Разработка и заполнение баз данных

И.В. Жильцов

2016 г.

Что представляет из себя база данных?

– набор информации, имеющей отношение к какому-либо предмету или явлению, например:

– Имя, адрес электронной почты, номер телефона, рекомендации по диете, название организации, куда нужно отправлять счёт за курсы;
– Демографические и социоэкономические характеристики ВИЧ-инфицированных больных, госпитализированных в конкретный стационар;
– Демографические данные и исходы заболевания у пациентов с коинфекцией ВИЧ и ВГС;
– Для ВИЧ-инфицированных пациентов: демографические данные, схемы АРТ и лечения ОИ, СПИД-ассоциированные заболевания, лабораторные данные, побочные эффекты препаратов, коинфекции.

Демографические и социоэкономические характеристики ВИЧ-инфицированных больных, госпитализированных в конкретный стационар:

Patient ID	Birth	Gender	Exposure group	Ethnicity	Employed
1	01/02/1987	m	1	1	Student
2	06/07/1967	f	2	1	PT
3	05/03/1952	m	1	1	Student
4	31/10/1945	f	2	1	FT
5	26/09/1969	f	3	2	FT
1000	06/09/1969	m	1	2	Unemployed

Совокупность информации, структурированной таким образом, чтобы сделать возможной обработку указанной информации при помощи ЭВМ.

Как собрать хорошие данные?

Ключевое условие – хороший дизайн исследования.

– Определите цель сбора данных и продумайте, как будете их использовать;

– Какую информацию вы планируете получить на основе собранных данных?

– Каковы предметы исследования?

– Какую информацию о каждом предмете исследования вам необходимо собрать и хранить (*переменные*)?

– Собираемые данные являются результатами независимых измерений либо повторных замеров в одной и той же группе?

Пилотное исследование

Провести такое исследование до начала сбора данных –хорошая идея на любой случай.

– Поговорите с потенциальными пользователями результатов исследования;
– Обсудите вопросы, на которые нужно получить ответы;
– Набросайте образец формы, которую будет нужно заполнять;
– Прикиньте, как будут оформляться отчёты;
– Если возможно, используйте для работы тщательно продуманные базы данных (двумерные таблицы), аналогичные тем, которые будут применяться в ходе выполнения основного исследования.

Программное обеспечение, используемое для создания баз данных:

Базы данных: MS Access, DBase

Двумерные таблицы: MS Excel, Open Office Calc

Статистическая обработка: SAS, SPSS, STATA, Statistica, MedCalc

Базы данных:

Позволяют создавать большие массивы данных и гибко управлять ими.

– Позволяют работать со ссылками: можно из двух и более связанных таблиц, содержащих необходимую информацию, собрать одну таблицу с требуемыми данными;

– Информация не дублируется, что уменьшает вероятность ошибок ввода данных;

– Возможен поиск данных по поисковым запросам;

– Позволяют оформлять формы для ввода данных в виде реально используемых бумажных форм;

– Легкость организации процедур верификации данных;

– Наилучший вариант для долговременного хранения данных

Двумерные таблицы проще, с ними легче работать.

– Возможны ограничения по размеру (например, в MS Excel 2003 и более ранних версиях – не более 256 переменных и 65536 строк);

– Неудобное извлечение данных;

– Скудные возможности по верификации данных, отсутствие защиты от повреждения данных (например, при сортировке);

– Позволяют производить простую статобработку непосредственно в таблице;

– Имеют ряд функций, общих с базами данных.

Программы для статистической обработки данных:

– Имеют общие черты и с базами данных, и с двумерными таблицами;

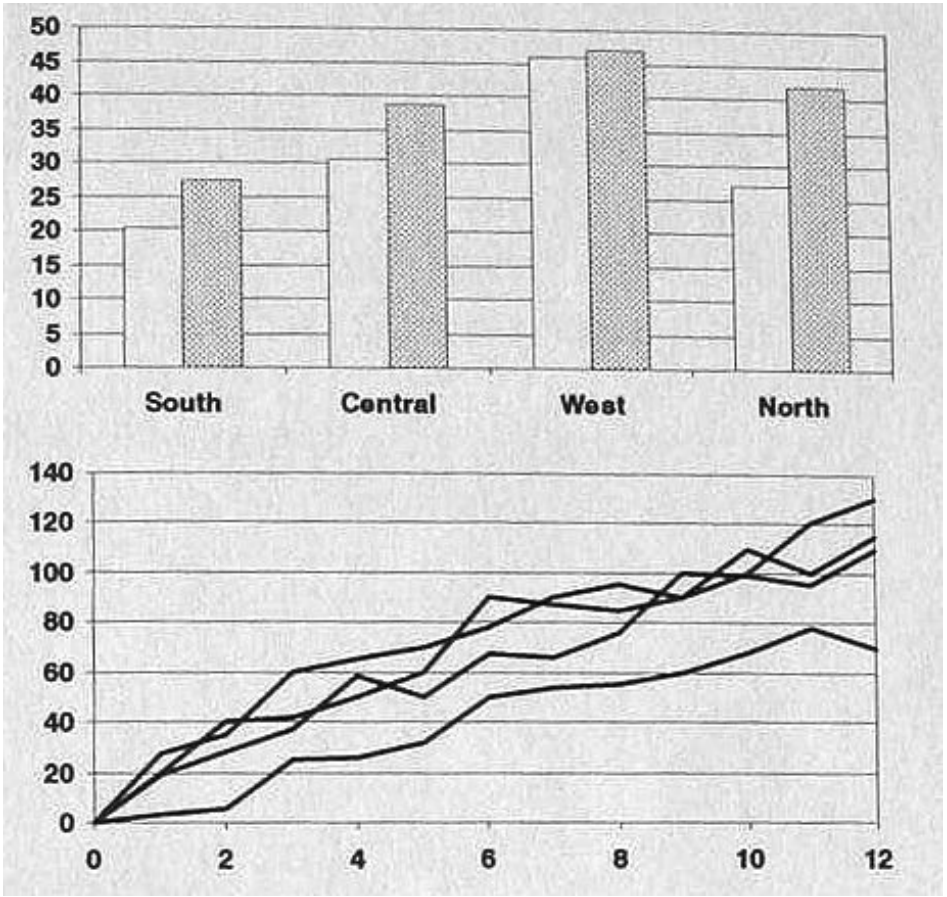
– …но манипуляции с данными требуют знания интерфейса соответствующего статпакета;

– Можно вводить данные непосредственно в форму для обсчёта;

– Обычно допускают простой импорт двумерных таблиц с данными из других программ.

Два основных типа данных:

– числовые (количественные);
– категориальные (качественные)



Качественные (категориальные) данные:

– **Бинарные** (жив/мёртв, мужчина/женщина, заболевание развилось/не развилось)

– **Номинальные** (две и более категории, не ранжируемые по порядку: например, группы риска какого-либо заболевания)

– **Порядковые** (ранжируемые): две и более категории, которые по своей природе допускают ранжирование (выстраивание в определённом порядке) – степени тяжести заболевания, стадии заболевания и т.д.

