Системы оценки достоверности научных доказательств и убедительности рекомендаций: сравнительная характеристика и перспективы унификации

Н. С. Андреева[1], О. Ю. Реброва[2], Н. А. Зорин[3], М. В. Авксентьева[1, 4], В. В. Омельяновский[1]

[1] Национальный центр по оценке технологий в здравоохранении (НЦ ОТЗ), Москва, Россия

[2] Российский национальный исследовательский медицинский университет (РНИМУ) им. Н. И. Пирогова Минздрава России, Москва, Россия

[3] Научный центр экспертизы средств медицинского применения (НЦЭСМП) Министерства здравоохранения Российской Федерации, Москва, Россия

[4] Первый Московский государственный медицинский университет (МГМУ) им. И. М. Сеченова Минздрава России, Москва, Россия

Принятие решений о применении медицинских технологий и их включении в клинические руководства должно быть основано на комплексном анализе всех имеющихся научных доказательств их эффективности и безопасности. В данном обзоре описаны системы оценки достоверности научных доказательств и убедительности рекомендаций, принятые в настоящее время известными международными агентствами по оценке медицинских технологий и организациями, разрабатывающими клинические руководства (SIGN, OCEBM, GRADE, NICE, NHMRC). Проведен также сравнительный анализ критериев, используемых для оценки достоверности доказательств (таких, как качественная и количественная характеристики и согласованность доказательств) и для оценки убедительности рекомендаций (таких, как обобщаемость доказательств, соотношение между пользой и вредом от вмешательства, стоимость лечения, ценности и предпочтения пациентов, применимость рекомендаций в условиях национальной системы здравоохранения). Кроме того, проанализированы принципы классификации доказательств эффективности медицинских технологий по уровням достоверности и классификации клинических рекомендаций по уровням убедительности. В заключительной части обзора рассмотрены перспективы введения единой системы оценки достоверности доказательств и убедительности рекомендаций за рубежом и в России.

КЛЮЧЕВЫЕ СЛОВА: уровни достоверности доказательств, уровни убедительности рекомендаций, GRADE, системы оценки, клинические руководства.

ВВЕДЕНИЕ

Одним из принципов доказательной медицины (ДМ) является положение о том, что принятие решений о применении тех или иных медицинских технологий должно быть основано на комплексном анализе всех имеющихся научных доказательств их эффективности и безопасности, а не на мнении экспертов и личном клиническом опыте врачей. Это касается и разработки клинических руководств, которые должны соответствовать принципам ДМ, причем авторы руководств должны отчетливо излагать степень своей уверенности в достоверности научных доказательств и в обоснованности рекомендаций. Для обозначения степени уверенности в достоверности научных доказательств и в обоснованности клинических рекомендаций были созданы соответствующие системы оценок и уровней, получившие название «уровни достоверности доказательств» (**levels of evidence**) и «уровни убедительности рекомендаций» (**grades of recommendation**). Те и другие обозначаются обычно либо римскими цифрами, либо латинскими буквами и призваны информировать пользователей руководств (прежде всего, врачей, но также средних медицинских работников, менеджеров здравоохранения и пациентов) о степени научной обоснованности рекомендуемого варианта действий. Формализованный систематизированный подход к оценке достоверности доказательств эффективности медицинских технологий и убедительности рекомендаций способствует предотвращению ошибок в суждениях, их критическому восприятию и распространению среди специалистов в области здравоохранения [10].

В настоящее время римские цифры и латинские буквы, обозначающие достоверность доказательств и убедительность рекомендаций, включаются и в руководства, подготовленные и/или изданные в России. При этом не все врачи осознают, что до сих пор отсутствует единая система их оценки, а зачастую и смысл процесса, предшествующего расстановке

цифр и букв, остается неясным. Как следствие, представления о том, что подразумевается под «убедительными доказательствами» и доказательствами, «основанными на лучших научных данных», в руководствах существенно варьируют. Специалисты, пытавшиеся в последние годы вдумчиво использовать упомянутые системы оценки, сталкиваются со значительными трудностями, связанными с отсутствием единых подходов к формированию таких систем.

Целью данного обзора было проанализировать системы оценки достоверности научных доказательств и убедительности рекомендаций, разработанных и используемых в настоящее время за рубежом; результаты проведенного анализа могут помочь созданию аналогичных систем оценки в России.

ОСНОВНЫЕ ТЕРМИНЫ И ПОНЯТИЯ

Уровни достоверности доказательств (УДД) отражают степень уверенности в том, что найденный эффект от применения медицинской технологии является истинным. Согласно эпидемиологическим принципам, достоверность доказательств определяется по трем основным критериям: по качественной и количественной характеристикам и по согласованности доказательств (вставка) [1, 2]. **Качественная характеристика доказательств (quality)** – это сводный показатель методологического качества всех доступных исследований, отвечающих на поставленный клинический вопрос (например, какова эффективность применения конкретной медицинской технологии). Под **методологическим качеством исследования** подразумевают степень, с которой дизайн исследования, методы его проведения и анализа данных предупреждают возникновение и минимизируют влияние систематических и случайных ошибок, способных привести к искажению истинного размера эффекта и, соответственно, снизить достоверность результатов. Важным является понятие **иерархии дизайнов исследований**, отражающее тот факт, что одни типы дизайнов подвержены более сильному влиянию систематических ошибок, чем другие, и, следовательно, результаты таких клинических исследований заведомо обладают меньшей достоверностью. На данный момент доказано, что наибольшей достоверностью обладают результаты рандомизированных клинических исследований (РКИ), а также их систематических обзоров и метаанализов.

