СТАТЬИ

УДК 519.25:[57+61]

КАК ПРОПУСКИ В МЕДИЦИНСКИХ ДАННЫХ МОГУТ ВЛИЯТЬ НА РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ?

Мантрова А.И.

ФГБОУ ВО «Омский государственный медицинский университет», Омск, e-mail: 0329110315@mail.ru

В статистике и в математике интенсивно ведется разработка методов анализа данных с пропусками. Но в экспериментальной и клинической медицине эта проблема недооценивается. Наиболее опасным видом неслучайных пропусков является гибель части особей (пациентов или подопытных животных) в опытной группе от изучаемой патологии. Обычно затем выжившие особи опытной группы сравниваются по интересующему исследователя показателю с контрольной группой, в которой гибели не было. Различие между этими двумя группами традиционно трактуется в медицине как следствие изучаемой патологии. Такая трактовка является некорректной, поскольку различие между группами может быть вызвано не только изучаемой патологией, но и пропусками в данных, вызванными этой патологией. Проблема неслучайных пропусков и обусловленного ими искажения результатов и выводов исследований наиболее актуальна в тех отраслях медицины, где выше летальность: в реаниматологии, травматологии, неотложной кардиологии и т.п. В обзоре принято деление всех медико-биологических измерений на разрушающие и неразрушающие. Рассмотрены современные статистико-математические, а также методические и организационные способы борьбы со смещениями, вызываемыми неслучайными пропусками. Подчеркивается необходимость внедрения этих методов в медицинские исследования.

Ключевые слова: пропуски, данные с пропусками, неслучайные пропуски, смещение, медицинские исследования

HOW MISSING VALUES IN MEDICAL DATA CAN IMPACT RESEARCH RESULTS?

Mantrova A.I.

Omsk State Medical University, Omsk, e-mail: 0329110315@mail.ru

In statistics and in mathematics, methods for analyzing of missing data are being intensively developed. But in experimental and clinical medicine this problem is underestimated. The most dangerous case of missing not at random is the death of a part of individuals (patients or experimental animals) in the experimental group of the pathology under study. Usually, then, the surviving individuals of the experimental group are compared for the indicator of interest to the indicator with a control group in which there was no death. The difference between these two groups is traditionally interpreted in medicine as a consequence of the pathology under study. This interpretation is incorrect, since the difference between the groups can be caused not only by the pathology being studied, but also by missing values caused by this pathology. The problem of missing not at random and the resulting bias of the results and conclusions of research is most relevant in those branches of medicine where mortality is higher: in resuscitation, traumatology, emergency cardiology, etc. The review adopted the division of all biomedical measurements into destructive and non-destructive. Modern statistical-mathematical, as well as methodological and organizational methods of dealing with biases caused by missing not at random are considered. It emphasizes the need to introduce these methods in medical research.

Keywords: missing values, missing data, missing not at random, bias, medical research

При статистической обработке результатов медицинских исследований основную массу данных составляют обычные *полные* наблюдения, когда от подопытных животных или обследуемых пациентов получают результаты, представляющие собой определенные цифровые значения. Но иногда в цифровых данных возникают пропуски, то есть такие наблюдения, о которых известен лишь факт их существования, но ничего не известно об их цифровых значениях. Например, пропуск в данных возникает в том случае, когда взятый от животного или человека биоматериал оказывается случайно утраченным до его исследования. Нечто среднее между полными наблюдениями и пропусками представляют собой *цензурированные* наблюдения – об их цифровых значениях известна лишь часть информации. Например, если онкологический больной, у которого изучалось влияние нового вида лечения на выживаемость, через 2 года после начала лечения погиб в автокатастрофе, то его результат выживаемости можно записать как «не менее 2 лет» или «2 года*», где звездочка (согласно ГОСТ 27.504-84 и последующим стандартам) представляет собой пометку о том, что данное наблюдение является цензурированным [1].

