VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

POČÍTAČOVÁ PODPORA PRO MONITORING A HODNOCENÍ KVALITY DAT V KLINICKÉM VÝZKUMU

COMPUTER-AIDED DATA QUALITY MONITORING AND ASSESSMENT IN CLINICAL RESEARCH

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. Branislav Šiška
AUTHOR

VEDOUCÍ PRÁCE — doc. Ing. Daniel Schwarz, Ph.D.
SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Biomedicínské a ekologické inženýrství**

Ústav biomedicínského inženýrství

Student: Bc. Branislav Šiška ***ID:*** 155605

Ročník: 2 ***Akademický rok:*** 2017/18

NÁZEV TÉMATU:

Počítačová podpora pro monitoring a hodnocení kvality dat v klinickém výzkumu

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši problematiky kvality dat v klinickém výzkumu s vazbou na existující řešení pro elektronický sběr dat (EDC informační systémy – Electronic Data Capture). 2) U vybraného informačního systému navrhněte rozšíření palety funkcí o automatické monitorování kvality dat: a) detekci anomálních záznamů pomocí statistických metod, b) detekci anomálních záznamů pomocí metod strojového učení (machine learning) a rozpoznávání vzorů (pattern recognition), a to včetně návrhu hodnocení úspěšnosti detekce. 3) Zabývejte se možnostmi předzpracování dat pomocí transformací datových záznamů s proměnnými různých datových typů na numerické vektory. 4) Navržené naprogramujte, přičemž využijte SQL databázi a jeden ze skriptovacích jazyků (např. Python, PHP) nebo jiné vývojové prostředí (např. Matlab, R). Využijte anonymizovaná data z již uzavřených zdravotnických registrů nebo z neintervenčních klinických studií.

DOPORUČENÁ LITERATURA:

[1] Stephen L George & Marc Buyse: Data fraud in clinical trials, DOI: 10.4155/CLI.14.116.

[2] Švihálková H: Aplikace shlukovacích metod na data klinických registrů [online]. Brno, 2011. Dostupné z: <http://is.muni.cz/th/208192/prif_m/>. Diplomová práce. Masarykova univerzita, Přírodovědecká fakulta. Vedoucí práce Daniel Klimeš.

Termín zadání: 5.2.2018 ***Termín odevzdání:*** 18.5.2018

Vedoucí práce: doc. Ing. Daniel Schwarz, Ph.D.

Konzultant:

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

ABSTRAKT

Diplomová práca sa zaoberá monitoringom a hodnotením kvality dát v klinickom výskume. Obvyklý spôsob na identifikáciu odľahlých hodnôt využíva jednorozmerné štatistické metódy pre každú premennú z formulára klinickej štúdie. Metóda popísaná v diplomovej práci vstupuje priamo do databáze štúdie a detekuje odľahlé hodnoty pomocou strojového učenia a viacrozmerného štatistického prístupu, ktorý transformuje všetky premenné z formulára do jednej, reprezentujúcej odpovedajúci záznam pacienta. Navrhnutý algoritmus je navrhnutý v programovom prostredí Matlab.

KĽÚČOVÉ SLOVÁ

EDC systémy, klinické štúdie, falšovanie dát, stopovanie zmien, kvalita dát, monitoring, hodnotenie kvality

ABSTRACT

The diploma thesis deals with the monitoring and evaluation of data in clinical research. Usual methods to identify incorrect data are one-dimensional statistical methods per each variable in the register. Proposed method enters directly into database and finds out outliers in data using machine learning combined with multidimensional statistical methods that transform all column variables of clinical register to one, representing one record of patient in the register. Algorithm of proposed method is written in Matlab.

KEYWORDS

EDC systems, clinical trials, data fraud, audit trail, data quality, monitoring, quality assessment

Šiška, B. *Počítačová podpora pro monitoring a hodnocení kvality dat v klinickém výzkumu*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2018. 67 s. Diplomová práce. Vedoucí práce: doc. Ing. Daniel Schwarz, Ph.D.

PREHLÁSENIE

Prehlasujem, že svoju diplomovú prácu na téma Počítačová podpora pro monitoring a hodnocení kvality dat v klinickém výzkumu som vypracoval samostatne pod vedením vedúceho diplomovej práce a s použitím odbornej literatúry a ďalších informačných zdrojov, ktoré sú citované a uvedené v zozname použitých zdrojov na konci práce.
Ako autor uvedenej diplomovej práce ďalej prehlasujem, že v súvislosti s vytvorením tejto práce som neporušil autorské práva tretích osôb, nezasiahol som nedovoleným spôsobom do cudzích autorských práv osobnostných a som si plne vedomý následkov porušenia ustanovenia § 11 a nasledujúcich autorského zákona č. 121/2000 Sb., vrátane možných trestnoprávnych dôsledkov vyplývajúcich z ustanovení časti druhej, hlavy VI. diel 4 Trestného zákonníka č. 40/2009Sb.

V Brne dňa Podpis autora

POĎAKOVANIE

Ďakujem vedúcemu mojej diplomovej práce doc. Ing. Danielovi Schwarzovi, Ph.D. za odbornú pomoc, rady a čas, ktorý si našiel pri spracovávaní mojej diplomovej práce. Tak isto ďakujem svojej rodine za nekončiacu podporu a dodávanie motivácie do života.

ZOZNAM OBRÁZKOV

ZOZNAM TABULIEK

ÚVOD

V dnešnej dobe sú súčasné štandardy liečby a doporučené farmakoterapeutické postupy založené na dôkazoch – Evidence based medicine. Pred uvedením nového liečivého prípravku na trh, musí prípravok najprv prejsť registráciou, kedy regulačné úrady podrobne skúmajú predložené dokumenty. Jedná sa hlavne o bezpečnosť, účinnosť a kvalitu hodnoteného prípravku. Súčasťou registrácie je aj súhrn údajov o prípravku, ktorý je kľúčovým zdrojom informácií o danom liečivom prípravku pre lekárov a zdravotných odborníkov. Úspešný registračný proces musí preukázať, bezpečnosť a úspešnosť nového liečiva ešte pred jeho uvedením do klinickej praxe. Práve na to slúžia klinické štúdie, ktoré v súlade s existujúcimi celosvetovo uznávanými pravidlami preukazujú tieto vlastnosti. Výsledkom týchto štúdií je jednoznačne definovaná cieľová skupina pacientov, pre ktorých bude mať liečba najväčší prínos. Obsahuje aj bezpečnostný profil, kde sú zaznamenané očakávané nežiadúce účinky daného liečiva a vyhodnotenie pomeru prínosu a rizík pri použití liečiva v praxi. Ďalej sa sleduje farmakovigilancia, ktorá skúma výskyt nežiadúcich účinkov a sledovanie rizík liečiva po uvedení do klinickej praxe [1][2].

Klinický register je systém jasne definovaných údajov zdravotných, alebo demografických od pacientov so špecifickými zdravotnými charakteristikami, ktoré sú uchované v centrálnej databáze. Klinické registre môžu slúžiť ako monitorovací nástroj na zlepšenie lekárskej starostlivosti[1]. Pri tvorbe registrov sú však často údaje o pacientovi zadané nepresne, alebo sú niektoré polia dokonca vynechané. Chýbajúce dáta tvoria v registroch 6% – 50%. To spôsobuje nepresnosti pri výpočtoch, čo môže viesť ku nesprávnym konečným výsledkom celej štúdie. Chybné dáta, alebo inak anomálne údaje delíme na systematicky vzniknuté a náhodné. Systematické anomálie môžu byť dôsledkom chýb zberného programu, nejasne definovaných parametrov, alebo zneužitia zberu dát. Náhodné chyby vznikajú nepresným prepisom dát, alebo preklepmi.

V Českej republike je najvyšší regulačným orgánom, zodpovedným za proces registrácie nových liečiv Štátny ústav pre kontrolu liečiv – SÚKL. Na Európskej úrovni za proces registrácie odpovedá Európska lieková agentúra – European Medicine Agency – EMA. V Spojených štátoch amerických je to zase Úrad pre kontrolu potravín a liečiv – US Food and Drug Administration – FDA. Každoročne je v Českej republike podľa autorov knihy [2] predložených Štátnemu ústavu pre kontrolu liečiv približne 300 žiadostí o povolenie klinickej štúdie liečivého prípravku. Z tohto počtu je približne 3% z akademického prostredia. Ostatné štúdie sú predkladané súkromnými spoločnosťami, prevažne zo zahraničia. Realizáciu klinickej štúdie je potrebné dôkladne premyslieť ešte pred jej zahájením, aby sa predišlo najhoršiemu možnému scenáru, a to predčasné ukončenie už zahájenej štúdie z dôvodu neschopnosti zvládnutia zadávateľa. Všetko úsilie a finančné prostriedky by tak vyšli na zmar.

Cieľom práce je navrhnúť algoritmus na zistenie anomálií pri zadávaní dát do databáze klinických štúdií. Algoritmus bude vedieť odhaliť chyby vzniknuté z nepozornosti, ako aj cielené zadávanie fiktívnych hodnôt.

1 KLINICKÉ ŠTÚDIE

Termínom klinická štúdia sa označujú projekty, týkajúce sa klinického výskumu. Ide hlavne o klinické hodnotenie humánnych liečivých prípravkov, alebo skúšky zdravotníckych prostriedkov. Môžu to však byť aj štúdie, ktorých cieľom je hodnotenie účinnosti a bezpečnosti diagnostických, preventívnych a liečebných metód [2].

Klinické hodnotenie humánnych liečivých prípravkov predstavuje systematické testovanie účinnosti, bezpečnosti a akosti liečebného prípravku, na ktorom sa podieľajú pacienti, alebo zdravotní dobrovoľníci, pod vedením skúšajúcich doktorov. Tento proces predstavuje do teraz najdokonalejší spôsob ako získať dôkazy a dáta o účinnosti a bezpečnosti daného liečivého prípravku. Pretože sú však najčastejšími iniciátormi a realizátormi farmaceutické spoločnosti a výrobcovia zdravotníckych technológií, ktoré tieto štúdie dokladajú k registrácií a uvedeniu nového liečiva na trh, môžu mať na konečné výsledky týchto štúdií vplyv nie len vedecké metódy, ale aj obchodné záujmy. Ak má byť výsledok klinickej štúdie považovaný za vierohodný, čo býva hlavným cieľom sponzora, je nutné doložiť ku popisu metód štatistickej analýzy nazbieraných dát aj metodiku zaistenia ich kvality.

1.1 História

K rozmachu klinických štúdií došlo hlavne v druhej polovici 20. storočia. V tomto období totiž vzniklo veľké množstvo nových liečivých látok, pre do vtedy neliečiteľné choroby. Dnes je počet liečiv oveľa väčší, avšak stále sa hľadajú nové a vyvíjajú viac vhodné prípravky s vhodnejšími vlastnosťami pre pacientov. Nový prípravok by mal zlepšiť celkovú kvalitu života pacienta, jeho bezpečnosť či znížiť nákladnosť liečby. V súčasnosti nové účinné látky môžu vzniknúť týmito spôsobmi [2][3]:

- Modifikácia chemickej štruktúry už známeho liečiva
- Využitie novo objavených prírodných látok
- Objavenie nových vlastností už známych chemických látok
- Cielená syntéza nových chemických látok
- Systematický screening vybraných molekúl s určitou biologickou aktivitou

Nové molekuly s vyhovujúcimi biologickými vlastnosťami často využívajú počítačové technológie liekového designu Computer Aided/Assisted Drug Design CAAD, ktorý napomáha pri hľadaní a modelovaní nových molekúl [4].

1.2 Cyklus klinickej štúdie

Proces kým sa novo vyvinutý prípravok dostane k pacientovi je veľmi dlhý a náročný, a musí absolvovať radu klinických štúdií. Vstup nového liečivého prípravku je možný až po úspešnej klinickej štúdií, ktorá završuje aplikovaný klinický výskum. Laboratórne a predklinické skúšky prevádzané na zvieracích modeloch, ktoré sa uskutočňujú ešte pred tým ako sa liečivo začne podávať prvým ľudským subjektom sa neustále rozširujú. Stále precíznejší a detailnejší prístup k prevedení klinických štúdií však významne zvyšuje finančnú a časovú náročnosť jednotlivých projektov. Bezpečnosť a účinnosť nového liečiva sa postupne hodnotí v niekoľkých fázach. Každá fáza má dopredu definované hypotézy a ciele. Prechod liečiva do nasledujúcej fázy je možný len v prípade, ak úspešne splní fázu predchádzajúcu.

Na vývoji nového liečiva sa podieľajú široké medzioborové tímy so špecializáciou z oblastí medicíny, farmácie, biológie, technológie, etiky, ekonómie, štatistiky a legislatívy. Nároky na vývoj nového liečivého prípravku tak stále rastú, čo predstavuje aj stále náročnejší proces vývoja. To môže byť spolu s finančnou rizikovosťou jeden z dôvodov, prečo môžeme v posledných rokoch sledovať pokles počtu novo udelených registrácií liečiv a celkové predĺženie doby ich vývoja [5][6]. Od vývoja jedného liečiva až po jeho úspešnú registráciu s uvedením na trh trvá približne od 10 do 15 rokov, pričom náklady sa pohybujú približne od 800 miliónov do 1 miliardy USD [2][7][8].

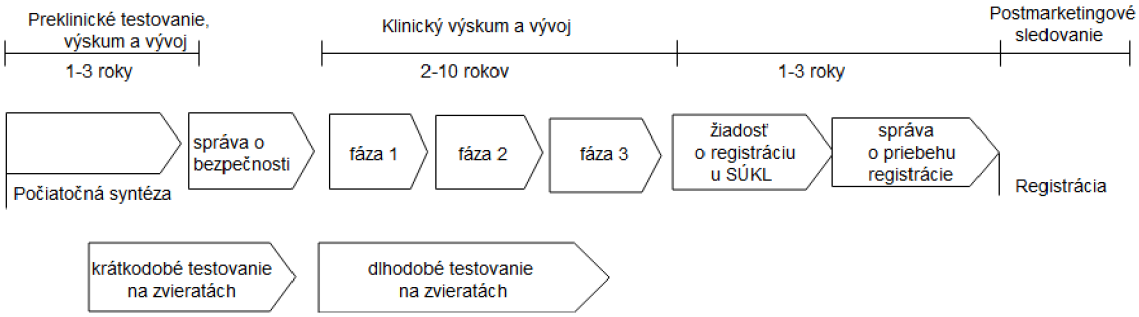
Dnes vývoj nového liečiva začína obvykle testovaním molekúl s dobrým farmakologickým účinkom v laboratóriách, kde sa hľadá molekula s dostatočným liečebným potenciálom a súčasne nízkou toxicitou. Takto vybraná molekula prechádza preklinickými štúdiami, ktorých cieľ je preukázať farmakologickú účinnosť a zistiť údaje o jej toxicite, alebo tolerabilite a tiež zistiť bezpečnosť pri interakcii s živým organizmom. V tejto časti ešte nemôžme hovoriť o pojme liečivo. Aj cez snahu obmedziť testy na zvieracích modeloch na maximum, čo sa vďaka in vitro metodike čiastočne darí, hrajú pokusné zvieratá nenahraditeľnú rolu. Hlavne preto, že pred podaním človeku je nutné overiť bezpečnosť a účinnosť prípravku na orgánové systémy, ktoré sú prítomné len v komplexne živom organizme. Preklinické štúdie teda obecne hodnotia, aká bezpečná a účinná nová látka je a či sa dá predpokladať jej klinické využitie. Ak sa tento predpoklad potvrdí, vývoj postupuje do 1. fáze klinickej štúdie. Počet molekúl, ktoré postúpi do 1. fáze klinickej štúdie je približne 10% z celkového počtu molekúl testovaných na zvieracích modeloch.

Legislatíva v Českej republike vyžaduje, aby klinické štúdie fázy 1 až 3 boli ešte pred ich začatím povolené Štátnym ústavom pre kontrolu liečiv SÚKL, niekedy tiež označovaným ako regulačná autorita. Klinická štúdia fázy 4 už nemusí byť schválená SÚKLom, ale musí mu byť oznámená.

Fáza 1 predstavuje podanie liečivého prostriedku človeku po prvý krát. Zaujímavé dáta sú bezpečnosť, znášanlivosť, tolerancia a nežiadúce účinky daného liečivého prostriedku. Testy prebiehajú väčšinou na zdravých dobrovoľníkoch a hľadá sa hodnota ideálnej dávky. Samotné testovanie prebieha v špecializovaných jednotkách tzv. klinickofarmakologických jednotkách, kde sú dobrovoľníci pod dohľadom lekárov. Ak je vyvíjané liečivo samo o sebe toxické, napríklad cytostatiká pri liečbe nádorových ochorení, fáza 1 už prebieha rovno na pacientoch a nie na zdravých dobrovoľníkoch.

Vo fáze 2 sa uskutočňuje pilotná štúdia. Zisťuje sa, ako na nový liečivý prípravok reaguje málo početná skupina, väčšinou ide o desiatky chorých pacientov. Po prvý krát sa okrem bezpečnosti sleduje aj účinnosť.

Vo fáze 3 prebiehajú rozsiahle randomizované klinické štúdie. Testy prebiehajú na veľkom súbore pacientov. Stanovuje sa najmä podrobná účinnosť liečivého prípravku na špecifické ochorenie, napríklad pri priamom porovnaní s iným. Ďalej sa sledujú výskyt a sila nežiadúcich účinkov a celkový bezpečnostný profil. Na konci fázy 3 sú výsledky spolu s výsledkami fáze 2 predložené regulačným autoritám. Slúžia ako nutné podklady pre povolenie použitia liečivého prípravku v klinickej praxi.



Obrázok 1 – vývoj nového liečiva, upravené a prevzaté z [2]

Po registrácii liečiva prebieha fáza 4. Overujú sa vlastnosti liečiva v reálnom prostredí klinickej praxe. Testovacia vzorka by mala byť menej selektovaná ako v predchádzajúcich fázach a čo najväčšia. Pre neintervenčné štúdie platí, že by žiadna riadená selekcia náboru subjektov ani nemala nastať a rozdelenie testovacej vzorky, by malo čo najviac odpovedať rozdeleniu skutočnej populácie. Veľký dôraz je kladený na sledovanie nežiadúcich účinkov, keďže s veľkým súborom testovacích subjektov sa môžu prejaviť do teraz nezaznamenané vedľajšie účinky. Ak by bol v tejto fáze spozorovaný nový nežiadúci účinok, je veľmi dôležité ihneď výsledky ohlásiť relevantným štátnym orgánom [2][9].

Dnes sa stále viac upúšťa od tohto tradičného delenia na 4 fázy a častejšie dochádza ku kombinácií jednotlivých fáz štúdií. Klinické hodnotenie liečiv je štandardizované a regulované legislatívnymi normami, ktoré napomáhajú k získaniu relevantných dát. Regulácia vývoja súčasne prispieva ku skvalitneniu výskumu a zvýšeniu bezpečnosti pacientov a zdravých dobrovoľníkov vo všetkých fázach klinických štúdií. Grafický prehľad jednotlivých fáz a vývoja liečiva je zobrazený na Obrázok 1.

1.2.1 Zabezpečenie a kontrola akosti

Zabezpečovanie a kontrola akosti – quality assurence/quality control – QA/QC je povinná legislatívna požiadavka pre osoby a organizácie uskutočňujúce klinické štúdie. Systém zabezpečenia a kontroly akosti je dôležitý nástroj, ktorý ma zaistiť, aby v priebehu klinickej štúdie neprichádzalo k porušovaniu etických princípov a aby získané dáta boli dostatočne presné a vierohodné.

Základom kontrolného systému sú písomné dokumenty, tzv. štandardné operačné procesy – SOP. Tieto dokumenty popisujú proces získavania, spracovania a dokumentácie údajov. Sú spísané v súlade so správnou klinickou praxou – good clinical practise – GCP a príslušnými právnymi predpismi. Dôvodom vzniku GCP bola séria neúspešných a neefektívnych klinických štúdií v minulosti. GCP reguluje a popisuje návod, ako by mala byť klinická štúdia vykonaná, popisuje role a zodpovednosti pre sponzorov, investigátorov a monitorov. Obsahuje tiež dokument o ochrane osobných údajov pre subjekty a dobrovoľníkov klinických štúdií.

Hlavná úloha dokumentov SOP je popísať, ako postupovať pri jednotlivých konkrétnych činnostiach v rámci klinickej štúdie a tieto procesy štandardizovať. V každom SOP dokumente je popísané na čo pri samotnej realizácií kroku nezabudnúť, poprípade šablóny, ktoré slúžia na uľahčenie spracovania dokumentov.

Z časového hľadiska sa jednotlivé fázy štúdie delia na:

- Fáza pred zahájením štúdie
- Fáza priebehu štúdie
- Fáza po skončení štúdie

Vo fáze pred zahájením klinickej štúdie by mali byť použité SOP na základné dokumenty ako protokol klinickej štúdie, písomné informácie pre pacientov a dokument informovaný súhlas pacienta. Ďalej dokument s informáciami pre skúšajúcich a na záznamy subjektov štúdie – case report form – CRF.

Vo fáze priebehu štúdie sa SOP používajú na činnosti spojené s monitorovaním. Napríklad postupy pri monitorovacej návšteve, postupy pre zadanie a opravu CRF a overovanie jednotlivých záznamov oproti zdrojovým dátam – source data verification – SDV. Ďalej sa používajú pri zaznamenávaní a hlásení nežiadúcich príhod – farmakovigilancií, alebo pri hlásení o uskutočnení štúdie na SÚKL. Po skončení SOP ošetrujú hlavne uzatvorenie klinickej štúdie v každom centre, uchovaním základnej dokumentácie, informovaní SÚKL o ukončení danej štúdie a nakoniec vytvorením súhrnnej záverečnej správy.