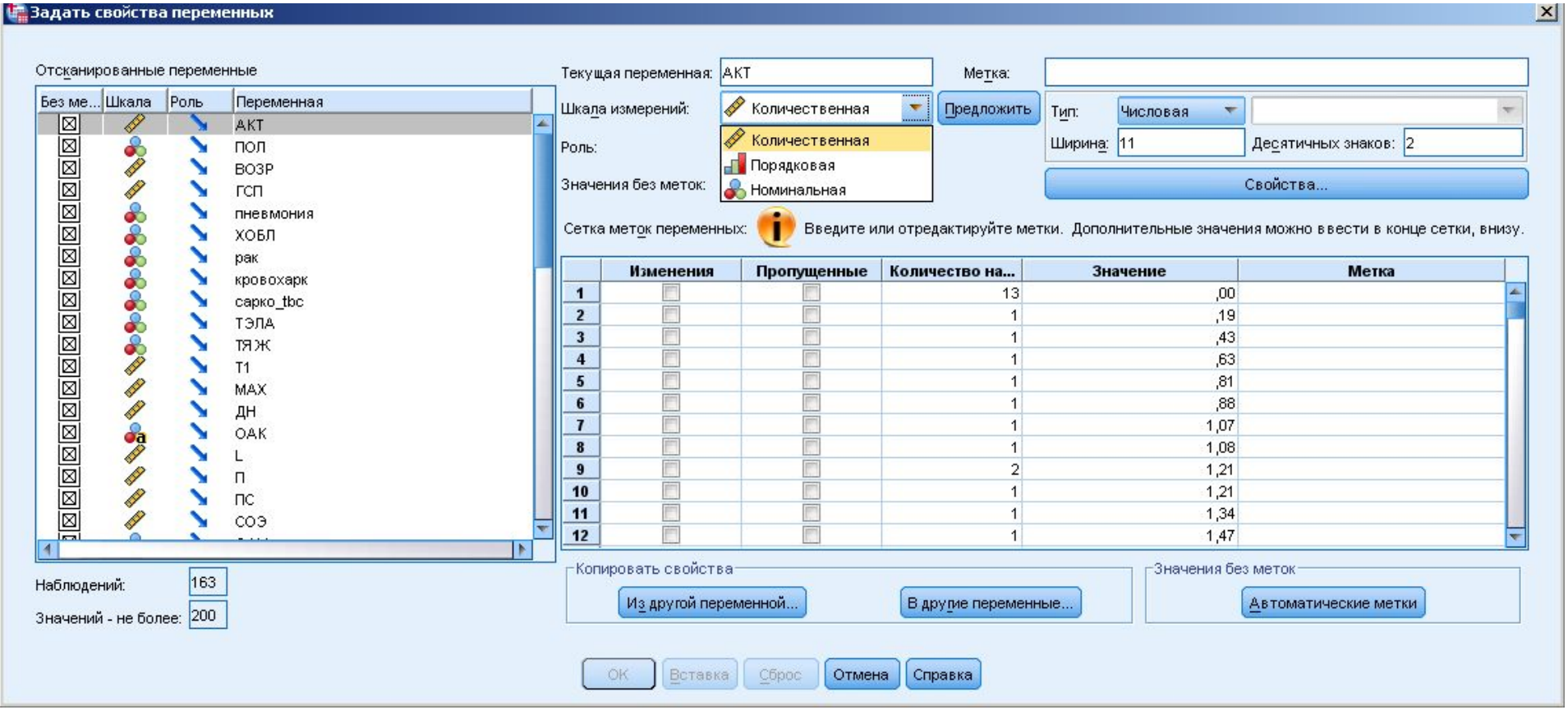
Количественные (числовые) данные:

– Дискретные: могут принимать только определённые значения в определённом диапазоне (например, индекс качества жизни или количество половых партнёров);

– Непрерывные: могут принимать любое значение в рамках измеряемого диапазона (например, вес, рост, уровень CD4- лимфоцитов и т.п.);

– Цензурированные: могут быть измерены только в определённом диапазоне (например, число копий РНК ВИЧ в единице объёма плазмы крови, время дожития и т.д.);

– Прочие типы данных: ранги, доли, частоты, отношения.


Задать свойства переменных
Отсканированные переменные
Без ме...
Шкала
Роль
Переменная
АКТ
ПОЛ
ВОЗР
ГСП
пневмония
ХОБЛ
рак
кровохарк
сарко_tbc
ТЭЛА
ТЯЖ
Т1
МАХ
ДН
ОАК
L
П
ПС
СОЭ
Наблюдений: 163
Значений - не более: 200
Текущая переменная: АКТ
Метка:
Шкала измерений: Количественная
Предложить
Тип: Числовая
Ширина: 11
Десятичных знаков: 2
Свойства...
Роль:
Значения без меток:
Количественная
Порядковая
Номинальная
Сетка меток переменных:
Введите или отредактируйте метки. Дополнительные значения можно ввести в конце сетки, внизу.
Изменения
Пропущенные
Количество на...
Значение
Метка
1 13 ,00
2 1 ,19
3 1 ,43
4 1 ,63
5 1 ,81
6 1 ,88
7 1 1,07
8 1 1,08
9 2 1,21
10 1 1,21
11 1 1,34
12 1 1,47
Копировать свойства
Из другой переменной...
В другие переменные...
Значения без меток
Автоматические метки
OK
Вставка
Сброс
Отмена
Справка

Поля и форматы данных:

– Текстовые: текст, комбинация текста и цифр либо цифры, не нуждающиеся в обработке (имя, адрес, телефонный номер, пол, группа/фактор риска и т.д.). Возможна разбивка на меньшие поля: Имя, Фамилия и т. д.;

– Числовые: числа, предназначенные для статобработки, а также коды и категории, (возраст, уровень CD4-лимфоцитов, вес, кодировка групп, бинарные переменные – «да/нет» кодируется как «1/0»);

– Дата: любая информация, которая должна храниться в виде даты (дата установления диагноза, выполнения исследования, наступления ожидаемого исхода и т.д.). Может быть представлена в разных форматах: дд/мм/гг, мм/дд/гг, дд/месяц/гггг и т.д. *Старые версии программ могут автоматически изменять форматы дат. Кроме того, форматы дат в PC и MAC*

Поля и форматы данных:

– **Поля бинарных данных**: в некоторых программах есть формат ячеек, позволяющий хранить только данные вида «да/нет» (также «вкл/выкл», «истина/ложь» и др.).

– **Поля с выпадающим списком**: некоторые программы имеют формат, позволяющий вносить в поля базы данных только значения, имеющиеся в выпадающем списке (открывается при помещении курсора в соответствующую ячейку), например, коды исследуемых групп, степени тяжести заболевания и т.п.

Практика сбора высококачественных данных:

– Будьте последовательны

Многие проблемы проистекают от непоследовательности при сборе и оформлении одинаковых данных.
***Например*: данные одновременно хранятся в формате дд/мм/гг и мм/дд/гг, пол обозначается как «м/ж», «0/1» и «1/2» в одной базе данных.**

– Старайтесь не трансформировать числовые переменные в категориальные до этапа анализа данных:

***Например*: не выделяйте возрастные группы, группы по уровню какого-либо показателя (CD4+ >200/мкл и т.п.)**

Практика сбора высококачественных данных:

– Пропуски данных: для многих переменных неизбежны.

Придумайте общую стратегию работы с подобными данными.
Не оставляйте соответствующие ячейки пустыми: ***многие статпрограмы автоматически заполняют пустые ячейки нулевыми значениями*.**

Вместо этого используйте специальный код: например, м – 1, ж – 2, данных нет – 9. Старайтесь, чтобы эти коды нельзя было перепутать с данными (например, не используйте код «999» для уровня CD4+ лимфоцитов). Если неизвестен день – выбирайте 1-е число, если месяц – выбирайте июнь и т.д. ***Не оставляйте отсутствующие поля даты пустыми*!**

Практика сбора высококачественных данных:

Простая проверка данных.

Хорошая привычка – проводить простую проверку правильности введения данных в базу перед статанализом.

