ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

Katedra biomedicínské techniky

Štatistické metódy pre riešenie neúplných dát v hodnotení zdravotníckych technológií

Statistical methods for dealing with incomplete data in health technology assessment

Diplomová práca

Študijný program:	Biomedicínska a klinická technika
Študijný obor:	Systémová integrácia procesov v zdravotníctve

Autor diplomové práce:	Bc. Ľubica Ladányiová
Vedúci diplomové práce:	Ing. Vojtěch Kamenský

Kladno 2019

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení:	**Ladányiová**	Jméno:	**Ľubica**	Osobní číslo:	**425682**
Fakulta:	**Fakulta biomedicínského inženýrství**				
Garantující katedra:	**Katedra biomedicínské techniky**				
Studijní program:	**Biomedicínská a klinická technika**				
Studijní obor:	**Systémová integrace procesů ve zdravotnictví**				

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Statistické metody pro řešení neúplných dat v hodnocení zdravotnických technologií.

Název diplomové práce anglicky:

Statistical methods for dealing with incomplete data in health technology assessment.

Pokyny pro vypracování:

Cílem diplomové práce je porovnání statistických metod pro řešení chybějících či neúplných klinických dat pro potřeby HTA. Analyzujte současný stav používaných metod pro řešení chybějících či neúplných dat v ČR a ve světě. Na základě analýzy současného stavu vyberte vhodné metody pro aplikaci u klinických dat. Následně tyto statistické metody aplikujte na reálná klinická data pacientů s postižením povrchní stehenní tepny a porovnejte vliv jednotlivých metod na výsledky HTA analýzy. Výstupem práce bude zhodnocení výhod a nevýhod použití jednotlivých metod a doporučení pro využívání těchto metod pro HTA.

Seznam doporučené literatury:

[1] Zvárová, J., Lhotská, L., Přibík, V., Biomedicínská informatika: Data a znalosti v biomedicíně a zdravotnictví, Nakladatelství Karolinum, 2010, ISBN 978-80-246-1805-0

Jméno a příjmení vedoucí(ho) diplomové práce:

Ing. Vojtěch Kamenský

Jméno a příjmení konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **18.04.2019**
Platnost zadání diplomové práce: **20.09.2020**

prof. Ing. Peter Kneppo, DrSc.
podpis vedoucí(ho) katedry

prof. MUDr. Ivan Dylevský, DrSc.
podpis děkana(ky)

PREHLÁSENIE

Prehlasujem, že som diplomovú prácu s názvom „Štatistické metódy pre riešenie neúplných dát v hodnotení zdravotníckych technológií“ vypracovala samostatne a použila som k tomu úplný výpis citácií použitých prameňov, ktoré uvádzam v zozname priloženom k diplomovej práci.

Nemám závažný dôvod proti použitiu tohto školského diela v zmysle § 60 Zákona č. 121/2000 Zb., o práve autorskom, o právach súvisiacom s právom autorským a o zmene niektorých zákonov (autorský zákon), v znení neskorších predpisov.

V Prahe …..……..……...

Bc. Ľubica Ladányiová

POĎAKOVANIE

V prvom rade patrí poďakovanie vedúcemu práce Ing. Vojtěchovi Kamenskému, za jeho odborný dohľad, trpezlivý prístup a podnetné pripomienky, ktorými mi vypomáhal počas celej doby spracovávania diplomovej práce. Ďakujem rodine, priateľom a kolegom za podporu a rady.

ABSTRAKT

Štatistické metódy pre riešenie chýbajúcich dát v hodnotení zdravotníckych technológií

Cieľ práce: Využitie štatistických metód na riešenie chýbajúcich hodnôt, simulácia mechanizmu neúplných dát na reálnych údajoch zo štúdie na klinicko-ekonomické zhodnotenie intervencií povrchovej stehennej tepny. Porovnanie vplyvu štatistických metód na výsledky nákladovej analýzy a následné zhodnotenie metód.

Úvod: Pomocou štatistických metód môžeme riešiť problém chýbajúcich hodnôt. Metódy sa rozlišujú podľa prístupu na metódy odstraňujúce nekompletné prípady a na metódy doplňujúce chýbajúce hodnoty.

Metódy: Na plnej dátovej matici bol pomocou softvéru R nasimulovaný stav chýbajúcich hodnôt mechanizmu MCAR v rozsahu 20 %. K porovnaniu vplyvu štatistických metód boli využité štyri metódy. (1) Metóda analýzy kompletných prípadov, ktorá pracuje s odstránením neúplných prípadov z merania. (2) Imputácia aritmetickým priemerom, využívajúc doplnenie dát pomocou priemeru. (3) Imputácia EM logaritmom a (4) imputácia MCMC logaritmom sú metódy mnohonásobného doplnenia. Posledné dve menované metódy boli spracované v programe R. Následne boli spočítané priemerné náklady pre jednotlivé intervencie. Z priemerných nákladov bola vypočítaná nákladová efektivita pre štyri klinické efekty. Zhodnotenie vplyvu štatistických metód bolo pomocou porovnania výsledkov s výsledkami zdrojových dát.

Výsledky: Analýza kompletných prípadov bola štatisticky významne odlišná od zdrojových dát pre intervenciu PTA/s a bypass. Pre metódy jednoduchej imputácie aritmetickým priemerom, mnohonásobnej imputácie EM algoritmom a MCMC algoritmom neboli zistené štatisticky významné rozdiely oproti zdrojovým dátam. V analýze nákladovej efektivity sa analýza kompletných prípadov líšila v prípade PTA o 1, 74 %, PTA/s 4,77 % a bypass o 3,80 %, hodnota ICER 1: 24,94 %, 8,12 %, ICER 2 22,0 %. Imputácia aritmetickým priemerom sa líšila oproti zdrojovým dátam PTA 0,11 %, PTA/s 1,43 %, bypass 0,37 %. Mnohonásobná imputácia Em algoritmom bol rozdiel pre PTA 0,38 %, PTA/s 1,18 %, bypass 1,13 %, ICER 1: 3,94 %, 2, 06 %, ICER 2: 1,27 %. Mnohonásobná imputácia mala najmenšie rozdiely so zdrojovými dátami: PTA 1,15 %, PTA/s 0,04 %, ICER 1: 3,90 %, 1,35 %, ICER 2: 0,27 %.

Záver: Ako vhodnejšie metódy k riešeniu chýbajúcich dát sú metódy mnohonásobnej imputácie. Využitie analýzy kompletných prípadov sa nedoporučuje, zmenšuje veľkosť vzorky.

Kľúčové slová: Chýbajúce hodnoty, analýza kompletných prípadov, imputácia aritmetickým priemerom, mnohonásobná imputácia, EM algoritmus, Markovove reťazce Monte Carlo, analýza nákladovej efektivity, hodnotenie zdravotníckych technológií

ABSTRACT

Statistical methods for dealing with missing data in health technology assessment

Objectives: The use of statistical methods to deal with missing values, simulation of the mechanism of incomplete data on real-life clinical-economic study on the evaluation of surface femoral artery interventions. Comparison of the impact of statistical methods on the results of cost effectiveness analysis and subsequent evaluation of the methods.

Introduction: By using statistical methods, we can solve the problem of missing values. Methods are distinguished based on the approach, on methods, that remove incomplete cases, and methods that complement missing values.

Methods: Using the full data matrix and software R, the state of missing values of the MCAR mechanism was simulated – mechanism "missing completely at random" in the range of 20 %. Four methods were used to compare the impact of statistical methods. (1) Complete Case study, which works with the removal of incomplete measurements. (2) Mean imputation that adds data using a value of the mean for each variable. (3) Imputation of EM algorithm and (4) imputation of MCMC algorithm are model-based multiple imputation methods. Last two mentioned methods were processed in the R program. Following the average cost was calculated for individual interventions. Cost- effectiveness analysis for four the clinical effects was calculated from the average cost taken from each method. Subsequently, the impact of statistical methods was evaluated by comparing their results with the results from the source data.

Results: Complete case analysis was statistically significantly different in the intervention PTA / s bypass. Single mean imputation, EM multiple imputation and multiple imputation MCMC were not statistically significant different with source data.

After cost-effectiveness analysis the differences between Complete case vs source data were: for PTA by 1, 74 %, PTA/s 4,77 % and bypass by 3.80 %, ICER 1: 24,94 %, 8,12 %, ICER 2: 22,0 %. Single mean imputation vs. source data: PTA by 0,11 %, PTA/s 1,43 % and bypass 0,37 %. Multiple imputation by EM algorithm was a difference for PTA 0,38 %, PTA/s 1,18 %, bypass 1,13 %, ICER 1: 3, 94 %, 2, 06 %, ICER 2: 1, 27 %. Multiple imputation had the smallest differences with source data: PTA 1,15 %, PTA/s 0,04 %, ICER 1: 3,90 %, 1,35 %, ICER 2: 0, 27 %.

Conclusion: More suitable methods for dealing with missing data are methods of multiple imputations. Complete case analysis is not recommended as it reduces samples.

Keywords: Missing values, complete case study, mean imputation, multiple imputation, EM algorithm, Markov chain Monte Carlo, cost effectiveness analysis

Obsah

Zoznam použitých skratiek

Skratka	Význam
BVC	Body výkon celkom
CBA	Cost benefit analysis
CCA	Cost consequence analysis
CEA	Cost effectiveness analysis
CE	Cost effectiveness
CEA	Cost effectiveness
CMA	Cost minimization analysis
COI	Cost of ilness
COT	Cost of treatment
CUA	Cost utility analysis
EM	Expectation maximization
GUI	Graphical user interface
HTA	Health technology assesment
ICER	Incremental cost effectivenesss ratio
KL	Kontrastná látka
MAR	Missing values at random
MCMC	Markov chain Monte Carlo
MCAR	Missing values competely at random
MNAR	Missing not at random
MI	Multiple imputation
ML	Maximum likelihood
OŠD	Ošetrovacie dni
PTA	Perkutánna transluminárna angioplastika
PTA/s	Perkutánna transluminárna angioplastika s následnou implantáciou stentu
QUALY	Quality adjustment life year
VIM	Visualization of imputed values
ZULP	Zvlášť účtovaný lekársky prípravok
ZUM	Zvlášť účtovaný materiál
ZUM_C	Zvlášť účtovaný materiál celkom

Úvod

Chýbajúce hodnoty v dátach sú problém, ktorý riešia výskumníci počas celého priebehu vedeckej činnosti. Ak by sme žili v ideálnom svete, tak by všetky výsledky boli perfektné a všetky hodnoty kompletné. Výskyt neúplných dát pre štúdie je však častý [1]. Problém neúplných dát sa týka nielen spoločenských vedných oborov, ale aj technickej a lekárskej praxe [1, 2].

Príčiny vzniku sú rôzne a samozrejme záleží na dizajne práce, akým sa výskum spracováva. Konkrétne, či sa jedná o prospektívnu štúdiu, kde sa chýbajúcim dátam môžeme vyhnúť pravidelnými kontrolami a tým pádom sa dá úplne predchádzať ich vzniku. V retrospektívnych štúdiách, kde výsledky boli namerané a výskum ukončený, je potreba vzniknutý problém riešiť alternatívnym spôsobom. Najčastejšie príčiny neúplných dát sú zapríčinené ľudským faktorom alebo technikou, ako napríklad nesprávne nastavenie prístroja a podobne [3]. V klinických skúškach môže nastať situácia, kedy pacient predčasne ukončí štúdiu, ale výsledky je potreba aj napriek tomu vyhodnotiť [4]. Pre spracovanie akejkoľvek analýzy, ktorá vyhodnocuje výskum, potrebujeme kompletné údaje [3,7]. Bez tohto predpokladu nie je možné vypočítať výsledky a štúdie vyhodnotiť.

Tento problém sa začal riešiť pomocou štatistických metód. Za otca štatistických metód pre riešenie chýbajúcich dát môžeme považovať pána Donalda Rubina, amerického štatistika, ktorý sa touto problematikou zaoberal už od sedemdesiatych rokov minulého storočia. Z jeho štúdií, tvrdení a teoretických poznatkov, vychádza veľa súčasných prác a metód. Práve on tvrdil, že sa problém chýbajúcich hodnôt častokrát rieši ad hoc metódou, ktorá má za následok nepresné výsledky a následne nevhodne vyvedené závery [2].

Cieľom diplomovej práce je porovnať vplyv jednotlivých štatistických metód. Vytvoriť prehlaď metód k riešeniu chýbajúcich dát, ktorý by priblížil čitateľovi danú problematiku komplexne a pomohol mu s výberom metódy. V práci sú využité reálne dáta, ktoré slúžili ku klinicko-ekonomickému zhodnoteniu troch intervencií.

K dosiahnutiu cieľa boli stanovené dielčie úlohy:

1. Simulácia chýbajúcich hodnôt na zdrojových dátach obsahujúcich plné výsledky.
2. Aplikácia vybraných štatistických metód na riešenie chýbajúcich údajov.
3. Vyhodnotenie HTA analýzy a porovnanie vplyvu jednotlivých metód na výsledok.
4. Zhodnotenie využitia metód a vytvorenia odporučení k výberu a využitiu.

V práci je pozornosť venovaná hlavne na rozdiely medzi štatistickými metódami. Rozdiely v interpretácii HTA analýzy nie sú rozoberané.

1 Súčasný stav problematiky

Základy štatistických metód pre riešenie chýbajúcich dát vznikli v sedemdesiatych rokoch minulého storočia. Rubin vo svojej knihe rozlišuje dva základné prístupy k riešeniu neúplných dát [6]:

1. Metódy založené na odstránení neúplného vzorku.
2. Metódy založené doplnení chýbajúcich hodnôt.

V priebehu rokov sa vytvorili ďalšie pokročilé, sofistikované techniky na riešenie chýbajúcich dát, ktoré sa snažia chýbajúci údaj nasimulovať. Využíva sa pri nich predpoklad normálneho rozloženia dát či zložité viacnásobné Monte Carlo simulácie.

Účelom tejto kapitoly je vytvoriť prehľad dostupných štatistických metód, charakterizovať základné limity pre ich využitie ako aj vyhodnotenie ich pozitívnych a negatívnych vplyvov. Získané informácie boli základom pre rozhodovanie vo výbere štatistických metód použitých na doplnenie dát v zdrojových dátach.

1.1 Základné typy štúdií v HTA

Hodnotenie zdravotníckych technológií HTA bolo definované Goodmanom [7] ako systematické hodnotenie vlastnosti, efektu alebo iného dopadu zdravotníckych technológií. Účelom je informovať o technológiách a ich využitiu v zdravotníctve. Vytvoriť štúdiu je možné nielen z pohľadu pacienta, poskytovateľa zdravotnej starostlivosti, inštitúcie ale aj z pohľadu dopadu na spoločnosť. Pojem zdravotníckej technológie však neoznačuje len prístrojové vybavenie ale aj liečivé prípravky, diagnostické a liečebné postupy, preventívne postupy [7].

K vyhodnoteniu používame odporučené postupy, ktorými vypracovávame rôzne typy analýz. K práci využívame nielen primárnárne dáta či údaje získané zbieraním hodnôt ale aj sekundárne pramene z dostupných publikovaných štúdií. Veľkú časť štúdií hodnotenia zdravotníckych technológií tvorí komparatívna analýza, ktorá vzájomne porovnáva napríklad klinické efekty danej liečebného postupu či zdravotníckej technológie.

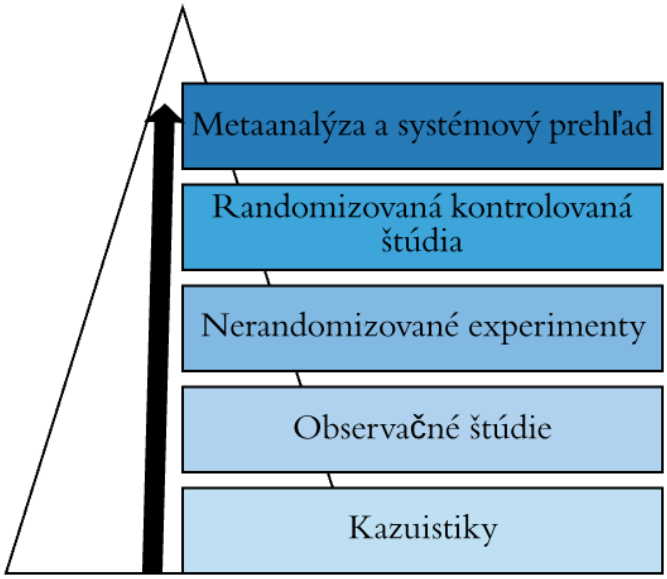
Sekundárne zdroje môžeme podľa sily dôkazu rozdeliť na:

1. Metaanalýza a systematický prehľad – využívajú dáta z randomizovaných kontrolovaných štúdií, či iný publikovaných výsledkov, ktoré následne štatisticky spracovávajú – výsledok má najvyššiu vedeckú váhu,
2. Randomizované kontrolované štúdie – štúdia, ktorá paralelne prebieha v dvoch skupinách a z toho je jedna skupina liečená s placebom a druhá s liečivou látkou.

Rozdelenie do skupín je náhodné. A účastníci nevedia, do ktorej skupiny sú zaradení.

3. Nerandomizované experimenty – paralelné štúdie v dvoch skupinách,
4. Observačné štúdie – pozorovacia štúdia – nie je vytvorená kontrolná skupina s placebo efektom,
5. Kazuistiky – prípadové štúdie.

Sila výsledkov vedeckých štúdií je prezentovaná na obrázku č. 1.



Obrázok 1: Sila rastu vedeckých štúdií [vlastné spracovanie na základe 53]

Veľkú časť výskumnej práce v HTA tvoria nákladové analýzy [7].

1. **COI** – Cost of illness – Analýza nákladov na ochorenie – číselné vyčíslenie všetkých nákladov, (priame zdravotnícke náklady, nepriame zdravotnícke náklady, nezdravotnícke náklady) náklady, ktoré nie je možné číselne vyjadriť (úroveň, stresu, bolesť, pocit pohodlia).
2. **CMA** – Cost minimalization analysis – Analýza minimalizácie nákladov - porovnáva dve varianty o vybraných parametroch, následne je zvolená varianta s nižšími nákladmi - náklady sú číselne vyhodnotené.
3. **CBA** – Cost benefit analysis – Analýza nákladov a prínosov – Analýza hodnotí cenu diagnostických či terapeutických prístrojov / metód a výšku nákladov, náklady sú vyjadrené v peňažných jednotkách.
4. **CEA** – Cost effectivenes analysis – Analýza nákladovej efektivity – metóda na porovnávanie dvoch alebo viacerých alternatívnych technológií s mierou/ účinnosťou efektu – všetky parametre sú kvantitatívne vyjadrené.
5. **CUA** – Cost utility analysis – Analýza nákladovej efektivity – Používa indikátor QALY – prepočet rokov života za predpokladu plnej kvality života – hodnota

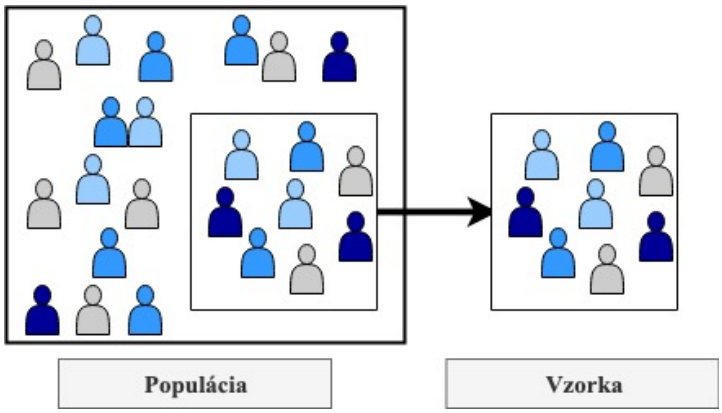
je získavaná z dotazníkových štúdií, niektoré údaje sa zhodnocujú podľa rôznych hodnotiacich škál kvalitatívneho charakteru. Nie všetky hodnoty majú kvantitatívne – číselné vyjadrenie.

Z časového hľadiska spracovania dát rozlišujeme štúdie na [8]:

1. **Prospektívne** - Metódy si zvolíme pred samotným začiatkom štúdie. V tomto prípade je ľahšie prechádzať neúplným dátam, poprípade pri odstránení neúplných dát je možné doplniť ďalšie kazuistiky, merania, či počty experimentov.
2. **Retrospektívne** – Využívajú dáta zozbierané v minulosti, namerané hodnoty primárne ani nemuseli byť určené k vedeckému spracovaniu. Pri tomto type štúdie je častý výskyt chýbajúcich hodnôt. Ak je klinická štúdia ukončená, dodatočné meranie už nie je možné. Pri tomto type štúdií je vhodné využiť štatistické metódy a postupy k ich doplneniu a následne je potom možné hodnoty adekvátne vedecky spracovať.

Pred samotným použitím štatistických metód na doplnenie dát je potrebné si definovať základné pojmy a termíny, ktoré sa v štatistike používajú [9]:
Premenná – znaky, ktoré sú zaznamenané meraním, operáciou, v literatúre sa môžeme stretnúť aj s označením vektor,
Populácia – základný súbor, množina ľubovoľných štatistických jednotiek,
Vzorka – výberový súbor, dátový súbor, dáta, podmnožina celej populácie, taktiež dátová matica. Vzťah medzi vzorkou a populáciou je graficky znázornení na obrázku č. 2.

Údaje, s ktorými pracujeme si môžeme predstaviť ako dátovú maticu so zaznamenanými hodnotami. Vyjadrené premenné nemajú len číselný charakter, ale môžu byť zapísané aj kategoricky - slovne. Z matematického hľadiska rozdeľujeme štatistické znaky na presný číselný údaj, reálne číslo alebo hodnoty na škále [9].



Obrázok 2: Populácia vs Vzorka [vlastné spracovanie]

Základné typy premenných si môžeme rozdeliť na [10]:

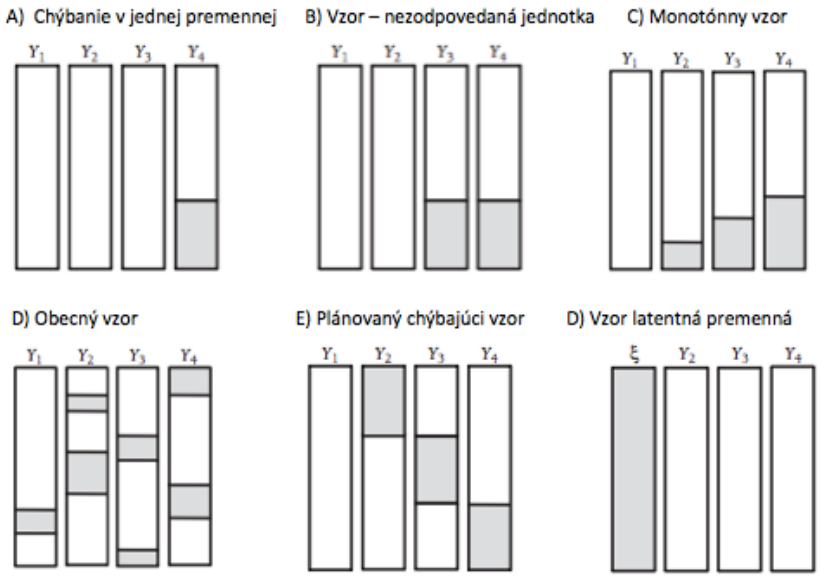
1. **Kvantitatívne premenné jednotky –** numerické premenné – vyjadrené sú číslom a je možné ich zmerať, delia sa na
   - **diskrétne premenné** – majú konečné alebo spočítateľné množstvo variant (počet výskytu ochorenia, vek pacienta v rokoch)
   - **spojité premenné –** majú nespočetne veľké množstvo realizácií z množiny reálnych čísiel alebo z ich ľubovoľnej podmnožiny (priemerný príjem v KČ)

2. **Kvalitatívne kategoriálne premenné** – vyjadrené slovne, nie je možné ich zmerať. Podľa možných tried rozlišujeme - využívame napríklad v dotazníkoch
   - **binárne premenné –** možné sú len dve možní varianty výsledkov (pohlavie, výskyt ochorenia)
   - **množné premenné –** vyskytujú sa vo viacerých hodnotách (rodinný stav, dosiahnuté vzdelanie)

K výberu vhodnej metódy na riešenie chýbajúcich dát je potrebné poznať mechanizmus vzniku chýbajúcich dát. Rubin vo svojej knihe definoval tri základné mechanizmy ich vzniku MAR – Missing at Random – hodnoty chýbajú náhodne, MCAR – Missing completely at random – hodnoty chýbajú úplne náhodne a MNAR Missing not at random – hodnoty nechýbajú náhodne. Princíp mechanizmov je vysvetlený v ďalšej časti kapitoly. Nie všetky štatistické postupy sú vhodne pre každý typ mechanizmu [6]. Jednotlivé mechanizmy opisujú vzájomné vzťahy medzi nameranými premennými a ich pravdepodobnou príčinou ich neúplnosti [3].

K pochopeniu problematiky dát je potrebné rozlišovať aj vzor chýbajúcich dát [2]. Chýbajúci vzor údajov vystihuje konfiguráciu pozorovaných hodnôt a ich chýbanie v súbore sledovaných parametrov [8].