Po zavedení systému zabezpečenia akosti je ale rovnako dôležité zaistiť aj jeho kontrolu. Inštitúcie podieľajúce sa na realizácii štúdie majú nastavené pravidelné kontroly a audity. Nástrojom na kontrolu kvality akosti sú periodické kontroly kvality. Cieľom je overiť, či priebeh štúdie je v súlade s protokolom, správnou klinickou praxou a štandardnými operačnými postupmi. Hodnotia sa oblasti ako nábor pacientov do štúdie, dokumentácia klinickej štúdie, splnenie regulačných požiadavkou, etické a bezpečnostné požiadavky, študijný tím, dodržovanie protokolu a správne zaobchádzanie s hodnotenými liečivami.

O každej kontrole musí existovať písomný záznam. Ak by sa objavil systémový problém, môže nastať až úprava niektorého zo súvisiacich pracovných postupov systému zabezpečenia a kontroly akosti. Pri takomto zásahu je nutné informovať osobu zodpovednú za udržovanie systému zabezpečovania a kontroly akosti. Ďalšou možnosťou ako zaistiť dodržiavanie kontroly kvality je interný audit. Posudzujú sa rovnaké aspekty klinickej štúdie ako pri periodickej kontrole. Rozdiel je však v tom, že zatiaľ čo periodické kontroly uskutočňuje najčastejšie projektový manažér štúdie, pri internom audite je to osoba, ktorá je na danej klinickej štúdií nezávislá a teda sa na nej aktívne nepodieľa. Pre zaistenie úplnej nezávislosti môže byť na audit najatá aj osoba z tretej strany. Vtedy ide o externý audit. Je to síce finančne najnáročnejšia možnosť, ale z hľadiska odhalenia nedostatkov v procese klinickej štúdie, či v samotnom systéme QA/QC najúčinnejšia.

Auditovanie by malo prebiehať podľa písomných postupov, v ktorých je zdokumentované čo a ako auditovať, frekvencia uskutočnenia auditov, ich forma a obsah. Každý audit by mal zohľadňovať typ, rizikovosť pre subjekty a zložitosť danej klinickej štúdie. Všetky výsledky a pozorovania sú zaznamenané v správe z auditu – audit report [10].

1.2.2 Dokumentácia

Dokumenty, bez ktorých sa žiadna klinická štúdia nezaobíde sú tzv. základné dokumenty – essential documents. Sú to dokumenty, ktoré dovoľujú samostatne, alebo ako celok hodnotiť danú klinickú štúdiu a kvalitu získaných údajov. Okrem základných dokumentov musí zadávateľ klinickej štúdie vytvoriť aj ďalšie, ktoré mapujú bezproblémové vedenie štúdie. Môžu to byť napríklad monitorovací plán, validačný plán, plán štatistických analýz, inštrukcie o zachádzaní s liečivými prípravkami, inštrukcie pre hlásenia pri sledovaní bezpečnosti, inštrukcie pre obsluhu systému na zber elektronických dát, návod na randomizáciu subjektov, zoznam subjektov štúdie a zoznam členov výskumného tímu.

Najdôležitejším dokumentom klinickej štúdie je protokol. Obsahuje všetky dôležité informácie o štúdií, ako jej popis, prečo sa daná štúdia realizuje, jej ciele a plán, metodika, postup zberu a spracovania dát. Tento dokument je väčšinou tvorený celým tímom odborníkov. Tím je tvorený minimálne lekárom, ktorý je odborníkom na danú problematiku, štatistického odborníka a projektového manažéra. Niekedy tento tím rozširuje aj monitor danej štúdie. Protokol je záväzný a preto ho musia dodržovať všetky osoby, ktoré sa na klinickej štúdii podieľajú. Dokument je veľmi obsiahli a väčšinou presahuje 100 strán.

Ďalším dôležitým dokumentom je písomný informovaný súhlas subjektov v danej štúdií. Tento dokument musí byť podpísaný každým subjektom štúdie a dokazuje, že v danej štúdii je subjekt dobrovoľne a bol oboznámený z celým priebehom, jej účelom a dôležitosťou štúdie. Text musí byť pre pacienta zrozumiteľný, bez zložitých a odborných výrazov. Lekár mu vysvetlí, čo sa od neho bude vyžadovať a čo jeho účasť v danej klinickej štúdii obnáša. Subjektu je vždy ponechaný dostatočný čas na preskúmanie dokumentu a jeho prípadné doplňujúce otázky sú zodpovedané zo strany prítomného lekára.

Pre každú klinickú štúdiu platí, že je potrebné zaistiť, aby osoby, ktoré budú zadávať dáta do CRF mali potrebné vzdelanie a aby vedeli ako s daným systémom CRF správne pracovať. Za týmto účelom je tvorený podrobný, ale praktický dokument užívateľský návod. Obsahuje základné informácie o systéme na zber dát, ako sa do systému správne prihlásiť s menom a heslom, ako uložiť zadané dáta a ako sa zo systému správne odhlásiť. Návod opisuje ako sa zakladá nový subjekt v systéme a akým spôsobom sa zakladá pre subjekt primárny kľúč, ktorý ho jednoznačne identifikuje po celú dobu klinickej štúdie. Ďalej zdôrazňuje akým formátom sa zadávajú časové údaje, desatinné čísla či textové polia. Pri eCRF je v príručke telefonický či emailový kontakt na podporu – helpdesk, kde sa môžu užívatelia v prípade nezrovnalosti, či problému obrátiť [2].

1.2.3 Tvorba CRF

Na získanie dát vo formáte, ktorý umožňuje ich následné spracovanie a štatistické zhodnotenie je potrebné sledované parametre o účinnosti a bezpečnosti skúmaného liečiva zbierať v dokumente nazvanom záznam subjektu štúdie – case report form – CRF. Tento formulár je definovaný ako vytlačený, alebo elektronický dokument navrhnutý k zaznamenaniu všetkých protokolom vyžadujúcich informácií, ktoré budú následne predané zadávateľovi o každom subjekte danej štúdie. CRF musí byť vytvorený, testovaný a validovaný ešte pred samotným zahájením zberu dát. Keďže úpravy CRF v priebehu klinickej štúdie sú veľmi zložité a môžu narušiť integritu zhromažďovaných dát, je dôležité už prvotný návrh CRF vytvoriť zodpovedne.

Dnes sa na zber dát najčastejšie využíva elektronický CRF pomocou niektorého zo systémov electronic data capture – EDC. Ide o spojenie počítačového programu a databázy, ktoré slúžia na vytvorenie a správu jednotlivých eCRF. Systémy EDC odstraňujú nevýhody CRF, ako napríklad náklady na tlač formulárov, distribúciu a ich následný zber z centier prevádzajúcich klinické štúdie. Odpadá tiež zložitá kontrola úplnosti veľkého množstva dokumentov a ich následné uskladnenie a archivácia. Ďalšou výhodou je, že dáta sú už v elektronickej podobe a tak odpadá nutnosť ručne digitalizovať dáta z tlačených dokumentov CRF pre systém managementu dát – data management system – DMS [2].

Samotná digitalizácia dát býva veľmi často zdrojom chýb z nepozornosti, alebo z nedostatočnej čitateľnosti dát v CRF. Pri elektronickom CRF je aj proces kontroly a čistenia dát výrazne rýchlejší a jednoduchší. Pri rôznych nezrovnalostiach a ďalších otázkach môže byť eCRF opäť lepšia možnosť, keďže väčšina programov obsahuje nápovedu a zadávanie prebieha priamo pomocou nástrojov, ktoré sú dostupné v EDC. Elektronický CRF má však v určitých prípadoch aj svoje nevýhody. Tie sa prejavia napríklad pri menších klinických štúdiách s menším počtom centier. Pri takýchto štúdiách je vhodnejšie a jednoduchšie použiť klasický tlačený formát CRF. Kvalitne navrhnutý CRF, či eCRF môže výrazne redukovať chyby pri zadávaní dát, ich následnej kontrole, validácií a pri konečnom vyhodnotení. Ideálne, design a tvorba CRF prebieha súčasne s tvorbou protokolu štúdie. Pri tvorbe je potrebná znalosť všetkých dát, ktoré sa budú v danej klinickej štúdii zbierať. Častým javom je, že do štúdie sa zbierajú aj nadbytočné dáta s úmyslom ich neskoršieho vyhodnotenia. To však môže paradoxne viesť k zníženiu kvality dát, keďže tento proces zvyšuje nároky na pracnosť kontroly správnosti dát. Pre správny rozsah zbieraných dát je preto dôležitá spolupráca medzi manažérmi dát a klinickými odborníkmi.

Ďalšie dôležité dokumenty popisujúce klinickú štúdiu sú dáta management plán, ktorý ucelene popisuje systém spracovania dát, plán štatistických analýz, ktorý obsahuje podrobnejší popis plánovaných metód štatistických analýz oproti tomu, aký je uvedený v protokole. Hlavným zmyslom tohto dokumentu je zachovať integritu štúdie tak, že postup štatistických analýz je naplánovaný ešte pred tým ako sú dostupné samotné dáta. Ak by sa jednotlivé metódy vyberali na základe skúmaných dát, mohlo by dôjsť ku skresleniu výsledkov a ich interpretácii. Pretože rôzne analýzy vedia poskytnúť rôzne výsledky. Zadávateľ by si tak mohol vybrať metódu, ktorá poskytuje pre neho najprijateľnejšiu možnosť.

CRF má veľký vplyv na kvalitu zbieraných dát a môže tak nepriamo ovplyvniť aj výsledky celej klinickej štúdie. Nedostatočná pozornosť pri tvorbe CRF sa prejavuje vyššou časovou náročnosťou zberu a čistenia dát. V krajnom prípade môže dôjsť až ku nenaplneniu cieľov štúdie, napríklad pri zbieraní nedostačujúcich dát.

Keďže každá štúdia pre naplnenie svojich cieľov vyžaduje odlišné typy údajov, pri ich procese výberu a zakomponovania do CRF môžu byť veľmi nápomocné šablóny, ktoré vychádzajú zo skúseností predchádzajúcich, už zrealizovaných klinických štúdií. Pri formulovaní otázok do CRF je potrebné klásť dôraz na jednoduchosť, zrozumiteľnosť, jednoznačnosť a logickú nadväznosť jednotlivých otázok. Najčastejšie sa CRF zhromažďujú tieto údaje [2] [11]:

- Dátum, fáza a identifikácia štúdie
- Identifikácia subjektu
- Základné demografické informácie ako vek, pohlavie, výška a váha
- Charakteristika subjektu ako osobná anamnéza, zvyklosti či tehotenstvo
- Informácie o diagnóze primárneho ochorenia
- Zaraďovacie a vyraďovacie kritériá
- Dávkovanie, administratívne informácie o liečivom prípravku
- Dĺžka trvania liečby
- Súbežná medikácia či terapia
- Typ, dĺžka, trvanie, intenzita a následky zvoleného opatrenia
- Dôvod odslepenia, či predčasného ukončenia daného subjektu v štúdii

Existujú rôzne sady už vytvorených šablón s údajmi, ktoré sa vyskytujú v klinických štúdiách najčastejšie. Tieto šablóny sú dostupné v globálnej knihovni – global library. Ich použitie výrazne uľahčuje návrh obsahu CRF. Ďalšou výhodou je prítomnosť už naprogramovaných validovaných kontrol vybraných polí edit checks, či sady queries. Keďže globálna knihovňa je formátovaná štandardizovane, je možné ľahšie porovnávať a vyhodnocovať dáta naprieč niekoľkými rôznymi klinickými štúdiami, ktoré využívali rovnaké formuláre globálnej knihovne.

Pri využívaní globálnej knihovne sa však môžu zhromažďovať aj duplicitné údaje. Sú to také údaje, ktoré môžeme odvodiť, alebo vypočítať už z doteraz zozbieraných údajov v CRF. Ako príklad môže poslúžiť výpočet BMI z údajov o výške a váhe, ktoré sú zbierané vo formuláre osobnej analýzy. Nezaznamenávanie duplicitných hodnôt tak šetrí čas pri vyplňovaní formuláru, ale zvyšuje kvalitu dát a efektivitu ich spracovania. Okrem duplicitných údajov je snaha v CRF obmedziť aj otázky otvoreného typu. Ide o otázky, v ktorých nie je pevne nastavený formát odpovedí a ktoré sa následne veľmi ťažko štatisticky hodnotia. Najdôležitejšie parametre CRF sú identifikácia a označenie štúdie, verzia CRF, označenie fázy štúdie, identifikácia subjektu štúdie a samotné študijné dáta.

Výsledná kvalita dát je podmienená usporiadaním a podobou otázok, v akej sú dané otázky prezentované zadávateľovi dát vo formuláre CRF. Správne usporiadanie a rozvrhnutie otázok znižuje chybovosť a uľahčuje samotné vypĺňanie CRF. Otázky by mali byť krátke a výstižné, pri zachovaní ich jednoznačného významu. Terminológia použitá v otázkach by mala byť čo najviac prispôsobená osobám, ktoré budú daný formulár vyplňovať. Pre minimalizáciu chýb je nutné jednoznačne odlíšiť otázky kladené pacientovi a otázky kladené skúšajúcemu lekárovi. Pre zaistenie požadovanej kvality dát v tlačenej verzii CRF je nutné vyplňujúcu osobu naviesť a jednoznačne špecifikovať požadovaný formát údajov. To je pri tlačenej verzii CRF ošetrené najčastejšie priloženými inštrukciami pre správne vyplnenie odpovedí. Je nutné dodržovať vizuálnu jednotnosť v celom dokumente CRF spolu s rovnakými základnými typografickými pravidlami. Napríklad, ak je ako separátor použitá bodka, musí byť na oddelenie desatinnej časti vo všetkých ostatných otázkach použitý rovnaký formát. Pre zvýraznenie textu sa preferuje zmena veľkosti či štýlu písma pred farebným zvýraznením.

Typy odpovedí na otázky vo formulári CRF sa delia na:

- Otvorené – text, čísla a alfanumerické znaky
- Uzatvorené – výber z dvoch a viac možností
- Kombinácia otvorených a uzatvorených otázok

Pre otvorené odpovede musí byť v CRF vždy vyhradený dostatočný priestor a čo najviac upresnený požadovaný formát odpovedí, pre následné zjednodušenie spracovania. Pri uzatvorených odpovediach sa vyberá z predom známej konečnej množiny možností. Je však dôležité určiť správny spôsob označenia odpovede, to znamená označenie krížikom, krúžkom, či prešKrtnutím nehodiacej sa možnosti. Najprehľadnejšia a najpoužívanejšia je metóda označenia krížikom. Táto metóda sa používa aj preto, že pri následnom elektronickom spracovaní tlačeného CRF sa môže použiť program optical character recognition – OCR, ktorý dokáže z naskenovaného formulára výsledky rozpoznať a automaticky preniesť do elektronickej databáze. Kombinácia otvorených a uzatvorených otázok sa najčastejšie používa vo forme doplňujúceho textu ku zaškrtávacej odpovedi „iné“. Pre jednotlivé otázky sa pri rozsiahlych CRF odporúča viacúrovňové číslovanie pred jednoúrovňovým. Následnou kombináciou čísla strany formuláru a čísla otázky je možné jednoznačne identifikovať každú otázku v CRF.

Pri vyplňovaní otázky, však môže nastať možnosť, keď okienko zostane nevyplnené. Môže to spôsobiť problémy pri spracovaní dát. Hoci by takýchto odpovedí malo byť čo najmenej, je dôležité nastaviť CRF aj pre túto možnosť. Prázdne pole môže nastať pri prehliadnutí otázky, alebo vyplňujúci túto otázku nechal bez odpovede. Riešením v prípade uzatvorenej otázky sa najčastejšie odpoveď dopĺňa podľa variant:

- NA – not applicable
- ND – not done
- UNK – unknown

Ktorá konkrétna hodnota bude priradená k danému poľu sa odvíja od významu otázky a záleží aj od konkrétneho prípadu. Ďalšou vhodnou elimináciou prázdnych odpovedí sú logicky súvisiace otázky. Takéto otázky sú v CRF vizuálne zaradené do jedného bloku, poprípade oddelené ohraničením. Na začiatok bloku sa umiestňuje uzatvorená otázka, ktorá jednoznačne určí dokončenie bloku, alebo jeho mienené nevyplnenie.

Elektronický CRF poskytuje oproti tlačenej verzii rozsiahlejšie možnosti kontrol dodržiavania doporučeného formátu dát. To zabezpečuje automatická validácia, ktorá prebieha v reálnom čase. Základné typy formátov odpovedí eCRF sú:

- Textové pole – Text field: možnosť zadať číslo, či alfanumerické znaky, nastavuje sa validačná kontrola pre očakávaný vstup, ktorá jednoznačne špecifikuje počet desatinných miest, či počet zadaných znakov
- Poznámka – Note box: možnosť zadať text vo formáte dlhšieho reťazca, snaha čo najviac obmedziť
- Radio button: uzatvorená odpoveď, neumožňuje nechať odpoveď prázdnu
- Rolovací zoznam – Drop–down list: respondent vyberá jednu odpoveď z navrhovaných odpovedí, umožňuje nechať odpoveď prázdnu
- Check box: vyplnenie nula až *N* odpovedí
- Analogová škála – Analog scale: typ poľa s posuvníkom, používa sa napríklad pri subjektívnom hodnotení bolesti od 0 do 10
- Sekvenčné pole – Log: viacnásobný záznam rovnakého typu odpovedí, môžeme pridávať riadky a stĺpce, počet takto dynamický pridaných položiek je možné limitovať maximálne povolenou hodnotou

1.2.4 Tvorba databáze

Keďže databáza klinickej štúdie môže byť veľmi zložitá, je dôležité celý proces tvorby podrobne zdokumentovať. Najprv sa rozhodne, v akom systéme bude databáza vytvorená a následne sa určia potrebné atribúty. Najčastejšie sa používajú tieto [2][12]:

- Primárny kľúč: jednoznačne charakterizujúci základný údaj, popisuje subjekt v databáze, pri väzbe údajov na daný primárny kľúč, sú údaje jednoznačne zviazané s vybraným subjektom, v systémoch sa používa automaticky generovaný, alebo ako kombinácia čísla centra a poradového čísla subjektu, primárny kľúč je vždy unikátny
- Tabuľka: je charakterizovaná ako skupina logicky navzájom súvisiacich premenných, tabuľka obsahuje viacero premenných, ktoré určujú podstatu databáze, k parametrom sa v priebehu štúdie pridávajú hodnoty a tak vzniká databáza
- Premenné: organizácia premenných je volená tak, aby logicky odpovedala štruktúre dotazníku, pre každú premennú sa definuje jej typ
- Odvodené premenné: sú veľmi podobné štandardným premenným, netvoria ich dáta zadávané do CRF, ale vznikajú výpočtom vzorca, alebo iným odvodením, používajú sa tiež na prevod jednotiek, aby dáta boli v odpovedajúcom formáte
- Návštevy: odpovedajú najčastejšie fyzickým návštevám subjektu štúdie v priebehu klinickej štúdie u skúšajúceho doktora, pri ktorej sa prevedú nejaké laboratórne úkony
- Prístupové údaje: je nutné vytvoriť prístupové údaje pre všetkých užívateľov, ktorý budú do databáze vstupovať

1.2.5 Validačný plán

Dokument obsahuje zoznam a popis všetkých typov kontrol, ktoré budú na zozbieraných dátach vykonané v záujme zaistiť čo najlepšiu čistotu a kvalitu dát. Plán je delený podľa typu plánovaných kontrol:

- Kontroly formátu: overuje sa či je hodnota parametru v rámci definovanom pre dané prednastavené pole
- Kontroly súvislostí: tvoria najväčšiu skupinu kontrol a tvoria základ správnosti dát, sú založené na vzájomnom porovnaní viacerých premenných, napríklad porovnanie, či má systolický tlak väčšiu hodnotu ako diastolický, alebo či údaje o tehotenskom teste nie sú uvedené pri testovanom subjekte mužského pohlavia
- Self evidence correction: prípadná oprava dát je prevedená datamanažérom, napríklad, keď skupina skúšajúcich doktorov pochopí otázku v CRF zle, odpoveď je síce pravdivá, ale nezapadá do prednastavenej databáze, v tomto prípade sa vytvorí v databáze nová premenná, do ktorej sa nakopírujú pôvodné hodnoty premennej, zdrojové dáta zostanú v pôvodnej premennej a zároveň sú dáta prístupné pre analýzu dát, odpadá tak niekedy náročná úprava dát priamo so skúšajúcimi doktrormi

- Manuálne kontroly: používajú sa, ak je veľmi náročné naprogramovať automatickú kontrolu počítačom, manuálne sa hodnotí u každého jednotlivého prípadu, čo je časovo náročné, preto je treba počet týchto kontrol minimalizovať
- Špecifické kontroly: niektoré EDC systémy už obsahujú rutinné kontroly pre opakujúce sa, ale špecifické dáta, napríklad laboratórny modul, ktorý vyhodnocuje, či sú namerané hodnoty vo fyziologickom rozmedzí, ak je hodnota abnormálna, vyžaduje sa od lekára potvrdenie, či je zistená odchýlka významná, alebo nie

1.2.6 Kódovanie

Kódovaním sa myslí proces štandardizácie odborných lekárskych termínov. Je nutné uviesť zoznam premenných, ktoré sa budú kódovať a tiež typ slovníka, ktorý bude na kódovanie použitý. Najčastejšie používanými pre kódovanie liečivých prípravkov sú WHO Drug dictionary a pre kódovanie nežiadúcich účinkov MedDRA – Medical Dictionary for Regulatory Activities.

