BIG DATA VE ZDRAVOTNICTVÍ – PERSPEKTIVY PROBLEMATIKY

J. Hendl

Anotace

Existují různé verze definice pojmu Big data. Nejsou si však zcela nepodobné. Jedna z nich říká, že se jedná o „množiny dat, které jsou tak veliké a komplexní, že je nelze zpracovat v rámci běžných databázových systémů." (Wikipedia) Obtíže způsobuje také potřeba včasné odezvy na jejich relativně rychlou změnu.

Big data řešení představuje kombinaci speciálních databázových technologií, vhledu do řešené problematiky, analýzu dat a jejich vizualizaci. V mnoha oblastech se v současnosti experimentuje s metodami, které umožňují shromažďovat a zpracovávat ohromné objemy dat, přitom se zjišťuje, zda neobsahují skryté datové konfigurace, které mohou indikovat řešení specifických problémů, které dovolují provést predikce nebo odhalit změny.

V zdravotnictví se setkáváme s problémy souvisejícími s velkými objemy strukturovaných a nestrukturovaných dat velmi často v důsledku digitalizace existujících dat a generování nových dat. To zahrnuje lékařské záznamy o pacientech, radiologické snímky, data o klinických pokusech, data o lécích, data z humánní genetiky a o populacích, genomiky atd. Novější formy dat se týkají 3D zobrazování, data z genomických a biometrických senzorických zařízení a dat z „chytrých" telefonů a elektronických náramků.

Příspěvek upozorňuje na trendy a perspektivy řešení problematiky Big data.

Klíčová slova

Big data, data mining, lékařské rozhodování, digitalizace, strukturovaná a nestrukturovaná data

1. Úvod

V posledních letech dochází v celém světě k dosud nevídanému nárůstu množství a různorodosti generovaných dat. Objem informací se ve světě zdvojnásobuje každé dva roky. Někdy je tento jev označován jako „datové tsunami". Podporuje ho rozšiřování sociálních médií, technických prostředků komunikace a snímání dat (mobilní telefony, senzory, internet of things, IoT). Souvisí také s využíváním dat v oblastech jako komerční sféra nebo zdravotnictví. Příklady jsou mikroblogování na Twitteru, kde se zpracovává denně 12 TB dat, nebo Facebook, který přijímá denně 500 milionů zpráv. Digitální universum obsahuje v současnosti přibližně 3 zetabytů. Jestliže v roce 2000 se uchovávalo 25% dat, dnes je to 90%. Cisco skupina předpokládá, že v roce 2015 bude připojeno k internetu 25 miliard hardwarových prvků. Rychle vznikají velké a komplexní soubory dat, která nazýváme Big data.

V našem příspěvku uvedeme vymezení a popis vybraných aspektů konceptu Big data a způsoby, jak se tento koncept uplatňuje. V další části se budeme zabývat problémem využití velkých dat v medicíně.

2. Big data – vlastnosti a koncepty

Big data informatické aplikace se týkají především zvládání velkých objemů dat, ale také mixu různých typů dat (různorodost dat) a toho, jakou roli při jejich vzniku hraje čas (rychlost). Tyto charakteristiky Big data se obvykle v anglické literatuře nazývají 3V (volume, variety, velocity). V tabulce 1 uvádíme jejich charakteristiky.

Aspekt	Charakteristika
Objem	Hlavní aspekt, v posledních letech se nesmírně zvětšilo množství generovaných dat. Nepředstavuje však hlavní obtíže.
Různorodost	Mnoho rozličných formátů dat, od strukturovaných dat po nestrukturovaná data.
Rychlost	Rychlost změny dat. Zvyšuje se množství dat, které se musí rychle uložit a zpracovat.

Tabulka 1 – Základní vlastnosti Big data: objem, různorodost, rychlost

Objevují se různé definice pojmu Big data. Opírají se obvykle o uvedené tři vlastnosti. Jedna z přijímaných definic říká, že Big data jsou data s různorodým formátem, o velkém objemu a rychle se měnící, což v souhrnu způsobuje, že je nelze spravovat pomocí konvenčních databázových prostředků. Vazbu mezi vlastnostmi Big data a technologickými změnami ukazuje tabulka 2.

Aspekt	Možnosti a technologie
Objem	Virtualizace ukládání do cloudů v datových centrech, množství prvků připojených k internetu.
Různorodost	Existuje potřeba analyzovat i nestrukturovaná data. Uplatnění databází NoSQL.
Rychlost	Miliony připojených chytrých telefonů a senzorů zvyšují objem i rychlost změn.