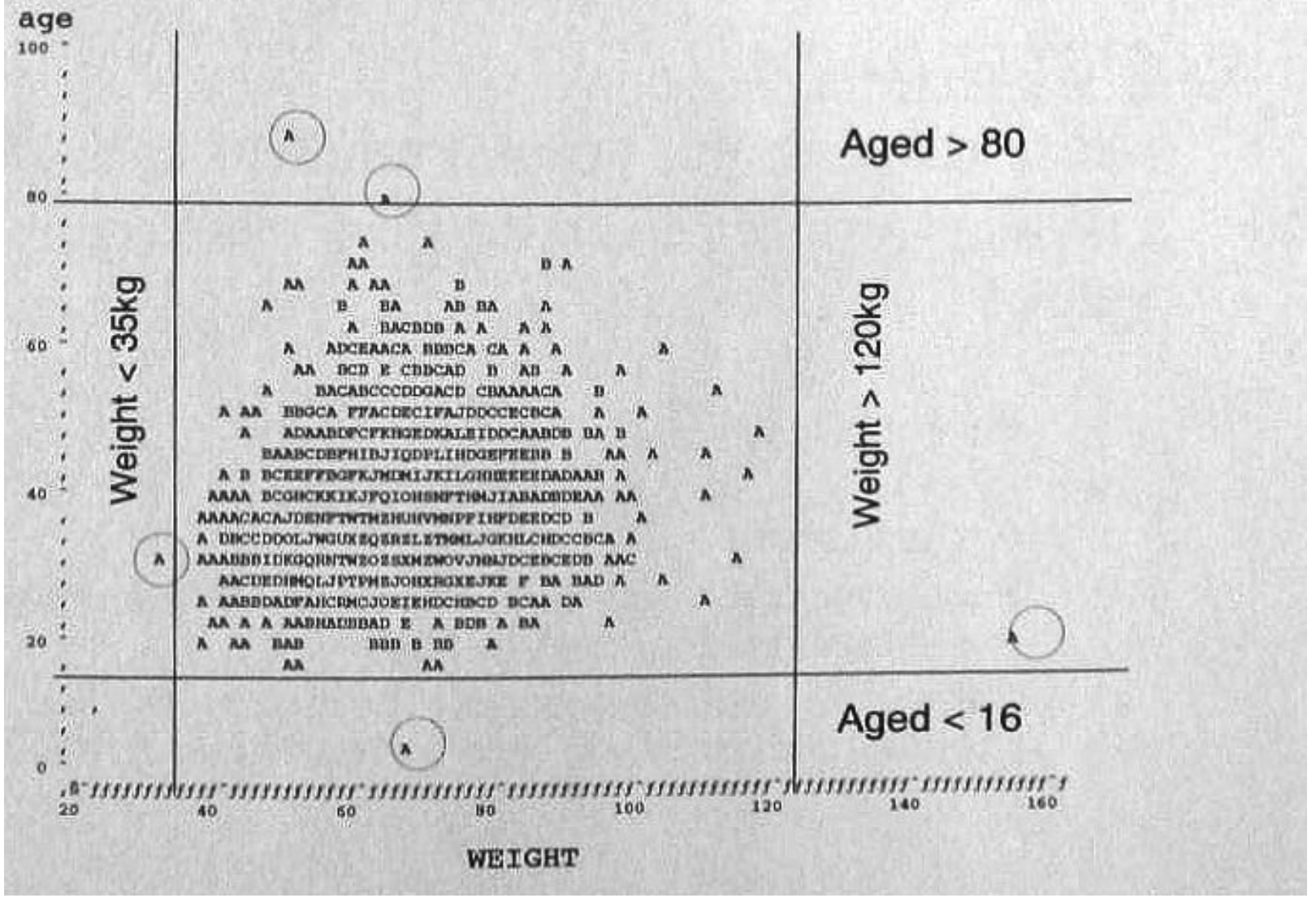
– есть ли среди введенных дат явно несообразные/ непоследовательные?

– все ли введенные даты событий больше дат рождения пациентов? Не заходят ли они за дату смерти?

– если данные кодированы, нет ли в базе непредусмотренных кодов?

– не выглядят ли непрерывные данные явно запредельными?

Plot of age and weight at recruitment



Практика сбора высококачественных данных:

При всякой возможности избегайте внесения «просто текста», оставляя его обработку «на потом».

– данные из длинного текста потом неудобно извлекать;
– при этом возможно появление множественных кодировок одних и тех же данных.

Diagnosis	*N*	
Kaposi	1	N=7
Kaposis Sarcoma	1	
Kaposis sarcoma	2	
Kaposi's sarcoma	1	
Kaposis	2	
NHL	1	N=6
Non Hodgkin Lymphoma	1	
Non-Hodgkin Lymphoma	3	
Non-hodgkins lymphoma	1	
Toxopl. Brain	1	N=3
Toxoplasm brain	1	
Toxoplasmos	1	

Практика сбора высококачественных данных:

Не смешивайте числа и текст.

Например, при заполнении числовых ячеек не пишите там «>200000» или «<75»: некоторые программы интерпретируют такие записи как отсутствующие данные. Вместо этого следует пользоваться *специальными заменителями*, например, «200001» для первого случая или «74» для второго случая.

Практика сбора высококачественных данных:

Что делать, если собрано много значений одной переменной для одного и того же случая?

Такое часто бывает в «продольных» исследованиях, где выполняется мониторинг уровня CD4+ лимфоцитов, вирусной нагрузки в плазме крови и т.п. в исследуемой выборке.

**Нужно создавать по переменной для каждого из значений, полученных в одинаковый момент времени: например, CD4_1, CD4_2, CD4_3 и т.д. При этом каждой такой переменной должна быть сопоставлена переменная даты. Аналогично производится разложение по переменным сложных диагнозов, например, множества сопутствующих заболеваний при ВИЧ-инфекции (*в ячейке можно указывать дату*

Метод 1 сложнее, но практически не имеет ограничений.

Patient ID	AIDS1	DateAIDS1	AIDS2	DateAIDS2	.....	AIDS9	DateAIDS9
1	24	01/01/1999	2	01/01/1999			
2	3	05/04/1988					
3							
4	1	01/06/1997	5	06/08/1997		24	02/03/2004
5							
1000	18	01/09/1992					

Метод 2 проще и требует меньше места, но позволяет вводить только значения, предусмотренные разработчиком базы.

Patient ID	Oes cand	PCP	KS	NHL	Toxo	.....	Wasting
1	08/01/98						
2					06/15/95		
3							
4		08/03/97		12/01/92			
5							
1000							

Данные могут храниться в двух форматах:

1. **Формат «высокий столбец»: каждая запись для одного пациента, соответствующая *определённому моменту времени,* указывается в отдельной строке.**

Eg Single file of CD4 counts

ID	*Date*	*CD4 count*
100	1/1/99	200
100	1/4/99	250
100	1/8/99	500

2. **Формат «широкая строка»: для каждого пациента отводится одна линейка таблицы.**

ID	CD4-v1	CD4-d1	CD4-v2	CD4-d2	CD4-v3	CD4-d4
100	200	1/1/99	250	1/4/99	500	1/8/99

Оба формата подразумевают уникальные идентификаторы для каждого пациента, ввиду чего легко транспонируются специальными программами в любой требуемый вид.

«Высокий» формат экономит место, но требует, чтобы программа поддерживала достаточное количество линеек. Кроме того, *возможны ошибки в указании идентификаторов пациентов*, что приводит к потере данных.

«Широкий» формат менее чувствителен к вводу идентификатора пациента (вводится только один раз), но требует, чтобы программа поддерживала достаточное количество столбцов. Кроме того, внесение каждого непредусмотренного значения требует переделки базы. Если для одного пациента было внесено больше данных, чем для другого,

Если в ходе исследования производится модификация/расширение базы данных, необходимо вести журнал изменений, а также хранить окончательные версии каждой из модификаций, помечая их таким образом, чтобы можно было точно установить дату и версию модификации. Не стирайте старые версии, замещая их новыми!!!

