Liet. matem. rink., T. 41, spec. nr., 469–477

Lietuvos vaikų įgimtų raidos anomalijų statistinis tyrimas

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

1. Įvadas

Šis taikomasis darbas remiasi duomenimis, paimtais iš Lietuvos Žmogaus Genetikos Centro (LŽGC) palaikomos ir atnaujinamos duomenų bazės LIRECA. Šioje duomenų bazėje kaupiami duomenys apie naujagimių įgimtas raidos anomalijas (ĮRA) Lietuvoje nuo 1992 metų. Tyrimo tikslas – ištirti ĮRA-ų paplitimo dėsningumus ir dažnumo kitimo tendencijas bei jį įtakojančius veiksnius panaudojant logistinės regresijos (logistic regression) modelius.

Šie modeliai pastaruoju metu vis dažniau taikomi įvairiose srityse, tame tarpe ir genetikoje (žiūr., pavyzdžiui, Umbach and Weinberg (1997). Jie leidžia atlikti žymiai subtilesnę ir lankstesnę kokybinių požymių statistinę analizę negu tradiciniai metodai, kurie remiasi Pirsono χ^2 kriterijumi, specialiais (ranginiais) koreliacijos ir asociacijos matais bei, jeigu stebėjimų ir požymių reikšmių nėra perdaug, tiksliuoju Fišerio testu (exact Fisher test). Didesnio požymių skaičiaus išsamesnę statistinę analizę šiais metodais atlikti sudėtinga, o kartais tiesiog neįmanoma. Nežiūrint to, logistinės regresijos modeliai kol kas nėra itin populiarūs Lietuvoje statistinių metodų taikytojų tarpe. To priežastis, matyt, yra šių modelių sudėtingesnis matematinis aprašymas ir dalykinė interpretacija bei atitinkamos literatūros, ypač lietuvių kalba, trūkumas. Todėl pirmoje darbo dalyje pateikiamas logistinės regresijos modelio aprašymas, trumpai aptariant jo interpretaciją ir kai kuriuos jo realizacijos statistinės analizės sistemoje SAS aspektus.

Antroje dalyje pateikiami ĮRA-ų pasitaikymo dažnį įtakojančių veiksnių, nagrinėtų šiame darbe aprašymas ir statistinio tyrimo rezultatų aptarimas.

Deja, daugelio rodiklių, registruojamų duomenų bazėje LIRECA ĮRA-as turintiems naujagimiams, pasiskirstymas tarp ĮRA-ų neturinčių naujagimių nėra žinomas. Tai žymiai apriboja tyrimo galimybes bei suteikia jo rezultatams savitą interpretaciją. Vien duomenų baze LIRECA paremtas tyrimas gali atsakyti tik į klausimą, kokie požymiai tam tikrą ĮRA-ą turinčius naujagimius skiria nuo kitas ĮRA-as turinčių naujagimių. Todėl tyrime papildomai naudojami statistiniai duomenys apie kasmetinį (nuo 1993 iki 1997) naujagimių skaičių pagal Lietuvos rajonus. Tai leidžia daryti gana bendras preliminarias išvadas apie kai kuriuos ĮRA-ų paplitimo dėsningumus ir dažnumo kitimo tendencijas tarp visų Respublikos naujagimių.

Nuodugnesniam tyrimui reiktų duomenis sukauptus bazėje LIRECA papildyti duomenimis iš kitų dviejų LŽGC-e kaupiamų duomenų bazių: genetinio konsultavimo ir patologinių anatominių tyrimų. Bet tai – ateities uždavinys.

2. Logistinė regresija

Regresinius modelius nuo kitų daugiamatės statistikos metodų skiria tai, kad jie ne tik leidžia nustatyti dominantį kintamąjį įtakojančius veiksnius, bet ir jų pagalba prognozuoti to kintamojo reikšmes. Regresinėje analizėje paprastai tiek prognozuojamas kintamasis, tiek ir prognozuojantys kintamieji (toliau juos vadinsime *prediktoriais* arba *kovariantais* ("covariants")) būna kiekybiniai. Dichotominiams arba ranginiams dydžiams prognozuoti taikomi specialūs regresiniai modeliai, dažniausiai logistinė regresija arba probit-regresija (probit-regression). Pirmasis ypač populiarus, nes turi glaudų ryšį su optimalaus klasifikavimo taisykle Gauso skirstinių mišinių modeliuose.