1.2.7 Dokončenie plnenia databáze

Uzamknutie databáze je proces, kedy sa k určitému dátumu znemožní akákoľvek úprava dát v danej databáze. Predstavuje posledný krok spracovania dát z pozície data managementu pred ich poslaním na štatistickú analýzu [2]. Pred uzamknutím databáze býva overené, či databáza splňuje podmienky na uzatvorenie a to:

- Databáza obsahuje dáta všetkých subjektov štúdie
- V databáze sú všetky dáta vrátane externých, napríklad laboratórne dáta
- Všetky dáta v tlačenej podobe sú digitalizované
- Všetky plánované validácie boli prevedené
- Všetky preklady relevantných premenných boli prevedené
- Všetky kódovania relevantných premenných boli prevedené a validované
- Bola vykonaná kontrola dát za účasti klinických odborníkov
- Prípadne zistené nezrovnalosti vyriešené
- Všetky queries sú zodpovedané
- Prebehla kontrola kvality dát na požadovanej úrovni prístupnej chybovosti
- Dokumentácia všetkých procesov data managementu je dokončená
- Plán štatistickej analýzy je vypracovaný a schválený
- Odchýlky od protokolu sú zdokumentované a odsúhlasené
- Samotný zadávateľ súhlasí s uzamknutím databáze

Podľa platných legislatívnych noriem je potrebné oficiálne dokumenty z klinických štúdií archivovať po dobu niekoľkých rokov. To platí ako pre tlačené verzie dokumentov finálnych verzií s podpismi, tak aj pre finálne elektronické dokumenty. Samotná archivácia nastane až po odovzdaní záverečnej klinickej správy, ktorá formálne ukončuje danú klinickú štúdiu.

1.3 EDC systém

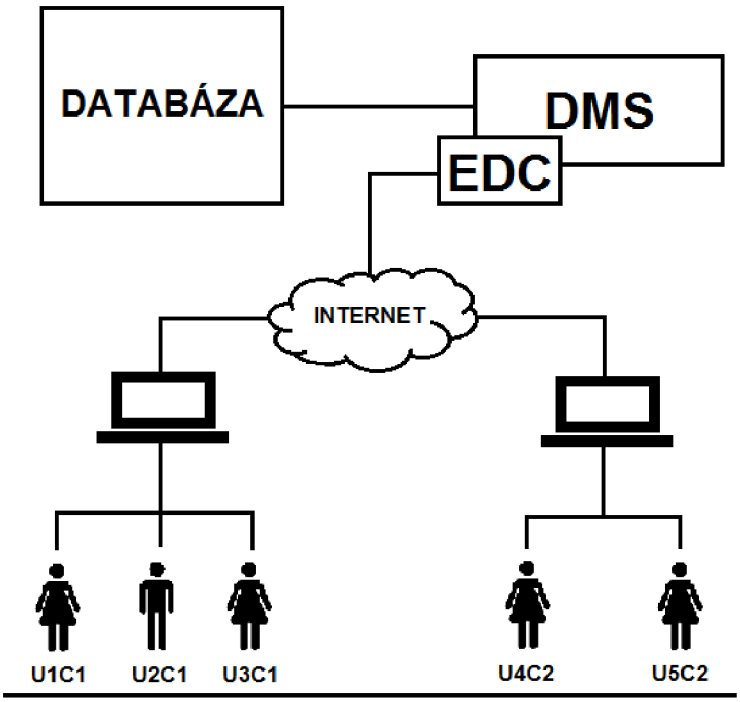
Použité termíny Data Management System – DMS, alebo Clincal Data Management System – CDMS sú označenia pre komplexný softwarový nástroj, ktorý poskytuje pokrytie všetkých nárokov na funkcionalitu a management pri správe dát klinickej štúdie. DMS vytvára samotnú databázu a elektronickú verziu CRF, zadávanie a monitorovanie dát pomocou jednotlivých validačných nástrojov, obsahuje modul pre management queries a možnosť exportu dát [2].

Electronic Data Capture sa označujú systémy DMS/CDMS, ktoré slúžia len na zadávanie dát online, alebo off–line formou prostredníctvom počítača, mobilného zariadenia, či tabletu. Ďalšie označenia pre tieto systémy sú Remote Data Entry – RDE, alebo Remote Data Capture – RDC a považujú sa za ekvivalentné. Popisujú technológie vzdialeného zadávania dát do databáze, najčastejšie cez webový prehliadač z miesta, ktoré je odlišné od miesta uloženia databázy. Systém prístupu užívateľov je zobrazený na Obrázok 2. Existuje možnosť použiť aj ďalší systém Electronic Diary – ED. Slúži na subjektívne monitorovanie zdravotného stavu testovaného subjektu, alebo sa používa pri zadávaní dát na dennej či týždennej báze bez prítomnosti pacienta, pomocou elektronického zariadenia, ktorým je subjekt vybavený.

Vhodný výber databázového systému sa posudzuje na základe všetkých aspektov štúdie, na základe prvého návrhu CRF a keď sú známe ciele a usporiadanie štúdie. Správne zvolený DMS pomáha k zvyšovaniu kvality dát pri akceptovateľnej časovej a finančnej náročnosti daného systému. Je tiež dôležité brať ohľad na technické požiadavky vybraného DMS ako napríklad architektúru počítačových systémov, konfiguráciu siete, systém zálohovania dát a ich archiváciu či vyžadované hardwarové požiadavky na mieste vykonávania štúdie. Pri voľbe použitia vhodného systému tak treba brať ohľad hlavne na:

- Rozsah a zložitosť CRF a samotnej štúdie
- Počet zapojených centier
- Veľkosť dátového úložiska
- Splnenie požiadavkou zabezpečenia a legislatívy
- Nároky na spracovanie externých dát

Vhodný DMS by tak mal predovšetkým byť užívateľsky jednoducho používaný, podporovať kódovanie dát, zahrňovať možnosť importu externých dát, byť zabezpečený cez systém užívateľských oprávnení, mať možnosť uzamknutia databáze na niekoľkých úrovniach, napríklad na úrovni štúdie, subjektu, či určitého poľa, mal by obsahovať možnosť automatického generovania databázovej štruktúry, mať možnosť jednoduchého – single data entry, ako aj dvojitého – double data entry zadávania dát, jednoduchý spôsob opravy dát, audit trails a mal by obsahovať systém pre reporting a export dát vo formáte Statistical Analysis System – SAS, či inom vyhovujúcom štatistickom formáte ako CSV, XLS či R.



Obrázok 2 – popis princípu prístupu pre užívateľov U1,U2 a U3 z centra C1 a užívateľov U4 a U5 z centra C2, prevzaté a upravené z [2]

Audit trails patrí k jednej z najdôležitejších funkcionalít systému. Slúži na záznam všetkých zmien prevedených v databáze v priebehu klinickej štúdie a tiež obsahuje časovú stopu, kedy bola daná zmena urobená. Systém teda zaznamenáva kto, čo a kedy v databáze zmenil. V každom momente je tak možné dohľadať prevedenú zmenu spolu s časovým údajom. Je zásadný pre možnosť spätnej rekonštrukcie celej štúdie, čo je jedným z požiadavkou na transparentnosť štúdie. Ďalšou dôležitou vlastnosťou každého DMS je určenie jednoznačných prístupových práv a definovanie užívateľských rolí pre zadávateľa, monitora, dátového manažéra a tiež osoby, ktoré sú inak zapojené do klinickej štúdie. Platí teda, že osoby majú prístup len k tým dátam, ktoré nevyhnutne potrebujú k svojej práci. Za každým databázovým systémom je spravidla relačná databáza, ktorá predstavuje súbor tabuliek s príslušnými atribútmi a záznamami. Jednotlivé tabuľky sú prepojené pomocou primárnych a cudzích kľúčov, sú konzistentné voči definovaným pravidlám a dodržujú integritu obmedzení.

1.4 Validácia a zadávanie dát v praxi

Zaistenie validity a konzistencie zozbieraných dát je jednou z hlavných úloh data managementu. Kontrolujú chyby, ktoré sa môžu objaviť v priebehu klinickej štúdie. Validácia prebieha podľa pripraveného validačného plánu, v ktorom je definované ako budú kontroly prevedené a ako budú procesy kontroly otestované a zdokumentované. Definícia, programovanie a implementácia kontrol sa nazýva aj edit checks a začína po schválení finálnej verzie CRF a tvorbe databáze. Najčastejšie chyby a edit checks používané na ich odhalenie je možné rozdeliť na:

- Chýbajúce hodnoty: sú všetky povinné hodnoty, ktoré nemôžu zostať nevyplnené
- Kontrola predpísaného formátu: vybrané polia s textovým, alebo numerickým formátom odpovede majú presne definovanú formu
- Kontrola rozsahu a numerických hodnôt: numerické polia sa môžu často obmedziť intervalom minimálnej a maximálnej hodnoty
- Logická inkonzistencia: kontrola vzájomnej kombinácie premenných, ktorá je reálne veľmi nepravdepodobná
- Odchýlka od protokolu: porovnáva dáta uvedené v CRF s požiadavkami z protokolu danej štúdie

Pred spustením databáze pre zadávanie reálnych dát je nutné systém otestovať. Chýbajúce, alebo zle implementované edit checks môžu mať negatívny dopad na celkovú kvalitu dát. Na odhalenie chýb sa používajú zdokumentované a detailne nastavené postupy testovania. Na základe výsledkov sú objekty a jednotlivé celky opravené a znovu otestované [2][13].

Cyklus zadávania dát do systému EDC sa v elektronickej podobe začína založením a distribúciou účtov pre osoby, ktoré budú zodpovedné za zadávanie dát v mieste, kde bude daná štúdia prebiehať. Užívatelia dostanú svoje prihlasovacie údaje emailom, ktorý obsahuje unikátne prihlasovacie meno, heslo a webovú adresu pre prihlásenie do rozhrania EDC systému. Pred samotným začatím štúdie je odporúčané overenie prístupu do EDC systému. Problémom môžu byť bezpečnostné brány vnútornej siete zdravotníckeho zariadenia ako je firewall a podobne. Odporúča sa tiež zaškolenie užívateľov na prácu so systémom. Tým odpadajú následné časté otázky na užívateľskú podporu. Školenie môže byť vedené lektorom, alebo môže byť uskutočnené cez e–learning. V e–learningovej forme si môže užívateľ vyskúšať aj novo získané vedomosti vo forme skúšobných úloh. Čas investovaný do dôkladného zaškolenia práce so systémom sa neskôr mnohonásobne vráti pri čistení dát a vo výslednej kvalite dát.

Nezrovnalosť zistená po zadaní dát vedie ku generácii dotazu – query do miesta prevádzania štúdie týkajúceho sa správnosti dát. Úlohou je získať objasňujúce informácie o podozrivej hodnote v CRF. Dokumentuje sa proces vyriešenia queries, čas zadania a získania odpovede. Proces riešenia je úprava zistených nezrovnalostí a verifikácia zadaných dát. Môže sa stať, že zadávaná hodnota presahuje jej povolený rozsah, avšak od zadávateľa je potvrdená ako platná.

Stav query môže byť:

- Otvorený: pri kontrole bola zistená nezrovnalosť a query je otvorené až do kým sa daná chyba neopraví
- Zatvorený: po úprave dát na prípustnú hodnotu
- Re–query: v prípade neuspokojivej korekcie sa opakuje celý proces
- Overené: po úprave dát je nutná kontrola datamanagerom

Pred samotným uzatvorením štúdie musia byť všetky vytvorené queries uzatvorené, zdokumentované a verifikované. Z jednotlivých počtov odoslaných queries do centier sa dajú vyvodiť výsledky, ako dané centrá pristupujú k vyplňovaniu formulárov.

Záverečná kontrola dát môže prebiehať dvomi spôsobmi:

- Kontrola dát vybraných parametrov u všetkých subjektov: ide o kontrolu tzv. kritických premenných, ktoré sú nenahraditeľné pre splnenie danej štúdie
- Kontrola všetkých premenných u náhodného vzorku subjektov: počet chýb je porovnaný s celkovým počtom polí, tento pomer vyjadruje percento chybných hodnôt

2 KVALITA DÁT

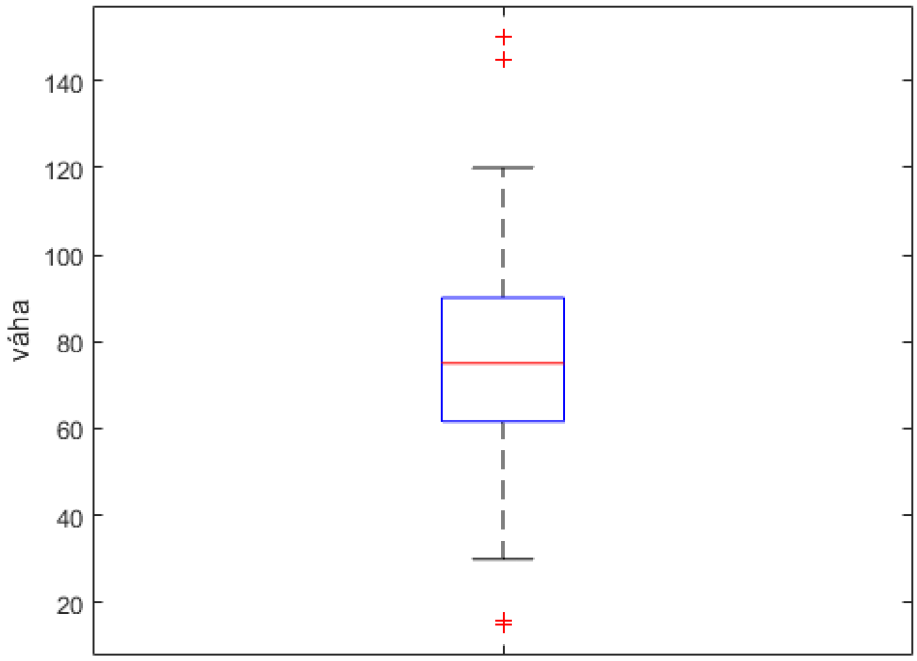
Pre kvalitné dáta platí, že neobsahujú odľahlé hodnoty – outliery. Tieto hodnoty môžeme definovať ako malé množstvo dát, ktoré obsahujú extrémnejšie hodnoty ako ostatné namerané dáta v danom vzorku. Niekedy však môžu odľahlé hodnoty vystupovať ako valídne dáta, čo by znamenalo, že testovaný súbor nemá normálne rozdelenie [14].

Kvalita dát sa vyhodnocuje pomocou algoritmov využívajúce:

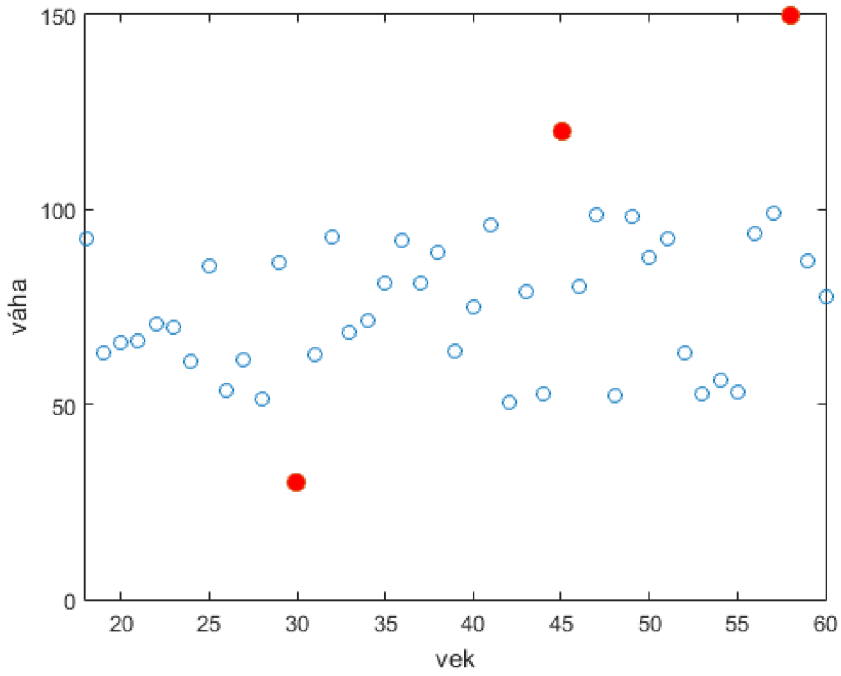
- Detekciu odľahlých hodnôt pomocou vizualizácie
- Detekciu odľahlých hodnôt pomocou štatistických metód
- Detekciu odľahlých hodnôt metódami strojového učenia

2.1 Detekcia odľahlých hodnôt pomocou vizualizácie

Poskytuje najjednoduchší spôsob na detekciu odľahlých hodnôt. Na zobrazenie sa používajú krabicový graf – box plot ako je zobrazený na Obrázok 3, alebo bodový graf – scatterplot zobrazený na Obrázok 4 nižšie.



Obrázok 3 – krabicový graf s odľahlými hodnotami označenými znakmi ako červené +



Obrázok 4 – bodový graf s odľahlými hodnotami označenými červenou farbou

2.2 Detekcia odľahlých hodnôt pomocou štatistických metód

Detekcia pomocou štatistických metód dokáže jasne určiť, či testovaná hodnota je, či nie je odľahlá hodnota. Veľká nevýhoda však je, že tieto metódy predpokladajú normálne rozdelenie dát, alebo aspoň rozdelenie podobné normálnemu, napríklad Cauchyovo rozdelenie [14].

V prípadoch, kedy táto podmienka nie je splnená, sa používa lineárna, alebo nelineárna transformácia. Novo získané dáta z nelineárnej transformácie získame logaritmickou transformáciou ako :

$$y'_i = \log y_i \text{ ,(1)} \qquad y'_i = \ln y_i \text{,(2)} \qquad y'_i = \log(b_0 + b_1 y_i) \text{ ,(3)}$$

kde y_i je pôvodná hodnota a y'_i je výsledná hodnota získaná pomocou parametrov b_1 a b_2, ktoré sú zistené pomocou regresie. Keďže logaritmus záporných čísiel nie je definovaný, tak sa pri záporných hodnotách používa vhodná voľba koeficientov b_0 a b_1. Tiež treba dbať na to, že $\log 1 = 0$ a logaritmus medzi 0 a 1 sú záporné čísla. Namiesto logaritmickej transformácie sa pre tieto obmedzenia môže použiť transformácia odmocninou, ktorú podľa [15] dostaneme ako

$$y'_i = \sqrt{y_i}\,, \tag{4}$$

kde y'_i je výsledná hodnota a y_i je pôvodná hodnota.

Na detekciu odľahlých hodnôt sa využívajú nasledujúce štatistické metódy:

- Kategoriálny test: pre kategoriálne premenné, za odľahlé sú považované hodnoty vyskytované s menšou frekvenciou ako stanovená minimálna frekvencia, často sa používa percento výskytu 0,05
- Normálne obojstranné a jednostranné testy: za odľahlé sú označené hodnoty, ktoré sú vzdialené od priemeru o x–násobok smerodajnej odchýlky, často sa používa hodnota x=3
- Grubbov obojstranný a jednostranný test: pre každú hodnotu je spočítaná Grubbova štatistika a ako odľahlá hodnota je označená hodnota, ktorej Grubbova štatistika je väčšia ako jej kritická hodnota
- Percentilový obojstranný a jednostranný test: za odľahlú hodnotu je považovaná hodnota, ak spadá do oblasti nad horný zadaný percentil, alebo pod dolný zadaný percentil
- Tukey obojstranný a jednostranný test: používa na výpočet vzorce (5) až (8), kde hodnotu outlierového koeficientu volí sám užívateľ, typicky od 1 do 5

Patria sem aj kritériá:

- Chauvenetovo kritérium: za odľahlú je považovaná hodnota, ktorá spĺňa kritérium, že pravdepodobnosť obdržania vypočítanej odchýlky od priemeru je menšia ako 1/(2n)
- Peircovo kritérium: podobné Chauvenetovému kritériu, ale pre výpočet sa používa tabuľka s R hodnotami, bližšie popísaná v [16].

Odľahlé hodnoty sa delia na outliery – mild outliers a extrémy – extreme outliers [17]. Pri nenormálnom rozdelení pre mild outliers platí, že hodnoty sú menšie ako

$$Q_1 - o.c \cdot IQR \text{ ,(5)} \qquad \text{alebo sú väčšie ako} \qquad Q_3 + o.c \cdot IQR \text{ ,(6)}$$

kde Q_1 je prvý kvartil, Q_3 je tretí kvartil, $o.c$ je outlierový koeficient (často sa využíva hodnota 1,5) a IQR je interkvartilové rozpätie počítané ako rozdiel tretieho a prvého kvartilu. Pre extreme outliers platí, že hodnoty sú menšie ako

$$Q_1 - 2 \cdot o.c \cdot IQR\text{,(7)} \qquad \text{alebo sú väčšie ako} \qquad Q_3 + 2 \cdot o.c \cdot IQR\text{.(8)}$$

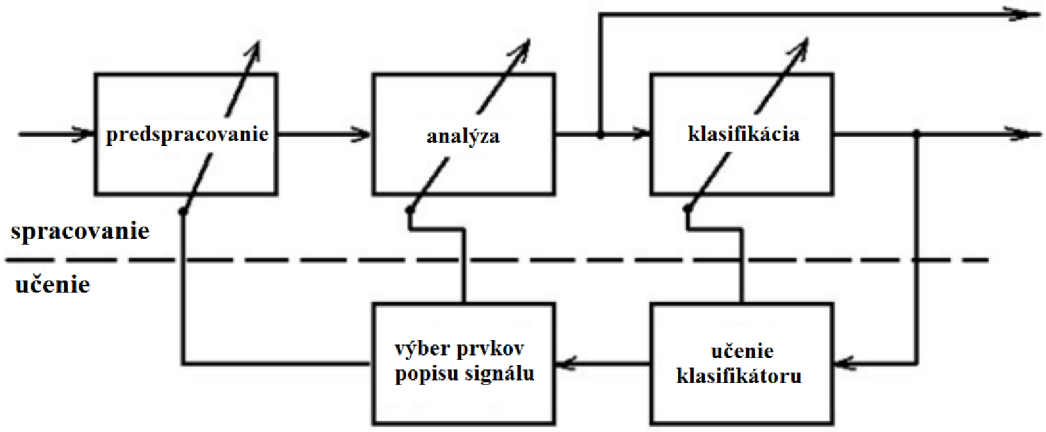
Odľahlé hodnoty spôsobujú, že pri testovaní hypotéz sa nemôžu použiť parametrické testy z dôvodu nenormality dát a nehomogenity rozptylu. Pri použití neparametrických dvoj výberových, alebo mnohonásobne výberových testov ako Kruskal–Wallis ANOVA test, odľahlé hodnoty majú na výsledok testu len malý, alebo nepozorovateľný vplyv.

