УДК 681.3+519.2:616.006-73.916

Использование современных информационных технологий для анализа данных катамнеза больных раком грудной железы

Л. Я. Васильев, Е. Б. Радзишевская, Я. Э. Викман, О. М. Гладкова, В. З. Гертман
Институт медицинской радиологии им. С. П. Григорьева АМН Украины, Харьков

Резюме

В работе представлены некоторые результаты научных исследований, проводимых на массивах данных формализованной медицинской информации с использованием современных информационных технологий. Массивы создавались при помощи специально разработанного программного комплекса, позволяющего накапливать информацию любого вида и объема и выбирать из нее подмассивы любой структуры, пригодные для дальнейшей обработки в стандартных программных средах.

В качестве массива данных рассматривались данные катамнеза больных раком грудной железы I-II стадии. Использование технологии Data Mining позволило выделить показатели, имеющие информационную ценность для формирования отдаленного прогноза течения заболевания. Такими показателями являются индекс массы тела и сторона поражения. Для всестороннего обоснования выдвинутой гипотезы использовались традиционные методы статистики.

Ключевые слова: рак грудной железы, технология Data Mining, информативные показатели.

Клин. информат. и Телемед. 2005. Т.2. №1. с.56–61

Введение

Проблема научного анализа медицинских данных с каждым днем становится все более актуальной, т. к. усложняются задачи, требующие неочевидных решений, найти которые можно лишь после обработки достаточно больших информационных массивов. При этом большое число исследовательских медицинских учреждений хранит в своих архивах огромное количество историй болезней, информация из которых, занесенная в соответствующем виде в компьютер, могла бы стать бесценным источником для выявления новых неочевидных закономерностей и связей.

Сказанное выше, послужило причиной разработки и создания в Институте медицинской радиологии им. С. П. Григорьева АМН Украины автоматизированного комплекса «База данных онкологических больных», предназначенного для хранения и научного анализа медицинской информации.

При планировании к системе были предъявлены следующие требования:

- *Полная формализация хранимой информации.*
- *Гибкая структура,* позволяющая в любой момент времени добавлять новые виды исследований и модифицировать уже существующие без вмешательства в программный код.
- *Возможность выбора информации любого уровня.*
- *Совместимость с существующим программным обеспечением, предназначенным для статистической обработки данных.*

Эксплуатация программного комплекса в течение последних 5 лет позволила провести ряд научно-исследовательских работ, связанных с математической обработкой формализованной медицинской информации.

В частности, была проведена работа по многофакторной оценке эффективности схем комплексного лечения больных раком молочной железы с использованием катамнестических данных. Результаты применения современных информационных технологий, в том числе, технологий поиска скрытых закономерностей — Data Mining — позволили получить некоторые нетривиальные результаты, часть которых приведена в предлагаемой статье.

Материалы и методы

Проводился ретроспективный анализ историй болезни 165 женщин с первичным раком грудной железы (РГЖ) I–II стадии, проживающих в Харькове и Харьковской области и проходивших лечение к клинике ИМР АМНУ в 1993–1994 гг. Данный временной диапазон был выбран с целью проведения анализа 10-летней выживаемости и отдаленных результатов лечения у больных исследуемой группы.

Данные из историй болезни вносились в электронную базу данных с помощью программного комплекса «База данных онкологических больных». Из 228 архивных историй болезни для ввода были отобраны и введены в электронную базу 176, из которых 165 содержали полный объем запланированной для исследования информации и использовались в расчетах.

Для математической обработки результатов исследования использовался программный комплекс WizWhy технологии Data Mining [3] – современной мультидисциплинарной области, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и пр. Одним из методов систем Data Mining являются алгоритмы ограниченного перебора, предназначенные для поиска логических закономерностей в данных.

Для проверки гипотез, выдвинутых при помощи системы WizWhy использовались традиционные методы прикладной статистики для сравнения несвязанных выборок (критерии Манна-Уитни и Колмогорова-Смирнова для анализа количественной вариации и анализ таблиц сопряженности для анализа альтернативной вариации) [3].

Поскольку полученные результаты касаются отдаленных прогнозов развития заболевания (появления отдаленных метастазов) дополнительно проводился анализ выживаемости по Каплану-Майеру [3], который подтвердил полученные выводы.

Результаты исследования

В исследуемой группе возраст больных колебался в пределах от 26 до 81 года с медианой 54 года.

