Український Журнал
ХІРУРГІЇ

УДК 618.19-006.6:001.891.57+519.87

РУСИН А.В.[1, 2], РУСИН В.І.[1], ОДОШЕВСЬКА О.М.[1], ДЕВІНЯК О.Т.[1]
[1]Ужгородський національний університет, медичний факультет, м. Ужгород, Україна
[2]Закарпатський обласний клінічний онкологічний диспансер, м. Ужгород, Україна

ПОБУДОВА МАТЕМАТИЧНОЇ МОДЕЛІ ДЛЯ ОПТИМІЗАЦІЇ ПРОГРАМИ СКРИНІНГУ РАКУ ГРУДНОЇ ЗАЛОЗИ

Резюме. *У статті наведена робота з оптимізації анкети-опитувальника для скринінгу раку грудної залози, побудована на її основі математична модель, здатна визначати ризик розвитку раку грудної залози, встановлений вплив факторів анамнезу та способу життя на імовірність виникнення раку.*
Матеріали та методи. *Статистичний аналіз та моделювання проводили у середовищі R 3.0.1. Для визначення формули для розрахунку ризику розвитку раку грудної залози використовували модель L1-регуляризованої логістичної регресії, для оптимізації границь між класами виконувався аналіз операційної характеристики моделі.*
Результати та обговорення. *На основі даних анкетування 321 жінки була оптимізована анкета-опитувальник для скринінгу раку грудної залози, підтверджений негативний вплив основних факторів на ризик розвитку раку грудної залози. Крім факторів високого ризику ідентифіковано також ряд сприятливих факторів, що знижують ризик розвитку раку: годування груддю понад 3 місяці, припинення менструацій до 45 років. Новостворена модель для визначення ризику раку грудної залози на основі анкетування жіночого населення характеризується високою точністю підгонки (97,5 %) та прогнозування (94,1 %), що дозволило створити комп'ютерну програму для визначення ризику раку грудної залози для використання в клініках.*
Висновок. *Використання запропонованого математичного моделювання істотно покращує ефективність анкетного скринінгу, спрощує анкетування за рахунок зменшення кількості запитань, є більш точним та швидким способом визначення груп ризику раку грудної залози.*
Ключові слова: *фактори ризику, рак грудної залози, модель прогнозування, математичне моделювання.*

Вступ

Як відомо, етіологія раку грудної залози (РГЗ) залишається до цього часу неясною [4], а основними проблемами, що стоять перед вітчизняним лікарем, є визначення ризику захворювання на РГЗ у цілого контингенту населення або в індивідуальному випадку, раннє виявлення хвороби (передракового стану чи РГЗ), визначення обсягу і характеру необхідних лікувальних заходів [1].

Моделі прогнозування ризику дозволяють лікарю і фахівцям громадської охорони здоров'я оцінювати індивідуальний ризик розвитку РГЗ із використанням відомих епідеміологічних і клінічних факторів [16].

Існуючі моделі розрахунку ризику РГЗ базуються на комбінації факторів ризику та вираховують ризик РГЗ на певний період часу чи на все життя жінки. Кілька моделей було розроблені з цією метою, найуживанішими з них є: модель Гейла [14], що була змінена [21] та доповнена [9, 13], модель Клауса [10], модель Тайрера — Кужика (модель IBIS) [20], модель BOADICEA [19] та ін.

У даний час більшість моделей мають помірну дискримінаційну здатність, з площею під кривими в діапазоні від 0,55 до 0,70 [17, 22], що обмежує їх використання в клініці.

Суттєвим є і той факт, що у 60 % РГЗ виникає спорадично за відсутності відомих факторів ризику [15].

Онкоепідеміологічне тестування є інформативним, дешевим, простим, безпечним та неінвазивним методом для виявлення серед здорового населення груп осіб, які мають фактори ризику та потребують проведення уточнюючої діагностики [3, 7]. Багатьма дослідниками були складені анкети для визначення ризику захворювання на РГЗ серед здорового жіночого населення [2, 5, 6].

