Další velká datová sada ÚZIS zveřejněna

smis-lab.cz/2024/11/07/dalsi-velka-datova-sada-uzis-zverejnena

7. 11. 2024

Tomáš Fürst, Vít Karásek, Arnošt Komárek, Robert Straka

Po dlouhé době jsme se dočkali další velké datové sady z ÚZISu. Na první pohled to vypadá jako svatý grál, protože popis polí naznačuje, že data obsahují na úrovni jednotlivců informace o pětiletce narození, pohlaví, kalendářních týdnech všech covidových vakcín, kalendářním týdnu úmrtí a příčině úmrtí (covid/necovid). Navíc data obsahují informace o testech a covidových hospitalizacích. Takovou datovou sadu dosud (pro nás zcela nepochopitelně) nezveřejnila žádná země na světě.

Při podrobnějším pohledu ovšem **data vzbuzují mnoho pochybností.**

1. **Soubor obsahuje 12 597 668 řádků**, což je podezřele mnoho. I když jsou vyloučeny druhé, třetí a další infekce, které jsou v popisu dat označeny jako duplicity, zůstane tam 12 125 969 řádků, což je pořád moc. Tolik lidí v ČR není, ani když se započítají všichni mrtví od roku 2020.
2. **Soubor neobsahuje (anonymizovaný) identifikátor člověka**, nejde tedy spárovat řádky, které patří stejnému člověku. Nejde proto propojit údaj o očkování s údajem o úmrtí a jeho příčinou. Proto také nelze zjistit, jaké duplicity jsou důvodem předchozího bodu.

3. Při zkoumání sloupce „DatumUmrtiLPZ” zjistíme, že čísla tam uvedená přibližně odpovídají průběhu úmrtí dle ČSÚ. Až do konce roku 2022 nepřesahuje rozdíl v týdenních počtech úmrtí oproti ČSÚ jedno procento, ovšem ve druhé polovině roku 2024 už roste k desítkám procent. **Údaje o úmrtích jsou tedy použitelné maximálně do konce roku 2023**. Nicméně i rozdíly oproti ČSÚ ve statistikách týdnů let 2020–2023 jsou podezřelé a neměly by tam být.
4. Daleko větší problém však je, že celkem **1 556 198 řádků souboru neobsahuje žádné údaje** (krom DCCI=0), tedy ani údaje o narození ani o pohlaví. Podobně pak máme 63 606 řádků které obsahují pouze datum úmrtí, znovu bez údajů o narození a pohlaví. Podstatná část těchto úmrtí nastala v roce 2020 a při srovnání s daty od ČSU je zřejmé, že k těmto úmrtím skutečně došlo. Tato úmrtí nejen že nelze spárovat s příčinou úmrtí, ale nejde je ani přiřadit k dané věkové skupině a pohlaví. Není tedy možné vypočítat žádnou míru úmrtnosti, protože ta je vždy dramaticky závislá na věku a pohlaví.

Tato datová sada je tedy pro analýzu bezcenná. Její zkoumání přitom již stálo několik statistiků hodiny času. O výše popsaných problémech budeme ÚZIS prostřednictvím žadatele informovat a požádáme o zveřejnění dat ve stejném formátu jako zde s tím, že rozhodné období bude prodlouženo minimálně do konce roku 2023 a bude přidán sloupec, zda bylo úmrtí způsobeno covidem či nikoliv.