УДК 004.05:61

О.В. Красько[1], В.В. Роубо[1], Н.Н. Савва[2]

ОЦЕНКА КАЧЕСТВА ДАННЫХ В ЭПИДЕМИОЛОГИЧЕСКИХ РЕГИСТРАХ

Рассматриваются вопросы количественной оценки качества данных регистров, в основе построения которых лежит база данных. Оценки базируются на предложенных авторами индексах стандартизированности, валидности, своевременности обновления, информационной полноты, востребованности для практического использования при разработке и сопровождении регистров на основе баз данных. Предлагаются оценки стоимости информации при создании и сопровождении регистров.

Введение

Эпидемиологические регистры представляют собой коллекцию записей о пациентах с определенной нозологией. Основные цели создания таких регистров – организация помощи больным путем научно обоснованного расчета и планирования профилактических, диагностических, лечебных, реабилитационных, паллиативных и других мероприятий на основе анализа заболеваемости, выживаемости и смертности. Кроме того, результаты анализа данных эпидемиологических регистров служат доказательной аргументацией для внесения изменений в систему здравоохранения и законодательство, а также являются основой для дальнейших научных исследований в области выявления причинно-следственных связей развития болезни.

Как правило, создание подобных регистров – это первый шаг к построению эффективных социальных программ в здравоохранении и обществе в целом. Статистическая информация, поставляемая регистром для конечных пользователей (медицинских работников, Министерства здравоохранения, государственных и социальных органов), должна отражать состояние популяции (народонаселения) по показателям заболеваемости, распространенности и смертности определенного заболевания, по обеспечению медицинской помощью, а также данные мониторинга заболеваемости этой же популяции и др.

Бумажная технология ведения регистров (рис. 1) приводила к ошибкам человека, дублированию информации, пропускам в данных из-за различных исходных форматов и, как следствие, к искажению эпидемиологических данных. На рис. 1 видно, что запрос на определенные данные направлялся ответственному лицу, которое имело доступ к коллекции записей о пациентах (районного, областного и других уровней). Запрос мог формулироваться неоднозначно. Соответственно информация, передаваемая для результирующего документа, могла быть неточной, неоднозначной, нечеткой, содержать дубли и пропуски в данных, ее ручная обработка могла еще больше искажать информацию.

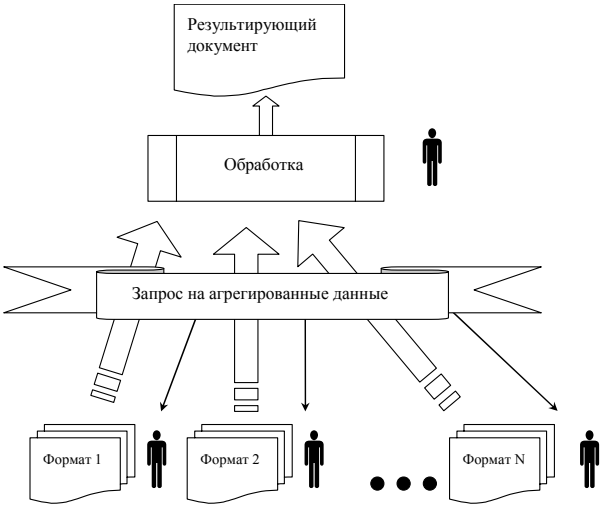

Рис. 1. Схема регистрации до появления информационных технологий

В связи с развитием коммуникационных и компьютерных технологий происходят изменения в порядке ведения регистров (рис. 2). В настоящее время информационные технологии в медицине (равно как и в других областях, где ведется статистический учет) позволяют переходить от бумажных коллекций записей к электронным формам ввода и хранения информации. Единый информационный подход позволяет создавать не только электронные регистры (субрегистры) отдельных медицинских учреждений, в том числе межцентровые, но и так называемые популяционные (областные, республиканские, национальные, европейские и пр.), целью которых является регистрация всех случаев определенных заболеваний в конкретной возрастной когорте, встречающихся на определенной территории. Примерами могут являться Детский канцер-субрегистр Республики Беларусь, регистр Европейского сообщества иммунодефицитов (European Society for Immunodeficiency, http://www.esid.org/esid_registry.php), Мировой регистр типированных доноров для трансплантации костного мозга и т. п.

