УДК 61:002.51.6:681.31:51:31

СУЧАСНІ ТЕХНОЛОГІЇ ОБРОБКИ Й АНАЛІЗУ МЕДИЧНИХ ДАНИХ

О.В. Гойко

Національна медична академія післядипломної освіти імені П.Л.Шупика

У статті висвітлюється проблема наукового аналізу медичних даних з використанням сучасних технологій. Даються деякі рекомендації щодо вибору методу обробки та програмного забезпечення, підготовки даних для комп'ютерного аналізу та обробки даних медичних спостережень.

Ключові слова: комп'ютерний аналіз, сучасні технології аналізу даних, математичні методи обробки, пакети прикладних програм.

СОВРЕМЕННЫЕ ТЕХНОЛОГИИ ОБРАБОТКИ И АНАЛИЗА МЕДИЦИНСКИХ ДАННЫХ

О.В. Гойко

Национальная медицинская академия последипломного образования имени П.Л. Шупика

В статье освещается проблема научного анализа медицинских данных с использованием современных технологий. Даются некоторые рекомендации относительно выбора метода обработки и программного обеспечения, подготовки данных для компьютерного анализа и обработки данных медицинских наблюдений, а также представления и интерпретации полученных результатов.

Ключевые слова: компьютерный анализ, современные технологии анализа данных, математические методы обработки, пакеты прикладных программ.

MODERN TECHNOLOGIES PROCESSING AND ANALYSIS OF MEDICAL DATA

Gojko O.V.

National Medical Academy of Postgraduate Education named after P.L.Shupyk

In the article the problem of scientific analysis of medical data lights up with the use of modern technologies. Some recommendations are given in relation to the choice of method of treatment and software, preparation of information for a computer analysis and treatment of these medical supervisions.

Key words: computer analysis, advanced data mining technology, mathematical methods of processing, application packages.

Вступ. Проблема наукового аналізу медичних даних з кожним днем стає все більш актуальною, оскільки питання, що виникають у лікарів при прийнятті правильного рішення, стають все складнішими. Знайти ці рішення можна лише після обробки й аналізу достатньо великих інформаційних масивів.

В основі обробки й аналізу даних лежать математичні методи, що здебільшого є незмінними вже протягом багатьох десятиліть. Відповідно незмінними залишаються й загальні принципи та послідовність дій при обробці даних, проте технологія обробки даних міняється, і досить істотно. Листок паперу, олівець та калькулятор, якими користувалися раніше, відійшли в минуле, їм на зміну приходять нові сучасні технічні засоби, які удосконалюються надзвичайно швидкими темпами. В останні роки обробка й аналіз будь-якої інформації стають не просто неможливими, а й недопустимими без використання комп'ютерів і відповідного сучасного програмного забезпечення.

Комп'ютерний аналіз медичних даних припускає деяке математичне перетворення даних за допомогою певних програмних засобів, а отже користувачу необхідно мати уявлення не лише про математичні методи обробки даних, а й про відповідні програмні засоби.

Незважаючи на те, що статистичні методи за останні півстоліття істотно не змінилися, завдяки використанню комп'ютерів значно розширилося коло застосовування цих методів.

Разом з тим, відповідне програмне забезпечення за цей час істотно змінилося. Зі зміною поколінь ЕОМ мінялися й покоління програмних засобів обробки даних. І якщо можливості перших ЕОМ з аналізу даних не перевершували можливості сучасних середніх калькуляторів, то в 70-і роки вже з'явилися пакети прикладних програм, які містили практично всі ті математичні методи обробки, які входять до складу й сучасних пакетів (SSP, BMDP, SPSS, Statistica тощо). Подальший розвиток пакетів здійснювався шляхом вдосконалювання технології й аналізу (табл. 1).

