Analiza pojavnosti ne-Hodgkinovog limfoma u bolesnika sa Sjoegrenovim sindromom metodama analize doživljenja

Šinka, Petra

Master's thesis / Diplomski rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:876788

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Petra Šinka

ANALIZA POJAVNOSTI NE-HODGKINOVOG LIMFOMA U BOLESNIKA SA SJOEGRENOVIM SINDROMOM METODAMA ANALIZE DOŽIVLJENJA

Diplomski rad

Voditelj rada:
prof. dr. sc.
Anamarija Jazbec

Zagreb, rujan 2022.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Sredinom prošlog stoljeća dogodio se najznačajniji razvoj analize doživljenja, skupine statističkih metoda koje se koriste kada se uz promatrani događaj od interesa proučava i vrijeme potrebno do pojave tog događaja. Počeci analize doživljenja datiraju još iz 17. stoljeća te je ona dugo vremena bila vezana isključivo uz istraživanje mortaliteta. Primjena joj se u zadnjih nekoliko desetljeća proširila znatno izvan granica medicinskih istraživanja pa se u današnje vrijeme ove statističke metode koriste i u raznim drugim područjima i industrijama.
Najveći doprinos razvitku analize doživljenja dali su 1958. godine matematičari Edward Lynn Kaplan i Paul Meier predloživši način za procjenu funkcije doživljenja te 1972. godine statističar David Cox, koji je razvio Coxov regresijski model. U prva dva poglavlja ovog rada bit će obrađene osnove analize doživljenja te već spomenuti Coxov regresijski model.
U trećem poglavlju ovog rada obrađena će teorija biti primijenjena na podatke o skupini pacijenata dijagnosticiranih s Primarnim Sjoegrenovim sindromom. Poznato je da u usporedbi s općom populacijom, bolesnici koji pate od ove bolesti imaju povećan rizik od razvoja ne-Hodgkinovog limfoma. Iz tog će razloga promatrani događaj od interesa biti razvitak limfoma, a ispitat će se i utjecaj nekoliko lijekova i drugih faktora na očekivano vrijeme do razvijanja ne-Hodgkinovog limfoma.

Poglavlje 1

Analiza doživljenja

1.1 Osnovni pojmovi

Pretpostavimo da želimo proučavati pojavu nekog događaja u grupi subjekata. Kada nam vrijeme do samog događaja ne bi bilo važno, pojavu događaja mogli bismo modelirati pomoću binarne logističke regresije. Na primjer, u analizi smrtnosti nakon otvorene operacije srca nije toliko bitno je li pacijent umro tijekom operacije ili nakon dva mjeseca, već je li umro ili ne (u nekom određenom razdoblju). Za druge je pak probleme vrijeme do nekog događaja izrazito značajno. Analiza koja bi jednostavno brojala sve smrti (događaje) odbacila bi vrijednu informaciju o vremenu te time lišila istraživače korisnih zaključaka. Tu stupa na snagu analiza doživljenja.

Analiza doživljenja (engl. *survival analysis*) jest skup statističkih procedura za analizu podataka kod kojih je *vrijeme do događaja* promatrana varijabla od interesa. Pod *vrijeme* mislimo na godine, mjesece, tjedne ili dane od početka istraživanja do trenutka kad se događaj dogodio (ili do kraja istraživanja). *Događaj* uglavnom ima negativnu konotaciju, budući da je često događaj od interesa smrt, pojava bolesti, povratak bolesti nakon remisije i slično. Upravo se iz tog razloga u analizi doživljenja umjesto „događaj" često kaže neuspjeh (engl. *failure*), a za vrijeme kažemo *vrijeme doživljenja*. Međutim, vrijeme doživljenja može biti npr. „vrijeme potrebno za povratak na posao nakon kliničkog zahvata", u kojem je slučaju neuspjeh zapravo pozitivan događaj.

Neki od prednosti analize doživljenja za proučavanje vremena doživljenja nad ostalim statističkim procedurama su sljedeći:

1. Vrijeme do događaja može imati neobičnu distribuciju. Budući da se radi o

vremenu, ono je restringirano da bude nenegativno pa se radi o asimetričnoj distribuciji koja nikada neće biti normalna.

2. Vjerojatnost doživljenja određenog vremena često je relevantnija nego očekivano vrijeme doživljenja (ono može biti teško za procijeniti ukoliko je velik broj cenzuriranih[1] podataka).

3. Funkcije koje se koriste u analizi doživljenja smislene su i dobro prilagođene potrebama analize varijable odaziva (vremena do događaja).

Osim što je često korištena u eksperimentima u industriji, analiza doživljenja ipak najveću primjenu ima u medicinskim istraživanjima. No, ova je vrsta statističke analize posvuda prisutna i upotrebljiva bilo kada kad su faktori od interesa vrijeme te pojava događaja. Prva dva poglavlja izrađena su prema knjizi [5].

1.2 Cenzuriranje

U većini istraživanja mora se uzeti u obzir ključni analitički problem zvan **cenzuriranje**. Ono se javlja kada imamo *neku* informaciju o vremenu doživljenja, ali ne znamo *točno* vrijeme. Najčešći oblik cenzuriranja je tzv. **desno cenzuriranje**, kod kojeg je točno vrijeme doživljenja veće od neke poznate vrijednosti.

Tri najčešća razloga za desno cenzuriranje navedena su ovdje:

1. događaj se nije dogodio prije završetka istraživanja,

2. izgubi se kontakt s osobom za vrijeme istraživanja te tako nisu dostupni daljnji podaci potrebni za analizu,

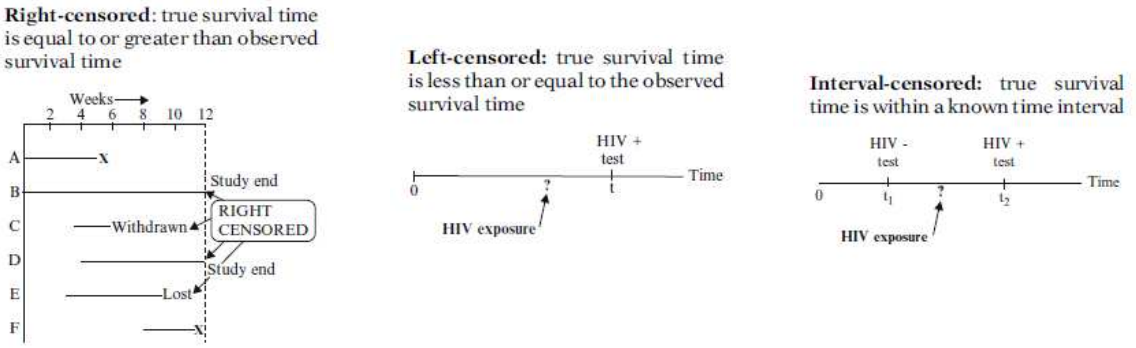
3. osoba je povučena iz istraživanja zbog smrti (a smrt nije bila promatrani događaj od interesa) ili zbog nekog drugog razloga.

Podaci mogu biti i **lijevo cenzurirani**. U ovom se slučaju događaj dogodio prije nekog poznatog vremena. Na primjer, ukoliko promatramo osobe dok ne postanu HIV pozitivne, možemo zabilježiti neuspjeh kada osoba već na prvom testiranju ispadne HIV pozitivna. Tada ne znamo točno vrijeme prvom izlaganju virusu, te stoga ne znamo točno vrijeme kad se neuspjeh dogodio.

[1] Pojam „cenzuriranja" bit će objašnjen u sljedećem poglavlju.

Treća vrsta cenzuriranja jest **intervalno cenzuriranje**. Kao što samo ime govori, pravo vrijeme događaja je unutar nekog poznatog vremenskog intervala. Kao primjer uzmimo opet analizu bolesnika s HIV-om. Osoba se može testirati na HIV dva puta, od kojih je prvi put ispala HIV negativna, a drugi HIV pozitivna. Dakle, znamo da je točno vrijeme događaja u intervalu između prvog i drugog testiranja, ali ne znamo njegovu točnu vrijednost.

Da sumiramo, ukoliko je podatak desno cenzuriran u trenutku t, točno vrijeme doživljenja je stupilo nekad nakon tog trenutka. Za lijevo cenzurirane podatke u trenutku t događaj se dogodio prije tog trenutka, dok se za intervalno cenzurirane podatke događaj dogodio negdje unutar intervala $\langle t_1, t_2 \rangle$.



(a) Desno cenzuriranje (b) Lijevo cenzuriranje (c) Intervalno cenzuriranje

Slika 1.1: Prikaz različitih vrsti cenzuriranja (izvor slike jest knjiga [5])

1.3 Funkcija doživljenja i funkcija hazarda

U analizi doživljenja obično varijablu odaziva označavamo s T, s obzirom da ona predstavlja vrijeme do događaja (engl. *time to event*). Umjesto da statistički model gradimo u terminima očekivanog vremena događaja, praktičnije je definirati ga pomoću funkcije doživljenja. Sve formule i izvodi izvađeni su iz knjige [6].

Definicija 1.3.1. *Za slučajnu varijablu T definiramo* ***funkciju doživljenja*** $S : [0, +\infty\rangle \to [0, 1]$ *kao*

$$S(t) \coloneqq \mathbb{P}(T > t) = 1 - F(t),$$

gdje je F funkcija distribucije slučajne varijable T.

Ako je događaj smrt, $S(t)$ je vjerojatnost da se smrt dogodi nakon vremena t, tj. vjerojatnost da osoba preživi barem do vremena t. Vrijedi $S(0) = 1$ jer svi subjekti dožive vrijeme $t = 0$. Također, funkcija doživljenja je nerastuća te poprima vrijednosti u intervalu $[0, 1]$.

Definicija 1.3.2. *Za slučajnu varijablu T definiramo* ***funkciju hazarda*** $h : [0, +\infty\rangle \to [0, +\infty\rangle$ *s*

$$h(t) \coloneqq \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

te ***kumulativnu funkciju hazarda*** $H : [0, +\infty\rangle \to [0, +\infty\rangle$ *s*

$$H(t) \coloneqq \int_0^t h(u)\, du.$$

Kao što vidimo, funkcija hazarda definirana je preko uvjetne vjerojatnosti, ali ona je zapravo omjer nenegativnog i pozitivnog broja te stoga može poprimiti vrijednosti u intervalu $[0, +\infty\rangle$.

Funkciju hazarda malo je teže interpretirati nego funkciju dožiljenja zbog kompliciranije definicije, međutim, u suštini, funkcija hazarda $h(t)$ daje trenutni potencijal po jedinici vremena da se događaj dogodi, uz uvjet da je subjekt doživio trenutak neposredno prije t. Primijetimo da, za razliku od funkcije doživljenja koja se fokusira na preživljavanje (tj. uspjeh), funkcija hazarda fokusirana je na neuspjeh (događaj se dogodi). Tako se na neki način funkcija hazarda može gledati kao funkcija koja daje suprotnu stranu informacije od funkcije doživljenja.

Za neprekidnu slučajnu varijablu T vrijedi:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t)}{\mathbb{P}(T \geq t) \cdot \Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{S(t) \cdot \Delta t} \\ &= \frac{\partial F(t)}{\partial t} \cdot \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)}, \end{aligned}$$

gdje je f funkcija gustoće slučajne varijable T.

Budući da vrijedi

$$\frac{\partial \ln S(t)}{\partial t} = \frac{\partial S(t)/\partial t}{S(t)} = \frac{\partial [1 - F(t)]/\partial t}{S(t)} = -\frac{f(t)}{S(t)},$$

funkcija hazarda također se može prikazati kao

$$h(t) = -\frac{\partial \ln S(t)}{\partial t}.$$

Ako idemo unatrag, integral od $h(t)$ je

$$\int_0^t h(u)\, du = -\ln S(t),$$

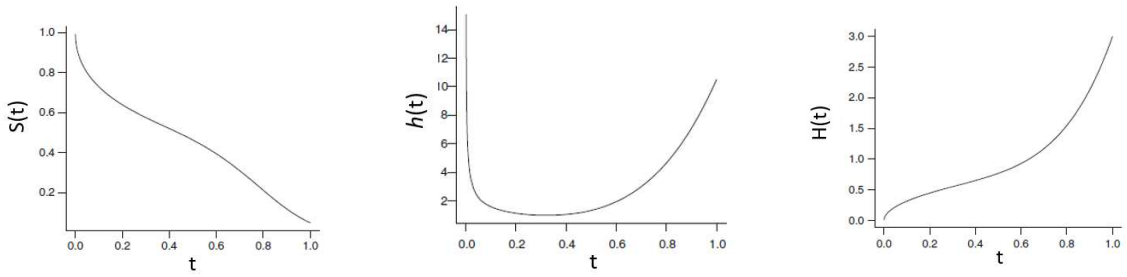
a kako je po definiciji prethodni integral upravo jednak kumulativnom hazardu, imamo

$$H(t) = -\ln S(t)$$

te

$$S(t) = e^{-H(t)}.$$

Upravo smo vidjeli da znajući jednu od triju funkcija S, h ili H možemo izvesti i druge dvije. Te tri funkcije na različite načine opisuju istu distribuciju.



(a) Funkcija doživljenja (b) Funkcija hazarda (c) Kum. funkcija hazarda

Slika 1.2: Primjer funkcije doživljenja, hazarda te kumulativne funkcije hazarda za istu distribuciju (slike su iz knjige [5])

Iz grafa funkcije doživljenja možemo dobiti kvantile distribucije vremena doživljenja, kao i medijan vremena doživljenja:

$$T_{medijan} = S^{-1}(0.5).$$

Napomena 1.3.3. *Kada je T diskretna slučajna varijabla, funkcija hazarda dana je s*

$$h(t_j) = \mathbb{P}(T = t_j \mid T \geq t_j) = \frac{\mathbb{P}(T = t_j)}{\mathbb{P}(T \geq t_j)} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}, \quad j = 1, 2, \ldots$$

pa budući da je $S(t_0) = 1$, funkciju doživljenja možemo izraziti preko funkcije hazarda:

$$S(t) = \prod_{t_j \leq t} \frac{S(t_j)}{S(t_{j-1})} = \prod_{t_j \leq t} [1 - h(t_j)], \quad t \geq 0.$$

1.4 Kaplan-Meierov procjenitelj

Standardni procjenitelj funkcije doživljenja zove se Kaplan-Meierov procjenitelj (ili na engleskom, *Product-Limit estimator*). Ime je, naravno, dobio po matematičarima Edwardu L. Kaplanu i Paulu Meieru koji su 1958. godine u svom radu predložili spomenuti procjenitelj. Radi se o neparametarskoj statistici koju ćemo definirati nakon što uvedemo potrebne oznake. Od ovog poglavlja na dalje teorija je izvađena iz knjige [6].

