



**KAUNO TECHNOLOGIJOS UNIVERSITETAS
MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS
TAIKOMOSIOS MATEMATIKOS KATEDRA**

Erika Butkutė

**NEAPIBRĖŽTŲ SVEIKATOS DUOMENŲ
IR IŠGYVENAMUMO ANALIZĖ TAIKANT
BAJESO METODĄ IR KOKSO MODELĮ**

Magistro darbas

**Vadovas
prof. dr. R. Alzbutas**

KAUNAS, 2014



KAUNO TECHNOLOGIJOS UNIVERSITETAS
MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS
TAIKOMOSIOS MATEMATIKOS KATEDRA

TVIRTINU
Katedros vedėjas
doc. dr. N. Listopadskis

NEAPIBRĖŽTŲ SVEIKATOS DUOMENŲ
IR IŠGYVENAMUMO ANALIZĖ TAIKANT
BAJESO METODĄ IR KOKSO MODELĮ

Taikomosios matematikos magistro baigiamasis darbas

Vadovas
prof. dr. R. Alzbutas
2014 06 05

Recenzentas
dr. K. Berškienė
2014 06 05

Atliko
FMMM - 2 gr. stud.
E. Butkutė
2014 05 30

KAUNAS, 2014

KVALIFIKACINĖ KOMISIJA

Pirmininkas: Juozas Augutis, profesorius (VDU)

Sekretorius: Eimutis Valakevičius, profesorius (KTU)

Nariai: Vytautas Janilionis, docentas (KTU)

Kristina Šutienė, docentė (KTU)

Jonas Valantinas, profesorius (KTU)

Zenonas Navickas, profesorius (KTU)

Arūnas Barauskas, dr., direktoriaus pavaduotojas (UAB „Danet Baltic“)

Butkutė E. Uncertain health data and survival analysis applying Bayes method and Cox model: Master's work in Applied Mathematics / Supervisor prof. R. Alzbutas; Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology .– Kaunas, 2014. – 70 p.

SUMMARY

Increasing number of clinical trials contributed to the increased need of analysis of censored survival data. Cox proportional hazards model has become a great success in this area, which was published in 1972 (Cox, 1972). Regression analysis of survival data has become possible. This model was proposed in order to evaluate effects of the different covariates, which influence the time to event (survival) data. It is discussed PHM right-censored data for mortality basing on the Bayesian approach. The main aspect of the Bayes rule is prediction. Improvement of the new numerical algorithms, such as Gibbs sampling allowing obtain a sample from posterior value, motivated to start using the Bayesian methods in the survival analysis. It is selected baseline hazard simulation using the polygonal function for the assessment of survival curves in order to avoid inaccurate and misleading results. This paper is intended to show primarily the importance of including the covariates to the prediction of survival distribution. Also it is presented review of Bayesian MCMC method for Cox model and described the baseline hazard simulation by using the polygonal function. Finally, it is showed how this model can be adapted in using the software WinBUGS. Method is illustrated with examined practical tasks.

TURINYS

IVADAS	9
1 BENDROJI DALIS	12
1.1 IŠGYVENAMUMO ANALIZĖS METODIKA IR CENZŪRUOTI DUOMENYS	12
1.1.1 IŠGYVENAMUMO TEORIJOS APŽVALGA	12
1.1.2 IŠGYVENAMUMO DUOMENYS. CENZŪRAVIMAS	14
1.1.3 DUOMENŲ CENZŪRAVIMAS IŠ DEŠINĖS	15
1.1.4 KAPLANO – MEJERIO METODAS	16
1.2 BAJESO METODAS IR JO TAIKYMAS	18
1.2.1 BAJESO METODAS	18
1.2.2 BAJESINIS MODELIAVIMAS	19
1.2.3 BAJESO METODO TAIKYMAS IŠGYVENAMUMO ANALIZEI	20
1.3 KOKSO MODELIS IR JO TAIKYMAS	21
1.3.1 PROPORCINGO PAVOJINGUMO MODELIS	23
1.3.2 MARTINGALO TAIKYMAS KOKSO MODELyje	24
1.3.3 KOKSO MODELIS IŠ BAJESO PERSPEKTYVOS	28
1.3.4 BAZINIO PAVOJINGUMO MODELIAVIMAS NAUDOJANT DAUGIAKAMPIO FUNKCIJĄ	29
1.4 PROGRAMINIŲ PRIEMONIŲ PASIRINKIMAS IR APŽVALGA	30
2 TIRIAMOJI DALIS	32
2.1 PACIENTŲ, SERGANČIŲ LEUKEMIJA IŠGYVENAMUMO DUOMENŲ ANALIZĖ	32
2.1.1 PACIENTŲ, SERGANČIŲ LEUKEMIJA IŠGYVENAMUMO FUNKCIJŲ VERTINIMAS KAPLANO – MEJERIO METODU	34
2.1.2 PACIENTŲ, SERGANČIŲ LEUKEMIJA IŠGYVENAMUMO ANALIZĖ NAUDOJANT PROGRAMINĘ ĮRANGĄ WINBUGS	37
2.2 PACIENTŲ, SERGANČIŲ GERKLŲ VĖŽIU IŠGYVENAMUMO DUOMENŲ ANALIZĖ	42
2.2.1 PACIENTŲ, SERGANČIŲ GERKLŲ VĖŽIU, LIGOS IŠPLITIMO DUOMENŲ IŠGYVENAMUMO FUNKCIJŲ VERTINIMAS KAPLANO – MEJERIO METODU	43
2.2.2 PACIENTŲ, SERGANČIŲ GERKLŲ VĖŽIU LIGOS IŠPLITIMO DUOMENŲ IŠGYVENAMUMO ANALIZĖ NAUDOJANT PROGRAMINĘ ĮRANGĄ WINBUGS	44
DARBO APŽVALGA IR IŠVADOS	49
LITERATŪRA	52

1 PRIEDAS. K-M IŠGYVENAMUMO FUNKCIJOS REZULTATAI NAUDOJANT PROGRAMINĘ PRIEMONĘ MEDCALC.....	53
2 PRIEDAS. PROGRAMINĖS ĮRANGOS WINBUGS APRAŠYMAS	57
3 PRIEDAS. PROGRAMINĖS ĮRANGOS WINBUGS KODAS	67

LENTELIŲ SĄRAŠAS

1 lentelė. Leukemija sergančių pacientų remisijos duomenys	33
2 lentelė. Pacientų, sergančių leukemija ir gydytų preparatu 6-mp, išgyvenamumo funkcijos vertinimas.	34
3 lentelė. Pacientų, sergančių leukemija ir gydytų placebo, išgyvenamumo funkcijos vertinimas.....	35
4 lentelė. Programinės įrangos WinBUGS rezultatai, aposteriorinė statistika	37
5 lentelė. Programinės įrangos WinBUGS rezultatai. DIC, kai bazinio pavojingumo vertinimui naudojama daugiakampio funkcija	38
6 lentelė. Programinės įrangos WinBUGS rezultatai. Aposteriorinė statistika, jei bazinio pavojingumo vertinimui naudojamas gama apriorinis procesas.....	38
7 lentelė. Programinės įrangos WinBUGS rezultatai. DIC, kai bazinio pavojingumo vertinimui naudojamas gama procesas	38
8 lentelė. Programinės įrangos WinBUGS rezultatai, aposteriorinė statistika	44
9 lentelė. Programinės įrangos WinBUGS rezultatai. DIC, kai bazinio pavojingumo vertinimui naudojama daugiakampio funkcija	45

PAVEIKSLŲ SĄRAŠAS

1 pav. Rizikos kitimas	13
2 pav. Ligonų stebėjimo laikas.....	16
3 pav. Progresyvus cenzūravimas	16
4 pav. Pacientų, sergančių leukemija ir gydytų preparatais 6-mp ir placebo, Kaplano – Mejerio išgyvenamumo kreivės.....	36
5 pav. Programinės įrangos WinBUGS rezultatai susiję su kovariantės koeficientais. Pateikiamas pacientų grupės gydomas vaistais aposteriorinis tankis	39
6 pav. Programinės įrangos WinBUGS rezultatai. Prognozuojami išgyvenamumo modeliai	40
7 pav. Programinės įrangos WinBUGS rezultatai. Bazinio pavojingumo modeliavimas naudojant daugiakampio funkciją.....	40
8 pav. Programinės įrangos WinBUGS rezultatai. Parametro tau stačiakampė diagrama.....	41
9 pav. Pacientų, sergančių gerklų vėžiu, ligos stadijų, Kaplano – Mejerio išgyvenamumo kreivės.....	43
10 pav. Programinės įrangos WinBUGS rezultatai susiję su kovariantės koeficientais. Ligos stadijų aposteriorinis tankis	46
11 pav. Programinės įrangos WinBUGS rezultatai. Prognozuojami išgyvenamumo modeliai, stadijos 2 ir stadijos 3	47
12 pav. Programinės įrangos WinBUGS rezultatai. Prognozuojami išgyvenamumo modeliai, stadijos 3 ir stadijos 4	47
13 pav. Programinės įrangos WinBUGS rezultatai. Prognozuojami ligos stadijų išgyvenamumo modeliai.....	48

ĮVADAS

Proporcingo pavojingumo modelis (PHM) arba kitais žodžiais proporcingo rizikingumo (mirtingumo) modelis išgyvenamumo duomenų analizės kontekste apima plačiai žinomą Kokso modelį. Šis modelis Kokso (1972) buvo pateiktas siekiant įvertinti poveikį išgyvenamumui tokių kovariančių, kurios kinta priklausomai nuo laiko ir nagrinėjamų įvykių skaičiaus.

Šio darbo tikslas yra pritaikyti Bajeso metodą atliekant išgyvenamumo modelio parametrų vertinimą ir tikslinimą. Siekiant šio tikslo buvo sprendžiami šie uždaviniai:

- Susipažinti su sveikatos duomenų analizei taikomais metodais.
- Išplėtoti Kokso modelį bei išgyvenamumo ir proporcingo pavojingumo analizės metodiką.
- Susipažinti su programinės įrangos WinBUGS taikymu.
- Programinės įrangos WinBUGS pagalba atlikti neapibrėžtų sveikatos duomenų ir išgyvenamumo analizę taikant Bajeso metodą ir Kokso modelį.

Šiame darbe visų pirma yra aptariamas proporcingo pavojingumo modelis PHM. Šis modelis buvo plačiai naudojamas, po to kai Koksas (1972) pristatė jį vienaveiksmių duomenų kontekste. Tai reiškia, jog kiekvienas tyrimo subjektas gali patirti įvykį ne daugiau kaip vieną kartą, pavyzdžiui - asmens mirtis galima tik kartą. Tokie duomenys yra vadinami vienaveiksmiais išgyvenamumo duomenimis. Be to, darbe aptariamas Bajeso metodo taikymas taip iš dešinės cenzūrotų mirtingumo duomenų ir išgyvenamumo analizei.

Gerai žinoma, kad Bajeso metodas turi pranašumą prieš klasikinius tikimybinius metodus nagrinėjant cenzūrotus duomenis ir mažas imtis. Priežastis, kodėl iki pastarųjų metų Bajeso metodai nebuvo plačiai naudojami išgyvenamumo analizėje yra ta, kad realiems tyrimams aposteriorinį pasiskirstymą esant cenzūravimui yra labai sunku išgauti tiesiogiai. Tačiau, naujų skaitinių algoritmų tobulinimas, tokių kaip Gibss'o atrankos metodas, leidžiantis gauti imtį iš aposteriorinio skirstinio, sudarė galimybes pradėti naudoti Bajeso metodus išgyvenamumo analizėje. Bajeso įverčių, kurie nusakomi daugiau negu vienu parametrų lygmeniu, apskaičiavimas yra pilnai įgyvendinamas, naudojant imitacinį modeliavimą. Gibss'o modeliavimas literatūroje išskiriamas kaip viena iš geriausiai žinomų Monte Karlo Markovo grandinės (MCMC) metodo realizacija. MCMC metodas tapo įprastu įrankiu Bajeso analizėje. Be to, specializuota programinė įranga, skirta įgyvendinti MCMC metodus yra nemokama. Taigi, atsižvelgiant į PHM kontekstą, šiame darbe aprašomas ir tokios nemokamos WinBUGS programinės įrangos naudojimas, ir MCMC metodo taikymas sudėtingų statistinių modelių analizei.

BUGS (Bayesian inference Using Gibbs Sampling) tipo specializuota programinė įranga yra skirta įgyvendinti MCMC metodu pagrįstas tikimybių modelių analizes, kuriose parametrai yra laikomi atsitiktiniais dydžiais. BUGS taikymo projektas 1989 metais buvo pradėtas vykdyti Medicinos mokslinių tyrimų tarybos Biostatistikos skyriuje Kembridže. Nuo to laiko, šio tipo programinė įranga tapo viena iš populiariausių statistinio modeliavimo paketų, su tuo metu esančiais virš 30.000 registruotų WinBUGS vartotojų iš viso pasaulio. Nežiūrint to, per pastaruosius 20 metų BUGS tėra tik viena nedidelė dalis iš visos didžiulės Bajeso idėjų taikymų įvairovės. Atitinkamos programinės įrangos svarba buvo pripažinta „JK mokslinių matematikos tyrimų tarptautinėje apžvalgoje“ [20]. Čia pripažįstama, kad WinBUGS gali tapti gana galingu ir lanksčiu įrankiu išgyvenamumo analizėje. Be to, modeliavimas yra pakankamai greitas (net ir sudėtingų modelių su dideliais duomenų kiekiais), o esami modeliai gali būti lengvai pritaikomi.

Išgyvenamumo modelį galima nagrinėti taikant pavojingumo (rizikingumo ar mirtingumo intensyvumo) funkciją. Apibendrintu atveju, nagrinėjant PHM, bazinę pavojingumo funkciją galima sieti su gama procesu. Tačiau, atskirais atvejais, priklausomai nuo duomenų, tai gali sąlygoti netikslus ir klaidinančius išgyvenamumo modeliavimo rezultatus. Šiai problemai spręsti taikomas bazinio pavojingumo modeliavimas naudojant daugiakampio metodą, kuris buvo pasiūlytas Beamonte ir Bermúdez (2003) [7]. Beamonte ir Bermúdez (2003) parodė, kad bazinio pavojingumo apskaičiavimas vertinimui naudojant daugiakampio funkciją vietoje įprastos laipsninės funkcijos yra tik truputį sudėtingesnis ir turi tą privalumą, kad modelis yra tolydus.

Darbe bazinio pavojingumo funkcijos apskaičiavimui ir palyginimui pradžioje naudojamas gama procesas. Taip pat darbe pristatomas ir alternatyvus metodas - Bajeso pusiau parametrinis metodas, kurio pagalba sudaromas PHM hierarchinis modelis. Neparametrinė Kokso modelio dalis yra ne neigiama daugiakampio funkcija. Programinės įrangos WinBUGS pagalba šie modeliai nesudėtingai palyginami, naudojant Informacijos nukrypimo kriterijų (DIC).

Šiame darbe taip pat siekiama parodyti, kaip svarbu vertinti kovariantes prognozuojant išgyvenamumo pasiskirstymą. Be to, pateikiama Bajeso metodo MCMC realizacijos apžvalga Kokso modeliui ir aprašomas bazinio pavojingumo modeliavimas naudojant daugiakampio funkciją. Galiausiai dar pateikiama, kaip šis modelis gali būti pritaikytas. Naudojant programinę įrangą WinBUGS, metodo taikymas demonstruojamas sprendžiant kelis praktinius uždavinius.

Darbo praktinėje dalyje buvo nagrinėjamas viešai publikuotas pacientų, sergančių leukemija ir gydytų 6–merkaptopurino (6-mp) ir placebo preparatais, remisijos laikas. Pirmiausia, abiem pacientų grupėms, gydytoms skirtingais preparatais, išgyvenamumo funkcijos vertinamos Kaplano – Mejerio metodu. Vykdamas medicininis tyrimus, tai dažnai naudojamas metodas, siekiant įvertinti pacientų dalį, išgyvenusių tam tikrą laiką po gydymo proceso. Tolesnis tyrimas vykdomas

programinės įrangos WinBUGS pagalba. Kokso modeliui atliekamas neapibrėžto bazinio pavojingumo vertinimas naudojant daugiakampio funkciją ir, įvertinus kovariančių poveikį remisijai, gaunamos prognozuojamos išgyvenamumo funkcijos. Taip pat, pagal analogišką metodiką atliekamas išgyvenamumo funkcijų vertinimas pacientams sergantiems gerklų vėžiu. Tik antrajame tyrime ligos baigtis klasifikuojama keturiomis stadijomis.

Šiame darbe nagrinėjami uždaviniai svarbūs tuo, jog atliekant sveikatos duomenų ir išgyvenamumo analizę reikia tinkamai atsižvelgti į galimus duomenų ribotumus ir rezultatų neapibrėžtumą. Atliekant vertinimą dėl gydymo preparatų svarbu taikyti tokią analizę, kuri įgalintų atskirti preparatų poveikį išgyvenamumui atsižvelgiant į neapibrėžtumą dėl ribotų duomenų. Parinkus tam tinkamą statistinį metodą ir suradus parametrų įverčius, jau galima prognozuoti išgyvenamumo ribas ir taip praktikoje nustatyti preparatų efektyvumą, pritaikyti tinkamiausią gydymą, planuoti tolesnius klinikinius ar laboratorinius tyrimus.

1 BENDROJI DALIS

1.1 IŠGYVENAMUMO ANALIZĖS METODIKA IR CENZŪRUOTI DUOMENYS

Išgyvenamumo analizės metodai leidžia tuo pat metu tyrinėti pilnus ir nepilnus stebėjimus, todėl tokia analizė dažniausiai ir taikoma medicinoje.

Nagrinėjant išgyvenamumo ir rizikos funkcijas, pastebima tolesnė tyrimo eiga, t. y. atskleidžiami rizikos veiksniai. Tai galima atlikti konstruojant ir analizuojant Kokso regresijos modelius. Regresijos modeliais siekiama atskleisti ne išgyvenamumo funkcijos, o intensyvumo funkcijos priklausomybę nuo tam tikrų veiksnių. Taikant Kokso regresijos modelį galima vertinti riziką lemiančius veiksnius ir interpretuoti rizikos laipsnį. Intensyvumo funkcija Kokso proporcingųjų intensyvumų modelyje priklauso nuo dviejų dauginamųjų: nuo bazinės intensyvumo funkcijos ir nuo rizikos veiksnių.

Rizikos veiksniais gali būti ir kategoriniai kintamieji. Nežinomi regresijos modelio koeficientai nustatomi remiantis dalinio tikėtinumo metodu, nes dažniau skaičiuojamos tikimybės tik neišlikusiųjų, o ne visų stebimųjų. Kadangi intensyvumo funkcija yra netiesinė funkcija, tai geriausios modelio koeficientų reikšmės, tenkinančios dalinio tikėtinumo lygtį, randamos iteraciniais metodais.

1.1.1 IŠGYVENAMUMO TEORIJS APŽVALGA

Analizės metodų, skirtų laiko iki baigties taško analizei ir modeliavimui, visuma vadinama išgyvenamumo analize[1]. Išgyvenamumo duomenys apibrėžiami kaip stebėjimo duomenys iki bet kurio baigties taško, nebūtinai mirties. Analizuojant tokius duomenis, terminas „gyvavimo laikas“ suprantamas kaip laikas nuo stebėjimo pradžios iki baigties įvykio[1].

Tegul T apibrėžiamas, kaip populiacijos individo išgyvenimo nuo stebėjimo pradžios iki baigties taško laikas. Tuomet T priklauso nuo populiacijos charakteristikų, individualių organizmo savybių ir daugelio kitų priežasčių, todėl jis laikomas atsitiktiniu dydžiu. Šio atsitiktinio dydžio skirstinio funkcija $P\{T \leq t\}$ žymima $F(t)$, tankis $p(t)$. Skirstinio funkcija

$$F(t) = P\{T \leq t\},$$

tai tikimybė neišgyventi daugiau nei t (laiko vienetų), o funkcija

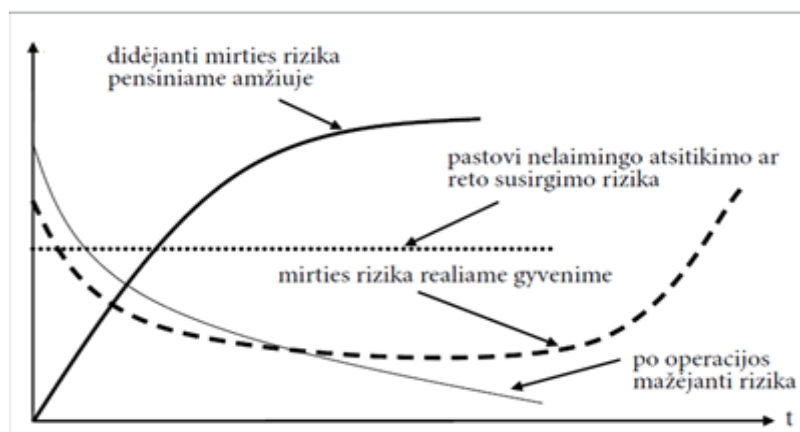
$$S(t) = 1 - F(t) = P\{T > t\},$$

tai tikimybė išgyventi daugiau nei t (laiko vienetų). Funkcija $S(t)$ vadinama populiacijos išgyvenamumo funkcija. Ši funkcija kinta tarp nulio ir vieneto ir yra nedidėjanti.

Medicininis požiūris svarbu įvertinti baigties taško (t. y. mirties ar susirgimo) riziką. Todėl vertinama baigties taško atsiradimo tikimybė momentu t su sąlyga, kad iki momento t baigties taško nebuvo. Ši rizika bendroju atveju yra laiko t funkcija ir žymima $\lambda(t)$, t. y.

$$\lambda(t) = p(t)/S(t) = [d(-S(t))/dt]/S(t) = d(-\ln(S(t)))/dt.$$

Rizikos funkcijos $\lambda(t)$ kitimas priklauso nuo baigties taško apibrėžimo bei tiriamo paciento charakteristikų. Kaip pavyzdys, mirties rizika priklauso nuo paciento amžiaus, biologinių rodiklių (t. y. paveldėjimo ar susirgimų), įrangos veikimo bei kitų veiksnių. Pensinio amžiaus asmenų mirties rizikos funkcija $\lambda(t)$ didėja, didėjant t . Paprasčiausiai taip yra, nes rizika numirti senatvėje su amžiumi didėja. O nelaimingų atsitikimų ar retų susirgimų rizika yra pastovi. Mirties po operacijos rizikos funkcija $\lambda(t)$ yra mažėjanti funkcija. Mirties nuo gyvenimo pradžios iki senyvo amžiaus rizikos funkcija yra vonios formos, kas reiškia, jog kūdikystės laikotarpiu ji mažėja, vėliau iki tam tikro amžiaus išlieka pastovi, o senatvėje didėja. Visus šiuos aspektus puikiai iliustruoja rizikos kitimo 1 paveikslas [1]:



1 pav. Rizikos kitimas

Konkrečios populiacijos išgyvenamumo funkcija vertinama naudojant parametrinį arba neparametrinį išgyvenamumo funkcijos modelį. Naudojant parametrinį metodą, daroma prielaida, jog populiacijos išgyvenamumo funkcija yra tam tikro tipo (kaip eksponentinio ar Veibulo), tačiau su nežinomais parametrais, kuriuos reikia įvertinti. Todėl dažniausiai naudojamas išgyvenamumo funkcijos neparametrinis įvertis yra Kaplano - Mejerio kreivė [1].