Для определения методологического качества исследований используют различные системы оценки, в основном это вопросники или шкалы, разработанные для отдельных типов дизайна клинических испытаний. В 2002 г. исследователи Американского агентства по проведению и оценке качества исследований в области здравоохранения (AHRQ) в результате систематического поиска нашли 20 систем оценки методологического качества, используемых для систематических обзоров, 49 систем для РКИ и 19 для обсервационных исследований [2]. В настоящее время не существует единой, всеми принятой системы для оценки методологического качества исследований. При этом уровень методологического качества одних и тех же исследований может меняться в зависимости от того, какая шкала или вопросник были использованы для оценки. Например, Juni с соавт. сравнили 25 разных шкал для оценки методологического качества РКИ, изучавших эффективность низкомолекулярного и стандартного гепарина в предотвращении послеоперационного тромбоза, и продемонстрировали, что исследованиям с высоким методологическим качеством, определенным по одной шкале, может быть присвоено низкое методологическое качество при использовании других шкал [11].

Критерии оценки достоверности доказательств

Качественная характеристика доказательств: сводный показатель методологического качества всех доступных исследований.

Количественная характеристика (объем) доказательств: размер эффекта, количество исследований, суммарный размер выборки пациентов.

Согласованность доказательств: степень совпадения результатов различных исследований.

Количественная характеристика, или объем доказательств[1] (quantity) зависит от трех составляющих:

- размер (величина) эффекта;
- количество исследований;
- суммарный размер выборки пациентов в исследованиях.

Теоретически чем больше размер эффекта, тем ниже вероятность того, что возникшие при проведении исследования систематические и случайные ошибки могли привести к получению ложного результата. Очевидно также, что чем больше количество исследований (если только они высокого методологического качества), тем выше достоверность наблюдаемого эффекта. Кроме того, чем больше величина суммарной выборки пациентов, тем у́же доверительный интервал для оценки эффекта и, следовательно, выше точность этой оценки. Таким образом, количественная характеристика доказательств определяет степень уверенности в том, что наблюдаемый эффект был не случайным, а вызванным воздействием изучаемого вмешательства.

Согласованность доказательств (consistency) демонстрирует степень совпадения результатов различных исследований, посвященных изучению эф-

[1] Единой русскоязычной терминологии пока не сложилось, поэтому в ряде случаев мы приводим несколько возможных синонимов.

фективности одной и той же медицинской технологии (проведенных в разных популяциях, с одинаковым или различным дизайном). Высокая степень совпадения результатов разных исследований говорит о том, что полученный эффект является воспроизводимым, и это повышает степень достоверности результатов.

Таким образом, оценка достоверности доказательств начинается с оценки методологического качества отдельных исследований и совокупности исследований и дополняется оценкой количественных характеристик и согласованности доказательств. При этом оценка достоверности доказательств является хотя и обязательным, но недостаточным условием для вынесения клинических рекомендаций. **Клинические рекомендации** характеризуются как разработанные с помощью систематизированных методов положения, целью которых является содействие практикующим врачам и пациентам в выборе целесообразного метода лечения в конкретных клинических обстоятельствах. Исходя из этого определения, экспертные группы, отвечающие за составление рекомендаций, должны явным образом учитывать не только достоверность доказательств, но и ряд дополнительных факторов: соотношение между пользой и вредом от применения медицинской технологии, обобщаемость доказательств (распространение доказательств эффективности на конкретные популяции больных и условия клинической практики), ценности и предпочтения пациентов, стоимость лечения [3,4]. **Уровни убедительности рекомендаций** (УУР), в отличие от УДД, отражают не только степень уверенности в достоверности эффекта вмешательства, но и степень уверенности в том, что следование рекомендациям принесет больше пользы, чем вреда в конкретной ситуации.

На рис. 1 изображен процесс разработки клинических руководств от оценки методологического качества исследований до формирования клинических рекомендаций через оценку достоверности доказательств. Этапы определения достоверности доказательств и составления рекомендаций должны проводиться последовательно, но лучше отдельно друг от друга с присвоением отдельного УДД и отдельного УУР. Несмотря на то что высокий УДД в большинстве случаев ассоциируется с высоким УУР, доказательства с определенным УДД не всегда подразумевают точно такой же УУР. Например, в высококачественных РКИ было продемонстрировано, что продолжительный прием оральных антикоагулянтов приводит к снижению риска возникновения повторного тромбоза у пациентов с тромбозом глубоких вен в анамнезе. Однако продолжительный прием антикоагулянтов также ведет к повышению риска развития кровотечений, возникновению необходимости в регулярном контроле показателей свертываемости крови и появлению дополнительных затрат на лечение [18–21]. В результате для части пациентов преимущество от применения антикоагулянтов будет нивелироваться недостатками лечения, и выбор пациентов (принимать или не принимать антикоагулянты) будет неоднозначным. Следовательно, рекомендации, предлагающие продолжительное применение оральных антикоагулянтов для всех пациентов, скорее всего, будут иметь низкую степень убедительности, несмотря на высокую степень достоверности доказательств [12, 22].

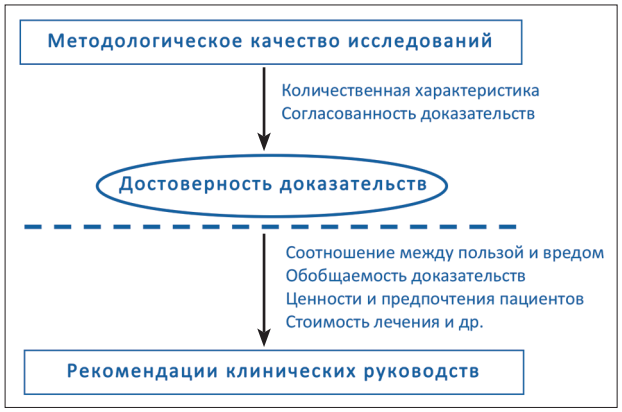

Рис. 1. **Процесс разработки клинических руководств от оценки методологического качества исследований до формирования клинических рекомендаций через оценку достоверности доказательств (адаптировано из [2]).**