Kadangi darbe tiriamas įvairių veiksnių ryšys su įgimtų anomalijų atsiradimu, tai mus dominantys kintamieji yra dichotominiai: ar naujagimis turi ar neturi konkrečią ĮRA-ą. Todėl jų statistinei analizei taikome logistinės regresijos modelį. Pateiksime jo trumpą aprašymą (išsamų temos išdėstymą galima rasti Agresti (1990), Christensen (1990) ir Santer& Duffy (1989); trumpoje apžvalgoje (Fienberg (2000)) aptariami pagrindiniai rezultatai, paskutiniai pasiekimai ir vystymosi tendencijos kokybinių požymių statistinėje analizėje).

Tegu Y žymi atsitiktinį dydį įgyjantį, dvi reikšmes: 0 ir 1. Jeigu $Y = 0$, tai, pavyzdžiui, reiškia, kad tiriamos ĮRA-os nėra, o $Y = 1$ reiškia, kad ji, deja, yra. Kai $Y = 1$ (kartais atvirkščiai, kai $Y = 0$), sakoma, kad įvyko *įvykis*, o priešingu atveju – *neįvykis*. Vienintelė tikimybinė charakteristika, kuri nusako tą įvykį, yra jo tikimybė $p = \mathbf{P}(Y = 1)$. Tegu $x \in \mathbf{R}^m$ yra prediktorių vektorius. Natūralu laikyti, kad prediktoriai įtakoja įvykio tikimybę, t.y., $p = p(x) = \mathbf{P}(Y = 1|x)$. Logistinės regresijos modelyje ši priklausomybė aprašoma logistine funkcija, priklausančia nuo nežinomo parametro $\beta \in \mathbf{R}^m$,

$$p(x) = p(x|\beta) = \frac{\exp\{\beta^T x\}}{1 + \exp\{\beta^T x\}}.$$

Atvirkštinė logistinei funkcija yra vadinama logit-funkcija. Ji susieja tikimybę $p(x)$ su prediktoriais tokiu būdu:

$$\operatorname{logit}\big(p(x)\big) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta^T x.$$

Dydis po logaritmo ženklu vadinamas šansu arba šansais ("odds"). Jis parodo, kiek kartų įvykio tikimybė yra didesnė (mažesnė) už neįvykio tikimybę.

Tegu $\{(x_j, Y_j),\ j = 1, \ldots, N\}$ yra tyrimo duomenys (dydžio N imtis). Nežinomam parametrui $\beta \in \mathbf{R}^m$ įvertinti taikomas didžiausio tikėtinumo (DT) metodas. Log-tikėtinumo funkcija turi tokį pavidalą

$$L(\beta) = \sum_{j=1}^{N} Y_j \beta^T x_j - \sum_{j=1}^{N} \ln\big(1 + \exp\{\beta^T x_j\}\big).$$

Didžiausio tikėtinumo įvertinys (DTĮ) $\hat{\beta}$ randamas iš lygties

$$L(\hat{\beta}) = \max_{\beta} L(\beta).$$

Jis apskaičiuojamas naudojant iteratyvias procedūras. Dažniausiai taikomi metodai yra: tradicinis Niutono, Fišerio šerdies (Fisher scoring) bei svertinis mažiausių kvadratų (weighted least squares). Šiuos metodus naudoja ir SAS procedūra PROC LOGISTIC (žiūr. SAS/STAT (1997)).

Kai DT įvertis $\hat{\beta}$ suskaičiuotas, įvykio sąlyginė tikimybė, kai prediktorius x įgyja reikšmę x_0, prognozuojama dydžiu

$$\hat{p}(x_0) = p(x_0|\hat{\beta}),$$

o pats įvykis tokiu būdu:

$$\hat{Y} = \begin{cases} 0, & \text{kai} \quad \hat{p}(x_0) \leqslant r, \\ 1, & \text{kai} \quad \hat{p}(x_0) > r, \end{cases} \tag{1}$$

kur r yra pasirinkta *kritinė reikšmė* (cut-point).

Parinkto modelio prognozavimo kokybei įvertinti skaičiuojami ranginės koreliacijos koeficientai: c (žiūr. formulę (2)), Somers'o D, Goodman'o-Kruskal'o γ ir Kendall'o τ, tarp stebėjimų Y_j, $j = 1, \ldots, N$, ir atitinkamų prognozuojamų įvykio tikimybių $\hat{p}(x_j)$, $j = 1, \ldots, N$, o taip pat kryžminio patikrinimo (cross-validation) metodu įvertinami įvykio ($Y = 1$) ir neįvykio ($Y = 0$) prognozės klaidų skaičiai ir kitos jais nusakomos charakteristikos: sensityvumas, specifiškumas, sąlyginės neteisingai prognozuoto įvykio ("false negatives") ir neįvykio ("false positives") tikimybės ir pan. Įvykio ir neįvykio prognozės klaidų skaičius priklauso nuo pasirinktos kritinės reikšmės r. Paprastai skaičius r parenkamas taip, kad minimizuotų bedrą klaidingų sprendimų skaičių (klaidingo klasifikavimo tikimybę), tačiau, kadangi neretai klaidų prognozuojant įvykį ir neįvykį „nuostoliai" gali labai skirtis, tai racionaliau yra r parinkti taip, kad jis minimizuotų riziką

