Liet. matem. rink., T. 43, spec. nr., 565–570

Informacijos šaltinių apie Lietuvos vaikų įgimtų anomalijų paplitimą palyginamoji analizė

Jurgita ŽIDANAVIČIŪTĖ (VGTU), Marijus RADAVIČIUS,
Jurgis SUŠINSKAS (MII), Algirdas UTKUS (VU)
el. paštas: zjurga@one.lt, mrad@ktl.mii.lt

1. Įvadas

Vienas iš Lietuvos Žmogaus genetikos centro (LŽGC), įkurto 1991 metais, uždavinių yra duomenų apie paveldimas ligas ir įgimtas raidos anomalijas (IRA) Lietuvoje kaupimas ir analizė, siekiant įvertinti jų dažnį, paplitimo tendencijas bei galimas priežastis. LŽGC nuo 1993 metų pradėjo pildyti IRA registrą (jo tarptautinis akronimas LIRECA), kuriam duomenis apie visus IRA Lietuvoje atvejus pateikia gimdymo stacionarai, gydytojai akušeriai ginekologai, neonatologai ir LŽGC gydytojai genetikai. Per metus užregistruojama apie 500–600 paveldimų ligų ir IRA atvejų. Duomenų bazė (DB) LIRECA yra pastoviai palaikoma ir atnaujinama, tačiau pastaraisiais metais užregistruotų IRA atvejų skaičius ženkliai sumažejo, ypatingai po 1997 m. Kadangi realiai IRA dažniai negali staigiai keistis, galima daryti prielaidą, kad DB LIRECA užregistruojami ne visi atvejai. Juolab, kad nei šios DB palaikymui, nei duomenų rinkimui valstybė lėšų neskiria.

Iškyla aktualus uždavinys surasti būdus, kaip įvertinti duomenų iškraipymo mastą ir realius ĮRA dažnius. Tam tikslui šiame darbe buvo pasinaudota kita duomenų baze, surinkta vieno iš autorių, kurioje sukaupti jau mirusių vaikų su ĮRA autopsijos (skrodimo) duomenys, apimantys laikotarpį nuo 1982 iki 1991 metų. Ją toliau vadinsime PA (patoanatomine) duomenų baze. Kadangi skrodimas yra standartinė reglamentuota procedūra, tai galima tikėtis, kad autopsijos duomenys, jeigu ir nepilnai atspindi realią padėtį, tai šis neatitikimas nėra susijęs nei su laiku, nei su vieta, o yra susijęs tik su įgimto defekto pasekmių rimtumu (mirtimi) ir diagnozės (vizualaus) nustatymo lengvumu. Todėl tų duomenų pagrindu galima nustatyti tokio tipo įgimtų anomalijų paplitimą respublikoje ir jų dažnių kitimo tendencijas iki 1992 metų bei apskaičiuoti prognozę tolimesniems metams. Gautos prognozės ir DB LIRECA duomenų 1992–1997 m. sugretinimas leidžia įvertinti jų registravimo bazėje ypatumus ir parinkti juos aprašantį modelį. Palyginę pastarųjų metų surinktus duomenimis su prognoze, gauta sudaryto modelio pagrindu, galėtume atsakyti į klausimą, ar nuogąstavimai apie pablogėjusią registravimo kokybę bazėje LIRECA yra pagrįsti ir su kokiais faktoriais tai gali būti susiję.

Atliktas tyrimas remiasi apibendrintais tiesiniais modeliais: Puasono bei logistine (binomine) regresija.

2. Apibendrintas tiesinis modelis

Apibendrinti tiesiniai modeliai (generalized linear model [1]) praplečia tradicinę tiesinių modelių klasę ir dėl to pritaikomi platesnei duomenų analizės sričiai. Jie yra labai patogūs sudėtingoms statistinėms hipotezėms apie kokybinių požymių tarpusavio sąryšius tikrinti. Jais galima aprašyti tiriamojo kintamojo y vidurkio kitimo priklausomybę nuo įtakojančių veiksnių, naudojant ir netiesinę sąryšio funkciją, o kintamojo y stebėjimų skirstinį modeliuoti skirstiniais iš eksponentinių skirstinių šeimos.

