

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



APLIKÁCIA METÓD DATA MININGU NA
ZOSKUPOVANIE EPIZÓD ZDRAVOTNEJ
STAROSTLIVOSTI

DIPLOMOVÁ PRÁCA

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

**APLIKÁCIA METÓD DATA MININGU NA
ZOSKUPOVANIE EPIZÓD ZDRAVOTNEJ
STAROSTLIVOSTI**

DIPLOMOVÁ PRÁCA

Študijný program: Ekonomická a finančná matematika
Študijný odbor: 9.1.9. Aplikovaná matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Vedúci práce: Mgr. Henrieta Tulejová, MSc.



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Matej Poliak
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov: Aplikácia metód data miningu na zoskupovanie epizód zdravotnej starostlivosti / *Application of data mining methods to health care episode grouping*

Cieľ: Vybrať a popísať metódy data miningu a použiť ich na zoskupenie zdravotníckych výkonov k vybranej epizóde zdravotnej starostlivosti. Stručne ukázať použitie dosiahnutých výsledkov pri ďalších analýzach.

Vedúci: Mgr. Henrieta Tulejová, MSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.
Dátum zadania: 25.01.2013

Dátum schválenia: 04.02.2013
prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

PodĎakovanie Touto cestou sa chcem poĎakovať vedúcej mojej diplomovej práce Mgr. Henriete Tulejovej, MSc. za ochotu, odborné rady a podnetné pripomienky, ktoré boli veľmi cenné pri písaní tejto práce. VĎaka patrí spoločnosti Dôvera zdravotná poisťovňa, a. s. za poskytnutie dát, ktoré boli základom k vypracovaniu práce. Ďalej by som sa chcel poĎakovať JUDr. MUDr. Zuzane Hromjakovej za veľmi prínosný medicínsky pohľad na niektoré otázky. V neposlednej rade patrí poĎakovanie mojej rodine a priateľom za neustálu podporu.

Abstrakt v štátnom jazyku

Poliak, Matej: Aplikácia metód data miningu na zoskupovanie epizód zdravotnej starostlivosti [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; vedúci: Mgr. Henrieta Tulejová, MSc., Bratislava, 2014, 78 s.

V práci sa venujeme zoskupovaniu epizód zdravotnej starostlivosti na základe pohľadu na zdravotnícke administratívne dáta. Využívame pritom predovšetkým data miningovú metódu, analýzu asociačných pravidiel. V práci najprv uvádzame základné pojmy analýzy asociačných pravidiel obohatené o niekoľko nami neskôr používaných tvrdení. Nasleduje úvod do problematiky epizód zdravotnej starostlivosti z pohľadu, akým na ne nazeráme v práci. Popíšeme bodovací systém ktorý využíva niekoľko množín asociačných pravidiel na ohodnotenie zdravotníckych výkonov podľa sily ich príslušnosti k skúmanej epizóde, implantácii bedrového kĺbu. Výsledky verifikujeme porovnaním so vstupom od lekára. Nakoniec uvádzame niekoľko praktických príkladov využitia zoskupených epizód zdravotnej starostlivosti.

Kľúčové slová: Epizódy zdravotnej starostlivosti, Analýza asociačných pravidiel, Data mining

Abstract

Poliak, Matej: Application of Data Mining Methods to Healthcare Episode Grouping [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Henrieta Tulejová, MSc., Bratislava, 2014, 78 p.

In our work we deal with health care episode grouping based on claims data using association rule analysis, a data mining method. First, we introduce the basic notation of association rule analysis together with a few theorems which are being used through the later parts of the work. A description of our view on the health care episode problematics follows. We then describe a scoring system that uses several sets of association rules to rank patients' claims based on the perceived strength of their relationship to the episode of interest, hip joint replacement surgery. We validate the results by comparing them to an input provided by a physician. Lastly, we present a few examples of using grouped health episodes in practice.

Keywords: Health Care Episodes, Association Rule Analysis, Data Mining

Obsah

Zoznam obrázkov	9
Zoznam tabuliek	11
Zoznam použitých symbolov	13
Úvod	14
1 Analýza asociačných pravidiel	16
1.1 Základné definície	17
1.2 Príklady použitia pravidiel	18
1.3 Meranie zaujímavosti pravidiel	21
1.3.1 Opak pravidla	21
1.3.2 Ďalšie tvrdenia o miere lift	24
1.3.3 Nezávislosť položiek	26
1.4 Hľadanie asociačných pravidiel	27
2 Dáta, epizódy a asoc. pravidlá	28
2.1 Dáta zdravotnej poisťovne	28
2.2 Zoskupovanie epizód	30
2.3 Dátový prístup	33
2.4 Asociačná analýza	34
2.4.1 Úloha asociačnej analýzy o produktoch	35
2.4.2 Využitie viacerých množín asociačných pravidiel	38
3 Zoskupovanie epizód starostlivosti	40
3.1 Dáta	40
3.1.1 Výber dát	40
3.1.2 Anonymizácia dát	42
3.1.3 Základný popis dát	43
3.1.4 Čistenie dát	46
3.1.5 Zoskupovanie hospitalizácií	49
3.1.6 Zoskupovanie ambulantných kontaktov	51

3.1.7	Ďalšie informácie pre zoskupené hospitalizácie a kontakty	51
3.2	Asociačné pravidlá	52
3.2.1	Asociačné pravidlá pre diagnózy	53
3.2.2	Asociačné pravidlá pre odbornosti	56
3.2.3	Asociačné pravidlá pre produkty a ich skupiny	57
3.2.4	Asociačné pravidlá pre kombináciu diagnózy, odbornosti a produktu	59
3.2.5	Zhrnutie a vyhodnotenie pridelenia bodov	60
3.3	Verifikácia výsledkov	62
3.4	Výsledky pre inú epizódu	65
4	Ďalšie použitie výsledkov	67
4.1	Základná analýza epizódy	67
4.2	Závažnosť epizódy	70
4.3	Analýza poskytovateľov starostlivosti	71
4.3.1	Identifikácia extrémne drahých poskytovateľov	71
4.3.2	Vytváranie zhlukov poskytovateľov	72
	Záver	75
	Zoznam použitej literatúry	77

Zoznam obrázkov

1	Databáza transakcií pre príklad 1.3.	20
2	Príklad pre popisované množiny položiek: E , $\neg E$, $P_V \Rightarrow E$, $E \Rightarrow P_K$, $E_{K1} \Rightarrow E$, E_{K2} nezávislé s E , $E \Rightarrow E_{K3}$ pričom $E_{K2}, E_{K3}, E_{K4} \not\Rightarrow E$ a $E_{K1}, E_{K2}, E_{K3}, E_{K4} \Rightarrow P_K$	36
3	Histogram mier lift v nájdených pravidlách pre diagnózy.	54
4	Histogram mier lift v nájdených pravidlách pre kombináciu diagnóza- odbornosť-produkt.	59
5	Počet kontaktov na pacienta, resp. náklady na pacienta v kontaktoch podľa minimálneho počtu bodov ich pridelenia k epizóde.	61
6	Histogramy podielu nákladov pacientov, ktoré priradíme k epizódam pri hranici bodov 4 pre pacientov s epizódou, resp. pacientov referenčnej skupiny.	62
7	Priradené body za diagnózy, odbornosti a produkty vs. správna klasifi- kácia ako kombinácia súvisiaca (zelený bod), resp. nesúvisiaca (červený bod) s epizódou.	63
8	Rozhodovací strom.	64
9	Počet kontaktov na pacienta, resp. náklady na pacienta v kontaktoch podľa minimálneho počtu bodov ich pridelenia k epizóde implantácie kolenného kĺbu.	65
10	Histogramy podielu nákladov pacientov pre epizódu implantácie kolen- ného kĺbu, podiel nákladov priradených k epizóde pre pacientov s epi- zódou, resp. referenčných pacientov.	66
11	Priemerné náklady v rôznych vrstvách starostlivosti v roku pred a v roku po hlavnej hospitalizácii (bez hlavnej hospitalizácie).	67
12	Výkony priradené k epizóde jednotlivých pacientov v čase. Čierne bodky predstavujú ambulantné kontakty, červené čiary hlavnú hospitalizáciu, modré čiary iné hospitalizácie a zelené čiary kúpeľné pobyty.	68
13	Histogram počtu kontaktov podľa času od hlavnej hospitalizácie pre všetkých pacientov s epizódou.	69
14	Histogramy pôvodných a logaritmickej nákladov pacientov na epizódu.	70

-
- 15 Histogram priemerných nákladov vzoriek v permutačnom teste škálovaných faktorom 1, 25. Priemer skúmaného poskytovateľa je vyznačený bodom. 73
- 16 Dendrogram pre analýzu zhlukov. Čím nižšie sú dva objekty spojené, tým väčšia je ich podobnosť. Prerušovaná čiara predstavuje zvolenú úroveň „zostrihania” stromu. 74
- 17 Náklady na poskytovateľov po vrstvách a zoskupenie do zhlukov. . . . 74

Zoznam tabuliek

1	Zhrnutie anonymizácie. Hrubé písmená predstavujú skutočné písmená, c vyjadruje číslicu, p písmeno, ooo trojmiestny kód odbornosti a skupina hviezdíček predstavuje náhodne generované číslo.	42
2	Zoznam premenných v obidvoch hlavných používaných tabuľkách dát. .	44
3	Výsek dát zobrazujúci typický výskyt odporúčajúceho lekára.	45
4	Štruktúra tabuľky <i>Data_a</i>	47
5	Štruktúra tabuľky <i>Data_h</i>	47
6	Konverzia niektorých najdôležitejších dátumov na používaný formát. .	47
7	Príklad pacienta s chýbajúcimi obidvomi dátumami pri hospitalizácii. Piaty riadok je očividne chybou po hospitalizácii z prvého riadku a môžeme ho vylúčiť.	48
8	Príklad spájania záznamov hospitalizácií, pred a po spojení.	49
9	Príklad dvoch zoskupených hospitalizácií pozostávajúcich z viacerých čiastkových. V prvej ($Grup = 132$) nasledujú čiastkové hospitalizácie po sebe, v druhej ($Grup = 1423$) sa vyskytlo zduplikovanie jednej čiastkovej hospitalizácie.	50
10	Zhrnutie počtu čiastkových hospitalizácií v zoskupenej hospitalizácii. .	50
11	Najčastejšie sa vyskytujúce diagnózy s mierou lift nad 2,5 pre pravidlá Diagnózy I. Kódy znamenajú: M16 — artróza bedrového kĺbu, Z96 — prítomnosť iných funkčných implantátov, T84 — komplikácie vnútorných ortopedických protetických pomôcok, implantátov a štepov, S73 — vyklíbenie, vyvrtnutie a natiahnutie kĺbu a väzov bedier, M18 — artróza prvého karpometakarpálneho kĺbu, S71 — otvorená rana bedier a stehna.	54
12	Časť nájdených pravidiel Diagnózy II. Názov položky obsahuje „p**“, kde hviezdíčky predstavujú hranicu príslušného pásma nákladov v percentách, o ktorom pravidlo hovorí.	55

13	Diagnózy vyskytujúce sa v zoskupených hospitalizáciách na základe ktorých boli pacienti priradení k epizóde. Najčastejšie sa vyskytujúce sú M16 — artróza bedrového kĺbu, Z96 — prítomnosť iných funkčných implantátov a M17 — artróza kolenného kĺbu.	56
14	Najčastejšie sa vyskytujúce odbornosti s mierou lift nad 1, 5 pre pravidlá Odbornosti I. Kódy znamenajú: 27 — fyziatria, balneológia a liečebná rehabilitácia, 11 — ortopédia, 183 — doprava pacientov, 25 — anestéziológia a intenzívna medicína, 218 — nemocničná krvná banka, 45 — reumatológia.	57
15	Skupiny produktov s mierou lift nad 2, 5. Kódy znamenajú: MX a MC — špeciálny zdravotnícky materiál, ZK — rehabilitačné a kompenzačné zdravotnícke pomôcky, RT — transfúzne lieky, HK — kúpeľná liečba, RB — tkanivá a bunky pri transplantáciách.	58
16	Príklad nájdených pravidiel z kombinácií diagnóz, odborností a produktov. Vyskytujú sa diagnózy M16 (artróza bedrového kĺbu) a I10 (esenciálna hypertenzia), odbornosti 20 (všeobecné lekárstvo), 11 (ortopédia), 1 (vnútorné lekárstvo), 10 (chirurgia) a 49 (kardiológia). Kódy C07AB07 a C10AA05 predstavujú ATC skupiny liekov, zvyšné kódy sú podskupiny zdravotnej starostlivosti.	60
17	Zhrnutie pridelovania bodov na základe asociačných pravidiel.	60
18	Porovnanie pôvodnej a správnej klasifikácie; počty kombinácií zo skráteného zoznamu, resp. náklady na príslušné kombinácie.	63
19	Výstup lineárnej regresie pre jednoduchý model závažnosti epizódy. . .	71
20	Poskytovatelia s najväčšími a najmenšími priemernými nákladmi na epizódu.	72

Zoznam použitých symbolov

supp	support, miera výskytu asociačného pravidla
supp _{ABS}	absolútny support, miera výskytu asociačného pravidla
conf	confidence, miera sily asociačného pravidla
lift	lift, miera zaujímavosti asociačného pravidla
ZS	zdravotná starostlivosť
VZS	vykázaná zdravotná starostlivosť, resp. dáta VZS
Data_a	tabuľka dát mimo hospitalizácií
Data_h	tabuľka dát hospitalizácií

Úvod

Epizódy zdravotnej starostlivosti sú pomerne novým konceptom, ktorý sa vo svete postupne začína používať pri vyhodnocovaní efektivity poskytovania zdravotnej starostlivosti alebo ako základná jednotka pri platbách za zdravotnú starostlivosť. Epizódou zdravotnej starostlivosti rozumieme množinu zdravotníckych výkonov, ktoré boli jednému pacientovi vykonané v súvislosti s konkrétnou udalosťou v jeho zdravotnom stave, teda množinu výkonov ktorú spája jeden problém. Môže pritom ísť napríklad o zdravotnú starostlivosť týkajúcu sa akútneho stavu od prvotnej hospitalizácie cez kontrolné vyšetrenia a rehabilitácie až po pacientovo vyliečenie sa. Epizóda sa tiež môže týkať pacientovej chronickej choroby, zahŕňať lieky a bežné návštevy u špecializovaného lekára, ale aj komplikácie a prípadné hospitalizácie súvisiace s touto chorobou. Trvanie epizódy môže byť od určeného počtu dní pre akútne stavy u ktorých sa očakáva ústup potreby ďalšej starostlivosti po istom čase cez dopredu neučené trvanie, až prakticky nekonečné pre chronické choroby. Koncept epizód zdravotnej starostlivosti je zaujímavý tým, že vo všeobecnosti zahŕňa starostlivosť poskytovanú viacerými poskytovateľmi zdravotnej starostlivosti (nemocnicami, ambulanciami, atď.), aj keď často sa určuje jeden poskytovateľ, ktorému sa pripisuje zodpovednosť za priebeh epizódy.

Podľa [9] je napriek veľkému záujmu zo strany nákupcov zdravotnej starostlivosti (zdravotných poisťovní a analogických inštitúcií v zahraničí) skutočná implementácia často v začiatkoch či experimentálnej fáze. Čo sa týka hodnotenia poskytovateľov zdravotnej starostlivosti, epizódy už sú používané viacerými zahraničnými nákupcami starostlivosti, predovšetkým v USA a ďalší plánujú využitie do budúcnosti. Ako príklad sa dajú uviesť inštitúcie za vládny programami Medicare a Medicaid zdravotného poistenia v USA [4]. Projekty uvažujúce platbu poskytovateľom starostlivosti na základe epizód sú v testovacej fáze a väčšinou v menšom rozsahu, špecializované na malý počet konkrétnych ochorení.

Na zostavovanie epizód starostlivosti existuje niekoľko tzv. episode grouperov. Zoskupovaním epizód starostlivosti rozumieme priradenie jednotlivých pacientových zdravotníckych výkonov k rôznym epizodám — ide teda o činnosť vyžadujúcu znalosť medicínskych faktorov ohľadne epizód starostlivosti, ako aj znalosť dát ktoré sú o pacientoch k dispozícii. Pri grouperoch sa spravidla jedná o komerčné programy s obmedzenou

dostupnou dokumentáciou. Taktiež ďalšia literatúra ohľadne epizód pozostáva predovšetkým z článkov ako [9], ktoré prinášajú isté odporúčania a témy na zváženie ohľadne tvorby epizód, konkrétnejšie informácie, na základe ktorých by sa dalo epizódy zoskupovať v praxi, ale neobsahujú. Komerčné zoskupovače epizód taktiež pracujú s dátami aké sa zbierajú ohľadne zdravotnej starostlivosti v zahraničí, tie však môžu byť rozdielne od dát, aké sa v súčasnosti zbierajú v slovenskom prostredí. Cieľom práce je preto zostaviť a verifikovať systém na zoskupovanie epizód zdravotnej starostlivosti v slovenských podmienkach. Na rozdiel od zahraničných komerčných grouperov, ktoré sú postavené predovšetkým na základe lekárskeho znalostí o poskytovanej starostlivosti, náš spôsob zoskupovania epizódy čerpá informácie v zásade výlučne z dát. To znamená, že pravidlá, na základe ktorých sú priradované jednotlivé výkony u konkrétného pacienta k jednotlivým epizódam starostlivosti, získavame vhodnou metódou zo zdravotníckych dát o istej množine pacientov, pričom našim vstupom je výber metód a postupov, prípadne nastavenie niektorých parametrov.

V práci najprv rozoberáme metódu data miningu, analýzu asociačných pravidiel. Po všeobecnom úvode postavenom na základných článkoch týkajúcich sa problematiky pridávame konkrétnejšie poznatky ktoré priamo slúžia potrebám práce. Ďalšia kapitola sa zaoberá širším popisom problému ktorý riešime, prostriedkov ktoré máme k dispozícii a úvah, ako k riešeniu problému pristupujeme. Popisujeme teda dáta ktoré máme k dispozícii, bližšie rozoberáme epizódy starostlivosti a uvažujeme nad možnosťami použitia analýzy asociačných pravidiel. V tretej kapitole podrobne popisujeme konkrétny postup použitý pri zoskupení epizód starostlivosti zameranom na konkrétnu dvojicu epizód — implantáciu umelého bedrového, resp. kolenného kĺbu. V poslednej kapitole stručne ukazujeme niekoľko možností praktického použitia dosiahnutých výsledkov.

1 Analýza asociačných pravidiel

Analýza asociačných pravidiel je metóda používaná na hľadanie vzťahov v databázach [5]. Za prelomový článok v problematike sa považuje [1]. V tomto článku z roku 1993 autori popisujú formuláciu problému a uvádzajú tiež algoritmus na jeho praktické riešenie. V literatúre sa hlavné pojmy v problematike často definujú formálne mierne odlišnými spôsobmi, ktoré sú však pre praktické použitie ekvivalentné. My preberáme pojmy predovšetkým z [5] a predstavujeme ich v najbližších odsekoch.

Majme množinu *položiek*, $I = \{i_1, i_2, \dots, i_m\}$ a množinu, resp. databázu *transakcií*, $\mathcal{D} = \{t_1, t_2, \dots, t_n\}$. Transakcie môžu obsahovať údaje ako svoju identifikáciu, čas realizácie, atď. podľa toho, čo predstavujú. Podstatná súčasť transakcie t_i je však množina $T_i \subseteq I$, pričom množinu všetkých T_i označujeme $D = \{T_i, i \in 1, 2, \dots, n\}$ a často pojmom databáza transakcií prípadne transakcia pre skrátenie označujeme D , resp. T_i v prípadoch, keď prípadné zvyšné informácie o transakciách nie sú podstatné. Pre niektoré ďalšie úvahy je potrebné, aby každá položka bola obsiahnutá aspoň v jednej transakcii, prípadne aby žiadna položka nebola obsiahnutá vo všetkých transakciách. Takéto položky tak či tak nedávajú veľa dodatočnej informácie, preto predpokladáme databázu, ktorá položky týchto typov neobsahuje.

Pravidlom rozumieme tvrdenie zapísané formou $X \Rightarrow Y$, pričom platí $X, Y \subseteq I$. Množiny, ktoré sú podmnožinami I vo všeobecnosti nazývame množiny položiek. V praxi sa často hľadajú pravidlá pre ktoré platí $X \cap Y = \emptyset$, z pohľadu teórie to ale nie je nutný predpoklad. V praxi sa tiež často hľadajú iba pravidlá s $|Y| = 1$, teda uvažovaná množina Y má práve jeden prvok. Množina X zas môže byť prázdna, môže mať jeden, alebo aj viac prvkov. Množinu X voláme ľavá strana, množinu Y pravá strana pravidla. Nás zaujímajú pravidlá spĺňajúce isté kritériá, napr. *silné* a *významné* pravidlá (presná definícia nasleduje neskôr), pričom silné aj významné pravidlo $X \Rightarrow Y$ znamená „v transakciách ktoré obsahujú X , teda všetky položky z X sa často nachádza aj položka Y “. Pravidlo spĺňajúce vopred dané podmienky na svoje parametre nazývame *asociačným pravidlom*.

1.1 Základné definície

Aby sme mohli definovať pojmy významnosti a sily pravidla, zavádzame niekoľko definícií.

Definícia 1.1. *Pojmom (relatívny) support množiny položiek $X \subseteq I$ označujeme pomer*

$$\text{supp}(X) = \frac{|\{T \in D \mid X \subseteq T\}|}{|D|}.$$

Poznámka: Často sa tiež používa definícia $\text{supp}_{\text{ABS}}(X) = |\{T \in D \mid X \subseteq T\}|$, pričom pri takejto definícii ide horíme o *absolútnom supporte*. Platí však:

$$\text{supp}(X) = \frac{\text{supp}_{\text{ABS}}(X)}{|D|} = \frac{\text{supp}_{\text{ABS}}(X)}{n}.$$

Keďže n je pre konkrétnu databázu konštantné, vždy vieme ľahko prejsť od jednej definície k druhej. Tiež $\text{supp}(X) \in [0, 1]$ a $\text{supp}_{\text{ABS}}(X) \in \mathbb{N}^0$, je teda väčšinou zrejmé, o ktorej definícii sa hovorí. Z týchto dôvodov sa často nezdôrazňuje, či sa hovorí o absolútnom alebo relatívnom supporte. Pri výpočtoch však výrazom $\text{supp}(X)$ označujeme vždy relatívny support.

Pojem miery support teda vyjadruje v koľkých, respektíve v akej časti transakcií sa vyskytujú všetky položky z množiny X . Poznamenáme, že vždy $\text{supp}(\emptyset) = 1$, keďže prázdna množina je podmnožinou každej množiny. Často sa používa konštrukcia $\text{supp}(X \cup Y)$, $X, Y \subseteq I$, pričom platí $\text{supp}(X \cup Y) \leq \text{supp}(X)$ (a $\text{supp}(X \cup Y) \leq \text{supp}(Y)$). Množina $X \cup Y$ je totiž nadmnožina množiny X (resp. Y) a preto počet množín T_i ktoré sú jej nadmnožinou (teda sú nadmnožinou X a navyše aj Y), nemôže byť väčší.

Pomocou miery support sa definuje základná miera sily asociačného pravidla:

Definícia 1.2. *Mierou confidence asociačného pravidla $X \Rightarrow Y$, kde $X, Y \subseteq I$ a $\text{supp}(X) > 0$ označujeme podiel:*

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}.$$

Poznámka: Keďže $\text{supp}(X \cup Y) = \frac{\text{supp}_{\text{ABS}}(X \cup Y)}{n}$ a $\text{supp}(X) = \frac{\text{supp}_{\text{ABS}}(X)}{n}$, platí

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{\text{supp}_{\text{ABS}}(X \cup Y)}{\text{supp}_{\text{ABS}}(X)}.$$

Keďže $0 \leq \text{supp}(X \cup Y) \leq \text{supp}(X)$, vždy $\text{conf}(X \Rightarrow Y) \in [0, 1]$.

Miera confidence pravidla je pomer prípadov (počtu transakcií) v ktorých pravidlo platí ku počtu prípadov, v ktorých sa pravidlo dá uplatniť (počtu transakcií ktoré obsahujú jeho ľavú stranu). Confidence vraví, aká je pravdepodobnosť že v transakcii sa nachádza Y ak sa dá pravidlo aplikovať (ak sa v transakcii nachádza X). Pravidlo nič nehovorí o takom prípade, keď sa v transakcii nenachádzajú prvky X . O tom, ako často sa dá pravidlo použiť, hovorí pojem *support pravidla*:

Definícia 1.3. *Pojem support pravidla $X \Rightarrow Y$ definujeme ako:*

$$\text{supp}(X \Rightarrow Y) = \text{supp}(X),$$

kde $\text{supp}(X)$ je *support ľavej strany pravidla podľa definície 1.1.*

Poznámka: Ako podotýka [3], v literatúre (napr. aj v [1]) sa častejšie používa definícia $\text{supp}_{\text{ALT}}(X \Rightarrow Y) = \text{supp}(X \cup Y)$. Takáto definícia vyjadruje, v akej časti transakcií je pravidlo platné, pričom naša definícia hovorí, v akej časti sa pravidlo dá použiť. Medzi týmito definíciami existuje z definície confidence jednoduchý vzťah:

$$\begin{aligned} \text{supp}_{\text{ALT}}(X \Rightarrow Y) &= \text{supp}(X \cup Y) = \text{supp}(X) \cdot \text{conf}(X \Rightarrow Y) = \\ &= \text{supp}(X \Rightarrow Y) \cdot \text{conf}(X \Rightarrow Y) \end{aligned}$$

Podľa [3] je naša definícia lepšie použiteľná pri alternatívnych mierach sily pravidla, preto ju aj používame my.