2.3 Detekcia odľahlých hodnôt metódami strojového učenia

Strojové učenie je podoblasť umelej inteligencie, ktorá má schopnosť učiť sa z okolitého prostredia a na tieto atribúty adekvátne reagovať bez toho, aby bola na tieto úlohy explicitne naprogramovaná. Pomáha sledovať, rozhodovať a predvídať udalosti [18].

Metódy strojového učenia sa delia do skupiny bez učiteľa, kde pre vstupné dáta nie je zadaný presný výstup a do skupiny s učiteľom, kde naopak poznáme výstup. Tento proces umožňuje, že jednotlivé klasifikátory pozorujú najprv testovacie dáta. Z týchto dát vytvárajú predikčné funkcie s vhodne vybranými príznakmi, na základe ktorých sa klasifikujú dané testovacie dáta [19]. Presnosť klasifikácie sa určí podľa vzorcov (22) – (25) spomenutých v kapitole Hodnotenie úspešnosti klasifikácie.

Extrakcia príznakov je obvykle prvým krokom, ktorý je nutné urobiť pre popis každého pozorovania. V tomto prípade subjektu v skúmanej klinickej štúdií tzv. príznakovým vektorom. Jeho dĺžka sa redukuje v kroku, ktorý sa označuje ako výber príznakov. Pri viacrozmerných dátach býva redukcia dimenzionality podstatnou časťou analýzy dát.



Obrázok 5 – obecné schéma strojového učenia, upravené a prevzaté z [37]

2.3.1 Analýza hlavných komponent

Jedna zo základných metód, ktorá na zjednodušenie analytických výpočtov využíva extrakciu premenných, tiež nazývaná ako Principal Component Analysis – PCA. Okrem extrakcie sa používa na vizualizáciu viacrozmerných dát a tiež k odhaleniu skrytých (latentných) premenných, ktoré často pomáhajú pri interpretácii dát. Výpočet je podľa [20] definovaný ako:

$$(\mathbf{A} - \lambda_k \mathbf{I})\boldsymbol{u_k} = 0\,, \tag{9}$$

kde $(\mathbf{A} - \lambda_k \mathbf{I})$ je charakteristická rovnica, používa sa na výpočet vlastných hodnôt λ_k.

Tie získame z rovnice :

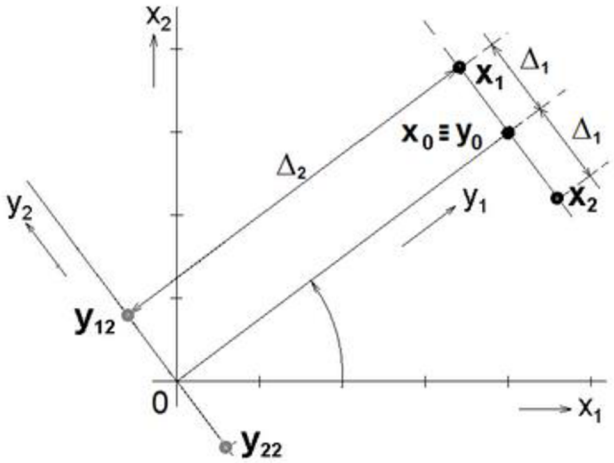
$$|\mathbf{A} - \lambda_k \mathbf{I}| = 0\,, \tag{10}$$

kde $|\mathbf{A} - \lambda_k \mathbf{I}|$ je determinant charakteristickej rovnice. Vlastné vektory $\boldsymbol{u_k}$ súvisia s vlastnými hodnotami λ_k. Vlastné hodnoty predstavujú rozptyl odpovedajúci hlavným komponentám. Každá takáto hodnota odpovedá jednej komponente, takže vo výsledku dostaneme toľko vlastných hodnôt, koľko je definovaných premenných. Najväčšia vlastná hodnota a k nej príslušný vlastný vektor odpovedá prvej komponente. Táto komponenta vyjadruje najväčší podiel variability v dátach.

Podľa [21] je PCA vysvetlená na nasledujúcom príklade. Reálne objekty sú popísané vektormi v dvojrozmernom priestore súradnicami x_1 a x_2. Pri vyjadrení zadaných vektorov v inej súradnicovej sústave, ktorých súradnice sú y_1 a y_2 platí, že súradnice y_1 a y_2 sú dané lineárnou kombináciou pôvodných súradníc x_1 a x_2. Pre dvojrozmerný priestor teda platí:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 \\ y_2 &= a_{21}x_1 + a_{22}x_2 \end{aligned} \quad = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{11}$$

Lineárna transformácia tak spôsobí, že nová súradnicová sústava (y_1,y_2) je oproti pôvodnej (x_1,x_2) len otočená okolo svojho počiatku. Veľkosť tohto otočenia závisí na hodnotách parametrov a_{11}, a_{21} a a_{22}. Pre ortogonálnosť novej súradnicovej sústavy je nutné, aby bol skalárny súčin transformačných vektorov $a_1 = (a_{11}, a_{12})$ a $a_2 = (a_{21}, a_{22})$ nulový. Ak nie je žiadúce, aby došlo k predĺženiu či skráteniu mierky na osách, tak obidva transformačné vektory by mali byť normované. Veľkosť ich modulu by mala teda byť jednotková. Ukážka princípu PCA je na Obrázok 6 nižšie.



Obrázok 6 – princíp PCA, prevzaté z [19]

Na analýzu hlavných komponent sa teda používajú všetky kritické parametre pre zistenie ich variability a hlavne korelácie parametrov, aby sa zistilo, ktoré parametre je možné vyradiť a ktoré budú použité na ďalšiu analýzu [1].

2.3.2 Faktorová analýza

Factor Analysis – FA pracuje s rozborom štruktúry vzájomných závislostí premenných za predpokladu, že sú závislosti dôsledkom pôsobenia nezmerateľných faktorov pôsobiacich v pozadí, tiež nazývané ako spoločné faktory. Do určitej miery sa faktorová analýza dá považovať za rozšírenie metódy analýzy hlavných komponent PCA. Vychádza však zo snahy vysvetliť závislosti jednotlivých premenných. Nevýhoda je nutnosť určiť počet spoločných faktorov ešte pred samotnou analýzou. Jej prednosti sú však úspornosť a obecnosť. Spoločné faktory vyvolávajú koreláciu medzi premennými a chybové faktory prispievajú k rozptylu pozorovaných premenných. Faktorová analýza predpokladá, že korelácia medzi premennými je výsledkom pôsobenia spoločných faktorov a nie vzájomného vzťahu medzi premennými. Faktorový model je možné popísať v maticovej podobe ako :

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{L}^T + \boldsymbol{E}\,, \tag{12}$$

kde $\boldsymbol{X}$ je dátová matica o rozmere $n \times p$, $\boldsymbol{F}$ je matica o rozmere $n \times m$, ktorej stĺpce sú spoločné faktory $F_1, F_2, \ldots m$, $\boldsymbol{L}$ je matica faktorových záťaží o rozmere $p \times m$ a $\boldsymbol{E}$ je matica chýb o rozmere $n \times p$, ktorej stĺpce sú špecifické faktory $e_1, e_2, \ldots e_p$ [20].

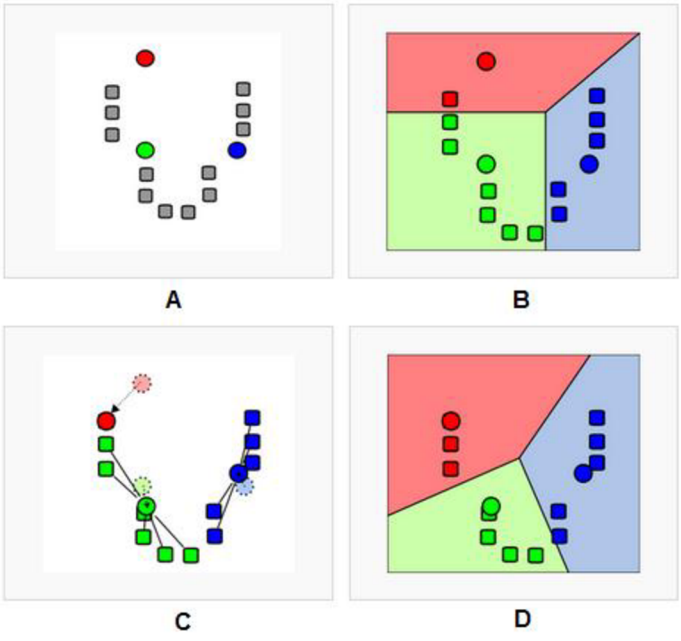
Nutná podmienka pre použitie faktorovej analýzy ako aj pre analýzu hlavných komponent je korelácia medzi pôvodnými premennými. V opačnom prípade by faktorová analýza nemala čo objasniť a výsledok analýzy hlavných komponent by boli komponenty rovnaké s pôvodnými premennými.

2.3.3 K-means

Jedna z klasifikačných metód bez učiteľa tiež nazývaná ako metóda k–priemerov hľadá také skupiny vo viacrozmernom priestore, kedy je skupinová podobnosť čo najväčšia. Zhluk sa tvorí minimalizáciou celkovej sumy štvorca vzdialeností vo vnútri skupín. Výsledkom je teda vytvorenie *K* skupín, ktoré sú od seba čo najviac oddelené [21].

Proces algoritmu začína zvolením počiatočného rozkladu do *K* zhlukov. Toto rozdelenie je často náhodné. Určia sa centroidy pre všetky zhluky v danom rozklade. Postupne sa hodnotia pozície všetkých objektov, kedy ak má objekt najmenšiu vzdialenosť k vlastnému centroidu, tak ostane na danom mieste. Ak nemá, tak sa presunie do zhluku, ktorému je najviac podobný. Následne sa opäť prepočítajú centroidy každého *K* zhluku. To sa opakuje až do momentu, keď by už prípadný ďalší posun nezlepšil zvolené metriky podobnosti. Takto sa v *K* skupinách presúvajú objekty podľa toho, aby sa minimalizovala variabilita vo vnútri skupín a maximalizovala variabilita medzi jednotlivými skupinami. Ide o iteratívny proces.

Obrázok 7 A zobrazuje 15 objektov. Tri centroidy sa umiestnili náhodne, každý inou farbou. Objekty sa následne priradia k centroidom s najmenšou vzdialenosťou. Takto vznikli tri zhluky, čo je zobrazené na obrázku B. V ďalšom kroku zobrazenom na C sa prepočítajú centroidy zhlukov tak, aby centroid daného zhluku vyjadroval ťažisko objektov, ktoré patria pod daný centroid. Kroky z obrázku B a C sa opakujú do kým nenastane ustálenie sústavy zobrazené na obrázku D.



Obrázok 7 – jednotlivé kroky k–means metódy, prevzaté z [22]

2.3.4 K-medoids

Pre metódu k–medoidov platí, že oproti k–means už zástupcom stredu nie je centroid, ale reprezentatívny objekt nazývaný medoid. Metóda postupne hľadá *K* reprezentatívne objekty. Tie sú definované ako objekty zhluku, ktorých priemerná nepodobnosť ku všetkým objektom v zhluku je minimálna. Zhluk je definovaný ako súbor takýchto objektov, ktoré sú priradené ku rovnakému medoidu. Jedná sa o robustnejšiu metódu ako K-means.

Postupne sa vyselektujú *K* reprezentatívne objekty. Prvý je objekt s čo najmenšou sumou nepodobnosti voči ostatným objektom. Tento objekt sa následne umiestni, čo najviac centrálne. Postupne sa po iteráciách vyberajú ďalšie objekty, ktoré čo najviac znižujú sumu nepodobnosti k najpodobnejšiemu vybranému objektu. Metóda sa zastaví až po nájdení *K* reprezentatívnych objektov. Miera nepodobnosti (vzdialenosti) je vyjadrená pomocou metriky podobnosti. Príklady možného použitia sú spomenuté v kapitole Metriky podobnosti a vzdialenosti.

V druhej fáze algoritmu sa zhlukovanie zlepší. Zrovnajú sa všetky páry objektov a postupne sa pre každý medoid a objekt zisťuje hodnota metriky podobnosti pre danú konfiguráciu. Ak sa kritérium zlepší, tak sa daný objekt stáva novým medoidom namiesto starého. Takto to pokračuje až do kým už nedochádza k žiadnemu zlepšeniu metriky podobnosti.

2.3.5 Dbscan

Vyššie spomenuté metódy rozdeľujú objekty na základe ich vzájomnej vzdialenosti. Najčastejšie tak v trojrozmernom priestore vznikajú zhluky tvaru gule. Na objavenie zhlukov rôznych tvarov sa používajú metódy založené na hustote. Tieto metódy vytvárajú zhluky tak, aby pre každý objekt v danom zhluku platilo, že v jeho okolí je minimálny počet ďalších objektov.

Jedna z metód založených na hustote je metóda Dbscan – density based spatial clustering. Je založená na zväčšovaní jednotlivých zhlukov, kým je zachovaná požadovaná úroveň hustoty pre všetky získané zhluky. Algoritmus objavuje zhluky kontrolovaním okolia všetkých objektov v databáze. Po nájdení základného objektu vytvorí pre neho nový zhluk. Po prehľadaní všetkých objektov, v krokoch spája základné objekty so všetkými objektami, ktoré sú z nich hustotou priamo dosiahnuteľné. Do jednotlivých zhlukov sú tak zaradené len tie objekty, ktoré sú hustotou dosiahnuteľné z ktoréhokoľvek základného objektu v danom zhluku. Algoritmus sa ukončí za podmienky, že už neexistuje objekt, ktorý by sa mohol pridať do akéhokoľvek zhluku [22].

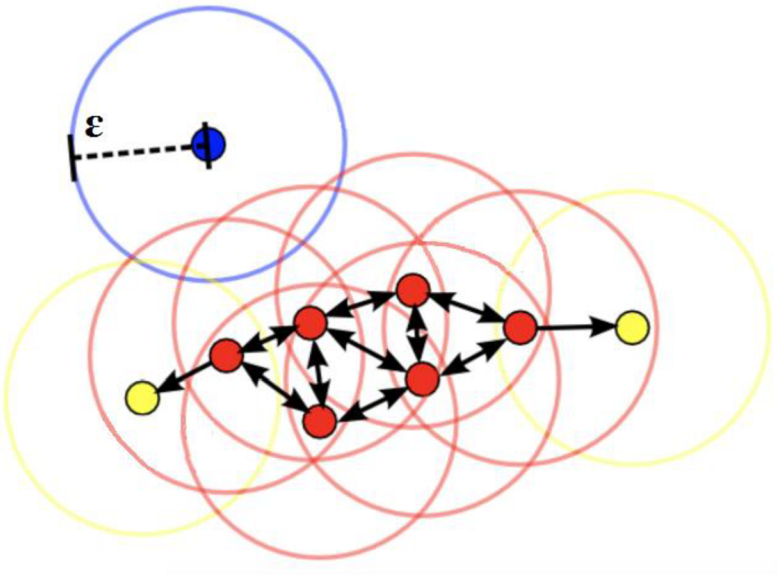
Ak predpokladáme, že ε sú pevne dané čísla potom platia nasledujúce definície:

- ε okolie objektu : vzdialenosť od stredu objektu, vyjadrená kružnicou s polomerom ε
- objekt v jadre : objekt, v ktorom okolí je aspoň m objektov vrátane seba
- priamo dosiahnuteľný objekt r : ak q je objekt jadra a r leží v jeho ε okolí

- hustotne dosiahnuteľný objekt r : ak od objektu q existuje reťaz objektov $q, q_1, \ldots q_x$, v ktorej sú všetky $q, q_1, \ldots q_x$ objekty jadra a každý objekt z reťaze leží v ε okolí svojich susedou z reťaze, objekt q_1 je tak priamo hustotne dosiahnuteľný z q, objekt q_2 z objektu q_1, ...
- hustotne spojené objekty r a t : ak existuje objekt jadra q tak, že r aj t sú hustotne dosiahnuteľné z objektu q
- odľahlé objekty : za odľahlé sú považované objekty, ktoré nie sú dosiahnuteľné z iného objektu

Algoritmus ako prvé určí, objekty jadra. Pre každý takto definovaný objekt následne uloží odkaz na jeho susedou v okolí ε. Pre zistenie, ktoré objekty patria do jadra je odporúčané mať objekty uložené v štruktúre s priestorovým indexom. Táto metóda uľahčuje hľadanie susedných objektov. Každý objekt jadra je základom svojho zhluku. Odkazy na susedov vytvoria orientovaný graf. Prehľadávaním tohto grafu do hĺbky, alebo šírky sa vytvoria jednotlivé zhluky.

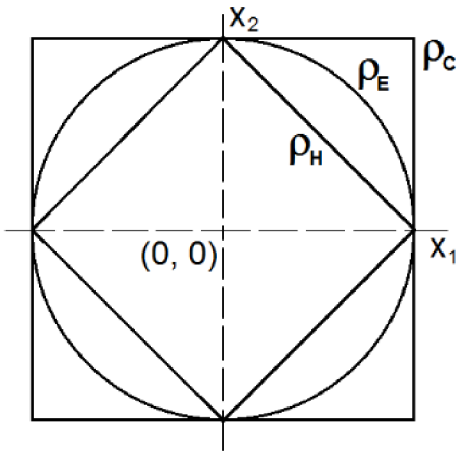
Názorná ukážka je ukázaná na Obrázok 8 nižšie. Objekty v jadre sú označené ako červené plne vyplnené kruhy. Priamo dosiahnuteľné objekty sú označené ako žlté plne vyplnené kruhy. Odľahlý objekt je zobrazený modrým plne vyplneným kruhom. V okolí každého objektu je zobrazená kružnica v odpovedajúcej farbe, ktorá vyjadruje ε okolie objektu. Reťaz dosiahnuteľných objektov je zobrazená pomocou čiernych šípok [23].



Obrázok 8 – Dbscan, prevzaté a upravené z [25]

2.4 Metriky podobnosti a vzdialenosti

Použitie konkrétnej metriky podobnosti, alebo vzdialenosti závisí na riešenej úlohe. Pri použití metriky na klasifikáciu je rozhodujúcim kritériom kvalita výsledkov klasifikácie. Ďalšie z možných kritérií sú výpočetná náročnosť či charakter rozloženia dát. Obecne tak neexistuje presný postup pre výber optimálnej metriky podobnosti [25].



Obrázok 9 – Geometrické miesta bodov s rovnakou vzdialenosťou od súradnicového počiatku v dvojrozmernom priestore. ρ_E- Euklidova metrika, ρ_C- Čebyševova metrika, ρ_H- City block, prevzaté z [25]

Pomocou metrík podobností sa definuje vzdialenosť (podobnosť) medzi dvoma pozorovaniami. Tieto pozorovania môžu byť popisom jednotlivých subjektov (príznakových vektorov), alebo centroidov či medoidov v zhlukovacích metódach. Pre maticu *X(m,n)*, ktorá je vyjadrená ako *m* riadkových vektorov $x_1, x_2, \ldots, x_m$ sú vzdialenosti medzi vektormi x_s a x_t definované nasledovne [26] :

2.4.1 Euklidová metrika

Metrika s najnázornejšou geometrickou interpretáciou. Geometrickým miestom bodov s rovnakou Euklidovou vzdialenosťou od daného bodu je v trojrozmernom priestore guľa a v dvojrozmernom kruh. Metrika vďaka kvadrátu rozdielu kladie väčší dôraz na väčšie rozdiely medzi súradnicami ako v lineárnom prípade.

$$d_{st} = \sqrt{(x_s - x_t)(x_s - x_t)'} \tag{13}$$

2.4.2 Metrika kosínovej podobnosti

Metóda je založená na výpočte skalárneho súčinu a predpokladá, že vektory sú normované. Hodnoty podobnosti sú rovné kosínusu uhlu medzi vektormi.

$$d_{st} = 1 - \frac{x_s x_t'}{\sqrt{(x_s x_s')(x_t x_t')}} \tag{14}$$

2.4.3 Minkowského metrika

Pre túto metriku platí, že je viac obecná ako Euklidova a City block metrika. Namiesto druhej odmocniny je použitá obecná odmocnina. Táto úprava zvyšuje váhu vplyvu členov s väčším rozdielom dielčích súradníc obidvoch objektov. Čím je mocnina väčšia, tým je dôraz na väčšie rozdiely medzi súradnicami väčší.

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} \left| x_{sj} - x_{tj} \right|^{p}} \tag{15}$$

2.4.4 City block metrika

Svoje pomenovanie metóda získala, pretože výpočet v dvojrozmernom priestore pripomína vzdialenosť, ktorú by prešiel automobil z jedného miesto do druhého pri použití pravouhlo zastavanom mestskom prostredí.