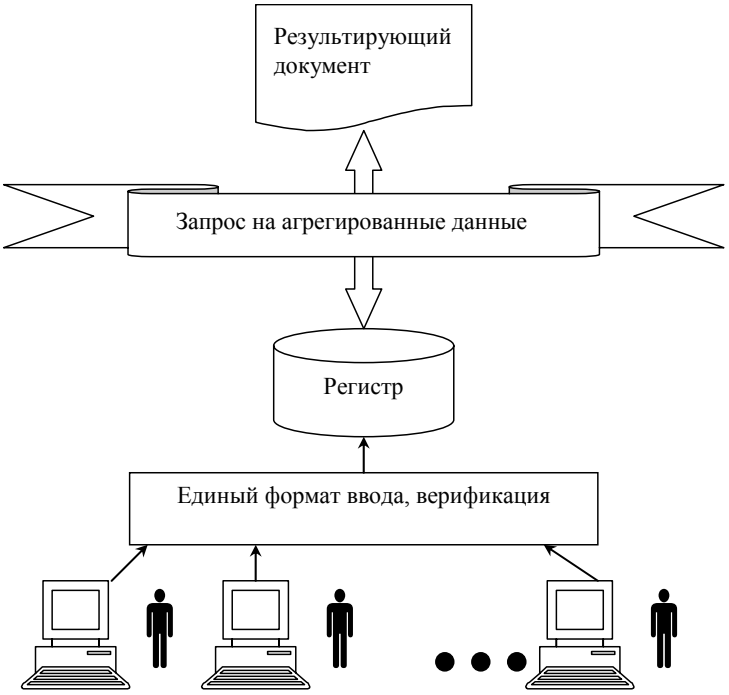

Рис. 2. Схема регистрации с использованием единого информационного подхода

В настоящей статье речь пойдет о качестве информации регистров, в основе построения которых лежит база данных. Информация, которая заносится в электронный регистр и хранится в нем, должна иметь определенные свойства, чтобы служить определенным целям предметной области (в данном случае здравоохранения) и общества в целом. Часть требований к проектированию информационного наполнения регистра будет определяться современными тенденциями к проектированию баз данных, а часть – онтологией предметной области.

Для того чтобы использовать единую терминологию в рамках тематики медицинских регистров, приведем в соответствие информационные термины (используемые в информационных системах) и термины регистров, характерных для медицины: *запись* – случай заболевания, связанный с пациентом и его характеристиками; *атрибут* – характеристика пациента, которая необходима для успешного функционирования регистра (возраст, пол, клинические данные, лабораторные исследования и пр.) [1, 2].

1. Основные свойства информации в регистрах

Единый информационный подход позволяет увязать в общую систему как различные субрегистры, так и различных потребителей информации, однако надо иметь в виду, что формализация уже имеющихся данных в каждом случае – нетривиальная задача, для решения ко-

торой при повторном создании абстрактной логической схемы данных, при повторном кодировании справочников, а также реинжиниринге бизнес-логики необходимо руководствоваться общими свойствами данных регистров (субрегистров).

Чтобы понять, как оценивать качество данных в регистрах, необходимо обратиться к свойствам данных, характерных для таких систем [3–9]. Несмотря на различие методологических подходов к выбору информативных характеристик для регистрации случаев, существуют общие принципы, на которых должны базироваться информационные характеристики регистров [2].

Стандартизированность как свойство данных объясняется тем, что сгенерированные статистические данные на основе регистра должны быть представлены в соответствии со стандартами, принятыми в мире (или стране, если нет международных стандартов, а имеются только республиканские).

Валидность (аккуратность, точность) связана, с одной стороны, с точностью учета случаев заболеваемости и смертности в интересах здравоохранения и общества (потребителей информации), а с другой – с безошибочностью заполнения и особенностями функционирования спроектированной для регистра информационной системы.

