

ПРИЛОЖЕНИЕ

к Рекомендации Коллегии
Евразийской экономической комиссии

от 20 г. №

РУКОВОДСТВО

**по обеспечению качества данных, полученных в условиях
реальной клинической практики для целей регулирования
обращения лекарственных препаратов**

I. Общие положения

1. Настоящее Руководство является рамочным документом, который позволяет подготовить специальные рекомендации по каждому из аспектов обращения лекарственных препаратов таким образом, чтобы обеспечить надлежащее качество полученных в условиях реальной клинической практики (РКП) данных медицинских исследований, с заданными параметрами и характеристиками.

2. Настоящее Руководство содержит определения, положения и рекомендации, которые могут применяться к любому источнику данных в целях характеристики, оценки и обеспечения качества данных для обоснования принятия решений в сфере обращения лекарственных средств уполномоченными органами государств – членов Евразийского экономического союза (далее соответственно – регуляторные решения, государства-члены, Союз). Система качества данных, описанная в настоящем Руководстве позволяет обеспечить первичное и вторичное использование данных медицинских исследований, а также

использование метаданных и вспомогательной информации (например, информация об управлении мастер-данными (основными данными) и базовые справочные данные), необходимых для обоснования принятия регуляторных решений. На рисунке 1 в обобщенном виде представлены 4 вида основных характеристик системы качества данных (СКД).

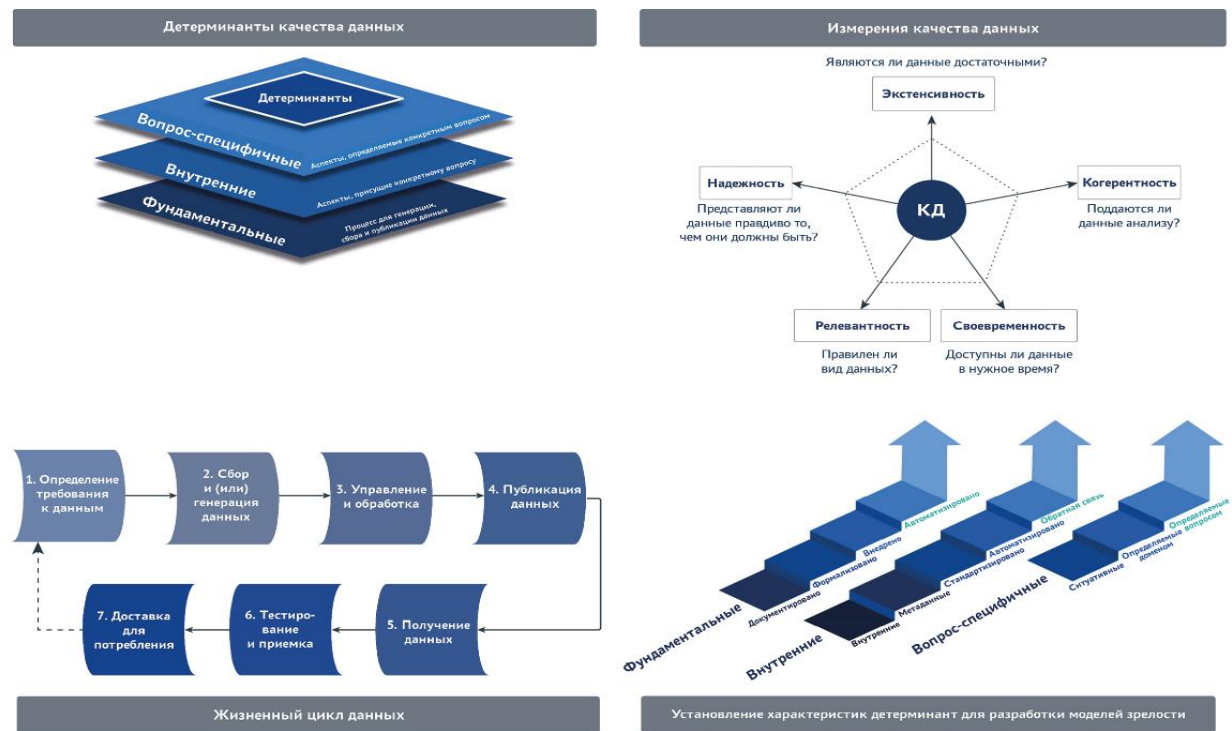


Рисунок 1. Основные характеристики системы качества данных медицинских исследований

3. Настоящее Руководство применяется уполномоченные органами (экспертными организациями) государств-членов и субъектами обращения лекарственных средств в целях регулирования обращения лекарственных препаратов, но допускается его применение иными сторонами, участвующими в применении (использовании) лекарственных препаратов (например, научные медицинские сообщества, держатели источников данных, страховые компании, исследователи и ассоциации пациентов).

4. Поскольку определения, методы и характеристики системы качества данных зависят от типов и источников данных, система качества

данных позволяет обеспечить единый подход к выявлению, идентификации и дальнейшей разработке процедур оценки качества данных и рекомендаций в отношении существующих и новых типов данных.

5. Целями создания системы качества данных при принятии регуляторных решений являются обеспечение:

согласованности процессов, связанных с качеством получаемых данных;

развития горизонтальных систем качества данных;

условий для более эффективного и автоматизированного использования данных.

6. В отличие от вертикальной системы качества данных, в которой все процессы разрабатываются исходя из сложившейся ситуации и предназначены для конкретного случая использования, горизонтальная система качества данных позволяет получить набор определенных функциональных возможностей пригодный для различных случаев практического применения.

II. Сфера применения

7. Настоящее Руководство распространяется на широкий спектр регуляторных решений:

планирование и проведение клинических исследований;

сбор биоаналитических данных;

доклинические исследования;

сбор сведений о спонтанных сообщениях о нежелательных явлениях;

данные химического и производственного контроля и др.

III. Определения

8. Для целей настоящего Руководства используются понятия, которые означают следующее:

«актуальность» – особый аспект своевременности, который учитывает степень свежести данных (например, актуальность и непосредственную полезность);

«внутренняя детерминанта» – характеристика, присущая конкретному набору данных (например, форматирование данных), которую можно рассматривать как детерминанту, зависящую от конкретного набора данных;

«вопрос-специфичная для контекста детерминанта» – аспект качества данных, который напрямую зависит от его предполагаемого или фактического использования и который невозможно определить в отрыве от контекста (конкретного вопроса или конкретного подхода к анализу);

«вторичное использование данных» – использование данных, которые изначально были собраны для других целей (например, для статистического учета, медицинских научных исследований и др.) и которые могут включать в себя медицинские записи, электронные медицинские данные;

«детерминанты качества данных» – факторы, влияющие на качество данных или их характеристику и подразделяющиеся на три категории: фундаментальные детерминанты, внутренние детерминанты, вопрос-специфичные для контекста детерминанты;

«доступность данных» – возможность доступа к данным для общественного использования с точки зрения их обнаруживаемости, экспортируемости и удобства использования;

«измерения качества данных» – аспекты качества данных разделенные на 5 различных групп (экстенсивность, когерентность, своевременность, релевантность и надежность) каждая из которых отвечает на определенные вопросы, касающиеся данных и может быть разделена на субизмерения;

«качество данных» – сумма нескольких особенностей (аспектов, характеристик) данных, начиная от их соответствия реальности и заканчивая их представлением;

«когерентность (согласованность, последовательность)» – измерение, которое выражает, насколько разные части итогового набора данных последовательны в отношении представления и значений этих данных;

«краткость данных» – характеристика данных, которые должны быть представлены в сжатом виде, в том числе свойство данных, заключающееся в том, что они содержат только существенную, а не ложную информацию;

«метаданные» – набор данных, который описывает другие данные («данные о данных») и предоставляет информацию об их контексте относительно назначения, местонахождения, ключевых переменных, способа формирования, формата и владельца набора данных, публикуемый в каталогах данных, с целью обеспечения возможности поиска данных и проверки их пригодности для конкретных задач без раскрытия самих данных;

«метрики качества данных» – индикаторы, которые можно применять к источнику данных для оценки одного или нескольких измерений качества;

«модель зрелости» – система оценки процессов, технологий и структуры организации или функционального подразделения, в которой

определены фундаментальные детерминанты на каждом уровне, служащие ориентиром для постоянного повышения качества данных с течением времени, и которая обеспечивает структурированный подход к оценке эффективности управления процессами, политиками и практиками обеспечения качества данных в организации или отдельном подразделении;

«надежность» – измерение, характеризующее, насколько точно данные отражают то, что они непосредственно измеряют;

«неизменяемость данных» – свойство данных которое подразумевает, что данные никогда не удаляются и не изменяются с того момента как только данные «зафиксированы» (например, внесены в базу данных), но могут только дополняться (в том числе новой информацией, призванной опровергнуть или заменить предыдущие данные), при этом данные, введенные в систему (от которых могут зависеть другие данные или действия), нельзя изменять без явного указания нового состояния информации и сохранения сведений о предыдущем состоянии;

«первичное использование данных (первичное использование (электронных) медицинских данных)» – обработка медицинских записей для оказания медицинских услуг с целью оценки, поддержания или восстановления состояния здоровья человека, которому они принадлежат, в тех конкретных целях, для которых они были собраны;

«правильность» (accuracy) – степень расхождения между полученными данными и реальными данными;

«прецизионность (точность)» – степень приближения, с которой данные отражают реальность (например, возраст человека может быть указан в годах или месяцах);

«прослеживаемость» – характеристика данных, которая относится к категории надежности, поскольку она связывает измеряемые и фактические данные;

«своевременность» – наличие данных в нужное время для принятия регуляторных решений (сбор и предоставление данных в приемлемые сроки);

«система качества данных» – совокупность понятий, принципов и рекомендаций, предназначенная для оценки и управления качеством данных, охватывающая широкий спектр источников данных в целях определения характеристик, оценки и обеспечения качества данных при принятии регуляторных решений;

«сущность» – совокупность однородных значений, относящихся к конкретной переменной (например, масса тела), также называемая «уровень строки»;

«управление мастер-данными» – система, которая помогает объединять данные из разных систем, обеспечивать их безопасность и очищать (удалять дубликаты и др.), чтобы работать с нужной информацией;

«фундаментальная детерминанта» – характеристика генерации данных, которая связана с формированием данных (процессы генерации, сбора, обработки и перевода данных в доступную форму) и влияет на качество данных, но не является частью самих данных (например, программные системы, обучение, процессы аудита);

«целостность данных» – обеспечение и поддержание надежности и согласованности данных на протяжении всего жизненного цикла данных;

«экстенсивность (расширяемость)» – измерение, определяющее количество доступных данных.

IV. Качество данных исследований и структура системы качества данных

9. Данные исследований считаются качественными если:

их получение, соответствует целям и потребностям пользователей в отношении проводимых медицинских исследований необходимых для принятия регуляторных решений;

они корректно отражают реальность, для фиксации которой они предназначены.

10. Система качества данных для принятия регуляторных решений состоит из двух частей, характеризующих разные стадии процесса составления спецификаций.