Tabulka 2 – Technologie a Big data

Big data se objevují v mnoha oblastech zpracování informací. Příkladem z oblasti internetu jsou údaje o logování na stránkách Google nebo Facebook. Aplikace Big data nejsou nikdy izolovaným řešením. Vyžaduje se určitá infrastruktura. Jde o to propojit technické a softwarové aspekty za účelem využívání informací. Data se musí shromáždit, organizovat a integrovat. Aplikace integrovaných systémů typu Big data se objevují v obchodě, v řízení výroby, ve zdravotnictví, řízení dopravy atd. Například v procesu výroby redukuje aplikace Big data počet výpadků. V dopravě takové aplikace řídí dopravu a zmírňují pravděpodobnost zácpy ve městě, snižují spotřebu nebo zabraňují zvýšení nečistot v ovzduší.

V přehledu uvádíme pro orientaci jednotky objemu informací, které se používají v informatice a v oblasti Big data:

- Kilo- znamená 1,000; kilobyte je tisíc bytů.
- Mega- znamená 1,000,000; megabyte je milion bytů.
- Giga- znamená 1,000,000,000; gigabyte je miliarda bytů.
- Tera- znamená 1,000,000,000,000; terabyte je trilion bytů.
- Peta- znamená 1,000,000,000,000,000; petabyte je 1,000 terabytů.
- Exa- znamená 1,000,000,000,000,000,000; exabyte je 1,000 petabytů.
- Zetta- znamená 1,000,000,000,000,000,000,000; zettabyte je 1,000 exabytů.
- Yotta- znamená 1,000,000,000,000,000,000,000,000; yottabyte je 1,000 zettabytů.

Data jsou přirozeným zdrojem, který je možné využívat. Základní funkční vlastnosti při práci se soubory typu Big data jsou popsány schématem na obr. 1.

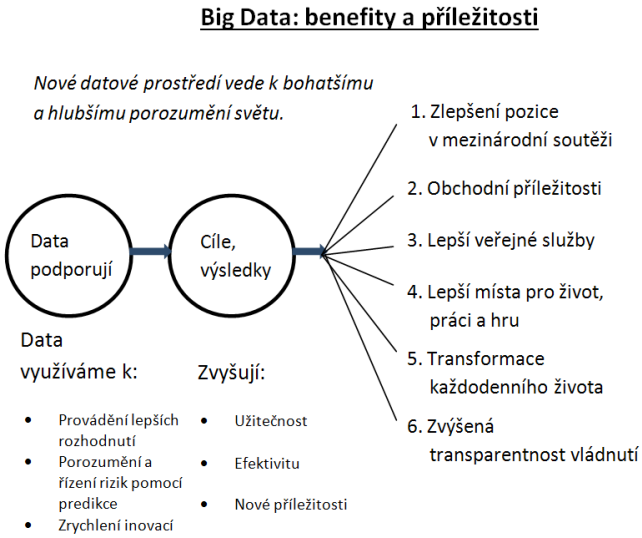
Obrázek 1 – Základní funkční vlastnosti při práci se soubory typu Big data

Data musí být nejdříve získána a zachycena, pak musí být organizována a integrována. V závislosti na problému se analyzují, nakonec se na základě výsledků provádí rozhodnutí a doporučená akce. Například firma Amazon může na konci procesu zpracování doporučit zákazníkovi nějakou knihu na podkladě jeho dřívějších objednávek nebo mu poskytne slevový kupon. Přitom upravuje systém svá doporučení k akci podle chování milionů podobných zákazníků. Ačkoliv se zdá, že tento proces je přímočarý, obsahuje nuance, které celý proces komplikují. Jestliže například kombinujeme více zdrojů dat, musíme je validizovat a rozhodnout, zda jejich integrace má smysl pro celé řešení.

Aplikacím typu Big data se předpovídá velká budoucnost s dopadem i do státní politiky. Například Alex Pentland (12), znalec scény Big data, odborník v oblasti informatiky a sociálních vztahů, prohlásil, že „s Big data se můžeme začít dívat na podrobnosti sociálních interakcí a na to, jak se odehrávají. Nejsme

již omezeni průměry, jakými jsou tržní indexy nebo výsledky voleb. To je velká změna. Schopnost vidět detaily trhu nebo politických revolucí a schopnost je predikovat nebo je ovlivňovat je skutečně jako Prometheův oheň – můžeme jej využít pro dobro nebo zlo, takže s Big data přicházejí zajímavé časy. Technologie Big data neznamenají evoluci, ale revoluci."

Big data zvyšují příležitostí propojovat a sdílet informace, vytvářet nové znalosti a podporovat inovace. Pomáhají jedincům využívat informační technologie k pozitivním změnám. Stát má podporovat šíření technologií Big data. Na úrovni státu a jeho politik lze možnosti v tomto směru popsat schématem na obrázku 2.



