ANALÝZA FAKTORŮ ASOCIOVANÝCH S VÍCEČETNÝMI PŘÍČINAMI SMRTI V ČESKU V ROCE 2018 POMOCÍ XGBOOST REGRESE A METODY SHAP

Bety Ukolova[1] – Boris Burcin[2]

AN ANALYSIS OF THE FACTORS ASSOCIATED WITH MULTIPLE CAUSES OF DEATH IN CZECHIA IN 2018 USING XGBOOST REGRESSION AND SHAP VALUES

Abstract

This study focuses on the factors that are associated with recording multiple causes as the cause of death in Czechia. An XGBoost multiple regression is used in the analysis and its results are interpreted with SHAP values. The most significant factors associated with the number of causes of death, ranked in order of importance, are the place of death, the region, and the underlying cause of death. Age and autopsy also contribute, albeit to a lesser extent. Several important interactions were identified as well.

Keywords: multiple causes of death, death certificate, mortality, Czechia
Demografie, 2024, **66(1): 24–38**
DOI: https://doi.org/10.54694/dem.0331

ÚVOD

Úmrtí je zřídkakdy důsledkem jediné příčiny. Konkrétně v Česku umírá s tímto počtem příčin úmrtí přibližně jedna desetina osob (*ČSÚ*, 2018). Avšak převládající přístupy k analýze úmrtnosti založené na základní příčině úmrtí (chorobě, která iniciuje řetězec morbidních stavů, jež přivodí smrt) tuto skutečnost nikterak nereflektují. Alternativu může přinášet vícečetný přístup, jenž začleňuje všechny příčiny úmrtí, které jsou uvedeny na listě o prohlídce zemřelého (LPZ). V důsledku toho může nabývat pohled na úmrtnostní profily populací nového rozměru, neboť struktura úmrtnosti podle základních příčin je obohacena o struktury podle bezprostředních příčin smrti (stavů, které předcházejí pouze úmrtí samotnému), dále podle předchozích příčin (stavů, které vznikají v důsledku základních příčin úmrtí, ale přímo neústí ve smrt) a přispívajících chorob (morbidních stavů bez přímé vazby na základní příčinu úmrtí, avšak přítomných v okamžiku úmrtí) (*ÚZIS,* 2006). Vícečetný přístup tak může nalézat uplatnění nejen při studiu souvislostí mezi příčinami úmrtí a při přehodnocení zátěže populací chorobami s ohledem na ty, jež jsou

1) Katedra demografie a geodemografie, Přírodovědecka fakulta Univerzity Karlovy, Praha. Kontakt: elizaveta.ukolova@natur.cuni.cz.
2) Katedra demografie a geodemografie, Přírodovědecka fakulta Univerzity Karlovy, Praha. Kontakt: boris.burcin@natur.cuni.cz.

uváděny jako jiné než základní příčiny, ale je vhodný i k identifikaci problémů se zaznamenáváním příčin úmrtí (*Ausstats*, 2006; *Lindahl – Johansson*, 1994). Správnost a úplnost vyplnění LPZ totiž ovlivňuje proces automatizovaného kódování základních příčin úmrtí, které jsou fundamentálním zdrojem informací o zdravotním stavu populací (*ÚZIS*, 2021). Nezbytnost přesného vyplňování formuláře LPZ je zřejmá rovněž na individuální úrovni, neboť základní příčina úmrtí je směrodatnou informací, s níž mohou pracovat např. pojišťovny, ale i další subjekty. Taktéž mohou onemocnění a morbidní stavy uvedené na LPZ podávat svědectví o historii zdravotního stavu zemřelého pro jeho rodinné příslušníky (*NCHS*, 2003; *Flagg – Anderson*, 2021; *Curtin – Tolson – Arias – Anderson*, 2019).

Problematičnost vícečetných příčin úmrtí vychází najevo společně s identifikací významných sociodemografických faktorů, které ovlivňují způsoby vyplňování LPZ. Existují-li, platí, že zjištěné rozdíly v úmrtnostních poměrech nemusejí vůbec pramenit ze skutečných disparit ve zdravotním stavu populací. Již nejstarší studie věnované vícečetným příčinám úmrtí upozorňují na významné geografické rozdíly jak v úplnosti, tak ve způsobu zapisování příčin úmrtí (*Rosenberg*, 1986; *Guralnick*, 1966). Tento rys trvá (*Wall*, 2005; *Desesquelles et al*., 2012), přičemž autoři také nacházejí důkazy, že místo úmrtí, národnost/etnický původ či (ne)provedení pitvy ovlivňují jak délku zaznamenávaných chorobných řetězců, tak i pravděpodobnost uvádění některých významných příčin úmrtí jako základní (*Wall*, 2005).

V českém prostředí dosud nebyly faktory kódování vícečetných příčin úmrtí podrobeny analýze. Přitom však je Česko státem s výraznými regionálními rozdíly jak v délce života, tak i ve struktuře úmrtnosti podle příčin (*Pachlová*, 2014). Mohou se na tomto stavu podílet i rozdílná schémata u zaznamenávání příčin úmrtí? Kromě efektu regionu zemřelého na počet onemocnění a chorobných stavů uvedených na LPZ je níže zkoumán i efekt pohlaví, rodinného stavu, vzdělání, pitvy, místa úmrtí, věku a kapitoly Mezinárodní klasifikace nemocí (MKN), do níž spadá příčina vybraná jako základní. Tento příspěvek ukazuje kromě identifikace nejdůležitějších faktorů i interakce mezi nimi a rovněž se podílí na snaze nalézt odůvodnění opakovaně dokumentovaného poklesu průměrného počtu uváděných příčin úmrtí u zemřelých v nejstarším věku (*Dessesquelles*, 2012; *Pechholdová*, 2014). Analýza vychází z individuálních dat o zemřelých v Česku v roce 2018, poskytnutých Českým statistickým úřadem (ČSÚ) a Ústavem zdravotnických informací a statistiky (ÚZIS). Původní datový soubor čítal 112 920 zemřelých, z nichž analýza byla provedena nad 106 795 osobami. Vynechány z analýzy byly (i) osoby zemřelé na vnější příčiny smrti (tj. mající základní příčinu z XIX kapitoly MKN, která sdružuje poranění, otravy a některé jiné následky vnějších příčin v rozmezí kódů S00–T98) a (ii) osoby zemřelé před dokončením věku nula. Opodstatnění nezahrnutí prvé subpopulace tkví ve specifických pravidlech pro kódování jejích příčin smrti, poněvadž zemřelí v důsledku vnějších příčin by měli podle pravidel WHO mít uvedeny vždy aspoň dva kódy, jeden vyjadřující způsob zranění (např. pád), druhý jeho následek (např. zlomenina) (*ÚZIS*, 2006). Tudíž se jedná o zemřelé, pro něž je odlišný minimální počet uváděných příčin smrti než pro ostatní. Analogický postup byl následován pro děti zemřelé v kojeneckém věku. U nich však vícero kódů může vyjadřovat zdravotní komplikace u matky kojence (*WHO*, 2016).

Pro identifikaci klíčových faktorů zapisování vícečetných příčin smrti jsou konstruovány regresní modely, jejichž výsledky jsou interpretovány pomocí Shapleyových hodnot. Je využita metoda XGBoost regrese, jež je přímo určená pro velké datové soubory, neboť umožňuje výpočetně vysoce efektivní modelování právě rozsáhlých datových struktur (*Mahesh*, 2020).