11. Первая часть (общая система) обеспечивает согласованный подход к качеству данных, охватывающий широкий спектр типов данных и расширяемый с учетом новых случаев использования полученных данных в процессе принятия регуляторных решений. Такой подход обеспечивает общую основу для разных аспектов качества данных, применяемых к разным типам данных и сценариям их использования: определения, измерения качества данных и примеры метрик, охватывающих эти измерения. Кроме того, в рамках такого подхода следует определить общие принципы применения процессов оценки качества данных и представить набор моделей зрелости, способствующих повышению уровня автоматизации принятия регуляторных решений на основе полученных данных.

12. Вторая часть (специализированная система) предназначена для детализации и расширения общей системы в отношении конкретных типов данных или регуляторных решений. Эта часть является основой для подготовки применимых на практике руководств, которые в

дальнейшем будут изменяться (актуализироваться) по мере изменения данных и развития медицинских технологий.

13. Настоящее Руководство содержит положения по формированию системы качества данных для уполномоченных органов государств-членов которая предназначена для применения к данным исследований реальной клинической практики и доказательств, полученных на их основе. В настоящем Руководстве приводятся общие принципы, определения и принципы, касающиеся качества данных применительно к регулированию обращения лекарственных препаратов. Систему качества данных следует регулярно пересматривать в целях совершенствования ее применения при принятии регуляторных решений.

V. Общие подходы, лежащие в основе поддержания и оценки качества данных

1. Детерминанты (факторы) качества данных для генерации доказательств

14. Структура данных, которые потенциально могут быть использованы в регуляторных целях, распространяется на разнообразные источники данных, каждый из которых формируется в ходе разных процессов и подходит для разных первичных и вторичных случаев использования. При оценке общего качества набора данных на этапе принятия регуляторных решений важно различать, что именно влияет на качество, а также что и на каком этапе можно измерить или контролировать. В системе качества данных подобные элементы, приведенные на рисунке 2 и связанные с качеством данных, называются «детерминантами (факторами)» (далее – детерминанты) и делятся на три категории:

а) фундаментальные детерминанты. Влияют на качество данных, но не являются частью самих данных. Они не зависят от содержания набора данных и не могут быть полностью выведены из него. Чтобы при принятии регуляторных решений данным можно было доверять, инфраструктура и процессы, лежащие в их основе и обеспечивающие сбор, хранение, преобразование и передачу данных, следует разработать таким образом, чтобы не нарушалось соответствие между данными и параметрами реального объекта, который они представляют. Примерами фундаментальных детерминант являются:

сертифицированные системы программного обеспечения для сбора и обработки данных;

процедуры, программы обучения и процессы аудита, позволяющие обеспечить надлежащую запись и документирование данных;

процессы валидации и верификации стадий обработки данных;

б) внутренние детерминанты. Их можно получить на основе набора данных и, возможно, некоторых внешних общих знаний, но без учета контекста, в котором эти данные были сгенерированы, а также контекста, в котором данные будут использоваться (например, научный вопрос или регуляторное решение). Примерами внутренних детерминант являются когерентное (согласованное) или некогерентное (несогласованное) форматирование, наличие ошибок (например, усечение) или правдоподобность данных;

в) вопрос-специфичные для контекста детерминанты. Примерами таких детерминант являются приемлемость полноты набора данных или степень их приближенности, (например, дата, выраженная в днях или месяцах) необходимые для получения ответа на конкретный вопрос.

15. В целом фундаментальные детерминанты оказывают непосредственное влияние на качество данных. Когда их невозможно

контролировать, единственным вариантом является контроль внутренних аспектов качества данных. Пределы подобного контроля ограничены его способностью обеспечивать соответствие цели, если вопрос (или совокупность типичных вопросов) не сформулирован.



Рисунок 2. Детерминанты качества данных

2. Качество данных в ходе получения (сбора) доказательств

16. Данные, пригодные и доступные для получения (сбора) доказательств, получают в процессе (являющимся частью более широкого «жизненного цикла»), специфичном для типа данных, процессов и организаций, которые производят эти данные. В отношении уже собранных данных (вторичное использование) оценку пригодности для конкретной цели следует проводить до начала этого процесса.

В качестве примера можно привести общий жизненный цикл, представленный на рисунке 3, и включающий в себя:

определение требований к данным: какие данные требуются и какими должны быть их характеристики. На этом этапе для первичных

данных могут быть определены элементы, непосредственно связанные с получением (сбором) доказательств;

получение (сбор) данных: получение данных, отражающих наблюдаемую реальность;

управление данными и их обработка: включая передачу, нормализацию и очистку данных;

публикация данных: обеспечение доступности данных для пользователей;

сбор и агрегирование данных: получение данных от одного или нескольких пользователей;

тестирование и приемка: оценка пригодности полученных данных для целевых пользователей;

предоставление для использования: передача данных в целях осуществления конкретных действий (например, анализа) с этими данными.

17. Не все вышеперечисленные этапы присутствуют во всех рабочих процессах обработки данных (например, данные, полученные с датчиков или из социальных сетей, могут собираться исходя из того, что доступно, а не на основе конкретных требований), при этом могут применяться дополнительные этапы, а их порядок может отличаться.

18. Для целей оценки и управления качеством данных следует определить, какие именно детерминанты применимы на каждом этапе этого процесса, а также какое воздействие они могут оказывать. Например, если внутренние аспекты качества данных можно измерить, то такие измерения следует использовать для повышения надежности на этапе сбора и генерации данных, либо для оценки качества на момент публикации. Интеграция с дополнительными данными потребует повторной оценки. Специфические детерминанты качества, зависящие от

конкретного вопроса, следует оценивать каждый раз, когда данные перепрофилируются для ответа на новый вопрос, для которого они изначально не собирались или не предназначались.

19. Проверки качества данных происходят на различных стадиях процесса генерации доказательств и могут включать итеративные циклы обратной связи, как показано пунктирной линией на рисунке 3.

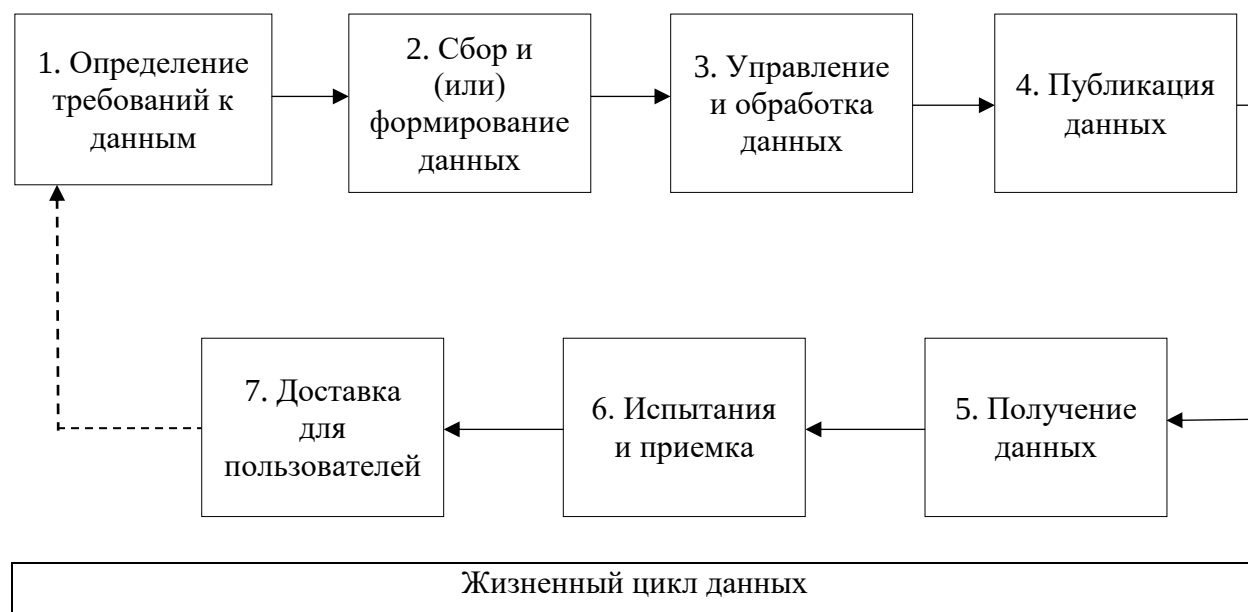


Рисунок 3. Типичный алгоритм обработки данных в процессе формирования доказательств

Качество данных в сравнении со стандартизацией

20. С точки зрения принятия регуляторных решений качество данных отличается от стандартизации данных: данные, не пригодные для цели с точки зрения ответа на регуляторный вопрос, не станут пригодными при стандартизации, а нестандартизированные данные допускается использовать при принятии регуляторного решения. Качество данных также применяется к отдельным и нестандартным источникам данных.

21. На внедрение систем и процессов, обеспечивающих качество данных, в значительной степени влияет (а в некоторых случаях полностью определяет) принятие стандартов, а также рекомендации по управлению данными и наличие таких ресурсов, как систематизация и системы управления мастер-данными, которые лежат в основе смысловой совместимости. Следует отметить, что процессы стандартизации данных могут изменять оригинальную информацию и их смысл.

22. В связи с этим рекомендации в отношении конкретных стандартов и процессов стандартизации не включены в настоящую систему, хотя принятие стандартов определяет уровни зрелости внедрения.

Первичное и вторичное использование данных

23. При применении руководящих принципов и метрик возникает важное различие между первичным и вторичным использованием данных. Если системы разрабатываются для сбора и обработки данных для определенной первичной цели, или существует набор установленных требований для вторичного использования, то внутренние и вопрос-специфичные аспекты качества данных могут быть учтены уже во время сбора и генерации данных. Таким образом, можно разработать системы и процессы, которые гарантируют определенный уровень качества, необходимый для генерации доказательств. Фактически, большую часть решений по анализу и спецификации процесса допускается принимать на этой стадии, а последующий анализ следует посвятить обобщению результатов и оценке степени неопределенности. Как правило, это не относится к вторичному использованию данных (запланированному или спонтанному), при котором критерии качества для использования могут

не совпадать с критериями, применимыми к первоначальным целям сбора данных. В этих случаях качество данных часто можно контролировать только с помощью внутренних детерминант (факторов).

Публикация и использование данных

24. На протяжении всего жизненного цикла данные обрабатываются в двух различных контекстах. В рамках одного (опубликование) данные генерируются или собираются, обрабатываются (иногда гармонизируются) и становятся доступными. Примерами являются объединение данных из нескольких источников с целью формирования набора данных, доступного для общего пользования (например, в каталоге) или генерацию данных носимыми устройствами и их «публикацию» через средства программного интерфейса. В рамках другого (использование) данные собирают и обобщают для проведения анализа. Одним из примеров может служить исследование, в рамках которого собираются, объединяются и согласовываются данные из различных источников для ответа на конкретный вопрос.

25. Эти два контекста могут накладываться друг на друга (например, когда данные собираются для конкретного первичного использования) или могут быть совершенно разными (например, когда данные собираются и публикуются в каталоге для ряда возможных предусмотренных или непредусмотренных вариантов использования, обычно для вторичного анализа).

26. Это различие следует учитывать, поскольку цель и потенциал для оценки качества между этими двумя контекстами меняются. Даже внутренние аспекты качества для одного и того же набора данных могут различаться. Детальная спецификация оценки качества может быть разработана отдельно для контекстов опубликования или использования

данных, например, для каталога данных – с точки зрения приемлемого минимального качества для рутинного использования, или для получения данных – с точки зрения минимальной пригодности для решения конкретного вопроса.

