**ŽILINSKÁ UNIVERZITA V ŽILINE**

FAKULTA RIADENIA A INFORMATIKY

Obrázok, na ktorom je text, písmo, čierny, dizajn

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

**Softvérové spracovanie biomedicínskych údajov**

**Študijný program: Biomedicínska informatika**

Semestrálna práca

Vypracovali: Michal Bežo Akademický rok: 2024/2025

Filip Halimovič Študijná skupina: 5ZIB11

Kevin Laurinčík

Radoslav Žilka

Obsah

[Úvod 4](#_Toc198486173)

[Zadanie semestrálnej práce 4](#_Toc198486174)

[Cieľ a úlohy projektu 4](#_Toc198486175)

[Technické požiadavky 5](#_Toc198486176)

[Pochopenie údajov a problému 5](#_Toc198486177)

[Hereditárna hemochromatóza(medicínske pozadie) 5](#_Toc198486178)

[Genetické mutácie HFE génu 5](#_Toc198486179)

[Relevancia pre biomedicínsku informatiku 5](#_Toc198486180)

[Dataset 6](#_Toc198486181)

[Príprava a predspracovanie údajov 6](#_Toc198486182)

[Načítanie a počiatočná kontrola údajov 6](#_Toc198486183)

[Čistenie a štandardizácia údajov 6](#_Toc198486184)

[Kategorizácia premenných 7](#_Toc198486185)

[Príprava údajov na analýzu 7](#_Toc198486186)

[Použité metódy a techniky 8](#_Toc198486187)

[Štatistická analýza 8](#_Toc198486188)

[Hardy-Weinbergova rovnováha (HWE) 8](#_Toc198486189)

[Percentuálna distribúcia genotypov 8](#_Toc198486190)

[Korelačná analýza medzi mutáciami a diagnózami 8](#_Toc198486191)

[Analýza diagnóz podľa MKCH-10 9](#_Toc198486192)

[Vizualizačné techniky 9](#_Toc198486193)

[Softvérové a programové nástroje 9](#_Toc198486194)

[Reprodukovateľnosť a modularita 9](#_Toc198486195)

[Návrh a popis prototypu riešenia 10](#_Toc198486196)

[Architektúra riešenia 10](#_Toc198486197)

[Spracovanie vstupných údajov 11](#_Toc198486198)

[Používateľské rozhranie 11](#_Toc198486199)

[Analýza výsledkov 12](#_Toc198486200)

[Hardy-Weinbergova rovnováha 12](#_Toc198486201)

[Genotypové zastúpenie 13](#_Toc198486202)

[Závislosť mutácií a diagnóz 13](#_Toc198486203)

[Genotypové grafy podľa veku a pohlavia 14](#_Toc198486204)

[Vývoj typu diagnóz počas rokov 15](#_Toc198486205)

[Vyhodnotenie výsledkov 16](#_Toc198486206)

[Záver 18](#_Toc198486207)

**Zoznam obrázkov**

[Obrázok 1: HW\_Test\_C282Y 12](#_Toc198486158)

[Obrázok 2: HW\_Test\_H63D 12](#_Toc198486159)

[Obrázok 3: HW\_Test\_S65C 12](#_Toc198486160)

[Obrázok 4: Percentuálne zastúpenie genotypov – C282Y 13](#_Toc198486161)

[Obrázok 5: Percentuálne zastúpenie genotypov – H63D 13](#_Toc198486162)

[Obrázok 6: Percentuálne zastúpenie genotypov – S65C 13](#_Toc198486163)

[Obrázok 7: Korelačná matica mutácií a pečeňového ochorenia 13](#_Toc198486164)

[Obrázok 8: Chi\_kvadrát\_test korelácie 14](#_Toc198486165)

[Obrázok 9: Genotyp\_C282Y\_Vek 14](#_Toc198486166)

[Obrázok 10: Genotyp\_H63D\_Vek 14](#_Toc198486167)

[Obrázok 11: Genotyp\_S65C\_Vek 15](#_Toc198486168)

[Obrázok 12: Genotyp\_C282Y\_Pohlavie 15](#_Toc198486169)

[Obrázok 13: Genotyp\_H63D\_Pohlavie 15](#_Toc198486170)

[Obrázok 14: Genotyp\_S65C\_Pohlavie 15](#_Toc198486171)

[Obrázok 15: Typy diagnóz počas rokov 15](#_Toc198486172)

# Úvod

Moderná medicína čoraz viac využíva metódy výpočtového spracovania dát na podporu diagnostiky, predikcie a personalizovanej liečby. Genetické údaje patria medzi najkomplexnejšie a najcitlivejšie typy medicínskych informácií, preto je ich správna analýza kľúčová pre identifikáciu rizikových faktorov rôznych ochorení, vrátane tých dedičných.