ИСТОРИЯ РАЗВИТИЯ СИСТЕМ ОЦЕНКИ УРОВНЕЙ ДОСТОВЕРНОСТИ ДОКАЗАТЕЛЬСТВ И УБЕДИТЕЛЬНОСТИ РЕКОМЕНДАЦИЙ

Первая концепция уровней достоверности доказательств и убедительности рекомендаций была представлена в 1979 г. Канадской рабочей группой по профилактической медицине (Canadian Task Force on the Periodic Health Examination) [5]. С тех пор были созданы и использованы на практике многочисленные системы по оценке качества и достоверности доказательств и убедительности рекомендаций. При этом разработкой и внедрением новых систем занимались не только крупные ассоциации по оценке медицинских технологий, но и многие специализированные медицинские сообщества. В 2002 г. при выполнении систематического обзора американским агентством AHRQ было найдено 40 имевшихся на тот период подходов к оценке достоверности доказательств [2]. В продолжение данной работы канадская рабочая группа COMPUS (Canadian Optimal Medication Prescribing and Utilization Service) дополнила этот список еще 10 методами оценки достоверности доказательств, появившимися за период с 2000 по 2005 гг. [6]. С годами происходило постепенное усложнение принципов оценки достоверности доказательств. Только 30 %

систем оценки (7 из 23), опубликованных до 2000 г., в той или иной степени учитывали как качественную и количественную характеристики, так и согласованность доказательств; среди систем оценки, опубликованных с 2000 по 2002 гг., уже 82 % (9 из 11) систем рассматривали все три фактора [2]. Кроме того, если в ранних системах оценки качество доказательств зависело только от положения исследований в иерархии дизайнов, то в настоящее время определение методологического качества каждого из исследований является обязательным компонентом оценки. В процессе развития методов этой оценки позднее всех остальных критериев стали учитывать согласованность доказательств [2].

Из 40 систем оценки, проанализированных в 2002 г. американским агентством AHRQ, только 12 систем (30 %) были созданы не для написания клинических рекомендаций, а исключительно в качестве инструмента доказательной медицины для проведения систематического анализа достоверности данных. Бóльшая часть систем объединяли оценку достоверности доказательств с оценкой убедительности рекомендаций [2]. Первая шкала УУР, разработанная Канадской рабочей группой по профилактической медицине в 1979 г., была представлена вместе с УДД и основана в первую очередь на оценке достоверности доказательств, однако авторы уже в то время отметили возможность понижения степени уверенности рекомендации в зависимости от «бремени страданий», вызванных вмешательством [5].

На данный момент существование большого количества разных шкал УДД и УУР приводит в замешательство как работников здравоохранения, принимающих решения, так и практикующих врачей, а, следовательно, снижает качество медицинского обслуживания пациентов. Беспорядок усугубляется использованием в разных шкалах различных обозначений уровней: с помощью букв (A, B, C и т.д.), чисел (I, II, III и т.д.) или их комбинаций (Ia, Ib, IIa и т.д.). Например, при оценке целесообразности назначения антивирусной терапии пациентам с острым гепатитом **C** для достижения клиренса вирусной РНК присвоенные УУД и УУР были обозначены как **1+** и **A** (SIGN, 2006 г. [7]), **B** и **I** (American Association for the Study of Liver Diseases, 2009 г. [8]), **B** и **2** (European Association for the Study of Liver, 2011 г. [9]). Очевидно, что в сложившейся ситуации системы уровней в том разнообразии, которым они представлены сейчас, не способны выполнять возложенную на них функцию: кратко и оперативно информировать о достоверности доказательств эффективности медицинских технологий и об убедительности рекомендаций. Все большее число соответствующих международных организаций признает необходимость введения единого подхода к оценке достоверности доказательств и убедительности рекомендаций.

В России еще в конце 1990-х годов было признано необходимым при разработке отечественных клинических руководств использовать принципы доказательной медицины и опираться только на результаты методологически строгих исследований [23, 24]. Однако в значительной части современных отечественных клинических руководств приводятся положения, не обоснованные научными доказательствами, а системы оценки достоверности доказательств и убедительности рекомендаций остаются в России невостребованными и неразвитыми [17]. Большинство врачебных сообществ и ассоциаций предпочитает опираться в своей практике на уровни достоверности доказательств и убедительности рекомендаций, взятые из зарубежных клинических руководств, зачастую забывая о необходимости их адаптации к условиям реальной практики в России. Основными причинами, не позволяющими напрямую использовать зарубежные руководства и результаты зарубежных клинических исследований, являются: популяционные различия между контингентами больных; несопоставимость условий лечения в лечебных учреждениях; различия факторов риска, тяжести течения и исходов заболевания; различия в затратах на лечение и распределении бюджета здравоохранения; отличия в подготовке медицинских специалистов [17]. С развитием в России доказательной клинической медицины, с увеличением числа высококачественных отечественных клинических исследований, с возрастающей ролью анализа эффективности затрат на лечение в условиях рыночной экономики неизбежно будет расти потребность в совершенствовании и использовании инструментов оценки достоверности доказательств и убедительности рекомендаций.

ХАРАКТЕРИСТИКА СИСТЕМ ОЦЕНКИ ДОСТОВЕРНОСТИ ДОКАЗАТЕЛЬСТВ И УБЕДИТЕЛЬНОСТИ РЕКОМЕНДАЦИЙ

Для сравнительного анализа, представленного в настоящем обзоре, были выбраны наиболее известные системы оценки достоверности доказательств и убедительности рекомендаций: SIGN, OCEBM, GRADE, NICE и NHMRC. При анализе принципов оценки достоверности доказательств основное внимание было уделено трем основным критериям: качественная и количественная характеристики доказательств и согласованность доказательств. При анализе принципов оценки убедительности рекомендаций за основополагающие критерии были приняты: обобщаемость доказательств, соотношение между пользой и вредом от вмешательства, стоимость лечения, ценности и предпочтения пациентов, применимость рекомендаций в условиях национальной системы здравоохранения. Далее

представлено краткое описание выбранных систем оценки (согласно последней версии, доступной в мае 2012 г.)[2].

