UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

HODNOTENIE NEMOCNÍC NA ZÁKLADE REHOSPITALIZÁCIÍ

DIPLOMOVÁ PRÁCA

2016 Bc. Norbert Füle

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

HODNOTENIE NEMOCNÍC NA ZÁKLADE REHOSPITALIZÁCIÍ

DIPLOMOVÁ PRÁCA

Študijný program: Ekonomicko-finančná matematika a modelovanie
Študijný odbor: 1114 Aplikovaná matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Vedúci práce: Mgr. Henrieta Tulejová, MSc.

2016 **Bc. Norbert Füle**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Norbert Füle
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Hodnotenie nemocníc na základe rehospitalizácií
Hospital Performance Measurement Based on Readmissions

Cieľ: Cieľom práce je vyvinúť prediktívny model na základe rehospitalizácií, ktorý bude možné použiť na hodnotenie nemocníc.

Vedúci: Mgr. Henrieta Tulejová, MSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 10.02.2015

Dátum schválenia: 11.02.2015 prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou by som sa chcel poďakovať kolektívu pracovníkov poisťovne Dôvera. Svojej vedúcej práce Mgr. Henriete Tulejovej, MSc. za príležitosť, odbornú pomoc a trpezlivosť; taktiež MUDr. Anne Dankovej za podporné konzultácie z pohľadu zdravotníka.

Abstrakt v štátnom jazyku

FÜLE, Norbert: Hodnotenie nemocníc na základe rehospitalizácií [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Henrieta Tulejová, MSc., Bratislava, 2016, 53 strán.

Cieľom tejto diplomovej práce bolo vytvoriť model pre hodnotenie nemocníc, založený na rehospitalizáciách. Predstavili sme základy štatistických metód použitých v našej práci. Následne sme opísali metodiku pre spracovanie dát, použitý prediktívny model a metodiku pre hodnotenie nemocníc. Nakoniec sme prezentovali výsledky a možné vylepšenia nášho modelu.

Kľúčové slová: rehospitalizácie, viacúrovňová regresia, bayesovská inferencia, markovovské Monte Carlo simulovanie

Abstract

FÜLE, Norbert: Hospital Performance Measurement based on Readmissions [Master's Thesis], Comenius University in Bratislava, Faculty of mathematics, physics, and informatics, Department of Applied Mathematics and Statistics; Thesis supervisor: Mgr. Henrieta Tulejová, MSc., Bratislava, 2016, 53 pages.

The aim of this master's thesis is to create a model for hospital measurement, based on readmissions. Firstly, we introduce the basics of underlying statistical methods used in our work. Secondly, we describe the methods for data processing, used predictive model, and methodology of hospital measurement. Finally, we present obtained results and possible improvements to this model.

Keywords: readmissions, multilevel regression, bayesian inference, Markov chain Monte Carlo

Obsah

Úvod

Oblasť zdravotnej starostlivosti prechádza v posledných rokoch výraznou transformáciou. Podľa [8], z dôvodu zvýšeného dopytu pacientov pre kvalitnejšiu zdravotnú starostlivosť sú poskytovatelia zdravotnej starostlivosti pod tlakom, aby prinášali lepšie výsledky. Dynamika nákladov za zdravotnú starostlivosť sa mení kvôli dlhšej priemernej dĺžke života, defenzívnym medicínskym praktikám, ale aj čoraz väčšiemu prenikaniu chronických ochorení. Prostredie sa stáva čoraz komplexnejším, a preto si vyžaduje systematické použitie dát pre napomáhanie manažmentu, plánovaniu a hodnoteniu.

Využitie analytických metód môžeme vidieť aj na vládnej úrovni. V Spojených štátoch amerických prijatý zákon *Patient Protection and Affordable Care Act* vyžadoval federálne ministerstvo zdravotníctva zaviesť program na zníženie počtu rehospitalizácií. Rehospitalizácie predstavujú negatívny dôsledok zdravotnej starostlivosti a sú definované ako hospitalizácie vzniknuté do 30 dní od konca poslednej hospitalizácie daného pacienta. Cieľom tohoto programu bolo poskytnúť motiváciu nemocniciam, aby zaviedli do výkonu zdravotnej starostlivosti také postupy, ktoré by tieto nákladné a vo veľkom množstve zbytočné rehospitalizácie znížili. Napríklad, v čase implementácie tohto programu bolo približne 20% pacientov programu Medicare znovu prijatých do nemocnice do mesiaca od posledného prepustenia z nemocnice [18].

Centers for Medicare and Medicaid Services (ďalej ako CMS) je federálna agentúra, ktorá manažuje programy Medicare a Medicaid v Spojených štátoch [4]. CMS využíva hodnotenia založené na viacerých metrikách pri hodnotení zdravotnej starostlivosti - napr. mortalitu, komplikácie, ale aj rehospitalizácie, a iné. Podstatou týchto hodnotení je myšlienka, že pacienti, ktorým sa dostane vysokokvalitnej starostlivosti počas ich hospitalizácií a prechodu do prostredia neakútnej starostlivosti, budú mať lepšie šance na prežitie, schopnosť fungovať po rekonvalescencii a kvalitu života.

Táto práca vychádza zo zadania zdravotnej poisťovne Dôvera. Zadaním bolo vytvoriť a implementovať model pre hodnotenie nemocníc na základe rehospitalizácií. Po krátkej analýze literatúry v tejto oblasti sme sa rozhodli vychádzať práve z modelu a metodiky používanej CMS.

Všetky hodnotenia založené na rehospitalizáciách pre CMS vyvíja, implementuje, a udržiava Yale New Haven Health Services Corporation/Center for Outcomes Rese-

arch and Evaluation. Naša práca vychádza z metodiky CMS pre rehospitalizácie z ľubovoľného dôvodu v celej nemocnici (z angl. "Hospital-Wide All-Cause Readmission Measure"). Vychádzali sme primárne z dokumentácie metodológie pre toto hodnotenie [5] a článku [16].

Pretože súvisiaca metodika bola vytvorená odborníkmi v oblasti analytiky v zdravotnej starostlivosti, vyvíja sa a každoročne zdokonaľuje už niekoľko rokov, rozhodli sme sa lokalizovať postup z [5] so zámerom v niektorých miestach navrhnúť vlastné riešenia. Zdrojový kód je poskytovaný na vyžiadanie a je v jazyku SAS. Našim hlavným prínosom bolo teda upraviť existujúcu metodiku pre naše dáta (čo sa ukázala byť netriviálna úloha), naprogramovať spracovanie dát, štatistickú inferenciu a spracovanie pre výsledné hodnotenie nemocníc v jazyku `R` [22]. Niektoré naše rozhodnutia a úpravy sme tiež konzultovali s tímom pracovníkov poisťovne Dôvera ([9], [21], [26]).

Našu prácu začíname zhrnutím a popisom použitých štatistických, numerických a programovacích metód. V druhej kapitole predstavíme dáta, z ktorých sme vychádzali, ich spracovanie a použité algoritmy. V tretej kapitole uvádzame prediktívny model a metodiku pre hodnotenie nemocníc. Prácu zakončujeme základnou diagnostikou modelu a prezentáciou výsledkov.

1 Teoretické zhrnutie využitých metód

V tejto kapitole predstavíme teoretické zhrnutie štatistiských a výpočtových metód využitých v našej diplomovej práci. Časť o regresných modeloch sme spracovali na základe [11]. Časť o hamiltonskom Monte Carlo simulovaní sme prebrali z [20], časť o algoritme No-U-Turn je spracovaná z [15]. V časti o programe Stan sme vychádzali z jeho dokumentácie [25]. Napokon, v časti o diagnostike konvergencie sme vychádzali hlavne z [6].

1.1 Regresné modely

1.1.1 Lineárna a logistická regresia

Lineárnou regresiou sa rozumie metóda odhadu strednej hodnoty odhadovanej (závislej) premennej Y pomocou lineárnej kombinácie súboru iných premenných $X_1, X_2, ...$, nazývaných prediktory. Príkladom lineárneho regresného modelu, je model

$$Y = X \cdot \beta + \epsilon, \tag{1}$$

kde ϵ predstavuje akýsi chybový člen.

V prípade, kedy závislá premenná Y je binárnou, t.j. nadobúda len hodnoty 0 a 1, nie je vhodné na jej odhad použiť lineárnu regresiu.

Štandardný spôsob na modelovanie binárnej náhodnej premennej je v štatistike logistická regresia. Podstatou logistickej regresie je modelovanie pravdepodobnosti, že premenná Y nadobudne hodnotu 1. Na tento účel je potrebné transformovať kombináciu prediktorov, aby táto transformácia nadobúdala hodnoty v intervale $(0, 1)$. Docielime to použitím tzv. funkcie logit:

$$logit(x) = \log\left(\frac{x}{1-x}\right). \tag{2}$$

Modelujeme teda

$$logit(Pr[Y = 1]) = X \cdot \beta, \tag{3}$$

čo môžeme upraviť na

$$Pr[Y = 1] = logit^{-1}(X \cdot \beta), \tag{4}$$

kde inverzná funkcia funkcie logit má tvar:

$$logit^{-1}(x) = \frac{1}{1 + e^{-x}}. \tag{5}$$

1.1.2 Viacúrovňová regresia

Niektoré druhy dát môžu byť inherentne štrukturované takým spôsobom, že medzi určitými skupinami týchto dát sú rozdiely, ktoré v bežnej regresii nevieme zachytiť. Napríklad pri modelovaní na dátach o ľuďoch z celého Slovenska môžu v danom modeli existovať rozdiely medzi regiónmi. Jednou z možností, ako sa s takýmto štrukturovaním vysporiadať sú v klasickej regresii dummy premenné. Ďalšou z možností je využitie viacúrovňovej regresie, pri ktorej môžeme modelovať regresné parametre tak, že budú odzrkadľovať rozdielnu štruktúru.

Viacúrovňovú regresiu definujeme v súlade s [11] ako regresiu, ktorej niektoré parametre - regresné koeficienty - sú dané pravdepodobnostným modelom. Tento druhoúrovňový model má vlastné parametre - tzv. hyperparametre modelu - ktoré sú tiež odhadované z dát. Viacúrovňovú regresiu predstavíme na príklade.

Uvažujme štúdiu s dátami o študentoch veľkého počtu škôl. Predikujeme y (napr. záverečná známka študenta) v závislosti od x (napr. výsledok prípravnej skúšky) a parametra α, ktorý sa podľa školy líši (napr. "úroveň" školy). Zároveň odhadujeme parametre regresného modelu, v ktorom sa snažíme vysvetliť α na základe parametrov daných konkrétnej škole. Uvažujme preto pre školy spoločný intercept a, ktorý nám vstúpi do druhej časti modelu spolu s vysvetľujúcou premennou u_j (napr. priemerný príjem rodičov študentov danej školy). Potom regresiu na úrovni študentov a regresiu na úrovni škôl považujeme za dve úrovne viacúrovňovej regresie. Náš model potom môžeme zapísať nasledovne, pričom koeficienty v obidvoch častiach odhadujeme súčasne:

$$y_i = \alpha_{j[i]} + \beta \cdot x_i + \epsilon_i, \quad i = 1, .. \, n, \quad \text{(študenti)} \tag{6}$$

$$\alpha_j = a + b \cdot u_j + \eta_j, \quad j = 1, .. \, J. \quad \text{(školy)} \tag{7}$$

V tomto modeli x_i zodpovedá prediktoru na úrovni študentov a u_j prediktoru na úrovni škôl. ϵ_i a η_j sú nezávislé chybové členy pre zodpovedajúce úrovne. Pod $\alpha_{j[i]}$ označujeme parameter pre školu j, ktorej prislúcha študent i. Počet J (v tomto prípade škôl) v regresii vyššej úrovne je bežne výrazne menší ako n, t.j. veľkosť vzorky modelu

nižšej úrovne (v tomto prípade študentov).

Dve kľúčové časti viacúrovňovej regresie sú premenlivé koeficienty, a model pre tieto premenlivé koeficienty. Klasická regresia vie niekedy tiež obsiahnuť takéto premenlivé koeficienty využitím indikátorových (dummy) premenných. Vlastnosť, ktorá odlišuje viacúrovňové modely od klasickej regresie, je ale práve modelovanie premenlivosti medzi skupinami.

Viacúrovňové modely je možné nájsť v literatúre aj ako modely s náhodnými efektmi, či modely so zmiešanými efektmi. Náhodnými efektmi sa nazývajú modelované regresné koeficienty, v zmysle že sa považujú za náhodné dôsledky procesu súvisiaceho s modelom, ktorý ich predikuje. Na druhej strane, fixné efekty zodpovedajú buď parametrom, ktoré nie sú premenlivé, alebo parametrom, ktoré sú premenlivé, ale nie sú sami modelované. Modely so zmiešanými efektmi potom zahŕňajú náhodné aj fixné efekty.