Своевременность является компромиссом между валидностью и полнотой информации.

Полнота определяется онтологией предметной области, а также потребностью общества в различных статистических данных.

Разграниченность данных также связана с онтологией предметной области. При включении в регистр пациент подписывает информированное согласие для хранения и предоставления данных в регистре. Персонифицированная информация доступна только определенным лицам, неперсонифицированная включается в отчетность и может являться предметом дальнейшей статистической обработки.

Востребованность данных соотносится с конечными потребителями информации.

Стоимость информации определяется онтологией предметной области, результаты исследований, заносимые в регистр, могут требовать затрат, несопоставимых с пользой от его ведения.

Ниже рассмотрено более подробно каждое из этих свойств.

1.1. Стандартизированность

В настоящее время соответствие регистра международным стандартам, вовлечение данных в межцентровые и международные исследования означает использование единого подхода к определению заболевания, сопутствующей информации о заболевании и об исходах заболевания. Это свойство данных нужно обеспечивать в момент их занесения в регистр. Оно распространяется на определение заболевания, т. е. стандарты выявления наличия определенного заболевания, классификаторы заболеваний, дату постановки диагноза, регион, стандартизированные исходы заболевания и пр.

Необходимо иметь в виду, что часть требований к стандартизации не имеет непосредственного отношения к данным, которые содержатся в электронном регистре, однако определяет их надежность. Сами значения показателей могут заноситься в регистр, но способ их измерения остается за кадром и влиять на него может только квалификация врача, устанавливающего диагноз. Поэтому наряду с данными в регистрах часто хранят информацию и о способах измерения (получения) данных.

Индекс стандартизированности регистра определяется как

$$k_s = \frac{r}{m},$$

где r – количество атрибутов, значения которых заносятся в регистр по результатам стандартизированных процедур или на основе стандартизированных справочников или классификаторов; m – число атрибутов записи.

Чем выше индекс стандартизированности, тем больше регистр соответствует нормам и стандартам, а данные из него могут использоваться для сравнения с аналогичными данными регистров других регионов (стран) и быть пригодными для межцентровых исследований.

При межцентровых исследованиях возможно сравнение информации из регистров (субрегистров) на основе значений атрибутов, представленных стандартизированными справочниками. Анализируя частоту использования значений стандартизированного справочника в различных субрегистрах, можно определить гомогенность выборок для возможности дальнейшего сравнения данных и их агрегации [10].

1.2. Валидность

Валидность определяет аккуратность, точность и непротиворечивость данных, касающихся одного случая заболевания. Исторически ведение регистров (субрегистров) предполагало собственное кодирование атрибутов, поэтому при компьютеризации и централизации регистров важное место отводится повторному проектированию логической организации данных и повторному кодированию данных, анализу и восстановлению утерянных данных, определению их внутренней непротиворечивости [3].

Валидность данных может определяться через индекс сложности валидации:

$$k_v = \frac{1}{m}\sum_{p=1}^{P}\left(C_{q_p}^2\right),$$

где q_p – количество атрибутов, вовлеченных в одно правило валидации, $q_p > 1$; $C_{q_p}^2$ – количество комбинаций из q_p по 2; P – количество правил валидации; m – число атрибутов записи.

Например, дата рождения не должна быть позже даты постановки диагноза. Одновременно атрибут <возраст при установлении диагноза> может являться перекрестным для валидации всех трех атрибутов.

Во многих случаях использование справочников в качестве возможных значений атрибутов помогает решить проблему точности и аккуратности данных, однако при этом усложняется реализация правил и ограничений автоматизируемых операций приложения для ввода и верификации данных в регистр, поскольку использование определенных значений из одного справочника ведет к ограничениям на использование данных других справочников. Разработка множества правил валидации опирается на опыт врачей-специалистов. При проектировании приложения возможны два варианта реализации валидации: при вводе данных и при формировании отчетов. Первый вариант намного предпочтительнее, поскольку большинство ошибок связано с вводом данных и может быть своевременно исправлено до стадии формирования отчетов[1].