SIGN (SCOTTISH INTERCOLLEGIATE GUIDELINES NETWORK) [25]

Шотландская межколлегиальная организация по разработке клинических рекомендаций (SIGN) была основана в 1993 г. с целью формирования рекомендаций для национальной системы здравоохранения Шотландии. Клинические рекомендации SIGN предназначены для широкого круга работников здравоохранения и охватывают различные клинические области. Система SIGN в том виде, в котором используется сейчас, была принята в 2000 г.

Разработка клинических рекомендаций SIGN начинается с определения дизайна и методологического качества исследований, каждому из которых присваивается УДД (**levels of evidence**). Оценка методологического качества проводится с помощью вопросников, разработанных для каждого типа дизайна исследований (за основу были взяты вопросники MERGE Комитета по охране здоровья Нового Южного Уэльса, Австралия). Для предотвращения возможной предвзятости при оценке методологического качества каждое исследование проходит оценку по крайней мере двух независимых экспертов. Любые разногласия в оценке разрешаются в процессе обсуждения всей экспертной группой или приглашенным независимым экспертом. Оценка качества когортных и случай-контроль исследований осуществляется экспертами, обладающими опытом в соответствующей клинической области. Шкала УДД имеет 8 категорий, начинается с **1++** (наиболее достоверные доказательства: высококачественные систематические обзоры РКИ, РКИ с очень низким риском систематических ошибок) и заканчивается **4** (наименее достоверные доказательства: мнения экспертов).

После присвоения каждому из исследований УДД мультидисциплинарная экспертная группа переходит непосредственно к процессу формирования рекомендаций, который включает три этапа. На первом этапе принимается решение о **сводном** УДД для совокупности исследований (**overall levels of evidence**). При этом эксперты учитывают качественные и количественные характеристики доказательств, согласованность доказательств, обобщаемость результатов исследований и степень применимости доказательств к целевой популяции пациентов. Во время второго этапа эксперты должны прокомментировать и принять во внимание ряд факторов, связанных с выполнением рекомендаций: доказательства потенциального вреда от выполнения рекомендаций; объем ресурсов, требующихся для реализации рекомендаций; соотношение между преимуществами и недостатками рекомендуемых мер для отдельных подгрупп пациентов; практическая осуществимость рекомендаций. В результате, опираясь на **сводный** УДД и анализ возможных последствий применения рекомендаций, экспертная группа присуждает им один из четырех УУР (**grades of recommendation**): **A, B, C** или **D**. Эти уровни убедительности не отражают клиническую значимость рекомендаций, а показывают, с какой вероятностью при выполнении рекомендаций будет достигнут нужный клинический эффект. Поэтому на третьем этапе эксперты имеют возможность выбрать ключевую рекомендацию (**key recommendation**), которая, по их мнению, окажет наибольшее влияние на состояние здоровья и качество жизни пациентов. Кроме того, рабочая группа по составлению рекомендаций имеет право отметить так называемые «принципы надлежащей практики» (**good practice points**), необходимость соблюдения которых настолько очевидна, что не требует научных доказательств, т.е. их научное доказательство было бы абсурдным.

Несмотря на то, что экспертная группа предоставляет полное обоснование принятого решения, четких критериев для определения **сводного** УДД и УУР в системе SIGN не существует.

К сильным сторонам системы SIGN относится прозрачность процесса оценки методологического качества исследований по вопросникам с присвоением УДД каждому исследованию. В результате УДД для отдельных исследований являются хорошо воспроизводимыми, в отличие от сводных УДД и УУР, для определения которых не прописаны четкие критерии. В 2009 г. организация SIGN приняла решение перейти на систему оценки качества доказательств и убедительности рекомендаций GRADE. На сайте организации размещено положение об основных принципах оценки GRADE, которые SIGN планирует использовать в будущем [26].

OCEBM (OXFORD CENTER FOR EVIDENCE-BASED MEDICINE) [27–30]

Работа Оксфордского центра доказательной медицины (OCEBM) направлена на распространение и развитие принципов доказательной медицины. Шкала уровней УДД OCEBM, впервые представленная в 1998 г., была подвергнута ревизии в 2009 и 2011 гг. Последний вариант шкалы УДД (**levels of evidence**), как и все предыдущие версии, был создан в качестве инструмента, призванного помочь практикующим врачам и пациентам самостоятельно ориентироваться в быстро нарастающем объеме медицинских дан-

[2] На сайте журнала (www.hta-rus.ru/journal) приведено приложение к данной статье, где дана подробная характеристика всех проанализированных систем оценки.

ных. Шкала УДД OCEBM представлена в нескольких вариантах, и выбор одного из них зависит от поставленного клинического вопроса, который может быть связан с распространенностью медицинской проблемы, эффективностью диагностического теста или скрининга, прогнозом заболевания, определением пользы или потенциального вреда от лечения. УДД присваивается отдельным исследованиям в первую очередь на основании дизайна. При этом не исключается возможность понижения уровня достоверности на основании неудовлетворительного методологического качества исследования, а также при неточности измерения эффекта, незначительном размере эффекта, гетерогенности результатов исследований или косвенности доказательств; с другой стороны, возможно повышение уровня достоверности при значительном размере эффекта. Однако конкретных критериев, насколько и в каких случаях должен быть снижен или поднят УДД, авторы системы оценки OCEBM не приводят. Шкала УДД OCEBM имеет 5 категорий; например, при оценке терапевтической пользы вмешательства УДД измеряется от **1** (наиболее достоверные доказательства: систематические обзоры РКИ или N-из-1[3] исследований) до **5** (наименее достоверные доказательства: имеется лишь обоснование механизма действия).