Таблиця 1. Хронологія розвитку пакетів прикладних програм для обробки й аналізу даних

Роки	Основні пакети аналізу даних	Операційні системи
1970-1985	SSP, BMDP, SAS	
1985-1995	Statgraphics, STATA, SAS, Systat, STADIA, МЕЗОЗАВР, САНІ, Евріста, Клас-майстер тощо	DOS
1995-2009	Statgraphics Plus, SAS, SPSS, Statistica, Excel тощо	Windows

Удосконалювання технічних засобів призводить до зміни технології обробки даних. У ті порівняно недавні часи, коли обробка даних здійснювалася вручну, найбільш трудомістким процесом був етап самих статистичних обчислень і розрахунків за різними формулами. І цілком природно, що на цьому етапі була зосереджена увага фахівців, а тому пропонувалися різні спрощені варіанти розрахунків, більш прості методи, спеціально пристосовані для ручного рахунку тощо. Потім, з появою перших комп'ютерних пакетів, технологія ґрунтувалася на принципі командного рядка й вимагала досить пристойних знань не лише статистики, а й володіння комп'ютером на рівні програміста. Далі розвиток пішов шляхом використання меню готових процедур, що різко знизило вимоги як до знання статистики, так і до володіння комп'ютером. Останнім часом продовжується поліпшення інтерфейсу з користувачем, активніше використовується графічний підхід, особливо важливого значення набуває візуалізація даних, що ще більше полегшує обробку даних.

Завдяки використанню комп'ютерів, обчислювальний етап став найменш трудомістким, полегшилися й інші етапи обробки даних. На перше місце по відносній трудомісткості вийшли такі етапи, як: освоєння статистичного пакета, підготовка даних до аналізу, попередній аналіз даних; інтерпретація результатів. Все це в цілому привело до зміни технології обробки й аналізу даних. При цьому для застосування основних методів обробки даних від виконавця потрібно лише виконання певних статистичних правил і грамотне використання обраного ним пакета. Лікареві не потрібно заглиблюватися в складність математичних визначень, а варто лише зрозуміти, для чого і як ці методи використовуються.

На практиці для лікаря на сьогоднішній день обробка й аналіз даних зводиться до вирішення наступних завдань:

• мати уявлення про основні статистичні методи;

• освоїти відповідний пакет, що буде використовуватися для аналізу й обробки даних;

• провести власне цей аналіз та вміти правильно інтерпретувати отримані результати.

Що стосується першого завдання, а саме - уявлення про основні статистичні методи, то слід нагадати, що саме від коректності й грамотності застосування статистичних методів залежить правильність зроблених висновків і відповідно об'єктивність в прийнятті рішення.

Довгий час обробка й аналіз медичних даних залишалися привілеєм фахівців, тому що глибоке розуміння сучасних методів аналізу даних вимагає серйозної математичної підготовки. Ідеальним варіантом є випадок, коли людина, добре знаючи математичну статистику, застосовує комп'ютерні методи для аналізу своїх даних. Однак для того, щоб глибоко знати статистичні методи, необхідна спеціальна математична підготовка в обсязі вузівського курсу, що нереально для лікаря, оскільки підготовка в галузі прикладної статистики у медичних вузах для лікарів-дослідників явно недостатня. Звичайно, ці прогалини слід заповнити і необхідні уявлення про основні статистичні методи медичним працівникам необхідно мати. Не дивлячись на те, що є чимало підручників [1-3, 5, 8, 10, 11], в яких описані статистичні методи, на нашу думку, найбільш вдало цей матеріал викладено у навчальному посібнику Мінцера О.П. і співавт. "Оброблення клінічних і експериментальних даних у медицині", що виданий у 2003 році [14]. В ньому досить повно і детально і, найголовніше, зрозумілою для медиків мовою викладені різні аспекти обробки й аналізу медичних даних, наводяться медичні приклади використання тих або інших методів.

Основні математичні методи обробки й аналізу даних, які найчастіше використовуються при типових медичних дослідженнях, подаються у табл. 2.