1.3. Своевременность

Потребители информации нуждаются в свежих данных за текущий период, однако специфика валидации каждого нового случая заболевания или летального исхода должна быть верифицирована по определенному протоколу после проведения медицинского обследования. Своевременность данных может оцениваться через индекс своевременности обновления данных k_t, который определяется периодичностью обновления записей, а также частотой агрегирования данных для получения результирующих документов (статистической отчетности) для потребителей информации. Каждое обновление – это время, потраченное на опрос или обследование пациента. Чем больше данных обновляется и чем чаще это делается, тем выше затраты на сопровождение регистра.

Индекс своевременности обновления можно рассчитать следующим образом:

$$k_t = \frac{T}{z},$$

[1]На сегодняшний момент оптимальным подходом в проектировании валидации такого рода является использование языка XML для записи правил валидации. Это позволяет быстро изменять, добавлять и удалять фрагменты бизнес-логики при проектировании приложения [11].

где T – время до получения новых данных о случае заболевания (как правило, речь идет о наблюдениях и исходах (результатах лечения) некоторого случая заболевания); z – частота запросов на статистические отчеты (раз в месяц, раз в год и т. п).

Значение $k_t > 1$ свидетельствует о том, что или частота запросов избыточна (потребитель не получит изменений в данных; выполнение операций, связанных с генераций результатов запроса, будет лишним, избыточным), или изменения в популяции фиксируются реже, чем это требуется конечным потребителям информации. В любом случае сложность сопровождения регистра для медика-регистратора определяется количеством регулярно обновляемой информации, поэтому, как правило, регистр не должен иметь большого количества атрибутов частого обновления.

1.4. Полнота информации

Полнота информации о случае заболевания – важное свойство, которое определяется потребностью конечного потребителя в эпидемиологических данных.

Проектирование регистра с расчетом на появление информации о некоторых атрибутах в будущем редко оправдано, поскольку затраты на создание регистра возрастают, требования же конечного пользователя к информации могут изменяться и атрибуты остаются невостребованными. Индекс заполнения (полноты атрибута) i может определяться следующим образом:

$$k_{fill}^{i} = \frac{n_{fill}^{i}}{n},$$

где n_{fill}^{i} – количество записей в регистре с имеющейся информацией по атрибуту i; n – общее количество записей в регистре.

По индексу заполнения можно определить, насколько действительно доступна и полна информация по некоторой характеристике случая заболевания.

Информационная полнота регистра в целом может быть определена как среднее:

$$k_{fill} = \frac{1}{m}\sum_{i=1}^{m} k_{fill}^{i},$$

где m – общее число атрибутов записи.

Сравнение различных регистров (субрегистров) по данному показателю может оказаться полезно при последовательном анализе данных, например характеристик популяций, проживающих в различных регионах. Низкое значение информационной полноты может свидетельствовать как о плохой организации регистрации в определенной субпопуляции (регионе), так и о сложности или высокой стоимости получения информации по данной характеристике в субпопопуляции.

1.5. Разграниченность данных

При установлении наличия заболевания пациенту предлагается подписать так называемое информированное согласие об использовании неперсонифицированной части его данных в исследованиях, отчетах и пр. При проектировании приложения необходимо учитывать права доступа к персонифицированной и неперсонифицированной информации. Индекс доступности данных может определяться как

$$k_a = \frac{m_a}{m},$$

где m_a – количество атрибутов неперсонифицированной информации; m – общее число атрибутов записи.

Низкое значение индекса обычно имеют специализированные регистры; популяционные регистры имеют индекс, близкий к единице.

1.6. Востребованность данных

При проектировании регистров конечный потребитель далек от первичных атрибутов (характеристик). Одной из ошибок проектирования является несогласование с конечным потребителем необходимого и достаточного объема данных на ранних стадиях проектирования и эксплуатации регистра. Заполнение всех информационных характеристик порой избыточно и, как результат, не востребовано конечным потребителем. Это создает ненужные стоимостные, временные и человеческие затраты.