Статистические методы анализа данных с пропусками и анализа цензурированных данных идеологически и терминологически близки. Но для статобработки цензурированных данных к настоящему времени создано уже немало эффективных методов, чего, к сожалению, нельзя сказать о данных с пропусками. Актуальность проблемы влияния пропусков в данных на результаты

и выводы медицинских исследований отмечается в ряде публикаций [2–4]. Особую остроту эта проблема имеет в тех случаях, когда пропуски в данных вызваны гибелью животных или людей от изучаемой патологии и носят неслучайный характер.

Цель исследования: провести обзор литературных данных по проблеме влияния неслучайных пропусков на результаты медицинских исследований, а также gроанализировать способы решения этой проблемы.

Материалы и методы исследования: поиск, обработка и анализ отечественной и зарубежной литературы по тематике настоящего обзора.

Поиск публикаций осуществлялся в информационно-телекоммуникационной сети Интернет: в зарубежных электронных базах данных «EMBASE» и «PUBMED» (в базу «PUBMED» как ее главная составная часть входит «MEDLINE»), а также в российской научной электронной библиотеке «e-LIBRARY.RU». Поскольку в данной области медицинской статистики еще не сложилась общепринятая русскоязычная и англоязычная терминологии, ниже приведены термины, использованные для поиска публикаций:

– данные с пропусками – англ. missing data,

– пропуски или пропущенные значения (наблюдения) – англ. missing values,

– цензурированные значения (наблюдения) – англ. censoring values,

– цензурирование (процесс формирования цензурированных данных и данных с пропусками) – англ. censoring,

– случайные пропуски – англ. missing at random, missing-at-random,

– неслучайные пропуски – англ. missing not at random,

– неслучайное цензурирование (то есть неслучайный механизм формирования пропусков и цензурированных наблюдений) – англ. informative censoring,

– смещение (в результатах и выводах исследования) – англ. bias,

– потеря части данных – англ. loss of data, dropout,

– импутация (замещение пропущенного значения его оценкой) – англ. imputation.

Результаты исследования и их обсуждение

Проблему неслучайных пропусков в данных, вызванных изучаемой патологией, можно проиллюстрировать следующим примером [5]. Предположим, в опыте на крысах исследуется влияние какой-либо тяжелой патологии на содержание в мозге некоторого вещества X. Животные были в случайном порядке разделены на две группы: 1) опытную, у которой моделировали изучаемую патологию, приведшую к гибели (а говоря языком статистики – к цензурированию, ведущему к образованию пропусков в данных) 40 % крыс; 2) контрольную, у которой моделировали не изучаемую патологию, а только сопутствующие экспериментальные воздействия (наркоз, фиксация и т.п.). В контрольной группе гибели животных, естественно, не было.

Пусть среднее содержание вещества X в мозге крыс опытной группы оказалось на 55 % ниже, чем в контрольной (значимость различий между группами $P < 0,05$). Обычно эти результаты интерпретируются так: «изучаемая патология привела к статистически значимому снижению содержания вещества X на 55 % по сравнению с контрольной группой». В экспериментах с пропусками в данных подобные формулировки являются некорректными. Можно ведь предположить, что снижение показателя произошло не вследствие патогенетических и компенсаторных процессов, запущенных изучаемой патологией, а просто из-за того, что животные с высокими содержаниями вещества X оказались менее устойчивыми и погибли (представим, что у 40 % погибших крыс как раз и были самые высокие уровни вещества X в группе), а животные со средними и низкими содержаниями выжили. В этом гипотетическом случае, если у выживших особей содержание вещества X в мозге осталось таким же, каким оно было до моделирования патологии, то за счет отсева (пропусков) высоких значений возникает иллюзия снижения показателя, и это кажущееся снижение может быть ошибочно истолковано как «открытие нового звена в патогенезе изучаемой патологии».

Если же в рассмотренном примере погибли крысы с самыми низкими уровнями показателя, тогда истинное снижение вещества X под влиянием патологии составляет по сравнению с контрольной группой не 55 %, а намного большую величину. То есть в этом случае тоже происходит искажение результатов, но уже в другую сторону. Причина искажений в обоих случаях состоит в том, что группу выживших особей сравнивают с контрольной группой, в которой перемешаны «потенциально выжившие» и «потенциально погибшие» особи.