1.1.2 IŠGYVENAMUMO DUOMENYS. CENZŪRAVIMAS

Išgyvenamumo duomenų modeliavimas ir analizė yra viena iš seniausių statistikos sričių. Išgyvenamumo duomenys susiję su gyvenimo laikotarpiu arba, apskritai, su laukimo laiku pradedant nuo kažkokio pradinio įvykio 0 laiku (pavyzdžiui - gimimas, gydymo pradžia ar įdarbinimas į tam tikrą darbą) iki kažkokio paskutinio reikšmingo įvykio (pavyzdžiui - mirtis, atkrytis ar invalidumo pensija). Taigi pagrindiniai duomenys bus idealiu atveju nepriklausomas neneigiamas atsitiktinis kintamasis T .

Biomedicinos programose, ypač klinikiniuose tyrimuose, iškyla svarbus klausimas, kaip tiriami „laiko iki įvykio“ duomenys, kuomet kai kurie asmenys vis dar gyvi tyrimo ar analizės pabaigoje. Taigi reikšmingas įvykis, yra neįvykusi mirtis. Todėl gauname cenzūravimo iš dešinės duomenis. Tam, kad siūlomo išgyvenamumo modelio įverčiai pateiktų nešališkus rezultatus, svarbi prielaida, jog asmenys, kurie yra cenzūruoti, taip pat būtų tolesnės mirties rizikos grupėje, kaip ir tie, kurie vis dar gyvi ir necenzūruoti. Rizika, nustatyta bet kuriuo laiko momentu (asmenys vis dar gyvi ir necenzūruoti) turėtų būti pavyzdinė visiems gyventojams, kurie gyvi tuo pačiu metu. Jei tai yra tas atvejis, cenzūravimo procesas vadinamas neinformatyviu. Statistiškai, jei cenzūravimo procesas nepriklauso nuo išgyvenamumo laiko, tuomet automatiškai turėsime neinformatyvų cenzūravimą.

Nepriklausomas cenzūravimas beveik visada išreiškiamas per neinformatyvų cenzūravimą. Pagrindinis tikslas analizuojant cenzūruotus išgyvenamumo duomenis yra įvertinti išgyvenamumo trukmės priklausomybę nuo kovariančių. Kitas tikslas yra įvertinti išgyvenamumo trukmės pagrindinį pasiskirstymą.

1.1.3 DUOMENŲ CENZŪRAVIMAS IŠ DEŠINĖS

Tarkime nagrinėjant n dydžio imties iš tam tikros populiacijos (t. y. ligonių kontingento) išgyvenamumą, kiekvienas *i-tasis* individas stebimas C_i laiką (2 pav.). Dydis C vadinamas cenzūravimo momentu. Tam tikram skaičiui pacientų stebėjimo metu įvyko baigties taškas (t.y. mirė, susirgo), todėl tikslūs jų gyvavimo laikai U_1, U_2, \dots, U_k yra žinomi. Likusiems $n - k$ pacientams stebėjimo laikotarpiu baigties taškas neįvyko. Tai reiškia, kad tam tikru laiko momentu jie dingo iš stebėjimo (t.y. išvyko, nustojo lankytis pas gydytoją ir t. t.) arba tiesiog visą stebėjimo laiką jų baigties taško nesulaukta. Todėl šių ligonių tolesnis gyvavimas yra prognozuojamas ir tokie neišsamūs stebėjimo duomenys vadinami cenzūruotais. Cenzūruotų pacientų stebėjimo laikai yra žinomi ir lygūs C_1, C_2, \dots, C_{n-k} . Apibendrinant išgyvenamumo duomenis, galima tvirtinti, kad tyrimo metu nustatome dvimačio atsitiktinio dydžio reikšmes [1]:

$$(t_1, I_1), (t_2, I_2) \dots (t_n, I_n),$$

čia I – cenzūravimo indikatorius;

$I = 1$, jei žinomas laikas nuo stebėjimo pradžios iki baigties taško, $I = 0$ – cenzūruotas stebėjimas.

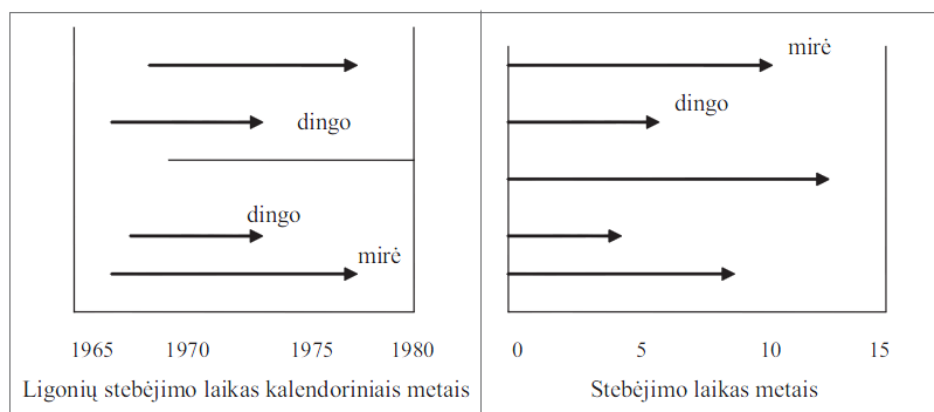
Žinomas tik laikas, iki kurio baigties taškas neįvyko. Tuomet t_j apibrėžiami:

$$t_j = \begin{cases} U_j, I_j = 1, \\ C_j, I_j = 0. \end{cases} \quad (1.1)$$

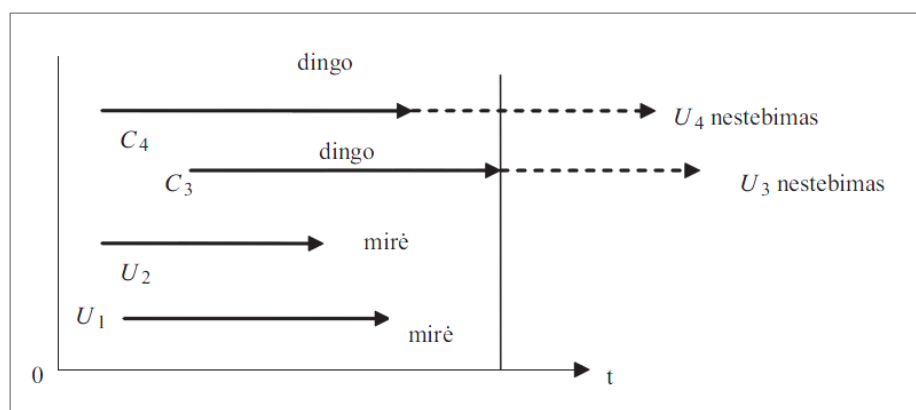
Cenzūruotos iš dešinės imtys gali būti klasifikuojamos [1]:

- I tipo imtis, jei visi ligoniai stebimi vienodą laiką ir cenzūravimas atliekamas tyrimo pabaigoje, t. y. $C_j \equiv C$.
- II tipo cenzūruota imtis iš dešinės, jei visi ligoniai stebimi vienodą laiko tarpą. Cenzūravimas atliekamas tik pasiekus tam tikrą fiksuotą baigties taškų skaičių. Šiuo atveju cenzūravimo laikas yra atsitiktinis. Tokios imtys dažniausiai naudojamos atliekant patikimumo tyrimus.
- III tipas imtis (progresyvus cenzūravimas), jei pacientai į tyrimą įtraukiami skirtingu laiku (pavaizduota 3 pav.). Cenzūravimas vyksta tyrimai baigiantis, t. y. visiems individams tuo pačiu neatsitiktiniu laiku.

- Atsitiktinis cenzūravimas: cenzūravimo laikai C_1, C_2, \dots, C_n yra atsitiktiniai dydžiai, nepriklausantys nuo gyvavimo laikų U_1, U_2, \dots, U_n . I tipo cenzūravimas taip pat yra atskiras atsitiktinio cenzūravimo atvejis.



2 pav. Ligonių stebėjimo laikas



3 pav. Progresyvus cenzūravimas

1.1.4 KAPLANO – MEJERIO METODAS

Siekiant ištirti tam tikrų požymių įtaką taikomi įvairūs išgyvenamumo analizės metodai. Vienas iš jų yra Kaplano - Mejerio statistinis įvertinimas (Kaplan – Maier). Kaplano – Mejerio kreivė, tai dažniausiai naudojamas išgyvenamumo funkcijos neparametrinis įvertis. Medicininuose tyrimuose dažnai naudojamas, siekiant įvertinti pacientų dalį išgyvenusią tam tikrą laiką po gydymo proceso. Tipiškas taikymas yra pacientų grupavimas į kategorijas.

Kaplano - Mejerio formulė pateikia grafinius stebėjimų vaizdavimus, kuriais numatoma tolimesnė tyrimo eiga. Ši prognozė priklauso nuo išgyvenamumo funkcijos $S(t)$. Todėl grafikuose

pateikiamos išgyvenamumo funkcijos sukauptos reikšmės. Kaplano - Mejerio metodu brėžiamas grafikas yra laiptuotos kreivės tipo, į kurį patenka visi tyrime dalyvavę cenzūruoti ir necenzūruoti stebėjimai. Grafike šuoliukais žymima tos vietos, kuriose yra necenzūruotų stebėjimų. Kaplano - Mejerio metodu galima palyginti išlikimo funkcijas skirtingose grupėse nepriklausomai nuo laiko intervalo suskirstymo į smulkesnius intervalus.

Cenzūruotiems duomenims išgyvenamumo funkcijos $S(t)$ dažniausiai naudojamas įvertis yra Kaplano – Mejerio kreivė $\hat{S}(t)$. Tarkime t_j laikai, apibrėžti pagal (1.1) formulę, išdėstyti didėjimo tvarka: $0 < t_1 < \dots < t_j < \dots < t_k$, k – skirtingų t_j reikšmių skaičius, $t_0 = 0$. Taigi Kaplano – Mejerio išgyvenamumo funkcijos $S(t)$ įvertinimas apibrėžiamas [3]:

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \hat{p}(T > t_j | T \geq t_j)$$

Jei mirtis (ar įvykis) laiku t_j neįvyksta, tai

$$\hat{p}(T > t_j | T \geq t_j) = 1.$$

Jei viena ar daugiau nesėkmių įvyksta laiku t_j , tai

$$\hat{p}(T > t_j | T \geq t_j) = \frac{n_j - d(t_j)}{n_j}.$$

Išgyvenamumo funkcija esamuoju laiko momentu priklauso nuo išgyvenamumo funkcijos ankstesniuoju laiko momentu t.y. žymi išgyvenusiųjų dalį. Ši dalis skaičiuojama pagal žemiau pateiktą Kaplano - Mejerio formulę [3]:

$$\hat{S}(t) = \prod_{t_j \leq t} \frac{n_j - d_j}{n_j},$$

čia t_j - mirties (ar įvykio) momentas; n_j - pacientai (asmenys) rizikos grupėje, t.y. stebėtų iki momento t_j skaičius; d_j - baigčių momentu t_j skaičius.

1.2 BAJESO METODAS IR JO TAIKYMAS

Bajeso teorema žinoma nuo 1763, kuomet ją savo moksliniame rašte aprašė Thomas Bayes, kurio vardu ši tikimybių teorema ir pavadinta. Tačiau pirmą kartą Bajeso metodo sąvoka paminėta tik Pearl (1985). Šiuo metu Bajeso metodas yra labai plačiai naudojamas įvairiems sprendimo priėmimo uždaviniams spręsti. Šio darbo tikslas yra panaudoti Bajeso metodą medicinos srityje.

Bajeso metodas išgyvenamumo duomenų analizėje sulaukia vis daugiau dėmesio, dėl progreso kompiuterinėse programose ir modeliavimo technikos. Todėl šis metodas atliekant išgyvenamumo duomenų analizę tampa vis svarbesniu įrankiu medicinos ir sveikatos srityje.

1.2.1 BAJESO METODAS

Bajeso taisyklė grindžiama apibrėžiant tikimybinį modelį stebimiems duomenims D , atsižvelgiant į nežinomų parametrų θ vektorių, taip privedant prie tikėtumo funkcijos $L(\theta|D)$. Daroma prielaida, kad parametro θ reikšmės nėra fiksuotos. Jos yra atsitiktinės ir turi apriorinį pasiskirstymą apibrėžiamą, kaip $\pi(\theta)$. Norint įvertinti θ reikia nustatyti aposteriorinį pasiskirstymą, kuris gaunamas pagal Bajeso teoremą. Parametro θ aposteriorinis pasiskirstymas apibrėžiamas [4]:

$$\pi(\theta|D) = \frac{L(\theta|D)\pi(\theta)}{\int_{\Theta} L(\theta|D)\pi(\theta)d\theta}, \quad (2.1)$$

kur Θ reiškia parametro intervalą θ . Iš (2.1) formulės aišku, kad $\pi(\theta|D)$ yra proporcinga tikėtumo funkcijai, padaugintai iš jau žinomo skirstinio,

$$\pi(\theta|D) \propto L(\theta|D)\pi(\theta),$$

ir tokiu būdu tai siejama su proporcija stebėtų duomenų per $L(\theta|D)$, ir proporcija apriorinės informacijos įvertinant per $\pi(\theta)$.

Dydis $m(D) = \int_{\Theta} L(\theta|D)\pi(\theta)d\theta$ yra $\pi(\theta|D)$ normavimo konstanta, dažnai dar vadinama duomenų ar prognozuojamo apriorinio pasiskirstymo ribiniu pasiskirstymu. Tačiau daugelis modelių ir taikymų, $m(D)$ neturi analitinio sprendinio uždaros formos, kaip ir $\pi(\theta|D)$. Todėl

naudojamas vienas iš populiariausių atrankos skaičiavimo metodų pagal $\pi(\theta|D)$ yra Gibbs'o atranka. Tai yra naudingas atrankos metodas, leidžiantis gauti imtį iš $\pi(\theta|D)$ be informacijos apie normavimo konstantą $m(D)$.

Pagrindinis Bajeso taisyklės aspektas yra prognozavimas. Prognozavimas yra dažnai svarbus tikslas sprendžiant regresijos problemas, ir paprastai svarbus pasirenkant modelį. Būsimo stebėjimo vektoriaus z prognozuojamas aposteriorinis pasiskirstymas atsižvelgiant į duomenis D apibrėžiamas, kaip [4]

$$\pi(z|D) = \int_{\Theta} f(z|\theta)\pi(\theta|D)d\theta, \quad (2.2)$$

kur $f(z|\theta)$ yra imties z tankis, ir $\pi(\theta|D)$ - aposteriorinis skirstinys esant θ . Galima matyti, jog (2.2) formulė yra tik $f(z|\theta)$ aposteriorinė tikimybė ir todėl atranka pagal (2.2) formulę yra vykdoma lengvai per Gibbs'o atrankos metodą pagal $\pi(\theta|D)$. Ši Bajeso metodo ypatybė, kai (2.2) formulė nurodo prognozavimą ir prognozuotus pasiskirstymus yra lengvai apskaičiuojama, kai imtys pagal $\pi(\theta|D)$ yra turimos.

1.2.2 BAJESINIS MODELIAVIMAS

Bajeso metodai gali būti sudėtingi, ypač renkantis pateiktą ankstesnę informaciją, analizuojant turimus duomenis ir formuluojant matematinės lygtis, kurios visiškai apibrėžia modelį.

Analizuojant ankstesnio tyrimo duomenis, naudojant Bajeso metodus, būtų tikslingiau naudotis konkrečiu pavyzdžiu, kuriuo remiantis būtų lengviau suprantama, metodo ir jo rezultatų, įgyvendinimo ir skirtingų tyrimų plano reikalingumo svarba. Ypač tuomet verta apvarstyti, ar ankščiau atliktas tyrimas buvo užbaigtas sėkmingai ir su mažesnėmis sąnaudomis.

Per paskutinį dešimtmetį, pagrįsti Monte-Karlo atrankos metodai, įvertinantys didesnės dimensijos aposteriorinius integralus buvo sparčiai vystomi. Gibbs'o modelis galimai yra vienas iš geriausiai žinomų MCMC metodų atrankos algoritmų Bajeso metodo analizės literatūroje. Pagrindinis Gibbs'o atrankos metodo orientyras kalbant apie problemas Bajeso metodo išvados yra Gelfand ir Smith (1990). Tačiau Casella ir George (1992) pateikia puikią literatūros apžvalgą, skirtą Gibbs'o modelio įgyvendinimui [6].

Tegul $\theta = (\theta_1, \theta_2, \dots, \theta_p)'$ yra parametrų p -dimensijų vektorius, o $\pi(\theta|D)$ - jų aposteriorinis pasiskirstymas, atsižvelgus į D duomenis. Tuomet Gibbs'o atrankos metodo struktūra apibrėžiama [4]:

Žingsnis 1. pasirinkti pradinį tašką $\theta_0 = (\theta_{1,0}, \theta_{2,0}, \dots, \theta_{p,0})'$ ir $i=0$.

Žingsnis 2. sugeneruoti $\theta_{i+1} = (\theta_{1,i+1}, \theta_{2,i+1}, \dots, \theta_{p,i+1})'$:

- sugeneruoti $\theta_{1,i+1} \sim \pi(\theta_1 | \theta_{2,i}, \dots, \theta_{p,i}, D)$;
- sugeneruoti $\theta_{2,i+1} \sim \pi(\theta_2 | \theta_{1,i+1}, \theta_{3,i}, \dots, \theta_{p,i}, D)$;
-
- sugeneruoti $\theta_{p,i+1} \sim \pi(\theta_p | \theta_{1,i+1}, \theta_{2,i+1}, \dots, \theta_{p-1,i+1}, D)$.

Žingsnis 3. nustatyti, kai $i=i+1$ ir eiti į žingsnį 1.

Taigi kiekvienas θ parametras yra jam būdingoje tvarkoje ir ciklas šioje schemoje turi sugeneruoti p atsitiktinius kintamuosius. Gelfand ir Smith (1990) [4] parodo, kad tam tikromis taisyklingomis sąlygomis, vektorių seka $\{\theta_i, i=1,2,\dots\}$ turi pastovų pasiskirstymą $\pi(\theta|D)$.

Tyrimui naudojama programinė įranga turi įvertinti tikslus taip, kaip tikimasi ir turėtų būti parengta, naudojant išsamų metodą analizuojant klinikinius tyrimus. Rekomenduojama, kad programinės įrangos išvystymas apimtų svarstomą problemą kiekvienam programinės įrangos vykdymo kartojimui. Tai turėtų padėti išvengti programinės įrangos gaunamų klaidų, kurios analizę padaro netikslingą ir pateikia netikslų rezultatą. Bajeso metodo modeliavimą puikiai vykdo programinė įranga WinBUGS.

1.2.3 BAJESO METODO TAIKYMAS IŠGYVENAMUMO ANALIZEI

Lygindami klasikinius metodus ir Bajeso metodus atsakysime į klausimą *kodėl išgyvenamumo analizei pasirenkamas būtent Bajeso metodas?*

Pateiksime Bajeso metodo teigiamus aspektus lyginant su klasikinais tikimybiniais metodais išgyvenamumo analizėje. Gerai žinoma, kad išgyvenamumo modelius paprastai yra gana sunku pritaikyti, ypač esant sudėtingoms cenzūruojančioms sistemoms. Naudojant Gibbs'o modelį ir kitus MCMC metodus, pritaikyti sudėtingus išgyvenamumo modelius yra gana paprasta, ir programinės įrangos tokios kaip BUGS prieinamumas ganėtinai palengvina šį įgyvendinimą. Be to, MCMC atrankos metodas leidžia daryti tikslią išvadą apie bet kokią imties dydį, nesiimant asimptotinių

skaičiavimų. Klasikiniai metodai, pavyzdžiui, dispersijos apskaičiavimui, paprastai reikalauja asimptotinių argumentų, kuriuos gali būti gana sudėtinga išvesti ir kai kuriuose modeliuose tai tiesiog neįmanoma. Tuomet, visada kyla problema ar modelio (imties) dydis yra pakankamai didelis, kad būtų veiksminga asimptotinė aproksimacija. Priešingai, Bajeso metodu, dispersijos įverčiai, taip pat kaip bet kuri kita aposteriorinė suvestinė tampa kaip šalutinis Gibbs'o modelio rezultatas, ir todėl juos įprasta gauti, kai yra prieinami modeliai iš aposteriorinio pasiskirstymo.

Bajeso metodas taip pat leidžia apjungti apriorinę informaciją natūraliu būdu, kur to negali padaryti klasikiniai metodai. Tai yra galingas įrankis išmatuojantis ir įvertinantis apriorinius duomenis, ypač registruotus duomenis. Be to, daugeliui modelių, klasikinių metodų išvados gali būti gauta kaip specialus Bajeso išvados atvejis su daugeliu neinformatyvių apriorinių tipų. Pavyzdžiui, su pastoviu aprioriniu pasiskirstymu, aposteriorinis metodas atitinka maksimalų tikimybės įvertinimą. Šia prasme, klasikinių metodų išvados gali būti laikomos specialiu Bajeso išvados pavyzdžiu.

Bajeso metodo išvada taip pat turi keletą pranašumų prieš klasikinius metodus savo modelio kūrimo prieinamumu ir lankstumu bei duomenų analizės įrankiais. Tikslus Bajeso metodo koeficientų arba modelio kriterijų skaičiavimus galima gauti per Gibbs modelį. Klasikiniu atveju nėra vieningos metodikos, kuri lygintų modelius, nes paprastai tai reikalauja asimptotinių argumentų. Kitos duomenų analizės priemonės, pavyzdžiui, prognostiniai skirstiniai ir liekanos pagal Bajeso metodą gali būti lengviau apskaičiuojamos išgyvenamumo modeliams. Neapibrėžta kovariantė arba duomenys yra dar viena sritis, kurioje Bajeso metodas turi aiškius pranašumus prieš klasikinį metodą, nes trūkstamos reikšmės iš esmės Bajeso sistemoje laikomos parametrais ir prideda tik vieną papildomą lygmenį Gibbs'o modelyje.

Taigi, skaičiavimo algoritmai šiek tiek keičiasi, jei trūksta duomenų. Esant trūkstamų duomenų problemoms, variantiškumo įverčius paprastai yra gana sudėtinga apdoroti ir reikalaujama naujų sprendimų, tačiau Bajeso metodo atveju tai yra šalutinis Gibbs'o modelio rezultatas.