Данные и метаданные

27. Метаданными считаются данные более высокого уровня о данных более низкого уровня, которые предоставляют информацию о их назначении и происхождении (например, характеристики источников, этапы обработки данных, цепочка происхождения и определения элементов данных). Когда данные состоят из числовых или неструктурированных сведений (например, изображений), метаданные, как правило, предоставляются в качестве дополнения к набору данных. В целом граница между данными и метаданными не является четкой: некоторая информация, выступающая в качестве метаданных в одном контексте (например, поставщик приборов для испытания), может рассматриваться как данные в другом (например, при оценке систематической ошибки измерения).

28. При принятии регуляторных решений метаданные, как правило, следует рассматривать так же, как и данные. Если какое-либо изменение в метаданных потребует пересмотра выводов, сделанных на их основе, то с точки зрения оценки качества данных его следует рассматривать как данные. В контексте качества данных метаданные не следует рассматривать исключительно как набор метрик (в том числе метрик качества данных) и краткое описание наборов данных. Они должны также включать характеристику источников, процессов и определения элементов данных.

29. Все виды метаданных часто публикуются в каталогах данных, цель которых – обеспечить возможность поиска данных и проверки их пригодности для конкретных задач без раскрытия самих данных

3. Неизменяемость (иммутабельность) данных

30. Данные о каких-либо измеряемых или наблюдаемых аспектах реальности могут меняться со временем, отражая как фактическое изменение наблюдаемых сущностей, так и изменение доступной информации в некоторый момент времени. Например, масса тела человека может изменяться как вследствие изменений в самом организме, так и в результате замены предыдущих измерений более точными измерениями.

31. Важно отличать текущую доступность данных (или текущие знания о реальности) от данных, которые были доступны в определенный момент времени (конкретная запись знаний о реальности). Последние данные, отражающие то, что было известно в определенный момент времени, «неизменяемы» (иммутабельны) в том, что они (по определению) не могут измениться.

32. Для любого регуляторного решения доказательства следует основывать на данных, предназначенных для фиксации состояния знаний на определенный момент времени. Другими словами, данные, используемые для поддержки регулирования, должны быть неизменяемыми (иммутабельными). Это не означает, что доказательства никогда не могут быть обновлены: любое обновление следует рассматривать как отдельный (пусть и связанный) набор данных. Это является основополагающим принципом, на котором следует основывать принятие регуляторных решений.

33. Следствием этого принципа является то, что данные, используемые для принятия решений, должны иметь версии и оставаться неизменными в пределах каждой конкретной версии.

4. Данные в сравнении с информацией

34. В самом строгом определении данные представляют собой необработанные факты или наблюдения (например, полученные (собранные) с помощью приборов), а информация – это выводы и новые факты, основанные на этих данных, после того как они были поняты и обработаны в своем контексте (например, ответ пациента на вмешательство в отличие от набора отдельных показаний прибора, наблюдаемых для этого ответа).

35. Система качества данных сосредоточена на получении (сборе) доказательств, которые могут быть использованы при принятии регуляторных решений, и поэтому она выходит за рамки разграничения между данными и информацией.

5. Система отсчета (валидация в сравнении с верификацией)

36. Некоторые аспекты качества данных могут быть измерены относительно разных референсов, содержащихся в одном и том же наборе данных или существующих за пределами набора данных либо в качестве общего референса или внешнего «золотого стандарта», либо в качестве реального факта в реальном мире. Например, масса тела человека может быть верифицирована на предмет качества на основе ее фиксации в данных (например, как недостающее значение), на основе знания естественного диапазона массы или верифицирована на соответствие массе, указанной в первичном документе.

37. В некоторых системах оценка качества внутри набора данных называется верификацией, а оценка по отношению к исходной записи (или внешнему «золотому стандарту») – валидацией.

Понятие валидации не следует путать с валидацией как формой проверки степени согласованности полученных результатов (согласно пункту 102).

6. Детализация (структурированность) данных и качество данных

38. В отношении структурированных данных качество данных обычно оценивается на разных уровнях детализации:

уровень значения соответствует конкретной точке данных (например, массе тела). Его также называют уровнем ряда, когда речь идет обо всех значениях, относящихся к одному и тому же физическому объекту (явлению). Понятие «физический объект (явление)» (далее – сущность) используется для обозначения субъекта, характеризующегося определенным набором знаний. В информационной записи (например, об образце), каждое значение в записи будет отражать измерение переменной, относящейся к одному и тому же субъекту. Сущность как правило обозначается в записи при помощи идентификатора;

уровень переменной («уровень столбца») охватывает одну точку данных для всей совокупности индивидуумов (например, масса тела как переменная в таблице демографических данных клинического исследования). Метрики качества данных на уровне значений для бинарных или категориальных данных в большинстве случаев легко перевести в «уровень столбца» (например, путем преобразования значений этих данных в проценты);

уровень набора данных охватывает всю совокупность связанных наблюдений. Как правило набор данных, предназначенный для решения конкретного вопроса будет состоять из однородных данных (например, конкретных измерений) или из разрозненных типов данных, которые связаны между собой тем, что они измеряют (прямо или косвенно) одни и те же сущности. В этом смысле разные части наборов данных «связаны» между собой, поскольку в них содержатся общие ссылки на одни и те же сущности.

В некоторых контекстах в рамках одного набора данных можно провести дополнительное различие между его частями, относящимися к схожим сущностям. Когда проводится такое различие эти части называют табличным уровнем, поскольку они обычно отображаются в отдельных таблицах.

39. Концепция детализации также применима к неструктурированным данным, но определение ее уровней обычно зависит от конкретного типа данных и поэтому не рассматривается в настоящей общей системе.

40. Система качества данных будет сосредоточена на самом низком возможном уровне, то есть в случае структурированных данных – на уровне значений. Однако некоторые метрики могут быть определены только на более высоком уровне. Например, правдоподобность одной записи о человеке с массой тела 300 кг может не вызвать нарушения метрики, но если 80 % записей содержат информацию о наличии людей с массой тела 300 кг, то нарушение произойдет.

VI. Измерения и метрики качества данных

41. Определение измерений и метрик качества данных основано на общем определении измерений, метрик и мер:

измерение представляет собой один или несколько связанных аспектов или особенностей реальности (например, для физического объекта – его линейные размеры или прочность);

метрика представляет собой способ оценки значения конкретной особенности (например, абсолютной длины, измеряемой в метрах при некоторых оговоренных обстоятельствах);

мера представляет собой только один вариант использования метрики (например, 2 метра). Несколько мер можно объединить, чтобы получить более общие метрики (например, среднюю длину). Меры допускается указывать без единицы измерения. Допускается использовать в качестве мер величины, которые не являются непрерывными.

42. Метрики качества данных можно определить, как индикаторы, которые, при применении к источнику данных, позволяют оценить одно или несколько измерений качества данных. Единичная метрика качества данных может использоваться в качестве индикатора для нескольких измерений (в соответствии с примерами для когерентности (согласованности, последовательности) в таблице 3).

43. Для некоторых метрик могут быть определены пороговые значения приемлемости (например, максимальный процент недостающих значений). В случае непреднамеренного вторичного использования такие пороговые значения приемлемости могут быть определены только в зависимости от конкретного вопроса. Однако, в тех случаях, когда данные собираются для первичного использования или для некоторых четко определенных вторичных целей, можно определить пороговые значения (например, минимальные и (или) максимальные), которые применяются уже в момент сбора данных.

44. Особенности (аспекты, характеристики) качества данных следует классифицировать по независимым измерениям, позволяющим раскрыть качество данных (то есть разные измерения должны описывать разные отдельные вопросы качества данных). Несколько систем данных предлагают организацию качества данных в измерениях, которые схожи между разными системами, но часто не имеют единых точных определений. Это затрудняет согласованную оценку качества данных при объединении данных из нескольких источников.

45. В настоящем Руководстве представлен набор измерений качества данных (приведен на рисунке 4), имеющих значение с точки зрения принятия регуляторных решений, а также приводятся их точное определение, возможные метрики и примеры. Цель состоит в том, чтобы устранить неоднозначность и создать полезный источник справочных данных, который поможет свести различные концепции качества, полученные из различных источников к общему основанию. Это будет полезно для формирования метрик и моделей зрелости в целях обеспечения надлежащей доказательной базы.



Рисунок 4. Измерения качества данных

1. Надежность

46. Надежность отвечает на вопрос о том, насколько точно данные отражают наблюдаемую реальность. При рассмотрении определения качества как «соответствия назначению» надежность отражает степень правильности и достоверности данных.

Субизмерения надежности

47. Исходя из определения надежности выделяются следующие субизмерения:

правильность (accuracy) (далее – реальность). Определение правильности объединяет в себе меру количества неверной информации в наборе данных (данные, систематически не отражающие реальность) и

формальную величину правильности измерений (например, разницу между измерениями и реальным значением). Например, масса тела человека может быть указана неверно из-за ошибки транскрипции данных или того, что человек взвешивался полностью одетым, что дает систематическое превышение массы тела на 1 – 2 кг;

прецизионность (точность). Определение прецизионности (точности) включает в себя воспроизводимость значений при повторных измерениях, поскольку оно отражает, насколько «грубым» является несоответствие между данными и характеристикой, которую предполагается измерить.

Другие концепции качества данных, связанные с надежностью

48. Другие концепции связаны с релевантными аспектами качества данных, которые подпадают под измерение или широко используются, но не строго соответствуют представленному определению измерения.

49. С надежностью строго связана концепция правдоподобия, определяемая как вероятность того, что некоторая информация является истиной. Правдоподобие может служить средством обнаружения ошибок: если некоторая комбинация информации маловероятна (или невозможна) в реальном мире, это свидетельствует о наличии проблем с правильностью. Например, масса тела человека, превышающая 300 кг, возможна, но масса тела многих или всех людей в наборе данных, превышающая это значение, неправдоподобна (если только фундаментальные детерминанты (факторы) не указывают на обратное) и, вероятно, указывает на ошибки в измерении или обработке данных. Правдоподобие возникает в результате сравнения элемента данных с типичными или необходимыми характеристиками сущности, которую

она должна представлять, и поэтому ее трудно измерить как чисто внутреннюю характеристику, поскольку она зависит от наличия фоновых знаний или внешнего «золотого стандарта».

50. Прослеживаемость (преемственность или происхождение данных) по отношению к данным, представляет собой знание о том, как эти данные были сформированы, из какого источника они были получены и какую обработку прошли, прежде чем предстали в своем нынешнем виде.

Подходы в отношении надежности

51. Надежность фундаментально зависит от имеющихся систем и процессов, используемых для первичного сбора данных, их обработки и хранения на последующих стадиях процесса формирования доказательной базы как для первичного, так и вторичного использования.

52. При отсутствии ошибок правильность не снижается в процессе агрегирования данных. Напротив, прецизионность (точность) может снизиться при гармонизации данных с общей моделью данных (ОМД), поскольку для конкретного использования данных в той или иной модели может потребоваться менее прецизионное (точное) представление, чем в исходных источниках. Например, общая модель данных, позволяющая ставить временные метки в секундах, в то время как источник данных использует миллисекунды, или общая модель данных, использующая общую терминологию, которая в отдельных областях менее специализирована, чем терминология, используемая в первоисточнике.