Chýbanie jednej premennej je relatívne vzácny jav, vyskytnúť sa napríklad môže v experimentálnych štúdiách [2]. S prípadom, že hodnota chýba v rovnakom pomere v dvoch premenných jednotkách, sa môžeme stretnúť pri štúdiách, ktoré porovnávajú dve technológie s nízkou cennou oproti technológiám s vyššou cenou. Monotónny chýbajúci vzor je typický pre dlhé štúdie, kde účastníci odchádzajú a už sa nepočíta s ich opätovným návratom . Napríklad ukončenie účasti na klinickom testovaní nového liečiva z dôvodu výskytu vážnych nežiadúcich účinkov. Z literatúry vyplýva, že monotónny chýbajúci vzorec výrazne znižuje matematickú komplexnosť sofistikovaných matematických metód [11]. Enders je názoru, že najbežnejší vzor akým chýbajú údaje je obecný vzor, kde hodnoty chýbajú úplne náhodne. [2]. Vzor chýbajúcich dát, ktorý je plánovaný sa vyskytuje pri zhromažďovaní veľkého počtu dotazníkov pre danú štúdiu. Vzor latentnej premennej chýba pre celú jednu sledovanú vzorku, v bežnej praxi sa nevyskytuje často [12]. Jednotlivé vymenované vzory môžeme vidieť na obrázku č. 3.



Obrázok 3: Typy vzorov chýbajúcich hodnôt [6]

1.2 Mechanizmy vzniku chýbajúcich hodnôt

Pred samotným použitím štatistických metód je potrebné poznať charakter chýbajúcich dát, mechanizmus ich vzniku a vzájomní vzťahy medzi chýbajúcimi premennými jednotkami [6].

1.2.1 Hodnoty chýbajú náhodne MAR (Missing at Random)

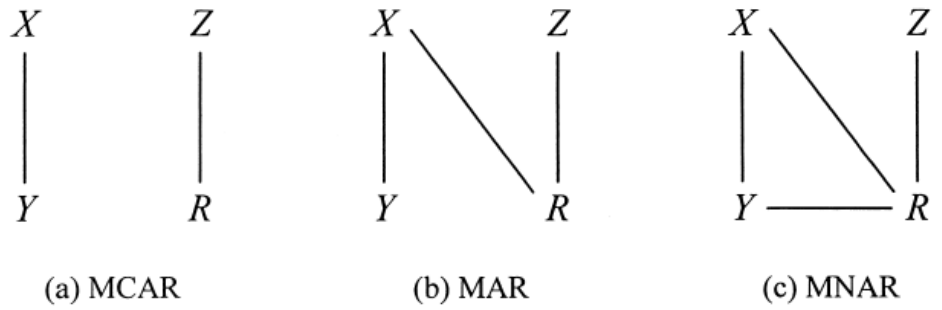
Náhodné rozloženie chýbajúcich dát – Chýbajúce hodnoty v mechanizme MAR môžu byť označené ako ignorovateľné, väčšinou s nimi a k s nimi štúdia počíta. V praxi sa to prejavuje ako preferovanie odpovede u jedného pohlavia. Ako príklad si môžeme uviesť uvedenie hmotnosti a veku u ženy. Muži sú ochotnejší tieto údaje vyplňovať[34].

1.2.2 Hodnoty chýbajú úplne náhodne MCAR (Missing Completely at Random)

Údaje chýbajú úplne náhodne, pokiaľ hodnoty nezávisia na akýkoľvek dátach, ktoré sú pozorované alebo chýbajúce. V praxi sa s mechanizmom MCAR môžeme typicky stretnúť pri strate dát, zlyhaní prístroja či nevyplnení dotazníka či karty pacienta [34].

1.2.3 Hodnoty nechýbajú náhodne MCAR (Missing Not at Random)

Ak mechanizmus vzniku chýbajúcich dát nie je možné zaradiť k MAR alebo MCAR, zaraďujeme ho do kategórie NMAR. Typickým príkladom môže byť zverejnenie nákladov (napríklad nákupná cena prístroja, či spotrebného materiálu), čím sú náklady vyššie, tým je väčšia pravdepodobnosť, že tieto údaje nebudú zverejnené [34].



Obrázok 4: Mechanizmus vzniku MCAR, MAR, MNAR: Grafické znázornenie jednotlivých mechanizmov [6]

X – kompletne dostupná pozorovaná premenná

Y – čiastočne neúplná premenná

Z – súčasť príčiny chýbajúcich dát

R – chýbanie dát

1.3 Prehľad metód pre riešenie neúplných dát

V tejto kapitole je vytvorený prehľad štatistických metód pre riešenie chýbajúcich dát. Štatistické metódy pre riešenie chýbajúcich dát si môžeme rozdeliť do troch hlavných skupín, metódy založené na odstránení vzorku, metódy založené na doplnení a sofistikované špeciálne štatistické metódy. Metódy založené na odstránení neúplného vzorku

1.3.1 Metódy založené na vynechaní vzorky

Analýza kompletných prípadov - Litwise deletion - Complate case analysis

Metóda je založená na odstránení prípadov, ktoré neobsahujú všetky namerané hodnoty či údaje. S týmto typom riešenia chýbajúcich dát sa môžeme stretnúť pomerne často. Túto metódu môžeme použiť aj pre porovnanie výsledkov zo štúdie, ktorá obsahovala kompletné údaje s analýzou, ktorá sledovala rovnaký zámer ale nebola úplná [13]. Výsledky je možné následne porovnať a zhodnotiť.

Výhody: Môže sa použiť na akúkoľvek analýzu a nie sú potrebné žiadne špeciálne výpočtové metódy, jednoduchá metodika, prosté vylúčenie zo vzorku, nie je nutné poznať charakter chýbajúcich dát – aplikovateľné na všetky mechanizmy (MAR, MNAR, MCAR).

Nevýhody: Môže vylúčiť veľkú časť výsledkov. Ako príklad si môžeme uviesť súbor údajov s 500 ľuďmi a 10 premenných, každá má chýbajúce údaje v 2,5 %, potom vylúčeným neúplných prípadov nám vznikne 180 jednotlivcov, čim sa zbavíme 320 ostatných. Funguje dobre s mechanizmom MCAR [14]. Znižuje sa váha výsledku, lebo sa nám zmenšuje pozorovaný súbor.

Analýza dostupných prípadov - Pairwise deletion - Available case analysis

Analýza s pokúša minimalizovať stratu, ku ktorej dochádza pri vymazaní z analýzy kompletných prípadov. Využíva sa pri tom korelácia – vzájomný vzťah medzi dvoma premennými. Pre každý pár, ktoré sú známe sa zohľadni korelačný koeficient.

Výhody: Metóda maximalizuje všetky dostupné údaje [12, 15]. Technika je preferovanejšia než Litwise deletion, pretože nedochádza k radikálnemu zmenšovaniu sledovanej vzorky.

Nevýhody: Je potreba poznať charakter chýbania dát, metóda vhodná pre mechanizmus MCAR. Literatúra uvádza vznik štandardnej chyby vzniknutej použitím počítačovej techniky [6].

1.3.2 Metódy založené doplnení chýbajúcich hodnôt

Jednoduchá imputácia aritmetickým priemerom

Metóda vypočítava jedinú náhradnú hodnotu pre každý chýbajúci údaj. Používajú sa napríklad jednoduché štatistické výpočty ako aritmetický priemer.

Výhody: Nie je nutné odstraňovať údaje zo vzorku.

Nevýhody: Väčšina výsledkov používa predbežné odhady parametrov, vytvárajú štandardné chyby, pridávajú ďalšie nepresnosti k odhadom parametrov [2]. Obecne sa využívanie tejto metódy neodporúča [12, 9, 13].

Obecné nevýhody imputačných technik podľa Allisona sú, že vedú k podhodnoteniu štandardných chýb a tým prispievajú k nadhodnoteniu štatistických testov [5]. Môžeme si to v praxi predstaviť, že doplnená dáta sú úplne určené modelom aplikovaným na pozorované údaje, to znamená, neobsahujú žiadnu chybu [5].

Regresná imputácia – Podmienená priemerná imputácia

Metóda nahrádza chýbajúcu hodnotu hodnotou s predpokladaným skóre z regresnej rovnice. Základná myšlienka je použitie informácie z úplnej premennej na index v neúplných premenných. Medzi premennými môže byť korelácia, hodnoty vypožičiavajú informácie z pozorovaných údajov. Metóda je zložená z nasledujúcich krokov. Odhadnutie súboru regresných rovníc a predpovedanie neúplnej premennej z úplných premenných. Vygenerovanie predpokladaných hodnôt pre neúplné premenné a spracovania dát.

Výhody: Metóda má presnejšie výsledky ako jednoduchá imputácia [2].

Nevýhody: Môže viesť k výrazným odchýlkam vo výsledkom [2].

Stochastická regresná imputácia

Metóda využíva regresné rovnice k predikcii neúplných premenných z úplných premenných a zároveň je rozšírená o krok, ktoré rozširuje skóre s normálovým reziduálnym výrazom. Pridaním rezídua sa obnoví variabilita údajov a je možné efektívnejšie eliminovať štatistickú odchýlku.

Výhody: Jediná imputačná metóda k nepravidelným odhadom parametrov v mechanizme MAR. Z publikovaných štúdií vyplýva, že metóda prináša nezaujaté hodnoty [3].

Nevýhody: Oslabuje štandardné chyby, tým zvyšuje možnosť ich výskytu [2].

Imputácia Hot-Deck

Hot Deck imputácia je súbor techník, ktoré doplňujú chýbajúce hodnoty bodmi od „podobných prípadov“ [17]. Z dostupných literárnych zdrojov vypláva, že metóda ešte nie je dostatočne metodicky podložená, na rozdiel od ostatných metód [18]. Stretneme sa tu s pojmom „darca“ a „príjemca“. Dochádza tu k nahradeniu dát príjemca od darcu. Darca môže byť vybraný náhodne zo súboru potenciálnych darcov – tzv. „donor pool“. Je možná aj varianta, že darca je identifikovaný a hodnoty sú priradené metódou „nearest neighbour“. Tieto metódy sa nazývajú deterministické, pretože neexistuje náhodný výber „darca“.

Výhody: Výsledok doplnenia dát môže byť použitý na sekundárne spracovanie, metóda sa vyhýba problému nekonzistencie medzi vzorkami, metóda je menej citlivá na výber správneho mechanizmu vzniku chýbajúcich dát [18].

Nevýhody: Metóda je obmedzená – nedostatok metodických zdrojov, nedostatočne popísaný postup pre vytváranie darcov a príjemcov [18], metóda podceňuje štandardné chyby[2], produkuje predpojaté odhady korelácií [16].

1.3.3 Štatistické metódy s využitím modelu

Metóda maximálnej vierohodnosti

Metóda maximálnej vierohodnosti (z angl. Maximum likelihood) patrí medzi moderné štatistické postupy, ktoré majú široké uplatnenie [18, 19]. V analýze je východiskovým bodom určenie distribučnej funkcie danej populácie. V spoločenských a behaviorálnych vedách sa predpokladá normálne rozloženie dát. Cieľom metódy ML je identifikovať parametre populácie, ktoré majú najvyššiu pravdepodobnosť [2].

Existujú dve hlavné metódy ML:

Priama maximálna vierohodnosť – priama maximalizácia mnohorozmerného normálneho rozloženia, pravdepodobnostná funkcia pre predpokladaný lineárny model.

Algoritmus maximalizácie očakávania, tzv EM algoritmus (Expectatiom Maximization) – odhady strednej hodnoty a tzv. kovariančnej matrix, ktorí možno použiť na získanie konzistentných odhadov požadovaných parametrov. Metóda očakáva maximalizáciu, jednotlivé kroky sa niekoľko krát za sebou opakujú, až kým sa nedosiahne maximálnej vierohodnosti .

Výhody: Odhady maximálnej vierohodnosti sú konzistentné s efektívnymi premennými [20], metóda vhodná pre mechanizmus MAR, pri mechanizme MCAR je výsledok lepši oproti tradičným technikám[2]

Nevýhody: Vyžaduje špecializovaný software, časovo náročná [14], vyžaduje veľkú vzorku dát, metóda je vhodná len pre lineárne a logaritmické modely [5], v rámci mechanizmu NMAR prináša odchýlky vo výsledkoch [2].

Bayesovská simulačná metóda

V Bayesovskej simulačnej metóde je potreba definovať si pojem parameter. Parameter je definovaný ako pevná charakteristika populácie. Cieľom metódy je odhadnúť skutočnú hodnotu parametru a vytvoriť interval založený na odhade. Metóda zobrazuje parameter ako náhodnú premennú a popisuje jej tvar. Bayesovský vierohodný interval určuje pravdepodobnosť, že vybraný parameter sa vyskytuje medzi dvoma určitými hodnotami – prisudzuje pravdepodobnosť parametru, nie dáta.

Metóda sa skladá z troch hlavných krokov:

1. Prior Distributution - Učenie hlavnej distribúcie vybraného parametru – popisuje subjektívne presvedčenie o relatívnej pravdepodobnosti parametru.

2. Využitie funkcie pravdepodobnosti pre súhrn dôkazov o rôznych hodnotách parametrov.
3. Posterior Distribution – výsledok z pravdepodobnosti určuje hornú a dolnú hranicu, vytvorenie novej distribúcie, ktorá popisuje relatívnu pravdepodobnosť parametrov.

Výhody: Zohľadňuje „nepresnosti“, doplnenie chýbajúcich hodnôt na základe parametrov [21].

Nevýhody: Zdĺhavá a nároční metód, nutnosť zdokumentovať proces výberu distribúcie [21].

Metóda s využitím algoritmu Markovovych reťazcov Monte Carlo

Metóda s využitím Markovovych reťazcov Monte Carlo (MCMC) je základnou metódou využívajúca algoritmus a vychádza z bayesovských základov [22]. Metóda viacnásobného doplnenia je podľa autorov metódou voľby ku komplexnému doplneniu dát [23–25]. Monte Carlo proces je založený na stochastickom postupe odhadu náhodných čísiel [26]. Čísla sú vypočítavané pomocou algoritmov. Metóda je iteratívna, čo znamená výpočet je realizovaný vo viacnásobných opakovaniach.

Výhody: Komplexnosť, vhodné využitie pri dobrej distribúcii dát, metóda voľby [6, 24, 27].

Nevýhody: Náročnosť na spracovanie, nutnosť rozboru a analýzy dát [6, 24, 27].

Metóda je dostupná v balíčku MICE v programe R. Kde je možné chýbajúce dáta dopočítať.

1.3.4 Rozdelenie štatistických metód podľa počtu výpočtov

Jednotlivé štatistické metódy je možné rozdeliť na:

- **Jednoduché imputácie** – do tejto skupiny patria všetky metódy, ktoré hodnoty dopočítavajú v jednom kroku (doplnenie pomocou aritmetického priemeru, výpočet regresie, stochastická regresia, Hot Deck imputácia)
- **Mnohonásobné imputácie** – metódy pracujú na dopočítavaní chýbajúcich hodnôt pri tvorbe viacerých datasetov, následne sú dáta analyzované a finálny výsledok je po spojení z viacerých dopočtov, k vyvareniu viacerých datasetov sa využívajú algoritmy (napríklad algoritmus EM, či MCMC).

Detailnejší popis mnohonásobnej imputácie, je v nasledujúcej časti kapitoly.

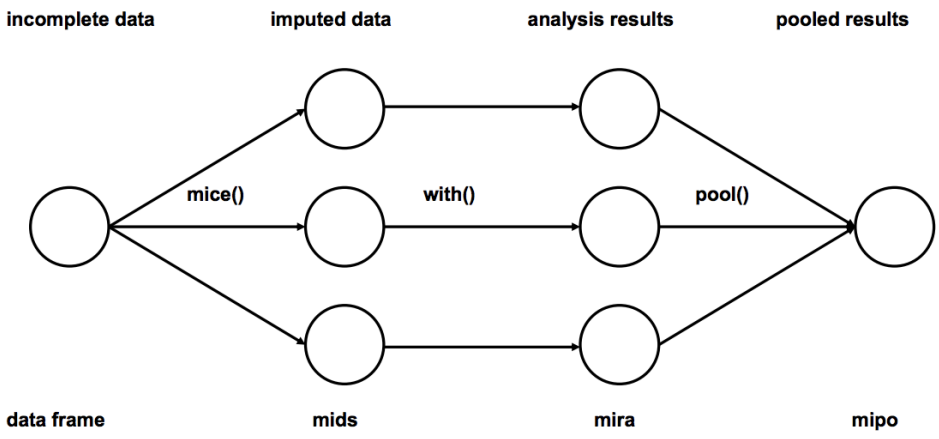
Mnohonásobná imputácia

Mnohonásobná imputácia – MI (z angl. Multiple imputation) rieši problém jednoduchej imputácie pridaním ďalšej formy chyby založenej na odchýlke odhadu parametrov v rámci doplnenia dát, tzv. „between imputation error“. Doplnené hodnoty vychádzajú z distribúcie, takže obsahujú určitú variabilitu. Metóda nahradzuje chýbajúci položku dvoma alebo viacerými prijateľnými hodnotami, čo predstavuje lepšie rozmiestnenie výsledkov [5]. Metóda je založená na modelovaní. Cieľom nie je len doplniť chýbajúce hodnoty, ale spracovať neúplné údaje s istou časťou rozptylu hodnôt na dosiahnutie platných štatistických záverov [16].

Metóda zahrňuje tri kroky, ktoré sú graficky prezentované na obrázku č. 5. Spustenie imputačného modelu definovaného vybranými premennými na vytvorenie imputovaných dátových súborov – chýbajúce hodnoty sú doplnené v čase m na vytvorenie m kompletných dáta setov. Správne voľby modelov si vyžadujú identifikáciu premenných s chýbajúcimi dátami, je nutné vypočítať podiel chýbajúcich hodnôt pre každú premennú. Následne by malo byť posúdené, či sa v údajoch vyskytujú rožne vzory chýbajúcich hodnôt.

Výhody: Viacnásobná imputácia sa môže použiť s akýmkoľvek údajmi a modelom s konvenčným štatistickým softwarom. Štúdie ukazujú, že dáta typu MAR, viacnásobné použitie MI môže viesť ku konzistentnému odhadu [19].

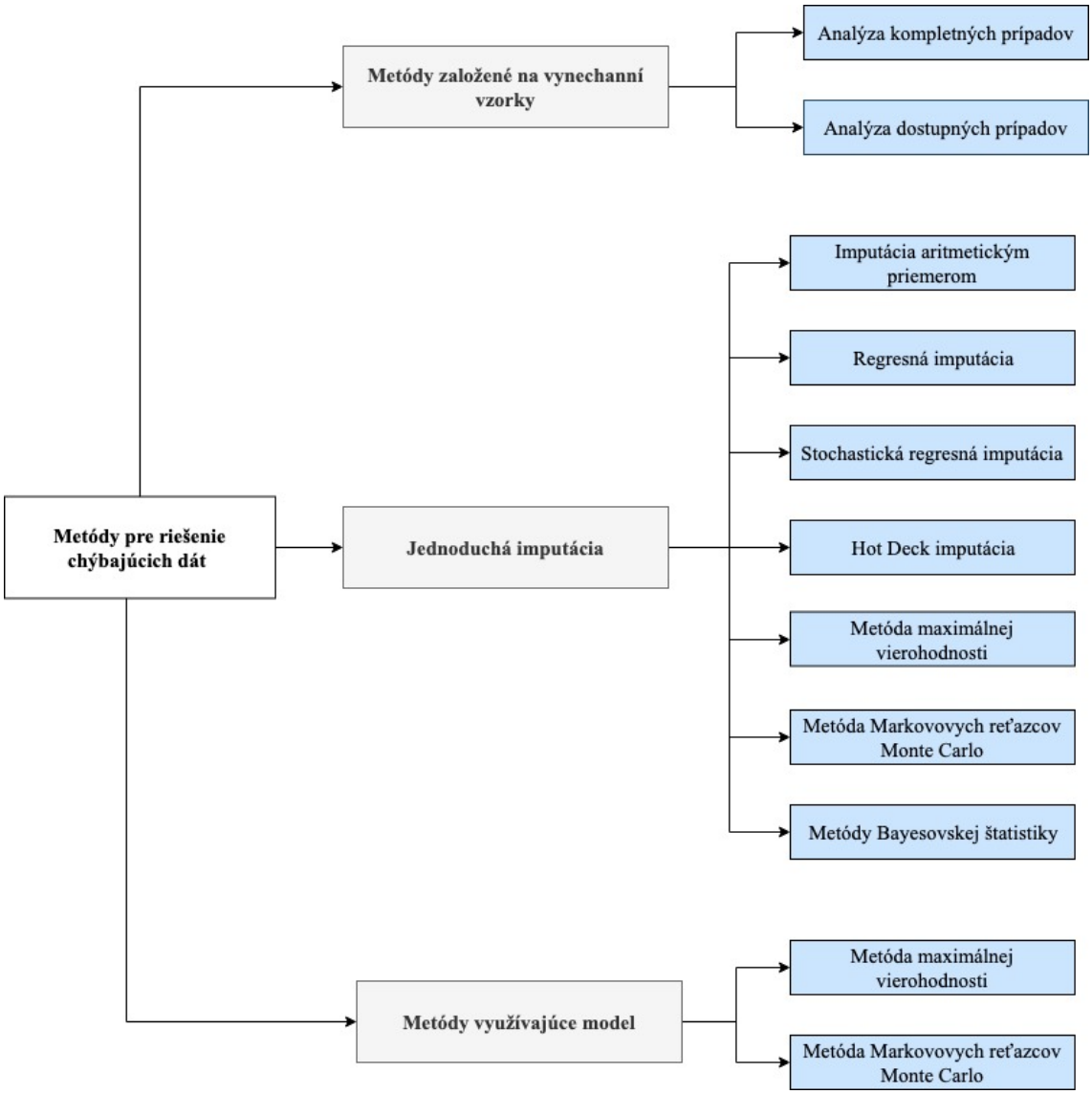
Nevýhody: Metóda je matematicky náročnejšia a viacerí autori dospeli k záverom, že pri viacnásobnom opakovaní celého procesu sa výsledky nemusia zhodovať nakoľko je výsledok podmienený stochastickým procesom [23, 28, 30].



Obrázok 5: Schéma výpočtu viacnásobnej imputácie [24]

Na obrázku č. 6 vidíme základné rozdelenie štatistických postupov k riešeniu chýbajúcich či neúplných dát

.



Obrázok 6: Základné štatistické metódy pre riešenie chýbajúcich dát [vlastné spracovanie]

1.4 Využitie štatistických metód k riešeniu chýbajúcich hodnôt

Aj napriek tomu, že štatistické postupy pre riešenie chýbajúcich dát sú staré už štyridsať rokov, tak v prostredí Českej republike je množstvo literárnych prameňov zaoberajúcich sa využitím štatistickým metód značne obmedzených. Pomocou literárne rešerše boli nájdené nasledujúce práce, či odborné publikácie:

Problém chýbajúcich dát v dotazníkových šetreniach – Autor: Iva Pecáková – článok prináša stručný náhľad do problematiky chýbajúcich dát, mechanizmy vzniku. Autorka kladne hodnotí výsledky mnohonásobnej imputácie s bayesovským prístupom na dáta chýbajúce v dotazníkovej štúdií [29].

Imputácia chýbajúcich hodnôt v rozsiahlych dátových súboroch – Autor: Markéta Nárožná – diplomová práca, práca bola spracovaná z pohľadu študentky katedry matematickej analýzy a aplikovanej matematiky Univerzity Palackého v Olomouci. Autorka využíva metódu viacnásobnej imputácie a hot deck imputácie. Metódy boli využité pre reálne dáta z prostredia životných podmienok. Z hodnotenia autorky vyplýva, že je potrebné poznať dôkladne vzťahy medzi premennými [31].

Analýza chýbajúcich hodnôt: porovnávanie vhodnosti tradičných metód naprieč mechanizmami – Autor: Ivan Petrúšek – diplomová práca pre fakultu sociálnych vied UK v Prahe – autor pomocou simulácie vytvoril všetky mechanizmy a využíval nasledovné metódy: analýzu kompletných prípadov, jednoduchú imputáciu aritmetickým priemerom, regresnú metódu a metódu stochastickej regresie. Výsledky ukazujú, že metódy pre mechanizmus MNAR vedú k vychýleniu populačných parametrov [32]. Mechanizmus MAR ukázal, že pri štatistickom testovaní hypotéz je väčší predpoklad pravdepodobnosti prvého druhu. Obecne najlepšia metóda sa javila stochastická regresná analýza

Najnovšia publikácia pochádza z pera pána Petrúška, ktorá bola vydaná Sociologickým ústavom ČR. Monografia pojednáva o základných poznatkoch, rozbore mechanizmu chýbajúcich hodnôt [22]. Autor porovnáva metódy pri skúmaní determinantov politickej znalosti a príjmu. V knihe je dostupný prehľad mechanizmov. Štatistické metódy sú využité na troch konkrétnych kazuistikách.