1.3 KOKSO MODELIS IR JO TAIKYMAS

Kokso modelio naudojimas ir sėkmingumas dabar yra didelis pasiekimas išgyvenamumo laiko duomenų analizėje su kovariantėmis. Kopenhagos Statistikos tyrimų padalinyje, įsteigtame 1978 metais, išgyvenamumo duomenų analizė buvo viena iš pagrindinių mokslinių tyrimų sričių, buvo atlikti taikomieji medicinos projektai, naudojant Kokso modelį.

Proporcingo pavojingumo modelio variantas pateiktas Kokso yra semiparametrinis ta prasme, kad bazinė rizikos funkcija $\lambda_0(t)$ nėra modeliuojama kaip t parametrinė funkcija. Dėl Kokso modelio dalinio tikėtinumo funkcijos išvystymo, $\lambda_0(t)$ yra leidžiama pasirinkti vertes, nes ji nepatenka vertinant lygtis su parametrais. Kokso modelio dalinio tikėtinumo metodas vykdomas taip: tarkime, kad duomenys gaunami stebint n asmenų, ir sakome, jog iš jų turimi skirtingi įvykių laikai d ir $n - d$ yra iš dešinės cenzūruotas išgyvenamumo laikas. Dėl patogumo, tarkime, jog kiekvieną kartą miršta tik vienas asmenuo, taigi duomenys neturi jokių sąryšių. Pažymime nurodytus skirtingus išgyvenamumo laikus $y_{(1)}, y_{(2)}, \dots, y_{(d)}$, ir $y_{(j)}$ yra j -tasis nurodytas išgyvenamumo laikas. Asmenys, kurie yra rizikos grupėje laiku $y_{(j)}$ pažymimi R_j . Šiai asmenų aibei jokie įvykiai neįvyksta ir išgyvenamumo laikas necenzūruojamas, tik prieš $y_{(j)}$. Dydis R_j vadinamas rizikos aibe. Kokso modelio dalinio tikėtinumo funkcija parametrai β aprašoma: [4]

$$PL(\beta|D) = \prod_{j=1}^d \frac{\exp(x_{(j)}''\beta)}{\sum_{i \in R_j} \exp(x_i''\beta)}, \quad (2.3)$$

kur $x_{(j)}$ reiškia $p \times 1$ individo kovariantės vektorius, kuris patiria įvykį nurodytu išgyvenamumo laiku $y_{(j)}$. (2.3) formulės vardiklį esantis sumavimas yra visų asmenų verčių, kurie patiria riziką laike t , \exp suma. Reikia atkreipti dėmesį, kad sandauga yra imama asmenų, kuriems įvykių laikai buvo užregistruoti. Asmenų, kuriems išgyvenamumo laikas cenzūruotas duomenys nenaudojami (2.3) formulės skaitiklyje, bet jie tikrai naudojami įvykio laikų rizikos aibės sumavime, kurie įvyksta prieš cenzūruotą stebėjimą. Be to, ši tikėtinumo funkcija priklauso tik nuo įvykio laikų eilės (reitingavimo), nes taip nustatoma kiekvieno įvykio laiko rizika. Todėl parametras β priklauso tik nuo išgyvenamumo laikų sudaryto eiliškumo.

Regresijos koeficientas β proporcingo pavojingumo modelyje interpretuojamas taip: $\exp(\beta)$ parodo, kiek kartų padidėja rizikos funkcija, kovariantės reikšmei padidėjus 1. Jei $\beta > 0$, didėjant x , individo rizika didėja; o jei $\beta < 0$, didėjant x , rizika mažėja. Dydis $\exp(\beta)$ dar vadinamas pavojaus rizika.

Programose, skirtose išgyvenamumo duomenims apdoroti, pateikiami regresijos koeficientų įverčiai, jų standartinės paklaidos bei kriterijaus, skirto β_i reikšmingumui tikrinti, p reikšmės. Statistiniuose paketuose taip pat pateikiami $\lambda_0(t)$ bei $S_0(t)$ įverčiai taškuose $t_1 < t_2 < \dots < t_i$, čia t_i – skirtingi gyvavimo ar stebėjimo laikai.

Interpretuojant Kokso modelio rezultatus, svarbu ir pavojaus rizikos $\exp(\beta_i)$ patikimumas, todėl statistiniuose paketuose pateikiami dydžiai $\exp(\beta_i)$ su pasikliautinaisiais intervalais. Jei $\exp(\beta_i)$ pasikliautinąjo intervalo ribos didesnės nei 1, i-toji kovariantė didėdama reikšmingai didina baigties taško riziką. Jei $\exp(\beta_i)$ pasikliautinąjo intervalo ribos mažesnės nei 1, i-tajai kovariantei didėjant, rizika reikšmingai mažėja. Jei į $\exp(\beta_i)$ pasikliautinąjį intervalą patenka 1, i-toji kovariantė išgyvenamumui reikšmingos įtakos neturi.

1.3.1 PROPORCINGO PAVOJINGUMO MODELIS

Daugėjant klinikinių tyrimų 1950-aisiais ir 1960-aisiais, žymiai padidėjo poreikis analizuoti cenzūruotus išgyvenamumo duomenis, o dideliu laimėjimu šioje srityje tapo Kokso proporcingo pavojingumo modelis, paskelbtas 1972 m. (Cox, 1972). Dabar, išgyvenamumo duomenų regresinė analizė tapo įmanoma.

Rizikos funkcija apskirtai priklauso ir nuo laiko ir nuo kovariantės aibės (nes kai kurios iš jų gali būti taip pat priklausomos nuo laiko). PHM atskiria šiuos komponentus, nurodant, kad rizika asmeniui laiku t , kurio kovariantės vektorius Z yra apibrėžiama [4]:

$$\lambda(t|Z) = \lambda_0(t) \exp\{G(Z, \beta)\},$$

kur $\lambda_0(t)$ - neneigiama bazinio pavojingumo funkcija, o β regresijos koeficiento parametras. Antra dalis aprašyta \exp forma, nes turėtų būti teigiama. Šis modelis reiškia, kad rizikos santykis dviem asmenims yra pastovus laike, ir kovariantės laikui bėgant nesikeičia. Įprasta manyti, jog kovariančių rezultatas yra multiplikatyvus, todėl rizikos funkciją galima apibrėžti, kaip [4]

$$\lambda(t|Z) = \lambda_0(t) \exp(Z' \beta), \quad (3.1)$$

kur $\eta = Z' \beta$ yra pastovus prognozavimo parametras. Apibrėžtas (3.1) modelis reiškia, kad rizikos santykis dviem asmenims priklauso nuo skirtumo tarp prognozavimo bet kuriuo laiku.

1.3.2 MARTINGALO TAIKYMAS KOKSO MODELyje

Skaičiavimo proceso duomenų, įskaitant išgyvenamumo duomenis, analizė paprastai remiasi intensyvumo modeliavimu. Tegul atsitiktinis dydis T reiškia individo išgyvenamumo laiką, tuomet išgyvenamumo kreivė apskaičiuojama pagal $S(t) = P(T > t)$. Tai naudinga, kai išgyvenamumo duomenys padalyti į laiko ir į daugelį mažų intervalų. Tarkime, kai intervalo ilgis yra lygus Δt , kur Δt yra nedidelis. Darant prielaidą, kad T yra absoliučiai vientisas (t.y., turi tikimybinį tankį), žiūrima į tuos, kurie išgyveno iki tam tikro laiko t , ir apsvarstomas įvykio tikimybės apskaičiavimas, vykstantis per tam tikrą baigtinį laiko intervalą $[t, t + \Delta t)$. Bazinio pavojingumo dažnis $\lambda_0(t)$, įprastas visiems subjektams, yra apibrėžiamas kaip tokia riba: [7]

$$\lambda_0(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \quad (3.2)$$

(3.2) lygtyje apibrėžiamas $\lambda_0(t)$ gali būti aiškinamas kaip asmens staigios mirties (arba įvykio) rodiklis, su sąlyga, kad šis asmuo išgyveno iki laiko t . Visų pirma, $\lambda_0(t)\Delta t$ yra apytikslė mirties tikimybė, esanti $[t, t + \Delta t)$, atsižvelgiant į išgyvenamumą iki laiko t . Nors yra paprasta apskaičiuoti išgyvenamumo kreivę, sunkiau yra įvertinti pavojingumo rodiklį kaip sutartinę laiko funkciją. Kas vis dėlto yra gana lengva, tai įvertinti bazinį suvestinį rizikos dažnį, apibrėžiamą, kaip:

$$\lambda_0(t) = \int_0^t \lambda_0(s) \quad (3.3)$$

Kokso modelis aprašo rizikos funkciją (taip pat vadinamą intensyvumo funkcija) objektui i su kovariantėmis $Z_i = (Z_{i1}, \dots, Z_{ip})$, kaip [8]

$$\lambda(t|Z) = \lambda(t|Z=0) \exp[\beta'Z] = \lambda_0(t) \exp(\beta'Z) \quad (3.4)$$

Tai nežinomos bazinio pavojingumo rodiklio $\lambda_0(t)$ (neparametrinės dalies) ir eksponentinės funkcijos nežinomo regresijos koeficiento $\beta'Z = \sum_j \beta_j Z_{ij}$ (parametrinės dalies) rezultatas. Paprastai, β ir Z yra laikomos pastoviomis per laiko momentą t .

Šio modelio parametrai yra p -erdvinės regresijos parametras β ir neparametrinis bazinio pavojingumo dažnis $\lambda_0(t)$, kuris yra laikomas integraliu vietos mastu:

$$\int_0^{T_{\max}} \lambda_0(t) dt < \infty \text{ su } \max T \text{ reiškia laiko momentą, kuomet tyrimas yra nutraukiamas.}$$

Pavojingumo funkcija $\lambda(t|Z)$ individui su kovariantišku vektoriumi lygiu nuliui yra $\lambda_0(t)$. Tai yra pavojaus funkcija, kai nėra kovariančių, o tai paaiškina, kodėl ji vadinama bazinio pavojingumo funkcija. Pagrindinė prielaida, esanti (3.4) yra tai, kad santykinė rizika yra pastovi su laiku. Ypatingas atvejis, kur $p = 1$ ir Z_1 nekinta laikui bėgant, pavyzdžiui, gydymo rodiklis. Santykinės rizikos koeficientas aprašomas formule:

$$\frac{\lambda(t|Z_1 + 1)}{\lambda(t|Z_1)} = e^\beta. \quad (3.5)$$

Racionalus koeficientas (3.5) nepriklauso nuo laiko, nes tik bazinis intensyvumas atspindi priklausomybę nuo laiko. Jei yra daugiau nei viena kovariantė modelyje, gausime tą patį rezultato tipą, jei palyginsime du individus su tuo pačiu kovariantišku vektoriumi, išskyrus tai, kad jie skiriasi Z_1 (su verte 1) kaip ir formulėje (3.5). Pateiktas (3.4) modelis reiškia, kad mirties tikimybė kitame laiko intervale, atsižvelgiant į tai, kad pacientas gyveno iki šio laiko ir turi duotas reikšmes kovariantiškam vektoriumi, T klasifikavimui ir t.t., gali būti rasta dauginant bazinį pavojingumą $\lambda_0(t)$ iš paprasto logaritmo, pakeltu nepriklausomų kintamųjų tiesinės kombinacijos laipsniu. Kitaip tariant, tam tikro asmens mirties tikimybė įtakojama to kaip dažnai pacientai miršta ir tam tikro asmens individualiomis savybėmis. Jeigu paimsime tiesinės kombinacijos antilogaritmą, padauginsime, o ne pridėsime kovariančių reikšmes. Šiame modelyje, kovariantės turi didinimo arba proporcingą poveikį mirties tikimybei – tai yra, sąvoką "proporcingo pavojingumo" modelis.

Skaičiavimo proceso duomenų analizė, įskaitant išgyvenamumo duomenis, paprastai grindžiama intensyvumo modeliavimu. Jei turime n objektų, kurie tiriami objektui i , $i = 1, 2, \dots, n$, tai $I_i(t)$ yra intensyvumo procesas skaičiavimo procesui duotam kovariantiniam vektoriumi $Z_i = (Z_{i1}, \dots, Z_{ip})$, o $Y_i(t)$ rizikos rodiklis, t.y. objektų aibė, kuri vis dar objekto mirties i rizikoje tuo metu T_i , (t.y. gyvas ir necenzūruotas tuo laiko momentu tiesiog prieš laiką t). Be to, stebime procesą $N_i(t)$, kad suskaičiuotume mirčių skaičių, kuris įvyko $[0, t]$ intervale. Šis procesas yra pastovus ir lygus nuliui tarp mirčių ir pereina per vieną vienetą kiekvienos mirties momentu. Jei stochastinis procesas $\{N_i(t), t \geq 0\}$ yra skaičiavimo procesas, ir tuomet atitinka šias sąlygas [9]:

- $N_i(t) \geq 0$,
- $N_i(t)$ ima sveikąsias reikšmes;
- Jei $s < t$, $N_i(t) - N_i(s)$ parodo mirčių skaičių, kurios įvyko intervale $[s, t]$.

Vadinasi, naujos mirties koeficientas yra $I_i(t) = Y_i(t)\lambda(t|Z_i)$. Intensyvumas $I_i(t)$ gali būti apibūdinamas kaip tikimybė, jog svarbus įvykis įvyko per mažą laiko intervalą $[t, t + dt)$, turint omenyje, kad tai neįvyko anksčiau. Tai maždaug

$$dN_i(t) \approx \lambda(t|Z_i)dt = \lambda_0(t) \exp[\beta'Z_i] = I_i(t) \quad (3.6)$$

kur $dN_i(t)$ yra $N_i(t)$ padidėjimas per mažą intervalą $[t, t + dt)$ (t.y. stebimų mirčių skaičius, įvykusių $[t, t + dt)$). Taigi, $I_i(t)$ yra dauginimosi intensyvumas, kuris gali būti modeliuojamas

$$I_i(t) = Y_i(t)\lambda(t|Z_i) = Y_i(t)\lambda_0(t) \exp[\beta'Z_i] \quad (3.7)$$

kur intensyvumo procesas yra stebėto proceso ir nestebėtos funkcijos rezultatas. Taigi intensyvumo procesas $N_i(t)$ pagal modelį (3.4) yra:

$$I_i(t)dt = Y_i(t) \exp[\beta'Z_i] d\Lambda_0(t),$$

kur $d\Lambda_0(t)$ reiškia momentinę tikimybę, kad objektas rizikos momentu t patirs įvykį kitame laiko intervale $[t, t + dt)$.

Pristatant martingalo apibrėžimą vėlesniam laikui, pradedama nuo paprasto intensyvumo proceso apibrėžimo, kuris formuluojamas taip [10]:

$$I_i(t)dt = P(dN_i(t) = 1|past) = P([N_i(t + dt) - N_i(t)] = 1|past)$$

arba

$$I_i(t)dt = E(dN_i(t) = 1|past)$$

Dauginimosi intensyvumas $I_i(t)$ gali būti įterptas į praeities funkciją, t.y.

$$E[dN_i(t) - I_i(t)dt|past] = 0$$

Įvedamas sekantis procesas, iš (3.6) formulės

$$M_i(t) = N_i(t) - \int_0^t I_i(s) ds \quad (3.8)$$

kuris gali būti suvokiamas kaip

KLAIDA = STEBĖJIMAS – MODELIS

Procesą (3.8) dar galima užrašyti ir taip

$$E[dM_i(t)|past] = 0.$$

Taigi, paprasta intuityvi intensyvumo (greičio) proceso (3.7) sąvoka yra lygiavertė teiginiui, kad skaičiavimo procesas atėmus integruotą intensyvumo procesą yra martingalas.

Naudinga įvertinti integralo transformaciją [12]

$$S(t) = \prod_{(0,t]} \{1 - d\Lambda(s)\}$$

kadangi $d\Lambda_i(t) = \int_0^t I_i(s) ds$ yra intensyvumas atitinkantis pavojaus funkciją $I_i(t)$, tai

$$\prod_{(0,t]} \{1 - d\Lambda(s)\} = \exp\left(-\int_0^t I_i(s) ds\right)$$

Tarkime, kad asmenys buvo stebimi iki mirties ir tyrime cenzūruoti. Taigi turimi stebėti duomenys $D = \{N_i(t), Y_i(t), Z_i; i = 1, 2, \dots, n\}$ ir nežinomas parametras $\beta, \Lambda_0(t)$. Pagal neinformatyvų cenzūravimą, duomenų tikimybė skaidosi daugikliais, su viena išraiška, priklausomai tik nuo cenzūravimo proceso ir sekančios išraiškos:

$$L_i(D|\beta, \Lambda_i(t)) = \exp\left(-\int_{t \geq 0} I_i(t) dt\right) \prod_{t \geq 0} [I_i(t)]^{dN_i(t)}; i = 1, 2, \dots, n. \quad (3.9)$$

Taigi, pagal tikimybės modelį, kintamieji D turi savo jungtinį pasiskirstymą:

$$L(D|\beta, \Lambda(t)) = \prod_{i=1}^n L_i(D|\beta, \Lambda_i(t)) \quad (3.10)$$

1.3.3 KOKSO MODELIS IŠ BAJESO PERSPEKTYVOS

Pusiau parametrinis požiūris į Bajeso išvadą rizikos regresijos modeliuose paprastai įgauna (3.4) modelį. Sekant skaičiavimo procesą, aukščiau pristatytas žymėjimas objektui i ($i = 1, 2, \dots, n$) reiškia procesą, skaičiuojantį mirtis, įvykstančias iki laiko t , o $dN_i(t)$ yra mažas $N_i(t)$ padidėjimas per intervalą $[t, t + dt)$. $N_i(t)$ ir $dN_i(t)$ lygūs 1, jeigu įvykis įvyksta $[0, t)$ ir $[t, t + dt)$, atitinkamai, ir 0 priešingu atveju. Pagal neinformatyvų cenzūravimą, stebimo skaičiavimo proceso tikimybė yra proporcinga (3.9) ir išplaukia iš Puasono formos. Nors $dN_i(t)$ yra daugiausia vienas visiems i , t ; begalybės skaičiavimo procesas didėja, $dN_i(t)$, prisideda prie tikimybės taip pat kaip su nepriklausomais Puasono atsitiktiniais dydžiais su vidurkiais $I_i(t)dt$ per intervalą $[t, t + dt)$ [12]. Apibrėžus modelį pagal šią sistemą, intensyvumas gali būti laikomas tame intervale pastoviu. Efektyviems tolesniems skaičiavimams, įdiegiamas duomenų padidinimo metodas, grindžiamas tuo, kad $dN_i(t)$ yra nepriklausomi Puasono atsitiktiniai dydžiai pagal tikimybės išraišką (3.10), t.y.

$$dN_i(t) \sim \text{Puasono}(I_i(t)dt) \quad (3.11)$$

Kai kintamieji D stebimi, toliau domina aposteriorinis pasiskirstymas $P(\beta, \Lambda_0(t)|D)$. Bajeso teorema nusako $P(\beta, \Lambda_0(t)|D) \propto P(D, \beta, \Lambda_0(t))$. Tokiu būdu, tikimybės modelis gali būti išreikštas kaip jungtinis modelio parametrų aposteriorinis pasiskirstymas pagal Bajeso metodą, t.y.

$$P(\beta, \Lambda_0(t)|D) \propto L(D|\beta, \Lambda_0(t))P(\beta)P(\Lambda_0(t)) \quad (3.12)$$

Pagrindinis dėmesys skiriamas parametrui β . Kitas svarbus dydis yra bazinio pavojingumo funkcija, $\Lambda_0(t)$, kuri yra geriausiai suprantama kaip procesas laikui bėgant. Kadangi (3.12) yra sudėtingos formos, ji dažniausiai patogiai sutrumpinama naudojant modeliavimą. Taigi gali būti

atlikta naudojant Gibbs'o modelį. Įgyvendinant Gibbs'o modelį, koduojamą nuo nulio, reikia nustatyti ir tada sukonstruoti veiksmingą modeliavimo metodą kiekvienam iš dviejų pilnai susijusių sąlyginių aposteriorinių pasiskirstymų $P(\beta|D, \Lambda_0(t)), P(\Lambda_0(t)|D, \beta)$. Tai yra lygiai tie patys veiksmiai, kuriuos WinBUGS programinė įranga atlieka automatiškai. Įvairios Bajeso (3.11) modelio formulės iš esmės skiriasi neparimetrine $\Lambda_0(t)$ sąlyga.

1.3.4 BAZINIO PAVOJINGUMO MODELIAVIMAS NAUDOJANT DAUGIAKAMPIO FUNKCIJĄ

Naudojant idėjas iš Beamonte ir Bermúdez (2003) [7], apsvarstomas dauginimosi intensyvumo modelis (3.4) :

$$I_i(t) = Y_i(t)\lambda_0(t)\exp[\beta'Z_i]$$

Neparimetrinė modelio dalis, $\lambda_0(t)$, turėtų būti neneigiama daugiakampio funkcija su viršūnėmis, esančiomis ties momentais $a_0 = 0 < a_1 < \dots < a_T < a_{T_{\max}+1}$, kur daugiakampis atitinkamai paima reikšmes $\tau_0 = 0 < \tau_1 < \tau_2 < \dots < \tau_{T_{\max}} < \tau_{T_{\max}+1}$, ir tampa pastoviu po momento $a_{T_{\max}+1}$ bei gali būti užrašytas kaip:

$$\lambda_0(t) = \begin{cases} \tau_j + \frac{(\tau_{j+1} - \tau_j)(t - a_j)}{a_{j+1} - a_j}, & a_j \leq t \leq a_{j+1}, j = 1, \dots, T_{\max} \\ \tau_{T_{\max}+1}, & t \geq T_{\max} + 1 \end{cases} \quad (3.13)$$

Privalumas naudojant (3.13), vietoje įprastos laipsninės funkcijos tas, kad $I_i(t)$ bus tolydus. Gama apriorinis procesas įgyja nepriklausomą kumuliacinės rizikos didėjimą. Tai yra nerealu daugelyje taikomųjų nustatymų, ir neleidžia pasiskolinti vertės tarp gretimų intervalų. Sulyginimo idėja apdorojant apriorinį vektoriui τ , kuris gali būti apibrėžtas kaip auto-koreliuotas pirmos eilės procesas [15]. Tuomet

$$\tau_{j+1} = \tau_j \exp(e_j), j = 1, \dots, T_{\max} \quad (3.14)$$

Parametrų įverčiai ir 95% patikimų intervalų gali būti gauti iš MCMC pavyzdžių. Reikia laikytis atsargumo, nes per didelių parametrų pasirinkimas gali paveikti galutinių parametrų įverčius.

