Editorial

Big data, „big“ problémy nebo účinné nástroje lepší medicíny?

Bedřich Friedecký

V databázi Medline bylo možné vyhledat v roce 2010 po zadání hesla „Big data medicine“ 44 publikací, v roce 2016 již jich bylo 644. Při zadání hesla „Big data analytics“ se objevila v roce 2010 publikace jedna, ale v roce 2016 jich bylo již 139. Jde o novou módu nebo o nový nástroj hodnocení dat (výsledků, významnosti, validity)? Pojem „big data“ (ekvivalentně metadata, velká data, veledata) se podle obvyklé definice vztahuje na soubory dat o velikosti mimo možnosti zpracování běžným software v rozumném čase. Ke zpracování „big dat“ jsou nutné datové sklady o kapacitě peta (10^{15}), exa (10^{18}) a snad v budoucnu i zettabytů (10^{21}) informací. Obecně lze věc chápat tak, že denně a nepřetržitě vznikají obrovská kvanta dat, která by bylo možné efektivně využít, pokud by se podařilo najít k tomu vhodné způsoby. V našem případě jde o možnosti a limity využití v medicíně.

Množství informací samo o sobě není úměrné jejich skutečnému významu a často je neefektivní přítěží a zdrojem problémů, které nebyly v éře před moderními informačními technologiemi známy. Běžně používáme zdroje dat, v nichž se reflektují politické, ekonomické a sociologické efekty. Jsou zdroje dat, které mají podobu komunikačních - sociálních kanálů a sítí a přitom paradoxně izolují jedince od sebe navzájem (Facebook, Tweet). Odcizování strategických dat (vypadá to, že občas překvapivě snadné), manipulace s nimi (hackerství), problém skutečného, či údajného ovlivňování veřejného mínění, data snižující integritu jedince (redukující ho na objekt reklamy a sledující jeho osobní data k nevyjasněným účelům), patří zřejmě k negativním efektům zacházení s velkými soubory dat.

Velký hadronový srážeč LHC CERN produkuje ve stavu experimentu cca 100 TB/s dat. Vynucuje si cenzuru dat a k vyhodnocení se použije cca 100/s dat (nepatrný zlomek procenta), ostatní se musí odfiltrovat.

Taková cenzura dat je známa již z dob druhé světové války, kdy britskou kontrarozvědkou dešifrovaná data z německého systému Enigma bylo nutné rovněž filtrovat systémem ULTRA. Není důvod myslet si, že v současnosti tento jev přestal existovat. Tyto aspekty naznačují nejednoznačnou a kontroverzní (a zdaleka ne jedinou) stránku problému, do kterého v současnosti medicína vstupuje.

Pátrání po využití „big dat“ v medicíně nás nevyhnutelně zavede do oblasti „P4“ medicíny (precizní, prediktivní, personalizované, participační), tedy do oblasti optimalizovaných, personalizovaných, na konkrétní skupiny pacientů orientovaných diagnostických a terapeutických postupů. Laboratorní vyšetření v této oblasti představují zejména omické metody - genomika, proteomika, metabolomika, mikrosomika, sekvence a jiné [1-6]. Oblastmi medicíny, kde má využití analýzy „big dat“ za sebou významné výsledky, jsou například onkologie a neurodegenerativní choroby. Třeba pod heslem „big data cancer“ bylo možno najít při psaní článku 1141 odkazů na Medline. Projevuje se úsilí o využívání „big dat“ v kardiologii [7] a zdá se, že v budoucnosti se práce s nimi stane regulérní součástí nejen lékařského výzkumu, ale i rutinní zdravotní péče.

Máme již nyní akceptovat práci s „big daty“ při vzdělávání? Orientace v přívalu dat v soudobé medicíně (občas používaný výraz tsunami of data), elektronizace zdravotní péče (Electronic health rekord - EHS) si nejspíš vynutí i vzdělávání v této oblasti u zdravotních pracovníků, úměrně k jejich pracovním zařazením, a to včetně „středních“ zdravotnických pracovníků [8] bez ohledu na současné utilitaristické trendy k urychlování vzdělání a k určitému ignorování moderní IT techniky ze strany některých kategorií zdravotníků.

Tématem, úzce souvisejícím s „big daty“ jsou biobanky biologických materiálů. Význam dobře zřízených a správně fungujících biobank je zřejmě nedocenitelný. Aktuálním sdělením na toto téma je zpráva vídeňského kongresu o jejich problémech [9].