53. Внутренние аспекты надежности трудно измерить в рамках системы, ориентированной исключительно на данные, однако измерения правдоподобия могут служить средством для обнаружения некоторых классов ошибок. Надежность не зависит от конкретного вопроса, хотя

для каждого вопроса, в зависимости от данных, устанавливается пороговое значение приемлемой надежности.

Таблица 1

Примеры метрик надежности

Субизмерение	Метрическая группа	Абстрактная метрика	Оценка	Пример
Правдоподобие (косвенный показатель правильности) Включает в себя в качестве примера только правильность, поскольку это наиболее распространенное применение правдоподобия. Теоретически, правдоподобие может распространяться и на другие измерения надежности (например, масса, выраженная в граммах в виде целого числа, скорее всего, будет неточной, если все полученные значения массы представляют собой	атемпоральное правдоподобие (правдоподобие, оцениваемое вне зависимости от времени)	значения и распределение данных согласуются с внутренними измерениями (measurements) или локальными сведениями	валидация	рост и масса тела имеют положительное значение
				число уникальных субъектов на вмешательство соответствует ожиданиям (относительно применимого «золотого стандарта»)
		значения и распределения данных для независимых измерений (measurements) одного и того же факта согласуются	верификация	оральная и аксиллярная температура примерно схожи
				уровень глюкозы при измерении в сыворотке крови подобен уровню глюкозы, измеренной

Субизмерение	Метрическая группа	Абстрактная метрика	Оценка	Пример
числа заканчивающиеся тремя нулями).				в капиллярной крови, с помощью глюкометра
		логические ограничения между значениями согласуются с общепринятыми знаниями об измеряемом объекте (процессе)	верификация	пол пациента соответствует половому контексту, характерному для заболевания или состояния (беременность, рак предстательной железы)
		значения повторного измерения одного и того же факта демонстрируют ожидаемую вариабельность	верификация	значения массы тела схожи, если измерения проводят разные медсестры в одном и том же медицинском учреждении с использованием одного и того же оборудования
		значения и распределения данных согласуются с проверенными	валидация	значения HbA _{1c} , полученные в стационаре и референсной

Субизмерение	Метрическая группа	Абстрактная метрика	Оценка	Пример
		референтными стандартами		лаборатории государства-члена, статистически совпадают при одинаковых условиях
		эквивалентные значения идентичных измерений (measurements) получены из двух независимых баз данных, содержащих одни и те же наблюдения с одинаковой достоверностью	валидация	значение даты рождения в истории болезни не совпадает со значением в регистровой записи того же пациента
		две или более зависимые базы данных дают схожие значения для идентичных переменных (например, база данных 1 извлечена из базы данных 2)	валидация	значение стадии рака в истории болезни пациента не соответствует коду в онкологическом реестре, в который включен этот пациент

Субизмерение	Метрическая группа	Абстрактная метрика	Оценка	Пример
		рассчитанные значения данных согласуются с общеизвестными значениям	валидация	рост и масса тела пациента, приводящие к получению неправдоподобного значения индекса массы тела (ИМТ) (например, менее 5), предполагают, что получены неточные данные о росте, массе тела или и том и другом
	временное правдоподобие	наблюдаемые или выводимые значения соответствуют ожидаемым временным свойствам	верификация	дата выписки наступает после даты поступления
		последовательность значений, представляющих переходы состояний, соответствует ожидаемым свойствам	верификация	дата введения первичной вакцины предшествует дате введения бустерной вакцины
		наблюдаемые или выводимые значения	валидация	продолжительность пребывания на

Субизмерение	Метрическая группа	Абстрактная метрика	Оценка	Пример
		имеют схожие временные характеристики в отношении одного или нескольких внешних компараторов («золотой стандарт»)		амбулаторной процедуре соответствует страховым данным для схожих популяций (не более 1 дня)
		последовательности значений, представляющие переходы состояний, схожи с внешним эталоном («золотым стандартом»)	валидация	последовательность иммунизации соответствует указаниям по вакцинации, принятым в государствах-членах
		ожидается, что меры плотности значений данных будут получены на основе внешних данных и соотнесены с фактором, учитывающим параметры времени	валидация	количество иммунизаций в месяц показывает ожидаемый рост вне сезона гриппа

Если «золотой стандарт» недоступен, метрику можно «верифицировать» на основе сравнения со схожей метрикой из другого источника. Это могут быть эквивалентные значения для идентичных измерений из двух независимых баз данных. Это необязательно «золотой стандарт», но может быть наилучшим доступным вариантом. При отсутствии «золотого стандарта» следует использовать альтернативные методы.

2. Экстенсивность (расширяемость)

54. Экстенсивность (расширяемость) объединяет два типичных измерения, встречающихся в системе качества данных – полноту покрытия данных и само покрытие. Здесь они объединены, поскольку оба относятся к количеству доступных данных.

55. Измерение экстенсивности (расширяемости) определяется количеством данных. Если рассматривать определение качества как «пригодный для цели», то экстенсивность (расширяемость) отражает, насколько достаточно данных. Экстенсивность (расширяемость) объединяет два типичных измерения, встречающихся в системе качества данных – полноту покрытия данных, а также их покрытие. Здесь они объединены, поскольку оба относятся к количеству доступных данных.

Субизмерения экстенсивности (расширяемости)

56. Если рассматривать количество доступной информации, то его можно выразить в процентах по отношению к общему количеству информации, которое может быть доступно. Различие между полнотой и покрытием вытекает из определения количества полностью доступной информации.

57. Полнота измеряет количество доступной информации по отношению к общему количеству информации, которая могла бы быть доступна с учетом процесса фиксации и формата данных. Данные, отсутствующие в наборе данных (по систематическим причинам, например, таким как информация, имеющаяся в источнике данных, но не включенная в модель данных) либо конкретные записи, не доступные для определенного поля, называются «недостающими». Процент недостающих значений для обязательного поля (например, пол) в наборе

данных является показателем полноты данных. Существует принципиальное различие между недостающими данными, о существовании которых известно (например, дата рождения пациента), и недостающими данными, о существовании которых неизвестно (например, наличие сопутствующих патологий). Довольно часто в модели данных определение переменной как «требуемой» подразумевает, что о существовании относительных значений известно, и поэтому, когда такие данные являются недостающими, речь идет об «известных недостающих». В целом при отсутствии явного отрицания может оказаться невозможным отличить «недостающие известные» от «недостающих неизвестных». Если ожидается, что некоторая точка данных будет зафиксирована, невозможность отличить «недостающие известные» от «недостающих неизвестных» является проблемой надежности (невозможно оценить, соответствуют ли данные той реальности, которую они должны отражать).

58. Покрытие измеряет количество доступной информации по отношению к тому количеству информации, что существует в реальном мире, независимо от того, находится ли она внутри процесса фиксации и формата данных или нет. Покрытие может быть нелегко измерить, поскольку общий объем информации может быть не поддающимся определению или недоступным. Примером покрытия может служить процентная доля некоторой популяции (например, страны или конкретной демографической группы), имеющейся в наборе данных. Для методов получения (сбора) доказательств покрытие называют «наблюдаемостью».

Другие концепции качества данных, связанные с экстенсивностью (расширяемостью)

59. Двумя концепциями, которые часто связывают с экстенсивностью (расширяемостью) являются репрезентативность и степень недостаточности данных. Хотя эти понятия в определенной степени описывают количество доступных данных, они в большей степени используются для характеристики того, насколько данные отражают реальность. Репрезентативность определяется как наличие у данных тех же характеристик в группе пациентов, что и у популяции, которую они призваны представлять (например, является ли совокупность индивидов, присутствующих в наборе данных, репрезентативной в отношении исследуемой популяции). Степень недостающих данных характеризует то, как неполные данные влияют на покрытие набора данных.

Подходы к оценке экстенсивности (расширяемости)

60. Экстенсивность (расширяемость) собираемой информации зависит от особенностей процесса сбора данных. Однако при объединении разных наборов данных для вторичного использования нет никакой гарантии полноты общего набора данных. На внутреннем уровне допускается использовать метрики для оценки степени полноты данных. Метрики, оценивающие, сколько данных присутствует в наборе данных по отношению к тому, что могло бы присутствовать в данной модели данных, довольно просты в вычислении. Метрики, оценивающие полноту данных по отношению к популяции, которую они должны измерять, более сложны и могут предусматривать использование «золотого стандарта». Полнота по отношению к схеме легко определяется, в то время как покрытие зависит от некоторых

предположений, которые могут быть заданы только по отношению к конкретному вопросу. Пороговые значения, используемые в качестве критериев приемлемости, также могут быть заданы для конкретного вопроса (например, 80 % полноты).

Примеры метрик экстенсивности (расширяемости)

Субизмерение	Метрическая группа	Абстрактное правило	Оценка	Пример
Полнота	отсутствующие требуемые значения	отсутствующие значения определяются конкретной схемой и не имеют динамики	верификация	раса или пол субъекта не должны приобретать нулевых значений в динамике
		отсутствующие значения определяются конкретной схемой единожды	верификация	идентификатор визита не указан (отсутствует) в некоторых записях
	оцениваемые недостающие значения	недостающие значения оцениваются на основе общих ожиданий	верификация	внезапное уменьшение количества кодов диагнозов связано с нарушением передачи данных исполнителем
		оценка недостающих значений по отношению к надежному источнику информации	валидация	в текущей переменной количество идентификаторов визитов в два раза меньше, чем в валидированной базе данных. Уменьшение количества кодов МКБ-IX после перехода на МКБ-X

Субизмерение	Метрическая группа	Абстрактное правило	Оценка	Пример
Покрытие		покрытие популяции	верификация	доля (в процентах) целевой популяции, представленная в базе данных

3. Когерентность (согласованность, последовательность)

61. Обнаружение непоследовательности часто является способом измерения надежности данных.

62. Измерение когерентности (согласованности, последовательности) отвечает на вопросы:

можно ли анализировать набор данных как «единое целое» или необходимы дополнительные действия, например, объединение нескольких наборов данных;

одинаков ли формат значений (например, даты) во всем наборе данных;

одинакова ли прецизионность (точность) значений (например, во всех ли случаях возраст округляется (аппроксимируется) до целых лет);

согласованы ли ссылки на сущности так, чтобы информация об одной и той же сущности была надлежащим образом «связана» между разными частями набора данных.

63. Если рассматривать определение качества как «пригодный для цели», то когерентность (согласованность, последовательность) относится к анализируемости данных.