Таблиця 2. Основні математичні методи обробки й аналізу даних, які використовуються при типових медичних дослідженнях

Джерело інформації, завдання дослідження	Методи обробки й аналізу, які використовуються для реалізації поставлених завдань
Опитувальники, анкети, тести Обстеження стану здоров'я; думка того, кого обстежують; історії хвороби Виявлення прихованих взаємозв'язків	Кореляційний аналіз, метод автоматичної класифікації тощо Факторний аналіз
Скринінгові обстеження	Дискримінантний аналіз, кластерний аналіз, методи розпізнавання образів
Історії хвороби Клінічні обстеження, лікування та реабілітація хворих; ускладнення при лікуванні. Дослідження ефективності різних процедур, вивчення зв'язків між процедурами та їхніми наслідками	Кореляційний аналіз, дисперсійний аналіз, регресійний аналіз Методи оцінювання гіпотез, регресійний аналіз
Медико-статистичні дані Дослідження захворюваності, динаміка захворюваності, виявлення періодичності захворюваності	Методи аналізу випадкових процесів, спектральний аналіз, математичне моделювання
Експерименти Лабораторні експерименти й досліди на тваринах при заданих умовах	Методи планування експериментів, регресійний аналіз, дисперсійний аналіз, багатомірний статистичний аналіз, методи математичного моделювання
Клінічні дослідження Порівняльні лікувальні дослідження, аналіз виживаності й спадковості з урахуванням належності пацієнта до певної групи, вивчення дозування препаратів Розробка методів діагностики	Дисперсійний аналіз, регресійний аналіз, дискримінантний аналіз, методи оцінювання гіпотез Дискримінантний аналіз, кластерний аналіз, методи розпізнавання образів
Дослідження механізмів дії патогенних факторів	Дисперсійний аналіз, регресійний аналіз, методи математичного моделювання
Клінічні лабораторні дані Зберігання, збір і передача клінічної інформації, аналіз якості й надійності лабораторних досліджень, догляду за пацієнтами	Дисперсійний аналіз, регресійний аналіз

Однак, досить часто заздалегідь важко визначити, який метод дасть найкращий результат. Тому варто передбачити можливість застосування різних способів обробки даних, що при використанні комп'ютерного аналізу стає набагато простішим.

Що стосується другого завдання, а саме - освоєння пакета аналізу, за допомогою якого буде здійснюватися обробка й аналіз даних, то слід зазначити, що це один з найбільш трудомістких етапів обробки даних. Сьогодні вже ніхто не проводить статистичний аналіз експериментальних даних вручну, зараз з цією метою використовуються різноманітні комп'ютерні пакети прикладних програм.

Більшість комп'ютерних статистичних програм не є чисто медичними прикладними програмами, оскільки більшість методів статистичного аналізу є універсальними й можуть застосовуватися не лише в різних галузях медичної статистики, але й у найрізноманітніших галузях людської діяльності. Наприклад, з погляду формальної логіки статистичний прогноз інфекційної захворюваності й прогноз курсу долара - це та ж сама задача, а тому вона може вирішуватися за допомогою одних і тих же пакетів прикладних програм.

На сьогоднішній день число пакетів для обробки інформації досягає кількох десятків, серед яких зарубіжні пакети, такі, як: SYSTAT, STATGRAPHICS, BMDP, SPSS, SAS, CSS, Statistica, а також вітчизняні: STADIA, ЕВРІСТА, МЕЗОЗАВР, САНІ, КЛАСС-МАСТЕР, СІГАМД тощо.

Основну частину наявних пакетів для обробки даних можна віднести до трьох категорій: спеціалізовані пакети, пакети загального призначення і професійні пакети.

Спеціалізовані пакети, як правило, містять методи з одного-двох розділів статистики або методи, що

використовуються в конкретній предметній галузі (наприклад, *Мезозавр*—програма аналізу часових рядів). Спеціалізовані пакети застосовуються для вирішення вузького кола завдань з використанням спеціальних методів статистичного аналізу. Експлуатація цих програм вимагає високого рівня підготовки користувача в галузях певних розділів статистики.

Пакети загального призначення або універсальні (Statistica, SPSS, Діастат, STADIA, STATGRAPHICS, SYSTAT, Excel) є найбільш зручними для користувача-початківця завдяки відсутності орієнтації на специфічну предметну галузь, широкому діапазону статистичних методів і дружньому інтерфейсу користувача. Вони більш доступні для практики й можуть використовуватися широким колом фахівців різного профілю. Практично всі задачі, що стосуються обробки й аналізу медико-біологічних досліджень, можуть бути вирішені за допомогою універсальних пакетів, зокрема Statistica [4, 5] та Excel [12].