Z analýzy zahraničných zdrojov vyplýva, že využitie štatistických dát na riešenie chýbajúcich či neúplných dát sa môže uplatniť v klinických odvetviach. Matthieu Powney sa vo svojej práci venoval spracovaniu prehľadu klinických štúdií, ktoré nemali kompletné dáta s využitím týchto metód [33]. Ako vidíme v tabuľke číslo jedna, v praxi najčastejšie využívanou metódou, je najjednoduchšia metóda a to odstránenie neúplných súborov dát.

Carides vo svoje práci používa metódy založené na regresii na odhadnutie priemerných nákladov na ľavú ventrikulárnu disfunkciu a náklady na liečbu kolitídy [34]. V randomizovanej štúdii bolo náhodne zaradených 2569 pacientov s výskytom infarktu myokardu. Vzorka pacientov pochádzal z Kanady, Belgicka a Anglicka. Bolo porovnávanie nákladov na liečbu s využitím enalaprilu a s využitím placeba. Časť dát bolo cenzurovaných. K využitiu dát, ktoré boli primárne určené k použitia z pohľadu pacienta sa využitím regresnej metódy využili dáta získané z veľkého vzorku pacientov k odhadu priemerných lekárskych nákladov na liečbu. Metóda využívala základný vzťah – náklady na liečbu a dobu prežitia. Podľa Caridesa je metóda funkčná a jej využitím sa zvyšuje efektivita. Aplikácia metód zvýraznila vplyv odľahlých hodnôt.

Štatistická regresná metóda bola využitá aj v prípade štúdie o rakovine vaječníkov. Ako autor uvádza, pacienti s menej agresívnou formou ochorenia mali tendenciu zbierať informácie o nákladoch s nižšou sadzbou ako pacienti s agresívnou formou, ktorí majú náklady na život vyššie.

Tabuľka 1: Výsledky metanalýzy – výskyt neúplných dát a využitie štatistických metód v zdravotníctve [22]

			Využitie štatistických metód pre doplnenie hodnôt				*Vhodnosť metodiky*		
Parametre		*Počet štúdií*	*Analýza kompletných prípadov*	*Jednoduchá imputácia*	*Mnohonásobná imputácia*	*Kombinácia metód*	*Áno*	*Nie*	*Nejasné*
Počet pacientov	**1 - 100**	8	17	8	0	4	18	9	13
	101 - 200	2	8	2	2	5	9	11	1
	201 - 300	2	4	2	0	4	8	4	0
	301 - 400	1	1	1	1	1	3	3	1
	400 +	1	1	1	1	4	6	3	2
Klinická oblasť	**Duševné poruchy**	13	2	4	2	1	7	2	3
	Nádorové ochorenia	11	4	1	1	0	5	4	2
	Reumatológa	10	4	1	0	1	3	5	1
	Infekčné ochorenia	8	4	1	0	1	4	2	1
	Ochorenia kardiovaskulárneho systému	7	1	2	0	0	3	3	2
	Stomatológia	6	3	1	0	2	3	3	0
	Neurológia	6	2	0	0	0	2	2	2
	Anestéziológia a liečba bolesti	6	3	0	0	0	2	2	2
	Ostatné	33	8	4	1	4	15	10	4

Lineárna regresia bola použitá k vyhodnoteniu mezných nákladov v piatich rokoch. Autori sa prikláňajú k použitiu viac ako jednej metódy na rovnaký súbor dát, konečný výsledok má lepšiu štatistickú hodnotu. Lineárny regresný model má podľa nich však limity vážne obmedzenia, z výsledkov nebolo možné skúmať podmienené rozloženia akumulácie nákladov vzhľadom k špecifikáciám prežitia [35].

Do retrospektívnej nákladovej štúdie na liečbu bolo vybraných 773 pacientov s diagnostikovanou rakovinou pľúc, prostaty, tlstého čreva a pŕs v Amerike. Zdravotné údaje obsahovali všetky ambulantné výkony, terapeutické a diagnostické výkony. Platby Medicare boli použité ako náhrada nákladov na priame lekárske služby, ale nezahrňovali nepriame zdravotnícke náklady. Autori sa snažili určiť vplyv nezávislých premenných na celkové zdravotnícke náklady na rakovinu po dobu dvoch rokov. V štúdii použili deskriptívnu analýzu a regresnú štatistickú metódu. Z výsledkov vyplýva, že využitie regresnej metódy je štatisticky významný, špeciálne keď sa pracuje s veľkou vzorkou. Metóda je však náročná na spracovanie a jej praktické využitie je podľa Baera diskutabilné [36].

Štatistická metóda lineárnej regresie bola použitá aj u nákladovej analýzy skríningového programu u rôznych foriem rakoviny. Autor uvádza, že dlhodobé štúdie sú často neúplné, niektorí pacienti neboli sledovaní do konca liečby. Ako uvádza štúdia, regresná metodika by bola obzvlášť cenná pri identifikácii nákladovo efektívnych intervenčných alebo preventívnych programov. Výsledky sa ukázali byť konzistentné a autori navrhujú metodiku použiť, pokiaľ databáze nákladov obsahujú iba celkové náklady pre osoby, ktoré boli prítomné celé sledovanie. Odhad je efektívnejší, ak sú údaje o nákladoch zaznamenávané vo viacerých časových intervaloch [37].

Cieľom štúdie od Olsena bolo zistiť vplyv štatistických metód aplikovaných na štúdie s neúplnými údajmi na skreslenie účinku liečby v randomizovanej kontrolovanej štúdie na osteoartrózu kolena. Olsen simuloval chýbajúce údaje v dôsledku nedokonalého vyplnenia dotazníku k hodnoteniu bolesti a fyzickej funkcie. Boli využité nasledujúce metódy: ignorácia chýbajúcich dát, viacnásobná imputácia s dvoma metodickými prístupmi. Z výsledkov vyplynulo, že najlepšia imputačná metóda z hľadiska malého skreslenia výsledkov je viacnásobná regresná imputácia. Ignorovanie neúplných vzoriek bolo zhodnotené ako nevýkonné. Podľa autora, je viacnásobná imputačná metóda vhodná k zníženiu štatistického skreslenia [38].

Burtom vo svojej práci využíva štatistické metódy pre vyhodnotenie štúdie, ktorá porovnáva chemoterapiu so štandardnou paliatívnou liečbou u pacientov s pokročilou formou karcinómu pľúc. Štúdia obsahovala údaje od 115 pacientov, nie všetky však obsahovali úplné údaje o nákladoch. 85 pacientov nebolo reprezentatívnym vzorom, preto bolo nutné použiť štatistické metódy. Autor využil viacnásobnú imputáciu k imputácii hodnôt nezohľadnených v priamych zdravotníckych nákladoch, čo umožnilo vypočítať celkové náklady na ochorenie a vypracovať nákladovú efektivitu u všetkých

pacientov. Z výsledkov štúdie vyplýva, že vyžitie neúplného vzorku pacientov prinesie horší výsledok ako viacnásobná imputácia chýbajúcich dát u neúplných údajov skúmaných pacientov. Autor doporučuje metódu viacnásobnej imputácii k zvýšeniu štatistickej významnosti výsledku štúdie [39].

U štúdie k hodnoteniu vzťahu medzi hodnotou krvného tlaku a mortalitou u osôb vyšších nad 85 rokov bola využitá. Využitiu len štúdií s úplnou skladbou dát by mohlo viesť ku skresleniu výsledkov, preto sa autori rozhodli chýbajúce viacnásobnou imputáciou. Výsledky však ukázali len malé rozdiely s porovnaním s hodnotami u kompletných prípadov (vylúčenie neúplných vzorov). Podľa Buurena odhady rizika nie sú citlivé na chýbajúce údaje a využitie zložitejších modelových riešení neprináša lepšie výsledky [40].

Simons sa vo svojej štúdii zaoberá využitím metódy mnohonásobnej imputácie u chýbajúcich dát u dotazníku EQ-5D-3L. Vyhodnocuje dopad imputovania hodnôt verzus imputácia indexových hodnôt. Obe metodiky boli porovnané. Z analýzy autori vyviedli závery, že veľmi závisí na zistenom chýbajúcom dátovom vzorku. U veľkých veľkostiach vzorku – nad 500 respondentov s primárne chýbajúcimi dátami sú výsledky imputácie hodnôt alebo využitie indexových hodnôt podobné. Imputácia sa stáva presnejšia, pokiaľ je chýbajúci charakter dominantnou zložkou nezodpovedanej časti. . Využitie indexových hodnôt je efektívnejšie pre malý súbor [41].

Z literárnej rešerši vyplýva, že pre potreby klinického hodnotenia, či HTA sa využívajú klasické metódy k doplneniu chýbajúcich dát, hodnotenie využitých metód a ich hodnotenie je spracované v tabuľke č. 2. Aj napriek rozdielnym názorom na dané metódy sa autori zhodujú v jednom, že ideálne je metódam aktívne predchádzať, pretože ani jedna nemá 100% výsledok.

Tabuľka 2: Prehľad publikovaných štúdií s využitím štatistických metód [vlastné spracovanie]

Názov štúdie	*Autor*	*Rok vydania*	*Použité metódy*	*Zhodnotenie autora*
A regression-based method for estimating mean treatment cost in the presence of right-censoring	Carides et al [34]	2000	Imputácia s využitím regresie	Numerická hodnota ukazuje dobrý odhad chýbajúcich hodnôt.
Cost-effectivenes in clinical trials: using multiple imputation to deal with incomplete cost data	Burton et al [42]	2007	Mnohonásobná imputácia vs Analýza kompl. prípadov	Mnohonásobná imputácia má lepšie výsledky ako analýza kompletných príp.
Handling missing values in cost effectiveness analysis that use data from cluster randomized trials	Diáz-Ordaz et al. [43]	2012	Mnohonásobná imputácia	Model podhodnocuje neistotu, tendencia skreslenia distribúcie .
Bayesian estimation of cost-effectiveness from censored data	Heitjan et al. [44]	2004	Bayesovská metóda	Metóda je numericky náročná, výsledok závisí na designu štúdie, presný výsledok .
Linear regression of censored medical cost	Lin D. et al [37]	2000	Imputácia s využitím regresie	Regresná metóda sa javí ako nedostatočná, nutnosť použiť iné štatistické metódy.
Missing...presumed at random: cost-analysis of incomplete data	Briggs et al. [45]	2003	Bayesovská metóda Mnohonásobná imputácia	Pre mechanizmus MAR záleží na rozsahu chýbajúcich dát, u malého rozsahu je vhodná aj metóda jednoduchej imputácie. MI vytvára najmenší absolútny rozdiel medzi hodnotami.

A guide to handling missing data in cost effectiveness analysis conducted within randomised contolled trials	Rita et al.[46]	2014	Prehľad metód	Výsledok štúdie po doplnení dát by mal byť porovnaný s výsledkom z alternatívnej metódy. Metódy maximálnej vierohodnosti sú nevhodné ak sú náklady pre klinické efekty rozdielne.
Multipe imputaton of missing blood pressure covariates in survival analysis	Buureen et al.[40]	1999	Mnohonásobná imputácia	Očakávaný priemerný tlak krvi je nižší než u sledovaných dát, vypočítané dáta sú nepresné a s odchýlkami.
Multiple imputation to deal with missing EQ-5D-3L data“ Should we impute individual domains or the actual index	Simons et al.[49]	2014	Mnohonásobná imputácia	Hodnoty z MI boli podobné, dôležitá je veľkosť pozorovaného vzorku, problémy s imputáciou u malého počtu respondentov.
The impact of using different imputation method for missing quality of life score on the estimation of the cost-effectiveness of lung-volume-reduction surgery	Blough et al. [47]	2009	Mnohonásobný imputácia Regresná metóda	Nie je vhodná pre mechanizmus MCAR, obe metódy vytvárajú podobné výsledky, rozdiely vo výsledkoch sú malé.

2 Metódy

V nasledujúcej kapitole sú popísané podrobné postupy použité pre riešenie chýbajúcich hodnôt. Pre riešenie problému bola využitá štúdia pre porovnanie ekonomicko-klinického zhodnotenia endovaskulárnej a chronickej liečbe pacientov s postihnutím povrchnej stehennej tepny. Zdrojové dáta boli poskytnuté vedúcim práce Ing. Vojtěchom Kamenským. Jednalo sa o ekonomicko klinické zhodnotenie metód v terapii povrchovej stehennej tepny. V práci boli porovnávané tri metódy - perkutánna transluminárna angioplastika (PTA), perkutánna transluminárna angioplastika s následnou implantáciou stentu (PTA/s) a Bypass. Pre vypracovanie diplomovej práci sú využívané ekonomické dáta z pohľadu plátcu zdravotnej starostlivosti.

Na základe prehľadu dostupných štatistických metód pre riešenie chýbajúcich údajov boli vybrané nasledujúce metódy:

1) Analýza kompletných prípadov
2) Jednoduchá imputácia aritmetickým priemerom
3) Mnohonásobná imputácia algoritmom maximálnej vierohodnosti tzv. Expectation maximization algorithm (EM algoritmus)
4) Mnohonásobná imputácia pomocou Markovovych reťazcov Monte Carlo tzv. Markov chain Monte Carlo (MCMC algoritmus)

S analýzou kompletných prípadov a jednoduchou imputáciou aritmetickým priemerom sa môžeme stretnúť bežne v každej vedeckej praxi. Pre vypracovanie je dostačujúci bežný softvér na osobnom počítači. Tieto metódy sú spracovávané v tabuľkovom procesore Excel. Doplnenie chýbajúcich hodnôt mnohonásobnou imputáciou je potrebný štatistický softvér. V našom prípade je zvolený program R.

2.1 Zdrojové dáta

Zdrojové dáta obsahovali celkovo 134 sledovaných pacientov, 65 sledovaných pacientov liečených metódou PTA, 26 pacientov liečených metódou PTA/s a 43 pacientov liečených pomocou metódy bypass. Údaje v úplnej matici mali nasledovný charakter:

- Pohlavie – kvalitatívna binárna premenná – muž/žena
- Vek – kvantitatívna spojitá premenná, vyjadrená v rokoch
- Rok liečby - kvantitatívna spojitá premenná
- Metóda – kategorická premenná, PTA, PTA/s a Bypass
- ZUM celkom – zvlášť účtovaný materiál (ZUM) – kvantitatívna pomerovo spojitá premenná, vyjadrená v KČ
- Body ošetrovacie dni – kvantitatívna pomerovo spojitá premenná, vyjadrená v bodoch

- Body výkon celkom – kvantitatívna pomerovo spojitá premenná, vyjadrená v bodoch
- ZULP / kontrastná látka = zvlášť účtované liečebné prostriedky – kvantitatívna pomerovo spojitá premenná, vyjadrená v KČ

2.2 Program R

Program R je programovací jazyk, ktorý sa využíva pre štatistickú spracovanie dát. Pomocou R je možné údaje spracovávať a ukladať aj graficky znázorniť [48]. Program je podporovaný s operačným systémom Linux, Windows aj MacOS. Licencia je dostupná k voľnej inštalácii na stránkach http:/r-project.org. Po otvorení má program R Studio pracovné okno rozdelené, v úseku script sa zapisujú jednotlivé kroky v R kódoch, v dolnej časti je konzola, kde vidíme výsledky zadaných kódov. V pravej časti môžeme nájsť úsek s načítanými dátami, výsledky analýz či vytvorené grafy. Pre spracovanie práce sú využívané hlavne nasledujúce balíčky:

- **AMELIA II** – slúži k diagnostike chýbajúcich dát, ich nahradeniu a následnej analýze nahradených hodnôt [50] . Amelia II nahrádza chýbajúce hodnoty mechanizmom jednoduchej imputácie a viacnásobnej imputácie. Doplnok je možný riadiť pomocou R kódov v R Console, dostupná je aj funkcia Graphical User Interface (GUI), užívateľské prostredie v podobe okna, ktoré má uľahčiť spracovanie dát. K balíčku je dostupná aj tzv. vignettes – podrobný manuál práce s balíčkom v programe R.
- **MICE –** obsahuje komplexné funkcie k práci s dátovými maticami. Pomocou balíčku je možné nielen charakterizovať, vyhodnotiť charakter chýbajúcich údajov, ale aj nasimulovať vybrané mechanizmy chýbania dát. MICE umožňuje vykonať viacnásobnú imputáciu. Táto metóda je metódou voľby pre komplexné riešenie chýbajúcich dát [6, 24, 25, 27]**.**
- **VIM** – balíček VIM (z ang. Visualization and Imputation of Missing Values) je sofistikovaný nástroj k analýze chýbajúcich hodnôt ich charakteru ale aj doplneniu[51]. Balíček obsahuje pokročilé nástroje k vizualizácii chýbajúcich a doplnených hodnôt. Balíček je dostupný aj vo verzii GUI [52].
- **RCMRD** – balíček R Commander sa používa na štatistické spracovanie údajov, načítanie dát, ich úpravu, štatistické funkcie ako test normality, numerické prehľady, štatistické testy napr. párový test, ANOVA, Kruskal – Wallisov test a mnoho ďalších funkcií pre grafické spracovanie a tvorbu grafov. Rcmdr funguje v užívateľskom prostredí GUI – Graphical User Interface.

2.3 Simulácia chýbajúcich hodnôt

Kľúčový prvok pre hodnotenie chýbajúcich dát je ich generovanie podľa konkrétneho mechanizmu [27, 53, 54]. Zo zdrojových dát sú pre simuláciu chýbajúcich hodnôt vybrané údaje, ktoré budú použité pre ekonomické vyjadrenie nákladov na ošetrenie:

- **ZUM celkom = ZUM_C** – zvlášť zúčtovaný materiál – údaje sú pre PTA, PTA/s, Bypass
- **Ošetrovacie dni = BVC** – údaje sú zaznamenané pre PTA, PTA/s, bypass
- **Kontrastná látka** = **KL** – využíva sa len pri metóde PTA a PTA/s

Z každej dátovej matice je odstránené 20% zaznamenaných údajov. Simuluje sa mechanizmus MCAR – náhodné chýbanie dát. Tento mechanizmus vylučuje pravdepodobnosť závislosti chýbajúcich hodnôt so zámerným neuvedením hodnoty [6, 52, 55]. Mechanizmus bol zvolený z dôvodu, že sa jedná o ekonomické údaje, ktoré zaznamenávajú osoby, ktoré zákrok vykazujú a dá sa predpokladať, že príčiny chýbajúcich dát nie sú z nevôli zaznamenať tento údaj.

Obecné matematické vyjadrenie je možné zapísať pomocou matice. Pre štúdiu s plnými hodnotami použijeme maticu *Y,* maticu s chýbajúcimi hodnotami nazveme *M*.

Ak $Y = (y_{ij})$ označuje dátovú maticu s plnými hodnotami tak nech $M = (m_{ij})$ je indikátorová matica chýbajúcich hodnôt, kde :

$$m_{ij} = \begin{cases} 1 & y_{ij}\ chýba \\ 0 & y_{ij}\ je\ prítomná \end{cases} \qquad (2.1.)$$

Pre mechanizmus, ktorý vedie k vzniku chýbajúcich hodnôt, je potrebné dodržať podmienku rozdelenia pravdepodobnosti matice *M* za podmienky *Y*, potom :

$$f(M|Y, \emptyset)\,, \qquad (2.2.)$$

kde ∅ značí neznáme parametre.

Pre zvolené metódy jednoduchej a viacnásobnej imputácie hodnôt je potreba nasimulovať mechanizmus MCAR – hodnoty chýbajú úplne náhodne. V tomto mechanizme chýbajúce hodnoty nezávisia na pozorovaných ani na chýbajúcich hodnotách *Y*.

$$f(M|Y, \emptyset) = f(M|Y, \emptyset)\ \ pre\ všetky\ Y, \emptyset \qquad (2.3.)$$

Na simuláciu chýbajúcich hodnôt podľa presného mechanizmu je vhodné použiť funkciu balíčka MICE v programe R. Balíček si nainštalujeme kódom:

```
>install.packages(mice)
```

Aby sme boli schopný importovať súbory vo formáte .xls a meniť názvy riadkov je nutná inštalácia prídavných balíčkov readxl a dplyr. Príkaz funguje analogicky pre všetky balíčky programu R:

```
>install.packages(„názov balíčka“)
```

Postup zadania jednotlivých kódov v programe R môžeme vidieť nižšie.

Spustenie jednotlivých balíčkov:

```
>require(mice)

>require(readxl)

>require(dplyr)
```

Vyhľadanie súboru uloženého s názvom „Full_data_PTA“ na lokálnom úložisku v priečinku „DP vypracovanie“:

```
>getwd()

>file <- 'DP vypracovanie/Full_data_PTA.xlsx'

>data <- read_xlsx(file, sheet='PTA')
```

Označenie polohy stĺpcov v tabuľe kde, kde sa nachádzajú údaje pre: ZUM, Ošetrovacie dni a kontrastná látka:

```
>data2 <- data %>% select(3:5)
```

Odstránenie údajov, zvolený mechanizmus MCAR, parameter prop 0,2 určuje 20% odstránenia údajov:

```
>data3 <- ampute(data2, prop=0.2, mech = 'MCAR')
```

Uloženie novej matice údajov pre metódu PTA:

```
>data4 <- data3$amp

>write.csv(data4,'DP vypracovanie/amputed_pta.csv')
```

Pre vytvorenie novej dátovej matice s chýbajúcimi hodnotami 20% s mechanizmom MCAR má pre metódu PTA/s a bybass analogický charakter. Metóda bypass nemá v nákladoch kontrastnú látku, preto sa simulácia týka len premenných jednotiek ZUM celkom a ošetrovacie dni.

```
>require(mice)

>require(readxl)

>require(dplyr)
```

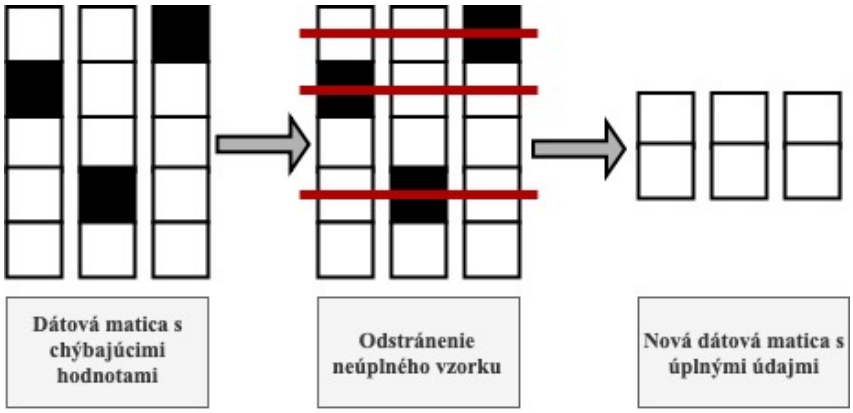
```
>getwd()
>file <- 'DP vypracovanie/Full_data_PTAS.xlsx'
>datas <- read_xlsx(file, sheet='PTAS')
>datas2 <- datas %>% select(3:5)
>datas3 <- ampute(datas2, prop=0.2, mech = 'MCAR')
>datas4 <- datas3$amp
>write.csv(datas4,'DP vypracovanie/amputed_ptas.csv')
>getwd()
>file <- 'DP vypracovanie/Full_data_BYPASS.xlsx'
>datab <- read_xlsx(file, sheet='Bypass')
>datab2 <- datab %>% select(3:4)
>datab3 <- ampute(datab2, prop=0.2, mech = 'MCAR')
>datab4 <- datab3$amp
>write.csv(datab4,'DP vypracovanie/amputed_bypass.csv‘
```

Pre kontrolu simulácie chýbajúcich dát je využitý balíček VIM nasledujúcimi príkazmi:

```
>library(vim)
>aggr(Amputed_PTA, col=mdc(1:2), numbers=TRUE,
sortVars=TRUE, labels=names(Amputed_PTA), cex.axis=.7,
gap=3, ylab=c("Proporcie chýbajúcich dát","Vzor chýbajúcich
dát"))
>aggr(Amputed_PTAS, col=mdc(1:2), numbers=TRUE,
sortVars=TRUE, labels=names(Amputed_PTAS), cex.axis=.7,
gap=3, ylab=c("Proporcie chýbajúcich dát","Vzor chýbajúcich
dát"))
>aggr(Amputed_bypass, col=mdc(1:2), numbers=TRUE,
sortVars=TRUE, labels=names(Amputed_bypass), cex.axis=.7,
gap=3, ylab=c("Proporcie chýbajúcich dát","Vzor chýbajúcich
dát"))
```

2.4 Analýza kompletných prípadov

Analýza kompletoch prípadov pracuje s prípadmi, ktoré majú platné hodnoty u všetkých premenných – to znamená, že neúplné prípady sú z analýzy odstránene. Pre lepšiu predstavu je proces znázornený na obrázku č. 7. Výsledkom je kompletná dátová matica, ktorá obsahuje však menej prípadov, avšak plná matica dovoľuje nasledujúce analytické spracovanie. Spracovanie analýz kompletných prípadov nevyžaduje žiadne špeciálne matematické spracovanie a k jej vypracovaniu je dostačujúci aj tabuľkový procesor excel.