Obrázek 2 – Přínosy technologií Big data pro stát

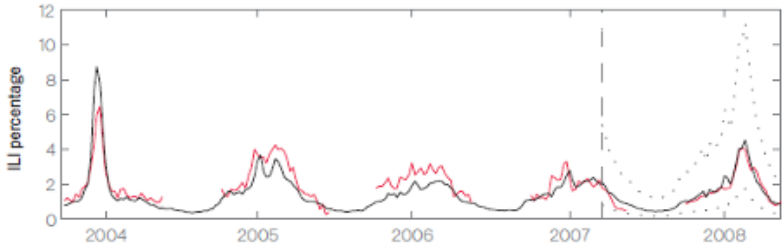
3. Big data ve zdravotnictví

Důkladnou analýzou uplatnění konceptů Big dat ve zdravotnictví se zabývaly asi jako první týmy výzkumníků z McKinsey institutu ve své obecně pojaté zprávě z roku 2011 (10) a speciální zprávě z roku 2013 (6). Provedly predikce vývoje pro USA a pro některé oblasti aplikací i pro Evropu. V obou zprávách se upozorňuje, že ve zdravotnictví vývoj směřuje k digitalizaci medicínských záznamů o pacientovi (EMR, Electronic Medical Record), farmaceutické firmy a další organizace ukládají informace o výzkumu a činnostech do elektronických databází, státní organizace urychlují změny tak, že uložená data poskytují zájemcům z oblasti zdravotnictví ve využitelné formě. Celý proces dosáhl určitého hraničního bodu, který vyžaduje uplatnění nových informačních technologií. Cesty řešení pro Evropu naznačuje Zillnerová et al. (19).

Odborníci z farmaceutického průmyslu, pojišťovny, stát a poskytovatelé zdravotnických služeb začínají analyzovat velké soubory dat, aby získaly pomocí nich nové poznatky a vhled do složitých problémů, se kterými se musí vyrovnat. Většinou se jedná o první kroky. Například výzkumníci analyzují data, aby rozpoznali, které ošetření je efektivní za daných podmínek, identifikují vzorce vedlejších účinků nebo důvody pro opětovné přijímání pacienta do nemocnice. Získávají tak vhled, který může zlepšit zdravotní péči a snížit náklady. Existují informační technologie, které jim v tomto směru pomáhají.

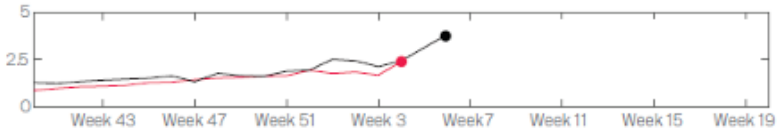
Na možnosti uplatnění analýzy velkých dat upozornil širší zdravotnickou veřejnost a zástupce politického života projekt internetové firmy Google v souvislosti s možnostmi monitorování výskytu chřipkového onemocnění (5). Epidemie sezonní chřipky jsou velkým problémem systému veřejného zdravotnictví, které ročně ve světě vede v 250 000–500 000 případech k úmrtí. Nové typy virů chřipky, proti kterým neexistuje imunita, mohou vést k pandemii s mnoha miliony úmrtí. Včasná detekce aktivity onemocnění vede k rychlejší reakci a může redukovat dopad jak sezónní tak pandemické chřipky. Jeden ze způsobů včasné detekce představuje monitorování vyhledávacího chování uživatelů se vztahem ke zdraví při použití internetu. Google společnost navrhla metodu analýzy velkého počtu Google dotazů s cílem monitorovat nemoci podobné chřipce. Některé dotazy jsou silně korelované s procentem návštěvy lékaře, při kterých pacient prezentuje symptomy chřipky. To umožnilo navrhnout algoritmus pro odhad aktivity chřipky v jednotlivých oblastech USA s odstupem jednoho dne. Tento postup umožňuje využít Google dotazy k detekci chřipkové epidemie v oblastech s větší populací, která má přístup k internetu.

Služba se označuje zkratkou GFT (Google Flu Trends). GFT porovnává své predikce s historickou základní úrovní chřipkové aktivity pro danou oblast a pak rozhoduje, zda se jedná u aktuální aktivity o minimální, nízkou, střední, vysokou nebo intenzivní aktivitu. Získané odhady podle firmy Google velmi dobře korelují s konvenčními epidemiologickými daty (CDC, Centers for Disease Control and Prevention), jak na národní tak na oblastní úrovni.