$$R(r) = K_1\big(1 - \text{sen}\,(r)\big) + K_2\big(1 - \text{spe}\,(r)\big) = K_1 \text{fp}\,(r) + K_2 \text{fn}\,(r),$$

kur K_1 yra „nuostoliai" nenuspėjus įvykio, K_2 yra „nuostoliai" nenuspėjus neįvykio, o $\text{sen}\,(r)$, $\text{spe}\,(r)$, $\text{fp}\,(r)$ ir $\text{fn}(r)$ yra atitinkamai prognozavimo taisyklės (1) su kritine reikšme r įvertintas sensityvumas, specifiškumas, neteisingai prognozuoto įvykio ir neįvykio tikimybės. Grafiškai šį uždavinį galima išspręsti panaudojant ROC (Receiver Operating Characteristic) kreivę. ROC kreivė tai grafikas $\text{sen}\,(r)$ atžvilgiu $\text{fn}\,(r) = 1 - \text{spe}(r)$, $0 < r < 1$. Optimali r reikšmė r^* yra ta, kuriai taškas $(\text{sen}\,(r^*), \text{fn}\,(r^*))$ yra tiesės $u = (K_2/K_1)t + h$, einančios virš taškų (0, 0) ir (1, 1) ir liečiančios ROC kreivę lietimosi su ja taškas. ROC kreivė turi ir kitą interpretaciją. Jos ribojamas plotas yra lygus

ranginės koreliacijos koeficientui c, tad kuo jis didesnis (arčiau 1) tuo geresnė logistinio modelio prognozė. Koreliacijos koeficientas c apibrėžiamas lygybe

$$c = \frac{d + s/2}{v}, \tag{2}$$

kur d, s ir v yra atitinkamai suderintų (concordant), susietų (tied) ir visų galimų porų dvejetų skaičius (žiūr. SAS/STAT (1997)). Dvi poros $(Y_j, \hat{p}(x_j))$ ir $(Y_i, \hat{p}(x_i))$ vadinamos suderintomis, jeigu

$$z = \big(\hat{p}(x_j) - \hat{p}(x_i)\big)(Y_j - Y_i) > 0,$$

nesuderintomis (nonconcordant), jeigu $z < 0$, ir susietomis, jeigu $z = 0$.

Įvairioms hipotezėms apie modelio parametrus tikrinti naudojami tikėtinumo santykio ir χ^2 kriterijai. Pasikliautiniams intervalams sudaryti taikoma DTĮ asimptotinė teorija (taip vadinami Valdo (Wald) pasikliautiniai intervalai) bei tikėtinumo santykio monotoniškumas atžvilgiu nežinomų parametrų ir jo skirstinio χ^2 aproksimacija (taip vadinami „tikslūs" pasikliautiniai intervalai).

Aiškesnę negu patys parametrai $\beta = (\beta_1, \ldots, \beta_d)$ interpretaciją turi kita per juos išreiškiama charakteristika. Tai – *šansų santykis* (odds ratio). Logistinės regresijos modelyje jis nusakomas taip:

$$O_i = \exp\{\beta_i\}, \quad i = 1, \ldots, d.$$

Jis parodo, kiek kartų pasikeičia (padidėja arba sumažėja) šansai (t.y. įvykio ir neįvykio tikimybių santykis), kai i-ojo prediktoriaus $x_{(i)}$ reikšmė padidėja per vienetą. Jeigu įvykio tikimybė maža, o šiame tyrime būtent taip ir yra, tai šansų santykis apytiksliai lygus atitinkamų įvykių tikimybių santykiui. Tokiu būdu O_i apytiksliai parodo, kiek kartų pasikeičia (padidėja arba sumažėja) įvykio tikimybė, kai i-ojo prediktoriaus $x_{(i)}$ reikšmė padidėja per vienetą. Šansų santykio įvertis gaunamas tiesiogiai iš parametro β didžiausio tikėtinumo įverčio $\hat{\beta}$:

$$\hat{O}_i = \exp\{\hat{\beta}_i\}, \quad i = 1, \ldots, d.$$

Jeigu prediktorių reikšmės kartojasi daug kartų, tai patogiau naudoti logistinį modelį su pasikartojimais. Tuomet laikoma, kad Y_j turi binominį skirstinį su parametrais n_j ir $p(x_j)$,

$$Y_j \sim B\big(n_j, p(x_j)\big), \quad j = 1, \ldots, N,$$

kur x_j, $j = 1, \ldots, N$, yra visos prediktorių įgyjamos skirtingos reikšmės, o n_j yra reikšmės x_j pasikartojimų skaičius. Šį modelį toliau vadinsime *binominiu logistinės regresijos* arba *binominiu log-tiesiniu* modeliu. Antruoju pavadinimu norima pabrėžti tą faktą, kad šis modelis yra dažnai kokybinių požymių analizėje taikomo *log-tiesinio* modelio atskiras atvejis (Agresti (1990), Christensen (1990)). Beje, jis yra atskiras atvejis

ir kito pastaruoju metu gana populiaraus *apibendrintojo tiesinio* ("generalized linear") modelio (šį modelį reikia skirti nuo bendrojo tiesinio ("general linear") modelio) (žiūr. Christensen (1990)).

3. Įgimtų raidos anomalijų statistinio tyrimo rezultatai

Tyrimui buvo išskirtos tokios dažniau Lietuvoje pasitaikančių ĮRA-ų grupės (žiūr. Utkus (2000)):

- širdies-kraujagyslių,
- DVD (dauginiai defektai),
- nervinio vamzdelio,
- CNS (centrinės nervų sistemos),
- chromosominės, pagrindinė jų dalį sudaro Dauno sindromas,
- virškinimo sistemos,
- urogenitalinės sistemos,
- galūnių defektų, tame tarpe galūnių redukcijos,
- lūpos/gomurio defektų.

Tiriant bendrus ĮRA-ų paplitimo dėsningumus ir dažnumo kitimo tendencijas Lietuvoje naudojome binominės logistinės regresijos (arba logtiesinį binominį) modelį. Logistinės regresijos modelio adekvatumo patikrinimui buvo taikomas Hosmerio ir Lemešovo suderinamumo kriterijus (Hosmer and Lemeshow Goodness-of-Fit Test, žiūr. SAS/STAT (1997)). Šio kriterijaus p-reikšmė HL p bei įvykio ir jo tikimybės prognozės ranginės koreliacijos koeficientas c, jei tik jie suskaičiuoti, bus nurodomi kiekvienam parinktam modeliui, kartais tiesiog skliausteliuose po tiriamos ĮRA-os pavadinimo.

Pagrindiniai prognozavimo kintamieji (prediktoriai, kovariantai) buvo naujagimio gimimo data, nusakoma metais, ir vieta, kurioje gyveno naujagimio motina, nusakoma administraciniu Lietuvos padalinimu (rajonu). Iš pastarojo kintamojo buvo sudaryta visa eilė išvestinių rodiklių:

- *Etnolingvistinė grupė* (etnos) pagal prof. Zigmą Zinkevičių (Z. Zinkevičius (1998), taip pat žiūr. V. Kučinskas (2001)). Jis siūlo tokį skirstymą: žemaičiai – pietų (sž = 1), vakarų (wž = 2) ir šiaurės (nž = 3), aukštaičiai – pietų (sa = 4), vakarų (va = 5) ir rytų (ea = 6).
- Dichotominis skirstymas į etnolingvistines grupes *eza: žemaičiai – aukštaičiai.*
- *Urbanizacijos lygis* (urbalyg): 0 = kaimas, 1 = miestas, 2 = didmiestis.
- *Zona*, atspindi vietos nutolimą nuo Baltijos jūros. Buvo sudarytos 6 maždaug vienodo pločio zonos, 1-oji prie pat Baltijos jūros, 6-oji – labiausiai nuo jos nutolusi, prie Baltarusijos sienos.

Į tyrimą taip pat buvo įtraukti papildomi kintamieji aprašantys šių (ir kitų) rodiklių sąveiką. Juos žymėsime sąveikaujančių rodiklių vardais sujungtais žvaigždute (įprastas žymėjimas aprašant modelius statistiniuose paketuose). Pavyzdžiui, Etnos*Zona reikš

kintamąjį, aprašantį kintamųjų Etnos ir Zona sąveiką. ĮRA-os tikimybės modelivimui buvo taikomi tik taip vadinami *hierarchiniai* modeliai. Tai modeliai tenkinantys sąlygą: jeigu modelio prediktorių sąraše yra kurių nors prediktorių sąveika, tai jame turi būti ir patys prediktoriai.