Индекс востребованности i-го атрибута можно рассчитать как отношение количества запросов за год с использованием i-го атрибута к общему числу запросов за единицу времени (раз в месяц, раз в год и т. п.):

$$k_{need}^{i} = \frac{u_i}{u},$$

где u_i – количество запросов с использованием i-го атрибута за единицу времени; u – общее число запросов к регистру за единицу времени.

2. Стоимость получения и сопровождения информации

Необходимость определения стоимости данных обосновывается следующими причинами:

1. Часть данных эпидемиологического регистра может быть получена от пациента, например исторические данные о нахождении пациента в неблагоприятных условиях (радиации, химического воздействия и др.). Как правило, эти данные нуждаются в объективном контроле, который занимает много времени и средств (запросы в соответствующие органы о проживании, работе в неблагоприятных условиях и пр.), их количество в регистре должно быть обусловлено целями создания регистра, необходимостью иметь такие данные и их целесообразностью. Формализация ответов пациента в этом случае (например, срок работы во вредных условиях труда) необходима для последующего анализа.

2. Каждый клинико-лабораторный атрибут нового случая в регистре требует проведения определенных тестов, которые подтверждают наличие заболевания или определяют его стадию. Некоторые из тестов являются дорогостоящими. Необходимо понимать, что не все субрегистры в состоянии обеспечивать наличие такой информации. Как правило, дорогостоящие тесты используются при проведении клинических исследований, но не при сопровождении эпидемиологических регистров.

3. Как правило, в проектировании регистра принимают участие врачи-специалисты по заболеваниям определенной нозологии, а не экономисты, поэтому существуют определенные сложности в оценке стоимости получения информации.

4. Создание и сопровождение регистра ограничено бюджетными возможностями организации, министерства либо ведомства, поэтому важно заранее оценить приблизительную стоимость проектов по регистрации.

Стоимость получения единицы данных (записи в регистре) может быть рассчитана как

$$C_G = \frac{1}{R}\left(\sum_{w=1}^{W} C_w\right) + \sum_{s=1}^{S} C_s ,$$

где R – распространенность заболевания в исследуемой популяции (количество имеющих заболевание, деленное на размер исследуемой популяции); C_w – стоимость теста подтверждения заболевания; W – количество тестов, необходимых для подтверждения заболевания при занесении в регистр; C_s – стоимость запросов (обращений в соответствующие организации/органы) по историческим данным нового пациента, $s = 1, \ldots, S$; S – количество атрибутов регистра, требующих запросы в соответствующие органы для объективного контроля.

Стоимость сопровождения информации (актуализация) связана в основном с популяционными эпидемиологическими регистрами. Если в клинических регистрах (исследовательских) основная доля затрат приходится на получение информации, то в популяционных регистрах основная доля затрат – это сопровождение и актуализация. Данная стоимость может варьироваться в зависимости от количества обновляемых атрибутов; ее можно рассчитать по формуле $C_A = n\sum_{i=1}^{U} C_i$, где C_i – стоимость обновления i-го атрибута, $i = 1, \ldots, U$; U – количество обновляемых атрибутов регистра; n – общее количество записей в регистре. Если учесть периодичность запросов к регистру и индекс своевременности обновления регистра, можно рассчитать стоимость сопровождения регистра за единицу времени (месяц, год).

Заключение

В настоящей работе рассмотрены количественные характеристики оценки качества данных в регистрах. При разработке регистров определенных нозологий существуют дополнительные свойства данных, связанные с манифестацией и протеканием заболевания, лечением и состоянием пациента в течение длительного срока после его окончания, однако их изложение выходит за рамки данной статьи. Приведенные же свойства данных универсальны, их целесообразно оценивать как при создании, так и в процессе эксплуатации регистров различного назначения.

Количественные характеристики качества данных помогают оценить целесообразность создания регистра, выявить технические ошибки при занесении данных и дублирование информации, а также пригодность информации для статистической обработки в интересах конечных пользователей. Контроль качества данных используется как на локальном уровне, так и при проведении мультинациональных исследований, в которых принимают участие регистры разных стран.

