Analýza dat v neurologii

LXIII. Pozor na interpretaci ekologických (korelačních) studií – II.

L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková

Institut biostatistiky a analýz
Masarykova univerzita, Brno

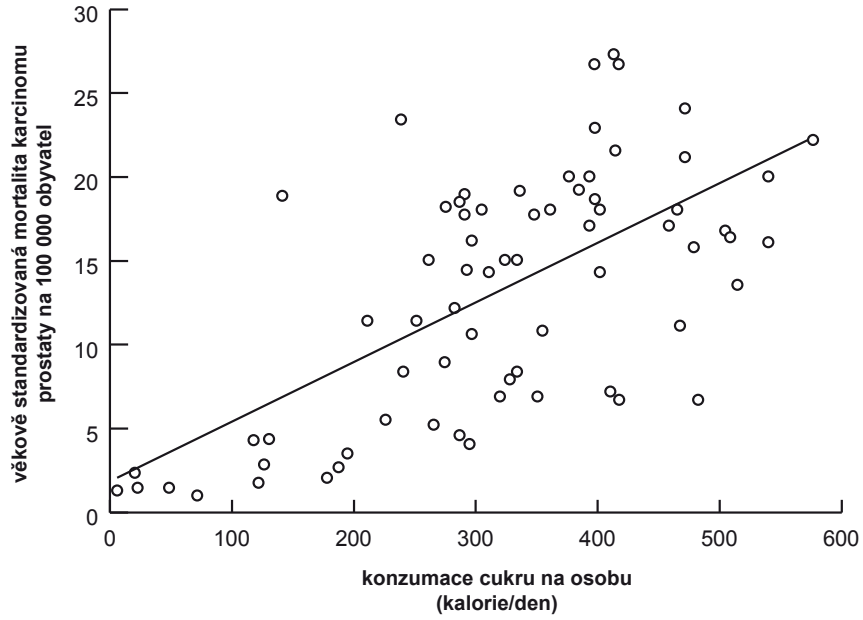
✉
doc. RNDr. Ladislav Dušek, Ph.D.
Institut biostatistiky a analýz
MU, Brno
e-mail: dusek@iba.muni.cz

V předchozím díle seriálu jsme otevřeli problematiku tzv. ekologických (korelačních) studií jako zvláštního typu observačních studií „expozice–účinek". Jejich nejvýznamnějším specifikem je sledování vztahu mezi expozicí a jejím následkem na skupinové, někdy až populační úrovni. Do analýzy zde typicky vstupují celé kohorty či populace, reprezentované agregovanými hodnotami zkoumaných charakteristik jako je např. konzumace cukru, expozice UVB záření apod. Na straně následku (efektu) potom nejčastěji vystupují epidemiologické parametry, typicky incidence, mortalita či prevalence určitých chorob. Krátká rešerše literatury uvedená v minulém díle doložila, že tento typ sledování je velmi často využíván i v současném výzkumu, ačkoli sledování vztahů mezi parametry bez možnosti korelovat individuální data vyvolává mnoho otázek a často i oprávněnou kritiku. Zejména pokud interpretace takových studií nerespektuje jejich objektivní limitace.

Faktem je, že ekologické studie patří mezi experimentální plány popisné (observační). Často také bývají řazeny mezi studie analytické. Na rozdíl od intervenčních experimentů zde nezasahujeme do přirozeného vývoje událostí a hodnotíme, nejčastěji retrospektivně, výskyt sledovaných jevů. Retrospektivní přístup přináší sám o sobě řadu interpretačních limitů, avšak ekologické

Ekologické studie analyzují agregovaná data kohort nebo populací, z nichž každá vystupuje ze statistického hlediska jako jeden objekt. Počet hodnot v analýze četností je zde tedy odvozen od počtu kohort/populací, nikoli od počtu individuálních jedinců jako v běžné analýze tabulek četností. Používané statistické metody jsou nicméně shodné s analýzami kontingenčních tabulek generovaných ze souborů jednotlivých pacientů. Jako typický příklad zde citujeme studii autorů Colli a Colli (2006) zabývající se vztahem mezi mortalitou karcinomu prostaty, stravovacími návyky a expozicí slunečnímu světlu v 71 zemích. Jednotlivé proměnné byly zpracovány jako spojité (kvantitativní) a každá země je tak zastoupena jednou hodnotou pro danou proměnnou. Studie využila standardní statistické metody pro zpracování spojitých dat, tj. průměr, minimum, maximum a percentily pro popisnou analýzu, korelační koeficienty a lineární regresní modely (model přímky) pro analýzu vztahů mezi proměnnými.