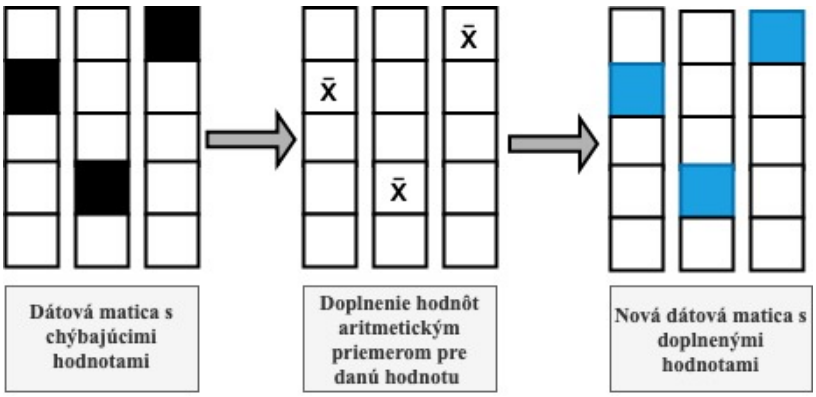


Obrázok 7 : Grafické znázornenie analýzy kompletných prípadov [vlastné spracovanie]

2.5 Jednoduchá imputácia aritmetickým priemerom

V tejto metóde sú pre prípady s chýbajúcimi kvantitatívnymi premennými údaje doplnené hodnotou aritmetického priemeru [6]. Aritmetický priemer sa vypočíta zo všetkých platných hodnôt. Matematicky môžeme výpočet zapísať nasledovne:

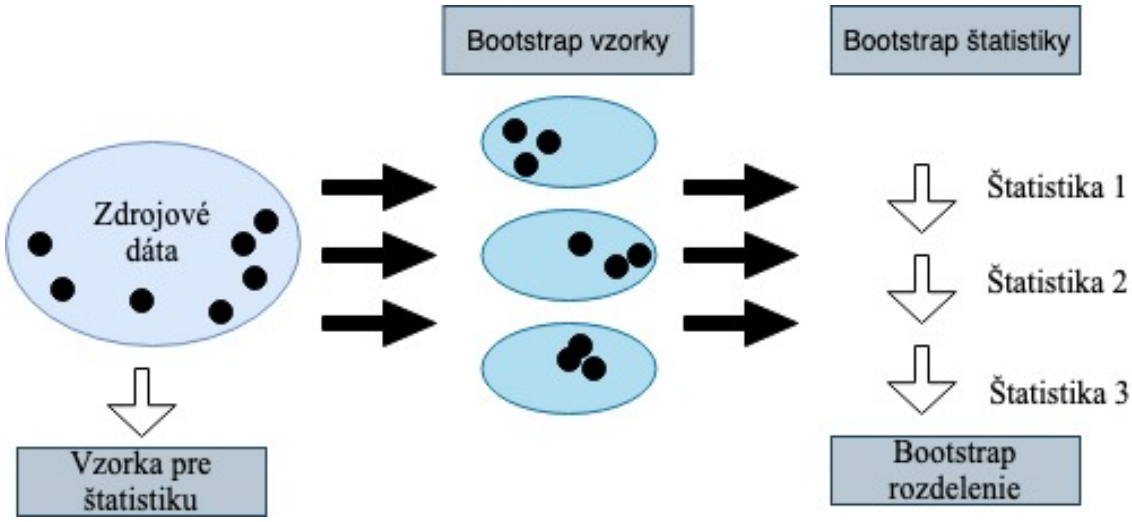
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad (2.5.)$$



Obrázok 8 : Grafické znázornenie jednoduchej imputácie aritmetickým priemerom [vlastné spracovanie]

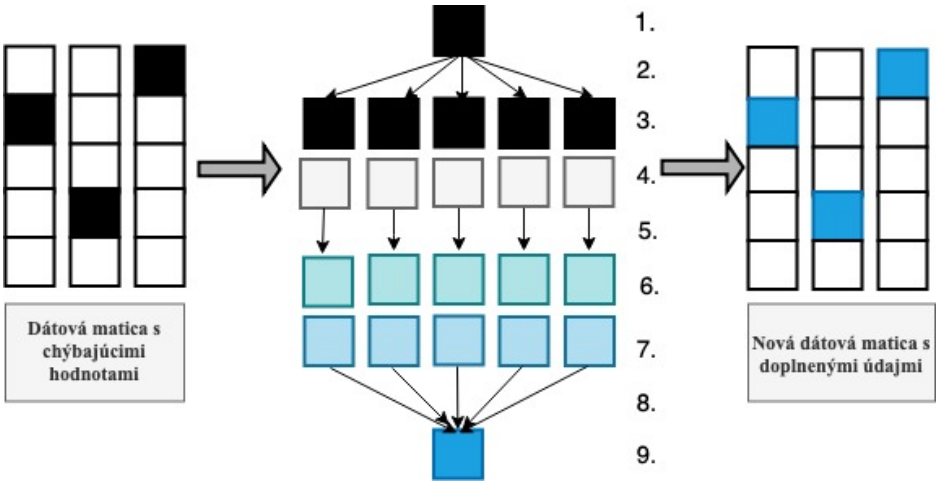
2.6 Mnohonásobná imputácia s využitím EM algoritmus

Označenie EM logaritmu pochádza z anglického spojenia expectation maximalizations. Skladá sa z E kroku a M kroku. Metóda je iteratívna, čo označuje po sebe opakujúci sa proces. E krok odhaduje tzv. postačujúcu štatistiku – sumarizácia všetkých informácií určený k výpočtu parametru [50]. Vo výpočte sa používajú hodnoty vektorov priemerov a pomocou regresnej metodiky sú hodnoty nahradzované [12]. Následne hodnoty získané regresnými výpočtami vstupujú do kroku M. M krok odhaduje parametre pomocou metódy maximálnej vierohodnosti. Pre výpočet chýbajúcich hodnôt pomocou EM logaritmu je využitý balíček AMELIA II v programe R. AMELIA II je oproti pôvodnej verzii balíčka doplnená o metódu bootstrap – intenzívna metóda pre štatistickú analýzu [56]. Bootstrapping je typ prevzorkovania, kde sa veľké množstvo menších vzoriek rovnakej veľkosti opakovane odoberá s nahradením z jednej originálnej vzorky [57]. Vysvetlenie metódy bootstrap môžeme vidieť na obrázku č. 9.



Obrázok 9: Grafické znázornenie metódy bootstrap [vlastné spracovanie]

Po využití metódy bootstrap je následne pomocou EM algoritmu vytvorených mnoho dátových matíc s doplnenými hodnotami. Následne je možné opakovať proces doplnenia hodnôt na pôvodnej matici s chýbajúcimi hodnotami. Balíček AMELIA II umožňuje vyhodnocovať jednotlivé imputácie a následne je možné jednotlivé datasety skombinovať do finálnej doplnenej dátovej matice. Celý proces využitia metódy EM algoritmu pomocou balíčka AMELIA II je znázornený na obrázku č. 10.



Obrázok 10: Grafické znázornenie metódy AMELIA II - 1. nekompletný dataset, 2. bootstrap, 3. bootstrap dáta, 4. EM algoritmus, 5. doplnené dátové matice, 6. analýza, 7. rozlíšenie výsledkov 8. kombinácia, 9. finálny výsledok [vlastné spracovanie]

Po nainštalovaní balíčka AMELIA II boli v programe R použité nasledujúce príkazy k vypracovaniu:

Načítanie balíčkov a dát:

```
>require(Amelia)

>require(readxl)

>require(dplyr)

>getwd()

>file <- 'DP vypracovanie/DATA_DP_LL.xlsx'

>data <- read_xlsx(file, sheet='Missing Data_PTA')
```

R kódy pre intervenciu PTA:

Bounds – ohraničenie, je nutné zadať, aby systém generoval len pozitívne hodnoty, nakoľko chýbajúce dáta sú náklady – tzn. nemôžu byť negatívne.

```
>bounds <- matrix(c(1L, range(data$ZUM_C, na.rm=T), 2L, range(data$BVC, na.rm=T), 3L, range(data$KL, na.rm=T)), nrow=3, byrow=T)
```

Príkaz m nám určuje počet vytvorených nasimulovaných datasetov:

```
>amel_PTA <- amelia(data, m=10, bounds=bounds)
```

Grafické znázornenie doplnených hodnôt v porovnaní s pôvodným datasetom s chýbajúcimi hodnotami. Doplnené premenné sú označené nasledovne:

ZUM_C = ZUM celkom

BVC = Body výkon celkom

KL = Kontrastná látka

```
>plot(amel_PTA)
>compare.density(amel_PTA, var= "ZUM_C")
>compare.density(amel_PTA, var= "BVC")
>compare.density(amel_PTA, var= "KL")
```

Následne je aplikované doplnenie dát ešte v jednom opakovaní v počte 10 vytvorených doplnených datasetov:

```
>amel2_PTA <- amelia(data, m=10, bounds=bounds)
```

Pre vytvorenie výsledného datasetu, ktorý kombinuje oba výsledky z dvoch opakovaní 10-tich vložení bol zadaný nasledovný R kód:

```
>result_PTA<- ameliabind(amel_PTA, amel2_PTA)
```

Pre uloženie nového súboru vo formáte ccv:

```
>setwd("~/Desktop")
>write.amelia(obj=result_PTA,file.stem="AMELIA_PTA",separat
e = TRUE)
```

Doplnenie chýbajúcich hodnôt pre metódu PTA/s a Bypass bolo analogické s rozdielom, že hodnoty KL – kontrastná látka sa pre metódu bypass nesimulovali:

R kódy pre intervenciu PTA/s:

```
>require(Amelia)
>require(readxl)
>require(dplyr)
>getwd()
>file <- 'DP vypracovanie/DATA_DP_LL.xlsx'
>data_PTAS <- read_xlsx(file, sheet='Missing Data_PTAS')
```

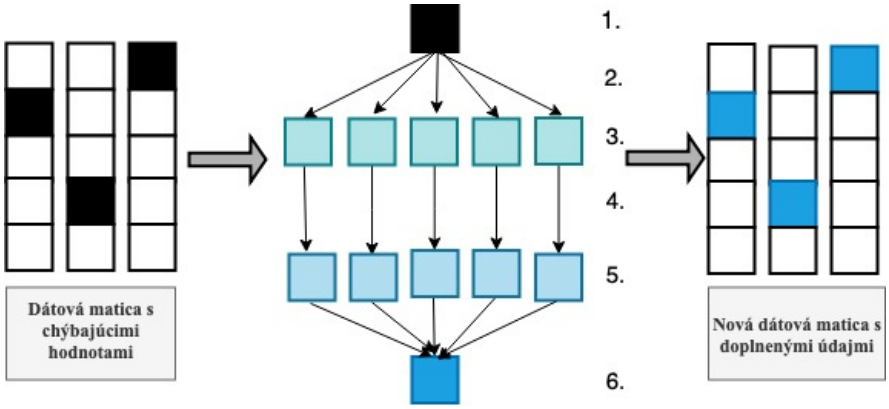
```
>bounds <- matrix(c(1L, range(data$ZUM_C, na.rm=T), 2L,
range(data$BVC, na.rm=T), 3L, range(data$KL, na.rm=T)),
nrow=3, byrow=T)

>amel_PTAS <- amelia(data, m=10, bounds=bounds)

>plot(amel_PTAS)

>compare.density(ame_PTAS, var= "ZUM_C")

>compare.density(amel_PTAS, var= "BVC")

>compare.density(amel_PTAS, var= "KL"):

>amel2_PTAS <- amelia(data_PTAS, m=10, bounds=bounds)

>result_PTAS<- ameliabind(amel_PTA, amel2_PTA)

>setwd("~/Desktop")

>write.amelia(obj=result_PTAS,file.stem="AMELIA_PTAS",separ
ate = TRUE)
```

R kódy pre intervenciu bypass:

```
>require(Amelia)

>require(readxl)

>require(dplyr)

>getwd()

>file <- 'DP vypracovanie/DATA_DP_LL.xlsx'

>data_BYPASS <- read_xlsx(file, sheet='Missing Data_BYPASS')

>bounds <- matrix(c(1L, range(data$ZUM_C, na.rm=T), 2L,
range(data$BVC, na.rm=T)), nrow=3, byrow=T)

>amel_BYPASS <- amelia(data_BYPASS, m=10, bounds=bounds)

>plot(amel_BYPASS)

>compare.density(ame_BYPASS, var= "ZUM_C")

>compare.density(amel_BYPASS, var= "BVC")

>amel2_BYPASS <- amelia(data_BYPASS, m=10, bounds=bounds)

>result_BYPASS<- ameliabind(amel_BYPASS, amel2_BYPASS)

>setwd("~/Desktop")

>write.amelia(obj=result_BYPASS,file.stem="AMELIA_BYPASS",s
eparate = TRUE)
```

2.7 Mnohonásobná imputácia s využitím Markovovych reťazcov Monte Carlo

Metóda Markovovych reťazcov Monte Carlo sa zaraďujú k bayesovskej štatistike [24]. Základná myšlienka metódy je generovanie výberu z pravdepodobnostného rozdelenia, ktoré sa nazýva Markovove reťazce. Markovov reťazec je stochastický proces, to znamená, že každé opakovanie prinesie odlišný výsledok a predpokladá sa, že závisí iba na súčasnom stave, nie na minulom stave [40]. Algoritmus využívaný v balíčku MICE pracuje s predpokladom, že vedomie podmieneného rozdelenia jednotlivých premenných v dátovej matici, dovoľuje určiť ich spoločné rozdelenie [24]. Metóda MICE sa zaraďuje medzi iteratívne, prvá slučka postupne nahradzuje chýbajúce hodnoty a druhá slučka opakuje proces až po dobu, kedy sa nedokončí požadovaný počet iterácií (opakovaní), ktorý je potreba si zvoliť. Balíček MICE je navrhnutý tak, že je nutné dodržať postup pre vypracovanie: viacnásobne generovať doplnené datasety, následne datasety štatisticky spracovať a vhodné datasety s vyhovujúcimi hodnotami použiť do finálneho súboru. Postup je zobrazený na obrázku č. 11:



Obrázok 11: Grafické znázornenie použitia balíčka MICE - 1. Dátová matica s chýbajúcimi hodnotami, 2. Využitie príkazu Mice(), 3. doplnené datasety, 4. štatistická analýza, 5. výsledok štatistickej analýzy, 6. príkaz Pool() k vytvoreniu finálnej dátovej matice [vlastné spracovanie]

Pre vypracovanie bolo zvolené generovanie 10 doplnených dátových matíc s 50-timi opakovaniami doplnenia.

V programe R boli zadané nasledujúce príkazy pre metódu PTA:

Pre spustenie a načítanie dát:

```
>require(mice)

>require(dplyr)

>require(glue)
```

```
>library(readxl)
>getwd()
>file <- 'DP vypracovanie/DATA_DP_LL.xlsx'
>data_PTA <- read_xlsx(file, sheet='Missing Data_PTA')
```

Pre doplnenie chýbajúcich kvantitatívnych údajov pomocou metódy MCMC je nutné definovať m = počet vytvorený nových doplnených súborov, maxit = počet opakovania výpočtu pre jednotlivé datasety:

```
>Imputed_data_PTA <- mice(data_PTA, m=10, maxit = 50, method
= 'pmm', seed = 500 , diagnostics = TRUE , printFlag = TRUE)
```

Pre zobrazenie priebehu iterácii bol zadaný R kód:

```
>plot(Imputed_data_PTA)
```

Využitie funkcie ttest pre štatistické spracovanie vytvorených datasetov pre jednotlivé chýbajúce premenné:

ZUM_C = ZUM celkom

BVC = Body výkon celkom

KL = Kontrastná látka

```
>ttest_PTA<-with(Imputed_data_PTA, t.test(ZUM_C + BVC + KL))
```

Vytvorenie konečného doplneného dátového súboru:

```
> MICE_FINAL_PTA <- summary(pool(ttest_PTA))
```

Uloženie novovytvoreného súboru:

```
>write.csv2(MICE_FINAL_PTA,"Desktop/DP vypracovanie/
MICE_FINAL_PTA.xlsx")
```

Intervencia PTA/s v programe R mala nasledujúce R kódy:

```
>require(mice)

>require(dplyr)

>require(glue)

>library(readxl)

>getwd()

>file <- 'DP vypracovanie/DATA_DP_LL.xlsx'

>data_PTAS <- read_xlsx(file, sheet='Missing Data_PTAS')

>Imputed_data_PTAS <- mice(data_PTA, m=10, maxit = 50,
method = 'pmm', seed = 500 , diagnostics = TRUE , printFlag
= TRUE)

>plot(Imputed_data_PTAS)

>ttest_PTAS <-with(Imputed_data_PTAS, t.test(ZUM_C + BVC +
KL))

> MICE_FINAL_PTAS <- summary(pool(ttest_PTAS))

>write.csv2(MICE_FINAL_PTAS,"Desktop/DP vypracovanie/
MICE_FINAL_PTAS.xlsx")
```

Intervencia bypass: analogické zadávanie kódov, s rozdielom v simulovaných hodnotách, doplňujú sa hodnoty pre ZUM celkom a body výkon celkom:

```
>getwd()

>file <- 'DP vypracovanie/DATA_DP_LL.xlsx'

>data_BYPASS <- read_xlsx(file, sheet='Missing Data_BYPASS')

>Imputed_data_BYPASS <- mice(data_BYPASS, m=10, maxit = 50,
method = 'pmm', seed = 500 , diagnostics = TRUE , printFlag
= TRUE)

>plot(Imputed_data_BYPASS)

>ttest_BYPASS <-with(Imputed_data_BYPASS, t.test(ZUM_C +
BVC))

> MICE_FINAL_BYPASS <- summary(pool(ttest_BYPASS))

>write.csv2(MICE_FINAL_BYPASS,"Desktop/DP vypracovanie/
MICE_FINAL_BYPASS.xlsx")
```

2.8 Výpočet nákladov

K simulácii chýbajúcich hodnôt a následnému spracovaniu boli použité náklady z perspektívy platca zdravotného poistenia. Náklady, ktoré vstupujú do výpočtu celkových nákladov môžeme rozdeliť na tie, vyjadrené v KČ – ZUM Celkom (ZUM_C) a ZULP / náklady na kontrastnú látku (KL) a na náklady vyjadrené v bodoch – body za vykázané ošetrovacie dni (OŠD) a celkové body vykázané za výkon (BVC). Výpočet nákladov je možné vyjadriť vzťahom.

$$Celkové\ náklady = ZUM\_C + KL + OŠD * hodnota\ bodu + BVC * hodnota\ bodu \quad (2.10.)$$

Presnú hodnotu bodu v danom roku je možné získať z úhradovej vyhlášky. V práci bola počítaná hodnota bodu za rok 2010 0,91 a v rokoch 2011-2013 0,90.

2.9 Štatistická analýza nákladov

Pre každú zvolenú metodiku riešenia chýbajúcich hodnôt boli vypočítané celkové náklady pre každú klinickú metódu terapie stehennej tepny. Pre jednotlivé výsledky je spracovaná popisná štatistika. Pre potvrdenie či zamietnutie normálneho rozloženia dát je využitý Shapiro-Wilkov test normality.

Shapiro-Wilkov test normality

Pomocou testu analyzujeme rozptyl – porovnanie teoretických a empirických kvantilov. Shapiro Wilkov test využíva, tzv. Q-Q plot – kvantil kvantilový graf, kde sa vyhodnocujú vzdialenosti jednotlivých bodov od regresnej priamky [9].

Nakoľko predpokladáme, že výsledné hodnoty nebudú mať normálne rozloženie, k štatistickému porovnaniu vypočítaných nákladov jednotlivých metód doplnenia dát je zvolený neparametrický štatistický test. Pre odvodenie výsledku neparametrického testu nie je nutné špecifikovať typ rozdelenia .

Párový Wilcoxonov test

Zvolený typ testu sa využíva v porovnaní dvoch výberov v danom kvantitatívnom znaku, v našom prípade výpočet nákladov z výstupov metód riešenia chýbajúcich dát. Vypracovanie pracuje s rozdielmi medzi jednotlivými porovnávanými pármi. A stanovuje sa či je rozdiel štatisticky významný. Sledujeme, či sa charakteristika jednotlivých výsledkov mení [10].

2.10 Analýza nákladové efektivity

Analýza nákladovej efektivity (CEA z ang. Cost effectivenes analysis) slúži k ohodnoteniu vynaložených nákladov k dosiahnutému efektu. Metoda je vhodná k porovnaniu dvoch alebo viacerých technológií. V našom prípade PTA vs PTA/s a Bypass. Ukazateľ CEA je tzv. kritérium efektívnosti, ktoré je možné vyhodnotiť pomocou nákladov na jednotku výstupu (nákladová efektivita) alebo v prevrátenej forme vzorca – efektívnosť na peňažnú jednotku.

$$CE = \frac{Cost_{int}}{Effect_{int}} \tag{2.11.}$$

Cost $_{int}$ = náklady na intervenciu (KČ)

Effect $_{Int}$ = efekt intervencie (jednotka efektu)

Hodnotením zisťujeme, ktorá zdravotnícka technológia je dominantná, to znamená je menej nákladná a prináša vyšší efekt. Ak je hodnotená zdravotnícka technológia nákladnejšia oproti komparatérovi a zároveň generuje väčší efekt, tak sa na vyhodnotenie vypočítava hodnota ICER (z ang. Incremental cost-effectivenes ratio) [7].

ICER vypočítame nasledujúcim vzťahom:

$$ICER = \frac{Cost_{int} - Cost_{Comp}}{Effect_{int} - Effect_{Comp}} \tag{2.12.}$$

Cost $_{int}$ = náklady na intervenciu (KČ)

Cost $_{int}$ = náklady na komparátora (KČ)

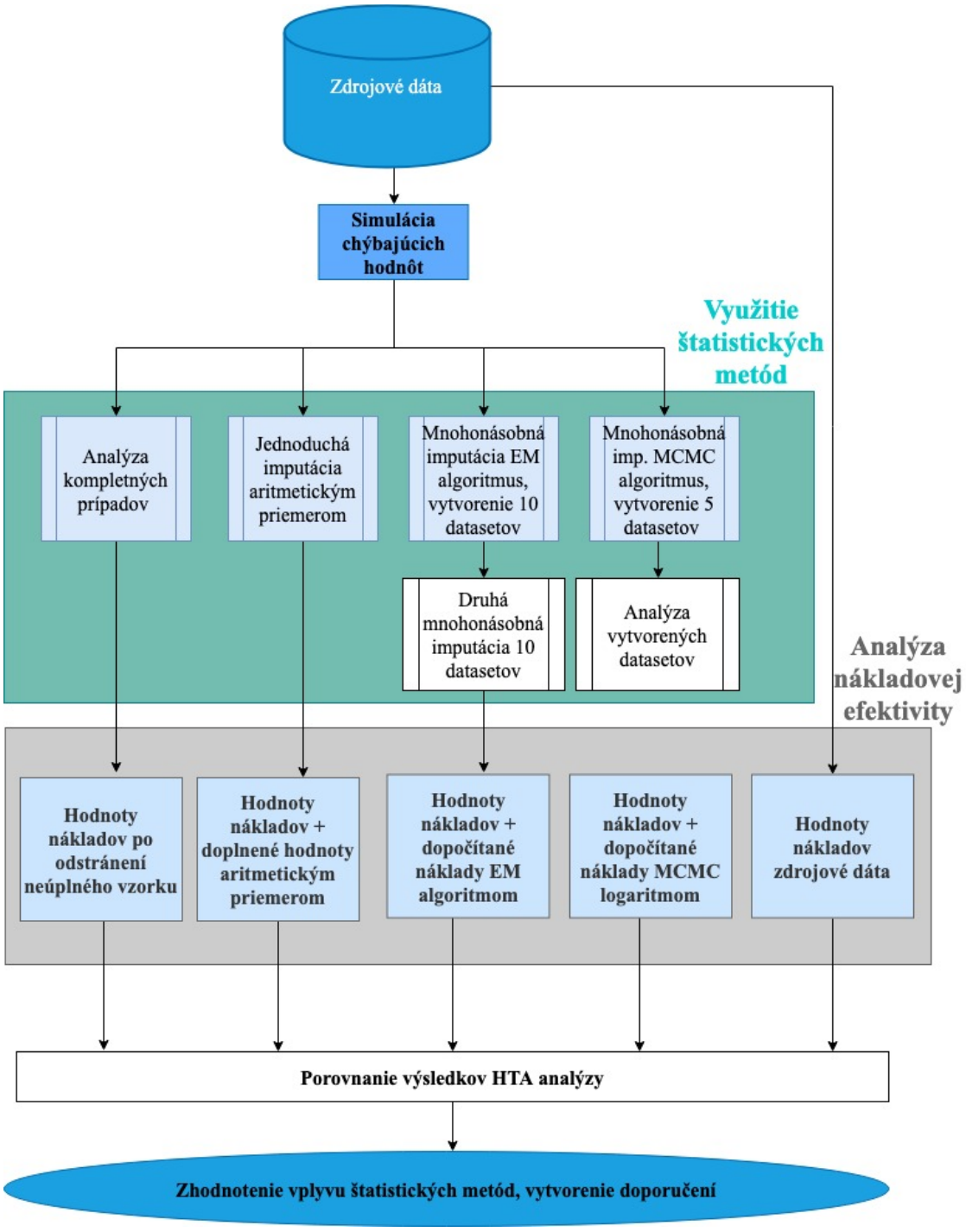
Effect $_{Int}$ = efekt intervencie (jednotka efektu)

Effect $_{Int}$ = efekt komparátora (jednotka efektu)

Ukazateľ ICER vyjadruje množstvo monetárnych jednotiek. ktoré je treba vynaložiť k získaniu ďalšej jednotky výstupu [7].