Skaičiavimai atlikti su statistinės duomenų analizės sistema SAS (LŽGC licenzija). Logistinė regresinė analizė atlikta su SAS procedūra PROC LOGISTIC (SAS/STAT (1997)).

Pirmame tyrime be duomenų bazės LIRECA buvo taip pat naudojami statistiniai duomenys apie 1993–1997 metais Lietuvoje gimusius vaikus pagal rajonus (viso 208044). Gauti rezultatai pateikti tik toms ĮRA-oms, kurioms pavyko nustatyti statistiškai reikšmingą (p-reikšmė mažesnė už 0,1) efektą nors vieno iš ankščiau minėtų prediktorių. Pirmasis skaičius po kintamojo pavadinimo nurodo p-reikšmę, o antrasis – didžiausio tikėtinumo įvertį $\hat{\beta}$ arba šansų santykio įvertį $\hat{O}$, jeigu kintamasis yra kiekybinis arba dichotominis (t.y., turi tik vieną laisvės laipsnį).

Bendras ĮRA-ų sntykinis dažnumas (iš viso 3346 atvejai) ir jo kitimo tendencijos statistiškai reikšmingai skiriasi tarp rajonų: *metai* ($p = 0,0326$, $\hat{\beta} = -0,0384$), *rajonas* ($p < ,0001$), *metai*rajonas* ($p = 0,0307$). Hosmer'io ir Lemeshov'o testo p-reikšmė HL $p = 0,1793$, o $\gamma = 0,562$. Tiriant metų ir teritorinių kintamųjų įtaką buvo išskirti tokie statistiškai reikšmingi rodikliai: *zona* ($p = 0,0004$), *etnolingvistinė grupė* (*etnos*) ($p = 0,0069$), *etnos*zona* ($p = 0,0012$), *metai*etnos* ($p = 0,0415$), *metai* ($p = 0,0608$, $\hat{\beta} = -0,0291$). Parinktam modeliui HL $p = 0,3184$, o $c = 0,538$. Vadinasi, ĮRA-ų santykinis dažnumas kasmet po truputį mažėja.

Širdies ir kraujagyslių anomalijų santykinis dažnumas (viso 690 atvejų) statistiškai reikšmingai skiriasi tarp *rajonų* ($p = 0,0006$) ir priklauso nuo *metų* ($p = 0,0042$, $\hat{O} = 0,925$), HL $p = 0,2291$, oc $= 0,570$. Tiriant metų ir teritorinių kintamųjų įtaką buvo išskirti tokie statistiškai reikšmingi rodikliai: *etnolingvistinė grupė* (*etnos*) ($p = 0,0105$), *zona* ($p = 0,0851$), *metai*zona* ($p = 0,0040$), *metai*etnos* ($p = 0,0043$) ir *etnos*zona* ($p = 0,0175$). Šiam modeliui HL $p = 0,0978$, o $c = 0,556$.

Dauginių defektų (*DVD*) santykiniai dažnumai (viso 365 atvejai) taip pat statistiškai reikšmingai skiriasi tarp rajonų ($p = 0,0465$); HL $p = 1,0$, o $c = 0,570$. Statistiškai rekšmingas jų ryšys su laiko-teritoriniais rodikliais *etnos* ($p = 0,0767$), *metai*zona* ($p = 0,0787$) ir beveik „reikšmingas" su metais ($p = 0,1014$, $\hat{\beta} = -0,0694$); HL $p = 0,3632$, o $c = 0,552$.

Statistiškai reikšmingos *rajonų* ar kitų teritorinių kintamųjų įtakos *nervinio vamzdelio* anomalijų (viso jų yra 354) santykiniam dažnumui nėra, bet ryški jo mažėjimo tendencija laike, rodiklio *metai* $p < ,0001$, $\hat{O} = 0,838$, HL $p = 0,0310$ (modelis nesuderintas), $c = 0,527$.

Labai panaši situacija ir su *CNS* anomalijom, kurių iš viso yra 139. Rodiklio *metai* $p = 0,0044$, $\hat{O} = 0,839$, HL $p = 0,4120$, $c < 0,520$.

Chromosominių anomalijų dažnumui (297 atvejai) statistiškai reikšminga yra *zona* ($p = 0,0138$), *etnos*zona* ($p = 0,0226$) ir *metai* ($p = 0,0803$, $\hat{O} = 1,074$), tarp rajonų statistiškai reišmingų skirtumų nėra. Parinktam modeliui HL $p = 0,8621$, o $c = 0,511$.