Науковці наголошують на необхідності додаткових досліджень для ідентифікації високих прогностичних маркерів ризику РГЗ з наступним

включенням їх в більш точну модель оцінки ризику РГЗ [8].

Метою дослідження є оптимізація анкети-опитувальника для скринінгу РГЗ, побудова на її основі математичної моделі, здатної визначати ризик розвитку РГЗ, а також встановлення впливу факторів анамнезу та способу життя пацієнтки на імовірність виникнення РГЗ.

Матеріал та методи

Статистичний аналіз та моделювання проводили у середовищі R 3.0.1. Для визначення формули для розрахунку ризику розвитку РГЗ, використовували модель логістичної регресії. Два класи: норма та РГЗ слугували бінарним відгуком моделі, а відповіді на питання анкети скринінгу РГЗ — можливими предикторами (27 питань, 40 предикторів).

Числові характеристики стандартизувались таким чином, щоб нульовий рівень предиктора відповідав першому квартилю характеристики. Для вибору найменшої кількості предикторів, потрібних для проведення класифікації із максимальною точністю, використовували покроковий метод побудови моделі логістичної регресії. Так, починаючи із нульової моделі, на кожному кроці вносився або вилучався лише один предиктор таким чином, щоб мінімізувати інформаційний критерій Акаіке (AIC, [12]). Для побудови та перехресної валідації моделі *L1*-регуляризованої логістичної регресії використовували додатковий програмний пакет glmnet [11]. Для оптимізації граничного рівня виконувався аналіз операційної характеристики моделі (ROC-аналіз [18]).

Індекс маси тіла (ІМТ) вираховували за формулою Кетле (ІМТ = вага(кг)/зріст($м^2$)).

Результати та обговорення

За період з 2011 по 2012 рік проведено анкетування 321 жінки за раніше розробленою комп'ютерною програмою на основі анкети-опитувальника за відомими факторами ризику. Ми створили й протестували прототип, спрямований на пересічних жінок для розрахунку ризику РГЗ в клінічних умовах.

Результати попереднього дослідження показали, що найбільш цінними предикторами РГЗ є поява незвичних симптомів при самообстеженні, наявність маститу, хронічних захворювань щитоподібної залози в анамнезі та сильних або частих стресових ситуацій, а також ранній початок менструацій. З іншого боку, встановлено, що місцевість проживання, штучна менопауза, наявність гіпертонічної хвороби та діабету не проявляють статистичного зв'язку із доброякісними чи злоякісними процесами в ГЗ.

Змінними в моделі прогнозування ризику розвитку РГЗ були: вік, вага, зріст, паління, вживання алкоголю, вік менархе, наявність передменструального синдрому (ПМС), протизаплідні заходи, вік менопаузи, вік першої вагітності, кількість пологів, тривалість лактації, наявність абортів/викиднів, попередні захворювання грудних залоз (мастит, травма, доброякісні стани), хронічні захворювання щитоподібної залози, печінки, матки та/або яєчників, регулярність проходження огляду в гінеколога та наявність симптомів при самообстеженні, наявність родичок з РГЗ та випадки раку іншої локалізації у сім'ї.

У результаті валідації моделі було виконано 26 кроків, а кінцева модель містила 22 предиктори. Точність моделі становила 97,2 %, а точність при 10-кратній перехресній валідації — 93,0 %. Коефіцієнти моделі наведені в таблиці 1.

Помітно, що більшість факторів мають високий рівень статистичної значимості. Однак три предиктори: шкідливі звички — алкоголь, перша вагітність — після 40 років та — виділення з соска при самообстеженні істотно відрізняються від інших, маючи значні коефіцієнти із ще більшими стандартними похибками та p-величинами > 0,99. Це свідчило про нестабільність моделі і було зумовлено перехресними кореляціями між предикторами.