V [11] autori uvádzajú, že fixné efekty môžeme považovať za špeciálne prípady náhodných efektov, v ktorých variancia vyššej úrovne je 0 alebo ∞. Preto v ich koncepcii sú všetky regresné parametre náhodné, a výraz viacúrovňový pokrýva všetky prípady. Autori preto považujú výrazy fixné, náhodné, a zmiešané za mätúce a často zavádzajúce, preto sa ich použitiu vyhýbajú.

1.1.3 Zhrnutie regresných modelov

V predchádzajúcej častiach sme teda predstavili niekoľko druhov regresných modelov. Najjednoduchším druhom regresného modelu je lineárna regresia. Štandardnými spôsobmi na odhadovanie parametrov lineárnej regresie sú metóda najmenších štvorcov a metóda maximálnej vierohodnosti. Modelovanie funkcie strednej hodnoty závislej premennej, narozdiel od priameho modelovania strednej hodnoty, nám vnáša do odhadovania pridanú náročnosť. V prípade logistickej regresie ale stále dokážeme pre odhad parametrov použiť určitú úpravu metódy najmenších štvorcov [23].

Viacúrovňová regresia nám ale vnáša do odhadovania parametrov dodatočnú komplexitu. Táto komplexita si vyžaduje buď výraznú úpravu metódy najmenších štvorcov (napríklad tzv. metódu penalizovaných najmenších štvorcov [1]) alebo iný prístup k odhadovaniu parametrov. Takýmto prístupom je napríklad bayesovská štatistická inferencia.

Bayesovský štatistik musí mimodátovú informáciu o parametri $\theta \in \Theta$ dáta-generujúceho rozdelenia pravdepodobnosti $p(X_1^n \mid \theta)$ vyjadriť vo forme apriórneho rozdelenia pravdepodobnosti $p(\theta)$. Prior, ako sa apriórne rozdelenie stručnejšie nazýva, vyjadruje štatistikovu neistotu o hodnote parametra θ. Po tom, ako je získaná realizácia x_1^n výberu X_1^n, modifikuje bayesiánec svoju apriórnu informáciu pomocou bayesovej vety

$$p(\theta \mid x_1^n) = \frac{p(x_1^n \mid \theta)p(\theta)}{\int_\Theta p(x_1^n \mid \theta)p(\theta)d\theta}, \tag{8}$$

a získa tak aposteriórne rozdelenie $p(\theta \mid x_1^n)$ parametra θ pri daných dátach x_1^n. Aposteriórne rozdelenie (alebo posterior) vyjadruje štatistikovu neistotu o θ po tom, ako boli pozorované dáta. Posterior samozrejme závisí aj od modelu - teda od apriórneho rozdelenia $p(\theta)$ a dáta-generujúceho rozdelenia $p(X_1^n \mid \theta)$ [14].

Odhady v bayesovskej inferencii začínajú s nejakými počiatočnými odhadmi ako štartovacími hodnotami, a pokračujú v iteráciách. V každej iterácii sú simulované výbery z aposteriórneho rozdelenia a tieto sú následne opravené, až kým sa nedostanú k cieľovému alebo stacionárnemu rozdeleniu. Táto časť je kľúčová a zároveň odlišná od klasickej štatistickej inferencie. Našim cieľom je získať rozdelenie, nie bodový odhad. Takýto proces simulácie sa nazýva markovovské Monte Carlo simulovanie (z angl. Markov chain Monte Carlo) [6].

1.2 Simulácia výberu z aposteriórneho rozdelenia parametrov

V nasledujúcej časti predstavíme algoritmus markovovského Monte Carlo simulovania, ktorý sme použili pri simulovaní hodnôt parametrov modelov použitých v našej práci.

Algoritmy markovovského Monte Carlo simulovania sú postavené na vytvorení takého markovovského reťazca, ktorého limitným rozdelením bude aposteriórne rozdelenie odhadovaného parametra. Hlavnou myšlienkou je simulovať dostatočne veľa realizácií markovovského reťazca, aby sa priblížil k limitnému rozdeleniu. Takýto počet iterácií sa nazýva zahrievací čas. Potom ďalšie realizácie po zahrievacom čase aproximujú náhodný výber z aposteriórneho rozdelenia.

Príkladom takéhoto algoritmu je Metropolisov-Hastingsov algoritmus, ktorý generuje realizácie a následne ich prijíma alebo odmieta v závislosti od pravdepodobnosti vypočítanej na základe daného aposteriórneho rozdelenia. Z definície Metropolisovho-Hastingsovho algoritmu sú realizácie skonštruovaného markovovského reťazca v po sebe

nasledujúcich dátových bodoch ale vysoko korelované. Na druhej strane, realizácie v každom k-tom dátovom bode, kde k je dostatočne veľké, sú približne nezávislé. Preto môžeme za približne náhodný výber z aposteriórneho rozdelenia považovať realizácie markovovského reťazca až v dostatočne od seba vzdialených dátových bodoch. V ďalšej časti opíšeme metódu, ktorá sa takémuto správaniu (akejsi "korelovanej náhodnej prechádzke") dokáže vyhnúť.

1.2.1 Algoritmus hamiltonského Monte Carlo simulovania

V tejto časti zhrnieme teoretický základ za špeciálnym typom algoritmu Markovovského Monte Carlo simulovania, založeného na koncepte hamiltonskej dynamiky z mechaniky.

Hamiltonská dynamika, hamiltonián Pod pojmom hamiltonská dynamika sa rozumie reformulácia klasickej mechaniky, sformalizovaná do rozdielnej matematickej podoby. V hamiltonskej dynamike je fyzikálny systém opísaný funkciou $H(q,p)$, ktorá sa nazýva hamiltonián. Pod vektorom q sa rozumie vektor polohy a pod vektorom p sa rozumie vektor hybnosti. To, ako sa q a p menia v čase opisuje systém diferenciálnych rovníc:

$$\begin{aligned}\frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i},\\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i},\end{aligned}$$

pre $i = 1, ..., d$, kde d je počet dimenzií vektorov p, q.

Pri hamiltonskom Monte Carlo simulovaní sa bežne používajú hamiltoniány, ktoré sa dajú zapísať ako

$$H(q,p) = U(q) + K(p), \tag{9}$$

kde $U(q)$ označuje potenciálnu energiu a $K(p)$ kinetickú energiu systému.

V nefyzikálnych aplikáciách markovovských Monte Carlo simulácií bude poloha q zodpovedať premennej, ktorú skúmame. $U(q)$ je definovaná ako záporná hodnota logaritmu hustoty rozdelenia premennej q (plus nejaká arbitrárna konštanta), ktorej výber z rozdelenia budeme simulovať. Kinetická energia $K(p)$ je zas bežne definovaná ako $K(p) = \frac{1}{2}p^T M^{-1} p$. V tomto výraze M predstavuje symetrickú, kladne definitnú maticu, ktorá je často diagonálna a skalárnym násobkom identitnej matice. Táto forma

$K(p)$ potom zodpovedá zápornej hodnote logaritmu hustoty (spolu s nejakou arbitrárnou konštantou) viacrozmerného normálneho rozdelenia so strednou hodnotou 0 a kovariančnou maticou M.

Taktiež je potrebné poznamenať, že hamiltonská dynamika má niektoré vlastnosti, ktoré sú potrebné pre jej použitie v markovovskom Monte Carlo simulovaní. Tieto vlastnosti neuvádzame, čitateľ ich môže nájsť v [20].

Počítačová implementácia, metóda leapfrog Aby sme mohli systém diferenciálnych rovníc hamiltonskej dynamiky vyriešiť numericky, musíme ho aproximovať, a to diskretizáciou času. T.j. zvolíme si nejaký malý krok ϵ, začneme v čase 0 a iteratívne počítame približné hodnoty stavu systému v časoch $\epsilon, 2\epsilon, 3\epsilon$, atď. Najčastejšia metóda na diskretizáciu systému diferenciálnych rovníc je Eulerova metóda. Neskôr uvedieme tzv. metódu leapfrog, ktorá sa v Hamiltonskom Monte Carlo simulovaní používa bežne a na ktorú v ďalších častiach nadviažeme.

Autor [20] uvažuje pre jednoduchosť hamiltonián v tvare $H(q,p) = U(q) + K(p)$. Ďalej uvažuje $K(p) = \frac{1}{2}p^T M^{-1} p$ a maticu M diagonálnu, so zložkami $m_1, ..., m_d$ spĺňajúcimi $K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$.

Iteračné vzorce pre metódu leapfrog vieme potom zapísať ako:

$$p_i(t + \frac{\epsilon}{2}) = p_i(t) - \frac{\epsilon}{2} \cdot \frac{\partial U}{\partial q_i}(q(t)), \tag{10}$$

$$q_i(t + \epsilon) = q_i(t) + \epsilon \cdot \frac{p_i(t + \frac{\epsilon}{2})}{m_i}, \tag{11}$$

$$p_i(t + \epsilon) = p_i(t + \frac{\epsilon}{2}) - \frac{\epsilon}{2} \cdot \frac{\partial U}{\partial q_i}(q(t + \epsilon)). \tag{12}$$

Metóda funguje spôsobom, že začneme vypočítaním polovičného kroku pre premenné hybnosti. Potom použitím nových hodnôt hybnosti vypočítame celý krok pre premenné polohy. Nakoniec vypočítame ďalší polovičný krok pre premenné hybnosti využitím nových hodnôt premenných polohy.

Pravdepodobnosť a hamiltonián, kanonické rozdelenie Aby sme mohli využiť hamiltonskú dynamiku v algoritmoch Monte Carlo simulovania výberu z určitého rozdelenia, potrebujeme previesť hustotu tohoto rozdelenia na funkciu potenciálnej energie a taktiež zaviesť premenné hybnosti. Potom dokážeme simulovať markovovský reťazec,

v ktorom v každej iterácii aktualizujeme hybnosť a potom prijmeme alebo odmietneme kandidáta na polohu, nájdeného pomocou hamiltonskej dynamiky.

Pravdepodobnostné rozdelenie, ktorého realizácie chceme simulovať, môžeme prepojiť s potenciálnou energiou na základe konceptu kanonického rozdelenia zo štatistickej mechaniky. Nech $E(x)$ je funkcia energie pre stav x nejakého pomyselného fyzikálneho systému. Kanonické rozdelenie nad stavmi tohoto systému má potom funkciu hustoty

$$P(x) = \frac{1}{Z} \exp\left(\frac{-E(x)}{T}\right),$$

Tu, T označuje teplotu systému a Z je normalizačná konštanta, ktorá zabezpečuje integráciu tejto funkcie na 1. Inými slovami, ak nás zaujíma rozdelenie s hustotou $P(x)$, môžeme ho získať ako kanonické rozdelenie s $T = 1$ a určením $E(x) = -\log P(x) - \log Z$, kde Z je akákoľvek vhodná kladná konštanta.

V našom prípade je hamiltonián $H(q, p)$ funkciou energie hamiltonského systému pre stav, ktorý je kombináciou polohy q a hybnosti p. Preto združenú hustotu rozdelenia nad stavmi tohoto systému definujeme ako

$$P(q, p) = \frac{1}{Z} \exp\left(\frac{-H(q, p)}{T}\right). \tag{13}$$

Pokiaľ je hamiltonián separovateľný na $H(q, p) = U(q) + K(p)$, združená hustota sa dá rozložiť na

$$P(q, p) = \frac{1}{Z} \exp\left(\frac{-U(q)}{T}\right) \exp\left(\frac{-K(p)}{T}\right), \tag{14}$$

a je zrejmé, že poloha q a hybnosť p sú nezávislé a každé z nich má vlastné kanonické rozdelenie pre funkcie energií $U(q)$ a $K(p)$. Polohu q budeme používať na reprezentáciu premenných, ktoré skúmame a hybnosť p zavedieme len preto, aby sme zabezpečili chod hamiltonskej dynamiky.

V bayesovskej štatistike je bežným cieľom výskumu aposteriórne rozdelenie parametrov modelu, preto tieto parametre stotožníme s polohou q. Potom môžeme aposteriórne rozdelenie vyjadriť ako kanonické rozdelenie s $T = 1$, keď použijeme funkciu potenciálnej energie definovanú ako

$$U(q) = -\log[\pi(q)L(q|D)], \tag{15}$$

kde $\pi(q)$ predstavuje hustotu apriórneho rozdelenia parametrov modelu a $L(q|D)$ vierohodnostnú funkciu pre dáta D.