У 32% больных была первая стадия распространенности процесса, и, соответственно, у 68% – вторая. Пораженные регионарные лимфоузлы имели 44% поступивших больных. Всем больным проводилось оперативное вмешательство, причем подавляющему большинству (51%) – радикальная мастэктомия по Пейти, и комплексное лечение по стандартной схеме. Доминирующей гистоформой был признан дольковый, частью протоковый инфильтрирующий рак (48%).

В результате проведенного лечения рецидивы возникли у 15% больных, отдаленные метастазы – у 23%, летальность составила 3% (5 больных).

Срок возникновения отдаленных метастазов колебался от 7,2 месяцев до 107,7 месяцев (медиана 34 месяца).

На начальных этапах исследования применительно к данным, содержащих полную информацию о течении заболевания у больных РГЖ I–II стадии был проведен поиск неочевидных закономерностей методом ограниченного перебора, предназначенным для поиска логических закономерностей, в указанном массиве данных. В результате была выдвинута гипотеза о зависимости наличия/отсутствия отдаленных метастазов от *индекса массы тела по Кетле* (ИМТ), рассчитываемого по формуле:

$$\text{ИМТ} = \frac{\text{вес (в кг)}}{\text{рост}^2\text{(в метрах)}}$$

Считается, что нормальное соотношение веса тела и роста характеризуется ИМТ<25, интервал 25–30 характеризует 1-ю степень ожирения, ИМТ>30 – 2-ю. Выяснилось, что у больных с ИМТ выше нормы отдаленные метастазы не появляются с вероятностью P=0,897. Особенно очевидной данная закономерность является в подгруппе больных с туморотрицательными аксиллярными лимфоузлами. В этом случае вероятность благоприятного исхода составляет P=1,00.

Для всестороннего изучения данного утверждения были использованы традиционные методы статистического анализа, результаты которого приведены ниже. В частности, факт достоверности отличий между группами больных с отдаленными метастазами и без них по признаку «ИМТ» подтверждается критерием Манна-Уитни высоким уровнем значимости (p<0,001).

Из приведенной табл.1 видно, что в группе без отдаленных метастазов наибольший процент пациенток (42,5%) имели ИМТ выше нормы (1-ая стадия ожирения), а в группе с отдаленными метастазами наибольшее количество больных (52,78%) имели ИМТ ниже нормы.

В подгруппе больных *без первичных метастазов в регионарные лимфоузы* ситуация является еще более явной:

Табл. 1. Распределение больных по признаку наличия/отсутствия отдаленных метастазов в зависимости от индекса массы Кетле.

ИМТ		Без отдаленных метастазов	С отдаленными метастазами	Всего
Ниже нормы	абсолютное количество	32.00	19.00	51.00
	процент по столбику	26.67	52.77	32.69
Норма	абсолютное количество	36.00	11.00	47.00
	процент по столбику	30.00	30.56	30.13
Выше нормы	абсолютное количество	51.00	6.00	57.00
	процент по столбику	42.50	16.67	36.54
Значительно выше нормы	абсолютное количество	1.00	0.00	1.00
	процент по столбику	0.83	0.00	0.64
Всего	абсолютное количество	120.00	36.00	156.00
	процент по столбику	100.00	100.00	100.00

Табл. 2. Распределение больных по признаку наличия/отсутствия отдаленных метастазов в зависимости от ИМТ у больных без первичных метастазов в регионарные лимфоузлы.

ИМТ		Без отдаленных метастазов	С отдаленными метастазами	Всего
Ниже нормы	абсолютное количество	23.00	11.00	34.00
	процент по столбику	31.94	73.33	39.08
Норма	абсолютное количество	18.00	4.00	22.00
	процент по столбику	25.00	26.67	25.29
Выше нормы	абсолютное количество	30.00	0.00	30.00
	процент по столбику	41.67	0.00	34.48
Значительно выше нормы	абсолютное количество	1.00	0.00	1.00
	процент по столбику	1.39	0.00	1.15
Всего	абсолютное количество	72.00	15.00	87.00
	процент по столбику	100.00	100.00	100.00

Табл. 3. Распределение больных по признаку наличия/отсутствия отдаленных метастазов в зависимости от стороны поражения.

Признак наявности/отсутсвия отдаленных метастазов		Правая сторона	Левая сторона	Всего
Без отдаленных метастазов	абсолютное количество	50.00	70.00	120.00
	процент по столбику	69.44	83.33	76.92
С отдаленными метастазами	абсолютное количество	22.00	14.00	36.00
	процент по столбику	30.56	16.67	23.08
Всего	абсолютное количество	72.00	84.00	156.00
	процент по столбику	100.00	100.00	100.00

ни одна из пациенток с отдаленными метастазами не страдала избыточной массой тела (табл. 2).