Для встановлення адекватних коефіцієнтів створювали окрему модель логістичної регресії із тим же набором предикторів, однак впровадивши у процес побудови *L1*-регуляризацію (*least absolute shrinkage and selection operator, LASSO*). Суть *L1*-регуляризації полягає в додаванні до цільової функції регресії штрафу за складність моделі, пропорційного до норми вектора коефіцієнтів.

Краща модель була знайдена при параметрі *L1*-регуляризації $\lambda = 0{,}0005$ і характеризувалася точністю підгонки (97,5 %) та середньою точністю при 10-кратній перехресній валідації (94,1 %). Тобто, крім виправлення коефіцієнтів, за допомогою *L1*-регуляризації було також досягнуто покращення моделі. Матриця похибок моделі наведена в таблиці 2.

Чутливість моделі (*Sens)* (за результатами перехресної валідації) становить 90,2 %, а специфічність (*Spec)* — 96,8 %. Значення коефіцієнтів регуляризованої моделі наведені в таблиці 3.

Крім факторів високого ризику ідентифіковано також ряд сприятливих факторів, що знижують ризик розвитку РГЗ. До них належать: годування груддю понад 3 місяці (2 фактори) та припинення менструацій до 45 років.

Щодо інших факторів, то ожиріння (високі значення ІМТ) може істотно підвищити ризик розвитку РГЗ. Вік менше 45 років знижує ризик, а після 45 — підвищує. Однак невдовзі після 45 років відбувається припинення менструацій, що нівелює нарахований ризик. Таким чином, лише пізня менопауза, а також старший вік є факторами, що підвищують ризик. Ранній початок менструацій також чинить певний несприятливий вплив, додаючи до зсуву по 1,15 одиниці за кожен рік, що відділяє вік початку менструацій від віку 14 років.

Примітно, що тривале непроходження медоглядів підвищує ризик виявлення РГЗ.

Слід розуміти, що вибірка, взята для побудови моделі, не відображає структуру популяції. У вибірці частка жінок із раком становила 41,1 %, тоді як у по-

Таблиця 1. Коефіцієнти моделі покрокової логістичної регресії

Фактор впливу	Коефіцієнт	Стандартна похибка	P-величина
Зсув	–7,4165	1,7478	0,000022
ІМТ	0,4234	0,1155	0,000247
Вік пацієнтки	2,4698	0,6869	0,000324
Стрес	2,8101	0,8639	0,001142
Мастит	7,4729	1,9964	0,000182
Травми грудної залози	4,2094	1,3723	0,002159
Хрон. захворювання щитоподібної залози	2,469	1,5706	0,115943
Шкідливі звички — алкоголь	19,8271	2305,48	0,993138
Шкідливі звички — паління	3,8996	1,483	0,008552
Шкідливі звички — паління та алкоголь	5,7234	2,3667	0,015592
Початок менструацій	1,2385	0,457	0,00673
Менстр. регулярність — припинилися	–5,3076	1,665	0,001434
Перша вагітність — до 23 років	1,8578	1,1019	0,0918
Перша вагітність — після 28 років	6,5833	1,9687	0,000826
Перша вагітність — після 40 років	24,5302	5455,05	0,996412
Годування груддю — годувала 3–12 місяців	–3,7956	1,2367	0,002147
Годування груддю — годувала більше 12 місяців	–4,9935	1,4134	0,000411
Самообстеження — виділення із соска	21,6967	3757,481	0,995393
Самообстеження — поява у грудних залозах вузлів	8,1804	1,9614	0,000030
Медогляд — ніколи не проходила	4,8673	1,9593	0,012983
Медогляд — проходила більше року тому	4,4882	1,4382	0,001804
Медогляд — проходила менше року тому	2,7104	1,2266	0,027125
РГЗ чи рак яєчників у родичок	3,3277	1,3569	0,014192