1.2 Príklady použitia asociačných pravidiel

Definovali sme teda mieru support, ktorá hovorí o tom, ako často sa dá pravidlo použiť, resp. z akej veľkej časti dát sme pravidlo vyvodili — táto miera sa dá chápať aj ako miera štatistickej významnosti pravidla. Ďalej sme definovali mieru confidence, ktorá

vyjadruje, ako často pravidlo platí. Keď hovoríme o asociačných pravidlách, myslíme tým pravidlá ktoré spĺňajú určené limity na tieto miery, preto ich považujeme za použiteľné aj vierohodné. Teraz uvedieme niekoľko príkladov použitia asociačných pravidiel a predstavíme koncept, akým sa meria zaujímavosť pravidla.

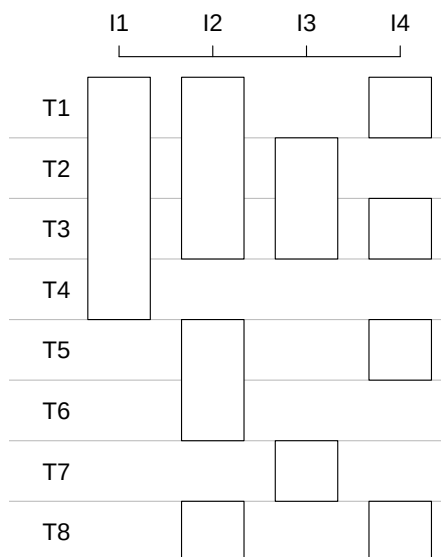
Príklad 1.1. Klasický príklad (podľa [1]) predstavuje analýza nákupného koša v maloobchodnom predaji. Jednotlivé transakcie predstavujú obsah jedného nákupu jedného zákazníka, pričom množina položiek zodpovedá množine predávaných tovarov. Pravidlo ako $\{\text{chlieb, lekvár}\} \Rightarrow \{\text{maslo}\}$ chápeme ako pozorovanie, že nákupy obsahujúce chlieb aj lekvár tiež často obsahujú maslo. Rôzne použitia takýchto pravidiel môžu byť:

- keď nájdeme všetky pravidlá ktorých pravá strana je produkt I_1 , zameranie sa na položky ľavých strán týchto pravidiel môžu pomôcť predaju I_1 ,
- pohľad na pravé strany pravidiel ktoré majú v ľavej strane produkt I_2 môže napovedať, predaj akých produktov bude ovplyvnený skončením predaja I_2 ,
- keď nájdeme pravidlá ktoré majú I_3 v ľavej a I_4 v pravej strane, zistíme, ktoré produkty treba predávať spolu s I_3 aby sa zvýšil predaj I_4
- po nájdení všetkých pravidiel obsahujúcich položky predávané na regáloch A a B môžeme zisťovať, či majú tieto regále súvis a či teda by mali byť umiestnené blízko seba

Príklad 1.2. Ďalším príkladom (podľa [5]) môže byť analýza príjmu obyvateľov na základe informácií ako vek, vzdelanie, atď. Transakciami tu sú jednotliví ľudia, pričom položky tu sú jednotlivé kúsky informácií, ktoré máme o danej osobe. Podstatná tu je diskretizácia číselných premenných ako vek a príjem do pásiem, pričom každé pásmo tvorí jednu položku. Keď príjem rozdelíme do troch pásiem na nízky, stredný a vysoký, vzniknú tým tri položky, pričom v každej transakcii bude práve jedna z týchto položiek. Ako výstup sa môžeme pozrieť na pravidlá ktoré majú v pravej strane napr. nízky príjem a tým zistíme, ktoré kombinácie faktorov sa často spájajú s nízkym príjmom.

Príklad 1.3. Ako ďalší prípad ukážeme príklad malej databázy s $n = 8$ (počet transakcií) a $m = 4$ (počet položiek) na ktorej predvedieme miery support, confidence. Databázu zobrazíme pomocou diagramu kde jednotlivé transakcie sú v riadkoch, položky

v stĺpcoch a obdĺžnikmi je zakreslené, ktoré transakcie obsahujú jednotlivé položky (obr. 1).



Obr. 1: Databáza transakcií pre príklad 1.3.

Teda máme množiny položiek $T_1, \dots, T_8$, položky I_1, I_2, I_3, I_4 , pričom napr. $T_1 = \{I_1, I_2, I_4\}$ alebo $T_2 = \{I_1, I_2, I_3\}$.

Položka I_1 sa objavuje v štyroch transakciách z ôsmich, preto $\text{supp}(\{I_1\}) = \frac{4}{8} = 0,5$. Položky I_1 a I_4 sa objavujú zároveň transakciách T_1 a T_3 , teda $\text{supp}(\{I_1, I_4\}) = \frac{2}{8} = 0,25$. Môžeme teda vypočítať $\text{conf}(\{I_1\} \Rightarrow \{I_4\}) = \frac{\text{supp}(\{I_1, I_4\})}{\text{supp}(\{I_1\})} = \frac{0,25}{0,5} = 0,5$. Teda položka I_4 sa vyskytuje v 50% transakcií ktoré obsahujú položku I_1 .

Podobne by sme zistili že takisto $\text{conf}(\{I_1\} \Rightarrow \{I_3\}) = 0,5$. Tiež napr. $\text{conf}(\{I_1, I_3\} \Rightarrow \{I_2\}) = 1$, čiže vo všetkých transakciách obsahujúcich I_1 aj I_3 sa nachádza aj I_2 . Z našej definície $\text{supp}(\{I_1, I_3\} \Rightarrow \{I_2\}) = \text{supp}(\{I_1, I_3\}) = 0,25$.

Ako významné a silné pravidlá (teda tie ktorými by sme sa ďalej zaoberali) by sme brali pravidlá spĺňajúce minimálnu hodnotu mier support s confidence, napr. $\text{minsup} = 0,2$ a $\text{minconf} = 0,5$. Existujú však ďalšie miery vhodnosti pravidiel podľa ktorých sa dajú ďalej filtrovať čo najsilnejšie pravidlá. Meraniu zaujímavosti pravidiel sa venuje nasledujúca podkapitola.

1.3 Meranie zaujímavosti pravidiel

Článok [5] hovorí o tzv. *apriórnej confidence* pravidla $X \Rightarrow Y$, ktorá sa definuje ako $\text{conf}(\emptyset \Rightarrow Y)$. Jedná sa o mieru confidence pravidla bez znalosti o ľavej strane pravidla. Keď sa pozrieme na pravidlo $\{I_1\} \Rightarrow \{I_4\}$ z príkladu 1.3, toto má hodnoty confidence aj apriórnej confidence zhodne 0,5. To znamená, že pravidlo $\{I_1\} \Rightarrow \{I_4\}$ nie je príliš zaujímavé — položka I_4 sa vyskytuje s rovnakou pravdepodobnosťou bez ohľadu na to, či sa v transakcii nachádza položka I_1 alebo nie. Zaujímavé pre nás teda su pravidlá, v ktorých sa miery confidence a apriórna confidence líšia. Pravidlo $\{I_1\} \Rightarrow \{I_3\}$ je príkladom takého pravidla, pretože $\text{conf}(\{I_1\} \Rightarrow \{I_3\}) = 0,5$, zatiaľ čo $\text{conf}(\emptyset \Rightarrow \{I_3\}) = 0,375$. Pomer týchto dvoch čísel sa používa ako miera zaujímavosti pravidla:

Definícia 1.4. Mierou lift pravidla $X \Rightarrow Y$ nazývame pomer

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{conf}(\emptyset \Rightarrow Y)}.$$

Poznámka: Keďže $\text{conf}(X \Rightarrow Y) = \text{supp}(Y)$, platí $\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{conf}(\emptyset \Rightarrow Y)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \text{supp}(Y)}$, ide teda o pomer supportu množiny $X \cup Y$ skutočného a takého, aký by bol za predpokladu nezávislosti X a Y . Zaujímavé môžu byť pravidlá s hodnotou výraznejšie vyššou alebo nižšou ako 1. Hodnota nižšia ako 1 totiž znamená, že spolu s X sa Y vyskytuje o dosť menej často ako v celej databáze. Zaujímavé môže byť preto aj pravidlo $X \Rightarrow Y$ s nulovou confidence (teda nulovým lift) — intuícia môže napovedať, že v tomto prípade z X vyplýva „neplatí Y “ so 100%-nou istotou.

Z definície 1.4 tiež vyplýva, že miera lift je symetrická, teda $\text{lift}(X \Rightarrow Y) = \text{lift}(Y \Rightarrow X)$ pre ľubovoľné množiny položiek X, Y .

1.3.1 Opak pravidla

Intuícia z komentára za definíciou 1.4 nás privádza k ďalšiemu pojmu ktorý budeme používať, a to pojmu *opaku* množiny položiek. Na množine položiek X nás často zaujímajú všetky transakcie, ktoré danú množinu X obsahujú. Keď hovoríme o opaku množiny X , značíme $\neg X$, vzťahujú sa k nemu všetky transakcie, o ktorých nie je pravda, že obsahujú množinu X . S opakom množiny pracujeme prakticky ako s každou množinou položiek, formálne však nejde podmnožinu množiny položiek, preto naň nejde pou-

žiť definíciu miery support akú máme pre množinu položiek. Zavádzame preto novú definíciu:

Definícia 1.5. Mieru support pre opak $\neg X$ množiny položiek $X \subseteq I$ definujeme ako

$$\text{supp}(\neg X) = \frac{|\{T \in D \mid X \not\subseteq T\}|}{|D|}.$$

Formálne nie je definované ani zjednotenie opaku množiny položiek s množinou položiek. Intuitívne preto môže byť zrejmé, čo môže byť myslené výrazom $\text{supp}(\neg X \cup Y)$ — ide o počet transakcií, ktoré obsahujú všetky prvky Y , neplatí ale, že obsahujú všetky prvky X . Pre množiny položiek $Y, Z \subseteq I$ rozpíšeme definíciu supportu ich zjednotenia:

$$\text{supp}(Z \cup Y) = \frac{|\{T \in D \mid (Z \cup Y) \subseteq T\}|}{|D|} = \frac{|\{T \in D \mid (Z \subseteq T) \wedge (Y \subseteq T)\}|}{|D|},$$

kde $(Y \cup Z) \subseteq M \Leftrightarrow (Y \subseteq M) \wedge (Z \subseteq M)$ je zrejmá množinová vlastnosť. Dosadením za $(Z \subseteq T)$ analogický výraz pre $\neg X$ z definície 1.5 získavame konzistentnú definíciu pre $\text{supp}(\neg X \cup Y)$:

Definícia 1.6. Nech $\neg X$ je opak množiny položiek a Y je množina položiek. Potom definujeme:

$$\text{supp}(\neg X \cup Y) = \frac{|\{T \in D \mid (X \not\subseteq T) \wedge (Y \subseteq T)\}|}{|D|}.$$

Poznámka: Analogickým spôsobom s vynechaním menovateľa v zlomkoch v definíciách by sa definoval absolútny support. Tiež podobným spôsobom ako sme dostali definíciu 1.6 by sme prišli k definícii „zjednotenia” dvoch opakov množín položiek, alebo kombinácie viacerých množín a opakov množín.

O opaku množiny v práci využívame tvrdenie, ktoré nasleduje po pomocnej leme:

Lema 1.7. Majme množiny $M_i, X, Y \subseteq J$ pre $i = 1, \dots, n$ a $\mathcal{M} = \{M_i \mid i = 1, \dots, n\}$. Potom platí:

$$\begin{aligned} \{M \in \mathcal{M} \mid Y \subseteq M\} = & \{M \in \mathcal{M} \mid (Y \subseteq M) \wedge (X \subseteq M)\} \cup \\ & \cup \{M \in \mathcal{M} \mid (Y \subseteq M) \wedge (X \not\subseteq M)\}, \end{aligned}$$

a

$$\{M \in \mathcal{M} \mid (Y \subseteq M) \wedge (X \subseteq M)\} \cap \{M \in \mathcal{M} \mid (Y \subseteq M) \wedge (X \not\subseteq M)\} = \emptyset.$$

Dôkaz. Pre každú množinu $M \in \mathcal{M}$ platí práve jedno z dvojice tvrdení $X \subseteq M$ a $X \not\subseteq M$. Preto je prienik dvoch množín vyžadujúcich tieto vlastnosti prázdny — pre žiadnu množinu z $\mathcal{M}$ neplatia obe tvrdenia. Z faktu že, aspoň jedno z tvrdení platí pre každý prvok $\mathcal{M}$ vyplýva, že každý prvok $\{M \in \mathcal{M} \mid Y \subseteq M\}$ sa nachádza aspoň v jednej z množín $\{M \in \mathcal{M} \mid (Y \subseteq M) \wedge (X \subseteq M)\}$ a $\{M \in \mathcal{M} \mid (Y \subseteq M) \wedge (X \not\subseteq M)\}$. $\square$

Veta 1.8. *Pre množiny položiek $X, Y \subset I$ platí:*

$$\text{supp}(Y) = \text{supp}(Y \cup X) + \text{supp}(Y \cup \neg X).$$

Dôkaz. Podľa lemy 1.7 platí:

$$\begin{aligned} \{T \in D \mid Y \subseteq T\} &= \{T \in D \mid (Y \subseteq T) \wedge (X \subseteq T)\} \cup \\ &\cup \{T \in D \mid (Y \subseteq T) \wedge (X \not\subseteq T)\}. \end{aligned}$$

Keďže množiny na pravej strane rovnosti sú podľa lemy 1.7 disjunktné, súčet počtu och prvkov sa rovná počtu prvkov ľavej strany rovnosti. Po predelení rovnosti o veľkostiach množín počtom prvkov databázy n , získame dokazovanú rovnosť. $\square$

Dôsledok 1.9. *Pre množiny položiek $X, Y \subset I$ platí:*

$$\text{conf}(Y \Rightarrow \neg X) = 1 - \text{conf}(Y \Rightarrow X).$$

Dôkaz. Podľa vety 1.8 platí $\text{supp}(Y \cup \neg X) = \text{supp}(Y) - \text{supp}(Y \cup X)$. Po vydelení oboch strán rovnosti výrazom $\text{supp}(Y)$ dostávame:

$$\text{conf}(Y \Rightarrow \neg X) = \frac{\text{supp}(Y \cup \neg X)}{\text{supp}(Y)} = \frac{\text{supp}(Y)}{\text{supp}(Y)} - \frac{\text{supp}(Y \cup X)}{\text{supp}(Y)} = 1 - \text{conf}(Y \Rightarrow X).$$

$\square$

1.3.2 Ďalšie tvrdenia o miere lift

So zavedením opaku pravidla dostávame ďalší pohľad na mieru zaujímavosti pravidla lift. Ako sme už naznačili, nezávislosť dvoch množín X a Y je teoreticky ekvivalentná tvrdeniu $\text{lift}(X \Rightarrow Y) = 1$ ($= \text{lift}(Y \Rightarrow X)$). Nezávislosť inými slovami môžeme intuitívne formulovať ako: „ Y nastáva s rovnakou pravdepodobnosťou bez ohľadu na to, či X nastáva alebo nie“. Formálne zapísané: $\text{conf}(X \Rightarrow Y) = \text{conf}(\neg X \Rightarrow Y)$. Intuíciu potvrdí nasledujúce tvrdenie:

Veta 1.10. *Pre množiny položiek $X, Y \subset I$ platí:*

$$\text{lift}(X \Rightarrow Y) = 1 \Leftrightarrow \text{conf}(X \Rightarrow Y) = \text{conf}(\neg X \Rightarrow Y).$$

Dôkaz. Dôkaz je zrejímavý po ekvivalentných úpravách vychádzajúc z definícií $\text{conf}(X \Rightarrow Y)$ a $\text{conf}(\neg X \Rightarrow Y)$. Využijeme tiež z vety 1.8, že $\text{supp}(\neg X \cup Y) = \text{supp}(Y \cup \neg X) = \text{supp}(Y) - \text{supp}(Y \cup X)$ a pre $Y = \emptyset$ dostávame z rovnakého tvrdenia $\text{supp}(\neg X) = \text{supp}(\emptyset \cup \neg X) = \text{supp}(\emptyset) - \text{supp}(\emptyset \cup X) = 1 - \text{supp}(X)$. Nasleduje séria ekvivalentných krokov:

$$\begin{aligned} \text{conf}(X \Rightarrow Y) &= \text{conf}(\neg X \Rightarrow Y), \\ \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} &= \frac{\text{supp}(\neg X \cup Y)}{\text{supp}(\neg X)}, \\ \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} &= \frac{\text{supp}(Y) - \text{supp}(X \cup Y)}{1 - \text{supp}(X)} \quad / \cdot \text{supp}(X)[1 - \text{supp}(X)], \\ \text{supp}(X \cup Y) \cdot [1 - \text{supp}(X)] &= [\text{supp}(Y) - \text{supp}(X \cup Y)] \cdot \text{supp}(X), \\ \text{supp}(X \cup Y) - \text{supp}(X \cup Y) \text{supp}(X) &= \text{supp}(Y) \text{supp}(X) - \text{supp}(X \cup Y) \text{supp}(X), \\ \text{supp}(X \cup Y) &= \text{supp}(Y) \text{supp}(X), \\ \frac{\text{supp}(X \cup Y)}{\text{supp}(Y) \text{supp}(X)} &= 1. \end{aligned}$$

Posledné tvrdenie z definície 1.4 znamená, že $\text{lift}(X \Rightarrow Y) = 1$. Dostali sme ho ekvivalentnými úpravami z pravej strany dokazovanej ekvivalencie, tá teda platí. $\square$

Poznámka: Rovnakým spôsobom by sa dalo ukázať, že platí

$$\text{lift}(X \Rightarrow Y) > 1 \Leftrightarrow \text{conf}(X \Rightarrow Y) > \text{conf}(\neg X \Rightarrow Y)$$

a

$$\text{lift}(X \Rightarrow Y) < 1 \Leftrightarrow \text{conf}(X \Rightarrow Y) < \text{conf}(\neg X \Rightarrow Y).$$

Tušenie nám navráva, že čím väčší je rozdiel medzi $\text{conf}(X \Rightarrow Y)$ a $\text{conf}(\neg X \Rightarrow Y)$, tým extrémnejšia by mala byť hodnota miery lift. Nasledujúca veta nám tušenie potvrdí.

Veta 1.11. *Pri fixnej množine položiek $X \subseteq I$ je miera $\text{lift}(X \Rightarrow Y)$ rastúcou funkciou miery $\text{conf}(X \Rightarrow Y)$ a klesajúcou funkciou miery $\text{conf}(\neg X \Rightarrow Y)$, pokiaľ sú obe miery confidence nenulové.*

Dôkaz. Vychádzame z dôkazu vety 1.10, kde sme pomocou vety 1.8 rozpísali $\text{conf}(\neg X \Rightarrow Y) = \frac{\text{supp}(Y) - \text{supp}(X \cup Y)}{1 - \text{supp}(X)}$. Z tohto výrazu vyjadríme $\text{supp}(Y)$:

$$\begin{aligned} \text{supp}(Y) &= \text{supp}(X \cup Y) + \text{conf}(\neg X \Rightarrow Y)[1 - \text{supp}(X)] = \\ &= \text{conf}(X \Rightarrow Y) \text{supp}(X) + \text{conf}(\neg X \Rightarrow Y)[1 - \text{supp}(X)], \end{aligned}$$

pričom sme využili definíciu $\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$. Ďalej dostávame nové vyjadrenie pre mieru lift:

$$\begin{aligned} \text{lift}(X \Rightarrow Y) &= \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)} = \\ &= \frac{\text{conf}(X \Rightarrow Y)}{\text{conf}(X \Rightarrow Y) \text{supp}(X) + \text{conf}(\neg X \Rightarrow Y)[1 - \text{supp}(X)]}. \end{aligned}$$

Keďže uvažujeme pevnú množinu X , môžeme $\text{supp}(X) \in (0, 1)$ označiť ako konštantu $\lambda \in (0, 1)$. Ďalej $\text{conf}(X \Rightarrow Y)$ a $\text{conf}(\neg X \Rightarrow Y)$ sú čísla, ktoré závisia iba na voľbe Y , tie označíme $\text{conf}(X \Rightarrow Y) =: \xi$ a $\text{conf}(\neg X \Rightarrow Y) =: \eta$. Pre funkciu lift teda dostávame vyjadrenie pomocou konštanty λ a premenných ξ a η :

$$\text{lift}(X \Rightarrow Y) = l(\xi, \eta) = \frac{\xi}{\lambda\xi + (1 - \lambda)\eta}, \text{ kde } \xi, \eta \in [0, 1].$$

Vypočítame parciálne derivácie funkcie $l(\xi, \eta)$:

$$\begin{aligned}\frac{\partial l(\xi, \eta)}{\partial \xi} &= \frac{\partial}{\partial \xi} \left[\frac{\xi}{\lambda \xi + (1 - \lambda) \eta} \right] = \frac{(\lambda \xi + (1 - \lambda) \eta) \frac{\partial}{\partial \xi} \xi - \xi \frac{\partial}{\partial \xi} (\lambda \xi + (1 - \lambda) \eta)}{(\lambda \xi + (1 - \lambda) \eta)^2} = \\ &= \frac{\lambda \xi + (1 - \lambda) \eta - \lambda \xi}{(\lambda \xi + (1 - \lambda) \eta)^2} = \frac{(1 - \lambda) \eta}{(\lambda \xi + (1 - \lambda) \eta)^2} \geq 0, \\ \frac{\partial l(\xi, \eta)}{\partial \eta} &= \frac{\partial}{\partial \eta} \left[\frac{\xi}{\lambda \xi + (1 - \lambda) \eta} \right] = \frac{-\xi \frac{\partial}{\partial \eta} (\lambda \xi + (1 - \lambda) \eta)}{(\lambda \xi + (1 - \lambda) \eta)^2} = \frac{(\lambda - 1) \xi}{(\lambda \xi + (1 - \lambda) \eta)^2} \leq 0.\end{aligned}$$

Ostré nerovnosti platia, ak $\text{conf}(\neg X \Rightarrow Y) \neq 0$, resp. $\text{conf}(X \Rightarrow Y) \neq 0$. $\square$

Poznámka: Analogické tvrdenie platí aj pre mieru lift ako funkciu premenných $\text{supp}(X \cup Y)$ a $\text{supp}(\neg X \cup Y)$, keďže pri fixnej množine X ide o rastúce lineárne funkcie premenných $\text{conf}(X \Rightarrow Y)$ a $\text{conf}(\neg X \Rightarrow Y)$: $\text{supp}(X \cup Y) = \text{conf}(X \Rightarrow Y) \text{supp}(X)$, $\text{supp}(\neg X \cup Y) = \text{conf}(\neg X \Rightarrow Y) \text{supp}(\neg X)$.

1.3.3 Overovanie nezávislosti množín položiek

V predchádzajúcom texte spomíname situáciu, keď sú dve množiny položiek navzájom nezávislé. Teoreticky sú dve množiny položiek nezávislé práve vtedy, keď $\Pr(X) \cdot \Pr(Y) = \Pr(X \cup Y)$, kde $\Pr(P)$ označíme pravdepodobnosť výskytu množiny položiek P . Práve miera support je odhadom pravdepodobnosti výskytu množiny položiek na základe dát. Z toho plyní tvrdenie o nezávislosti položiek ak miera lift je rovná jednej, keďže $\text{lift}(X \Rightarrow Y) = \text{lift}(Y \Rightarrow X) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \text{supp}(Y)} = 1 \Leftrightarrow \text{supp}(X \cup Y) = \text{supp}(X) \text{supp}(Y)$.

Nezávislosť množín položiek by sme mohli chcieť štatisticky otestovať na základe odhadu pre pravdepodobnosti výskytu, miery support. Podľa [10] je jednoduchým a používaným testom na overenie nezávislosti tzv. *chí-kvadrát* test. Pre dve množiny položiek množiny môžu v transakcii T_i nastať štyri situácie: $X \subseteq T_i \wedge Y \subseteq T_i$, $X \not\subseteq T_i \wedge Y \subseteq T_i$, $X \subseteq T_i \wedge Y \not\subseteq T_i$ a $X \not\subseteq T_i \wedge Y \not\subseteq T_i$. Chí-kvadrát test využíva tabuľku počtov jednotlivých situácií. Zapísané v našom značení ide o maticu:

$$\begin{pmatrix} \text{supp}_{\text{ABS}}(X \cup Y) & \text{supp}_{\text{ABS}}(\neg X \cup Y) \\ \text{supp}_{\text{ABS}}(X \cup \neg Y) & \text{supp}_{\text{ABS}}(\neg X \cup \neg Y) \end{pmatrix}.$$

Chí-kvadrát test testuje na základe tejto matice dvojicu hypotéz:

$$H_0 : X \text{ a } Y \text{ sú nezávislé} \quad \text{vs.} \quad H_1 : X \text{ a } Y \text{ sú závislé.}$$

Nakoniec, nezávislosť X a Y zamietame na zvolenej hladine významnosti α , ak p-hodnota chí-kvadrát testu je menšia ako α .