Есть еще вариант, что гибель части особей происходит таким образом, что среднее значение (медиана, средняя арифметическая) исследуемого показателя или другие его статистические параметры в группе не меняются. В принципе такое возможно,

и тогда мы действительно получим изменения, обусловленные только патологическим процессом. Но в этом случае получается, что исследуемый показатель, хотя и изменяется при данной патологии (в силу каких-то второстепенных, побочных процессов), но никак не влияет на вероятность для организма выжить или умереть. Поэтому, исследуя такой показатель, вряд ли можно обнаружить важное звено патогенеза изучаемого заболевания или добиться снижения смертности от него. Получается, что при пропусках, вызванных гибелью особей от изучаемой патологии, ценность того или иного показателя как объекта для исследования тем выше, чем сильнее влияют на него пропуски, вызванные изучаемой патологией (неважно, отсеивают они наибольшие значения показателя или наименьшие). Ситуации, когда различия между группами вызваны только влиянием пропусков или только изучаемым патологическим процессом, представляют собой логически возможные крайности. В реальности эти два фактора чаще всего действуют совместно, а итоговые изменения представляют собой сумму их влияний [4, 5]. Проблема в том, как статистически оценить долю каждого из этих факторов в итоговой сумме.

Неслучайный характер пропусков, вызванных изучаемой патологией, был обоснован в предыдущем абзаце логически, исходя из того, что ни один исследователь не станет изучать заведомо никчемные показатели, никак не влияющие на вероятность выжить или умереть. Но неслучайность пропусков можно установить и статистически: например, в работе [2] это было сделано по результатам многомерного анализа (англ. «multivariate analysis»).

Очевидно, влияние неслучайных пропусков, вызванных изучаемой патологией, на результаты исследования будет тем больше, чем выше летальность от данной патологии. Поэтому проблема пропусков и вызванных ими смещений в результатах исследований наиболее актуальна в отраслях медицины, имеющих дело с экстремальными и терминальными состояниями, то есть в травматологии [6, 7], реаниматологии [8–10], неотложной кардиологии [11, 12] и т.п.

Также весьма актуальна проблема потери части данных в исследованиях по экспериментальной фармакологии [13–15] и клинической фармакологии [16, 17]. В США в 2010 г. был выпущен специальный нормативный акт [17], в котором определены 3 градации пропусков в данных по клинической фармакологии в зависимости от степени их случайности (полностью случайные пропуски, случайные пропуски и неслучайные пропуски) и даны указания по их статистической обработке [16, 17]. Однако специалисты в этой области продолжают выражать озабоченность тем, что пропуски в данных по клиническим испытаниям лекарств до сих пор часто или не признаются серьезной проблемой, или считаются неприятностью, которую лучше игнорировать [12].

В клинической наркологии признано, что пропуски в данных являются важной методологической проблемой и что необходимо проводить исследования с использованием различных методов статистической обработки данных с пропусками [18].

Еще одним примером может служить исследование в области клинической гинекологии [19], в котором статистически анализировались дневники симптомов эндометриоза, предоставляемые больными женщинами. Часть данных отсутствовала, а те, которые были получены, можно расценивать как цензурированные данные или как данные с пропусками. Для их статистической обработки были применены специальные методы [19]. Известно, что пропуски в данных можно рассматривать как крайний (предельный) случай цензурированных наблюдений [1].

Как отечественными, так и зарубежными авторами признано, что не существует универсального алгоритма для статистико-математической обработки данных с пропусками [16, 20]. Несмотря на многолетнюю разработку методов анализа данных с пропусками, пока не найдено удовлетворительных решений для многих медицинских задач. Как следствие, например, в работе [18] параллельно использовалось сразу 6 методов обработки данных с пропусками.