DATA

Každé zemřelé osobě v Česku je přiřazen List o prohlídce zemřelého. Jedná se o formulář, na němž jsou mimo příčin smrti uvedeny i sociodemografické charakteristiky zemřelého a tyto informace se posléze předávají mezi několika institucemi. Pro daný kontext je relevantní to, že je přes matriční úřad odeslána na ČSÚ část listu bez příčin úmrtí a přes ÚZIS na ČSÚ část listu s nimi. Na ČSÚ jsou poté části opětovně spojeny a vzniká statistika příčin úmrtí. Tento proces s sebou mimo jiné přináší i výběr základní příčiny úmrtí (*ÚZIS*, 2021).

Jak již bylo výše naznačeno, základní příčina úmrtí je stav nebo onemocnění, které stály na začátku řetězce

zdravotních komplikací vedoucích ke smrti. Výběr příčiny odpovídající této definici se provádí v souladu s mezinárodními pravidly, která jsou implementována v softwaru IRIS. Počet uváděných příčin úmrtí tudíž předurčuje trasu skrze rozhodovací pravidla tohoto softwaru. Jediná příčina naopak oprošťuje systém od možností volby či odvození základní příčiny úmrtí (*ÚZIS*, 2021).

Sekce Listu o prohlídce zemřelého k vyplnění příčin úmrtí je rozdělena na dvě části, kdy do jedné se vpisuje sled chorobných stavů vedoucích ke smrti (tedy základní, předchozí a bezprostřední příčiny úmrtí) a do druhé se uvádí stavy, které ke smrti pouze přispěly, avšak nepatří do hlavní posloupnosti (přispívající příčiny úmrtí). Prvá část je tvořena čtyřmi řádky, jež jsou vázány pravidlem „jako důsledek". V souladu s ním by se měl na nejspodnější řádek sekce uvádět iniciátor řetězce a stavy, které vyvolal, postupně nad něj (*ÚZIS,* 2021). Druhá část na řádky členěna není, avšak pořadí uvedení by mělo vypovídat o významnosti přispívajících chorob v procesu úmrtí. *ÚZIS* (2021) shrnuje, že správné, úplné a co nejpřesnější vyplňování příčin úmrtí hraje naprosto klíčovou roli pro produkci kvalitní statistiky.

Příčiny úmrtí v individuálních datech o zemřelých poskytnutých ČSÚ jsou dvojího druhu. Prvé představují původní, jen digitalizované záznamy dat z LPZ, a kromě samotné diagnózy nesou i informace o místě zápisu dané příčiny úmrtí. Druhé jsou již zpracovanými záznamy, které sice neobsahují redundantní informace či chybné kódy (např. diagnózy nekompatibilní s pohlavím zemřelého), avšak při jejich zpracovávání se z nich vytrácí i informace o místě zápisu. Analýza zde tudíž vychází z původních záznamů, neboť smyslem příspěvku je zkoumat právě faktory určující zaznamenávání příčin v jejich původní podobě. Navíc cílové proměnné v modelech tvoří kromě celkového počtu příčin úmrtí i délka chorobného řetězce (počet příčin v první části) a počet přispívajících chorob, které byly identifikovány právě díky znalosti místa zápisu.

METODY

Provedená analýza je postavena na modelech vícenásobné regrese, jejíž principy jsou vyloženy jinde (např. v monografii *Härdle – Simar*, 2012). V tabulce 1 jsou prezentovány prediktory, vymezení jejich druhů a kategorií. Rovněž jsou zde definovány i závisle proměnné jednotlivých modelů. Výsledky regrese jsou interpretovány za pomoci Shapleyových hodnot, jež rozkládají vliv určitých kombinací prediktorů na příspěvky k predikci u každého pozorování pro jednotlivé proměnné. V literatuře jsou Shapleyovy hodnoty (*Shapley,* 1953), definovány jako „*mean marginal contribution of each feature value across all possible values in the feature space*" (*Lundberg – Lee*, 2017: 1). Smyslem této metody je rozdělit mezi jednotlivé prediktory ten díl predikce, který pramení z rozdílu mezi regresním modelem obsahujícím nezávisle proměnné a mezi základním modelem bez nezávisle proměnných. Každý z prediktorů je v důsledku tohoto reprezentován hodnotou, kterou přispívá ke změně hodnoty cílové proměnné, když na ni působí jejich koalice. Výpočet Shapleyových hodnot tudíž obnáší postupné konstruování všech možných regresních modelů z hlediska uvažovaných vysvětlujících proměnných. Shapleyovy hodnoty jsou dopočítávány vzorcem (*Lundberg – Lee*, 2017):

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\left(|F|-|S|-1\right)!}{|F|!} \times \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right)\right]$$

Vzorec je definicí Shapleyovy hodnoty pro proměnnou *i* působící v modelu o *S* nezávisle proměnných, kdy tento počet je podmnožinou *F*, která pokrývá celkový počet všech možných proměnných, které mohou do modelu vstoupit. Je zřejmé, že uvedený vzorec sestává ze součtu součinů dvou komponent. Prvá komponenta představuje pravděpodobnost, že proměnná příspěvkem v dané konstelaci koalice skutečně přispěje, tedy pravděpodobnost, že se daná koalice vytvoří způsobem, jakým se vskutku vytvořila. Druhá komponenta součinu je dána rozdílem dvou predikcí, kdy prvá je důsledkem působení proměnných v modelu, kam ta *i-tá* ještě nevstoupila ($f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right)$) a druhá představuje predikci na základě proměnných včetně oné *i-té* ($f_S(x_S)$).

Shapleyova analýza patří do souboru metod aditivních příspěvků prediktorů (*Lundberg – Lee*, 2017), jejichž vysvětlující modely lze vyjádřit ve tvaru:

$$g\left(z'\right) = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i^{'}$$

Kde φ_0 představuje hodnotu predikce v základním modelu (je to tedy prostý průměr vysvětlované proměnné). Výraz φ_i v sumě figuruje pro Shapleyovu hodnotu vypočítanou podle výše uvedeného vzorce a výraz z_i' představuje indikátorovou funkci, která vrací hodnotu 1, pokud je proměnná *i* začleněna do modelu, jinak hodnotu 0. Model vyjadřuje, že predikce je rozložitelná na dvě složky: (i) průměr, tedy výsledek, když nepůsobí žádné nezávisle proměnné, a (ii) specifický příspěvek vznikající v důsledku působení kombinace proměnných.

Shapleyovy hodnoty disponují řadou vlastností, díky nimž se otevírá široký prostor pro rozličné operace. Například je zcela správné je sčítat. Další vlastnosti, jež však sahají nad rámec tohoto příspěvku, lze najít např. v *Lundberg a Lee* (2017). Neopomenutelnou výhodou Shapleyových hodnot je, že mají přímou interpretaci vázanou na predikovanou hodnotu, jak vyplývá z metodiky jejich výpočtu, v porovnání s jinými statistickými ukazateli, jež lze pro interpretaci působení proměnných v modelu použít (např. standardizované beta koeficienty). To je také opodstatněním využití Shapleyových hodnot v tomto článku.