Система OCEBM не предусматривает оценку достоверности совокупности доказательств и убедительности рекомендаций. К сильным ее сторонам относятся простота в использовании и то, что в ней учтен весь спектр клинических вопросов, которые могут возникнуть перед врачом или пациентом в условиях реальной практики. Недостатком системы является отсутствие конкретных критериев для определения методологического качества исследований, а также для определения финального УДД. Однако сами авторы указывают на то, что данная система служит лишь дополнением к традиционным системам критической оценки доказательств и может быть использована врачами и пациентами лишь в качестве эвристического метода для быстрого поиска ответов на возникающие клинические вопросы.

GRADE (GRADING OF RECOMMENDATIONS ASSESSMENT, DEVELOPMENT AND EVALUATION) [31–45]

Система классификации и оценки качества рекомендаций GRADE была создана международной группой экспертов из различных ведущих организаций по оценке медицинских технологий, включая NICE, AHRQ, NHMRC. Рабочая группа GRADE сформировалась в 2000 г. как неофициальное объединение исследователей, заинтересованных в устранении недостатков имеющихся систем оценки. Система GRADE была разработана для написания систематических обзоров и вынесения рекомендаций, касающихся альтернативных подходов к лечению (в том числе отсутствия лечения и современных стандартов лечения). Данная система оценки применима для решения широкого спектра клинических вопросов, включая диагностику, скрининг, профилактику и терапевтическое лечение, а также вопросов общественного здравоохранения.

К особенностям системы GRADE относится классификация исходов лечения по степени их значимости для пациентов с использованием специально разработанной для этого шкалы. В классификации исходов выделяются (в порядке убывания значимости) критически важный (**critical**), важный (**important**) и маловажный (**of limited importance**) исходы. Достоверность доказательств и УДД определяются отдельно для **каждого критически важного и важного исхода для совокупности исследований** (рис. 2). Обоснованием такого подхода (ориентированного на исход) служит тот факт, что достоверность доказательств как для отдельного исследования, так и для совокупности исследований зависит от того, какой исход оценивается. Например, если в серии РКИ без «ослепления» определяли степень выраженности болевого синдрома и общую смертность, то очевидно, что результаты первого из этих двух исходов были более подвержены влиянию систематической ошибки, связанной с отсутствием «ослепления», и имеют более низкую достоверность доказательств.

УДД по системе GRADE включают 4 категории: высокий, средний, низкий и очень низкий уровни. По умолчанию, доказательства, основанные на результатах РКИ, относятся к высокому уровню достоверности, а доказательства, основанные на результатах обсервационных исследований, к низкому УДД. Однако выделяют 5 факторов, которые могут перевести достоверность доказательств на более низкий уровень, и 3 фактора, которые могут поднять достоверность доказательств на более высокий уровень. К факторам, приводящим к понижению УДД, относятся:

1) риск возникновения систематических ошибок;

2) несогласованность результатов между исследованиями;

3) косвенность доказательств;

4) неточность определения размера эффекта;

5) публикационное смещение[4].

[3] В N-из-1 исследовании пациент проходит сопряженные друг с другом лечебные периоды: период экспериментальной терапии, за которой следует период стандартной терапии или плацебо. Периоды повторяются циклически.

[4] Публикационное смещение – систематическая ошибка, связанная с погрешностью отбора исследований, возникающая из-за предпочтения публиковать исследования с положительными (статистически значимыми) результатами.

К факторам, отвечающим за повышение УДД, относятся:

1) существенный размер эффекта;

2) дозозависимый эффект;

3) неучтенные вмешивающиеся факторы, исключение которых уменьшило бы размер найденного эффекта.

В зависимости от степени выраженности перечисленных факторов достоверность доказательств может быть повышена/понижена на один или два уровня. В качестве заключительного шага при оценке достоверности доказательств определяется итоговая достоверность доказательств и сводные УДД,