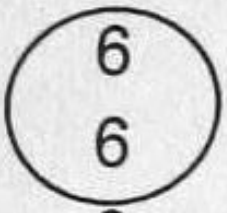
Найдите все ошибки, допущенные при заполнении представленной базы данных ☺️

Patient	Gender	Date of birth	AIDS	First VL
1	2	14/09/1969	Yes	7,500
2	F	01/05/1982	PCP	49
3	—	5/9/69*		≤50
4	M	1954	No	100000
5	M	02/05/1974	1	4.86
6	—	12/20/56	02/05/1998	?
6	d/k	19/10/1964	Pneumocystis, 2001	≥750000
8	Male	01/01/1935		123561
9	male	02/05/1958	28/02/1957	—
10	—	09/09/2006		*
2	female	31/6/29 — 2 old		

* needs to be checked

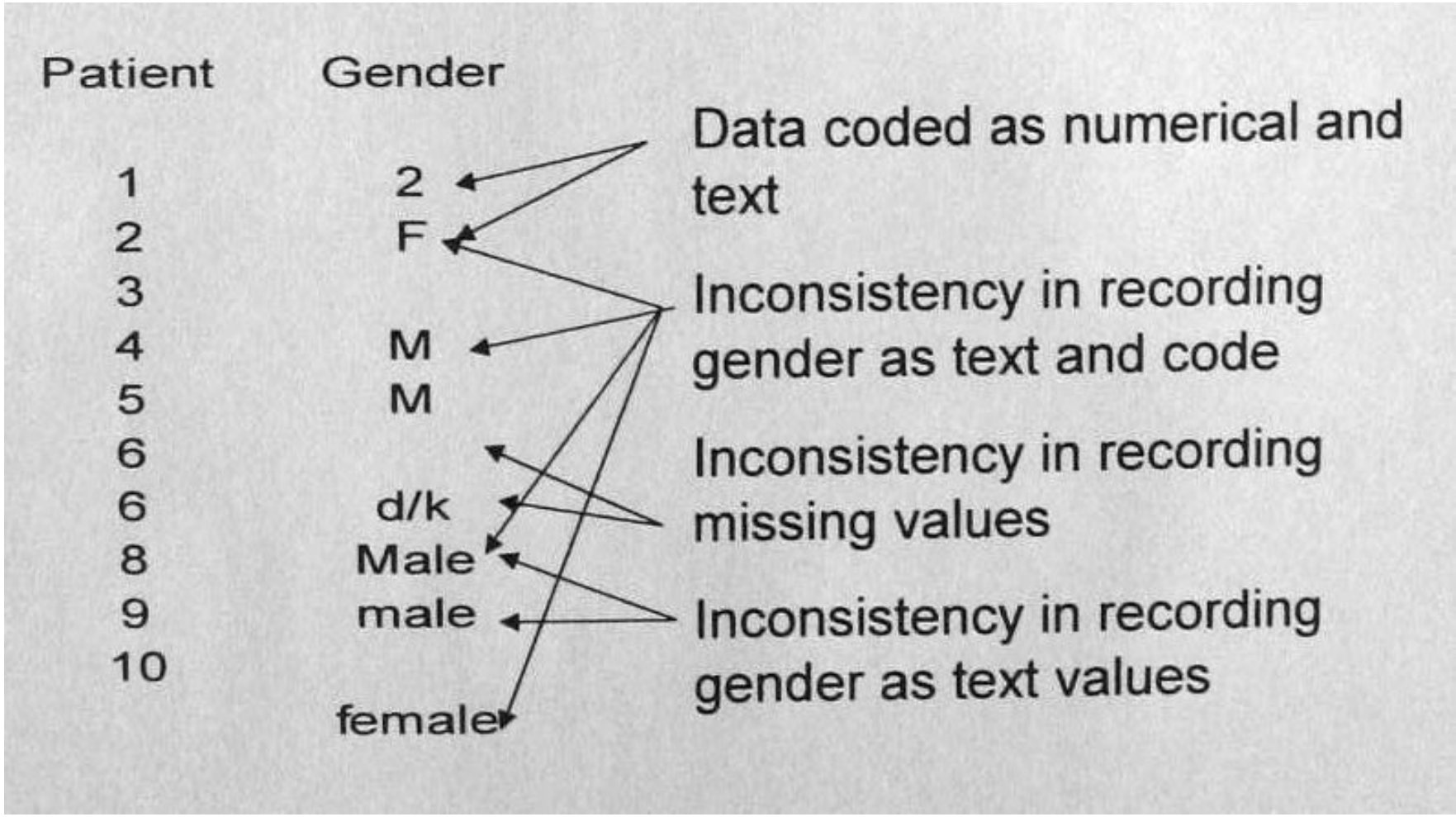
Patient

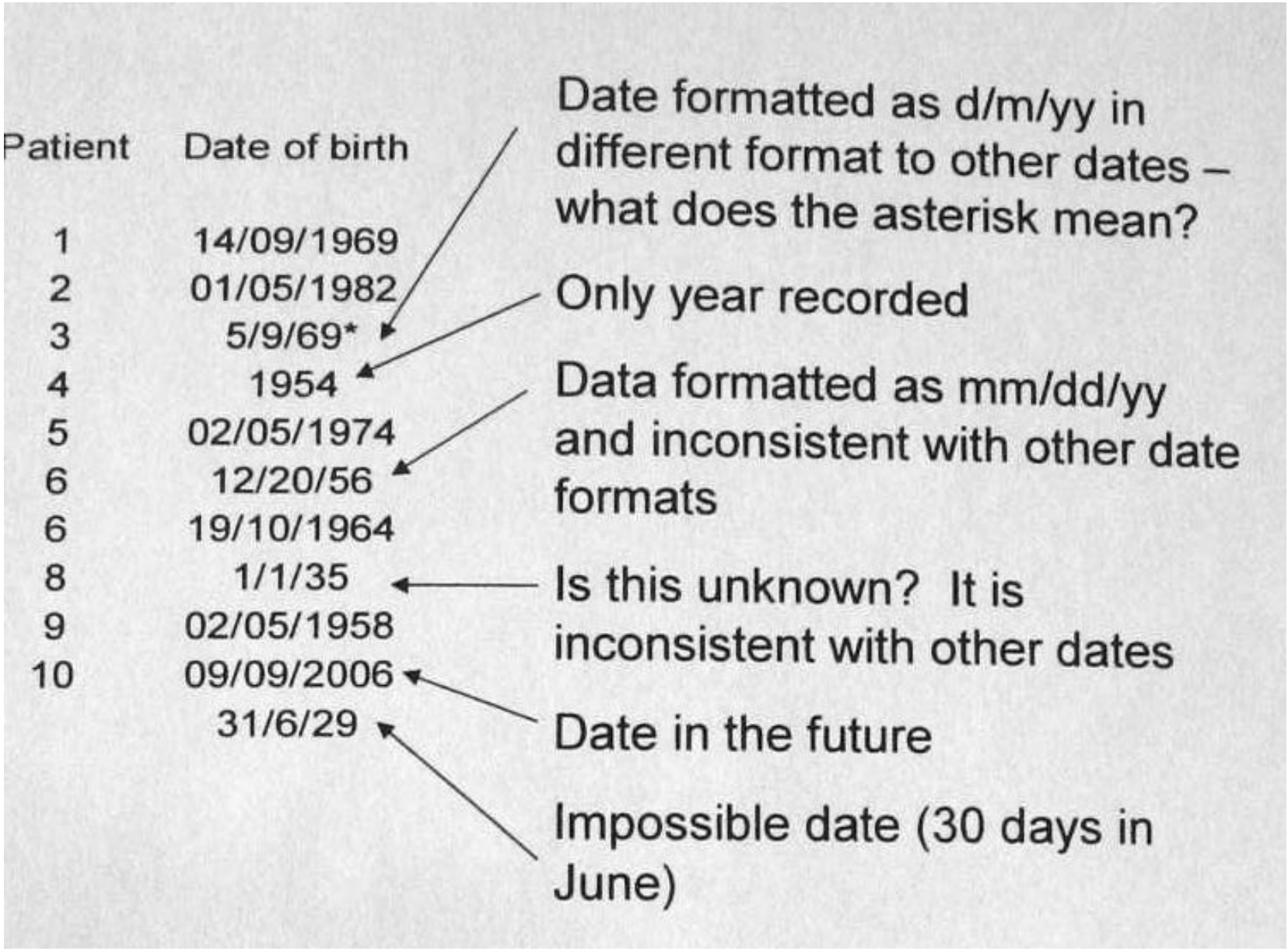
1
2
3
4
5
6
6
8
9
10



2 patients with same ID – is it the same patient (and other data should be similar) or a different patient?

