

Тихова Г.П.
Пропуск данных в выборке:
как решать проблему и как ее избежать

*ГБОУ ВПО «Петрозаводский государственный университет»,
185910, Петрозаводск*

Статья посвящена проблеме пропуска данных в клинических исследованиях и испытаниях. Рассмотрены три механизма, ответственных за возникновение пропущенных данных в выборке. Подробно рассмотрен каждый из них, его влияние на репрезентативность выборки и величину смещения результатов. Указаны пути снижения вероятности и количества пропущенных данных на этапе планирования исследования и на стадии статистической обработки и формулирования заключений.

Ключевые слова: *клиническое исследование, пропуск данных.*

Для цитирования: Тихова Г.П. Пропуск данных в выборке: как решать проблему и как ее избежать. *Регионарная анестезия и лечение острой боли.* 2016; 10 (3): 205–209. DOI: 10.18821/1993-6508-2016-10-3-205-209.

Для корреспонденции: Тихова Галина Петровна, научный сотрудник лаборатории клинической эпидемиологии Института высоких биомедицинских технологий, ГБОУ ВПО «Петрозаводский государственный университет», Петрозаводск, e-mail: tikhovag@gmail.com.

Tikhova G.P.

DATA MISSING: HOW TO SOLVE AND HOW TO ESCAPE THE PROBLEM

Petrozavodsk State University named after O.V. Kuusinen, 185910, Petrozavodsk

The article is devoted to the problem of missing data in clinical trials and clinical studies. The author considered three mechanisms of generating of missing data in collected sample. Each mechanism type is reviewed in details in terms of its effects on sample representativeness and the magnitude of result bias. The ways to reduce probability and amount of missing data are pointed in the phase of planning and on the stage of statistical data processing and inference.

Key words: *clinical study, missing data. MAR, MCAR, MNAR.*

For citation: Tikhova G.P. Data missing: how to solve and how to escape the problem. *Regionarnaya anesteziya i lechenie ostroy boli* (Regional Anesthesia and Acute Pain Management, Russian journal) 2016; 10 (3): 205–209. (In Russ.). DOI: 10.18821/1993-6508-2016-10-3-205-209.

For correspondence: Galina Tikhova, senior researcher, Laboratory of clinical epidemiology, Institute of highest biomedical technologies, Petrozavodsk State University, 195910, Petrozavodsk, e-mail: tikhovag@gmail.com.

Conflict of interest. The authors declare no conflict of interest.

Funding. The study had no sponsorship.

Received 21 July 2016
Accepted 10 August 2016

Прежде чем мы начнем разговор о проблеме пропуска данных, нам надо договориться о терминах и, в частности, дать определение самому предмету обсуждения. Что такое пропуск данных в клиническом исследовании? Если в исследовании измеряются (определяются, рассчитываются) несколько переменных, описывающих наблюдаемых пациентов, то отсутствие значения одной и более переменных в конкретном наблюдении называется пропуском данных. Эта проблема очень часто встречается в клинических исследованиях и даже в хорошо

спланированных и четко проведенных клинических испытаниях. В больших мультицентровых клинических исследованиях процент пропуска данных иногда достигает более 10% [1]. На этапе планирования клинического испытания его возможная доля учитывается при расчете необходимого объема выборки и мощности статистических методов, применяемых для анализа данных. К сожалению, эта проблема довольно редко обсуждается в публикациях по обработке клинических данных. Вероятно, по умолчанию считается, что она неизбежна и в целом

не очень существенно влияет на чистоту исследования и величину смещения результатов. Однако ни первое, ни второе предположения совершенно не верны. Попытаемся привести убедительные аргументы в защиту тезиса о важности и распространенности проблемы пропуска данных в рутинной практике клинических исследований.

Кто виноват?

Перечислим некоторые аспекты клинических исследований, которые могут значимо влиять на количество пропущенных данных:

1. Пропуск данных в связи с прекращением действия исследуемой процедуры (препарата, фактора риска) из-за развития побочных эффектов или неэффективности лечения у некоторых пациентов из набранной выборки. Для каждого такого пациента причины выбывания из исследования до его завершения могут быть различны и, кроме того, они могут быть не единственными, а комбинированными друг с другом. Однако это может быть и одна и та же причина, скажем, непереносимость препарата пациентами с определенной сопутствующей патологией. Таким образом, наша итоговая выборка будет смещена (не репрезентативна), т. е. она не будет содержать определенную категорию больных, которые являются частью целевой исследуемой популяции.

