Analýza dat v neurologii

XLI. Simpsonův paradox

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 tomto díle seriálu volně navážeme na díly předchozí a pokusíme se vysvětlit statistický paradox nazvaný po Edwardu H. Simpsonovi, který jej jako první popsal (Simpson, 1951). Ačkoli jde o obecnější fenomén, navážeme při jeho vysvětlování na předchozí výklad věnovaný odhadu poměru šancí a relativního rizika (díly XXXV–XXXIIX seriálu) a zejména na výklad věnovaný vlivu tzv. zavádějících faktorů (díly XXXIX–XL seriálu). Zkoumáme-li vztah dvou náhodných veličin, typicky vztah mezi expozicí nějakým faktorem a výskytem zkoumaného jevu, musíme mít vždy na paměti, že existuje nezanedbatelné riziko zkreslení výsledků. Zejména u retrospektivních studií, např. u studie případů a kontrol, kde při výběru subjektů nelze přímo ovlivňovat rovnoměrné zastoupení všech charakteristik srovnávaných skupin. Tyto znaky potom mohou v důsledku rozdílné prevalence ve skupinách významně ovlivnit výskyt sledovaného jevu, a zkomplikovat tak jednoduché zobecnění našich pozorování. V této souvislosti proto hovoříme o zavádějících faktorech. Dokonce se může i stát, že v rámci určitých kategorií hodnot zavádějícího faktoru vykazuje výskyt zkoumaného jevu opačný trend než v jiných kategoriích.

Právě výrazný rozdíl ve výstupech dílčích kategorií většího souboru, případně až rozpor mezi výstupy z těchto dílčích podsouborů a z celého souboru dat byl nazván **Simpsonův paradox**. Název je to trefný, neboť popisuje skutečně nečekaný, paradoxní výsledek analýz. Jeho podstatu se pokusíme obecně vysvětlit na jednoduchém příkladu, v němž máme za úkol srovnat hodnotu markeru *X* ve dvou skupinách subjektů označených *A* a *B*. Pracujeme jednak s celkovým souborem zařazených jedinců a jednak s podsoubory, které vznikly rozdělením celého souboru na tři skupiny podle zvoleného (stratifikačního) faktoru. Paradox spočívá v situaci, kdy v každém ze tří dílčích podsouborů vykazují jedinci ze skupiny *A* hodnotu markeru *X* větší než jedinci ze skupiny *B*, avšak na spojených datech dostaneme opačný výsledek. Jinak řečeno, tři dílčí porovnání dvou skupin subjektů konzistentně ukazují nějaký výsledek, přičemž po spojení všech pozorování ve větší celek dostáváme opačný výstup.

Čtenáře nyní jistě napadne legitimní pochybnost, zda je takový rozpor mezi dílčími pozorováními a jejich shrnutím vůbec možný. Doložme si tedy takovou situaci přímo na jednoduchém výpočtu. Zkoumáme účinnost dvou léků L_1 a L_2, a to ve dvou srovnatelných nemocnicích. Podle vstupní hypotézy by léky měly zabraňovat výskytu určitého nežádoucího účinku. V první nemocnici test dopadne ve prospěch léku L_1, který ochrání více pacientů než lék L_2 (30 vs 25 %). Výsledky ve druhé nemocnici jsou ještě průkaznější, opět ve prospěch léku L_1 (100 vs 75 %). Zdá se tedy, že lék L_1 je jednoznačně lepší než lék L_2. Pokud se ale na výsledky podíváme po spojení souborů z obou nemocnic, tak se superiorita L_1 již jako jednoznačná nejeví. Konkrétní celkový výsledek totiž vychází 48,1 % ochráněných pacientů u léku L_1 a 58,3 % pacientů u léku L_2. Jak je to možné?

