Obsah

Seznam příloh

1 Úvod

Cílem tohoto projektu je analyzovat data týkající se onkologické péče a vyvodit z nich klíčové poznatky. Analyzovaná data byla získána z databázového výpisu za období let 2019 až 2022. Pro zpracování a analýzu dat byla využita programovací platforma Python, konkrétně knihovny pandas pro manipulaci s daty a matplotlib pro vytváření přehledných vizualizací. Tento dokument podrobně popisuje celý proces analýzy, od počáteční přípravy a čištění dat až po finální interpretaci výsledků.

2 Cíle projektu

Cílem projektu je analyzovat data pacientů, léčby a neutropenie za účelem:

1. Stanovení úmrtnosti pacientů podle časového období od poslední léčby.
2. Incidence febrilní neutropenie u pacientů na systémové léčbě.
3. Úmrtnost podle diagnózy.

3 Použité technologie

- Programovací jazyk: **Python**
- Knihovny: **pandas** pro manipulaci s daty, **matplotlib** pro vizualizaci dat.

4 Základní popis analyzovaného vzorku populace

Analyzovaný dataset obsahuje informace o pacientech s onkologickými diagnózami v období let 2019–2022. Hlavní charakteristiky vzorku populace zahrnují:

- Počet pacientů: **2000**
- Věkové rozložení: datum narození v období **1933–1995**
- Rozložení pohlaví: muži **469** z toho **156** úmrtí; ženy **1531** z toho **195** úmrtí
- Průměrný počet léčeb: **12.74** na pacienta

5 Importování dat

Data byla načtena z Excel souboru pomocí knihovny pandas.

```
4   data = pd.read_excel("MOU_data.xlsx", sheet_name=None)
5   pacienti = data["data_pacienti"]
6   lecba = data["data_lecba"]
7   neutropenie = data["data_neutropenie"]
```

Obrázek 1: Import dat

6 Kontrola dat

Před samotnou analýzou jsem provedl pečlivou kontrolu kvality dat, abych zajistil, že výstupy analýzy budou přesné a relevantní. Kontrola zahrnovala:

6.1 Kontrola nulových řádků

Nulové hodnoty se vyskytly pouze ve sloupci **datum_umrti**, což je v pořádku a odpovídá očekávání.

```
10  #otestování nulových řádků
11  print(pacienti.isnull().sum())
12  print(lecba.isnull().sum())
13  print(neutropenie.isnull().sum())
```

```
id                  0
pohlavi             0
rok_narozeni        0
datum_diagnozy      0
topografie          0
morfologie          0
diagnoza            0
klinicke_stadium    0
datum_umrti      1649
```

Obrázek 2: Kontrola nulových řádků

6.2 Kontrola duplicity

Kontrola duplicit neodhalila žádné opakované záznamy.

```
15  print(pacienti.duplicated().sum())
16  print(lecba.duplicated().sum())
17  print(neutropenie.duplicated().sum())
```

Obrázek 3: Kontrola duplicity

6.3 Kontrola validity dat

Výsledky kontroly validity dat ukazují, že data narození spadají do očekávaného rozsahu. Stejně tak jsou v pořádku i data diagnózy a úmrtí.

```
21  print(pacienti[["rok_narozeni", "datum_diagnozy", "datum_umrti"]].describe())
```

Obrázek 4: Kód pro kontrolu validity dat

	rok_narozeni	datum_diagnozy	datum_umrti
count	2000.00000	2000	351
mean	1963.15000	2020-12-27 11:17:31.200000	2021-07-30 11:08:43.076923136
min	1933.00000	2019-01-01 00:00:00	2019-07-24 00:00:00
25%	1952.00000	2020-01-27 18:00:00	2020-11-05 00:00:00
50%	1962.00000	2021-01-18 00:00:00	2021-10-10 00:00:00
75%	1974.00000	2021-12-05 12:00:00	2022-04-24 00:00:00
max	1995.00000	2022-12-01 00:00:00	2022-12-30 00:00:00
std	13.29157	NaN	NaN