V rámci tejto semestrálnej práce sa zaoberáme spracovaním genetických údajov týkajúcich sa mutácií génu HFE, ktoré sú úzko späté s dedičnou formou hemochromatózy – autozomálne recesívnym ochorením spôsobujúcim zvýšenú absorpciu železa a jeho ukladanie do rôznych orgánov, predovšetkým pečene. Cieľom práce je vytvoriť dátový a softvérový prototyp, ktorý umožní analyzovať a predikovať genetické riziko pacienta na základe jeho genotypu, veku, pohlavia a klinickej diagnózy.

Projekt si kladie za cieľ:

* očistiť a spracovať reálne genetické údaje z dodaného datasetu
* analyzovať výskyt mutácií a ich súvislosti s demografickými a klinickými údajmi
* vizualizovať získané informácie pre jednoduché porozumenie
* navrhnúť a vyvinúť funkčný prototyp aplikácie pre genetickú predikciu hereditárnej hemochromatózy

# Zadanie semestrálnej práce

Semestrálna pozostáva zo spracovania reálneho problému formou prípadovej štúdie a návrhu softvérového prototypu. Zadanie je tematicky zamerané na analýzu genetických mutácií génu HFE, ktorý zohráva významnú úlohu pri vzniku dedičnej hemochromatózy (hereditary hemochromatosis, HH).

## Cieľ a úlohy projektu

**Primárny cieľ**

Vytvoriť nástroj na vyhodnotenie pravdepodobnosti genetického rizika hereditárnej hemochromatózy na základe genetických údajov o mutáciách génu HFE a ich súvislostí s inými premennými (vek, pohlavie, diagnózy).

**Kľúčové úlohy**

* Predspracovanie údajov: Detekcia a riešenie chýbajúcich alebo neštandardných údajov v datasete. Zabezpečiť konzistentné označenie genotypov a odstrániť záznamy, ktoré môžu narušiť analýzu.
* Overenie Hardy-Weinbergovej rovnováhy: Overiť, či rozdelenie genotypov pre jednotlivé mutácie (C282Y, H63D, S65C) spĺňa genetický predpoklad rovnováhy v populácii.
* Analýza genotypov: Vyhodnotiť výskyt jednotlivých genotypov, zistiť počet prenášačov a pacientov s genetickou predispozíciou.
* Vzťah k diagnózam: Skúmať korelácie medzi výskytom mutácií a klinickými diagnózami (najmä ochorenia pečene – K76.0, K75.9).
* Vizuálna analýza: Vytvoriť grafy znázorňujúce rozdelenie genotypov a ich vzťah k veku, pohlaviu a diagnózam podľa klasifikácie MKCH-10.
* Analýza vývoja v čase: Preskúmať časové trendy v diagnostike na základe roku vyšetrenia a aktuálnosti diagnóz podľa najnovších štandardov MKCH-10 (WHO/NCZI).
* Vývoj softvérového prototypu: Navrhnúť a vytvoriť prototyp softvéru, ktorý umožní nahratie údajov, ich spracovanie a vizualizáciu výsledkov.
* Prezentácia a dokumentácia: Zostavenie odbornej dokumentácie, ktorá popisuje celý proces riešenia, a príprava prezentácie funkčného prototypu.

### Technické požiadavky

* Implementácia môže byť realizovaná v ľubovoľnom programovacom jazyku (preferenčne Python).
* Vizualizácie údajov musia byť zrozumiteľné a informatívne.
* Práca musí byť vypracovaná skupinovo (do 4 členov), pričom je potrebné dodržať termíny nahlásenia skupiny, odovzdania a obhajoby.

***Bonusové hodnotenie:***

Za aktívne využívanie systému Git, dokumentovanie vývoja prostredníctvom commitov a spoluprácu členov prostredníctvom repozitára je možné získať až 3 bonusové body.

