

Відомості про авторів / Сведения об авторах / About the Authors

Розвадовский Андрей Федорович – младший научный сотрудник, ФГБУН «Морской гидрофизический институт РАН»; email: rozvadovsky@yandex.ru.

Rozvadovskiy Andrey Fedorovich – junior researcher, FSIS «Marine Hydrophysical Institute RAS», Sevastopol; e-mail: rozvadovsky@yandex.ru.

Лазебный Владимир Семенович – кандидат технических наук, доцент Национального технического университета Украины «Киевский политехнический институт»; тел.: (067)387-69-10; e-mail: smartvs@mail.ru.

Lazebny Vladimir Semenovich – Candidate of Technical Sciences, Ass. Professor, National Technical University of Ukraine «KPI», Kiev, tel.: (067)387-69-10; e-mail: smartvs@mail.ru.

Пилинский Владимир Владимирович – кандидат технических наук, профессор Национального технического университета Украины «Киевский политехнический институт»; тел.: (095) 276-92-70; e-mail: pww@ukr.net.

Pilinsky Vladimir Vladimirovich – Candidate of Technical Sciences, Professor, National Technical University of Ukraine «KPI», Kiev, tel.: (095) 276-92-70; e-mail: pww@ukr.net.

УДК 681.5

Н. В. САВЧЕНКО, Д. А. ПАНАСЕНКО**ИНТЕЛЕКТУАЛЬНИЙ МЕТОД ОЦІНКИ СТАНУ ЗДОРОВ'Я ЛЮДИНИ**

На основі побудови формалізованих моделей для захворювань печінки (хронічний гепатит і цироз), описана залежність результуючого діагнозу від результатів обстеження пацієнта, яка реалізована в програмі. Представлені етапи отримання моделей, а також наведені експериментальні вагові коефіцієнти, використовувані при побудові моделі. Всього моделей із зазначеними характеристиками 33.

Ключові слова: хронічний гепатит, цироз, діагноз, математичне моделювання, модель множинної регресії, програмний продукт.

Вступ. В даний час проблеми зі здоров'ям, різного характеру, зустрічаються практично у кожній людини. Отже, необхідно з високою точністю та швидкістю визначати, на що хворий пацієнт, і проводити в найкоротші терміни всі дії, спрямовані на його лікування. Для того щоб коректно поставити діагноз, потрібно проводити ряд діагностик, які на сьогоднішній день не володіють відносно доступними й ефективними технологіями. Не є винятком і діагностика захворювань печінки (хронічний гепатит і цироз). Саме тому завдання вдосконалення методів діагностики різних захворювань печінки як і раніше є актуальною як з точки зору практичних застосувань, так і з позиції теоретичної медичної науки, розвиваючи математичне моделювання різних захворювань.

Виходячи з цього, автоматизована побудови формалізованих моделей для захворювань печінки (хронічний гепатит і цироз).

В якості основного підходу обраний метод регресійного аналізу сукупності лінійних моделей, що пов'язують можливі значення діагнозу з показниками, отриманими на основі обстеження пацієнта. Лінійний регресійний аналіз дозволяє будувати лінійні залежності між наборами пояснювальних змінних та які пояснюють. Використання нелінійних методів поки недоцільно через низьку якість вихідних даних. Використання методів регресійного аналізу для побудови моделей різних процесів, у тому числі і в медицині, не є новим [1]. Повний аналіз даного питання проведено в [2]. Серед нових напрямків щодо використання методів регресійного аналізу в медицині можна вказати

логістичні моделі [2], а також виділимо деякі підходи [3 - 4], що представляють інтерес з точки зору розвитку методів даного дослідження. Дану задачу часто відносять до такого напрямку як – доказова медицина [5, 6].

На відміну від зазначених робіт, дана робота передбачає програмну реалізацію побудови набору моделей, які в сукупності використовуються при виявленні діагнозу.

Це пов'язано з тим, що наявні дані, на даний момент не дозволяють побудувати модель, що задовольняє всім вимогам до неї за достовірністю та надійністю пророкувань діагнозу на її основі. Як вихід з даної ситуації і пропонується побудова набору моделей, використовуючи програму і які є кращими зі складу тих моделей, які можна побудувати на основі даних, якими володіємо. По мірі наповнення бази даних по пацієнтам число моделей в кінцевому наборі буде зменшуватися до тих пір, поки не залишиться одна модель. Також слід розуміти, що в міру розвитку медичної науки і практики з'являються нові діагностичні показники, і в даних по пацієнтам, зібраних до появи цих показників, які відповідають цим показникам графі будуть порожні. Наявність таких перепусток обумовлено, насамперед, тим, що поки не у всіх медичних установах можуть бути виконані всі показники, які перераховані нижче. Таким чином, проблема прогалин у даних буде існувати довго, що вказує на важливість даної роботи.