Princip výpočtu Shapleyových hodnot naznačuje, že se může jednat o výpočetně velmi náročnou metodu, a to i z toho důvodu, že je nutné vyzkoušet všechny možné modely pro uvažovaný počet proměnných. Za účelem zefektivnění procesu lze uplatnit celou řadu algoritmů strojového učení. Zde byl užit XGBoost pro regresní úlohy, konkrétně se pracovalo v prostředí R s balíčkem „xgboost" a posléze s „SHAPforxgboost". Balíček umožňuje nalezení nejvhodnějších parametrů modelu (na základě posouzení výše střední čtvercové chyby), jeho natrénování a produkci predikcí, eventuálně potom výpočet Shapleyových hodnot. Obecně je „boosting" označením pro proces postupného vylepšování modelu začleňováním vysvětlujících proměnných a „xg" figuruje pro „extrémní gradient" (*Jordan – Mitchell*, 2015). To vyjadřuje způsob, jakým algoritmus nalézá minimum ztrátové funkce, která popisuje, jak přesně (či nepřesně) algoritmus modeluje datovou matici. Ztrátová funkce v sobě zahrnuje vzdálenosti mezi predikcemi a hodnotami závisle proměnné v trénovací datové množině, jež je určena právě k nalézání nejvhodnějších parametrů. Ty jsou předpokladem pro malé rozdíly mezi predikcemi a známými hodnotami závisle proměnné. Více se lze o principech a matematické podstatě regresního strojového učení dozvědět např. v *Mahesh* (2020) nebo *Zhou* (2021).

Aplikace balíčku „xgboost" umožnila nalézt efektivně optimální parametry modelu a následně díky vyzkoušení všech možných modelů i spočítat Shapleyovy hodnoty. Ale protože není cílem tohoto příspěvku návrh výkonného predikčního modelu pro počet příčin úmrtí uváděných na LPZ, nejsou dále ani interpretovány vlastnosti modelu mnohonásobné regrese (jeho přesnost, míra „natrénování" ani jiné).

Tab. 1: Specifikace proměnných v modelech vícenásobné regrese a mediánové Shapleyovy hodnoty za každou kategorii prediktoru / Specification of variables in multiple regression models and median Shapley values for each feature category

Proměnná *Variable*	**Kategorie** *Category*	**Zkratka** *Abbreviation*	**Typ** *Type*	**Mediánové Shapleyovy hodnoty** *Median Shapley values*
Pohlaví / *Sex*	Muž / *Man*	M	Kategorické prediktory *Class variables predictors*	0,035
	Žena / *Woman*	Ž		–0,033
Rodinný stav / *Marital status*	Svobodní / *Single*	SV		–0,011
	Vdané/ženatí / *Married*	VD/Ž		0,016
	Rozvedení / *Divorced*	ROZ		–0,020
	Ovdovělí / *Widowed*	OVD		–0,002
	Jiné/neuvedené / *Other/not specified*	NE		–0,160

(pokračování / *continued*)

Tab. 1: Specifikace proměnných v modelech vícenásobné regrese a mediánové Shapleyovy hodnoty za každou kategorii prediktoru / Specification of variables in multiple regression models and median Shapley values for each feature category

Proměnná *Variable*	**Kategorie** *Category*	**Zkratka** *Abbreviation*	**Typ** *Type*	**Mediánové Shapleyovy hodnoty** *Median Shapley values*
Nejvyšší ukončené vzdělání *Education*	Základní / *Primary*	ZŠ	Kategorické prediktory *Class variables predictors*	–0,018
	Středoškolské bez maturity / *Secondary without A-level examination*	SŠB		0,055
	Středoškolské s maturitou / *Secondary education with A-level examination*	SŠs		–0,003
	Vyšší odborné, včetně vyučení *Short-cycle tertiary education*	VOŠ		0,049
	Vysokoškolské / *Tertiary*	VŠ		–0,025
	Nezjištěné / *Not specified*	NE		–0,008
Region místa úmrtí *Region of the place of death*	Jihočeský kraj	JHČ		–0,589
	Jihomoravský kraj	JHM		–0,034
	Karlovarský kraj	KAR		0,178
	Královéhradecký kraj	KRH		0,007
	Liberecký kraj	LIB		0,065
	Moravskoslezský kraj	MSZ		–0,177
	Olomoucký kraj	OLO		–0,395
	Pardubický kraj	PAR		0,463
	Hlavní město Praha	PHA		0,078
	Plzeňský kraj	PLZ		0,024
	Středočeský kraj	STC		–0,187
	Ústecký kraj	UST		–0,383
	Kraj Vysočina	VYS		–0,287
	Zlínský kraj	ZLN		0,474
Pitva / *Autopsy*	Provedena / *Performed*	ANO		–0,151
	Neprovedena / *Not performed*	NE		0,020
Místo úmrtí / *Place of death*	Doma / *At home*	doma		–0,678
	Ve zdravotnickém zařízení lůžkové péče *In a medical facility for inpatient care*	nemocnice I		0,341
	V jiném zdravotnickém zařízení *In some other medical facility*	nemocnice II		–0,227
	Na ulici, veřejném místě *On the street, in a public place*	ulice		–0,590
	Při převozu do zdravotnického zařízení *During transport to the medical facility*	převoz		–0,577
	V zařízení sociálních služeb *In a social services facility*	soc. služby		–0,498
	Jinde/nezjištěno / *Elsewhere/not identified*	ne		–1,087

(pokračování / *continued)*

Tab. 1: Specifikace proměnných v modelech vícenásobné regrese a mediánové Shapleyovy hodnoty za každou kategorii prediktoru / Specification of variables in multiple regression models and median Shapley values for each feature category

Proměnná ***Variable***	**Kategorie** ***Category***	**Zkratka** ***Abbreviation***	**Typ** ***Type***	**Mediánové Shapleyovy hodnoty** ***Median Shapley values***
Základní příčina úmrtí (nejvíce zastoupené kategorie) *Underlying cause of death (the most represented categories)*	Novotvary / *Neoplasms*		C00–D48	0,182
	Nemoci oběhové soustavy *Diseases of the circulatory system*		I00–I99	–1,370
	Nemoci dýchací soustavy *Diseases of the respiratory system*		J00–J99	0,162
	Nemoci endokrinní, výživy a přeměny látek *Endocrine, nutritional, and metabolic diseases*		E00–E90	0,497
	Nemoci trávicí soustavy *Diseases of the digestive system*		K00–K93	–0,322
Věk / *Age*	1–110		–	–
Počet příčin smrti uvedený na LPZ* / *The number of causes of death listed on the death certificate*	1–9			
Počet příčin smrti uvedený na LPZ do prvé části / *The number of causes of death listed in the part one of the death certificate*	1–9			
Počet příčin smrti uvedený na LPZ do druhé části / *The number of causes of death listed in the part two of the death certificate*	1–9			

Pozn.: * Jedná se o všechny možné skutečně uvedené hodnoty, celkový počet příčin smrti uvedený na hlášení o smrti může přesáhnout tuto hodnotu. Na jeden řádek sekce příčin smrti v hlášení o smrti lze vyplnit i více příčin smrti než je jedna.
Note: ** These are all the possible values actually reported, the total number of causes of death reported on a death certificate may exceed this value. More than one cause of death can be entered on one line of the cause of death section of the death report.*
Zdroj: Autoři.
Source: *Authors.*