Je vytvorená linearizáciou Euklidovej metriky, čo prináša zníženie významu členov s väčším rozdielom medzi dielčími súradnicami oboch vektorov. Oproti euklidovej metrike je tiež výpočetne menej náročná. Pre zachovanie kladnej výslednej hodnoty vzdialenosti je nevyhnutná absolútna hodnota. Geometrickým miestom s rovnakou vzdialenosťou je štvorec vo vnútri Euklidovej kružnice.

$$d_{st} = \sum_{j=1}^{n} \left| x_{sj} - x_{tj} \right| \tag{16}$$

2.4.5 Mahalanobisová metrika

Táto metrika berie v úvahu koreláciu medzi premennými a je nezávislá na rozsahu hodnôt premenných. Pracuje s rozdielnou variabilitou a korelačnou štruktúrou v dátach. Vzdialenosti medzi objektami počíta v systéme súradníc, ktorého osy nemusia byť na seba kolmé.

$$d_{st} = \sqrt{(x_s - x_t)\boldsymbol{C}^{-1}(x_s - x_t)'} \ , \tag{cc}$$

kde **C** je kovariančná matica.

2.4.6 Čebyševová metrika

Metrika sa používa pri výpočetne náročných prípadoch, kedy je náročnosť výpočtu podľa euklidovsky orientovaných metrík neprijateľná. Geometrickým miestom bodov s rovnakou čebyševovskou vzdialenosťou od daného bodu je kocka, čo je zobrazené na Obrázok 9.

$$d_{st} = max_j\{\left| x_{sj} - x_{tj} \right|\} \tag{17}$$

2.4.7 Pearson korelačný koeficient

Koeficient je vyjadrený podľa[27] ako:

$$S_{PC}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{\boldsymbol{x}_{d1}^T \boldsymbol{x}_{d2}}{||\boldsymbol{x}_{d1}|| ||\boldsymbol{x}_{d2}||}, \tag{18}$$

kde diferenčné vektory sú vyjadrené ako $\boldsymbol{x}_{di} = (x_{i1} - \overline{x_\iota},\ x_{i2} - \overline{x_\iota}, \dots\ x_{in} - \overline{x_\iota},)^T$, x_{il} predstavujú *j*-tú súradnicu vektoru $\boldsymbol{x}_i$ a $\overline{x_\iota}$ je stredná hodnota určená zo súradníc vektoru $\boldsymbol{x}_i$ $(\overline{x_\iota} = \frac{1}{n}\sum_{j=1}^{n} x_{ij})$. Pearson korelačný koeficient nadobúda hodnoty z intervalu $\langle -1; 1 \rangle$. Z hodnôt Pearsonovho korelačného koeficientu sa následne vzdialenosť medzi vektormi určí ako:

$$D_{PC}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{1 - S_{PC}(\boldsymbol{x_1}, \boldsymbol{x_2})}{2}. \tag{19}$$

2.4.8 Spearmanov korelačný koeficient

Popisuje ako dobre vzťah medzi dvoma veličinami odpovedá monotónnej funkcii, ktorá môže byť nelineárna. Koeficient je definovaný ako:

$$d_{st} = 1 - \frac{(r_s - \overline{r_s})(r_t - \overline{r_t})'}{\sqrt{(r_s - \overline{r_s})(r_s - \overline{r_s})'}\sqrt{(r_t - \overline{r_t})(r_t - \overline{r_t})'}}, \tag{20}$$

kde r_s a r_t sú pozície vektorov x_s a x_t, $r_s = (r_{s1}, r_{s2}, \dots\ r_{sn})$. $\overline{r_s} = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$, $\overline{r_t} = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$.

2.5 Hodnotenie úspešnosti klasifikácie

Keďže nie je dôležité len zvoliť dobrú klasifikačnú metódu a klasifikátor natrénovať, ale aj schopnosť overiť si úspešnosť vybranej klasifikačnej metódy, tak sa na zistenie výsledkov a porovnaní s ostatnými metrikami ideálne využíva testovanie na dvoch dátových súboroch. Jeden súbor sa využije na naučenie klasifikátoru a druhý súbor na samotné testovanie úspešnosti klasifikácie. Dáta, ktoré sa použijú pri učení klasifikátoru sa potom označujú ako trénovacie.

Keďže v reálnych podmienkach vedeckého výskumu nie sú takéto dva nezávislé dátové súbory často dostupné, pristupuje sa k deleniu jedného dátového súboru na trénovací a testovací. Pri trénovacích dátach je dôležitá znalosť skutočného zaradenia objektov do daných tried. Porovnáva sa výsledok klasifikácie objektov so skutočnosťou, z ktorej vychádza matica zámen zobrazená na tabuľke 1 nižšie [28][31].

Tabuľka 1 – matica zámen

Trieda zaradenia	Výsledok klasifikácie	
	Pozitívny objekt	**Negatívny objekt**
Pozitívny objekt	TP	FN
Negatívny objekt	FP	TN

V matici zámen jednotlivé skratky vyjadrujú:

- TP – true positive : koľko výsledkov klasifikácie označených za pozitívne boli aj v skutočnosti pozitívne
- TN – true negative : koľko výsledkov klasifikácie označených za negatívne boli aj v skutočnosti negatívne
- FP – false positive : koľko výsledkov klasifikácie označených za pozitívne boli v skutočnosti negatívne
- FN – false negative : koľko výsledkov klasifikácie označených za negatívne boli v skutočnosti pozitívne

Z týchto hodnôt sa následne dajú odvodiť miery hodnotenia úspešnosti klasifikácie testovaných dát ako:

$$správnosť = \frac{TP+TN}{TP+TN+FP+FN} \tag{22}$$

$$chyba = \frac{FP+FN}{TP+TN+FP+FN} \tag{23}$$

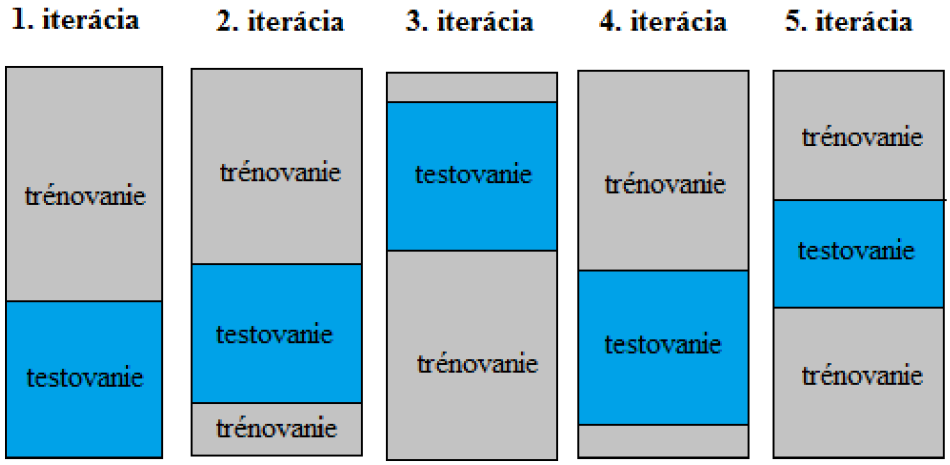
$$senzitivita = \frac{TP}{TP+FN} \tag{24}$$

$$specificita = \frac{TN}{TN+FP} \tag{25}$$

Správnosť vyjadruje podiel správne klasifikovaných objektov zo všetkých testovaných objektov. Chyba naopak udáva podiel chybne klasifikovaných objektov zo všetkých objektov. Senzitivita je podiel objektov správne klasifikovaných za pozitívne ku všetkým pozitívnym objektom a specificita vyjadruje podiel objektov správne klasifikovaných za negatívne ku všetkým negatívnym subjektom.

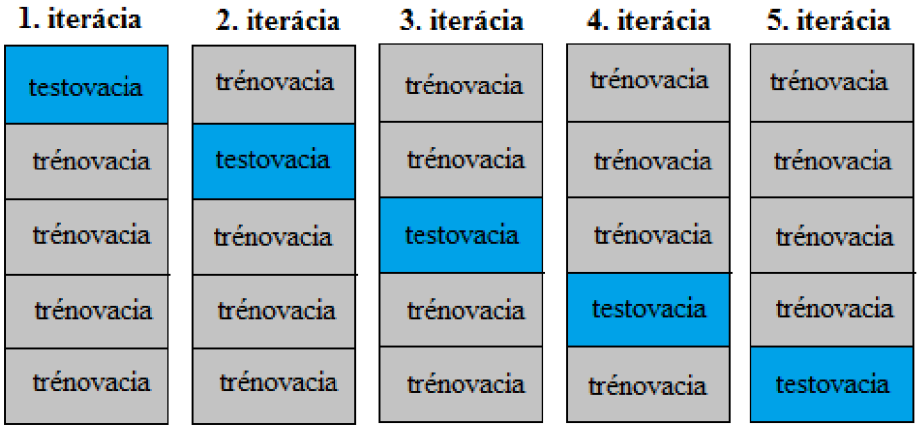
Pri absencii druhého nezávislého dátového súboru sa na učenie a testovanie používa ten istý dátový súbor, pričom sa využíva jeden zo štyroch základných prístupov:

- Resubstitúcia : jednoduchá a rýchla metóda, často vedie k nadhodnoteným výsledkom klasifikácie. Pri použití klasifikátoru s dobrými výsledkami resubstitúcie na inom dátovom súbore už nemusí byť úspešnosť taká vysoká.
- Náhodný výber s opakovaním – Bootstrap : metóda je založená na *N* krát náhodnom výbere objektov s opakovaním z pôvodného dátového súboru, ktoré sa použijú ako testovací súbor. Pri rozumnej veľkosti dát sa vyberie okolo 63% objektov určených na učenie a 37% objektov na testovanie. Nevýhoda metódy je opakovanie objektov v treňovacom súbore a výhodou je rýchlosť metódy.
- Predikčné testovanie externou validáciou – Hold out : približne jedna tretina dát je použitá na testovanie a ostávajúce dve tretiny sú použité na učenie klasifikátoru. Výhodou je nezávislý testovací a trénovací dátový súbor, v ktorých sa objekty neopakujú. Nevýhodou však môže byť menej dát na trénovanie a samotné testovanie. Výsledok je vysoko závislí na výbere trénovacích dát, čo viedlo k vytvoreniu modifikácií metódy. Jedna z nich využíva polovicu dát na trénovanie a druhú polovicu na testovanie. Následne sa dátové súbory prehodia a výsledok sa spriemeruje, čo však pri malých dátových súboroch nemusí na trénovanie stačiť. Preto sa v praxi viac používa *r* krát náhodné rozdelenie súboru na testovací a trénovací a výsledných *r* výsledkov sa následne spriemeruje. Nevýhodou tohto prístupu je vysoká časová náročnosť a prekryv trénovacích a testovacích dát. Ukážka zobrazená na Obrázok 10.



Obrázok 10 – hold out metóda – rozdelenie súboru na testovací a trénovací s opakovaním, upravené a prevzaté z [31]

- K–násobná krížová validácia – Cross validation : jeden z prípadov je k–fold metóda, ktorá dátový súbor rozdelí na *k* častí. Jedna sa vždy použije na testovanie a ostatných *k-1* je použitých na trénovanie. Iterácie prebiehajú tak, že každá časť je na testovanie použitá len raz. Odstraňuje nedostatky ako prekryv testovacích a trénovacích dát pri použití hold out metódy, nevýhoda je však časová náročnosť. Špeciálnym prípadom k-fold metódy je leave–one–out metóda, v ktorej platí $k=N$, takže v každej z N iterácii sa jeden objekt použije na testovanie a ostatných $N-1$ je použitých na trénovanie. Výsledok úspešnosti tak už nezávisí na rozdelení dát na trénovacie a testovacie. Táto metóda je na čas najnáročnejšia zo všetkých spomenutých a je vhodná pre malé súbory dát. Metóda leave–one–out je zobrazená na Obrázok 11.

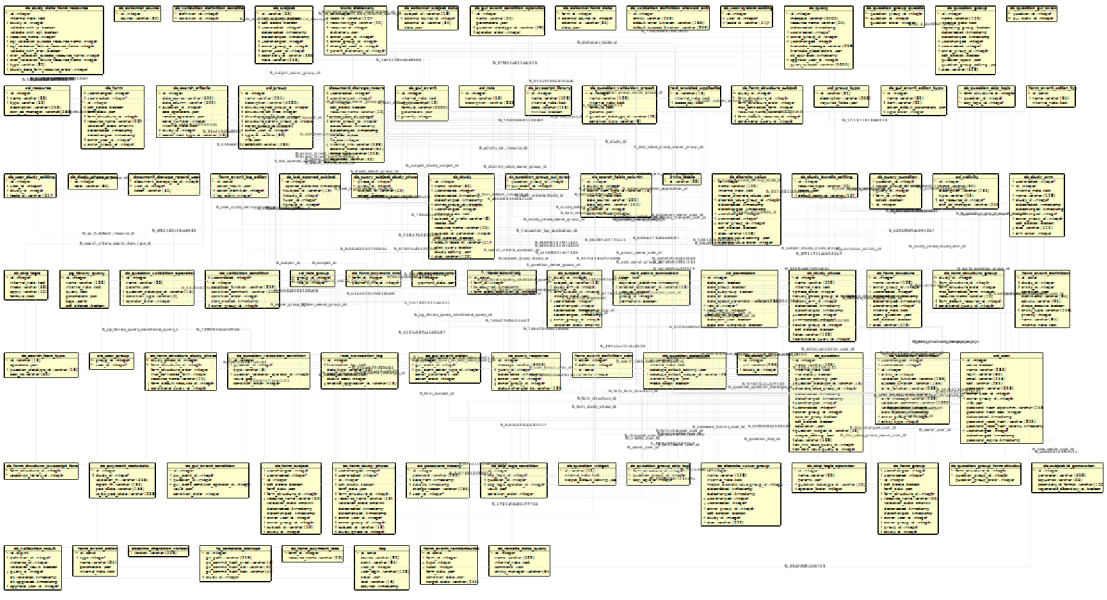


Obrázok 11 – krížová validácia leave–one–out, upravené a prevzaté z [31]

3 NÁVRH VLASTNÉHO SPRACOVANIA

3.1 Výber informačného systému

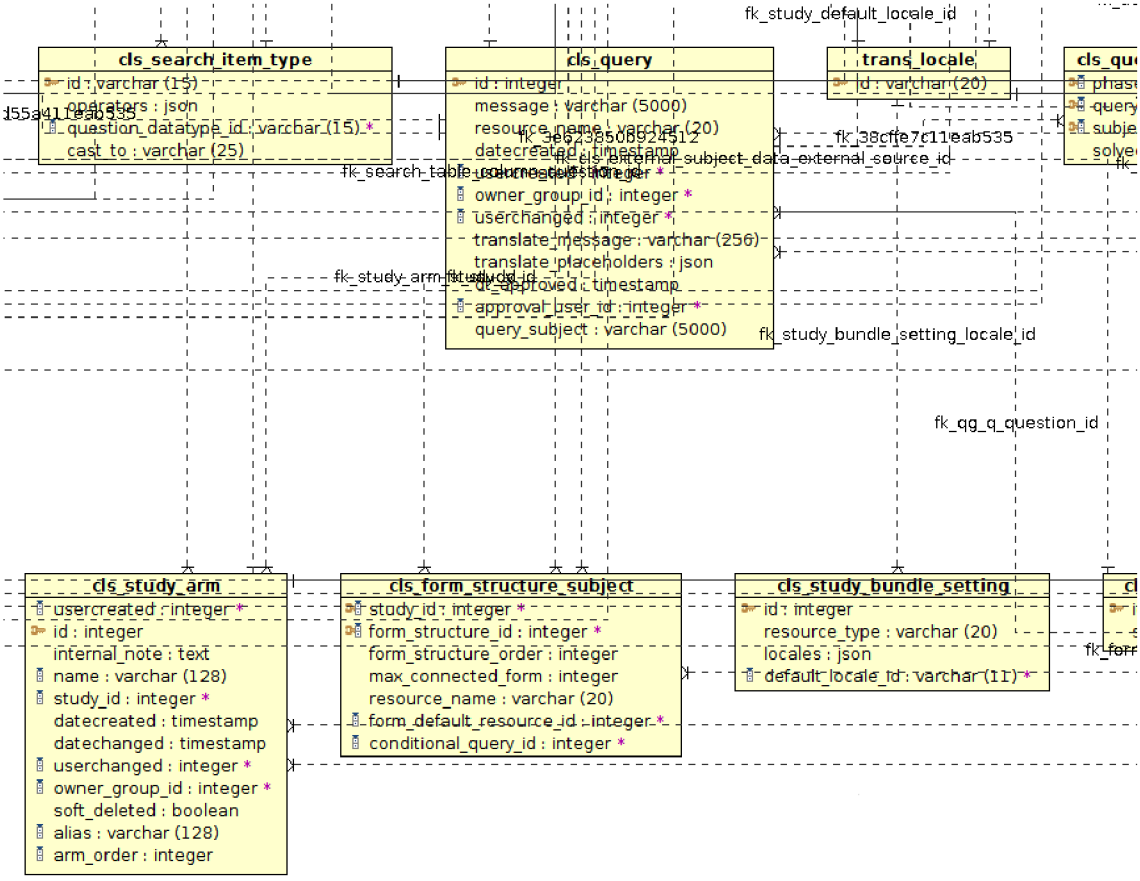
Za účelom vypracovania diplomovej práce bol vybraný CLADE–IS (Clinical Data Warehousing Information System) pre opakované snímanie elektronických dát (EDC), ktorý sa využíva v klinickej praxi. Software podporuje klasické randomizované štúdie (RCTs) ako aj design štúdii neintervenčných, popisovaných ako real–world data/real–word evidence (RWD/RWE). Systém zahŕňa všetky kroky potrebné pre organizáciu klinickej štúdie v súlade s dobrou klinickou praxou GCP. Výhodou systému je dostupnosť online cez webový prehliadač. Odpadá tak prácna inštalácia softwaru samotným užívateľom. Ďalšou výhodou je jeho univerzálnosť. Keďže systém vychádza z jedeného ERD diagramu, zobrazeného na Obrázok 12, schéma databáze zostane vždy zachovaná aj pri rôznych typoch klinických štúdií. Jednotlivé tabuľky predstavujú entity a väzby medzi nimi sú vyjadrené pomocou primárnych a cudzích kľúčov v pomere 1:1, vyjadrujúcej vzťah jedného pacienta a jedného centra. Ďalšia väzba 1:*N*, môže vyjadriť, že jeden lekár ošetruje viacero pacientov, ale každý pacient chodí na kontrolu len k danému lekárovi. Posledná väzba *N*:*M* vyjadruje, že určité množstvo pacientov berie určitú kombináciu liekov. Detail použitého ERD diagramu je zobrazený na Obrázok 13. Vývoj softwaru CLADE-IS je orientovaný na open source technológie. Pracuje s PostgreSQL databázovým systémom, na aplikačnej úrovni je to PHP/Symfony a ako celok beží na Linux Ubuntu LTS operačnom systéme [31].



Obrázok 12 – ERD diagram CLADE–IS, obrázok nie je čitateľný a slúži len na ukážku prepojenia entít a zložitosti zberného systému

Tento informačný systém bol vybraný hlavne pre jeho robustnosť, použiteľnosť a otvorenosť. Robustnosť v zmysle, že vie pracovať s veľkým rozsahom klinických štúdii. Má užívateľsky prívetivé prostredie, ktoré predstavuje dobrú použiteľnosť. Otvorenosť spočíva v dobrej komunikácii s ostatnými systémami, ktoré do informačného systému pridávajú rôzne rozšírené funkcie.

Počas absolvovania povinnej praxe som mal možnosť spolupracovať s vývojovým tímom CLADE–IS, ktorý prerástol do ďalšej spolupráce. Od tohto tímu som následne získal dátové súbory z troch uzavretých klinických štúdií bližšie popísaných nižšie.



Obrázok 13 – detail časti ERD diagramu CLADE–IS

3.2 Informácie o vybraných klinických štúdiách

Dáta zo štúdií boli poskytnuté ako surové dáta vo forme SQL dumpov a tiež vo forme slúžiacej na štatistickú analýzu ako .xlsx export súbor z CLADE–IS. Tento súbor je oproti dumpom značne zjednodušený a vyžaduje prevedenie mnoho transformačných krokov. Ukážka .xlsx súboru je pre lepšiu predstavu zobrazená na Obrázok 14 nižšie. V dodanom .xlsx súbore sa nachádza mnoho listov (dolná lišta MS Excel), z ktorých každý vyjadruje jeden formulár otázok. Každý formulár obsahuje premenné viacerých dátových typov. V prvých stĺpcoch je spravidla jednoznačný identifikátor záznamu pacienta. Ďalej nasledujú identifikačné údaje daného formulára, dátumy vytvorenia a prípadnej zmeny údajov a potom už jednotlivé skúmané parametre podľa typu formulára. V jednom z formulárov sa nachádza list s popisom každej premennej v štúdii a jej dátového typu.