Аналіз методів діагностики захворювань печінки. Для того щоб побудувати математичні моделі,

які описують залежність діагнозу від вхідних факторів (результати аналізу і попереднього обстеження пацієнта), необхідно оцінити, наскільки наявні дані дозволяють сформулювати бажану модель.

Великі набори спостережень мають ряд істотних недоліків:

1. За багатьма характеристиками для великих груп хворих вихідні дані відсутні по ряду причин. По-перше, збір даних вимагає тривалого періоду часу, а по ряду характеристик такі матеріали стали збиратися відносно нещодавно, тому для більш ранніх спостережень ці дані відсутні. По-друге, діагностичні обстеження хворих можуть проводитися не за всіма характеристиками, тобто набори аналізів в історіях хвороб неповні. Також слід враховувати, що дані можуть формуватися лише на основі карток хворих, тому історії хвороб відсутні.

2. Необхідно розуміти, що дані в ряді випадків недостатньо точні; особливо дані, в основі яких лежать суб'єктивні відчуття пацієнта і думки лікаря.

У зв'язку з вищевказаним виникає необхідність врахування проблеми низької якості вихідних даних у процесі моделювання. Для вирішення даної проблеми можна спробувати доповнити відсутні дані, покладаючись на сучасні математичні методи згладжування та екстраполяції даних. На жаль, даний метод не підходить, тому прогалини в даних можуть бути значними. Інший шлях – побудова цілої серії моделей, що спираються на різні групи з окремих показників, за якими можливо побудувати прийнятні моделі. У міру збільшення обсягу спостережень, отримані моделі будуть уточнюватися, окремі моделі будуть відкидатися зважаючи на появу більш якісних моделей, поки в кінцевому результаті не буде отримана одна чи кілька не покращуваних моделей.

Процедура побудови регресійних моделей.

Нижче представлена загальна процедура побудови регресійних моделей, яка реалізована до захворювань печінки. Процедура складається з наступних етапів:

1. Обчислюються коефіцієнти попарної кореляції між усіма характеристиками, які можуть бути представлені в таблиці, і насамперед коефіцієнти кореляції між діагнозом і рештою характеристик.

2. Зони всіх можливих значень модуля коефіцієнта кореляції (тобто відрізка $[0; 1]$) розбиваються на наступні сім зон:

– від 0 до 0,01, не включаючи праву межу. Це зона, при попаданні в яку коефіцієнти кореляції відповідні характеристики можна вважати практично незалежними. Стосовно до розглянутої задачі це означає недостатній обсяг вихідних даних, що не дозволяє залежності реалізуватися. Застосовуючи до діагностування при достатньо великому обсязі даних це означає, що захворювання практично не залежить від даної характеристики, і немає необхідності проводити лабораторний аналіз хворого за даним показником;

– від 0,01 до 0,1, не включаючи граничне значення. Це зона дуже слабкої залежності характеристик. Висновки в цілому аналогічні попереднім: в цілому показники незалежні. Однак, якщо одна з характеристик – діагноз, то в окремих рідкісних випадках захво-

рування і дана характеристика можуть виявитися залежними. Відкидати дані значення недоцільно, тому при великій кількості слабо корельованих з діагнозом змінних модель регресії, побудована за цим змінним, може виявитися досить якісною;

– від 0,1 до 0,3 (не включаючи) – зона слабкої залежності характеристик, яка типова для багатьох нетехнічних сфер діяльності. Характеристики, що потрапили в цю групу, зазвичай домінують і є основою формування регресійних моделей;

– від 0,0 до 0,7 (не включаючи) – зона середньої залежності. Показники, що потрапили в дану зону, вимагають уважного вивчення, оскільки таких показників зазвичай дуже мало, але вони досить сильно взаємопов'язані з діагнозом. Відзначимо, однак, що, як було зазначено вище, середня і сильна залежність може бути обумовлена також малим об'ємом аналізованих даних, спільних для даної характеристики і діагнозу;

– від 0,7 до 0,9 (не включаючи) – зона підвищеної залежності. Відзначимо, що подібні показники в медицині зустрічаються відносно рідко, оскільки багато захворювань, мабуть, надають неповний, частковий вплив на супутні їм фактори зважаючи на складний, комплексний характер самих захворювань (до подібних захворювань відносяться і захворювання печінки)