1.4 Hľadanie asociačných pravidiel

Na hľadanie asociačných pravidiel používame algoritmus *apriori* [2], konkrétne jeho implementáciu v balíku *arules* pre prostredie R [5]. Algoritmus nájde všetky pravidlá tvaru $X \Rightarrow Y$ v databáze, ktoré spĺňajú podmienky $X \cap Y = \emptyset$, $|Y| = 1$, $\text{supp}(X \Rightarrow Y) \geq \text{minsupp}$, $\text{conf}(X \Rightarrow Y) \geq \text{minconf}$ a $|X| + |Y| \leq \text{maxlen}$.

Algoritmus pracuje v dvoch fázach. V prvej fáze nájde všetky množiny položiek spĺňajúce kritéria maximálnej dĺžky a minimálnej miery support, z ktorých v druhej fáze vygeneruje pravidlá spĺňajúce navyše aj podmienku pre mieru confidence. Výpočtovo náročná je predovšetkým prvá fáza, kde treba preveriť veľké množstvo množín položiek; pre počet položiek m a dĺžku pravidla k existuje totiž $\binom{m}{k}$ množín položiek na overenie.

Algoritmus pre zjednodušenie problému využíva vlastnosť, že ak množina položiek X nespĺňa minimálnu hodnotu pre mieru support, ani žiadne iné množiny obsahujúce X hranicu nespĺňajú, keďže, ako spomíname v časti 1.1, platí $\text{supp}(X \cup Y) \leq \text{supp}(X) < \text{minsupp}$.

2 Dáta zdravotnej poisťovne, epizódy starostlivosti a asociačné pravidlá

V tejto kapitole sa venujeme niekoľkým rôznym témam, ktorých súlad však tvorí podstatu práce. Opisujeme dáta, s akými máme možnosť pracovať. Bližšie a presnejšie definujeme epizódy zdravotnej starostlivosti a ako ich vnímame v súvislosti s dostupnými dátami. Nakoniec opíšeme ideu použitia analýzy asociačných pravidiel ako hlavnej pomôcky pre pohľad na dáta, ktorý potrebujeme na zoskupovanie epizód zdravotnej starostlivosti.

2.1 Dáta zdravotnej poisťovne

K dispozícii máme dáta pochádzajúce od zdravotnej poisťovne. Zdravotné poisťovne na Slovensku majú o svojich poistencoch rôzne administratívne dáta, zdravotnícke informácie sú však značne obmedzené. Podľa [12] je databáza tzv. *vykázananej zdravotnej starostlivosti (VZS)* najdôležitejším dostupným zdrojom údajov o zdravotnom stave poistenca. Databáza vykázananej zdravotnej starostlivosti je súborom zdravotnej starostlivosti ako ju vykazujú poskytovatelia starostlivosti za účelom jej preplatenia. Je to zoznam výkazov na lieky, zdravotnícke pomôcky, rôzne výkony, hospitalizácie, atď, ide o kompletnú zdravotnú starostlivosť, ktorá bola pacientom poskytnutá. Obmedzujúcim faktorom je však štruktúra údajov, ktorá je prispôbena skôr na praktické vykazovanie zdravotnej starostlivosti a nie je ideálna na skúmanie. Medzi z nášho pohľadu najdôležitejšie údaje ktoré sú súčasťou vykázaných položiek patria:

- *identifikácia pacienta*, napríklad *rodné číslo*,
- *kód produktu* zdravotnej starostlivosti — identifikuje predmet výkazu. Unikátny kód má každý liek, pomôcka, výkon zdravotnej starostlivosti. Hospitalizácia sa vyказuje ako celok pod jedným produktom, aj keď niektoré výkony počas hospitalizácie sa vykazujú osobitne. Okrem kódu produktu máme k dispozícii jeho zaradenie do štruktúry produktov, napr. *skupiny* či *podskupiny* produktov, alebo *ATC skupiny* lieku,
- *kód diagnózy*, k liečeniu ktorej je určený predmet výkazu, podľa [11]. Používa sa

trojmiestny alebo päťmiestny kód, pričom päťmiestny je spresnením trojmiestneho,

- *dátum* vykonania výkonu, pri liekoch dátum predpisania aj vyzdvihnutia lieku, resp. pri hospitalizácii jej začiatok aj koniec,
- *kód typu zdravotnej starostlivosti*, či ide o liek, všeobecnú či špecializovanú ambulantnú starostlivosť, hospitalizáciu, atď,
- *náklady* na predmet zdravotnej starostlivosti. Existuje viacero typov nákladov, najdôležitejšie sú asi *vykázané* náklady ako ich vykáže poskytovateľ, alebo *schválené* náklady, keďže poisťovňa niekedy neschváli vykázané náklady v celej výške,
- *identifikácia lekára* a nižšej zložky a to vykonávajúcich, preskribujúcich, alebo odporúčajúcich. Nižšia zložka predstavuje menšiu jednotku v rámci poskytovateľa zdravotnej starostlivosti, ako napr. jednu ambulanciu. Kód nižšej zložky v sebe zahŕňa trojmiestny číselný kód *odbornosti* nižšej zložky.

Podľa [12] sa nedá vždy spoliehať na úplnú presnosť výkazov — napríklad kód diagnózy často nebýva vykazovaný jednoznačne, niektorý lekár môže používať trojmiestny kód, zatiaľ čo iný lekár by použil presnejší päťmiestny. Tiež podobné výkony môžu byť vykázané pod iným kódom ako urgentná diagnóza, teda v zásade rovnaký zákrok môže byť pod rôznymi diagnózami, pričom sa líšia okolnosti poskytnutia výkonu. Nakoniec nie je jasné, aké ďalšie diagnózy môžu súvisieť s primárnou diagnózou. Preto na kód diagnózy pozeráme ako na dôležitú, nie však absolútne spoľahlivú informáciu a používame ho v spojení s ostatnými informáciami.

Ďalší problém súvisí s hospitalizáciami, hospitalizácie používajú vždy kód produktu „*HOSP**“, pričom máme iba nepriame informácie o účele hospitalizácie z vykazanej diagnózy, odbornosti oddelenia, atď. Počas hospitalizácie sa síce zvyknú osobitne vykazovať niektoré výkony, z tých však tiež nevieme ľahko získať presné informácie o účele hospitalizácie.

Práve k hospitalizáciám sa viaže systém čakacích listín a tu sa dá získať istá dodatočná informácia o hospitalizácii. Ak sa jedná o niektorú z hospitalizácií k výkonom ktoré sú definované v [14], máme dodatočnú informáciu, ktorý typ to je. Túto informáciu máme iba pre plánované hospitalizácie, podobné operácie ktoré je potrebné

vykonať akútne takto identifikovať nevieme. Môžeme ale napríklad presne identifikovať plánované hospitalizácie k výmene bedrového alebo kolenného kĺbu, ktoré majú podľa [14, príloha] kódy „WL3031”, resp. „WL0302”. Tieto hospitalizácie môžu byť podľa [12] dobrým príkladom pre náš výskum — epizódy okolo nich môžu zahŕňať problémy pred plánovanou hospitalizáciou, pričom na starostlivosti po plánovanej hospitalizácii zas môžeme skúmať rehabilitáciu či ďalšie komplikácie. Rozhodli sme sa preto zvoliť epizódy súvisiace so spomínanými dvomi hospitalizáciami ako predmet skúmania v práci.

Čakacie listiny sú jeden z mála spôsobov mimo vykázananej zdravotnej starostlivosti, ako získať o pacientoch medicínske informácie (ďalšími spôsobmi môžu byť napr. výnimky na lieky alebo zdravotnícke pomôcky ktoré sa vedú tiež mimo VZS, tie nám ale prídu značne limitované a málo použiteľné). Skutočne, aj na Slovensku pre účely pre-rozdeľovania financií medzi zdravotnými poisťovňami používaný systém kategorizácie pacientov podľa chronických chorôb do tzv. farmaceuticko-nákladových skupín používa výhradne informácie o vykázaných liekoch [16, prevažne § 27b]. Toto poukazuje na použitie vykázananej zdravotnej starostlivosti pre naše účely za oprávnené.

2.2 Zoskupovanie epizód starostlivosti

Epizóda zdravotnej starostlivosti sa definuje ako množina výkonov zdravotnej starostlivosti, ktoré sa viažu k jednej chorobe, úrazu alebo inej udalosti zhoršeného zdravotného stavu u jedného pacienta v priebehu istého obdobia, často od prvých vyšetrení po koniec liečby [9]. Keď sa pozrieme na epizódu starostlivosti, vidíme kompletnú zdravotnú starostlivosť, ktorá bola pacientovi poskytnutá v súvislosti s konkrétnym problémom v pacientovom zdraví. Výkony v rámci jednej epizódy môžu byť poskytované u jedného alebo viacerých lekárov, u jedného alebo viacerých poskytovateľov zdravotnej starostlivosti.

Pacient môže mať naraz viacero rôznych epizód ktoré spolu môžu alebo nemusia súvisieť. Príkladom nesúvisiacich epizód môže byť keď si pacient s astmou zlomí nohu — liečba astmy ako chronickej choroby dýchacieho traktu nesúvisí s liečbou zlomenej nohy. Epizódy, ktoré však spolu súvisia, ktoré môžu zdieľať liečbu alebo sa ovplyvňovať, sú na popis komplikovanejšie — vtedy vyvstáva otázka, či takúto kombináciu

epizód nedefinovať ako samostatnú, inú epizódu. Ďalej sa môžu epizódy ďalej vetviť do viacerých typov podľa závažnosti prípadu, či výskytu rôznych komplikácií.

Pri pohľade na dobre definované epizódy vidíme celú zdravotnú starostlivosť súvisiacu s jedným konkrétnym zdravotným problémom. Keď sa pozrieme na jeden konkrétny typ epizódy u viacerých skupín pacientov (napríklad pacientov rôznych regiónov alebo pacientov liečených v rôznych nemocniciach), môžeme porovnávať skutočné ceny, skutočnú dĺžku hospitalizácie, skutočné množstvo výkonov, ktoré bolo priemerne podané každej skupine pacientov na vyliečenie rovnakých zdravotných problémov. Tu je jedna z aplikácií epizód starostlivosti, vo férovejšom porovnávaní poskytovateľov zdravotnej starostlivosti a v následných možnostiach hľadanie neefektív alebo odmeňovania dobrej práce — pri pohľade na epizódu starostlivosti napríklad môže byť vidieť, ktorí pacienti potrebovali viac pooperačnej starostlivosti.

Pri zoskupovaní epizód starostlivosti narážame na nasledovné hlavné výzvy:

1. ako identifikovať tzv. „spúšťač“ epizódy, čiže výkon alebo výkony zdravotnej starostlivosti, na základe ktorých určíme, že pacient mal epizódu konkrétneho typu priradeného k spúšťaču epizódy. Môže ísť napr. o plánovanú alebo akútnu hospitalizáciu, vyšetrenie s istým nálezom, alebo, ak sa jedná o chronickú chorobu, aj o dlhodobú preskripciu liekov istého typu,
2. ako určiť časový horizont epizódy pred a po jej spúšťači, počas ktorého hľadáme spomedzi pacientových výkonov tie, ktoré možno k epizóde priradiť,
3. ktoré výkony v určenom časovom horizonte priradiť k identifikovanému typu epizódy,
4. ako určiť prípadnú triedu závažnosti epizódy, alebo ako identifikovať, či sa jedná o epizódu s komplikáciou, či epizódu súvisiacu s niektorou inou pacientovou epizódou,
5. ako k epizóde priradiť lekára alebo poskytovateľa zdravotnej starostlivosti, ktorý je za ňu zodpovedný.

V dokumente [4, časť 3] opisuje *Centers of Medicaid & Medicare Services* (obdobu štátnej zdravotnej poisťovne v USA zastrešujúca zdravotné poistenie pre dôchodcov a

sociálne odkázané vrstvy) svoje riešenia väčšiny bodov pri opise konštrukcie epizód pre niekoľko často sa vyskytujúcich pľúcnych a srdcových chorôb:

1. Spúšťač epizód je definovaný na základe medicínskych znalostí pomocou kódov diagnóz výkonov. Pre spustenie epizódy liečenej ambulantnou starostlivosťou sú nutné dve návštevy ambulancie v priebehu 30 dní. Epizódy pre hospitalizácie začínajú príslušnou hospitalizáciou. Chronické epizódy sú tu definované ako srdcové choroby a spúšťajú sa tiež príslušným vyšetrením, alebo automaticky pri spustení priradenej akútnej epizódy.
2. Pre akútne epizódy je stanovená dĺžka 90 dní od počiatku. Pre hospitalizácie sa berú navyše tri dni pred začiatkom hospitalizácie na zachytenie výkonov súvisiacich s diagnostikou. Chronická epizóda trvá 365 dní a obnovuje sa na ďalších 365 dní ak zaznamená opätovný výskyt spúšťača.
3. Výkony sa priradujú na základe kódu produktu a kódu diagnózy príslušných k epizóde. Niektoré produkty, ako napr. vyšetrenia u lekára sú zaradené k epizóde len ak majú priradenú diagnózu spojenú s epizódou — bez informácie o diagnóze by totiž mohli prislúchať k rozličným zdravotným stavom. Iné produkty, ktoré sú dostatočne späté s príslušnou epizódou sa k nej priradia vždy ak sa nájdu v určenom časovom okne pre epizódu. Všetky informácie o príslušnosti epizód, diagnóz a produktov pochádzajú od zdravotníckych expertov, sú teda medicínskeho pôvodu.
4. Závažnosť epizódy je určená v definícii epizódy výskytom dodatočných komplikácií, ako napr. infarkt myokardu pri srdcových chorobách.

My sa pri bode 1 zameriavame na dva typy operácii — implantáciu bedrového a kolenného kĺbu. Tieto vieme v dátach identifikovať podľa záznamov z čakacích listín, ako je spomínané v časti 2.1. Priradovaniu výkonov k epizóde (bod 3) sa ďalej venujeme vo zvyšku tejto kapitoly a v celej kapitole 3, pričom informácie o vzťahoch v dátach získavame z dát vykázananej zdravotnej starostlivosti pomocou analýzy asociačných pravidiel. Našou ambíciou je teda prísť k pohľadu na epizódy starostlivosti založenému na informáciách získaných z dát, čo je zásadný rozdiel oproti postupu založenému na medicínskych faktoch z [4]. Otázkami 2, 4 a 5 sa zaoberáme po vyriešení 3.

2.3 Dátový prístup k zoskupovaniu epizód starostlivosti

V nasledujúcich častiach práce je našim zámerom popísať postup, ako aplikujeme analýzu asociačných pravidiel na hľadanie zákonitostí v dátach, ktoré nám môžu pomôcť pri priradovaní výkonov k epizódam starostlivosti. Naš prístup je, ako je spomínané v predchádzajúcej časti, založený takmer výhradne na pohľade na dáta. V tejto podsekcii bližšie rozoberáme, možné dôsledky nášho dátového prístupu oproti medicínskemu pohľadu napr. z [4].

Prvou skutočnosťou ktorú si treba uvedomiť je, že ani pri dátovom tak aj pri medicínskom prístupe k zoskupovaniu epizód starostlivosti ide v zásade o neriadenú klasifikáciu. Ide o úlohu vytvoriť pravidlá na základe ktorých jednotlivý výkon buď priradíme alebo nepriradíme ku konkrétnej epizóde (resp. mohli by sme výkon priradiť k epizóde čiastočne, ak by napr. mohol byť podľa našich pravidiel priradený k viacerým rôznym epizódam). Pravidlá však vytvárame bez poznania správneho zaradenia, teda naše pravidlá nemôžeme vytvárať vzhľadom na nejakú dopredu známu klasifikáciu. Pri medicínskom prístupe je vytvorená klasifikácia na základe medicínskych znalostí o epizódach, ktoré predstavujú rôzne zdravotné stavy. Stále však ide len o akési kvalifikované „hádanie“, ktoré môže skutočnosť opisovať rôzne presne, pričom ale nevieme overiť, do akej miery. To, že zoskupovanie epizód na základe medicínskych informácií nemôže byť absolútne presné vyplýva napr. z faktu, že aj medzi výsledkami dvoch komerčných zoskupovačov epizód existujú isté rozdiely [9].

Na základe konzultácie s lekárom budeme mať možnosť získať medicínske priradenie výkonov k jednej z epizód. Toto priradenie použijeme ako verifikáciu výsledkov, hlavným cieľom je ale popísať zoskupovanie epizód bez medicínskeho vstupu. Potrebujeme sa preto pozrieť na dáta a použiť metódy, ktoré na nás budú pôsobiť čo najrozumnejšie. V priebehu postupu sa snažíme o rôzne spôsoby nepriameho overenia výsledkov, ktoré nezávisia na medicínskom vstupe.

Považujeme za pravdepodobné, že použitím dátového prístupu dosiahneme menšiu presnosť ako pri medicínskom prístupe. Ako uvidíme neskôr, musíme totiž arbitrárne voliť rôzne parametre (ako napr. hranicu pre support a confidence pri analýze asociačných pravidiel). Keďže týchto parametrov je pomerne veľa a porovnanie s medicínskym priradením môžeme vykonať až na úplnom konci dátového priradenia, môže byť výzvou

do budúcnosti skúmať detailnejšie účinky zmeny jednotlivých parametrov na výsledok.

Vieme si však predstaviť aj výhody dátového prístupu. Očividnou výhodou je to, že sa oň môžeme pokúšať aj my, teda len na základe dát a bez medicínskych znalostí. Zároveň náš postup ani nevyužíva medicínsku informáciu ktorú máme k dispozícii, ako napríklad názvy diagnóz. Keď dostaneme nejaké výsledky, môžeme sa informatívne pozrieť na ich interpretáciu, tú ale používame maximálne pri rozhodovaní o voľbe parametrov, ktoré by sme inak volili úplne arbitrárne. Inými slovami, náš proces je automatický. Na základe dát a spúšťača epizódy sám vytvorí priradenie výkonov. Toto priradenie teda vyžaduje iba malú námahu, hoci tá je kompenzovaná možnou menšou presnosťou.

Ako sme už spomenuli, náš dátový pohľad na zoskupovanie epizód starostlivosti je založený na analýze asociačných pravidiel. Podrobne tento prístup rozoberáme v nasledujúcej časti. Tu len rozoberieme pár základných myšlienok s ktorými pracujeme. V podkapitole 2.1 máme opísané dáta ktoré máme k dispozícii. Ako vieme z kapitoly 1, pomocou asociačných pravidiel vieme identifikovať, ktoré „položky” sú v silnom vzťahu, napríklad čo sa týka ich závislosti. V zásade našim prístupom bude pozrieť sa na rôzne aspekty dát, ako diagnózy, odbornosti lekárov, skupiny produktov zdravotnej starostlivosti, atď., prehlásiť ich za položky v analýze asociačných pravidiel a hľadať, ktoré sú do rôznej miery silno závislé na položke „pacient spadá do epizódy ktorú skúmame”. Výkony, ktoré sa dostatočne viažu k takto zisteným vlastnostiam, priradíme k epizóde. Presnejšie v nasledujúcej časti.

2.4 Analýza asociačných pravidiel pre zoskupovanie epizód starostlivosti

V tejto časti rozoberáme myšlienkový postup, ktorým sme sa dopracovali k formulovaniu úlohy analýzy asociačných pravidiel, ktorú používame ďalej. Väčšiu časť úvah ilustrujeme na konkrétnom type úlohy, pričom v závere časti odvodíme zo získaných záverov bližšiu predstavu o úlohách, aké používame v ďalších kapitolách.

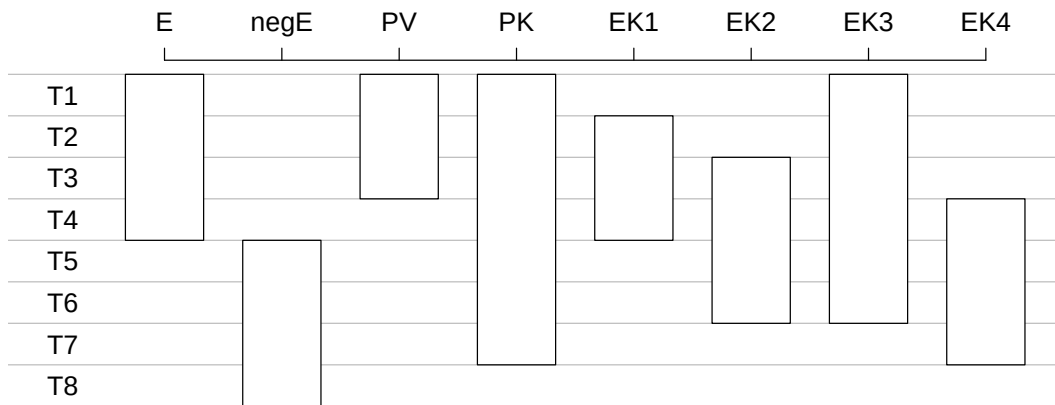
2.4.1 Úloha asociačnej analýzy o produktoch

Uvažujme množinu položiek $I = \{p_1, p_2, \dots, p_m, e\}$, kde $p_1, p_2, \dots, p_m$ je zoznam všetkých produktov zdravotnej starostlivosti, ktoré sa nachádzajú v našich dátach a e predstavuje epizódu, ktorú skúmame. Množina transakcií nech predstavuje množinu pacientov ktorú uvažujeme, pričom $T_i \subseteq I$ je množina všetkých produktov, ktoré pacient i v dátach zaznamenal aspoň raz a v množine sa tiež vyskytuje prvok e ak je u pacienta spustená skúmaná epizóda. Označme P ako množinu obsahujúcu jeden produkt, $P = \{p_k\}, k \in \{1, \dots, m\}$ a $E = \{e\}$. O týchto množinách položiek ďalej skrátene hovoríme ako o produkte, resp. o epizóde. V ďalšom postupe skúmame asociačné pravidlá v tvare $P \Rightarrow E$ a $E \Rightarrow P$. Na základe nich sa snažíme dospieť k tomu, ako určiť, ktoré produkty boli vykázané kvôli skúmanej epizóde — tie chceme k epizóde priradiť.

Nech množina $P_V = \{p_k\}, k \in \{1, \dots, m\}$ predstavuje produkt „významný“, ktorý sa viaže čisto k nami skúmanej epizóde, teda produkt, ktorý sa vyskytuje len ak je u pacienta prítomná skúmaná epizóda. Potom platí $\text{conf}(P_V \Rightarrow E) = 1$, pretože vždy keď má pacient produkt P_V , má aj epizódu E . Môžeme si tiež uvedomiť, že táto skutočnosť platí bez ohľadu na $\text{supp}(P_V) = \text{supp}(P_V \Rightarrow E)$. Takýto produkt by sme určite chceli k epizóde priradiť.

O hodnote $\text{conf}(E \Rightarrow P_V)$ tu ale bez ďalších informácií nevieme povedať nič — môžu sa vyskytnúť prípady, kedy pacient síce má epizódu E , produkt P_V ale použitý nebol. Hodnota $\text{conf}(E \Rightarrow P_V)$ nám však príliš nápomocná byť nemusí. Predstavme si „klamlivý“ produkt P_K , pre ktorý je $\text{conf}(E \Rightarrow P_K) = 1$, alebo všeobecnejšie, $\text{conf}(E \Rightarrow P_K)$ je blízke jednotke (pre skrátenie značíme $E \Rightarrow P_K$). Toto tvrdenie znamená, že veľká väčšina pacientov so skúmanou epizódou zaznamenala produkt P_K . Tento produkt ale mohli zaznamenať aj pacienti bez skúmanej epizódy — môžu existovať iné dôvody ako epizóda E , prečo majú pacienti produkt P_K .

Predstavme si preto, že v dátach máme zaznamenanú aj inú epizódu E_K , pre ktorú platí $E_K \Rightarrow P_K$. Uvažujeme ale E_K také, že neplatí tvrdenie $E_K \Rightarrow E$. Pravidlo $E_K \Rightarrow E$ totiž znamená, že spolu s E_K sa veľmi často vyskytuje E a teda automaticky sa často vyskytuje aj P_K . Pre úvahu má zmysel aj symbolický zápis: $[(E_K \Rightarrow E) \wedge (E \Rightarrow P_K)] \Rightarrow (E_K \Rightarrow P_K)$. V tomto prípade preto $E_K \Rightarrow P_K$ už nie je žiadnou dodatočnou informáciou k $E_K \Rightarrow E$ a $E \Rightarrow P_K$.