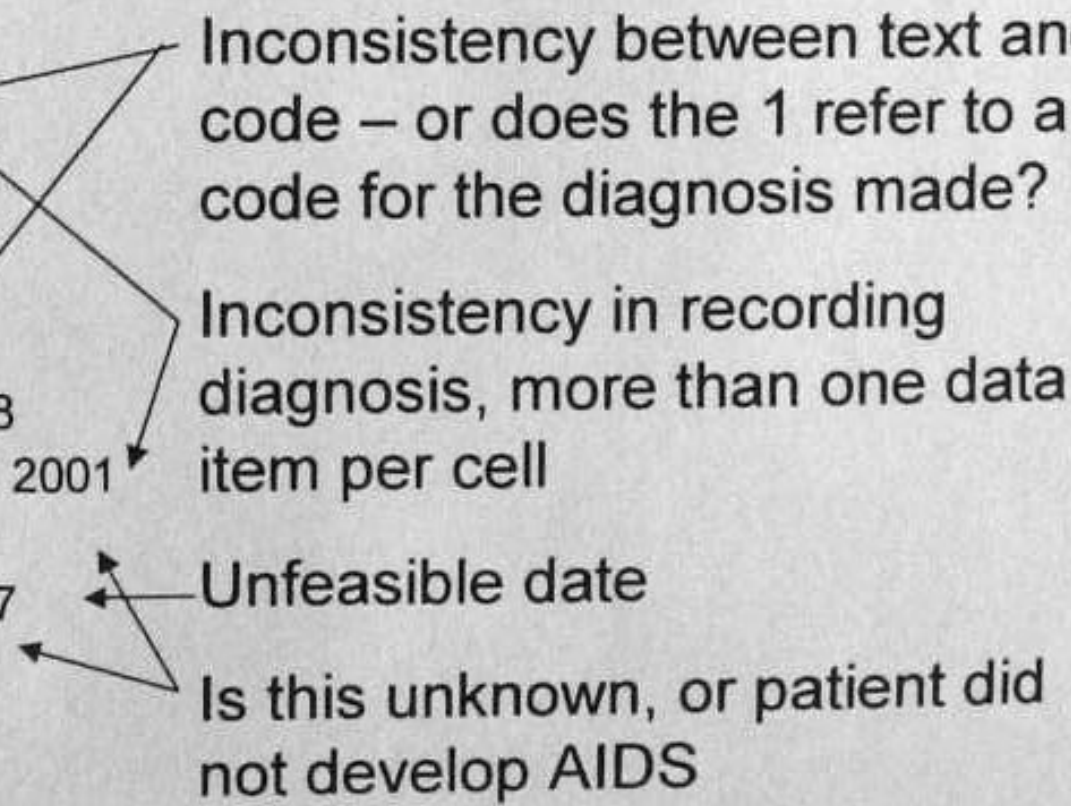
One line of data with no patient ID

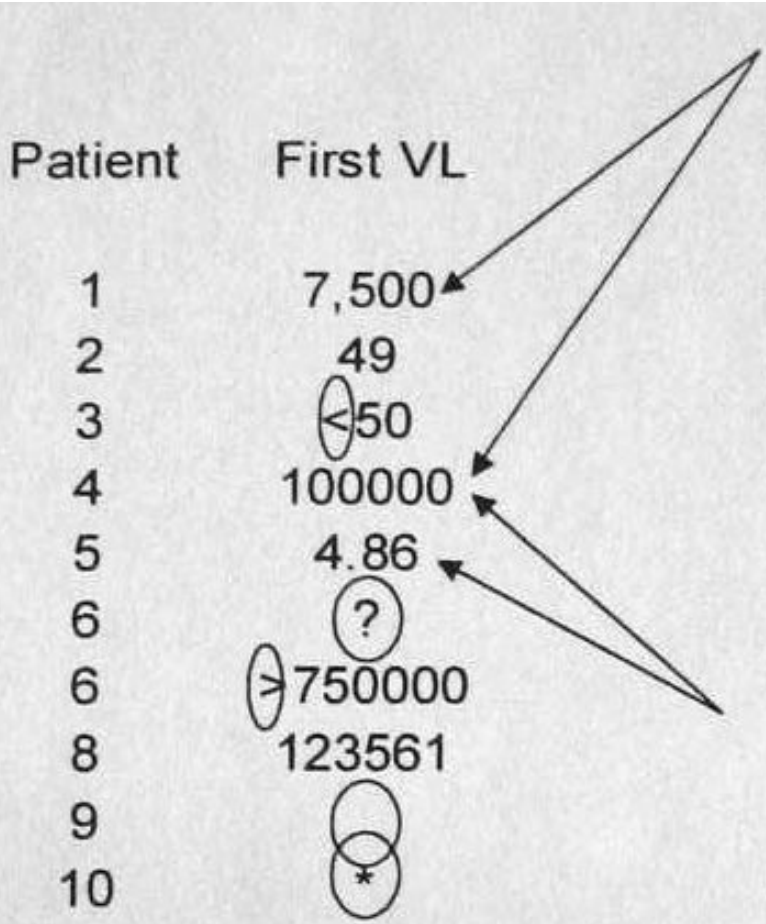

Patient
Gender
1
2
3
4
5
6
6
8
9
10
2
F
M
M
d/k
Male
male
female
Data coded as numerical and text
Inconsistency in recording gender as text and code
Inconsistency in recording missing values
Inconsistency in recording gender as text values


Patient
Date of birth
1
14/09/1969
2
01/05/1982
3
5/9/69*
4
1954
5
02/05/1974
6
12/20/56
6
19/10/1964
8
1/1/35
9
02/05/1958
10
09/09/2006
31/6/29
Date formatted as d/m/yy in different format to other dates – what does the asterisk mean?
Only year recorded
Data formatted as mm/dd/yy and inconsistent with other date formats
Is this unknown? It is inconsistent with other dates
Date in the future
Impossible date (30 days in June)

Is the purpose to collect date, diagnosis or whether AIDS has occurred?