Graf 1. Vztah mezi konzumací cukru a mortalitou karcinomu prostaty v 71 zemích.



Příkladem výstupu je vztah mezi zaznamenanou úrovní spotřeby cukru a mortalitou na karcinom prostaty (graf 1). Graf znázorňuje relativně silnou a statisticky významnou vazbu mezi těmito proměnnými (Pearsonův korelační koeficient $r = 0{,}71$, regresní koeficient $B = 0{,}037$; $p < 0{,}001$); každý bod v grafu je přitom agregací (průměrem) dat jedné země vstupující do výpočtů.

Obdobným způsobem byly zpracovány všechny proměnné a vytvořeny dva vícerozměrné modely popisující vztah charakteristik populací k mortalitě na karcinomu prostaty: 1. spotřeba cukru, cereálií, cibule, živočišného tuku a 2. spotřeba cereálií, cukru, cibule a index UV záření.

Ačkoli zde autoři nemohou exaktně prokázat kauzalitu vztahů, jsou obdobné ekologické studie užitečným nástrojem pro identifikaci vlivu faktorů ovlivňujících riziko zdravotních problémů dané kohorty/populace.

Příklad 1. Ekologické (korelační) studie – definiční příklad.

Vstupem do analýzy ekologických studií nemusí být pouze spojité proměnné, jak ukazuje příklad 1, ale také proměnné binární nebo kategoriální, kdy je každá populace/kohorta zastoupená svým zařazením do jedné kategorie – např. v populaci je více než 50 % pacientů s diabetem – ano/ne, alergie se vyskytuje u více než 30 % dětí - ano/ne, více než 70 % populace konzumuje 400 a více kalorií/den ve formě cukru – ano/ne, atd.

Následující příklad hodnotí vztah mezi výskytem alergie u více než 30 % dětí v hodnocené kohortě a více než roční expozicí alergenu na vzorku 96 populačních kohort. Analyzovaný vzorek je tak dán počtem kohort, nikoli počtem jedinců v kohortách. Data na individuální úrovni nejsou k dispozici.

Alergie u více než 30 % dětí	Expozice alergenu delší než 1 rok		Celkem
	ne	ano	
ne	40	10	50
ano	20	26	46
celkem	60	36	96

Alergie u více než 30 % dětí	Expozice alergenu delší než 1 rok		Celkem
	ne	ano	
ne	66,7 %	27,8 %	52,1 %
ano	33,3 %	72,2 %	47,9 %
celkem	100,0 %	100,0 %	100,0 %

Data jsou popsána standardně pomocí absolutních a relativních četností. Z výsledků popisné statistiky se jeví, že kohorty (populace) s expozicí alergenu delší než 1 rok jsou častěji asociovány s více než 30% výskytem alergií u dětí (72,2 vs. 33,3 %). Tuto hypotézu je možno testovat běžnými testy pro kontingenční tabulky jako je Pearsonův test dobré shody nebo Fisherův exaktní test. V případě použití Pearsonova testu zde zamítáme nulovou hypotézu o nezávislosti výskytu alergie a expozice alergenu ($\chi^2 = 13{,}63$; $p < 0{,}001$).

Závěr: V ekologické studii 96 kohort byla na hladině významnosti $p < 0{,}001$ zamítnuta hypotéza o nezávislosti výskytu alergie u více než 30 % dětí a expozice alergenu delší než 1 rok. Výskyt alergií byl statisticky významně častější u kohort s delší expozicí alergenu (72,2 vs. 33,3 %).