Neka u istraživanju sudjeluje n osoba, neka je V_i pravo vrijeme događaja subjekta i, C_i potencijalno vrijeme cenzuriranja istog tog subjekta, $i = 1, \ldots, n$. Ono je potrebno u slučaju da ne znamo pravo vrijeme događaja, tj. ukoliko je $V_i > C_i$. Označimo $T_i = \min(V_i, C_i)$ te ćemo tu veličinu dalje zvati vrijeme doživljenja. Kako bismo znali radi li se o cenzuriranom podatku ili ne, uvodimo indikatorsku varijablu δ_i koja je jednaka 1 ukoliko se događaj dogodio te 0 ukoliko je podatak cenzuriran. Parovi slučajnih varijabli (T_i, δ_i) svi su potrebni podaci za bilo koju analizu doživljenja.

Pretpostavimo da se događaji dogode u D različitih diskretnih točaka $t_1 < t_2 < \cdots < t_D$ te da u trenutku t_i se dogodi d_i događaja. Neka je Y_i broj pojedinaca koji su rizični u trenutku t_i (tj. oni za koje se događaj još nije dogodio do trenutka t_i, tj. preživjeli su do neposredno prije trenutka t_i). Za te je osobe vrijeme doživljenja, dakle, veće ili jednako t_i. Veličina d_i/Y_i procjena je uvjetne vjerojatnosti da se za pojedinca za kojeg se događaj nije dogodio neposredno do trenutka t_i događaj dogodi u trenutku t_i.

Sada konačno možemo definirati Kaplan-Meierov procjenitelj, koji je dan sljedećom formulom:

$$\hat{S}(t) = \begin{cases} 1, & t < t_1 \\ \prod\limits_{\substack{t_i \leq t \\ i=1,\cdots,D}} [1 - \frac{d_i}{Y_i}], & t \geq t_1 \end{cases}$$

.
Za vrijednosti t veće od najveće opažene vrijednosti ovaj procjenitelj nije dobro definiran (vidi [6]).

Kaplan-Meierov procjenitelj je step funkcija sa skokovima u opaženim trenucima događaja. Veličina skokova ne ovisi samo o broju opaženih događaja u trenucima t_i, već i o broju cenzuriranih podataka do trenutka t_i.

Varijanca ovog procjenitelja dobivena je Greenwoodovom formulom:

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)},$$

a standardnu grešku procjenitelja dobivamo kao korijen iz varijance, $\hat{V}[\hat{S}(t)]^{1/2}$.

Kumulativnu funkciju hazarda $H(t) = -\ln[S(t)]$ procjenjujemo procjeniteljem $\hat{H}(t) = -\ln[\hat{S}(t)]$.

1.5 Usporedba funkcija doživljenja

Pretpostavimo da želimo usporediti funkcije doživljenja za $K \geq 2$ populacija. Testiramo sljedeće hipoteze:

$$H_0 : S_1(t) = S_2(t) = \ldots = S_K(t), \quad \forall t \leq \tau$$
$$H_1 : \text{barem jedna od } S_j(t) \text{ je drugačija za neki } t \leq \tau,$$

gdje je τ najveće vrijeme u kojem sve grupe imaju barem jedan rizični subjekt.

Podaci na kojima testiramo navedene hipoteze sastoje se od nezavisnih uzoraka (s nekim podacima koji su cenzurirani) iz svake od K populacija. Neka su opet $t_1 < t_2 < \ldots < t_D$ različita vremena događaja u zajedničkom uzorku (uzorak koji dobijemo kad spojimo svih K uzoraka). U trenutku t_i opazimo d_{ij} događaja u j-tom uzorku od Y_{ij} rizičnih subjekata (iz j-tog uzorka u trenutku t_i), $j = 1, \ldots, K$, $i = 1, \ldots, D$. Neka je $d_i = \sum_{j=1}^{K} d_{ij}$ i $Y_i = \sum_{j=1}^{K} Y_{ij}$ ukupan broj događaja i ukupan broj rizičnih subjekata u trenutku t_i u zajedničkom uzorku.

Testna statistika dobivena je kao suma težinskih razlika između opaženog broja neuspjeha (događaja) i očekivanog broja neuspjeha pod H_0 u j-tom uzorku, a glasi:

$$Z_j(\tau) = \sum_{i=1}^{D} W(t_i) \left[d_{ij} - Y_{ij} \left(\frac{d_i}{Y_i} \right) \right], \quad j = 1, \ldots, K.$$

Ovdje je W težinska funkcija koju dodajemo kako bi se još više naglasile razlike između opaženih i očekivanih vrijednosti. Očekivani broj događaja u j-tom uzorku u trenutku t_i je udio rizičnih pojedinaca iz uzorka j u trenutku t_i, dakle Y_{ij}/Y_i, pomnožen s brojem događaja d_i u trenutku t_i.

Varijanca statistike $Z_j(\tau)$ dana je s

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i}\left(1 - \frac{Y_{ij}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i,\ j = 1, \ldots, K,$$

a kovarijanca od $Z_j(\tau)$ i $Z_g(\tau)$ procijenjena je s

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i}\frac{Y_{ig}}{Y_i}\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i,\ g \neq j,\ g = 1, \ldots, K.$$

Budući da je $\sum_{j=1}^{K} Z_j(\tau) = 0$, komponente vektora $(Z_1(\tau), \ldots, Z_K(\tau))$ čine linearno zavisan skup. Testna statistika je konstruirana odabirom bilo kojih $K-1$ Z_j-ova. Procijenjena kovarijacijska matrica ovih statistika dana je s $(K-1)\times(K-1)$ matricom Σ (koja sadržava odgovarajuće $\hat{\sigma}_{jg}$-ove). Testna statistika je sljedeća:

$$(Z_1(\tau), \ldots, Z_{K-1}(\tau))\Sigma^{-1}(Z_1(\tau), \ldots, Z_{K-1}(\tau))^{\intercal} \overset{H_0}{\sim} \chi^2(K-1).$$

Nekoliko je težinskih funkcija predloženo po literaturama. Česta je $W(t) = 1\ \forall t$, taj odabir funkcije dovodi do poznatog *log-rank* testa. Još neki popularni izbori težinske funkcije su $W(t_i) = Y_i$ te $W(t_i) = f(Y_i)$, gdje je f neka fiksna funkcija, predlaže se često drugi korijen.

Poglavlje 2

Coxov regresijski model

Često želimo usporediti vremena doživljenja za dvije ili više grupe, a još češće su nam poznate neke dodatne karakteristike subjekata koje bi mogle imati utjecaj na zavisnu varijablu. Coxova regresija koristi se kada želimo procijeniti utjecaj pojedinih nezavisnih varijabli na zavisnu koja sadrži dvije informacije - je li se događaj dogodio te vrijeme kad se taj događaj dogodio (ili cenzurirano vrijeme ako se događaj nije dogodio). Za slučaj da nam nije poznata informacija o vremenu, koristili bismo logističku regresiju. Model se zasniva na pretpostavci paralelnog hazarda koja će biti objašnjena u nastavku.

Prediktorske varijable mogu biti različite, ovisno o tipu istraživanja, od godina, spola, obrazovanja, do dijetnih navika, krvnog tlaka, statusa pušenja i slično. Nakon dodavanja potencijalnih prediktorskih varijabli, usporedba vremena doživljenja među grupama trebala bi biti točnija i manje pristrana nego procjena bez prediktorskih varijabli.

Kao i prije, s V ćemo označiti pravo vrijeme događaja, C predstavlja cenzurirano vrijeme, $T = \min(V, C)$. Imamo n subjekata, dakle podaci se sastoje od trojki $(T_j, \delta_j, \mathbb{X}_j)$, $j = 1, \ldots, n$, gdje je δ_j indikatorska varijabla koja ukazuje na to je li se događaj dogodio ili ne (1 ako se dogodio i 0 ako nije) te je $\mathbb{X}_j = (X_{j1}, \ldots, X_{jp})^\intercal$ vektor kovarijata za j-tog subjekta (p predstavlja broj kovarijata).

Neka je $h(t \mid \mathbb{X})$ funkcija hazarda u trenutku t za pojedinca s *vektorom rizika* (tj. vektorom prediktorskih varijabli) $\mathbb{X}$. Coxov regresijski model tada glasi

$$h(t \mid \mathbb{X}) = h_0(t) \exp\left(\sum_{k=1}^{p} \beta_k X_k\right),$$

gdje je $h_0(t)$ proizvoljna osnovna funkcija hazarda (tzv. *baseline*), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ vektor parametara. Umjesto eksponencijalne funkcije, može se koristiti i neka druga funkcija $c(\boldsymbol{\beta}^\intercal \mathbb{X})$, međutim eksponencijalna je najčešća upravo zbog svojstva da je pozitivna na cijeloj domeni (funkcija hazarda također mora biti nenegativna).

Ukoliko model logaritmiramo, dobit ćemo njegovu lineariziranu verziju:

$$\ln h(t \mid \mathbb{X}) = \ln h_0(t) + \sum_{k=1}^{p} \beta_k X_k.$$

Sada vidimo da regresijski koeficijenti za X_j, β_j, zapravo predstavljaju povećanje u log hazardu u bilo kojoj fiksnoj vremenskoj točki ako X_j uvećamo za 1 i sve ostale prediktorske varijable držimo konstantnima. To matematički možemo napisati na sljedeći način:

$$\beta_j = \ln h(t \mid (X_1, \ldots, X_j + 1, \ldots, X_p)) - \ln h(t \mid (X_1, \ldots, X_j, \ldots, X_p))$$

Iz svojstva logaritama sada slijedi da je prethodna jednadžba jednaka logaritmiranom omjeru hazarda u trenutku t, tj.

$$\beta_j = \ln \left(\frac{h(t \mid (X_1, \ldots, X_j + 1, \ldots, X_p))}{h(t \mid (X_1, \ldots, X_j, \ldots, X_p))} \right)$$

Sada slijedi da je omjer hazarda sa $X_j + d$ u odnosu na X_j kada su ostale kovarijate konstantne jednak $\exp(\beta_j d)$. Odnosno, efekt povećanja X_j za d je množenje hazarda događaja faktorom $\exp(\beta_j d)$ u svim vremenskim točkama, uz pretpostavku da je X_j linearno povezan s $\ln h(t)$.

Generalno, omjer hazarda za pojedinca s vektorom prediktorskih varijabli $\mathbb{X}^*$ u odnosu na pojedinca s prediktorima $\mathbb{X}$ jest

$$HR = \frac{h(t \mid \mathbb{X}^*)}{h(t \mid \mathbb{X})} = \frac{h_0(t) \exp(\boldsymbol{\beta}^\intercal \mathbb{X}^*)}{h_0(t) \exp(\boldsymbol{\beta}^\intercal \mathbb{X})} = \frac{\exp(\boldsymbol{\beta}^\intercal \mathbb{X}^*)}{\exp(\boldsymbol{\beta}^\intercal \mathbb{X})} = \exp\{\boldsymbol{\beta}^\intercal (\mathbb{X}^* - \mathbb{X})\},$$

gdje HR označava omjer hazarda (engl. *hazard ratio*).

Upravo smo vidjeli da se u izvodu omjera hazarda osnovna funkcija hazarda $h_0(t)$ poništava. Naime, značajno svojstvo Coxovog regresijskog modela je to da je *baseline* funkcija $h_0(t)$ nespecificirana, tj. ne znamo točnu distribuciju vremena doživljenja. Ono što znamo je način na koji prediktorske varijable utječu na tu distribuciju, tj. množenjem funkcije hazarda sa $\exp(\boldsymbol{\beta}^\intercal \mathbb{X})$. Zbog činjenice da je osnovna funkcija hazarda $h_0(t)$ nespecificirana, za Coxov regresijski model kaže se da je *semi-parametrijski*.

Napomena 2.0.1. *Primijetimo da smo Coxov model definirali pomoću fiksiranih kovarijata* $\mathbb{X}$*, međutim kovarijate mogu biti i ovisne o vremenu, tj. njihova vrijednost se mijenja kroz vrijeme. Tada vektor kovarijata u trenutku* t *označavamo s* $\mathbb{X}(t) = (X_1(t), \ldots, X_p(t))$*. Za vremenski ovisne kovarijate bitno je da je poznata vrijednost* $\mathbb{X}(t)$ *u svakom trenutku dok subjekt sudjeluje u istraživanju. Coxov regresijski model u ovom slučaju je oblika:*

$$h(t \mid \mathbb{X}) = h_0(t) \exp\left(\sum_{k=1}^{p} \beta_k X_k(t)\right).$$

2.1 Procjena parametara modela

Slučaj različitih vremena događaja

U ovom potpoglavlju pretpostavljamo da su vremena događaja i cenzuriranja nezavisna za svakog subjekta te da su se svi događaji dogodili u različitim trenucima. Neka su ponovno $t_1 < \ldots < t_D$ opažena vremena događaja te $X_{(i)k}$ k-ta kovarijata pridružena pojedincu čije je vrijeme događaja jednako t_i. S $R(t_i)$ označavamo skup svih rizičnih pojedinaca u trenutku t_i.

Parametri modela procijenjuju se malo modificiranom metodom maksimalne vjerodostojnosti (engl. *maximum likelihood estimation* ili MLE). Funkcija vjerodostojnosti u Coxovom modelu zove se funkcija „parcijalne" vjerodostojnosti zbog činjenice da formula za vjerodostojnost uzima u obzir vjerojatnosti samo za one subjekte koji su doživjeli neuspjeh, a ne i za one čija su vremena cenzurirana. Ona se tretira kao i obična funkcija vjerodostojnosti, a dana je sljedećom formulom:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left[\sum_{k=1}^{p} \beta_k X_{(i)k}\right]}{\sum_{j \in R(t_i)} \exp\left[\sum_{k=1}^{p} \beta_k X_{jk}\right]}$$

Vidimo da brojnik sadrži informaciju samo o onim pojedincima za koje se događaj dogodio, a nazivnik o svim rizičnim subjektima (zajedno i s onima koji će kasnije biti cenzurirani).