Klíčem k odpovědi je různý počet pacientů zařazených ke sledování v obou hodnocených nemocnicích, tedy tzv. nevybalancovaný design, který dává výstupům v dílčích pozorováních zcela rozdílnou váhu. Velmi malý počet opakování v dílčích sledováních pak způsobí, že celkový souhrn má sílu („váhu") dílčí výsledky zcela otočit. Lék L_1 byl v první nemocnici testován na 20 pacientech a v další nemocnici jen na sedmi. Účinnosti tedy dosáhl u šesti pacientů z 20 v první nemocnici (30 %) a v druhé nemocnici u všech sedmi pacientů (100 %). Lék L_2 byl v první nemocnici nasazen u osmi pacientů, kde uspěl dvakrát (25 %); ve druhé nemocnici byl pak indikován u 16 pacientů a uspěl 12krát (75 %). Pokud testy z nemocnic spojíme, pak lék L_1 byl účinný u 13 pacientů z 27 testovaných (48,1 %), avšak lék L_2 dosáhl účinku u 14 pacientů z celkových 24, tedy v 58,3 % případů.

Ačkoli výše uvedené rozdíly mezi nemocnicemi a léky by v důsledku malého počtu opakování byly stěží statisticky významné, je konečný výsledek jistě poučný. Různé počty opakování v dílčích pozorováních (zde nemocnicích), a tedy různá váha dílčích výsledků neumožňují paušální shrnutí celkových výsledků bez detailního rozboru. Pravděpodobnost zkreslení je v těchto případech velmi vysoká. Tento typ výsledku není v klinickém výzkumu a v klinické praxi nijak vzácný, podobné výstupy mohou nastat např. v těchto situacích:

- Jakékoli spojování vzájemně nezávislých dílčích experimentů s malými vzorky v jeden celek může vést přesně k situaci popsané v příkladu výše. Typickým příkladem mohou být klinické studie Fáze I a Fáze II (zaměřené na úvodní testy toxicity a bezpečnosti nových léků); ty bývají z etických důvodů často prováděny s omezeným počtem probandů. Podobným příkladem může být studium jevů (zejména vzácných jevů) prováděné na dílčích souborech v různých zdravotnických zařízeních.
- Vzácné nemoci (*rare diseases*, dle definice s prevalencí menší než 5×10^4) a pro ně určené léčivé přípravky (tzv. *orphan drugs*) lze hodnotit pouze v me-

zinárodních projektech, které mají potenciál nabrat dostatečně velký vzorek pacientů. Dílčí studie na úrovni menších regionů mohou vést k výrazně zkresleným výsledkům.
- Hodnocení nových léčebných přístupů zdravotními systémy vyspělých států v současnosti stále častěji využívá sledování výsledků v reálné klinické praxi daného státu. To je jistě zcela legitimní a správné. Avšak je třeba velmi opatrně interpretovat výsledky studií, které na národní úrovni usilují o opakování (nebo ověření) výsledků velkých randomizovaných studií, jež ale často pracovaly s významně většími počty probandů.

Výše uvedené příklady popisují relativně časté experimentální a klinické situace, kde zkreslující závěr vyplývající z nesprávné sumarizace dílčích pozorování (dílčích publikací apod.) může mít závažné důsledky. Z tohoto důvodu má znalost Simpsonova paradoxu velký význam. Ve výše uvedeném příkladu vedlo srovnávání dvou léků na dílčích, různě velkých podsouborech k zavádějícím závěrům. Spojená data prokázala opačný výsledek. Tento příklad ukázal, jak vážným problémem mohou být zobecněné a neověřené závěry učiněné na bázi dílčích studií. Nicméně jde o příklad značně specifický v tom, že dílčí soubory bylo rozumné spojit do většího, reprezentativnějšího. Stratifikačním faktorem zde byla jednotlivá testovací centra, nemocnice. Logickým závěrem tedy je, že ve srovnání léků L_1 a L_2 více věříme analýze celkového souboru s větším vzorkem.