2.11 Procesný graf vypracovávania práce

Grafické znázornenie postupu vypracovávania práce



Obrázok 12: Postup spracovania práce [vlastné spracovanie]

3 Výsledky

V tejto kapitole môžeme nájsť všetky výsledky pre metódy využité v riešení problematiky chýbajúcich údajov. Na začiatku kapitoly môžeme nájsť grafické znázornenie simulácie chýbajúcich údajov pre jednotlivé metódy terapie povrchovej stehennej tepny. Následne sú výsledky jednotlivých nákladov štatisticky spracované. V poslednej časti sú vyobrazené výsledky nákladovej analýzy a percentuálne rozdiely pre jednotlivé metódy riešenia chýbajúcich dát.

3.1 Simulácia chýbajúcich hodnôt

Zdrojové dáta využité pre spracovanie obsahovali anonymizované dáta v rozsahu 134 pacientov. Porovnávali sa výsledky dva intervenčné prístupy – chirurgický a endovaskulárny.

Metódou PTA bolo zaznamenaných 65 pacientov, pre metódu PTA/s 26 a pre metódu Bypass 43. Pre jednotlivých pacientov boli doložené nasledujúce premenné:

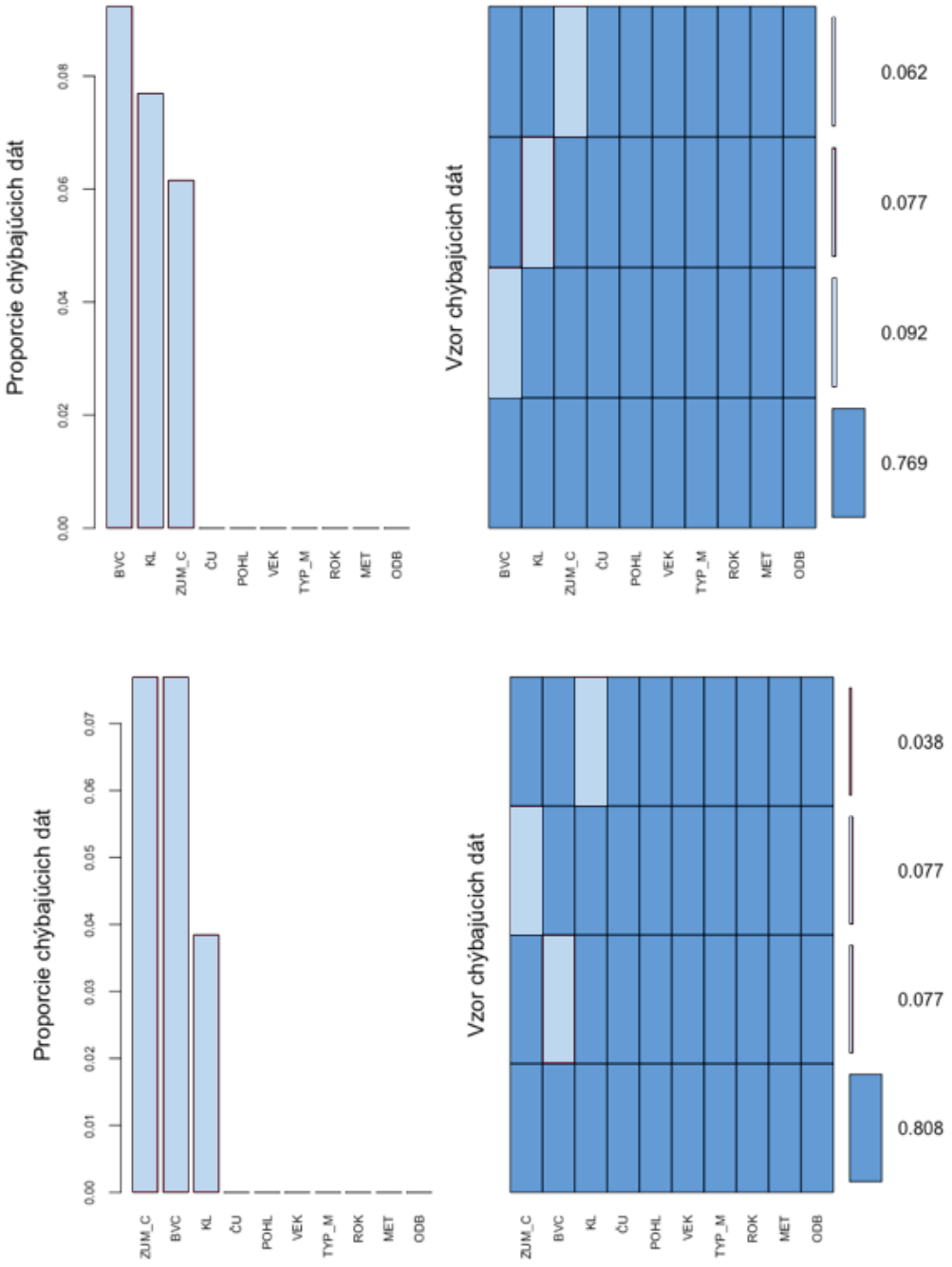
- BVC – Body vykázané za výkony celú liečbu,
- KL – Náklady na podanie kontrastnej látky,
- ZUM_C – ZUM Celkom – body vykázané za celú liečbu,
- ČÚ – Číslo účtu – identifikácia pacienta,
- VEK – Vek pacienta,
- TYP_M – Typ metódy – Chirurgická / Endovaskulárna,
- ROK – Rok, kedy prebehla liečba,
- MET – Metóda terapie – PTA, PTA/s, Bypass,
- OBC – Ošetrovacie body vykázané za celú liečbu.

Pre simuláciu chýbajúcich hodnoť boli zvolené premenné jednotky vyjadrujúce náklady a počet bodov za výkon:

- ZUM_C,
- BVC,
- KL.

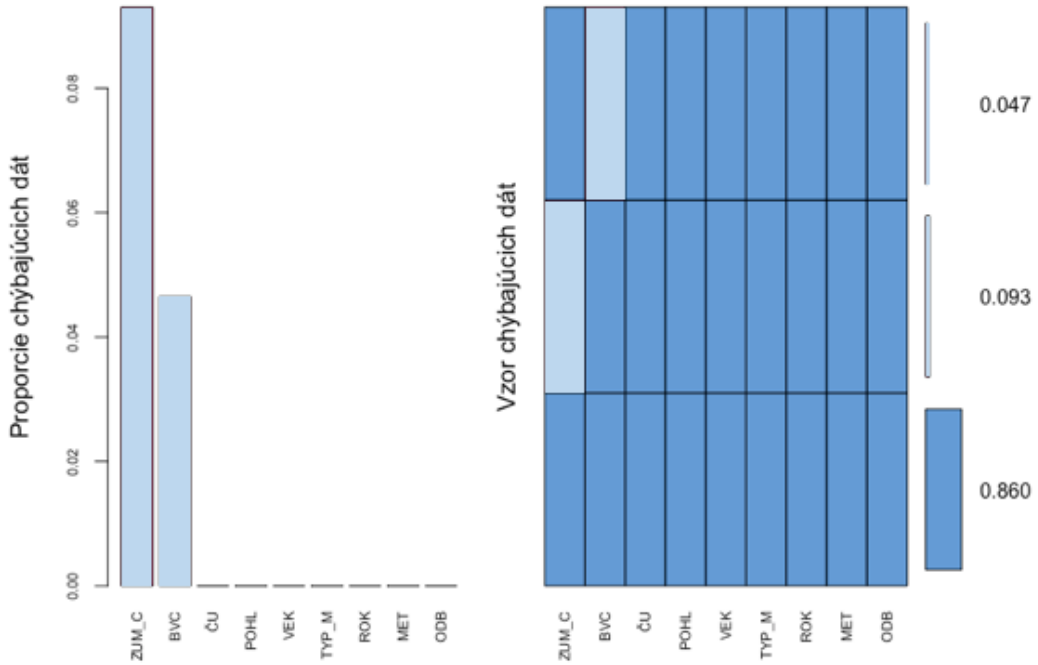
Pre všetky tri premenné bolo v programe R príkazom zvolený rozsah chýbajúcich dát 20 %. Rozsah chýbajúcich hodnôt sa vzťahuje na celú dátovú maticu. Tento rozsah bol vytvorený pomocou balíčka MICE. Pomocou balíčka VIM boli pre chýbajúce dáta vytvorené grafické znázornenia. Jednotlivé premenné slúžia výpočtom nákladov, a nie je medzi nimi nejaká predpokladaná závislosť. Mechanizmus chýbajúcich hodnôt bol zvolený MCAR – úplne náhodné chýbanie údajov. Pre každú intervenciu bola simulácia spustená individuálne. Jednotlivý výsledok náhodného odstránenia hodnôt je zobrazený na nasledujúcich obrázkoch. Kde ľavá časť prezentuje rozsah chýbajúcich hodnôt

v dátovej matice, na ose X vidíme premenné a osa Y veľkosť odstránených údajov. Pravá časť obrázkov vyobrazuje vzor chýbajúcich hodnôt v dátovej matici. Na obrázku číslo 13 a 14 vidíme akým spôsobom balíček MICE náhodne odstránil z troch vybraných premenných jednotiek údaje. Zvyšné údaje sú v nezmenenej forme, čo je viditeľné na ose Y. Z grafu je znateľné aj, že odstránenie je náhodné a aj pri zadaní rozsahu 20 % je výsledok odlišný. Pre PTA je rozsah chýbajúcich hodnôt 23 % a pre PTA/s 19 %.



Obrázok 14: Chýbajúce hodnoty v súbore pre PTA/s [vlastné spracovanie]

Pre chirurgickú intervenciu bypass sa k odstráneniu dát použili dve premenné. Hodnota pre kontrastnú látku sa v matici nevyskytovala. Po zadaní rozsahu k odstráneniu 20 % dát má výsledná dátová matica 14 % neúplných údajov. Tento výsledok je prezentovaný na obrázku číslo 15.



Obrázok 15: Chýbajúce hodnoty v súbore pre bypass [vlastné spracovanie]

3.2 Analýza kompletných prípadov

Analýza kompletných prípadov sa zaraďuje medzi metódy, ktoré odstraňujú neúplné prípady. V našom prípade to znamená odstránenie celej informácie pre každú premennú, pokiaľ sa v ňom vyskytoval chýbajúci údaj. Pre metódu PTA sa z dátovej matice odstránilo 15 prípadov, veľkosť matice sa znížil z 65 na 50. Pre intervenciu PTA/s po odstránení zostalo 21 kompletných prípadov z pôvodných 26 meraní. Veľkosť dátovej matice sa pre metódu bypass zmenšila z 43 prípadov na 38 prípadov s plnými dátami. V percentuálnom vyjadrení sa veľkosť dátovej matice pre PTA zmenšila o 23 % pre metódu PTA/s o 19 % a o 11 % pre metódu bypass.

3.3 Jednoduchá imputácia aritmetickým priemerom

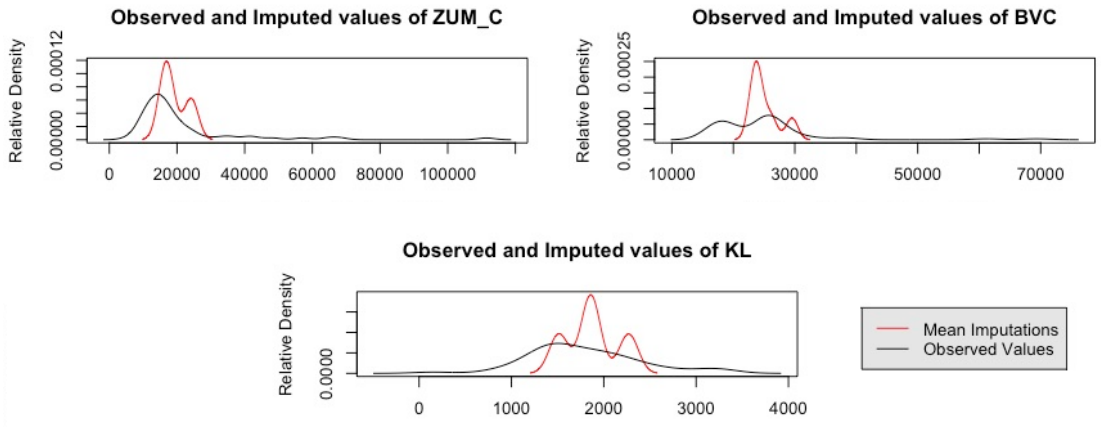
Metóda jednoduchej imputácie aritmetickým priemerom je založená na nahradení chýbajúcich hodnôt aritmetickým priemerom získaných zo zaznamenaných hodnôt. Konkrétne vypočítané hodnoty, ktoré boli použité na riešenie práce sú znázornené v tabuľke číslo 3.

Tabuľka 3: Použité priemerné hodnoty k doplneniu chýbajúcich dát

Metóda	*ZUM Celkom (KČ)*	*Body výkon celkom (Body)*	*Kontrastná látka (KČ)*
PTA	21 471	24 992	1 802
PTA/s	55 196	30 959	2 023
Bypass	18 287	34 227	

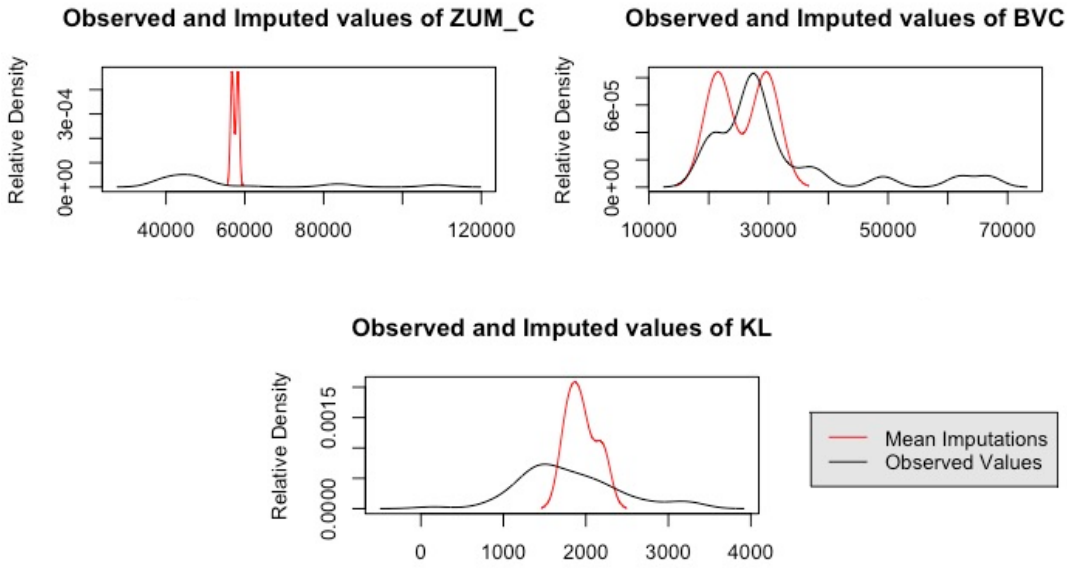
3.4 Mnohonásobná imputácia EM algoritmom

Metóda mnohonásobnej imputácie využíva metódu maximálnej vierohodnosti. Metóda sa snaží odhad parametrov prispôsobiť čo najviac pozorovaným hodnotám [22]. Konkrétne doplnenie údajov v matici neúplných dát prebiehalo v programe R, pomocou balíčka AMELIA II. Simulácia bola spustená dva krát a každý vytvorený set obsahoval 10 vytvorených datasetov. Finálny výsledok je vytvorený spojením všetkých datasetov. Pre každú intervenciu a jej premennými s chýbajúcimi hodnotami bol vytvorený graf hustoty pozorovaných jednotiek. Na grafe je vyobrazený aj priebeh priemerných dopočítaných hodnôt. Priebeh zaznamenaných hodnôt a doplnených hodnôt pre intervenciu PTA je prezentovaný na obrázku č. 16. Na obrázku vidíme tri grafy pre premenné ZUM_C, BVC a KL, na ose Y vidíme relatívnu hustotu rozloženia na ose X hodnoty (náklady) jednotlivých premenných. Tmavá linka prezentuje údaje, ktoré sú zaznamenané a červené priebeh doplnenia. Môžeme porovnať tvar línií hustôt. Na jednotlivých znázorneniach je viditeľný rozdiel v doplnení údajov. Pre premennú ZUM_C krivka doplnených hodnôt kopíruje tvar zaznamenaných údajov, obdobne tomu je aj v prípade premennej KL. Doplnené hodnoty pre premennú BVC majú mierne odlišný tvar krivky ako pôvodné zaznamenané hodnoty.



Obrázok 16: Porovnanie hustoty dostupných údajov a údajov doplnených pre intervenciu PTA [vlastné spracovanie]

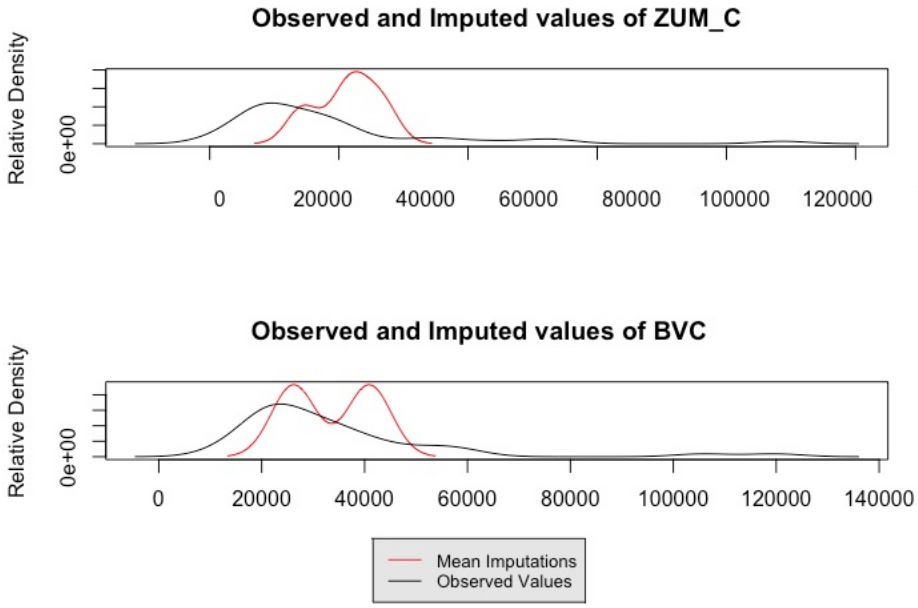
Na obrázku číslo 17 vidíme znázornenú hustotu rozloženia zaznamenaných údajov a hustotu rozloženia doplnených údajov pre intervenciu PTA/s. Z obrázku je vidieť, že u premennej ZUM_C sa doplnilo relatívne veľa hod v okolí v hodnote nákladov 6000 kč.



Obrázok 17: Porovnanie hustoty dostupných údajov a údajov doplnených pre intervenciu PTA/s

[vlastné spracovanie]

Pre intervenciu bypass vidíme na obrázku číslo 18 hustotu sledovaných a doplnených údajov. Bypass je jediná zvolená intervencia, kde simulácia doplnenia hodnôt prebieha len na dvoch premenných. Červená krivka znázorňuje hustotu doplnených hodnôt. Osa X vyjadruje náklady na jednotlivé premenné, osa Y je relatívna hustota hodnôt.

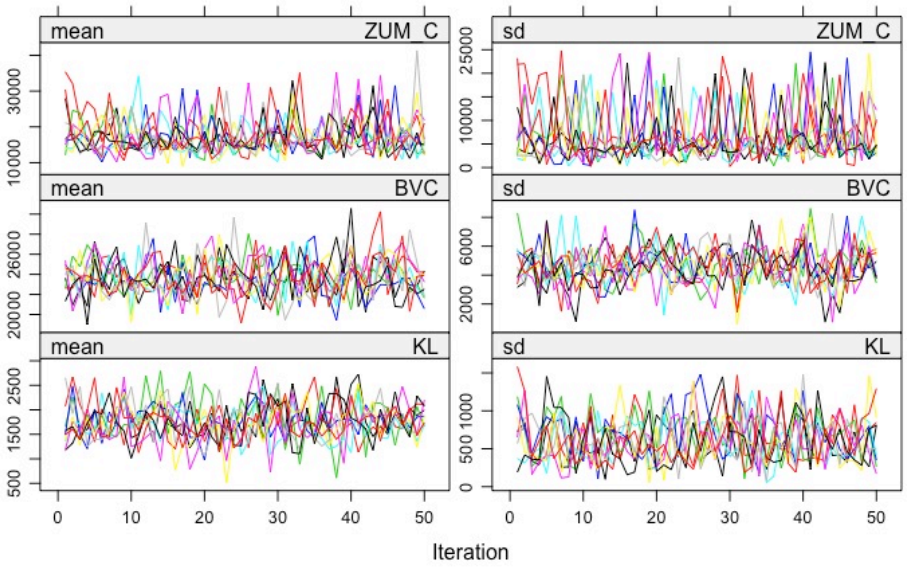


Obrázok 18: Porovnanie hustoty dostupných údajov a údajov doplnených pre intervenciu bypass

[vlastné spracovanie]

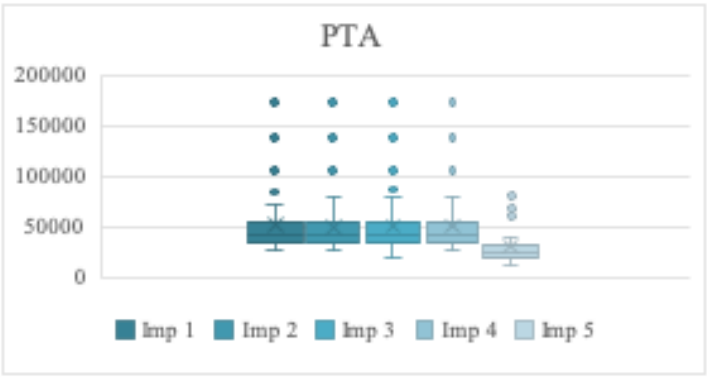
3.5 Mnohonásobná imputácia MICE algoritmom

Metóda mnohonásobnej imputácie využíva pre dopočítanie hodnôt algoritmus MICE – Markovove reťazce Monte Carlo. Pre spracovanie bol využitý štatistický program R. Metóda prebieha v troch krokoch. Kde v prvom kroku sa chýbajúce hodnoty nasimulujú stochastickým procesom. Pri spustení imputácie bolo definované vytvorenie 5 datasetov s výpočtom hodnoty v 50 iteráciách. Priebeh jednotlivých iterácií pre chýbajúce hodnoty intervencie PTA je znázornený na obrázku č. 19, kde farebné čiary znázorňujú trend jednotlivých opakovaní. Na pravej v stĺpcoch mean vidíme priemer hodnôt. V pravej časti je znázornená smerodajná odchýlka. Vizuál jednotlivých línií je typický pre metódu Monte Carlo pri využití veľkého počtu opakovaní.



Obrázok 19: Výpočty chýbajúcich hodnôt pomocou metódy MCMC pre intervenciu PTA [vlastné spracovanie]

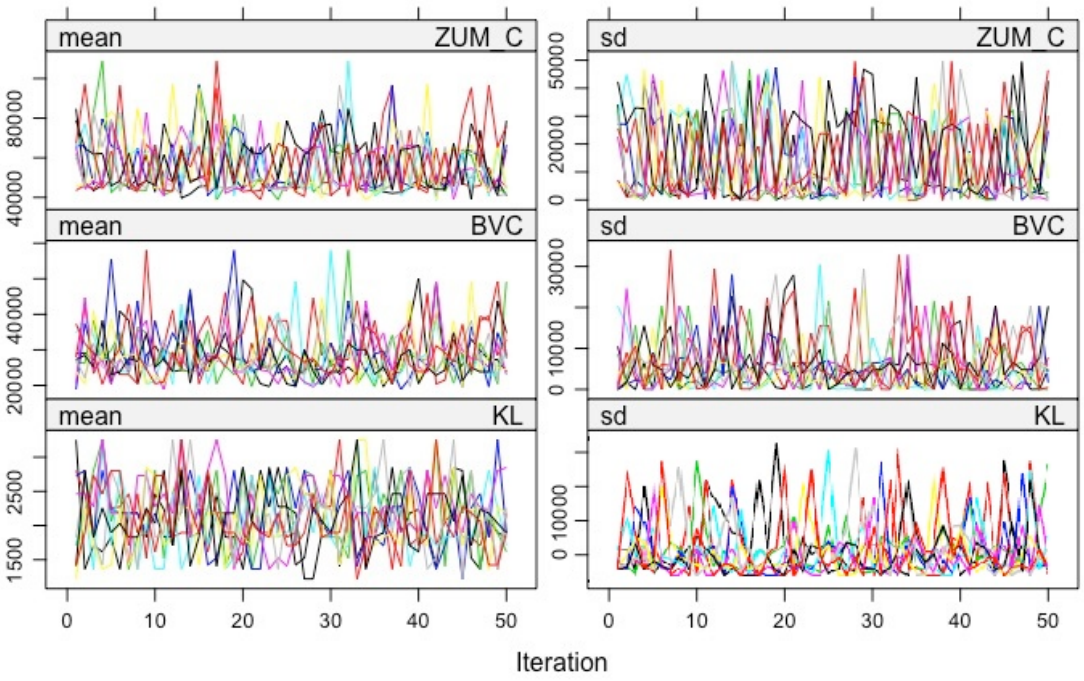
Ďalším nevyhnutným krokom vo vypracovaní je analýza vzniknutých piatich datasetov, ktoré nám vznikli po simulácii. V našom prípade bola využitá analýza pomocou grafu boxplot, v ktorom môžeme vidieť výsledné hodnoty, minimum, maximum a ich odľahlé hodnoty. Z grafu 1, že dataset označený Imp 5 sa líši svojím rozložením hodnôt od ostatných.