Algoritmus hamiltonského Monte Carlo simulovania Teraz máme definované potrebné teoretické základy pre uvedenie algoritmu hamiltonského Monte Carlo simulovania (ďalej ako "algoritmus HMC"). Algoritmus HMC je možné použiť len na simulovanie výberu zo spojitého rozdelenia, ktorého hustotu vieme vyjdariť (až na nejakú normalizačnú konštantu). Taktiež musíme vedieť vypočítať parciálne derivácie logaritmu tejto hustoty. Tieto derivácie teda musia existovať, možno s výnimkou množiny bodov s nulovou pravdepodobnosťou. V tých bodoch je ale možné použiť nejakú arbitrárnu hodnotu.

Algoritmus HMC simuluje výber z kanonického rozdelenia pre q a p, definované rovnicou 14. Parameter q má teda rozdelenie, ktoré nás zaujíma a ktoré je určené funkciou potenciálnej energie $U(q)$. Čo sa týka rozdelenia parametra p, ktoré je nezávislé na q, to si môžeme určiť podľa našej vôle, pomocou funkcie kinetickej energie $K(p)$. V súčasnosti je zvykom použiť pri algoritme HMC kvadratickú funkciu kinetickej energie, zvolenie ktorej má za následok, že p má normálne rozdelenie so strednou hodnotou 0. Komponenty p sa najčastejšie volia tak, aby boli navzájom nezávislé a i-ty komponent mal varianciu m_i. Potom funkcia kinetickej energie, ktorá určuje toto rozdelenie (za podmienky $T = 1$) je $K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$.

Každá iterácia algoritmu HMC má dva kroky. Prvý krok zmení len hybnosť, druhý môže zmeniť polohu aj hybnosť.

V prvom kroku náhodne vyberieme nové hodnoty hybnosti z ich zodpovedajúceho normálneho rozdelenia, nezávisle na hodnotách polohy.

V druhom kroku, počnúc stavom (q, p), je hamiltonská dynamika simulovaná použitím metódy leapfrog pre L krokov pri veľkosti krokov ϵ. Tu, L a ϵ sú parametre algoritmu HMC, ktoré je potrebné nakalibrovať pre jeho vhodný beh. Premenné hybnosti sú na konci tejto trajektórie prenásobené -1, čím dostávame kandidáta na nasledujúci stav (q^*, p^*). Tento kandidát je prijatý ako ďalší stav markovovského reťazca s pravdepodobnosťou

$$min(1, \mathrm{e}^{(-H(q^*,p^*)+H(q,p))}) = min(1, \mathrm{e}^{(-U(q^*)+U(q)-K(p^*)+K(p))}) \tag{16}$$

Ak je navrhnutý stav zamietnutý, tak za nasledujúci stav sa určí súčasný stav.

1.2.2 Algoritmus No-U-Turn

Algoritmus HMC je silno závislý od zvolenia vhodných parametrov L a ϵ. Ak je ϵ príliš veľké, simulácia bude nepresná a vyústi v malú akceptáciu nových kandidátov. Ak je ϵ príliš malé, algoritmus spraví veľa výpočtov pre veľmi malé kroky. Ak je L príliš veľké, algoritmus HMC bude generovať trajektórie, ktoré sa budú otáčať a vracať smerom späť. Ak je L príliš malé, po sebe nasledujúce vzorky budú príliš blízko pri sebe, čo spôsobí správanie podobné náhodnej prechádzke (ktoré sa snažíme pomocou hamiltonskej dynamiky obmedziť).

Použitie algoritmu HMC je teda limitované potrebou kalibrácie veľkosti kroku ϵ a počtom krokov L. Kalibrácia týchto parametrov si vyžaduje okrem expertízy v danej oblasti aj niekoľko predbežných behov algoritmu. Dosť problematický je výber L, pretože je náročné nájsť jednoduchú metriku na to, kedy považujeme vzniknutú trajektóriu za prikrátku, pridlhú, alebo akurát. Preto voľba často závisí na heuristike založenej na autokorelačných štatistikách v predchádzajúcich behoch.

Algoritmus No-U-Turn pre markovovské Monte Carlo simulovanie (z angl. "žiadne otočenie do protismeru", ďalej ako "algoritmus NUTS") eliminuje potrebu nastavenia týchto parametrov.

Parameter L, t.j. počet krokov je nastavovaný v algoritme NUTS počas výpočtu ďalšieho kroku v metóde leapfrog. Algoritmus NUTS vytvára dráhu vpred aj vzad v pomyselnom čase. Najprv spraví jeden krok, náhodne vpred alebo vzad, potom dva kroky, štyri kroky, atď. Toto zdvojnásobovanie krokov implicitne vytvára vyvážený binárny strom, ktorého vrcholy zodpovedajú stavu systému. Zdvojnásobovanie krokov končí, keď sa v ľubovoľnom vyváženom podstrome trajektória z ľavej do pravej časti začne vracať naspäť (t. j. táto fiktívna častica sa začne otáčať do protismeru). V tomto momente algoritmus NUTS zastaví simuláciu a zvolí nasledujúci stav z dovtedy vypočítaných stavov.

Parameter ϵ, t. j. veľkosť kroku je zvolený počas zahrievacieho času a počíta sa metódami stochastickej optimalizácie. Tieto metódy sú podrobne popísané v [15]

Oba tieto parametre sa v algoritme NUTS nastavia počas zahrievacieho času.

1.3 Program Stan

Pre bayesovskú štatistickú inferenciu používame program Stan. Stan je vyvíjaný v jazyku C++ a má nadstavby pre rôzne programovacie prostredia, vrátane prostredie jazyka `R`. Pre simuláciu výberu z aposteriórneho rozdelenia používa Stan algoritmus NUTS.

Stan nám umožňuje špecifikovať model v programovacom jazyku podobnom C++:

```
data {
  int<lower=0> n; ## pocet riadkov
  real x[n];      ## vektor prediktorov
  int y[n];       ## vektor zavislej premennej
}
parameters {
  real alpha;
  real beta;
}
model {
  for(i in 1:n) {
    y[i] ~ bernoulli_logit(alpha + x[i] * beta);
  }
}
```

Horeuvedený je príklad zdrovojého kódu pre jednoduchý model logistickej regresie $logit(Pr[Y = 1]) = \alpha + X \cdot \beta$.

V časti `data` špecifikujeme pre program Stan, aké vstupy má očakávať. Každá premenná v zdrojovom kóde musí mať vopred deklarovaný dátový typ. Premennou `n` sme označili počet riadkov dát, ktoré vstúpia do tohto modelu. Dátovým typom je `int`, z anglického integer, t. j. celé číslo. Ohraničenia na premenné sa uvádzajú v špicatých zátvorkách za dátovým typom, v tomto prípade sme počet riadkov ohraničili zdola nulou. Premenná `x` zodpovedá `n`-rozmernému vektoru prediktorov s dátovým typom `real`, z anglického real, t. j. reálnym číslam. Premenná `y` zas zodpovedá `n`-rozmernému vektoru predikovaných hodnôt. Dátový typ sme nastavili `int`, pretože pri logistickej regresii nadobúda táto premenná len hodnoty 0 a 1.

V časti `parameters` deklarujeme všetky parametre modelu, ktorých rozdelenie budeme simulovať. V tomto prípade sú to intercept α a parameter β prislúchajúci prediktoru `x`.

V prípade, že by sme chceli vykonať úpravy alebo transformácie dát alebo parametrov priamo v modeli, slúžia na to časti `transformed data` a `transformed parameters`. Napríklad pre výpočet logaritmu hodnôt `x` by sme najprv deklarovali novú premennú `real new_x[n]` a potom jej zložkám priradili logaritmus pomocou príkazu `new_x <- log(x)` .

V časti model už špecifikujeme samotný model. V našom prípade sa jedná o logistickú regresiu, kde premennú `y` predikujeme pomocou interceptu a prediktora `x`. V tejto časti tiež určujeme apriórne rozdelenia parametrov - nešpecifikovaný prior priradí parametrom neinformatívny prior, t. j. apriórne rovnomerné rozdelenie. Premenné `y` odhadujeme pomocou cyklu `for`, pretože funkcia `bernoulli_logit` nie je (narozdiel od funkcií iných rozdelení) vektorizovaná.

Takýto zdrojový kód pre model je programom Stan skompilovaný a môžeme ho použiť v prostredí jazyka `R`. V prostredí jazyka `R` je potrebné najprv spojiť všetky dáta, ktoré vstupujú do modelu príkazom `list()` tak, aby názvy premenných v `list`-e zodpovedali názvom v zdrojovom kóde modelu v časti `data`. Simuláciu výberu z aposteriórneho rozdelenia parametrov potom zavoláme príkazom:
`fit <- stan('kod_modelu.stan', data = data_pre_model)`.

1.3.1 Príklad

Na jednoduchom príklade teraz ukážeme, ako sa líši odhadovanie parametrov modelu logistickej regresie pomocou metódy najmenších štvorcov a simuláciou výberu z aposteriórneho rozdelenia pomocou markovovského Monte Carlo simulovania. Budeme odhadovať parametre modelu $logit(Pr[Y = 1]) = \alpha + X \cdot \beta$.

Najskôr sme naprogramovali jednoduchú funkciu na generovanie dát. Argumentami tejto funkcie sú veľkosť generovaného súboru, parametre modelu a disperzia normálneho rozdelenia so strednou hodnotou 0. Hodnoty X vygenerujeme z normálneho rozdelenia so strednou hodnotou 0 a varianciou 10. Z nich pri zadaných hodnotách parametrov modelu (intercept $\alpha = 2.54$ a $\beta = 1.66$) dopočítame hodnoty Y. Takto vygenerujeme

súbor 1000 dát.

Najprv odhadneme parametre logistického regresného modelu, daného týmito vygenerovanými dátami, upravenou metódou najmenších štvorcov, a to pomocou zabudovanej knižnice `glm`:

```
> summary(fit.mle)

Call:
glm(formula = y ~ 1 + x, family = binomial(link = "logit"), data = tab)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.90204  -0.01064   0.00002   0.01887   2.08626

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.2022     0.2966   7.425 1.13e-13 ***
x             1.3910     0.1459   9.535  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1374.61  on 999  degrees of freedom
Residual deviance:  190.45  on 998  degrees of freedom
AIC: 194.45

Number of Fisher Scoring iterations: 9
```