Příklad 2. Analýza kontingenčních tabulek v ekologických studiích.

studie jsou objektivními limity zatížené ještě více. Zatímco ve všech ostatních studiích je základní jednotkou analýzy jednotlivec, v ekologických studiích je to skupina osob, např. populace regionů či dokonce celých států. Již v minulém díle jsme rozebírali řadu faktorů, které mohou informační hodnotu takových pozorování silně omezit. Typickou ukázku ekologické studie představuje práce citovaná a dokumentovaná v příkladu 1 (Colli a Colli, 2006). Je patrné, že zde korelujeme charakteristiky životního stylu, agregované pro populace celých států, s integrálním ukazatelem, tedy mortalitou na nádory prostaty. Takto zobecněné analýzy jsou velmi náchylné ke zkreslení nebo dezinterpretaci výsledků; často proti sobě vystupují značně rozdílné faktory s množstvím pozaďových vlivů maskujících skutečný efekt.

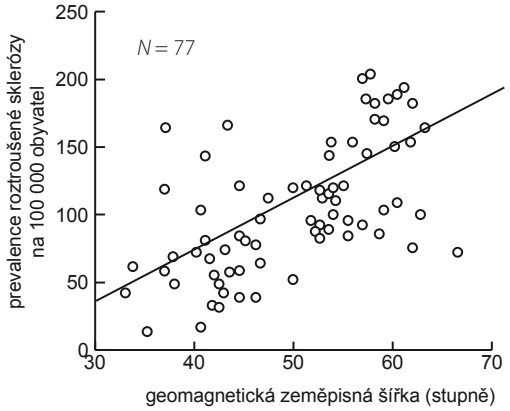
Při ekologických sledováních může řadu problémů generovat již samotný proces získávání dat ve formě charakteristik studovaných populací. Data za skupinu osob jsou totiž vždy určitým způsobem zobecněná. Může jít např. o průměry či mediány hodnot měřených na určitém vzorku jedinců, o celkovou kumulativní dávku danou expozicí nebo o populační prevalenci epidemiologického faktoru. Obecně můžeme charakteristiky populací vstupující do ekologických studií dělit následovně:

- Agregované proměnné získané původně měřením na vzorku jedinců a následně numericky vyjádřené pro celou skupinu (jako průměr, medián, podíl osob s nějakou charakteristikou apod.). Agregaci lze použít na straně zkoumaných rizikových faktorů (průměrná konzumace cukru v populaci) i pro vyjádření jejich důsledků (podíl obézních osob, průměrný body mass index). Již z popisu je patrné, že agregace hodnot maskuje inter-individuální variabilitu ve zkoumaných populacích. Rovněž samotný postup zvolený pro numerické vyjádření a agregaci hodnot může být podstatným zdrojem zkreslení a je třeba jej velmi pozorně interpretovat. Například může být velký rozdíl mezi průměrnou a mediánovou konzumací cukru apod.
- Charakteristiky sídel, pracovního prostředí či environmentální faktory představují velmi podstatnou skupinu proměnných vhodných pro ekologické studie. Na rozdíl od výše popsaných agregovaných proměnných zde často není možné získat hodnoty příslušné konkrétnímu jedinci. Jako příklad uveďme např. koncentraci polutantů ve vzduchu ve městě nebo obsah těžkých kovů v pitné vodě v určité oblasti.
- Souhrnné charakteristiky populací a společnosti, pro které neexistuje z definice možnost měření na individuální úrovni. Příkladem zde může být hustota osídlení nebo procento HDP investované ročně do zdravotnictví. Chceme-li takového faktory na straně expozice korelovat např. s nemocností či jinými epidemiologickými proměnnými, nutně pracujeme s celými populacemi.