Apibendrintą tiesinį modelį nusako:

- Stebėjimai y_i, $i = 1, 2, \ldots, n$, yra nepriklausomi atsitiktiniai dydžiai, kurių tikimybiniai skirstiniai priklauso apibendrintai vienparametrinei eksponentinių skirstinių šeimai, t.y. jų pasiskirstymo tankiai turi tokį pavidalą:

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right\}. \tag{1}$$

  Čia θ yra skaliarinis eksponentinių skirstinių šeimos parametras, ϕ yra papildomai įvestas (maišantis) mastelio (scale) parametras, suteikiantis papildomas galimybes priderinti modelio dispersiją prie realių duomenų, o $a(\varphi)$, $b(\theta)$, $c(y, \varphi)$ yra funkcijos, nusakančios konkretų skirstinį iš eksponentinių skirstinių šeimos.

- Tiesinė komponentė $\eta_i = x_i'\beta$, kur x_i yra prediktoriai (aiškinantieji kintamieji).

- Sąryšio (link) funkcija g susieja atsitiktinių dydžių y_i vidurkius μ_i su tiesinėm komponentėm η_i lygtimi $g(\mu_i) = x_i'\beta$.

Pasirinkdami konkretų sąryšio funkcijos g ir tikimybinio tankio $f(y)$ variantą, gauname apibendrinto tiesinio modelio atskirus atvejus.

- Tradicinis tiesinis modelis: $g(\mu) = \mu$.
- Logistinė regresija: skirstinys yra binominis $B(p, k)$, $g(\mu) = \operatorname{logit}(\mu/\mathrm{k}) = \log(\mu/(\mathrm{k} - \mu))$.
- Puasono regresija: skirstinys yra Puasono $g(\mu) = \log(\mu)$.

Logistinė (binominė) regresija ir Puasono regresija priklauso logaritminių tiesinių (log-linear) modelių grupei [3, 4, 8, 9].

3. Tyrimo rezultatai

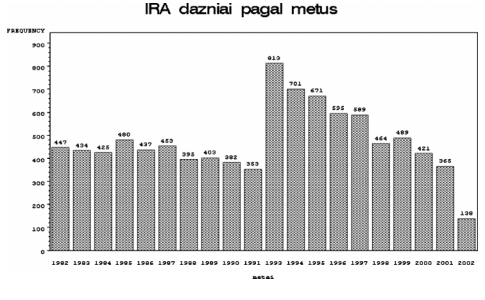
Tyrimui buvo naudojamos dvi ĮRA DB: LIRECA ir pato-anatominė. Kiekviena jų turi savo sudarymo ir duomenų registravimo ypatumus. Būtent jais galima būtų paaiškinti, kodėl net ir artimais, tik pora metų besiskiriančiais laikotarpiais (PA baigiasi 1991 m., nuo 1993 m. prasideda LIRECA) užregistruotų ĮRA proporcijos skirtingose bazėse labai skiriasi (žr. 1 pav.). PA duomenų bazėje užregistruoti tik vaikai (iki 14 metų). Vadinasi, vaikai, kurių ĮRA nėra rimtas gyvybinių sistemų pažeidimas ir ji nėra jų mirties priežastis, į šią bazę pakliūna tik atsitiktinai. Tuo tarpu bazėje LIRECA pagal jos sudarymo

taisykles registruojami visi ĮRA atvejai – visi naujagimiai, kuriems nustatyta kokia nors ĮRA forma. Kiekvienu tokiu atveju užpildoma detali anketa su duomenimis apie naujagimio motiną, tėvą, brolius ir seseris, apie jo vystymosi ir gimimo aplinką. Šiuo požiūriu PA DB yra daug skurdesnė už LIRECA. Atrinkus bendrus kintamuosius abiems DB, jos buvo sujungtos į vieną bendrą duomenų bazę, kuri ir buvo naudojama šiame tyrime.