Praktické aspekty práce laboratoří s „big daty“, jejich případný převod z počítačové grafiky na data klasické statistiky, korektní redukce jejich nadbytku na množství a formu, využitelné v rutinní práci, jsou například popsány v pracích, zabývajících se hmotnostní spektrometrií a proteomikou [10, 11]. V současnosti se začínají objevovat publikace, používající „big dat“ (big data approach) k prezentování výsledků z klinických laboratoří, dosažených v klasických oblastech hodnocení analytické kvality, při hodnocení preciznosti, systematických chyb, biologických, sezónních, metodických variabilit, referenčních intervalů a dalších [12, 13, 14]. Tyto studie s „big data přístupem“ jsou založeny na analýze dat, pocházejících z desítek až stovek laboratoří a často pořízené dlouhodobě trvajícími experimenty mají tak předpoklady větší validity dosažených výsledků, zejména pro harmonizační úsilí. V protikladu k tomu pozitivu je otázkou, jak se využívají obrovské soubory dat, získávané už po řadu let firmami prostřednictvím jimi organizovaných a všeobecně známých programů kvality. Jsou využívaná dostatečně a k obecným, především harmonizačním účelům?

Zásadní problém „big dat“? Zdá se, že zejména u omických metod, bez kterých se precizní a personalizovaná a prediktivní medicína nejen neobejde, ale na nichž je v podstatě založena, se tento přístup k datům stane standardním. Ovšem, „big data“ nelze jednostranně přeceňovat. Při chybném použití mohou mít škodlivé následky a naopak zhoršovat situaci [15]. Velikost dat není vždy přímo úměrná jejich významu. „Malá“ data, představovaná tradiční statistikou bývají často čistější, významnější, užitečnější. A jak nám zkušenost ukazuje, není klasická statistika silným nástrojem v rukách mnohých laboratorních pracovní-

ků. Používání „big dat“ by rozhodně nemělo být módní záležitostí, ozdobující publikované texty novými ornamenty, ale další a v některých případech nezbytnou možností racionálního využití získaných dat v medicínských procesech.

Už delší dobu bychom neměli být přehnanými optimisty při posuzování validity a reprodukovatelnosti výsledků publikovaných informací a přístup s „big daty“ může jejich slabiny buď zlepšit nebo naopak prohloubit [16]. Obojí je možné.

Literatura

1. **Hood L, Friend SH.** Precision, personalised, predictive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* 2011, 8:184-187
2. **Wu L, Sun Q, Desmeth Sugowara H, Xu Z, McCluskey K et al.** World data centre for microorganisms: an informatic infrastructure to explore and utilize preserved microbial strains worldwide. *Nucleic Acids Res* 2017. 45 (D1): D 611-D618 www.wdcm.org
3. **Vicini S, Fields O, Lai I, Litwack ED, Martin AH, Morgan TM et al.** Precision medicine in the age of big bata. The present and future role of large-scale unbiased sequincing in drug discovery and development. *Clin Pharmacol Ther* 2016, 99:198-207
4. **Wu PY, Chang CW, Kadeli CD, Venugopalan J, Hoffman R, Wang MD.** Advanced big data analysis for – omic data and electronic health records: toward precision medicine. *IEEE Trans Biomed* 2017, 64:263-273
5. **Gligorjevič V, Malod-Dogin N, Pržulj N.** Integrative methods for analyzing big data in precision medicine. *Proteomics* 2016, 16:741-758
6. **Alyass A, Turcotte M, Mayre D.** From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomic* 2015, 27:8-33
7. Clin Lab News Aug 2017. Big Data in Cardiovascular Disease
8. **O´Connor S.** Big data and data science in health care. What nurses and midwiwes need to know? *J Clin Nurs* 2017 doi:10. 1111/jicn. 14164
9. **Kinkorová J.** Mezinárodní kongres Evropský týden o biobankách. Vídeň 2016. *Čas Lék Čes* 2017, 152:211-212
10. **Awon MG, Saeed F.** MS-REDUCE: an ultrafast technique for reduction of big mass spectrometry data for high-throughput processing. *Bioinformatics* 2016, 32:1518-1526
11. **Awon MG Saeed F.** An out-of-core GPU based dimensionality reduction algorithm for big mass spektrometry data ans its application in bottom-up proteomics. *ACM BCB* 2017, 550-555
12. **De Grande LA, Goosens K, Van Uytphange K, Halsall I, Yoshimura NJ, Hens K, Thienpont LM.** Using big data to describe the effect of seasonal variation in thyroid stimulating hormone. *Clin Chem Lab Med* 2014, 55: e34-e36
13. **De Grande LA, Goosens K, Van Uytphange K, Stockl D, Thienpont LM.** The Empower project-a new way of assessing and monitoring test comparability and stability. *Clin Chem Lab Med* 2015, 53:1197-1204
14. **Zabell APR, Stone J, Randall KJ.** Using big data for LC-MS/MS analysis. *Clin Lab News* May 217
15. **Househ MS, Aldosari B, Alanti A, Kushniruk AW, Borycki EM.** Big data, big problems:A healthcare perspective. *Stud Health Technol Inform* 2017, 238:36-39
16. **Joanitis JPA.** Why most published research are false. *Plos Med* 2005, 2, e124.