VÝSLEDKY

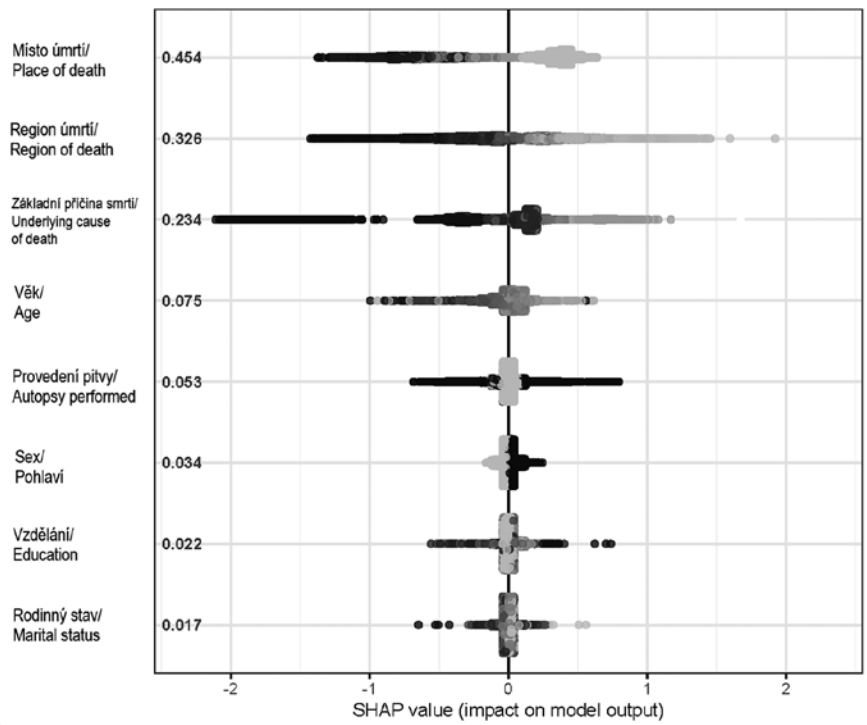
Před samotným modelováním byla provedena explorační analýza základního datového souboru. Byla porovnávána relativní zastoupení jednotlivých kategorií v souboru zemřelých s jedinou příčinou a s vícerem. Zemřelí s jedinou příčinou úmrtí jsou v průměru o čtyři roky mladší než zemřelí s alespoň dvěma příčinami, avšak věková hranice oddělující čtvrtinu nejstarších úmrtí je již rozdílná pouze o roky dva. Zemřelí s jedinou příčinou mají mnohem vyšší zastoupení v Jihočeském a Olomouckém kraji (1,92:1,00 a 2,06:1,00 představuje poměr podílu zemřelých s jedinou vs. s aspoň dvěma příčinami smrti). Zatímco se subpopulace neodlišují z hlediska pohlaví, tak u zemřelých s jedinou příčinou je vyšší zastoupení osob se SŠ vzděláním ukončeným maturitou (1,43:1,00), osob s „jiným" rodinným stavem (4,54:1,00) a nakonec pitvaných osob (1,20:1,00) než u zemřelých s aspoň dvěma příčinami smrti. Osoby s jedinou příčinou také umírají častěji doma (2,19:1,00), na veřejném místě (1,83:1,00), při převozu do zdravotnického zařízení (2,12:1,00) nebo jinde (2,26:1,00). Dále nalezneme odlišnosti mezi zemřelými s jedinou příčinou úmrtí a s aspoň dvěma podle základní příčiny úmrtí. V první zmíněné skupině se vyskytují téměř výhradně respirační a srdeční onemocnění (32,01:1,00, resp. 13,35:1,00), jsou-li opomenuta úmrtí v důsledku neznámých příčin úmrtí, které se u osob zemřelých s aspoň dvěma příčinami jako základní nekódují.

Na svislé ose grafu 1 jsou vyneseny vysvětlující proměnné nejvýkonnějšího modelu v pořadí podle „důležitosti" pro vysvětlení počtu příčin úmrtí uváděných na LPZ. Na vodorovné ose jsou příslušné Shapleyovy hodnoty pro každé pozorování a proměnnou. Je-li pozorování u některé z proměnných umístěno na úrovni nuly, znamená to, že tato proměnná není zdrojem odlišností predikovaného počtu zapsaných příčin úmrtí a toho, který by platil v základním modelu. Naopak pro pozorování umístěná daleko od nuly platí, že je jejich predikovaná hodnota vysvětlované proměnné vzdálena od základní hodnoty právě díky těm proměnným, v nichž nabývá pozorování vysokých Shapleyových hodnot.

V první řadě lze z grafu 1 zjistit stěžejní determinanty pro vysvětlení počtu příčin úmrtí na LPZ. Nejsilněji působí místo úmrtí, posléze region a kapitola základní příčiny úmrtí. S odstupem následuje věk, provedení pitvy a pohlaví. Vzdělání a rodinný stav mají spíše okrajový efekt. Je však nutné brát na zřetel, že relevance proměnné vzdělání v modelu může být ovlivněna početností případů úmrtí se vzděláním neznámým či neuvedeným (asi 65 % případů). Dále lze z grafu rovněž vypozorovat, že některé kategorie proměnných spíše snižují predikci (záporné Shapleyovy hodnoty), jiné naopak zvyšují (kladné Shapleyovy hodnoty). Proto jsou v tabulce 1 ve sloupci zcela vpravo uvedeny mediány Shapleyových hodnot jednotlivých skupin.

Úmrtí v nemocnicích a jiných zdravotnických zařízeních se váže s významně vyšším počtem příčin smrti. Zbylá místa úmrtí, v čele s neznámým, spíše počet příčin snižují. Velmi nesourodá je role regionu. Úmrtí ve Zlínském, Pardubickém a Karlovarském kraji bývají spojována spíše s více příčinami smrti, kdežto úmrtí v Jihočeském, Ústeckém a Olomouckém kraji naopak

Graf 1: Nejdůležitější proměnné pro vysvětlení počtu příčin úmrtí a Shapleyovy hodnoty pro jednotlivá pozorování, Česko, 2018 / The most important variables for explaining the number of causes of death and the Shapley values for individual observations, Czechia, 2018



Zdroj dat: Český statistický úřad, 2018.
Data source: *Czech Statistical Office, 2018.*

s relativně menším počtem. Pro třetí nejvýznamnější proměnnou platí, že zejména kardiovaskulární onemocnění a nemoci trávicí soustavy se často uvádějí s nízkým počtem dalších příčin smrti a naopak na LPZ osob zemřelých na endokrinologická a metabolická onemocnění nebo nemoci pohybového aparátu se uvádí kódů více. Dále jsou v tabulce 1 průměry podle věkových skupin a lze vidět, že s věkem zpočátku mediánové hodnoty Shapleyových hodnot pozvolna rostou a následně v nejstarších věkových skupinách stagnují. Otázka vlivu věku je podrobněji rozebírána v závěru textu. Poslední proměnnou s nezanedbatelným efektem na počet příčin úmrtí je provedení pitvy. Na základě středních hodnot nelze u role této proměnné evidentně docházet k žádným závěrům. Provedení pitvy v některých situacích vede k vyššímu, jindy naopak nižšímu počtu příčin úmrtí, a tak je její efekt modifikován dalšími faktory. Proto jsou dále zkoumány interakce proměnných.