Graf 1: Boxplot pre 5 datasetov intervencie PTA [vlastné spracovanie]

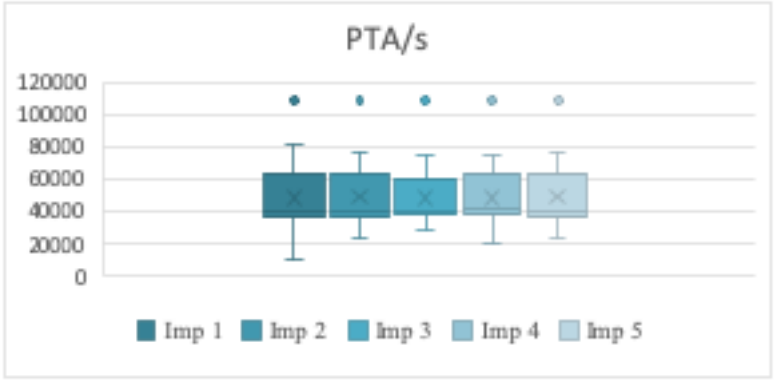
Pre získanie konečného výsledku bolo nutné výsledky zo zvolených imputácií nakombinovať do finálnej doplnenej dátovej matice. Pre intervenciu PTA sa použili datasety s označením Imp 1, Imp 2, Imp 3 a Imp 4.

Priebeh jednotlivých iterácií pre intervenciu PTA/s vidíme na obrázku číslo 20. Hodnoty na ose Y predstavujú škálu nákladov pre premenné, osa X charakterizuje priebeh výpočtu.



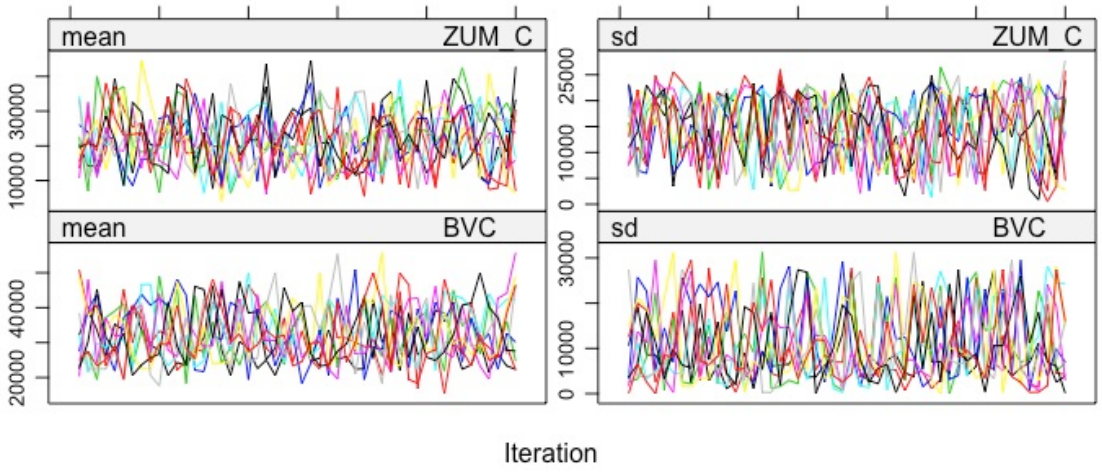
Obrázok 20: Výpočty chýbajúcich hodnôt pomocou metódy MCMC pre intervenciu PTA/s [vlastné spracovanie]

Analýza vygenerovaných datasetov aj v tomto prípade prebiehala pomocou grafu boxplot. Z jednotlivých výsledkov vidíme, že sa síce imputácie od seba líšia minimom a maximom ale rozloženie hodnôt je obdobné (graf 2). Finálna doplnená dátová matica pre intervenciu vznikla spojením datasetov Imp 1, Imp 2, Imp 3, Imp 4 a Imp 5.

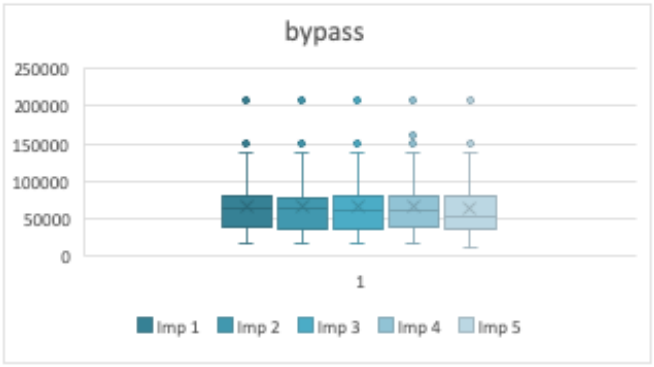


Graf 2: Boxplot pre 5 datasetov interveniciu PTA/s [vlastné spracovanie]

Výpočet chýbajúcich hodnôt pre intervenciu bypass prebiehal tiež pomocou 5 datasetov a 50 iterácií pre jednotlivé dopočty hodnôt. Priebeh výpočtov pomocou MCMC algoritmu hodnôt ZUM_C a BVC vidíme na obrázku číslo 21.



Obrázok 21: Výpočty chýbajúcich hodnôt pomocou metódy MCMC pre intervenciu bypass [vlastné spracovanie]



Graf 3: Boxplot pre 5 datasetov intervencie bypass [vlastné spracovanie]

Graf 3: Boxplot pre 5 datasetov intervencie bypass [vlastné spracovanie]

Rozloženie dopočítaných hodnôt je v prípade údajov pre bypass podobné, hodnoty sa od seba významne nelíšia, čo môžeme vidieť na grafe 3. Výsledný doplnený dataset vznikol spojeným všetkých piatich vytvorených dátových matíc.

3.6 Štatistické spracovanie súborov dát

Vyhodnotenie výpočtov bolo spracované pomocou tabuľkového procesoru Excel vo formáte .xlsx.

K štatistickému vyhodnoteniu výsledkov metód doplnenia dát bol využitý balíček programu R - Rcmdr. V programe R boli využité súbory vo formáte .xlsx a .csv. Jednotlivé dátové matice s hodnotami nákladov boli označené nasledovne:

- C1 – Zdrojové dáta – pôvodné nezmenené údaje,
- C2 – Analýza kompletných prípadov – vyškrtnuté neúplné prípady,
- C3 – Imputácia aritmetickým priemerom – údaje doplnené priemernou hodnotou,
- C4 – Imputácia EM algoritmom – údaje doplnené štatistickou metódou,
- C5 – Imputácia MCMC algoritmom – údaje doplnené štatistickou metódou.

3.6.1 Popisná štatistika

Spracovanie nákladov jednotlivých terapeutických metód s využitím štatistických metód k doplneniu dát sú popísané pomocou základnej charakteristiky popisnej štatistiky. Výsledky boli vyhodnotené pomocou funkcie analýza údajov – deskriptívna štatistika, ktorá je bežne dostupná v programe Excel. V tabuľke č. 4 vidíme spočítané hodnoty pre terapeutickú metódu PTA. Je tu vidieť kvantitatívne rozdiely medzi jednotlivými metódami, ako aj to, že sa metódou C2 znižuje veľkosť vzorku ale menia sa aj hodnoty pre minimum, medián či priemernú hodnotu. Suma uvádza súčet všetkých hodnôt, kde tiež môžeme pozorovať rozdiely. Najnižšia hodnota je v prípade analýzy kompletných prípadov

Tabuľka 4: Štatistický popis dátových matíc pre intervenciu PTA

Štatistický ukazovateľ	*C1 - Zdrojové dáta*	*C2 - Analýza komp. príp.*	*C3 - Imp aritm. priemer*	*C4 - Imp. EM algoritmus*	*C5 - Imp. MCMC algoritmus*
Priemer	49 444	50 307	49 392	49 255	48 876
Štandardná chyba	3 233	4 062	3 212	3 222	3 250
Medián	42 338	4 1101	42 102	41 738	40 877
Smerodajná odchýlka	26 065	28 720	25 893	25 979	26 205
Minimum	26 138	28 651	26 721	26 742	26 296
Maximum	187 455	187 455	187 455	187 455	187 455
Suma	3 213 884	2 515 347	3 210 499	3 201 568	3 176 925
Počet	65	50	65	65	65

Nasledujúca tabuľka obsahuje vypočítané hodnoty popisnej štatistiky pre terapeutickú metódu PTA/s.

Tabuľka 5: Štatistický popis dátových matíc pre intervenciu PTA/s

Štatistický ukazovateľ	*C1 - Zdrojové dáta*	*C2 - Analýza komp. príp.*	*C3 - Imp aritm. priemer*	*C4 - Imp. EM algoritmus*	*C5 - Imp. MCMC algoritmus*
Priemer	94 104	98 592	95 452	92 994	94 137
Štandardná chyba	7 818	9 403	7 733	7 915	7 801
Medián	76 280	77 053	78 074	75 245	75 819
Smerodajná odchýlka	39 864	43 090	39 431	40 356	39 778
Minimum	61 981	61 981	61 981	61 981	61 981
Maximum	233 090	233 090	233 090	233 090	233 090
Suma	2 446 693	2 070 442	2 481 757	2 417 843	2 447 558
Počet	26	21	26	26	26

Rozdiely hodnôt popisnej štatistiky pre intervenciu bypass vidíme v tabuľke číslo 6.

Tabuľka 6: Štatistický popis dátových matíc pre intervenciu bypass

Štatistický ukazovateľ	*C1 - Zdrojové dáta*	*C2 - Analýza komp. príp.*	*C3 - Imp aritm. priemer*	*C4 - Imp. EM algoritmus*	*C5 - Imp. MCMC algoritmus*
Priemer	62 673	60 237	62 443	61 964	62 620
Štandardná chyba	6 384	7 095	6 296	6 310	6 351
Medián	46 913	46 263	48 699	48 851	46 913
Smerodajná odchýlka	41 860	43 734	41 288	41 375	41 649
Minimum	24 569	24 569	24 569	24 569	24 569
Maximum	235 388	235 388	235 388	235 388	235 388
Suma	2 694 944	2 288 988	2 685 058	2 664 431	2 692 676
Počet	43	38	43	43	43

3.6.2 Testovanie normality hodnôt

Pre vyhodnotenie či údaje majú normálne rozloženie bol vypracovaný Shapiro-Wilkovov. Vyhodnotenie bolo spracované pre každú hodnotu intervencie PTA, PTA/s a Bypass. Pre štatistické spracovanie počítame s hladinou významnosti 5 %.

Na základe výsledkov p-hodnôt v tabuľkách 8, 9 a 10 nejde predpokladať normálne rozdelenie dát pre hodnoty metód PTA, PTA/s a bypass. Týmto je potvrdená správnosť výberu neparametrického testu v nasledujúcej časti práce.

Tabuľka 7: Test normality rozloženia dát pre metódu PTA

Shapiro-Wilkovov test PTA	*C1 - Zdrojové dáta*	*C2 - Analýza komp. príp.*	*C3 - Imp aritm. priemer*	*C4 - Imp. EM algoritmus*	*C5 - Imp. MCMC algoritmus*
W	0,63	0,61	0,59	0,61	0,62
p-hodnota	$1{,}31.10^{-11}$	$7{,}79.\ 10^{-12}$	$1{,}44.\ 10^{-12}$	$7.71.\ 10^{-12}$	$1{,}31.\ 10^{-11}$
α	0,05	0,05	0,05	0,05	0,05

Tabuľka 8: Test normality rozloženia dát pre metódu PTA/s

Shapiro-Wilkovov test PTA/s	*C1 - Zdrojové dáta*	*C2 - Analýza komp. príp.*	*C3 - Imp aritm. priemer*	*C4 - Imp. EM algoritmus*	*C5 - Imp. MCMC algoritmus*
W	0,73	0,77	0,75	0,72	0,73
p-hodnota	$1{,}60.\ 10^{-7}$	$2{,}55.\ 10^{-6}$	$2{,}69.\ 10^{-7}$	$1{,}10.10^{-7}$	$1{,}20.\ 10^{-7}$
α	0,05	0,05	0,05	0,05	0,05

Tabuľka 9: Test normality rozloženia dát pre metódu bypass

Shapiro-Wilkovov test Bypass	*C1 - Zdrojové dáta*	*C2 - Analýza komp. príp.*	*C3 - Imp aritm. priemer*	*C4 - Imp. EM algoritmus*	*C5 - Imp. MCMC algoritmus*
W	0,68	0,64	0,67	0,66	0,68
p-hodnota	$2,\ 10^{-10}$	$1{,}79.10^{-10}$	$1{,}68.\ 10^{-10}$	$1{,}25.\ 10^{-10}$	$2{,}04.\ 10^{-10}$
α	0,05	0,05	0,05	0,05	0,05

3.6.3 Porovnanie nových dátových súborov

K vyhodnoteniu rozdielu medzi pôvodnými údajmi zo zdrojových dát a novozískanými hodnotami bol zvolený párový Wilcoxonov test. K štatistickému vypracovaniu boli stanovené nasledujúce hypotézy:

H0: medián rozdielov je nulový

H1: medián rozdielov sa nerovná nuly

Wilcoxonov test testuje hypotézu rovnosti distribučných funkcií overením symetrického rozloženia náhodnej veličiny [9]. Pre zhodnotenie sa využili zdrojové dáta v porovnaní s údajmi doplnených pomocou jednotlivých metód. V tabuľke č. 10 vidíme vypočítané jednotlivé p-hodnoty pre štatistické porovnávanie metódy PTA.

Tabuľka 10: Vyhodnotenie párový Wilcoxonov test - PTA

Wilcoxonov párový test PTA	
C1 – Zdrojové dáta vs. C2 – Analýza kompletných prípadov	
p-hodnota 1	0,17
C1 – Zdrojové dáta vs. C3 – Imputácia aritmetickým priemerom	
p-hodnota 2	0,29
C1 – Zdrojové dáta vs.C4 – Imputácia EM logaritmom	
p-hodnota 3	0,75
C1 – Zdrojové dáta vs. C5 – Imputácia MCMC	
p-hodnota 4	0,712
α	0,05

Aj napriek rozdielnym hodnotám vypočítaných p-hodnôt pre metódu PTA nezamietame nulovú hypotézu na hladine významnosti 5 %, a tým je možné povedať, že medzi nákladmi nebol zistený štatisticky významný rozdiel. štatistický rozdiel.

V tabuľke č. 11 vidíme vypočítané hodnoty pre intervenciu PTA/s. Na základe hodnôt pre metódu PTA/s s využitím analýzy kompletných prípadov výsledná p-hodnota 1 zamieta nulovú hypotézu. Medzi nákladmi C1 a C2 bol zistený štatisticky významný rozdiel na hladine významnosti 5 %. P-hodnoty pre porovnanie zvyšných metód doplnenia chýbajúcich údajov nezamieta nulovú hypotézu, tzn. medzi nákladmi C2, C4 a C5 nebol zistený štatisticky významný rozdiel.

Tabuľka 11: Vyhodnotenie párový Wilcoxonov test – PTA/s

Wilcoxonov párový test PTA/s	
C1 – Zdrojové dáta vs. C2 – Analýza kompletných prípadov	
p-hodnota 1	0,03
C1 – Zdrojové dáta vs. C3 – Imputácia aritmetickým priemerom	
p-hodnota 2	0,27
C1 – Zdrojové dáta vs.C4 – Imputácia EM logaritmom	
p-hodnota 3	0,79
C1 – Zdrojové dáta vs. C5 – Imputácia MCMC	
p-hodnota 4	0,93
α	0,05

Výpočet rozdielu medzi metódami pre intervenciu bypass p-hodnoty 1 zamietame nulovú hypotézu. Medzi nákladmi zdrojových dát a nákladmi získanými analýzou kompletných prípadov bol zistený štatistický rozdiel na hladine významnosti 5 %. Výsledky p hodnôt 2, 3 a 4 nezamietajú nulovú hypotézu, čím môžeme tvrdiť, že medzi jednotlivými nákladmi nie je zistený štatistický rozdiel.

Tabuľka 12: Vyhodnotenie párový Wilcoxonov test - bypass

Wilcoxonov párový test Bypass	
C1 – Zdrojové dáta vs. C2 – Analýza kompletných prípadov	
p-hodnota 1	0,01
C1 – Zdrojové dáta vs. C3 – Imputácia aritmetickým priemerom	
p-hodnota 2	0,11
C1 – Zdrojové dáta vs.C4 – Imputácia EM logaritmom	
p-hodnota 3	0,67
C1 – Zdrojové dáta vs. C5 – Imputácia MCMC	
p-hodnota 4	0,83
α	0,05

3.7 Nákladová efektivita

Pri výpočte nákladovej efektivity sa využili priemerné náklady zo zdrojových dát a z finálnych dátových matíc po aplikácií vybraných metód. Nákladová analýza sa vyhodnocovala pre efekt primárna priechodnosť, technický úspech, ročne prežitie pacienta a záchrana končatiny v roku operácie. Hodnoty efektu pochádzajú z publikovanej práce Ing. Kamenského [58]. V tabuľke č. 13 je vypočítaná hodnota nákladovej efektivity, ako aj hodnoty ICER pre porovnanie intervencií, % rozdiel vyjadruje percentuálny rozdiel medzi hodnotou zdrojových dát a vypočítanou hodnotou pre jednotlivé využité metódy.

Tabuľka 13: Výsledky CEA analýzy, výpočet pre efekt primárna priechodnosť

Efekt = Primárna priechodnosť								
Metóda	*C (KČ)*	*Efekt E*	*CE*	*% rozdiel*	*ICER 1*	*% rozdiel*	*ICER 2*	*% rozdiel*
C1 - priemerné náklady - zdrojové dáta								
PTA	49 444	50	991					
Bypass	62 673	85	742		382			
PTA/S	94 104	67	1 407		2 627		-1 786	
C2 - priemerné náklady - Analýza kompletných prípadov								
PTA	50 307	50	1 008	**1,74**				
Bypass	60 237	85	713	**-3,89**	287	**-24,94**		
PTA/S	98 592	67	1 474	**4,77**	2 840	**8,12**	-2 179	**22,00**
C3 - priemerné náklady - Imputácia aritmetickým priemerom								
PTA	49 392	50	990	**-0,11**				
Bypass	62 443	85	739	**-0,37**	377	**-1,34**		
PTA/S	95 452	67	1 427	**1,43**	2 709	**3,14**	-1 876	**5,02**
C4 - priemerné náklady - Imputácia EM logaritmus								
PTA	49 255	50	987	**-0,38**				
Bypass	61 964	85	733	**-1,13**	367	**-3,93**		
PTA/S	92 994	67	1 390	**-1,18**	2 573	**-2,06**	-1 763	**-1,27**
C5 - priemerné náklady – Imputácia MCMC								
PTA	48 876	50	979	**-1,15**				
Bypass	62 620	85	741	**-0,08**	397	**3,90**		
PTA/S	94 137	67	1 407	**0,04**	2 662	**1,35**	-1 791	**0,27**

V tabuľke č. 14 vidíme výsledky CEA analýzy s využitím hodnôt efektu technický úspech. Z výsledkov vidíme, že aj s použitím inej hodnoty efektu sa nám výsledné poradie intervencií nemení a je dobre vidieť, že percentuálny rozdiel vo výsledku je v rovnakom pomere ako u efektu primárna priechodnosť. Najvyšší rozdiel medzi zdrojovými dátami a dátami získanými doplnením vidíme pre hodnotu ICER 1 s využitím nákladov C2. Rozdiel je až 24,94 %. Rozdiely medzi hodnotami získaných z mnohonásobných imputácií – C4 a C5 majú nemenšie rozdiely so zdrojovými dátami. Hodnoty pre nákladovú analýzu, ICER 1 a ICER 2 sa v oboch prípadoch odlišujú menej ako o 5 %.

Tabuľka 14: Výsledky CEA analýzy, výpočet pre efekt technický úspech

Efekt = Technický úspech								
Metóda	*C (KČ)*	*Efekt E*	*CE*	*% rozdiel*	*ICER 1*	*% rozdiel*	*ICER 2*	*% rozdiel*
C1 - priemerné náklady - zdrojové dáta								
PTA	49 444	86	578					
Bypass	62 673	99	631		952			
PTA/S	94 104	91	1 036		8 426		-3 655	
C2 - priemerné náklady - Analýza kompletných prípadov								
PTA	50 307	86	588	**1,74**				
Bypass	60 237	99	606	**-3,89**	714	**-24,94**		
PTA/S	98 592	91	1 086	**4,77**	9 110	**8,12**	-4 460	**22,00**
C3 - priemerné náklady - Imputácia aritmetickým priemerom								
PTA	49 392	86	578	**-0,11**				
Bypass	62 443	99	623	**-0,37**	939	**-1,34**		
PTA/S	95 452	91	1 024	**1,43**	8 691	**3,14**	-3 838	**5,02**
C4 - priemerné náklady - Imputácia EM logaritmus								
PTA	49 255	86	576	**-0,38**				
Bypass	61 964	99	623	**-1,13**	914	**-3,93**		
PTA/S	92 994	91	1 024	**-1,18**	8 253	**-2,06**	-3 608	**-1,27**
C5 - priemerné náklady – Imputácia MCMC								
PTA	48 876	86	572	**-1,15**				
Bypass	62 620	99	630	**-0,08**	989	**3,90**		
PTA/S	94 137	91	1 037	**0,04**	8 540	**1,35**	-3 665	**0,27**

Výsledky získané výpočtom CEA analýzy s efektom ročné prežitie pacienta vidíme, že aj keď sa nominálna hodnota výsledkov líši, pomerovo je rozdiel medzi zdrojovými dátami C1 a nákladmi pre C2-C5 rovnaký ako v prechádzajúcich dvoch prípadoch. Aj pri vysokých hodnotách klinického efektu ročné prežitie sa nám výsledné poradie a rozdiely nezmenili.