Pagrindinę chromosominių anomalijų dalį, 263 atvejai iš 297, sudaro *Dauno sindromas*. Todėl natūralu, kad ir šiai ĮRA-ai tarp rajonų statistiškai reišmingų skirtumų nėra. Reikš-

mingi rodikliai yra *zona* ($p = 0,0779$), *metai*zona* ($p = 0,0389$) ir *etnos*zona* ($p = 0,0500$), HL $p = 0,4896$, o $c = 0,519$.

Urogenitalinės sistemos anomalijoms (jų užregistruota 118) teritoriniai rodikliai įtakos neturi, bet yra ryški jų dažnėjimo tendencija, rodiklio *metai* $p = 0,0068$, $\hat{O} = 1,194$, HL $p = 0,7064$, nors $c < 0,502$.

Galūnių defektų, viso jų užregistruota 378, santykinis dažnis statistiškai reikšmingai skiriasi tarp *rajonų* ($p = 0,0011$) ir priklauso nuo *metų* ($p = 0,0948$, $\hat{O} = 0,941$), HL $p = 0,8215$, o $c = 0,605$. Skirtumus tarp rajonų gerai apašo rodiklis *zona* ($p <$,0001), tačiau kadangi į modelį įeina tik jis vienas (rodiklio metai p-reikšmė $p = 0,1041$ neperžiangia kritinės ribos 0,1), tai c šiek tiek mažesnis, $c = 0,570$, o HL $p = 0,9940$.

Lūpos/gomurio defektų, jų užregistruota 241, santykinio dažnumo statistiškai reikšmingų skirtumų tarp rajonų nėra, tačiau jis susijęs su kitais teritoriniais ir laiko-teritoriniais rodikliais: *zona* ($p = 0,0402$), *metai*etnos* ($p = 0,0604$), *metai*zona* ($p = 0,0678$), *metai* ($p = 0,0796$, $\hat{\beta} = 6,5258$), *metai*eza* ($p = 0,0868$, $\hat{\beta} = -3,226$) ir *metai*etnos*zona* ($p = 0,0886$). Rodiklio *metai*eza* įtraukimas į logstinės regresijos modelį reiškia, kad yra skirtingos nagrinėjamos ĮRA-os santykinio dažnio kitimo tendencijos laike *žemaičių ir aukštaičių etnolingvistinėse grupėse* (rodiklis eza). Parinktame modelyje HL $p = 0,8814$, $c = 0,518$.

Antrame tyrime buvo naudojama tik duomenų bazė LIRECA. Interpretuojant jame gautus rezultatus reikia turėti omenyje, kad šiuo atveju nagrinėjamas vieno ar kito prediktoriaus statistinis ryšys su kurios nors ĮRA-os santykiniu dažnumu kitų ĮRA-ų tarpe, t.y. tarp ĮRA-as turinčių Lietuvos naujagimių, o ne jos dažnumu aplamai tarp visų Lietuvos naujagimių.

Į analizę buvo įtraukti tokie papildomi prediktoriai: naujagimio lytis (lytis.), gimimo mėnuo (men), motinos (mamz) ir tėvo (tamz) amžius vaiko gimimo metu, ĮRA-os motinos (ams) ar tėvo (ats) šeimoje arba tarp brolių ar seserų (abs), o taip pat išvestiniai rodikliai: ar tėvas vyresnis už motiną (amzdf0), tėvo ir motinos amžiaus skirtumas, jei jis neigiamas, (amzdf1); bei jų sąveikos, amzdf2=-amzdf1* amzdf1, $(\text{tamz-mamz})^2$.

Dažniausiai pasitaikančių *širdies ir kraujagyslių* anomalijų santykiniam dažnumui (770 atvejai iš 3700, HL $p = 0,6008$, $c = 0,586$) statistiškai reikšmingi *zona* ($p = 0,0144$), *amzdf1* ($p = 0,0185$, $\hat{O} = 1,069$), *anomalijos tėvo šeimoje* (ats) ($p = 0,0271$, $\hat{O} = 0,703$), *lytis* ($p <, 0434$, $\hat{O} = 1,180$), *gimimo mėnuo* (men; šiuo atveju jis laikomas kiekybiniu kintamuoju) ($p = 0,0589$, $\hat{O} = 0,978$) ir *urbanizacijos lygis* (urbalyg) ($p = 0,0677$, $\hat{O} = 0,911$). Kol kas sunku duoti racionalų mėnesio numerio ryšio su AK1 dažnumu paaiškinimą.

