

O datech

Michal Kulich

10. února 2021

Pojednání na téma „Jaká data by chtěli statistici mít, aby s nimi byli spokojeni“ původně sepsané pro členy skupiny BISOP.

Hlavní je toto: chcete-li mít data, z kterých dokážete spočítat nebo odhadnout, jak to doopravdy je (s kovidem nebo s čímkoli jiným) musíte mít data

- (a) *úplná*, tj. nic k nim nejde přidat a nic nebylo ubráno, nebo
- (b) data *neúplná*, která ale tvoří *reprezentativní* vzorek úplných dat.

Reprezentativnost znamená, že každý kousek dat, který by do nich mohl být vložen, má stejnou pravděpodobnost, že v nich skutečně bude.

Jestliže data nejsou úplná ani reprezentativní, pak *lžou* — nejde z nich zjistit skutečnost. Na tento fakt se neustále v praxi „zapomíná“, protože získat reprezentativní data je složité a nepohodlné. Mnohem pohodlnější je vzít prostě nějaká data, která jsou zrovna někde na disku v počítači, a na princip reprezentativnosti zapomenout. Výsledek je, že značná část práce, kterou lidé s takovými daty dělají, není nic víc než vaření z vody nebo černá magie bez ohledu na to, jak komplikované věci s takovými pseudodaty provádějí.

Data, která má ÚZIS, jsou přesně tohoto druhu. Výsledky testů jsou známy pouze od lidí, kteří se dobrovolně nechají otestovat nebo je jim test předepsán. Přitom se počítají dohromady PCR testy s testy antigenními, což zase porušuje jiný zásadní princip — konsistenci dat (všechny položky jsou měřeny stejnými metodami za stejných podmínek). O lidech, kteří nepřišli na test, nevíme nic. Podíl lidí, kteří se nenechali otestovat se odhaduje z počtu hospitalizovaných kovid+ lidí, kteří předtím neměli test. To ovšem také není zrovna reprezentativní výběr z těch, kdo se nenechali otestovat, že. A tak dále, není třeba dávat další příklady.

To nutně neznamena, že data, která ÚZIS sbírá, jsou zcela bezcenná. Nikoli, jsou velmi potřebná, ale musí se s nimi zacházet opatrně a při vědomí, co za data to je. Některé věci (hrubé krátkodobé časové trendy) se na nich vypořizovat dají, ale opravdu jen hrubě a s velkou nejistotou, jestli data neukazují chiméry. Mezi všemi, kdo s těmi daty pracují, jsem neviděl nikoho, kdo by tohle přiznal: buď si toho nejsou vědomi anebo jim to nevadí a nehodlají přiznávat jak málo toho vlastně vědí.

Nemáme žádný „seznam dat, která potřebujeme“. Potřebujeme aspoň nějaká data, která netrpí těmi problémy zmíněnými výše, tj. jsou sbírána na pečlivě vybraných vzorcích podle předem připraveného plánu a naprosto konzistentními metodami, které minimalizují možnost chyb a omylů. Takhle se mají dělat pořádné studie. Přitom každá taková studie musí mít jasně stanovený účel, který determinuje, jaká data je potřeba sbírat, kdy a na kom.

Neexistuje nic jako univerzální data, která geniálně nasbíráme a pak v nich nalezneme odpovědi na všechny problémy světa. Každá data jsou dobrá jen pro něco a specifikace toho „něčeho“ musí předcházet jejich sběru.

Takže voláme po tom, aby byly takové studie naplánovány a zahájeny a aby ten úkol byl svěřen lidem, kteří vědí, o co jde a mají s tím zkušenosti. To neznamena jen statistiky — jde o týmovou interdisciplinární práci, na které se musí podílet mnoho lidí s různými odbornostmi — a mezi nimi statistik, který na to má kvalifikaci. Ano, bylo by to drahé a komplikované, ano, trvalo by to nějakou dobu, ale on covid je taky dost drahý a taky může trvat ještě nepříjemně dlouho. Určitě by se to vyplatilo.

Mimochodem, epidemiologové mají po léta dost důkladně rozpracovanou metodiku observačních studií pro hledání rizikových faktorů vzniku a šíření nemocí, infekčních i jiných. Jenže epidemiologie v Česku zamrzla někde na úrovni padesátých let a neumí nic lepšího než řešit lokální outbreaky salmonely nebo hepatitidy A. Na covid tohle nestačí. Ono tedy ani v cizině to není žádná sláva — jako by v rozčilení zapomněli na to, co umějí, a poklesli na tutéž bídnou úroveň, na které jsme my.

Další druh studií, které by šly dělat, jsou vyhodnocení vlivu různých opatření, omezení a pod. Třeba otevření škol: udělá se to v omezené míře na vybraných lokalitách nebo územích a pak se pečlivě studuje, jak se to projeví, hned, na místě, ne až z nepřesných čísel, která po nějaké době doputují do ÚZISu. A přitom máte kontrolu — můžete porovnávat situaci na otevřených školách a školách s distanční výukou a lépe vidíte, co otevření škol opravdu způsobí či nezpůsobí.

Co se týče užitečnosti takových plánovitě sbíraných dat pro modeláře — máte jednak parametry do svých modelů, které jsou získané lépe než sečtením hrušek s jablky, nebo zpětným přizpůsobováním modelů datům, aby seděly. Ani jednomu ani druhému se věřit nedá. Za druhé, kdyby existovala objektivní data o účinnosti různých opatření, mohli byste porovnat odpovědi, které poskytuje model s realitou a podle toho modely vylepšovat a přizpůsobovat.