Analýza dat v neurologii

I. Úvod a vymezení pojmů

doc. RNDr. Ladislav Dušek, Dr.
Institut biostatistiky a analýz,
Masarykova univerzita, Brno
e-mail: dusek@cba.muni.cz

Tento článek je úvodem k seriálu, ve kterém se společně s dalšími odborníky pokusíme v uvolněném stylu přibližovat základy analýzy klinických dat. Neklademe si za cíl nahradit existující učebnice, témata budou volena především s ohledem na jejich význam pro praxi. Uvolněný styl je nutný proto, abychom z této problematiky sejmuli nálepku formalizmu plného termínů a matematických formulí, běžné medicíně poměrně cizích. Analýzu dat neboli „statistiku" budeme prezentovat jako nástroj, který má svým uživatelům pomáhat, a ne je limitovat v myšlení nebo děsit ve spaní. Vstupním předpokladem této aktivity je, že lékař si nemusí umět vše sám spočítat, měl by ale umět výsledky analýz přečíst a interpretovat. Neboť jako každý nástroj, i analýzu dat lze použít nesprávně nebo ji dokonce zneužít, a tato nebezpečí je dobré znát. Bude-li zájem, nezůstaneme pouze u základů, připraveny jsou kapitoly o moderních metodách vícerozměrné analýzy, modelování, dolování znalostí, strojového učení a další.

Úvodem je nutné zdůraznit, že data v neurologii se nijak specificky neodlišují od klinických dat obecně. Metody tedy mohou být dokumentovány na číselných příkladech, které nutně nemusí souviset s neurologickým výzkumem. To ale platí pouze pro jednoduché jednorozměrné analýzy. I neurologie se totiž dostává pod tlak rostoucího množství vyšetřovaných parametrů, které dovedou podstatně zkomplikovat vícerozměrnou strukturu klinických dat. Zde pak analýza často opouští pozici pouhého nástroje k potvrzení hypotéz a stává se strategickou disciplínou přinášející nové otázky a výzvy. Tím je myšleno především dělení parametrů na ovlivňující a ovlivňované, zkoumání jejich vztahů v prediktivních analýzách a formalizace v prognostických modelech.

Hodnotné a využitelné informace však nezískáme analýzou náhodně sebraných údajů od několika pacientů. Náhoda v sobě nese i možnost chybných výsledků, což může mít v klinické praxi zvláště závažné důsledky. A tak, jako ostatně pro všechny postupy v medicíně, existují i pro analýzu dat jasná kritéria určující kvalitu a výslednou hodnotu. Za nejzávaznější lze označit výstupy prospektivních **klinických studií**. Jejich plánované a optimalizované provedení by mělo minimalizovat pravděpodobnost náhodných chyb a zajistit kontrolu možných zkreslení. Metodologie klinických studií představuje základnu pro získání skutečně průkazných a objektivních závěrů s možností zpětné kontroly. Hovoříme o **medicíně založené na průkaznosti** (**„evidence-based medicine"**) a bez jejího uplatnění si již nelze vývoj současné medicíny vůbec představit. I problematice klinických studií bude v tomto seriálu věnována pozornost, nicméně zpočátku se zaměříme na **popisnou analýzu dat** a z ní vyplývající pravidla srovnávacích analýz.

V úvodním sdělení se dále budeme zabývat daty, neboť bez nich by nebyla ani jejich analýza. Jakkoli téma „data" vypadá obyčejně a zastarale, je zcela zásadní a determinuje veškeré snažení od jejich sběru až po interpretaci. Data mohou být různá a nesou různou informační hodnotu, ke které se lze probojovat různými metodami. Neznalost v této oblasti může zcela znehodnotit celou práci. Je to skutečný základní kámen všech analýz, který zde využijeme k definici běžně užívaných pojmů.

Daty rozumíme v nejobecnějším smyslu údaje získávané o sledovaných subjektech nebo experimentálních jednotkách (pacientech, buňkách, apod). Aby data mohla být analyzována, musí být získávána v strukturované podobě, tedy jako definované **parametry** (proměnné, sledované veličiny), a v této souvislosti hovoříme o **parametrickém záznamu**. Je-li na jednom subjektu sledován jeden parametr, vzniká při měření více subjektů **jednorozměrný parametrický záznam**. Jde-li na jednom subjektu o současné měření více parametrů pro společnou analýzu, hovoříme o **vícerozměrném parametrickém záznamu**.

Analýzy z důvodu reprodukovatelnosti provádíme na definované skupině subjektů (**výběrový soubor**, **datový soubor**) a provedená šetření jsou označována jako **výběrová šetření**. Pojem **výběr** tak implikuje skutečnost, že analyzovaný soubor je většinou konečný (tedy omezený počtem vybraných subjektů), ale závěry analýzy jsou vztahovány na potenciálně nekonečnou skupinu subjektů (**cílová skupina**, např. všech pacientů s danou diagnózou). Hodnocený soubor dat je tak výběrem z cílové populace a tuto v analýze zastupuje.