# Pochopenie údajov a problému

## Hereditárna hemochromatóza(medicínske pozadie)

Hereditárna hemochromatóza (HH) je geneticky podmienené ochorenie spôsobené narušenou reguláciou vstrebávania železa v tenkom čreve. Ide o najčastejšiu autozomálne recesívnu poruchu metabolizmu železa v európskej populácii, pričom najvýznamnejšie mutácie sa vyskytujú v géne HFE, konkrétne mutácie C282Y, H63D a S65C. V dôsledku týchto mutácií dochádza k zvýšenej absorpcii železa, čo vedie k jeho hromadeniu v orgánoch ako sú pečeň, srdce a pankreas. Následkom môže byť cirhóza pečene, kardiomyopatia, diabetes mellitus typu 2 a iné závažné stavy.

Klinicky sa ochorenie prejavuje zvyčajne až v dospelosti, častejšie u mužov. Včasná detekcia na základe genetickej predispozície však umožňuje preventívne opatrenia a monitorovanie pacienta.

## Genetické mutácie HFE génu

* **C282Y:** Najčastejšia a najvýznamnejšia mutácia. Homozygotné jedince majú vysoké riziko vzniku HH.
* **H63D:** Menej riziková, ale jej kombinácia s C282Y (zložený heterozygot) zvyšuje pravdepodobnosť ochorenia.
* **S65C:** Raritnejšia mutácia, no môže zvyšovať riziko v kombinácii s ostatnými.

**Genotyp môže byť:**

1. WT (wild type) – bez mutácie,
2. heterozygotný – jedna alela nesie mutáciu,
3. homozygotný – obe alely sú mutované,
4. zložený heterozygot – kombinácia dvoch rôznych mutácií na rôznych alelách.

## Relevancia pre biomedicínsku informatiku

Spracovanie genetických údajov a ich prepojenie s demografickými a klinickými informáciami má potenciál zlepšiť:

1. diagnostiku ochorení,
2. stratifikáciu pacientov podľa rizika,
3. personalizované preventívne stratégie.

V tejto práci je ambíciou vytvoriť dátovo-informačný model, ktorý na základe týchto údajov určí pacientov s genetickou predispozíciou a umožní klinicky lepšie rozhodovanie.

## Dataset

**Dataset SSBU25\_dataset.xls obsahuje:**

* Výsledky genetických testov pre mutácie C282Y, H63D a S65C
* Vek pacienta,
* Pohlavie pacienta,
* Diagnózy podľa klasifikácie MKCH-10,
* Rok vyšetrenia alebo záznamu.

**Tieto údaje budú analyzované s cieľom:**

1. Získať prehľad o rozložení genotypov v populácii,
2. Identifikovať skupiny s genetickým rizikom,
3. Zistiť, či existuje štatisticky významná súvislosť medzi genotypom a výskytom špecifických diagnóz.

# Príprava a predspracovanie údajov

Pred samotnou analýzou genetických údajov je nevyhnutné zabezpečiť ich kvalitu a konzistenciu. Tento proces zahŕňa čistenie dát, štandardizáciu formátov, kategorizáciu premenných a prípravu údajov na ďalšie spracovanie.

## Načítanie a počiatočná kontrola údajov

Dataset bol poskytnutý vo formáte Excel (.xls) a obsahuje informácie o genetických mutáciách génu HFE, demografické údaje (vek, pohlavie) a klinické diagnózy pacientov. Na načítanie údajov bol použitý nástroj pandas, ktorý umožňuje efektívne spracovanie tabuľkových dát.

Po načítaní datasetu bola vykonaná počiatočná kontrola údajov s cieľom identifikovať chýbajúce hodnoty, neštandardné formáty a potenciálne nekonzistencie. Zistilo sa, že niektoré záznamy obsahujú neúplné alebo nesprávne údaje, ktoré bolo potrebné upraviť alebo odstrániť.

## Čistenie a štandardizácia údajov

**Čistenie údajov zahŕňalo nasledujúce kroky:**

1. Odstránenie chýbajúcich alebo neúplných záznamov: Záznamy s chýbajúcimi hodnotami v kľúčových premenných (napr. genotyp, vek, pohlavie) boli odstránené, aby sa zabezpečila integrita analýzy.
2. Štandardizácia formátov: Hodnoty v stĺpcoch boli preformátované do jednotného formátu (napr. pohlavie bolo kódované ako 'M' pre mužov a 'F' pre ženy).
3. Korekcia neštandardných zápisov: Niektoré genotypy boli zapísané rôznymi spôsobmi (napr. 'C282Y homozygous', 'Homozygote C282Y'). Tieto zápisy boli zjednotené na štandardné označenia (napr. 'C282Y/C282Y').

