentální (léčebné) intervence, patří jistě mezi nejzajímavější a zároveň nejobtížnější oblasti biostatistiky. Analýza dat zde má velkou šanci objevit zcela nové a pro praxi velmi podstatné vztahy mezi sledovanými veličinami. Jelikož výsledné vztahy mohou být velmi komplikované, má na tomto místě velký význam přehledné grafické znázornění výsledku. Zvláštní formu takových interakčních grafů navrhl ve své publikaci z roku 1985 Minja Paik a my jejich význam přibližujeme v příkladu 1. V podstatě jde o vykreslení odhadů šancí (nikoliv však jejich poměrů, tedy *OR*) na

V retrospektivní studii zkoumáme vztah léčby novým preparátem a výskytu komplikací choroby, přičemž jako možný zavádějící faktor je hodnocen polymorfizmus určitého genu (I-II-III-IV). Cílem analýzy je zjistit, zda jde o faktor významně modifikující účinek sledované léčby na výskyt komplikací.

Vizualizace interakcí mezi léčbou, polymorfizmem genu a výskytem komplikací onemocnění



Graf pro vizualizaci interakce dvou faktorů a jejího vlivu na hodnocený cílový parametr znázorňuje šanci na výskyt sledovaného jevu v jednotlivých kategoriích obou faktorů. Relativní četnost výskytů těchto kombinací je popsána poměrem ploch kruhů v grafu. Vliv analyzovaného faktoru je zobrazen v jednotlivých kategoriích hodnot zavádějícího faktoru a dále celkově (šance v kategoriích zavádějícího faktoru jsou propojeny čárkovanou čarou, související relativní četnosti jsou vyznačeny stejným barevným odstínem plochy kruhu).

1. Spočteme hrubé poměry šancí (crude *OR*) v jednotlivých kategoriích modifikujícího faktoru i celkově.

   OR_{I} = 1,333 (1,006; 1,768); p = 0,046
   OR_{II} = 1,167 (0,803; 1,694); p = 0,419
   OR_{III} = 0,882 (0,573; 1,360); p = 0,571
   OR_{IV} = 0,304 (0,185; 0,497); p < 0,001
   OR_{celkem} = 1,130 (0,951; 1,343); p = 0,165

2. Otestujeme homogenitu *OR* mezi úrovněmi modifikujícího faktoru pomocí testu dle Breslowa a Daye: p = 0,001 → zamítáme hypotézu o homogenitě odhadů *OR* mezi kategoriemi modifikujícího faktoru

3. Výsledky popsané v bodech 1 a 2 ukazují na silný vliv modifikujícího faktoru, který v podstatě činí celkové hodnocení poměru šancí (tedy bez ohledu na polymorfizmus genu) nesmyslným. Za těchto okolností poskytne výpočet nezkresleného odhadu poměru šancí (*adjusted OR*) napříč úrovněmi modifikujícího faktoru pomocí Mantelovy-Haenszelovy metody statisticky nevýznamný výsledek:

   OR_{MH} = 0,954 (0,795; 1,145); p = 0,644

Shrnutí: V analýze byla prokázána významná heterogenita vlivu léčby novým preparátem na výskyt komplikací v kategoriích variantami sledovaného genu. Vliv léčby je nutné hodnotit odděleně dle varianty genu, neboť jeho polymorfizmus zde vystupuje jako faktor modifikující účinek této léčby. Dílčí poměry šancí kalkulované v rámci kategorií zavádějícího faktoru, ukazují na různý účinek léčby dle variant polymorfizmu. Z výsledků analýzy vyplývá silná interakce mezi oběma analyzovanými faktory, což dokládá i grafické znázornění.

Příklad 3. Příklad vlivu kategoriálního faktoru modifikujícího účinek na výskyt komplikací léčby.

výskyt zkoumaného jevu v závislosti na hodnotách působící expozice. Tento vztah je vykreslen jednak pro hrubé odhady kalkulované na celém souboru dat a jednak zvlášť pro *strata* daná kategoriemi zavádějícího faktoru. I při relativně vysokém počtu kategorií (*strat*) tak na první pohled poznáme faktor, který významně ovlivňuje účinek expozice, a je tedy faktorem modifikujícím účinek.