Предложенные оценки успешно используются в Республиканском научно-практическом центре детской онкологии и гематологии. Так, при проектировании регистра первичных иммунодефицитов Республики Беларусь этот подход позволил оптимизировать его структуру на этапе создания в 2007 г. Плановая оценка внесенных данных в работе Детского канцер-субрегистра Республики Беларусь, функционирующего с 1999 г., позволяет своевременно верифицировать информацию и максимально исключить влияние «человеческого фактора» на результаты обработки данных.

Авторы рассматривали свойства данных на примере эпидемиологических регистров. Однако описанный подход к оценке качества данных может быть распространен и на другие системы регистрации, например, в экологии, социальной службе, при демографических исследованиях и др.

Обеспечение качества данных за счет грамотного проектирования и сопровождения регистров – это не только залог эффективных решений в медицине, но и гарантия повышения качества жизни каждого человека и общества в целом.

Список литературы

1. Дейт, К.Дж. Введение в системы баз данных / К.Дж. Дейт. – 8-е изд. – М. : Издательский дом «Вильямс», 2006. – 1328 с.

2. Registries for Evaluating Patient Outcomes: A User's Guide. Prepared by Outcome DEcIDE Center (Outcome Sciences, Inc. dba Outcome) under Contract No. HSA290200500351 TO1 ; eds. R.E. Gliklich, N.A. Dreyer ; AHRQ Publication. – No. 07-EHC001-1. – UK, Rockville : Agency for Healthcare Research and Quality, 2007.

3. Bray, F. Evaluation of data quality in the cancer registry: principles and methods Completeness / F. Bray, D.M. Parkin // European Journal of Cancer. – 2009. – Vol. 45, № 5. – P. 747–764.

4. Хоффманн, Э. Наблюдение и описание международной миграции: проблемы качества данных при использовании регистрационных записей государственной иммиграционной служ-

бы как источника статистической информации / Э. Хоффманн // Вопросы статистики. – 2007. – № 2. – С. 12–16.

5. Надлежащая клиническая практика : ГОСТ Р 52379–2005. – Введ. 01.04.2006. – М. : Изд-во стандартов, 2005. – 34 с.

6. Batini, C. Data quality: concepts, methodologies and techniques / C. Batini, M. Scannapieco. – Springer, 2006. – 262 c.

7. Olson, J.E. Data Quality, the accuracy dimension / J.E. Olson. – Morgan Kaufmann, 2003. – 294 p.

8. Kennedy, L. Global registries for measuring pharmacoeconomics and quality-of-life outcomes: focus on design and data collection, analysis and interpretation / L. Kennedy, A.M. Craig // Pharmacoeconomics. – 2004. – Vol. 22, № 9. – P. 68–551.

9. Evaluation and implementation of public health registries / D.J. Solomon [et al.] // Public Health Rep. – 1991. – Vol. 106, № 2. – P. 50–142.

10. Agresti, A. Categorical Data Analysis / A. Agresti. – 2nd ed. – N.Y. : John Wiley & Sons, 2002. – 710 p.

11. Кэгл, К. XML / К. Кэгл. – СПб. : Питер, 2007. – 784 с.

Поступила 19.04.2010

[1]*Объединенный институт проблем информатики НАН Беларуси, Минск, Сурганова, 6*
e-mail: krasko@newman.bas-net.by

[2]*Республиканский научно-практический центр детской онкологии и гематологии, Минский р-н, пос. Лесное-2*
e-mail: nsavva@mail.ru

O.V. Krasko, V.V. Roubo, N.N. Savva

DATA QUALITY EVALUATION IN EPIDEMIOLOGICAL REGISTRY

The paper addresses the problem of evaluation of data quality in epidemiological registry databases. Proposed indices of «standardness», validity, timeliness, completeness, needfulness are useful for practical aspects of registration based on modern information technology. Also, evaluation of cost-efficacy of information registry is briefly discussed.