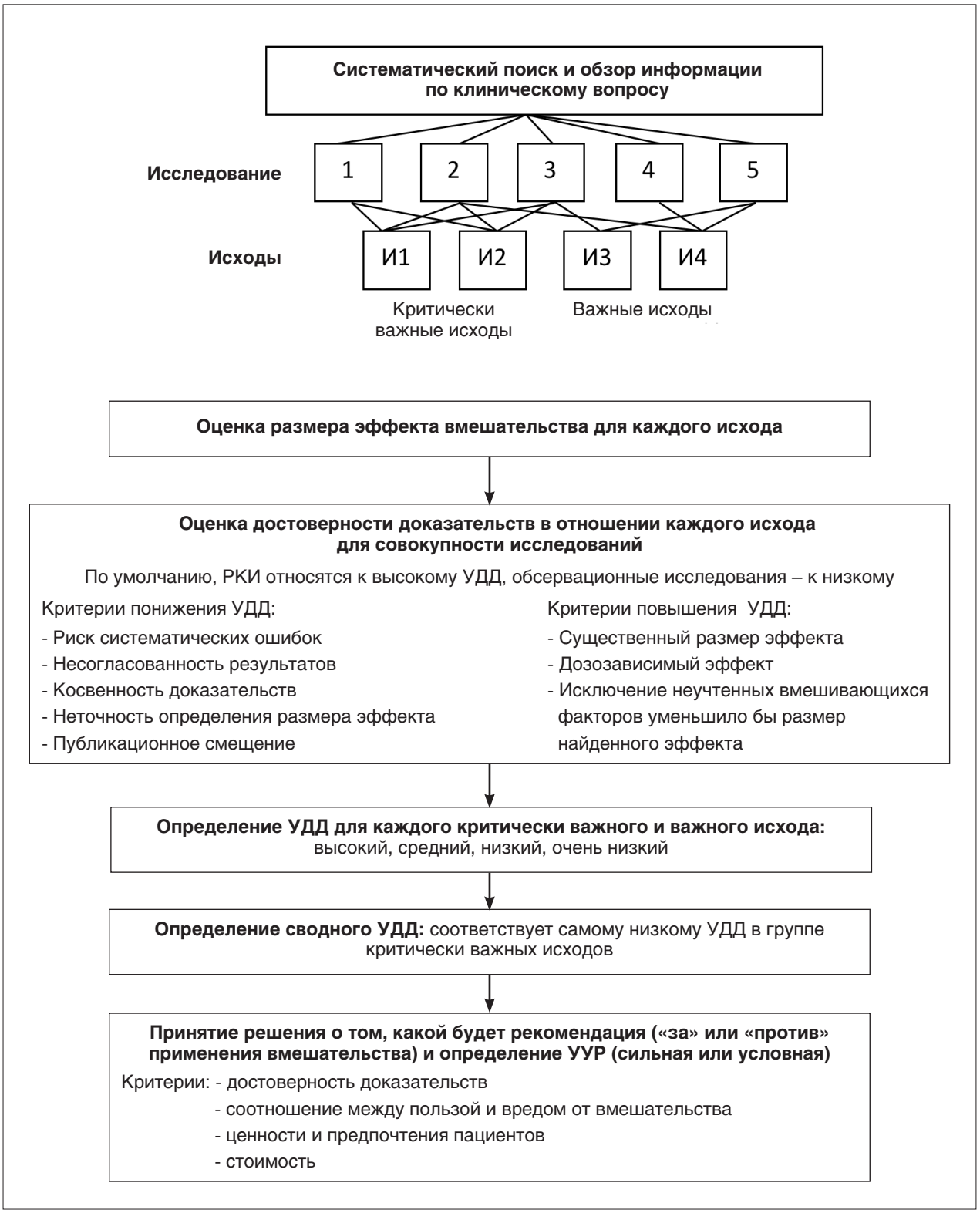

Рис. 2. **Схематическое изображение процесса оценки достоверности доказательств и убедительности рекомендаций по системе GRADE (адаптировано из [36]).**

которые соответствует самым низким УДД, присвоенным в группе крайне важных исходов. Следует обратить внимание на то, что система GRADE не рассматривает и не оценивает опубликованные систематические обзоры и метаанализы. Это объясняется тем, что даже методологически безупречные систематические обзоры могут опираться на данные не только качественно проведенных исследований с совпадающими результатами, но и слабых исследований с высоким риском систематических ошибок и несопоставимыми результатами, вследствие чего достоверность выводов, сделанных в таком обзоре, будет низкой, несмотря на его высокое методологическое качество. Авторы системы GRADE предоставляют четкие критерии и подробные указания, необходимые как для оценки каждого из факторов, влияющих на достоверность доказательств, так и для присвоения УДД. Для определения методологического качества исследований авторы используют критерии оценки риска возникновения систематических ошибок, совпадающие с критериями Кокрановского сотрудничества.

УУР в системе GRADE отражает степень уверенности в том, что благоприятные эффекты рассматриваемого вмешательства в целом превосходят его нежелательные последствия. Решение о том, какой будет рекомендация («за» или «против» вмешательства) и какой уровень убедительности ей будет присвоен, принимает экспертная группа, отличная от рабочей группы, проводившей систематический обзор и оценку качества доказательств. Шкала УУР включает только две категории: сильная рекомендация (**strong recommendation**) и рекомендация слабой силы/условная (**weak/conditional recommendation**). Экспертная группа присваивает категорию сильной рекомендации в случае полной уверенности в том, что ожидаемая польза от применения вмешательства превосходит его нежелательные последствия. Категория условной рекомендации присваивается в случаях меньшей уверенности экспертной группы в благоприятном соотношении между ожидаемыми преимуществами и недостатками вмешательства. При вынесении решения о степени убедительности рекомендаций экспертная группа рассматривает четыре ключевых фактора:

- достоверность доказательств;
- баланс между положительными эффектами и нежелательными явлениями;
- ценности и предпочтения пациентов;
- стоимость лечения.

Предполагается, что сильной рекомендации будут придерживаться практически все врачи и пациенты, обладающие полной информацией о вмешательстве; условной рекомендации будут придерживаться большинство врачей и пациентов, обладающих полной информацией о вмешательстве, но всё же значительная часть людей сделает альтернативный выбор. Условная рекомендация подразумевает также, что врачам следует более тщательно рассмотреть и учесть жизненные ценности и предпочтения пациентов, прежде чем рекомендовать им применение вмешательства.

К основным плюсам системы GRADE относятся прозрачность и четкая последовательность процесса оценки вмешательства, подробное описание критериев достоверности доказательств как для одного исхода, так и для совокупности исходов, требование полного обоснования всех принятых решений, учет клинической значимости рассматриваемых исходов. Сильной стороной системы GRADE является также применение символов и слов вместо букв или цифр для обозначения уровней достоверности доказательств и убедительности рекомендаций, что упрощает их восприятие. Однако для тех организаций, которые хотят сохранить буквенные и цифровые обозначения, авторы GRADE предлагают использовать буквы (от A до D) и цифры (1 и 2) для обозначения УДД и УУР, соответственно. Категоризация шкалы убедительности рекомендаций только на два уровня (сильная и условная рекомендации) тоже значительно облегчает восприятие рекомендаций и их применение в клинической практике.