Obr. 2: Príklad pre popisované množiny položiek: E , $\neg E$, $P_V \Rightarrow E$, $E \Rightarrow P_K$, $E_{K1} \Rightarrow E$, E_{K2} nezávislé s E , $E \Rightarrow E_{K3}$ pričom $E_{K2}, E_{K3}, E_{K4} \not\Rightarrow E$ a $E_{K1}, E_{K2}, E_{K3}, E_{K4} \Rightarrow P_K$.

Keby teda máme v dátach všetky typy epizód ktoré môže pacient mať, mohli by sme postupovať nasledovným spôsobom:

1. identifikovať všetky množiny položiek zodpovedajúce epizódam, E_i , také, že
 - (a) $E_i \not\Rightarrow E$, alebo
 - (b) E_i a E sú nezávislé,
2. identifikovať množinu položiek zodpovedajúce produktom P , že $E \Rightarrow P$,
3. pre každé E_i zistiť, či platí $E_i \Rightarrow P$,
4. ak zistíme, že pacient má E a nemá žiadnu z E_i pre ktoré platí podmienka z 3., vyhlásime, že P sa u pacienta viaže k epizóde E .

Tento postup by sme zopakovali pre P zodpovedajúce postupne všetkým produktom $p_1, \dots, p_m$. V kroku 1. by sme zvolili medzi kritériom „ $\not\Rightarrow$ ” resp. nezávislosti, podľa požadovanej interpretácie (kritérium nezávislosti je silnejšie ako kritérium „ $\not\Rightarrow$ ”). Kritérium $E_i \not\Rightarrow E$ vlastne znamená, že neplatí $\text{conf}(E_i \Rightarrow E) \geq c$, hranica pre mieru confidence, teda platí $\text{conf}(E_i \Rightarrow E) < c$, čo je podľa dôsledku 1.9 ekvivalentné $\text{conf}(E_i \Rightarrow \neg E) > 1 - c$, teda vedeli by sme takéto pravidlá nájsť klasickým postupom analýzy asociačných pravidiel, kde hľadáme pravidlá s mierou confidence nad určenou hranicou. Kritérium nezávislosti by sme zas mohli overiť pomocou chí-kvadrát testu z časti 1.3.3, alebo jednoduchšie, cez určenú hranicu pre mieru lift.

Popísaný postup je možno myšlienkovu zaujímavý, v praxi pre nás ale nie je uplatniteľný. Museli by sme totiž nejako identifikovať rôzne epizódy E_i , čo by bolo náročné

a potencionálne nespoľahlivé. Museli by sme ešte riešiť problém, ako sa vysporiadať so situáciou v bode 3., ak zistíme, že pacient má niektoré z epizód E_i relevantné pre produkt P .

Z tohto dôvodu sa vrátime k pravidlu $P \Rightarrow E$ a pozrieme sa na jeho mieru zaujímavosti lift. Platí:

$$\text{lift}(P \Rightarrow E) = \frac{\text{conf}(P \Rightarrow E)}{\text{conf}(\emptyset \Rightarrow E)} = \frac{\text{conf}(P \Rightarrow E)}{\text{supp}(E)}.$$

My pracujeme s pevnou epizódou E , preto je $\text{supp}(E) < 1$ konštanta. Preto maximálna hodnota miery lift pre pravidlá $P \Rightarrow E$ je $\frac{1}{\text{supp}(E)} > 1$ a toto maximum sa nadobúda pre $\text{conf}(P \Rightarrow E) = 1$. Našou ideou je preto zistiť, ktoré produkty majú dosť vysokú mieru lift pravidla $P \Rightarrow E$ a túto vedomosť skombinovať s podobnými pravidlami o diagnózach, odbornostiach, atď., o čom hovorí ďalšia podkapitola.

K miere lift uvedieme ešte jednu poznámku. Ako uvádzame neskôr, v kapitole 3.1, pracujeme s dátami, ktoré obsahujú všetkých pacientov s epizódou E , ale iba časť, náhodnú vzorku, zo všetkých pacientov bez epizódy E ktorých máme k dispozícii. Preto hodnota lift s ktorou pracujeme, nezodpovedá hodnote lift, ktorú by sme získali z kompletných dát. Toto je dobre vidieť po rozpísaní definície pre mieru lift dvomi spôsobmi:

$$\text{lift}(P \Rightarrow E) = \text{lift}(E \Rightarrow P) = \frac{\text{conf}(P \Rightarrow E)}{\text{conf}(\emptyset \Rightarrow E)} = \frac{\text{conf}(E \Rightarrow P)}{\text{conf}(\emptyset \Rightarrow P)}$$

Z každého spôsobu rozpísania poznáme iba jeden z dvoch výrazov zlomku. Poznáme $\text{conf}(\emptyset \Rightarrow E) = \text{supp}(E)$, lebo máme všetkých pacientov s E , ale nepoznáme $\text{conf}(P \Rightarrow E)$, lebo nepoznáme výskyt P vo všetkých dátach. Podobne nepoznáme $\text{conf}(\emptyset \Rightarrow P) = \text{supp}(P)$, ale zas poznáme $\text{conf}(E \Rightarrow P) = \frac{\text{supp}(E \cup P)}{\text{supp}(E)} = \frac{\text{supp}_{\text{ABS}}(E \cup P)}{\text{supp}_{\text{ABS}}(E)}$, keďže obidva supporty sa zisťujú iba na základe transakcií obsahujúcich E .

Vo vete 1.10 sa hovorí o hodnotách $\text{conf}(E \Rightarrow P)$ a $\text{conf}(\neg E \Rightarrow P)$. Už sme spomínali, že $\text{conf}(E \Rightarrow P)$ poznáme. Hodnotu $\text{conf}(\neg E \Rightarrow P)$ síce nepoznáme, vieme však získať túto hodnotu aproximovanú. V našich dátach totiž máme náhodnú vzorku z dát kde platí $\neg E$, hodnota $\text{conf}(\neg E \Rightarrow P)$ z našich dát je preto odhadom hodnoty $\text{conf}_C(\neg E \Rightarrow P)$ z celkových dát. Preto podľa poznámky za vetou 1.10, keď

$\text{lift}(P \Rightarrow E) = \text{lift}(E \Rightarrow P) > 1$ potom aj $\text{conf}(E \Rightarrow P) > \text{conf}(\neg E \Rightarrow P)$. Ak usúdime, že nerovnosť platí aj pre miery z celkových dát na základe toho, že platí pre odhad jednej z mier, potom $\text{conf}(E \Rightarrow P) > \text{conf}_C(\neg E \Rightarrow P)$ a zas na základe vety 1.10 aj celkovo $\text{lift}_C(P \Rightarrow E) > 1$.

Záverom, ak pozorujeme $\text{lift}(P \Rightarrow E) > 1$, môžeme považovať aj hodnotu z kompletných dát $\text{lift}_C(P \Rightarrow E)$ za väčšiu ako 1. Treba však brať do úvahy, že pre mieru lift by sme z iného počtu dát mohli dostať inú hodnotu.

2.4.2 Využitie viacerých množín asociačných pravidiel

V predchádzajúcej časti sme potvrdili, že najjednoduchšou a potencionálne efektívnou možnosťou na nájdenie vzťahov medzi produktmi ZS a skúmanou epizódou je pozeráť sa na mieru lift pravidla $P \Rightarrow E$, kde E je jednozložková množina položiek predstavujúca skúmanú epizódu. Vieme, že nás zaujímajú pravidlá s mierou lift vyššou ako 1, konkrétnu spodnú hranicu pre lift ktorú prakticky aplikujeme ale určíme pri pohľade na dáta.

Namiesto produktov môžeme pri hľadaní asociačných pravidiel uvažovať iné premenné, alebo ich kombinácie. Ako príklad označíme O jednozložkovú množinu položiek predstavujúcu odbornosť a D množinu predstavujúcu diagnózu. Môžeme sa pozeráť na pravidlá o diagnózach $D \Rightarrow E$, odbornostiach $O \Rightarrow E$, alebo aj kombináciách viacerých premenných, napr. $D \cup O \cup P \Rightarrow E$. Tiež za množinu transakcií nemusíme brať pacientov, množinou transakcií môžu byť napr. ambulantné kontakty. Pravidlo $D \cup O \cup P \Rightarrow E$ na množine ambulantných kontaktov by sa dalo interpretovať nasledovne: keď pacient navštívi ambulanciu odbornosti O s diagnózou D a vykoná sa produkt P , často má epizódu E .

Na základe vety 1.11 máme presnejšiu interpretáciu pravidla $X \Rightarrow E$ v závislosti od miery lift : čím väčšia je hodnota miery lift , tým častejšie sa vyskytuje X spolu s E oproti tomu, ako často sa vyskytuje X bez E . Symbolicky zapísané: s pevným E vedie ku zvýšeniu $\text{lift}(X \Rightarrow E) = \text{lift}(E \Rightarrow X)$ zvýšenie $\text{supp}(E \cup X)$ alebo zníženie $\text{supp}(\neg E \cup X)$.

Keďže sme prišli na to, že napr. výlučne na základe asociačnej analýzy na produktoch zdravotnej starostlivosti sa ťažko dopracujeme k dôveryhodným výsledkom, prinášame

myšlienku spojiť výsledky analýz pre rôzne pohľady na dáta. Nájdeme rôzne množiny asociačných pravidiel nad rôznymi množinami položiek (nad spomínanými diagnózami, odbornosťami, atď., alebo kombinovane) a prípadne rôznymi množinami transakcií (pacienti, kontakty). Každé pravidlo dostane určený počet bodov napríklad na základe hodnoty jeho miery lift. Každý ambulantný kontakt pacienta dostane počet bodov na základe toho, z koľkých množín pravidiel sa neho dajú pravidlá uplatniť — teda podľa toho, nakoľko sedí diagnóza, odbornosť a ďalšie použité premenné, prípadne ich kombinácie. Po priradení bodov preskúmame ich pridelené počty a určíme minimálnu hranicu bodov na to, aby sme daný kontakt prehlásili za príslušný k epizóde.

Nevýhodou prístupu je, že musíme v podstate arbitrárne rozhodnúť, za ktoré množiny pravidiel priradíme aký počet bodov a ako priraďovať body každej z množiny pravidiel. Tu nás spočiatku môže viesť predovšetkým intuícia. Ako sme spomínali, výsledky porovnáme s medicínskym zaradením na jednej epizóde a tým overíme použiteľnosť postupu.

3 Zoskupovanie epizód starostlivosti

V tejto kapitole najprv predstavíme presnú formu dát, akú využívame na zoskupovanie epizód starostlivosti prihliadajúc na pozorovania z kapitoly 2, pričom sa venujeme aj spracovaniu a základnému prieskumu týchto dát. Ďalej sa bližšie venujeme úvahám, akými sme postupovali pri volení presného postupu zoskupovania epizód starostlivosti. Nakoniec sa snažíme vyvodiť pozorovania o dosiahnutých výsledkoch.

3.1 Dáta

3.1.1 Výber dát

Ako spomíname v časti 2.1, pracujeme prevažne s dátami *vykázanej zdravotnej starostlivosti*. Epizódy starostlivosti ktoré skúmame sú *implantácia bedrového kĺbu*, ktorú vieme identifikovať z čakacích listín na hospitalizácie podľa kódu „WL0301” a *implantácia kolenného kĺbu* s kódom „WL0302”.

Pre naše účely sme vybrali všetkých dostupných pacientov, ktorí podstúpili aspoň jednu z dvoch typov operácií v priebehu roku 2012, presnejšie všetkých pacientov ktorých nejaká hospitalizácia obsahujúca jeden z dvoch spomínaných kódov z čakacích listín zasahovala aspoň jedným dňom do roku 2012. Konkrétne ide o 825 pacientov s implantáciou bedrového kĺbu a 632 pacientov s implantáciou bedrového kĺbu. Dokopy 5 pacientov zaznamenalo v priebehu roku 2012 implantáciu aj bedrového aj kolenného kĺbu.

Pre potreby analýzy asociačných pravidiel používame tiež referenčnú skupinu pacientov. Aby sme totiž zistili, aké aspekty zdravotnej starostlivosti sa vyskytujú významne častejšie u pacientov s nami skúmanými epizódami, potrebujeme inú skupinu pacientov na porovnanie. Konkrétne sme ako referenciu vybrali pacientov, ktorí zaznamenali ľubovoľnú hospitalizáciu v priebehu roku 2012. Takýchto sme mali k dispozícii vyše 150000. Toto je vcelku vysoké číslo vzhľadom k počtom pacientov so skúmanými epizódami. Navyše, celá vykázaná zdravotná starostlivosť všetkých týchto pacientov by tvorila obrovské množstvo dát, problematické čo sa týka ich spracovania. Preto sme sa rozhodli použiť náhodne vybratú vzorku z týchto pacientov, presnejšie vzorku veľkosti 2581. Takto tvoria pacienti s implantáciou bedrového kĺbu približne 20% zo všetkých

4033 pacientov a pacienti s implantáciou kolenného kĺbu asi 15%.

Zámerne sme z referencie nevylučovali pacientov, ktorí by mohli mať implantácie bedrového alebo kolenného kĺbu v priebehu iných období okrem roku 2012. Máme na to tri hlavné dôvody:

1. v porovnaní so všetkými hospitalizovanými pacientmi ich je malý počet (okolo jedného percenta), teda v našom prípade je iba malá šanca, že by nejako vážnejšie mohli ovplyvniť výsledky,
2. v praxi by sme v niektorých prípadoch nemuseli mať možnosť takéto vylúčenie vykonať z dôvodu prístupnosti dát,
3. pre analýzu asociačných pravidiel toto teoreticky nepredstavuje významnú prekážku — aj ak v referenčnej vzorke máme niekoľko pacientov so štruktúrou výkonov ako pacienti s epizódou, tým že predstavujú iba časť referencie zostanú výkony charakteristické pre epizódu stále častejšie sa vyskytujúce u pacientov s epizódou.

Pre všetkých pacientov, aj pacientov so skúmanými epizódami, aj referenčných, sme získali údaje vykazanej zdravotnej starostlivosti za roky 2011 až 2013. Pre každého pacienta teda máme všetky dostupné dáta viažuce sa k zdravotnej starostlivosti aspoň jeden kalendárny rok pred a jeden rok po výkone spúšťajúcom epizódu starostlivosti. Podľa [12] totiž operácii implantácie kĺbu môže predchádzať obdobie ťažkostí kvôli ktorým treba operáciu uskutočniť. Po operácii zas môže nastať doba rehabilitácií, prípadne dodatočných komplikácií. Kvôli týmto dôvodom sme teda zvolili rovnaký čas pred aj po novej dobe vlastnej operácie. Tu vlastne narážame na otázku, akú dlhú dobu treba priradiť trvaniu epizódy, teda na bod 2 z kapitoly 2.2. To, akú dobu nakoniec priradíme epizóde pred a po jej spustení sme sa podujali osvetliť v neskorších kapitolách. Vtedy sa tiež venujeme otázke, či nami vybraný rok času ktorý máme zachytený v dátach pred a po spustení epizódy je dostatočne dlhá doba. V literatúre zaoberajúcej sa epizódami sme nenatrafili na popis epizód podobných stavov akými sa tu zaoberáme, preto nemáme dobrý odhad na potrebnú dĺžku obdobia, nevieme, do akej miery ide o chronický stav.

3.1.2 Anonymizácia dát

S dátami pracujeme v anonymizovanej forme. K zoznamu pôvodných identifikátorov všetkých pacientov boli náhodným spôsobom bez návratu priradené identifikátory, s ktorými pracujeme my. Naše identifikátory sú vo formáte „P*****”, kde hviezdičky predstavujú spomínané náhodne vybrané číslo od 1 000 000 do 9 999 999. Podobným spôsobom s vlastnými špecifikami sme anonymizovali databázu lekárov a zdravotníckych zariadení. V pôvodných dátach sa vyskytujú kódy lekárov vo formáte „pccccooo”, kde „pcccc” má formát jedného písmena a piatich číslíc a predstavuje identifikátor lekára, pričom „ooo” je kód jeho odbornosti. Tento identifikátor nahrádzame náhodne vygenerovaným identifikátorom formátu „L*****”, kde hviezdičky predstavujú podobne ako pri pacientoch náhodne vygenerované číslo v rozpätí 10 000 až 99 999. Zdravotnícke zariadenia majú kódy podľa úrovne organizácie. Hlavná zložka predstavuje celého poskytovateľa zdravotnej starostlivosti (ide v podstate o firmu v zmysle zákona, bez ohľadu na to, či pod firmou je niekoľko nemocníc alebo jediná ambulancia) a má kód vo formáte „pcccc”, ktorý je podobne ako u lekárov zložený z písmena a piatich číslíc. My každej hlavnej zložke priradíme kód formátu „Z*****”, kde hviezdičky predstavujú znova náhodné číslo medzi 10 000 a 99 999. Kód nižšej zložky, ktorá predstavuje napríklad jednotlivú ambulanciu alebo nemocničné oddelenie vznikne v pôvodných dátach spojením kódu hlavnej zložky so šiestimi ďalšími znakmi formátu „oooccc”, kde „ooo” je kód odbornosti nižšej zložky a „ccc” je číslo navzájom odlišujúce nižšie zložky rovnakej odbornosti v rámci jednej hlavnej zložky. Zhrnutie prinášame v tabuľke:

	Pôvodný formát	Anonymizovaný formát
Pacient	ccccccc	P *****
Lekár	pccccooo	L *****ooo
Hlavná zložka	pcccc	Z *****
Nižšia zložka	pccccoooccc	Z *****oooccc

Tabuľka 1: Zhrnutie anonymizácie. Hrubé písmená predstavujú skutočné písmená, *c* vyjadruje číslicu, *p* písmeno, *ooo* trojmiestny kód odbornosti a skupina hviezdičiek predstavuje náhodne generované číslo.

3.1.3 Základný popis dát

Dáta máme rozdelené na dve hlavné tabuľky, jednu obsahujúcu údaje o hospitalizáciách (tú ďalej nazývame „Data_h”) a druhú pozostávajúcu z ostatnej zdravotnej starostlivosti (ďalej ako „Data_a”). Hlavným dôvodom je okrem ďalších špecifik a prídavných premenných spojených s hospitalizáciami spôsob vykazovania hospitalizácií. Hospitalizácia na jednom oddelení presahujúca cez jeden mesiac totiž pozostáva z viacerých riadkov, jeden riadok pre každý mesiac hospitalizácie. Takéto hospitalizácie treba spájať a my sme mali možnosť dostať tabuľku hospitalizácií už v spojenej forme. Podotknúť však treba, že pri hospitalizácii môžu nastať prekľady medzi oddeleniami, tieto sú následne vykázané ako dve hospitalizácie, ktorých dátum konca jednej je zhodný s dátumom začiatku druhej hospitalizácie. Takéto hospitalizácie preto ešte spracúvame osobitne v ďalšej časti kapitoly. Tiež podľa [12] sa môžu vyskytnúť rôzne osobitne vykázané výkony vykázané ako samostatná hospitalizácia prekrývajúca sa dátumom s pôvodnou. Rovnako sa počas hospitalizácie vykazujú rôzne výkony a lieky samostatne, tie podľa ich dátumov tiež pri spracovaní dát priradujeme k prvotnej hospitalizácii.

V tabuľke 2 prinášame zoznam premenných v obidvoch tabuľkách. Nie všetky premenné sa vyskytujú v obidvoch tabuľkách, na ich mieste nájdeme prázdne miesta.

Niektoré premenné sme už zbežne popísali pri zbežnom pohľade v časti 2.1. Tu máme priestor na ich podrobnejšie rozobratie.

Diagnózy sme rozobrali v časti 2.1, pripomenieme len, že ide o významné vodítko k účelu vykonávania zdravotnej starostlivosti, ktoré však netreba preceňovať — rovnaké stavy môžu byť vykázané inak, resp. k jednému zdravotnému môže byť pridružených viac diagnóz. Pri hospitalizácii sa vykazuje aj diagnóza pri prepustení. Tá je prítomná vo všetkých výkazoch hospitalizácii a asi v 20% záznamov sa líši od diagnózy pri prijatí. Zmena diagnózy môže znamenať napríklad komplikáciu.

Dátumy máme pri niektorých výkonoch dva. Pri liekoch a zdravotníckych pomôckach ide o dátum predpisu (dátum 2) a dátum vyzdvihnutia (dátum 1). V tomto prípade je pre nás zaujímavý dátum predpisu, napr. keď chceme liek zoskupiť s inými výkonmi u rovnakého lekára v daný deň a dať tak dokopy starostlivosť tvoriacu tzv. *ambulantný kontakt*. Pri hospitalizáciách, kúpeľoch a podobnej starostlivosti ide o začiatok a koniec pobytu, pričom, ak chýba pri zázname začiatok (resp. koniec), treba ho, ako sme spo-

Data_a	Data_h	Význam
IDP	IDP	Identifikátor pacienta
DAT1	DATZ	Dátum 1 – výkonu, prevzatia lieku, začiatku hosp.
DAT2	DATK	Dátum 2 – predpisu lieku, konca hosp.
DIAG	DIAG	Kód diagnózy (5-miestny ak je k dispozícii)
	DIAGP	Kód diagnózy pri ukončení hospitalizácie
CENA	CENA	Schválená cena výkonu
TYPZS		Kód typu zdravotnej starostlivosti
LEK_V		Kód vykonávajúceho lekára
AMB_V	ODD	Kód vykonávajúcej nižšej zložky
LEK_P		Kód preskribujúceho lekára
AMB_P		Kód preskribujúcej nižšej zložky
LEK_O		Kód odporúčajúceho lekára
AMB_O		Kód odporúčajúcej nižšej zložky
ODB_V	ODB	Kód vykonávajúcej odbornosti
ODB_P		Kód preskribujúcej odbornosti
ODB_O		Kód odporúčajúcej odbornosti
PROD		Kód produktu zdravotnej starostlivosti
SKUP		Skupina a druh zdravotnej starostlivosti
PODSKUP		Kód podskupiny zdravotnej starostlivosti
ATC		Kód ATC skupiny lieku
MNOZ		Kód množstva
	PROD_H	Kód produktu z čakacej listiny hospitalizácie

Tabuľka 2: Zoznam premenných v obidvoch hlavných používaných tabuľkách dát.

mínali vyššie, hľadať v predchádzajúcom (nasledujúcom) mesiaci. Pre ostatné výkony máme typicky vykázaný iba dátum 1 a ten aj používame.

Lekárov a nižšie zložky sledujeme vo všeobecnosti troch typov. Napríklad pri liekoch môže liečbu odporučiť špecializovaný lekár, pričom liek na základe odporúčania môže predpisovať aj všeobecný lekár. To nám dáva odporúčajúceho a preskribujúceho lekára spolu s ich ambulanciami. Vykonávajúca nižšia zložka je tomto prípade bola lekáreň, v ktorej si pacient liek vyzdvihne a políčko vykonávajúceho lekára zostava prázdne. V tabuľke 3 ukazujeme typicky príklad tejto situácie. V prvých riadkoch vidíme odporúčajúceho lekára — špecialistu aj ambulanciu s kódom odbornosti 031, ktorý podľa [13, str. 17] predstavuje odbornosť *hematológia a transfuziológia*. Preskribujúci lekár je odbornosti 020, *všeobecné lekárstvo* a vykonávajúca nižšia zložka má odbornosť 160, čiže *lekáreň*. V tabuľke si tiež môžeme všimnúť že dátum 2 je často zhodný s dátumom 1, avšak v niektorých prípadoch je skorší. To zodpovedá spomínanému významu

dátumov pri liekoch, kde dátum 2 je dátum preskripcie lieku a dátum 1 je dátumom vyzdvihnutia. Pri hospitalizácii zaznamenávame iba vykonávajúce oddelenie, na ktorom je pacient hospitalizovaný.

Pri pohľade na odbornosti prioritizujeme odbornosť nižšej zložky — ak napríklad pacienta ošetruje nejaká pohotovostná zložka, zaujíma nás táto informácia, kým ošetrojúci lekár môže mať rôzne, aj nesúvisiace zameranie. Ak nemáme k dispozícii odbornosť zložky, berieme odbornosť lekára.

IDP	DAT1	DAT2	DIAG	CENA	TYPZS	LEK_V	AMB_V	LEK_P	AMB_P
405	P1008920	04.01.11	04.01.11	I70	39.50	701	Z90983160301	L50821020	Z17092020201
415	P1008920	25.02.11	25.02.11	I70	39.50	701	Z90983160301	L50821020	Z17092020201
433	P1008920	14.04.11	14.04.11	I70	19.75	701	Z90983160301	L50821020	Z17092020201
434	P1008920	19.04.11	14.04.11	K76	10.67	701	Z90983160301	L50821020	Z17092020201
449	P1008920	26.05.11	26.05.11	I87	19.75	701	Z90983160301	L50821020	Z17092020201
461	P1008920	24.06.11	23.06.11	I87	39.50	701	Z90983160301	L50821020	Z17092020201

LEK_O	AMB_O	ODB_V	ODB_P	ODB_O	PROD	SKUP	PODSKUP	ATC	MNOZ
405	L32497031	Z67654031201	160	20	31	C96118	LC	LC99 B01AB11	2
415	L32497031	Z67654031201	160	20	31	C96118	LC	LC99 B01AB11	2
433	L32497031	Z67654031201	160	20	31	C96118	LC	LC99 B01AB11	1
434	L25882048	Z63682048201	160	20	48	C84370	LC	LC99 A05BA03	1
449	L32497031	Z67654031201	160	20	31	C96118	LC	LC99 B01AB11	1
461	L32497031	Z67654031201	160	20	31	C96118	LC	LC99 B01AB11	2

Tabuľka 3: Výsek dát zobrazujúci typický výskyt odporúčajúceho lekára.