Субизмерения когерентности (согласованности, последовательности)

64. Когерентность (согласованность, последовательность) — сложное и имеющее свои особенности измерение, которое включает в себя следующие субизмерения:

когерентность (согласованность, последовательность) формата, которая позволяет охарактеризовать выражены ли данные одинаково во всем наборе данных (например, позволяет охарактеризовать, что набор

данных, в котором даты представлены одновременно в виде ДД-ММ-ГГГГ и ММ-ДД-ГГГГ, не подходит для интегрированного анализа);

структурная или реляционная когерентность (относительная согласованность, относительная последовательность), которые являются синонимами и могут характеризовать нетабличные данные. Если возникает необходимость в расширениях системы качества данных для конкретных типов данных, следует рассмотреть возможные различия между структурной или реляционной когерентностью в зависимости от того идентифицируются ли одни и те же сущности одинаковым образом во всем наборе данных. Значение структурной когерентности (согласованности, последовательности) заключается в том, что ссылки связаны правильными сущностями (например, таблица аннотаций образцов ссылается на правильное значение в таблице результатов);

семантическая когерентность (смысловая согласованность), которая позволяет определить означает ли одно и то же значение для конкретного случая тоже самое, что и для всего набора данных. Например, означает ли показатель «анурия» состояние полного прекращения мочевыделения или характеризует определенную меру количества выделяемой мочи. Семантическая когерентность предполагает, что одно и то же понятие меры должно иметь одинаковую прецизионность (точность) во всем наборе данных;

уникальность, которая характеризует свойство, когда одна и та же информация не дублируется, а появляется в наборе данных только один раз. В отношении уникальности следует отметить, что понятия «информация» и «данные» отличаются. Два разных измерения (measurements), в результате которых получены одни и те же данные, не будут представлять собой дублирование информации (и такие измерения, скорее всего, будут отличаться по ценности, если включить в

них метаданные). В тоже время одно и то же измерение, представленное дважды, будет являться дублированием информации. Такая проблема характерна для данных, агрегированных из разных источников. Данные, содержащие некоторую избыточность, получают более низкую оценку в измерении уникальности, но эти дополнительные записи могут способствовать улучшению других измерений, таких как надежность;

соответствие, которое определяется по отношению к конкретному стандарту или модели данных и является наиболее эффективным способом оценки когерентности. Оно подразделяется на форматное, структурное и семантическое соответствие. Например, при проверке на соответствие оценивается, является ли формат представления данных согласованным, то есть соответствует ли он общему целевому стандарту (например, стандарту представления даты в формате ДД-ММ-ГГГГ);

валидность которая определяется, когда референтная модель специфична для оцениваемого набора данных и является частным случаев когерентности (согласованности, последовательности. Например, если в файле с данными применяется схема, устанавливающая, что все даты рождения в файле должны быть в формате ДД-ММ-ГГГГ, то файл может быть непосредственно оценен как валидный или нет.

Подходы к оценке когерентности (согласованности, последовательности)

65. Когерентность (согласованность, последовательность) данных в источнике в значительной степени зависит от таких основополагающих факторов, как синхронизация процессов и систем в организации, генерирующей данные, или при агрегировании нескольких наборов данных при обязательстве такой организации (организаций)

использовать внутренние или внешние стандарты данных. Когерентность (согласованность, последовательность) данных, агрегированных и переработанных для вторичного использования, зависит от наличия общих стандартов и референтных данных. Внутренние аспекты когерентности (согласованности, последовательности) набора данных могут быть улучшены в основном на этапе обработки, связанном со стандартизацией данных. Однако повышение когерентности (согласованности, последовательности) предполагает аппроксимацию (приближение) или уточнение смысла данных. Например, для уточнения информации часто требуется доступ к исходной системе и процессам. Некоторые аспекты семантической когерентности (смысловой согласованности) сложно оценить с помощью метрики, поэтому их можно оценить только в отношении конкретного вопроса и стратегии анализа.

Таблица 3

Пример метрик когерентности (согласованности, последовательности)

Субизмерение	Метрическая группа	Абстрактное правило	Оценка	Пример
Когерентность (согласованность, последовательность) формата	синтаксические ограничения	значения данных соответствуют внутренним ограничениям форматирования	верификация	пол обозначается только одним символом ASCII
	допустимые значения	значения данных соответствуют допустимым значениям или диапазонам	верификация	для животного в поле «Пол» могут быть указаны только значения «М», «Ж», или «Н»
		значения данных соответствуют ограничениям представления, основанным на внешних стандартах	валидация	значения для основного языка соответствуют стандартам ИСО
Реляционная когерентность (относительная согласованность,	референтная когерентность (согласованность, последовательность)	значения данных соответствуют реляционным ограничениям	верификация	номер медицинской записи пациента связан с другими таблицами, как и ожидалось

Субизмерение	Метрическая группа	Абстрактное правило	Оценка	Пример
относительная последовательность)		уникальные (ключевые) значения данных не дублируются	верификация	номер медицинской записи присваивается только одному пациенту
		значения данных соответствуют реляционным ограничениям, основанным на внешних стандартах	валидация	значения данных соответствуют всем требованиям, исключающим наличие пустых значений, в общем формате обмена данными между учреждениями
	когерентность (согласованность) схемы	прослеживаемость изменений в модели данных или управления версиями модели данных	верификация	данные версии 1 не содержат информацию о времени выписки пациента из медицинского учреждения
	вычислительная когерентность (согласованность)	рассчитанные значения соответствуют программным спецификациям	верификация	значения индекса массы тела (ИМТ), рассчитанные в базе данных и вручную, совпадают

Субизмерение	Метрическая группа	Абстрактное правило	Оценка	Пример
		результаты расчетов, полученные на основе опубликованных алгоритмов, дают значения, которые совпадают с валидными значениями, предоставленными из внешних источников	валидация	рассчитанные перцентили индекса массы тела дают результаты, совпадающие с результатами исследования и данными предоставленными внешними референтными источниками
Семантическая когерентность (смысловая согласованность)	когерентность (согласованность) прецизионности (точности)	прецизионность (точность) значений соответствует целевому стандарту	верификация	например, используются две десятичных знака в дробной части, которые как правило отличны от нуля
	семантическая когерентность (смысловая согласованность)	использование списков кодов последовательно во всех данных	верификация	например, уровень кодирования MedDRA для показания не меняется в пределах набора данных
Уникальность		один и тот же субъект представлен с одинаковыми идентифицирующими данными	верификация	Елена Петрова также фигурирует в базе как Алена Петрова с той же датой рождения

Субизмерение	Метрическая группа	Абстрактное правило	Оценка	Пример
		один и тот же субъект представлен с несколькими идентифицирующими данными	верификация	записи об Иване Петрове выглядят как записи о разных лицах, а не как записи об одном и том же человеке
		записи данных субъектов сопоставимы по уникальным ключам (кодам)	валидация	идентификационный код и дата рождения Елены Петровой совпадает с датой рождения и кодом Алены Петровой

4. Своевременность

66. Параметр «Своевременность», позволяет оценить отражают ли данные целевую реальность на момент их использования.

67. В случае, если качество данных рассматривается через определение их пригодности для цели, то параметр своевременность позволяет оценить, насколько точно (близко) данные отражают целевую реальность на момент их использования.

Субизмерения своевременности

68. Субизмерением своевременности является актуальность. Следует учитывать, что отсутствие актуальности не означает отсутствия своевременности (например, исторические данные могут быть не актуальными, но при этом быть своевременными для ретроспективных исследований).

Другие концепции качества данных, связанные со своевременностью

69. В контексте системы качества данных задержка получения данных (ситуация, при которой данные фиксируются позже заявленного момента), относится к параметру надежность (поскольку позволяет оценить соответствуют ли данные реальности на тот момент, в который они должны были ее отразить).

Подходы к оценке своевременности

70. Своевременность определяется системами и процессами, используемыми для сбора и предоставления данных.

Пример метрик своевременности

Субизмерение	Метрическая группа	Абстрактное правило	Ссылка
Актуальность	не применимо	среднее время обновления базы данных (или временная метка), которое как показатель актуальности отражает более узкий аспект оценки своевременности (продолжительность времени фактической записи данных в базу данных, а не момент времени сбора данных)	верификация
		последнее обновление базы данных (или временная метка)	верификация

5. Релевантность

71. В целях оценки качества данных релевантность определяется как степень, в которой набор данных способен предоставить элементы данных, пригодные для получения ответа на конкретный исследовательский вопрос. Такой подход к определению релевантности является более узким и ориентированным по сравнению с расширенным понятием релевантность, которое применяется в отношении источника данных, генерирующего валидные доказательства, позволяющие ответить на конкретный исследовательский вопрос, исходя из дизайна исследования. Релевантность также может быть определена как степень, того, в какой мере что-либо связано с происходящим процессом, изучаемым вопросом или имеет отношение к поставленной цели.

72. Чтобы разграничить эти два значения, в данном Руководстве используются дополнительные указания к понятию «релевантность как измерение качества данных» (релевантность в значении определения,

приведенного в пункте 66 настоящего Руководства) и «релевантность вопросу» (более общее значение).

73. Релевантность как измерение качества позволяет получить ответ на вопрос содержит ли набор данных значения (или элементы данных), необходимые для решения конкретного вопроса с использованием конкретного метода. Различие между экстенсивностью (расширяемостью) и релевантностью связано с тем, что они позволяют получить ответы на разные вопросы. Экстенсивность (расширяемость) позволяет ответить на вопрос какое количество данных имеется в наличии (каким количеством данных располагает исследователь), а релевантность дает ответ на вопрос какими именно данными исследователь располагает.

74. Если рассматривать определение качества как «пригодный для цели», то релевантность как измерение качества данных описывает, насколько полно данные охватывают те аспекты реальности, которые предполагается измерить.

75. В настоящей системе качества данных релевантность по отношению к вопросу определяется «вопрос-специфичными детерминантами (факторами)», которые применяются ко всем измерениям качества данных.

76. Ранее введенные измерения аспектов качества данных основаны на ряде определяющих их вопросов:

насколько точно данные отражают реальность;

какое количество данных имеется;

возможно ли проанализировать данные в целом;

доступны ли данные в нужное время.

Для характеристики измерения релевантности как одного из параметров качества данных в вышеприведенный перечень следует включить вопрос «какой именно тип данных имеется в наличии?».

77. Учитывая вышеописанный контекст, релевантность может быть определена только в отношении исследовательского вопроса и стратегии анализа данных (определение предположений, решений и методов для ответа на конкретный вопрос). Однако в некоторых случаях можно определить совокупность часто задаваемых исследовательских вопросов, которые можно охарактеризовать с точки зрения релевантности в контексте конкретного типа источника данных. Это называется релевантностью для данного домена (области) исследовательского вопроса (далее – домен (область)).

78. Пример метрик релевантности проводится в таблице 5. Данный пример приводится для пояснения того, что именно может быть отнесено к параметру релевантности. Поскольку не все переменные являются эквивалентными, то для конкретных сценариев использования и (или) типов данных следует указать их фактические метрики, характеризующие качество этих данных.

Таблица 5

Пример метрик релевантности

Субизмерение	Группа метрик	Абстрактное правило	Ссылка	Пример
не применимо	не применимо	количество переменных (столбцов) доступных в данном наборе данных сравнивается с количеством необходимых переменных	верификация	не применимо

VII. Общие рекомендации и модели зрелости

79. Выбор наборов данных для их использования при принятии регуляторных решений требует понимания того, насколько такие данные соответствуют критериям точности, экстенсивности (расширяемости), когерентности (согласованности), своевременности и релевантности. Такие измерения качества формируются на протяжении всего жизненного цикла данных – от их генерации и обработки до агрегирования и, в конечном итоге, анализа. При этом данные, первоначально собранные для одних целей, могут быть использованы в других целях при соблюдении этических и правовых требований.

80. Выбор показателей качества и методов проверки в значительной степени зависит от типов данных и целей их использования. Вместе с тем такие измерения и проверки следует организовать в соответствии с четкой (когерентной, согласованной) структурой, которая способствует обеспечению единообразия и выявлению пробелов.