– від 0,9 до 0,95 – зона сильної залежності, яка застосовна до медичної сфери, де вимоги до характеру залежностей між змінними менш жорсткі в порівнянні, наприклад, з технічними сферами, відповідні величини зазвичай вважаються практично лінійно залежними, і тому одна з цих величин може бути відкинута. Дане положення (про лінійної залежності) неприйнятне, зокрема, для унікальних поєднань вихідних даних, що відображають специфічний характер залежностей між величинами;

– від 0,95 до 1,0 – зона дуже сильної залежності. У цьому випадку досліджувані величини практично лінійно залежні при будь-яких наборах вихідних даних.

3. Із стовпця, зіставленого для запису діагнозу, вибираються рядки, в яких коефіцієнт кореляції потрапляє в зону середньої і вище залежностей. Чинники, зіставлені цим рядкам, і вибираються в якості пояснюючих параметрів шуканої регресійної моделі.

4. Аналізуються коефіцієнти взаємної кореляції між різними парами базових чинників. Якщо зустрічаються пари чинників з сильною або дуже сильною залежністю, то один з цих чинників виводиться із складу базового набору чинників.

5. Будуються моделі множинної регресії для різних поєднань цих чинників. При побудові моделей необхідно одночасно відмітити виведення для кожної моделі також наступних параметрів: об'єм вибірки, використаної при побудові моделі (беруться тільки ті набори з початкового набору даних, в яких усі поля для усіх аналізованих чинників заповнені (не порожні)); вірогідність довіри до моделі (коефіцієнт значущості моделі), обчислена на основі використання розподілу Стюдента.

6. З отриманого складу моделей виділяються мо-

делі, найбільш ефективні з точки зору коефіцієнта детермінації R^2 : чим ближче коефіцієнт R^2 до одиниці, тим точніше (з меншою помилкою) побудована модель множинної регресії описує діагноз. Найбільш типові порогові значення для R^2 – 0,95, рідше – 0,9.

7. З виділених моделей робиться повторний вибір на основі порівняння коефіцієнтів значущості (вірогідності на основі критерію Фишера), що описують рівень довіри до моделі; а саме наскільки отримані характеристики моделі можуть бути пояснені випадковим поєднанням значень чинників, що входять в модель. Типові порогові значення в переважному числі додатків приймаються рівними 0,05, рідше – 0,1. Виділений в результаті набір моделей і є шуканим.

8. Виписуються рівняння для усіх отриманих моделей множинної регресії в порядку зростання основного (першого) критерію порівняння моделей. Кожне з цих рівнянь може бути використане для визначення діагнозу, що має два варіанти (або перше захворювання, або друге), коли одержувані в кожному конкретному випадку значення змінної, які відповідають діагнозу, замінюються найближчим з чисел – 0 або 1. Тоді значення 0 відповідає першому захворюванню, а значення 1 – другому. Аналогічні процедури будуються і при багатозначному варіанті результуючого діагнозу.

9. В результаті реалізації процедури, описаної в пункті 8, в частині моделей результат може виявитися рівним 0, а для іншої частини – 1. В цьому випадку для отримання результуючого діагнозу може бути використана одна з наступних процедур.

А. В якості результуючого значення Y береться середнє значення, отримане усереднюванням значень y_i , з вагами, пропорційними коефіцієнтам довіри q_i .

$$Y = \frac{\sum_{i=1}^n q_i \cdot y_i}{\sum_{i=1}^n q_i}, \quad (1)$$

де n – число моделей в остаточному переліку; y_i – значення діагнозу в i -й моделі;

$$q_i = 1 - p_i, \quad (2)$$

де p_i – коефіцієнт значимості моделі.

Потім значення Y округлюється до 0 або 1 (аналогічно пункту 8), і отримане значення і визначає остаточний діагноз.

Б. Якщо усі y_i рівні 0 чи 1, то Y дорівнює 0 або 1 відповідно. Якщо ж частина y_i рівні 0, а частина – 1, то призначається додаткове обстеження пацієнта з метою діагностування захворювання або уточнення значень факторів, що входять в вибрані моделі.