Typ zdravotnej starostlivosti je trojmiestny kód, ktorý je podľa [12] nápomocný k určeniu základného typu zdravotnej starostlivosti, teda rozlíšeniu napríklad medzi starostlivosťou všeobecného či špecializovaného lekára, laboratórnymi vyšetreniami, liekmi či ďalšími, menej častými typmi, medzi ktorými môže byť napríklad služba prvej pomoci, domáca ošetrovateľská starostlivosť, či ďalšie. Na tento ukazovateľ sa pozeráme skôr pri dodatočných analýzach dát.

Pre produkt zdravotnej starostlivosti máme vykázaný jeho kód spolu s ďalším rozdelením do podskupiny a ATC skupiny. ATC skupina predstavuje tzv. *anatomicko-terapeuticko-chemickú skupinu*, [15]. ATC skupina je medzinárodným štandardom pre klasifikáciu liečiv a identifikuje účinnú látku v lieku. Preto namiesto jednotlivých produktov sa v prípade liekov môžeme zaoberať ATC skupinou, kde pod jedným kódom máme zoskupené lieky s rovnakou účinnou látkou. Medzi produktmi s rovnakou ATC skupinou pre naše účely nemá zmysel rozlišovať, preto pri liekoch namiesto kódu produktu používame kód ATC skupiny. Pre ostatné produkty zas používame kódy podskupiny ich zaradenia. Existujú však produkty, ktoré nemajú priradenú podskupinu, pri

nich používame kód produktu. Pre produkty pri ktorých máme priradenú podskupinu (a nejde o lieky), tá vo väčšine prípadov dostatočne presne identifikuje typ výkonu bez zbytočného rozlišovania napr. medzi výrobcom zdravotníckej pomôcky, kde analogické pomôcky od rôznych výrobcov sú produkty pod rôznymi kódmi.

Pozerať tiež môžeme na druh a skupinu produktu ZS (tie tvoria v dátach jednu premennú, ktorej ďalej pre skrátenie hovoríme skupina). Skupina často (ale nie vždy) tvorí prvé dva znaky kódu podskupiny ZS. Oproti podskupine má skupina ZS tú výhodu, že vykázaná je prakticky vždy. Z 1 708 990 záznamov s nenulovou cenou sú iba 4 záznamy bez priradenia skupiny ZS, pričom to je zjavne chyba, keďže produkty z týchto záznamov majú v ostatných prípadoch skupinu priradenú vždy. Naproti tomu, záznamov bez vykázananej podskupiny s nenulovou cenou máme v dátach 34 890. Skupina je navyše v niektorých prípadoch vcelku informatívna sama o sebe — pre kód *HK*, teda skupinu *Kúpeľná liečba* sme zaznamenali v dátach iba dva kódy produktov, „KUPEL1” a „KUPEL2”, pričom rozdiel medzi nimi je podľa [12] iba z pohľadu vykazovania.

Kód produktu z čakacej listiny spomínaný v kapitole 2.1 máme pri hospitalizáciách poznamenaný osobitne, keďže z hľadiska pôvodu dát ide o úplne iný zdroj ako produkt ZS (aj keď všetky hospitalizácie majú rovnaký kód produktu a ten preto v dátach nemáme). Ako sme spomínali, na základe tohto kódu identifikujeme hospitalizácie, ktoré sú spúšťačmi našich epizód.

V tabuľkách 4 a 5 prinášame prehľad používaných premenných aj s ich dátovými typmi a príkladmi hodnôt pre obidve dátové tabuľky.

3.1.4 Čistenie dát

Prvotné dáta ktoré sme dostali k dispozícii obsahujú chyby. Stručne tu opíšeme kroky, ktoré používame na odstránenie najviac očividných chýb.

Pre zjednodušenie práce s dátumami zavádzame premenné *PocD*, ktoré ďalej používame namiesto dátumov. Tieto premenné nahrádzajú dátum počtom dní od 31. 12. 2010, teda prvý deň ktorý máme v dátach, 1. 1. 2011 má hodnotu $PocD = 1$. Vytvoríme teda premenné *PocD1* a *PocD2*, z premenných *DAT1* a *DAT2* pre tabuľku *Data_a* a premenné *PocDZ* a *PocDK* pre *DATZ*, resp. *DATK* z tabuľky *Data_h*. V tabuľke 6 ponúkame stručné zhrnutie hodnôt *PocD* pre používané dátumy. Rok 2012

```
'data.frame':      1875076 obs. of  26 variables:
 $ IDP      : chr  "P1006726" "P1006726" "P1006726" "P1006726" ...
 $ DAT1     : chr  "09.05.11" "09.05.11" "09.05.11" "16.05.11" ...
 $ DAT2     : chr  "09.05.11" "09.05.11" "" "16.05.11" ...
 $ DIAG     : chr  "J03" "J03" "J03.9" "J45.8" ...
 $ CENA     : num  2.64 1.45 0 46.1 17.39 ...
 $ TYPZS    : chr  "701" "701" "102" "701" ...
 $ LEK_V    : chr  "" "" "L89353008" "" ...
 $ AMB_V    : chr  "Z83891160301" "Z83891160301" "Z36480008201" "Z98769160301" ...
 $ LEK_P    : chr  "L89353008" "L89353008" "" "L37448003" ...
 $ AMB_P    : chr  "Z36480008201" "Z36480008201" "" "Z61406003201" ...
 $ LEK_O    : chr  "" "" "" "" ...
 $ AMB_O    : chr  "" "" "" "" ...
 $ ODB_V    : chr  "160" "160" "8" "160" ...
 $ ODB_P    : chr  "8" "8" NA "3" ...
 $ ODB_O    : chr  NA NA NA NA ...
 $ PROD     : chr  "C45997" "C09961" "8" "C45958" ...
 $ SKUP     : chr  "LC" "LC" "VV" "LC" ...
 $ PODSKUP : chr  "LC99" "LC99" "VV01" "LC99" ...
 $ ATC      : chr  "J01CE02" "R03CC02" "" "R03AK06" ...
 $ MNOZ     : num  1 1 1 1 1 1 1 1 1 1 ...
```

Tabuľka 4: Štruktúra tabuľky *Data_a*.

```
'data.frame':      12415 obs. of  13 variables:
 $ IDP      : chr  "P1006726" "P1006726" "P1007758" "P1007758" ...
 $ DATZ     : chr  "31.07.12" "31.07.12" "28.07.12" "18.05.13" ...
 $ DATK     : chr  "05.08.12" "05.08.12" "31.07.12" "20.05.13" ...
 $ DIAG     : chr  "080.0" "Z38.0" "X23.0" "W57.4" ...
 $ DIAGP    : chr  "080.0" "Z38.0" "X23.0" "W57.4" ...
 $ CENA     : num  442 347 450 438 530 ...
 $ PROD_H   : chr  "" "" "" "" ...
 $ ODD      : chr  "Z93391009101" "Z93391051101" "Z67972007101" "Z67972007101" ...
 $ ODB      : chr  "9" "51" "7" "7" ...
```

Tabuľka 5: Štruktúra tabuľky *Data_h*.

bol priestupný, má teda 366 dní.

Dátum	PocD
1. 1. 2011	1
1. 1. 2012	366
31. 1. 2012	731
31. 1. 2013	1096

Tabuľka 6: Konverzia niektorých najdôležitejších dátumov na používaný formát.

Najprv sa venujeme dátam o hospitalizáciách, *Data_h*. Odstránili sme 43 záznamov, ktoré nemali ani dátum začiatku, ani dátum konca. Ide zjavne o chybné záznamy ktoré nenesú žiadnu ďalšiu informáciu, keďže všetky majú nulové ceny a vždy sa tiež nájde korektne vykázaná hospitalizácia s rovnakou diagnózou a ďalšími prítomnými údajmi, teda ich vylúčime.

Ďalej sme našli 90 pacientov, ktorí majú rôzny počet hospitalizácii s chýbajúcim

IDP	DATZ	DATK	DIAG	DIAGP	CENA	PROD_H	ODD	ODB	PocDZ	PocDK
P9293776	29.02.12	04.09.12	A16.2	A16.2	560		Z76125003116	3	425	613
P9293776	24.07.13	25.07.13	G40.9	G40.9	0		Z20085201302	201	936	937
P9293776	25.07.13	26.07.13	G40.9	G40.9	1078		Z20085004102	4	937	938
P9293776	26.07.13	14.08.13	F10.3	F03	1190		Z20085005102	5	938	957
P9293776			A16.2	A16.2	0		Z76125003116	3	NA	NA

Tabuľka 7: Príklad pacienta s chýbajúcimi obidvoma dátumami pri hospitalizácii. Piaty riadok je očividne chybou po hospitalizácii z prvého riadku a môžeme ho vylúčiť.

dátumom začiatku a konca. Keď sú u pacienta obidve čísla rovnaké, existuje možnosť, že ide o hospitalizácie prekračujúce mesiac, ktoré kvôli chybe neboli spojené (hospitalizácie rozdelené do viacerých mesiacoch popisujeme v časti 3.1.3). Keď sú však čísla rôzne, ide určite o nejakú chybu. Manuálne sme prešli hospitalizácie spomínaných 90 pacientov a pre každú hospitalizáciu bez začiatku alebo konca sme vykonali jeden z krokov:

- Ak nájdeme hospitalizáciu bez konca a v ďalšom mesiaci bez začiatku, tieto dve spojíme,
- Ak nájdeme hospitalizáciu bez konca v poslednom mesiaci v dátach alebo bez začiatku v prvom mesiaci, začiatok, resp. koniec nastavíme na začiatok, resp. koniec príslušného mesiaca,
- Ak nájdeme hospitalizáciu bez začiatku/konca mimo a) a b) tak, že sa dá pripojiť k inej hospitalizácii v rámci krátkej doby, pripojíme ju,
- Ak nájdeme hospitalizáciu mimo c) nenulovej ceny, určíme začiatok/koniec na začiatok/koniec príslušného mesiaca,
- Ak nájdeme hospitalizáciu mimo d), teda nulovej ceny a nepripojiteľnú k žiadnej inej, vymažeme ju.

Snažili sme sa týmto čo najlepšie využiť informácie ktoré máme a v čo najväčšej miere opraviť chybné záznamy. Prioritou je ale, aby sme mali vyplnené dátumy pri všetkých hospitalizáciách.

Zostávajúcich 108 záznamov ktoré ešte nemajú vyplnené všetky dátumy tvorí podľa pohľadu na ich dátumy a nižšie zložky 54 dvojíc hospitalizácií presahujúcich mesiac ktoré z nejakého dôvodu neboli spojené, my ich preto spojíme. Spojenie vykonávame tak, že vo výslednom, spojenom, riadku spojíme čo najviac informácie z dvoch spájaných riadkov — v jednom z nich napr. môže chýbať diagnóza pri prepustení.

# povodne zaznamy:											
IDP	DATZ	DATK	DIAG	DIAGP	CENA	PROD_H	ODD	ODB	PocDZ	PocDK	
P9332169	30.11.12		M16	M16	0	WL0301	Z63663011101	11	700	NA	
P9332169		03.12.12	M16.0	M16.0	770	WL0301	Z63663011101	11	NA	703	
P9391650	29.01.12		O80.0		0		Z67972009101	9	394	NA	
P9391650		02.02.12	O80	O80.0	398		Z67972009101	9	NA	398	
P9601982	27.06.13		D24	D24	0		Z26527319101	319	909	NA	
P9601982		01.07.13	D24	D24	801		Z26527319102	319	NA	913	
# spojene zaznamy:											
IDP	DATZ	DATK	DIAG	DIAGP	CENA	PROD_H	ODD	ODB	PocDZ	PocDK	
P9332169	30.11.12	03.12.12	M16.0	M16.0	770	WL0301	Z63663011101	11	700	703	
P9391650	29.01.12	02.02.12	O80.0	O80.0	398		Z67972009101	9	394	398	
P9601982	27.06.13	01.07.13	D24	D24	801		Z26527319101	319	909	913	

Tabuľka 8: Príklad spájania záznamov hospitalizácií, pred a po spojení.

Výsledné hospitalizačné záznamy majú už všetky dátumy vyplnené. Aj zvyšné premenné okrem *PROD_H* majú hodnotu vo všetkých záznamoch. Premenná *PROD_H* je neprázdna len pri hospitalizácii, ktorá bola na čakacej listine, konkrétne v 1 863 z 12 415 záznamov. Premenná *CENA* je nulová v 2 045 záznamoch. Ide väčšinou o záznamy ktoré skončili prekladmi a cena za celú hospitalizáciu je vykázaná pri záznamoch po prekladoch. Ako narábame s prekladmi, o tom hovoríme neskôr.

Pri tabuľke *Data_a* je situácia jednoduchšia. Identifikujeme a spojíme záznamy ako kúpeľná starostlivosť, ktoré, podobne ako hospitalizácie, môžu začínať a končiť v inom mesiaci a vtedy sú na viacerých riadkoch. Pri spájaní sme natrafili, podobne ako pri hospitalizáciách na 12 chybných záznamov, z nich 6 sme vedeli opraviť a zanechať. Po takomto spojení zostanú 3 záznamy iných typov starostlivosti, ktoré nemá zmysel zlučovať, pričom nemajú dátum 1 a majú dátum 2. Týmto záznamom priradíme dátum 2 aj ako dátum 1. Zostali ešte 4 záznamy, ktoré nemajú žiadny dátum, tie vylúčime, keďže dátumy nemáme z kade získať.

3.1.5 Zoskupovanie hospitalizácií

Pri popise dát o hospitalizáciách sme spomínali, že pri hospitalizácii môžu nastať rôzne preklady medzi oddeleniami, prípadne aj medzi nemocnicami. Preklad môžeme identifikovať tak, že u pacienta nájdeme čiastkovú hospitalizáciu (teda jeden záznam v dátach o hospitalizáciách), ktorá začína v rovnaký deň, ako iná čiastková hospitalizácia končí. Navyše, niektoré výkony v rámci hospitalizácie môžu byť vykazované ako samostatné výkony, ktoré máme v druhej dátovej tabuľke, *Data_a*. Potrebujeme teda zlúčiť všetky

čiastkové hospitalizácie do jednej zoskupenej hospitalizácie a tiež k nej priradiť ostatné výkony počas jej trvania.

Najprv zoskupíme čiastkové hospitalizácie tak, že prejdeme hospitalizačné záznamy každého pacienta v poradí podľa dátumu začiatku a postupne zaraďujeme čiastkové hospitalizácie do spoločnej skupiny ak sa začiatok novej čiastkovej hospitalizácie vyskytuje pred najneskorším koncom čiastkovej hospitalizácie zo skupiny. Technicky vytvárame pre tabuľky *Data_h* novú premennú *Grup*, pričom každému riadku priradíme hodnotu podľa priradenej skupiny.

IDP	DATZ	DATK	DIAG	DIAGP	CENA	PROD_H	ODD	ODB	PocDZ	PocDK	Grup
P1115247	29.06.12	08.07.12	S32	S32	755.00		Z74631011101	11	546	555	132
P1115247	08.07.12	12.07.12	S32.0	S32.0	745.00		Z42733013101	13	555	559	132
P1115247	12.07.12	13.07.12	S32.0	S32.0	0.00		Z42733613101	613	559	560	132
P1115247	13.07.12	16.07.12	S32.0	S32.0	0.00		Z42733013101	13	560	563	132
P1115247	16.07.12	19.07.12	Z47.8	Z47.8	0.00		Z74631011101	11	563	566	132
P2464776	20.03.11	22.03.11	K56.7	K56.7	0.00		Z40038202102	202	79	81	1423
P2464776	20.03.11	20.03.11	K56	K56	0.00		Z40038010102	10	79	79	1423
P2464776	20.03.11	22.03.11	K56	K56	423.83		Z40038202102	202	79	81	1423
P2464776	22.03.11	23.03.11	K56	K56	0.00		Z40038010102	10	81	82	1423

Tabuľka 9: Príklad dvoch zoskupených hospitalizácií pozostávajúcich z viacerých čiastkových. V prvej (*Grup* = 132) nasledujú čiastkové hospitalizácie po sebe, v druhej (*Grup* = 1423) sa vyskytlo zduplikovanie jednej čiastkovej hospitalizácie.

Týmto sme 12 415 hospitalizačných záznamov zaradili do 8 639 skupín, pričom 6 361 spojených hospitalizácií je tvorených jediným záznamom.

Počet záznamov v skupine	1	2	3	4	5	6	7	8
Výskyt	6361	1290	588	328	44	20	6	2

Tabuľka 10: Zhrnutie počtu čiastkových hospitalizácií v zoskupenej hospitalizácii.

Ako vidíme v tabuľke 9, mohli sa vyskytnúť nejaké duplicitné záznamy. Vo všeobecnosti ale nevieme, či ide o duplicitný záznam, alebo nejakú osobitne vykázanú starostlivosť. V našom príklade je cena duplicitného záznamu nulová. Ukazuje sa, že pri duplicitných je často nenulová cena iba v jednom zázname. Preto duplicitné záznamy nevyklúčujeme a ak je cena nenulová, považujeme ich za osobitne vykázanú starostlivosť. Nás v ďalších krokoch zaujíma, či a prípadne v akej časti nákladov sa vyskytujú jednotlivé diagnózy či odbornosti v zoskupenej hospitalizácii. Preto ponechanie riadku v zoskupenej hospitalizácii s rovnakou diagnózou, nižšou zložkou, atď. a nulovou cenou naše ďalšie úvahy neovplyvní.

V druhom kroku vytvoríme premennú *Grup* pre tabuľku *Data_a* a tie výkony, ktoré podľa dátumu spadajú do trvania zoskupenej hospitalizácie u pacienta, dostanú priradený kód tejto skupiny. Tu je dôležité, podľa akého dátumu toto priradenie uskutočnime. Pre lieky či zdravotnícke pomôcky je totiž podstatný dátum predpisu. Niektoré lieky však môžu byť podávané priamo lekárom a pri tých zas chceme uvažovať dátum výkonu. Z tohto dôvodu vytvoríme pre *Data_a* novú premennú *typ*, ktorá má hodnotu „L” ak je vykonávajúcou zložkou lekáreň alebo výdajňa zdravotníckych pomôcok a inak má hodnotu „V”. Kúpeľnú starostlivosť ale k hospitalizáciám nezoskupujeme, jej priradíme typ „H”. Typ „L” teda priraďujeme na základe dátumu predpisu, typ „V” na základe dátumu výkonu.

3.1.6 Zoskupovanie ambulantných kontaktov

V tabuľke ostatnej starostlivosti sme v časti 3.1.5 priradili 430 593 výkonov z celkových 1 875 076 k hospitalizáciám. Zvyšok výkonov typov „V” a „L” chceme zoskupiť do ambulantných kontaktov, teda do skupín s rovnakým pacientom, dátumom a ambulanciou. Pre každý takýto ambulantný kontakt vytvoríme unikátny kód, ktorý priradíme do existujúcej premennej *Grup*. Tá bola doteraz prázdna pre výkony nepriradené k hospitalizáciám. Kód je číselný podobne ako pri zoskupených hospitalizáciách, číslujeme ale do 100 001. Pre kúpeľnú starostlivosť vytvoríme vlastné skupiny číslované od 990 000. Celkovo takto vzniklo 590 027 ambulantných kontaktov a 1 602 skupín pre kúpeľnú starostlivosť.

3.1.7 Ďalšie informácie pre zoskupené hospitalizácie a kontakty

V častiach 3.1.5 a 3.1.6 sme každému riadku v dátach priradili hodnotu nejakej skupiny, do ktorej patrí podľa nejakého spôsobu uvažovania. Pre každú skupinu nás ďalej bude zaujímať:

1. ktorého pacienta je to skupina,
2. aké sú na skupinu náklady,
3. aké sú náklady skupiny na každú diagnózu,
4. aké sú náklady skupiny na každú odbornosť,
5. aké sú náklady skupiny na každý produkt ZS,

6. aké sú náklady skupiny na každú skupinu ZS.

Vytvoríme preto pomocné premenné, napríklad pre prípad 3. maticu s riadkami zodpovedajúcim jednotlivým skupinám, stĺpcami zodpovedajúcim diagnózam, kde hodnoty matice sú náklady pre príslušnú skupinu a príslušnú diagnózu.

Na základe kapitol 2.1 a 3.1.3 používame údaje nasledovne:

3. diagnózy berieme trojmiestne, teda z päťmiestnej diagnózy použijeme prvé 3 znaky,
4. primárne berieme odbornosť odporúčajúcej nižšej zložky; inak pri type „L” odbornosť preskribujúceho lekára a pri „V” odbornosť vykonávajúceho lekára; pri hospitalizáciách berieme odbornosť vykonávajúceho lekára,
5. ako produkt používame pri liekoch ATC skupinu a pri zvyšku podskupinu produktu ZS; ak tá nie je k dispozícii, tak kód produktu.

3.2 Asociačné pravidlá pre epizódu implantácie bedrového kĺbu

V tejto podkapitole sa venujeme konkrétnemu postupu, ako vytvárame asociačné pravidlá na základe časti 2.4 pre naše dáta popísané v časti 3.1. Podrobne popisujeme postup spolu s pohľadom na priebežné výsledky, na základe ktorých určujeme rozličné potrebné parametre. Všeobecný postup je nasledovný:

1. určíme, aký pohľad na dáta ideme použiť — aké množiny položiek, aké transakcie,
2. vytvoríme príslušnú množinu asociačných pravidiel,
3. pohľadom na miery pravidiel z vytvorenej množiny navrhujeme spôsob pridelovania bodov pre danú množinu pravidiel.

V tejto časti sa venujeme epizóde implantácie bedrového kĺbu, ktorú máme v dátach pod kódom „WL0301”. Ako množinu pacientov s ktorými pracujeme používame v časti 3.1.1 opísanú množinu pacientov so skúmanou epizódou spojenú s referenčnou množinou. Pre úlohy kde uvažujeme ako množinu transakcií množinu pacientov je preto $\text{supp}(\{\text{WL0301}\}) = 0.2465$, pomer pacientov s epizódou zo všetkých uvažovaných pacientov. Podľa kapitoly 3.2 je ďalej pre pravidlá tvaru $X \Rightarrow \{\text{WL0301}\}$ maximálna hodnota miery lift $\frac{1}{\text{supp}(\{\text{WL0301}\})} = 4.0561$.

3.2.1 Asociačné pravidlá pre diagnózy

Na základe skúmania samotných diagnóz sme sa rozhodli udeliť každej hospitalizácii, ambulantnému kontaktu alebo kúpeľnému pobytu (ďalej skrátene *skupine výkonov*) maximálne 4 body. Popíšeme teda postupne 4 spôsoby priradenia bodov pozerajúce sa na diagnózy, pričom na základe každého z nich môže skupina výkonov dostať maximálne 1 bod.

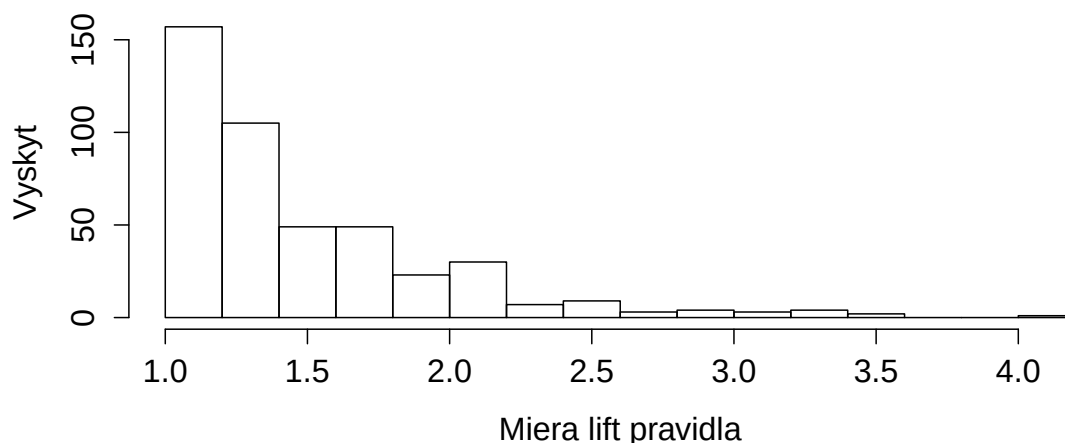
Diagnózy I. Ako základný pohľad na výskyt diagnóz u pacientov sa pozrieme, ktoré diagnózy sa vyskytujú u pacientov so skúmanou epizódou častejšie ako u referenčnej skupiny. Každému pacientovi z množiny s ktorou pracujeme priradíme všetky kódy diagnóz, ktoré sme pre neho v našich dátach zaznamenali. Pridáme ešte položku „WL0301” pacientom so skúmanou epizódou. Takto dostávame databázu, nad ktorou môžeme hľadať asociačné pravidlá.