Obr. 1: Výstup z odhadu parametrov pomocou funkcie `glm`

Takto sme dostali bodový odhad oboch parametrov modelu. Obidva boli vyhodnotené ako signifikantné na úrovni 0.001. Naproti tomu, výstup pri odhade parametrov pomocou programu Stan vyzerá odlišne.

```
> print(fit.bayes, digits = 3)
Inference for Stan model: model.
2 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=2000.

         mean se_mean    sd    2.5%     25%     50%     75%   97.5% n_eff  Rhat
alpha   2.246   0.014 0.285   1.726   2.035   2.235   2.434   2.828   431 1.001
beta    1.416   0.006 0.144   1.152   1.315   1.412   1.508   1.715   570 1.003
lp__  -96.153   0.038 0.928 -98.689 -96.561 -95.848 -95.494 -95.245   591 1.000

Samples were drawn using NUTS(diag_e) at Sun Mar 27 20:42:31 2016.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Obr. 2: Výstup z odhadu parametrov pomocou modelu Stan

Pre každý parameter dostávame nasledovné koeficienty:

- `mean` je priemerom simulovaných hodnôt cez všetky reťazce,
- `se_mean` určuje tzv. Monte Carlo chybu. Jedná sa o odhad neistoty vyplývajúci z toho, že k dispozícii máme len obmedzený počet simulácií výberu z aposteriórneho rozdelenia daného parametra. Hodnota Monte Carlo chyby by sa mala pohybovať okolo $1 - 5\%$ zo smerodajnej odchýlky simulovaných hodnôt, ktorú nájdeme vo vedľajšom stĺpci `sd`,
- ďalšie stĺpce určujú kvantily intervalu vierohodnosti pre daný parameter. Tieto intervaly nie sú intervalmi spoľahlivosti, ako ich poznáme z klasickej štatistiky, ale jednoducho znamenajú, že na základe výsledkov modelu je určitá pravdepodobnosť (daná rozdielom kvantilov), že skutočná hodnota parametra padne do tohoto intervalu,
- `n_eff` vyjadruje tzv. efektívnu veľkosť vzorky. V prípade nulovej autokorelácie medzi simuláciami by sa táto hodnota rovnala počtu simulácií; jedná sa teda o akúsi mieru autokorelácie. Rovnosť `n_eff` a počtu simulácií ale nie je podmienkou, výrazný rozdiel môže ale naznačovať problémy s modelom alebo že by bolo vhodnejšie použiť dlhšie reťazce,
- koeficient `Rhat` popíšeme bližšie v ďalšej časti.

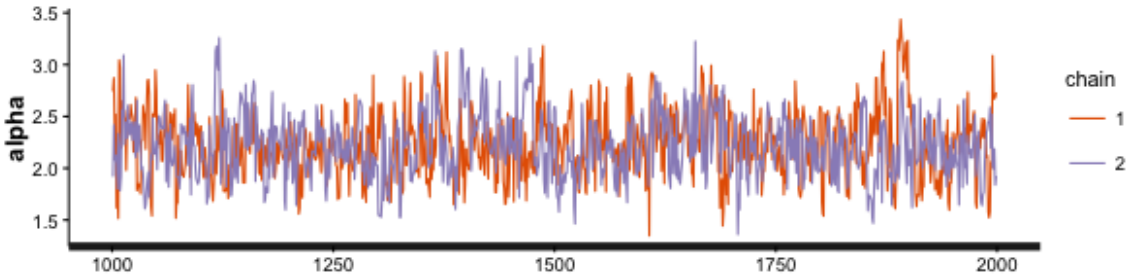
Na základe horeuvedených výsledkov môžeme vidieť, že bodové odhady parametrov z prvého prístupu sú porovnateľné s priemerom simulovaných hodnôt z aposteriórneho rozdelenia daných parametrov.

1.4 Diagnostika konvergencie k cieľovému rozdeleniu

V tejto časti uvedieme niekoľko metód, ako skontrolovať, či nám markovovský reťazec, simulovaný pri výbere z rozdelenia parametrov modelu, dokonvergoval k cieľovému rozdeleniu. Tieto metódy sú založené na porovnávaní viac markovovských reťazcov simulovaných nezávisle od seba.

1.4.1 Traceplot

Najjednoduchšia možnosť je pomocou tzv. traceplotov. Traceplot je graf vývoja simulovaných hodnôt výberu z rozdelenia parametra v čase a pre rôzne reťazce, pričom uvažujeme iterácie po zahrievacom čase.



Obr. 3: Príklad traceplotu

Na obrázku je uvedený príklad traceplotu parametra pre 2 reťazce. Čo na traceplote hľadáme, je akási "tlstá, chlpatá húsenica" alebo niečo, čo by sme mohli označiť za "trávnaté" [6]. Cieľom je, aby sa odhady z rôznych reťazcov pohybovali okolo rovnakej hodnoty a aby sa navzájom miešali. Pri tejto základnej diagnostike je potrebné skontrolovať traceploty všetkých parametrov.

1.4.2 Konvergenčné kritérium $\hat{R}$

Ďalšia metóda na kontrolu konvergencie je tzv. Gelmanov a Rubinov "potential scale reduction factor", označovaný $\hat{R}$ [2]. Koeficient $\hat{R}$ nám poskytuje odhad konvergencie

založený na variancii parametra v rámci reťazca a variancii parametra naprieč reťazcami.

Podľa [2], $\hat{R}$ vypočítame nasledovným spôsobom. Uvažujme m nezávislých simulovaných markovovských reťazcov s n iteráciami po zahrievacom čase. Označme ψ_{jt} hodnotu parametra ψ v j-tom reťazci v t-tej iterácii. Ďalej označme $\bar{\psi_{j\cdot}}$ priemer hodnôt parametra ψ v j-tom reťazci zo všetkých iterácií a $\bar{\psi_{\cdot\cdot}}$ priemer hodnôt parametra ψ naprieč všetkými reťazcami, taktiež zo všetkých iterácií. Odhad variancie parametra ψ označíme $\hat{\sigma}_+^2$ a vypočítame ako vážený priemer variancie v rámci reťazcov, označenej ako W a variancie naprieč reťazcami, označenej $\frac{B}{n}$,

$$\begin{aligned} W =& \frac{1}{m(n-1)} \cdot \sum_{j=1}^{m} \sum_{t=1}^{n} \left(\psi_{jt} - \bar{\psi_{j\cdot}}\right)^2 \\ \frac{B}{n} =& \frac{1}{m-1} \cdot \sum_{j=1}^{m} \left(\bar{\psi_{j\cdot}} - \bar{\psi_{\cdot\cdot}}\right)^2. \end{aligned}$$

Potom

$$\hat{\sigma}_+^2 = \frac{n-1}{n} \cdot W + \frac{B}{n}.$$

Do variancie je zahrnutá aj korekcia a výsledný odhad variancie dostávame ako

$$\hat{V} = \hat{\sigma}_+^2 + \frac{B}{mn}.$$

Potom koeficient $\hat{R}$ vypočítame ako

$$\hat{R} = \frac{\hat{V}}{W}$$

Ak je $\hat{R}$ veľké, znamená to, že odhad variancie $\hat{\sigma}_+^2$ sa dá znížiť väčším počtom simulácií, alebo že väčší počet simulácií zvýši W, pretože dovtedy nasimulované reťazce zatiaľ nenasimulovali dostatok rôznych hodnôt z cieľového rozdelenia. Ak je $\hat{R}$ blízke 1, môžeme tvrdiť, že každá z m množín n iterácií je blízko k cieľovému rozdeleniu. V modeloch sa teda snažíme získať hodnotu $\hat{R}$ rovnú 1, prípadne nanajvýš 1.1 [6]. Hodnotu 1 tiež $\hat{R}$ nadobúda teoreticky pre počet simulácií blížiacich sa k nekonečnu.

2 Dáta a ich spracovanie

V tejto kapitole predstavíme dáta, ktoré sme mali k dispozícii pre použitie v našej práci. Ďalej predstavíme spôsoby ich spracovania a použité algoritmy.

2.1 Administratívne a klinické dáta

V oblasti zdravotníctva existujú dva druhy dát. Administratívne dáta sú všetky dáta, ktoré zdravotnícky personál vykazuje zdravotným poisťovniam pre potreby zúčtovania platieb za zdravotnú starostlivosť. Tieto dáta má povinnosť vykazovať a odosielať poisťovni každý poskytovateľ zdravotnej starostlivosti, ktorý má s poisťovňou zmluvný vzťah. V týchto dátach nájdeme hlavne informácie o čase a mieste návštevy pacienta u lekára, v lekárni, či inom zdravotníckom zariadení. Ďalej informácie týkajúce sa zdravotníckeho výkonu, diagnózu súvisiacu s výkonom, informácie o výške platieb od zdravotnej poisťovne, v niektorých prípadoch aj lekára alebo ambulanciu, ktorý návštevu odporučil (napríklad keď všeobecný lekár predpisuje lieky na odporúčanie špecializovaného lekára, ktoré by v inom prípade nemohol).

Každé ochorenie a symptóm má pridelený kód, ktorý slúži na jeho identifikáciu. Medzinárodný klasifikačný systém, ktorého úprava, nazývaná medzinárodná klasifikácia chorôb (skrátene MKCh-10), používa na Slovensku, predstavuje systematicky triedený a hierarchicky usporiadaný zoznam chorôb a číslo 10 predstavuje desiatu revíziu [19].

Za klinické dáta sa považujú všetky dáta, ktoré súvisia s diagnostikou a liečbou pacienta. Tieto dáta sú uložené v zdravotnej dokumentácii pacienta. Zdravotná poisťovňa k nim má prístup len za účelom revíznej činnosti, pre bežný chod poisťovne nie sú potrebné. Jedná sa najmä o hodnoty merané počas vyšetrenia a diagnostiky (napr. hodnoty systolického/diastolického tlaku, pre diagnózu diabetes mellitus napr. hodnoty glykémie/glykovaného hemoglobínu). Prístup k takýmto dátam majú okrem samotných pacientov iba lekári, ktorí ich potrebujú pri analýze zdravotného stavu pacienta a výkone práce.

V našej práci využívame iba administratívne dáta. Tie sme spracúvali v súlade so zákonom o ochrane osobných údajov [27].

2.2 Dátové zdroje

Všetky dáta potrebné k našemu modelu nám boli poskytnuté v 3 tabuľkách. Z dôvodu, že v dátach sa vyskytujú osobné údaje chránené zákonom o ochrane osobných údajov [27], pre ilustráciu dátových zdrojov uvádzame len hlavičky použitých dát.

2.2.1 Hospitalizácie

Tabuľka s hospitalizáciami zahŕňa všetky vykázané údaje o príjmoch, prekladoch a prepusteniach pacientov z nemocníc. Ďalej údaje o niektorých pomocných výkonoch (napríklad laboratórne výkony, transfúzie krvi, transplantácie, iná nákladná zdravotná starostlivosť) a dáta o tzv. osobitne hradených výkonoch. Z tabuľky s dátami k hospitalizáciám sme využili:

- ICP
  Identifikačné číslo poistenca, ktoré používa poisťovňa. Tento identifikátor je rovnaký naprieč použitými tabuľkami.
- DATUMN, DATUMP
  Dátum nástupu a dátum prepustenia z nemocnice, prípadne prekladu na iné oddelenie v rámci jednej nemocnice.
- DIAGN, DIAGP
  MKCh-10 kód diagnózy vykázanej pri nástupe a prepustení z nemocnice alebo oddelenia. Tieto kódy sme tiež zoskupovali do CCS kódov.
- AMBV
  Kód ambulancie alebo oddelenia, na ktorom sa pacient nachádzal.
- SKUPDRUHPPZS
  Skupina/druh poskytnutej zdravotnej starostlivosti. Tento atribút sme využili iba kvôli filtrovaniu dát relevantných k hospitalizáciám.
- PRODUKT
  Produkt poskytnutej zdravotnej starostlivosti. Jedná sa o unikátny identifikátor výkonu, zdravotnej pomôcky, lieku, a pod. Taktiež primárne využitý pri filtrovaní dát.

- OPEROV

  Príznak prítomnosti operácie pacienta, t.j. príznak, či pacient podstúpil ľubovoľný operačný zákrok.

- AKUTNA

  Príznak akútnej hospitalizácie, t.j. príznak, či sa jednalo o hospitalizáciu vopred naplánovanú, alebo akútnu.

2.2.2 Poistenci

Tabuľka s údajmi o poistencoch obsahuje dáta o tom, akým spôsobom hradí poistenec verejné zdravotné poistenie, dátumy poistenia, odchodu z poisťovne, narodenia a úmrtia. Z tabuľky s údajmi o poistencoch sme využili stĺpce:

- ICP

  Identifikačné číslo poistenca.

- VEK

  Vek poistenca k 1.1.2014, určený z dátumu narodenia.

- DATUMRTIA

  Dátum úmrtia poistenca. Ak je poistenec nažive, tento atribút nemá hodnotu.

2.2.3 Vykázaná zdravotná starostlivosť

Tabuľka o vykázanej zdravotnej starostlivosti obsahuje informácie o každom vykázanom výkone, poskytnutej zdravotnej pomôcke, či liekom, atď. Ku každému riadku sa viaže dátum, vykázaná diagnóza, informácie o výkone, prípadný odporúčací lekár, a pod. Taktiež údaje o výške úhrady za daný výkon alebo vyšetrenie. Z tabuľky s vykázanou zdravotnou starostlivosťou sme využili nasledujúce stĺpce:

- ID

  Primárny kľúč tabuľky

- ICP

  Identifikačné číslo poistenca.

- DATVYKONUAKCEPT

  Dátum kontaktu u poskytovateľa zdravotnej starostlivosti. V tomto prípade to môže byť dátum vyšetrenia u lekára, ale aj dátum vyzdvihnutia liekov na predpis v lekárni.

- KODDIAGAKCEPT

  MKCh-10 kód diagnózy súvisiacej s vyšetrením alebo výkonom. Tieto kódy boli v ďalšom spracovaní zoskupené.

2.3 Identifikácia hospitalizácií

Dáta o hospitalizáciách pacientov sa nevykazujú spôsobom, kedy jednej hospitalizácii zodpovedá jeden riadok. Hospitalizácie sú vykazované fragmentovane - preklad na iné oddelenie alebo do inej nemocnice znamená ďalší riadok. Nový mesiac tiež znamená vytvorenie nového riadku. Taktiež sa v dátach môžu vyskytnúť rôzne prídavné výkony (napríklad transfúzie krvi, atď.). Celý proces od prijatia pacienta do nemocnice až po jeho prepustenie pre nás predstavuje jeden logický celok, ktorý nám bude do modelu vstupovať ako jeden riadok. Takto rozdrobené dáta bolo potrebné algoritmicky spojiť.

Algoritmus pre identifikáciu hospitalizácií v dátach sme konzultovali a vyvíjali spolu s [21]. Skript na základe logických pravidiel priradí riadkom z tabuľky s dátami o hospitalizáciách číslo hospitalizácie, do ktorej daný riadok patrí. Takto vieme následne spojiť dáta podľa tohoto identifikátora.

Skript sme najprv aplikovali napísaný v jazyku `R`. Jeho beh ale trval na našich dátach približne hodinu, preto sme ho prepísali do jazyka C++. Takto upravený skript sme tiež vedeli spustiť v prostredí jazyka `R`, využitím knižnice Rcpp [10]. Prepísaním sa beh skriptu zrýchlil na niekoľko desiatok sekúnd.

Aplikovaním tohoto algoritmu teda priradíme k riadkom arbitrárne číslo hospitalizácie, ku ktorej daný riadok patrí. Cez tieto čísla potom pospájame naše dáta tak, aby sme všetky relevantné informácie pre jednu hospitalizáciu obsiahli v jednom riadku, ktorý bude po určitých úpravách vstupovať do modelu.

Spôsob, akým budú hospitalizácie pripisované nemocniciam sme určili po konzultáciách s [26]. Ak sa jednalo o hospitalizáciu, pri ktorej bol pacient liečený vo viacerých

rôznych nemocniciach, túto hospitalizáciu pripíšeme obidvom nemocniciam. To znamená, že ak bol pacient počas hospitalizácie liečený v dohromady dvoch nemocniciach, táto hospitalizácia pre nás bude predstavovať dva riadky s rovnakými atribútmi, ktoré sa budú líšiť jedine nemocnicou.

2.4 Zoskupovanie kódov diagnóz

Clinical Classifications Software (ďalej ako CCS) je schéma kategorizácie diagnóz a procedúr [7], založená na medzinárodnej verzii MKCh-10 kódov (označované ICD-10, z angl. International Classification of Diseases). Tieto kódy sú zoskupené do menšieho množstva klinicky významných kategórií, ktoré sú niekedy použiteľnejšie pre prezentáciu rôznych deskriptívnych štatistík. Túto kategorizáciu sme pre malé množstvo MKCh-10 kódov, ktoré sa líšili od medzinárodnej verzie ICD-10, rozšírili.

Zoskupenie MKCh-10 kódov diagnóz do CCS kódov diagnóz nám prináša niekoľko výhod. V prvom rade takto zoskupené kódy sú prijateľnejšie ako vstupy do modelu, pretože sa zníži počet rôznych kódov. Niektoré MKCh-10 kódy môžu tiež byť zriedka vykazované.