1.4 PROGRAMINIŲ PRIEMONIŲ PASIRINKIMAS IR APŽVALGA

Bajeso metodai nebuvo plačiai naudojami išgyvenamumo analizėje iki pastarųjų kelerių metų, nes tokiems realiems modeliams aposteriorinį pasiskirstymą pagal cenzūravimą tiesiogiai išgauti yra labai sunku. Tam tikslui ir buvo kuriamos ir tobulinamos programinės įrangos. Dažniausiai naudojamos:

- BATS programinė įranga skirta atlikti Bajeso metodo laiko eilučių analizę;
- MATLAB ir MINITAB programinės įrangos, kurių skaičiavimo algoritmai yra puikus įvadas Bajeso analizei atlikti;
- ir plačiausiai yra naudojamas BUGS programinės įrangos paketas, kurio pagalba galima analizuoti bendrus hierarchinius modelius naudojant MCMC metodą.

BUGS yra specializuotas programinės įrangos paketas, skirtas įgyvendinti MCMC metodu pagrįstas pilnų tikimybių modelių analizes, kuriose visi nežinomieji yra kaip atsitiktiniai dydžiai. Šis BUGS projektas pradėtas vykdyti Medicinos mokslinių tyrimų tarybos Biostatistikos skyriuje Kembridže 1989 metais. Nuo to laiko programinė įranga *WinBUGS* (*angl. the Windows version of Bayesian inference Using Gibbs Sampling*) plačiausiai ir tikslingiausiai naudojama sveikatos tyrimų analizės kontekste [17].

WinBUGS yra tokia programinė įranga, įgalinanti analizuoti išgyvenamumo duomenis, kuomet taikant Bajeso skirstinių vertinimo metodą naudojama Monte Karlo Markovo grandinė. Modeliai gali būti aprašomi formuluotėmis programinės įrangos BUGS kalba arba pavaizduoti grafine aplinka vadinama DoodleBUGS. WinBUGS apdoroja modelio rodiklius ir konstruoja analizuojamo objekto modelio atvaizdavimą.

Ši programinė įranga siūlo vartotojo sąsają, dialogo langus ir meniu komandas, per kurias galima analizuoti modelį naudojant MCMC metodus. Galima aprašyti programas, kurios kartu sudaro vientisą sąsają su WinBUGS specializuotais reikalavimais. Taip pat galima sąsaja su WinBUGS žemesniame lygmenyje, įtraukiant naujus objekto tipus, kurie gali būti naudojami WinBUGS nežinant modulių, kuriuose jie būtų įgyvendinti. Nei vienam iš šių išplėtojimo tipų nereikalingos priegios prie WinBUGS programos teksto perkompiliavimo iš naujo.

Vartotojui patogi „punkto ir spragtelėjimo“ aplinka leidžia naudoti Bajeso pilnojo tikimybės modelio statistines metodikas (MCMC metodo) klases. Konceptualus programinės įrangos modelis yra pagrįstas tikimybinio modelio konstravimu, kuris analogiškai vaizduojamas grafiniu modeliu. Grafiniai modeliai sudaryti iš mazgų (viršūnių) ir kraštinių (briaunų).

Statistinė modelio analizė atliekama, naudojant įvairius modeliavimo metodus, žinomus kaip Monte Karlo Markovo grandinės metodas (MCMC). Šis metodas leidžia performuoti imtį atsižvelgiant į nagrinėjamą aposteriorinę informaciją. Pagrindinis metodas yra Gibbs'o atranka, kurioje kiekviena iteracijos nauja reikšmė, kiekvieno nepastebėto stochastinio mazgo yra imama iš atitinkamo parametro sąlyginio pasiskirstymo. Tai reiškia, kad jo pasiskirstymas, priklauso nuo visų kitų parametrų ir duomenų. Šis pasiskirstymas turi būti konstruojamas paprasčiausiai iš mazgų (t.y. mazgų, kuriems svarbus yra pirminis mazgas). Šia prasme Gibbs'o atranka yra tiesiog lokalaus skaičiavimo seka diagramoje.

WinBUGS programos teksto struktūra taip pat analogiška grafiniam modeliui. Šiuo programinės įrangos projektavimo metodu siekiama sukurti visiškai ištesiamas modulinės sistemas. Programinė įranga susideda iš daug komponentų, kurie nesusiję iki įkėlimo ar net atlikimo laiko. Kiekvienas programinės įrangos komponentas turi aiškią sąsają, apibūdinančią tikslų vykdymus, kurie gali būti panaudoti ir kituose komponentuose. Komponento sąsaja aprašoma kodu kompiuteriui apdorojamame formate, vadinamu simbolių failu. Šis failas sąsajas leidžia tikrinti nuosekliai ir kompiliavimo metu ir bet kuriuo pasirinktu metu. Tokiu būdu programinės įrangos patikimumas yra geresnis. Norint aprašyti papildomus komponentus į failą įtraukiami nauji metodai ar programos. Atliekant šį veiksma nereikalingas jokios programinės įrangos dalies perkompiliavimas iš naujo.

Taip pat pasirinkta ir darbe naudojama *MedCalc programinė priemonė* [18]. MedCalc yra statistinė programinė priemonė, skirta biomedicininiams tyrimams atlikti. Patogi tuo, jog greitai ir lengvai atlieka išgyvenamumo analizę ir sugeneruoja Kaplano – Mejerio kreivę.

Atliekant klinikinius tyrimus dažnai domimasi laiku, kuriuo pacientams dalyvaujantiems tyrime įvyksta kažkoks konkretus įvykis arba baigtis (t.y. mirtis, nebedalyvauja tyrime ir t.t.). Pacientai pradedami stebėti nuo tam tikro pradinio taško ir laikas registruojamas, kai pasireiškia konkretus laukiamas įvykis. Paprastai tyrimas baigiamas dar iki tol, kol visi pacientai šio įvykio nesulaukia ir likusių pacientų baigtis lieka nežinoma. Be to nežinomas rezultatas ir tų pacientų kurie pasišalina iš tyrimo. Todėl dėl visų šių atvejų tolimesnis laikas yra registruojamas (duomenys cenzūruojami). Programinės priemonės MedCalc pagalba, šie duomenys gali būti nesunkiai analizuojami taikant išgyvenamumo lenteles ar Kaplano – Mejerio kreivę, nes tai yra dažniausiai naudojamas metodas išgyvenamumo charakteristikoms apibrėžti.

2 TIRIAMOJI DALIS

Darbe nagrinėtų metodų ir modelių taikymas demonstruojamas sprendžiant kelis praktinius uždavinius. Programinės įrangos WinBUGS pagalba atliekamas neapibrėžtų sveikatos duomenų ir išgyvenamumo analizė taikant Bajeso metodą ir Kokso modelį.

Pirmu atveju nagrinėjami dviejų pacientų grupių, sergančių leukemija ir gydytų 6-mp ir placebo preparatais remisijos duomenys ir pacientų išgyvenamumas. Antru atveju nagrinėjamas pacientų, sergančių gerklų vėžiu išgyvenamumas, tik šiuo atveju ligos baigtis klasifikuojama keturiomis stadijomis.

2.1 PACIENTŲ, SERGANČIŲ LEUKEMIJA IŠGYVENAMUMO DUOMENŲ ANALIZĖ

Šių dienų gydytojai, parinkdami gydymą pacientams, vis labiau vadovaujasi taip vadinama įrodymais pagrįsta medicina. Dauguma klinikinių tyrimų atliekami siekiant įsitikinti, kad naują vaistinį preparatą žmonėms vartoti ar naudoti yra saugu ir veiksminga. Kiekvienai ligai rekomenduojami gydyti vaistai skirstomi pagal tai, kiek yra atlikta jų veiksmingumą patvirtinančių klinikinių tyrimų. Norint nustatyti, ar naujas vaistas efektyvus, ar ne, vykdomi tyrimai su dviem pacientų grupėmis. Viena grupė pacientų gauna tiriamojo vaisto, o kita – placebo. Grupės visiškai vienodos, skiriasi tik gydymas. Pacientai į grupes parenkami atsitiktiniu būdu, ir visi turi vienodą tikimybę patekti į bet kurią iš grupių. Apie tai, kad vaistinės medžiagos vartojamame preparate gali nebūti, perspėjama iš anksto, tai gali sumažinti placebo veiksmingumą. Dalyvaujant klinikiniuose tyrimuose, tiek gydytojai, tiek pacientai prisideda prie medicinos mokslo progreso. Gydytojai sužino daugiau apie naujausius medikamentus, o pacientams suteikiama galimybė nemokamai pasinaudoti naujausiais gydymo būdais (kurie nėra prieinami kitais būdais), gauti kvalifikuotą sveikatos priežiūrą, ir kartu išbandyti naujų vaistų veiksmingumą. Tyrimas vykdomas informuoto paciento sutikimu. Atliekant klinikinius tyrimus, dėmesio centre yra pacientų saugumas bei privatumas. Privaloma laikytis tam tikrų reikalavimų, t.y. privalu laikytis biomedicininų tyrimų etikos ir, kiek įmanoma, užtikrinti, kad bet kokia numatoma rizika neviršytų įprastinio gydymo ar pačios ligos keliamos rizikos. Taigi toks tyrimas vykdomas tik tuo atveju, jei numatoma nauda pateisina riziką.

Šiame darbe nagrinėjamas pacientų, sergančių leukemija, dviejų grupių remisijos laikas. Duomenys viešai publikuoti Gehan (1965a) [20]. Tyrime dalyvauja 42 pacientai. Pirmos grupės gydymui taikomas 6 – merkaptopurino preparatas, o antrai grupei - placebo. 6 – merkaptopurinas (6-mp) yra citostatinis preparatas, kuris naikina leukemines ląsteles neleisdamas joms dalintis. O placebo vadinamas išoriškai į vaistą panašus, bet veikliosios cheminės medžiagos neturintis preparatas. Asmenys nežino, ar jiems skirtas gydymas, ar placebo. Placebu kontroliuojamas tyrimas būna koduotas arba dvigubai koduotas, t. y. jei pacientas žino apie tai, kokį vaistą vartoja – vadinama atviru tyrimu. Jei pacientas nežino, bet gydytojas žino apie vartojamą preparatą – koduotas tyrimas. Ir jei nežino nei pacientas, nei gydytojas, kuriai iš grupių priklauso konkretus pacientas – dvigubai koduotas tyrimas. Nagrinėjamu atveju tyrimas yra dvigubai koduotas. Nepaisant to, kad kai kuriems tyrime dalyvaujantiems asmenims gali būti skiriama placebo, dalis jų vis tiek pasijunta geriau. Taip nutinka, nes žmogus įtiki, jog naujas gydymas jam padės (tai dar vadinama placebo efektu). Kitaip tariant, skiriamas gydymas turi dvejopą poveikį, t.y. fiziologinį ir psichologinį. Psichologinis poveikis turi skirtis nuo tiriamojo vaistinio preparato fiziologinio poveikio, t.y. vaistas turi būti veiksmingesnis nei placebo. Jei tiriamasis preparatas nėra veiksmingas, jo vartojimo rezultatai panašūs į placebo.

Nagrinėjamas periodas yra remisijos laikas savaitėmis, pateikiamas 1 lentelėje:

1 lentelė. Leukemija sergančių pacientų remisijos duomenys

1 Grupė (n = 21 pacientas) Gydymas preparatu 6-mp	6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+
2 Grupė (n = 21 pacientas) Gydymas placebo	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Žymėjimas $t+$ nurodo stebimo laiko t cenzūravimą iš dešinės. Iš 42 stebimų remisijos laikų, 12 buvo cenzūruojami, 17 skirtingų remisijos atvejų buvo stebimi, ir praeitas laikas buvo cenzūruotas.

2.1.1 PACIENTŲ, SERGANČIŲ LEUKEMIJA IŠGYVENAMUMO FUNKCIJŲ VERTINIMAS KAPLANO – MEJERIO METODU

Tyrime pacientų išgyvenamumo funkcijos skirtingose grupėse vertinamos Kaplano – Mejerio metodu. Šiuo metodu gaunami išgyvenamumo funkcijos rezultatai ir grafikas. Siekiama išsiaiškinti rizikos faktorius, kurie turi įtakos tyrimui. Gaunami pacientų, sergančių leukemija ir gydytų preparatais 6-mp ir placebo, remisijos laikų Kaplano – Mejerio įverčiai.

Skirtingi remisijos laikai t_j (savaitėmis) yra $t_0 = 0, t_1 = 6, t_2 = 7 \dots t_{21} = 36$. Rezultatai pateikti 2 lentelėje.

2 lentelė. Pacientų, sergančių leukemija ir gydytų preparatu 6-mp, išgyvenamumo funkcijos vertinimas.

Laikas (savaitės)	Bendras pacientų skaičius	Remisijos skaičius momentu t	Cenzūruotų reikšmių momentu t skaičius	Išgyvenamumo funkcija
0	21	0	0	1
6	21	3	1	0.857
7	17	1	1	0.807
10	15	1	2	0.753
13	12	1	0	0.690
16	11	1	3	0.628
22	7	1	0	0.538
23	6	1	5	0.448
>23	-	-	-	-

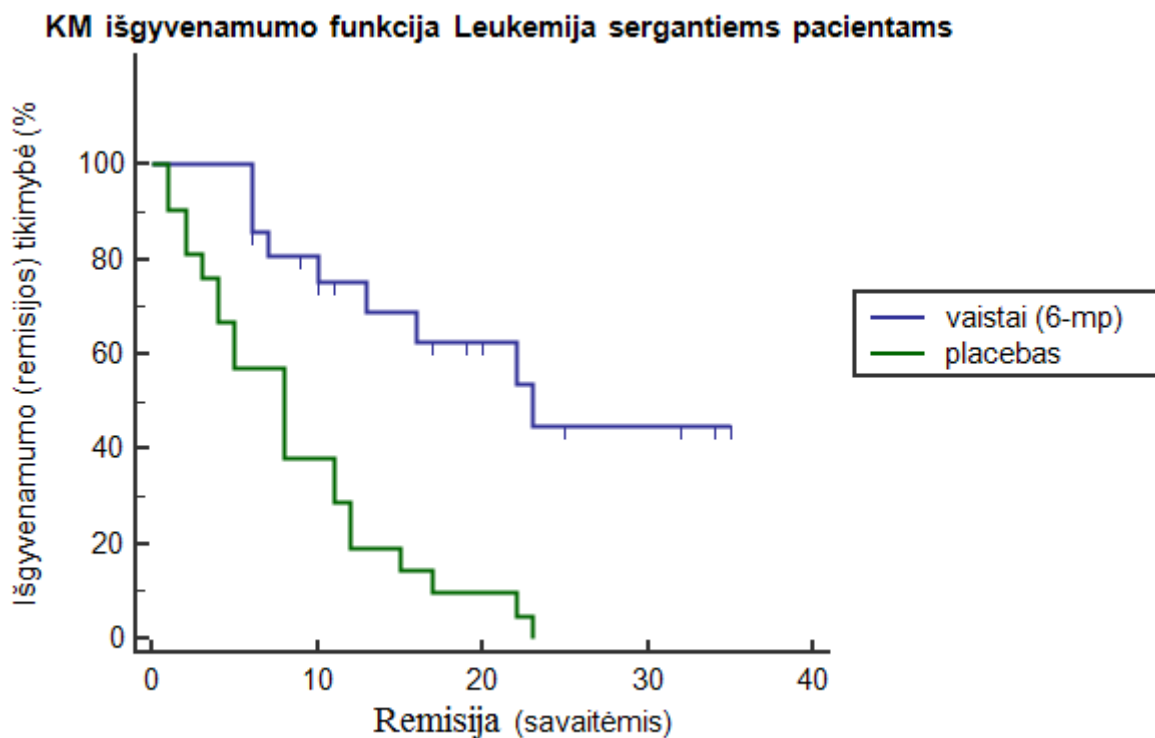
Remiantis 2 lentelėje pateiktu išgyvenamumo funkcijos Kaplano – Mejerio įverčiu, galima teigti, kad pacientų, sergančių leukemija ir gydytų preparatu 6-mp tikimybė remisijai neatsirasti per 6 savaites po gydymo yra apie 0,857; per 10 savaičių – 0,753; per 23 savaites – 0,448.

3 lentelė. Pacientų, sergančių leukemija ir gydytų placebo, išgyvenamumo funkcijos vertinimas

Laikas (savaitės)	Bendras pacientų skaičius	Remisijos skaičius momentu t	Cenzūrotų reikšmių momentu t skaičius	Išgyvenamumo funkcija
0	21	0	0	1
1	21	2	0	0.90
2	19	2	0	0.81
3	17	1	0	0.76
4	16	2	0	0.67
5	14	2	0	0.57
8	12	4	0	0.38
11	8	2	0	0.29
12	6	2	0	0.19
15	4	1	0	0.14
17	3	1	0	0.10
22	2	1	0	0.05
23	1	1	0	0

Remiantis 3 lentelėje pateiktu išgyvenamumo funkcijos Kaplano –Mejerio įverčiu, galima teigti, kad pacientų, sergančių leukemija ir gydytų placebo tikimybė remisijai neatsirasti per 1 savaitę po gydymo yra apie 0,90; per 11 savaičių – 0,29; per 22 savaites – 0,05.

Naudojant programinę priemonę MedCalc, Kaplano – Mejerio metodu įvertintos išgyvenamumo funkcijos abiem pacientų grupėms pateikiamos grafiškai 4 pav.



4 pav. Pacientų, sergančių leukemija ir gydytų preparatais 6-mp ir placebo, Kaplano – Mejerio išgyvenamumo kreivės

Taikant šį metodą šių pacientų grupių išgyvenamumo funkcijos skiriasi, tačiau nėra aišku kaip reikšmingai. Visgi galima pastebėti, kad lyginant su placebo ir jo galimai pozityviu poveikiu 6-mp preparatas efektyviau stabdo vėžio remisiją.

2.1.2 PACIENTŲ, SERGANČIŲ LEUKEMIJA IŠGYVENAMUMO ANALIZĖ NAUDOJANT PROGRAMINĘ ĮRANGĄ WINBUGS

Tolesniame tyrimo etape programinės įrangos WinBUGS pagalba taikant darbe nagrinėtus metodus ir modelius atliekama išgyvenamumo analizė. Pateikiama išgyvenamumo statistinių duomenų santrauka ir duomenų rinkinio aposteriorinio pasiskirstymo grafiniai vaizdavimai.

Siekama parodyti, kaip svarbu vertinti kovariantes prognozuojant išgyvenamumo pasiskirstymą. Grupės pacientų skaičiui priskiriamos kovariantės apibrėžiamos Z . Modeliavimas vykdomas su

- $Z_i = 0.5$, jei pacientas gydomas placebo;
- $Z_i = -0.5$, jei pacientas gydomas vaistu 6-mp.

Išgyvenamumo kreivių įvertinimui pasirenkamas bazinio pavojingumo modeliavimas naudojant daugiakampio metodą. Tam apibrėžiamos daugiakampio funkcijos viršūnės esančios ties momentais $a=c(2,4,6,8,10,12,14,16,18,20,22,24,26,2,30,32,34,36)$.

Aposteriorinio pasiskirstymo grafinis vaizdavimas gali parodyti problemas, susijusias su MCMC algoritmo vykdymu. Tokiam vaizdavimui programinė įranga WinBUGS turi daug priemonių. Viena iš jų yra paprasta rezultatų santrauka, pateikianti aposteriorinį vidurkį, medianą ir standartinį nuokrypį su 95% pasikliautinumo intervalu. Aposteriorinės informacijos radimo procesas yra iteracinis. Todėl vykdant modeliavimą atliekama 10000 skaičiavimų iteracijų.

4 lentelė. Programinės įrangos WinBUGS rezultatai, aposteriorinė statistika

Viršūnė	Vidurkis	Standartinis nuokrypis	MC klaida	2.5%	Mediana	97.5%	Iteracijos
beta	1.42	0.4168	0.008083	0.6489	1.405	2.276	10000
sigma	0.1647	0.1581	0.0116	0.001176	0.1287	0.5447	10000

Iš 4 lentelėje pateiktų rezultatų matome, jog placebo gydomos pacientų grupės palyginti su 6-mp grupe santykinė rizika yra $\exp(1,42) = 4.14$. Tai parodo, kiek kartų padidėja rizikos funkcija, kovariantės reikšmei padidėjus 1.

Toliau pateikiamos ir lyginamos dviejų modelių DIC vertės, t.y. DIC, kai bazinio pavojingumo vertinimui naudojama daugiakampio funkcija ir DIC su bazinio pavojingumo vertinimui naudojamu gama procesu.

Programinė įranga WinBUGS automatiškai realizuoja DIC modelio palyginimo kriterijų. Taip vertinamas Informacijos nukrypimo kriterijų (angl. Deviance Information Criterion) ir su juo susiję statistiniai duomenys. Šis kriterijus įvertina modelio sudėtingumą ir palygina skirtingus modelius. P_D yra „efektyvus parametru skaičius“, gaunamas $P_D = \bar{D} - \hat{D}$. Taigi, P_D yra nuokrypio ir aposteriorinio vidurkio nuokrypio skirtumo, aposteriorinis vidurkis. Stebimo taško nuokrypis \bar{D} apibrėžiamas $(-2 \times \log(\text{tikimybė}))$. Tikimybė apibrėžiama, kaip $p(y|\theta)$ ir \hat{D} yra įvesties įvertis, pastaroji remiasi aposteriorinio vidurkio reikšmės parametrais. Tuomet $DIC = \bar{D} + P_D = \hat{D} + 2P_D$. Modelis su mažiausia DIC verte geriau prognozuoja tos pačios struktūros duomenų rinkinio atkartojimą.