Pre hľadanie pravidiel potrebujeme určiť minimálne hodnoty mier support a confidence pre aké sa budú pravidlá hľadať. Vo finálnom výbere pre nás relevantných pravidiel samozrejme môžeme tieto kritéria ľubovoľne sprísniť. Pre minimum miery confidence zvolíme 0.24. Keďže $\text{supp}(\{\text{WL0301}\}) = 0,2465$, pre $\text{conf}(X \Rightarrow \{\text{WL0301}\}) = 0,2465$ je miera lift tohto pravidla rovná 1. Teda voľbou minima confidence ako 0,24 určite neprídeme o žiadne pravidlá v hľadanom tvare s mierou lift nad 1.

Ako minimum miery support určíme vcelku arbitrárne hodnotu 0,001. Znamená to, že vo výsledných pravidlách nemáme diagnózy, ktoré sa vyskytujú u menej ako 0,1% pacientov v našej databáze. Také diagnózy považujeme za príliš raritné na to, aby sme sa nimi zaoberali.

S týmito podmienkami sme získali 446 pravidiel v tvare $D \Rightarrow \{\text{WL0301}\}$, kde D je množina položiek obsahujúca jednu diagnózu. Oproti celkovému počtu unikátnych kódov diagnóz, 1543, je to vcelku vysoké číslo a určite ho chceme znížiť dodatočnou podmienkou pre mieru lift. Pozrieme sa preto na výskyt rôznych hodnôt miery lift v nájdených pravidlách (obr. 3).

Stanovením spondej hodnoty lift na 1,5 sa počet nájdených pravidiel zmenší na 160. Rozhodneme sa tieto pravidlá akceptovať s tým, že body udelíme podľa pásiem. Celý jeden bod dostanú diagnózy s mierou lift nad 2,5, 0,7 bodu dostanú diagnózy s mierou



Obr. 3: Histogram mier lift v nájdených pravidlách pre diagnózy.

lift medzi 2 a 2,5 a nakoniec diagnózy pre ktoré je lift v pásme medze 1,5 a 2,5 dostanú 0,4 bodu.

lhs	rhs	confidence	lhs.support	lift
{D=M16}	=> {WL0301}	0.7040472	0.349337261	2.855723
{D=Z96}	=> {WL0301}	0.8114754	0.143740795	3.291468
{D=T84}	=> {WL0301}	0.8076923	0.015316642	3.276124
{D=S73}	=> {WL0301}	0.6279070	0.012665685	2.546887
{D=M18}	=> {WL0301}	0.6500000	0.011782032	2.636499
{D=S71}	=> {WL0301}	0.6190476	0.012371134	2.510952

Tabuľka 11: Najčastejšie sa vyskytujúce diagnózy s mierou lift nad 2,5 pre pravidlá Diagnózy I. Kódy znamenajú: M16 — artróza bedrového kĺbu, Z96 — prítomnosť iných funkčných implantátov, T84 — komplikácie vnútorných ortopedických protetických pomôcok, implantátov a štepov, S73 — vyklíbenie, vyvrtnutie a natiahnutie kĺbu a väzov bedier, M18 — artróza prvého karpometakarpálneho kĺbu, S71 — otvorená rana bedier a stehna.

Body pridelieme všetkým zoskupeným hospitalizáciám, kúpeľným pobytom a ambulantným kontaktom. Body pridelujeme na základe pomeru nákladov jednotlivých diagnóz vykázaných pre danú skupinu výkonov. Napríklad, ak 80% nákladov hospitalizácie je vykázaných na diagnózu, ktorá je bodovaná 1 bodom a 20% jej nákladov je na diagnózu ktorá získala 0,7 bodov, výsledný počet bodov z tejto časti pre hospitalizáciu je $0,8 \cdot 1 + 0,2 \cdot 0,7 = 0,94$.

Diagnózy II. V predchádzajúcej časti sme sa pozerali len na to, či pacient danú diagnózu v nami skúmanom období zaznamenal, alebo nie. V tejto časti sa pozeráme na to, akú časť pacientových nákladov tvoria jednotlivé diagnózy. Pre potreby vytvárania asociačných pravidiel vytvoríme pre každú diagnózu 4 pásma definované hranicami 5%, 10%, 20% a 50%. Ak pre pacienta predstavujú náklady na danú diagnózu z celkových

nákladov aspoň takú časť ako je hranica pásma, priradíme mu položku — kombináciu pásma-diagnóza.

Množinou transakcií pre vytváranie asociačných pravidiel je teda rovnaká množina pacientov, položky ale predstavujú kombinácie pásma $\times$ diagnóza (plus položka príslušnosti k skúmanej epizóde). Pravidlá vytvoríme, rovnako ako v minulej časti, pre minimálne hodnoty confidence 0,24 a support 0,001. Interpretovať pravidlá s vysokou mierou lift môžeme nasledovne: pacienti so skúmanou epizódou majú o dosť častejšie vysoký podiel nákladov na danú diagnózu ako referenčná skupina.

lhs	rhs	confidence	lhs.support	lift
{p05-D=M17} =>	{WL0301}	0.4256757	0.043593520	1.726606
{p05-D=Z96} =>	{WL0301}	0.9341317	0.049189985	3.788981
{p50-D=M16} =>	{WL0301}	0.9399684	0.186450663	3.812656
{p20-D=M16} =>	{WL0301}	0.9328622	0.250073638	3.783832
{p10-D=M16} =>	{WL0301}	0.9279176	0.257437408	3.763776
{p05-D=M16} =>	{WL0301}	0.9185268	0.263917526	3.725685

Tabuľka 12: Časť nájdených pravidiel Diagnózy II. Názov položky obsahuje „p***“, kde hviezdíčky predstavujú hranicu príslušného pásma nákladov v percentách, o ktorom pravidlo hovorí.

Počet pravidiel s mierou lift nad 1,5 je 54. Každá diagnóza sa môže nachádzať až v štyroch pravidlách, deje sa to ale iba pre diagnózu M16, artróza bedrového kĺbu. Diagnózy obodujeme tak, že priradíme 0,25 bodu za každé pravidlo s mierou lift nad 1,5, v ktorom sa diagnóza nachádza.

Diagnózy III. V predchádzajúcej časti sme sa pozerali na podiel ceny diagnózy na celkových nákladoch pacienta. Podobne môžeme skúmať jednoduchý počet výkonov s danou diagnózou. Vytvoríme podobné pásma s hranicami 5%, 10%, 20% a 50%, pričom do pásma zaradíme diagnózy, ktoré predstavujú aspoň daný podiel počtu výkonov u pacienta. Kombináciu pásma-diagnóza spĺňajúcu hranicu podielu počtu výkonov priradíme pacientovi ako položku pre analýzu asociačných pravidiel.

Ďalej je úloha veľmi podobná minulej časti. Pre mieru lift nad 1,5 existuje 73 pravidiel. Každá diagnóza zas môže byť súčasťou štyroch pravidiel, znova sa vo všetkých pravidlách vyskytuje iba diagnóza M16. Body priradíme podobne ako v predchádzajúcej časti, 0,25 za každé nájdené pravidlo obsahujúce diagnózu.

Diagnózy IV. Posledný bod za diagnózu sme sa rozhodli priradiť každému pacientovi individuálne. Konkrétne zistíme pre každého pacienta množinu diagnóz, ktoré boli

prítomné pri jeho zoskupenej hospitalizácii na základe ktorej bol zaradený k epizóde. Každá skupina výkonov pacienta obsahujúca niektorú z takto zistených diagnóz dostane pridelený celý bod.

D=D62	D=I26	D=I95	D=I97	D=J95	D=M16	D=M17	D=M87	D=T84	D=Z96	D=Z98	D=Z99
6	3	27	21	3	828	56	3	9	81	4	7

Tabuľka 13: Diagnózy vyskytujúce sa v zoskupených hospitalizáciách na základe ktorých boli pacienti priradení k epizóde. Najčastejšie sa vyskytujúce sú M16 — artróza bedrového kĺbu, Z96 — prítomnosť iných funkčných implantátov a M17 — artróza kolenného kĺbu.

3.2.2 Asociačné pravidlá pre odbornosti

Z toho dôvodu, že diagnózy považujeme za o niečo lepšie vodítko k tomu, na čo bol určený daný výkon zdravotnej starostlivosti, na základe odborností sme sa rozhodli priradiť menší počet bodov ako za diagnózy. Odbornosť nižšej zložky ale môže v súvislosti s inými informáciami stále byť významnou vedomosťou, preto za odbornosť priradíme skupinám výkonov 3 body. V tejto časti používame odbornosti také, ako je popísané v časti 3.1.7.

Odbornosti I. Základné pravidlá pre odbornosti vytvárame rovnako ako základné pravidlá pre diagnózy. Množina transakcií predstavuje množinu pacientov. Množina položiek je množina odborností, pričom položku priradíme tomu pacientovi, ktorého aspoň jeden výkon má danú odbornosť.

Po pohľade na vytvorené pravidlá sme stanovili spodnú hranicu miery lift na 1, 2. Odbornostiam s mierou lift medzi 1, 2 a 1, 5 pritom priradíme 0, 7 bodu a odbornostiam s mierou lift nad 1, 5 celý jeden bod, čím priradíme body celkovo 34 odbornostiam. Body pre odbornosti tiež priradíme ambulantným kontaktom (a hospitalizáciám a kúpeľom) na základe podielu nákladov danej odbornosti na nákladoch celého kontaktu.

Odbornosti II a III. Ďalšie body za odbornosti priradíme analogickým spôsobom ako pre pravidlá Diagnózy II a III. Teda je to odbornostiam, ktoré tvoria u pacientov istú časť nákladov, resp. istú časť celkového počtu výkonov, so splnenou podmienkou na mieru zaujímavosti pravidla lift. Hlavnou zmenou ktorú sme spravili je, že pre

lhs	rhs	confidence	lhs.support	lift
{0=27}	=> {WL0301}	0.4259947	0.555228277	1.727900
{0=11}	=> {WL0301}	0.4736842	0.514874816	1.921336
{0=183}	=> {WL0301}	0.4117647	0.380559647	1.670181
{0=25}	=> {WL0301}	0.5329341	0.344329897	2.161662
{0=218}	=> {WL0301}	0.4689266	0.156406480	1.902038
{0=45}	=> {WL0301}	0.4442105	0.139911635	1.801786

Tabuľka 14: Najčastejšie sa vyskytujúce odbornosti s mierou lift nad 1,5 pre pravidlá Odbornosti I. Kódy znamenajú: 27 — fyziatria, balneológia a liečebná rehabilitácia, 11 — ortopédia, 183 — doprava pacientov, 25 — anestéziológia a intenzívna medicína, 218 — nemocničná krvná banka, 45 — reumatológia.

podiel počtu výkonov sme hranicu najvyššej kategórie posunuli z 50% na 40%, inak by podmienku nespĺňala žiadna odbornosť. Body takto priradíme 23 kombináciám pásmodiagnóza pre Odbornosti II a 14 kombináciám v časti Odbornosti III.

3.2.3 Asociačné pravidlá pre produkty a ich skupiny

Produktom zdravotnej starostlivosti vo viacerých spôsoboch zaradenia do podskupín a ATC skupín, resp. do skupín, sme sa rozhodli prideliť podobne ako odbornostiam maximálne 3 body.

Hlavným rozdielom oproti práci s diagnózami a odbornosťami je spôsob, akým prideľujeme body kontaktom. Pri diagnózach a odbornostiach malo zmysel prideľovať body podľa časti nákladov, ktoré predstavovala napr. pre danú hospitalizáciu príslušná diagnóza alebo odbornosť — môže sa totiž ľahko stať, že celá hospitalizácia bude vykázaná s jednou diagnózou alebo na jednej odbornosti. Keď si ale predstavíme situáciu, že pri hospitalizácii máme vykázané produkty typické pre veľký počet hospitalizácií spolu s jedným produktom špecifickým pre danú hospitalizáciu súvisiacu s epizódou podľa ktorého ideme prideľovať body, tieto body chceme prideliť celej hospitalizácii. Keď teda v skupine výkonov odhalíme produkt alebo skupinu na základe ktorej prideľujeme istý počet bodov, pridelíme dané body celej skupine výkonov.

Skupiny I. Prvú sadu pravidiel pre skupiny vytvárame rovnako ako pravidlá Diagnózy I a Odbornosti I, teda položky priradujeme pacientom, ak sa skupina produktov predstavujúca položku vyskytuje aspoň raz v pacientových výkazoch. Pri hranici miery lift 1,5 sme našli 14 významných podskupín, z nich 5 má mieru lift nad 2,5. Skupinám s hodnotou lift nad 2,5 priradíme jeden bod, pravidlám s hodnotou lift medzi 1,5 a

2,5 pridáme pol bodu.

lhs	rhs	confidence	lhs.support	lift
{S=MX}	=> {WL0301}	0.6419077	0.376730486	2.603676
{S=ZK}	=> {WL0301}	0.6797853	0.329307806	2.757313
{S=RT}	=> {WL0301}	0.6487148	0.240648012	2.631286
{S=HK}	=> {WL0301}	0.7189189	0.217967599	2.916045
{S=MC}	=> {WL0301}	0.7909091	0.032400589	3.208048
{S=RB}	=> {WL0301}	0.6666667	0.001767305	2.704102

Tabuľka 15: Skupiny produktov s mierou lift nad 2,5. Kódy znamenajú: MX a MC — špeciálny zdravotnícky materiál, ZK — rehabilitačné a kompenzačné zdravotnícke pomôcky, RT — transfúzne lieky, HK — kúpeľná liečba, RB — tkanivá a bunky pri transplantáciách.

Skupiny II. Druhé pravidlá pre skupiny konštruujeme analogicky s pravidlami Diagnózy II a Odbornosti II, teda položky tvoria skupiny a pásma, pričom podstatné je prekročenie hranice pásma danou skupinou zdravotnej starostlivosti u pacienta. Hranice volíme znova 5%, 10%, 20% a 50%, pričom do všetkých pásiem s hodnotou lift pravidla nad 1.5 zas padá len jedna skupina, MX (špeciálny zdravotnícky materiál). Celkovo sme takto našli 16 pravidiel.

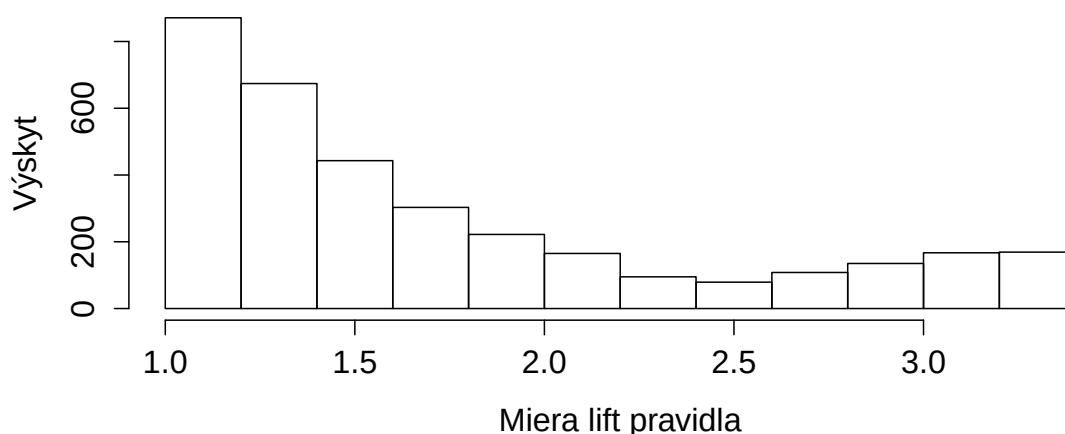
Skupiny na základe podielu počtu výkonov neskúmame. Niektoré skupiny totiž môžu obsahovať veľa lacnejších položiek (napr. lieky), niektoré zas málo drahých (špeciálny zdravotnícky materiál). Preto sa nám počet výkonov nejaví ako dobrá miera.

Produkty. Posledný bod v tejto časti udeľujeme produktom, a to na základe ich výskytu. Merať podiel počtu výkonov nechceme robiť z podobného dôvodu ako pri skupinách a merať podiel nákladov nám príde zmysluplnejšie pri skupinách z intuitívneho dôvodu, že produktov je omnoho viac druhov a možno medzi nimi môžu vznikať substitúcie.

Produkty zaraďujeme ako to je zhrnuté v časti 3.1.7 do ATC skupín pri liekoch, prípadne podskupín produktu v zvyšných dátach, pokiaľ je podskupina definovaná. Pravidlá na základe takto zadelených produktov vytvárame analogicky s pravidlami Diagnózy I. Vznikne takto 429 pravidiel s mierou lift nad 1,5 (z nich 167 s mierou lift nad 2,5). Týmto pravidlám priradíme pol bodu pre lift do 2,5 a celý bod pre lift nad 2,5.

3.2.4 Asociačné pravidlá pre kombináciu diagnózy, odbornosti a produktu

Posledný bod sme sa rozhodli priradiť na základe kombinácie diagnózy, odbornosti a produktu, avšak tentokrát nad transakciami tvorenými skupinami výkonov. Každú skupinu výkonov priradíme ako položky všetky diagnózy, odbornosti aj produkty, ktoré sú v skupine vykázané. Tiež pridáme položku „WL0301“, ak ide o skupinu výkonov pacienta so skúmanou epizódou. Hľadáme pravidlá v tvare $D \cup O \cup P \Rightarrow \{WL0301\}$, kde D , O a P sú jednoprvkové množiny položiek predstavujúce diagnózu, odbornosť, resp. produkt zdravotnej starostlivosti. Minimálnu hodnotu miery support sme znížili na 0,0001, keďže pracujeme s väčšou množinou transakcií a zaujímajú nás aj málo časté kombinácie položiek. Minimálnu hodnotu miery confidence sme stanovili na 0,05 s tým, že pravidlá s nízkou mierou lift neskôr vyfiltrujeme.



Obr. 4: Histogram mier lift v nájdených pravidlách pre kombináciu diagnóza-odbornosť-produkt.

Pri pohľade na obr. 4 nemáme veľký dôvod meniť mieru používanú hranicu miery lift 1,5, zbavíme sa pomocou nej, ako sa zdá, veľkého počtu málo zaujímavých pravidiel, pričom dosť pravidiel zostane. Všetkým kontaktom ktorých aspoň jeden výkaz spĺňa všetky podmienky na diagnózu, odbornosť aj produkt ľubovoľného zostávajúcего pravidla pridelíme celý bod. Takýchto pravidiel spĺňajúcich hranicu miery lift sa našlo celkom 1646.

lhs	rhs	confidence	lhs.support	lift
{D=M16, O=20, P=RAFLM}	=> {WL0301}	0.8387097	0.0001595048	2.842523
{D=I10, O=11, P=C07AB07}	=> {WL0301}	0.7674419	0.0001106243	2.600985
{D=M16, O=1, P=C07AB07}	=> {WL0301}	0.7222222	0.0001389235	2.447728
{D=M16, O=20, P=C07AB07}	=> {WL0301}	0.8836478	0.0008181053	2.994826
{D=M16, O=10, P=VV06}	=> {WL0301}	0.8571429	0.0001800861	2.904996
{D=M16, O=10, P=VV03}	=> {WL0301}	0.8553459	0.0004090527	2.898906
{D=M16, O=49, P=C10AA05}	=> {WL0301}	0.7291667	0.0001234876	2.471264

Tabuľka 16: Príklad nájdených pravidiel z kombinácií diagnóz, odborností a produktov. Vyskytujú sa diagnózy M16 (artróza bedrového kĺbu) a I10 (esenciálna hypertenzia), odbornosti 20 (všeobecné lekárstvo), 11 (ortopédia), 1 (vnútorné lekárstvo), 10 (chirurgia) a 49 (kardiológia). Kódy C07AB07 a C10AA05 predstavujú ATC skupiny liekov, zvyšné kódy sú podskupiny zdravotnej starostlivosti.

3.2.5 Zhrnutie a vyhodnotenie pridelenia bodov

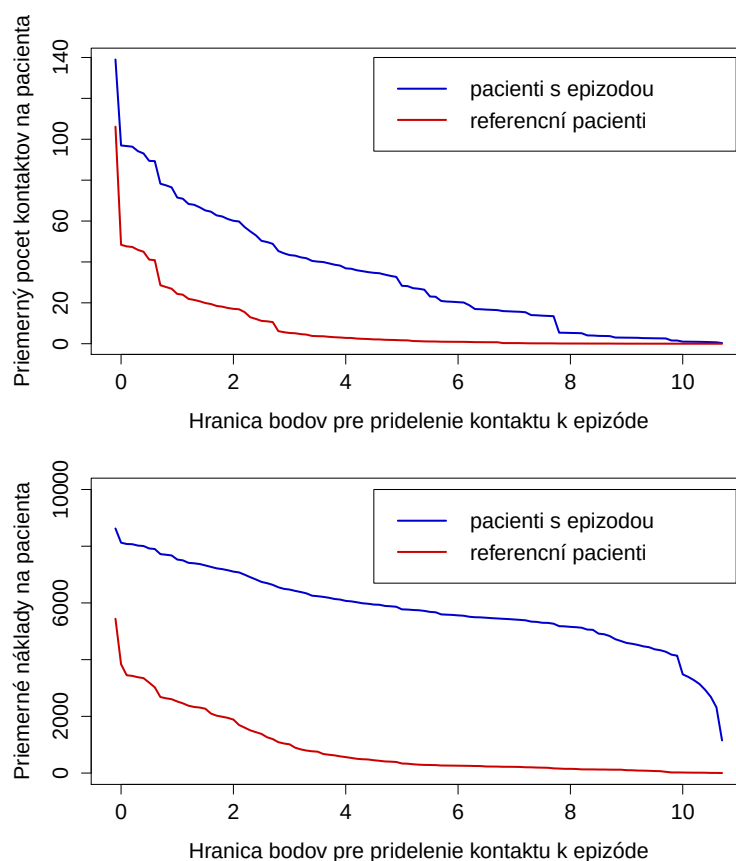
V tabuľke 17 prinášame zhrnutie spôsobov, akými sme sa v predchádzajúcich častiach rozhodli prideľovať asociačným pravidlám body.

Položky	Pravidlá založené na	Hranica lift	Bodovanie
Diagnózy	I výskyte	1,5	0,4, 0,7 a 1 bod podľa miery lift
	II podiele ceny	1,5	0,25 bodu za každé pásmo
	III podiele počtu výkonov	1,5	0,25 bodu za každé pásmo
	IV spúšťajúcej hospitalizácii	—	1 bod za splnenie podmienky
Odbornosti	I výskyte	1,2	0,7 alebo 1 bod podľa miery lift
	II podiele ceny	1,5	0,25 bodu za každé pásmo
	III podiele počtu výkonov	1,5	0,25 bodu za každé pásmo
Skupiny	I výskyte	1,5	0,5 alebo 1 bod podľa miery lift
	II podiele ceny	1,5	0,25 bodu za každé pásmo
Produkty	výskyte	1,5	0,5 alebo 1 bod podľa miery lift
Kombinácia	výskyte	1,5	1 bod za splnenie podmienky

Tabuľka 17: Zhrnutie prideľovania bodov na základe asociačných pravidiel.

Každý skupine výkonov sme teda pridelili počet bodov od 0 po 11. S narastajúcim počtom pridelených bodov pritom rastie aj naša dôvera v to, že skupina výkonov patrí k skúmanej epizóde. Chceme teraz určiť jednu hranicu počtu bodov, ktorú keď skupiny výkonov presiahnu, považujeme ich za prislúchajúce k epizóde. Tu nám znova posluží referenčná skupina pacientov — rozumná hranica počtu bodov by mala byť taká, nad akou referenčná skupina má už minimálne množstvo starostlivosti, zatiaľ čo u pacientov s epizódou je množstvo starostlivosti nad hranicou ešte vcelku veľké.

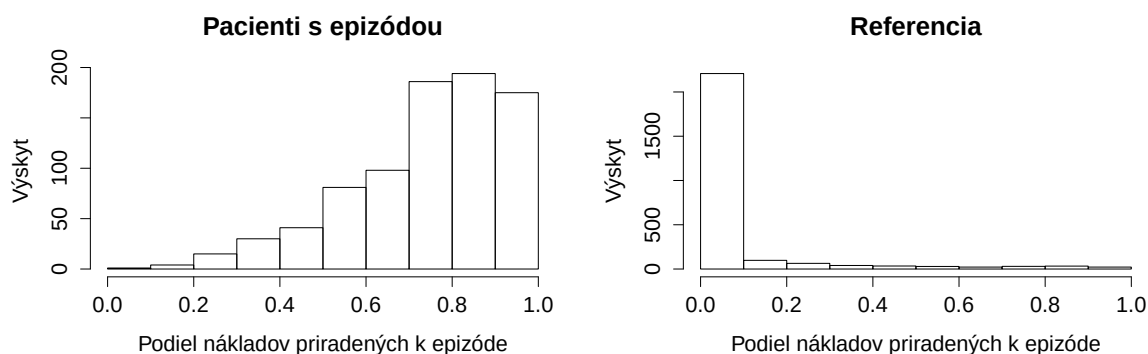
Tu sa pozrieme na obrázok 5. Na osi x je pomýšľaná hranica bodov nad ktorú považujeme skupiny výkonov za prislúchajúce k epizóde. Na y-ovej osi je k danej hranici bodov zobrazený priemerný počet priradených skupín výkonov, resp. ich cena na pa-



Obr. 5: Počet kontaktov na pacienta, resp. náklady na pacienta v kontaktoch podľa minimálneho počtu bodov ich pridelenia k epizóde.

cienta. Podľa počtu kontaktov nám príde vhodná hranica okolo 3, podľa ceny sa zdá byť vhodnou hranicou skôr číslo 5 — sú to čísla, kedy hodnoty pre referenciu začnú klesať napohľad veľmi pomaly. My sa rozhodneme pre hodnotu v strede, teda 4. Hodnota 4 má tiež dobrú interpretáciu, keďže za diagnózy pridelujeme maximálne 4 body. Aj skupina výkonov s plným počtom bodov za diagnózu ešte potrebuje aspoň nejakú časť bodu za iné premenné, aby bol počet bodov ostro väčší ako 4 a bola priradená k epizóde.