81. В таблицах 6 и 7 показано, как детерминанты (факторы), определяющие качество (фундаментальные, внутренние или вопрос-специфичные) влияют на разные измерения качества как данных, так и метаданных. Эти таблицы содержат рекомендации относительно того, какие метрики и действия применяются на том или ином этапе жизненного цикла данных. Например, измерение экстенсивности (расширяемости) определяется исключительно фундаментальными детерминантами (основополагающими факторами) на этапе создания. На более поздних этапах жизненного цикла данных внутренние показатели данных позволяют лишь частично оценить степень надежности (метрики правдоподобия).

82. Эти таблицы также служат основой для разработки моделей зрелости, предназначенных для характеристики качества данных для

принятия регуляторных решений. Модели зрелости содержат рекомендации относительно того, как детерминанты (факторы) могут быть охарактеризованы на последовательных уровнях зрелости. Более высокие уровни зрелости позволяют получить самые убедительные доказательства наиболее эффективным способом.

83. Представлены три разные модели зрелости, соответствующие трем детерминантам (факторам), чтобы показать, как зрелость развивается в отношении установления характеристики процессов, внутренних аспектов (метрик) и определения целевых вопросов. Эти модели предназначены для применения к разным этапам и действиям, составляющим общую систему получения (сбора) доказательств.

84. Представленные модели зрелости являются абстрактными в том смысле, что они содержат классы или рекомендации, которые следует дополнить деталями реализации для конкретных типов данных и вариантов использования.

85. Для установления характеристик и внедрения процессов, чтобы достичь более высокого уровня зрелости, требуется время как держателям источников данных, так и экспертам уполномоченных органов (экспертных организаций), чтобы они могли понять какое влияние оказывает модель более высокого уровня зрелости. Это также зависит от контекста, например, от региона распространения заболевания, частоты заболевания, системы здравоохранения и др. Система качества данных непрерывно актуализируется исходя из новых материалов, посвященных вариантам применения данных и представляющим особый интерес с точки зрения регулирования, что позволяет выполнять надлежащую клиническую оценку данных в сфере регулирования лекарственных средств.

Таблица 6

Оценка влияния детерминант (факторов) на параметры качества данных

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Релевантность
Фундаментальная		первичное. Собранные в соответствии с установленными протоколами данные, могут быть достаточными для принятия регуляторных решений			первичное. Как правило, гарантировано дизайном процесса сбора данных
	первичное и вторичное. Надежность данных (во всех аспектах) зависит от существующих систем и процессов, используемых для генерации (сбора) данных. На надежность влияют обработка и преобразование данных на более	первичное и вторичное. Протокол сбора данных определяет, какие данные будут собираться	первичное и вторичное. Зависит от организации процессов, генерирующих данные, и от степени соблюдения внутренних или внешних стандартов данных	первичное и вторичное. Определяется исключительно используемыми системами и процессами	

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Релевантность
	поздних этапах, например, при стандартизации в соответствии с общей моделью данных				
	вторичное. Прецизионность (точность) может снизиться в процессе преобразования и гармонизации данных	вторичное. Невозможно гарантировать полноту сводного набора данных или степень его покрытия для разных вариантов использования. Детерминанту можно только оценить или контролировать	вторичное. Опирается на общие стандарты и справочные данные. Для повышения когерентности (согласованности) может потребоваться документация по процессам создания данных		вторичное. Обычно оценивается для конкретного использования или класса использований при отборе наборов данных
Внутренняя	первичное и вторичное. Показатели правдоподобия могут использоваться для выявления (ограниченного) класса проблем, связанных с надежностью.	первичное и вторичное. Показатели полноты, основанные на модели данных, легко внедрить	первичное и вторичное. Когерентность (согласованность) может быть измерена исключительно на основе данных (с возможным доступом к	первичное и вторичное. Некоторые аспекты своевременности могут наблюдаться в наборах данных (например, даты событий для	первичное и вторичное. Релевантность данных не зависит от самого набора данных

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Релевантность
	Прямые измерения правильности требуют доступа к источнику данных		независимым от наборов данных референтным данным)	определения актуальности). Сам по себе набор данных, как правило, не может показать, насколько актуальна содержащаяся в нем информация	
		вторичное. Меры покрытия более сложны и могут требовать сопоставления с «золотым стандартом»	вторичное. Когерентность (согласованность) может быть в значительной степени улучшена только на основе набора данных и элементов, не зависящих от данных (например, картирование с общим стандартом). Для полного решения вопроса о когерентности (согласованности)		

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Релевантность
			может требоваться доступ к дополнительной информации о процессах. Когерентность (согласованность) должна оцениваться каждый раз, когда «интегрируется» новый источник данных		
Вопрос- специфичная	Первичное. Процессы и системы сбора данных, как правило, разрабатываются для того, чтобы дать ответ на конкретный вопрос и достичь необходимых целевых показателей по всем параметрам качества данных, которые предполагает данная цель.				
	вторичное. Пороговое значение приемлемой надежности может быть определено только для конкретного вопроса и метода	вторичное. Покрытие и полнота зависят от конкретного вопроса: метрики могут быть определены только для конкретного вопроса и метода или конкретной области. Для полноты, как правило, вопрос	вторичное. Некоторые показатели семантической когерентности (смысловой согласованности) (когерентности распределения данных или когерентности	вторичное. Приемлемая своевременность зависит от конкретного вопроса и его более широкого регуляторного применения (например,	вторичное. Релевантность может быть определена только в отношении одного или нескольких вопросов

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Релевантность
		определяет совокупность пороговых значений приемлемости и общих метрик	абстракции) может быть измерена только в отношении конкретного вопроса и метода	утверждение или мониторинг)	

Таблица 7

Оценка влияния детерминант (факторов) на измерения качества метаданных

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Актуальность
Фундаментальная		первичное. В отношении первичных данных экстенсивность (расширяемость) метаданных может быть охарактеризована по источнику	первичное. Когерентность (согласованность) метаданных зависит от наличия общих стандартов и терминологии		первичное. Как правило, обеспечивается за счет дизайна процесса сбора данных
		первичное и вторичное. Надежность метаданных зависит от процессов их сбора		первичное и вторичное. Своевременность метаданных зависит	

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Актуальность
		по всей цепочке обработки данных одним из ключевых аспектов обеспечения надежности является фиксация метаданных как можно ближе к источнику		исключительно от процессов, поддерживающих их сбор	
			вторичное. В отношении вторичных данных когерентность (согласованность) зависит от наличия широко согласованных стандартов и общих ресурсов, таких как онтологии или службы справочных данных	вторичное. Если данные перепрофилируются и используются в разных системах, своевременность метаданных должна обеспечиваться дизайном (метаданные должны быть синхронизированы с данными)	вторичное. Релевантные метаданные могут требоваться и контролироваться последующей системой, но не могут быть гарантированы у источника
Внутренняя	первичное и вторичное. Некоторые метаданные (например, сводная статистика) могут	первичное и вторичное. Внутренние меры для мета-качества данных схожи с мерами для данных (например,	первичное и вторичное. Когерентность (согласованность) метаданных зависит только от	первичное и вторичное. Оценка своевременности как аспекта данных обычно зависит от	первичное и вторичное. Релевантность метаданных не зависит от самого набора данных

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Актуальность
	быть сгенерированы на основе набора данных, когда данные и метаданные рассматриваются как единое целое, прослеживаемость можно оценить и по внутренним мерам	полнота и недостающие поля). В отличие от данных, оценка метаданных может не требовать ссылок на «золотые стандарты» (например, отсутствие значений метаданных не связано с формированием выборки из популяции)	конкретных метаданных и элементов, не зависящих от данных (например, совместно используемых справочных данных)	метаданных (например, временных меток)	
Вопрос- специфичная	Первичное. Требования к метаданным разработаны для конкретного вопроса и обычно достаточны для его решения				
	первичное и вторичное. Метаданные должны быть в целом надежными, независимо от конкретного вопроса (не все собранные метаданные могут		первичное и вторичное. Когерентность (согласованность) метаданных не зависит от конкретного вопроса	первичное и вторичное. Своевременность метаданных не зависит от конкретного вопроса	

Детерминанта (фактор)/ измерение	Надежность	Экстенсивность (расширяемость)	Когерентность (согласованность)	Своевременность	Актуальность
	быть релевантны для всех вопросов)				
		вторичное. Определение того, какие метаданные необходимы, в конечном счете зависит от вопроса (или совокупности типичных вопросов)			вторичное. Релевантность метаданных зависит исключительно от вопроса (или совокупности типичных вопросов)

1. Фундаментальные детерминанты (основополагающие факторы): рекомендации и уровни зрелости

86. Для управления качеством данных следует дать характеристику систем и процессов, лежащих в основе генерации и обработки данных (фундаментальные детерминанты (основополагающие факторы)). На рисунке 5 приведен набор определенных уровней зрелости, каждый из которых представляет собой прогрессирующую иерархию рекомендаций по установлению характеристик фундаментальных детерминант (основополагающих факторов) с целью обозначить направление совершенствования для надлежащего и эффективного установления характеристик этих аспектов качества данных.

87. Рекомендуется как можно ранее внедрять концепцию надлежащей работы с данными «Находимость, Доступность, Совместимость и Повторное использование» (FAIR) (далее – принципы НДСП (FAIR)) как в отношении ее полного, так и частичного применения к данным и метаданным параллельно с внедрением моделей зрелости. Принципы НДСП (FAIR) связаны с возможностями автоматической обработки (то есть способностью вычислительных систем находить, получать доступ, взаимодействовать и повторно использовать данные без вмешательства человека, или с минимальным его участием). Такой подход необходим в связи с постоянным ростом объема, сложности и скорости генерации информации.