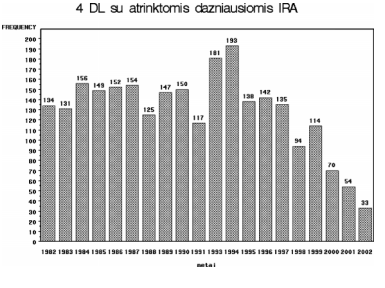
Kaip matosi iš 1 pav., PA DB ĮRA dažniai kinta pakankamai iš lėto, ko negalime pasakyti apie DB LIRECA. Joje, ypač nuo 1997 m., matomas ryškus ĮRA dažnių sumažėjimas. Darant prielaidą, kad ĮRA dažniai yra toks procesas, kuris staigiai kisti negali, galime iškelti hipotezę, kad dalis naujagimių ĮRA atvejų DB LIRECA neužregistruojami. Tiesa, ĮRA dažnių sumažėjimą paskutiniaisiais metais taip pat galime sieti ir su gimstamumo mažėjimu Lietuvoje, todėl parenkant logaritminį tiesinį modelį buvo į tai atsižvelgta, panaudojant informaciją apie naujagimių skaičių ir modeliuojant ne ĮRA dažnių vidurkį, o jų proporciją tarp naujagimių (santykinį vidurkį).

Kaip jau buvo užsiminta, abi bazės skiriasi dar ir diagnozavimo tikslumo prasme. Aišku, kad skrodimo metu ĮRA diagnozuoti daug paprasčiau, ypač, jeigu ji yra (vienintelė) mirties priežastis. Vadinasi, PA bazėje lyginant su LIRECA turėtų būti santykinai daugiau sunkiai diagnozuojamų ĮRA atvejų. Iš kitos pusės, bazėje LIRECA santykinai daugiau turėtų būti betarpiškai gyvybei nepavojingų ĮRA atvejų. Todėl, norint palyginti abi DB ir nustatyti DB LIRECA adekvatumo realiai padėčiai intervalą, buvo išskirtos tos ĮRA rūšys, kurios, nesant registravimo pablogėjimui, vienoje ir kitoje bazėje turėtų turėti tas pačias dažnių kitimo tendencijas. Tokiomis ĮRA visų pirma galėtų būti tos, kurios yra lengvai ir patikimai diagnozuojamos. Remiantis ekspertine nuomone bei vaikų mirtingumo dėl ĮRA analize kaip lengvai ir patikimai diagnozuojamos buvo išskirtos tokių tipų ĮRA: nervinio vamzdelio, spina bifida, anencephaly, virškinimo, Dauno. Toliau nagrinėsime tik šio tipo ĮRA. Jų sumarinių dažnių kitimo grafikas pateiktas 2 pav.

Iš jo matyti, kad tarp abiejų DB yra lyg ir neblogas atitikimas, tačiau po 1997 metų DB LIRECA registruotų ĮRA dažniai staigiai mažėja. Ar šis ĮRA dažnių sumažėjimas yra statistiškai reikšmingas buvo patikrinta, parenkant binominės (logistinės) regresijos modelį. Naudojome tokius kintamuosius: y, metai m, matuojami atžvilgiu bazinių metų 1992 ($m =$ metai $- 1992$), gimusių m-aisiais metais naujagimių skaičius g_m (duomenys paimti iš [5–7]), yra duomenų bazės tipas b, $b = 0$ bazei PA ir $b = 1$ bazei LIRECA, bei



1 pav. ĮRA dažniai duomenų bazėse PA ir LIRECA.