## Kategorizácia premenných

Pre účely analýzy boli premenné kategorizované nasledovne:

**Vek:** Pacienti boli rozdelení do vekových kategórií (napr. 0–19, 20–39, 40–59, 60+), aby sa umožnila analýza výskytu mutácií v rôznych vekových skupinách.

**Genotyp:**

Genotypy boli klasifikované do kategórií:

* WT/WT: bez mutácie
* Heterozygot: jedna mutovaná alela
* Homozygot: obe alely mutované
* Zložený heterozygot: kombinácia dvoch rôznych mutácií

**Diagnózy:** Diagnózy boli kódované podľa klasifikácie MKCH-10 a rozdelené do skupín podľa typu ochorenia (napr. ochorenia pečene, metabolické poruchy).

## Príprava údajov na analýzu

Po čistení a kategorizácii boli údaje pripravené na ďalšiu analýzu:

* **Overenie Hardy-Weinbergovej rovnováhy:** Pre každú mutáciu bola vypočítaná frekvencia alel a genotypov, aby sa overila genetická rovnováha v populácii.
* **Analýza výskytu mutácií:** Bola vykonaná analýza výskytu jednotlivých genotypov v populácii a ich rozdelenie podľa veku, pohlavia a diagnóz.
* **Vizualizácia údajov:** Pripravené údaje boli použité na vytvorenie grafov a tabuliek, ktoré znázorňujú rozdelenie genotypov a ich vzťah k demografickým a klinickým premenným.

Týmto spôsobom bola zabezpečená kvalita a konzistencia údajov, čo je nevyhnutné pre spoľahlivú analýzu a interpretáciu výsledkov. V nasledujúcej kapitole sa budeme venovať použitým metódam a technikám analýzy údajov.

# Použité metódy a techniky

Spracovanie genetických údajov a ich analýza si vyžaduje kombináciu bio informatických, štatistických a softvérových nástrojov. V tejto práci sme sa zamerali na overenie genetických hypotéz, výpočty frekvencií, testovanie genetickej rovnováhy a vizualizáciu dát za účelom ich biologickej a klinickej interpretácie.

## Štatistická analýza

### Hardy-Weinbergova rovnováha (HWE)

Jednou z kľúčových analytických metód bolo overenie Hardy-Weinbergovej rovnováhy pre každú zo skúmaných mutácií (C282Y, H63D, S65C). Táto rovnováha popisuje genetickú distribúciu genotypov v populácii pri predpoklade, že nedochádza k selekcii, migrácii, mutácii ani náhodným chybným faktorom.

Postup:

* Spočítanie absolútnych počtov genotypov pre každú mutáciu.
* Výpočet frekvencie alel (p, q).
* Očakávané frekvencie genotypov podľa HWE:
  + - p² pre homozygot „divokého typu“ (napr. WT/WT),
    - 2pq pre heterozygota,
    - q² pre homozygota s mutáciou.
* Porovnanie pozorovaných a očakávaných počtov pomocou χ² testu nezávislosti.

Tento test umožňuje overiť, či genetická distribúcia v populácii zodpovedá teoretickému predpokladu rovnováhy.

### Percentuálna distribúcia genotypov

Pre každú mutáciu sme vyčíslili:

* Percento nosičov (heterozygotov),
* Percento homozygotov (mutovaných),
* Percento jedincov bez mutácie (wild type),
* Počet pacientov so zloženými heterozygotmi.

Účelom bolo identifikovať skupiny pacientov s vysokým genetickým rizikom.

### Korelačná analýza medzi mutáciami a diagnózami

Použili sme jednoduché kontingenčné tabuľky a výpočet:

* Relatívneho rizika (RR) a
* Chi-kvadrát testu medzi výskytom mutácií a konkrétnymi diagnózami (napr. K76.0, K75.9).

Týmto spôsobom sme zisťovali, či existuje signifikantná súvislosť medzi výskytom HFE mutácií a ochoreniami pečene, prípadne inými metabolickými poruchami.