5 lentelė. Programinės įrangos WinBUGS rezultatai. DIC, kai bazinio pavojingumo vertinimui naudojama daugiakampio funkcija

	\bar{D}	\hat{D}	P_D	DIC
dN	204.489	200.011	4.478	208.968
Bendras	204.489	200.011	4.478	208.968

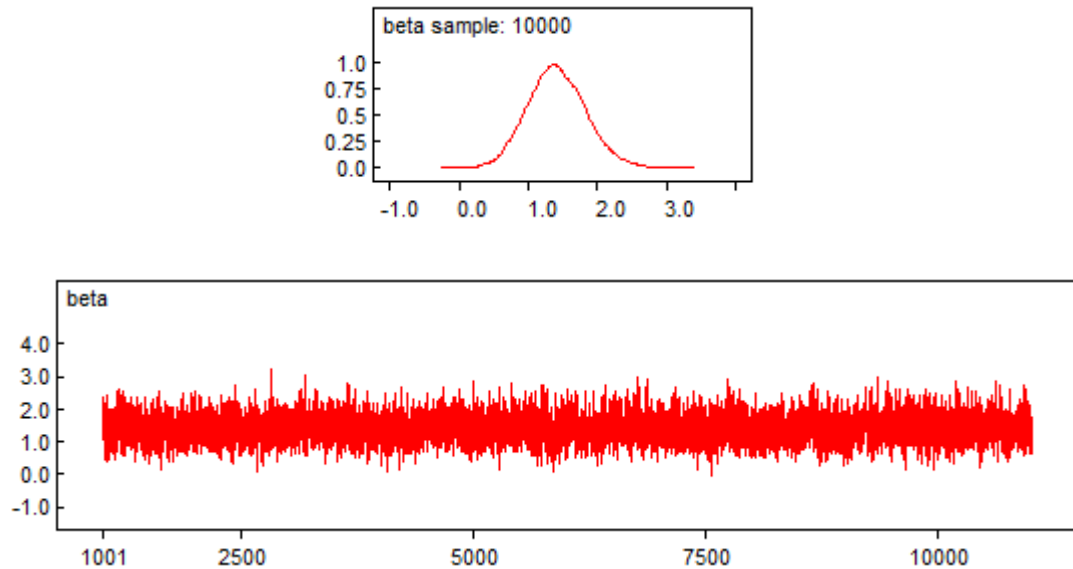
6 lentelė. Programinės įrangos WinBUGS rezultatai. Aposteriorinė statistika, jei bazinio pavojingumo vertinimui naudojamas gama apriorinis procesas

Viršūnė	Vidurkis	Standartinis nuokrypis	MC klaida	2.5%	Mediana	97.5%	Iteracijos
beta	1.535	0.4155	0.004894	0.7508	1.526	2.373	10000

7 lentelė. Programinės įrangos WinBUGS rezultatai. DIC, kai bazinio pavojingumo vertinimui naudojamas gama procesas

	\bar{D}	\hat{D}	P_D	DIC
dN	212.756	192.919	19.837	232.593
Bendras	212.756	192.919	19.837	232.593

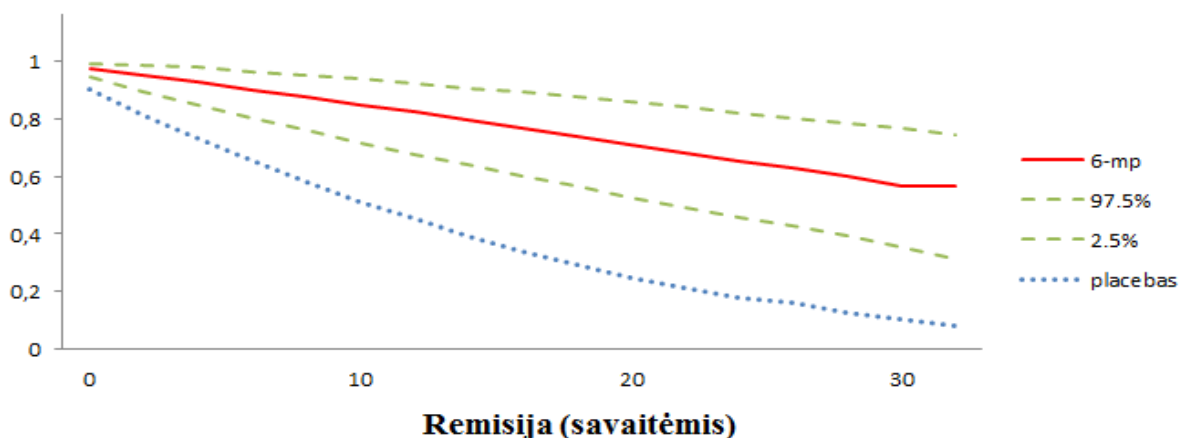
Lyginamos DIC vertės pateiktos 5 lentelėje ir 7 lentelėje. Matome, jog pirmuoju atveju $DIC=208.968$, o su gama procesu $DIC=232.593$. Modelis su mažiausia DIC verte geriau prognozuoja tos pačios struktūros duomenų rinkinio atkartojimą. Taigi šiuo atveju DIC vertė, kai neparametrinėje Kokso modelio dalyje bazinio pavojingumo vertinimui naudojama daugiakampio funkcija yra tinkamesnė nei vertinimui su aprioriniu gama procesu.



5 pav. Programinės įrangos WinBUGS rezultatai susiję su kovariantės koeficientais. Pateikiamas pacientų grupės gydomos vaistais aposteriorinis tankis

Išsamesnis aposteriorinio pasiskirstymo vaizdavimas pateikiamas tankio funkcijos grafiku. Sugeneravus MCMC, 5 pav. vaizduoja galutinės imties vystymą, naudojant kovariantės koeficientų reikšmes abiem pacientų grupėms ir jo ribinio pasiskirstymo Monte Karlo vertinimą. Visa modelio hipermetrija rodo priimtina susiliejimą stacionariam pasiskirstymui.

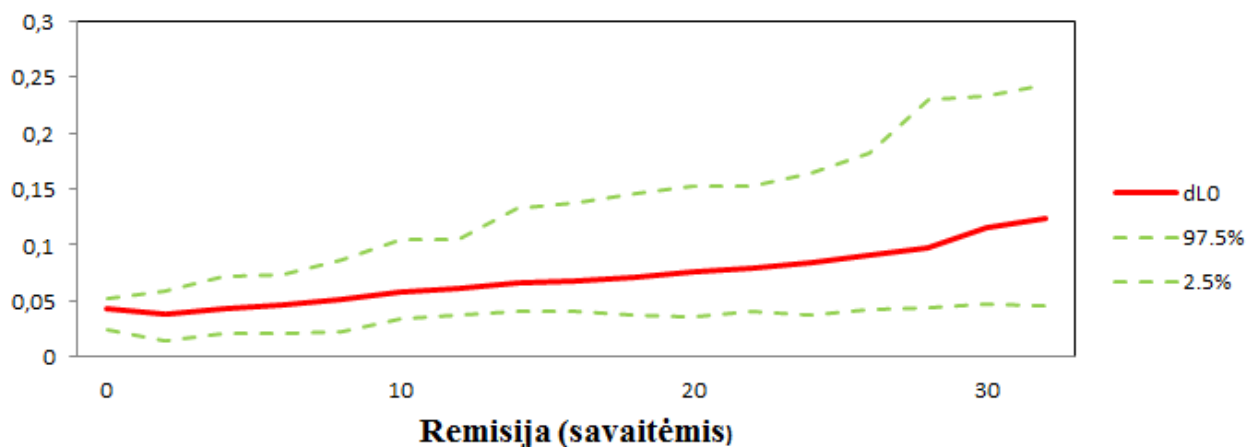
Išgyvenamumo vertinimas dviem pacientų grupėms, gydytoms 6-mp ir placebo preparatais



6 pav. Programinės įrangos WinBUGS rezultatai. Išgyvenamumo vertinimas

Atlikus bazinio pavojingumo modeliavimą naudojant Kokso modelį, daugiakampio funkciją ir įvertinus remisijos duomenų ir kovariančių poveikį išgyvenamumo funkcijoms gaunamos 6 pav. prognozuojamos išgyvenamumo funkcijos abiem leukemija sergančių pacientų grupėms gydytoms preparatais 6-mp ir placebo. Tikimybė, kad praėjus tam tikram laikui pacientas išgyvens vis artėja prie nulio, kai vartojami vaistai, matome toks artėjimas sulėtėja, t.y. mirtingumas sumažėja, nes kovojama su liga.

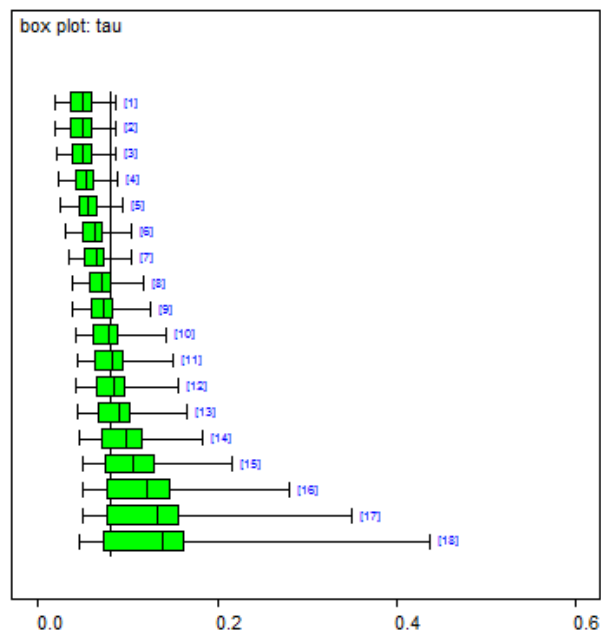
Neapibrėžto bazinio pavojingumo vertinimas naudojant daugiakampio funkciją



7 pav. Programinės įrangos WinBUGS rezultatai. Bazinio pavojingumo vertinimas naudojant daugiakampio funkciją

Bazinio pavojingumo modeliavimo atveju, egzistuoja nuo kitų parametrų ir laiko labai nepriklausantis mirimo intensyvumas, 7 pav. vaizduojamas kaip tiesė, su galimo neapibrėžtumo ribomis. Tai tarsi pastovus „mirimo greitis“, tai reiškia, jog kasmet dalis pacientų miršta pastoviai. Jis gali keistis priklausomai nuo to, kai yra liga – didėja, jei naudojami vaistai – sugrįžta į pastovų lygį arba net pradeda mažėti.

Žemiau vaizduojama stačiakampė diagrama (angl. boxplot), kurioje pateikiamas apibendrintas visų taško elementų aposteriorinis pasiskirstymas.



8 pav. Programinės įrangos WinBUGS rezultatai. Parametro tau stačiakampė diagrama

Iš šios diagramos galima spręsti apie bendrą matuojamo kintamojo imties centrą, išsibarstymo, minimalios bei maksimalios reikšmių vaizdą, t.y. parodo grafinį skaitinių imties charakteristikų išsidėstymą. Diagramoje esantis stačiakampis („dėžė“) braižomas nuo apatinio kvartilio iki viršutinio kvartilio ir padalintas brūkšniu į dvi dalis ties mediana.

2.2 PACIENTŲ, SERGANČIŲ GERKLŲ VĖŽIU IŠGYVENAMUMO DUOMENŲ ANALIZĖ

Nagrinėjamas pacientų, sergančių gerklų vėžiu praktinis uždavinys, tik šiuo atveju ligos baigtis klasifikuojama keturiomis stadijomis. Kardaun (1983) pateikė ataskaitą su viešai publikuojamais duomenimis apie 90 vyrų diagnozuotą gerklų vėžį Nyderlandų ligoninėje 1970-1978 metais [21].

Ataskaitoje pateiktas laikas yra intervalai tarp gydymo pradžios ir mirties/tyrimo pabaigos. Taip pat užregistruotas nustatytos pacientų diagnozės amžius, diagnozės metai ir ligos stadijos. Tačiau šiame tyrime didžiausias dėmesys skiriamas šios ligos stadijų išgyvenamumo trukmės priklausomybei vertinant kovariantes, todėl diagnozės amžiaus ir diagnozės metų duomenys į tyrimą neįtraukiami.

Paprastai skiriamos 4 vėžio stadijos, t. y. pirma arba antra stadija - maži ar neišplitę už organo ribų navikai, trečia - išplitę į gretimas struktūras ir ketvirta stadija - išplitę į kitus, toli nuo pirminio židinio esančius organus navikai. Šios 4 ligos stadijos apibrėžtos tarptautinėmis klasifikacijos taisyklėmis. Ligos išplitimui (t.y. stadijai) apibūdinti naudojama T.N.M. simbolių sistema:

- T reiškia naviko dydį;
- N – sritinius limfmazgius;
- M – vėžio metastazės toli nuo pirminio naviko organuose.

Ligos stadija nustatoma, kai, ištyrus pacientą, įvertinamos TNM reikšmės.

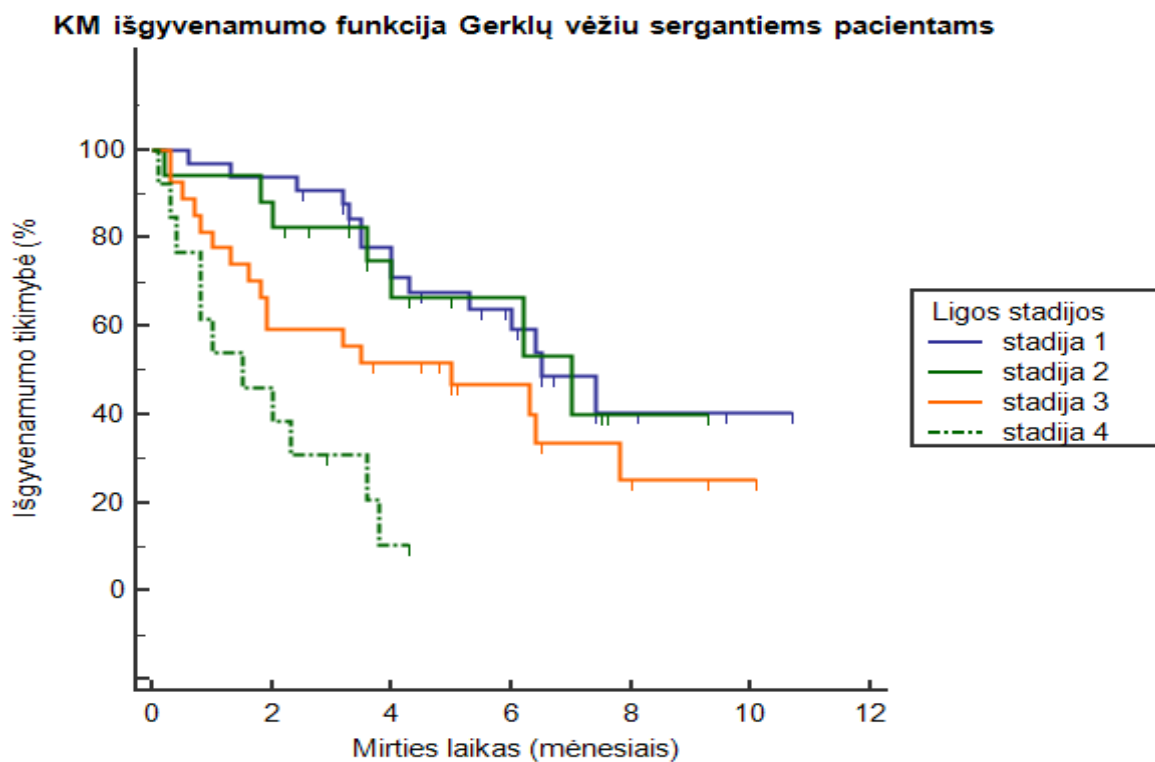
Šiame darbe ligos stadijų žymėjimas apibrėžiamas 1, 2, 3 ir 4 nuo mažiausiai iki labiausiai rimtos:

- Stadijai 1 priklauso 33 pacientai;
- Stadijai 2 – 17;
- Stadijai 3 – 27;
- Stadijai 4 – 13 pacientų.

Iš tiriamos pacientų grupės, 50 vyrų mirė nuo vėžio nesulaukę tyrimo pabaigos, todėl iš stebėtų gyvavimo laikų 40 buvo cenzūruoti iš dešinės.

2.2.1 PACIENTŲ, SERGANČIŲ GERKLŲ VĖŽIU, LIGOS IŠPLITIMO DUOMENŲ IŠGYVENAMUMO FUNKCIJŲ VERTINIMAS KAPLANO – MEJERIO METODU

Naudojant programinę priemonę MedCalc, Kaplano – Mejerio metodu apskaičiuotos šių keturių ligos stadijų išgyvenamumo funkcijos pateikiamos grafiškai (išgyvenamumo lentelės rezultatai pateikiami 1 priede).



9 pav. Pacientų, sergančių gerklų vėžiu, ligos stadijų, Kaplano – Mejerio išgyvenamumo kreivės

Taikant šį metodą šių pacientų stadijų išgyvenamumo funkcijos skiriasi, tačiau nėra aišku kaip reikšmingai. Todėl tolesniame tyrimo etape vertinamas išgyvenamumo funkcijos neapibrėžtumas.

2.2.2 PACIENTŲ, SERGANČIŲ GERKLŲ VĖŽIU LIGOS IŠPLITIMO DUOMENŲ IŠGYVENAMUMO ANALIZĖ NAUDOJANT PROGRAMINĘ ĮRANGĄ WINBUGS

Tolesniame tyrimo etape, programinės įrangos WinBUGS pagalba taikant darbe nagrinėtus metodus ir modelius atliekama išgyvenamumo analizė. Pateikiama išgyvenamumo statistinių duomenų santrauka ir duomenų rinkinio aposteriorinio pasiskirstymo grafiniai vaizdavimai.

Siekiami parodyti, kaip svarbu vertinti kovariantes prognozuojant išgyvenamumo pasiskirstymą. Šiam tyrimui apibrėžiamos kovariantės:

$Z_1 = 1$, jei paciento ligos stadija 2;

$Z_2 = 2$, jei paciento ligos stadija 3;

$Z_3 = 3$, jei paciento ligos stadija 4.

Išgyvenamumo funkcijos vertinimui pasirenkamas bazinio pavojingumo modeliavimas naudojant daugiakampio metodą. Tam apibrėžiamos daugiakampio funkcijos viršūnės esančios ties momentais:

$a=c(0.2,0.5,0.8,1.1,1.4,1.7,2,2.3,2.6,2.9,3.2,3.5,3.8,4.1,4.4,4.7,5,5.3,5.6,5.9,6.2,6.5,6.8,7.1,7.4,7.7,8,8.3,8.6,8.9,9.2,9.5,9.8,10.1,10.7)$.

Aposteriorinės informacijos radimo procesas yra iteracinis. Todėl vykdant modeliavimą atliekama 10000 skaičiavimų iteracijų.

8 lentelė. Programinės įrangos WinBUGS rezultatai, aposteriorinė statistika

Viršūnė	Vidurkis	Standartinis nuokrypis	MC klaida	2.5%	Mediana	97.5%	Iteracijos
beta.dis[1]	-0,0984	0,3245	0,0063	-0,7487	-0,0934	0,5286	10000
beta.dis[2]	0,3434	0,1184	0,0028	0,1148	0,3424	0,5828	10000
beta.dis[3]	0,5109	0,09158	0,0041	0,3345	0,5087	0,6944	10000
sigma	0,5053	0,1825	0,0061	0,2283	0,4781	0,9238	10000

8 lentelės rezultatų santraukoje pateikiamas aposteriorinis vidurkis, mediana ir standartinis nuokrypis su 95% pasikliautinumo intervalu.

9 lentelė. Programinės įrangos WinBUGS rezultatai. DIC, kai bazinio pavojingumo vertinimui naudojama daugiakampio funkcija

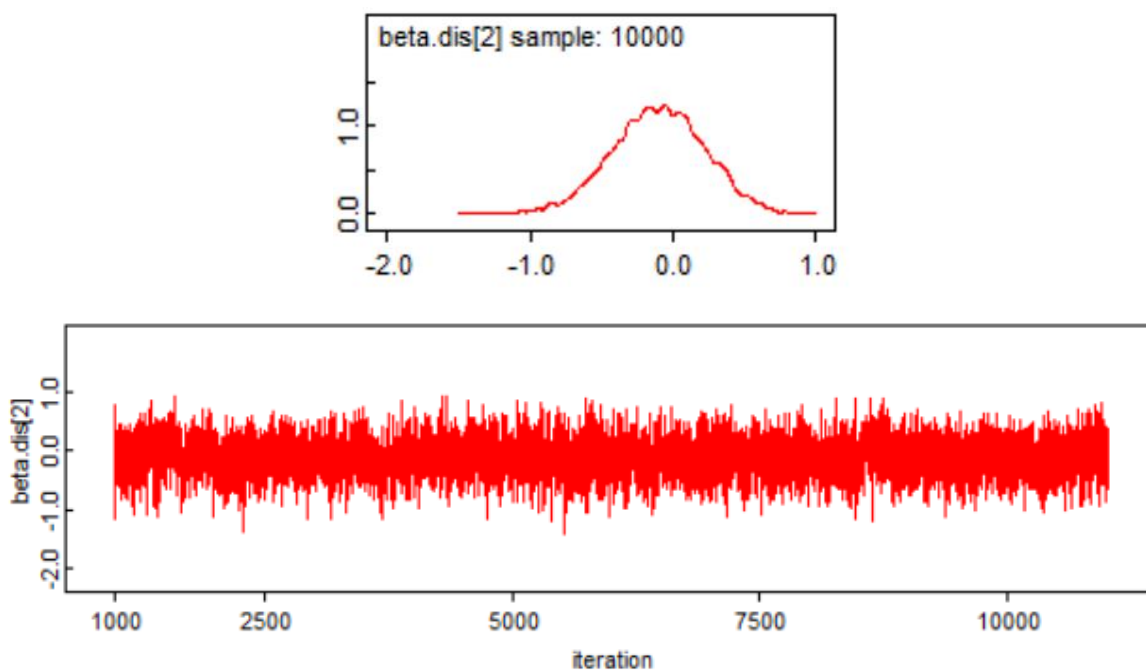
	\bar{D}	\hat{D}	P_D	DIC
dN	795.8	778.5	17.28	813.1
Bendras	795.8	778.5	17.28	813.1

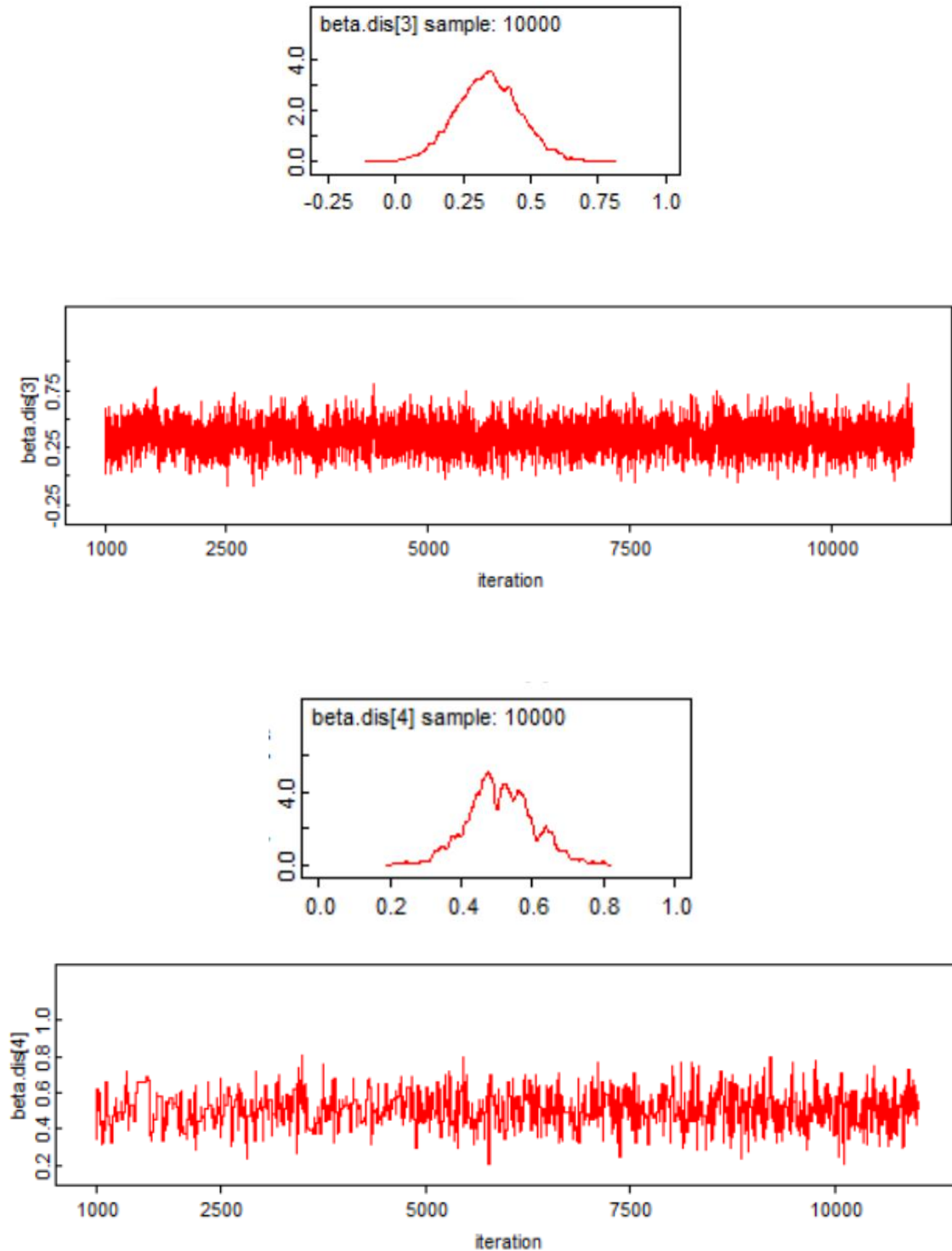
10 lentelė. Programinės įrangos WinBUGS rezultatai. DIC, kai bazinio pavojingumo vertinimui naudojamas gama procesas

	\bar{D}	\hat{D}	P_D	DIC
dN	823.7	770.1	53.6	877.3
Bendras	823.7	770.1	53.6	877.3

Lyginamos informacijos nukrypimo kriterijaus vertės 9 lentelėje ir 10 lentelėje. Matome, jog pirmuoju atveju $DIC = 813.1$, o su gama procesu $DIC = 877.3$. Modelis su mažiausia DIC verte geriau prognozuoja tos pačios struktūros duomenų rinkinio atkartojimą. Taigi šiuo atveju DIC vertė, kai neparametrinėje Kokso modelio dalyje bazinio pavojingumo vertinimui naudojama daugiakampio funkcija yra tinkamesnė nei vertinimui su aprioriniu gama procesu.