Одним из распространенных методических подходов является импутация (замещение) пропущенного значения его статистической оценкой, то есть значением, полученным по сохранившимся (непропущенным) значениям, которые наиболее близки к пропущенному. Среди таких методов весьма перспективной представляется множественная импутация (англ. «multiple imputation»), примененная, в частности, в работах [7, 21, 22]. Идеологически близок к ней минимаксный метод, основанный на анализе логически возможных крайних вариантов влияния пропусков на результаты исследования [5, 23, 24]. В работе [6] основной принцип анализа пропущенных значений заключался в том, чтобы не заменять их вычисленными оценочными значениями или наиболее близкими из сохранившихся (непропущенных) значений, а также не исключать пропуски из статобработки,

а включить пропуски в математическую модель в качестве отдельной категории. Существуют и другие статистико-математические способы обработки пропущенных данных [16]. Но, кроме сложных методов, есть полезные простые рекомендации. Например, в экспериментах с неслучайными пропусками, вызванными изучаемой патологией, корреляционный анализ внутри каждой группы статистически более корректен, чем сравнение опытных групп с контрольной [25, 26].

Проблему данных с пропусками можно рассматривать и в контексте нарушения рандомизации [5]. В медико-биологическом исследовании, если измерение показателя не требует забоя животного (такие измерения называются неразрушающими), то его измеряют у каждой особи до моделирования заболевания и после, получая таким образом связанные (парные) наблюдения, которые обрабатывают статистическими методами, предназначенными для таких наблюдений. Если же для измерения изучаемого показателя необходима эвтаназия животного (такие измерения называются разрушающими), тогда с соблюдением правил рандомизации формируются опытная и контрольная группы, отличающиеся лишь наличием у опытных животных изучаемой патологии. Это есть вынужденный прием, пойти на который заставляет невозможность измерить показатель дважды у одной и той же особи. Корректность и высокая эффективность такого приема в экспериментах с полными выборками привели к его механическому переносу и на те эксперименты, в которых часть особей не доживает до забоя из-за тяжести исследуемой патологии. При этом рандомизированность эксперимента грубо нарушается. Принадлежность показателя к разрушающим или неразрушающим не является его неотъемлемым свойством, а может меняться в зависимости от характера исследования и принципа измерения [5].

Проблему искажающего влияния неслучайных пропусков на результаты исследования можно решать не только статистическими, но и в ряде случаев методическими и организационными способами. Под методическими способами понимается прежде всего применение биопсийных и неинвазивных методик (в частности, методов так называемой дистантной химии), что позволяет перевести изучаемый показатель из категории разрушающих в категорию неразрушающих. В качестве примера можно привести неинвазивное определение фосфоросодержащих метаболитов *in vivo* при помощи ядерного магнитного резонанса.

Рассмотрим примеры организационного решения проблемы неслучайных пропусков [5]. Предположим, нужно изучить, как изменяется содержание вещества X в крови у мужчин в процессе старения. Для этого из одной и той же популяции с соблюдением правил рандомизации формируются группы мужчин в возрасте 50–55, 55–60 и 60–65 лет и в крови у них определяется содержание вещества X. Ясно, что различие среднего уровня вещества X в этих выборках может быть обусловлено не только его возрастными изменениями, но и пропусками: естественной смертью части мужчин, которая увеличивается в каждой следующей возрастной группе, а также гибелью их от несчастных случаев или выбыванием из-под наблюдения. Существует альтернативная организация работы – измерение вещества X у каждого взятого в исследование 50-летнего мужчины вплоть до достижения им 65 лет (или до смерти). Однако при этом необходимо, чтобы все пробы крови анализировались в одинаковых условиях.

В тех случаях, когда в опытах на животных возможно количественное варьирование тяжести изучаемой патологии, можно сформировать несколько опытных групп, различающихся по степени тяжести патологии, а следовательно и по проценту летальности (проценту пропусков, вызванных изучаемой патологией). Зависимость изменений величины измеряемого показателя (измерение которого является разрушающим) от величины процента пропусков можно затем установить при помощи статистических методов, в частности регрессионного и дисперсионного анализа. Например, можно использовать двухфакторный регрессионный анализ, где один фактор – степень тяжести патологии, второй фактор – процент пропусков, а отклик – уровень показателя, требующего эвтаназии животных (разрушающее измерение).