Patient	AIDS
1	Yes
2	PCP
3	
4	No
5	1
6	02/05/1998
6	Pneumocystis, 2001
8	
9	28/02/1957
10	



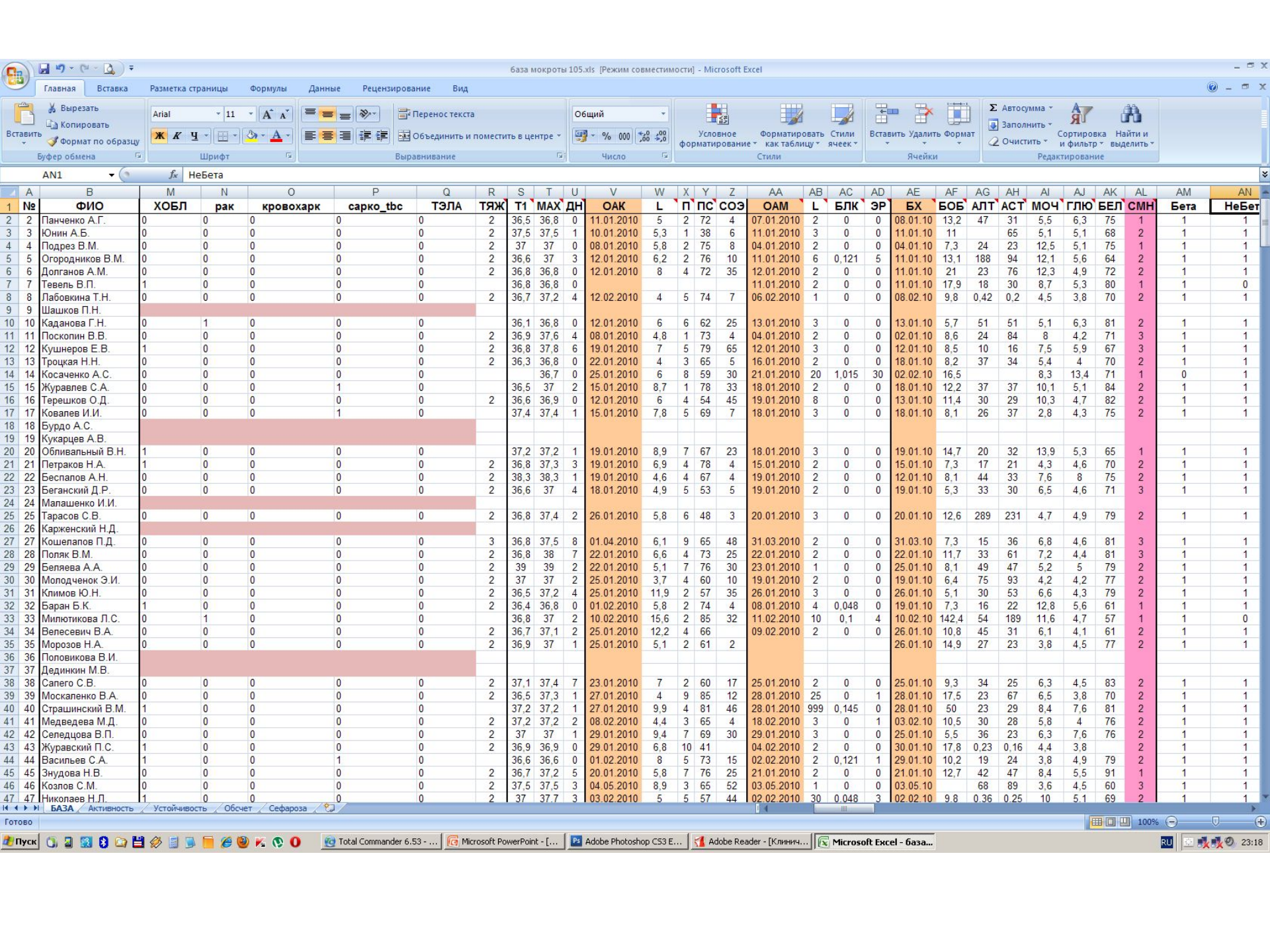


Inconsistency between text and numbers – ',' converts value to text – value will be lost

> And < converts value to text – value will be lost – develop strategy for values outside range

Data inconsistent - recorded as $\log_{10}$copies/ml and copies/ml

Are these unknown, what is difference between '?' and missing? What does the asterisk mean