Obrázek 3 – GFT modelová data a přerušovaná CDC data o chřipkové aktivitě

Obrázek 3 ukazuje GFT data v oblasti (Středoantlantická oblast v USA) a CDC přerušovaná data s dosaženou korelací 0,96. Také je uveden 95% pás predikce.



Obrázek 4 – Podrobné srovnání predikcí chřipkové aktivity v letech 2007-2008

Obr. 4 ukazuje podrobné srovnání predikcí chřipkové aktivity v letech 2007 – 2008. V 5. týdnu byl detekován strmý nárůst signálu, který byl později potvrzen daty z CDC institutu.

Prvotní motivace pro GFT spočívala v úsilí o včasnou identifikaci aktivity nemoci a rychlou reakci. V jedné zprávě se dokazovalo, že GFT signál detekoval zvýšený výskyt chřipky až o 10 dní dříve než tuto skutečnost ohlásila služba CDC.

Google signál GFT je příkladem „kolektivní inteligence", které je možné využít k identifikaci trendů a k výpočtu predikcí. Chování lidí na internetu ukazuje jejich vůli a potřeby bez omezení. Odborník T. W. Malone ze Sloan School of Management na MIT se ve svém komentáři vyjádřil: „Zdá se, že takový způsob využití dat vytvořených nezáměrně uživateli Googlu nám odhalí vzorce chování, které by jinak byly neviditelné."

V přehledu uvádíme, že Big data ve zdravotnictví tvoří:

- Klinická data získávaná z elektronických záznamů ambulancí, nemocnic, zobrazovacích center, laboratoří, lékáren atd.
- Data o nákladech na zdravotnickou péči.
- Biometrická data získaná monitorovacími přístroji na dálku. Jde o váhu, krevní tlak, úrovně glykemie atd.
- Data získaná od pacientů o úrovni spokojenosti se zdravotním stavem, vlastním monitorování spánku, jídla, pohybu.
- Genomické informace, které se zlevňují v důsledku rozvoje technologií.
- Další determinanty jako socioekonomický status nebo faktory prostředí.

McKinseyho zpráva (10) předpokládá, že Big data technologie mohou redukovat náklady a zvýšit efektivitu v těchto oblastech: 1) klinické rozhodování při určování diagnózy a volbě terapeutického postupu, 2) ve výzkumu a 3) ve veřejném zdravotnictví. Stručně popíšeme jednotlivé perspektivní aplikace v jmenovaných oblastech.

1. Big data procedury pomáhají řešit problémy srovnávání efektivnosti a finanční náročnosti diagnostických a terapeutických postupů. Tvoří tak důležitý úkol v souvislosti se snižováním celkových nákladů při zachování úrovně lékařské péče.
2. V oblasti výzkumu jde o prediktivní modelování s cílem snížit náklady na neperspektivní léčiva a přístroje, vývoj prostředků a algoritmů ke zlepšení schémat klinických pokusů, aby bylo možné individualizovat terapii a získávání pacientů s cílem urychlit přenos způsobu terapie

do terénu, analýzu výsledků klinických pokusů a záznamů od pacientů s cílem identifikovat potřebu dalšího sledování a odhalit vedlejší účinky dříve než se dostane produkt na trh.

3. Procedury Big data se využijí ve veřejném zdravotnictví s cílem zlepšit popis a analýzu šíření onemocnění s cílem urychlit reakci zdravotnického systému, urychlit vývoj cílené vakcinace, transformovat velké množství dat do akčních informací s cílem zlepšit identifikaci potřeb, efektivnost služeb a predikování krizí v systému.

Oblasti, kde mohou Big data a jejich analýza nejvíce přispět k zdravotnické péči jsou:

- identifikace pacientů, kteří jsou největšími konzumenty zdravotnických zdrojů nebo kteří mají velké riziko vedlejších účinků,
- poskytování informací jedincům, aby se mohli poučeně rozhodnout a efektivněji využívat informace pro zlepšení svého zdraví, změnu chování a vlastní sledování,
- identifikace postupů, programů a intervencí, které nepřinášejí prokazatelně benefity nebo stojí příliš mnoho,
- redukce opakovaných přijetí tím, že se identifikují environmentální faktory nebo faktory životního stylu a navrhne se odpovídající úprava režimu,
- zlepšení výsledků při domácím monitorování zdravotního stavu,
- ovlivňování zdraví populace detekováním zranitelných míst v populacích pacientů během epidemií nebo přírodních katastrof,
- propojení klinických, finančních a dat o řízení s cílem analyzovat využití zdrojů efektivně a v reálném čase.