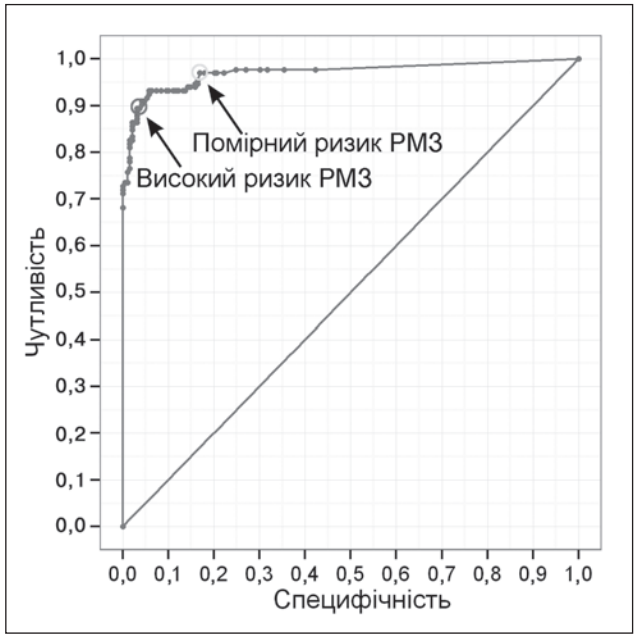

Рисунок 1. Операційна характеристика результатів перехресної валідації моделі прогнозування ризику РГЗ

пуляції захворюваність становить близько 45,3 жінки на 100 000 за рік, тобто 0,0453 %. Тому, якщо модель прогнозує наявність РГЗ, то насправді імовірність раку не 95,2%, як випливає із матриці похибок, а згідно з теоремою Байєса

Таблиця 2. Матриця похибок моделі при перехресній валідації

		Спостережено	
		Норма	**Рак**
Прогнозовано	**Норма**	183	13
	Рак	6	119

$$P(AX) = Sens \times P(A) / Sens \times P(A) + (1 - Spec) \times \times (1 - P(A)) = 0{,}902 \times 0{,}000453 / 0{,}902 \times 0{,}000453 + + 0{,}032 \times 0{,}999547 = 0{,}0126.$$

Тобто внаслідок того, що РГЗ є відносно рідкісним явищем, при скринінгу всієї популяції лише 1,26 % серед пацієнток, яким модель визначить високий ризик РГЗ, справді матиме в цей рік РГЗ.

Оскільки чутливість моделі при валідації становить 90,2 %, близько 10 % хворих на РГЗ при скринінгу буде пропущено. Щоб зменшити цю кількість, слід впровадити проміжний клас «помірний ризик РГЗ», для якого зменшити в моделі граничний рівень імовірності для прогнозування РГЗ (який для класу «високий ризик РГЗ» становить 50 %). Для оптимізації граничного рівня виконувався аналіз операційної характеристики моделі (ROC-аналіз [18], рис. 1).

Найближчий до максимальної чутливості локальний оптимум операційної характеристики було вибрано як граничний рівень. Координати оптимуму — чутливість 97,0 %, специфічність 83,1 % — спостерігаються

при граничному рівні 12,3 %. Тобто, якщо відгук моделі становить ймовірність понад 50 %, зараховуємо пацієнта до групи високого ризику, якщо менше 50 %, але більше 12,3 % — до групи помірного ризику, і якщо менше 12,3 % — до групи низького ризику РГЗ. Також слід відзначити, що площа під кривою операційної характеристики 0,974, що значно перевищує характеристики існуючих моделей для скринінгу РГЗ.

Крім того, із рис. 1 помітно, що при специфічності 100 % (практично безпомилкове передбачення наявності РГЗ) можна досягти показника чутливості 73,5 % (при граничному рівні прогнозу 95,6 %). Тобто пацієнтки, для яких модель спрогнозує показник ймовірності вище 95,6 %, майже гарантовано мають РГЗ, причому таким способом можна виявити майже три чверті всіх хворих на РГЗ.

Порівняння результатів прогнозування за допомогою моделі *L1*-регуляризованої логістичної регресії із прогнозом, що був сформований на основі цієї ж анкети, однак за допомогою адитивної формули та з коефіцієнтами, визначеними на основі експертної думки онкологів та мамологів, наведено в таблиці 4.