Budući da je maksimiziranje funkcije vjerodostojnosti ekvivalentno maksimiziranju njenog logaritma, označimo $LL(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta})$. Sada je:

$$LL(\boldsymbol{\beta}) = \sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k X_{(i)k} - \sum_{i=1}^{D} \ln\left[\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k X_{jk}\right)\right].$$

Kako bismo našli maksimum ove funkcije, tražit ćemo stacionarne točke njenih parcijalnih derivacija $U_l(\boldsymbol{\beta}) = \partial LL(\boldsymbol{\beta})/\partial\beta_l,\ l = 1, \ldots, p$.

$$U_l(\boldsymbol{\beta}) = \sum_{i=1}^{D} X_{(i)l} - \sum_{i=1}^{D} \frac{\sum_{j \in R(t_i)} X_{jl} \exp\left[\sum_{k=1}^{p} \beta_k X_{jk}\right]}{\sum_{j \in R(t_i)} \exp\left[\sum_{k=1}^{p} \beta_k X_{jk}\right]}$$

Parametre modela $\boldsymbol{\beta}$ dobivamo rješavanjem sustava jednadžbi $U_l(\boldsymbol{\beta}) = 0,\ l = 1, \ldots, p$ nekim od iterativnih metoda.

Tri su glavna testa za testiranje hipoteza o parametrima $\boldsymbol{\beta}$. Neka je $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\intercal$ (parcijalni) MLE procjenitelj vektora parametara $\boldsymbol{\beta}$ te neka je $\mathbf{I}(\boldsymbol{\beta})$ negativna $p \times p$ matrica drugih derivacija logaritmirane funkcije parcijalne vjerodostojnosti evaluirana u vektoru parametara $\boldsymbol{\beta}$, $\mathbf{I}(\boldsymbol{\beta}) = [I_{gl}(\boldsymbol{\beta})]_{g,l=1,\ldots,p}$, gdje je $I_{gl}(\boldsymbol{\beta}) = -\partial U_l(\boldsymbol{\beta})/\partial\beta_g$. Za testiranje hipoteze $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ koristimo tri statistike:

$$\begin{aligned}
\chi^2_W &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\intercal \mathbf{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{H_0}{\sim} \chi^2(p) \\
\chi^2_{LR} &= 2[LL(\hat{\boldsymbol{\beta}}) - LL(\boldsymbol{\beta}_0)] \overset{H_0}{\sim} \chi^2(p) \\
\chi^2_{SC} &= \mathbf{U}(\boldsymbol{\beta}_0)^\intercal \mathbf{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{U}(\boldsymbol{\beta}_0) \overset{H_0}{\sim} \chi^2(p),
\end{aligned}$$

gdje je $\mathbf{U}(\boldsymbol{\beta}) = (U_1(\boldsymbol{\beta}), \ldots, U_p(\boldsymbol{\beta}))^\intercal$ vektor parcijalnih derivacija evaluiran u $\boldsymbol{\beta}$, a navedene statistike su testne statistike za Waldov, *likelihood ratio* i *score* test. Statistike imaju hi-kvadrat distribucije za velike uzorke.

Slučaj s više istih vremena događaja

U prošlom potpoglavlju prikazane su funkcije parcijalne vjerodostojnosti u slučaju kad su se svi događaji dogodili u različitim trenucima. U praksi je zbog načina bilježenja vremena taj slučaj rijedak, tj. više pojedinaca imaju ista vremena događaja. Neka su različita opažena vremena označena s $t_1 < \ldots < t_D$, neka je d_i broj događaja u trenutku t_i te neka je $\mathbb{D}_i$ skup svih pojedinaca kojima se događaj dogodio u trenutku t_i. Sa $\mathbf{s}_i$ označit ćemo sumu vektora $\mathbb{X}_j$ po svim subjektima koji dožive događaj u t_i, tj. $s_i = \sum_{j \in \mathbb{D}_i} \mathbb{X}_j$, a R_i predstavljat će skup rizičnih subjekata u trenutku t_i.

Više je različitih prijedloga za konstruiranje funkcije parcijalne vjerodostojnosti, a prvi je Breslowov:

$$L_1(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(\boldsymbol{\beta}^\intercal \mathbf{s}_i\right)}{\left[\sum_{j \in R_i} \exp\left(\boldsymbol{\beta}^\intercal \mathbb{X}_j\right)\right]^{d_i}}.$$

Svaki od d_i događaja ovdje je tretiran posebno te doprinosi funkciji vjerodostojnosti. Kad ima malo istih vremena događaja, ova je aproksimacija dosta dobra i implementirana je u većini statističkih paketa.

Efronov prijedlog je sljedeći:

$$L_2(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(\boldsymbol{\beta}^\intercal \mathbf{s}_i\right)}{\prod_{j=1}^{d_i} \left[\sum_{k\in R_i} \exp\left(\boldsymbol{\beta}^\intercal \mathbb{X}_k\right) - \frac{j-1}{d_i} \sum_{k\in \mathbb{D}_i} \exp\left(\boldsymbol{\beta}^\intercal \mathbb{X}_k\right)\right]},$$

što je bliže pravoj funkciji parcijalne vjerodostojnosti temeljenoj na diskretnom modelu. Kada je broj istih vremena događaja mali, Breslowova i Efronova vjerodostojnost vrlo su bliske. Isti se testovi provode koji su navedeni u prošlom potpoglavlju, jedino s drugačijom funkcijom parcijalne vjerodostojnosti.

2.2 Lokalni testovi

Ponekad nas zanima testiranje hipoteza za podskup parametara $\boldsymbol{\beta}$. Nulta hipoteza u tom je slučaju jednaka $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta_{10}}$, gdje je $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_2^\intercal)^\intercal$. Ovdje je $\boldsymbol{\beta}_1$ $q \times 1$ vektor parametara $\boldsymbol{\beta}$ od interesa i $\boldsymbol{\beta}_2$ je vektor ostalih $p - q$ parametara $\boldsymbol{\beta}$.

Waldov test za navedenu hipotezu baziran je na procjenitelju najveće parcijalne vjerodostojnosti od $\boldsymbol{\beta}$, vektoru $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\intercal, \hat{\boldsymbol{\beta}}_2^\intercal)^\intercal$. Particionirajmo negativnu matricu drugih derivacija logaritmirane funkcije parcijalne vjerodostojnosti (matricu $\mathbf{I}$) u obliku:

$$\mathbf{I}_p = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{pmatrix},$$

gdje je $\mathbf{I}_{11}$ $q \times q$ podmatrica matrice $\mathbf{I}$ s obzirom na derivacije po varijablama vektora $\boldsymbol{\beta}_1$, $\mathbf{I}_{22}$ je podmatrica derivacija po $\boldsymbol{\beta}_2$, a $\mathbf{I}_{12}$ i $\mathbf{I}_{21}$ su matrice miješanih drugih derivacija. Waldova testna statistika je:

$$\chi^2_W = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^\intercal \left[\mathbf{I}^{11}(\hat{\boldsymbol{\beta}})\right]^{-1} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \overset{H_0}{\sim} \chi^2(q).$$

U gornjoj formuli $\mathbf{I}^{11}(\hat{\boldsymbol{\beta}})$ predstavlja gornju $q \times q$ podmatricu od $\mathbf{I}_p^{-1}(\hat{\boldsymbol{\beta}})$.

Test omjera vjerodostojnosti (engl. *likelihood ratio*) testira hipotezu $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta_{10}}$ testnom statistikom:

$$\chi^2_{LR} = 2\left\{LL(\hat{\boldsymbol{\beta}}) - LL\left[\boldsymbol{\beta}_{10}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{10})\right]\right\} \overset{H_0}{\sim} \chi^2(q),$$

gdje je s $\hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{10})$ označen MLE procjenitelj od $\boldsymbol{\beta}_2$ baziran na logaritmiranoj vjerodostojnosti s prvih q parametara $\boldsymbol{\beta}$ fiksiranih na vrijednosti $\boldsymbol{\beta}_{10}$.

Istu hipotezu možemo testirati i testom skorova sa statistikom:

$$\chi^2_{SC} = \mathbf{U}_1[\boldsymbol{\beta}_{10}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{10})]^\intercal \left[\mathbf{I}^{11}(\boldsymbol{\beta}_{10}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{10}))\right] \mathbf{U}_1[\boldsymbol{\beta}_{10}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{10})] \overset{H_0}{\sim} \chi^2(q).$$

Ovdje je $\mathbf{U}_1[\boldsymbol{\beta}_{10}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{10})]$ $q \times 1$ vektor skorova za $\boldsymbol{\beta}_1$, evaluiran u hipotetskoj vrijednosti $\boldsymbol{\beta}_{10}$ i restringiranom MLE procjenitelju od $\boldsymbol{\beta}_2$. Sve navedene statistike imaju hi-kvadrat distribuciju za velike uzorke.

2.3 Interakcije

U statističkim analizama često se uzima produkt nekih nezavisnih varijabli kako bi se utvrdilo jesu li one u interakciji. Za dvije nezavisne varijable kažemo da su u interakciji ako veza između bilo koje od tih varijabli i između zavisne varijable ovisi o vrijednosti druge nezavisne varijable u interakciji. U praksi, to znači da je teže predvidjeti posljedice mijenjanja vrijednosti varijable, osobito ako su varijable u interakciji kompliciranе za izmjeriti.

Interakciju varijabli X_1 i X_2 označavamo s $X_1 \times X_2$, a u Coxovom regresijskom modelu prisustvo interakcije izgledalo bi ovako:

$$h(t \mid (X_1, X_2)) = h_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2\right).$$

Izlučimo li glavne varijable, dobijemo isti model prikazan na dva načina:

$$h(t \mid (X_1, X_2)) = h_0(t) \exp\left((\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2\right)$$
$$\text{i } h(t \mid (X_1, X_2)) = h_0(t) \exp\left(\beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2\right).$$

Vidimo da interakcije neće postojati ukoliko je u obje jednadžbe koeficijent β_3 jednak nuli. Tada nijedna nezavisna varijabla neće imati doprinos na utjecaj druge nezavisne varijable na zavisnu varijablu. Prisustvo interakcija u modelu testiramo uz nultu hipotezu $H_0 : \beta_3 = 0$ nekim od lokalnih testova iz prošlog potpoglavlja.

2.4 Proporcionalnost hazarda

Coxov regresijski model još se zove i Coxov model proporcionalnog hazarda, dakle pretpostavka proporcionalnosti hazarda toliko je bitna da često stoji i u naslovu modela. Pretpostavka zahtijeva da je omjer hazarda za bilo koja dva subjekta konstantan kroz vrijeme, tj. hazardi su proporcionalni. Ključnu ulogu u testiranju je li pretpostavka modela zadovoljena imaju vremenski zavisne varijable.

Za fiksiranu kovarijatu X_1 stvorimo na umjetan način varijablu ovisnu o vremenu $X_2(t)$ kao interakciju naše varijable i funkcije ovisne o vremenu:

$$X_2(t) = X_1 \times g(t),$$

gdje je $g(t)$ neka poznata funkcija, najčešće se uzima $g(t) = \ln t$. Sada je Coxov model dan s

$$h(t \mid X_1) = h_0(t) \exp\left(\beta_1 X_1 + \beta_2 (X_1 \times g(t))\right),$$

dakle ukoliko usporedimo subjekte s vrijednostima kovarijata X_1 i X_1^*, njihov pripadni omjer hazarda je

$$\frac{h(t \mid X_1)}{h(t \mid X_1^*)} = \exp\left(\beta_1 [X_1 - X_1^*] + \beta_2 g(t)[X_1 - X_1^*]\right),$$

što ovisi o t ako β_2 nije jednak nuli. Stoga je test s hipotezom $H_0 : \beta_2 = 0$ test za testiranje pretpostavke proporcionalnog hazarda.

Što kad pretpostavka nije zadovoljena?

Kada pretpostavka proporcionalnosti hazarda nije zadovoljena i promatrana varijabla X_1 je dihotomna, jedan pristup rješavanju problema neproporcionalnosti je uvođenje nove varijable ovisne o vremenu u model. Varijablu definiramo kao

$$X_2(t) = \begin{cases} 0, & \text{ako je } t \leq \tau \\ X_1, & \text{ako je } t > \tau \end{cases}$$

Sada imamo model proporcionalnog hazarda dan po slučajevima sa

$$h(t \mid X(t)) = \begin{cases} h_0(t) \exp\left(\beta_1 X_1\right), & \text{ako je } t \leq \tau \\ h_0(t) \exp\left[(\beta_1 + \beta_2) X_1\right], & \text{ako je } t > \tau \end{cases}$$

Ovdje je $\exp(\beta_1)$ omjer hazarda za subjekta s kovarijatom $X_1 = 1$ u odnosu na subjekta s $X_1 = 0$, do trenutka τ, a nakon trenutka τ taj je omjer jednak $\exp(\beta_1 + \beta_2)$.

Dakle omjer hazarda se u „točki promjene" τ poveća $\exp(\beta_2)$ puta. Kako bismo odredili optimalnu točku promjene, isprobamo nekoliko vrijednosti i odaberemo onu koja maksimizira logaritmiranu parcijalnu funkciju vjerodostojnosti. Za vrijednosti koje isprobavamo kao potencijalne vrijednosti τ uzimamo vremena događaja subjekata.

Stratifikacija modela

Prošlo rješenje bilo je samo za dihotomne varijable. Generalno, podatke možemo podijeliti u grupe (tzv. strate) po kovarijati (ili više njih) koja ne zadovoljava pretpostavku proporcionalnog hazarda. Ako je kovarijata po kojoj klasificiramo neprekidna, grupe ćemo stvarati na temelju podjele stratifikacijske kovarijate u neke smislene intervale. Sada primijenimo Coxov model s ostalim kovarijatama na svaku grupu posebno, dakle kovarijata po kojoj smo grupirali i koja ne zadovoljava pretpostavku proporcionalnog hazarda neće biti u modelu. Subjekti u j-tom stratumu imaju proizvoljnu osnovnu funkciju hazarda $h_{0j}(t)$ te je efekt ostalih prediktorskih varijabli na funkciju hazarda te grupe dan modelom

$$h_j[t \mid \mathbb{X}(t)] = h_{0j}(t) \exp\left[\boldsymbol{\beta}^\intercal \mathbb{X}(t)\right]$$

za $j = 1, \ldots, s$ strata. U ovom su modelu regresijski koeficijenti isti za sve grupe, a osnovne funkcije hazarda mogu biti različite i nimalo povezane.