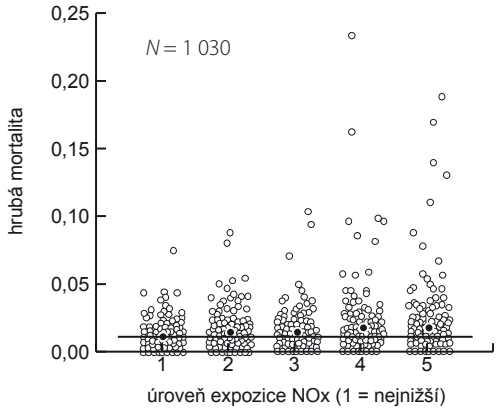
Z výše uvedeného jistě vyplývá i vysvětlení, proč jsou ekologické studie stále tak frekventované nejen v epidemiologické literatuře. Zejména jde o situace, kdy individuální data nelze získat anebo by jejich získání bylo neetické (např. studie cíleně vystavující vybrané osoby polutantům ve vzduchu za účelem srovnání zdravotních rizik s kontrolou není proveditelná). V řadě případů je důvodem pro ekologickou studii i fakt, že studovaný problém cíleně vyžaduje zobecnění na populační úrovni („community-level studies"), např. při hodnocení efektu různých preventivních programů. Analýzu vztahů na komunitní úrovni rovněž usnadňuje fakt, že kalkulačně nejsou postupy statistického hodnocení nijak odlišné od analýzy individuálních dat. V příkladu 1 hodnotíme trend dvou spojitých proměnných, v bodovém grafu však nejsou zaneseny jednotlivé osoby, ale celé státy. Obdobně příklad 2 ukazuje, jak lze do běžné frekvenční tabulky vkládat data celých populací s využitím jejich populačních charakteristik. Další výpočet se již technicky neliší od obecného postupu hodnocení kontingenčních tabulek.

Ekologické studie jsou využívány ve všech oblastech medicíny, zde prezentované příklady jsou ukázkou z odborné neurologické literatury.

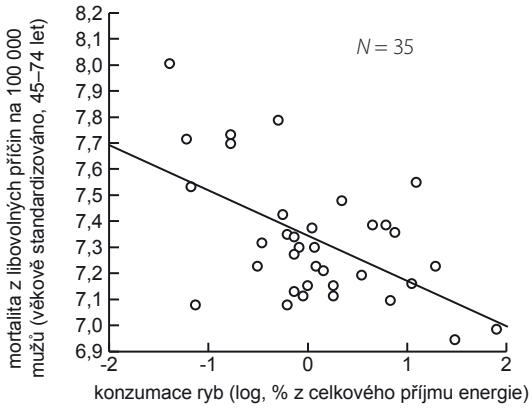
Graf 1. Vztah prevalence roztroušené sklerózy a geomagnetických disturbancí na evropských lokalitách*.



Graf 2. Vztah mezi mortalitou na cévní mozkovou příhodu a expozicí NOx na lokalitách v Sheffieldu$.



Graf 3. Vztah mezi konzumací ryb a mortalitou z libovolných příčin u mužů v různých zemích#.



*Sajedi SA, Abdollahi F (2012)

$Haining R et al (2010)

#Zhang J et al (1999)

Příklad 3. Tři příklady publikovaných ekologických studií.

Ačkoli jsou studie založené na individuálních datech obecně považovány za více spolehlivé, nelze ekologické sledování paušálně odsoudit jako zavádějící. Ostatně analýza (celo)populačních charakteristik může být z epidemiologického hlediska více relevantní než individuální záznamy o nemoci u vybrané skupiny osob. Ekologické sledování přináší do asociačních studií populační kontext a je vhodné pro sledování interakce více rizikových faktorů zároveň, jako zdroj nových hypotéz o rizikových faktorech chorob a také pro studium vzácných chorob, kde individuální data nemusí být v dostatečně velkých souborech k dispozici. Jistým důkazem potřebnosti je i fakt, že ekologická sledování jsou stále velmi často publikována ve významných klinických časopisech, vč. časopisů neurologických. Ukázku vybraných výstupů takových studií přináší příklad 3.

Výhodami ekologických studií jsou totiž zejména velké velikosti vzorku a široké spektrum korelovaných faktorů. Velmi často tak tyto analýzy vedou k objevu faktorů vzájemně modifikujících svůj vliv na etiopatogenezi nemocí („risk-modifying factors"). Velké počty osob zahrnutých do regionálních či celostátních populací a kohort představují výhodu zejména při studiu vzácných chorob, kde není analýza epidemiologických charakteristik na bázi individuálních sběrů dat často dobře proveditelná. Obdobně jsou ekologická sledování výhodná při studiu chorob s dlouhou latencí, kde by prospektivně organizované studie nebyly časově reálně proveditelné. Jistou výhodou těchto studií bývá i jejich nízká cena, často totiž pracují s rutinně pořizovanými daty bez nutnosti dalších nákladů.