Takéto zoskupenie nám taktiež zníži chybovosť z vykazovania MKCh-10 kódov. Môže sa stať, že rozdielni lekári vykážu rovnakú diagnózu rozdielnymi, ale podobnými MKCh-10 kódmi, napríklad zo zvyku.

V ďalšom, keď sa budeme odkazovať na diagnózy, vždy sa tým budú myslieť CCS kódy diagnóz (nie MKCh-10 kódy), pokiaľ nie je uvedené inak.

2.5 Výber dát

V našej práci sme sa obmedzili na hodnotenie nemocníc pre rok 2014. To znamená, že využijeme dáta o vykázanej zdravotnej starostlivosti od 1. januára 2013 do 31. decembra 2014. Dáta potrebujeme aj z roku 2013, aby sme mali prístup k administratívnym dátam do 12 mesiacov dozadu od akejkoľvek hospitalizácie v roku 2014.

Do modelu vstupujú dáta, ktoré spĺňajú nasledovné podmienky podľa [5], pričom pri každej podmienke uvádzame ekvivalent v našom modeli a prípadné vysvetlenie, prečo a ako sme podmienku upravili:

1. pacient je poistený v systéme MediCare,

- prirodzeným ekvivalentom sa javí zahrnúť pacientov poistených v poisťovni Dôvera, najmä z dôvodu dátovej dostupnosti,

2. pacient starší ako 65 rokov

   - MediCare je program, v ktorom výrazná väčšina poistencov sú seniori starší ako 65 rokov. Túto podmienku sme preto vypustili a model rozšírili na všetkých poistencov poisťovne Dôvera bez ohľadu na ich vek,

3. pacient prepustený nažive z nefederálnej nemocnice akútnej zdravotnej starostlivosti,

   - pretože systém zdravotnej starostlivosti je v Spojených štátoch amerických organizovaný inak, ako na Slovensku, medzi uvažované nemocnice sme zaradili len všeobecné a fakultné nemocnice; podmienku prepustenia nažive sme zachovali,

4. pacient nebol presunutý do inej nemocnice akútnej zdravotnej starostlivosti,

   - preklady medzi nemocnicami máme pokryté spájaním prekladov do logických celkov, predstavujúcich jednu pretrvávajúcu hospitalizáciu,

5. pacient poistený v systéme MediCare, časť A, 12 mesiacov pred dátumom hospitalizácie (vrátane),

   - pretože verejné zdravotné poistenie na Slovensku platí v prípade zmeny poisťovne od začiatku do konca kalendárneho roka, táto podmienka bude pre hospitalizácie z určitého roka splnená, ak bol pacient poistený v poisťovni Dôvera aj predchádzajúci rok.

V súlade s [5] tiež vylučujeme nasledovné hospitalizácie:

1. pacient prijatý do nemocníc pre liečbu rakoviny, financovaných na základe "Prospective Payment System",

   - pretože v našom prostredí majú špecializované nemocnice výrazne rozdielnu štruktúru pacientov ako všeobecné a fakultné, toto pravidlo sme zovšeobec-

nili práve na špecializované nemocnice; do nášho modelu teda vstupujú len všeobecné a fakultné nemocnice,

2. pacient bez prihlásenia do systému MediCare aspoň 30 dní po skončení hospitalizácie,

    - toto pravidlo sme nahradili úmrtím poistenca do 30 dní, z vyššieuvedeného dôvodu pri zmene zdravotnej poisťovne,

3. pacient prepustený napriek odporúčaniu lekára,

    - jedná sa o prípady, kedy sa pacient rozhodne odísť z nemocnice napriek termínu odporúčaného lekárom; tieto prípady sa z našich dát zistiť nedajú, toto pravidlo sme boli nútení vypustiť,

4. pacient prijatý na psychiatrickú diagnózu, rehabilitáciu, alebo liečbu rakoviny,

    - tieto prípady sme odstránili v súlade so zdrojovou metodikou.

2.6 Rehospitalizácia ako negatívny dôsledok

Rehospitalizácie sme definovali v úvode ako hospitalizácie vzniknuté do 30 dní od konca poslednej hospitalizácie daného pacienta a predstavujú negatívny dôsledok zdravotnej starostlivosti. Ale nie každú identifikovanú rehospitalizáciu považujeme v našom modeli za negatívny dôsledok. Napríklad plánované hospitalizácie môžu byť rehospitalizáciami (lebo ďalšia hospitalizácia pacienta je naplánovaná do 30 dní), ale považujú sa za súčasť zdravotnej starostlivosti. Za rehospitalizácie v našom modeli preto nepovažujeme ľubovoľné rehospitalizácie, ale tzv. neplánované rehospitalizácie. Na identifikáciu neplánovaných rehospitalizácií používa CMS určitý algoritmus [5], ktorý stojí na troch základných princípoch:

- niektoré špecifické druhy starostlivosti sa vždy považujú za plánované (transplantácie, udržiavacia chemoterapia/imunoterapia, rehabilitácia),

- inak je plánovaná rehospitalizácia definovaná ako neakútna rehospitalizácia kvôli naplánovanej procedúre,

- hospitalizácie na akútne choroby alebo komplikácie nie sú nikdy plánované.

Podľa [5] sú pri určovaní potenciálne plánovaných hospitalizácií v dátach prehľadávané CCS kódy procedúr (nie diagnóz). Procedúrou sa rozumie akýkoľvek vykonaný zákrok počas hospitalizácie. Procedúry sa ale v našom prostredí nevykazujú, len diagnózy. Preto sme museli tento algoritmus upraviť. Namiesto prehľadávania kódov procedúr a označovania rehospitalizácií za potenciálne plánovaných využívame príznak akútnej hospitalizácie, vykazovaný cez HospiCOM.

HospiCOM je vykazovací systém vyvinutý pre poisťovňu Dôvera, ktorý umožňuje online sledovanie hospitalizácií a plynulé manažovanie plánovanej zdravotnej starostlivosti. Ak pacient potrebuje akútnu pomoc, je bezodkladne hospitalizovaný. Ak ide o plánovanú starostlivosť, nemocnica pacienta zaeviduje do systému HospiCOM a jeho poradie určí tzv. ticket [17]. Preto vieme vďaka príznaku akútnej hospitalizácie určiť, či sa jednalo o plánovanú alebo neplánovanú hospitalizáciu. Tým nám odpadá krok prehľadávania hospitalizácií. Následne sme v súlade s [5] z plánovaných hospitalizácií filtrovali niektoré akútne ochorenia a komplikácie. Použitý algoritmus potom vyzerá nasledovne:

1. hospitalizácie z dôvodu udržiavacej chemoterapie alebo pôrodov sme označili za plánované,

2. hospitalizácie, ktoré boli označené za akútne, sme označili za neplánované,

3. zvyšné hospitalizácie (t.j. plánované podľa systému HospiCOM) sme označili za neplánované práve vtedy, pokiaľ sa jednalo o hospitalizácie zo skupiny akútnych diagnóz alebo komplikácií podľa [5].

Použitím tohto algoritmu vzniká predpoklad, že všetkým rehospitalizáciám sa zabrániť nedá. To znamená, že hodnotenie nemocníc bude v konečnom dôsledku založené na porovnávaní s akýmsi národným priemerom rehospitalizácií.

Naša klasifikácia rehospitalizácií je z veľkej časti založená na príznaku akútnosti. Správnosť vykazovania sme preto konzultovali s [26]. Z konzultácií vyplynulo, že sa stávajú prípady, kedy je tento príznak vykazovaný nesprávne. Vždy sa ale jedná o prípady, kedy má ísť o plánovanú hospitalizáciu, ktorá je označená za neplánovanú (napr. z dôvodu preskočenia poradovníka). Napriek tomu, že tieto prípady sa dejú relatívne často,

sme klasifikáciu ponechali tak ako je. Dôvodom je, že ak je hospitalizácia označená ako akútna, do nášho modelu vstupuje ako negatívny jav a preto má takéto konanie za následok zvýšenie počtu rehospitalizácií danej nemocnice. Opačný prípad sa v praxi nevyskytuje. Preto sme sa rozhodli po konzultácii použitie príznaku akútnej hospitalizácie ponechať, pretože nesprávne vykázanie môže mať za následok len zhoršenie hodnotenia danej nemocnice.

Našu úpravu algoritmu sme taktiež konzultovali s [9]. Z konzulácií vyplynulo, že sme zvolili vhodné úpravy algoritmu.

2.7 Zhrnutie

Pri úprave dát pre použitie v modeli teda vychádzame z dát o hospitalizáciách za rok 2014. Dáta o hospitalizáciách sme spojili tak, aby jeden riadok obsahoval atribúty jednej hospitalizácie. V hospitalizáciách sme identifikovali relevantné rehospitalizácie a tak sme dostali dáta, ktoré nám v ďalšom vstúpia do prediktívneho modelu.

3 Prediktívny model

V tejto kapitole opíšeme prediktívny model na modelovanie pravdepodobnosti rehospitalizácie, použité prediktory a technické detaily.

Základnou myšlienkou v použitom modeli je predikovanie pravdepodobnosti, že daná hospitalizácia je rehospitalizáciou. Dáta máme spracované tak, že jednej hospitalizácii zodpovedá v našom modeli jeden riadok. Závislá premenná v modeli bude mať hodnotu 1, ak hospitalizácia v danom riadku je rehospitalizáciou podľa algoritmu z kapitoly 2.6 a nastala do 30 dní od poslednej hospitalizácie súvisiaceho pacienta. Inak bude mať závislá premenná hodnotu 0. Preto budeme závislú premennú modelovať logistickou regresiou. V modeli je tiež zahrnutý efekt nemocníc, preto bude regresia zároveň viacúrovňová. Odhad rozdelení parametrov modelu získame pomocou programu Stan.

3.1 Prediktory

Do modelu nám vstupujú nasledovné prediktory:

3.1.1 Vek poistenca

Vek poistenca je jedným z najčastejších prediktorov v modelovaní v súvislosti so zdravotníctvom. Do nášho modelu vstupuje vek ako absolútna hodnota, ktorú sme pre danú hospitalizáciu získali z tabuľky o poistencoch.

3.1.2 Príčina hospitalizácie

Príčinou hospitalizácie sa v našom modelovaní rozumie hlavná diagnóza súvisiaca s hospitalizáciou. Tento prediktor sme z dát určili ako poslednú vykázanú diagnózu v danej hospitalizácii. Predpoklad je, že pri prepúšťaní pacienta do prostredia neakútnej starostlivosti (napr. aj domov) je už známa príčina, pre ktorú bol pacient hospitalizovaný. Pri začiatku hospitalizácie sú síce vykazované diagnózy, môže sa ale jednať o symptómy ochorenia, ktoré zatiaľ nie je známe. Tento prediktor vstúpi do modelu ako skupina dummy premenných, ich počet bude zodpovedať počtu rôznych diagnóz.

3.1.3 Komorbidity

V našom modeli ďalej nájdeme 31 prediktorov, ktoré tvoria dummy premenné skupín diagnóz. Jedná sa o tzv. komorbidity, t.j. zdravotné obmedzenia alebo ochorenia, ktoré sa u pacienta vyskytovali navyše popri hlavnej príčine hospitalizácie. Tieto prediktory slúžia na zachytenie chorobnosti pacientov, rôznorodosti skupiny pacientov v danej nemocnici a sú klinicky relevantné pri predikovaní rehospitalizácií [5]. Pre každú skupinu sa prehľadáva výskyt v tabuľke vykázanej zdravotnej starostlivosti a predchádzajúcich hospitalizáciách, do 12 mesiacov od začiatku danej hospitalizácie. V súvislosti s dátami o vykázanej zdravotnej starostlivosti sa prehľadávajú nielen kontakty u všeobecných lekárov, či specialistov, ale aj výbery liekov v lekárňach, a pod. Pre označenie pacienta ako komorbidného v jednej z týchto skupín stačí jediný výskyt v dátach.

3.1.4 Intercept špecifický pre nemocnicu

Intercept, ktorý sa líši naprieč nemocnicami, je prediktor druhej úrovne. Tento prediktor je teda zároveň modelovaný osobitnou regresiou. Intercept sa skladá z dvoch zložiek: posunu, tzv. offsetu a koeficientu zachytávajúceho varianciu pomeru hospitalizácií medzi nemocnicami.

Offset, ktorý je zložkou interceptu je tzv. národný priemer rehospitalizácií. Voči tomuto priemeru budu neskôr v metodike porovnávané nemocnice. Tento priemer sa vypočíta ako pomer rehospitalizácií naprieč všetkými nemocnicami za daný rok. Pre rok 2014 má tento koeficient hodnotu 6.758%

Potom, koeficient zachytávajúci varianciu pomeru hospitalizácií medzi nemocnicami, bude vycentrovaný okolo hodnoty 0.

3.2 Rozdelenie dát do skupín

Autori [16] uvádzajú, že jeden model pre všetky hospitalizácie by nebol informatívny v súvislosti so zlepšovaním nemocníc. Taktiež by nezachytával rozdiely, ako vplývajú prediktory na rôzne ochorenia. Pri vývoji modelu nemocnice tiež nemali dostatočné množstvo hospitalizácií na to, aby mohol existovať osobitný model pre každé ochorenie. Preto autori identifikovali 5 skupín, ktorých ochorenia mali podobné pomery rehospitalizácií a morbidity, ale tiež mali podobnú organizáciu zdravotnej starostlivosti,

podobné tímy lekárov, a generovali dostatočne veľkú vzorku pre väčšinu nemocníc. Jedná sa o tieto skupiny:

1. kardiorespiračná
2. kardiovaskulárná
3. nervová
4. operačná
5. medicínska

Prvé tri skupiny sú pomenované podľa sústavy, s ktorou súvisia. Operačnú skupinu tvoria všetky operačné zákroky a medicínsku skupinu tvoria zvyšné diagnózy.

Náš spôsob zaraďovania hospitalizácií do skupín sa tiež líši od pôvodného. Narozdiel od [5], nemusíme hľadať v dátach kódy procedúr, ktoré by mohli byť operáciou, pretože naše dáta priamo obsahujú príznak prítomnosti operácie pacienta. Hospitalizácie s týmto príznakom preto zaradíme do operačnej skupiny. Zvyšné hospitalizácie zaradíme do zodpovedajúcich skupín podľa hlavného dôvodu hospitalizácie, zoznam ktorých sa zhoduje s [5].

3.3 Viacúrovňová logistická regresia

Pravdepodobnosť rehospitalizácie teda modelujeme viacúrovňovou logistickou regresiou. Zároveň máme 5 osobitných modelov, ktoré môžu mať parametre pri rovnakých prediktoroch rôzne. Viacúrovňovú regresiu zapíšeme ako:

$$\begin{aligned} logit(Pr[Y_i = 1]) &= \alpha_j + A_{j[i]} \cdot \theta + D_{j[i]} \cdot \gamma + X_{j[i]} \cdot \beta \\ \alpha_j &= \mu + \eta_j \\ \eta &\sim N(0, \sigma_\eta^2), \end{aligned}$$

kde α_j predstavuje intercept špecifický pre nemocnicu j. $A_{j[i]}$, $D_{j[i]}$ a $X_{j[i]}$ zas predstavujú vek pacienta, dôvod hospitalizácie a komorbidity v hospitalizácii i, ktorá prislúcha nemocnici j. Parameter θ prislúcha veku a je teda jednorozmerný. Parameter γ prislúcha dôvodu hospitalizácie a jeho rozmer závisí od počtu rôznych CCS kódov diagnóz

v danej skupine. Napokon, parameter β prislúcha matici s komorbiditami a preto je 31-rozmerný.

V ďalšom uvedieme zdrojový kód programu Stan, ktorý predstavuje implementáciu použitého regresného modelu.