Для исключения влияния неоднородности по возрастному фактору и по частоте встречаемости первичных метастазов в регионарные лимфоузлы в подгруппах с ИМТ в норме и ниже нормы (подгруппа ИМТ−) и с ИМТ выше нормы (подгруппа ИМТ+) был проведен дополнительный анализ.

В подгруппе ИМТ− возраст больных колебался в пределах 29−79 лет с медианой 52 года; в подгруппе ИМТ+ медиана составила 58,5 лет в возрастных пределах от 26 лет до 81 года. Таким образом, принципиальных различий по возрастному признаку не наблюдалось. Статистические критерии для сравнения двух несвязанных выборок также не обнаружили достоверных различий по возрасту между группами.

Таким образом, нельзя сказать, что ИМТ достоверно зависит от возраста больной, что, учитывая повышенный риск неблагоприятного течения заболевания у лиц молодого возраста, могло бы быть причиной связи между ИМТ и отдаленными метастазами.

Следующим этапом исследования была проверка однородности изначального статуса подгрупп ИМТ+ и ИМТ− по частоте встречаемости первичных метастазов в регионарные лимфоузлы. Оказалось, что в подгруппе ИМТ− без метастазов в регионарные лимфоузлы было 57,14% больных (с метастазами − 42,86%), а в подгруппе ИМТ+ без метастазов − 53,45%, с метастазами − 46,55.

Существующее различие в частотах не является статистически значимым, что не дает основания предполагать наличие более благоприятного изначального статуса у больных группы ИМТ+ по сравнению с больными группы ИМТ−.

Обнаруженную зависимость нельзя также объяснить различием в распределении по стадиям распространенности процесса, т.к. среди больных с повышенной массой тела первую стадию имели лишь 10,3% всех больных.

Таким образом, проведенный анализ дает основания считать, что у больных с повышенным индексом массы тела вероятность возникновения отделенных метастазов ниже, чем у больных с индексом массы тела по Кетле в норме и ниже нормы. Отсутствие первичных метастазов в регионарные лимфоузлы у больных подгруппы ИМТ+ еще больше усиливает благоприятный прогноз.

Числовой мерой обнаруженной закономерности являются показатели инцидентности. Так, в подгруппе ИМТ− показатель инцидентности (отношения количества больных с отдаленными метастазами к общей численности дан-

ной группы, т.е. риск развития отдаленных метастазов) составил 0,31; в подгруппе ИМТ+ инцидентность составляет 0,12, а в подгруппе ИМТ+ без первичных метастазов в регионарные лимфоузлы – еще ниже (0,03).

В результате применения технологии Data Mining была обнаружена также взаимосвязь между появлением отдаленных метастазов и *стороной поражения*.

При левой стороне поражения вероятность благоприятного исхода составляет P=0,833. Соответствующие данные представлены в таблице 3.

Анализ таблиц сопряженности подтвердил высокую значимость полученного результата (p=0,03 – критерий χ–квадрат).

Анализ групп на однородность не обнаружил достоверной зависимости стороны поражения от признаков «возраст», и «первичное метастазирование в регионарные лимфоузлы», что видно также и из приведенных таблиц 4 и 5.

Результаты, приведенные в табл.4 не дают основания считать, что больные с правой стороной поражения имели менее «выгодный» возрастной ценз, то есть, были моложе альтернативной подгруппы.

Даже в вызывающем сомнения возрастном диапазоне 40–50 лет, где левая сторона поражения наблюдалась значительно реже, а, следовательно, меньшим по абсолютной численности должно было быть и количество отдаленных метастазов, соотношение исходов было следующим. При правой стороне поражения отдаленные метастазы не появились у 15, а появились у 9; при левой – не появились у 13, появились у 3.

По данным, приведенным в таблице, нельзя сказать, что больные с левой стороной поражения имели принципиально более низкий уровень поражения регионарных лимфоузлов.

С точки зрения однородности вызывал сомнения фактор стадийности, т.к. если у больных со II-ой стадией процесса левая сторона была поражена у 52%, а правая – у 48% (т.е., практически, пополам), то у больных с I-ой стадией левая сторона поражена у 66%. Таким образом 68% больных с левой стороной поражения имели I-ю стадию процесса, что могло послужить причиной более благоприятного исхода с точки зрения возникновения отдаленных метастазов для этих больных.

Однако, более детальный анализ показал, что даже при II-ой стадии процесса при левой стороне поражения благоприятный исход наблюдается у 80,4% больных, в то время как при правой стороне – у 65,5%. На приведенных ниже диаграммах показана зависимость появления отделенных метастазов от стороны поражения при I и II стадиях РГЖ (рис. 1).