Nicméně objektivní pravdou zůstává fakt, že ekologické studie nejsou a nemohou být posledním stupněm při prokazování kauzality vztahu mezi expozicí (rizikovým faktorem) a následkem. Interpretace kauzality vztahů čistě na bázi populačních charakteristik není v naprosté většině případů přípustná. Pro interpretaci výsledků ekologických studií je důležité sledovat jejich vstupní hypotézu. Hypotéza, zda konzumace cukru statisticky souvisí (koreluje) s mortalitou na nádory prostaty (příklad 1), není totožná s otázkou, zda tato konzumace tuto mortalitu způsobuje. Většina korelačních studií je ovšem prováděna na základě úvahy, která jakýsi příčinný vztah předpokládá. Studie sice může zmíněný vztah analyticky indikovat a označit ho za statisticky významný, avšak pro konečný důkaz kauzality jsou nutné další experimenty a analýzy. Tyto musí vyloučit zkreslení v důsledku náhodné chyby, systematická zkreslení, vliv matoucích („confounding") faktorů a také možnost ekologického zkreslení („ecological fallacy", viz díl 62 seriálu).

Průkazu kauzality vztahů jsme již věnovali díl 59 našeho seriálu. Jeho obsah zde můžeme pouze shrnout do závěru, že exaktní průkaz

kauzality vztahů vyžaduje kombinaci více experimentálních přístupů, víceúrovňovou analýzu dat („multi-level analysis“) zahrnující jak data ekologická, tak nutně i data individuální. Pro exaktní průkaz kauzality je vyžadována celá škála kritérií, od velmi exaktní analýzy reprodukovatelnosti sledování, kvantifikace trendu „dávka–účinek“ až po literární metaanalýzy za účelem posouzení věrohodnosti a koherence různých sledování. Nejde tedy o jednoduchý postup, který by ekologická sledování mohla sama o sobě naplnit. Ačkoli souhrnnou sadu kritérií kauzality publikoval Austin Bradford Hill již v roce 1965, stále se v mezinárodní literatuře objevují na dané téma nové metodické rozbory (Grant 2009, Howick et al 2009, Glass et al 2013) a téma je i kriticky rozvíjeno (Rothman a Greenland, 2005).

Tato neustále živá metodická debata jen dokazuje význam tématu a v konečném důsledku i význam ekologických analytických studií. Snad se nám v tomto a předchozím díle seriálu podařilo přesvědčit čtenáře, že jde sice o analýzy interpretačně limitované, avšak stále držící své pevné místo v současném, zejména epidemiologickém, výzkumu.

Literatura

Colli JL, Colli A. International comparisons of prostate cancer mortality rates with dietary practices and sunlight levels. Urol Oncol 2006;24(3):184–94.

Glass TA, Goodman SN, Hernán MA, et al. Causal inference in public health. Annu Rev Public Health 2013;34:61–75. doi: 10.1146/annurev-publhealth-031811-124606.

Grant WB. How strong is the evidence that solar ultraviolet B and vitamin D reduce the risk of cancer? An examination using Hill's criteria for causality. Dermatoendocrinology 2009:1(1):17–24.

Haining R, Li G, Maheswaran R, et al. Inference from ecological models: Estimating the relative risk of stroke from air pollution exposure using small area data. Spatial Spatiotemporal Epidemiol 2010;1(2–3):123–31. doi: 10.1016/j.sste.2010.03.006.

Hill AB. The Environment and Disease: Association or Causation? Proc R Soc Med 1965:58:295–300.

Howick J, Glasziou P, Aronson JK. The evolution of evidence hierarchies: what can Bradford Hill's ‚guidelines for causation' contribute? J R Soc Med 2009:102(5):186–94. doi: 10.1258/jrsm.2009.090020.

Rothman KJ, Greenland S. Causation and causal inference in epidemiology. Am J Public Health 2005;95(Suppl 1): S144–50.

Sajedi SA, Abdollahi F. Geomagnetic disturbances may be environmental risk factor for multiple sclerosis: an ecological study of 111 locations in 24 countries. BMC Neurol 2012;12:100. doi: 10.1186/1471-2377-12-100.

Zhang J, Sasaki S, Amano K, et al. Fish consumption and mortality from all causes, ischemic heart disease, and stroke: an ecological study. Prev Med 1999;28(5): 520–9.