2. Пропуск данных в связи с их неопределенностью, т.е. невозможностью точно определить значение конкретной переменной для данного пациента на некотором этапе исследования. Например, пациент с кардиопатологией не может выполнить необходимый нагрузочный тест или возрастной пациент не в состоянии заполнить необходимый опросник. Большинство таких случаев пропуска данных обусловлено состоянием здоровья пациента на момент сбора данных для исследования.

3. Пропуск данных, обусловленный неадекватной длительностью исследования или слишком большим набором переменных, регистрируемых в каждом наблюдении пациента. Это в большей степени касается проспективных когортных исследований, которые могут длиться несколько лет, а иногда даже десятилетий. Известен случай утраты огромной клинической базы данных по больным диабетом в Новом Орлеане во время нашумевшего урагана «Катрина». Это исследование проводилось более 10 лет и после урагана его фактически пришлось начинать с нуля. Но неприятности такого рода могут встретиться и в краткосрочном исследовании с большим количеством переменных, определяемых у пациента при каждом наблюдении. Существует негласное эмпирически полученное правило:

- для продольных клинических исследований: чем длиннее запланирован период исследования, тем более вероятен пропуск данных;
- для поперечных клинических исследований: чем больше переменных описывает каждое наблюдение пациента, тем более вероятен пропуск данных.

При планировании исследования необходимо найти баланс между желаемой детализацией информации об исследуемой процедуре (препарате, факторе риска) и допустимой долей пропущенных данных.

4. Пропуск данных, обусловленный выбором составной конечной точки [2]. Нередко в качестве первичной конечной точки выбирается некий индекс, который рассчитывается по нескольким регистрируемым переменным, например, индекс массы тела (для его расчета необходимо знать вес и рост пациента), коэффициент атерогенности (вычисляется только при известных значениях ЛПВП и ЛПНП), есть расчетные интегральные индексы, требующие и большего количества исходных переменных. В случае такой конечной точки пропуск значения хотя бы одной исходной переменной приводит к невозможности получить величину исследуемого показателя для данного пациента. Чем большее количество исходных переменных требуется для вычисления конечной точки, тем более вероятен пропуск ее значения.

Однако все же почему мы должны так внимательно относиться к проблеме пропуска данных и грамотно выбирать методы ее решения в каждом конкретном случае? Потому что неуправляемый, неконтролируемый или просто игнорируемый пропуск данных является источником существенного смещения результатов и, следовательно, может приводить к ошибкам в заключениях или вообще к ложным выводам.

Попытки решения этой проблемы клинических исследований предпринимались в течение всего прошедшего столетия, но настоящий прорыв произошел в 1970-х гг., когда интерес к этой проблеме привлек внимание специалистов по математической статистике [3]. В результате появился первый вариант таксономии пропущенных данных, опубликованный в 1976 г. в работе Rubin [4]. В дальнейшем предложенный вариант был развит в универсальную систему классификации [3], используемую в настоящий момент во всех протоколах клинических испытаний и исследований. Rubin сформулировал понятие *механизма пропуска данных* и выделил 3 его варианта, которые определяют тактику борьбы с этой проблемой и возможность с ней справиться в конкретном клиническом исследовании. Механизм пропуска данных описывает взаимосвязь

(если таковая существует) между переменными и вероятностью пропуска их значений в данном исследовании. В работе Rubin выделено 3 варианта таких взаимосвязей, и они обозначены английскими аббревиатурами, широко используемыми сейчас в литературе: MAR, MCAR и MNAR.

Механизм пропуска данных MAR

Тип пропущенных данных по механизму MAR (Missing At Random, «случайный пропуск») встречается в исследованиях в том случае, когда тенденция (склонность) изучаемой переменной Y к пропуску ее значений при регистрации всего пула признаков не зависит от ее собственной величины, но связана с величинами других показателей, регистрируемых одновременно с ней. Другими словами, не существует связи между значением переменной Y и вероятностью того, что это значение будет пропущено во время очередного наблюдения после фиксации значений всех остальных регистрируемых переменных. Лучше всего объяснить этот механизм на конкретном примере, хотя и несколько искусственном.