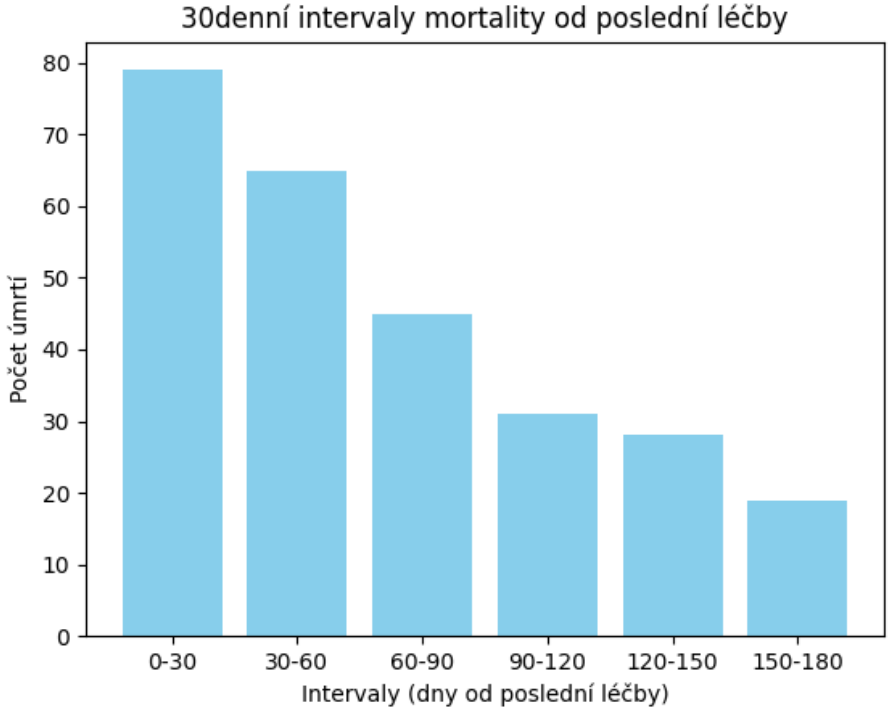
Obrázek 5: Výsledek validity dat

7 Analýza

7.1 Analýza úmrtnosti pacientů podle času od poslední léčby

```
11  #vytvoření tabulky s datem poslední léčby
12  posledni_lecba = (
13    lecba.groupby("id")["datum_aplikace"]
14    .max()
15    .reset_index()
16    .rename(columns={"datum_aplikace": "posledni_lecba"})
17  )
18
19  #spojení pacientů s tabulkou která obsahuje datum poslední léčby
20  pacienti = pd.merge(pacienti, posledni_lecba, on="id")
21
22  #výpočet počtu dnů mezi poslední léčbou a úmrtím
23  pacienti["rozdil_dny"] = (
24    pd.to_datetime(pacienti["datum_umrti"]) - pd.to_datetime(pacienti["posledni_lecba"])
25  ).dt.days
```