Как отмечалось в начале настоящего обзора, в медико-биологических исследованиях кроме неслучайных пропусков могут возникать также случайные пропуски в данных, например нечаянная утрата части биоматериала еще до проведения измерений. Для того, чтобы подобные пропуски были действительно случайными (не влияли на результаты исследования), нужно проводить обработку материала в рандомизированном порядке, как и все остальные этапы работы.

Заключение

Таким образом, из проведенного обзора литературы можно сделать вывод, что в последние годы интенсивно ведется разра-

ботка методов статистико-математического анализа данных с пропусками. Но наибольшую опасность, которая в медицине недооценивается, представляют неслучайные пропуски, вызванные гибелью части особей от изучаемой патологии. Отличие такой группы от контрольной группы традиционно трактуется только как следствие изучаемой патологии, но это отличие может быть вызвано также и пропусками. Эта проблема тем острее, чем выше летальность от изучаемой патологии. Необходимо внедрение в медицинские исследования современных статистико-математических, а также методических и организационных способов борьбы со смещениями, вызываемыми неслучайными пропусками.

Список литературы

1. Золин П.П. Цензурированнные данные и данные с пропусками в медицинских исследованиях // Патол. физиология и эксперим. терапия. 2010. № 4. С. 49–52.

2. Bech C.N., Brabrand M., Mikkelsen S., Lassen A. Risk factors associated with short term mortality changes over time, after arrival to the emergency department. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine. 2018. vol. 26. P. 29. DOI: 10.1186/s13049-018-0493-2.

3. Fuchs P.A., Del Junco D.J., Fox E.E., Holcomb J.B., Rahbar M.H., Wade C.A., Alarcon L.H., Brasel K.J., Bulger E.M., Cohen M.J., Myers J.G., Muskat P., Phelan H.A., Schreiber M.A., Cotton B.A. Purposeful variable selection and stratification to impute missing Focused Assessment with Sonography for Trauma data in trauma research. Journal of Trauma and Acute Care Surgery. 2013. vol. 75, no. 1 (Suppl. 1). P. S75–S81. DOI: 10.1097/TA.0b013e31828fa51.

4. Золин П.П., Лебедев В.М., Конвай В.Д. Математическое моделирование биохимических процессов с применением регрессионного анализа: монография. Омск: Издательство Омского государственного университета, 2009. 344 с.

5. Золин П.П. Проблема цензурированных выборок в медико-биологических исследованиях. Омск: Омская государственная медицинская академия, 1997. Деп. в ВИНИТИ 26.12.1997. № 3789-В97. 22 с.

6. Lefering R., Huber-Wagner S., Nienaber U., Maegele M., Bouillon B. Update of the trauma risk adjustment model of the TraumaRegister DGU™: the Revised Injury Severity Classification, version II. Critical Care. 2014. vol. 18. P. 476. DOI:10.1186/s13054-014-0476-2.

7. Seleno N., Vogel J., Liao M., Hopkins E., Byyny R., Moore E., Gravitz C., Haukoos J. Denver trauma organ failure score outperforms traditional methods of risk stratification in trauma. Academic Emergency Medicine. 2012. vol. 19. Suppl. 1. P. S144.

8. Золин П.П. Статистическая обработка цензурированных выборок при изучении экстремальных и терминальных состояний // Патогенез, клиника и терапия экстремальных и терминальных состояний. Омск, 1998. С. 39–43.

9. Золин П.П. Проблема цензурированных выборок в экспериментальной медицине // Патол. физиология и эксперим. терапия. 2000. № 1. С. 23–25.

10. Золин П.П. Постреанимационные нарушения обмена гипоксантина и их коррекция: дис. ... канд. мед. наук. Омск, 2002. 250 с.

11. Fabian-Jessing B.K., Vallentin M.F., Secher N., Hansen F.B., Dezfulian C., Granfeldt A., Andersen L.W. Animal models of cardiac arrest: A systematic review of bias and reporting. Resuscitation. 2018. vol. 125. P. 16–21. DOI: 10.1016/j.resuscitation.2018.01.047.