Dauginių defektų (*DVD*) santykiniam dažnumui (412 atvejai iš 3990, HL $p = 0,5646$, $c = 0,605$) statistiškai reikšmingi *amzdf0* ($p = 0,0004$, $\hat{O} = 0,308$), *amzdf1* ($p = 0,0015$, $\hat{O} = 0,626$) ir *amzdf2* ($p = 0,0211$, $\hat{O} = 0,970$), t.y. tai, kad motina vyresnė už tėvą, *urbanizacijos lygis* (urbalyg, šiuo atveju tai kategorinis kintamasis) ($p = 0,0036$), *etnolingvistinės grupės numeris* (etnos; šiuo atveju skaitoma, kad jis yra kiekybinis rodiklis) ($p = 0,0254$, $\hat{O} = 1,196$), *anomalijos tėvo šeimoje* (ats) ($p = 0,0430$, $\hat{O} = 1,413$) ir *zona* ($p = 0,0774$). Etnolingvistinės grupės numeris atspindi tam tikrą tendenciją, tačiau jo ryšio su DVD prasminga interpretacija neaiški.

Nervinio vamzdelio anomalijų santykiniam dažnumui (377 atvejai iš 4362, HL $p = 0,2496$, $c = 0,632$) statistiškai reikšmingi *metai* ($p < ,0001$, $\hat{O} = 0,872$), *lytis* ($p < ,0001$, $\hat{O} = 1,686$), *anomalijos motinos šeimoje* (ams) ($p = 0,0079$, $\hat{O} = 0,554$) ir *urbanizacijos lygis* (urbalyg) ($p = 0,0520$, $\hat{O} = 1,130$).

CNS anomalijų dažnumui, 161 atvejai iš 4174, (HL $p = 0,1537$, $c = 0,596$) statistiškai reikšmingi *zona* ($p = 0,0144$), *motinos amžius* (mamz) ($p = 0,0489$, $\hat{O} = 0,994$) ir *metai* ($p = 0,0439$, $\hat{O} = 0,916$).

Chromosominių anomalijų santykiniam dažniui (405 atvejai iš 3990, HL $p < ,0001$, $c = 0,648$) statistiškai reikšmingi *motinos amžius* (mamz) ($p < ,0001$, $\hat{O} = 1,101$), *anomalijos motinos šeimoje* (ams) ($p = 0,0480$, $\hat{O} = 0,673$) ir *AMZDF2* ($p = 0,0481$, $\hat{O} = 1,005$). Deja, modelis neadekvatus. Pagrindinę chromosominių ĮRA-ų dalį, 333 atvejai iš 405, sudaro *Dauno sindromas*. Todėl natūralu, kad jam gavosi labai panašūs rezultatai: (HL $p < ,0001$, $c = 0,659$) statistiškai reikšmingi *motinos amžius* (mamz) ($p < ,0001$, $\hat{O} = 1,110$), *anomalijos motinos šeimoje* (ams) ($p = 0,0252$, $\hat{O} = 0,597$) ir *AMZDF2* ($p = 0,0758$, $\hat{O} = 1,004$). Labiausiai tikėtina parinkto logistinės regresijos modelio neatitikimo duomenims priežastis yra neįtraukti į modelį svarbūs įtakojantys faktoriai (prediktoriai). Pavyzdžiui, pastebimas žymus Dauno sindromo dažnumo sumažėjimas pradedant 2000-aisiais. Motinos amžiaus įtaka Dauno sindromui yra gerai žinoma.

Virškinimo sistemos anomalijų santykiniam dažnumui (124 atvejai iš 4013, HL $p = 0,7552$, $c = 0,550$) statistiškai reikšmingas tik *tėvo amžius* (tamz) ($p = 0,0182$, $\hat{O} = 0,963$).

Urogenitalinės sistemos anomalijų santykiniam dažnumui (180 atvejai iš 4362, HL $p = 0,5618$, $c = 0,678$) statistiškai reikšmingi *metai* ($p < ,0001$, $\hat{O} = 1,191$), *lytis* ($p < ,0001$, $\hat{O} = 0,509$), *zona* ($p = 0,0380$) ir *žemaičių-aukštaičių etnolingvistinė grupė* (eza) ($p = 0,0660$, $\hat{O} = 0,246$).