Obrázek 6: Rámec pro vizualizaci



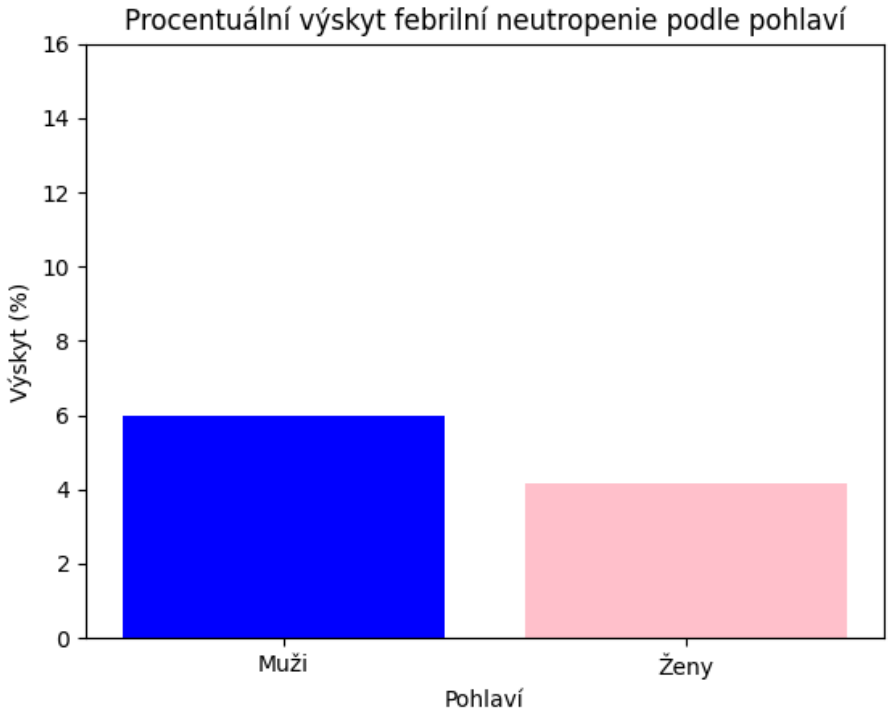
Obrázek 7: Graf mortality podle časového intervalu od poslední léčby

Z grafu je patrné, že úmrtnost pacientů v analyzovaném 180denním období od poslední léčby postupně klesá, což může naznačovat pozitivní vliv léčby a nebo zlepšení zdravotního stavu pacientů v čase.

7.2 Analýza incidence febrilní neutropenie u pacientů na systémové léčbě.

```python
#získání data poslední léčby pro každého pacienta
posledni_lecba = (
    lecba.groupby("id")["datum_aplikace"]
    .max()
    .reset_index()
    .rename(columns={"datum_aplikace": "posledni_lecba"})
)

#spojení tabulek pacienti a poslední léčby
pacienti_lecba = pd.merge(pacienti, posledni_lecba, on="id", how="left")

#spojení pacientů s daty o neutropenii
pacienti_neutropenie = pd.merge(pacienti_lecba, neutropenie, on="id", how="left")

#přidání příznaku neutropenie (True/False)
pacienti_neutropenie["neutropenie"] = ~pacienti_neutropenie["datum_neutropenie"].isna()

#výpočet výskytu neutropenie podle pohlaví
neutropenie_pohlavi = (
    pacienti_neutropenie.groupby("pohlavi")["neutropenie"].mean() * 100
)
```