## Analýza diagnóz podľa MKCH-10

Pre každého pacienta boli diagnózy (v podobe kódov MKCH-10) analyzované a zoskupené do kategórií:

* Choroby pečene (napr. K76.0 – steatóza, K75.9 – chronická hepatitída),
* Metabolické poruchy (napr. E70–E90),
* Iné príznaky a abnormality (napr. R70 – abnormality krvného testu).

Zároveň bola analyzovaná časová os výskytu diagnóz, aby sa sledoval vývoj klasifikácie a frekvencie výskytu jednotlivých skupín ochorení. Pri kontrole aktuálnosti kódov boli využité oficiálne zdroje NCZI a WHO pre revíziu MKCH-10.

## Vizualizačné techniky

Na vytvorenie prehľadných a informačne bohatých vizualizácií boli použité nasledujúce knižnice v prostredí Python:

**Matplotlib** a **Seaborn**:

* Histogramy veku v kontexte mutácií,
* Boxploty na porovnanie distribúcie medzi skupinami (napr. muži vs. ženy),
* Heatmapy pre korelácie medzi genotypmi a diagnózami,
* Časové línie diagnóz podľa roku vyšetrenia.

**Plotly**: Interaktívne vizualizácie pre prezentáciu.

## Softvérové a programové nástroje

**Programovací jazyk: Python**

* Pandas – načítanie a spracovanie údajov,
* NumPy – výpočty frekvencií a štatistických parametrov,
* Scipy – štatistické testy,
* Matplotlib / Seaborn – vizualizácie.

**Vývojové prostredie:**

* Pycharm (vývoj prototypu),
* GitHub – verziovanie a tímová spolupráca (repozitár: <https://github.com/KevinLaurincik/SSBU_Sem> ).

## Reprodukovateľnosť a modularita

Kód je organizovaný do modulov a funkcií, ktoré umožňujú:

* Znovu použitie spracovateľských funkcií,
* Jednoduché testovanie častí aplikácie,
* Zmeny vstupných dát bez potreby zásahu do logiky systému.

Týmto sa uzatvára prehľad analytických metód, ktoré boli použité na spracovanie genetických a klinických údajov. V nasledujúcej kapitole popíšeme návrh a architektúru samotného softvérového prototypu.

# Návrh a popis prototypu riešenia

Softvérový prototyp vyvinutý v rámci tejto práce slúži ako nástroj na spracovanie genetických údajov, vizualizáciu ich rozloženia a analýzu rizika rozvoja hereditárnej hemochromatózy (HH). Cieľom návrhu bolo vytvoriť modulárnu, intuitívnu a jednoducho rozšíriteľnú aplikáciu, ktorá umožní:

* načítanie dát zo vstupného Excel súboru,
* čistenie a predspracovanie dát,
* analýzu frekvencie výskytu genotypov,
* testovanie Hardy-Weinbergovej rovnováhy,
* vizualizáciu údajov (vek, pohlavie, diagnózy vs. mutácie),
* export výsledkov do CSV a grafickej podoby.

## Architektúra riešenia

Softvérová aplikácia vyvinutá v jazyku Python, zameraná na analýzu biomedicínskych dát. Jej architektúra je modulárna a zameraná na spracovanie genetických a diagnostických údajov, vrátane vizualizácie výsledkov.

**Logická vrstva**

| **Vrstva** | **Funkcia** |
| --- | --- |
| Prezentačná (GUI) | Interakcia s používateľom |
| Aplikačná | Riadenie toku údajov, spracovanie výsledkov, zber vstupov |
| Dátová (core) | Čistenie dát, výpočty HWE, výpis štatistík a generovanie grafov |

**Analytické moduly**

Tieto moduly vykonávajú štatistické a genetické analýzy:

* **CorrelationAnalysis.py**: Analyzuje koreláciu medzi rôznymi premennými v dátach.
* **DiagnosisAnalysis.py**: Spracováva a analyzuje diagnostické údaje pacientov.
* **GenotypeAnalysis.py**: Zameriava sa na analýzu genotypových dát.
* **HardyWeinbergTest.py**: Vykonáva test rovnováhy podľa Hardy-Weinbergovho princípu.

**Vizualizačný modul**

* **Plotter.py**: Zodpovedá za generovanie grafov a vizualizácií výsledkov analýz.