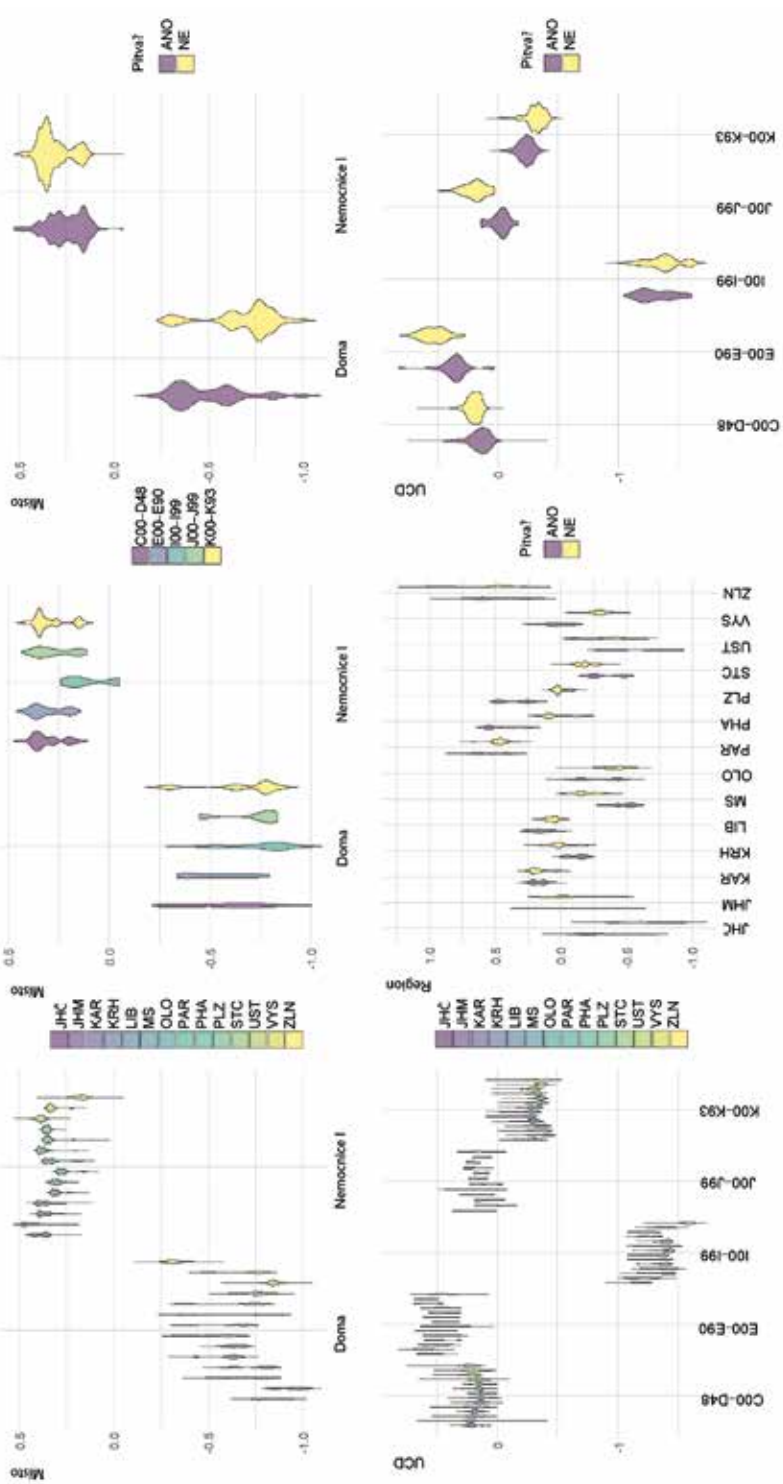
Na grafu 2 jsou houslové grafy pro tři nejdůležitější determinanty podle kategorií dalších proměnných, které by mohly prezentovat vztah mezi počtem příčin úmrtí a sledovaným prediktorem. Při konfrontaci dvou nejčastějších míst úmrtí (ve zdravotnickém zařízení a doma) a proměnné „region" se ukazuje, že pro geografické celky, kde úmrtí ve zdravotnickém zařízení zvyšuje počet příčin pouze málo, zároveň platí, že úmrtí doma tolik tuto proměnnou nepodhodnocuje v porovnání se základním modelem. Lze tudíž vymezit regiony, kde je efekt místa úmrtí obecně menší (Zlínský, Olomoucký a Liberecký kraj) a kde naopak větší (zejména Jihomoravský a Jihočeský kraj). Dále byla nejvlivnější proměnná diferencována podle dominantních skupin základních příčin úmrtí. Pro kardiovaskulární choroby platí, že jejich spojení s výrazně nižším počtem příčin úmrtí je příznačné zejména pro úmrtí ve zdravotnických zařízeních. Následující nejdůležitější kategoriální proměnná, pitva, má rovněž odlišný vliv na počet příčin úmrtí v závislosti na místu úmrtí, jak bylo ostatně naznačeno i výše. Z toho vychází, že provedení pitvy u úmrtí doma vede k vyššímu počtu příčin na LPZ, kdežto její provedení u zemřelých ve zdravotnických zařízeních se pojí se spíše nižším počtem příčin úmrtí, než když není pitva provedena.

Následně byl zkoumán efekt základní příčiny úmrtí v závislosti na regionu (graf 2). Vzhledem k tomu, že těžiště houslových grafů jednotlivých regionů v grafu 2 nekopírují tutéž konstelaci pro všechny dominantní skupiny základních příčin úmrtí, lze docházet k závěru, že to, jak působí základní příčina úmrtí na počet příčin úmrtí zjevně může být rovněž regionálně podmíněno, avšak tato interakce nabývá spíše slabšího rozměru. Ovšem takto tomu není při zohlednění pitvy. Bylo zjištěno, že zejména v Praze a Plzeňském kraji je absence pitvy předzvěstí nižšího počtu příčin úmrtí. Opačný efekt pitvy je vykazován v Moravskoslezském a Středočeském kraji. Nakonec byl sledován rozdíl vlivu provedení pitvy v závislosti na základní příčině smrti. Nepitvaní zemřelí na endokrinologická a metabolická, či respirační onemocnění jsou spojováni se spíše vyšším počtem příčin úmrtí, přičemž tomu tak není u ostatních dominantních skupin základních příčin úmrtí.

Posledním bodem analýzy je zkoumání vlivu věku na počet příčin úmrtí. Předchozí studie vesměs docházely ke zjištěním, že průměrný počet zapisovaných příčin úmrtí s věkem pozvolně vzrůstá, potom přibližně kolem modálního věku při úmrtí dosahuje svého maxima a záhy nastává pokles (*Pechholdová*, 2014; *Desesquelles et al*., 2016). O tom vypovídá i analýza založená na Shapleyových hodnotách. Vysvětlení tohoto průběhu však bývá dvojí. Jednak bývá pokles počtu příčin úmrtí na LPZ v nejvyšších věkových skupinách spojován s hypotézou, že se jej dožívají zpravidla osoby trpící méně onemocněními (s nižší mírou multimorbidity). Druhým nabízeným vysvětlením je, že u osob zemřelých v pokročilém věku (kdy je smrt již „normální") mohou osoby vyplňující LPZ vykazovat nižší míru snahy dopátrat se morbidních stavů, jimiž daná osoba před smrtí prošla. V tomto případě by pokles počtu příčin úmrtí na LPZ reflektoval neúplnost jeho vyplnění, jež se s věkem projevuje postupně stále více.