2 pav. Lengvai ir patikimai diagnozuojamų ĮRA dažniai autopsijų DB ir DB LIRECA.

iš jų išvestinius naujus dydžius: laikotarpio po 1997 metų indikatorių $\mathbf{1}\{m > 5\}$, tiesinio trendo po 1992 ir po 1997 metų kintamuosius m_+ ir $(m-5)_+$. Čia s_+ žymi skaičiaus s teigiamą dalį, $s_+ = \max(s, 0)$.

Testuojamas binominės regresijos modelis atrodė taip:

$$\operatorname{logit}(\mu_{\mathrm{m}}/\mathrm{g_m}) = \beta_0 + \beta_1 \mathrm{m} + \beta_2 \mathrm{b} + \beta_3 \mathrm{m}_+ + \beta_4 1\{\mathrm{m} > 5\} + \beta_5 (\mathrm{m}-5)_+. \qquad (2)$$

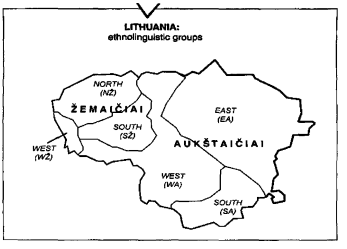
Naudojant statistinio paketo SAS procedūrą GENMOD [9], buvo gauti modelio (2) nežinomų parametrų didžiausio tikėtinumo įverčiai ir patikrintos hipotezės apie kiekvieno iš parametrų statistinį reikšmingumą. Išmetę iš modelio (2) statistiškai nereikšmingus prediktorius ir įvertinus likusiųjų prediktorių parametrus gavome galutinį modelį:

$$\operatorname{logit}(\mu_{\mathrm{m}}/\mathrm{g_m}) = -6,0056 + 0,4311\mathrm{b} - 0,2105\beta_5(\mathrm{m}-5)_+. \qquad (3)$$

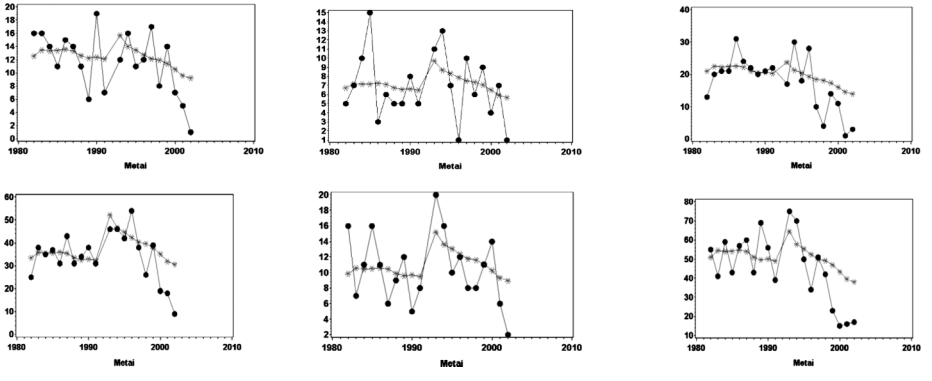
Visiems šio modelio parametrams nulinė hipotezė atmetama su reikšmingumo lygmeniu 0,001 (visos p-reikšmės mažesnės už 0,001). Parinkto apibendrinto tiesinio modelio adekvatumas nustatomas pagal nuokrypio (deviance) reikšmę. Gerai parinkto modelio atveju nuokrypis turėtų būti artimas vienetui [1,9]. Modelio (3) nuokrypis lygus 1,6196, kas, matyt, rodo nelabai gerą modelio ir duomenų atitikimą. Tiksliai į klausimą apie modelio adekvatumą (goodness-of-fit) atsako tikėtinumų santykio kriterijus *LR*, lyginantis modelį (3) su pilnu baziniu modeliu (saturated model). Šiuo atveju *LR* p-reikšmė mažesnė už 0,001 (suskaičiuota su SAS procedūra CATMOD [9]). Vadinasi, modelis netinkamas, santykinių dažnių nukrypimų nuo vidurkio negalima paaiškinti vien atsitiktinumu. Tai tipiškas reiškinys biologiniuose ir socialiniuose moksluose [1]. Tokiais atvejais rekomenduojama vietoje mastelio parametro φ standartinės reikšmės lygios 1 įstatyti jo įvertinį, apskaičiuojamą kaip kvadratinė šaknis iš nuokrypio ir laisvės laipsnių santykio. Atlikus atitinkamus perskaičiavimus, rezultatas iš esmės nepakito. Parinkto modelio (3) patikimai nenulinė koeficiento β_2 prie bazės požymio b reikšmė liudija, kad mums nepavyko išskirti ĮRA tipus, pagal kuriuos būtų visiškas atitikimas tarp duomenų bazėse AP ir LIRECA sukauptos informacijos. Tačiau gauti rezultatai patvirtino nuogąstavimus apie pablogėjusią registracijos kokybę pastaraisiais metais. Apie tai sprendžiame pagal paskutinįjį lygties (3) narį, kuris atspindi pastovią santykinių dažnių mažėjimo tendenciją po 1997 metų.