**Používateľské rozhranie**

* **app\_ui.py**: Definuje grafické používateľské rozhranie aplikácie.
* **app.py**: Obsahuje logiku aplikácie a spája používateľské rozhranie s analytickými modulmi.

**Hlavný spúšťací skript**

* **main.py**: Inicializuje aplikáciu a spúšťa hlavný proces.

**Dátový súbor**

* **SSBU25\_dataset.xls**: Obsahuje biomedicínske údaje, ktoré sú predmetom analýzy.

## Spracovanie vstupných údajov

Funkcia DataLoader načíta a skontroluje vstupný .xls alebo .xlsx súbor. Po úspešnom načítaní sa údaje:

* vyčistia (odstránia chýbajúce alebo neštandardné hodnoty),
* kategorizujú (napr. vekové skupiny, genotypy),
* pripravia na ďalšiu analýzu formou Pandas DataFrame.

Čistenie prebieha aj automaticky, pomocou funkcie clean\_dataset(), ktorá zabezpečí:

* zjednotenie formátu názvov mutácií (napr. C282Y/C282Y),
* odstránenie záznamov s neplatným kódom diagnózy alebo chýbajúcimi dátami,
* správne rozpoznanie genotypu podľa kombinácie mutácií.

## Používateľské rozhranie

Grafické rozhranie je navrhnuté pre jednoduché a intuitívne používanie. Po spustení aplikácie môže používateľ:

* vybrať vstupný súbor s údajmi,
* prehliadať a filtrovať údaje podľa typu mutácie,
* vybrať konkrétne analýzy alebo vizualizácie, ktoré chce spustiť,
* zobraziť výsledky v tabuľke alebo grafoch,

Používateľ má tiež možnosť aktivovať funkcie ako:

* Test HWE pre každú mutáciu (s výstupom do konzoly alebo CSV),
* Analýza rozdelenia genotypov podľa pohlavia alebo veku,
* Vyhľadanie pacientov s genetickou predispozíciou (napr. C282Y homozygot).

**Tok dát a interakcia modulov**

1. **Načítanie dát**: Aplikácia načíta údaje z SSBU25\_dataset.xls.
2. **Analýza**: Používateľ vyberie typ analýzy cez grafické rozhranie (app\_ui.py), ktoré následne spustí príslušný analytický modul.
3. **Vizualizácia**: Výsledky analýzy sú vizualizované pomocou Plotter.py.
4. **Zobrazenie výsledkov**: Vizualizácie a výsledky sú prezentované používateľovi prostredníctvom grafického rozhrania.

Vývoj prebiehal v repozitári SSBU\_Sem( <https://github.com/KevinLaurincik/SSBU_Sem> ), kde možno sledovať históriu zmien, jednotlivé commity a priebeh vývoja. Používanie Gitu umožnilo:

* verzionovanie jednotlivých funkčných modulov,
* sledovanie progresu tímu počas vývoja,
* jednoduché vrátenie k predchádzajúcim funkčným verziám,
* prípadné rozšírenie funkcionality (napr. o predikčný model pomocou ML).

# Analýza výsledkov

Cieľom tejto kapitoly je zhodnotiť výstupy softvérového prototypu z hľadiska spracovania genetických údajov a identifikácie súvislostí medzi HFE mutáciami a klinickými parametrami pacientov. Výsledky analýzy boli získané automatickým spracovaním údajov prostredníctvom navrhnutej aplikácie.

## Hardy-Weinbergova rovnováha

**C282Y**

**Obrázok, na ktorom je text, snímka obrazovky, písmo, číslo

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.**

Obrázok 1: HW\_Test\_C282Y

**H63D**

**Obrázok, na ktorom je text, snímka obrazovky, písmo, rad

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.**

Obrázok 2: HW\_Test\_H63D

**S65C**

Obrázok, na ktorom je text, písmo, snímka obrazovky

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 3: HW\_Test\_S65C

## Genotypové zastúpenie

**C282Y**

**Obrázok, na ktorom je text, štvorec, diagram, rad

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.**

Obrázok 4: Percentuálne zastúpenie genotypov – C282Y

**H63D**

**Obrázok, na ktorom je snímka obrazovky, text, rad, štvorec

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.**

Obrázok 5: Percentuálne zastúpenie genotypov – H63D

**S65C**

**Obrázok, na ktorom je text, vývoj, rad, snímka obrazovky

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.**

Obrázok 6: Percentuálne zastúpenie genotypov – S65C

## Závislosť mutácií a diagnóz

Obrázok, na ktorom je snímka obrazovky, štvorec, elektrická modrá, rad

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 7: Korelačná matica mutácií a pečeňového ochorenia