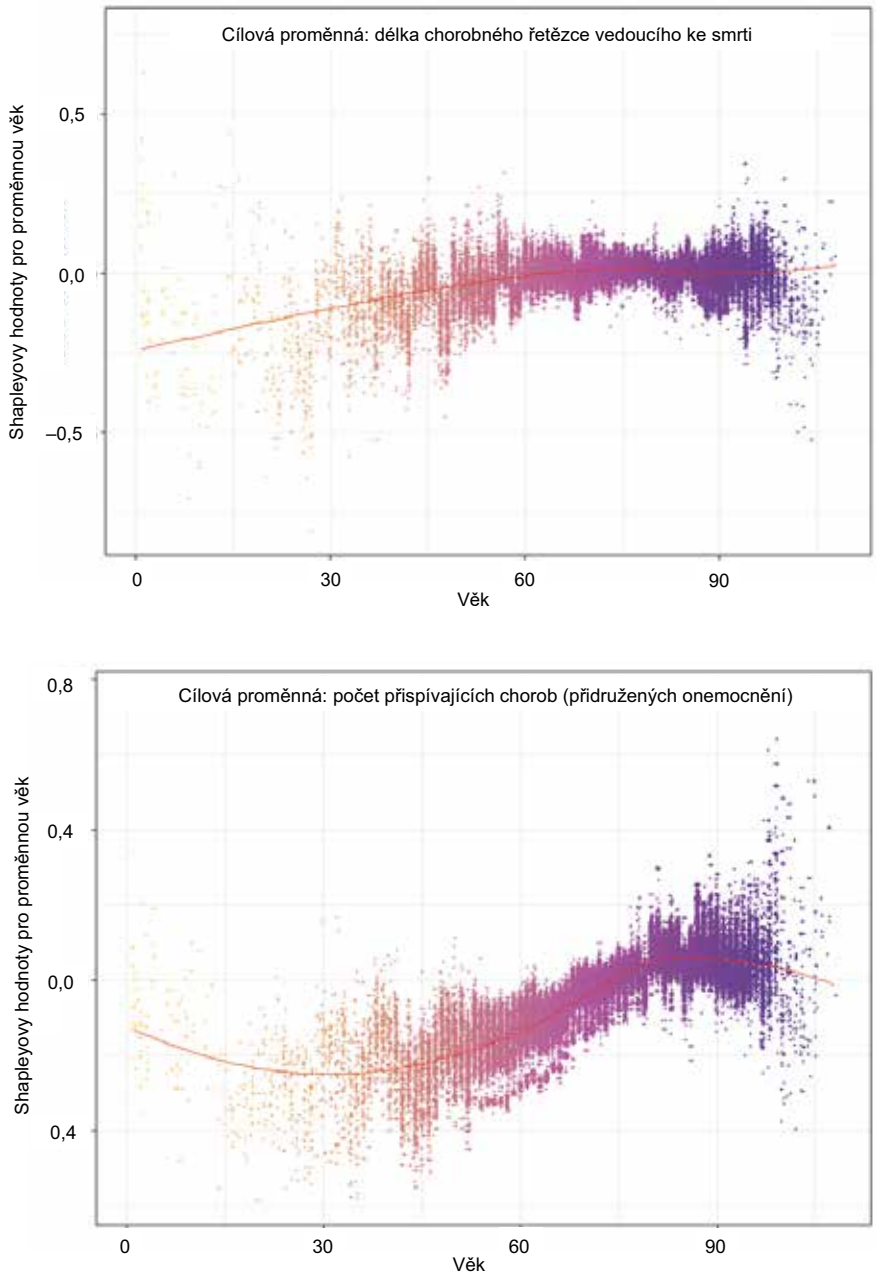
To, že v datech ČSÚ existuje sada proměnných určujících místo zápisu každé z příčin smrti, skýtá prostor pro bližší vhled do podstaty poklesu průměrného počtu příčin smrti v nejvyšších věcích. Lze se totiž domnívat, že nižší morbidita v okamžiku smrti se projeví hlavně v zapisování onemocnění do druhé části LPZ a tendence k nižší ochotě dopátrat se chorobných řetězců vedoucích ke smrti spíše naopak v první části LPZ. Na grafu 3 jsou vyneseny Shapleyovy hodnoty v závislosti na věku zemřelého.

Graf 2: Interakce nejvýznamnějších prediktorů (Shapleyovy hodnoty podle více faktorů)
Interactions of factors (Shapley values by multiple factors)



Zdroj dat: Český statistický úřad, 2018.
Data source: *Czech Statistical Office, 2018.*

Graf 3: Efekt věku na délku chorobného řetězce (první) a efekt věku na počet přidružených onemocnění (druhý), Česko, 2018 / Age effect on the length of chain of morbid events leading to death (first) and age effect on the number of contributory causes of death (second), Czechia, 2018



Zdroj dat: Český statistický úřad, 2018.
Data source: *Czech Statistical Office, 2018.*

Průběh příspěvků proměnné „věk" k počtu příčin úmrtí nabývá odlišné podoby podle uvažované části LPZ. Při začlenění jenom kódů uváděných do druhé části je efekt věku vyhraněnější. V kontrastu s tím se působení věku na predikovanou délku chorobných řetězců tak razantně nemění, a navíc se v řetězcích neprojevuje pokles ve stáří znamenající slabší, či dokonce negativní působení věku na základní hodnotu predikce. Došli jsme tedy k závěru, že prvotní interpretace častějšího výskytu menšího počtu příčin smrti v LPZ v nejstarších věkových skupinách by mohla směřovat spíše k tezi o nižší multimorbiditě u osob dožívajících se těchto věků. Samozřejmě se však objevují i další potenciální vysvětlení, např. že osoby vyplňující LPZ u zemřelých ve vysokém věku mají důraznější tendenci přisuzovat všem chorobám přímý podílna úmrtí.

DISKUZE A ZÁVĚR

Nejvýznamnějším faktorem ovlivňujícím počet příčin smrti je místo úmrtí, region a základní příčina úmrtí. Z hlediska významnosti lze s odstupem jmenovat ještě věk a provedení pitvy. Nejvíc se zvyšuje počet zápisů na LPZ u úmrtí v nemocnicích a ve Zlínském, Pardubickém a Karlovarském kraji. Kardiovaskulární onemocnění jako základní příčina úmrtí působí na počet příčin spíše slabě. Vliv jednotlivých faktorů je však zřejmě součástí komplexního mechanismu, který vede k tomu, že zcela jednoznačný vliv některých z nich v izolovaném pohledu mění směr působení při započtení dalších prediktorů. Příkladem jsou interakce zohledňující provedení pitvy, region nebo kapitolu základních příčin úmrtí. Konkrétně bylo zjištěno, že pitva a Zlínský, Olomoucký a Liberecký kraj předznamenávají nižší disparitu v počtu příčin úmrtí na LPZ podle místa úmrtí, takže rozdíl mezi úmrtími doma a ve zdravotnickém zařízení je menší. Rovněž bylo zjištěno, že nižší počet příčin úmrtí spojený s úmrtími v důsledku nemocí oběhové soustavy je typický spíše pro zemřelé ve zdravotnických zařízeních. V závěru příspěvku byl zkoumán efekt věku. Bylo zjištěno, že pokles počtu příčin úmrtí ve vyšších věcích je způsoben zejména ubýváním přispívajících onemocnění, nikoli zkracováním posloupností chorobných stavů ústících ve smrt.