Procjenu parametara i testiranje hipoteza provodimo kao i prije, gdje je logaritmirana funkcija parcijalne vjerodostojnosti dana s

$$LL(\boldsymbol{\beta}) = LL_1(\boldsymbol{\beta}) + \ldots + LL_s(\boldsymbol{\beta}),$$

gdje je $LL_j(\boldsymbol{\beta})$ logaritmirana funkcija parcijalne vjerodostojnosti koja koristi podatke iz j-tog stratuma.

Poglavlje 3

Primjena analize doživljenja na istraživanje o pSS-u i NHL-u

3.1 Uvod u problem i podatke

Primarni Sjoegrenov sindrom (pSS) je autoimuna, više-organska bolest koja se u većini slučajeva očituje osjećajem suhoće usta i očiju. Do navedenih simptoma dolazi uoči napada imunološkog sustava na tkivo žlijezda slinovnica te suznih žlijezda. U žljezdanim tkivima dolazi do upale, a posljedično do smanjene proizvodnje suza i sline. Sjoegrenov sindrom najčešće se javlja u srednjoj životnoj dobi, između 40 i 60 godina, a poznato je da su čak 90% oboljelih osobe ženskog spola. Bolest je nazvana po švedskom oftalmologu Henriku Sjoegrenu koji je početkom dvadesetog stoljeća prvi opisao ovu bolest.

Poznato je da u usporedbi s općom populacijom, bolesnici sa pSS-om imaju povećan rizik od razvoja **ne-Hodgkinovog limfoma (NHL-a)**, tumora koji nastaje zloćudnom preobrazbom stanica limfocitnog reda. Pod nazivom ne-Hodgkinov limfom, podrazumijeva se više od 30-ak vrsta limfoma, koji se međusobno razlikuju po tipu stanica iz kojih su nastali, brzini rasta, osjetljivosti na liječenje i prognozi. U svijetu NHL čine 5% svih malignih bolesti i uzrok su $3,4\%$ smrti od malignih bolesti. Stopa incidencije NHL-a u Republici Hrvatskoj u 2007. godini iznosila je 10 bolesnika na 100 000 stanovnika.

Tokom prošle godine u Kliničkom bolničkom centru u Zagrebu provedeno je jedno

zanimljivo istraživanje[1] na temu ovih dviju bolesti. Ono je bilo potaknuto gore navedenom činjenicom o većem riziku razvoja NHL-a ukoliko bolesnik već boluje od pSS-a. Budući da u Hrvatskoj još nema objavljenih epidemioloških podataka o učestalosti NHL-a u osoba sa pSS-om, cilj ovog istraživanja bio je odrediti stopu incidencije NHL-a kod bolesnika sa pSS-om liječenih u KBC-u Zagreb te ustanoviti obolijevaju li ti bolesnici od NHL-a više od opće populacije RH. U tu su svrhu pregledani medicinski kartoni bolesnika liječenih u KBC-u Zagreb u razdoblju od 2011. do 2021. godine te su u obzir uključene osobe koje su ispunjavale zajedničke klasifikacijske kriterije za pSS Američkog reumatološkog društva i Europske lige protiv reumatizma iz 2016. godine. Napomenimo da je istraživanje provedeno na Zavodu za kliničku imunologiju i reumatologiju na KBC-u Zagreb, zato su osobe ispunjavale kriterije gornjih centara te se u prikupljenoj bazi podataka nalaze i podaci o lijekovima protiv reumatoidnog artritisa. U ovom radu koristit ću navedenu bazu podataka, dakle sve osobe u bazi boluju od pSS-a, a neke od njih su razvile i NHL. Međutim, fokus neće biti na usporedbi stopa incidencija NHL-a u kohorti s općom populacijom, već na vremenu potrebnom za razvoj NHL-a u bolesnika sa pSS-om te na pitanju postoje li još neki čimbenici koji bi mogli utjecati na to vrijeme.

Promatrani događaj od interesa je „razvoj NHL-a". Za neke osobe događaj se dogodi do kraja istraživanja, a za neke ne. U oba je slučaja početak mjerenja vremena doživljenja datum dijagnosticiranja pSS-a, a za bolesnike koji razviju NHL ono završava u tretnutku njegovog dijagnosticiranja. Za pojedince koji nisu razvili NHL, vrijeme doživljenja je cenzurirano, tj. uzimamo da ono isto traje od dijagnosticiranja pSS-a pa do kraja istraživanja, ali uzimamo u obzir činjenicu da se za te osobe događaj nije dogodio. Svi grafovi i izračuni napravljeni su u programskom jeziku SAS (poveznica je na [1]).

3.2 Kratki opis varijabli

Ispišimo i objasnimo prvo sve prediktorske (nezavisne) varijable.

Protutijela:

- SSA,
- RF.

[1] Istraživanje su proveli Matea Martinić, Mirna Reihl Crnogaj, Branimir Anić te Miroslav Mayer, a sažetak istraživanja može se vidjeti na [8].

SSA ili Sjoegren specifična antitijela vrsta su autoantitijela (usmjerena su protiv antigena vlastitog tijela umjesto protiv stranih antigena) tipična za Sjoegrenov sindrom, ali ne i specifična jer mogu biti prisutna i u drugim sindromima i stanjima tijela. RF ili reumatoidni faktor su protutijela na protutijela koja stvaraju komplekse (spoj dvaju protutijela). Javljaju se u reumatoidnom artritisu, ali i u drugim autoimunim i kroničnim upalnim te malignim bolestima. Mogu biti odraz „restrikcije klonova" u Sjoegrenovoj bolesti pa samim time povećavaju rizik limfoma u Sjoegrenu. Budući da pojedinac može ili ne mora razviti navedena antitijela, obje varijable su dihotomne (razvio je antitijela ili ih nije razvio).

Lijekovi koje neki od ispitanika uzimaju:

- antimalarik,
- azatioprin,
- metotreksat,
- ciklofosfamid.

Prva su tri lijeka specifična za reumatoidni artritis, a ciklofosfamid se često koristi sam ili zajedno s nekim drugim lijekovima za liječenje tumora, ali liječnici ga mogu prepisati i u slučaju nekih drugih stanja. Svi se ovi lijekovi prepisuju kod težih oblika Sjoegrenove bolesti, a kao jači oblik imunosupresije mogu i povećati rizik od limfoma. Varijable su također dihotomne - osoba koristi lijek ili ne.

Ostale varijable:

- spol - dihotomna varijabla
- dob_dijagnoze - s koliko je godina osobi dijagnosticiran pSS; kontinuirana varijabla

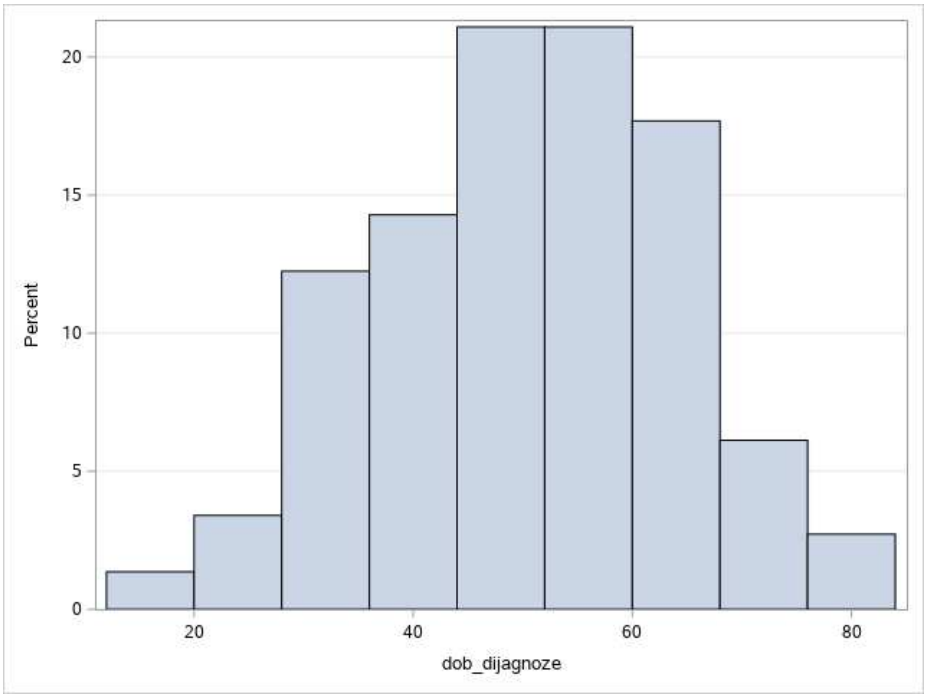
Zavisne varijable:

- t_dijagnoza - vrijeme (u godinama) od trenutka dijagnoze do razvitka NHL-a ili do kraja istraživanja; kontinuirana varijabla
- limfom - indikatorska varijabla (daje informaciju je li se događaj dogodio (razvoj limfoma) ili je podatak cenzuriran); dihotomna varijabla

3.3 Opisna statistika i analiza stanja

Tablica 3.1: Deskriptivne statistike za varijable dob_dijagnoze i t_dijagnoza (ispis iz SAS-a)

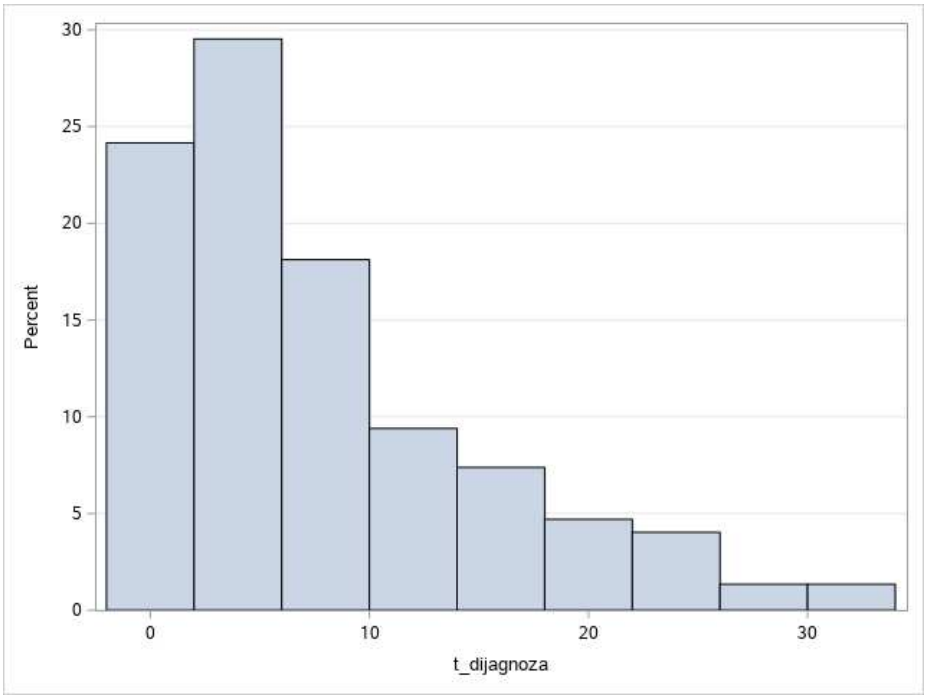
Variable	Mean	Std Dev	Variance	N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
dob_dijagnoze	49.993	13.826	191.157	147	14.000	40.000	50.000	60.000	83.000
t_dijagnoza	7.215	7.336	53.818	149	0.000	2.000	5.000	10.000	32.000



Slika 3.1: Histogram varijable dob_dijagnoze

Ovo potpoglavlje započinjemo analizom kontinuiranih varijabli. Najmlađa osoba kojoj je dijagnosticiran pSS imala je svega 14 godina, a najstarija 83 godine. Pola ispitanika ovu je bolest razvilo do 50. godine života. Ukoliko pogledamo histogram, vidimo da je distribucija opažene dobi u kojoj se razvio pSS otprilike zvonolikog oblika. Također, iz činjenice da su medijan (50 godina) i aritmetička sredina (49.993 godina) skoro jednake vrijednosti možemo naslutiti da se radi o simetričnoj razdiobi. Kolmogorov-Smirnovljev test daje p-vrijednost veću od 0.15 pa na standardnoj ra-

zini značajnosti od 5% ne odbacujemo nultu hipotezu o normalnoj distribuiranosti varijable dob_dijagnoze.