Prvním krokem je logicky popis situace v cílové populaci pomocí odhadů z výběrových šetření, prováděný v tzv. **popisné, exploratorní analýze**. Smysl vyplývá již z názvu, jde o grafické a početní techniky vedoucí k vyjádření informace z dat ve srozumitelné, správné a rozsahem akceptovatelné podobě. Přesněji řečeno, často nepřehledné záznamy o jednotlivých subjektech hodnocení (**primární data**) jsou nahrazeny vypočítanými hodnotami (**sumárními statistikami**). Popis musí pravdivě odpovídat primárním datům bez ztráty podstatné informace. Vezmeme-li jako příklad primárních dat hmotnost u 100 pacientů, které z nějakého důvodu hodnotíme v jednom

Tab. 1. Definiční přehled základních typů dat a jejich sumárních statistik.

Nominální data	**Kvalitativní přiřazení příslušnosti ke kategorii, data ve formě jmenných seznamů položek**	
příklad	diagnóza jako jmenný seznam položek nebo kódů	Matematické operace
dosažitelná informace	relace typu rovnost/nerovnost; data nemají kvantitativní charakter, nelze je číselně porovnávat	$X_1 = X_2$; $X_1 \neq X_2$
statistika středu	*modus:* nejčastější hodnota (položka)	
Ordinální data	**Vzestupné nebo sestupné uspořádání intenzity vlastností objektu ve formě odstupňovaných kategorií**	
příklad	skóre rozsahu/obtížnosti onemocnění I < II < III < IV stupnice výsledků vyšetření: + < ++ < +++	Matematické operace $X_1 = X_2$; $X_1 \neq X_2$
dosažitelná informace	relace typu větší než / menší než, srovnání údajů dle velikosti vzestupně nebo sestupně	$X_1 < X_2$; $X_1 > X_2$
statistika středu	*medián:* středová hodnota, kdy 50 % všech hodnot je menší než ona = frekvenční střed	
Intervalová data	**Data vyjadřující rozdíl v intenzitě vlastnosti, umožňují kvantifikaci intervalu mezi dvěma hodnotami**	
příklad	teplota vyjádřená v °C číselné stupnice definičně zahrnující nulu	Matematické operace
dosažitelná informace	kvantifikace číselného rozdílu: informace o kolik se liší dvě hodnoty	$X_1 = X_2$; $X_1 \neq X_2$
statistika středu	*aritmetický průměr:* $(X_1 + X_2 + \ldots + X_n) / n$	$X_1 < X_2$; $X_1 > X_2$
		Rozdíl: $X_1 - X_2$
Poměrová data	**Data umožňující interpretovat i podíl hodnot vyjadřujících intenzitu vlastnosti, nulová hodnota není často slučitelná s existencí subjektu**	
příklad	hmotnost pacienta, koncentrace hemoglobinu	Matematické operace
dosažitelná informace	relace typu kolikrát větší/menší; data umožňující nejvyšší míru kvantifikace	$X_1 = X_2$; $X_1 \neq X_2$
statistika středu	*geometrický průměr:* $(X_1 . X_2 .. \ldots . X_n)^{1/n}$	$X_1 < X_2$; $X_1 > X_2$
		Rozdíl: $X_1 - X_2$
		Podíl: X_1/X_2

*Data lze také jednodušeji rozdělit na **diskrétní** (nabývají jen určitých hodnot: binárně Ano/Ne kategorie) a **spojitá** (nabývají teoreticky všech hodnot v definovaném intervalu)*

souboru, pak běžnou formou sumarizace je například výpočet aritmetického průměru doplněný uvedením hmotnosti nejlehčího a nejtěžšího jedince. Průměr zde vystupuje jako tzv. **statistika středové polohy** (centrální tendence) a rozsah (daný minimem a maximem) vyjadřuje vzájemnou odlišnost jedinců a je **ukazatelem variability** (disperze) výběrového souboru. Takto jednoduše několik málo sumárních statistik nahradí i tisíce primárních údajů a pojem **sumarizace dat** nabývá konkrétní podoby.

Přínos popisné analýzy je ale podmíněn adekvátně zvolenou sumarizací, špatná volba sumární statistiky znehodnotí celou práci! A tím se dostáváme k poslednímu bodu tohoto sdělení. Výběr adekvátní sumární statistiky se totiž řídí především typem analyzovaných dat, která v zjednodušené podobě shrnuje tab. 1. Z přehledu vyplývá, že s rostoucí mírou kvantifikace viditelně rostou informační možnosti analýzy. Matematické operace využitelné pro nižší informační stupeň mohou být použity i pro stupně vyšší, avšak vždy znamenají jistou ztrátu informace. Naopak to ale možné není, tedy nelze počítat průměr z ordinálních nebo dokonce nominálních dat.

Zásadním poučením z tab. 1 je, že různé informační typy dat mají definičně určené různé statistiky středu, které je nutné respektovat. Věc sice triviální, nicméně často zapomínaná. Nikoho by jistě nenapadlo vzít slovní výpis nalezených diagnóz a počítat jejich průměr. U ordinálních stupnic toto ale běžně vidíme jako velmi zavádějící výstup, který je zdůvodňován například dlouhodobými zvyklostmi. Přitom legitimním ukazatelem středu takových dat je nezpochybnitelně medián. Použití statistik, jejich výhodám a nevýhodám se budeme věnovat v příštím díle, který přinese již konkrétní numerické příklady. Na závěr úvodní části uveďme jednu velmi příjemnou věc – o typu sbíraných dat rozhoduje do značné míry sám badatel a má tedy takto získané i získatelné výsledky pod kontrolou již od samotného počátku sledování.