```
data {
  int<lower=0> n; ## pocet riadkov
  int<lower=0> h; ## pocet nemocnic
  int<lower=0> c; ## pocet prediktorov v arrayi komorbidit
  int<lower=0> d; ## pocet diagnoz
  int<lower=0> k; ## pocet roznych komorbidit
  int y[n];       ## vektor zavislej premennej
  int X[n,c];     ## matica komorbidit
  int A[n];       ## vektor s vekmi
  int D[n];       ## vektor diagnoz
  int<lower=0> hz[n];
  real NWRR;      ## narodny priemer hospitalizacii
  real prior_mu;
  real prior_sigma;
}
```

V časti `data` sme deklarovali všetky dátové premenné vstupujúce do modelu. Taktiež vektor `hz`, ktorý nám predstavuje akýsi prevodník medzi `h`-rozmerným vektorom parametra η (definovaným v nasledujúcej časti) a `n`-rozmerným vektorom prislúchajúcim dátam. Vektor `hz` priradí každému riadku číslo od 1 do `h`, ktoré bude predstavovať nemocnicu prislúchajúcu danému riadku. Okrem toho sme tiež deklarovali strednú hodnotu a varianciu normálneho rozdelenia, ktoré sme uvalili na parametre modelu β, γ a θ ako apriórne rozdelenie. Dôvodom bolo obmedziť potenciálne extrémne hodnoty parametrov. Na tieto parametre sme napokon uvalili apriórne rozdelenie $N(0, 20)$.

```
parameters {
  real beta[k];
  real gamma[d];
  real theta;
```

```
  real eta[h];
  real<lower=0> sig_eta;
}
```

V časti `parameters` sú deklarované parametre modelu, ktorým budeme simulovať výber z ich aposteriórneho rozdelenia. Okrem parametrov stanovených v modeli je deklarovaná aj variancia koeficientu špecifického pre nemocnicu.

```
transformed parameters {
  vector[n] eta_n;
  vector[n] coal_n;
  {
    int j;
    real s;
    for (i in 1:n) {
      eta_n[i] <- eta[hz[i]];

      s <- 0;
      j <- 1;

      while (X[i,j] > 0) {
        s <- s + beta[X[i,j]];
        j <- j + 1;
      }

      coal_n[i] <- s + A[i] * theta + gamma[D[i]];
    }
  }
}
```

Nasleduje časť `transformed parameters`, v ktorej vykonávame úpravy parametrov modelu, hlavne z dôvodu rýchlejšieho behu. Prvou úpravou je vytvorenie parametra `eta_n`, ktorý je len vektorovým rozšírením parametra `eta` tak, aby každá zložka tvorila koeficient špecifický pre nemocnicu prislúchajúci danému riadku. Parameter `coal_n`

predstavuje súčet $A_{j[i]} \cdot \theta + D_{j[i]} \cdot \gamma + X_{j[i]} \cdot \beta$. Prediktor hlavnej diagnózy sme transformovali spôsobom, že každej diagnóze sme priradili číslo od 1 po k, kde k je počet diagnóz v danej skupine. Namiesto maticového násobenia, kedy by sme prediktor hlavnej diagnózy uvažovali ako množinu dummy premenných, vieme takouto transformáciou v zdrojovom kóde priamo adresovať parameter zodpovedajúci danej diagnóze. Čiže ak sme nejakú diagnózu transformovali na napr. `5`, potom `D[i] = 5` a `gamma[D[i]] = gamma[5]`. Pre zrýchlenie behu Stan programu sme tiež upravili výpočty súvisiace s komorbiditami tak, aby program neobsahoval maticové násobenie, pretože v týchto prediktoroch sa nachádza pomerne veľké množstvo núl. Maticu dummy premenných komorbidít sme transformovali na akýsi zoznam s číslami od 1 do 31. Pri výpočte prechádzame tento zoznam a spočítavame parametre `beta[.]` adresované daným číslom (podobne ako pri parametri `gamma`), až kým nenarazíme na koniec zoznamu. Tento postup sa ponáša na veľmi primitívnu implementáciu násobenia sparse matice.

```
model {
  beta ~ normal(prior_mu, prior_sigma);
  gamma ~ normal(prior_mu, prior_sigma);
  theta ~ normal(prior_mu, prior_sigma);
  eta ~ normal(0, sig_eta);
  y ~ bernoulli_logit(NWRR + eta_n + coal_n);
}
```

Špecifikáciu samotného modelu nájdeme v časti `model`. Prvé tri riadky určujú zodpovedajúcim parametrom apriórne rozdelenie. Štvrtý riadok predstavuje druhú úroveň regresného modelu, prislúchajúcu koeficientom špecifickým pre nemocnice a posledný riadok prvú úroveň, prislúchajúcu modelovaniu pravdepodobnosti rehospitalizácie.

```
generated quantities {
  vector[h] srr;
  vector[n] pred;
  vector[h] temppred;
  vector[h] tempexp;
```