Slika 3.2: Histogram varijable t_dijagnoza

Prije nego što počnemo analizirati najzanimljiviju varijablu - t_dijagnoza, tj. vrijeme od dijagnosticiranja pSS-a do razvitka NHL-a (ili do cenzuriranja), napomenimo da je u istraživanju sudjelovalo 154 ispitanika te je od 154 osobe samo njih 9 razvilo limfom. Možda ovaj broj na prvu ne djeluje značajno niti veliko, međutim u medicini je 9/154 velik udio i nikako neznačajan. Međutim, to znači da imamo jako puno cenzuriranih podataka, tj. da varijabla t_dijagnoza najčešće označava vrijeme od dijagnoze pSS-a pa sve do trenutka cenzuriranja. Svako cenzuriranje u ovom istraživanju je desno cenzuriranje te je uslijedilo zbog završetka istraživanja (u trenutku zadnjeg pregleda), a za 5 osoba ovaj je podatak izgubljen. Stoga, u većini slučajeva varijabla t_dijagnoza zapravo označava koliko je vremena prošlo od dijagnoze pSS-a pa do zadnjeg pregleda, a u nekoliko slučajeva vrijeme proteklo od dijagnoze pSS-a pa do razvitka NHL-a. Distribucija ove varijable pozitivno je asimetrična, 25% vremena manje je ili jednako od 2 godine te je mnogo vrijednosti jednako 0, odnosno pojedinci su kratko sudjelovali u istraživanju, neki čak ni nepunu godinu. Prosječno vrijeme od dijagnoze pSS-a pa do razvitka NHL-a ili do cenzuriranja (zadnjeg pregleda) je 7.215 godina, a jednoj je

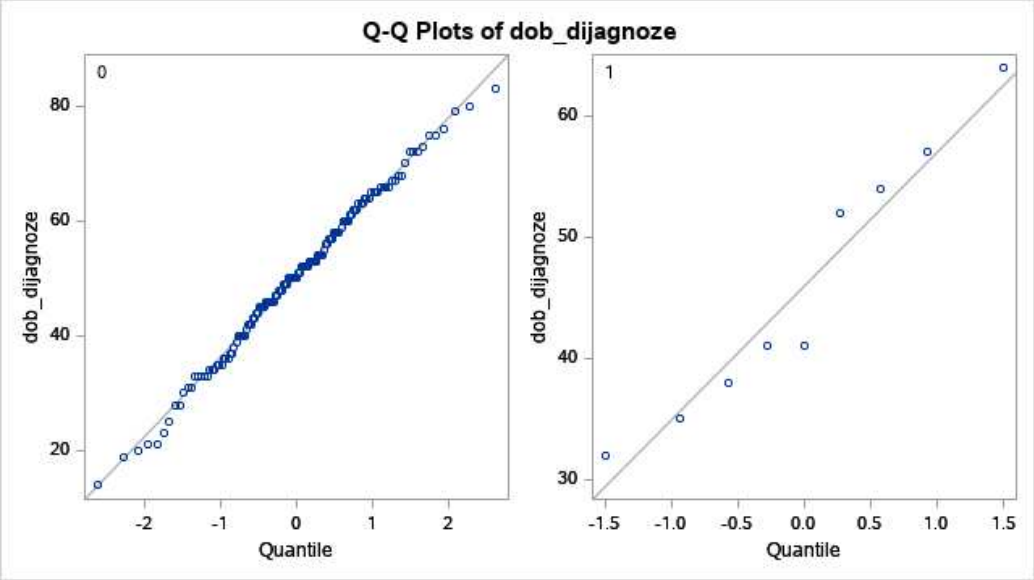
osobi prošlo 32 godine od dijagnoze pSS-a i još uvijek nije razvila NHL. Pogledajmo u sljedećoj tablici (br. 3.2) kakva je situacija sa subjektima koji su razvili limfom.

Tablica 3.2: Deskriptivne statistike za varijablu t_dijagnoza za osobe koje su razvile NHL

Analysis Variable : t_dijagnoza								
Mean	Std Dev	Variance	N	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
6.3333333	7.4330344	55.2500000	9	0	0	4.0000000	7.0000000	21.0000000

Kao što je već napomenuto, samo je 9 osoba koje su razvile limfom. Nekima je iste godine kada im je dignosticiran pSS ujedno dijagnosticiran i NHL, a ostalima je između ta dva događaja prošlo više godina, nekome čak dvadeset i jedna godina. U prosjeku je između te dvije dijagnoze prošlo 6.33 godina.

Usporedimo sad ove dvije varijable po grupama - jedna grupa sastojat će se od osoba koje su razvile limfom, a druga od osoba koje nisu. Već smo vidjeli da varijabla dob_dijagnoze prati normalnu distribuciju pa ćemo za testiranje prosječne dobi dijagnoze pSS-a između dvije grupe koristiti t-test. T-test za dva nezavisna uzorka koristi se za utvrđivanje jesu li očekivane vrijednosti dviju populacija statistički značajno različite pa se za nultu hipotezu uzima jednakost očekivanja tih dviju populacija. Pretpostavke testa su zadovoljene - uzorci su nezavisni i normalno distribuirani (može se vidjeti i iz kvantil-kvantil grafa (engl. *QQ plota*) sa slike 3.3). Treća pretpostavka t-testa je homogenost varijanci dviju populacija. Ukoliko taj uvjet nije zadovoljen, svejedno se može provesti t-test, ali s tzv. Satterthwaite korekcijom.



Slika 3.3: Kvantil-kvantil grafovi za grupe - lijevo bez razvijenog limfoma i desno s razvijenim limfomom - grupiranje podataka oko pravca označava normalnost

Tablica 3.3: Rezultati t-testa za varijablu dob_dijagnoze (ispis iz SAS-a)

Variable: dob_dijagnoze

limfom	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		138	50.2536	13.9813	1.1902	14.0000	83.0000
1		9	46.0000	11.0454	3.6818	32.0000	64.0000
Diff (1-2)	Pooled		4.2536	13.8355	4.7599		
Diff (1-2)	Satterthwaite		4.2536		3.8694		

limfom	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		50.2536	47.9002	52.6071	13.9813	12.5036	15.8581
1		46.0000	37.5098	54.4902	11.0454	7.4607	21.1604
Diff (1-2)	Pooled	4.2536	-5.1540	13.6613	13.8355	12.4099	15.6342
Diff (1-2)	Satterthwaite	4.2536	-4.3976	12.9048			

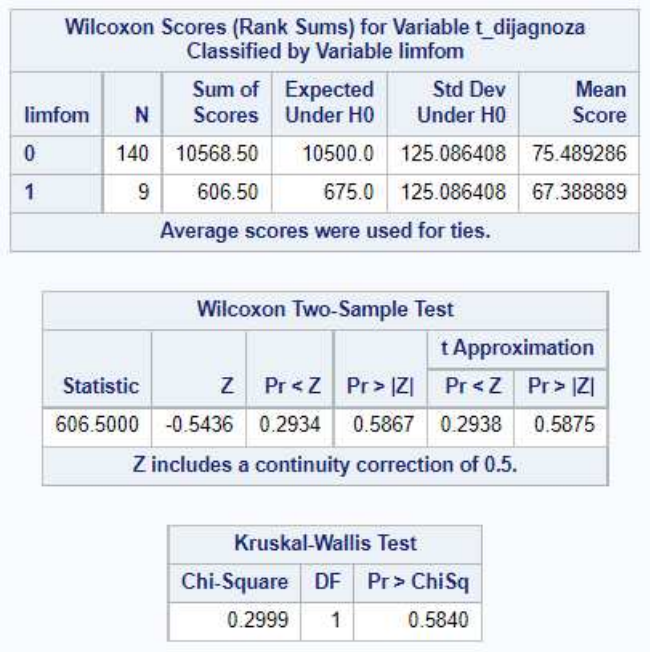
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	145	0.89	0.3730
Satterthwaite	Unequal	9.7531	1.10	0.2980

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	137	8	1.60	0.4871

Rezultati t-testa prikazani su u tablici 3.3. Prvo provjerimo je li zadovoljena jednakost varijanci populacija (ljudi sa i bez razvijenog limfoma) pa ovisno o tome biramo jednu od dviju metoda za izračun t statistike. P-vrijednost nulte hipoteze o jednakosti varijance ispada 0.4871 pa tu hipotezu ne odbacujemo te gledamo p-vrijednost *pooled* metode. Ona iznosi 0.3730 pa ni ovdje ne odbacujemo nultu hipotezu o jednakoj očekivanoj dobi razvijanja pSS-a između osoba sa i bez razvijenog limfoma, odnosno te dvije grupe ljudi otprilike imaju jednaku očekivanu dob razvitka Sjoegrenovog sindroma.

Za varijablu t_dijagnoza vidjeli smo da nije normalno distribuirana pa za nju ne možemo provesti t-test. Mogli bismo joj uzeti logaritam i provjeriti prati li ta transformacija varijable normalnu razdiobu, međutim i ova opcija otpada jer je mnogo vrijednosti varijable t_dijagnoza jednako nuli pa bismo imali problema prilikom logaritmiranja. Stoga ćemo za usporedbu vremena doživljenja za grupe ljudi sa i bez limfoma koristiti neparametarski Mann-Whitneyjev U-test (još se naziva i Wilcoxonov T-test). Ovaj test također zahtijeva nezavisnost uzorka te umjesto originalnih podataka pri računanju testne statistike koristi rangove pridružene podacima (poredanim u jedan zajednički uzorak). Test ispituje nultu hipotezu koja pretpostavlja jednakost medijana dviju populacija iz kojih dolaze uzorci. U promatranom istraživanju p-vrijednost ispada 0.5867 (tablica br. 3.4) pa ne odbacujemo nultu hipotezu o istoj distribuciji, tj. smatramo da osobe sa i bez limfoma imaju otprilike jednako vrijeme od uspostave dijagnoze pSS-a do otkrića limfoma/zadnjeg pregleda (tj. cenzuriranja).

Tablica 3.4: Rezultati Mann-Whitneyjevog testa za varijablu t_dijagnoza (ispis iz SAS-a)



Wilcoxon Scores (Rank Sums) for Variable t_dijagnoza Classified by Variable limfom

limfom	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	140	10568.50	10500.0	125.086408	75.489286
1	9	606.50	675.0	125.086408	67.388889

Average scores were used for ties.

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
606.5000	-0.5436	0.2934	0.5867	0.2938	0.5875

Z includes a continuity correction of 0.5.

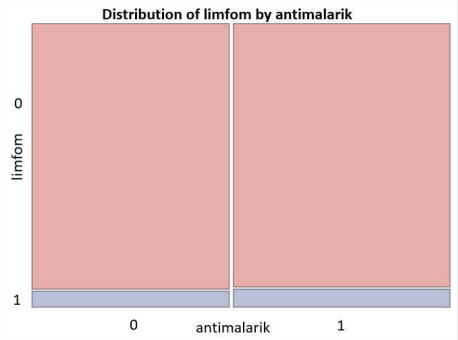
Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
0.2999	1	0.5840

Konačno dolazimo i do ostalih, dihotomnih varijabli. Prilikom njihove analize provest ćemo χ^2-testove za provjeru nezavisnosti svake od varijabli s varijablom limfom (govori jesu li osobe razvile limfom ili ne). Tako ćemo odmah moći vidjeti iz kontingencijskih tablica i raspodjelu frekvencija po kategorijama obiju varijabli, a i testirati povezanost varijabli statističkim testom. Testna statistika χ^2-testa računa se kao suma po ćelijama kontingencijske tablice kvadriranih razlika opaženih i očekivanih vrijednosti, podijeljenih s očekivanim vrijednostima. Na sljedećih nekoliko stranica vidjet ćemo za svaku varijablu ispis iz SAS-a koji prikazuje rezultate χ^2-testa. Prva slika bit će kontingencijska tablica za varijable limfom i drugu varijablu od interesa, druga slika bit će mozaik plot, grafički prikaz kontingencijskih tablica te na trećoj slici bit će izračunate testne statistike i p-vrijednosti. Budući da se na svakoj slici vidi da je određeni postotak očekivanih vrijednosti manji od 5, uzet ćemo p-vrijednost Fisherovog egzaktnog testa koji je pouzdaniji u ovakvom slučaju.

Frequency Expected Percent Row Pct Col Pct	Table of limfom by antimalarik		
	antimalarik		
limfom	0	1	Total
0	67 66.711 44.97 47.86 94.37	73 73.289 48.99 52.14 93.59	140 93.96
1	4 4.2886 2.68 44.44 5.63	5 4.7114 3.36 55.56 6.41	9 6.04
Total	71 47.65	78 52.35	149 100.00
Frequency Missing = 5			

(a) Tablica frekvencija



(b) Mozaični prikaz distribucije limfoma i antimalarika

Statistics for Table of limfom by antimalarik

Statistic	DF	Value	Prob
Chi-Square	1	0.0395	0.8425
Likelihood Ratio Chi-Square	1	0.0396	0.8423
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0392	0.8430
Phi Coefficient		0.0163	
Contingency Coefficient		0.0163	
Cramer's V		0.0163	

WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

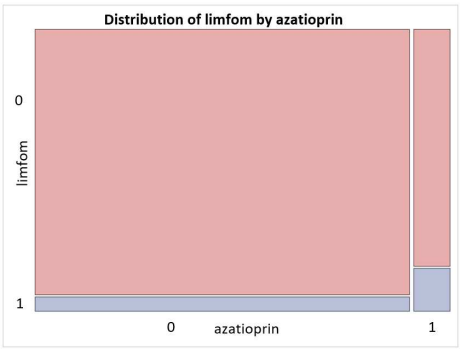
Fisher's Exact Test	
Cell (1,1) Frequency (F)	67
Left-sided Pr <= F	0.7038
Right-sided Pr >= F	0.5592
Table Probability (P)	0.2631
Two-sided Pr <= P	1.0000

(c) χ^2-test i Fisherov egzaktni test

Slika 3.4: Rezultati analize: tablica frekvencija (a), mozaični grafički prikaz distribucije limfoma (b), χ^2 i Fisherov egzaktni test za varijablu antimalarik (ispis iz SAS-a)

Frequency Expected Percent Row Pct Col Pct	Table of limfom by azatioprin		
	azatioprin		
limfom	0	1	Total
0	128 126.79 86.49 92.09 94.81	11 12.209 7.43 7.91 84.62	139 93.92
1	7 8.2095 4.73 77.78 5.19	2 0.7905 1.35 22.22 15.38	9 6.08
Total	135 91.22	13 8.78	148 100.00
Frequency Missing = 6			

(a) Tablica frekvencija



(b) Mozaični prikaz distribucije limfoma i azatioprina

Statistics for Table of limfom by azatioprin

Statistic	DF	Value	Prob
Chi-Square	1	2.1599	0.1417
Likelihood Ratio Chi-Square	1	1.6169	0.2035
Continuity Adj. Chi-Square	1	0.7432	0.3886
Mantel-Haenszel Chi-Square	1	2.1453	0.1430
Phi Coefficient		0.1208	
Contingency Coefficient		0.1199	
Cramer's V		0.1208	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	128
Left-sided Pr <= F	0.9670
Right-sided Pr >= F	0.1803
Table Probability (P)	0.1474
Two-sided Pr <= P	0.1803

(c) χ^2-test i Fisherov egzaktni test

Slika 3.5: Rezultati analize: tablica frekvencija (a), mozaični grafički prikaz distribucije limfoma (b), χ^2 i Fisherov egzaktni test za varijablu azatioprin (ispis iz SAS-a)

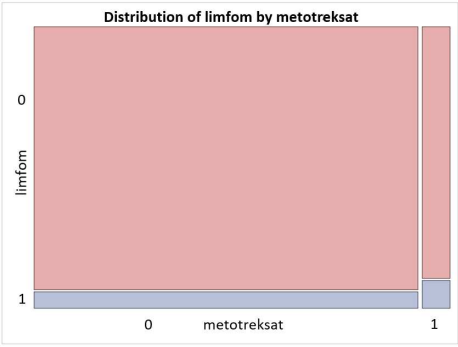
Frequency
Expected
Percent
Row Pct
Col Pct