Помітно, що використання математичного моделювання істотно покращує ефективність скринінгу за допомогою анкети-опитувальника. Між прогнозами обох моделей існує сильна кореляція (поліхорний кореляційний коефіцієнт $\varphi = 0{,}928$).

На базі моделі була створена комп'ютерна програма на мові програмування C# для визначення ризику РГЗ для використання в клініці, що є зрозумілою та зручною у використанні (рис. 2).

Висновки

1. Новостворена модель для визначення ризику раку грудної залози на основі анкетування жіночого населення характеризується високою точністю (97,5 %) та середньою точністю при 10-кратній перехресній валідації на рівні 94,1 %.

Таблиця 3. Коефіцієнти моделі L1-регуляризованої логістичної регресії

Фактор впливу	Коефіцієнт	Експонента коефіцієнта
Зсув	–6,8511	0,001058294
ІМТ	0,387639	1,473497632
Вік пацієнтки	2,251965	9,506400398
Стрес	2,621142	13,75141255
Мастит	6,861981	955,2576166
Травми грудної залози	3,841325	46,58713819
Хрон. захворювання щитоподібної залози	2,200601	9,030440064
Шкідливі звички — алкоголь	6,649803	772,6322287
Шкідливі звички — паління	3,551073	34,85069129
Шкідливі звички — паління та алкоголь	5,077335	160,3461575
Початок менструацій	1,158169	3,184096242
Менстр. регулярність — припинилися	–4,77038	0,008477165
Перша вагітність — до 23 років	1,627452	5,090886655
Перша вагітність — після 28 років	5,972894	392,6401644
Перша вагітність — після 40 років	9,462918	12873,39376
Годування груддю — годувала 3–12 місяців	–3,49157	0,030452978
Годування груддю — годувала понад 12 місяців	–4,50698	0,011031705
Самообстеження — виділення із соска	7,596227	1990,67166
Самообстеження — поява в грудних залозах вузлів	7,56149	1922,707422
Медогляд — ніколи не проходила	4,470936	87,43853092
Медогляд — проходила більше року тому	4,087886	59,61372761
Медогляд — проходила менше року тому	2,494377	12,11418881
РГЗ чи рак яєчників у родичок	2,988478	19,85544345

Таблиця 4. Порівняння статистичних показників адитивної моделі та моделі логістичної регресії*, %

Статистичний показник	Адитивна модель, високий ризик	Модель логістичної регресії, високий ризик	Адитивна модель, помірний та високий ризик	Модель логістичної регресії, помірний та високий ризик
Точність	88,2	94,1	62,6	88,8
Чутливість	78,0	90,2	98,5	97,0
Специфічність	95,2	96,8	37,6	83,1

Примітка: * — показники для моделі логістичної регресії засновані на результатах 10-кратної перехресної валідації.