	A	B	C	D	E	F	G	H
1	SUBJECT_ID	SECONDARY_ID	SITE_ID	FORM_ID	FORM_STRUCTURE_NAME	DATE_CREATED	DATE_CHANGED	hem_nd_1118
2	DAR-0000007	IN-006-005	16	6601	Laboratoty, VS FU Phase 8	2017-06-12	2017-06-12	No
3	DAR-0000009	IN-001-001	18	6820	Laboratoty, VS FU Phase 8	2017-06-20	2017-06-22	No
4	DAR-0000011	IN-001-003	18	6981	Laboratoty, VS FU Phase 8	2017-06-22	2017-06-22	No
5	DAR-0000031	IN-014-006	22	6596	Laboratoty, VS FU Phase 8	2017-06-10	2017-06-12	Yes
6	DAR-0000034	IN-009-003	23	6746	Laboratoty, VS FU Phase 8	2017-06-19	2017-06-19	Yes
7	DAR-0000042	IN-009-004	23	6825	Laboratoty, VS FU Phase 8	2017-06-21	2017-06-21	Yes
8	DAR-0000043	IN-009-005	23	6751	Laboratoty, VS FU Phase 8	2017-06-19	2017-06-19	Yes

... A1P10F48G317 A1P10F49 A1P10F49G319 A1P10F49G321 A1P10F50 A1P10F50G333 He ...

Obrázok 14 – ukážka dát (z .xlsx súboru) jedného z formulárov štúdie

3.2.1 Štúdia číslo 1 – D

Štúdia o výsledkoch pacientov so symptomatickou anémiou spojenou s chronickým zlyhaním obličiek na dialýze zrovnávaných s pacientami bez dialýzy, ktorý začali používať novú liečbu. Počet pacientov 2500, z toho 1250 podstupuje dialýzu a 1250 nepodstupuje dialýzu.

3.2.2 Štúdia číslo 2 – I

Register zameraný na diagnózu pacientov s chronickou myeloidnou leukémiou. Pacienti sa sledujú v dlhom časovom horizonte a pozorujú sa nežiadúce účinky. U pacientov sa hodnotia liečebné odpovede – hematologické, cytogenetické a molekulárne. V súčasnej dobe sú zapojené 4 česká centra. Pri analýzach sa sledujú časy do úmrtia, do progresie, ale aj do dosiahnutia jednotlivých liečebných odpovedí.

3.2.3 Štúdia číslo 3 – R

Neintervenčné klinické hodnotenie opisujúce, ako pacienti vnímajú antikoagulačnú liečbu a liečebný komfort spojený s liečbou novými liekmi pre prevenciu cievnej mozgovej príhody u nechlopňovej fibrilácii predsiení. V štúdii je zahrnutých cca 9000 pacientov z 11 krajín strednej a východnej Európy (Rusko, Poľsko, Rumunsko, Maďarsko, Rakúsko, Česká republika, Bulharsko, Estónsko, Slovinsko, Srbsko a Izrael).

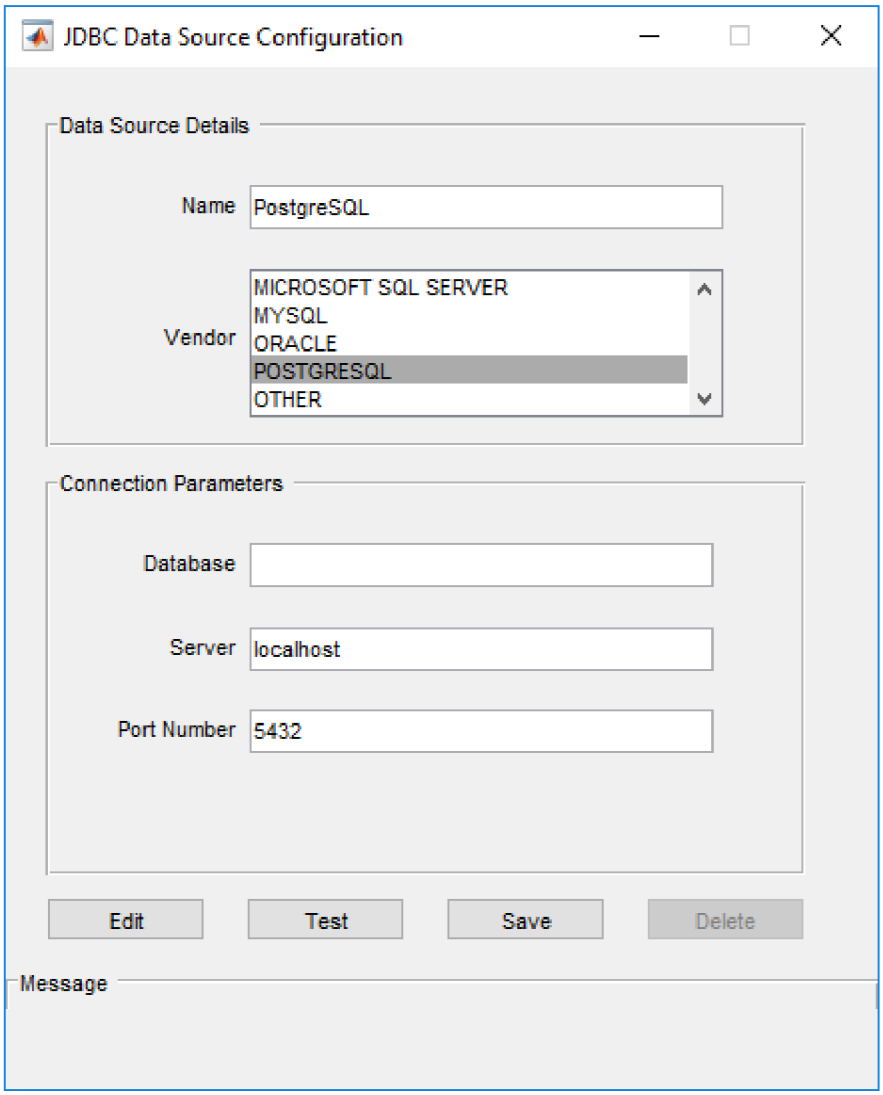
3.3 Načítanie a predspracovanie dát

Algoritmus potrebné dáta nenačítava z dostupných SQL dumpov a xlsx súborov, ale z dôvodu zaistenia čo najvyššej aktuálnosti načítava dáta priamo z databázy. V databáze boli z tabuliek vytvorené dátové pohľady, ktoré obsahujú potrebné dáta organizované po jednotlivých formulároch. Príklad vytvorenia takéhoto pohľadu pomocou SQL príkazu:

```
CREATE OR REPLACE VIEW custom.cls_form_view_fs6 AS  SELECT s.id AS subject_id,
    s.secondary_id,
    fsp.id AS form_id,
    gr.id AS group_id,
    ss.datecreated,
    ((fsp.form_data -> 'Q124'::text) ->> 'value'::text) AS control_datamanager_inf,
    ((fsp.form_data -> 'Q125'::text) ->> 'value'::text) AS control_doctor_inf
   FROM (((((cls_subject s
     JOIN cls_form_study_phase fsp ON ((((fsp.subject_id)::text = (s.id)::text) AND
(fsp.form_structure_id = 6))))
     JOIN cls_study_phase ph ON ((ph.id = fsp.study_phase_id)))
     JOIN cls_subject_study ss ON (((ss.subject_id)::text = (s.id)::text)))
     JOIN acl_group gr ON ((gr.id = ss.owner_group_id)))
     JOIN cls_form_subject fss ON (((fss.subject_id)::text = (s.id)::text)))
  WHERE ((ss.study_id = 612) AND (NOT s.test_subject) AND (NOT s.soft_delete))
  ORDER BY s.id, fsp.id;
```

Z príkazu je možné vidieť, že dáta sa skladajú z rôznych tabuliek, ktoré boli spojené pomocou príkazu *JOIN* a referenčná integrita bola zaistená cez jednoznačný identifikátor záznamu pacienta.

Na pripojenie do databáze sa využíva Java Database Connectivity – JDBC Postgresql Matlab ovládač, ktorý umožňuje prístup do PostgreSQL databázy priamo z príkazového riadku. Najprv je nutné ovládač stiahnuť z oficiálnych stránok [33] a pridať ho do cesty statickej Java triedy. V zložke s Matlab konfiguračnými súbormi sa vytvorí textový súbor s názvom `javaclasspath.txt`. Do tohto súboru sa následne vloží cesta k stiahnutému JDBC Postgresql ovládaču. Následne sa za pomoci aplikácie Database Explorer v Matlabe vyberie a nakonfiguruje JDBC zdroj dát. Pri nastavení sa volí meno zdroju dát, typ databázového systému a parametre pripojenia ako meno databáze, server na ktorom sa nachádza, a port na ktorom sa daná databáza vyskytuje. Ukážka pripojenia cez JDBC je zobrazená na Obrázok 15 nižšie. Následne sa môžu spúšťať ANSI SQL otázky priamo z Matlab command window.



Obrázok 15 – pripojenie sa z Matlabu do databáze pomocou JDBC ovládača

Testované boli dva prístupy načítania dát. V prvom sa databázové dáta z jednotlivých formulárov načítali do jednej veľkej matice, podľa odpovedajúcich záznamov pacientov. Keďže veľa záznamov obsahovalo nevyplnenú hodnotu v niektorej z premenných, čo by následne znižovalo veľkosť dátovej matice bol použitý druhý prístup. Ten dáta načíta po jednotlivých formulároch, čo zaisťuje vyplnenejšie záznamy pacientov a poskytuje výhodu pri viacrozmernej analýze. Súhrny prázdnych hodnôt pre načítanie formulárov do jedného súboru a pre prístup po jednotlivých formulároch sú zobrazené v tabuľke 2 a tabuľke 3 nižšie. Aby následná štatistická analýza mala zmysel, do ďalšej analýzy prešli len formuláre s minimálnou veľkosťou desať záznamov pacientov.

Tabuľka 2 – načítanie všetkých formulárov štúdií do jedného veľkého súboru

	Štúdia D	**Štúdia I**	**Štúdia R**
Počet nevyplnených polí celkovo	290774	2952741	1301362
Počet záznamov pacientov x premenných	427 x 882	13022 x 238	9491x258
Počet vymazaných záznamov pacientov	427	13022	9491

Z tabuľky 2 je jasne viditeľné, že pri prístupe cez spájanie formulárov do jedného veľkého súboru sa nevyplnili všetky polia záznamu pacienta ani v jednom prípade z testovaných štúdií. Lepšie výsledky dosiahol prístup po jednotlivých formulároch. Výsledky sú zobrazené v tabuľke 3 nižšie.

Tabuľka 3 – načítanie dát pre prístup po formulároch, boli vybrané najväčšie formuláre jednotlivých štúdií

	Štúdia D	**Štúdia I**	**Štúdia R**
Počet nevyplnených polí celkovo	0	0	2
Počet záznamov pacientov x premenných	1285x 16	2379 x 12	5327x26
Počet vymazaných záznamov pacientov	0	0	1

Keďže surové dáta nie sú dostatočne pripravené na ďalšiu analýzu, je potrebné ich predspracovať. Dáta zadané zadávateľom sa môžu líšiť podľa typu na:

- Kategoriálne a binárne: premenné sú rozdelené do skupín, v prípade binárnych (dummy) len do dvoch, napríklad muž/žena, alebo prítomnosť/neprítomnosť sledovaného znaku
- Ordinálne: je možné porovnať, či je jedna hodnota väčšia ako druhá
- Spojité: patria sem pomerové dáta, ktoré majú absolútnu nulu, napríklad výška pozorovaného subjektu a intervalové dáta, ktoré nemajú zmysluplnú nulu, napríklad teplota

Chýbajúce hodnoty môžu byť označené ako NA, alebo hodnotou, ktorá sa nezhoduje s možnými hodnotami vybraných premenných. V diplomovej práci boli záznamy pacientov, ktoré obsahovali prázdnu premennú vymazané. Táto metóda je najjednoduchšia, ale treba si uvedomiť, že jej použitím sa stráca určitá informácia a preto treba byť pri jej použití opatrný. Z doplňovacích metód je možné použiť doplnenie priemeru z hodnôt premenných, ktoré sú k dispozícii, alebo metódu mnohonásobného regresného modelu na vzorky bez chýbajúcich hodnôt.

3.4 Transformácia na numerické hodnoty a ich štandardizácia

Keďže pre ďalšiu štatistickú analýzu sú potrebné len numerické hodnoty, pre nenumerické premenné nasledovala ich transformácia. Pre premenné typu integer a real nebola nutná úprava. Premenná typu date síce je v numerickom tvare, avšak pre ďalšiu analýzu bola použitá transformácia. Z viacerých metód bola vybraná datenum transformácia, ktorá premennú typu date prevedie na numerickú hodnotu, ktorá reprezentuje počet dní od dátumu 0.január.0000. Keďže sa vo vyplnených formulároch môže hodnota dátumu objaviť vo viacerých formátoch, je nutné pokryť každý z nich.

Pre kvantifikovateľne vymenované premenné sa z usporiadaných hodnôt vytvorí očíslovaná škála pre každú hodnotu danej premennej. Takto sa prevedú číselníkové premenné discrete values. Nekvantifikovateľné premenné sa prevedú na binárne premenné. Zo všetkých typov tak boli vynechané len premenné typu string a text, ktoré najčastejšie predstavujú pole poznámky. Možná transformácia tohto typu by bola možná cez ASCI tabuľku, ktorá každému písmenu priradí odpovedajúce číslo. Metódu však práca nevyužíva z dôvodu malého výskytu premenných tohto typu v dodaných štúdiách, čo je spôsobené nepovinným vyplnením tejto hodnoty.

Pri skladaní matice dát bolo tiež nutné vyriešiť otázku uloženia rôznych dátových typov v jednej matici. Algoritmus po prvotnom načítaní formulára do matice typu cell po odstránení nevyplnených záznamov a transformácií pracuje s maticou typu double.

Keďže každá zo skúmaných premenných môže pochádzať z iného rozloženia a rozsahu, ďalším krokom procesu je štandardizácia premenných ich rozpätím na hodnoty od nula do jedna. Na konci procesu predspracovania dát tak každý riadok vybraného formulára klinického registru predstavuje jeden plne vyplnený záznam pacienta a každý odpovedajúci stĺpec formulára predstavuje numericky reprezentovanú premennú. Ukážka formulára po štandardizácii je zobrazená na Obrázok 16.

3241x72 double

	1	2	3	4	5	6	7	8	9
1	0.5556	0.1250	0.2222	0.1448	0.5556	0.1250	0.2222	0.4257	0.3921
2	0.6914	0.1250	0.4444	0.0619	0.6914	0.1250	0.4444	0.2905	0.3271
3	0.6296	0.3750	0.8889	0.0720	0.6296	0.3750	0.8889	0.2905	0.2621
4	0.5432	0.2500	0.3333	0.0309	0.5432	0.2500	0.3333	0.4257	0.2889
5	0.7284	0.1250	0.4444	0.1081	0.7284	0.1250	0.4444	0.1892	0.3959
6	0.6914	0.5000	0.5556	0.0788	0.6914	0.5000	0.5556	0.2500	0.2965
7	0.5062	0.3750	0.2222	0.1303	0.5062	0.3750	0.2222	0.3176	0.4380
8	0.6914	0.3750	0.7778	0.0634	0.6914	0.3750	0.7778	0.1554	0.0028
9	0.6543	0.5000	0.4444	0.1391	0.6543	0.5000	0.4444	0.2095	0.4112
10	0.6420	0.6250	0.8889	0.2099	0.6420	0.6250	0.8889	0.4730	0.3309
11	0.9259	0.5000	0.6667	0.1644	0.9259	0.5000	0.6667	0.1824	0.3080

Obrázok 16 – výrez z formuláru po štandardizácii dát

3.5 Modifikovaný k-means

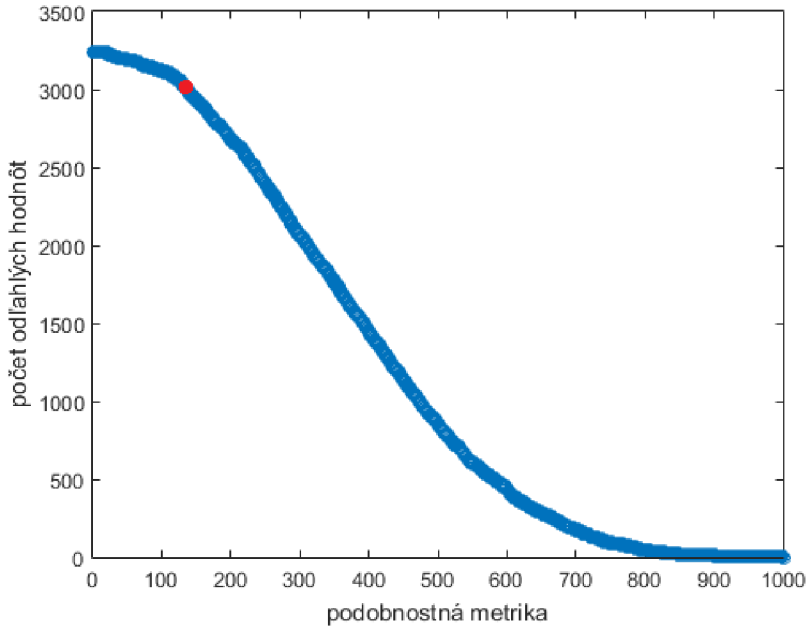
Na klasifikáciu dát bola použitá metóda k-means s niekoľkými vlastnými úpravami. Pre každý formulár registra je vytvorený jeden centroid. Od tohto centroidu sa následne spočíta podobnosť pre každý záznam pacienta daného formulára. Jeden riadok záznamu v registri, tak už nie je reprezentovaný súborom premenných, ale jednou hodnotou vybranej metriky podobnosti. Oproti metóde k-means sa tak už nevyužívajú iterácie, pretože cieľom nie je prevedenie zhlukovej analýzy, ale určenie tých pozorovaní, ktoré sú od centroidu vzdialené viac, ako definovaná prahová hodnota vzdialenosti. Takto vzdialené pozorovania je potom možné považovať za anomálne dáta, alebo aspoň za kandidátov na anomálne dáta.

Pre klasifikáciu anomálii v dátach sa použila vždy kombinácia troch metrík, ktoré predstavovali najväčšiu presnosť pri trénovaní na konkrétnej štúdii . Popis testovaných metrík a použitých vzorcov na výpočet je spomenutý v kapitole Metriky podobnosti a vzdialenosti.

3.6 Určenie prahu

Na určenie hranice podobnosti, za ktorou sa daný záznam v klinickom registri bude označovať za odľahlý sa testovalo viacero metód. Najlepšie výsledky dosahovala hodnota hranice určená ako 1,5 násobok interkvartilového rozpätia pripočítaná k tretiemu kvartilu analyzovanej premennej spomenutej vo vzorcoch (5) – (8).

Ďalšia z možných metód určenia prahu je založená na detekcii zlomu v krivke závislostí počtu nájdených odľahlých hodnôt na hodnote prahu. Tento prah je iterovaný automaticky od minima do maxima danej metriky podobnosti s konfigurovateľným krokom iterácie. Po tejto krivke sa po jednotlivých vzorkách posúva zvolené okno, ktoré určí miesto najväčšieho zlomu. V tomto mieste je následne zvolený prah odľahlej hodnoty. Pre lepšiu predstavu je detekcia zlomu v krivke zobrazená na nasledujúcom Obrázok 17.



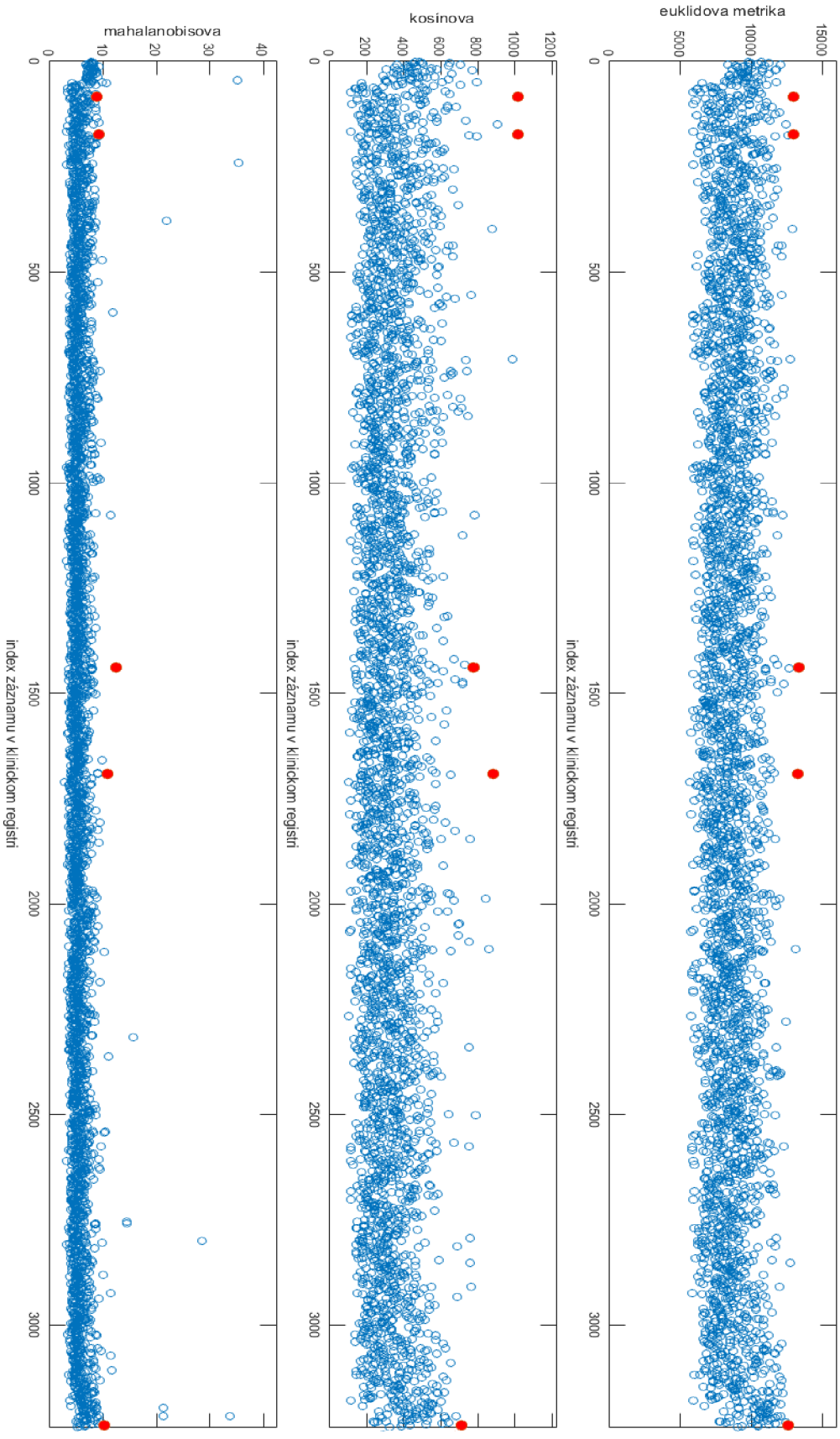
Obrázok 17 – detekcia zlomu v krivke závislosti počtu nájdených odľahlých hodnôt na hodnote prahu, metrika podobnosti zoradená od minima do maxima

Posledná z testovaných možností na zistenie optimálneho prahu určí za odľahlé hodnoty tie záznamy, ktoré sa v histograme vzdialeností vyskytovali do desiateho percentilu, alebo nad deväťdesiatym percentilom skúmanej vzdialenosti.