Norint detaliau ištirti šį reiškinį, natūralu į modelį įtraukti ir vietos faktorių. Bendroje (apjungtoje LIRECA ir AP) duomenų bazėje smulkiausias vietos vienetas yra rajonas. Kadangi rajonas yra kategorinis kintamasis, o Lietuvoje yra virš 40 rajonų, tai pridėjus šį požymį bendro modelio analizė darosi sudėtinga. Visų pirma požymių kryžminės dažnių lentelės turi daug nulių, dėl ko žymiai pablogėja chi-kvadrat ir kitų statistikų p-reikšmių aproksimacijų tikslumas. Kitas keblumas yra didelis parametrų skaičius. Dėl to bendras modelis su papildomu požymiu „rajonas“ yra praktiškai nesuskaičiuojamas.

Siekiant sumažinti parametrų skaičių modelyje, buvo tikrinamos įvairios hipotezės. Jos visos buvo atmestos. Pavyzdžiui, hipotezės, kad laikotarpyje iki 1997 metų lengvai ir patikimai diagnozuojamų ĮRA proporcijos pagal rajonus nepriklauso nuo bazės, p-reikšmė $p < 0,0001$. Todėl parametrų skaičių sumažinome apjungdami rajonus pagal



3 pav. Rajonų suskirstymas į etnografines grupes pagal Z. Zinkevičių [10].



4 pav. ĮRA dažnių prognozė vakarų, pietų ir šiaurės žemaičių bei vakarų, pietų ir rytų aukštaičių regionams atitinkamai.

etnografinį požymį (aukštaičiai – žemaičiai). Šio požymio pasirinkimas nėra visai atsitiktinis. Tarp Lietuvos gyventojų etnografinių grupių yra genetinių skirtumų [2], kas gali įtakoti ĮRA dažnius, ir be to etnografiniai regionai sudaro vientisas sritis (žr. 3 pav.). Regionams buvo patikrinta ta pati hipotezė kaip ir rajonams: ar laikotarpyje iki 1997 metų tiriamų ĮRA proporcijos pagal rajonus priklauso nuo bazės. Kadangi šį kartą hipotezė nebuvo atmesta ($p = 0,2374$), duomenų iki 1997 metų pagrindu galima sudaryti bendrą abiem bazėm modelį ir juo remiantis apskaičiuoti tiriamų ĮRA dažnių prognozę tolimesniems metams.