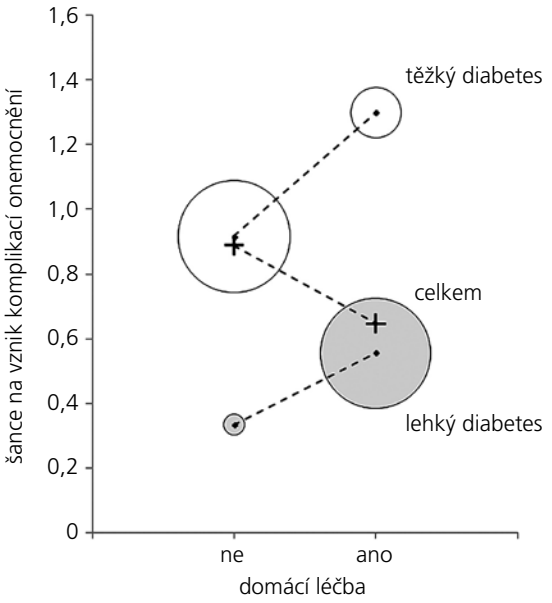
Velkou roli zde hraje fakt, o němž jsme ve výše uvedeném výčtu tří možných výsledků analýzy vůbec neuvažovali, a sice relativní váha jednotlivých podskupin (*strat*). Schematické obrázky v příkladu 1 zachycují i velikost *strat*; vzájemný poměr ploch kruhů u datových bodů odráží poměr počtu pozorování v různých kombinacích kategorií expozice × zavádějící faktor. Jedná se o velmi užitečnou dimenzi zobrazení, neboť *strata* vytvářená kategoriemi hodnot zavádějícího faktoru budou v praxi jen ojediněle stejně velká, a zejména u retrospektivních studií můžeme čekat velké rozdíly v jejich velikosti. Logicky potom na výpočet výsledného adjustovaného OR_{MH} budou mít větší vliv *strata* s větším počtem opakování.

Samotné grafické zviditelnění ovšem jako statistický důkaz významné interakce faktorů nestačí. Komplikovaný vliv faktoru modifikujícího účinek je nutné prokázat statistickými testy. Postup v tomto případě není příliš odlišný od normální asociační analýzy, kterou jsme již představili v dílech 39–40 našeho seriálu. Po odhadu hrubého poměru šancí (*OR*), při znalosti pravděpodobného vlivu zavádějícího faktoru, rozdělíme soubor do dílčích podskupin a budeme směřovat k odhadu adjustovaného OR_{MH}. Pokud ale test homogenity dílčích odhadů *OR* mezi *straty* (test dle Breslowa a Daye) prokáže významnou nehomogenitu dílčích odhadů *OR*, studujeme detailně, jak se vliv expozice na sledovaný klinický jev mezi *straty* liší. A v této fázi můžeme pomocí výše popsaného grafického znázornění odhalit i rozdílný „směr" tohoto vlivu, a tedy interakci vedoucí například i k Simpsonovu paradoxu, kdy stratifikované a nestratifikované odhady vedou k protichůdným závěrům.

Identifikaci faktoru modifikujícího účinek detailně dokumentují příklady 2 a 3. V příkladu 2 prokazujeme vliv faktoru modifikujícího účinek, který má pouze dvě kategorie, a celkový soubor je tak dělen na dvě *strata*. Příklad 3 pracuje se složitějším zadáním při dělení souboru na čtyři *strata*. Zejména u takového problému je

V retrospektivní studii zkoumáme vztah léčby v domácím prostředí a výskytu komplikací choroby, přičemž možným zavádějícím faktorem je u této choroby diabetes (DM) jako významná komorbidita. Cílem analýzy je zjistit, zda jde o faktor významně modifikující účinek domácí léčby na výskyt komplikací.

Vizualizace interakcí mezi léčbou, diabetem a výskytem komplikací onemocnění



Graf pro vizualizaci interakce dvou faktorů a jejího vlivu na hodnocený cílový parametr znázorňuje šanci na výskyt sledovaného jevu v jednotlivých kategoriích obou faktorů. Relativní četnost výskytů těchto kombinací je popsána poměrem ploch kruhů v grafu. Vliv analyzovaného faktoru je zobrazen v jednotlivých kategoriích hodnot zavádějícího faktoru a dále celkově (šance v kategoriích zavádějícího faktoru jsou propojeny čárkovanou čarou, související relativní četnosti jsou vyznačeny stejným barevným odstínem plochy kruhu). Graf ukazuje rozpor ve výsledku dílčích analýz a analýzy celého souboru dat, jde o typickou ukázku Simpsonova paradoxu.