Představme si ale nyní situaci, kdy stratifikačním faktorem bude charakteristika, jejíž kategorie nemá smysl slučovat, neboť popisují objektivně existující entity s významným vlivem na chování sledovaného jevu. Takovým stratifikačním faktorem může být např. klinické stadium nemoci, pohlaví pacientů, kategorie biomarkeru apod. I v takové situaci můžeme narazit na Simpsonův paradox, nicméně výsledná interpretace bude jiná než ve výše uvedeném příkladě s léky. Podívejme se na průzkum sledující výskyt závažných komplikací nějaké choroby v různých nemocnicích. Cílem je srovnat dvě nemocnice, N_1 a N_2. Představme si, že ke srovnání přistoupíme paušálně, tedy bez ohledu na spektrum léčených pacientů (cizím termínem bez ohledu na tzv. *case mix* nemocnice). Vlastní realizaci provedeme zdánlivě odpovědně a z každého zařízení vybereme posledních 1 000 léčených pacientů s danou chorobou. Z těchto dat vyplývá, že:
- v nemocnici N_1 se závažná komplikace vyskytla u 300 pacientů z 1 000, tedy u 30 %,
- v nemocnici N_2 se závažná komplikace vyskytla u 200 pacientů z 1 000, tedy u 20 %.

Závěr učiněný touto analýzou je jasný, nemocnice N_2 se jeví jako výrazně bezpečnější.

Představme si ale, že jde o nemoc, jejíž komplikace se významně častěji vyskytují u pacientů starších 50 let. Potom je samozřejmě mnohem správnější a spravedlivější srovnat nemocnice v rámci věkových kategorií, tedy tzv. *strat* vytvořených s pomocí věkových kategorií. Výsledky, které dostáváme, jsou následující:
- věková kategorie mladších pacientů (≤ 50 let):
- nemocnice N_1: 600 pacientů, 60 komplikací, tj. 10 %,
- nemocnice N_2: 900 pacientů, 117 komplikací, tj. 13 %.
- věková kategorie starších pacientů (> 50 let):
- nemocnice N_1: 400 pacientů, 240 komplikací, tj. 60 %,
- nemocnice N_2: 100 pacientů, 83 komplikací, tj. 83 %.

Je zřejmé, že dílčí pohled přes rozumně zvolené kategorie zavádějícího faktoru ukazuje opak původního výsledku, který jsme získali bez ohledu na tyto kategorie. Jelikož jsou aplikované kategorie věku klinicky i věcně relevantní, měly být použity hned na počátku analýzy. Vyšší podíl mladších pacientů totiž nemocnici N_2 v celkovém shrnutí zvýhodnil. Správný je závěr učiněný při respektování věkových kategorií a ten je, že nemocnice N_1 je z hlediska bezpečnosti lepší.

A kde se zde projevil Simpsonův paradox? Jednak v tom, že závěr získaný analýzou dílčích kategorií je opačný než celkové shrnutí. A dále také v tom, že větší velikost vzorku, kterou pro srovnání obou nemocnic máme k dispozici při ignorování věkových kategorií, nevede ke správnému výsledku.

Simpsonův paradox tak vlastně zpochybňuje jedno ze základních pravidel statistiky, a sice že větší soubory dat vedou ke spolehlivějším výsledkům. Pokud sloučením dílčích souborů dat ztratíme nějakou podstatnou informaci (např. výše, že nemocnice N_2 léčí významně větší množství mladších pacientů než nemocnice N_1), pak tato zavádějící proměnná může zkreslit výsledky spojeného, i když většího souboru dat. Riziko nastání tohoto paradoxu je třeba mít na paměti i v případě, když na dílčím souboru usilujeme o zopakování výsledku větší studie anebo dokonce toto zopakování výsledku automaticky předpokládáme. Takový předpoklad ovšem nemusí být vždy naplněn. Avšak tímto varováním téma Simpsonova paradoxu nekončíme, naopak v příštím díle se mu budeme detailněji věnovat při odhadu poměru šancí a relativního rizika.

Literatura

1. Hendl J. Přehled statistických metod zpracování dat. Praha: Portál 2004.
2. Simpson EH. The interpretation of interaction in contingency tables. J Roy Stat Soc B 1951; 13: 238–241.