Табл. 4. Распределение больных по признаку «сторона поражения» в зависимости от возраста больных.

Возраст больных, лет		Правая сторона	Левая сторона	Всего
до 40	абсолютное количество	9.00	10.00	19.00
	процент по стороне поражения	12.50	11.90	12.18
40-50	абсолютное количество	24.00	16.00	40.00
	процент по стороне поражения	33.33	19.05	25.64
51-60	абсолютное количество	18.00	31.00	49.00
	процент по стороне поражения	25.00	36.90	31.41
Старше 60	абсолютное количество	21.00	27.00	48.00
	процент по стороне поражения	29.17	32.15	30.77
Всего	абсолютное количество	72.00	84.00	156.00
	процент по стороне поражения	100.00	100.00	100.00

Табл. 5. Распределение больных по признаку «сторона поражения» в зависимости от первичного метастазирования в регионарные лимфоузлы.

Сторона поражения	Без первичного метастазирования		С первичным метастазированием		Всего	
	абсолютное количество	процент по стороне поражения	абсолютное количество	процент по стороне поражения	абсолютное количество	процент по стороне поражения
Правая	38.00	52.78	34.00	47.22	72.00	100.00
Правая	49.00	58.33	35.00	41.67	84.00	100.00
Всего	87.00	55.77	69.00	44.23	156.00	100.00

Нами также была проведена обработка данных для выявления возможных различий, связанных с локализацией процесса (квадрант локализации). По данному признаку выборка также оказалась однородной.

Таким образом, проведенный анализ дает основания считать, что у больных РГЖ I–II стадии с левой стороной поражения вероятность возникновения отдаленных метастазов ниже, чем у больных с пораженной правой стороной. Шанс возникновения отдаленных метастазов (показатель инцидентности) при левой стороне поражения составил 0,16, в то время как при правой – 0,32, таким образом, по нашим данным, при левой стороне поражения шансы возникно-

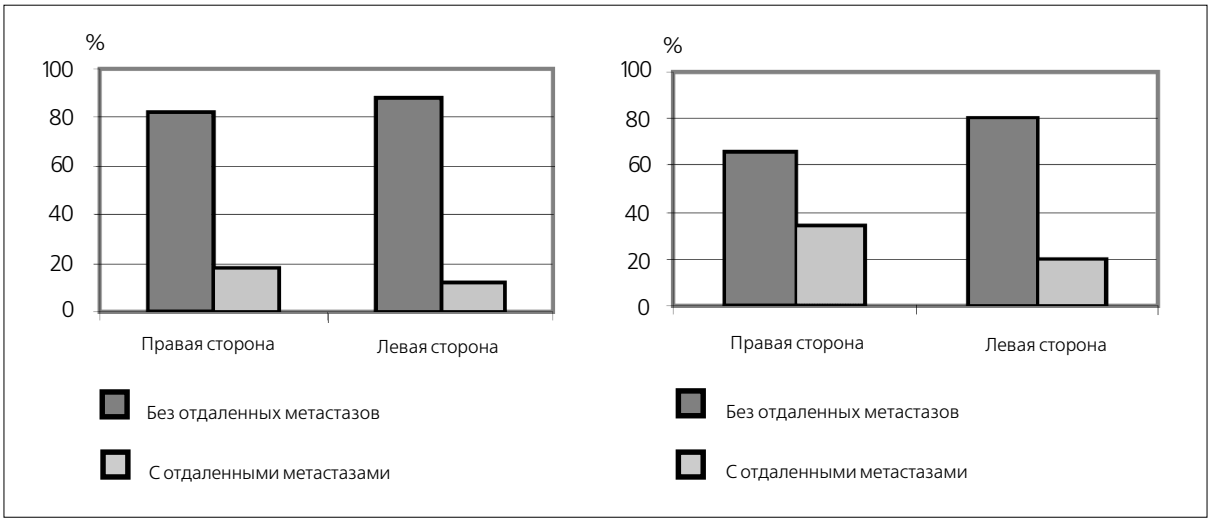

Рис.1. Зависимость появления отдаленных метастазов от стороны поражения при первой (слева) и при второй (справа) стадиях распространения процесса.

вения отдаленных метастазов в два раза ниже, чем при правой.