Професійні пакети призначені для користувачів, які мають справу із надзвичайно великими обсягами даних або вузькоспеціалізованими методами аналізу.

Особливістю будь-якого пакета статистичних програм є видача великої кількості інформації, що описує результат статистичного аналізу. Практично всі статистичні пакети забезпечують широкий набір засобів візуалізації даних: побудова графіків, дво- і тривимірних діаграм, а часто і різноманітні засоби ділової графіки.

Характеристики основних пакетів для обробки й аналізу даних наведені у табл. 3.

Таблиця 3. Характеристики основних статистичних пакетів

Характеристика	Statgraphics Plus	SPSS	Statistica	Excel
Фірма	Manugistics	SPSS	StatSoft	Microsoft
Версія	3, 2	15 ,0	6,0	2007
Рік розробки	2003	2007	2007	2007
Рік 1 версії	1983	1975	1990	1996
Обсяг пакета, МБ	14,5	26,3	16,3	
Доступність	4	3	2	1
Русифікованість	_	-/+	-/+	+
Число процедур	>250	>250	>250	19
Простота освоєння	3	4	2	1
Література	+	—	+	+
Навчання на кафедрі	—	—	+	+
Зручність роботи	2	4	3	1
Візуалізація	2	3	1	4

Цифри 1-4 у таблиці відображають експертну оцінку переваг одного пакета перед іншим (1 - вищий ступінь).

Так, пакет Statgraphics розроблявся для роботи в середовищі DOS, а потім був адаптований до операційної системи Windows і отримав нову назву Statgraphics Plus. За своїми характеристиками пакет займає проміжне місце між SPSS і Statistica.

Пакет SPSS створювався ще для "великих" машин і послідовно переводився для роботи в середовищі DOS, а потім Windows. Пакет добре відпрацьований, наближається за своїми можливостями до професійних пакетів, і реалізація статистичних процедур добре пристосована до практичної роботи.

Пакет Statistica спеціально створювався для роботи в середовищі Windows. Відрізняється найбільш розвиненим інтерфейсом і багатими графічними можливостями.

Електронна таблиця Excel найбільш поширена і, як правило, використовується при найпростішому статистичному аналізі даних. Важливою перевагою пакета Excel є його русифікованість, а також доступність, оскільки він встановлюється автоматично при інсталяції пакета MS Office. Тому пакет Excel найчастіше використовується при оформленні результатів роботи.

Слід зазначити, що всі ці пакети постійно оновлюються і з кожним роком з'являються їх нові версії.

При виборі пакета для аналізу даних можна виділити два аспекти: а) початковий вибір пакета аналізу; б) поточний вибір при переході на більш сучасний, більш потужний пакет. Підходи в обох випадках дещо відрізняються.

У першому випадку на вибір накладаються такі обмеження:

1. Можливості комп'ютера.

2. Можливості одержання установчої версії пакета.
3. Характеристики пакета.

По першому пункту - варто вибирати найсучасніші версії пакетів із тих, що можуть бути встановлені на наявний комп'ютер. Другий пункт очевидний - вибирати можна з тих пакетів, що доступні. Що стосується характеристик пакета, то тут варто розглянути такі аспекти: а) обчислювальні можливості, б) зручність роботи, в) складність освоєння.

Обчислювальні можливості. У випадку, коли необхідно обробляти медичні дані помірних обсягів (до декількох тисяч спостережень) стандартними статистичними методами, найкраще використовувати універсальні пакети. Якщо дивитися з позицій лікаря-дослідника, то всі сучасні універсальні статистичні пакети за своїми обчислювальними можливостями повністю відповідають можливим потребам (Statistica, SPSS, SAS, Statgraphics Plus, Systat та інші пакети, що працюють в операційній системі Windows). Проте завжди слід переконатися, що обраний пакет містить необхідні методи обробки.

Зручність роботи. Всі сучасні пакети досить зручні в роботі (коли вони вже освоєні).

Складність освоєння. За складністю освоєння пакети дещо розрізняються і тут варто віддати перевагу русифікованим пакетам або пакетам, з яких є доступна література або є можливість пройти курс навчання.