Na obr. 6 vidíme, akú časť nákladov sme priradili k epizódam u pacientov s epizódou, resp. u pacientov referenčnej skupiny. Výsledok vyzerá podľa predstáv, pre pacientov s epizódou často priradíme značnú časť nákladov, kým pre referenčnú skupinu sú náklady priradené k epizóde iba zriedkavo nad 10%.



Obr. 6: Histogramy podielu nákladov pacientov, ktoré priradíme k epizódam pri hranici bodov 4 pre pacientov s epizódou, resp. pacientov referenčnej skupiny.

3.3 Verifikácia výsledkov asociačnej analýzy

Podľa možností sa snažíme naše výsledky z predchádzajúcich častí verifikovať. Pre účely verifikácie máme možnosť vstupu lekára [8].

Pre tento účel sme vytvorili zoznam všetkých trojíc diagnóza-odbornosť-produkt vyskytujúcich sa v dátach, pričom ku každej trojici sme priradili sumu nákladov na výkony s touto kombináciou pre pacientov so skúmanou epizódou. Takto vzniklo 21 522 kombinácií. Z nich sme vybrali iba kombinácie ktoré dosiahli náklady nad 1000 Eur. Takto sa počet zostávajúcich kombinácií zmenšil na 531, pričom ale celkové náklady v zostávajúcich kombináciách predstavujú stále 79,9% nákladov všetkých kombinácií diagnóz, odborností a produktov. Ďalej máme pre účely porovnania k dispozícii priemerné počty pridelených bodov pre každú kombináciu, samostatne pre všetky štyri spôsoby priradovania bodov — diagnózy, odbornosti, produkty, aj pre kombináciu.

Zoznam 531 vybraných kombinácií (ďalej ako *skrátенý zoznam*) sme dali prehliadnúť lekárovi, ktorý nezávisle na našom bodovaní označil kombinácie ktoré môžu súvisieť s epizódou. Túto klasifikáciu ďalej berieme ako správnu a porovnávame s ňou naše výsledky.

Na skrátенý zoznam použijeme zároveň našu klasifikáciu, teda kombináciu zaradíme k epizóde na základe priemerného počtu bodov väčšieho ako 4. Túto klasifikáciu (ďalej ako pôvodná klasifikácia) môžeme porovnávať so správnou klasifikáciou.

V tabuľke 18 vidíme, koľko položiek zo skrátенého zoznamu je klasifikovaných obidvomi, iba jednou, prípadne žiadnou z porovnávaných klasifikácií. Pôvodná klasifikácia teda správne klasifikuje $280 + 185 = 465$ položiek, čiže 87,6% zo všetkých 531. Dôle-

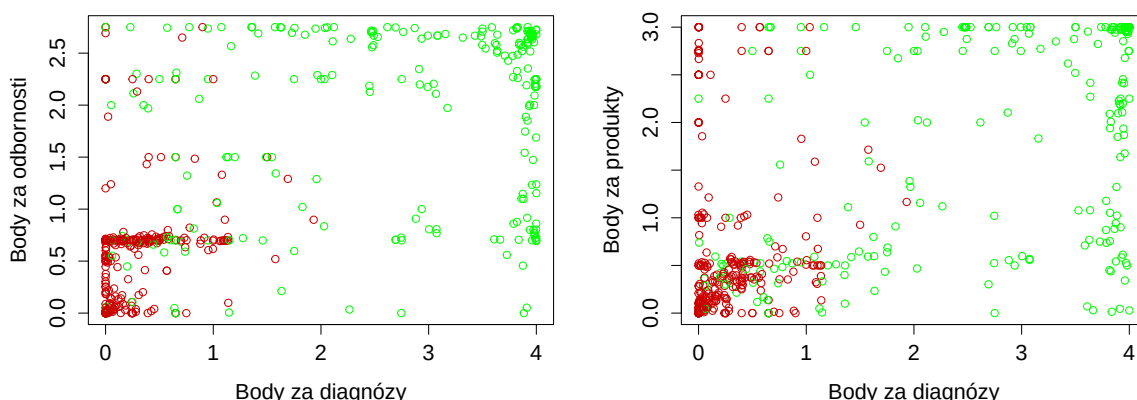
spravna	povodna	
	FALSE	TRUE
FALSE	280	22
TRUE	44	185

spravna	povodna	
	FALSE	TRUE
FALSE	795409,41	83828,40
TRUE	97444,02	3299253,11

Tabuľka 18: Porovnanie pôvodnej a správnej klasifikácie; počty kombinácií zo skráteného zoznamu, resp. náklady na príslušné kombinácie.

žitejší pre nás však je pohľad na náklady, ktorý prakticky dáva jednotlivým položkám váhy. Čo sa týka nákladov, naša klasifikácia sa zhoduje so správnou na 95,8%.

Tu môžeme rozmýšľať, o koľko lepšie sa dajú klasifikovať kombinácie zo skráteného zoznamu na základe známej skutočnej klasifikácie. Pre tento účel máme k dispozícii počty bodov udelené na základe asociačných pravidiel z diagnóz, odborností, produktov a kombinácie. Tieto počty bodov pre nás tvoria štyri premenné, pričom na základe nich sa snažíme čo najlepšie popísať správnu klasifikáciu ako vysvetľovanú premennú.



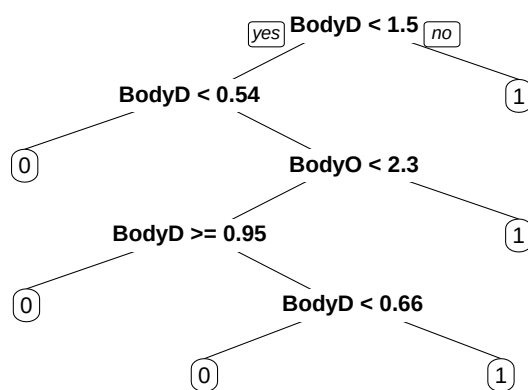
Obr. 7: Priradené body za diagnózy, odbornosti a produkty vs. správna klasifikácia ako kombinácia súvisiaca (zelený bod), resp. nesúvisiaca (červený bod) s epizódou.

Pre vytvorenie predstavy o takejto úlohe sa môžeme pozrieť napr. na obrázok 7. Vidíme body priradené kombináciám zo skráteného zoznamu, pričom zafarbenie je podľa ich správnej klasifikácie. Úlohou klasifikácie je rozdeliť priestor všetkých štyroch vysvetľujúcich premenných na dve oblasti tak, aby oblasti čo najlepšie zodpovedali vysvetľovanej klasifikácii. Naša pôvodná klasifikácia delí priestor nadrovinou $b_{diag} + b_{odb} + b_{prod} + b_{komb} = 4$.

Podľa [6] sa na hľadanie podobnej nadroviny ktorá čo najlepšie delí daný priestor používajú metódy lineárnej diskriminačnej analýzy. Na klasifikáciu akú potrebujeme sa tiež používa metóda klasifikačných stromov, ktorá vytvorí strom pravidiel, pričom

v každom vrchole stromu sa rozhodujeme na základe jednej premennej a výsledná klasifikácia je hodnota priradená listu, do ktorého sa dostaneme. My chceme zistiť, o koľko lepšie sú schopné tieto metódy klasifikovať položky na základe známej správnej klasifikácie.

Vážená lineárna diskriminačná analýza s nákladmi na položku ako váhami odhadne optimálnu nadrovinu deliacu priestor ako $1.225 \cdot b_{diag} + 0.6721 \cdot b_{odb} + 0.05274 \cdot b_{prod} + 0.2986 \cdot b_{komb} = 4$. Rozhodovací strom zostrojený nad podobne váženými bodmi je na obr. 8.



Obr. 8: Rozhodovací strom.

Môžeme si všimnúť, že výsledok diskriminačnej analýzy dáva väčšie váhy diagnózam a odbornostiam ako ostatným premenným a aj rozhodovací strom sa rozhoduje iba na základe týchto dvoch premenných — to môže nasvedčovať ich väčšiemu významu pre klasifikáciu.

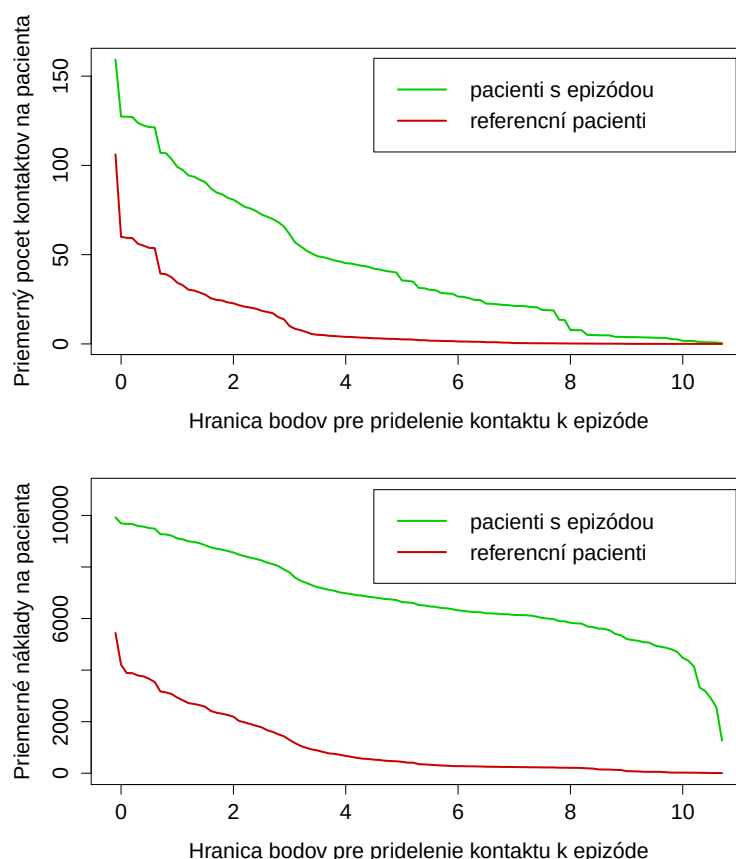
Dôležité ale je, ako sa klasifikácie na základe týchto modelov podobajú správnej klasifikácii. Výsledky sú také, že diskriminačná analýza správne zaradila 96,7% nákladov a rozhodovací strom dobre zaradil 97,3% nákladov. Toto je v porovnaní s 95,8% z pôvodnej klasifikácie celkom veľký nárast. Treba si ale uvedomiť, že zostrojený rozhodovací strom je značne „košatý“ a pre dáta pre ktoré nemáme k dispozícii správnu klasifikáciu by sme ťažko hľadali podobne jemnú štruktúru a nepovažujeme za pravdepodobné, že práve tento rozhodovací strom by bol rovnako presný pre iné dáta.

Záver je teda taký, že pôvodná klasifikácia správne klasifikuje 95,8% nákladov zo skráteného zoznamu pre ktorý máme správnu klasifikáciu, teda nákladov, ktoré predsta-

vujú asi 80% celkových nákladov. Pomocou správnej klasifikácie sme vytvorili modely ktoré ponúkli isté zlepšenie predikčnej sily modelu, máme ale pochybnosti o ich použiteľnosti na iných dátach. Používame preto pôvodný model klasifikácie. Ako vedľajší záver poukázali analýzy na väčšiu dôležitosť počtu bodov za diagnózy a odbornosti pre správnu klasifikáciu.

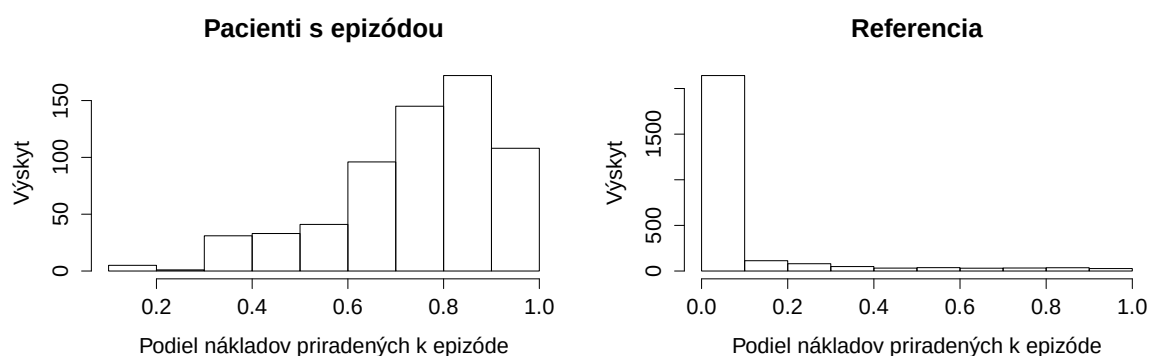
3.4 Výsledky pre epizódu implantácie kolenného kĺbu

Aby sme aspoň čiastočne overili aplikovateľnosť postupu prideľovania bodov na inú epizódu ako na tú na základe ktorej sme prideľovanie navrhovali, zopakujeme presnú procedúru prideľovania bodov pre druhú epizódu ktorú máme zachytenú v dátach, epizódu implantácie kolenného kĺbu. Použijeme rovnaké hodnoty všetkých parametrov, ako miery lift pri asociačných pravidlách a rovnaké pravidlá bodovania.



Obr. 9: Počet kontaktov na pacienta, resp. náklady na pacienta v kontaktoch podľa minimálneho počtu bodov ich pridelenia k epizóde implantácie kolenného kĺbu.

Po pridelení bodov sa pozrieme na analógie obrázkov 5 a 6 pre epizódu implantácie kolenného kĺbu, obrázky 9, resp. 10. Na základe spôsobu, akým sme určili hranicu 4 pre



Obr. 10: Histogramy podielu nákladov pacientov pre epizódu implantácie kolenného kĺbu, podiel nákladov priradených k epizóde pre pacientov s epizódou, resp. referenčných pacientov.

body z obrázku 5 by sme sa ľahko dopracovali k podobnému výsledku pre obr. 9. Tiež histogramy na obr. 10 vyzerajú dobre, keďže pacientom s epizódou znova priradujeme značnú časť nákladov a pacientom bez epizódy zriedka viac ako 10%.

Treba však poznamenať, že lepšie by sme náš postup overili na epizóde, ktorá je menej podobná pôvodne skúmanej epizóde implantácie bedrového kĺbu. Ako vidíme aj v tabuľke 12, diagnóza „artróza kolenného kĺbu“, ktorá sa ukazuje byť najčastejšou diagnózou pri implantácii kolenného kĺbu, sa vyskytuje aj v bodovaní pri epizóde implantácie bedrového kĺbu. Pohľad na koeficient korelácie medzi bodmi pridelenými pre skupiny výkonov na základe jednej a druhej epizóde naznačuje silnú koreláciu — jeho hodnota je 0,89.

Pre epizódu implantácie kolenného kĺbu sme ale nedosiahli ani protikladné výsledky, čo ďalej podporuje použiteľnosť našej metódy pridelovania bodov. Pre iné epizódy by možno bolo vhodné určiť iné hodnoty parametrov, napríklad v prípade, ak s nami používanými hodnotami vznikne v niektorom kroku extrémne veľa alebo extrémne málo asociačných pravidiel.

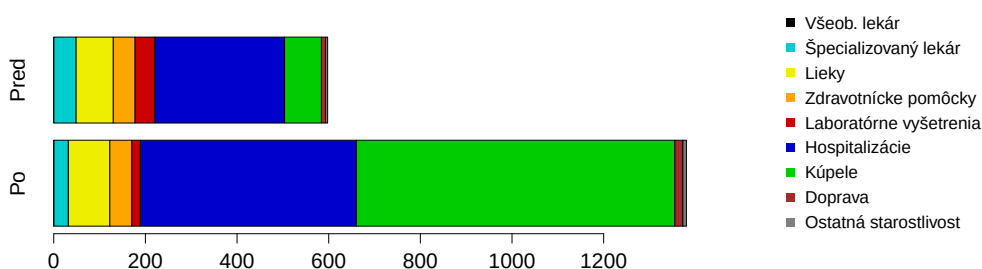
4 Ďalšie možnosti použitia výsledkov

V tejto časti naznačíme, ako by sa ďalej dalo pracovať so zoskupenými epizódami starostlivosti. Vyberáme niekoľko použití z množstva možností prevažne ako príklad. Rozsiahlejšie skúmanie je mimo obsahu tejto práce.

4.1 Základná analýza epizódy implantácie bedrového kĺbu

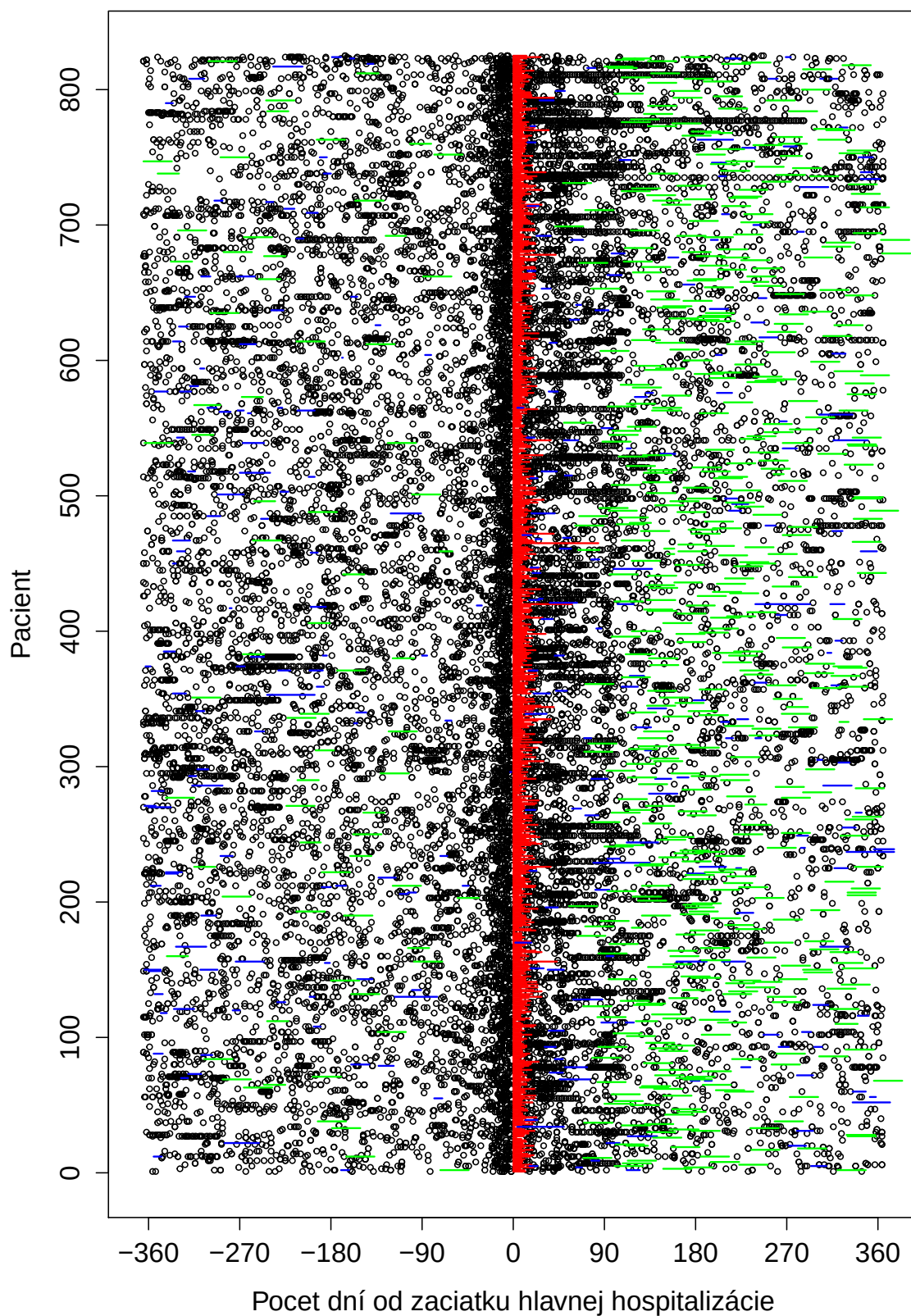
Ako základná možnosť použitia výsledkov nám pripadá pozrieť sa na vlastné zoskupené epizódy. Môžeme analyzovať ich priebeh, výkony ktoré obsahujú, náklady. Tu prinášame krátky pohľad na časť takýchto vlastností.

Doteraz sme pracovali s dátami za celé obdobie troch rokov, ktoré máme v dátach. Zámerom pre výber troch rokov dát bolo, aby sme pre každú hospitalizáciu v roku 2012 mali dáta pre aspoň jeden rok do minulosti a jeden rok do budúcnosti. Preto ďalej pracujeme iba s dátami, v ktorých dátum sa vyskytuje v rámci jedného roku, pred alebo po hlavnej hospitalizácii. Hlavnú hospitalizáciu určíme ako hospitalizáciu, na základe ktorej bola pacientovi priradená epizóda; ak zaznamenal pacient dve takéto hospitalizácie (stalo sa tak v 10 prípadoch) v priebehu sledovaného roku 2012, berieme prvú z týchto hospitalizácií. Analyzujeme tiež iba pacientov ktorým bola daná epizóda priradená a výkony iba tie, ktoré sme k epizóde priradili v predchádzajúcej kapitole.



Obr. 11: Priemerné náklady v rôznych vrstvách starostlivosti v roku pred a v roku po hlavnej hospitalizácii (bez hlavnej hospitalizácie).

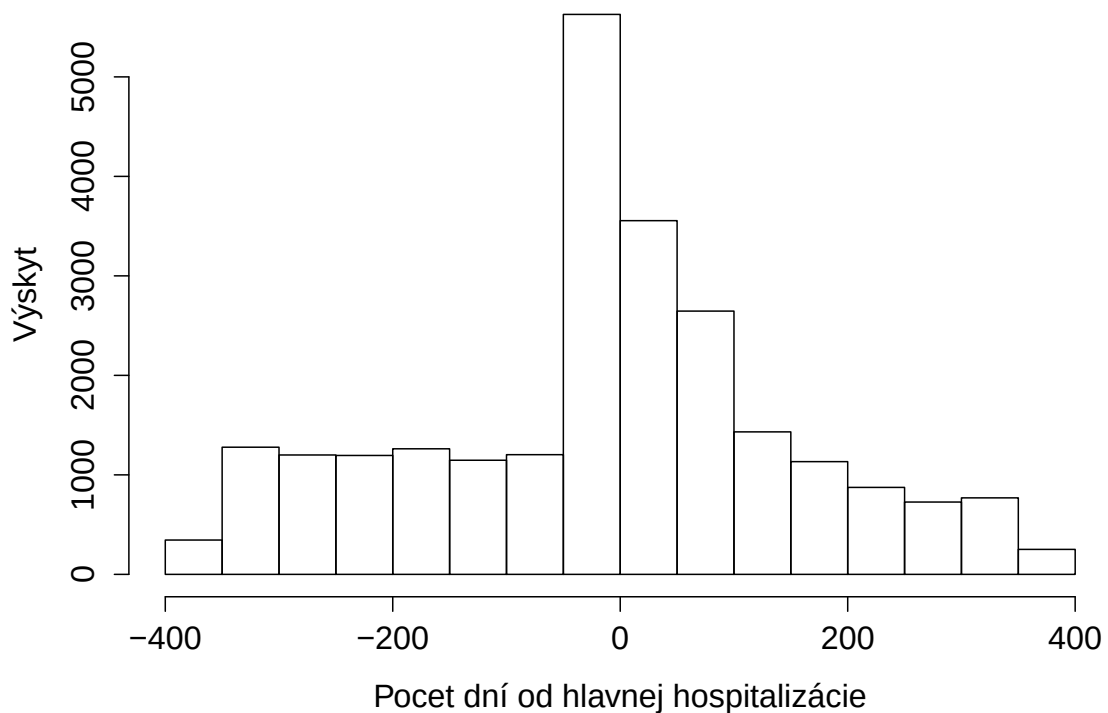
Na obr. 11 vidíme, aké priemerné náklady zaznamenali pacienti v rôznych tzv. vrstvách zdravotnej starostlivosti, počas roku pred a počas roku po hlavnej hospitalizácii. Vrstvy určíme z dát podľa premennej *typ zdravotnej starostlivosti*, prípadne



Obr. 12: Výkony priradené k epizóde jednotlivých pacientov v čase. Čierne bodky predstavujú ambulantné kontakty, červené čiary hlavnú hospitalizáciu, modré čiary iné hospitalizácie a zelené čiary kúpeľné pobyty.

premennej *skupina a druh ZS*. Na obrázku môžeme pozorovať predovšetkým očividný nárast objemu kúpeľnej starostlivosti. Ten sa dá interpretovať ako súčasť rehabilitácie v dobe po operácii [12].