Jedním z klíčových poznatků je, že v Česku existují významné regionální rozdíly ve způsobu vyplňování příčin úmrtí do LPZ. Pro další práci se nabízí vyhodnocení důsledků odlišných schémat pro zápis příčin úmrtí na regionální diferenciaci úmrtnosti. V tomto kontextu rezonuje již v úvodu vznesená otázka ohledně možného nadhodnocení (či podhodnocení) existujících regionálních rozdílů v úmrtnosti podle příčin úmrtí v Česku. Dalším poznatkem, jenž by neměl kvůli své „statistické nevýznamnosti" zapadnout, je to, že většina ryze sociodemografických charakteristik zemřelého (pohlaví, vzdělání a rodinný stav) výrazněji neovlivňuje počet příčin úmrtí uváděných na LPZ. Subpopulace podle těchto atributů se sice liší v rozličných aspektech úmrtnosti (např. v intenzitě), avšak nikoli ve vyplňování LPZ. Stojí však za zmínku, že interakce věku a vzdělání přinesla zjištění, že u vysokoškolsky vzdělané subpopulace zůstává až do nejvyšších věkových skupin vysoký počet příčin úmrtí, a tedy není sledován úbytek kódů v žádné z částí LPZ.

V kontextu dosud provedených studií se zdálo, že lze předpovědět roli pitvy (pitvané osoby by měly mít vyšší počet kódů), což se nepotvrdilo. Důvodem nesourodosti s výsledky předchozích studií může být i skutečnost, že ve stávajícím případě byla vynechána úmrtí způsobená vnějšími příčinami smrti, která ze zákona pitvě podléhají. Opodstatnění různorodosti, jež vzniká ve spojitosti s pitvou, si žádá další zkoumání. Zde je možné vyslovit hypotézu, že pitvy se v Česku provádí v zákonem stanovených případech, takže se mohou týkat spíše populace se specifickou úmrtností.

Příspěvek má i své limity. V první řadě se jedná o metodologická zjednodušení. Především je to modelování počtu příčin úmrtí jako spojité proměnné, avšak na LPZ lze samozřejmě spočítat pouze celkové počty chorob. Další zkreslení mohou vycházet z nedokonalostí datové základny, přesněji z kategorií „neznámé či jiné" přítomných a celkem často zastoupených zejména v sociodemografických proměnných. Konečně jako omezení lze vzít v úvahu i určitou primitivnost modelů. Pozornost by si jistě zasloužilo i například zkoumání determinant (ne)uvádění přispívajících chorob prostřednictvím logistické regrese. Takové zaměření by mohlo přispět k zodpovězení otázky, zda jsou příčiny úmrtí z druhé části LPZ relevantním datovým zdrojem pro výzkum morbidity.

Literatura

- Australian Bureau of Statistics. 2006. *Multiple Cause of Death Analysis*, 1997-2001. Dostupné z: https://www.abs.gov.au/Ausstats/abs@.nsf/Lookup/FDB92CC903BC3DC8CA256D6B0005A769.
- Curtin, S. C. et al. 2019. *Funeral directors' handbook on death registration and fetal death reporting: 2019 revision*. Dostupné z: https://stacks.cdc.gov/view/cdc/80634/cdc_80634_DS1.pdf.
- Český statistický úřad. 2018. *Data o zemřelých v Česku v roce 2018*. Datový soubor. Datum stažení: 20. 2. 2019.
- Désesquelles, A. F. et al. 2012. Analysing Multiple Causes of Death: Which Methods For Which Data? An Application to the Cancer-Related Mortality in France and Italy / Analyse des causes multiples de décès: quelles méthodes pour quelles données? L'exemple de la mortalité par cancer en France et en Italie. *European Journal of Population/Revue Européenne de Démographie*, 2012(28), s. 467–498. Dostupné z: http://www.jstor.org/stable/23274972. https://doi.org/10.1007/s10680-012-9272-3.
- Désesquelles, A. – Gamboni, A. – Demuru, E. 2016. We only die once… but from how many causes? *Population et Sociétés*, 534(6), s. 1–4. Dostupné z: https://www.cairn.info/revue-population-et-societes-2016-6-page-1.htm. https://doi.org/10.3917/popsoc.534.0001
- Flagg, L. A. – Anderson, R. N. 2021. Unsuitable Underlying Causes of Death for Assesing the Quality of Cause-of-death Reporting. *National Vital Statistics Reports*, 69(14). Dostupné z: https://www.ehdp.com/methods/nvsr69-14-508.pdf.
- Guralnick, L. 1966. Some problems in the use of multiple causes of death. *Journal of Chronic Diseases* 19(9), s. 979–990. https://doi.org/10.1016/0021-9681(66)90031-2.
- Härdle, W. K. – Simar, L. 2012. Regression Models. In: *Applied Multivariate Statistical Analysis*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17229-8_8.
- Israel, R. A. – Rosenberg, H. M. – Curtin, L. R. 1986. Analytical potential for multiple cause-of-death data. *American Journal of Epidemiology*, 124(2), s. 161–179. https://doi.org/10.1093/oxfordjournals.aje.a114375.
- Jordan, M. I. – Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), s. 255–260. https://doi.org/10.1126/science.aaa8415.
- Lindahl, B. I. B. – Johansson, L. A. 1994. Multiple cause-of-death data as a tool for detecting artificial trends in the underlying cause statistics: a methodological study. *Scandinavian Journal of Social Medicine*, 22(2), s. 145–158. https://doi.org/10.1177/14034948940220021l.
- Lundberg, S. M. – Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. Dostupné z: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
- Mahesh, B. 2020. Machine learning algorithms-a review. *International Journal of Science and Research*, 9(1), s. 381–386. Dostupné z: https://www.ijsr.net/archive/v9i1/ART20203995.pdf. https://doi.org/10.21275/ART20203995.
- National Center for Health Statistics. 2003. *Physicians Handbook on Medical Certification of Death.* Dostupné z: https://www.cdc.gov/nchs/data/misc/hb_cod.pdf.
- Pachlová, T. 2014. *Faktory ovlivňující regionální diferenciaci úmrtnosti v České republice*. Praha, Diplomová práce. Univerzita Karlova, Přírodovědecká fakulta, Katedra demografie a geodemografie. Vedoucí práce Burcin, Boris. Dostupné z: https://dspace.cuni.cz/handle/20.500.11956/70868.
- Pechholdová, M. 2014. Multiple cause-of-death data in the Czech Republic: an exploratory analysis. Demografie 56(4), s. 335–346. Dostupné z: https://www.czso.cz/documents/10180/20555381/13005314q4.pdf/e3140281-cec1-4b6b-82d3-590e442660d1?version=1.0.
- Shapley, L. S. 1953. A value for n-person games. In: Kuhn, H. – Tucker, A. (Eds.) *Contributions to the Theory of Games*. Princeton University Press, s. 307–317. https://doi.org/10.1515/9781400881970-018.
- Ústav zdravotnických informací a statistiky. 2006. *Informační systém List o prohlídce zemřelého.* Dostupné z: https://www.uzis.cz/res/file/registry/lpz/lpz-tiskopis.pdf.
- Ústav zdravotnických informací a statistiky. 2021. *Vyplňování informací na Listu o prohlídce zemřelého v ČR.* Dostupné z: https://www.uzis.cz/res/file/registry/lpz/lpz-instruktazni-video-2021.pdf.
- Wall, M. M. et al. 2005. Factors associated with reporting multiple causes of death. *BMC Medical Research Methodology*, 5, s. 1–13. DOI: https://doi.org/10.1186/1471-2288-5-4.
- Zhou, Z-H. 2021. *Machine learning*. Springer Nature. ISBN: 9811519676.