Варто зауважити, що без крайньої необхідності (неможливість забезпечити необхідну обробку даних) не бажано змінювати обраний і освоєний пакет аналізу, тому що це призведе до значного збільшення витрат праці.

Що стосується заміни пакета на більш сучасну версію, то тут є дві крайності:

1. Прагнення до постійного відновлення, установки найостанніших версій пакетів - як правило віднімає багато сил, не дозволяє виробитися корисним стереотипам дій, у той же час не приводить до суттєвого зростання можливостей.

2. З іншого, боку уподобання до застарілих пакетів найчастіше не дозволяє повною мірою використовувати можливості сучасної техніки і програмного забезпечення. Існує деякий емпіричний оптимум, що може визначатися зразковим терміном експлуатації пакета в 2-3 роки, після закінчення котрого доцільно здійснювати перехід до більш сучасних пакетів. При цьому перевагу слід надавати черговій версії того ж пакета, що використовувався раніше (наприклад, DOS-версію пакета Statgraphics можна замінити на версію пакета Statgraphics Plus для Windows). Спадкоємність значно полегшує процес освоєння пакета.

Аналіз даних з використанням статистичного пакета (робота з пакетом, власне технологія аналізу даних) включає наступні розділи:

1. Планування дослідження.
2. Підготовка даних до аналізу.
3. Вибір методу аналізу і його реалізація.
4. Інтерпретація та подання результатів.

Планування дослідження. Найкращим випадком є такий, коли ще до проведення дослідження існує певна ясність щодо передбачуваних до використання надалі методів обробки даних. У цьому випадку, як правило, вдається спланувати дослідження з урахуванням наступної обробки даних і уникнути ситуацій, коли виявляється, що якісь спостереження були зайвими, а якихось не вистачає для реалізації обраних методів аналізу.

Підготовка даних до аналізу. Це вкрай важливий і найчастіше недооцінюваний етап роботи. Сучасна технологія обробки даних починається саме з етапу підготовки даних до аналізу, метою якого є приведення даних до виду, що дозволяє провести наступну їхню обробку. Як правило, він включає: занесення даних, попереднє перетворення даних, візуалізацію даних з метою формування подання про досліджуваний матеріал.

Звичайно при проведенні медичного дослідження прагнуть урахувати максимальну кількість характеристик, які відіграють важливу роль при аналізі досліджуваного питання. При цьому дослідження складається з декількох серій спостережень, при яких в ідентичних умовах реєструються параметри окремих об'єктів (наприклад, хворих з якимось певним захворюванням). Маючи справу із серією спостережень, варто прагнути виразити їх у простій формі, що дозволила б безпосередньо або шляхом наступних обчислень зробити з них висновки.

Всі дані доцільно звести в єдину таблицю, в якій по рядках розташовані різні об'єкти спостереження (наприклад, хворі), а по стовпцях - параметри (наприклад, температура, частота серцевих скорочень, артеріальний тиск тощо). Всередині цієї таблиці об'єкти можуть бути об'єднані в кілька груп відповідно до загальної ознаки (наприклад, за віком, за патологією тощо). Однак на сьогоднішній день практично відпадає необхідність у попередньому структуруванні, побудові необхідних вибірок, ранжируванні тощо. Всі ці завдання в сучасних пакетах автоматизовані й виконуються безпосередньо при реалізації вибраного методу аналізу. На етапі підготовки даних дуже важливим моментом є візуалізація або перегляд даних. У медичних наукових дослідженнях графічне

подання допомагає спостерігати за тенденціями зміни, виявляти складні взаємодіючі фактори й спрощує зіставлення даних. Треба відзначити, що візуалізація даних при сучасних технологіях аналізу істотно полегшена. Разом з тим, графічне подання даних значно полегшує попередній аналіз інформації, метою якого є формування подання аналізованих даних і попередній вибір методів аналізу, за допомогою яких обчислюються елементарні статистичні характеристики (середнє значення, помилка середнього, середньоквадратичне відхилення), визначаються залежності між даними та статистично достовірні відмінності між групами тощо.