4. Big data v klinické péči o pacienta

Pojednáme výhody ze zpracování Big data s cílem zlepšit klinickou péči. Dobrým příkladem pro tyto efekty může být výzkum, kdy lékaři z Harvard Medical School a Harvard Pilgrim Health Care demonstrovali potenciál analýzy elektronických zdravotních záznamů pacienta při detekci a kategorizaci pacientů s podezřením na diabetes při zdravotní kontrole. Výzkumníci použili program k prozkoumání za čtyři roky nashromážděných záznamů pacientů z velké mnoha oborové ambulance a označili pacienty s podezřením na diabetes, přičemž využili jejich laboratorní výsledky, diagnostické kódy a dosavadní předepsané léky. Algoritmus úspěšně identifikoval pacienty s diabetem a rozlišil také mezi diabetem I. a II. typu, přičemž byl úspěšnější při identifikaci pacientů pomocí informací bez použití diagnostických kódů (17).

Koncept Big data hraje významnou roli v současném klinickém a medicínském výzkumu a uplatňuje se v klinických studiích. NIH (National Institue of Health) investoval 100 milionů dolarů do iniciativy BD2K (Big data to Knowledge). BD2K definuje biomedicínská Big data jako velké množiny dat generovaných ve výzkumu nebo jako velké datové množiny, které vzniknou agregováním menších datových množin.

Taková data vznikají v rámci studií **trendu a prevalence**, studií o rizikových faktorech a v studiích **genotypu a fenotypu**.

Jednou ze známých aplikací s využitím Big data je analýza prevalence a trendů nemocí. Například Elshalzly et al (4) prozkoumal 1,3 milionů dospělých jedinců v USA s poruchami lipoproteinů cholesterolu. Chan a McGarey (2) v přehledu popisují, jak takové množiny dat analyzovat, aby bylo možné dosáhnout závěrů týkajících se populací. Přitom si všímají sekulárních trendů, geografických oblastí, zdravotních rozdílností atd.

Klinická Big data se využívají také pro zjištění kauzality nebo statistických vztahů mezi rizikovými faktory a nemocí. Například Zhang et al (18) zkoumal klinická data 16 135 dospělých pacientů a věnoval se vztahu mezi glykemií dávkami inzulínu ve vztahu k mortalitě.

S pokrokem v technologii určování genotypu roste počet studií zkoumajících asociací na genetické úrovni genových expresí a genomických dat získaných od pacientů a kontrol. Například se využily klinická a genetická data od 5700 pacientů, kteří dostávají walfarin s cílem navrhnout algoritmus pro určení vhodné dávky walfarinu (9). Tato a další podobné studie jsou podobné studiím rizikových faktorů, ale pracují s mnohem většími objemy dat.

Velký počet studií využívá big data, aby bylo možné zdůvodnit zavedení nových metod a technik. Například Hill et al (7) navrhl rozhraní pro využití klinických dat pro odhad rizika různých nemoci, aby se usnadnila evaluace různých způsobů ošetření. Stephen et al. (13) navrhli pomocí klinických dat z datových úložišť algoritmus ke kategorizaci pediatrických pacientů s určitými respiračním poruchami do různých subtypů. Vzestup počtu takových studií ukazuje rostoucí zájem výzkumníků o analýzu velkých objemů klinických dat s cílem navrhnout pomocí Big data asistovaná klinická doporučení.

Využití velkých souborů dat se stalo častým jevem v mnoha oblastech medicínského výzkumu. Použité datové množiny se často liší různými parametry. Datové množiny s expresí genů získaných novými metodami sekvencování jsou obvykle značně veliké, kdežto data z klinických studií jsou ve srovnání s nimi menší. Sinha et al. (16) rozděluje velké datové soubory do tři různých typů, přičemž uvažuje p-dimenzionalitu (počet atributů) a n-počet případů:

1. Velké n a malé p,
2. Velké p a malé n,
3. Velké n a velké p.

Díky pokroku informačních technologiích lze dnes dobře pracovat se soubory prvního typu. Většina klinických dat jsou však spíše typu 2. Například Drai et al (3) zkoumal 69 afázických pacientů, které testoval pomocí až 6000 stimulačních vět (p=6000). Většina shlukovacích přístupů v takovém případě nefunguje dobře, protože problémový prostor není nasycen jednotlivými případy. Aby bylo možné takový problém řešit, je nutné provést pomocí kompresních metod určitou redukci dimenzionality.

Při volbě metod pro manipulaci s klinickými velkými soubory si všímáme tří oblastí:

- Technologie pro uložení a organizaci dat,
- Metodologie základní manipulace s daty před vlastní analýzou,
- Statistické postupy analýzy.