```
  for (i in 1:h) {
    temppred[i] <- 0;
    tempexp[i] <- 0;
  }

  for (i in 1:n) {
    pred[i] <- inv_logit(NWRR + coal_n[i] + eta_n[i]);
    temppred[hz[i]] <- temppred[hz[i]] + pred[i];
    tempexp[hz[i]] <- tempexp[hz[i]] + inv_logit(NWRR + coal_n[i]);
  }
  for (i in 1:h) {
    srr[i] <- temppred[i]/tempexp[i];
  }
}
```

Posledná časť `generated quantities` slúži na výpočet koeficientov pri behu programu Stan. Výhodou je, že v tejto časti sa definované hodnoty počítajú len raz za iteráciu, narozdiel od časti `transformed parameters`, v ktorom sú hodnoty prepočítavané v každom kroku metódy leapfrog. Hodnoty v tejto časti využijeme pri hodnotení nemocníc.

3.3.1 Metodika hodnotenia

Našim cieľom pri hodnotení je odhadnúť, ako sa líší predikovaný pomer hospitalizácií danej nemocnice od národného priemeru. Podľa [5], postup funguje na nasledovných princípoch:

- každej nemocnici vypočítame predikovaný a očakávaný počet hospitalizáciím
- predikovaný počet rehospitalizácií pre nemocnicu j označíme $pred_j$ a získame ho ako súčet odhadov pravdepodobností rehospitalizácie po spätnom dosadení dát do modelu s odhadnutými parametrami,
- očakávaný počet rehospitalizácií vypočítame analogicky, akurát nahradíme intercept špecifický pre nemocnicu len národným priemerom rehospitalizácií. To zna-

mená, že odhad pravdepodobností sa pre tieto dva prípady bude líšiť. Očakávaný počet rehospitalizácií potom predstavuje odhadnutý počet rehospitalizácií bez vplyvu nemocnice a predstavuje akúsi mieru chorobnosti pacientov danej nemocnice,

$$pred_j^c = \sum_i \alpha_j + A_{j[i]} \cdot \theta + D_{j[i]} \cdot \gamma + X_{j[i]} \cdot \beta, \tag{17}$$

$$exp_j^c = \sum_i \mu + A_{j[i]} \cdot \theta + D_{j[i]} \cdot \gamma + X_{j[i]} \cdot \beta, \tag{18}$$

- každej nemocnici j v modeli pre každú skupinu potom vypočítame tzv. štandardizovaný pomer rehospitalizácií (z angl. Standardised Readmission Ratio), ktorý označíme SRR_j^c, kde formálne $c = 1, ..., 5$ predstavuje označenie modelu v súvislosti s rozdelením do skupín. Tento podiel je naším odhadom, ako veľmi sa líši predikovaný počet rehospitalizácií danej nemocnice od počtu, očisteného od vplyv nemocnice v danej skupine:

$$SRR_j^c = \frac{pred_j^c}{exp_j^c}, \tag{19}$$

- pre každú nemocnicu j teda dostávame až 5 hodnôt SRR_j^c (pričom niektoré nemocnice môžu mať nula hospitalizácií v niektorých skupinách). Z týchto hodnôt vypočítame logaritmický priemer vážený počtom hospitalizácií v danej skupine. Teda

$$SRR_j = \exp^{\frac{\sum_{c=1}^{5} n_j^c \cdot \log(SRR_j^c)}{\sum_{c=1}^{5} n_j^c}}, \tag{20}$$

- tento priemer nám určuje mieru prebytku rehospitalizácií v danej nemocnici. Prenásobením hodnoty SRR národným priemerom rehospitalizácií dostávame tzv. rizikovo vážený pomer rehospitalizácií pre danú nemocnicu, označený $RSRR$, $RSRR_j = \mu \cdot SRR_j$. $RSRR$ predstavuje zároveň skóre, na základe ktorého budeme nemocnice hodnotiť.

Odhady $RSRR$ sa v [5] vypočítajú pomocou metódy bootstrap. Najprv sa zo všetkých nemocníc vyberie rovnaký počet pomocou výberu s opakovaním. Takto môžu byť niektoré nemocnice vybrané viackrát, ale do modelu vstúpia formálne ako rôzne.

Na týchto dátach sa odhadnú parametre viacúrovňovej logistickej regresie, a pre jednotlivé nemocnice sa vypočítajú odhady $RSRR$. Po určitom počte iterácií máme pre každú nemocnicu dostatočné množstvo odhadov na konštrukciu intervalového odhadu $RSRR$.

Keď sme sa tento postup replikovali, a to odhadovaním parametrov v prostredí jazyka `R` pomocou knižnice `lme4` [1], narazili sme sa prekážku v podobe pridlhého behu programu. Odhadnutie parametrov jedného modelu trvalo z dôvodu veľkého počtu dát niekoľko hodín, a preto by použitie metódy bootstrap bolo v našich podmienkach nerealistické.

Preto sme sa rozhodli zmeniť prístup a na odhadovanie parametrov použiť bayesovskú inferenciu. Pomocou algoritmu markovovského Monte Carlo simulovania dokážeme získať odhad rozdelenia jednotlivých parametrov a vďaka nim intervalový odhad hodnoty $RSRR$. Pre tento účel sme použili program Stan v prostredí jazyka `R`, predstavený v kapitole 1.3.

V práci sme pre každú skupinu simulovali 3 nezávislé markovovské reťazce s 2000 iteráciami, pričom zahrievací čas tvorilo 1000 iterácií. $RSRR$ sme vypočítali v každej iterácii mimo zahrievacieho času a údaje sme spojili naprieč reťazcami, čím sme dostali 3000 realizácií $RSRR$. Tieto realizácie tvoria simulovaný výber hodnôt $RSRR$ a pomocou neho potom nemocnice kategorizujeme do 5 skupín. Z realizácií $RSRR$ najprv vytvoríme 50 a 95%-né intervaly vierohodnosti, ktorých hranice označíme $(l_{50\%}, u_{50\%})$ a $(l_{95\%}, u_{95\%})$. Potom na základe toho, či sa v týchto intervaloch nachádza národný priemer rehospitalizácií μ zaradíme danú nemocnicu do skupiny:

- výrazne horšia ako národný priemer, ak $\mu < l_{95\%}$ a $\mu < l_{50\%}$
- horšia ako národný priemer, ak $\mu \in (l_{95\%}, u_{95\%})$ a $\mu < l_{50\%}$
- nelíšiaca sa od národného priemeru, ak $\mu \in (l_{95\%}, u_{95\%})$ a $\mu \in (l_{50\%}, u_{50\%})$
- lepšia ako národný priemer, ak $\mu \in (l_{95\%}, u_{95\%})$ a $u_{50\%} < \mu$
- výrazne lepšia ako národný priemer, ak $u_{95\%} < \mu$ a $u_{50\%} < \mu$

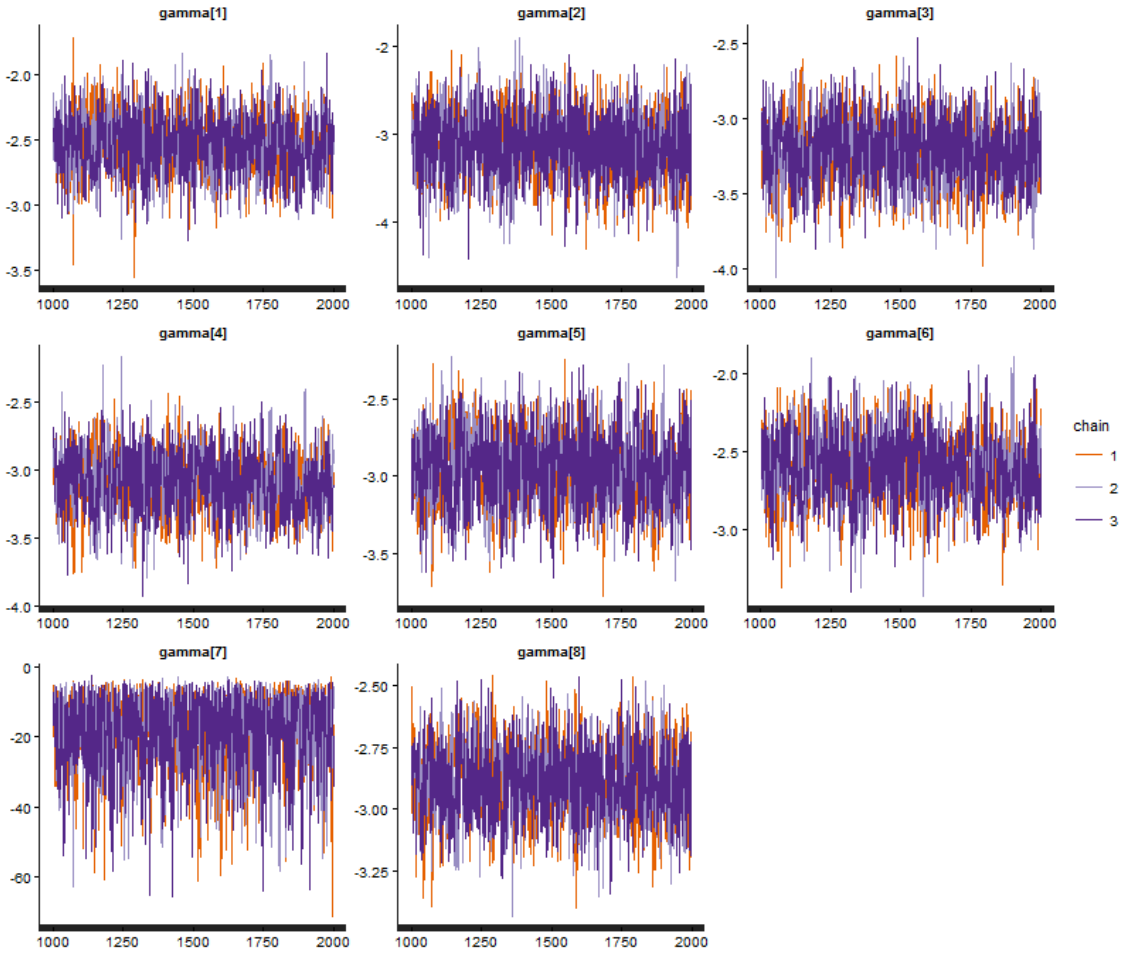
4 Diagnostika konvergencie, výsledné hodnotenia

V tejto kapitole opíšeme kontrolu konvergencie modelov a predstavíme anonymizované výsledky.

4.1 Diagnostika konvergencie modelov

Na základe podkapitoly 1.4 sme skontrolovali konvergenciu markovovských reťazcov pre každý parameter vo všetkých modeloch.

Jednoduchá vizuálna analýza traceplotov neodhalila parameter, ktorý by sa odlišoval od traceplotu uvedeného v podkapitole 1.4 ako príklad správnej konvergencie. Parametre sa medzi reťazcami navzájom dostatočne miešali. Doleuvedené sú na ilustráciu traceploty parametra γ z modelu pre kardiorespiračnú skupinu:



Obr. 4: Traceploty parametra γ

Traceploty všetkých parametrov z dôvodu príliš veľkého množstva parametrov neprikladáme k diplomovej práci.

Podobne dopadla analýza na základe parametrov $\hat{R}$. V tabuľke nižšie uvádzame deskriptívne štatistiky $\hat{R}$ pre jednotlivé modely:

	Minimum	Priemer	Maximum
Kardiorespiračná skup.	1.000	1.001	1.026
Kardiovaskulárna skup.	1.000	1.000	1.016
Nervová skup.	1.000	1.000	1.006
Operačná skup.	1.000	1.001	1.010
Medicínska skup.	1.000	1.001	1.009

Tabuľka 1: Štatistiky koeficientov $\hat{R}$ pre jednotlivé modely

Z tabuľky je zrejmé, že pre žiaden parameter koeficient $\hat{R}$ neprekročil hodnotu 1.026. Preto považujeme konvergenciu markovovských reťazcov za dostatočnú.

4.2 Hodnotenie nemocníc za rok 2014

4.2.1 Anonymizácia nemocníc

Predtým, ako uvedieme výsledné hodnotenia nemocníc, vyplývajúce z našej práce, bolo nutné nemocnice zanonymizovať. Anonymizáciu sme vykonali tak, že sme nemocnice rozdelili do 3 skupín podľa celkového počtu hospitalizácií v danom roku.

Z hodnotenia sme najprv vylúčili 12 nemocníc, ktorých celkový počet hospitalizácií v roku 2014 bol menší ako 500. Túto hranicu sme určili po konzultácii s [26]. Dôvodom vylúčenia bol príliš nízky počet hospitalizácií, ktorý nebol dostatočný a tiež spôsoboval signifikantný posun k lepším hodnotám $RSRR$. Po vylúčení nám zostalo 50 nemocníc, ktoré sme rozdelili do podobne veľkých skupín:

- do skupiny malých nemocníc (nižšie označených ako “S”) sme zaradili nemocnice, ktoré mali do 1500 hospitalizácií,
- do skupiny stredne veľkých nemocníc (označených ako “M”) sme zaradili nemocnice, ktoré mali od 1501 do 3000 hospitalizácií,

- do skupiny veľkých nemocníc (označených ako "L") sme zaradili nemocnice, ktoré mali nad 3001 hospitalizácií za daný rok.

V skupine malých a stredne veľkých nemocníc máme teda po 17 nemocníc a v skupine veľkých 16.

4.2.2 Hodnotenie nemocníc

Nemocnice sú hodnotené na základe 50% a 95%-ných intervalov vierohodnosti rizikovo váženého pomeru rehospitalizácií $RSRR$. Na základe výskytu hodnoty národného priemer rehospitalizácií $\mu = 6.758\%$ potom nemocnice zaradíme do jednej z 5 skupín. Hodnoty $RSRR$ uvádzame v prílohe A. V nasledujúcej tabuľke uvádzame zoskupené hodnotenia podľa veľkosti a kategórie nemocnice:

	Výrazne horšia	Horšia	Nelíši sa	Lepšia	Výrazne lepšia	Spolu
Malé	2	3	5	5	2	17
Stredne veľké	2	7	4	3	1	17
Veľké	1	1	9	3	2	16
Spolu	5	11	18	11	5	50

Tabuľka 2: Zoskupené hodnotenie nemocníc za rok 2014

Pretože výsledné odhady SRR (resp. $RSRR$) sú vážené počtom hospitalizácií, nepozorujeme koncentráciu hodnotení naprieč nemocnicami s výrazne odlišným počtom hospitalizácií za rok. Taktiež, počet výrazne horších a lepších približne zodpovedá skutočnému stavu [26].

4.2.3 Využitie v praxi

Model vytvorený a popísaný v našej práci slúži na identifikáciu nemocníc s priveľkým počtom rehospitalizácií.

Možné využitie modelu vidíme v rôznych incentívach pre nemocnice s dobrým hodnotením. V prípade dobre nastaveného systému ohodnotenia a vhodného manažmentu, je možné nemocnice motivovať k poskytovaní zdravotnej starostlivosti aj po prepustení. Napríklad, program zníženia rehospitalizácií v Spojených štátoch amerických sa tiež za-

sadil o to, ako nemocnice interagujú s pacientmi aj po prepustení. Niektoré nemocnice začali spolupracovať s komunitnými organizáciami, niektoré vytvorili kliniky pre následnú starostlivosť a niektoré začali telefonovať všetkým prepusteným pacientom, aby sa uistili, že ich stav je v poriadku[1]. Takéto opatrenia by mohli zvýšiť úroveň zdravotnej starostlivosti a zabrániť určitému percentu rehospitalizácií.