Table of limfom by metotreksat

limfom	metotreksat 0	metotreksat 1	Total
0	129 128.61 87.76 93.48 94.16	9 9.3878 6.12 6.52 90.00	138 93.88
1	8 8.3878 5.44 88.89 5.84	1 0.6122 0.68 11.11 10.00	9 6.12
Total	137 93.20	10 6.80	147 100.00

Frequency Missing = 7

(a) Tablica frekvencija



(b) Mozaični prikaz distribucije limfoma i metotreksata

Statistics for Table of limfom by metotreksat

Statistic	DF	Value	Prob
Chi-Square	1	0.2807	0.5963
Likelihood Ratio Chi-Square	1	0.2414	0.6232
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.2788	0.5975
Phi Coefficient		0.0437	
Contingency Coefficient		0.0437	
Cramer's V		0.0437	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	129
Left-sided Pr <= F	0.8839
Right-sided Pr >= F	0.4794
Table Probability (P)	0.3632
Two-sided Pr <= P	0.4794

(c) χ^2-test i Fisherov egzaktni test

Slika 3.6: Rezultati analize: tablica frekvencija (a), mozaični grafički prikaz distribucije limfoma (b), χ^2 i Fisherov egzaktni test za varijablu metotreksat (ispis iz SAS-a)

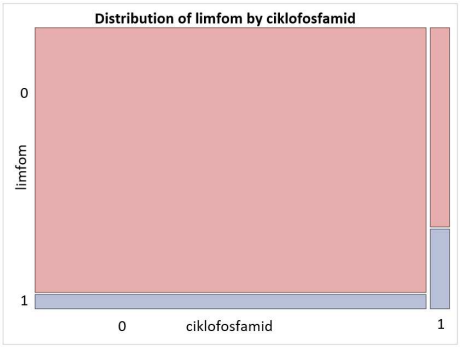
Frequency
Expected
Percent
Row Pct
Col Pct

Table of limfom by ciklofosfamid

limfom	ciklofosfamid 0	ciklofosfamid 1	Total
0	133 131.43 90.48 96.38 95.00	5 6.5714 3.40 3.62 71.43	138 93.88
1	7 8.5714 4.76 77.78 5.00	2 0.4286 1.36 22.22 28.57	9 6.12
Total	140 95.24	7 4.76	147 100.00

Frequency Missing = 7

(a) Tablica frekvencija



(b) Mozaični prikaz distribucije limfoma i ciklofosfamida

Statistics for Table of limfom by ciklofosfamid

Statistic	DF	Value	Prob
Chi-Square	1	6.4446	0.0111
Likelihood Ratio Chi-Square	1	3.7551	0.0526
Continuity Adj. Chi-Square	1	2.9959	0.0835
Mantel-Haenszel Chi-Square	1	6.4007	0.0114
Phi Coefficient		0.2094	
Contingency Coefficient		0.2049	
Cramer's V		0.2094	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	133
Left-sided Pr <= F	0.9950
Right-sided Pr >= F	0.0598
Table Probability (P)	0.0548
Two-sided Pr <= P	0.0598

(c) χ^2-test i Fisherov egzaktni test

Slika 3.7: Rezultati analize: tablica frekvencija (a), mozaični grafički prikaz distribucije limfoma (b), χ^2 i Fisherov egzaktni test za varijablu ciklofosfamid (ispis iz SAS-a)

Sa slika br. 3.4-3.7 zaključujemo da je najkorišteniji lijek antimalarik, koristi ga 78 osoba od njih 154, dakle malo više od polovice. Također, od 9 osoba s razvijenim limfomom njih 5 koristi antimalarik, dakle čini se da između antimalarika i razvitka limfoma nema neke povezanosti, što potvrđuje i p-vrijednost jednaka 1 (s velikom sigurnošću ne odbacujemo nultu hipotezu o nezavisnosti varijabli). Drugi najčešće korišten lijek je azatioprin (13 osoba ili 8.78% - slika 3.5(a)), zatim metotreksat (10 osoba ili 6.80%, slika 3.6(a)) te u konačnici ciklofosfamid (7 osoba ili 4.76% - slika 3.7(a)). Zanimljivo je da je ovo lijek koji se koristi inače najčešće za liječenje tumora te je u ovoj bazi njegova upotreba najrjeđa, a jedini je od svih varijabli kojem je p-vrijednost (koja iznosi 0.0598) na granici odbacivanja nulte hipoteze na nivou značajnosti od 5%, tj. možemo reći da bi mogla postojati određena povezanost između upotrebe ciklofosfamida i razvitka limfoma.

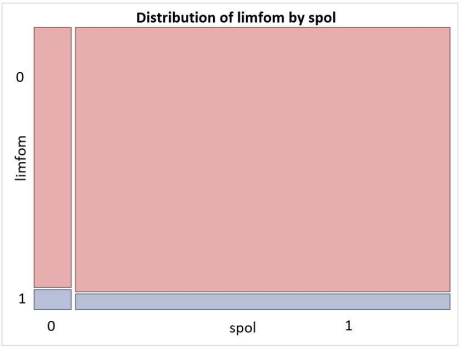
Na slici br. 3.8 muškarci su kodirani s 0, a žene s 1. Vidimo da veliku većinu osoba oboljelih od pSS-a zaista čine žene (140/154 ili 90.91%), kao što je i u uvodu ovog poglavlja već napisano da su 90% osoba oboljelih od pSS-a upravo žene. Međutim, ni ovdje ne možemo odbaciti hipotezu o nezavisnosti spola i razvitka limfoma. Isto tako nultu hipotezu o nezavisnosti ne možemo odbaciti niti za neko od protutijela, tj. nema povezanosti između razvijenih protutijela SSA ili RF i razvitka limfoma. Ono što vidimo je da je većina ispitanika razvila i protutijela SSA i RF (slike br. 3.9 i 3.10).

Frequency
Expected
Percent
Row Pct
Col Pct

Table of limfom by spol

limfom	spol 0	spol 1	Total
0	13 13.182 8.44 8.97 92.86	132 131.82 85.71 91.03 94.29	145 94.16
1	1 0.8182 0.65 11.11 7.14	8 8.1818 5.19 88.89 5.71	9 5.84
Total	14 9.09	140 90.91	154 100.00

(a) Tablica frekvencija



(b) Mozaični prikaz distribucije limfoma i spola

Statistics for Table of limfom by spol

Statistic	DF	Value	Prob
Chi-Square	1	0.0472	0.8280
Likelihood Ratio Chi-Square	1	0.0445	0.8328
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0469	0.8286
Phi Coefficient		-0.0175	
Contingency Coefficient		0.0175	
Cramer's V		-0.0175	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	13
Left-sided Pr <= F	0.5861
Right-sided Pr >= F	0.8090
Table Probability (P)	0.3951
Two-sided Pr <= P	0.5861

(c) χ^2-test i Fisherov egzaktni test

Slika 3.8: Rezultati analize: tablica frekvencija (a), mozaični grafički prikaz distribucije limfoma (b), χ^2 i Fisherov egzaktni test za varijablu spol (ispis iz SAS-a)

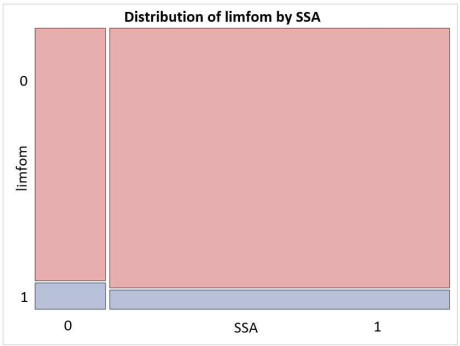
Frequency
Expected
Percent
Row Pct
Col Pct

Table of limfom by SSA

limfom	SSA 0	SSA 1	Total
0	19 19.451 15.57 16.81 90.48	94 93.549 77.05 83.19 93.07	113 92.62
1	2 1.5492 1.64 22.22 9.52	7 7.4508 5.74 77.78 6.93	9 7.38
Total	21 17.21	101 82.79	122 100.00

Frequency Missing = 32

(a) Tablica frekvencija



(b) Mozaični prikaz distribucije limfoma i protutijela SSA

Statistics for Table of limfom by SSA

Statistic	DF	Value	Prob
Chi-Square	1	0.1711	0.6791
Likelihood Ratio Chi-Square	1	0.1606	0.6886
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.1697	0.6804
Phi Coefficient		-0.0374	
Contingency Coefficient		0.0374	
Cramer's V		-0.0374	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	19
Left-sided Pr <= F	0.4808
Right-sided Pr >= F	0.8153
Table Probability (P)	0.2961
Two-sided Pr <= P	0.6520

(c) χ^2-test i Fisherov egzaktni test

Slika 3.9: Rezultati analize: tablica frekvencija (a), mozaični grafički prikaz distribucije limfoma (b), χ^2 i Fisherov egzaktni test za varijablu SSA (ispis iz SAS-a)

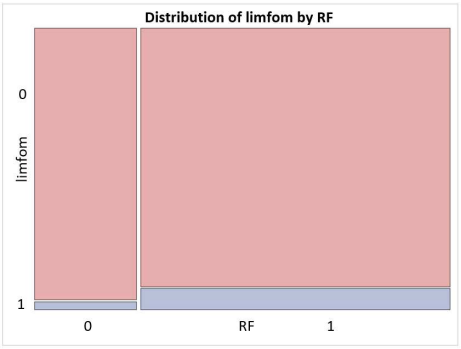
Frequency
Expected
Percent
Row Pct
Col Pct

Table of limfom by RF

limfom	RF 0	RF 1	Total
0	34 32.766 24.11 25.76 97.14	98 99.234 69.50 74.24 92.45	132 93.62
1	1 2.234 0.71 11.11 2.86	8 6.766 5.67 88.89 7.55	9 6.38
Total	35 24.82	106 75.18	141 100.00

Frequency Missing = 13

(a) Tablica frekvencija



(b) Mozaični prikaz distribucije limfoma i protutijela RF

Statistics for Table of limfom by RF

Statistic	DF	Value	Prob
Chi-Square	1	0.9686	0.3250
Likelihood Ratio Chi-Square	1	1.1343	0.2869
Continuity Adj. Chi-Square	1	0.3427	0.5583
Mantel-Haenszel Chi-Square	1	0.9617	0.3268
Phi Coefficient		0.0829	
Contingency Coefficient		0.0826	
Cramer's V		0.0829	

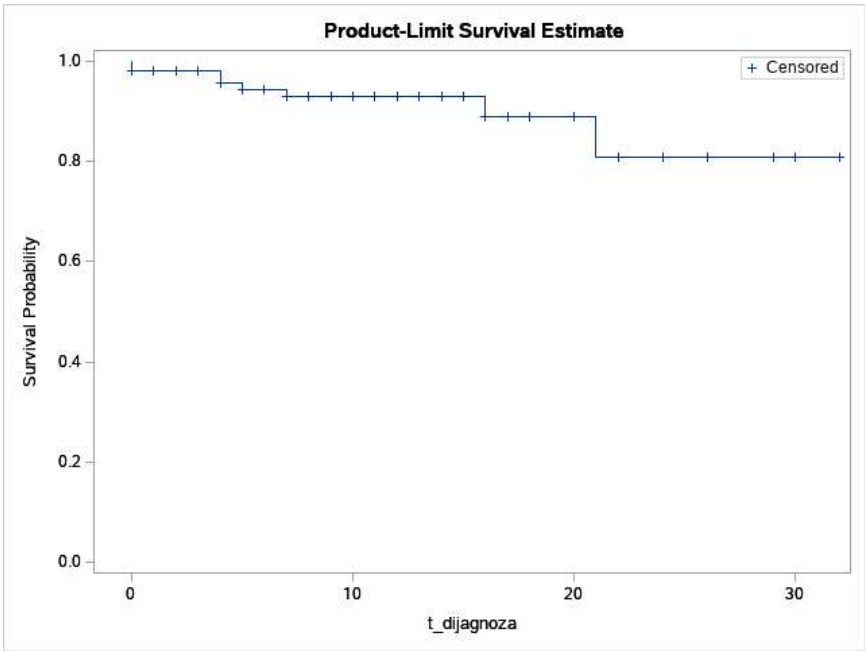
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	34
Left-sided Pr <= F	0.9298
Right-sided Pr >= F	0.2959
Table Probability (P)	0.2257
Two-sided Pr <= P	0.4513

(c) χ^2-test i Fisherov egzaktni test

Slika 3.10: Rezultati analize: tablica frekvencija (a), mozaični grafički prikaz distribucije limfoma (b), χ^2 i Fisherov egzaktni test za varijablu RF (ispis iz SAS-a)

3.4 Kaplan-Meierova procjena funkcija doživljenja



Slika 3.11: Kaplan-Meierova procjena funkcije doživljenja za cijeli uzorak

U ovom potpoglavlju grafički će biti prikazane Kaplan-Meierove (ili kraće KM) procjene funkcija doživljenja. Općenito, Kaplan-Meierova procjena funkcije doživljenja je step funkcija, sa skokovima u trenucima kad se događaj dogodio (odnosno, u našem slučaju, kad se razvio limfom). Ako u zadnjem trenutku KM procjena funkcije doživljenja nije pala do nule, znači da zadnja osoba ima cenzurirano vrijeme te će u tom slučaju vjerojatnost doživljenja ovog trenutka biti pozitivna.