V rámci Big data řešení se používají alternativní databáze jako Hadoop nebo NoSQL nebo zvláštní moduly známých statistických systémů jak SAS, SPSS nebo R.

Provádí se předzpracování pomocí expertů nebo se využívají speciální algoritmy pro kompresi dat nebo normalizaci. Tyto výpočetní metodologie mohou zanést do dat systematické zkreslení a způsobit problémy s integrací. V této fázi přípravy dat se také využívají různé techniky vizualizace.

Pro analýzu klinických dat se využívají metody jako mnohonásobná regresní analýza, logistická regresní analýza, shlukovací metody a metody typu regresních a klasifikačních stromů. Také se používají metody neurálních sítí, naivní Bayesova metoda, Markovovy modely časových řad, atd. Často se provádí vícečetné statistické testování, proto je nutné aplikovat korekce na hladinu významnosti a využívat teorii simultánního testování.

4. Předpoklady pro uplatnění konceptu Big data ve zdravotnictví

Připomínáme, že Big data mají velká omezení. Tato omezení zahrnují adekvátnost, přesnost, úplnost, povahu zdroje a další atributy kvality dat (15). Uživatelé velkých souborů dat se vypořádávají s velkým množstvím úkolů, které z toho vyplývají.

Je zapotřebí zvládnout problémy s datovými soubory v předchozím odstavci uvedeného typu 2 s velkou dimenzionalitou, ale menším počtem případů. Důležitým omezením je selekční systematická chyba a je nutné s ní počítat i u velkých souborů dat s velkým počtem případů. V tomto směru jsou omezené mnohé klinické studie. Také je nutné počítat se scházejícími hodnotami, což je u klinických studií častý případ. Rovněž stochastická závislost mezi jednotlivými případy i proměnnými hraje roli při analýze velkých souborů. Stává se to například tehdy, jestliže jednoho pacienta vyšetřujeme několikrát.

Upozorňuje se na to, že analýza velkých souborů není často optimální, protože výzkumníci nejsou obeznámeni s možnými nástroji a metodologiemi. Při analýze velkých dat je totiž nutné uvažovat nejenom aspekty jako způsob sběru dat, obhospodařování dat (curation), jejich extrakci, integraci, intepretaci, imputaci, ale také výběr vhodných statistických metod. Je nutné vyvíjet i nové adekvátní statistické metody.

Problémem Big data jsou různé formáty dat. Analýza dat ve zdravotnictví se musí přizpůsobit objemu, rychlosti a jejich různorodosti. Nepracuje se jenom se strukturovanými daty. Data mají dnes často multimediální a nestrukturovaný charakter. Různorodost dat představuje speciální problém velkých soborů dat. Strukturovaná data je možné elegantně uchovávat, organizovat

a analyzovat. Jedná se o výstupy z přístrojů a číselné zápisy do záznamu o pacientovi převedené do elektronické formy. Nestrukturovaná data vznikají v místě péče o pacienta, jde o poznatky doktorů a dalšího zdravotnického personálu, předpisy, výstupy ze zobrazovacích technik.

Do zdravotnictví pronikají datové výstupy z různých fitness přístrojů, genetiky a genomiky, sociálních medií a dalších zdrojů. V současnosti se zpracovává pouze malá část těchto dat.

Aplikace ve zdravotnictví, zvláště ty, které se týkají péče o pacienta, vyžadují efektivnější cesty, jak kombinovat a konvertovat různorodá data včetně automatizace konverze strukturovaných a nestrukturovaných dat.

Strukturovaná data z medicínských a zdravotních záznamů obsahují známé proměnné, jako pacientovo jméno, datum narození a další zakódované informace, které se dnes ukládají v standardních databázích. Potřeba kódování dat při ošetřování je pro lékaře a další ošetřující personál velkou překážkou pro přijetí elektronického záznamu. Na druhé straně většina informatiků má názor, že chyby se minimalizují digitálním vstupem a ne ručně provedenými poznámkami. Potenciál Big data aplikací je v tom, že kombinují tradiční data s novými formami dat, jak na individuální tak na populační úrovni. Již dnes jsou dostupné možnosti, kdy data z několika zdrojů podporují rychlejší a spolehlivější výzkum. Například farmaceutické firmy by mohly integrovat klinická data o populacích s genomickými daty, tento vývoj může urychlit schválení nových lékových terapií a jejich aplikaci správně vybraným pacientům (14).

Kvalita dat je hlavním úkolem ve zdravotnictví ze dvou důvodů: rozhodnutí se vztahem k životu a smrti závisí na přesné informaci a kvalitě medicínských dat, včetně těch nestrukturovaných. U nich právě platí větší prevalence chybovosti.