Obrázek 8: Rámec pro vizualizaci



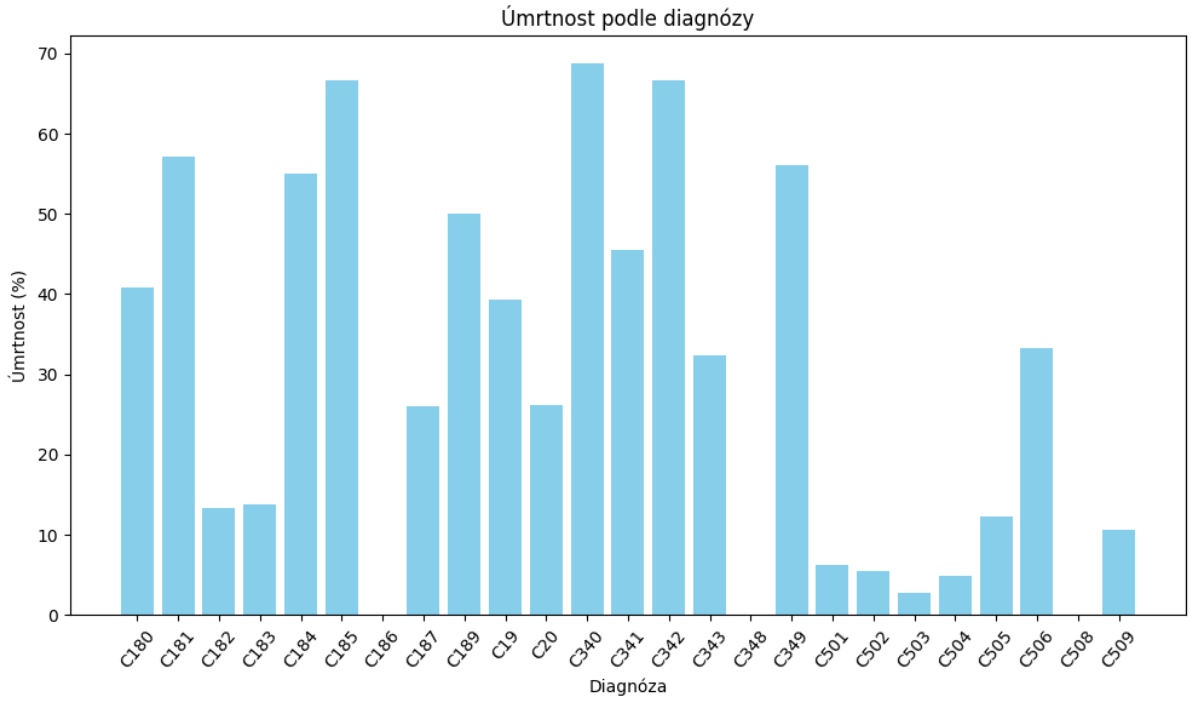
Obrázek 9: Incidence neutropenie podle pohlaví

Tato analýza zkoumá, jak často se u pacientů na systémové léčbě objevuje febrilní neutropenie (horečka a nízký počet bílých krvinek). Cílem je identifikovat možné rizikové faktory. Výsledky mohou pomoci zlepšit péči o pacienty a snížit riziko této komplikace. V rámci této analýzy jsme porovnávali výskyt (incidenci) sledovaného jevu mezi muži a ženami. Rozdíl mezi pohlavími nebyl signifikantní.

7.3 Analýza úmrtnost podle diagnózy.

```
#přidání sloupce úmrtí true/false
pacienti["umrti"] = ~pacienti["datum_umrti"].isna()

#skupinový počet pacientů a počet zemřelých podle diagnózy
umrtnost_podle_diagnozy = (
  pacienti.groupby("diagnoza", as_index=False)
  .agg(celkem_pacientu=("id", "count"), zemreli=("umrti", "sum"))
)

#výpočet úmrtnosti v procentech
umrtnost_podle_diagnozy["umrtnost (%)"] = (
  umrtnost_podle_diagnozy["zemreli"] / umrtnost_podle_diagnozy["celkem_pacientu"] * 100
)
```

Obrázek 10: Rámec pro vizualizaci



Obrázek 11: Úmrtnost podle diagnózy

Tato analýza se zaměřuje na úmrtnost pacientů v závislosti na jejich diagnóze. Cílem je identifikovat, u kterých diagnóz je úmrtnost nejvyšší, a zjistit možné příčiny těchto rozdílů. Výsledky mohou pomoci zlepšit léčebné postupy a cíleně snížit úmrtnost u nejrizikovějších skupin pacientů. Například u karcinomů prsu (C50X), kde je mortalita obecně nízká, s výjimkou diagnózy C506, která vykazuje vyšší úmrtnost.

8 Závěr

Projekt analýzy nemocničních dat přinesl cenné poznatky o trendech, rizikových faktorech a rozdílech v úmrtnosti či výskytu komplikací mezi různými diagnózami a skupinami pacientů. Výsledky ukázaly oblasti, kde je možné zlepšit péči o pacienty, například optimalizací léčebných protokolů nebo zavedením preventivních opatření. Analýzy tohoto typu slouží jako klíčový podklad pro rozhodování vedení nemocnice a mohou významně přispět ke zlepšení kvality poskytované zdravotní péče. V této práci jsou představeny pouze některé z indikátorů, které je možné sledovat.