Išsamesnis aposteriorinio pasiskirstymo vaizdavimas atlikus 10000 skaičiavimo iteracijų pateikiamas tankio funkcijos grafikais.

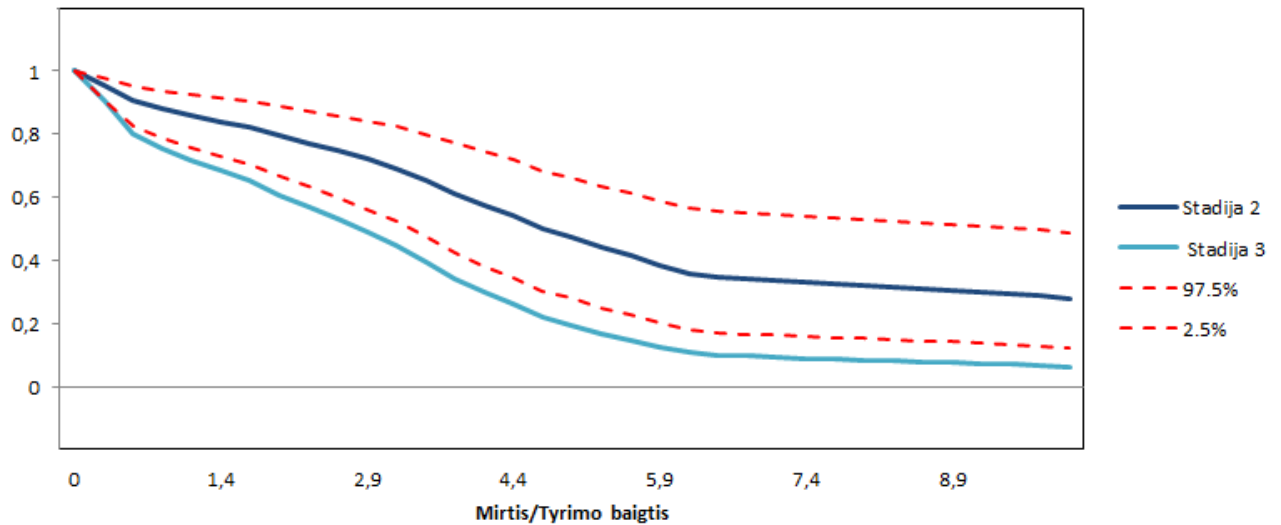




10 pav. Programinės įrangos WinBUGS rezultatai susiję su kovariantės koeficientais. Ligos stadijų aposteriorinis tankis

Sugeneravus MCMC, 10 pav. vaizduoja galutinės imties vystymą, naudojant kovariantės koeficientų reikšmes stadijoms ir jo ribinio pasiskirstymo Monte Karlo vertinimą. Visa modelio hypermetrija rodo priimtina susiliejimą stacionariam pasiskirstymui.

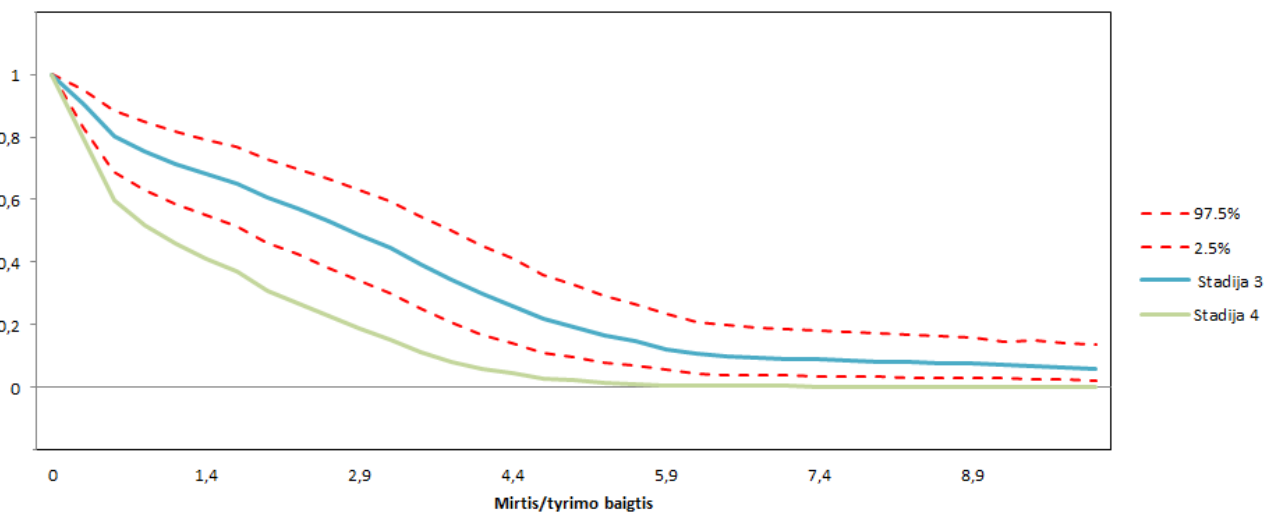
Išgyvenamumo vertinimas



11 pav. Programinės įrangos WinBUGS rezultatai. Išgyvenamumo vertinimas, stadijos 2 ir stadijos 3

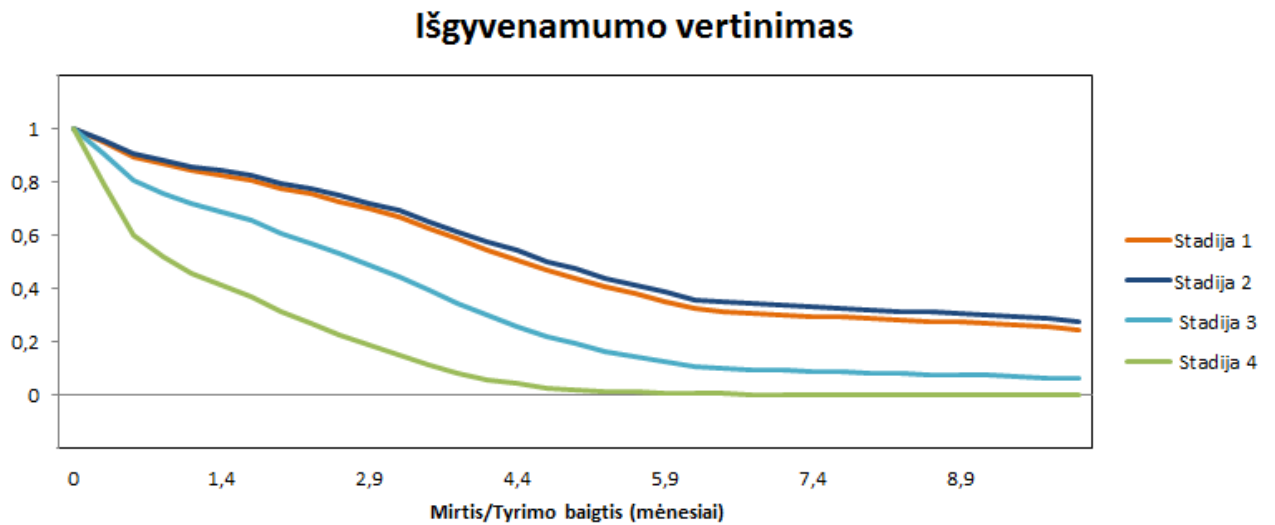
Vertinant išgyvenamumo priklausomybę nuo kovariančių nustatomas išgyvenamumo galimas pasiskirstymas. Tokiu būdu įvertinamas išgyvenamumo funkcijos neapibrėžtumas. Iš 11 pav. matoma, jog išgyvenamumas gerklų vėžio 2 stadijos ir 3 stadijos akivaizdžiai išsiskiria. 2 stadijos išgyvenamumo funkcija vaizduojama su galimo neapibrėžtumo ribomis.

Išgyvenamumo vertinimas



12 pav. Programinės įrangos WinBUGS rezultatai. Išgyvenamumo vertinimas, stadijos 3 ir stadijos 4

12 pav. matoma, jog išgyvenamumas gerklų vėžio 3 stadijos ir 4 stadijos akivaizdžiai išsiskiria. 3 stadijos išgyvenamumo funkcija vaizduojama su galimo neapibrėžtumo ribomis.



13 pav. Programinės įrangos WinBUGS rezultatai. Ligos stadijų išgyvenamumo vertinimas

Atlikus bazinio pavojingumo modeliavimą naudojant Kokso modelį, daugiakampio funkciją ir įvertinus remisijos duomenų ir kovariančių poveikį išgyvenamumo funkcijoms gaunamos 13 pav. skirtingų stadijų prognozuojamos išgyvenamumo funkcijos. Galima teigti, jog išgyvenamumas atskirose ligos stadijose akivaizdžiai išsiskiria ir tikimybė, kad praėjus tam tikram laikui pacientas išgyvens, tampa visiškai artima nuliui 4-oje ligos stadijoje praėjus maždaug 5-iems mėnesiams.

DARBO APŽVALGA IR IŠVADOS

Susipažinus su darbui aktualia teorine medžiaga galima teigti, jog Bajeso metodas gali būti pritaikomas išgyvenamumo analizei, kai turimi riboti ar cenzūruoti duomenys. Naudojant WinBUGS programinę įrangą išgyvenamumo neapibrėžtumo vertinimas su asmeniniam vartojimui skirtais kompiuterio resursais yra praktiškai įgyvendinamas. Taikant DIC kriterijų galima patvirtinti, kad neparametrinėje Kokso modelio dalyje bazinio pavojingumo vertinimui naudojama daugiakampio funkcija yra tinkamesnė nei vertinimui su aprioriniu gama procesu. Bajeso metodą realizavus jis leidžia akivaizdžiau klasifikuoti išgyvenamumo funkcijas nustatant reikšmingą kovariančių įtaką. Tai gali būti vertinama kaip gera alternatyva ieškant išgyvenamumo įverčių, palyginus su tais atvejais, kuomet taikomas maksimalaus tikėtimumo metodas arba kuomet modeliuojamas išgyvenamumo procesas taikant gama skirstinį.

Bajeso metodas taip pat turi keletą kitų pranašumų prieš klasikinius metodus, ypač modelių kūrimo galimybėmis, esant nevienarūšiams duomenims. Be to, remiantis gaunamu aposterioriniu pasiskirstymu Bajeso metodas leidžia padaryti akivaizdžią su išgyvenamumo klasifikavimu susijusią išvadą bet kokiam duomenų imties dydžiui. Tačiau, visgi iškyla problema, kad Bajeso metodo taikymo sudėtingumas ir galimai nekorektiškas naudojimas gali sąlygoti netinkamą duomenų analizę ir rezultatų interpretaciją.

Darbe nagrinėtų metodų ir modelių taikymas buvo demonstruojamas sprendžiant kelis praktinius uždavinius. Programinės įrangos WinBUGS pagalba buvo atlikta neapibrėžtų sveikatos duomenų ir išgyvenamumo analizė taikant Bajeso metodą ir Kokso modelį.

Pirmu atveju nagrinėtas dviejų pacientų grupių, sergančių leukemija ir gydytų 6-mp ir placebo preparatais remisijos duomenys ir pacientų išgyvenamumas. Pirmiausiai, abiem pacientų grupėms, gydytoms skirtingais preparatais, buvo įvertintos išgyvenamumo funkcijos taikant Kaplano – Mejerio metodą. Taikant šį metodą šių pacientų grupių išgyvenamumo funkcijos skiriasi, tačiau nėra aišku kaip reikšmingai. Visgi galima pastebėti, kad lyginant su placebo ir jo galimai pozityviu poveikiu 6-mp preparatas efektyviau stabdo vėžio remisiją. Tolesniame tyrimo etape buvo parodyta, kaip svarbu išgyvenamumo pasiskirstymo prognozavime atsižvelgti į kovariantes, priklausančias nuo laiko ir nagrinėjamų įvykių skaičiaus. Naudojant programinę įrangą WinBUGS buvo atliktas bazinio pavojingumo modeliavimas naudojant Kokso modelį, daugiakampio funkciją ir įvertinus remisijos duomenų ir kovariančių poveikį išgyvenamumo funkcijoms. Tikimybė, kad praėjus tam tikram laikui pacientas išgyvens vis artėja prie nulio, tačiau kai vartojami vaistai, toks artėjimas sulėtėja, t.y. mirtingumas sumažėja.

Antru atveju nagrinėtas pacientų, sergančių gerklų vėžiu išgyvenamumas, tik šiuo atveju ligos baigtis klasifikuojama keturiomis stadijomis. Kaplano – Mejerio metodu buvo apskaičiuotos išgyvenamumo funkcijos esant šiom keturioms ligos stadijomis. Tolesniame tyrimo etape taip pat buvo įvertinama išgyvenamumo priklausomybė nuo kovariančių ir nustatomas išgyvenamumo galimas pasiskirstymas. Tokiu būdu buvo įvertinamas išgyvenamumo funkcijos neapibrėžtumas. Į tai atsižvelgiant buvo patvirtinta, jog išgyvenamumas atskirose ligos stadijose akivaizdžiai išsiskiria. Be to, tikimybė, kad praėjus tam tikram laikui pacientas išgyvens, tampa visiškai artima nuliui 4-oje ligos stadijoje praėjus maždaug 5-iems mėnesiams.

PADĖKA

Norėčiau padėkoti savo darbo vadovui prof. dr. Robertui Alzbutui už paskatinimą studijuoti magistrantūroje, patarimus ir pagalbą rengiant šį darbą. Taip pat esu dėkinga Tomui Iešmantui už skirtą laiką ir naudingus patarimus rašant magistrinį darbą. Savo šeimai ir draugams dėkoju už nuolatinį palaikymą, kantrybę ir pagalbą.

LITERATŪRA

1. Vencloviėnė, J. Statistiniai metodai medicinoje. Kaunas, Vytauto Didžiojo universitetas, 2010.
2. Kaplan, E.L.; Meier, P. Non-parametric estimation from incomplete observations, 1958.
3. The Kaplan – Meier estimate of the survival function. Prieiga per internetą: <http://courses.nus.edu.sg/course/stacar/internet/st3242/handouts/notes2.pdf>.
4. J. G. Brahim; M. Chen; D. Sinha, Bayesian Survival Analysis. Springer – Verlag, New York, 2001.
5. Gelfand, A.E.; Smith, A.F.M. Sampling based approaches to calculating marginal densities, 1990.
6. Casella, G.; George, E.I. Explaining the Gibbs sampler, 1992.
7. Beamonte, E.; Bermúdez, J. D. A Bayesian semiparametric analysis for additive hazards models with censored observations, 2003.
8. Aalen, O. O. Nonparametric inference in connection with multiple decrement models, 1972.
9. Naes, T. The asymptotic distribution of the estimator for the regression Parameter in Cox's regression model, 1982.
10. Clayton, D. Bayesian analysis of frailty models. Technical Report, Medical Research Council Biostatistics Unit, Cambridge, 1994.
11. Gill, R.D. Product-integration, In Encyclopedia of Biostatistics. Wiley, Chichester, 2005.
12. Clayton, D. A Monte Carlo for Bayesian inference in frailty models. Biometrics, 1991.
13. Kalbfleisch, J. D. Nonparametric Bayesian analysis of survival time data. Journal of the Royal Statistical Society, 1978.
14. Kalbfleisch, J. D.; Prentice, R.L. The statistical analysis of failure time data. Wiley, New York, 1980.
15. Gamerman, D. Dynamic Bayesian models for Survival data, Applied Statistics, 1991.
16. Ntzoufras, I. Bayesian Modeling Using WinBUGS, 2009.
17. Prieiga per internetą: <http://www.mrc-bsu.cam.ac.uk/bugs/>.
18. Prieiga per internetą: <http://www.medcalc.org/>.
19. Prieiga per internetą: <http://www.cms.ac.uk/irm/irm.pdf>.
20. Kalbfleisch, J. D.; Prentice, R. L. The statistical analysis of failure time data. Wiley, New York, 1980.
21. Kardaun, O. Statistical analysis of male larynx-cancer patients – a case study. Statistical Nederlandica, 1983.

1 PRIEDAS. K-M IŠGYVENAMUMO FUNKCIJOS REZULTATAI NAUDOJANT PROGRAMINĘ PRIEMONĘ MEDCALC

Pacientų, sergančių leukemija ir gydytų 6-mp ir placebo preparatais išgyvenamumo lentelės:

Survival time	LAIKAS
Endpoint	REMISIJA
Factor codes	GRUPĖ

Cases summary

Factor	Number of events		Number censored		Total sample size
	N	%	N	%	
1	9	42,86	12	57,14	21
2	21	100,00	0	0,00	21
Overall	30	71,43	12	28,57	42

Mean and median survival

Factor	Mean	SE	95% CI for the mean	Median	95% CI for the median
1	23,287	2,827	17,746 to 28,829	23,000	13,000 to 23,000
2	8,667	1,377	5,967 to 11,366	8,000	4,000 to 11,000
Overall	15,339	1,212	12,963 to 17,715	12,000	8,000 to 17,000

Survival time	Factor				Overall	
	1		2			
	Survival Proportion	Standard Error	Survival Proportion	Standard Error	Survival Proportion	Standard Error
1	-	-	0,905	0,0641	0,952	0,0329
2	-	-	0,810	0,0857	0,905	0,0453
3	-	-	0,762	0,0929	0,881	0,0500
4	-	-	0,667	0,103	0,833	0,0575
5	-	-	0,571	0,108	0,786	0,0633
6	0,857	0,0764	-	-	0,714	0,0697
7	0,807	0,0869	-	-	0,690	0,0715
8	-	-	0,381	0,106	0,591	0,0764
9	-	-	-	-	-	-
10	0,753	0,0963	-	-	0,565	0,0773
11	-	-	0,286	0,0986	0,512	0,0788
12	-	-	0,190	0,0857	0,455	0,0796
13	0,690	0,107	-	-	0,426	0,0795
15	-	-	0,143	0,0764	0,398	0,0791
16	0,627	0,114	-	-	0,369	0,0784
17	-	-	0,0952	0,0641	0,341	0,0774
19	-	-	-	-	-	-
20	-	-	-	-	-	-
22	0,538	0,128	0,0476	0,0465	0,265	0,0765
23	0,448	0,135	0,000	0,000	0,189	0,0710
25	-	-	-	-	-	-
32	-	-	-	-	-	-
34	-	-	-	-	-	-
35	-	-	-	-	-	-
Endpoint: Observed n		9,0		21,0		
Expected n		19,3		10,7		
Observed/Expected		0,4675		1,9536		

Pacientų, sergančių gerklų vėžiu ligos stadijų išgyvenamumo lentelės:

Survival time	LAIKAS
Endpoint	MIRTIS
Factor codes	GRUPĖ

Cases summary

Factor	Number of events		Number censored		Total sample size
	N	%	N	%	
1	15	45,45	18	54,55	33
2	7	41,18	10	58,82	17
3	17	62,96	10	37,04	27
4	11	84,62	2	15,38	13
Overall	50	55,56	40	44,44	90

Mean and median survival

Factor	Mean	SE	95% CI for the mean	Median	95% CI for the median
1	7,083	0,656	5,797 to 8,370	6,500	5,300 to 7,400
2	6,428	0,863	4,736 to 8,119	7,000	4,000 to 7,000
3	5,127	0,805	3,549 to 6,705	5,000	1,800 to 7,800
4	1,928	0,418	1,110 to 2,747	1,500	0,800 to 3,600
Overall	5,748	0,301	5,158 to 6,338	6,000	3,800 to 7,000