Špatný překlad nebo převod rukou provedených poznámek je velmi častý.

Věrohodnost dat předpokládá simultánní souhru hardwarové architektury a algoritmů, metodologií a nástrojů, aby se vyhovělo požadavkům Big data. Architektura analytik a nástrojů pro strukturovaná a nestrukturovaná data se velmi liší od tradičních obchodních aplikací. Analytika v Big data v prostředí zdravotní péče se provádí v distribuovaném prostředí na několika serverech, přičemž se využívá paradigma paralelního zpracování a přístupu „rozděl a zpracuj". Podobně modely a techniky jako dolování z dat a statistické přístupy, algoritmy a vizualizační techniky vyžadují vzít v potaz charakteristiky Big data. Tradiční přístupy předpokládají, že uložená data jsou jistá, čistá a přesná. Důvěryhodnost jako výsledek takového zpracování ve zdravotnictví stojí v mnoha ohledech před podobnými problémy jako zpracování finančních dat, zvláště na straně plátců: jedná se o správného pacienta, nemocnici, plátce, typ platby, částku? Ostatní aspekty důvěryhodnosti jsou jedinečné pro zdravotní péči: Jsou diagnosy, ošetření, medikace, procedury a výsledky správně zachyceny?

Zlepšení koordinace péče, minimalizace chyb a redukce nákladů závisí na vysoké kvalitě dat, stejně tak bezpečnost a účinnost medikace, diagnostická správnost a přesné zacílení procesů nemoci pomocí ošetření (14).

5. Závěr

Možnosti Big data znamenají nové výzvy. Abychom realizovali potencionální přísliby velkých dat, je nutné se věnovat obecným požadavkům, jako zlepšení standardizace dat a interoperability, podpoře sdílení dat, dohledu nad daty, zlepšení analytických metod, zajištění kvalifikovaných pracovníků a zkoumání cest, jak využívat hodnotu Big data pro zlepšení práce kliniků a k prospěchu pacientů. Vytvoření vědomí o benefitech z využívání velkých dat ve zdravotnictví musí být spojeno se zvýšenou ochranou soukromí a bezpečnosti dat.

Využití Big data ve zdravotnictví přinese výhody jednotlivým lékařům v jejich ordinacích, skupinám uživatelů, nemocnicím, sítím nemocnic nebo jejich zřizovatelům. Potencionální benefity zahrnují detekci nemocí v jejich první fázi, když je možné je efektivně léčit, ovlivňování specificky určené oblasti zdraví jedinců a populací, detekci pojišťovacích podvodů. Pomocí analýzy Big data bude možné zodpovědět různé zajímavé otázky. Na základě historických dat bude možné predikovat/odhadovat určitý vývoj nebo stav, například trvání hospitalizace, vyhledávat pacienty, kteří potřebují určitý chirurgický zákrok, pacienty, kteří nebudou mít z chirurgického zákroku žádný profit, komplikace, pacienty s rizikem komplikací, pacienty s rizikem sepse nebo iatrogenních onemocnění, pacienty s progresí nemoci, kauzální faktory progrese určité nemoci.

MIT (Massachusetts Institute of Technology) uspořádal v roce 2014 diskusi o problémech s Big data na jednotkách intenzivní péče (1). V diskusi se projevil názor obecné platnosti, že doporučení EBM vychází převážně ze znáhodněných kontrolovaných klinických pokusů (RCT), které fungují jako zlatý standard evidence. Většina klinických situací však zůstává těmito standardy nepokryto. Klinikové musí stále jednat na základě teorie a zkušenosti. Teorie zůstane neadekvátní, pokud nebude potvrzena a systematicky převedena do praxe. V tomto bodě může podat pomocnou ruku analýza velkých souborů dat, které jsou na intenzivních jednotkách (a obecně v nemocnicích) shromažďovány. Zodpoví se tak mnohé otázky, které RCT nemohou zodpovědět. Naopak hypotézy generované v explorativním přístupu k velkým datům, mohou být potvrzovány pomocí RCT. RCT trpí řadou vlastností, které omezují jejich využití. Jedná se především o vysokou finanční náročnost a omezení výsledků na několik málo subpopulací. RCT mohu vyřešit pouze přesně určené výzkumné otázky. To je také omezující. V mnoha případech nelze RCT provést z etických důvodů. V tomto okamžiku mohu pomoci retrospektivní analýzy velkých souborů typu Big data.