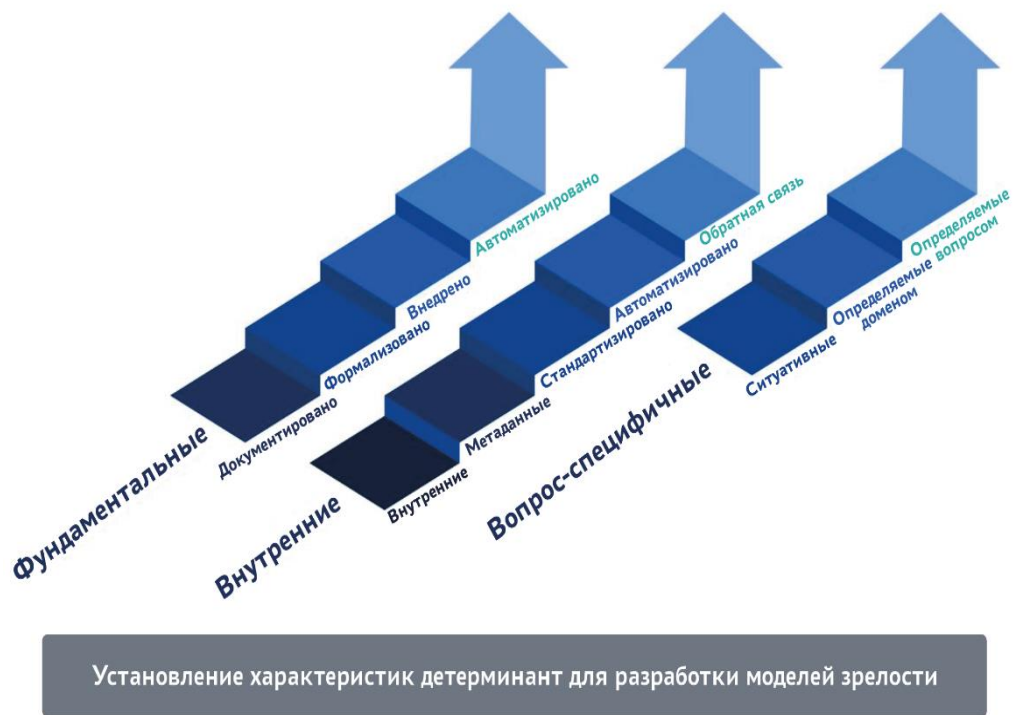


Рисунок 5. Модель зрелости для детерминант (факторов) качества данных

Уровень 1: документированные данные

88. Чтобы данные были пригодными для принятия регуляторных решений, процессы, связанные как минимум с генерацией и обработкой данных, должны быть документированы, достоверны, проверяемы (в соответствующих случаях это может распространяться на процедуры обучения) и соответствовать версиям. Это является фундаментальным фактором и обеспечивает надежность любой полученной информации. Документация должна охватывать детерминанты (факторы) прецизионности (точности), экстенсивности (расширяемости), когерентности (согласованности) и, если применимо, своевременности. В то время как некоторые из этих детерминант (факторов) зависят от конкретного вопроса, процессы и системы сбора данных как правило разрабатываются с учетом некоторых рутинных вопросов. Предоставление документации по обработке и преобразованию данных имеет решающее значение для обеспечения надежности и должно

осуществляться в отношении всех подобных операций, проводимых различными субъектами на протяжении всего жизненного цикла данных.

С точки зрения метаданных это означает, что метаданные (в той или иной форме) всегда должны сопровождать набор данных, к которому они относятся.

89. Для обеспечения качества данных, следует внедрить процедуры аудита или другие меры контроля.

90. Если система разработана для непрерывного сбора данных (в отличие от одноразового сбора), следует предусмотреть дополнительные процессы мониторинга и повышения эффективности.

Уровень 2: формализованные данные

91. Уровень 2 модели зрелости включает в себя первый уровень и расширяет его, таким образом, чтобы документация и метаданные во всех возможных случаях соответствовали отраслевым стандартам или системе менеджмента качества. То, какие отраслевые стандарты или программы применяются, зависит от конкретных типов данных и вариантов использования, и поэтому они могут быть определены только в рамках специализации конкретного исследования.

92. Уровень 2 модели зрелости следует считать минимальным уровнем приемлемой зрелости, хотя для новых типов данных допускаются исключения. Рекомендация по использованию стандартов распространяется и на метаданные.

Уровень 3: внедренные данные

93. Существуют системы, которые обеспечивают систематическое и целенаправленное выполнение процессов контроля качества данных в соответствии с отраслевыми стандартами. Следует создать

инфраструктуру для обеспечения и управления данными, включая поддержку стандартизации (например, управление справочными данными или управление мастер-данными). Снижая вероятность человеческой ошибки, такое внедрение может в целом повысить прецизионность (точность), надежность и когерентность (согласованность). Такое внедрение также может потребоваться, для обеспечения своевременности, и система должна обеспечивать сбор метаданных в соответствии с дизайном и как можно ближе к моментам генерации или сбора данных.

Уровень 4: автоматизированные данные

94. Операции и результаты работы вышеуказанных систем и инфраструктуры должны быть доступны в машиночитаемом формате, чтобы унифицировать данные и элементы качества данных для непосредственного последующего использования. Все данные и метаданные должны быть представлены в соответствии с принципами НДСП (FAIR), чтобы обеспечить полную автоматическую обработку параметров качества.

2. Внутренние детерминанты (факторы): рекомендации и уровни зрелости

95. Помимо документального подтверждения того, как собирались или генерировались данные, можно применять показатели, характеризующие внутренние аспекты качества данных. Их можно получить непосредственно из набора данных, но для их вычисления может потребоваться также использование каких-либо внешних источников знаний.

Уровень 0: внутренние данные

96. Строгих минимальных требований к качеству не существует, поскольку любые данные могут быть оценены до того, как они будут использованы для генерации доказательств. Этому начальному уровню присвоено значение «0», чтобы пояснить, что он соответствует данным «в исходном виде», независимо от их предлагаемого использования для принятия регуляторных решений. Тем не менее распространение данных без соответствующей оценки их качества не рекомендуется.

Уровень 1: метаданные

97. Данные предоставляются с совокупностью метрик качества в качестве метаданных. Некоторые из этих данных могут быть получены непосредственно из набора данных, другие – из общего процесса сбора данных (например, получение выборок). Метаданные также должны охватывать описание элементов данных, необходимых для их интерпретации (например, словари данных).

Уровень 2: стандартизированные данные

98. Данные сопровождаются стандартизированным набором метрик качества, которые можно сравнивать в различных наборах данных. Если применимо или возможно, стандарты должны охватывать справочные данные, которые могут быть использованы для оценки набора данных с точки зрения того, что они должны отражать (например, типичные популяционные распределения для оценки смещений). Метаданные используют общие определения, которые также обеспечивают сопоставимость и интеграцию различных наборов данных.

Уровень 3: автоматизированные данные

99. Оценка качества осуществляется в автоматизированном режиме (по меньшей мере для большого количества метрик). В целом это выполнимо только в том случае, если данные представлены стандартными способами (например, в формате общей модели данных), что позволяет применять к поступающим данным стандартную библиотеку тестов. Данные и метаданные должны соответствовать принципам НДСП (FAIR). В отношении основополагающих факторов принципы НДСП (FAIR) следует применять как можно раньше, по крайней мере частично. Уровень 3 предполагает полное соблюдение принципов НДСП (FAIR).

Уровень 4: данные с обратной связью

100. Существует экосистема данных, позволяющая пользователям данных оценивать их качество, обеспечивая обратную связь для совершенствования процесса сбора и подготовки данных, что обеспечивает непрерывный мониторинг и повышение качества данных. Следует обратить внимание, что порядок зрелости для уровней 2 и 3 может меняться в зависимости от конкретных типов данных.

3. Вопрос-специфичные детерминанты (факторы): рекомендации и уровни зрелости

101. В целом невозможно оценить релевантность набора данных или аспекты экстенсивности (расширяемости) и прецизионности (точности), не имея целевого вопроса и заданной стратегии анализа. Однако при рассмотрении вопроса о принятии большого массива данных для принятия регуляторных решений и их возможном использовании за пределами первичных сценариев применения важно четко определить, в

какой степени качество данных, включая их релевантность, может быть оценено априори.

Уровень 1: ситуативные данные
(данные ad-hoc, задаваемые специально данные)

102. Все параметры, являющиеся вопрос-специфичными, оцениваются только в «момент возникновения вопроса» в каждом конкретном случае.

Уровень 2: данные, определяемые областью

103. Для этого уровня выявлен ряд типичных вопросов, на основе которых можно определить метрики и пороговые значения, позволяющие обеспечить приемлемый уровень качества. Данные, публикуемые в каталогах данных, должны использовать такие метрики.

Уровень 3: данные, определяемые вопросом

104. Требования к конкретному вопросу точно кодифицированы и могут быть картированы по метрикам и пороговым значениям таким образом, чтобы можно было автоматически оценить релевантность набора данных для конкретного вопроса. На этом уровне контекст, в котором данные будут интерпретироваться для принятия решений, должен быть формализован и доступен для совместного использования. Это естественный уровень для первичных сценариев использования, в то время как для вторичного использования данных этот уровень следует рассматриваться как целевой уровень.

VIII. Подходы по внедрению системы качества данных

105. Настоящий раздел содержит рекомендации, которыми следует руководствоваться при внедрении системы качества данных и ее специализаций, чтобы способствовать достижению более высокого уровня зрелости.

1. Качество источника данных

106. В качестве общего правила, при разработке процессов сбора и генерации данных следует как можно раньше учитывать аспекты качества данных. Например, оценка качества, проводимая непосредственно перед выпуском продукции, может помочь исправить ошибку при сборе. Чем дальше данные уходят от исходного контекста, тем сложнее исправить ошибки. Это особенно актуально в отношении метаданных, поскольку полное понимание контекста создания данных возможно только в момент их формирования.

2. Роль управления мастер-данными и справочных данных в обеспечении качества данных

107. Наличие управления мастер-данными и справочных данных напрямую влияет на качество данных. Это часто является необходимым условием для обеспечения согласованности данных, а в некоторых вариантах генерации данных (например, данных о материалах) может повлиять на их надежность, поскольку разрозненная информация может привести к ошибочной информации. В более широком смысле управление мастер-данными и справочные данные позволяют автоматизировать ряд проверок качества данных и, следовательно, также влияют на надежность.

108. Общее управление мастер-данными и справочные данные позволяют решать вопросы согласованности, выходящие за рамки основного сценария использования, для которого были созданы эти данные, поскольку применение стандартов гарантирует поддержание определенного уровня семантической (смысловой) когерентности даже после этапов агрегирования данных.

3. Роль системы менеджмента качества и компьютеризированных систем в обеспечении качества данных

109. Внедрение системы менеджмента качества на более высоких уровнях зрелости требует формализации и внедрения систем и процессов для поддержания качества данных.

110. Система менеджмента качества (СМК) – это формализованный подход, принятый организацией, который документирует процессы, процедуры и обязанности для достижения целей и задач в области качества (например, надлежащей клинической практики (GCP), надлежащей лабораторной практики (GLP) или надлежащей производственной практики (GMP)). Эти цели в области качества достигаются посредством планирования качества, обеспечения качества, контроля качества и совершенствования качества. Общие стандарты (ИСО 9000), устанавливают принципы организации системы менеджмента качества для различных отраслей, в то время как в отношении конкретных отраслей или продуктов имеются более специализированные системы менеджмента качества. В зависимости от характера данных следует учитывать системы менеджмента качества, разработанные специально для отрасли медико-биологических наук и включающие в себя:

для данных клинических исследований – Правила надлежащей клинической практики Евразийского экономического союза, утвержденные Решением Совета Евразийской экономической комиссии от 3 ноября 2016 г. № 79, стандарт «Клинические исследования медицинских изделий, проводимые с участием человека в качестве субъекта»;

для данных, полученных с медицинских изделий или диагностических средств – стандарт «Изделия медицинские. Системы менеджмента качества. Требования для целей регулирования»;

для данных, полученных в ходе лабораторных исследований – Правила надлежащей лабораторной практики Евразийского экономического союза в сфере обращения лекарственных средств, утвержденные Решением Совета Евразийской экономической комиссии от 3 ноября 2016 г. № 81, приложение № 8 к Порядку формирования и ведения реестра химических веществ и смесей Евразийского экономического союза, утвержденному Решением Совета Евразийской экономической комиссии от 1 августа 2025 г. № 61;

для данных полученных от клинико-диагностических лабораторий – стандарты «Лаборатории медицинские. Частные требования к качеству и компетентности», «Общие требования к компетентности испытательных и калибровочных лабораторий».