BETY UKOLOVA
Je studentkou doktorského programu Demografie na katedře demografie a geodemografie na Přírodovědecké fakultě Univerzity Karlovy. Zaměřuje se na analýzu úmrtnosti a zdravotního stavu, dále na demografické metody.

BORIS BURCIN
Je absolventem Univerzity Karlovy, oboru ekonomická a sociální geografie na její Přírodovědecké fakultě v Praze, kde od roku 1990 působí jako odborný asistent na katedře demografie a geodemografie. Akademickou dráhu nastoupil po dvouleté praxi na poli demografické statistiky v tehdejším Federálním statistickém úřadu. Zabývá se otázkami úmrtnosti, asistované reprodukce a prognózováním populačního vývoje a je spoluautorem řady demografických studií analytického i prognostického zaměření pro řídící a plánovací praxi. V posledním desetiletí působí jako mezinárodní expert a konzultant pro Populační fond OSN (UNFPA) v oblasti populačního vývoje v postkomunistických zemích.

SUMMARY

The number of causes of death reported on a death certificate determines the options for selecting the underlying cause of death. However, research shows that the number of reported causes varies substantially, for example, by geographic location. Our aim is to identify the factors that are associated with recording multiple causes of death in Czechia and to quantify their impact on the predicted number of recorded causes of death. To achieve this, we employ XGBoost multiple regression and interpret the results using SHAP values. The most significant factors associated with the number of causes of death, ranked in order of importance, are the place of death, the region, and the underlying cause of death. Age and autopsy also contribute to this, albeit to a lesser extent. The biggest increase in the number of records on death certificates was observed for deaths occurring in hospitals and other medical facilities and for deaths occurring in the regions of Zlín, Pardubice, and Karlovy Vary. The underlying cardiovascular cause of death was associated with a lower number of entries compared to other major cause-of-death groups. A complex mechanism is likely behind the effect of individual factors, as an examination of the interactions between factors revealed several nuances. For example, when autopsies are performed in the Zlín, Olomouc, and Liberec regions, the disparities in cause-of-death recording by place of death become less pronounced. The SHAP values for the age effect revealed that the decrease in the number of causes of death recorded later in life is primarily due to a decline in contributing diseases rather than a decrease in the length of the chains of morbid events leading to death. This study provides ideas for further research, particularly in the area of contextualising existing regional disparities in cause-of-death recording by analysing regional differences in mortality in Czechia.

PŘÍLOHA / APPENDIX

Příloha 1: Četnosti proměnných za zemřelé v modelech vícenásobné regrese, Česko, 2018
Frequencies of variables for the deceased in multiple regression models, Czechia, 2018

Proměnná / *Variable*	**Kategorie / *Category***	**Počet / *Number***	**Podíl / *Share (%)***
Pohlaví / *Sex*	Muži / *Men*	53 229	49,8
	Ženy / *Women*	53 566	50,2
Rodinný stav / *Marital status*	Svobodní / *Single*	7 390	6,9
	Vdaní/Ženatí / *Married*	40 374	37,8
	Rozvedení / *Divorced*	15 328	14,4
	Ovdovělí / *Widowed*	43 686	40,9
	Jiné/neuvedené / *Other*	17	0,0
Nejvyšší ukončené vzdělání / *Education*	Základní nebo bez vzdělání *Primary or no education*	6 018	5,6
	Středoškolské bez maturity *Secondary without A-level examination*	10 668	10,0
	Středoškolské s maturitou / *Secondary with maturita-level examinitaion*	4 713	4,4
	Vyšší odborné, včetně vyučení *Short-cycle tertiary education*	156	0,2
	Vysokoškolské / *Tertiary*	3 222	3,0
	Nezjištěné / *Not specified*	82 018	76,8
Region bydliště / *Region of residence*	Jihočeský kraj	6 318	5,9
	Plzeňský kraj	5 908	5,5
	Středočeský kraj	12 924	12,1
	Ústecký kraj	8 830	8,3
	Vysočina	5 016	4,7
	Zlínský kraj	5 952	5,6
	Jihomoravský kraj	11 865	11,1
	Karlovarský kraj	3 296	3,1
	Královehradecký kraj	5 759	5,4
	Liberecký kraj	4 431	4,2
	Moravskoslezský kraj	12 980	12,2
	Olomoucký kraj	6 492	6,1
	Pardubický kraj	5 249	4,9
	Hlavní město Praha	11 775	11,0
Pitva / *Autopsy*	Provedena / *Performed*	14 562	13,6
	Neprovedena / *Not performed*	92 233	86,4

(pokračování / *continued)*

Příloha 1: Četnosti proměnných za zemřelé v modelech vícenásobné regrese, Česko, 2018
Frequencies of variables for the deceased in multiple regression models, Czechia, 2018

Proměnná / *Variable*	Kategorie / *Category*	Počet / *Number*	Podíl / *Share (%)*
Místo úmrtí / *Place of death*	Doma / *At home*	23 550	22,1
	Ve zdravotnickém zařízení lůžkové péče / *In a medical facility for inpatient care*	68 839	64,5
	V jiném zdravotnickém zařízení *In some other medical facility*	1 180	1,1
	Na ulici, veřejném místě *On the street, in a public place*	707	0,7
	Při převozu do zdravotnického zařízení *During transport to a medical facility*	575	0,5
	V zařízení sociálních služeb *In a social services facility*	9 445	8,8
	Jinde/nezjištěno *Elsewhere / not identified*	2 499	2,3
Základní příčina smrti *Underlying cause of death*	Nemoci endokrinní, výživy a přeměny látek / *Endocrine, nutritional, and metabolic diseases*	5 137	4,8
	Novotvary / *Neoplasms*	28 265	26,5
	Nemoci dýchací soustavy (respirační) *Diseases of the respiratory system (respiratory)*	8 306	7,8
	Nemoci oběhové soustavy (kardiovaskulární) / *Diseases of the circulatory system (cardiovascular)*	48 786	45,7
	Nemoci trávicí soustavy *Diseases of the digestive system*	4 920	4,6
	Ostatní / *Other*	11 381	10,7
Věková skupina / *Age group*	<15	98	0,1
	15–39	1 034	0,1
	40–64	15 463	14,5
	65–84	56 874	53,3
	85+	33 326	31,2