Na obr. 12 vidíme starostlivosť priradenú k epizóde pre jednotlivých pacientov v čase. Už pri pohľade na obrázok môžeme vysloviť isté závery o charaktere starostlivosti poskytovanej pre skúmanú epizódu. Potvrdzuje sa nám napríklad častejšia kúpeľná starostlivosť v priebehu roka po hospitalizácii (častejší výskyt zelených čiar). Tiež by sme mohli konštatovať, že starostlivosť je najviac koncentrovaná vo v okolí hlavnej hospitalizácie (najväčšia hustota čiernych bodov). Strojovým skúmaním by sme sa mohli snažiť napríklad hľadať vzory poskytovania zdravotnej starostlivosti a identifikovať tak pacientov s podobným priebehom epizódy.



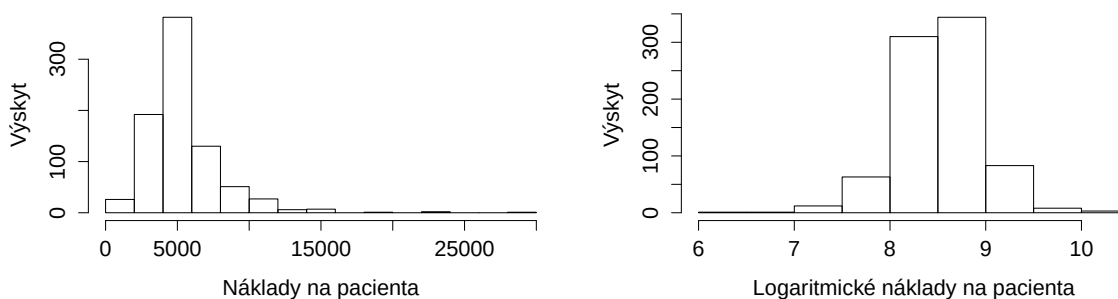
Obr. 13: Histogram počtu kontaktov podľa času od hlavnej hospitalizácie pre všetkých pacientov s epizódou.

Na obr. 13 si môžeme potvrdiť väčší výskyt ambulantných kontaktov pacientov v dobe okolo hlavnej hospitalizácie. Vidíme tiež, že počet kontaktov koncom roka po hospitalizácii je o niečo menší ako začiatkom roka pred hospitalizáciou. Toto môže svedčiť o dobe, kedy pacienti začnú pociťovať účinky operácie kĺbu a následnej rehabilitácie a vyžadujú preto menší objem zdravotnej starostlivosti.

Na základe obr. 13 tiež môžeme odpovedať na otázku 2 z kapitoly 2.3, teda aká je časové trvanie našej epizódy. Histogram nasvedčuje, že pacienti čerpajú starostlivosť súvisiacu s epizódou počas celého obdobia, ktoré máme v dátach, teda dĺžka epizódy je aspoň rok pred a rok po spúšťajúcej hospitalizácii. Preto pri hlbšom skúmaní epizódy by mohlo byť vhodné pracovať s dátami pokrývajúcimi aj väčší časový rozsah.

4.2 Závažnosť epizódy

V časti 2.3 sme tiež nechali otvorenú otázku 4, určovanie stupňa závažnosti epizódy. Na pokus o odhadnutie závažnosti epizódy môžeme ako najjednoduchšiu možnosť použiť pacientove náklady na epizódu. Pre ilustráciu postupu zostrojíme jednoduchý lineárny model. Ako vysvetľovanú premennú použijeme logaritmus nákladov na epizódu pre jednotlivých pacientov — podľa [7] je logaritmus zdravotníckych nákladov často bližší k normálnemu rozdeleniu ako samotné náklady. Nasvedčujú tomu aj histogramy z obr. 14.



Obr. 14: Histogramy pôvodných a logaritmických nákladov pacientov na epizódu.

Ako vysvetľujúce premenné použijeme dummy premenné znázorňujúce prítomnosť nákladov na niektorú z určených diagnóz a odborností u pacienta v arbitrárne určenej výške aspoň 10 Eur. Ako určené diagnózy resp. odbornosti vyberieme tie, ktoré v časti 3.2 dostali z pravidiel Diagnózy I, II, III resp. Odbornosti I, II a III dokopy aspoň 1 bod. Z pôvodných 29 premenných postupne vyradíme tie s p-hodnotou nad 0,05, až kým pridáme k finálnemu modelu (tab. 19). Priemerné náklady na pacienta sú 8,5077.

Pri serióznom skúmaní závažnosti by sme pravdepodobne použili zložitejší model aj sofistikovanejšie úvahy. Zaoberali by sme sa napríklad klasifikáciou do skupín podľa rizika vysokých nákladov, pričom tu zistené signifikantné diagnózy a odbornosti by

```

Residuals:
    Min       1Q   Median       3Q      Max
-2.2330 -0.2586  0.0009  0.2226  5.7267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
D=D62    0.51685    0.25447    2.031  0.04257 *
D=I95    0.34017    0.10412    3.267  0.00113 **
D=I97    0.56777    0.12887    4.406  1.2e-05 ***
D=M16    6.28331    0.15416   40.758 < 2e-16 ***
D=M17    0.09611    0.04298    2.236  0.02562 *
D=M45    0.72935    0.28251    2.582  0.01000 *
D=R26    0.30395    0.09924    3.063  0.00226 **
D=T84    0.27698    0.11509    2.407  0.01632 *
D=Z98    0.30951    0.15757    1.964  0.04983 *
O=11     1.41791    0.14436    9.822 < 2e-16 ***
O=27     0.77584    0.07951    9.757 < 2e-16 ***
O=611    0.21762    0.07521    2.893  0.00391 **

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5621 on 813 degrees of freedom
Multiple R-squared:  0.9957,    Adjusted R-squared:  0.9956
F-statistic: 1.572e+04 on 12 and 813 DF,  p-value: < 2.2e-16

```

Tabuľka 19: Výstup lineárnej regresie pre jednoduchý model závažnosti epizódy.

nám možno mohli slúžiť ako vodítko.

Aj samotný lineárny model by sa dal urobiť lepšie. Za zamyslenie by stálo, aké iné premenné použiť, alebo ako určiť hranicu nákladov na dummy premennú. Vylepšiť ho by tiež mohlo použitie demografických údajov o pacientov, ako napr. pohlavia a vhodných vekových skupín, či použitie interakčných premenných.

4.3 Analýza poskytovateľov starostlivosti

V časti 2.3 sme tiež položili otázku 5, otázku o priradení poskytovateľa zdravotnej starostlivosti zodpovedného za epizódu každého pacienta. Pri našich epizódach toto môžeme vyriešiť celkom jednoducho — keďže naše epizódy sú zoskupením starostlivosti viažúcej sa k hlavnej hospitalizácii, za zodpovedného poskytovateľa prehlásime nemocnicu vykonávajúcu túto hospitalizáciu. Podľa [12] môže mať totiž kvalita vykonanej operácie zásadný vplyv na ďalšiu potrebu čerpania starostlivosti. Uvádzame dva príklady analýz zameriavajúcich sa na takto pridelených poskytovateľov.

4.3.1 Identifikácia extrémne drahých poskytovateľov

Na základe priemerných nákladov na epizódu môžeme navzájom porovnávať zodpovedných poskytovateľov. V článku [7] používajú autori Monte Carlo permutačný test

na identifikovanie extrémne drahých poskytovateľov. Idea testu je nasledovná: mnohokrát vyberieme spomedzi všetkých pacientov vzorku takej veľkosti, aký je počet pacientov operovaných u skúmaného poskytovateľa. Keby sme mali zavedené skupiny závažnosti, vyberali by sme vzorky so zodpovedajúcim počtom pacientov v každej skupine závažnosti. Zo vzoriek vypočítame priemery. Ak je priemer nákladov skúmaného poskytovateľa extrémnou hodnotou v rámci priemerov zo vzoriek (napr. v najdrahších 5%), považujeme poskytovateľa za drahšieho. V [7] navyše autori porovnávajú náklady poskytovateľov s priermi vynásobenými faktorom 1,25, zisťujú teda, ktorí poskytovateľ je štatisticky význame drahší aspoň o 25% oproti priemeru.

Poskytovateľ	Počet	Priem. náklady	Poskytovateľ	Počet	Priem. náklady
Z66456	≥ 25	7661 Eur	Z32771	< 10	4278 Eur
Z67972	< 10	6792 Eur	Z80868	≥ 25	4251 Eur
Z63663	< 10	6725 Eur	Z69167	< 10	3841 Eur
Z65892	≥ 25	6364 Eur	Z74631	< 10	3687 Eur
Z20085	≥ 25	6149 Eur	Z66881	≥ 25	3666 Eur

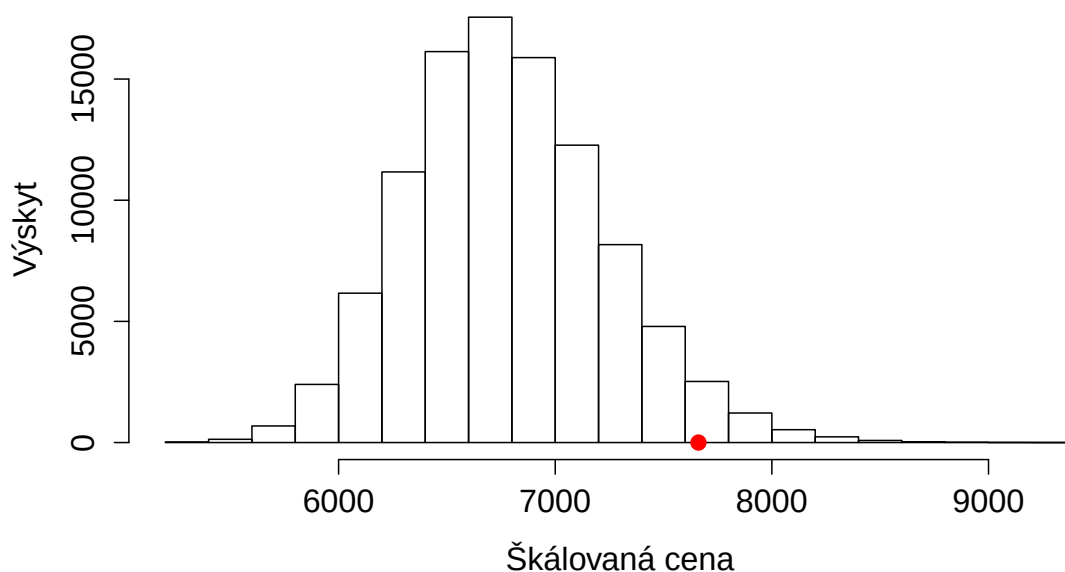
Tabuľka 20: Poskytovatelia s najväčšími a najmenšími priemernými nákladmi na epizódu.

V tabuľke 20 uvádzame niekoľko poskytovateľov s najväčšími, resp. najmenšími priemernými nákladmi na epizódu. Pre zachovanie anonymity údajov udávame iba interval v akom leží počet skúmaných pacientov s hlavnou operáciou u daného poskytovateľa. Celkom výrazne sa javí ako najdrahší poskytovateľ s kódom Z66456. Aplikujeme na neho preto popisovaný permutačný test.

Histogram testu pre 100000 vzoriek je na obr. 15. Iba 3,71% priemerov vzoriek je väčších, teda na hladine významnosti 5% môžeme zamietnuť hypotézu, že skúmaný poskytovateľ nie je aspoň o 25% drahší ako priemer.

4.3.2 Vytváranie zhlukov poskytovateľov

Ďalej nás môže zaujímať, ako by sa dali poskytovatelia zoskupiť do zhlukov na základe vzájomnej podobnosti v istých vlastnostiach. My sme sa pre demonštráciu rozhodli skúmať náklady poskytovateľov po vrstvách nákladov tak, ako boli používané aj v časti 4.1. Na zhlukovanie použijeme hierarchickú analýzu zhlukov, ktorá postupne zlučuje najpodobnejšie prvky, až pokým zlúči všetky prvky a tým vytvorí stromovú štruktúru, na ktorej vidíme, aká silná je príslušnosť jednotlivých objektov k zhlukom.

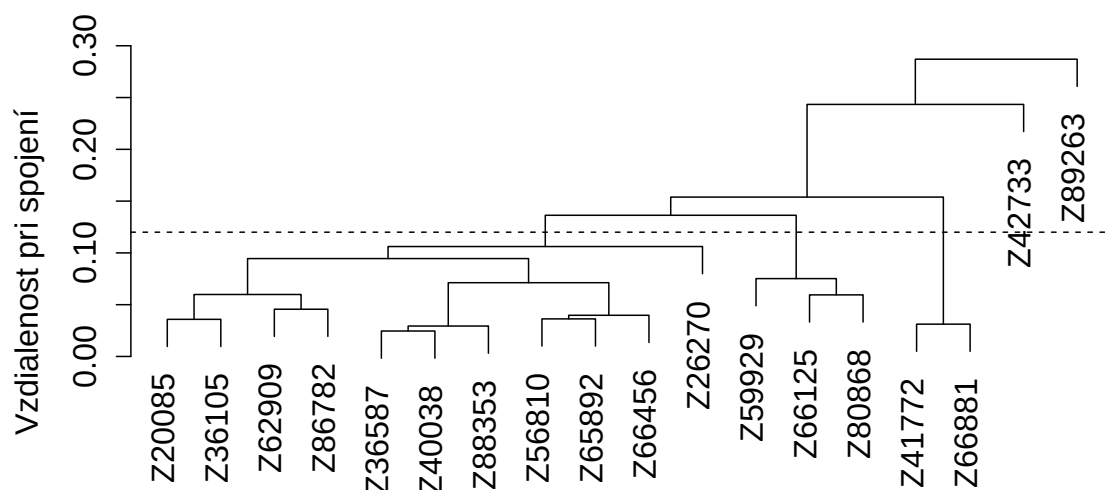


Obr. 15: Histogram priemerných nákladov vzoriek v permutačnom teste škálovaných faktorom 1, 25. Priemer skúmaného poskytovateľa je vyznačený bodom.

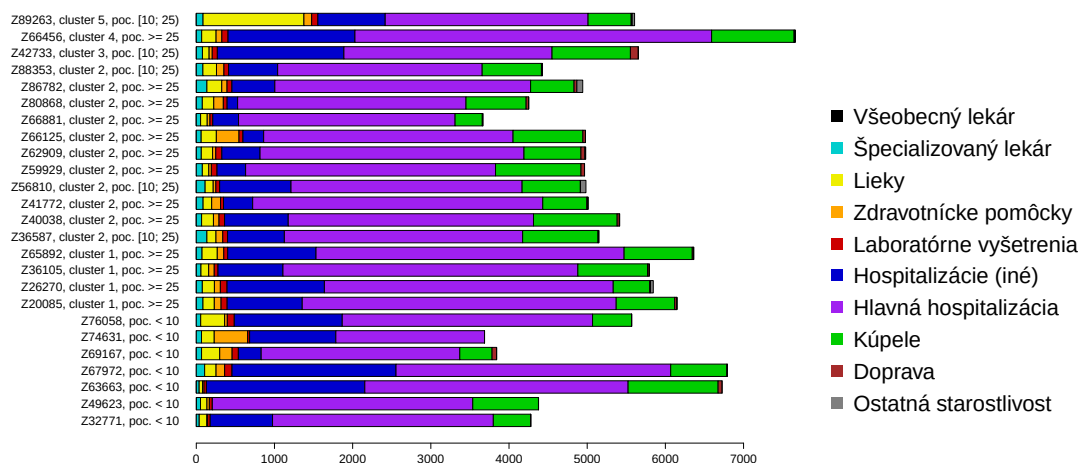
Podľa [6] je často vhodné pred zhlukovou analýzou dáta najprv normalizovať. My vykonáme normalizáciu v takom zmysle, že náklady na vrstvy starostlivosti každého poskytovateľa predelíme jeho celkovými nákladmi. Tým naše premenné dostali novú interpretáciu, a to časť nákladov poskytovateľa, ktorú tvorí daná vrstva. Najpodobnejší pre clustrovú analýzu budú teda poskytovatelia, ktorí majú najpodobnejšie zastúpenie jednotlivých vrstiev na svojich nákladoch. Z analýzy navyše vylúčime poskytovateľov, ktorí sú za menší počet ako 10 skúmaných pacientov — tým odstránime 7 poskytovateľov z 25. Pri zhlukovej analýze používame „average linkage” metódu, ktorú sme zvolili, lebo je to akási stredná cesta v rámci ostatných používaných metód ako „single linkage” a „complete linkage”.

Na obr. 16 vidíme dendrogram pre zhlukovú analýzu. Pohľadom na dendrogram sa rozhodneme, na akej úrovni ho „zostrihneme”, čo určí výsledný počet zhlukov. My opticky zvolíme výsledný počet zhlukov 5.

Výsledky zhlukovej analýzy vidíme na obr. 17. Zhluky môžeme z obrázku aj vizuálne interpretovať. Cluster 2 tvoria poskytovatelia s nižšími nákladmi na ďalšie hospitalizácie. Cluster 1 tvoria skôr drahší poskytovatelia s väčšími nákladmi na ďalšie hospitalizácie. Zhluky 3, 4, 5 sú samostatní poskytovatelia. Pozoruhodný je zhluk 5, ktorý má veľké náklady na lieky. Tiež stojí za povšimnutie zhluk 4, ktorý tvorí poskytovateľ



Obr. 16: Dendrogram pre analýzu zhlukov. Čím nižšie sú dva objekty spojené, tým väčšia je ich podobnosť. Prerušovaná čiara predstavuje zvolenú úroveň „zostrihania” stromu.



Obr. 17: Náklady na poskytovateľov po vrstvách a zoskupenie do zhlukov.

identifikovaný v časti 4.3.1 ako extrémne nákladový.

Záver

V práci sme popísali spôsob zoskupovania epizód zdravotnej starostlivosti založený čisto na pohľade na dáta. Postup sme vytvorili pre epizódu implantácie bedrového kĺbu a overili sme, že s rovnakým nastavením parametrov dostaneme kvalitatívne podobné výsledky pre epizódu implantácie kolenného kĺbu (časť 3.4). Výsledky sme v časti 3.3 verifikovali porovnaním s medicínskym priradením výkonov, pričom medicínske priradenie sa na starostlivosti pre ktorú ho máme k dispozícii (zodpovedajúcej 80% celkových nákladov) s naším priradením zhoduje na vyše 95% nákladoch.

K epizódam sme priradili ambulantné kontakty, hospitalizácie a kúpeľné na základe nami vytvoreného bodovania. Dokopy mohla každá skupina výkonov získať 11 bodov, presnejšie 4 body za diagnózu, 3 body za odbornosť nižšej zložky, 3 body za typ produktu zdravotnej starostlivosti a jeden bod za kombináciu premenných, viac bodov za silnejšiu relevanciu k skúmanej epizóde, pričom relevanciu sme skúmali predovšetkým pomocou analýzy asociačných pravidiel, pričom značnú pozornosť sme venovali voľbe konkrétneho tvaru problému analýzy asociačných pravidiel (časť 2.4).

V postupe sa vyskytlo veľa parametrov, ktorých hodnotu sme museli určiť bez možnosti priamej verifikácie ich dopadu. Hoci verifikácia nad celkovým výsledkom dopadla dobre, väčšia pozornosť venovaná nastaveniu týchto parametrov by mohla byť námetom pre budúci výskum.

V kapitole 4 sme uviedli príklady ďalších analýz, ktoré sa dajú vykonávať so zoskupenými epizódami starostlivosti. Za hlavnú výhodu analyzovania napríklad nákladov za jednu epizódu považujeme, že skúmame náklady čisto na starostlivosť súvisiacu s jedným problémom, čo nám očistí dáta od rôznych náhodných vplyvov. Ukazujeme napríklad možnosť identifikácie extrémne drahých poskytovateľov zdravotnej starostlivosti, alebo zhlukovanie poskytovateľov do skupín na základe podobných charakteristík ohľadne starostlivosti o skúmanú epizódu.

V priebehu práce uvádzame niekoľko ďalších námetov, ktoré by mohli slúžiť ako vylepšenia nášho postupu v budúcnosti. Niektoré z nich sú:

- vyberať referenčnú skupinu pre účely vytvárania asociačných pravidiel podobnejšiu skupine so skúmanou epizódou, napríklad s rovnakým demografickým zložením,

- skúmať dáta za dlhšie časové obdobie, keďže v časti 4.1 sme zistili, že nami zvolené obdobie dvoch rokov pravdepodobne nepokrýva skutočnú dĺžku trvania epizódy,
- identifikovať úrovně závažnosti epizódy. Toto môže byť veľmi dôležité pre ďalšie skúmanie, avšak môže byť ťažko vykonateľné bez ďalšieho medicínskeho vstupu.

Zoznam použitej literatúry

- [1] Agrawal, R., Imielinski, T., Swami, A.: *Mining Association Rules between Sets of Items in Large Databases*, ACM-SIGMOD Int'l Conference on Management of Data, Washington D.C (1993), dostupné na internete (18. 4. 2014): <http://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf>
- [2] Agrawal, R., Srikant, R.: *Fast Algorithms for Mining Association Rules*, Proc. 20th int. conf. very large data bases, VLDB (1994), dostupné na internete (18. 4. 2014): <http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf>
- [3] Borgelt, C.: *Find Frequent Item Sets and Association Rules with the Apriori Algorithm*, dostupné na internete (18. 4. 2014): <http://www.borgelt.net/doc/apriori/apriori.html>
- [4] Centers of Medicaid & Medicare Services: *2011 Medical Group Practice Supplemental QRUR User Guide June*, dostupné na internete (18. 4. 2014): <http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Downloads/QRUR-Supplemental-User-Guide.pdf>
- [5] Hahsler, M., Grün, B., Hornik, K.: *arules: Mining Association Rules and Frequent Itemsets*, R package version 1.0-13 (2013), dokumentácia na internete (18. 4. 2014): <http://cran.at.r-project.org/web/packages/arules/vignettes/arules.pdf>
- [6] Harman, R.: *Viacrozmerné štatistické analýzy 2*, prednášky, FMFI UK, Bratislava, 2013
- [7] Houchens, R. L., et al.: *The Use of an Episode Grouper For Physician Profiling in Medicare*, Medicare Payment Advisory Commission (2009), dostupné na internete (18. 4. 2014): http://www.medpac.gov/documents/Jun09_EpisodeGrouperStability_CONTRACTOR_JP.pdf
- [8] Hromjaková, Z.: *osobná komunikácia*, Dôvera, zdravotná poisťovňa, Bratislava, 2014

- [9] Hussey, P. et al.: *Using Episodes of Care as a Basis for Performance Measurement and Payment: Moving from Concept to Practice*, Health Aff 28 (2009), 1406-1417, dostupné na internete (18. 4. 2014): <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2925202/>
- [10] Somorčík, J.: Počítačová štatistika, prednášky, FMFI UK, Bratislava, 2013
- [11] *Systematický zoznam kódov diagnóz MKCH-10-SK-2013*, dostupné na internete (18. 4. 2014): <http://www.nczisk.sk/Standarty-v-zdravotnictve/Pages/Medzinarodna-klasifikacia-chorob-MKCH-10.aspx>
- [12] Tulejová, H.: osobná komunikácia, Dôvera, zdravotná poisťovňa, Bratislava, 2013, 2014
- [13] *Vestník úradu pre dohľad nad zdravotnou starostlivosťou*, č. 6/2007, dostupné na internete (18. 4. 2014): http://www.udzs-sk.sk/buxus/docs/vestniky/rocnik_2007/vestnik_6_2007_1_.pdf
- [14] *Vyhláška č. 151/2011 Z. z., ktorou sa mení a dopĺňa vyhláška Ministerstva zdravotníctva Slovenskej republiky č. 412/2009 Z. z., ktorou sa ustanovujú podrobnosti o zozname poistencov čakajúcich na poskytnutie plánovanej zdravotnej starostlivosti*
- [15] WHO Collaborating Centre for Drug Statistics Methodology: *ATC: Structure and principles*, dostupné na internete (18. 4. 2014): http://www.whocc.no/atc/structure_and_principles/
- [16] *Zákon č. 185/2012 Z. z., ktorým sa mení a dopĺňa zákon č. 580/2004 Z. z. o zdravotnom poistení a o zmene a doplnení zákona č. 95/2002 Z. z. o poisťovníctve a o zmene a doplnení niektorých zákonov v znení neskorších predpisov a o zmene a doplnení niektorých zákonov*