Kadangi neturėjome duomenų apie naujagimių skaičių pagal regionus (rajonus) iki 1992 metų, modelio sudarymui vietoje binominės regresijos teko naudoti Puasono regresiją. Rėmėmės Puasono teorema apie binominio skirstinio (su maža įvykio tikimybe) aproksimaciją Puasono skirstiniu bei prielaida, kad gimstamumo tendencijos visuose regionuose iš esmės sutampa. Atrinkus statistiškai reikšmingus kintamuosius ($p < 0,001$) ir įvertinus jų koeficientus gavome, kad

$$\mathrm{Log}\,(\mu_m/g_m) = \beta_0 + 0,45b + 1,4ea + 0,48nz - 0,17sa - 0,57sz + 1,06wa. \qquad (4)$$

Kintamieji ea, nz, sa, sz, wa yra indikatoriniai kintamieji atitinkantys regionus (etnografines grupes). Kintamasis $ea = 1$ rytų aukštaičiams ir $ea = 0$ kitais atvejais. Likusieji kintamieji apibrėžiami analogiškai tik atitinkamai šiaurės žemaičiams (nz), pietų aukštaičiams (sa), pietų žemaičiams (sz) ir vakarų aukštaičiams (wa).

Duomenų bazėje LIRECA užregistruotų ĮRA ir prognozuojamus modeliu (6) dažnius kiekvienoje etnografinėje grupėje galima palyginti 4 pav. Žiūrint į šiuos grafikus, matome, kad anksčiausiai, jau nuo 1997 m., registravimo blogėjimas prasidėjo rytų aukštaičių ir šiaurės žemaičių regionuose. Pietų ir vakarų aukštaičių regionuose ĮRA dažniai nukrypsta nuo bendros tendencijos tik nuo 2000 m. Vakarų žemaičių regione registravimo kokybės pablogėjimas atsiranda nuo 1999 m., o pietų žemaičių regione informacija buvo pakankamai gerai surenkama net iki 2001 m., išsiskiria tik paskutinieji 2002 m.

Standartizuotų liekanų analizė (nepateikta dėl vietos stokos) rodo, kad abejoti modelio adekvatumu nėra pagrindo.

Literatūra

[1] R. Christensen, *Log-Linear Models*, Springer, New York (1990).
[2] V. Kučinskas, Population genetics of Lithuanians, *Annals of Human Biology*, **28**, 1–14 (2001).
[3] *Log-Linear Analysis of Frequency Tables* [interaktyvus]. Electronic textbook, StatSoft. `www.statsoftnic.com/textbook/stloglin.html` (žiūrėta 2003 05 03).
[4] *Log-Linear Analysis of Frequency Tables* [interaktyvus]. Svam Software, Statistica. `www.svamsoftware.com/newsvam2/products/statistica/statadvance/stradvcomb2.html` (žiūrėta 2003 05 06).
[5] *Lietuvos miestų socialinė ir ekonominė raida*. Statistikos rinkinys. Vilnius (1991).
[6] *Moterys ir šeima*. Statistikos rinkinys. Vilnius (1993).
[7] *Natūralus gyventojų judėjimas* [interaktyvus]. Lietuvos Statistikos departamentas. `www.std.lt` (žiūrėta 2003 04 10).
[8] T.J. Santer, D.E. Duffy, *The Statistical Analysis of Discrete Data*, Springer, New York (1989).
[9] *SAS/STAT User's Guide Software* [interaktyvus]. `http://jeff-lab.queensu.ca/stat/sas/sasman/sas/stat` (žiūrėta 2003 05 15).
[10] Z. Zinkevičius, *The History of the Lithuanian Language*, Vilnius (1998).

A comparative analysis of information sources about the prevalence of congenital anomalies of Lithuanian children

J. Židanavičiūtė, M. Radavičius, J. Sušinskas, A. Utkus

The work is based on data from two data bases containing information about the prevalence of congenital anomalies of Lithuanian children. The first one is collected from autopsy data by one of the authors and covers the period 1981–1991. The second , called LIRECA, is maintained since 1993 and is based on questionnares filled in for each case of congenital anomaly of new-born. It is supposed that through absence of financial support this data base do not lately reflect the real situation. The aim of the work is to check this hypothesis on the ground of the comparative analysis of the both data bases and to estimate the extent of the disagreement.