Galūnių defektų santykiniam dažnumui (497 atvejai iš 4362, HL $p = 0,4602$, $c = 0,631$) statistiškai reikšmingi *anomalijos motinos* (ams) ($p < ,0001$, $\hat{O} = 2,325$) ir *tėvo* (ats) ($p < ,0001$, $\hat{O} = 1,873$) šeimoje, *zona* ($p < ,0001$), *gimimo mėnuo* (men, šiuo atveju tai yra kiekybinis kintamais) ($p = 0,0166$, $\hat{O} = 0,967$), *lytis* ($p < ,0001$, $\hat{O} = 0,796$) ir *etnolingvistinės grupės numeris* (etnos; šiuo atveju skaitoma, kad jis yra kiekybinis rodiklis) ($p = 0,0267$, $\hat{O} = 0,851$).

Lūpos/gomurio defektų santykiniam dažniui (277 atvejai iš 3709, HL $p = 0,5761$, $c = 0,578$) statistiškai reikšmingi *lytis* ($p = 0,0005$, $\hat{O} = 0,635$), *motinos amžius* (mamz) ($p = 0,0451$, $\hat{O} = 1,005$), *anomalijos motinos šeimoje* (ams) ($p = 0,0755$, $\hat{O} = 1,378$) ir *amzdf2* ($p = 0,0789$, $\hat{O} = 0,984$).

4. Išvados

Bendras ĮRA-ų dažnumas turi silpną, tačiau statistiškai reišmingą mažėjimo tendenciją laike. Šią tendenciją, matyt, galima paaiškinti LŽDC-o vykdomu profilaktiniu darbu. Nustatyti statistiškai reikšmingi bendro ĮRA-ų dažnumo skirtumai tarp Lietuvos rajonų,

tačiau atskiroms ĮRA-ų grupėms jie yra saviti arba jų iš viso nėra. Dauno sindromui ir chromosominėms ĮRA-oms, kurių didžiausią dalį ir sudaro Dauno sindromas, logistinės regresijos modelio parinkti nepavyko, Hosmerio ir Lemešovo suderinamumo kriterijaus p-reikšmė mažesnė už 0,0001. Labiausiai tikėtina to priežastis yra neįtraukti į modelį svarbūs rodikliai. Pavyzdžiui, pastebimas žymus Dauno sindromo dažnumo sumažėjimas bei jo dažnį įtakojančių faktorių pasikeitimas pradedant 2000-aisiais. Kitoms ĮRA-ų grupėms parinkti logistinės regresijos modeliai turi skirtngus kintamųjų sąrašus, dažniau pasitaikantys rodikliai yra motinos amžius, anomalijos motinos ar tėvo šeimoje, kūdikio lytis, urbanizacijos lygis. Pastebėti statistiniai dėsningumai reikalauja tolesnio tyrimo siekiant suteikti jiems prasmingą dalykinę interpretaciją.

Autoriai dėkingi prof. V. Kučinskui už galimybę naudotis LŽGC duomenų bazės LIRECA nekonfidencialia informacija ir duomenų analizės sistema SAS, Vilniaus Technikos Universiteto studentėms A. Čaplinskajai ir J. Židanavičiūtei padėjusioms atlikti šį darbą, o taip pat recenzentui, kurio pastabos žymiai pagerino straipsnio kokybę.

Literatūra

[1] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York (1990).
[2] R. Christensen, *Log-Linear Models*, Springer-Verlag, New York, Berlin (1990).
[3] S.E. Fienberg, Contingency tables and log-linear models: basic results and new developments, *Journal of the American Statistical Association*, **95**(450), 643–647 (2000).
[4] V. Kučinskas, Population genetics of Lithuanians, *Annals of Human Biology*, **28**, 1–14 (2001).
[5] Th.J. Santer, D.E. Duffy, *The Statistical Analysis of Discrete Data*, Springer-Verlag, New York, Berlin (1989).
[6] *SAS/STAT Software: Changes and Enhancments through Release 6.12*, SAS Institute Inc., Cary, NC, USA (1997).
[7] D.M. Umbach, C.R. Weinberg, Designing and analysing case-control studies to exploit independence of genotype and exposure, *Statistics in Medicine*, **16**, 1731–1743 (1997).
[8] A. Utkus, Lietuvos vaikų įgimtų raidos anomalijų etiologija, diagnostika, struktūra ir paplitimas, *Daktaro disertacija*, Vilnius (2001).
[9] Z. Zinkevičius, *The History of the Lithuanian Language*, Mokslo ir enciklopedijų leidimo institutas, Vilnius (1998).

Statistical analysis of congenital anomlies in children in Lithuania

M. Radavičius, J. Sušinskas, A. Utkus

Trends, prevalence and some influential (risk) factors of congenital anomlies in children in Lithuania are investigated using logistic regression model.