При совместном выполнении условий (повышенной индекс массы тела по Кетле и левая сторона поражения) вероятность благоприятного исхода повышается еще больше: из 33 больных отдаленные метастазы возникли у 3. Показатель инцидентности (показатель риска) при этом составил 0,09, а шансы возникновения отдаленных метастазов в 4,4 раза ниже, чем в остальной группе. Если добавляется условие отсутствия метастазов в регионарные лимфоузлы, то инцидентность снижается до 0,05, а шансы благоприятного исхода повышаются в 8 раз.

Выводы

Таким образом, использование современных информационных технологий, в том числе, технологий поиска скрытых закономерностей Data Mining, позволили выделить из массива данных катамнеза больных раком грудной железы I–II стадий показатели, которые являются диагностически информативными с точки зрения возникновения отдаленных метастазов. Такими показателями являются индекс массы тела по Кетле и сторона поражения. Показано, что вероятность благоприятного исхода заболевания повышается при повышенных значениях индекса массы тела и при левой стороне поражения.

Литература

1. Алфимов А. Е. Статистика и клинические исследования в онкологии // Матер. VII рос. онколог. конгресса – М., 2003. – С. 11–14.
2. Боровиков В. П., Боровиков И. П. «STATISTICA» – Статистический анализ и обработка данных в среде Windows. – М.: Филин, 1998. – 608 с.
3. Бююль Ахим, Цефель Петер Б. 89 SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. – СПб.: ДиаСофтЮП, 2001. – 608 с.
4. Генкин А. А. Новая информационная технология анализа медицинских данных программный комплекс ОМИС). – СПб.: Политехника, 1999. – 191 с.
5. Гланц Стивен. Медико-биологическая статистика / Пер. с англ. – М.: Практика, 1999. – 580 с.
6. Захарцева Л. М., Дроздов В. М., Нейман А. М. Определение прогностических факторов рака молочной железы // Матер. наук.-практ. конф. з міжнар. участю «Онкологія–XXI». – К., 2003. – С. 85–86.
7. Лищишина Е. М. Смертность от злокачественных новообразований в Украине: анализ избранных динамических кривых // Лікарська справа. – 1996. – № 10–12. – С. 161–163.
8. Марценюк В. П., Кравець Н. О. Медична інформатика. Методи системного аналізу. – Тернополь: Укрмедкнига, 2002. – 177 с.
9. Орлов А. И. О современных проблемах внедрения прикладной статистики и других статистических методов // Заводская лаборатория. – 1992. – № 1. – С. 67–72.

Usage of modern technologies for analysis of catamnesis data in patients with breast cancer

L. Vasilyev, E. Radzishevska, Y. Vikman, O. Gladkova, V. Gertman
Grigorev Institute for Medical Radiology, Kharkiv, Ukraine

Abstract

In work the results of investigations on data arrays of the formalized medical information with usage of modern information technologies are presented. The arrays formed with the help of a specially designed program complex permitting to store the information of any kind and volume and to select from subarrays of any structure, suitable for further processing in standard software.

As a data array the data of a catamnesis of the patients with breast cancer of I-II stages were considered. Usage of Data Mining technology has allowed to reveal the parameters having information value for forming of disease prognosis. Such parameters are mass index and side of lesion. For substantiation of the hypothesis the traditional methods of statistics were used.

Keywords: breast cancer, Data Mining technology, informative parameters.

Використання сучасних інформаційних технологій для аналізу даних хворих на рак грудної залози

Л.Я. Васильєв, Є.Б. Радзішевська, Я.Е. Вікман, О.М. Гладкова, В.З. Гертман
Институт медичної радіології ім. С. П. Григорьєва АМН України, Харків

Резюме

В роботі наведені деякі результати наукових досліджень, що проводились на масивах даних формалізованої медичної інформації з використанням сучасних інформаційних технологій. Масиви створювались за допомогою спеціально розробленого програмного комплексу, який дозволяє накопичувати інформацію будь-якого виду й обсягу та вибирати з неї підмасиви будь-якої структури, придатні для подальшої обробки в стандартних програмних середовищах.

В якості масиву даних було розглянуто дані катамнезу хворих на рак грудної залози I-II стадії. Використання технології Data Mining дозволило виділити показники, що мають інформаційну цінність для формування віддаленого прогнозу перебігу захворювання. Такими показниками є індекс маси тіла і сторона ураження. Для всебічного обґрунтування висунутої гіпотези було використано традиційні методи статистики.

Ключові слова: рак грудної залози, технологія Data Mining, інформативні показники.

Переписка

к. физ./мат. н. **Е. Б. Радзишевская**
Институт медицинской радиологии им. С. П. Григорьева АМН Украины
ул. Пушкинская, 82
Харьков, 61064, Украина
тел. +38 (057) 704-10-63
эл. почта: radz@kharkov.ua