По возможности процессы, связанные с качеством данных следует рассматривать в контексте стандартных систем менеджмента качества.

111. В современных условиях, исходя из цифровизации процессов на фундаментальные факторы качества данных влияют также компьютерные системы, которые используются для создания, изменения, хранения, архивирования, поиска или передачи данных. Жизненный цикл разработки программного обеспечения, включающий систему

обеспечения качества программного обеспечения, обеспечивает надлежащее проектирование, разработку и тестирование программного обеспечения. При внедрении систем качества для компьютеризированных систем следует учитывать положения Руководства по обеспечению целостности данных и валидации компьютеризированных систем (приложение к Рекомендации Коллегии Евразийской экономической комиссии от 19 сентября 2023 г. № 25) и серии стандартов «Требования и оценка качества систем и программного обеспечения», включая применение этих стандартов к моделям данных. Квалификация и валидация компьютерных систем обеспечивает надлежащее внедрение программного обеспечения, а также необходимые контроли процессов для его использования в соответствии со спецификациями, включая документацию, контроль доступа, управление поставщиками и аудиты.

112. Приведенная на рисунке 6 схема принятия решений содержит рекомендации по внедрению системы менеджмента качества в системы качества данных.

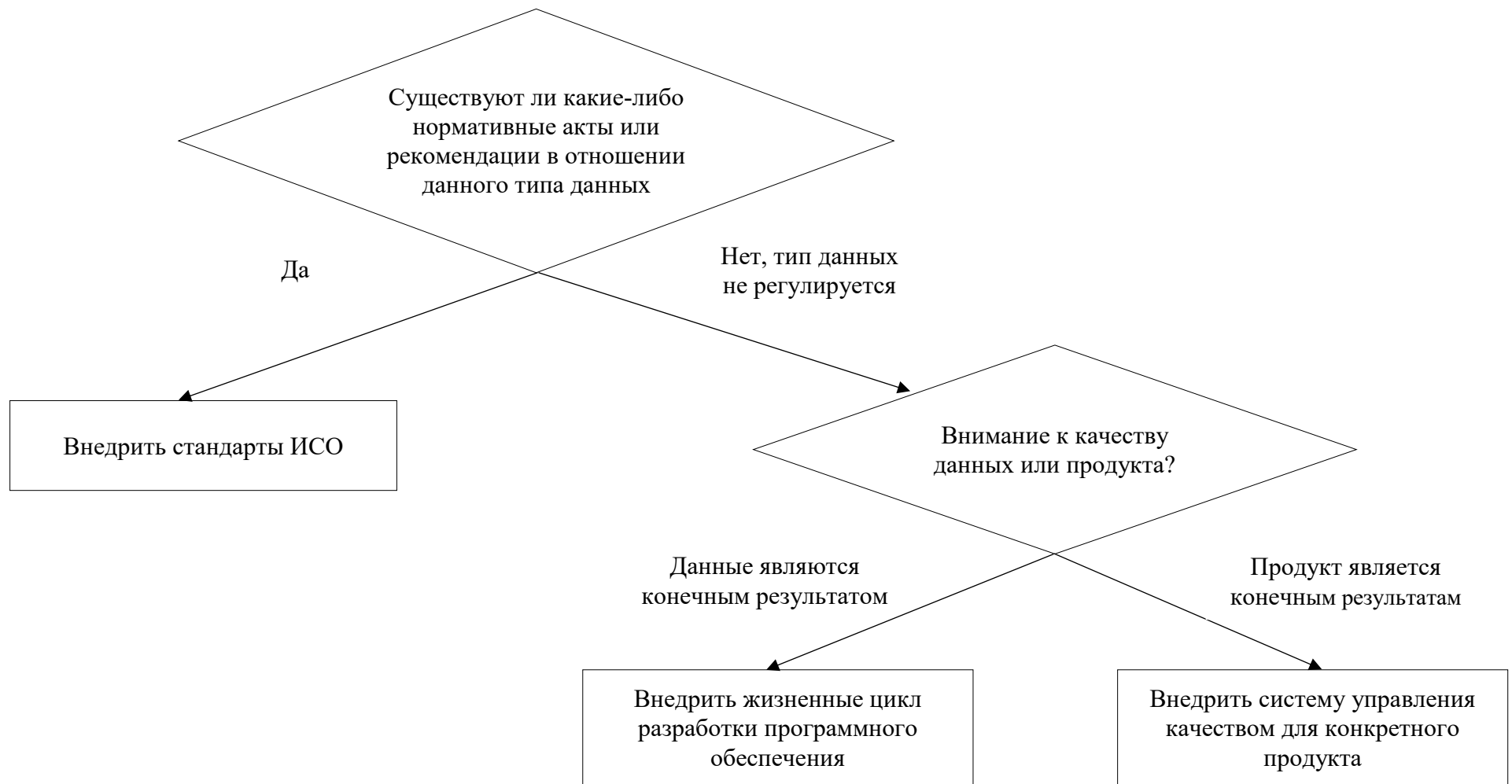


Рисунок 6. Схема принятия решений по внедрению системы менеджмента качества в рамках реализации системы качества данных

4. Роль ALCOA⁺ в обеспечении качества данных

113. ALCOA⁺ представляет собой систему обеспечения целостности данных, используемую во всей фармацевтической отрасли в таких сферах, как научные исследования, производство, испытания и цепи поставок. В ней содержится ряд принципов, которые следует соблюдать при обращении с данными и документацией. В частности, следует обеспечить соответствие данных основным принципам ALCOA (соотносимость (attributable), читаемость (legible), своевременность (contemporaneous), подлинность (original), точность (accurate)). Знак «+» в обозначении ALCOA указывает на соблюдение в отношении данных следующих дополнительных принципов: последовательность (consistent), устойчивость (enduring), доступность (available), прослеживаемость (traceable). Данные принципы взаимосвязаны с положениями Руководства по обеспечению целостности данных и валидации компьютеризированных систем.

114. В рамках настоящего Руководства следует применять принципы ALCOA⁺, которые связаны с фундаментальными детерминантами данных и влияют в первую очередь на их надежность, а также на экстенсивность, когерентность и своевременность.

115. Принципы ALCOA⁺ в значительной степени соответствуют параметрам системы качества данных, но носят более конкретный и практический характер. Например, определение точности по методу ALCOA описывает набор характеристик (поддающиеся проверке процессы кодирования, валидированная передача данных), которые должны быть обеспечены для того, чтобы «данные точно отражали сделанные наблюдения». Другие принципы, такие как читаемость и подлинность, также относятся к надежности (поскольку они позволяют

ответ на вопрос «отражают ли данные реальность?»), однако в системе качества данных они явно не сформулированы, поскольку являются предварительным условием для подачи документов в регуляторные органы.

116. В определенных случаях применение спецификаций, соответствующих принципам ALCOA⁺, позволяет достичь уровня зрелости 2 (формализованные данные) и выше для фундаментальных детерминант.

5. Роль внедрения контроля качества данных в обеспечении качества данных

117. Существуют разные варианты внедрения контроля качества данных (то есть тестирования).

118. Если истинные факты, которые отражают данные, известны и доступны, данные можно проверить с помощью валидации в сравнении с исходными записями, содержащими эти факты. Однако валидация может быть дорогостоящей и трудоемкой, и часто требует привлечения третьих сторон (арбитров), если факты не доступны в машиночитаемой структурированной форме. В качестве альтернативы данные можно проверять с помощью метрик внутреннего правдоподобия, а именно путем оценки набора данных на соответствие (приведено на рисунке 7):

другим данным в том же наборе данных: тест позволяет обнаружить логические или фактические противоречия (например, встраивание фоновых знаний о взаимосвязях между объектами и событиями). Например, время начала причинно-следственного эффекта должно наступать после вызывающего его вмешательства, или у пациентов-женщин не могут наблюдаться явления, характерные исключительно для мужчин;

внешним референтным диапазонам (или «золотые стандарты»): некоторая измеряемая величина не может превышать определенную величину, например, артериальное давление ни при каких условиях не может превышать 1000/500 мм рт. ст.;

правдоподобным тенденциям: отдельные данные могут быть валидными при индивидуальном наблюдении, но общая тенденция всех данных одного вида должна соответствовать ожидаемым распределениям или тенденциям. Например, маловероятно, что заболеваемость в популяции резко возрастет с 2 до 80 % за один год, или что клеточная линия, подвергшаяся экспозиции в ходе эксперимента, продемонстрирует меньший эффект, чем не подвергшаяся экспозиции контрольная группа. В данном случае данные оцениваются с учетом имеющихся исходных сведений о типичных характеристиках данных.

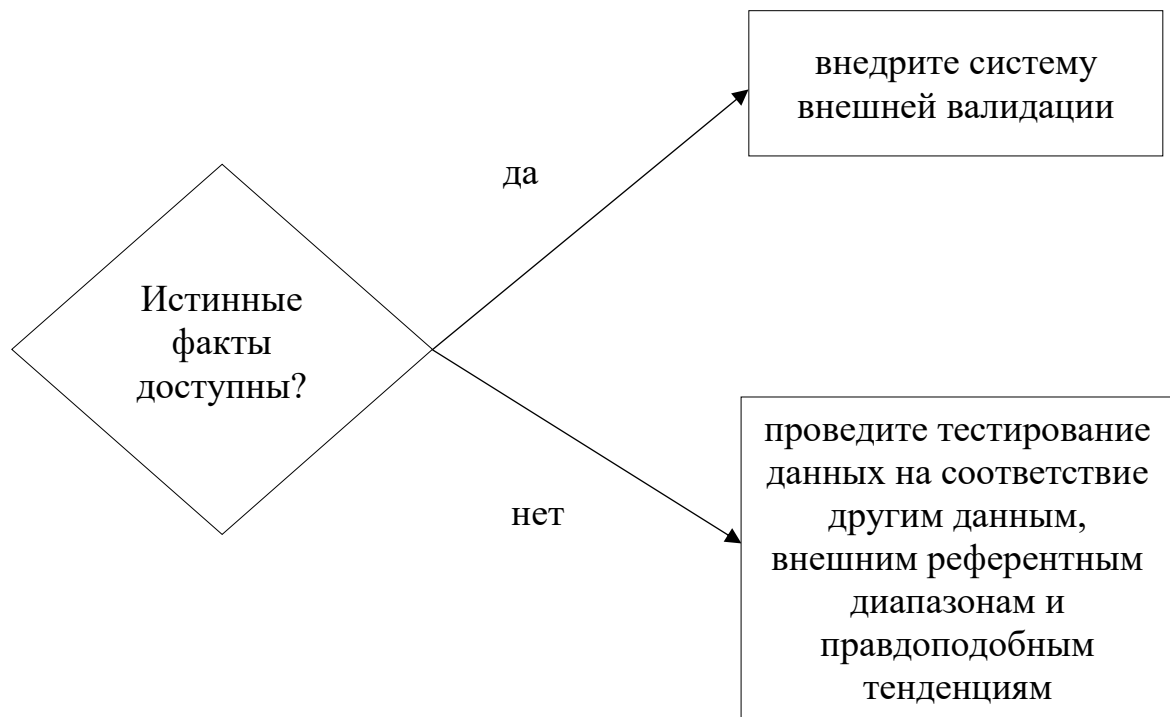


Рисунок 7. Обзор внешних референтных диапазонов