Таблиця 1 – Експериментальні вагові коефіцієнти

	b_0	b_1	b_2	b_3	b_5	b_6	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}
Y_1	1,06					-0,071	-0,35						
Y_2	1,2								-0,096		-0,15		
Y_3	0,17											0,05	0,057
Y_4	0,85										-0,19	0,03	
Y_5	0,29										-0,12	0,065	
Y_6	2,14	2,002	-0,013								-0,22		
Y_7	1,6	0,0007	0,011										
Y_8	1,54		-0,007			-0,16							
Y_9	1,65		-0,013										0,0098
Y_{10}	1,82	-0,002	-0,014										0,0023
Y_{11}	1,77	-0,0006	-0,009			-0,15							
Y_{12}	0,58						-0,045	0,0007				0,049	
Y_{13}	1,9		-0,011				-0,05						
Y_{14}	1,26						0,04	0,005	0,04				
Y_{15}	0,75		-0,007									0,055	
Y_{16}	1,56		-0,01					-0,012					
Y_{17}	0,91	-0,02	-0,008									0,057	
Y_{18}	2,2	-0,004	-0,017					-0,008					
Y_{19}	1,5						-0,017	-0,0065	0,041		-0,3		
Y_{20}	1,55				-0,02	-0,16						0,006	
Y_{21}	1,44							-0,007	0,06		-0,38		
Y_{22}	1,76	-0,0004	-0,01		-0,008								
Y_{23}	1,7		-0,009		-0,0084								
Y_{24}	1,008				0,006	-0,05	-0,04						
Y_{25}	1,2						-0,036	-0,004	-0,013				0,06
Y_{26}	1,05						-0,03	0,002	-0,19			0,01	
Y_{27}	0,8							0,001	-0,19			0,0006	
Y_{28}	2,3			0,3				-0,09					
Y_{29}	1,5			-0,5			-0,08						
Y_{30}	0,38			-0,75							0,125		
Y_{31}	0,6			-0,6									
Y_{32}	0,6			-0,6		0,000001							
Y_{33}	0,9	-0,006		-0,8									

Формування набору моделей вибору діагнозу по захворюваннях печінки. Моделі були побудовані з використанням програмної системи EViews 8.0. Загальна кількість таких моделей склало 160 штук.

Аналіз результатів посприяв отримати наступні результати.

Були перебрані різні поєднання вихідних факторів і для кожного з поєднань побудована відповідна модель множинної регресії, а також обчислені її необхідні характеристики: обсяг використаних даних, коефіцієнт детермінації, ймовірність помилки на основі F-статистики. При досить великому числі факторів (більше 4) моделі не виходили через відсутність спільних даних у вибраній сукупності факторів. У результаті розрахунку було побудовано 160 моделей, що охоплюють різні набори пояснюючих показників. За 87 набором показників моделі не відбулися – або значення коефіцієнта детермінації R^2 занадто малі (менше 0,1), або R^2 відсутня зважаючи на тривіальність отриманої моделі. Основна причина цього: для багатьох наборів параметрів зважаючи на малий обсяг загальних вихідних даних значення більшості з цих параметрів постійні на загальних даних. Крім того, є припущення, що вихідні дані являють собою суміш лінійно залежних змінних і накладених на них окремих викидів, відповідних нетиповому перебігу захворювання. На малих обсягах викиди мало ймовірні, і, отже, залежність діагнозу описується лінійною складовою. У міру поповнення вихідних даних ситуація за складом моделей може кардинально змінитися: моделі, які раніше виявилися несуттєвими, можуть опинитися в числі найбільш важливих, і навпаки.

Однак набір даних, який є в даний час призводить до зазначених результатів. Моделі були впорядковані в порядку зростання коефіцієнта r .

Далі представлено рівняння множинної регресії у загальному вигляді, а також для кожної з моделей в табл. 1 наведені вагові коефіцієнти, (позначення представлені після таблиці).

$$Y_i = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_k \cdot X_k. \quad (3)$$

Розшифровка позначень: Y_i – діагноз для i -ї моделі (0 – хронічний гепатит, 1 – цироз), X_1 – загальний білірубін при вступі, мкмоль/л, X_2 – протромбіновий індекс, %, X_3 – етанолів тест (0_1), X_4 – фібриноген В (0_1), X_5 – розмір хвоста підшлункової залози за даними ультразвукового дослідження (УЗД), мм, X_6 – стан шлунку за даними фіброгастроскопії (ФГС) (0 – норма, 1 – гастрит / гастродуоденіт, 3 – ерозії в шлунку або 12-палої кишці, 4 – виразка в шлунку або 12-палої кишці), X_7 – антитіла до Pseudomonasspp., мкг / мл, X_8 – антитіла до Streptococcuspp., мкн / мл, X_9 –

антитіла до Escherichiacoli, мкн / мл, X_{10} – антитіла до мікоплазми класу LgG (0 – ні, 1 – низький титр, 2 – високий титр), X_{11} – антитіла до вірусу гепатиту А класу LgG (0 – ні, 1 – низький титр, 2 – високий титр), X_{12} – активність запалення (1-2-3 ступінь), X_{13} – діаметр селезінкової вени по УЗД, мм, X_{14} – конкременти в жовчному міхурі по УЗД (0 – не має патології, 1 – осад, 2 – камені без ускладнень, 3 – ускладнена жовчнокам'яна хвороба).