Naša práca je na druhej strane obmedzená dátovým vykazovaním. Posun vykazovania k systému DRG by mohol výrazne napomôcť dátovej kvalite, pretože sa zjednoduší identifikácia hospitalizácií. Systém DRG (z angl. "diagnosis-related group") je klasifikačný systém hospitalizačných prípadov. V systéme DRG sa hospitalizačné prípady sa zaradia do DRG skupín na základe charakteristík pacienta a jeho hospitalizácie (veku a pohlavia pacienta, hlavnej a vedľajších diagnóz, poskytnutých zdravotných výkonov)[2]. Spracovanie dát je v súčasnosti obmedzené prehľadávaním fragmentovaných dát. Preto by úprava modelu na takéto vykazovanie priniesla nižšiu chybovosť a vyššiu spoľahlivosť.

[1]http://www.informationweek.com/regulations/readmissions-dont-measure-quality-harvard-doctor-says/d/d-id/1108727?

[2]http://www.hpi.sk/2015/04/pribeh-drg-1-co-to-je-a-na-co-to-sluzi/

Záver

V práci sme opísali metodológiu vývoja modelu pre hodnotenie nemocníc na základe rehospitalizácií. Táto metodológia vychádza z metodológie používanej v Spojených štátoch amerických pre hodnotenie nemocníc na federálnej úrovni.

Kapitola 1 obsahuje zhrnutie poznatkov, ktoré sme využili pri písaní našej práce a zahŕňa okrem štatistických metód aj teoretické základy algoritmu hamiltonského Monte Carlo simulovania a popis využitého technického riešenia.

V kapitole 2 sme sa venovali predstaveniu dátových zdrojov a vykonaných úprav. Najväčšími prekážkami pre nás tvorilo odlišné vykazovanie dát, ktoré si vyžiadalo veľa vlastných riešení pri spracovaní dát. V súvislosti s dátami bolo potrebné z vysoko fragmentovaných dát identifikovať samotné hospitalizácie, proces, ktorý opisujeme v kapitole 2.3. Ďalšou výzvou bolo lokalizovať algoritmus pre identifikáciu neplánovaných rehospitalizácií, ktorého úpravu sme predstavili v kapitole 2.6. Použitie prediktívneho modelu v metodológii si tiež pre nás vyžiadalo nový spôsob odhadovania parametrov v modeli. Využitie bayesovskej štatistickej inferencie bolo podmienené technickými limitáciami a výpočtovou kapacitou.

Kapitola 3 obsahuje podrobný popis regresného model pre odhadovanie pravdepodobnosti rehospitalizácie, technické riešenie v programe Stan a metodiku pre hodnotenie nemocníc. Nemocnice boli hodnotené na základe miery prebytku rehospitalizácií vyplývajúceho z prediktívneho modelu.

V kapitole 4 sme napokon uviedli základnú diagnostiku použitého modelu a samotné hodnotenie nemocníc. Spomenuli sme tiež príklad využitia a systém DRG, ktorého nasadenie by mohlo mať pozitívny vplyv na dátovú kvalitu v našej práci.

Zoznam použitej literatúry

[1] Bates, D., Maechler, M., Bolker, B., Walker, S.: *Fitting Linear Mixed-Effects Models Using lme4*. Journal of Statistical Software 67 (2015), 1-48.

[2] Brooks, S. P., Gelman, A.: *General Methods for Monitoring Convergence of Iterative Simulations*. Journal of Computational and Graphical Statistics 7(4) (1998), 434–455

[3] Carpenter, B., Gelman, A., Hoffman, M., et al.: *Stan: A probabilistic programming language*. Journal of Statistical Software (in press, 2016).

[4] Centers for Medicare & Medicaid Services: Home, dostupné na internete (25.4.2016): https://www.cms.gov/

[5] Centers for Medicare & Medicaid Services, *Hospital-Wide All-Cause Readmission Measure-Version 4.0: 2015 Measures Updates and Specifications Report*, dostupné na internete (28.2.2016): https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1219069855841

[6] Clark, M.: *Bayesian Basics: A Conceptual Introduction with Application in R and Stan*. Center for Statistical Consultation and Research, University of Michigan, 2015, dostupné na internete (28.2.2016): http://www3.nd.edu/~mclark19/learn/IntroBayes.pdf

[7] Clinical Classifications Software (CCS) for ICD-10-CM/PCS, dostupné na internete (28.2.2016): https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp

[8] Cortada, J. W., Gordon, D., Lenihan, B.: *The Value of Analytics in Healthcare*, dostupné na internete (24.4.2016): https://www.google.sk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiNkbrtuKfMAhWKVywKHW59D1cQFggvMAA&url=https%3A%2F%2Fwww.ibm.com%2Fsmarterplanet%2Fglobal%2Ffiles%2Fthe_value_of_analytics_in_healthcare.pdf&usg=AFQjCNFEvxMLTS69F5z1PMFMhkPXvuSRNg&sig2=ODmURChqHhGN7t-nn7t87g

[9] Danková, A., MUDr.: osobná komunikácia, Dôvera, generálne riaditeľstvo, Bratislava, 2015.

[10] Eddelbuettel, D., Francois, R.: *Rcpp: Seamless R and C++ Integration*. Journal of Statistical Software, 40(8) (2011), 1-18.

[11] Gelman, A., Hill, J.: *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New York, 1st printing, 2007.

[12] Gelman, A., Lee, D., Guo, J.: *Stan: A probabilistic programming language for Bayesian inference and optimization*, dostupné na internete (20.2.2016): http://www.stat.columbia.edu/~gelman/research/published/stan_jebs_2.pdf

[13] Goldstein, H.: *Multilevel Statistical Models*, John Wiley and Sons, West Sussex, štvrté vydanie, 2011.

[14] Grendár, M.: *Bayesovská štatistika*. Forum Statisticum Slovacum 3 (2009), 22-40.

[15] Hoffman, M. D., Gelman, A.: *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. Journal of Machine Learning Research 15 (2014), 1351-1381

[16] Horwitz, L. I., Partovian, C., Lin, Z., et al.: *Development and Use of an Administrative Claims Measure for Profiling Hospital-wide Performance on 30-Day Unplanned Readmission*. Annals of Internal Medicine 161 (2014), 66-75.

[17] HospiCOM a hlásenie hospitalizácií, dostupné na internete (28.2.2016): http://www.dovera.sk/lekar/hospicom

[18] Medicare's Hospital Readmission Reduction Program FAQ, dostupné na internete (24.4.2016): https://www.acep.org/Physician-Resources/Practice-Resources/Administration/Financial-Issues-/-Reimbursement/Medicare-s-Hospital-Readmission-Reduction-Program-FAQ/

[19] Medzinárodná klasifikácia chorôb - MKCh, dostupné na internete (28.2.2016): http://www.nczisk.sk/Standardy-v-zdravotnictve/Pages/Medzinarodna-klasifikacia-chorob-MKCH-10.aspx

[20] Neal, R. M.: *MCMC Using Hamiltonian Dynamics*, Handbook of Markov Chain Monte Carlo, Chapman and Hall/CRC, 2011.

[21] Poliak, M., Mgr.: osobná komunikácia, Dôvera, generálne riaditeľstvo, Bratislava, 2015.

[22] R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, dostupné na internete (24.4.2016): https://www.R-project.org/

[23] R: Fitting Generalized Linear Models, dostupné na internete (24.4.2016): https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html http://www.informationweek.com/regulations/readmissions-dont-measure-quality-harvard-doctor-says/d/d-id/1108727?

[24] Stan - Interfaces, dostupné na internete (28.2.2016): http://mc-stan.org/interfaces/

[25] Stan Development Team: *Modeling Language User's Guide and Reference Manual*.

[26] Tulejová, H., MSc., Mgr.: osobná komunikácia, Dôvera, generálne riaditeľstvo, Bratislava, 2015.

[27] *Zákon č. 122/2013 Z. z. o ochrane osobných údajov a o zmene a doplnení niektorých zákonov, ako vyplýva zo zmien a doplnení vykonaných zákonom č. 84/2014 Z. z.*

Prílohy

A. Hodnoty RSRR

Malé nemocnice

Veľkosť	$l_{95\%}$	$l_{50\%}$	Priemer	$u_{50\%}$	$u_{95\%}$	Hodnotenie
S	4,324%	5,028%	5,397%	5,757%	6,472%	Výrazne lepšia
S	4,940%	5,416%	5,707%	5,984%	6,532%	Výrazne lepšia
S	4,694%	5,526%	6,026%	6,495%	7,548%	Lepšia
S	4,955%	5,576%	5,939%	6,292%	7,004%	Lepšia
S	5,142%	5,789%	6,161%	6,517%	7,266%	Lepšia
S	5,214%	5,926%	6,322%	6,682%	7,509%	Lepšia
S	5,275%	5,866%	6,234%	6,578%	7,290%	Lepšia
S	5,487%	6,260%	6,667%	7,061%	7,907%	Nelíši sa
S	5,641%	6,458%	6,950%	7,400%	8,385%	Nelíši sa
S	5,790%	6,421%	6,751%	7,087%	7,752%	Nelíši sa
S	5,796%	6,457%	6,819%	7,177%	7,916%	Nelíši sa
S	5,849%	6,462%	6,831%	7,177%	7,963%	Nelíši sa
S	6,112%	6,783%	7,184%	7,556%	8,342%	Horšia
S	6,224%	7,137%	7,699%	8,221%	9,425%	Horšia
S	6,275%	7,067%	7,515%	7,945%	8,883%	Horšia
S	7,109%	7,796%	8,224%	8,620%	9,459%	Výrazne horšia
S	7,330%	7,985%	8,364%	8,721%	9,523%	Výrazne horšia

Tabuľka 3: Hodnotenie malých nemocníc za rok 2014

Stredne veľké nemocnice

Veľkosť	$l_{95\%}$	$l_{50\%}$	Priemer	$u_{50\%}$	$u_{95\%}$	Hodnotenie
M	5,162%	5,610%	5,867%	6,122%	6,625%	Výrazne lepšia
M	4,930%	5,532%	5,862%	6,189%	6,855%	Lepšia
M	5,379%	5,899%	6,200%	6,498%	7,071%	Lepšia
M	5,544%	5,999%	6,268%	6,529%	7,037%	Lepšia
M	5,780%	6,321%	6,630%	6,936%	7,543%	Nelíši sa
M	6,057%	6,669%	7,011%	7,341%	7,997%	Nelíši sa
M	6,067%	6,627%	6,919%	7,212%	7,799%	Nelíši sa
M	6,182%	6,737%	7,039%	7,321%	7,991%	Nelíši sa
M	6,209%	6,805%	7,175%	7,526%	8,246%	Horšia
M	6,285%	6,824%	7,135%	7,426%	8,058%	Horšia
M	6,286%	6,893%	7,210%	7,520%	8,215%	Horšia
M	6,378%	6,943%	7,280%	7,609%	8,278%	Horšia
M	6,658%	7,219%	7,561%	7,883%	8,548%	Horšia
M	6,695%	7,335%	7,673%	8,016%	8,701%	Horšia
M	6,711%	7,270%	7,625%	7,959%	8,696%	Horšia
M	6,801%	7,387%	7,690%	7,990%	8,602%	Výrazne horšia
M	6,967%	7,638%	8,028%	8,393%	9,126%	Výrazne horšia

Tabuľka 4: Hodnotenie stredne veľkých nemocníc za rok 2014

Veľké nemocnice

Veľkosť	$l_{95\%}$	$l_{50\%}$	Priemer	$u_{50\%}$	$u_{95\%}$	Hodnotenie
L	5,025%	5,450%	5,671%	5,894%	6,333%	Výrazne lepšia
L	5,243%	5,677%	5,936%	6,175%	6,699%	Výrazne lepšia
L	5,440%	5,847%	6,089%	6,324%	6,789%	Lepšia
L	5,587%	6,057%	6,343%	6,609%	7,170%	Lepšia
L	5,836%	6,228%	6,442%	6,652%	7,079%	Lepšia
L	5,804%	6,275%	6,538%	6,802%	7,305%	Nelíši sa
L	5,952%	6,357%	6,588%	6,810%	7,306%	Nelíši sa
L	5,983%	6,392%	6,624%	6,856%	7,291%	Nelíši sa
L	6,108%	6,480%	6,672%	6,864%	7,241%	Nelíši sa
L	6,113%	6,585%	6,848%	7,103%	7,628%	Nelíši sa
L	6,224%	6,688%	6,961%	7,211%	7,749%	Nelíši sa
L	6,236%	6,616%	6,836%	7,049%	7,475%	Nelíši sa
L	6,330%	6,754%	6,991%	7,219%	7,701%	Nelíši sa
L	6,369%	6,675%	6,855%	7,029%	7,371%	Nelíši sa
L	6,493%	6,900%	7,113%	7,322%	7,745%	Horšia
L	6,851%	7,237%	7,439%	7,633%	8,073%	Výrazne horšia

Tabuľka 5: Hodnotenie veľkých nemocníc za rok 2014