Основной недостаток системы GRADE заключается в ее сложности и трудозатратности. В связи с этим было разработано бесплатное программное обеспечение GRADEpro с интуитивным интерфейсом, призванное упростить процесс оценки доказательств и составления рекомендаций.

NICE (NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE) [46]

Британский Национальный институт здоровья и клинического совершенствования (NICE) является независимой организацией, выпускающей клинические руководства, рекомендации по использованию медицинских технологий (результаты оценки медицинских технологий), рекомендации по применению интервенционных вмешательств, рекомендации в области здоровья населения (профилактика заболеваний и пропаганда здорового образа жизни). Принципы системы оценок, разработанных NICE, в последний раз были пересмотрены и опубликованы на сайте института в 2009 г.

В процессе оценки достоверности доказательств NICE использует элементы системы GRADE, в частности достоверность доказательств оценивается отдельно для каждого исхода, при этом учитываются все основные факторы, влияющие на достоверность доказательств и рассматриваемые в системе GRADE. Однако система NICE имеет несколько довольно значимых отличий от GRADE, а именно:

1) оценка методологического качества исследований производится с помощью вопросников, разработанных для каждого типа дизайна;

2) проводятся обзор и оценка методологического качества исследований «затраты-эффективность»;

3) не присваиваются сводные УДД и УУР;

4) для обозначения убедительности рекомендаций используются словесные формулировки.

Согласно требованиям NICE, при разработке рекомендаций принимаются во внимание следующие факторы: достоверность доказательств; клиническая значимость рассматриваемых исходов; баланс между положительными эффектами и нежелательными последствиями от вмешательства; соотношение между положительными эффектами и экономическими затратами; ряд дополнительных факторов (например, доступность вмешательства для пациентов вне зависимости от пола, национальности, расы, возраста, религии, инвалидности и т.п.). Для того чтобы процесс перехода от оценки достоверности доказательств к составлению рекомендаций был максимально прозрачным, рабочая группа NICE описывает каждый из указанных факторов с обоснованием вынесенных суждений в приложении к отчетному документу. Система NICE различает три степени уверенности в целесообразности применения вмешательства:

- вмешательство **необходимо либо недопустимо** использовать (must or must not be used);
- вмешательство **следует либо не следует** (should or should not) использовать;
- вмешательство **может быть** использовано (could be used).

Рекомендация, указывающая на необходимость вмешательства, подразумевает обязательное его применение во всех соответствующих случаях в целях соблюдения норм безопасности и правил, регулирующих оказание медицинской помощи. Рекомендация, согласно которой вмешательство следует использовать, означает уверенность рабочей группы NICE в том, что применение вмешательства принесет больше пользы, чем вреда, для подавляющего большинства пациентов и будет экономически выгодным. Если же в рекомендации говорится, что вмешательство может быть использовано, это означает уверенность экспертной группы в том, что применение вмешательства принесет больше пользы, чем вреда для большинства пациентов и будет экономически выгодным, однако имеются и альтернативные варианты лечения, которые тоже будут экономически выгодными, либо пациенты могут выбрать менее эффективное, но более дешевое лечение. В последнем случае выбор стратегии лечения будет зависеть от жизненных ценностей и предпочтений самих пациентов.

К сильным сторонам системы NICE относятся высокая воспроизводимость результатов оценки методологического качества исследований (за счет использования вопросников), учет клинической значимости рассматриваемых исходов, четкость и прозрачность критериев GRADE, используемых для оценки достоверности доказательств. Кроме того, авторы системы NICE совсем отказались от обозначения уровней убедительности рекомендаций, которая отражается только в словесных формулировках, что с одной стороны, облегчает использование рекомендаций на практике, но с другой стороны, может снижать эффективность их распространения между организациями и практикующими врачами.

NHMRC (NATIONAL HEALTH AND MEDICAL RESEARCH COUNCIL) [47–48]

Австралийский Национальный институт здравоохранения и медицинских исследований (NHMRC) несет ответственность за составление рекомендаций по трем направлениям: клиническая практика, этические аспекты медицинских вмешательств, здоровье населения. Последний вариант системы оценок NHMRC был разработан на основании трех систем оценки: SIGN, GRADE и SORT (Strength of Recommendation Taxonomy) − и опубликован в 2009 г.

Система NHMRC предлагает рассматривать 5 доменов, влияющих на достоверность совокупности доказательств: доказательная база, согласованность доказательств, клинический вклад, возможность экстраполяции результатов на целевую популяцию пациентов, применимость в условиях австралийской системы здравоохранения. Каждый домен оценивается по шкале A (очень хорошо), B (хорошо), C (удовлетворительно) и D (плохо) согласно четким критериям, оформленным в виде наглядной таблицы («матрицы»).

Первый домен «Доказательная база» учитывает дизайн и методологическое качество исследований, а также объем доказательств (число исследований и величину выборки пациентов). Сначала каждому исследованию присуждается соответствующий уровень в иерархии дизайнов исследований (структура иерархии дизайнов зависит от изучаемого клинического вопроса: терапевтическое вмешательство, диагностические методы, прогноз или этиология заболевания, методы скрининга). Например, при оценке терапевтических вмешательств на самом высоком уровне в иерархии дизайнов помещаются систематические обзоры РКИ, а на самом низком уровне − описания серии случаев. Далее каждое исследование проходит оценку методологического качества, при этом разработчики рекомендаций могут выбрать инструмент оценки на свое усмотрение. Авторы NHMRC дают ряд советов, какие вопросники лучше использовать при оценке методологического качества исследований в зависимости от дизайна и поставленного клинического вопроса (например, вопросник GATE рекомен-

Таблица 1. **Сравнительная характеристика систем оценки достоверности доказательств и убедительности рекомендаций**