Survival time	Factor								Overall	
	1		2		3		4		Survival Proportion	Standard Error
	Survival Proportion	Standard Error	Survival Proportion	Standard Error	Survival Proportion	Standard Error	Survival Proportion	Standard Error	Survival Proportion	Standard Error
0,1	-	-	-	-	-	-	0,923	0,0739	0,989	0,0110
0,2	-	-	0,941	0,0571	-	-	-	-	0,978	0,0155
0,3	-	-	-	-	0,926	0,0504	0,846	0,100	0,944	0,0241
0,4	-	-	-	-	-	-	0,769	0,117	0,933	0,0263
0,5	-	-	-	-	0,889	0,0605	-	-	0,922	0,0282
0,6	0,970	0,0298	-	-	-	-	-	-	0,911	0,0300
0,7	-	-	-	-	0,852	0,0684	-	-	0,900	0,0316
0,8	-	-	-	-	0,815	0,0748	0,615	0,135	0,867	0,0358
1	-	-	-	-	0,778	0,0800	0,538	0,138	0,844	0,0382
1,3	0,939	0,0415	-	-	0,741	0,0843	-	-	0,822	0,0403
1,5	-	-	-	-	-	-	0,462	0,138	0,811	0,0413
1,6	-	-	-	-	0,704	0,0879	-	-	0,800	0,0422
1,8	-	-	0,882	0,0781	0,667	0,0907	-	-	0,778	0,0438
1,9	-	-	-	-	0,593	0,0946	-	-	0,756	0,0453
2	-	-	0,824	0,0925	-	-	0,385	0,135	0,733	0,0466
2,2	-	-	-	-	-	-	-	-	-	-
2,3	-	-	-	-	-	-	0,308	0,128	0,722	0,0472
2,4	0,909	0,0500	-	-	-	-	-	-	0,711	0,0478
2,5	-	-	-	-	-	-	-	-	-	-
2,6	-	-	-	-	-	-	-	-	-	-
2,9	-	-	-	-	-	-	-	-	-	-
3,2	0,878	0,0573	-	-	0,556	0,0956	-	-	0,687	0,0491
3,3	0,845	0,0637	-	-	-	-	-	-	0,675	0,0497
3,5	0,778	0,0744	-	-	0,519	0,0962	-	-	0,638	0,0514
3,6	-	-	0,749	0,110	-	-	0,205	0,120	0,613	0,0524
3,7	-	-	-	-	-	-	-	-	-	-
3,8	-	-	-	-	-	-	0,103	0,0940	0,599	0,0528
4	0,710	0,0819	0,665	0,126	-	-	-	-	0,560	0,0540
4,3	0,676	0,0847	-	-	-	-	-	-	0,547	0,0543
4,5	-	-	-	-	-	-	-	-	-	-
4,8	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	0,467	0,0995	-	-	0,531	0,0550

5	-	-	-	-	0,467	0,0995	-	-	0,531	0,0550
5,1	-	-	-	-	-	-	-	-	-	-
5,3	0,639	0,0879	-	-	-	-	-	-	0,514	0,0560
5,5	-	-	-	-	-	-	-	-	-	-
5,9	-	-	-	-	-	-	-	-	-	-
6	0,593	0,0927	-	-	-	-	-	-	0,494	0,0572
6,1	-	-	-	-	-	-	-	-	-	-
6,2	-	-	0,532	0,156	-	-	-	-	0,473	0,0584
6,3	-	-	-	-	0,400	0,105	-	-	0,452	0,0596
6,4	0,539	0,0987	-	-	0,333	0,107	-	-	0,409	0,0612
6,5	0,485	0,103	-	-	-	-	-	-	0,387	0,0616
6,7	-	-	-	-	-	-	-	-	-	-
7	-	-	0,399	0,164	-	-	-	-	0,361	0,0627
7,4	0,404	0,113	-	-	-	-	-	-	0,334	0,0637
7,5	-	-	-	-	-	-	-	-	-	-
7,6	-	-	-	-	-	-	-	-	-	-
7,8	-	-	-	-	0,250	0,108	-	-	0,297	0,0666
8	-	-	-	-	-	-	-	-	-	-
8,1	-	-	-	-	-	-	-	-	-	-
9,3	-	-	-	-	-	-	-	-	-	-
9,6	-	-	-	-	-	-	-	-	-	-
10,1	-	-	-	-	-	-	-	-	-	-
10,7	-	-	-	-	-	-	-	-	-	-
Endpoint: Observed n		15,0		7,0		17,0		11,0		
Expected n		22,6		10,0		14,1		3,3		
Observed/Expected		0,6647		0,6992		1,2070		3,2957		

2 PRIEDAS. PROGRAMINĖS ĮRANGOS WINBUGS APRAŠYMAS

DAGs ir Bajeso statistinė metodika

Grafiniai modeliai

Siekiant iliustruoti statistinius metodus, kuriuos naudoja WinBUGS, atsižvelgiama į paprastą vienmatės tiesinės regresijos modelį. Tariama, jog turimas stebėjimas y_i išmatuojamas $x_i, i = 1, \dots, N$, kur x_i yra gaminio bandymai (pvz. žinomi stebėjimo laikai). Jei daroma prielaida, kad tiesinis ryšys egzistuoja tai

$$y_i \sim N(\mu_i, \tau^{-1})$$

$$\mu_i = \alpha + \beta x_i,$$

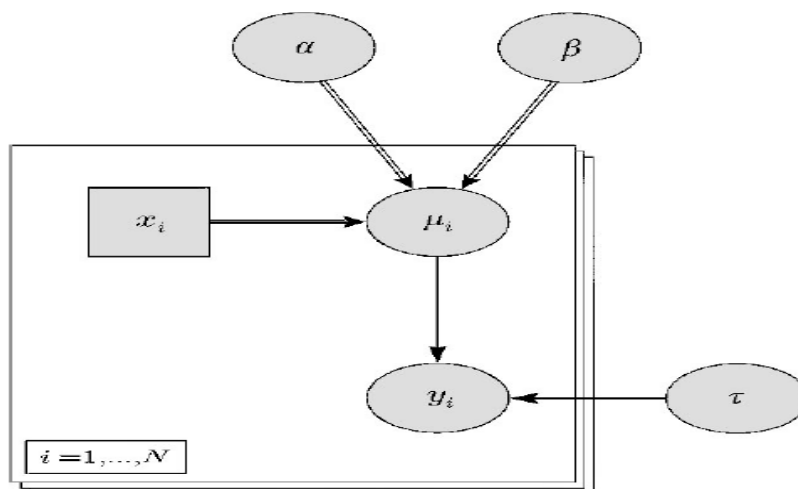
čia $i = 1, \dots, N$;

α ir β yra nežinomos koordinatų ašyje ir atitinkamai gradiento parametrai;

τ - liekanos dispersija, taip pat nežinoma.

Alternatyvus šio modelio vaizdavimas nurodomas neperiodiniu grafiku (DAG graph). Kiekvienas kiekis modelyje atitinka susikirtimo tašką (node) ir ryšiai tarp šių taškų rodo tiesioginę priklausomybę. Grafikas aiškus, nes kiekvienas ryšys yra vaizduojamas rodykle.

Šio grafiko žymėjimai apibrėžiami 1 pav. Stačiakampio formos taškai žymi žinomas konstantas. Elipsės formos taškai vaizduoja deterministinius ryšius (t.y. funkcijos) arba stochastinius kiekius, t.y. kiekius, kurie reikalauja pasiskirstymo prielaidos.



1 pav. Paprastos tiesinės regresijos kryptinis neciklinis grafas

WinBUGS programinė įranga yra suprojektuota Bajeso modeliams, todėl nežinomi parametrai, tokie kaip α , β ir τ yra stochastiniai dydžiai, nes kiekvienam turi būti priskirtas apriorinis pasiskirstymas. Stochastinė priklausomybė ir funkcinė priklausomybė atitinkamai apibrėžtos vienguba ir dvigubomis rodyklėmis. Pasikartojančiai struktūrai, tokiai kai ciklas nuo $i=1$ iki $i=N$, vaizduoja „lentelė“, kurios gali būti įterptos, jei modelis yra hierarchinis.

Mazgas v , turi būti mazgo w aukštesnės klasės elementas jei rodyklė išeinanti iš v taškų į w taškus, tada sakoma kad w yra žemesnės klasės elementas. Didžiausias dėmesys kreipiamas į stochastinius mazgus, t.y. nežinomus parametrus ir duomenis. Nustatant tikimybinis ryšius tarp jų, deterministiniai ryšiai nutraukiami ir į konstantas neatsižvelgiama. Taigi terminai aukštesnės klasės ir žemesnės klasės yra skirti atitinkamiems atsitiktiniams dydžiams.

DAGs naudojami, kad iliustruotai pavaizduotų labai plačią statistinių modelių klasę. Todėl, kai šie modeliai pasidaro sudėtingi, toks vaizdavimas tampa labai naudingas. DAGs grafai perduoda modelio esminę struktūrą netaikant sudėtingų lygčių. Tai pasiekama nukreipimu: taikomos pasiskirstymo prielaidos ir deterministiniai sąryšiai yra „paslėpti“.

DAG sudaro sąlyginių nepriklausomybės prielaidų seką, t.y. bet kokiam mazgui v , jei aukštesnės klasės yra žinomos, tuomet jokie kiti mazgai nepateikia papildomos informacijos apie v , išskyrus v . Sąlyginis nepriklausomumas, išreikštas per DAGs grafus, leidžia gauti modelio ypatybes net nepaisant to, kad jokia konkreti tikimybinė forma nebuvo apibrėžta. DAG pavaizdavimas labai palengvina sudėtingą pilnų tikimybes modelių analizę.

Bajeso modelio taikymas naudojant Gibbs atranką

Tarkime, kad turimi duomenys y ir nežinomi parametrai θ . Bajeso metodas statistikoje taikomas visų nežinomų parametrų vertinimui kaip su atsitiktiniais dydžiais ir kiekvienam priskiria tikimybes pasiskirstymą. Apibrėžiamas bendras tikimybes pasiskirstymas duomenims, t.y. gaunamas pilnas tikimybes modelis visiems tikėtiniams ir ne dydžiams.

Tradicinėje statistikoje dažnai pradedama nuo tikimybes modelio apskaičiavimo, t.y. tikimybes stebėtų duomenų, kad modelis, $P(\text{Data}|\text{Model})$. Yra lyginamos dviejų modelių santykinių dydžių tikimybes su modeliu, kuris labiau tikėtina, kad pateiks stebėtus duomenis, teikiant pirmenybę. Bajeso modelio pritaikymas, šį procesą vertina viena pakopa smulkiau, skaičiuojant $P(\text{Model}|\text{Data})$. Todėl pateikti modeliai gali būti įvertinti, tiesiogiai lyginanti jų tikimybes. Apskaičiuoti $P(\text{Model}|\text{Data})$ tikimybes reikalinga Bajeso teorema:

$$P(\text{Model}|\text{Data}) = \frac{P(\text{Data}|\text{Model})P(\text{Model})}{P(\text{Data})}$$

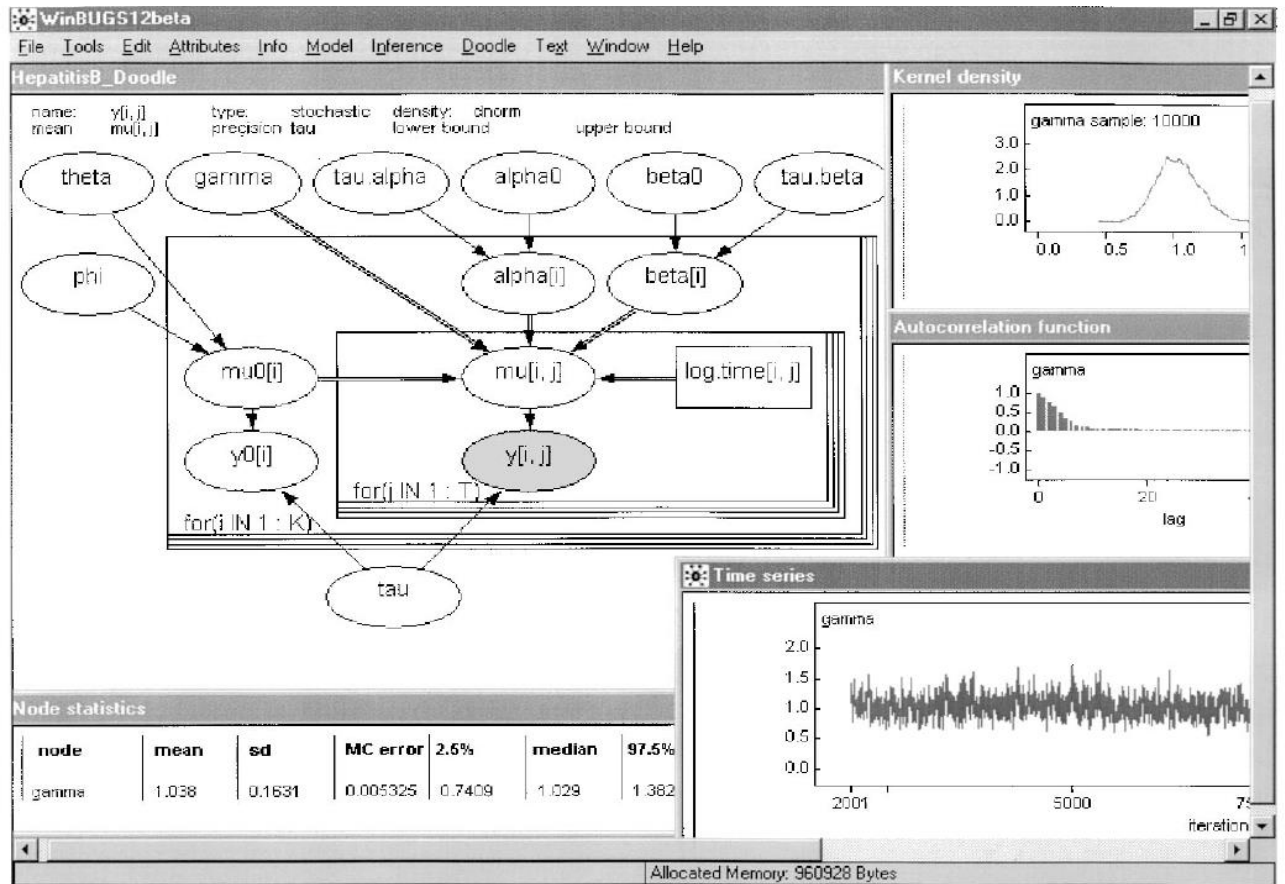
Kairės pusės išraiška nurodo duotų duomenų modelio tikimybę, ir taip gali būti apskaičiuojama tik po to, kai matomi duomenys. Tai paprastai vadinama aposteriorine modelio tikimybe. Dešinėje pusėje, $P(\text{Data}|\text{Model})$ yra tikimybė, ir $P(\text{Model})$ yra modelio tikimybė be žinomų duomenų. Lengviausias būdas yra kai šio modelio tikimybės duomenys žinomi, tuomet tai yra apriorinė modelio tikimybė. Ir galiausiai $P(\text{Data})$ yra trukmė, kuri garantuoja, kad aposteriorinė tikimybė būtų tarp 0 ir 1.

Reikalavimai ir grafinių modelių analizė

Su WinBUGS programine įranga modeliai gali būti nurodomi tekstine formuluote arba grafiškai. Abiem atvejais, duomenys ir nežinomiems parametrų pradinės reikšmės yra apibrėžti „S-Plus-like“ formatu arba stačiakampiame rinkinio formate. Tekstinis formatas apibrėžiamas naudojant deklaratyvią programavimo kalbą grupę, žinomą kaip BUGS kalba. Kitas kodas apibrėžia regresijos problemos tikimybę, pavaizduotai 1 pav.

```
for (i in 1:N) {
  y[i] ~ dnorm(mu[i], tau)
  mu[i] <- alpha + beta * x[i]
}
```

Žymėjimas \sim reiškia, kad „yra pasiskirstymas, kaip“. Toliau, visada turi būti nurodomas pasiskirstymo žymėjimas, šiuo atveju $\text{dnorm}(\cdot, \cdot)$. WinBUGS normalus pasiskirstymas yra parametrizuoti atsižvelgiant į tikslumą, o ne į dispersijas. Žymėjimas $<-$ skaitomas kaip „gauti“ ir nustato loginius (ar deterministinius) ryšius.



2 pav. WinBUGS programinės įrangos langas

2 paveiksle iliustruojamas WinBUGS programinės įrangos lango pavyzdys. Jame vaizduojama hepatito B skiepų modelio DoodleBUGS diagrama (aptarta Spiegelhalter (1996a)). Duomenys, $y[i, j]$, yra serijos matavimai, gauti iš ($K = 106$) Gambijos kūdikių nuo Hepatito B skiepų tyrimo. Grafikas vaizduoja hierarchinį atsitiktinių padarinių modelį, su i – kūdikių indeksas ir j – kartojamų priemonių kiekvienam kūdikiui indeksavimas. Daroma prielaida, jog $y[i, j]$ yra nepriklausomos sąlygos nuo jų vidurkio $\mu[i, j]$ ir nuo parametro τ , kuris nurodo matavimo proceso tikslumą. Pastebėtos pradinės „log-titre“ vertės $y0[i]$, priklausančios toms pačioms skaičiavimo paklaidoms ir taip pat priklausančios nuo τ . Kiekviena $\mu[i, j]$ reikšmė yra determinuota $\log.time[i, j]$ funkcija, gradiento parametrai atitinkamai ($\alpha[i]$ ir $\beta[i]$), klaidos kintamuosiuose („true“ pradinio log-titre: $\mu0[i]$) ir su ja susijęs γ koeficientas. Pasirinkta tiesinė forma ($\mu[i, j] <- \alpha[i] + \beta[i] * \log.time[i, j] + \gamma * \mu0[i]$), bet tiesiškumas nėra būtinas. Kai mazgas pasirenkamas, loginio mazgo funkcinė forma įvedama grafo viršuje. Parametrai $\alpha[i]$, $\beta[i]$ ir $\mu0[i]$ yra sudaryti nepriklausomai nuo gyventojų gimstamumo pasiskirstymo, parametrizuoto pagal: $\alpha0$ ir $\tau.alpha$; $\beta0$ ir $\tau.beta$; θ ir ϕ atitinkamai.

Grafiniai reikalavimai atliekami naudojant DoodleBUGS sąsają. 2pav. pavaizduota DoodleBUGS, kaip ekvivalentiškas Hepatito B imunizavimo modelis, aptartas Spiegelhalter (1996a). Sujungia atsitiktinius padarinius į didėjančios kreivės modelį su matavimo paklaida. Mazgai ir nukreiptos sąsajos yra sudaryti, naudojant paprastas operacijas. Išsami informacija apie pasiskirstymo prielaidas ar loginių funkcijų detales rodoma diagramoje, kai yra pasirinktas mazgas ($y [i, j]$). Visa tai, gali būti keičiama vartotojo. 2 pav. taip pat nurodo kai kuriuos rezultatus, kurie pagrįsti iteracijų 2001 – 12000 analize, naudojant vaizdavimo modelį. Tokie rezultatai gali būti gaunami bet kuriuo analizės metu. Atitinkami aposterioriniai modeliai paprastai paimami iš reikiamo stebimo objekto ir paveikiami atitinkamai.

Pagal abstraktaus grafinio modelio vaizdavimą nesudėtinga suprasti sąlygines nepriklausomybės prielaidas ir taip nustatyti kiekvieną sąlyginį pasiskirstymą, neatsižvelgiant į pagrindines prielaidas.

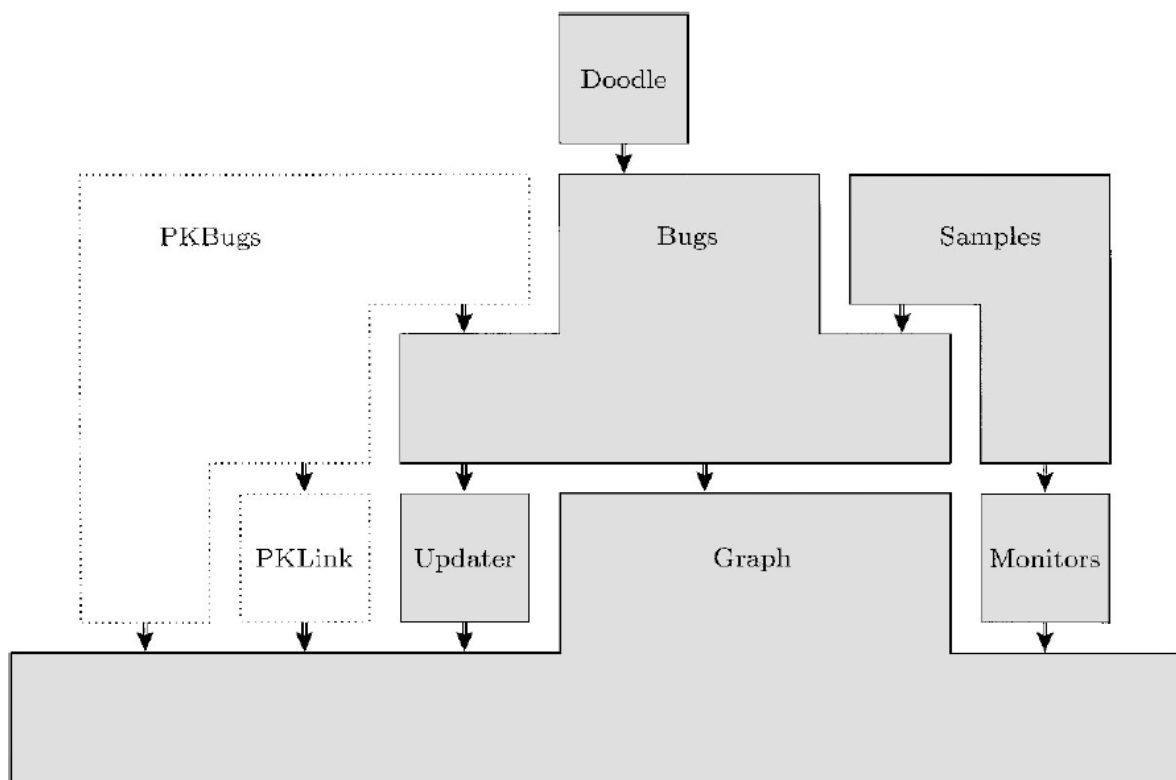
WinBUGS struktūra

WinBUGS programinė įranga buvo sukurta, naudojant BlackBox priemonę (Oberon mikrosistemos, Inc., Ciurichas: <http://www.oberon.ch>). Tai yra greitas taikomas kūrimo įrankis, skirtas išplėsti BlackBox struktūrą, kuri yra nepriklausomai ištesinama klasė. Programavimas vykdomas, naudojant Pascal kalbą, kuri jungia procedūras, objektinį orientavimą, ir modulinius pavyzdžius.

Programinė įranga parašyta naudojant BlackBox yra skirstoma į modulius. Modulis yra kompiliacijos pagrindas ir paprastai yra apibrėžiamas vienu ar daugiau tipų, apimančių jungimo procedūras, kai reikalinga informacija. Šie tipai gali būti eksportuoti greta konstantų, kintamųjų, ir procedūrų. Eksportuoti elementai sudaro sąsają, per kurią gali kartu veikti skirtingi moduliai. Priemonės BlackBox sąsajos atlieka pagrindinę funkciją. Kai modulis yra sudarytas iš sąsajų aprašymo, automatiškai sukuriama kompiliatorius. Toks sudarymas turi būti aiškus, užbaigtas, ir glaustas. Tai neturėtų būti neapibrėžta arba nustatomi nereikšmingi duomenys, arba pateikiamos klaidingos prielaidos.