Obrázok, na ktorom je text, snímka obrazovky, písmo, číslo

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 8: Chi\_kvadrát\_test korelácie

## Genotypové grafy podľa veku a pohlavia

Obrázok, na ktorom je snímka obrazovky, rad, diagram, štvorec

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 9: Genotyp\_C282Y\_Vek

Obrázok, na ktorom je snímka obrazovky, rad, diagram, vývoj

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 10: Genotyp\_H63D\_Vek

Obrázok, na ktorom je snímka obrazovky, rad, diagram, text

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 11: Genotyp\_S65C\_Vek

Obrázok, na ktorom je rad, text, vývoj, snímka obrazovky

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 12: Genotyp\_C282Y\_Pohlavie

Obrázok, na ktorom je rad, text, vývoj, diagram

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 13: Genotyp\_H63D\_Pohlavie

Obrázok, na ktorom je rad, vývoj, snímka obrazovky

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 14: Genotyp\_S65C\_Pohlavie

## Vývoj typu diagnóz počas rokov

Obrázok, na ktorom je rad, vývoj, diagram, svah

Obsah vygenerovaný umelou inteligenciou môže byť nesprávny.

Obrázok 15: Typy diagnóz počas rokov

# Vyhodnotenie výsledkov

Výsledky analýz ukazujú viaceré dôležité zistenia, ktoré podporujú cieľ semestrálnej práce – identifikáciu genetického rizika pre hereditárnu hemochromatózu (HH) na základe mutácií v géne HFE.

**Hardy-Weinbergova rovnováha**

Testovanie Hardy-Weinbergovej rovnováhy pre všetky tri analyzované mutácie (C282Y, H63D, S65C) ukázalo, že rozdelenie genotypov v prípade mutácií C282Y a H63D sa približne drží rovnováhy v populácii (Obrázky 1–2). Výsledky pre S65C (Obrázok 3) však naznačujú väčšiu odchýlku, čo môže byť spôsobené nižšou frekvenciou tejto mutácie alebo selektívnymi faktormi v skúmanej populácii.

**Genotypové zastúpenie**

Grafy percentuálneho zastúpenia genotypov (Obrázky 4–6) potvrdzujú, že:

* Najčastejším genotypom v prípade mutácie C282Y je heterozygotná forma (C282Y/WT), s nižším výskytom homozygotov.
* Pre mutáciu H63D je výskyt heterozygotov taktiež dominantný, ale homozygotov je relatívne viac ako pri C282Y.
* Mutácia S65C je raritná, pričom väčšina populácie je wild type.

Tieto výsledky zodpovedajú európskemu genetickému profilu populácie.

**Korelácia medzi mutáciami a diagnózami**

Korelačná matica (Obrázok 7) naznačuje, že žiadna z troch sledovaných mutácií HFE génu (C282Y, H63D, S65C) nemá silný lineárny vzťah s výskytom pečeňového ochorenia. Výskum by mal zvážiť iné faktory alebo možno použiť inú formu analýzy (napr. logistickú regresiu).

**Genotypy podľa veku a pohlavia**

Grafy rozdelenia genotypov podľa veku (Obrázky 9–11) ukazujú, že výskyt mutácií je relatívne rovnomerne rozložený naprieč vekovými skupinami, no zvýšená prevalencia homozygotov C282Y je zjavná najmä v strednom a staršom veku, čo korešponduje s oneskoreným nástupom symptómov HH.

Podľa pohlavia (Obrázky 12–14) možno pozorovať mierne vyšší výskyt rizikových genotypov u mužov, čo odráža aj klinickú realitu – u mužov sa HH prejavuje skôr a častejšie.

**Vývoj diagnóz v čase**

Graf (Obrázok 15) dokumentuje stabilný výskyt hlavných typov diagnóz v priebehu rokov. Mierny nárast pečeňových ochorení môže byť dôsledkom lepšej diagnostiky alebo zvýšeného povedomia o genetickom riziku.

Klasifikácia diagnóz sa pravidelne aktualizuje, aby lepšie zodpovedala aktuálnemu stavu poznatkov v medicíne. Z tohto dôvodu niektoré staršie kódy postupne zastarávajú a sú nahrádzané presnejšími alebo modernejšími označeniami, ktoré lepšie vystihujú konkrétne diagnózy.