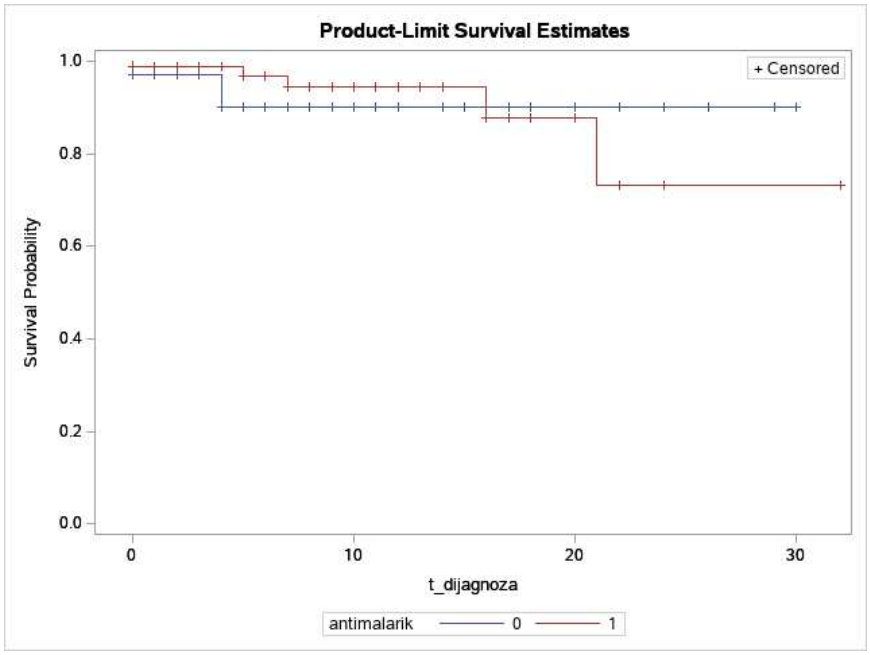
Na gornjem grafu (slika br. 3.11) prikazana je KM procjena za cijeli uzorak. Kao što već znamo, imamo jako puno cenzuriranih podataka pa će vjerojatnost doživljenja općenito biti vrlo velika. Naime, to je zbog činjenice da svaki razvitak HNL-a smanjuje procjenu funkcije doživljenja koja na početku iznosi 1 (100%), a u ovom istraživanju svega je 9 osoba od njih 154 razvilo limfom. U najvećem opaženom vremenskom trenutku (32 godine) vrijeme je cenzurirano, tj. osoba nije razvila limfom. Iz grafa sada vidimo da vjerojatnost da osoba ne razvije limfom 32 godine nakon dijagnoze pSS-a iznosi približno 80%.

Sljedećih 7 slika (br. 3.12-3.18) prikazuju KM procjene funkcija doživljenja po kategorijama dihotomnih varijabli. Budući da se radi o dihotomnim varijablama, bit će na svakom grafu prikazane po dvije funkcije doživljenja. Ukoliko se one vidno razlikuju izgledom, kao kad je npr. jedna funkcija dosta ispod druge, onda možemo naslutiti da grupa osoba s istom karakteristikom (tj. istom vrijednosti dihotomne varijable) ima različitu distribuciju vremena doživljenja u odnosu na drugu grupu ljudi. Slutnju možemo potvrditi pomoću p-vrijednosti *log-rank* testa koje su prikazane u tablici br. 3.5. Ukoliko je p-vrijednost manja od 0.05, odbacit ćemo nultu hipotezu o jednakosti dviju funkcija doživljenja, tj. zaključit ćemo da zaista postoji razlika u vremenu doživljenja između dvije grupe.

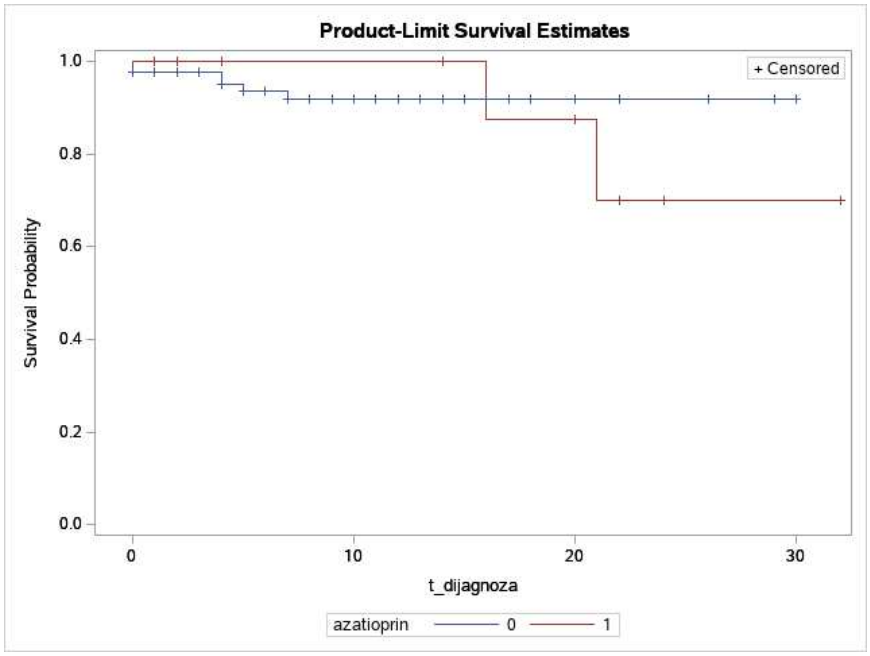
Tablica 3.5: Rezultati *log-rank* testa za usporedbu funkcija doživljenja po analiziranim dihotomnim varijablama

Varijabla	χ^2	df	p-vrijednost
antimalarik	0.0990	1	0.7531
azatioprin	0.1746	1	0.6760
metotreksat	0.0217	1	0.8828
ciklofosfamid	5.6699	1	0.0173
SSA	0.7479	1	0.3871
RF	0.4558	1	0.4996
spol	0.3568	1	0.5503

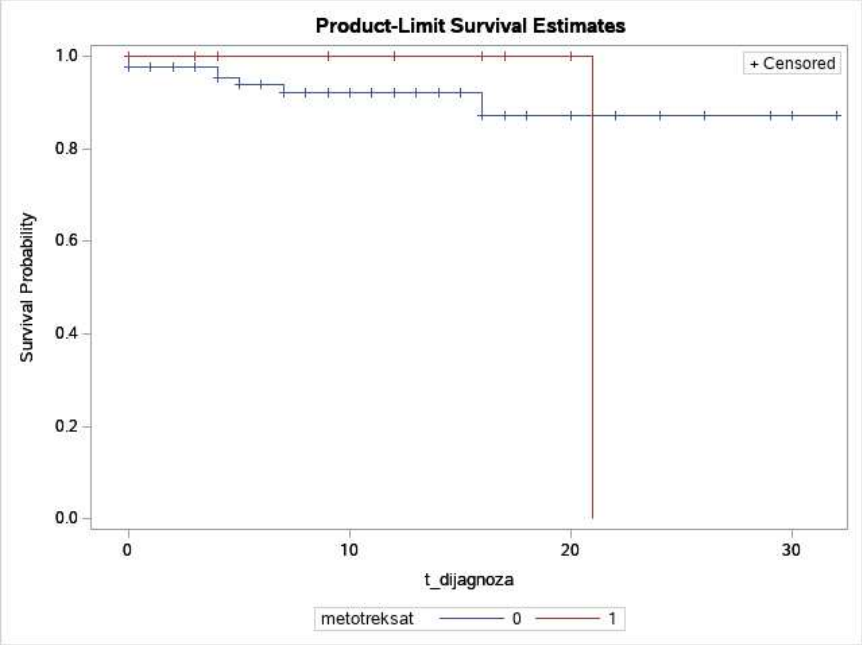
Kod lijekova antimalarik i azatioprin situacija je vrlo slična (slike br. 3.12 i 3.13). Kod oba lijeka vidimo da se funkcije doživljenja osoba koje uzimaju i koje ne uzimaju lijek isprepliću. Zbog isprepletanja funkcija doživljenja naslućujemo da ne možemo tvrditi da se funkcije doživljenja ovih dviju grupi značajno razlikuju. Tvrdnju potkrepljuju i velike p-vrijednosti *log-rank* testa u iznosu od 0.7531 za antimalarik i 0.6760 za azatioprin.



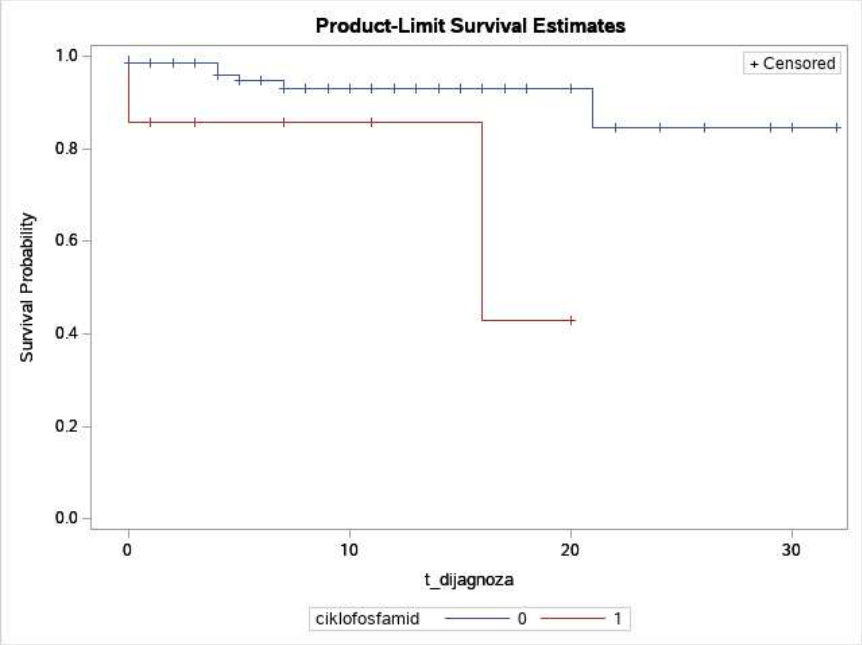
Slika 3.12: Kaplan-Meierova procjena funkcija doživljenja po varijabli antimalarik



Slika 3.13: Kaplan-Meierova procjena funkcija doživljenja po varijabli azatioprin



Slika 3.14: Kaplan-Meierova procjena funkcija doživljenja po varijabli metotreksat



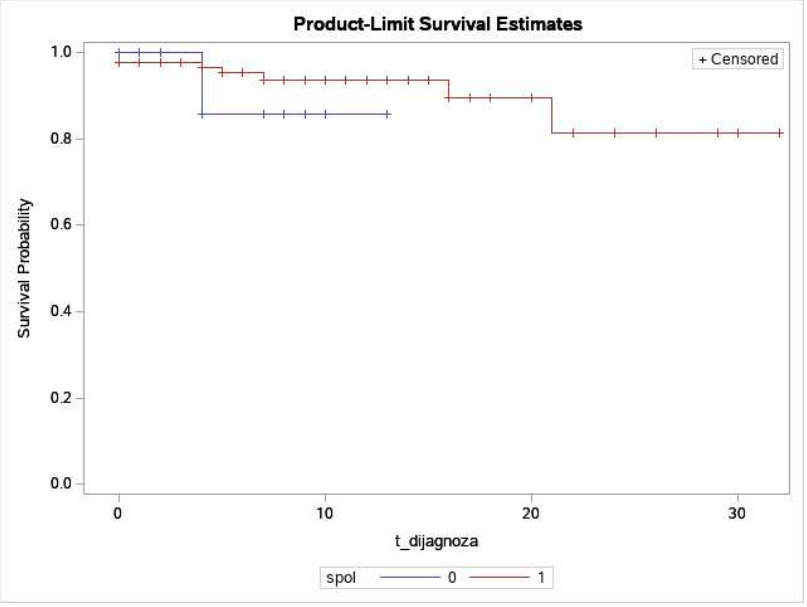
Slika 3.15: Kaplan-Meierova procjena funkcija doživljenja po varijabli ciklofosfamid

Za skupinu ljudi koja koristi metotreksat vidimo da funkcija doživljenja završava i naglo pada u nulu u trenutku $t = 21$ (slika br. 3.14). Razlog tomu je činjenica da

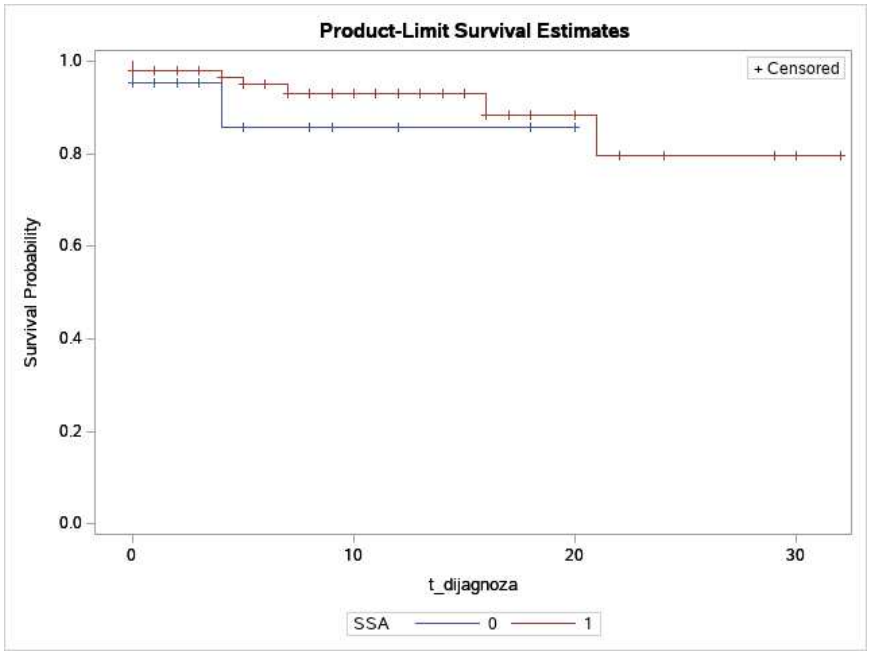
je $t = 21$ najveće opaženo vrijeme doživljenja za neku od osoba koje uzimaju metotreksat. Do tog trenutka sva su vremena cenzurirana, te je ostala jedna rizična osoba dvadeset i jednu godinu nakon dijagnoze pSS-a koja je te iste godine razvila limfom. Vjerojatnost ne razvijanja limfoma u prvih 30 godina od dijagnoze pSS-a za osobe koje ne koriste metotreksat nešto je veća od 80%.

Ciklofosfamid je jedina varijabla kod koje postoji značajna razlika između funkcija doživljenja između grupa te kod koje je p-vrijednost *log-rank* testa manja od standardne razine značajnosti 0.05. P-vrijednost iznosi 0.0173 te tako odbacujemo nultu hipotezu o jednakosti funkcija doživljenja. Zaključujemo sa slike br. 3.15 da je vjerojatnost doživljenja grupe ljudi koja ne koristi ciklofosfamid veća od grupe ljudi koji ga koriste (ti će ljudi vjerojatno prije razviti limfom).