3.7 Potvrdenie detekcie odľahlej hodnoty

Na prehlásenie záznamu v klinickom registri za vysoko odľahlý ho podľa navrhnutej metódy musia za odľahlý zaradiť všetky tri testované metriky podobnosti. Ak záznam zaradili za odľahlý len dve metriky podobnosti, ide o stredne odľahlý záznam. Vysoko a stredne odľahlé záznamy by mali byť následne z klinického registru odstránené, alebo by im mal dáta management venovať zvýšenú pozornosť. Pri zaradení záznamu za nízko odľahlý jednou metrikou podobnosti je nutné potvrdenie, či sa skutočne jedná o odľahlú hodnotu potrebné schváliť ručne. Pri potvrdení je možné si prehliadnuť doplňujúce informácie o danom zázname v odpovedajúcom formulári. Zobrazuje sa jednoznačný identifikátor záznamu pacienta, hodnoty a zobrazenie samotných metrík podobnosti.

Porovnanie troch metrík podobnosti pre skúmaný formulár je na Obrázok 18. Osa y predstavuje 3241 záznamov pacientov z formulára klinického registra a osa x predstavuje vzdialenosť centroidu od odpovedajúceho záznamu pacienta. Čím je hodnota tejto metriky väčšia, tým je sledovaný záznam pacienta menej podobný centroidu. Červenou farbou sú vyznačené záznamy pacientov, ktoré boli označené za odľahlé všetkými troma metrikami podobnosti. Algoritmus ich teda klasifikoval ako vysoko odľahlé záznamy pacientov.



Obrázok 18 – Výsledky metrík podobnosti – červené vysoko odľahlé záznamy pacientov

3.8 Správnosť metódy

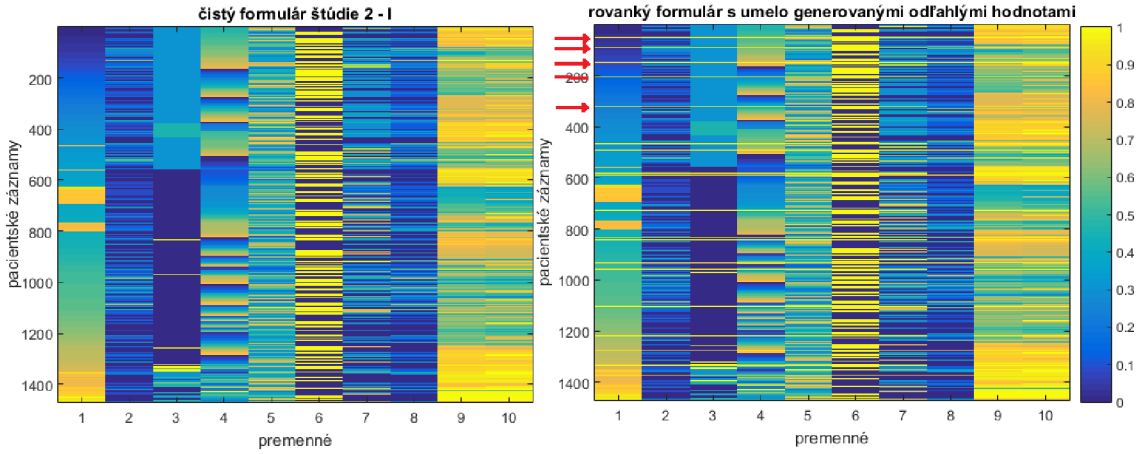
Pre hodnotenie kvality navrhnutej metódy klasifikácie je potrebné vedieť pozície skutočných odľahlých hodnôt od skúseného dátového manažéra klinickej štúdie. Tieto hodnoty sa však nepodarilo získať. Preto bol vytvorený umelo simulovaný dátový súbor, ktorý na definovaných miestach obsahoval anomálne dáta.

Na určenie správnosti metódy pre jednotlivú štúdiu sa načítal najväčší formulár z každej štúdie. Následne sa na každý formulár pustil algoritmus na detekciu odľahlých hodnôt. Záznamy pacientov s nájdenými vysoko, stredne a nízko odľahlými hodnotami boli z formulára odstránené. Dostávame tak, „čistý“ formulár bez odľahlých hodnôt.

Na takomto „čistom“ formulári sa následne umelo vygeneruje určitý počet vysoko, stredne a nízko odľahlých hodnôt. Dôležité je, že sú známe pozície, na ktorých boli tieto odľahlé hodnoty vygenerované. Tie sa použijú pri hodnotení úspešnosti detekcie a zostavovaní matici zámen. Po umelom vygenerovaní odľahlých hodnôt tak dostávame tzv. „nepravý“ formulár. Algoritmus na umelé generovanie odľahlých hodnôt je ukázaný na konci práce v časti Prílohy.

Pre vygenerovanie vysoko odľahlého záznamu sa odpovedajúce hodnoty upravili s veľkou zmenou na zhruba 90% premenných. Pre vygenerovanie stredne odľahlého záznamu sa odpovedajúce hodnoty upravili so strednou zmenou na zhruba 70% premenných. Pre vygenerovanie nízko odľahlého záznamu sa odpovedajúce hodnoty upravili s malou zmenou na zhruba 30% premenných. Určenie pozícií premenných, ktoré budú upravené sa generovalo náhodne. Preto sa mohlo stať, že na jednej premennej sa úprava previedla viac ako jeden krát. Vzorec na úpravu však vždy počíta s aktuálnou hodnotou premennej, ktorá bude vždy v intervale medzi maximálnou hodnotou 1 a minimálnou hodnotou 0.

Pre nepravý formulár sa umelo vygenerovalo 10 pozícií pre veľké zmeny, 30 pozícií pre stredné zmeny a 50 pozícií pre malé zmeny. To odpovedá 10 vysoko odľahlým, 30 stredne odľahlým a 50 nízko odľahlým záznamom pacientov. Keďže sú pozície odľahlých záznamov generované náhodne môžu sa aplikovať na jeden a ten istý riadok a teda rôzne typy odľahlých hodnôt tak môžu splynúť. Ukážka je zobrazená na Obrázok 19 nižšie.



Obrázok 19 – porovnanie „čistého“ formulára a na ňom vygenerované umelé hodnoty

Algoritmus následne rozdelí „nepravý“ formulár na trénovaciu a testovaciu časť, podľa metódy krížovej validácie leave one out. Za trénovaciu časť formulára sa určí vždy postupne jeden záznam pacienta. Ostatné záznamy formulára slúžili na testovanie. Princíp krížovej validácie je zobrazená na Obrázok 11.

Pre každú z dostupných štúdií bolo testovaných osem metrík podobnosti. Podľa počtu detekcie odľahlých hodnôt z „nepravého“ formulára, bolo následne z testovaných metrík podobnosti vybraných päť, ktorých výsledky najviac odpovedali počtu umelo generovaných odľahlých hodnôt. Tie boli následne použité na zistenie najlepšej testovacej trojice. Medzi sebou sa tak kombinovalo bez opakovania päť najlepších metrík podobnosti. Najlepšia trojica sa potom použila na detekciu odľahlých záznamov v danej klinickej štúdii.

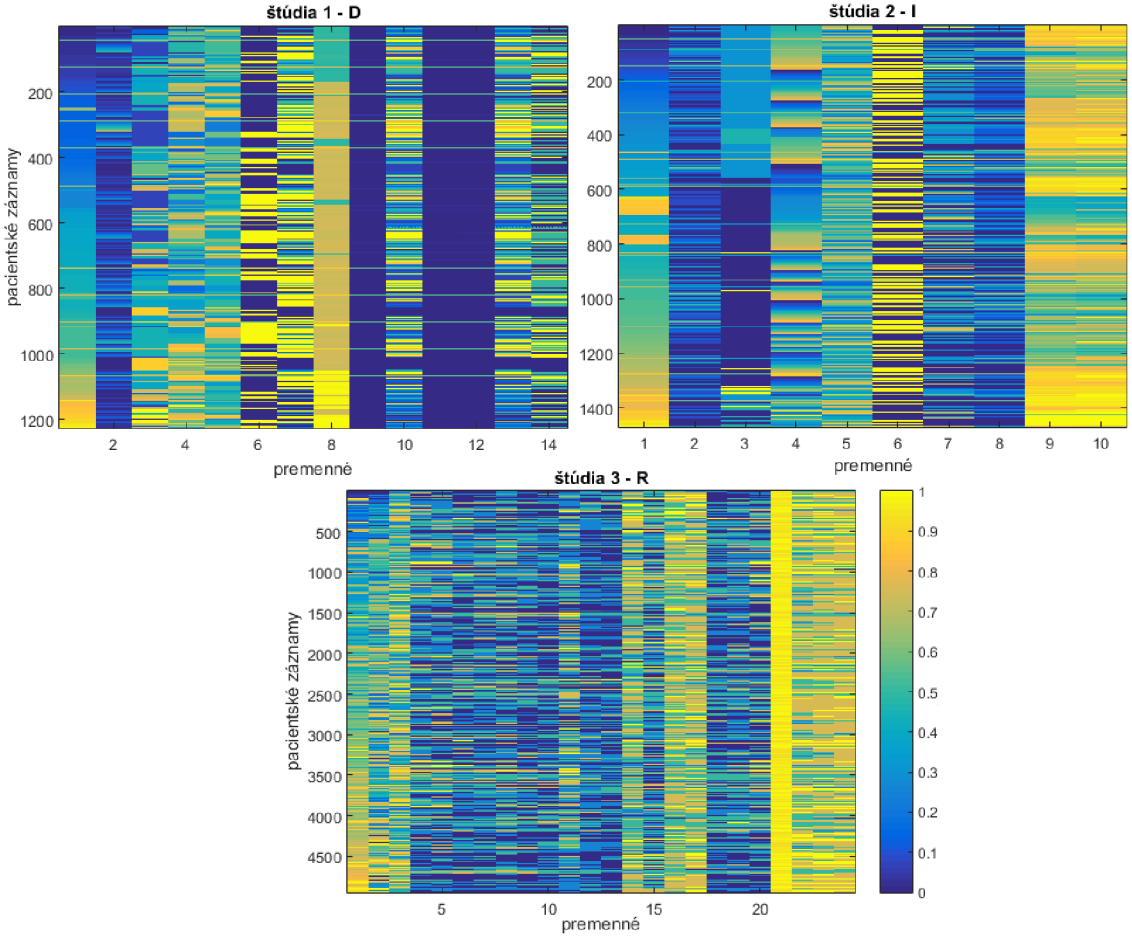
4 DOSIAHNUTÉ VÝSLEDKY

Algoritmus je univerzálny pre všetky databázy CLADE–IS, pretože sú založené na rovnakom ERD diagrame. V troch testovaných štúdiách je množstvo odľahlých záznamov úmerné veľkosti databáz.

Tabuľka 4 – nájdené odľahlé záznamy celkovo v testovaných štúdiách

Štúdia	Počet záznamov celkovo	Počet odľahlých záznamov	Priemerný počet odľahlých na formulár
1–D	257236	37645	316
2–I	199230	12979	432
3–R	45753	2644	139

Obrázok 20 zobrazuje najväčšie formuláre po umelom vygenerovaní odľahlých hodnôt pre každú z troch testovaných štúdií. Bol použitý príkaz *imagesc*, ktorý jednotlivým hodnotám priradí farebnú škálu a *colorbar*, ktorý farebnú škálu zobrazí.



Obrázok 20 – zobrazenie skúmaných formulárov pre každú štúdiu

4.1 Štúdia 1 – D

Pre štúdiu 1 – D bol vybraný ako reprezentatívny jej najväčší formulár.

Tabuľka 5 – počet generovaných a detekovaných odľahlých záznamov pre štúdiu 1–D

Metrika	Počet umelo generovaných odľahlých záznamov	Počet detekovaných odľahlých záznamov
Euklidová (EU)	90	28
Kosinová (KO)	90	3
Mahalanobisova (MA)	90	69
Minkowského (MI)	90	26
Spearman koef. (SP)	90	54
Pearson koef. (PE)	90	10
City block (CI)	90	49
Čebyševová (CHE)	90	4

V tabuľke 5 sa testovalo osem metrík podobnosti. Z nich sa vybralo päť metrík, ktorých hodnoty detekovaných odľahlých záznamov sa najviac priblížili umelo generovaným odľahlým záznamom. Z vybraných piatich metrík podobnosti sa následne vytvorili ich kombináciou trojice, ktorých úspešnosť sa následne testovala metódou leave–one–out. Výsledky úspešnosti detekcie pre vybrané trojice sú zobrazené v tabuľke 6 nižšie. Najvyššiu presnosť 95,84% detekcie dosiahla trojica City block, Euklidová a Minkowského metrika. Táto trojica bola potom použitá pri zistení celkového počtu odľahlých hodnôt v štúdii 1–D zobrazených v tabuľke 4.

Tabuľka 6 – kombinácie vybraných päť najúspešnejších metrík podobnosti pre štúdiu 1–D, metóda testovania leave–one–out

Trojice	Senzitivita	Specificita	Chyba	Presnosť
MA,SP,CI	0,4744	0,9591	0,0718	0,9282
MA,SP,EU	0,4487	0,9599	0,0726	0,9274
MA,SP,MI	0,4487	0,9599	0,0726	0,9274
MA,CI,EU	0,4744	0,9869	0,0457	0,9543
MA,CI,MI	0,4744	0,9869	0,0457	0,9543
MA,EU,MI	0,4487	0,9878	0,0465	0,9535
SP,CI,EU	0,3590	0,9713	0,0677	0,9323
SP,CI,MI	0,3590	0,9713	0,0677	0,9323
SP,EU,MI	0,1923	0,9721	0,0775	0,9225
CI,EU,MI	**0,3590**	**0,991**	**0,0416**	**0,9584**

4.2 Štúdia 2 – I

Pre štúdiu 2 – I bol vybraný ako reprezentatívny jej najväčší formulár.

Tabuľka 7 – počet generovaných a detekovaných odľahlých záznamov pre štúdiu 2–I

Metrika	Počet umelo generovaných odľahlých záznamov	Počet detekovaných odľahlých záznamov
Euklidová (EU)	90	61
Kosinová (KO)	90	9
Mahalanobisova (MA)	90	33
Minkowského (MI)	90	61
Spearman koef. (SP)	90	19
Pearson koef. (PE)	90	24
City block (CI)	90	30
Čebyševová (CHE)	90	72

V tabuľke 7 sa testovalo osem metrík podobnosti. Z nich sa vybralo päť metrík, ktorých hodnoty detekovaných odľahlých záznamov sa najviac priblížili umelo generovaným odľahlým záznamom. Z vybraných piatich metrík podobnosti sa následne vytvorili ich kombináciou trojice, ktorých úspešnosť sa následne testovala metódou leave–one–out. Výsledky úspešnosti detekcie pre vybrané trojice sú zobrazené v tabuľke 8 nižšie. Najvyššiu presnosť 94,42% detekcie dosiahla trojica Euklidová, Minkowského a Mahalanobisova metrika. Táto trojica bola potom použitá pri zistení celkového počtu odľahlých hodnôt v štúdii 2–I zobrazených v tabuľke 4.

Tabuľka 8 – kombinácie vybraných päť najúspešnejších metrík podobnosti pre štúdiu 2–I, metóda testovania leave–one–out

Trojice	Senzitivita	Specificita	Chyba	Presnosť
CHE,EU,MI	0,3165	0,9511	0,0830	0,9170
CHE,EU,MAH	0,4304	0,9504	0,0776	0,9224
CHE,EU,CI	0,3165	0,9504	0,0837	0,9163
CHE,MI,MAH	0,4304	0,9504	0,0776	0,9224
CHE,MI.CI	0,3165	0,9504	0,0837	0,9163
CHE,MAH,CI	0,2911	0,9511	0,0844	0,9156
EU,MI,MAH	**0,4304**	**0,9734**	**0,0558**	**0,9442**
EU,MI,CI	0,3165	0,9734	0,0619	0,9381
EU,MAH,CI	0,4304	0,9727	0,0565	0,9435
MI,MAH,CI	0,4304	0,9727	0,0565	0,9435

4.3 Štúdia 1 – R

Pre štúdiu 3 – R bol vybraný ako reprezentatívny jej najväčší formulár.

Tabuľka 9 – počet generovaných a detekovaných odľahlých záznamov pre štúdiu 3–R

Metrika	Počet umelo generovaných odľahlých záznamov	Počet detekovaných odľahlých záznamov
Euklidová (EU)	90	90
Kosinová (KO)	90	126
Mahalanobisova (MA)	90	262
Minkowského (MI)	90	90
Spearman koef. (SP)	90	339
Pearson koef. (PE)	90	294
City block (CI)	90	63
Čebyševová (CHE)	90	92

V tabuľke 9 sa testovalo osem metrík podobnosti. Z nich sa vybralo päť metrík, ktorých hodnoty detekovaných odľahlých záznamov sa najviac priblížili umelo generovaným odľahlým záznamom. Z vybraných piatich metrík podobnosti sa následne vytvorili ich kombináciou trojice, ktorých úspešnosť sa následne testovala metódou leave–one–out. Výsledky úspešnosti detekcie pre vybrané trojice sú zobrazené v tabuľke 10 nižšie. Najvyššiu presnosť 97,31% detekcie dosiahla trojica Euklidová, Minkowského a City block metrika. Táto trojica bola potom použitá pri zistení celkového počtu odľahlých hodnôt v štúdii 3–R zobrazených v tabuľke 4.

Tabuľka 10 – kombinácie vybraných päť najúspešnejších metrík podobnosti pre štúdiu 3–R, metóda testovania leave–one–out

Trojice	Senzitivita	Specificita	Chyba	Presnosť
EU,MI,CHE	0,2532	0,9722	0,0393	0,9607
EU,MI,CI	**0,2658**	**0,9846**	**0,0269**	**0,9731**
EU,MI,KO	0,2405	0,9689	0,0427	0,9573
EU,CHE,CI	0,2785	0,9714	0,0397	0,9603
EU,CHE,KO	0,2532	0,9570	0,0543	0,9457
EU,CI,KO	0,2658	0,9685	0,0427	0,9573
MI,CHE,CI	0,2785	0,9714	0,0397	0,9603
MI,CHE,KO	0,2532	0,9570	0,0543	0,9457
MI.CI,KO	0,2658	0,9685	0,0427	0,9573
CHE,CI,KO	0,2658	0,9587	0,0524	0,9476

5 DISKUSIA

Čas potrebný na výpočet algoritmu sa pri použití kombinácie viacrozmernej analýzy a prístupu priamo do databáze po jednotlivých formulároch znížil oproti prístupu s načítaním formulárov cez MS Excel pre všetky testované štúdie. Pre štúdiu 1 – D pri načítaní z MS Excelu trvalo nájdenie odľahlých hodnôt jednorozmernými štatistickými metódami približne 50 minút. Viacrozmernou analýzou to trvalo približne 20 minút. Pre štúdiu 2 – I pri načítaní z MS Excel 10 hodín a pri použití viacrozmernej metódy 2 hod. Pre štúdiu 3 – R sa čas zlepšil z 3 hodín pri načítaní z MS Excel na 1 hodinu pri použití viacrozmernej metódy s prístupom priamo do databázy.

Pri načítaní z MS Excel trvalo značnú dobu, kým sa načítal samotný súbor s dátami do prostredia Matlabu. S vyšším počtom formulárov v štúdii sa zvyšoval aj čas potrebný na ich načítanie. Táto vlastnosť odpadá použitím metódy prístupu k dátam priamo z databáze, kedy sú dáta k dispozícii skoro okamžite.

Najväčšia limitácia algoritmu na detekciu odľahlých hodnôt je jednoznačne mazanie záznamov pacientov, ak sa v nich vyskytuje čo i len jedna prázdna hodnota premennej. Priemerne sa zmazalo okolo 20% záznamov pacientov na formulár. V niektorých extrémnych prípadoch sa nevymazal ani jeden záznam pacienta, zatiaľ čo vo formulároch testovaných štúdií, ktoré obsahovali veľký počet premenných sa zmazalo až 90% záznamov pacientov. Pravdepodobne sa však jednalo o nepovinne vyplnený formulár, čo by vysvetľovalo vysoký počet nevyplnených polí. Kombinácia prázdnych hodnôt a veľkého počtu premenných tak zmazala aj dáta vhodné na analýzu. Tento nedostatok by mohla odstrániť jedna z doplňovacích metód. Napríklad doplnenie prázdnych polí priemerom z hodnôt premenných, ktoré sú k dispozícii, alebo metóda mnohonásobného regresného modelu na vzorky bez chýbajúcich hodnôt.