Modulis „black box“, veikia su jos aplinka tikrai per sąsajas. Akivaizdžiausia modulinio programavimo nauda yra ta, kad pagrindinė problema skaidoma į mažesnes, kurios sprendžiamos nepriklausomai ir bendra problema tampa gerokai labiau valdoma. Tačiau, sąveika tarp modulių vykdoma per paprastas sąsajas, kurios taip pat labai svarbios. Iki tol kol sąsaja išlieka ta pati, modulis gali būti pakeistas naujesnės versijos, nekeičiant programinės įrangos, o tai labai palengvina įrangos plėtojimą.

Moduliai, kurių funkcijų rinkiniai susiję vadinami posisteme. 3 pav. vaizduoja WinBUGS struktūrą, ir parodo kaip šešios pagrindinės posistemės kartu veikia.

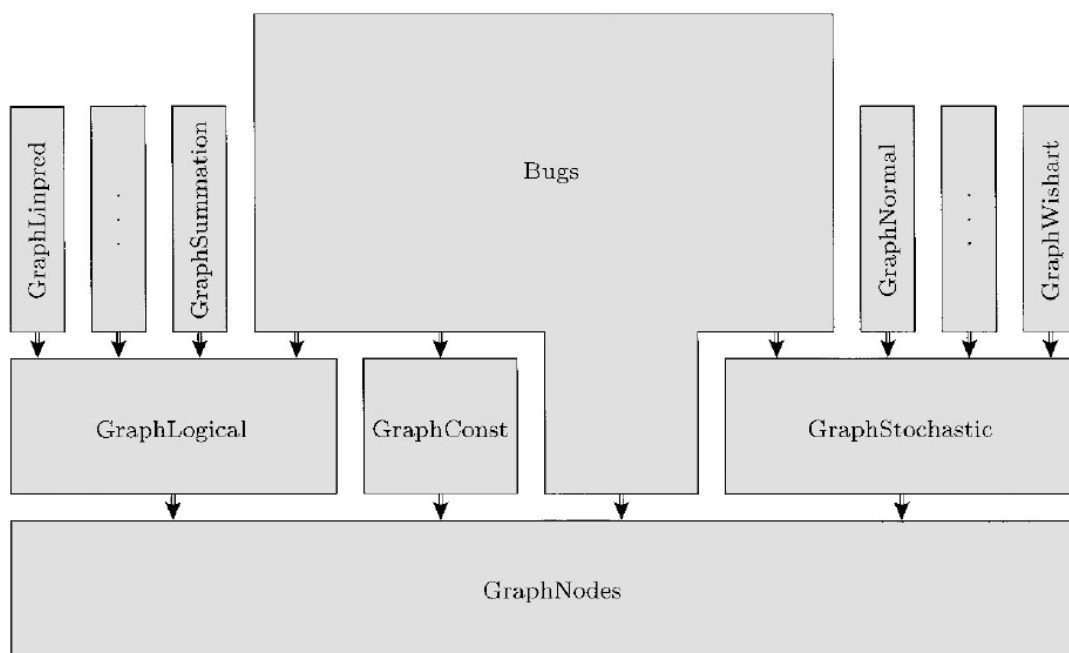


3 pav. WinBUGS programinės įrangos struktūra

3 pav. tamsesnė sritis rodo sąveiką tarp šešių pagrindinių posistemų (t.y. Graph, Updater, Monitors, Bugs, Samples, ir Doodle). Kita sritis parodo, kaip pritaikoma specializuota programa, t.y. PKBugs. Pavaizduotos formos ir posistemų dydžiai yra beprasmiški, tai tiesiog yra hierarchinė struktūra, kuri yra svarbi. Vaizduojamos rodyklės rodo vienos krypties ryšį tarp posistemų. Jei vienas posistemis nurodomas į kitą, tai naudojama pastarojo forma. Be to pastarasis posistemis neturi informacijos apie buvusio egzistavimą.

Grafo posistemės ir konstrukcija

Pagal susitarimą prie modulio pavadinimo gali būti pridedamos posistemės priešdėlis, kuriai modulis priklauso. Pavyzdžiui visi grafo posistemio moduliai gali prasidėti su „Graph“, t.y. „GraphNode“, „GraphStochastic“. 4 pav. pavaizduota grafo posistemės modulio sudarymas ir kaip Bugs posistemė tuo naudojasi.



4 pav. Graph posistemio modulio sudarymas

Grafo posistemė nustato hierarchijos tipą, kad sudarytų mazgus DAG grafuose. Pagrindinis tipo Mazgas apibrėžtas modulyje „GraphNodes“ ir vadinamas „GraphNodes“. Pagal susitarimą „GraphLogical.Node“, „GraphConst.Node“, ir „GraphStochastic.Node“ yra išplėsti, ir atitinkamai nurodo logiškus, pastovius ir stochastinius mazgus. Pastovūs mazgai nėra išplėsti. Logiški mazgai yra išplėsti į tam tikras funkcijų klases, tokias kaip tiesinė prognozė („GraphLinpred“) ir mazgų sumos („GraphSummation“). Jie padeda modelio specifikacijai. Tikimybiniai mazgai gali būti vienmačiai arba daugiamačiai (pavyzdžiui „Wishart“), ir tai yra „GraphStochastic“. Mazgas yra išplėstas į „GraphStochastic.Univariate“ ir „GraphStochastic.Multivariate“. Šie yra išplėsti toliau konkrečių tipų, daugelyje konkrečių pasiskirstymo modelių, pavyzdžiui „GraphNormal“ ir „GraphWishart“.

Grafinio modelio vaizdavimas, yra konstruojamas iš pradžių įvedus duomenų struktūros „vardą“, kuris apibrėžtas modulyje „BugsNames“. Kiekvienam kintamajam statistiniame modelyje yra „BugsNames“ tipo kintamasis. „Vardas“ naudojamas tam, kad būtų išsaugota vartotojo sukaupta informacija, ir mazgo objektų rinkinys, kuris tam naudojamas. Tai leidžia sudaryti žymėjimus tarp tekstinio ir vidinio modelio vaizdavimo, kad objektai, atitinkantys tam tikrus mazgus grafe, būtų lengvai pažymimi. Modulis „BugsIndex“ saugo „BugsNames“ sąrašą. Vardiniai kintamieji naudojami modeliui pavaizduoti.

Norint sukurti objektų mazgą naudojami faktoriniai objektai. Modulis „GraphNodes“ eksportuoja abstraktų tipą, pavadintą „Factory“, kuris turi vieną metodą, t.y. „New“. Šio tipo

kintamasis yra taip pat eksportuojamas. Kiekvienas konkretus pasiskirstymas ir konkretus funkcijos Grafo posistemės modulis, eksportuoja kintamąjį, kuris išplečia „GraphNodes.Factory“. Šio objekto „New“ metodas sukuria ir parodo (dinamiškai), tinkamo konkretaus pasiskirstymo ar konkrečios funkcijos mazgo tipą. Taigi mazgo objektas gali būti sukurtas tiesiog per „New“ metodo atitinkamą modulio faktoriinį objektą.

Bugs faktorinius objektus naudoja taip: pirmiausiai modelio reikalavimai išnagrinėjami gramatiškai ir sukuriamas „medžio“ vaizdavimas. Bugs posistemė pateikia mokymų failą, kuris apibrėžia modelio kalbos reikalavimus. Tai yra tiesiog taisomas tekstinis failas, kuriame pateikiami kiekvieno pasiskirstymo ir funkcijos vardai (žymėjimai). Pavyzdžiui „dnorm (s, s)“ reiškia normalųjį pasiskirstymą, kuris apibrėžiamas, naudojant du skaliaro (s) parametrus. Šalia kiekvieno identifikatoriaus yra procedūros paklausimas su sintakse „Module“. „Module“ yra modulio vardas, kuriame yra apibrėžtas tinkamas konkretus mazgo tipas (pavyzdžiui „GraphNormal“). Taip pat jame diegiama paprasta procedūra, kuri nustato faktoriinį objektą „GraphNodes“, lygų faktoriiniui objektui „Module“. Taigi, kai „New“ metodas vadinamas tinkamu, mazgo objektas grąžinamas dinamiškai ir todėl Bugs nereikalauja jokios informacijos apie pasiskirstymo specifikas ar konkretų funkcijos modulį. Mazgo objektai sukuriami tiesiog „paimant“ ir atliekant komandas priemonės faile ir toliau naudojant faktoriinį objektą eksportuoja į „GraphNodes“.

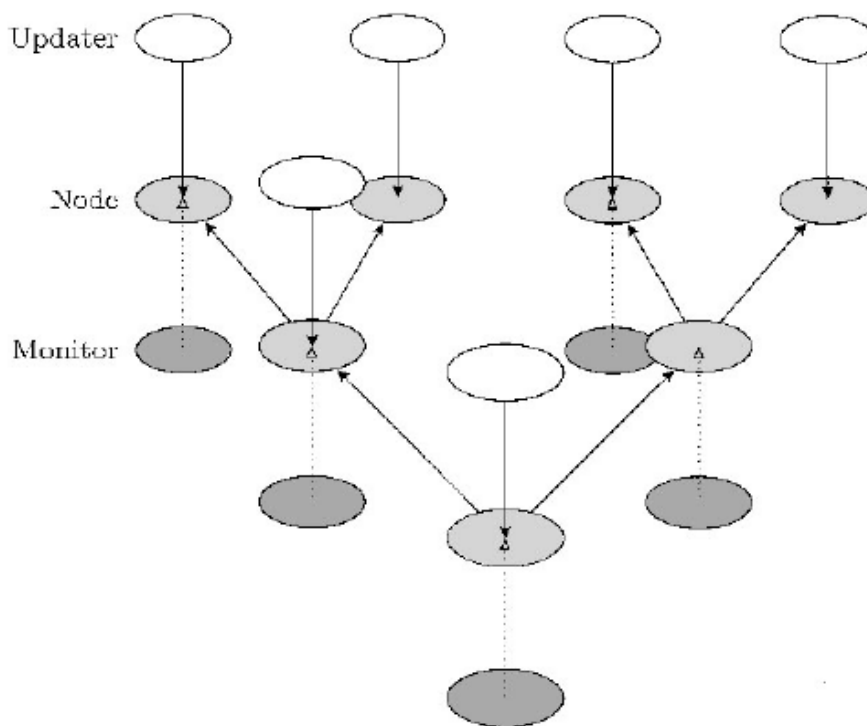
Kai tik visi mazgo objektai sukurti, įtraukiami atitinkamai į „BugsNames“. Vardiniai kintamieji saugomi „BugsIndex“, todėl grafe reikia apibrėžti nukreiptas sąsajas tarp mazgų. Tai atliekama kiekvienam mazgo objektui, kuriame įtraukiami nurodymai savo tiesioginiams aukštesnės klasės elementams. Tiesioginiai aukštesnių klasių elementų vardai nustatomi remiantis „medžio“ modelio pavaizdavimu, sudaryto pagal sintaksės analizatorių. Jie naudojami siekiant surasti aukštesnės klasės elementų objektus, kurie perduoda atitinkamą mazgą per nustatytą metodą.

Metodai

Pagrindinis bet kokios MCMC programinės įrangos tikslas turėtų būti, atrankos metodų panaudojimas. Tai yra efektyvu ir patikima. WinBUGS, metodai, pridėti prie mazgo objektų, leidžia programinei įrangai pasirinkti ir taikyti tinkamiausią atrankos schemą, kiekvieno sąlyginio pasiskirstymo. Jie gali būti skirstomi:

- Topologijos – metodai, kurie apibrėžia ar informuoja apie grafo struktūrą;
- Klasifikacijos – metodai, kurie nukreipia grafą ir pateikia informaciją apie tai, kaip stochastinio mazgo sąlyginis pasiskirstymas galėtų būti klasifikuojamas;
- Atnaujinimo – metodai, kurie įvertina svarbias funkcijas atnaujintų objektų, t.y. objektai, kurie gauna pavyzdžius iš pilno sąlyginio pasiskirstymo.

Atnaujinimo objektai, kurių pagrindinis tipas apibrėžiamas modulyje „UpdaterUpdaters“ yra atsakingi už MCMC modeliavimą. Jie atnauja grafą, apskaičiuodami naujas vertes mazgams. Yra atnaujinimų konkrečiam pasiskirstymui, t.y. kai mazgo sąlyga išreikšta uždara forma. Konkrečių objektų atnaujinimo sudarymas analogiškas mazgo objektų sudarymui. Kai tik tinkamo mazgo pilnas sąlyginis pasiskirstymas klasifikuojamas, įdiegta procedūra vykdoma iš priemonių failo. Taip nustatomas faktorinis objektas modulyje „UpdaterUpdaters“ yra lygus atitinkamam konkrečiam faktorinio objekto atnaujinimui. Ankstesnis gali būti naudojamas, kad sukurtų konkrečius atnaujinimo objektus neturint informacijos apie modulius, kuriuose jie yra apibrėžti.



5 pav. 3-D grafo pavaizdavimas

5 pav. iliustruoja 3-D grafą (įskaitant mazgo objektus), susijusius atnaujinimus ir kontrolės objektus. Yra trys sluoksniai: atnaujinimai (balta) sudaro viršutinį sluoksnį, mazgai (šviesiai pilka) formuoja antrą ir trečias sluoksnį sudaro kontrolės (tamsiai pilka). Rodyklės žymi kintamuosius, kurie sujungia objektus.

Kiekvienas atnaujinimas objekto rodykle į mazgą, kurį jis atnauja. Per šį atnaujinimą galima atitinkama lokali informacijos apie grafą. Norint gauti šį sujungimą, atnaujinimas turi metodą, vadinamą „Init“. Šis metodas gauna atitinkamą mazgo objektą kaip argumentą, o tai analogiška mazgų „Set“ metodui.

Svarbiausia procedūra, jungianti atnaujinimo objektus yra MCMC metodas. Tai sudaro atsitiktinę imtį iš mazgo pilno sąlyginio pasiskirstymo ir atitinkamai atnauja mazgo vertės sritį.

Bendriems atnaujinimams MCMC metodas yra paprastai apibrėžtas, nes nėra reikalingos jokios ypatingos aplinkybės. Tačiau, kai pilnas sąlyginis pasiskirstymas išreikštas uždara forma, tos formos parametrai, kažkoku būdu turi būti gauti. Tai atliekama, naudojant „LikelihoodForm“ metodą, kiekvienam iš pridodamo mazgo.

Kontrolės objektai (kurių pagrindiniai tipai yra apibrėžti Kontrolės posistemėje) yra taip pat sujungti į mazgus rodyklių. Kiekvienos Gibbs atrankos iteracijos pabaigoje, kiekvienas kontrolės objektas prijungiamas prie mazgo. Kadangi atnaujinimai yra automatiškai sukurti Bugs posistemės kiekvienam stochastiniam mazgui grafe, kontrolė yra suskirstoma į stochastinius ar loginius mazgus, per vykdymo laiką.

WinBUGS apibendrinimas

Šiuolaikinės skaičiavimo sąvokos ypač naudingos, kad greitai ir efektyviai išspręstų sudėtingas problemas. Objekto orientavimas yra pagrindas bendrų DAGs savybių, kad programinė įranga galėtų efektyviai nagrinėti pasirinktinai sudėtingus modelius. Programinės įrangos projektavimas orientuotas į komponentą, padidina jo patikimumą ir tampa lengvai ištesiamas. Kad moduliai gali būti įkelti pagal reikalavimą (t.y. vykdymo metu) reiškia tai, kad naujos funkcijos gali būti pridėtos ir nereikia bet kurios programinės įrangos dalies perkompiliuoti iš naujo.

Nors WinBUGS buvo suprojektuotas pirmiausiai DAGs grafams, kitokio tipo grafiniai modeliai, kaip sąlyginis autoregresinis modelis, taip pat yra leistini. Tai yra šiek tiek mažiau patogesnis ypatybių faktorizavimas, bet ganėtinai lengvai naudojamas.

Šiuo metu WinBUGS gali atlikti tik daugiamačių atnaujinimą, jeigu galima gauti tinkamą sąlyginį pasiskirstymą uždaroje formoje, pavyzdžiui Wishart. Daugiamatis atnaujinimas gali labai pagerinti bendrą MCMC modeliavimo efektyvumą ir būsima programinės įrangos versija yra skirta bendrų daugiamačių duomenų atrankos metodų naudojimui.

WinBUGS programinė įranga prieinama per <http://www.mrc-bsu.cam.ac.uk/bugs/> puslapį. Gali būti naudojama analizuoti daugybei problemų pavyzdžių ir bendrų problemų apriboto dydžio įvertinimo tikslais. Visa dokumentacija ir pavyzdžiai yra supakuoti kaip programinės įrangos dalis. Žemiau pateikta 1 lentelė rodo pasiskirstymą (kaip 1999 m. liepą) per 2000 registruotų vartotojų visame pasaulyje, pagal geografinę vietą, naudojimo tipą, ir prisijungimą. Programinė įranga yra pirmiausiai naudojama anglų kalba kalbančiose šalyse universitetų įstaigose mokomaisiais tikslais, nors daugiau kaip 40 % viso šios įrangos vartojimo yra realaus gyvenimo paraiškos

1 lentelė. Registruotų vartotojų pasiskirstymas

Location	Users	Type of use	Users	Affiliation	Users
North America	1064	Educational	1303	University	1436
Britain	336	Medical	616	Not-for-profit	279
Rest of Europe	481	Environmental	143	Company	277
Australia & New Zealand	119	Industrial	103	Personal	138
Rest of World	224	Financial	59	Other	94

3 PRIEDAS. PROGRAMINĖS ĮRANGOS WINBUGS KODAS

WinBUGS kodas Leukemijos modeliavimo atveju:

```

model
{
  for(i in 1:N) { #N nurodo pacientų skaičių
    for(j in 1:T) { # T remisijos momentų skaičius
#vertinama rizika, jei kyla stebimu momentu Y[N,T]=1, obs.t>=t ir priešingu atveju 0
      Y[i, j] <- step(obs.t[i] - t[j] + eps)

#eps=0.000001
#skaičiavimo procesas pereina per vieną vienetą kiekvienos mirties momentu, obs.t įvyko per laiko
intervalą [t[j],t[j+1]], jei t[j]<=obs.t<t[j+1]
#fail[i] apibrėžia - remisijos momentas=1, cenzūravimas=0
      dN[i, j] <- Y[i, j] * step(t[j + 1] - obs.t[i] - eps) * fail[i]
    }
  }
#Modelis
  for(j in 1:T) {
    for(i in 1:N) {
      dN[i, j] ~ dpois(Idt[i, j])# Tikėtinumo funkcija
      Idt[i, j] <- Y[i, j] * exp(beta * Z[i]) * dL0[j] # Intensyvumo
funkcija
    }
    #dL0[j] ~ dgamma(mu[j], c) # Gama proceso atveju
    #mu[j] <- dL0.star[j] * c # pavojingumo funkcija

# Išgyvenamumo funkcija 6-mp preparatu gydytai pacientu grupei
    S.treat[j] <- pow(exp(-sum(dL0[1 : j])), exp(beta * -0.5));
# Išgyvenamumo funkcija placebo gydytai pacientu grupei
    S.placebo[j] <- pow(exp(-sum(dL0[1 : j])), exp(beta * 0.5));

  }
  for(j in 1:T) {
# proporcingo pavojingumo modelis, kai bazinio pavojingumo vertinimui naudojama daugiakampio
funkcija
    tau[j+1] <- tau[j] * exp(e[j+1])
    e[j+1] ~ dnorm(0, tau.e)
    dL0[j] <- interp.lin(t[j],a[j],tau[j]) * step(a[j+1] - t[j] +
eps) * step(t[j] - a[j] -eps) + tau[j] * step(a[T+1] - t[j] + eps)
  }
  #c <- 0.001
  #r <- 0.1
  #for (j in 1 : T) { dL0.star[j] <- r * (t[j + 1] - t[j]) }

  tau[1] ~dgamma(0.01,0.01)
  beta ~ dnorm(0.0,0.000001) # apriorinis beta pasiskirstymas
  tau.e ~ dlnorm(0,0.01)
  sigma <- 1/sqrt(tau.e) # atsitiktinės sekos generavimas

```


WinBUGS kodas Gerklų vėžio modeliavimo atveju:

```

model
{
  for(i in 1:N) { #N nurodo pacientų skaičių
    for(j in 1:T) { # T mirties/tyrimo baigties momentų skaičius

#vertinama rizika, jei kyla stebimu momentu Y[N,T]=1, obs.t>=t ir priešingu atveju 0
      Y[i, j] <- step(obs.t[i] - t[j] + eps)

#eps=0.000001
#skaičiavimo procesas pereina per vieną vienetą kiekvienos mirties momentu, obs.t įvyko per laiko
intervalą [t[j],t[j+1]], jei t[j]<=obs.t<t[j+1]
#fail[i] apibrėžia - mirties/tyrimo baigties momentas=1, cenzūravimas=0
      dN[i, j] <- Y[i, j] * step(t[j + 1] - obs.t[i] - eps) *
fail[i]
    }
  }
#Modelis
  for(j in 1:T) {
    for(i in 1:N) {
      dN[i, j] ~ dpois(Idt[i, j]) # Tikėtinumo funkcija
      Idt[i, j] <- Y[i, j] * exp(beta.dis[stage[i]] * Z[i]) * dL0[j] #
Intensyvumo funkcija
    }
    #dL0[j] ~ dgamma(mu[j], c) # Gama proceso atveju
    #mu[j] <- dL0.star[j] * c # pavojingumo funkcija

# Gerklų vėžio 1 stadijos išgyvenamumo funkcija
    S.stage1[j] <- pow(exp(-sum(dL0[1 : j])), exp(beta.dis[1] *0));

#2 stadijos išgyvenamumo funkcija
    S.stage2[j] <- pow(exp(-sum(dL0[1 : j])), exp(beta.dis[2] *
1));

#3 stadijos išgyvenamumo funkcija
    S.stage3[j] <- pow(exp(-sum(dL0[1 : j])), exp(beta.dis[3] *2));

#4 stadijos išgyvenamumo funkcija
    S.stage4[j] <- pow(exp(-sum(dL0[1 : j])), exp(beta.dis[4] *
3));
  }
  for(j in 1:T) {
# proporcingos pavojingumo modelis, kai bazinio pavojingumo vertinimui naudojama
daugiakampio funkcija
    tau[j+1] <- tau[j] * exp(e[j+1])
    e[j+1] ~ dnorm(0, tau.e)
# naudojamos daugiakampio funkcijos viršūnės esančios ties momentais a
    dL0[j] <- interp.lin(t[j],a[j],tau[j]) * step(a[j+1] - t[j] +
eps) * step(t[j] - a[j] -eps) + tau[j] * step(a[T+1] - t[j] + eps)
  }
}

```