Вибір і реалізація методу аналізу у зв'язку з їх різноманіттям може виявитися завданням нетривіальним. Однак, використання комп'ютерного аналізу даних легко дозволяє спробувати вирішити завдання кількома подібними методами й вибрати той, котрий дає найкращий результат. І дійсно, у сучасних пакетах прикладних програм занесені дані досить просто обробити з використанням різних процедур, а потім можна вибрати метод, що дає найкращі результати. При цьому слід зазначити, що один раз занесені дані можуть бути оброблені різними методами й різними програмами.

Заключним етапом технології аналізу даних є інтерпретація і подання результатів аналізу. Дуже важливе значення мають повнота й рівень опису як самого аналізу, так і його результатів та їхньої інтерпретації. Читач повинен мати можливість ясно уявляти собі всю послідовність обробки даних, оцінити адекватність обраних методів аналізу й обґрунтованість сформульованих висновків. Так, при інтерпретації результатів статистичної обробки даних завжди необхідно пам'ятати про їхній імовірнісний зміст. Він полягає в тому, що не завжди отримані результати є точними, а лише статистичними оцінками істотних значень. Крім того, при перевірці статистичних гіпотез необхідно пам'ятати про статистичну значимість, тому що дослідження, як правило, проводяться лише на якійсь вибірці з генеральної сукупності (популяції). Питання поширення отриманих висновків тісно пов'язане з репрезентативністю аналізованих вибірок.

Висновок. Завдяки використанню комп'ютерів і широкому впровадженню сучасних комп'ютерних технологій докорінно змінився процес обробки й аналізу медичних даних. Застосування комп'ютерної техніки робить достатньо складні методи аналізу медичних даних більш доступними і наочними. Вже не потрібно вручну виконувати трудомісткі розрахунки, будувати таблиці і графіки - всю цю чорнову роботу взяв на себе комп'ютер, а людині залишилася лише творча робота: постановка задач, вибір методів їх вирішення та інтерпретація результатів.

Література

1. Айвазян С. А. Теория вероятностей и прикладная статистика: в 2 т. / С. А. Айвазян, В. С. Мхитарян - Юнити, 2001. - Т-1- 656 с.
2. Алексахин С.В. и др. Прикладной статистический анализ: учебное пособие для вузов / С.В. Алексахин - М.: ПРИОР, 2001. - 224 с.
3. Алексахин С.В. и др. Прикладной статистический анализ данных. Теория. Компьютерная обработка. Области применения: в 2-х кн. / С.В. Алексахин - М.: ПРИОР, 2002. - 688 с.
4. Боровиков В. Statistica. Искусство анализа данных на компьютере / В. Боровиков - СПб: Питер, 2001. - 656 с.
5. Гайдышев И. П. Анализ и обработка данных: специальный справочник / И.П. Гайдышев - Специальный справочник, 2001. - 752 с.
6. Гельман В.Я. Медицинская информатика: практикум / В.Я. Гельман - СПб: Питер, 2001. - 480 с.
7. Гланц С. Медико-биологическая статистика / С. Гланц. - Издательство "Практика", 1999. - 459 с.
8. Гмурман В. Теория вероятностей и математическая статистика / В. Гмурман. - Высшая школа, 2001. - 346 с.
9. Гойко О.В. Практичне використання пакета STATISTICA для аналізу медико-біологічних даних: навчальний посібник для студентів вищих навчальних закладів (Рекомендовано МОН України, ISBN 966-8326-31-8) / О.В. Гойко. - Київ, 2004. - 76 с.
10. Калинина В.Н., Математическая статистика: учебник / В.Н. Калинина, В.Ф. Панкин. - Высшая школа, 2001. - 336 с.
11. Кремер Н. Теория вероятностей и математическая статистика / Н. Кремер. - Юнити, 2001. - 543 с.
12. Лапач С.Н. и др. Статистические методы в медико-биологических исследованиях с использованием Excel / С.Н. Лапач. - Морион Лтд, 2000. - 320 с.
13. Лукьянова Е.А. Медицинская статистика / Е.А. Лукьянова. - РУДН, 2002. - 255 с.
14. Мінцер О.П. Оброблення клінічних і експериментальних даних у медицині: навч. посібник / О.П. Мінцер, Ю.В. Вороненко, В.В. Власов - К.: Вища шк., 2003. - 350 с.