Предположим, что среди пациентов блока интенсивной терапии некоторой клиники для определения нутритивного статуса рутинно применяется расчет индекса массы тела (ИМТ). В тех случаях, когда значение ИМТ выходит за нижнюю границу референсного интервала, больным проводятся дополнительные лабораторные исследования по определению уровня альбумина, трансферрина и преальбумина в крови. Однако больным, у которых ИМТ превышает нижний предел, дополнительные лабораторные исследования не проводятся. У нескольких больных с пониженным ИМТ также могут отсутствовать результаты лабораторных исследований, т.к. по различным непредвиденным обстоятельствам у них не получилось произвести забор крови (были переведены в другую клинику, отказались от дальнейшего обследования и т.п.). Таким образом, если исследователь решит провести ретроспективное клинико-эпидемиологическое исследование распространенности белково-энергетической недостаточности среди пациентов данного блока ИТ, он получит набор данных по альбумину, трансферрину и преальбумину пациентов, который определенно имеет пропуски. Зависят ли пропуски данных от значений самих изучаемых данных (альбумина, трансферрина или преальбумина)? Нет. Однако если мы обратимся к значениям ИМТ, то здесь определенно наблюдается связь вероятности отсутствия лабораторных данных с величиной ИМТ. Пропуск значений, например, альбумина, не связан с величиной самого альбумина, но связан с другой переменной, которая регистрируется

одновременно с ним, а именно ИМТ. Такой тип пропуска данных и называется MAR – «случайный пропуск».

Термин MAR несколько вводит в заблуждение, поскольку связь отсутствует только между вероятностью пропуска значения переменной Y и самими значениями этой переменной, однако эта вероятность может быть связана и весьма тесным образом со значениями других переменных, измеряемых одновременно с Y , т.е. она является не строго случайной, как мы увидели в примере, приведенном выше.

Механизм пропуска данных MCAR

Более строгое условие, чем MAR, определяет механизм пропуска данных типа MCAR (Missing Completely At Random, «полностью независимый пропуск»). В этом случае пропуски данных рассматриваются как абсолютно случайные события. Формальное определение механизма MCAR требует полной независимости вероятности пропуска данных переменной Y как от значений других регистрируемых переменных, так и от самой Y .

Если исследователь из вышеприведенного примера решит выделить группу пациентов с пониженным ИМТ и проанализировать данные по распространенности белково-энергетической недостаточности только у этой выделенной подгруппы, то он не заметит никакой связи пропуска данных с какими бы то ни было значениями переменных, в том числе и с ИМТ, они будут встречаться в выборке совершенно случайным образом. Такой массив данных демонстрирует пример пропуска данных по типу MCAR.

Надо отметить, что если исследователь предполагает в своей работе наличие механизма пропуска данных по типу MCAR, он имеет возможность проверить свое предположение. Для этого необходимо отделить записи пациентов с пропущенными данными по заданной переменной от данных пациентов, в которых пропуски отсутствуют, а затем рассчитать средние значения регистрируемых переменных в обеих группах и сравнить их с помощью статистического критерия. Если средние не имеют статистически значимого различия, то пропуски в переменной Y имеют механизм типа MCAR. Если же различие статистически значимо, это указывает на наличие некоторой связи между значениями переменных и вероятностью пропуска значений Y .

Механизм пропуска данных MNAR

Надо признаться, что MNAR (Missing Not At Random, «неслучайный пропуск») – самый неприятный и разрушительный для исследования механизм

пропуска данных, поскольку он может свести на нет все усилия исследователя достичь поставленной цели и сделать надежные заключения по результатам исследования. Пропуск данных типа MNAR имеет место в том случае, если вероятность пропуска значений переменной Y связана с собственным значением этой переменной даже тогда, когда остальные переменные набора данных зафиксированы на константах, иными словами, находятся под контролем и их влияние на вероятность пропуска исключено.

В качестве примера можно привести достаточно распространенную для когортных исследований ситуацию, когда пациенты принимают некоторый препарат, для которого проводится экологическое клиническое исследование частоты побочных эффектов. Однако не все согласившиеся принять участия в данном исследовании хорошо его переносят. Некоторые больные отказываются продолжать лечение или врач отменяет назначение именно по причине возникновения побочных эффектов у этих больных. В данной ситуации частота возникновения побочных эффектов исследуемой терапии не будет отражать реальной картины, т.к. пропуском данных конечной точки исследования управляет сама эта конечная точка. Это типичный пример механизма пропуска данных по типу MCAR, полное отсутствие случайности в пропуске значений исследуемой переменной, в данном случае частоты побочных эффектов, обусловленных приемом препарата.

Что делать?

Существует две очевидные линии борьбы с проблемой пропуска данных:

- разрабатывать дизайн и проводить исследование таким образом, чтобы максимально ограничить количество пропущенных данных. Есть большое число методик, которые успешно справляются с этой проблемой и позволяют существенно снизить риск пропуска данных. Эти разработки, к сожалению, не используются так широко, как этого можно было бы ожидать;
- применять в исследовании методы анализа, которые используют частичную информацию, заключенную в полученных данных, с тем, чтобы прогнозировать пропущенные значения и уменьшить ошибку смещения, порожденную пропуском.