база мокроты 105.xls [Режим совместимости] - Microsoft Excel

Главная | Вставка | Разметка страницы | Формулы | Данные | Рецензирование | Вид

AN1 | НеБета

№	ФИО	ХОБЛ	рак	кровохарк	сарко_tbc	ТЭЛА	ТЯЖ	Т1	МАХ	ДН	ОАК	L	П	ПС	СОЭ	ОАМ	L	БЛК	ЭР	БХ	БОБ	АЛТ	АСТ	МОЧ	ГЛЮ	БЕЛ	СМН	Бета	НеБет
2	Панченко А.Г.	0	0	0	0	0	2	36,5	36,8	0	11.01.2010	5	2	72	4	07.01.2010	2	0	0	08.01.10	13,2	47	31	5,5	6,3	75	1	1	1
3	Юнин А.Б.	0	0	0	0	0	2	37,5	37,5	1	10.01.2010	5,3	1	38	6	11.01.2010	3	0	0	11.01.10	11		65	5,1	5,1	68	2	1	1
4	Подрез В.М.	0	0	0	0	0	2	37	37	0	08.01.2010	5,8	2	75	8	04.01.2010	2	0	0	04.01.10	7,3	24	23	12,5	5,1	75	1	1	1
5	Огородников В.М.	0	0	0	0	0	2	36,6	37	3	12.01.2010	6,2	2	76	10	11.01.2010	6	0,121	5	11.01.10	13,1	188	94	12,1	5,6	64	2	1	1
6	Долганов А.М.	0	0	0	0	0	2	36,8	36,8	0	12.01.2010	8	4	72	35	12.01.2010	2	0	0	11.01.10	21	23	76	12,3	4,9	72	2	1	1
7	Тевель В.П.	1	0	0	0	0		36,8	36,8	0						11.01.2010	2	0	0	11.01.10	17,9	18	30	8,7	5,3	80	1	1	0
8	Лабовкина Т.Н.	0	0	0	0	0	2	36,7	37,2	4	12.02.2010	4	5	74	7	06.02.2010	1	0	0	08.02.10	9,8	0,42	0,2	4,5	3,8	70	2	1	1
9	Шашков П.Н.																												
10	Каданова Г.Н.	0	1	0	0	0		36,1	36,8	0	12.01.2010	6	6	62	25	13.01.2010	3	0	0	13.01.10	5,7	51	51	5,1	6,3	81	2	1	1
11	Поскопин В.В.	0	0	0	0	0	2	36,9	37,6	4	08.01.2010	4,8	1	73	4	04.01.2010	2	0	0	02.01.10	8,6	24	84	8	4,2	71	3	1	1
12	Кушнеров Е.В.	1	0	0	0	0	2	36,8	37,8	6	19.01.2010	7	5	79	65	12.01.2010	3	0	0	12.01.10	8,5	10	16	7,5	5,9	67	3	1	1
13	Троцкая Н.Н.	0	0	0	0	0	2	36,3	36,8	0	22.01.2010	4	3	65	5	16.01.2010	2	0	0	18.01.10	8,2	37	34	5,4	4	70	2	1	1
14	Косаченко А.С.	0	0	0	0	0			36,7	0	25.01.2010	6	8	59	30	21.01.2010	20	1,015	30	02.02.10	16,5			8,3	13,4	71	1	0	1
15	Журавлев С.А.	0	0	0	1	0		36,5	37	2	15.01.2010	8,7	1	78	33	18.01.2010	2	0	0	18.01.10	12,2	37	37	10,1	5,1	84	2	1	1
16	Терешков О.Д.	0	0	0	0	0	2	36,6	36,9	0	12.01.2010	6	4	54	45	19.01.2010	8	0	0	13.01.10	11,4	30	29	10,3	4,7	82	2	1	1
17	Ковалев И.И.	0	0	0	1	0		37,4	37,4	1	15.01.2010	7,8	5	69	7	18.01.2010	3	0	0	18.01.10	8,1	26	37	2,8	4,3	75	2	1	1
18	Бурдо А.С.																												
19	Кукарцев А.В.																												
20	Обливальный В.Н.	1	0	0	0	0		37,2	37,2	1	19.01.2010	8,9	7	67	23	18.01.2010	3	0	0	19.01.10	14,7	20	32	13,9	5,3	65	1	1	1
21	Петраков Н.А.	1	0	0	0	0	2	36,8	37,3	3	19.01.2010	6,9	4	78	4	15.01.2010	2	0	0	15.01.10	7,3	17	21	4,3	4,6	70	2	1	1
22	Беспалов А.Н.	0	0	0	0	0	2	38,3	38,3	1	19.01.2010	4,6	4	67	4	19.01.2010	2	0	0	12.01.10	8,1	44	33	7,6	8	75	2	1	1
23	Беганский Д.Р.	0	0	0	0	0	2	36,6	37	4	18.01.2010	4,9	5	53	5	19.01.2010	2	0	0	19.01.10	5,3	33	30	6,5	4,6	71	3	1	1
24	Малашенко И.И.																												
25	Тарасов С.В.	0	0	0	0	0	2	36,8	37,4	2	26.01.2010	5,8	6	48	3	20.01.2010	3	0	0	20.01.10	12,6	289	231	4,7	4,9	79	2	1	1
26	Карженский Н.Д.																												
27	Кошелапов П.Д.	0	0	0	0	0	3	36,8	37,5	8	01.04.2010	6,1	9	65	48	31.03.2010	2	0	0	31.03.10	7,3	15	36	6,8	4,6	81	3	1	1
28	Поляк В.М.	0	0	0	0	0	2	36,8	38	7	22.01.2010	6,6	4	73	25	22.01.2010	2	0	0	22.01.10	11,7	33	61	7,2	4,4	81	3	1	1
29	Беляева А.А.	0	0	0	0	0	2	39	39	2	22.01.2010	5,1	7	76	30	23.01.2010	1	0	0	25.01.10	8,1	49	47	5,2	5	79	2	1	1
30	Молодченок Э.И.	0	0	0	0	0	2	37	37	2	25.01.2010	3,7	4	60	10	19.01.2010	2	0	0	19.01.10	6,4	75	93	4,2	4,2	77	2	1	1
31	Климов Ю.Н.	0	0	0	0	0	2	36,5	37,2	4	25.01.2010	11,9	2	57	35	26.01.2010	3	0	0	26.01.10	5,1	30	53	6,6	4,3	79	2	1	1
32	Баран Б.К.	1	0	0	0	0	2	36,4	36,8	0	01.02.2010	5,8	2	74	4	08.01.2010	4	0,048	0	19.01.10	7,3	16	22	12,8	5,6	61	1	1	1
33	Милютикова Л.С.	0	1	0	0	0		36,8	37	2	10.02.2010	15,6	2	85	32	11.02.2010	10	0,1	4	10.02.10	142,4	54	189	11,6	4,7	57	1	1	0
34	Велесевич В.А.	0	0	0	0	0	2	36,7	37,1	2	25.01.2010	12,2	4	66		09.02.2010	2	0	0	26.01.10	10,8	45	31	6,1	4,1	61	2	1	1
35	Морозов Н.А.	0	0	0	0	0	2	36,9	37	1	25.01.2010	5,1	2	61	2					26.01.10	14,9	27	23	3,8	4,5	77	2	1	1
36	Половикова В.И.																												
37	Дединкин М.В.																												
38	Сапего С.В.	0	0	0	0	0	2	37,1	37,4	7	23.01.2010	7	2	60	17	25.01.2010	2	0	0	25.01.10	9,3	34	25	6,3	4,5	83	2	1	1
39	Москаленко В.А.	0	0	0	0	0	2	36,5	37,3	1	27.01.2010	4	9	85	12	28.01.2010	25	0	1	28.01.10	17,5	23	67	6,5	3,8	70	2	1	1
40	Страшинский В.М.	1	0	0	0	0		37,2	37,2	1	27.01.2010	9,9	4	81	46	28.01.2010	999	0,145	0	28.01.10	50	23	29	8,4	7,6	81	2	1	1
41	Медведева М.Д.	0	0	0	0	0	2	37,2	37,2	2	08.02.2010	4,4	3	65	4	18.02.2010	3	0	1	03.02.10	10,5	30	28	5,8	4	76	2	1	1
42	Селедцова В.П.	0	0	0	0	0	2	37	37	1	29.01.2010	9,4	7	69	30	29.01.2010	3	0	0	25.01.10	5,5	36	23	6,3	7,6	76	2	1	1
43	Журавский П.С.	1	0	0	0	0	2	36,9	36,9	0	29.01.2010	6,8	10	41		04.02.2010	2	0	0	30.01.10	17,8	0,23	0,16	4,4	3,8		2	1	1
44	Васильев С.А.	1	0	0	1	0		36,6	36,6	0	01.02.2010	8	5	73	15	02.02.2010	2	0,121	1	29.01.10	10,2	19	24	3,8	4,9	79	2	1	1
45	Знудова Н.В.	0	0	0	0	0	2	36,7	37,2	5	20.01.2010	5,8	7	76	25	21.01.2010	2	0	0	21.01.10	12,7	42	47	8,4	5,5	91	1	1	1
46	Козлов С.М.	0	0	0	0	0	2	37,5	37,5	3	04.05.2010	8,9	3	65	52	03.05.2010	1	0	0	03.05.10		68	89	3,6	4,5	60	3	1	1
47	Николаев Н.Л.	1	0	0	0	0	2	37	37,7	3	03.02.2010	5	5	57	44	02.02.2010	30	0,048	3	02.02.10	9,8	0,36	0,25	10	5,1	69	2	1	1

Листы: БАЗА | Активность | Устойчивость | Обсчет | Сефароза

Готово

Пуск | Total Commander 6.53 - ... | Microsoft PowerPoint - [... | Adobe Photoshop CS3 E... | Adobe Reader - [Клинич... | Microsoft Excel - база... | RU | 23:18

STATISTICA - [Data: База 1 (общая).sta (121v by 588c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help