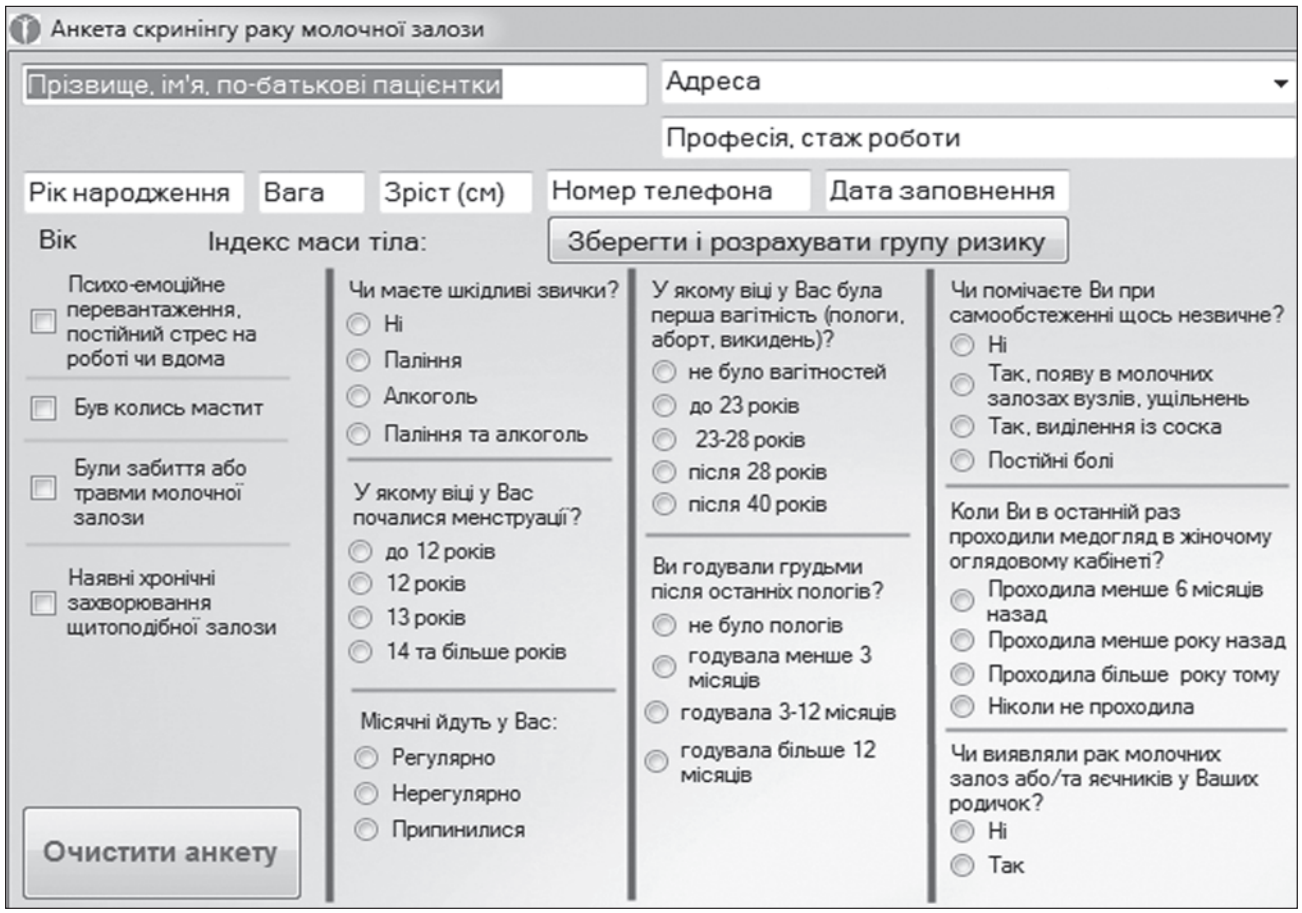

Рисунок 2. Інтерфейс програми для визначення ризику РГЗ на базі розробленої моделі

2. Використання запропонованого математичного моделювання істотно покращує ефективність анкетного скринінгу, спрощує анкетування за рахунок зменшення кількості запитань, є більш точним та швидким способом визначення груп ризику раку грудної залози.