Tabuľka 15: Výsledky CEA analýzy, výpočet pre ročné prežitie pacienta

Efekt = Ročné prežitie pacienta								
Metóda	*C (KČ)*	*Efekt E*	*CE*	*% rozdiel*	*ICER 1*	*% rozdiel*	*ICER 2*	*% rozdiel*
C1 - priemerné náklady - zdrojové dáta								
PTA	49 444	96	516					
Bypass	62 673	92	683		-3 227			
PTA/S	94 104	91	1 036		-8 757		-31 655	
C2 - priemerné náklady - Analýza kompletných prípadov								
PTA	50 307	96	525	**1,74**				
Bypass	60 237	92	656	**-3,89**	-2 422	**-24,94**		
PTA/S	98 592	91	1 086	**4,77**	-9 468	**8,12**	-38 356	**22,00**
C3 - priemerné náklady - Imputácia aritmetickým priemerom								
PTA	49 392	96	515	**-0,11**				
Bypass	62 443	92	680	**-0,37**	-3 183	**-1,34**		
PTA/S	95 452	91	1 051	**1,43**	-9 031	**3,14**	-33 009	**5,02**
C4 - priemerné náklady - Imputácia EM logaritmus								
PTA	49 255	96	514	**-0,38**				
Bypass	61 964	92	675	**-1,13**	-3 100	**-3,93**		
PTA/S	92 994	91	1 024	**-1,18**	-8 576	**-2,06**	-31 030	**-1,27**
C5 - priemerné náklady – Imputácia MCMC								
PTA	48 876	96	510	**-1,15**				
Bypass	62 620	92	682	**-0,08**	- 3 352	**3,90**		
PTA/S	94 137	91	1 037	**0,04**	-8 875	**1,35**	-31 516	**0,27**

Výsledky hodnôt nákladovej analýzy a hodnoty ICER 1, ICER 2 pre efekt záchrana končatiny v roku vidíme v tabuľke č. 16:

Tabuľka 16: Výsledky CEA analýzy, záchrana končetiny v roku operácie

Efekt = Záchrana končatiny v roku								
Metóda	*C (KČ)*	*Efekt E*	*CE*	*% rozdiel*	*ICER 1*	*% rozdiel*	*ICER 2*	*% rozdiel*
C1 - priemerné náklady - zdrojové dáta								
PTA	49 444	94	528					
Bypass	62 673	89	705		-2 815			
PTA/S	94 104	96	982		20 300		4 555	
C2 - priemerné náklady - Analýza kompletných prípadov								
PTA	50 307	94	537	**1,74**				
Bypass	60 237	89	678	**-3,89**	-2 113	**-24,94**		
PTA/S	98 592	96	1 029	**4,77**	21 948	**8,12**	5 559	**22,00**
C3 - priemerné náklady - Imputácia aritmetickým priemerom								
PTA	49 392	94	528	**-0,11**				
Bypass	62 443	89	702	**-0,37**	-2 777	**-1,34**		
PTA/S	95 452	96	996	**1,43**	20 936	**3,14**	4 784	**5,02**
C4 - priemerné náklady - Imputácia EM logaritmus								
PTA	49 255	94	526	**-0,38**				
Bypass	61 964	89	697	**-1,13**	-2 704	**-3,93**		
PTA/S	92 994	96	971	**-1,18**	19 881	**-2,06**	4 497	**-1,27**
C5 - priemerné náklady – Imputácia MCMC								
PTA	48 876	94	522	**-1,15**				
Bypass	62 620	89	704	**-0,08**	-2 924	**3,90**		
PTA/S	94 137	96	983	**0,04**	20 573	**1,35**	4 568	**0,27**

Pri porovnaní výsledkov nákladovej efektivity a hodnôt ICER pre jednotlivé klinické efekty a vypočítaných percentuálnych rozdielov medzi získanými hodnotami vidíme, že pre metódu analýzy kompletných prípadov, ktorý pracuje s vyradením neúplnej vzorky z výpočtov sa celkový rozdiel oproti ostatným metódam výrazne líši. Vo všetkých prípadoch je rozdiel oproti zdrojovým dátam o 24,94 % pre hodnotu ICER 1, 22 % pre hodnotu ICER 2. Výsledok nákladovej analýzy je pri metóde PTA/s, ktorý obsahovala najmenej údajov až 4,77 %. Najmenšie rozdiely sú medzi nákladmi získanými mnohonásobnou imputáciou MCMC, kde najväčší rozdiel je v hodnote 3,90 % pre hodnotu ICER 1 v porovnaní intervencie PTA verzus bypass. Rozdiely pre mnohonásobnú imputáciu EM algoritmom sa veľmi nelíšia od metódy imputácie

algoritmom MCMC. Najväčší rozdiel je v hodnote ICER 1 -3,93 %. Pre doplnenie dát aritmetickým priemerom sa výsledky odlíšili v hodnote pre ICER 2 o 5,02 %.

3.8 Zhrnutie výsledkov

V tejto časti sú zosumarizované všetky poznatky získané počas celého spracovania metód určených k riešeniu chýbajúcich dát od ich vypracovania, štatistického testovania až po výsledky HTA analýzy.

Analýza kompletných prípadov

- **Princíp:** Metóda odstraňuje vzorky pacientov s neúplnými dátami
- **Náročnosť na spracovanie:** Metódu nie je náročné spracovať, nie je potrebná znalosť dodatočných softvérov. Odstránenie neúplných vzoriek je možné spracovať rýchlo v programe Excel.
- **Rozdiel so zdrojovými dátami:** Výpočtom neparametrického testu bol zistený štatisticky významný rozdiel pre hodnoty získaným odstránením vzorku pre metódu PTA/s.
- **Výpočet nákladov:** Poradie intervencií oproti zdrojovým dátam nebolo zmenené, najnákladnejšia metóda stále ostala PTA/s, najlacnejšia metóda bola stále metóda PTA
- **Rozdiel výsledkov HTA so zdrojovými dátami:** Pri hodnotách ICER 1 a ICER 2 bol najvyšším v porovnaní s ostatnými metódami a to -24,94 %, 8,12 % a 22,00 %.
- **Odporúčanie:** Na základe štatistických testov a rozdielov vo výsledkoch nie je možné metódu odporučiť k aplikácii. Znižuje celkovú štatistickú silu výsledku, pretože dochádza k zmenšení celkovej dátovej matice.

Jednoduchá imputácia aritmetickým priemerom

- **Princíp:** Metóda založená na doplnení chýbajúcich hodnôt vypočítaným aritmetickým priemerom.
- **Náročnosť na spracovanie:** Výpočet aritmetického priemeru je veľmi ľahký, nie je potrebná žiadna špeciálna znalosť metodiky ani využitie špeciálneho softvéru.
- **Rozdiel so zdrojovými dátami:** Ani pre jednu intervenciu nebol zistený štatistický rozdiel medzi hodnotami nákladov získaných doplnením aritmetického priemeru aj keď výsledná p hodnota bola výrazne nižšia ako u metód mnohonásobnej imputácie.
- **Výpočet nákladov:** Poradie intervencií oproti zdrojovým dátam nebolo zmenené, najnákladnejšia metóda stále ostala PTA/s, najlacnejšia metóda bola stále metóda PTA

- **Rozdiel výsledkov HTA so zdrojovými dátami:** Rozdiel výsledkov v porovnaní so zdrojovými dátami mal lepši výsledok ako metóda analýzy kompletných prípadov, aj keď hodnota ICER 2 zaznamenala rozdiel 5,02 % oproti zdrojovým dátam.
- **Odporúčanie:** Na základe štatistických výsledkov a porovnaní výsledkov HTA analýzy nie je možné priamo odporučiť aj keď je výsledok lepší ako analýza kompletných prípadov.

Mnohonásobná imputácia EM algoritmom

- **Princíp:** Viacnásobné doplnenie hodnôt pomocou metódy na maximálnej vierohodnosti. Využíva dva základné kroky M – maximization (maximalizácia) E – expectation (očakávaní)
- **Náročnosť na spracovanie:** Hodnoty sú dopočítavané zložitým algoritmom, pri využití iterácií. Vypracovanie je potrebné pomocou špecializovaného softvéru. V programe R je možné si stiahnuť a nainštalovať balíček Amelia II, ktorá má dostupný aj popis práce. Práca s balíčkom je náročnejšia, nutnosť základného povedomia o práci s dátami a zadávaní príkazov do R. Ak sa nestanoví ohraničenie údajov pred samotným spustením imputácie (náklady môžu mať len pozitívnu hodnotu), tak program vypočítaval negatívne hodnoty chýbajúcich dát.
- **Rozdiel so zdrojovými dátami:** Ani pre jednu intervenciu nebol zistený štatistický rozdiel medzi hodnotami nákladov získaných viacnásobným doplnením.
- **Výpočet nákladov:** Poradie intervencií oproti zdrojovým dátam nebolo zmenené, najnákladnejšia metóda stále ostala PTA/s, najlacnejšia metóda bola stále metóda PTA
- **Rozdiel výsledkov HTA so zdrojovými dátami:** Rozdiel jednotlivých výsledkov CEA analýzy bol odlišný v intervale -0,38 % až 1,18 %. Hodnoty ICER 1 a ICER 2 sa líšili o menej ako 5 %.
- **Odporúčanie:** Výsledky štatistického testovania neparametrického testu sa zhodujú s percentuálnym porovnaním výsledkov HTA analýzy. Využitie metódy k doplneniu dát pre chýbajúce

Mnohonásobná imputácia MCMC algoritmom

- **Princíp:** Viacnásobné doplnenie hodnôt pomocou metódy Markovových reťazcov Monte Carlo, ktorý pracuje so znalosťami podmienených rozdelení jednotlivých premenných a tým určuje ich spoločné rozdelenie.
- **Náročnosť na spracovanie:** Metódu je možné spracovať pomocou balíčka MICE v programe R. Pri vypracovaní je nutné dodržať stanovený proces doplnenia údajov, tj. viacnásobné vytvorenie imputovaných datasetov, následne je nutné

vzniknuté datasety štatisticky spracovať. Až po spracovaní je možné výsledky nakombinovať. V prípade nedodržania postupu a nakombinovaní parciálnych výsledkov pomocou priemerných hodnôt dochádza k veľkému skresleniu výsledkov. Pre zadávanie príkazov je nutná znalosť práce s R. Metóda bola najviac časovo náročná na spracovanie.

- **Rozdiel so zdrojovými dátami:** Ani pre jednu intervenciu nebol zistený štatistický rozdiel medzi hodnotami nákladov získaných viacnásobným doplnením. P hodnota pre všetky intervencie mala v prípade MCMC najlepšie hodnoty (najviac sa približovali k číslu 1).
- **Výpočet nákladov:** Poradie intervencií oproti zdrojovým dátam nebolo zmenené, najnákladnejšia metóda stále ostala PTA/s, najlacnejšia metóda bola stále metóda PTA
- **Rozdiel výsledkov HTA so zdrojovými dátami:** Rozdiel medzi hodnotami získanými viacnásobným doplnením MCMC algoritmom v porovnaní s ostatnými hodnotami mali najlepší výsledok. Rozdiel všetkých výsledkov pre hodnotu analýzy CEA a ICER 1 a ICER 2 boli pod 5 %. Výsledky nákladovej efektivity sa u metódy PTA líšili o -1,15 %, PTA/s -0,08 % a 0,04 % pre intervenciu bypass.
- **Odporúčanie:** Metóda viacnásobného doplnenia pomocou MCMC algoritmom mala najlepšie výsledky pri štatistickom testovaní aj pri porovnaní rozdielov výsledkov HTA analýzy. Aj napriek náročnosti na spracovanie je vhodná na využitie pri probléme riešenia chýbajúcich dát.

4 DISKUSIA

Diplomová práca sa zameriava na zmapovanie štatistických metód pre riešenie chýbajúcich dát. V prvej časti práce sú teoretické základy a vysvetlenie obecných pojmov, ktoré je vhodné poznať k lepšej orientácii v problematike chýbajúcich hodnôt.

Pri spracovávaní prehľadu metód sa vyskytol nejeden problém. V prvom rade zistenie, že aj napriek tomu, že sa problematika adekvátneho doplnenia údajov vo svete rieši od sedemdesiatych rokov minulého storočia, a štatistika sama o sebe je veda, ktorá má veľkú históriu, tak v prostredí Českej a Slovenskej republiky nie je dostatok publikovaných vedeckých práci na túto tému. Medzi najaktívnejších autorov venujúcich sa analýze chýbajúcich údajov a ich spracovaniu považujem pána Ivana Petrúška zo Sociologického ústavu Akadémie vied ČR. Problematiku chýbajúcich dát otvoril už vo svojej diplomovej práci a je aj autorom monografie [22], ktorá má najpodrobnejšie informácie o štatistických metódach. V prostredí Českej republiky nebola nájdená žiadna iná publikácia, či vedecký článok, ktorá by popisoval tieto metódy v praktickom využití. Pre pochopenie metód som musela využiť zahraničné zdroje, na základe ktorých som zostavila ich porovnanie. Následne boli vybrané štyri metódy s odlišným prístupom k riešeniu chýbajúcich hodnôt.

Základná myšlienka správneho výberu metódy je založená na analýze mechanizmu chýbajúcich dát. Preto bol zvolený scenár pre simuláciu chýbajúcich hodnôt, mechanizmom založenom na princípe úplne náhodného chýbania dát [5, 6]. Pre zhodnotenie vplyvu štatistických metód boli využité zdrojové dáta pre tri intervencie terapie povrchovej stehennej tepny – PTA, PTA/s a bypass. Pre každú intervenciu bol zaznamenaný iný počet meraní a výsledkov. Vzhľadom na to, bolo zaujímavé sledovať rozdiely po využití štatistických metód, ktoré boli spôsobené iným rozsahom údajov.

Aby bol dodržaný predpoklad úplne náhodného chýbania hodnôt, tak pre odstránenie časti údajov bola využitá funkcia MICE v programe R. Rozsah odstránenia bol zvolený 20 %. V dôsledku rozdielu veľkosti dátovej matice, bolo pre každú intervenciu zistené, že po odstránení sa percento chýbajúcich hodnôt líši. Pre intervenciu PTA to bolo 23 %, PTA/s 19% a pre bypass 14 %. Takto vytvorené dátové matice boli následne využité pre ďalšie spracovanie.

Prvá využitá metóda bola analýza kompletných prípadov. Prípady, ktoré majú chýbajúcu časť údajov sú vylúčené z dátovej matice. Burton [42] vo svojej štúdií riešil problém chýbajúcich hodnôt v analýze nákladovej efektivity v klinickom hodnotení. Jeho výsledok je zhodný v porovnaní s výsledkom diplomovej práce. Táto metóda znehodnocuje výsledky a prikláňa sa k použitiu sofistikovaných metód. V našom prípade metóda zmenšila vzorku pre intervenciu PTA zo 65 prípadov na 50. Veľkosť PTA/s sa zmenšila z pôvodných 26 pacientov na 21 a metóda bypass z nameraných 43 na 38 pacientov. Aj Rubin[6] zhodnotil metódu ako nevýhodnú, pretože vedie ku strate a neefektívnemu využitiu údajov, ktoré už boli zaznamenané. Soley [14] uvádza, že pri využití tejto metódy dochádza k zmenšeniu sledovaného súboru, zníženiu bias a sily štatistického testu. V párovom neparametrickom testovaní bol zistený štatisticky významný rozdiel na hladine významnosti 5 % pre metódu PTA/s a bypass v porovnaní so zdrojovými dátami. Výsledná p hodnota pre metódu PTA/s bola 0,03 a pre bypass 0,01. Z výsledkov je znateľné, že pri dátových maticiach s menším rozsahom dát sa odstránenie

údajov prejaví väčším rozdielom v porovnaní s pôvodnými dátami. Z komparácie výsledkov HTA analýzy vyšli hodnoty pre metódu PTA o 1,74 % väčšie ako v prípade zdrojových dát. Hodnota nákladovej analýzy pre metódu bypass má hodnotu menšiu o 3,89% v porovnaní s pôvodnými dátami. PTA/s sa líši svojím výsledkom o 4,77%. Najväčšie rozdiely však boli zistené pre hodnoty ICER, pomer inkrementálnych nákladov a prínosov ICER 1 bol o 24,94 % menší oproti výsledkom z pôvodných dát. Hodnota ICER 2 bola zasa o 22 % vyššia.

Aj napriek výsledkom, ktoré ukazujú, že metóda nie je optimálna, je nutné konštatovať, že metóda má univerzálne využitie – nie je potrebná prídavná analýza údajov. A metóda je aj univerzálna – dá sa využiť vo všetkých odvetviach. Výber tejto metódy pre spracovanie diplomovej práce možno hodnotiť pozitívne. Rozdiely, ktoré vznikajú odstránením časti údajov, môžu motivovať ľudí k zamysleniu sa nad zmyslom riešenia chýbajúcich údajov a možnosťami ich doplnenia pomocou sofistikovaných metód.

Ako druhá bola využitá metóda doplnenia hodnôt pomocou vypočítaného aritmetického priemeru. Rubin a Little [6, 55] sa zhodujú v tvrdení, že využitie aritmetického priemeru zmenšuje prirodzený rozptyl hodnôt. V našom prípade sa jednoznačne nepreukázalo, že by metóda bola nevhodná na využitie. Hodnoty Wilcoxonovho testu majú v porovnaní so zložitejšími hodnotami horší výsledok. Štatisticky významný rozdiel s pôvodnými zdrojovými hodnotami sa nezistil. P hodnota mala pre intervenciu PTA hodnotu 0,29, PTA/s 0,27 a pre bypass 0,11. V porovnaní s výsledkom CEA analýzy sa hodnoty oproti zdrojovým dátam líšili relatívne málo. V prípade PTA 0,11 %, bypass 0,37 % a intervencii PTA/s 1,43 %. Hodnota ICER 1 sa líšila o 1,34 %, 3,34 % v porovnaní intervencií bypass vs PTA/s a 5,02 % pre hodnotu ICER 2. Celkovo sú rozdiely oproti zdrojovým dátam menšie ako v prípade analýzy kompletných prípadov. Avšak rozdiel 5,02 % hodnote ICER 2 je možné považovať už za významný. Každopádne, výhodou metódy je jej nenáročnosť na spracovanie. Aj táto metóda je podobne ako analýza kompletných prípadov veľmi univerzálna a dá sa využiť na všetky typy chýbajúcich hodnôt, ktoré sú kvantitatívne vyjadrené. Je však otázne, aký vplyv by metóda mala napríklad na výpočet krvného tlaku, kde je ťažké predpokladať, či každý biologický parameter má predpoklady pre priemerné hodnoty. Rozdiel v rozložení dát a celkového vplyvu doplnenia hodnôt aritmetickým priemerom by bolo vhodné skúmať na štúdiách, ktoré obsahujú väčší počet záznamov. Dá sa predpokladať, že rozdiely by boli signifikantnejšie a zmeny by bolo možné skúmať viac do hĺbky.

Treťou využitou metódou bola metóda viacnásobnej imputácie. Výpočet chýbajúcej hodnoty prebiehal pomocou EM algoritmu. Písmená v názve predstavujú dva kroky, ktorými sa metóda realizuje: E - expectation a M - maximalization. Na vypracovanie metódy bol využitý balíček Amelia II v programe R. Balíček má dostupný, prehľadne spracovaný popis metódy, kde sú vysvetlené aj základy práce so vzorovým súborom. Aj napriek tomu bolo spracovanie náročnejšie. V prvom rade bolo nutné

poznať aký typ premenných je potrebné nahradiť. Prvé pokusy neboli úspešné, pretože bolo nevyhnutné definovať logický rozsah hodnôt. Nutnosť definovania rozsahu doporučoval aj Rodwell [59] vo svojej štúdii, ktorá porovnávala doplnenie dát v malom rozsahu údajov. Bez tohto kroku generoval program záporné hodnoty, ktoré sa nám v našom prípade nehodili. Hodnoty prezentovali náklady, tzn. museli byť len kladné. Po upresnení parametrov už mali hodnoty potrebný charakter. Výsledné hodnoty pochádzajú z dvadsiatich súborov, ktoré boli vypočítané vo dvoch setoch simulácií a pomocou funkcie ameliablind spojené do výsledných matíc. Pre každú intervenciu prebiehalo dopĺňanie dát separátne.

Pri zisťovaní rozdielov nových údajov s pôvodnými dátami mali výsledné p hodnoty veľmi dobré výsledky: pre intervenciu PTA 0,75, PTA/s 0,75 a pre bypass 0,67. Ani pre jednu z intervenčných metód nebol zistený významný štatistický rozdiel na úrovni významnosti 5 %.

Rozdiely vo výsledku analýzy nákladovej efektivity boli taktiež lepšie v porovnaní s dvoma predchádzajúcimi metódami. Nákladová efektivita sa líšila pre PTA o 0,38 %, PTA/s o 1,13 % a pre bypass o 1,18 %. Hodnota inkrementálnych nákladov sa líšila v prípade ICER 1 – PTA vs bypass o 3,93 %. Táto hodnota predstavovala najvyšší rozdiel. Zvyšné rozdiely mali hodnotu 2,06 % a hodnota ICER 2 1,27 %. Ani jeden percentuálny rozdiel vo výsledkoch analýzy CEA nebol vyšší ako 5 %. Z čoho sa dá usúdiť, že metóda je vhodná na doplnenie dát pri mechanizme náhodného chýbania údajov. Toto tvrdenie sa zhoduje aj s ďalšími autormi vedeckých štúdií [20, 22, 60, 61].

Využitie metódy mnohonásobnej imputácie pomocou balíčka Amelia II má oproti analýze kompletných prípadov a využitiu aritmetického priemeru k doplneniu, nevýhodu v náročnosti na spracovanie. Metóda vyžaduje zručnosť a už predchádzajúce skúsenosti s prácou v programe R. Časová náročnosť je tiež vyššia, nakoľko bolo nutné spracovávať doplnenie pre každú intervenciu zvlášť. Aj napriek väčšej náročnosti sú výsledky dobrým dôvodom pre voľbu tejto metódy pre riešenie chýbajúcich dát.

Poslednou spracovávanou metódou bola metóda mnohonásobnej imputácie s využitím Markovovych reťazcov Monte Carlo. Metóda je súčasťou balíčka MICE, ktorý bol budovaný na poznatkoch Rubinovej knihy [6] a vychádza z jeho základných myšlienok. Princípom metódy nie sú len viacnásobné výpočty hodnôt, ale hlavne analýza doplnených hodnôt vo vytvorených datasetoch. Celý postup zadávania kódov v programe R je nastavený tak, že bez analýzy parciálnych hodnôt, nie je možné proces dokončiť. Balíček MICE oproti balíčku Amelia II nemá dostupný manuál a presný popis metód. Napriek tomu je najpoužívanejší v publikovaných prácach. Preto je na internete dostatok návodov ako s týmto balíčkom pracovať. MICE má aj komplexnú ponuku funkcií oproti balíčku Amelia II. Pomocou balíčka je možné simulovať chýbajúce údaje podľa všetkých troch mechanizmov. Výsledky doplnenia, či odstránenia údajov, je možné pomocou tohto balíčka spracovať aj graficky.

Prvé spracovanie metódy bolo v našom prípade tiež neúspešné. Výsledné hodnoty mali charakter a potrebný rozsah. Avšak pri štatistickom spracovaní a vyhodnotení rozdielov jednotlivých metód bol zistený problém v poslednom kroku spracovania. Finálna dátová matica pochádzala z piatich datasetov, ktoré boli dopočítavané 50-timi iteráciami. Na zlúčenie výsledkov bola využitá funkcia mean, ktorá spôsobila veľké skreslenie hodnôt. Výsledné hodnoty pre najzložitejšiu metódu mali výsledky veľmi podobné ako analýza kompletných prípadov. Po opätovnom prepracovaní metódy a využitia funkcie pool sa tento nedostatok odstránil.

V štatistickom testovaní rozdielov so zdrojovými dátami mala metóda viacnásobnej imputácie MCMC algoritmom najlepšie výsledky. Pre intervenciu PTA bola p hodnota 0,71, PTA/s 0,93 a pre bypass 0,83. Čím viac sa p hodnota blíži k 1, tým viac sú dva súbory menej odlišné.

Výsledné rozdiely v nákladovej efektivite oproti nákladovej efektivite zdrojových dát mali hodnotu pre PTA 1,15 % , PTA/s 0,08 % a bypass 0,04 %. Percentuálny rozdiel hodnôt ICER 1 a ICER 2 bol 3,9 %, 1,35 % a 0,27 %. Celkovo mala táto metóda zaznamenané najmenšie rozdiely v analýze nákladovej efektivity oproti ostatným štatistickým metódam. Aj napriek náročnému spracovaniu, je na základe výsledkov možné túto metódu odporučiť k riešeniu problému chýbajúcich dát.

Výsledky práce naznačujú, že využitie zložitejších metód môže mať pozitívny vplyv na výsledky spracovania dát. Výsledky HTA analýzy sa zásadne nezmenili a poradie intervencií zostalo nezmenené. Percentuálne rozdiely oproti zdrojovým dátam sú medzi využitými metódami signifikantné.

Využitie metód má svoje opodstatnenie, a to nielen v efektívnosti využitia dát, ale aj ďalšími prídavnými hodnotami. Riešenie problému vyžaduje bližšie zoznámenie sa s dátami a analýzou. Charakter hodnôt môže motivovať vedca k zamysleniu sa a k položeniu otázok. Prečo chýba odpoveď len u jedného typu otázky? Prečo tieto údaje neboli zaznamenané? Je problém v procese? Prečo je táto hodnota tak nízka? Je medzi hodnotami korelácia? Odpovede na tieto otázky môžu prispieť nielen k zlepšeniu procesu alebo dizajnu štúdie, ale môžu slúžiť aj k správnemu výberu metódy na doplnenie chýbajúcich dát. Odpoveď na otázku, ktorá štatistická metóda je najlepšia odpovedal pán Allison [5] výrokom: „*Najvhodnejšia metóda riešenia chýbajúcich hodnôt je predchádzanie ich vzniku*.“

Nakoľko je problematika chýbajúcich dát a ich možnosti riešenia v Českej republike veľmi málo rozoberaná, ďalší rozvoj problematiky diplomovej práce vidím vo využití metód na štúdie s väčším rozsahom dát alebo na štúdie, ktoré sa týkajú klinických údajov. Zaujímavé by mohlo byť aj využitie sofistikovaných metód na doplnenie kategorických premenných, ktoré sa nedajú kvantitatívne vyjadriť.

5 ZÁVER

Hlavný cieľ diplomovej práce bol stanovený ako porovnanie vplyvu štatistickým metód na riešenie neúplných hodnôt. Parciálne ciele práce mali za úlohu 1) nasimulovať chýbajúce dáta na zdrojových dátach pochádzajúcich z klinickej štúdie terapie povrchovej stehennej tepny, 2) aplikovať štatistické metódy na riešenie chýbajúcich údajov, 3) vyhodnotiť HTA analýzu a porovnať vplyv metód na výsledok, 4) zhodnotiť využité metódy a vytvoriť odporúčania.

Po spracovaní prehľadu dostupných metód boli vybrané metódy: analýza kompletných prípadov, imputácia aritmetickým priemerom, mnohonásobná imputácia EM algoritmom a mnohonásobná imputácia MCMC.