Add to Workbook Add to Report Add to MS Word

Arial 11 Vars Cases

Calc

	1 АКТ	2 ДПТ	3 ПОЛ	4 ВОЗР	5 РОЖ_1	6 ПНВ_2	7 АВИ_3	8 ВМН_4	9 ГМН_5	10 АНГ_6	11 ТР_7	12 ТУБ_8	13 ГРП_9	14 ЗДОР_1(	15 ГРУППИР	16 ГСГ	17 t>38	18 ДН>4	19 ДН>10	20 ГСП>1(	21 ГСП>2(	22 t>39	23 СМН>='	24 СМН>=2	25 СОЧ>=2	26 Легк
1	49,42	0,269	2	26	0	0	0	0	0	1	0	0	0	0	6	15	1	1	0	1	0	0	0	0	0	0
2	49,60	0,270	1	13	0	0	0	0	0	1	0	0	0	0	6	8	0	0	0	0	0	0	0	0	0	1
3	51,07	0,278	1	39	1	0	0	0	0	0	0	0	0	0	1	6	0	0	0	0	0	0	0	0	0	1
4	21,50	0,117	2	76	1	0	0	0	0	0	0	0	0	0	1	9	0	0	0	0	0	0	0	0	0	1
5	50,89	0,277	1	42	0	1	0	0	0	0	0	0	0	0	2	11	0	0	0	1	0	0	0	0	0	0
6	43,73	0,238	2	37	0	0	0	0	0	1	0	0	0	0	6	7	0	0	0	0	0	0	0	0	0	1
7	40,48	0,257	1	43	0	1	0	0	0	0	0	0	0	0	2	15	1	1	1	1	0	1	1	1	1	0
8	27,37	0,149	1	72	0	1	0	0	0	0	0	0	0	0	2	38	1	1	1	1	1	1	1	1	1	0
9	50,16	0,273	2	23	0	0	0	0	1	0	0	0	0	0	5	32	1	1	1	1	1	1	0	0	1	0
10	36,56	0,199	2	56	0	0	0	0	1	0	0	0	0	0	5	23	1	1	0	1	1	0	0	0	1	0
11	53,83	0,293	2	18	0	0	0	0	0	1	0	0	0	0	6	13	1	0	0	1	0	0	0	0	0	0
12	47,22	0,257	1	25	1	0	0	0	0	0	0	0	0	0	1	21	0	0	0	1	1	0	0	0	0	0
13	44,46	0,242	2	50	0	1	0	0	0	0	0	0	0	0	2	38	1	1	1	1	1	0	1	1	1	0
14	57,14	0,311	1	17	0	0	0	1	0	0	0	0	0	0	4	22	0	1	0	1	1	0	0	0	0	0
15	58,22	0,317	1	15	0	0	0	1	0	0	0	0	0	0	4	24	0	0	0	1	1	0	0	0	0	0
16	0,00	0,000	1	52	0	1	0	0	0	0	0	0	0	0	2	14	1	1	0	1	0	1	1	0	1	0
17	41,50	0,226	2	65	1	0	0	0	0	0	0	0	0	0	1	11	0	0	0	1	0	0	0	0	0	0
18	46,83	0,255	1	58	1	0	0	0	0	0	0	0	0	0	1	17	1	1	1	1	0	0	0	0	0	0
19	35,07	0,191	2	64	0	1	0	0	0	0	0	0	0	0	2	38	1	0	0	1	1	1	1	0	1	0
20	45,36	0,247	2	56	1	0	0	0	0	0	0	0	0	0	1	19	1	1	0	1	0	1	1	0	1	0
21	40,58	0,221	2	50	1	0	0	0	0	0	0	0	0	0	1	13	0	0	0	1	0	0	1	0	0	0
22	39,48	0,215	2	68	1	0	0	0	0	0	0	0	0	0	1	10	1	0	0	0	0	1	0	0	0	0
23	19,64	0,107	1	45	0	0	0	0	1	0	0	0	0	0	5	27	1	1	0	1	1	1	0	0	1	0
24	56,05	0,320	1	29	0	1	0	0	0	0	0	0	0	0	2	14	0	0	0	1	0	0	0	0	0	0
25	58,77	0,320	1	46	0	1	0	0	0	0	0	0	0	0	2	15	0	0	0	1	0	0	1	0	0	0
26	60,20	0,412	1	20	0	0	0	1	0	0	0	0	0	0	4	8	0	0	0	0	0	0	0	0	0	1
27	23,68	0,129	2	59	1	0	0	0	0	0	0	0	0	0	1	21	1	1	1	1	1	0	1	1	0	0
28	60,42	0,329	2	30	0	1	0	0	0	0	0	0	0	0	2	28	0	1	0	1	1	0	1	1	1	0
29	65,16	0,398	2	13	0	0	0	1	0	0	0	0	0	0	4	17	1	0	0	1	0	0	0	0	0	0
30	54,73	0,298	1	52	0	1	0	0	0	0	0	0	0	0	2	10	0	0	0	0	0	0	1	1	1	0
31	20,19	0,110	1	52	0	1	0	0	0	0	0	0	0	0	2	11	0	0	0	1	0	0	1	0	1	0
32	20,92	0,114	1	60	1	0	0	0	0	0	0	0	0	0	1	9	0	0	0	0	0	0	0	0	0	1
33	52,34	0,285	1	50	1	0	0	0	0	0	0	0	0	0	1	7	0	0	0	0	0	0	0	0	0	1
34	46,28	0,252	2	54	0	1	0	0	0	0	0	0	0	0	2	32	1	1	1	1	1	0	1	0	1	0
35	56,75	0,309	2	25	0	1	0	0	0	0	0	0	0	0	2	14	1	1	0	1	0	0	1	1	1	0
36	64,78	0,408	1	23	0	0	0	1	0	0	0	0	0	0	4	13	0	0	0	1	0	0	0	0	0	0
37	65,59	0,357	1	22	0	1	0	0	0	0	0	0	0	0	2	13	1	0	0	1	0	0	1	0	0	0
38	57,14	0,311	1	69	1	0	0	0	0	0	0	0	0	0	1	11	0	1	0	1	0	0	1	0	0	0
39	62,10	0,338	2	18	0	1	0	0	0	0	0	0	0	0	2	17	1	0	0	1	0	1	1	1	1	0
40	60,81	0,331	2	35	0	1	0	0	0	0	0	0	0	0	2	12	1	1	0	1	0	0	0	0	0	0
41	57,69	0,314	2	48	0	1	0	0	0	0	0	0	0	0	2	17	0	0	0	1	0	0	1	0	1	0
42	58,24	0,317	2	77	1	0	0	0	0	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	1
43	57,32	0,312	2	42	0	1	0	0	0	0	0	0	0	0	2	10	1	0	0	0	0	0	1	0	1	0

Ready | База 1 (обща | C356,V96 | Set:OFF | Weight:OFF | CAPS | NUM | ЗАП

Пуск | Total Commander 6.53 - ... | Microsoft PowerPoint - [... | Adobe Photoshop CS3 E... | Adobe Reader - [Клинич... | Microsoft Excel - база мо... | STATISTICA - [Data: ... | RU | 23:19

Резюме

Существует ряд правил построения электронных таблиц для обеспечения их максимальной совместимости с программами, выполняющими статистическую обработку:

1. Случаи располагаются в строках, переменные – в столбцах;

2. Случаи должны быть уникальными, т.е. каждая строка таблицы должна соответствовать одному уникальному пациенту. Соответственно, каждый случай должен иметь уникальный (неповторяющийся) идентификатор (порядковый номер);

3. Заголовки столбцов должны быть уникальными (неповторяющимися), короткими (не длиннее 10-12 символов) и, желательно, набранными латиницей (допустимо употребление цифр, дефисов и знаков подчеркивания);

5. Значения всех переменных, вносимые в таблицу, должны быть числовыми; символьные значения («да», «нет» и т.п.) не допускаются. В том случае, если переменные являются качественными/порядковыми, т. е. по природе своей требуют словесного описания, их необходимо формализовать, т.е. ввести схему цифрового кодирования описательных признаков и строго ее придерживаться в ходе заполнения базы;

6. Сложные качественные переменные необходимо разбивать на более простые с вариантами значений «1» (есть данное состояние) и «0» (нет данного состояния);

7. При заполнении переменных, содержащих даты, необходимо придерживаться единого формата (например, дд/мм/гггг). При внесении в ячейки таблицы цифровых значений необходимо следить за тем, чтобы точность указанных значений была единообразной в пределах одной переменной (например, всюду

8. По возможности следует избегать пустых ячеек на месте отсутствующих данных; в таких случаях лучше использовать специальные коды, резко отличающиеся от всех возможных значений учитываемого признака (например, 9999);

9. После заполнения электронную таблицу обязательно необходимо проверить на предмет неправильно внесенных данных. Обычно встречающиеся при этом ошибки:

– значения пропущены либо сдвинуты;

– случайное *изменение формата ячеек* (например, дата или текст вместо числа);

– *формат даты* не соответствует принятому для данной базы;

– значения дат *не соответствуют срокам выполнения исследования* (вариант: возраст пациентов выходит за рамки, оговоренные для исследования);

– *не соблюдена оговоренная точность* указания результатов замеров (указано больше либо меньше знаков после запятой, чем необходимо);

– *не все данные в номинальных либо порядковых переменных формализованы* (помимо числовых данных, в таблицу внесены текстовые);

– *идентификаторы случаев неуникальны* (повторяются);

– *нарушение принятой схемы кодировки* качественных либо порядковых переменных (указаны ошибочные коды).

Как правило, все подобные ошибки легко вычленяются при внимательном неоднократном осмотре электронной таблицы; возможна также автоматизированная проверка при помощи формул и специальных «проверочных» переменных.