Počet umelo generovaných odľahlých hodnôt v tabuľkách 5,7 a 9 predstavuje 10 vysoko odľahlých, 30 stredne odľahlých a 50 nízko odľahlých záznamov. Keďže pozície odľahlých záznamov boli generované náhodne, mohlo sa stať, že niektoré pozície sa navzájom prekryli. Je teda možné, že vysoko odľahlý záznam prekryl nízko odľahlý záznam, čo môže spolu so samotnou nepresnosťou metódy vysvetľovať rozdiely medzi počtom umelo generovaných a detekovaných odľahlých hodnôt.

Navrhnutá metóda dosahuje slušné výsledky pre tri zo štyroch ukazovateľov úspešnosti algoritmu zobrazené pre každú štúdiu v tabuľkách 6,8 a 10. Jediný parameter, ktorý je určite potreba zlepšiť je senzitivita metódy. Pri vývoji algoritmu bola úspešnosť umelého generovania odľahlých hodnôt najprv testovaná resubstitúciou. Hodnoty úspešnosti vychádzali veľmi uspokojivo s hodnotami specificity aj senzitivity okolo 98% pre každú z testovaných štúdií. To sa však zmenilo pri použití lepšieho prístupu na hodnotenie úspešnosti algoritmu, metódy leave–one–out. Tu sa ukázalo, ako resubstitúcia nadhodnocuje výsledky, pretože hodnoty senzitivity prudko klesli na 45% – 26%.

6 ZÁVER

Pri vyplňovaní formulárov sú často údaje o pacientoch zadané nepresne, alebo sú niektoré polia vynechané. To spôsobuje nepresnosti pri výpočtoch, čo môže viesť až ku nesprávnym záverečným výsledkom klinickej štúdie. Preto je dôležité týmto hodnotám prikladať zvýšenú pozornosť. Systematické anomálie vznikajú dôsledkom chýb zberného programu, nejasne definovaných parametrov, alebo zneužitia zberu dát. Náhodné chyby vznikajú nepresným prepisom dát, alebo preklepmi pri ich zadávaní. Navrhnutá metóda slúži na zistenie odľahlých záznamov pacientov v databáze klinických štúdii. Odhalí chyby zberného programu, chyby vzniknuté z nepozornosti a vďaka viacrozmernému prístupu aj cielené zadávanie fiktívnych hodnôt. Tieto hodnoty sú veľmi ťažko odhaliteľné, pretože sú vo fyziologickom rozmedzí.

Z výslednej správy s pozíciami odľahlých záznamov v klinickej štúdii môže oprávnená osoba pre zvolený záznam pacienta následne určiť, či ide o valídnu hodnotu, alebo je hodnota nesprávna a bude musieť byť pre ďalšiu analýzu upravená, či vymazaná. Ak by sa jednalo o chybu pri zadávaní, alebo prepise dát môže sa hodnota jednoducho opraviť a odpovedajúci záznam pacienta bude ponechaný v klinickej štúdii. Tento proces tak pomáha zvýšiť celkovú kvalitu zozbieraných dát.

Diplomová práca je delená do šiestich častí. V prvej je rozpísaná teória a pozadie klinických štúdií. V druhej je rozpísaná kvalita dát, metódy detekcie odľahlých hodnôt, metriky podobnosti a vzdialenosti, a hodnotenie úspešnosti klasifikácie. V tretej je popis navrhnutej metódy riešenia. V štvrtej sú obsiahnuté získané výsledky, za ktorou nasleduje piata časť s diskusiou, celkovým zhodnotením výsledkov a limitáciami navrhnutej metódy. V poslednej časti je celková práca zhodnotená ako záver.

LITERATURA

[1] Švihálková H: Aplikace shlukovacích metod na data klinických registrů [online]. Brno, 2011. Dostupné z: <http://is.muni.cz/th/208192/prif_m/>. Diplomová práce. Masarykova univerzita, Přírodovědecká fakulta. Vedoucí práce Daniel Klimeš.

[2] SVOBODNÍK, Adam, Regina DEMLOVÁ a Ladislav PECEN. *Klinické studie v praxi*. Brno: Facta Medica, 2014. ISBN 978-80-904731-8-8.

[3] Park BK, Boobis A, Clark S et al. *Managing the challange of chemically reactive metabolities in drug development*. Nat Rev Drug Discov 2011; 10(4): 292-306.

[4] Chodera JD, Mobley DL, Shirts MR et al. *Alchemical free energy methods for drug discovery: progress and chalanges*. Curr Opin Struct Biol 2011: 21(2): 150-160.

[5] Malaroye A., Reinventing clinical trails. Nature Biotechnology 2012; 30: 41-49.

[6] Mullard A. 2010 FDA drug approvals. Nat Rev Drug Discov 2011; 10(2): 82-85.

[7] DiMasi JA, Feldman L, Seckler A et al. Trends in risks associated with new drug developments: success rates for investigational drugs. Clin Pharmacol Ther 2010; 87(3): 272-277.

[8] Morgan S, Grootensdorst P, Lexchin J et al. The cost of drug development: a systematic review. et al. The cost of drug development: systematic review. Health Policy 2011;100(1): 4-17.

[9] Park BK, Boobis A, Clarke S et al. Managing the challenge of chemically reactive metabolities in drug development. Nat Rev Drug Discov. 2011; 10(4): 292-306. Dostupné z DOI: <http://doi:10.2038/nrd3408>

[10] Quality Assurance and Educational Standards for Clinical Trial Sites. J Oncol Practise 2008; 4(6): 280-282 Dostupné z WWW: <http://jop.ascopubs.org/content/4/6/280-full?sidpa9e414cd-8327-47cc-87aa-89959172e375.

[11] Rondel RK, Varley SA, Webb CF. Clinical data management. 2nd ed. Viley 2000. ISBN 0470-84636-4.

[12] McFadden E. Management of data in clinical trials. 2nd ed. Wiley: New York 2007. ISBN 978-0-470-04608-1.

[13] Rondel RK, Varley SA, Webb CF. Clinical data management. 2nd ed. Viley 2000. ISBN 0470-84636-4.

[14] ZVÁRA, Karel. Biostatistika. 2. vyd. Praha: Karolinum, 2004. ISBN 978-80-246-0739-9.

[15] LITTNEROVÁ, Simona. Mnohorozměrné statistické metody v hodnocení interakcí biologických společenstev a prostředí. Brno, 2008. Bakalářská práce. Masarykova univerzita, Fakulta přírodovědecká, Výzkumné centrum pro chemii životního prostředí a ekotoxikologii, Institut biostatistiky a analýz.

[16] M. ROSS, Stephen. Peirce's criterion for the elimination of suspect experimental data. Journal of Engineering Technology [online]. [cit. 2017-12-07]. Dostupné z: https://classes.engineering.wustl.edu/che473/handouts/OutlierRejection.pdf

[17] Forbelská, M.: Lineární statistické modely I. Skripta. Masarykova univerzita. Brno.

[18] R. Kohavi and F. Provost, \Glossary of terms," Machine Learning, vol. 30, no. 2-3, pp. 271-274, 1998.

[19] Portál matematická biologie [online]. [cit. 2017-12-07]. Dostupné z: http://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--vicerozmerne-metody-pro-analyzu-dat--volba-a-vyber-popisnych-promennych--extrakce-promennych-1

[20] Haruštiaková, D., Jarkovský, J., Littnerová, S., Dušek, L. Vícerozměrné statistické metody v biologii. Akademické nakladatelství CERM, s.r.o., Brno. (2012)

[21] Jolliffe, I.T. Principal Component Analysis. Springer, New York (2002).

[22] Software Carpentry advanced-numpy-lesson. Https://software-carpentry.org/ [online]. [cit. 2018-05-16]. Dostupné z: http://paris-swc.github.io/advanced-numpy-lesson/05-kmeans.html

[23] HAN, Jiawei. a Micheline. KAMBER. *Data mining: concepts and techniques*. San Francisco: Morgan Kaufmann Publishers, 2001. ISBN isbn1-55860-489-8.

[24] HEBELKA, Tomáš. Analýza dat z mikročipu pro zjišťování genové exprese [online]. Brno, 2010 [cit. 2018-05-16]. Dostupné z: https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=117320. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií, Ústav informačních systému.

[25] LUTINS, Evan. Medium. *DBSCAN: What is it? When to Use it? How to use it.* [online]. [cit. 2018-05-16]. Dostupné z: https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818

[26] *Analýza a hodnocení dat* [online]. [cit. 2018-05-16]. Dostupné z: https://is.muni.cz/www/98951/41610771/43823411/43823458/Analyza_a_hodnoc/44563155/56049312/Vicerozmerky_-_kap4_-_metriky_final.pdf

[27] Mathworks Documentation [online]. [cit. 2018-05-16]. Dostupné z: https://www.mathworks.com/help/stats/pdist.html

[28] HARUŠTIAKOVÁ, Danka, Jiří JARKOVSKÝ, Simona LITTNEROVÁ a Ladislav DUŠEK. *Vícerozměrné statistické metody v biologii* [online]. Brno: IBA MU, 2012 [cit. 2018-05-16]. Dostupné z: https://www.iba.muni.cz/res/file/ucebnice/jarkovsky-vicerozmerne-statisticke-metody.pdf

[29] Kuncheva, L.I. Combining Classifiers: Methods and Algorithms. John Wiley & Sons, Hoboken, New Jersey (2004).

[30] Bishop, C. Pattern Recognition and Machine Learning. Springer, New York. (2006)

[31] Kuncheva, L.I. Combining Classifiers: Methods and Algorithms. John Wiley & Sons, Hoboken, New Jersey (2004).

[32] BROŽOVÁ, Lucie, Daniel SCHWARZ, Ivo ŠNÁBL, et al. Czech Registry of Monoclonal Gammopathies – Technical Solution, Data Collection and Visualisation. Klinicka Onkologie [online]. 2017, 30(Suppl 2), 2S43-2S50 [cit. 2017-12-06]. DOI: 10.14735/amko20172S43. ISSN 0862495x. Dostupné z: https://www.linkos.cz/klinicka-onkologie-journal/search-for-articles/skupina/a/zobrazit/ids/5215/

[33] Mathworks Documentation. PostgreSQL JDBC for Windows [online]. [cit. 2018-05-16]. Dostupné z: https://www.mathworks.com/help/database/ug/postgresql-jdbc-windows.html

[34] Holčík, J. Analýza a klasifikace dat. Akademické nakladatelství CERM, s.r.o., Brno. (2012)

[35] VLČKOVÁ, Kateřina. Škálovaní [online], 4 [cit. 2017-12-06]. Dostupné z: https://is.muni.cz/el/1441/podzim2006/SZ2BP_ZPM/um/um/skalovani.pdf

[36] GEORGE, Stephen L a Marc BUYSE. Data fraud in clinical trials. Clinical Investigation [online]. 2015, 5(2), 161-173 [cit. 2017-12-07]. DOI: 10.4155/cli.14.116. ISSN 2041-6792. Dostupné z: http://www.future-science.com/doi/abs/10.4155/cli.14.116

[37] HOLČÍK J. Analýza a klasifikace signálů. Vysoké učení technické v Brně, 1992.

ZOZNAM POUŽITÝCH SKRATIEK

EMA	European Medicine Agency, európska lieková agentúra.
FDA	US Food and Drug Administration, úrad pre kontrolu potravín a liečiv.
SÚKL	Štátny ústav pre kontrolu liečiv
QA/QC	Quality Assurance/Quality Controll, zabezpečovanie a kontrola akosti
SOP	Standard Operation procedure, štandardné operačné procesy
CRF	Case Report Form, chorobopis
eCRF	Electronic Case Report Form, elektronický chorobopis
EDC	Electronic Data Capture, elektronický zber dát
BMI	Body Mass Index, index telesnej hmotnosti
WHO	World Health Organization, svetová zdravotnícka organizácia
DMS	Data Management Systém, systém managementu dát
GDP	Good Clinical Practice, dobrá klinická prax

PRÍLOHY

Vybrané časti navrhnutého algoritmu v programovom prostredí Matlab.

```
% transformacia premmenej z datumoveho formatu na pocet dni od 0.januara roku 0


for w=1:size(pre_form,2) % prechadza premenne formularu

            try % najde premennu s datumom a transformuje ju vo formate yyyy-MM-dd
hh:mm:ss

                temp=char(pre_form(:,w)); % nacita stlpec premennej s datumom

                date_col=datenum(temp,'yyyy-MM-dd hh:mm:ss'); % transformacia na
numericku hodnotu pre format datumu yyyy-MM-dd hh:mm:ss

                dd=num2cell(date_col); % transformcia pre lepsie ulozenie

                for k=1:size(pre_form,1) % pre odpovedajuce pozicie

                    pre_form(k,w)=dd(k,1); % prepise hodnoty datumu v povodnom formulari

                end

                disp(['Premenna s datumom vo formate yyyy-MM-dd hh:mm:ss bola
transformovana']); % ohlasi hlasku o uspesnej trasnformacii

            catch % ak nenajde premennu s datumom

                disp(['Premenna s datumom nie je vo formate yyyy-MM-dd hh:mm:ss']); %
ohlasi hlasku o neuspesnej transformacii

            end

            try % najde premennu s datumom a transformuje ju vo formate yyyy-MM-dd


                temp=char(pre_form(:,w)); % nacita stlpec premennej s datumom

                date_col=datenum(temp,'yyyy-MM-dd'); % transformacia na numericku
hodnotu pre format datumu yyyy-MM-dd

                dd=num2cell(date_col); % transformcia pre lepsie ulozenie


                for k=1:size(pre_form,1) % pre odpovedajuce pozicie

                    pre_form(k,w)=dd(k,1); % prepise hodnoty datumu v povodnom formulari

                end

                disp(['Premenna s datumom vo formate yyyy-MM-dd bola transformovana']);
% ohlasi hlasku o uspesnej trasnformacii


            catch % ak nenajde premennu s datumom

                disp(['Premenna s datumom nie je vo formate yyyy-MM-dd']); % ohlasi
hlasku o neuspesnej transformacii


            end

        end
```

```
% generovanie umelych odlahlych hodnot na cistom fomulari, na vstupe je formular so
standardizovanymi hodnotami

for i=1:size(form4,2) % prechadza premenne
    r = randi([1 size(form4,2)],1,24); % enereuje 24 nahodnych pozicii premennych

    step_high = round(size(form4,1)/ 10); % vygeneruje 11 outlierov pac. zaznamov

    for j=1:step_high:size(form4,1) % pre vysoko odlahle zaznamy
        if isempty(high_fake_position_x) % ak je matica s poziciami prazdna

            high_fake_position_x = j; % ulozi sa pozicia vysoko odlahleho zaznamu

        else % ak nie je prazdna

            high_fake_position_x=[high_fake_position_x, j]; % ulozi sa pozicia na koniec
        end
        addition = (1-form4(high_fake_position_x,r))/1.5 ; % vyrata pridavok na zaklade
typu odlahlej hodnoty/ velka zmena

        fake_form(high_fake_position_x,r)= form4(high_fake_position_x,r) + addition; %
na danych poziciach vo formulari prirata pridavok
    end

    high_fake_position_x=unique(high_fake_position_x,'stable'); % kedze pozcie boli
generovane nahodne, mozu sa opakovat

    r = randi([1 size(form4,2)],1,20); % vygenereuje 20 pozicii kde sa prevedu zmeny

    step_medium = round(size(form4,1)/ 30); % vygeneruje 31 outlierov pac. zaznamov

    for j=1:step_medium:size(form4,1) % pre stredne odlahle zaznamy

        if isempty(medium_fake_position_x) % ak je matica s poziciami prazdna

            medium_fake_position_x = j; % ulozi sa pozicia stredne odlahleho zaznamu
        else % ak nie je prazdna

            medium_fake_position_x=[medium_fake_position_x, j]; % pozicia na koniec
        end
        addition = (1-form4(medium_fake_position_x,r))/1.8 ; % vyrata pridavok na
zaklade typu odlahlej hodnoty/ stredna zmena

        fake_form(medium_fake_position_x,r)= form4(medium_fake_position_x,r) + addition;
% na danych poziciach prirata pridavok
    end

    medium_fake_position_x=unique(medium_fake_position_x,'stable'); % kedze pozcie boli
generovane nahodne, mozu sa opakovat

    r = randi([1 size(form4,2)],1,15); % vygenereuje 15 pozicii kde sa prevedu zmeny

    step_low = round(size(form4,1)/ 50); % vygeneruje 51 outlierov pacientskych zaznamov

    for j=1:step_low:size(form4,1) % pre nizko odlahle zaznamy

        if isempty(low_fake_position_x) % ak je matica s poziciami prazdna

            low_fake_position_x = j; % ulozi sa pozicia nizko odlahleho zaznamu
        else % ak nie je prazdna

            low_fake_position_x=[low_fake_position_x, j]; % ulozi sa pozicia na koniec
        end

        addition = (1-form4(low_fake_position_x,r))/ 3 ; % vyrata pridavok na zaklade
typu outlier

        fake_form(low_fake_position_x,r)= form4(low_fake_position_x,r) + addition; % na
danych poziciach prirata pridavok
    end

    low_fake_position_x=unique(low_fake_position_x,'stable'); % kedze pozcie boli
generovane nahodne, mozu sa opakovat
end
```

```
% vypocet matice zamen

% matica detected_outliers ma rozmery (pocet riadkov testovaneho formulara, 1)
a obsahuje 1 v pripade detekovanej odlahlej hodnoty, inak 0 na odpovedajucej pozicii

% matica fake_outliers ma rozmery (pocet riadkov testovaneho formulara, 1) a obsahuje 1
v pripade vygenerovanej odlahlej hodnoty, inak 0 na odpovedajucej pozicii

% TP = true positive
% TN = true negative
% FP = false positive
% FN = false negative


TP=0; TN=0; FP=0; FN=0;

for i=1:size(detected_outliers,1) % prechadza riadky matice

    if detected_outliers(i,1) ==1 && fake_outliers(i,1)==1 % ak na danom riadku bola
detekovana odlahla hodnota a zaroven na odpovedajucej pozicii aj umelo generovana

        TP = TP+1; % hodnota true positive sa navysi o 1

    elseif detected_outliers(i,1) ==0 && fake_outliers(i,1)==0 % ak na danom riadku
nebola detekovana odlahla hod. a zaroven na odpovedajucej pozicii ani umelo generovana

        TN= TN+1; % hodnota true negative sa navysi o 1

    elseif detected_outliers(i,1) ==1 && fake_outliers(i,1)==0 % ak na danom riadku bola
detekovana odlahla hodnota, ale na odpovedajucej pozicii nebola umelo generovana

        FP = FP+1; % hodnota false positive sa navysi o 1

    elseif detected_outliers(i,1) ==0 && fake_outliers(i,1)==1 % ak na danom riadku
nebola detekovana odlahla hod. a zaroven na odpovedajucej pozicii bola umelo generovana


        FN = FN+1; % hodnota false negative sa navysi o 1

    end

end

accuracy = (TP+TN)/ (TP+TN+FP+FN); % presnost metody
error = (FP+FN)/ (TP+TN+FP+FN); % chyba metody
specificity = TN /(TN + FP); % specificita metody
sensitivity = TP / (TP + FN); % senzitivita metody
```

```
% spajanie viacerych formularov do jedneho cez primarny a sekundarny kluc, formulare
klinickej studie su ulozene v zozname, spajanie je urcene pre pocitanie prazdnych poli

for i=1:size(study1_zoznam,1) % prebehne pre kazdy formular v zozname

    if i==1 % ak sa jedna o prvy formular

        spolu_study1=current_view.temp; % formular sa nacita sa do vyslednej matice

    else % pre ine ako prvy formular

        current_view=current_view.temp; % nacita sa do pomocnej premennej

        [x,y]=size(spolu_study1); % zistenie aktualnej velkosti vyslednej matice

        y1=y+1; % y pozicia pre novy zapis

        x1=x+1; % x pozicia pre novy zapis

        for j=1:size(current_view,1) % prechadza druhy formular

            idx = find(strcmp(spolu_study1(:,1),current_view(j,1))); % najde index
primarneho klucu z druheho formulara v prvom formulari

            if isempty(idx)  % nenasiel idx primarneho klucu z druheho formulara v prvom

                spolu_study1(x1,1)= current_view(j,1); % prida sa primarny kluc na
posledne miesto vo vyslednej matici

                spolu_study1(x1,2)= current_view(j,2); % prida sa novy sekundarny kluc
na posledne miesto vo vyslednej matici

                for k=3:size(current_view(j,:),2) % prechadza premenne druheho formulara

                    spolu_study1(x1,y1)= current_view(j,k); % k primarnemu
a sekundarnemu klucu z druheho formulara pripise odpovedajuce data do vyslednej matice

                    y1=y1+1; % zapocita sa nova y pozicia

                end

                x1=x1+1; % zapocita sa nova x pozicia

            elseif isequal(spolu_study1(idx(1),1), current_view(j,1)) &&
isequal(spolu_study1(idx(1),2), current_view(j,2)) % ak sa rovnaju primarny a sekundarny
kluc z druheho a prveho formulara

                for k=3:size(current_view(j,:),2) % prechadza premenne z 2. formulara

                    spolu_study1(idx(1),y1)= current_view(j,k); % k primarnemu
a sekundarnemu klucu z druheho formulara pripise odpovedajuce data do vyslednej matice

                    y1=y1+1; % zapocita sa nova y pozicia

                end

            end

            y1=y+1; % zapocita sa nova x pozicia

        end

    end

end
```