Список літератури

1. *Білинський Б.Т. Еволюція клінічних підходів до проблеми раку грудної залози на фоні прогресу онкологічної науки / Б.Т. Білинський // Онкологія. — 2010. — Т. 12, № 3. — С. 282-285.*
2. *Искусственные нейронные сети: прогнозирование вероятности развития рака молочной железы у женщин, имеющих факторы риска / [Ю.В. Думанский, В.В. Приходченко, Ю.Е. Лях, В.Г. Гурьянов] // Нейронауки: теоретичні та клінічні аспекти. — 2007. — Т. 3, № 1–2. — С. 106-109.*
3. *Приходченко В.В. Анкетный скрининг как метод первичного отбора групп риска заболеваний молочной железы (предварительное сообщение) / В.В. Приходченко // Медико-соціальні проблеми сім'ї. — 2007. — Т. 12, № 1–2. — С. 57-65.*
4. *Профилактика рака молочной железы / В.Ф. Семиглазов, Г.А. Дашян, В.В. Семиглазов // Практическая онкология. — 2011. — Т. 12, № 2. — С. 66-69.*
5. *Смоланка І.І. Профілактика і рання діагностика раку молочної залози / І.І. Смоланка, С.Ю. Скляр, І.І. Доценко // Жіночий лікар. — 2009. — № 5. — С. 40-45.*
6. *Факторы риска злокачественных и доброкачественных заболеваний молочной железы / И.А. Коноплева, В.Ф. Левшин, Е.Г. Пинносевич [и др.] // Советская медицина. — 1990. — № 12. — С. 93-96.*
7. *Харченко В.П. Скрининг и возможности совершенствования ранней диагностики рака молочной железы / В.П. Харченко, Н.И. Рожкова, Е.В. Меских // Вестник Московского онкологического общества. — 2006. — № 11. — С. 4-5.*
8. *Assessment of the accuracy of the Gail model in women with atypical hyperplasia / V.S. Pankratz, L.C. Hartmann, A.C. Degnim [et al.] // J. Clin. Oncol. — 2008. — Vol. 26(33). — P. 5374-5379.*
9. *Breast cancer risk assessment in the Czech female population — an adjustment of the original Gail model / J. Novotny, L. Pecen, L. Petruzelka [et al.] // Breast Cancer Res Treat. — 2006. — Vol. 95. — P. 29-35.*
10. *Claus E.B.* Autosomal dominant inheritance of early onset breast cancer / *E.B. Claus, N. Risch, W.D. Thompson //* Cancer. *— 1994. — Vol.* 73. — P. *643-651*
11. *Friedman J. Regularization Paths for Generalized Linear Models via Coordinate Descent / J. Friedman, T. Hastie, R. Tibshirani // Journal of Statistical Software. — 2010. — 33(1). — P. 1-22.*
12. *Pan W. Akaike's information criterion in generalized estimating equations / W. Pan // Biometrics. — 2001. — 57(1). — P. 120-125.*

13. *Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density / J. Chen, D. Pee, R. Ayyagari [et al.] // J. Natl. Cancer. Inst. — 2006. — Vol. 98. — P. 1215-1226.*
14. *Projecting individualized probabilities of developing breast cancer for white females who are being examined annually / M.N. Gail, L.A. Brinton, D.P. Byar [et al.] // J. Natl. Cancer Inst. — 1989. — Vol. 81. — P. 1879-1989.*
15. *Proportion of breast cancer cases in the United States explained by well-established risk factors / M.P. Madigan, R.G. Ziegler, J. Benichou [et al.] // J. Natl. Cancer. Inst. — 1995. — Vol. 87(22). — P. 1681-1685.*
16. *Pu X. Development and validation of risk models and molecular diagnostics to permit personalized management of cancer / Xia Pu, Y. Ye, X. Wu // Cancer. — 2014. — Vol. 120, Issue 1. — P. 11-19.*
17. *Risk prediction models of breast cancer: a systematic review of model performances / T. Anothaisintawee, Y. Teerawattananon, N. Wiratkapun [et al.] // Breast Cancer Res Treat. — 2012. — Vol. 133. — P. 1-10.*
18. *ROCR: visualizing classifier performance in R / T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer // Bioinformatics. — 2005. — Vol. 21(20). — P. 3940-3941.*
19. *The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions / A.C. Antoniou, A.P. Cunningham, J. Peto [et al.] // Br. J. Cancer. — 2008. — Vol. 98(8). — P. 1457-1466.*
20. *Tyrer J. A breast cancer prediction model incorporating familial and personal risk factors / J. Tyrer, S.W. Duffy, J. Cuzick // Stat Med. — 2004. — Vol. 23(7). — P. 1111-1130.*
21. *Validation studies for models projecting the risk of invasive and total breast cancer incidence / J.P. Costantino, M.H. Gail, D. Pee [et al.] // J. Natl. Cancer. Inst. — 1999. — Vol. 91(18). — P. 1541-1548.*
22. *Risk prediction models for colorectal cancer: a review / A.K. Win, R.J. Macinnis, J.L. Hopper, M.A. Jenkins // Cancer. Epidemiol. Biomarkers Prev. — 2012. — Vol. 21. — P. 398-410.*