Na základe výsledkov realizovaných na simulovaných dátach bolo štatistickým testom pre normalitu údajov potvrdené, že ani jeden dátový súbor nemá normálne rozdelenie. Neparametrickým párovým Wilcoxonovým testom boli skúmané rozdiely medzi zdrojovými údajmi a novými hodnotami. Pre analýzu kompletných prípadov bol v prípade intervencie PTA/S a bypass zistený štatisticky významný rozdiel od pôvodných údajov na hladine významnosti 5 %. Pre imputáciu aritmetickým priemerom, mnohonásobnú imputáciu EM algoritmom a mnohonásobnú imputáciu MCMC nebol zistený štatisticky významný rozdiel oproti zdrojovým dátam.

Následne bola spracovaná analýza nákladovej efektivity, ktorá bola vypočítaná pre štyri klinické efekty: primárna priechodnosť, technický úspech, ročné prežitie pacienta a záchrana končatiny v roku operácie.

Rozdiely vo výsledkoch nákladovej efektivity a hodnôt ICER boli percentuálne vyjadrené. Najväčšie rozdiely boli zaznamenané pri analýze kompletných prípadov. Imputácia aritmetickým priemerom mala lepšie výsledky ako analýza kompletných prípadov. Metódy mnohonásobného doplnenia – pomocou EM algoritmu a MCMC algoritmu mali najmenší percentuálny rozdiel s pôvodnými zdrojovými dátami.

Na základe vyhodnotenia výsledkov štatistických metód, bolo vytvorené odporúčanie k výberu metód pre riešenie chýbajúcich dát pre dosiahnutie dôveryhodného výsledku.

Chýbajúce, či neúplné dáta sú častým problémom vo vedných disciplínach. Diplomová práca prináša poznatky k spracovaniu a riešeniu tohto problému. Správnou voľbou metódy je možné zaznamenané údaje využiť efektívnejšie.

Zoznam použitej literatúry

[1] GRAHAM, John W. Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology* [online]. 2009, 60(1), 549–576. ISSN 0066-4308. Dostupné z: doi:10.1146/annurev.psych.58.110405.085530

[2] ENDERS, Craig K. *Applied missing data analysis* [online]. B.m.: Guilford Press, 2010.I SBN 9781606236390. Dostupné z: https://www.guilford.com/books/Applied-Missing-Data-Analysis/Craig-Enders/9781606236390

[3] IBRAHIM, Joseph G., Haitao CHU a Ming-Hui CHEN. Missing Data in Clinical Studies: Issues and Methods. *Journal of Clinical Oncology* [online]. 2012, 30(26), 3297–3303. ISSN 0732-183X. Dostupné z: doi:10.1200/JCO.2011.38.7589

[4] KNEIPP, S M a M MCINTOSH. Handling missing data in nursing research with multiple imputation. *Nursing research* [online]. nedatováno, 50(6), 384–9 [vid. 2019-04-16]. ISSN 0029-6562. Dostupné z: http://www.ncbi.nlm.nih.gov/pubmed/11725942

[5] ALLISON, Paul D. Missing Data. *Quantitative Applications in the Social Sciences* [online]. 2001, 104. ISSN 1468-5833. Dostupné z: doi:10.1136/bmj.38977.682025.2C

[6] LITTLE, Roderick J. A. a Donald B. RUBIN. *Statistical Analysis with Missing Data* [online]. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2002 [vid. 2018-05-15]. ISBN 9781119013563. Dostupné z: doi:10.1002/9781119013563

[7] GOODMAN, Clifford S. *HTA 101:* Introduction to health technology assesment [online]. 2004, Dostupné z: http://www.inahta.org/wp-content/uploads/2014/09/HTA-101_Goodman_2004.pdf

[8] KANG, Hyun. The prevention and handling of the missing data. *Korean Journal of Anesthesiology* [online]. 2013, 64(5), 402–406. ISSN 20056419. Dostupné z: doi:10.4097/kjae.2013.64.5.402

[9] NOVOVIČOVÁ Jana. Pravděpodobnost a matematická statistika [online]. Praha: Vydavatelství ČVUT, 1999. ISBN 80-01-01980-2. Dostupné z: http://euler.fd.cvut.cz/publikace/files/skripta3.pdf

[10] PROCHÁZKA, Bohumír. Stručná biostatistika pro lékaře. Vyd. 1. Praha: Karolinum, 2015. ISBN 978-80-246-2783-0.

[11] SCHAFER, J. L. (Joseph L.). *Analysis of incomplete multivariate data* [online]. B.m.: Chapman & Hall/CRC, 1997. ISBN 1439821860. Dostupné z: https://books.google.cz/books/about/Analysis_of_Incomplete_Multivariate_Data.html?id=3TFWRjn1f-oC&redir_esc=y

[12] DONG, Yiran a Chao-Ying Joanne PENG. Principled missing data method the researchers. *Springer Plus*. 2013, 2(222), 1–17.

[13] PIGOTT, Therese D. A Review of Methods for Missing Data. *Educational Research and Evaluation* [online]. 2001, 7(4), 353–383. ISSN 1380-3611. Dostupné z: doi:10.1076/edre.7.4.353.8937

[14] SOLEY-BORI, Marina. Dealing with missing data: Key assumptions and methods for applied analysis. *PM931 Directed Study in Health Policy and Management*. 2013, (4), 20.

[15] CHEEMA, Jehanzeb R. Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research. *Journal of Modern Applied Statistical Methods* [online]. 2014, 13(2), 53–75. ISSN 1538-9472. Dostupné

z: doi:10.22237/jmasm/1414814520

[16] SCHAFER, Josepn L. a John W. GRAHAM. Missing data: Our view of the state of the art. *Psychological Methods* [online]. 2002, 7(2), 147–177. ISSN 1082989X. Dostupné z: doi:10.1037//1082-989X.7.2.147

[17] RIEDEL, Marc a Wendy C. REGOECZI. Missing Data in Homicide Research. *Homicide Studies* [online]. 2004, 8(3), 163–192. ISSN 10887679. Dostupné z: doi:10.1177/1088767904265447

[18] ANDRIDGE, Rebecca R a Roderick J A LITTLE. A Review of Hot Deck Imputation for Survey Non-response [online]. nedatováno [vid. 2018-05-27]. Dostupné z: doi:10.1111/j.1751-5823.2010.00103.x

[19] JAKOBSEN, Janus Christian, Christian GLUUD, Jørn WETTERSLEV a Per WINKEL. When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology* [online]. 2017, 17(1), 1–10. ISSN 14712288. Dostupné z: doi:10.1186/s12874-017-0442-1

[20] ZENG, Donglin a D. Y. LIN. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* [online]. 2007, 69(4), 507–564. ISSN 13697412. Dostupné z: doi:10.1111/j.1369-7412.2007.00606.x

[21] LAMBERT, Paul C., Lucinda J. BILLINGHAM, Nicola J. COOPER, Alex J. SUTTON a Keith R. ABRAMS. Estimating the cost-effectiveness of an intervention in a clinical trial when partial cost information is available: a Bayesian approach. *Health Economics* [online]. 2008, 17(1), 67–81. ISSN 10579230. Dostupné z: doi:10.1002/hec.1243

[22] PETRÚŠEK, Ivan. *Analýza chybějících hodnot : srovnání metod při zkoumání determinantů politické znalosti a příjmu*. Prvé vydán. Praha: Sociologický ústav AVČR, 2017. ISBN 9788073302672.

[23] YUAN, Yang C. Multiple Imputation for Missing Data: Concepts and New Development. *SUGI 25: Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference* [online]. 2000, Paper 267-25. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.571.6854&rep=rep1&type=pdf

[24] VAN BUUREN, Stef a Karin GROOTHUIS-OUDSHOORN. MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software* [online]. nedatováno, VV(II). Dostupné z: http://www.jstatsoft.org/

[25] AZUR, Melissa J, Elizabeth A STUART, Constantine FRANGAKIS a Philip J LEAF. Multiple Imputation by Chained Equations: What is it and how does it work? [online]. Dostupné z: doi:10.1002/mpr.329

[26] Package „mice" Title Multivariate Imputation by Chained Equations [online]. 2019. Dostupné z: doi:10.18637/jss.v045.i03

[27] ZHANG, Zhongheng. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of translational medicine* [online]. 2016, 4(2), 30. ISSN 2305-5839. Dostupné z: doi:10.3978/j.issn.2305-5839.2015.12.63

[28] QUARTAGNO, Matteo a James R. CARPENTER. Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical Journal* [online]. 2019 ISSN 03233847. Dostupné z: doi:10.1002/bimj.201800222

[29] PECÁKOVÁ, Iva. Problém chýbajících dat v dotazníkových studiích. *Acta Oeconomica Pragensia*. 2014, (5), 66–78.

[30] TALJAARD, Monica, Allan DONNER a Neil KLAR. Imputation Strategies for

Missing Continuous Outcomes in Cluster Randomized Trials. *Biometrical Journal* [online]. 2008, 50(3), 329–345. ISSN 03233847. Dostupné z: doi:10.1002/bimj.200710423

[31] NÁROŽNÁ, Markéta. *DIPLOMOVÁ PRÁCE Imputace chybějících hodnot v rozsáhlých datových souborech*. B.m., 2013. Univerzita palackého v Olomouci.

[32] PETRÚŠEK, Ivan. *Analýza chýbajúcich hodnôt : porovnávanie vhodnosti tradičných metód naprieč mechanizmami*. B.m., 2015. Univerzita Karlova v Praze.

[33] POWNEY, M, P WILLIAMSON, J KIRKHAM a R KOLAMUNNAGE-DONA. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials* [online]. 2014, 15, 237. ISSN 1745-6215. Dostupné z: doi:10.1186/1745-6215-15-237

[34] CARIDES, G. W. A regression-based method for estimating mean treatment cost in the presence of right-censoring. *Biostatistics* [online]. 2000, 1(3), 299–313. ISSN 14654644. Dostupné z: doi:10.1093/biostatistics/1.3.299

[35] LIN, D Y, E J FEUER, R ETZIONI a Y WAX. Estimating medical costs from incomplete follow-up data. *Biometrics* [online]. 1997, 53(2), 419–34, ISSN 0006-341X. Dostupné z: http://www.ncbi.nlm.nih.gov/pubmed/9192444

[36] BAER, Onur, Joseph C GARDINER, Cathy J BRADLEY a Charles W GIVEN. Estimation from Censored Medical Cost Data [online]. nedatováno [vid. 2018-05-27]. Dostupné z: doi:10.1002/bimj.200210036

[37] LIN, D Y. Linear regression analysis of censored medical costs. *Printed in Great Britain Biostatistics* [online]. 2000, 1(1), 35–47 [vid. 2018-05-27]. Dostupné z: https://pdfs.semanticscholar.org/8400/182cdf89ce31e30ecc5acf45c497482df9c8.pdf

[38] OLSEN, I. C., T. K. KVIEN a T. UHLIG. Consequences of handling missing data for treatment response in osteoarthritis: A simulation study. *Osteoarthritis and Cartilage* [online]. 2012, 20(8), 822–828. ISSN 10634584. Dostupné z: doi:10.1016/j.joca.2012.03.005

[39] BURTON, Andrea, Lucinda Jane BILLINGHAM a Stirling BRYAN. Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. *Clinical Trials: Journal of the Society for Clinical Trials* [online]. 2007, 4(2), 154–161. ISSN 1740-7745. Dostupné z: doi:10.1177/1740774507076914

[40] VAN BUUREN, S, H C BOSHUIZEN a D L KNOOK. Multiple imputation od missing blood pressure covariates in survival analysis. *STATISTICS IN MEDICINE Statist. Med* [online]. 1999, 18, 681–694 [vid. 2018-05-27]. Dostupné z: http://www.stefvanbuuren.nl/publications/Multiple imputation - Stat Med 1999.pdf

[41] SIMONS, Claire L, OLIVER, Rivero-Arias , Ly-Mee YU a Judit SIMON. Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index? [online]. Dostupné z: https://link.springer.com/content/pdf/10.1007%2Fs11136-014-0837-y.pdf

[42] BURTON, Andrea, Douglas G. ALTMAN, Patrick ROYSTON a Roger L. HOLDER. The design of simulation studies in medical statistics. Statistics in Medicine [online]. 2006, 25(24), 4279–4292. ISSN 02776715. Dostupné z: doi:10.1002/sim.2673

[43] DÍAZ-ORDAZ, K, Michael G KENWARD a Richard GRIEVE. Handling missing values in cost effectiveness analyses that use data from cluster randomized trials. *J. R. Statist. Soc. A* [online]. 2014, 177(2), 457–474. Dostupné z: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12016

[44] HEITJAN, Daniel F., Clara Yuri KIM a Huiling LI. Bayesian estimation of cost-effectiveness from censored data. *Statistics in Medicine* [online]. 2004, 23(8),

1297–1309. ISSN 02776715. Dostupné z: doi:10.1002/sim.1740

[45] BRIGGS, Andrew, Taane CLARK, Jane WOLSTENHOLME a Philip CLARKE. *Missing....presumed at random: Cost-analysis of incomplete data* [online]. 2003. ISBN 1057-9230 (Print); 1057-9230 (Linking). Dostupné z: doi:10.1002/hec.766

[46] FARIA, Rita, Manuel GOMES, David EPSTEIN a Ian R WHITE. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. *PharmacoEconomics* [online]. 2014, 32(12), 1157–1170 ISSN 11792027. Dostupné z: doi:10.1007/s40273-014-0193-3

[47] BLOUGH, David K, Scott RAMSEY, Sean D SULLIVAN a Roger YUSEN. The quality impact of using diffierent imputation methods for missing quality of life scores on the estimation of the cost-effectivenes of lung volume reduction surgery. *Health Econ* [online]. 2009, 18, 91–101 Dostupné z: doi:10.1002/hec.1347

[48] VENABLES, W N, D M SMITH a R DEVELOPMENT CORE TEAM. *An introduction to R notes on R, a programming environment for data analysis and graphics* [online]. 2018. ISBN 3900051127 9783900051129. Dostupné z: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

[49] SIMONS, Claire L, @bullet OLIVER, Rivero-Arias @BULLET, Ly-Mee YU a Judit SIMON. Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index? [online]. nedatováno Dostupné z: https://link.springer.com/content/pdf/10.1007%2Fs11136-014-0837-y.pdf

[50] HONAKER, James, Gary KING a Matthew BLACKWELL. Amelia II: A Program for Missing Data. *Journal of Statistical Software* [online]. 2011, 45(7), 1–47. Dostupné z: doi:10.18637/jss.v045.i07

[51] TEMPL, M, P FILZMOSER, Matthias TEMPL a Peter FILZMOSER. *Visualization of missing values using the R-package VIM* [online]. nedatováno . Dostupné z: http://www.statistik.tuwien.ac.at

[52] ZHANG a ZHONGHENG. Missing data exploration: highlighting graphical presentation of missing pattern. *Annals of Translational Medicine* [online]. 2015, 3(22). ISSN 2305-5847. Dostupné z: doi:10.21037/8666

[53] SCHOUTEN, Rianne Margaretha, Peter LUGTIG a Gerko VINK. Generating missing values for simulation purposes: a multivariate amputation procedure Generating missing values for simulation purposes: a multivariate amputation procedure. *JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION* [online]. 2018, 88(15), 2909–2930. Dostupné z: doi:10.1080/00949655.2018.1491577

[54] WU, Wei, Fan JIA, Mijke RHEMTULLA a Todd D LITTLE. Search for efficient complete and planned missing data designs for analysis of change. *Behavior Research Methods* [online]. 2016, 48, 1047–1061. Dostupné z: doi:10.3758/s13428-015-0629-5

[55] LITTLE, T. D., T. D. JORGENSEN, K. M. LANG a E. W. G. MOORE. On the Joys of Missing Data. *Journal of Pediatric Psychology* [online]. 2014, 39(2), 151–162. ISSN 0146-8693. Dostupné z: doi:10.1093/jpepsy/jst048

[56] PRÁŠKOVÁ, Zuzana. *METODA BOOTSTRAP* [online]. 2004. Dostupné z: http://www.statspol.cz/robust/robust2004/praskova.pdf

[57] EFRON, Bradley, Robert J TIBSHIRANI, Boca RATON, London NEW a York WASHINGTON. *An Introduction to the Bootstrap* [online]. nedatováno [vid. 2019-04-27]. Dostupné z: https://cds.cern.ch/record/526679/files/0412042312_TOC.pdf

[58] KAMENSKÝ, Vojtěch. *Ekonomicko-klinické zhodnocení endovaskulární a chirurgické léčby u pacientů s postižením povrchní stehenní tepny*. B.m., 2014. b.n.

[59] RODWELL, Laura, Katherine J LEE, Helena ROMANIUK a John B CARLIN. Comparison of methods for imputing limited-range variables: a simulation study. *BMC Medical Research Methodology* [online]. 2014, 14(1), 57 [vid. 2019-04-16]. ISSN 1471-2288. Dostupné z: doi:10.1186/1471-2288-14-57

[60] DZIURA, James D., Lori A. POST, Qing ZHAO, Zhixuan FU a Peter PEDUZZI. Strategies for dealing with missing data in clinical trials: From design to analysis. *Yale Journal of Biology and Medicine*. 2013, 86(3), 343–358. ISSN 00440086.

[61] ZHANG, Zhongheng. Multiple imputation for time series data with Amelia package. *Annals of translational medicine* [online]. 2016, 4(3), 56, ISSN 2305-5839. Dostupné z: doi:10.3978/j.issn.2305-5839.2015.12.60

Zoznam obrázkov

Zoznam tabuliek

Zoznam grafov

Prílohy

Hodnoty nákladov pre intervenciu PTA

	C1 - Zdrojové dáta	C2 - Analýza komp. príp.	C3 - Imputácia aritm. Priemer	C4 - Imp. EM algoritmus	C 5 - Imputácia MCMC
1	35 529	35 529	35 529	35 529	35 529
2	44 214		51 659	44 539	51 726
3	45 762	45 762	45 762	45 762	45 762
4	30 333	30 333	30 333	30 333	30 333
5	187 455	187 455	187 455	187 455	187 455
6	59 425	59 425	59 425	59 425	59 425
7	42 379	42 379	42 379	42 379	42 379
8	29 797	29 797	29 797	29 797	29 797
9	29 082	29 082	29 082	29 082	29 082
10	43 230	43 230	43 230	43 230	43 230
11	49 807	49 807	49 807	49 807	49 807
12	35 969	35 969	35 969	35 969	35 969
13	53 082	53 082	53 082	53 082	53 082
14	33 384		33 153	32 570	32 570
15	37 973	37 973	37 973	37 973	37 973
16	33 796	33 796	33 796	33 796	33 796
17	37 332	37 332	37 332	37 332	37 332
18	44 661		42 102	38 294	38 294
19	68 385		66 527	67 449	64 876
20	57 271	57 271	57 271	57 271	57 271
21	50 255	50 255	50 255	50 255	50 255
22	31 101	31 101	31 101	31 101	31 101
23	26 138		26 721	26 742	26 340
24	52 159	52 159	52 159	52 159	52 159
25	56 512	56 512	56 512	56 512	56 512
26	41 825		40 515	49 848	33 703
27	47 044		46 145	43 262	38 767
28	113 737	113 737	113 737	113 737	113 737
29	42 338	42 338	42 338	42 338	42 338
30	34 424	34 424	34 424	34 424	34 424
31	34 845	34 845	34 845	34 845	34 845

32	42 777	42 777	42 777	42 777	42 777
33	85 730	85 730	85 730	85 730	85 730
34	39 158	39 158	39 158	39 158	39 158
35	43 563	43 563	43 563	43 563	43 563
36	52 463		56 404	47 947	58 950
37	40 189	40 189	40 189	40 189	40 189
38	37 968	37 968	37 968	37 968	37 968
39	46 840	46 840	46 840	46 840	46 840
40	52 698	52 698	52 698	52 698	52 698
41	33 511		33 569	33 017	33 947
42	147 268	147 268	147 268	147 268	147 268
43	40 767	40 767	40 767	40 767	40 767
44	36 813	36 813	36 813	36 813	36 813
45	43 269		43 102	38 796	46 038
46	39 616	39 616	39 616	39 616	39 616
47	36 794	36 794	36 794	36 794	36 794
48	36 819	36 819	36 819	36 819	36 819
49	39 665	39 665	39 665	39 665	39 665
50	55 778	55 778	55 778	55 778	55 778
51	52 615	52 615	52 615	52 615	52 615
52	39 474	39 474	39 474	39 474	39 474
53	39 759	39 759	39 759	39 759	39 759
54	40 877	40 877	40 877	40 877	40 877
55	41 326	41 326	41 326	41 326	41 326
56	37 823		47 634	41 738	44 592
57	37 800	37 800	37 800	37 800	37 800
58	66 641		41 956	51 232	29 881
59	62 410		62 242	67 290	65 745
60	74 581	74 581	74 581	74 581	74 581
61	27 786		33 587	33 670	26 296
62	28 651	28 651	28 651	28 651	28 651
63	44 908	44 908	44 908	44 908	44 908
64	69 323	69 323	69 323	69 323	69 323
65	68 982		69 837	69 828	69 853

Hodnoty nákladov pre intervenciu PTA/s

	C1 - Zdrojové dáta	C2 - Analýza komp. Príp.	C3 - Imputácia aritm. Priemer	C4 - Imp. EM algoritmus	C 5 - Imputácia MCMC
1	66 767		86 330	68 693	73 883
2	108 173	108 173	108 173	108 173	108 173
3	61 981	61 981	61 981	61 981	61 981
4	233 090	233 090	233 090	233 090	233 090
5	89 749		100 893	62 087	88 446
6	75 506	75 506	75 506	75 506	75 506
7	79 095	79 095	79 095	79 095	79 095
8	139 643	139 643	139 643	139 643	139 643
9	79 341	79 341	79 341	79 341	79 341
10	77 053	77 053	77 053	77 053	77 053
11	74 905		74 672	79 183	71 332
12	73 924	73 924	73 924	73 924	73 924
13	73 074	73 074	73 074	73 074	73 074
14	63 587	63 587	63 587	63 587	63 587
15	65 510		66 880	66 749	67 321
16	79 320		82 540	70 689	76 133
17	74 363	74 363	74 363	74 363	74 363
18	136 497	136 497	136 497	136 497	136 497
19	171 551	171 551	171 551	171 551	171 551
20	67 806	67 806	67 806	67 806	67 806
21	70 718	70 718	70 718	70 718	70 718
22	118 471	118 471	118 471	118 471	118 471
23	74 984	74 984	74 984	74 984	74 984
24	120 098	120 098	120 098	120 098	120 098
25	64 789	64 789	64 789	64 789	64 789
26	106 699	106 699	106 699	106 699	106 699

Hodnoty nákladov pre intervenciu bypass

	C1 - Zdrojové dáta	C2 - Analýza kompletných prípadov	C3 – Imp. aritm. priemer	C4 - Imp. EM algoritmus	C 5 – Imp. MCMC algoritmus
1	55 419	55 419	55 419	55 419	55 419
2	24 569	24 569	24 569	24 569	24 569
3	76 853		78 418	88 679	72 470
4	34 102	34 102	34 102	34 102	34 102
5	50 922	50 922	50 922	50 922	50 922
6	46 913	46 913	46 913	46 913	46 913
7	77 793		93 293	89 539	87 346
8	39 719	39 719	39 719	39 719	39 719
9	36 918	36 918	36 918	36 918	36 918
10	46 588	46 588	46 588	46 588	46 588
11	206 017	206 017	206 017	206 017	206 017
12	45 147	45 147	45 147	45 147	45 147
13	46 863	46 863	46 863	46 863	46 863
14	54 440	54 440	54 440	54 440	54 440
15	45 937	45 937	45 937	45 937	45 937
16	66 527		75 924	74 776	71 575
17	34 500	34 500	34 500	34 500	34 500
18	75 514	75 514	75 514	75 514	75 514
19	235 388	235 388	235 388	235 388	235 388
20	114 834	114 834	114 834	114 834	114 834
21	106 916	106 916	106 916	106 916	106 916
22	35 686	35 686	35 686	35 686	35 686
23	45 410	45 410	45 410	45 410	45 410
24	39 115	39 115	39 115	39 115	39 115
25	37 749	37 749	37 749	37 749	37 749
26	36 324	36 324	36 324	36 324	36 324
27	99 078	99 078	99 078	99 078	99 078
28	30 667	30 667	30 667	30 667	30 667
29	43 993	43 993	43 993	43 993	43 993
30	32 916	32 916	32 916	32 916	32 916
31	43 933	43 933	43 933	43 933	43 933
32	49 287	49 287	49 287	49 287	49 287
33	45 142	45 142	45 142	45 142	45 142

34	27 697	27 697	27 697	27 697	27 697
35	43 127	30 412	48 699	48 851	53 193
36	54 823	54 823	54 823	54 823	54 823
37	76 243	76 243	76 243	76 243	76 243
38	67 069	67 069	67 069	67 069	67 069
39	70 025	70 025	70 025	70 025	70 025
40	56 687		58 910	50 404	45 291
41	63 973	63 973	63 973	63 973	63 973
42	115 381		71 239	53 606	104 225
43	58 740	58 740	58 740	58 740	58 740