V našom datasete sme identifikovali dva takéto zastarané kódy týkajúce sa pečeňových ochorení, ktoré by v súčasnosti mali byť nahradené vhodnejšími alternatívami:

* **Kód K76.0 – Tuková degenerácia pečene**
  + *Počet pacientov:* 63
  + *Odporúčaný nový kód:* E66.9 – Nešpecifikovaná obezita
  + *Vysvetlenie:* K76.0 sa používal na označenie ukladania tuku v pečeňovom tkanive. Dnes sa však preferujú presnejšie diagnózy, ako napríklad nealkoholová tuková choroba pečene (NAFLD), ktorá je často spojená s obezitou či metabolickými poruchami.
* **Kód K75.9 – Nešpecifikované zápalové ochorenie pečene**
  + *Počet pacientov:* 236
  + *Odporúčaný nový kód:* K75.8 – Iné zápalové ochorenia pečene
  + *Vysvetlenie:* Pôvodný kód K75.9 je príliš všeobecný a dnes sa uprednostňuje konkrétnejšia identifikácia zápalového ochorenia (napr. autoimunitná hepatitída, vírusová hepatitída a pod.), čím sa zvyšuje presnosť diagnózy aj účinnosť liečby.

Zachytávanie a nahrádzanie zastaraných diagnóz v zdravotníckych databázach je kľúčové nielen pre kvalitné analytické spracovanie údajov a epidemiologické výskumy, ale aj pre zlepšenie samotnej zdravotnej starostlivosti. Pri digitalizácii údajov je preto nevyhnutné zabezpečiť, aby boli diagnózy aktualizované podľa platných klasifikačných štandardov.

# Záver

Táto semestrálna práca úspešne spojila bioinformatické, štatistické a softvérové prístupy na riešenie reálneho biomedicínskeho problému – genetickej predispozície na hereditárnu hemochromatózu. Vyvinutý softvérový prototyp umožnil systematické spracovanie genetických údajov, analýzu rozdelenia HFE mutácií a ich súvislosti s klinickými premennými ako vek, pohlavie a diagnóza.

Zistenia práce podporujú hypotézu o významnom vplyve mutácií C282Y a H63D na vznik HH, najmä u mužov v strednom a vyššom veku. Vizualizácie navyše poskytujú intuitívny pohľad na štruktúru populácie a môžu slúžiť ako nástroj v klinickej praxi pre rozhodovanie o ďalšom postupe.

**Zhrnutie dosiahnutých cieľov:**

* **Dátová analýza:** Úspešne sme vykonali predspracovanie údajov – identifikovali a odstránili neplatné alebo chýbajúce záznamy, štandardizovali genotypové označenia a kategorizovali premenné ako vek, pohlavie a diagnózy.
* **Genetická analýza:** Aplikovali sme Hardy-Weinbergov test na overenie genetickej rovnováhy a analyzovali percentuálne zastúpenie jednotlivých genotypov a alel. Vyhodnotili sme aj počet pacientov s genetickou predispozíciou a zloženými heterozygotmi.
* **Klinické súvislosti:** Vykonali sme korelačnú analýzu medzi mutáciami a výskytom vybraných diagnóz (najmä ochorení pečene) a skúmali sme vývoj diagnóz v čase vrátane identifikácie zastaraných kódov MKCH-10.
* **Vizualizácia údajov:** Vytvorili sme sériu grafických výstupov – histogramy, boxploty, heatmapy a časové série – ktoré poskytujú prehľadné zobrazenie genetických a klinických údajov.
* **Vývoj softvérového prototypu:** Navrhli a implementovali sme funkčnú desktopovú aplikáciu s používateľsky prívetivým GUI, ktorá umožňuje načítanie, spracovanie, analýzu a export výsledkov.

**Zhodnotenie riešenia:**

Vyvinutý prototyp sa ukázal ako efektívny a spoľahlivý nástroj na spracovanie komplexných biomedicínskych údajov. Jeho modulárna architektúra a využitie otvorených knižníc umožňujú flexibilitu a jednoduché rozšírenie o nové funkcionality. Z pohľadu používateľa je systém intuitívny a poskytuje kvalitné výstupy pre podporu klinických rozhodnutí.