12. Krantz M.J., Kaul S. The ATLAS ACS 2–TIMI 51 trial and the burden of missing data. J. Am. Coll. Cardiol. 2013. vol. 62/ no. 9. P. 777–781. DOI: 10.1016/j.jacc.2013.05.024

13. Золин П.П., Конвай В.Д., Ефременко Е.С., Старун А.С., Семочкин А.В., Жукова О.Ю., Мантрова А.И., Домрачев А.А. Рибоза не влияет на уровень средних молекул в сердце крыс в постреанимационном периоде // Естественные и технические науки. 2018. № 9. С. 16–18.

14. Золин П.П., Конвай В.Д., Ефременко Е.С., Старун А.С., Семочкин А.В., Жукова О.Ю., Мантрова А.И., Домрачев А.А. Изучение влияния рибозы на уровень молекул средней массы в крови реанимированных крыс // Современные проблемы науки и образования. 2018. № 6. [Электронный ресурс]. URL: http://www.science-education.ru/article/view?id=28211 (дата обращения: 11.02.2019).

15. Соколова Т.Ф., Золин П.П. Статистическая обработка цензурированных данных в оценке эффективности фармакологической коррекции посттравматических нарушений // Омский научный вестник. 2002. № 18. С. 53–55.

16. Little R.J., D'Agostino R., Cohen M.L., Dickersin K., Emerson S.S., Farrar J.T., Frangakis C., Hogan J.W., Molenberghs G., Murphy S.A., Neaton J.D., Rotnitzky A., Scharfstein D., Shih W.J., Siegel J.P., Stern H. The prevention and treatment of missing data in clinical trials // New Engl. J. Med. 2012. vol. 367. P. 1355–1360. DOI: 10.1056/NEJMsr1203730.

17. National Research Council. The prevention and treatment of missing data in clinical trials. Washington, DC: National Academies Press, 2010. 162 p.

18. Witkiewitz K., Falk D.E., Kranzler H.R., Litten R.Z., Hallgren K.A., O'Malley S.S., Anton R.F. Methods to analyze treatment effects in the presence of missing data for a continuous heavy drinking outcome measure when participants drop out from treatment in alcohol clinical trials. Alcohol Clin. Exp. Res. 2014. vol. 38. no. 11. P. 2826–2834. DOI: 10.1111/acer.12543.

19. Seitz C., Lanius V., Lippert S., Gerlinger C., Haberland C., Oehmke F., Tinneberg H.-R. Patterns of missing data in the use of the endometriosis symptom diary. BMC Womens Health. 2018. vol. 18. no. 1. P. 88. DOI: 10.1186/s12905-018-0578-0.

20. Куликова К.Ю. Статистический анализ медицинских данных при наличии пропусков // Процессы управления и устойчивость. 2014. Т. 1 (№ 1). С. 253–258.

21. Parvez B., Shah A., Muhammad R., Shoemaker M.B., Graves A.J., Heckbert S.R., Xu H., Ellinor P.T., Benjamin E.J., Alonso A., Shintani A.K., Roden D., Darbar D. Replication of a risk prediction model for ambulatory incident atrial fibrillation using electronic medical record. Circulation. 2012. vol. 126. P. 21.

22. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann. Transl. Med. 2016. vol. 4. no. 2. P. 30. DOI: 10.3978/j.issn.2305-5839.2015.12.63.

23. Золин П.П. Анализ данных с пропусками: на примере изучения метаболизма углеводов и липидов в печени в постреанимационном периоде // Омский научный вестник. 2009. № 1. С. 43–45.

24. Золин П.П., Лебедев В.М., Конвай В.Д. Регрессионные модели метаболизма: монография. Саратов: Ай Пи Эр Медиа, 2018. 310 с.

25. Zolin P.P., Konvai V.D. Disturbances of hypoxanthine metabolism in the liver of resuscitated rats. Bul. Exper. Biol. Med. 1997. vol. 124. no. 6. P. 1180–1182.

26. Золин П.П., Конвай В.Д. Механизмы нарушений метаболизма гипоксантина в печени реанимированных крыс // Бюл. экспер. биол. 1997. Т. 124. № 12. С. 629–631.