1. Spočteme hrubé poměry šancí (*crude OR*) v jednotlivých úrovních modifikujícího faktoru i celkově.

   $OR_{\text{lehký DM}}$ = 1,667 (1,111; 2,501); p = 0,010
   $OR_{\text{těžký DM}}$ = 1,418 (1,209; 1,663); p < 0,001
   OR_{celkem} = 0,727 (0,666; 0,794); p < 0,001!

2. Otestujeme homogenitu *OR* mezi úrovněmi modifikujícího faktoru pomocí testu dle Breslowa a Daye: p = 0,468 → nezamítáme hypotézu o homogenitě odhadů *OR* mezi kategoriemi modifikujícího faktoru.

3. Provedeme výpočet adjustovaného odhadu poměru šancí (adjusted *OR*) napříč úrovněmi modifikujícího faktoru pomocí Mantelovy-Haenszelovy metody:
   OR_{MH} = 1,452 (1,252; 1,684); p < 0,001 → adjustovaný poměr šancí je statisticky významný.

Shrnutí: Hrubý odhad poměru šancí je statisticky významný v obou kategoriích modifikujícího faktoru i při analýze sloučeného souboru dat. Avšak směr vlivu dle poměru šancí je u dílčích kategorií a u sloučeného souboru opačný, což je typická ukázka Simpsonova paradoxu. Jakoby domácí léčba působila jako rizikový faktor komplikací v rámci obou kategorií tíže diabetu, ale na celkovém souboru u ní pozorujeme trend zcela opačný, tedy protektivní účinek. Vliv domácí léčby v rámci kategorií tíže diabetu je nadto konzistentní, nezamítli jsme hypotézu homogenity těchto dílčích odhadů *OR*.

Adjustovaný odhad poměru šancí korigovaný na vliv tíže diabetu vedl k statisticky významnému výsledku, který potvrdil rizikovost domácí léčby vůči výskytu komplikací nemoci (OR_{MH} = 1,452).

Příčinou rozporu nekorigovaných analýz zde byl silně nevybalancovaný design a rozdílný výskyt různých kategorií tíže diabetu v kohortě s a bez domácí léčby. Správným výsledkem je odhad *OR* adjustovaný na vliv tíže diabetu.

Příklad 4. Simpsonův paradox na příkladu faktoru modifikujícího účinek se dvěma kategoriemi.

možnost přehledného grafického znázornění velmi cenná. Zvláštní výstup těchto analýz způsobený typickým projevem Simpsonova paradoxu přibližuje příklad 4, kde dílčí analýzy stratifikovaného souboru společně ukazují na určitý účinek léčby, ale po spojení dílčích souborů v celek získáváme výsledek opačný.

Na závěr zdůrazněme jeden velmi podstatný fakt, který se sice zdá samozřejmostí, nicméně bývá ve světě moderní vědy často opomíjen. Identifikace faktoru modifikujícího účinek samozřejmě předpokládá, že tento faktor experimentátor zná a o jeho účinku buď předem ví anebo jej předpokládá. Jen tak může být daný faktor zahrnut mezi sledované proměnné a správně zapojen do analýzy. Pokud výsledky ovlivňuje faktor, jenž není sledován vůbec, a tudíž ani zaznamenán v datovém souboru, pak jeho vliv samozřejmě odhalit nelze. Závažného zkreslení výsledku asociační analýzy se přitom můžeme při nevybalancovaném designu studie dočkat i od naprosto „obyčejného" parametru, jako je např. body mass index, pohlaví či věk pacientů. Je tedy velmi důležité, aby u asociačních studií byla data o zařazených pacientech co nejkomplexnější a subjekty byly řádně popsány, a to i běžnými demografickými, diagnostickými a klinickými charakteristikami. Tím umožníme prověření jejich možného vlivu na konečný výsledek studie. Jakkoli se tato poznámka v době nabité molekulárními a genetickými markery může jevit jako redundantní, stále ještě nejsme v poznání podstaty většiny jevů (onemocnění) tak daleko, abychom mohli zanedbat poctivý popis souboru dat. Právě progresivní genetické asociační studie, jež hledají vztahy mezi zcela novými markery a klinickými jevy, musí být schopny doložit, že jejich nálezy nejsou zkreslené zavádějícími faktory.

Literatura

Paik M. A graphic representation of a three-way contingency table: Simpson's paradox and correlation. American Statistician 1985; 39: 53–54.
Simpson EH. The interpretation of interaction in contingency tables. J Roy Stat Soc, B 1951; 13: 238–241.