Práce s Big data se stala častou vlastností biomedicínských studií. V dnešní době je možné generovat terabyty dat v relativně krátkém časovém okamžiku. Dostupnost velkých objemů dat je možné díky moderním technologiím ukládání a organizování dat. Výzkumníci při práci s velkými daty si musí osvojit nové techniky. Metodologie práce s Big data musí udržet krok se schopností jejich získávání a ukládání. Potřebě vzdělávání studentů medicíny v konceptu Big data se věnuje Moskowitz (11). Na negativní stránky přeceňování konceptu Big data v medicíně upozorňuje Huang (8).

Literatura:

[1.] BADAWI, O. et al. (2014) Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference. JMIR Med Inform 2014;2(2):e22) [FREE text] doi:10.2196/medinform.3447

[2.] CHAN L, MCGAREY P. (2012) Using large datasets for population-based health research. In: Gallin JI, Ognibene FP. eds. Principles and Practice of Clinical Research. 3rd ed. Maryland Heights, MO: Elsevier, Inc; p. 371–381

[3.] DRAI, D., GRODZINSKY, Y. (2006) A new empirical angle on the variability debate: quantitative neurosyntactic analyses of a large data set from Broca's aphasia. Brain Lang 2006;96(2):117-128. [doi: 10.1016/j.bandl.2004.10.016]

[4.] ELSHAZLY, M.B. et al. (2013) Non-high-density lipoprotein cholesterol, guideline targets, and population percentiles for secondary prevention in 1.3 million adults: the VLDL-2 Study (very large database of lipids). J Am Coll Cardiol 2013 Nov 19;62(21): 1960–1965. [doi: 10.1016/j.jacc.2013.07.045]

[5.] GINSBERG, J. et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457 (7232): 1012–1014.

[6.] GROVES, P. et al. (2013) The 'big data' revolution in healthcare – Accelerating value and innovation. USA: McKinsey Global Institute. [FREE text]

[7.] HILL, B. et al. (2013) Exploring the use of large clinical data to inform patients for shared decision making. Stud Health Technol Inform 2013;192:851-855.

[8.] HUANG, X. et al. (2015) Big data – a 21st century science Maginot Line? No-boundary thinking: shifting from the big data paradigm. BioData Mining (2015) 8: 7 [FREE text] DOI 10.1186/s13040-015-0037-5

[9.] INTERNATIONAL WARFARIN PHARMACOGENETICS CONSORTIUM, KLEIN, T.E. et al. (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med 2009;360(8):753-764 [FREE text] [doi: 10.1056/NEJMoa0809329]

[10.] MANYIKA, J. et al. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. USA: McKinsey Global Institute. [FREE text]

[11.] MOSKOWITZ, A. et al. (2015) Preparing a New Generation of Clinicians for the Era of Big Data. Harv Med Stud Rev. 2015 January; 2(1): 24–27. [FREE text]

[12.] PENTLAND, A. (2012) Reinventing society in the wake of big data [staženo 6.7. 2014, http://edge.org/conversation/reinventing-society-in-the-wake-of-big-data]

[13.] STEPHEN, R et al. (2003) Feasibility of using a large Clinical Data Warehouse to automate the selection of diagnostic cohorts. AMIA Annu Symp Proc 2003:1019 [FREE text]

[14.] RAGHUPATHI, W., RAGHUPATHI, V. (2014) Big data analytics in healthcare: promise and Potential. Health Information Science and Systems 2014, 2:3 [staženo: http://www.hissjournal.com/content/2/1/3] [FREE text]

[15.] SANDERS, C.M. et al. (2012) Understanding the limits of large datasets. J Cancer Educ 2012;27(4):664-669. [doi: 10.1007/s13187-012-0383-7]

[16.] SINHA, A. et al. (2009) Large datasets in biomedicine: a discussion of salient analytic issues. J Am Med Inform Assoc 2009;16(6):759–767 [FREE text] [doi: 10.1197/jamia.M2780]

[17.] TOH, S., PLATT, R.. (2013) Big data in epidemiology: too big to fail? Epidemiology. 2013 Nov;24(6):939. [FREE text]

[18.] ZHANG, Y, HEMOND, M.S. (2009) Uncovering the predictive value of minimum blood glucose through statistical analysis of a large clinical dataset. AMIA Annu Symp Proc 2009;2009:725–729 (FREE text: Medline)

[19.] ZILLNER, S. et al. (2014) Roadmap for Big Data Applications in the Healthcare Domain. [FREE text] [staženo20.2.2015: http://bigproject.eu/sites/default/files/2014_IEEE%20IRI%20HI_Towards%20Technology%20Roadmap%20Big%20Data_final.pdf]

Kontakt:

Prof. Jan Hendl
FSV UK – katedra sociologie
U Kříže 8 a 10
158 00 Praha 5 – Jinonice
e-mail: *jan.hendl@fsv.cuni.cz*