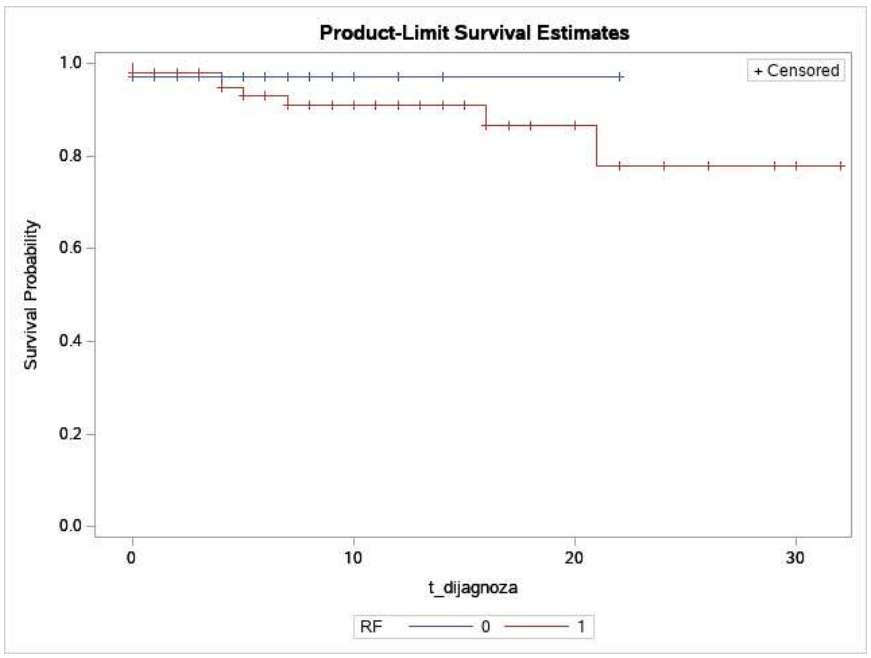
Za preostale varijable - spol i protutijela SSA i RF (slike br. 3.16, 3.17 i 3.18) - funkcije doživljenja isprepliću se negdje u početnim godinama nakon dijagnoze pSS-a te niti u daljnjim godinama nisu pretjerano različite. P-vrijednosti od 0.5503, 0.3871 i 0.4996 ukazuju na činjenicu da nećemo ni ovdje odbaciti nultu hipotezu o jednakosti funkcija doživljenja po kategorijama ovih varijabli.



Slika 3.16: Kaplan-Meierova procjena funkcija doživljenja po varijabli spol



Slika 3.17: Kaplan-Meierova procjena funkcija doživljenja po varijabli SSA



Slika 3.18: Kaplan-Meierova procjena funkcija doživljenja po varijabli RF

3.5 Coxova regresija

Provjerimo prvo je li zadovoljena pretpostavka proporcionalnosti hazarda kako bismo mogli dalje nastaviti s definiranjem modela. Pretpostavku ćemo testirati već spomenutom metodom, koristeći interakcije prediktorskih varijabli i varijabli ovisnih o vremenu, ovdje će to biti funkcija $\ln t$. Za provjeru hipoteze da je koeficijent uz interakciju jednak nuli (tj. da nema utjecaja te interakcije na vrijeme doživljenja) koristimo Waldov test. P-vrijednosti Waldovog testa za umjetno stvorene interakcije prikazane su u tablici 3.6. Vidimo da niti jedna interakcija nije statistički značajna na razini značajnosti od 5% te stoga ne odbacujemo pretpostavku proporcionalnosti hazarda. U modelu ostavljamo isključivo varijable koje smo promatrali i analizirali do sada, bez njihovih interakcija.

Tablica 3.6: P-vrijednosti Waldovog testa za sve prediktorske varijable

Parametar	anti.	azat.	meto.	ciklo.	spol	SSA	RF	dob_dijagnoze
p-vrijednost	0.924	0.998	0.995	0.965	0.385	0.834	0.999	0.082

U tablicama 3.7 i 3.8 prikazani su sljedeći podaci:

- ParEst - procijenjeni parametar $\hat{\beta}$ i StdErr - standardna greška
- χ^2 - chi-kvadrat statistika i p-value - p-vrijednost Waldovog testa za testiranje hipoteze da je parametar jednak nuli (faktor nema utjecaja na vrijeme doživljenja)
- HR - omjer hazarda i 95% HR CI - 95% pouzdani intervali za omjer hazarda.

Prva tablica prikazuje univarijatnu Coxovu regresiju, dakle samo jednu prediktorsku varijablu stavljamo u model te proučavamo njen utjecaj na vrijeme do razvitka limfoma. Ukoliko u model dodamo sve prediktorske varijable, utjecaj pojedinih varijabli na vrijeme doživljenja može se promijeniti zbog potencijalnih interakcija prediktorskih varijabli. Rezultati tog slučaja prikazani su u drugoj tablici, multivarijatni slučaj. Jedina je statistički značajna varijabla lijek ciklofosfamid, za koji smo već pomoću Kaplan-Meierove procjene funkcije doživljenja i rezultata *log-rank* testa vidjeli da se očekivano vrijeme do razvitka NHL-a razlikuje za osobe koje koriste i koje ne koriste lijek. Za tu je varijablu omjer hazarda daleko najveći i u slučaju univarijatne i u slučaju multivarijatne regresije. Ako uzmemo u obzir utjecaj svih prediktorskih varijabli, osobe koje koriste ciklofosfamid imaju 7.904 puta veći rizik razvitka limfoma u odnosu na osobe koje taj lijek ne koriste pa bi stoga bilo preporučljivo koristiti alternativne lijekove za supresiju imunološkog sustava. Napomenimo da ovi rezultati nisu najpouzdaniji s obzirom da su granice 95% pouzdanog intervala dosta velike,

međutim iz svega prezentiranog do sada definitivno možemo lijek ciklofosfamid smatrati najznačajnijim faktorom za razvijanje limfoma.

Tablica 3.7: Univarijatna Coxova regresija

Parametar	ParEst	StdErr	χ^2	p-value	HR	95% HR CI	
antimalarik	-0.212	0.678	0.098	0.754	0.809	0.214	3.052
azatioprin	0.356	0.857	0.173	0.678	1.428	0.266	7.656
metotreksat	0.157	1.067	0.022	0.883	1.170	0.144	9.470
ciklofosfamid	1.727	0.821	4.428	0.035	5.621	1.126	28.071
spol	-0.632	1.081	0.342	0.559	0.532	0.064	4.420
SSA	-0.693	0.821	0.712	0.399	0.500	0.100	2.501
RF	0.705	1.070	0.434	0.510	2.023	0.248	16.476
dob_dijagnoze	-0.010	0.026	0.152	0.697	0.990	0.941	1.042

Tablica 3.8: Multivarijatna Coxova regresija

Parametar	ParEst	StdErr	χ^2	p-value	HR	95% HR CI	
antimalarik	-0.305	0.744	0.168	0.681	0.737	0.171	3.169
azatioprin	-0.761	1.102	0.477	0.490	0.467	0.054	4.052
metotreksat	0.420	1.118	0.141	0.708	1.521	0.170	13.620
ciklofosfamid	2.067	1.022	4.089	0.043	7.904	1.066	58.618
spol	-1.096	1.142	0.922	0.337	0.334	0.036	3.132
SSA	-0.677	0.889	0.580	0.446	0.508	0.089	2.901
RF	0.807	1.150	0.492	0.483	2.242	0.235	21.373
dob_dijagnoze	-0.011	0.035	0.101	0.751	0.989	0.923	1.059

Nijedan od drugih faktora rizika nije statistički značajan te razlike između univarijatnog i multivarijatnog slučaja nisu velike. Osobe koje su razvile RF protutijela imaju 2.242 puta veći rizik razvijanja limfoma od osoba koje nisu razvile ova protutijela, iako taj rezultat nije statistički značajan. Pojedinci koji uzimaju lijekove antimalarik i azatioprin imaju manji hazard razvijanja NHL-a nego pojedinci koji ih ne uzimaju, a korištenje metotreksata povećava ovaj rizik 1.521 puta. Zanimljivo je da rizik razvijanja limfoma nije ovisan o dobi dijagnoze, tj. rizik je skoro jednak bilo da je osobi pSS dijagnosticiran ove ili sljedeće godine. Također, žene imaju 0.337 puta manji rizik razvijanja limfoma od muškaraca, a prema izloženom, osobe s razvijenim protutijelima SSA imaju upola manji rizik razvitka NHL-a u odnosu na pojedince bez razvijenih protutijela.

3.6 Zaključak

Pacijenti dijagnosticirani sa Sjoegrenovim sindromom (pSS) općenito imaju veći rizik razvoja ne-Hodgkinovog limfoma (NHL) u odnosu na opću populaciju. Kako bi se procijenilo vrijeme od dijagnoze pSS-a do razvitka limfoma te potencijalni utjecaj pojedinih faktora na to vrijeme, korištene su metode analize doživljenja. Proučavani faktori bili su imunosupresivni lijekovi antimalarik, azatioprin, metotreksat i ciklofosfamid, zatim spol osobe, dob kada je osobi dijagnosticiran pSS te protutijela SSA i RF. Od svih ovih faktora, jedini statistički značajni faktor ispao je lijek ciklofosfamid. Prvo je χ^2-testom pokazano da bi zaista mogla postojati povezanost između korištenja ciklofosfamida i razvitka limfoma, a zatim je Kaplan-Meierovom procjenom funkcija doživljenja i Coxovom regresijom utvrđeno da postoji statistički značajna razlika u vremenu doživljenja za osobe koje koriste i koje ne koriste ovaj lijek. Odnosno, osobe koje koriste ciklofosfamid imaju otprilike 7 puta veći rizik razvijanja limfoma od osoba koje ovaj lijek ne koriste. Ono što je zanimljivo je da se taj lijek inače koristi za liječenje tumora, ali kad se koristi kao imunosupresiv imao je upravo suprotan učinak. Stoga je prema svemu navedenom preporučljivo uzimati druge imunosupresivne lijekove, kao npr. antimalarik i azatioprin.

3.7 Ključni dijelovi koda

```
title "t-test";
proc ttest data=podaci;
        class limfom;
        var dob_dijagnoze;
run;

title "Mann-Whitneyjev test";
proc npar1way data=podaci wilcoxon;
        class limfom;
        var t_dijagnoza;
run;

title "Chi-kvadrat i Fisherov egzaktni test, spol";
proc freq data=podaci;
    tables limfom*spol/expected chisq cmh plot=MOSAIC;
run;    /* analogno i za ostale varijable */

title "Kaplan-Meier procjena funkcije dozivljenja za razvoj limfoma";
proc lifetest data = podaci method = km plots = (s);
    time t_dijagnoza *limfom(0);
run;

title "Univarijatna Coxova regresija, spol";
proc phreg data = podaci;
    model t_dijagnoza*limfom(0) = spol/rl;
run;    /* analogno i za ostale varijable */

title "Multivarijatna Coxova regresija";
proc phreg data = podaci;
    model t_dijagnoza*limfom(0) = spol azatioprin metotreksat
    ↪  antimalarik ciklofosfamid SSA dob_dijagnoze RF/rl;
run;
```

Bibliografija

[1] https://www.sas.com/fr_fr/software/on-demand-for-academics.html.

[2] https://www.zdravobudi.hr/clanak/reumatologija/sjogrenov-sindrom-1-dio-17489.

[3] https://www.cybermed.hr/centriaz/nehodgkinovlimfom/.

[4] https://www.onkologija.hr/ne-hodgkinov-limfom/ne-hodgkinov-limfom-statistika/.

[5] M. Klein D.G. Kleinbaum, *Survival Analysis - A Self-Learning Text*, Springer, 2012.

[6] M.L. Moeschberger J.P. Klein, *Survival analysis - Techniques for Censored and Truncated Data*, Springer, 2003.

[7] F.E.Harrell Jr., *Regression Modeling Strategies - With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer, 2015.

[8] B. Anić M. Mayer M. Martinić, M.R. Crnogaj, *Razvoj non-Hodgkinovog limfoma u bolesnika s primarnim Sjögrenovim sindromom: retrospektivna, kohortna studija provedena u Kliničkom bolničkom centru Zagreb*, (2021), https://hrcak.srce.hr/file/383832.

Sažetak

U ovom radu obrađena je analiza doživljenja, skupina statističkih metoda koje se koriste za obradu učestalosti nekog događaja te vremena potrebnog do pojave tog događaja. Prezentiran je i Coxov regresijski model kojim se procjenjuje utjecaj određenih faktora na vrijeme doživljenja. Ova je teorija zatim primijenjena na podatke o skupini ljudi s dijagnosticiranim Sjoegrenovim sindromom za koje se promatralo hoće li razviti ne-Hodgkinov limfom, tzv. događaj od interesa. Utjecaj lijekova i drugih faktora na vrijeme doživljenja modeliran je Coxovom regresijom.

Summary

This master's thesis gives a general theoretical introduction to the so-called survival analysis, a branch of statistics used for analyzing the expected duration of time until the event of interest occurs. Some factors might affect survival time, their influence is modeled using Cox regression. Discussed theory is then applied to the dataset of patients diagnosed with Sjoegren's syndrome for whom was observed whether they would develop non-Hodgkin's lymphoma or not. Cox regression is used to assess the effects that potentital risk factors, such as medications which suppress immune system, might have on the survival time.

Životopis

Rođena sam 25. rujna 1998. godine u Zagrebu. Pohađala sam Osnovnu školu Šestine (2005./2006.-2012./2013.) te Glazbeno učilište Elly Bašić (2008./2009.-2013./2014.), gdje sam svirala klavir. Završila sam Petu gimnaziju u Zagrebu, matematički smjer (2013./2014.-2016./2017.). Akademske godine 2017/2018. upisala sam inženjerski smjer na preddiplomskom studiju Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu te u akademskoj godini 2019./2020. stekla titulu prvostupnice matematike (univ. bacc. math.). Za vrijeme fakulteta držala sam instrukcije iz matematike i latinskog jezika te od 2019. godine radila kao biljeterka u Zagrebačkom gradskom kazalištu Komedija. Položila sam C2 stupanj engleskog jezika te B1 stupanj njemačkog jezika u Školi stranih jezika Vodnikova. U lipnju 2022. godine zaposlila sam se preko Student servisa Studentskog centra u Zagrebu u tvrtki True North, kao članica Data & AI tima. U slobodno vrijeme učim plesati u plesnoj školi Salsa de Fuego.