Висновки. На основі розроблених моделей, створено програмне забезпечення, що дозволяє проводити оперативну діагностику захворювань печінки.

Список літератури. 1. Медик В. А. Руководство по статистике в медицине и биологии. Т. 2. Прикладная статистика здоровья / В. А. Медик, Б. Б. Фишман, М. С. Токмачев. – М.: Медицина, 2001. – 352 с. 2. Леонов В. П. Статистика в кардиологии. 15 лет спустя / В. П. Леонов // Медицинские технологии. Оценка и выбор. – 2014. – № 1. – С. 17–28. 3. URL: http://www.biometrica.tomsk.ru/logit_1.htm. 4. Bruijne M. H. J. de. On the use of Cox regression in the presence of an irregularly observed time dependent covariate / M. H. J. de Bruijne, S. le Cessie, H. C. Kluine-Nelemas, H. C. van Houwelingen // Statistic in medicine. – 2001. – Vol. 20. No 24. – P. 3817–3829. 5. Marshall G. Prospective prediction in the presence of missing data / G. Marshall, B. Warner, S. MaWhinney, K. Hammermeister // Statistics in Medicine. – 2002. – Vol. 21, No 4. – P. 561–570. 6. Реброва О. Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTIKA / О. Ю. Реброва. – М.: Медиа Сфера, 2002. – 312 с. 7. Петров В. И. Клиническое мышление и доказательная медицина / В. И. Петров. 8. Власов В. В. Применение теории доказательств в медицине / В. В. Власов, Н. Н. Седова.

Bibliography (transliterated): 1. Medik V. A. Rukovodstvo po statistike v meditsine i biologii. T. 2. Prikladnaya statistika zdorov'ya. V. A. Medik, B. B. Fishman, M. S. Tokmachev. M.: Meditsina, 2001. 352 p. 2. Leonov V. P. Statistika v kardiologii. 15 letspustya / V. P. Leonov. Meditsinskie tekhnologii. Otsenka i vybor. 2014. No 1. pp. 17-28. 3. URL:http://www.biometrica.tomsk.ru/logit_1.htm. 4. Bruijne M. H. J. de. On the use of Cox regression in the presence of an irregularly observed time dependent covariate / M. H. J. de Bruijne, S. le Cessie, H. C. Kluine-Nelemas, H. C. van Houwelingen // Statistic in medicine. 2001. Vol. 20. No 24. pp. 3817-3829. 5. Marshall G. Prospective prediction in the presence of missing data / G. Marshall, B. Warner, S. MaWhinney, K. Hammermeister // Statistics in Medicine. 2002. Vol. 21. No 4. pp. 561-570. 6. Rebrova O. Yu. Statisticheskiy analiz meditsinskikh dannykh. Primenenie paketa prikladnykh programm STATISTIKA / O. Yu. Rebrova. M.: MediaSfera, 2002. 312 s. 7. Petrov, V. I. Klinicheskoe myshlenie i dokazatel'naya meditsina / V. I. Petrov. pp.15-32. 8. Vlasov V. V. Primenenie teorii dokazatel'stv v meditsine. V. V. Vlasov, N. N. Sedova.

Надійшла (received) 05.04.2016

Відомості про авторів / Сведения об авторах / About the Authors

Савченко Николай Владимирович – доцент кафедри Системи інформації, кандидат фізико-математических наук, доцент, НТУ «ХПІ», e-mail: nsavchenko77@mail.ru.

Savchenko Nikolay Vladimirovych – Associate Professor at the Department of Information systems, Candidate of Technical Sciences, Docent, NTU "KhPI", e-mail: nsavchenko77@mail.ru.

Панасенко Дмитрий Анатольевич – студент, НТУ «ХПІ», тел. (067) 596-11-82; e-mail: panasenkodmitriy@mail.ru.

Panasenko Dmitriy Anatoltyvych – student, NTU "KhPI", tel. (067) 596-11-82; e-mail: panasenkodmitriy@mail.ru.