Отримано 11.05.14

Русин А.В.[1,2], Русин В.И.[1], Одошевская Е.М.[1], Девиняк О.Т.[1]
[1]Ужгородский национальный университет, г. Ужгород, Украина
[2]Закарпатский областной клинический онкологический диспансер, г. Ужгород, Украина

ПОСТРОЕНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ ДЛЯ ОПТИМИЗАЦИИ ПРОГРАММЫ СКРИНИНГА РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

Резюме. В статье представлена работа по оптимизации анкеты-опросника для скрининга рака молочной железы, построенная на ее основе математическая модель, способная определять риск развития рака молочной железы, установлено влияние факторов анамнеза и образа жизни на вероятность возникновения рака.

Материалы и методы. Статистический анализ и моделирование проводили в среде R 3.0.1. Для определения формулы для расчета риска развития рака молочной железы использовали модель L1-регуляризированной логистической регрессии, для оптимизации границы между классами выполнялся анализ операционной характеристики модели.

Результаты и обсуждение. На основе данных анкетирования 321 женщины была оптимизирована анкета-опросник для скрининга рака молочной железы, подтверждено негативное влияние основных факторов на риск развития рака молочной железы. Кроме факторов высокого риска идентифицированы также ряд благоприятных факторов, снижающих риск развития рака: кормление грудью более 3 месяцев, прекращение менструаций до 45 лет. Новая модель для определения риска рака молочной железы на основе анкетирования женского населения характеризуется высокой точностью подгонки (97,5 %) и прогнозирования (94,1 %), что позволило создать компьютерную программу для определения риска рака молочной железы для использования в клиниках.

Вывод. Использование предложенного математического моделирования существенно улучшает эффективность анкетного скрининга, упрощает анкетирования за счет уменьшения количества вопросов, является более точным и быстрым способом определения групп риска рака молочной железы.

Ключевые слова: факторы риска, рак молочной железы, модель прогнозирования, математическое моделирование.

Rusyn A.V.[1,2], Rusyn V.I.[1], Odoshevska O.M.[1], Devinyak O.T.[1]
[1]Uzhgorod National University, Uzhgorod, Ukraine
[2]Transcarpathian Regional Clinical Oncological Hospital, Uzhgorod, Ukraine

CREATION OF MATHEMATICAL MODEL TO OPTIMIZE BREAST CANCER SCREENING PROGRAM

Summary. This paper describes the work on optimization of breast cancer screening questionnaire, mathematical model created on its base, which is capable of determining the risk of breast cancer developing; there is identified the impact of factors of history and lifestyle on cancer risk.

Materials and Methods. Statistical analysis and modeling were performed in R 3.0.1. L1-regularized logistic regression model was used to determine the formula for calculating the risk of breast cancer, the operating characteristics analysis of the model was carried out in order to optimize the border between classes.

Results and Discussion. Based on a survey of 321 women, questionnaire for breast cancer screening has been optimized, the negative impact of main factors on the risk of breast cancer is confirmed. In addition to high-risk factors a number of favorable factors that reduce the risk of cancer have been identified: breast-feeding for more than 3 months, ischomenia before 45 years. The new model for determining breast cancer risk based on a survey of the female population is characterized by high accuracy of fitting (97.5 %) and prediction (94.1 %), that made it possible to create a computer program for use in clinics, that is capable to determine the risk of breast cancer.

Conclusion. Using proposed mathematical modeling significantly improves the efficiency of questionnaire screening, makes survey easy due to reducing the number of questions and appears to be more accurate and fast way to determine groups at risk of breast cancer.

Key words: risk factors, breast cancer, prediction model, mathematical modeling.