Конечно, первый вариант решения проблемы гораздо предпочтительнее, но он требует внимания к нашей проблеме еще на стадии планирования клинического исследования, а также строгого

мониторинга выполнения заявленного протокола исследования. Если исследователь вспомнил о проблеме пропущенных данных, только столкнувшись с ней лицом к лицу на этапе статистической обработки измерений, то ему ничего не остается, как проанализировать механизм, породивший эти пропуски и затем воспользоваться второй линией защиты своего исследования от ошибок смещения. Для этого можно использовать два интуитивно понятных метода решения уже существующей проблемы:

- исключить запись (пациента) с пропущенными данными из обработки;
- внести фиктивные данные вместо пропущенных, спрогнозировав их специальными математическими методами аппроксимации, скользящего среднего, сплайнами или методом максимального правдоподобия.

Большинство статистических пакетов поддерживают эти методы решения проблемы, но какими бы они ни были современными, точными и продвинутыми, первая линия обороны от проблемы пропуска всегда даст лучшие и более надежные результаты. Следовательно, наилучший способ оградить себя от нее – озаботиться ею еще при разработке дизайна и определении конечных точек. Для этого надо:

- в первую очередь, проанализировать первичные конечные точки исследования, на основе которых делаются заключения и формируются выводы, на предмет вероятной невозможности определить (или зарегистрировать) их значения, когда другие переменные или показатели (обычно факторы воздействия) принимают конкретные значения;
- определить, может ли исследуемый фактор (факторы) воздействия принимать значения на всем протяжении своего допустимого интервала или возможны такие комбинации других изучаемых переменных, когда часть спектра его значений становится недоступной (не определена, нет возможности зарегистрировать) в ходе испытания (или исследования);
- если пропуск данных для какой-то переменной вероятен в связи с невозможностью определения ее значения при некоторых вполне реальных условиях (например, по состоянию здоровья пациента, мы говорили об этом в п.2), то можно попробовать заменить ее другой величиной, регистрация которой не будет зависеть от этих условий;
- найти оптимальный баланс между детализацией информации, достаточной для формулирования обоснованных выводов, и максимальной полнотой собранных данных. Необходимо, чтобы исследование было не только информативным, но и выполнимым в реальных условиях конкретной клиники.

Проблема пропуска данных при проведении клинического исследования – достаточно обширная тема, чтобы в деталях изложить ее полностью в одной публикации. Мы лишь обозначили некий контекст, основные моменты, на которые необходимо обратить пристальное внимание, чтобы избежать существенных ошибок смещения, и, следовательно, больших неприятностей от неправильных заключений, сделанных на основе искаженной картины, которая получилась из-за неконтролируемого или просто проигнорированного пропуска данных в собранной исходной выборке. Однако эта проблема на самом деле не так сложна, как может показаться на первый взгляд. Она просто требует к себе постоянного внимания на всех этапах работы. Если следовать этому правилу педантично и четко, анализировать свои данные на предмет выявления скрытых опасностей пропуска значений, начиная с разработки протокола и заканчивая статистической обработкой, то картина всегда будет ясной и ситуация с количеством пропущенных измерений будет оставаться под контролем.

Финансирование. Исследование не имело спонсорской поддержки.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Литература

1. Mirkes E.M., Coats, T.J., Levesley J., Gorban, A.N. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in Biology and Medicine*. 2016; 75: 203–16.
2. Тихова Г.П. Планируем клиническое исследование. Вопрос 2: Выбор конечных точек. *Регионарная анестезия и лечение острой боли*. 2014; 10(4): 67–70.
3. Enders C.K. *Applied Missing Data Analysis*. New York: Guilford Press; 2010
4. Rubin D.B. Inference and Missing Data. *Biometrika*. 1976; 63(3): 581–92.

References

1. Mirkes E.M., Coats, T.J., Levesley J., Gorban, A.N. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in Biology and Medicine*. 2016; 75: 203–216.
2. Tikhova G.P. Planning clinical research. Question #2: Choosing of Endpoint. *Regionarnaya anesteziya i lechenie ostroy boli [Regional anesthesia and acute pain management]*. 2014; 10(4): 67-70. (in Russian).
3. Enders C.K. *Applied Missing Data Analysis*. New York: Guilford Press; 2010.
4. Rubin D.B. Inference and Missing Data. *Biometrika*. 1976; 63(3): 581–92.

Поступила 21.07.16

Принята к печати 10. 08.16