Признак	SIGN	OCEBM	GRADE	NICE	NHMRC
Оценка достоверности доказательств					
Качественная характеристика	+	+ (нет четких критериев)	+	+	+
Количественная характеристика	+ (нет четких критериев)	+/– (нет четких критериев)	+	+	+
Согласованность	+ (нет четких критериев)	+ (нет четких критериев)	+	+	+
Итоговая оценка	1. УДД для отдельных исследований (8 уровней) 2. Сводные УДД (нет четких критериев)	1. УДД для отдельных исследований (5 уровней) 2. Нет сводного УДД	1. УДД для каждого исхода для совокупности исследований (4 уровня) 2. Сводные УДД для всех критически важных и важных исходов	1. УДД для каждого исхода для совокупности исследований (4 уровня) 2. Нет сводного УДД	Нет четкого понятия УДД; оценка совокупности доказательств по отдельным доменам: «Доказательная база», «Клинический вклад», «Согласованность доказательств»
Оценка убедительности рекомендаций					
Соотношение между пользой и вредом	+	–	+	+	+
Обобщаемость доказательств	+	–	+	+	+
Стоимость лечения	+	–	+	+	–
Ценности и предпочтения пациентов	–	–	+	–	–
Применимость рекомендаций в условиях национальной системы здравоохранения	+	–	–	–	+
Итоговая оценка	УУР (4 уровня: A, B, C, D)	–	УУР (2 уровня: сильная и условная рекомендации)	Нет УУР, но есть 3 степени уверенности в целесообразности использования вмешательства	УУР (4 уровня: A, B, C, D)

Примечание.
Качественная характеристика: *«+» – учитываются и дизайн, и методологическое качество исследований;*
«+/-» – учитывается только дизайн;
«-» – не учитывается.
Количественная характеристика: *«+» – учитывается хотя бы два из трех составляющих элементов (размер эффекта, количество исследований, суммарный размер выборки пациентов);*
«+/-» – учитывается только один из трех составляющих элементов;
«-» – не учитывают.
Для всех остальных критериев «+» обозначает, что критерий учитывается при вынесении оценки, «-» – не учитывается.

дуется для оценки исследований прогноза заболевания, вопросник SIGN или CASP − для оценки систематических обзоров и т.д.).

Домен «Клинический вклад» включает оценку точности определения эффекта, его выраженности и клинической значимости для пациентов (в том числе по сравнению с альтернативными стратегиями лечения), а также учитывает соотношение пользы и вреда от применения вмешательства. На взгляд авторов системы NHMRC, оценка именно этого домена является наиболее субъективной и требует активного обсуждения всеми участниками рабочей группы по составлению рекомендаций.

В системе NHMRC не существует УДД, а каждая рекомендация сопровождается оценками каждого из 5 указанных доменов. Путем сложения этих оценок определяют УУР (A, B, C или D); градации УУР отражают степень достоверности доказательств, лежащих в основе рекомендации. Так, рекомендации с уровнем убедительности A или B основаны на доказательствах, которым можно доверять в клинической практике; к рекомендациям с уровнем C или D нужно подходить с определенной долей осторожности и с учетом конкретных обстоятельств. Рекомендации могут получить уровень убедительности A или B только при условии, что два первых домена «Доказательная

база» и «Согласованность доказательств» также имеют оценки А или В.

К достоинствам системы NHMRC относится возможность наглядно оценить влияние каждого из 5 доменов на степень убедительности рекомендаций. Благодаря такому подходу различные работники системы здравоохранения могут в случае необходимости принимать собственные решения, придавая разный вес тому или иному домену в зависимости от конкретных задач, причем самых разнообразных – от составления рекомендаций по индивидуальной терапии пациента до принятия решения о возмещении расходов на лечение населению страны. К минусам данной системы нужно отнести отсутствие четкого разделения между процессами оценки достоверности доказательств и убедительности рекомендаций и отсутствие такого понятия как уровень достоверности доказательств.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СИСТЕМ ОЦЕНКИ ДОСТОВЕРНОСТИ ДОКАЗАТЕЛЬСТВ И УБЕДИТЕЛЬНОСТИ РЕКОМЕНДАЦИЙ

Проведенный сравнительный анализ наиболее известных систем оценки достоверности доказательств и убедительности рекомендаций показал, что, несмотря на ряд существенных различий, все системы опираются на одни и те же основополагающие критерии оценки (табл. 1). При оценке достоверности доказательств каждая из рассмотренных систем учитывает три основополагающих критерия: качественную характеристику (дизайн и методологическое качество исследований), количественную характеристику, или объем (размер эффекта, количество исследований, суммарный объем выборки пациентов) и согласованность доказательств. Основным различием в оценке достоверности доказательств является принципиально разный подход к присвоению УДД: в одних системах УДД присваивается каждому отдельному исследованию (OCEBM, SIGN), а **сводный УДД** – совокупности исследований (SIGN); в других системах УДД присваивается совокупности доказательств из всех исследований для каждого отдельного исхода лечения (GRADE, NICE), а **сводный УДД** определяется как наиболее низкий УДД среди всех критичных и важных исходов (GRADE). В австралийской системе NHMRC вместо присвоения УДД выносится ряд оценок для отдельных доменов для совокупности доказательств. При этом данная система оценки – единственная из рассмотренных, в которой нет четкого разделения между оценкой достоверности доказательств и принятием решения об убедительности рекомендаций.

Все рассмотренные системы оценки, кроме OCEBM, предназначены для составления клинических руководств, и их конечной целью является определение УУР (SIGN, GRADE, NHMRC) или степени уверенности в целесообразности применения медицинского вмешательства (NICE). Система оценки OCEBM, напротив, ограничивается оценкой достоверности доказательств и ставит перед собой цель помочь практикующим врачам и пациентам быстро и самостоятельно сориентироваться в существующих клинических исследованиях при поиске ответа на конкретный клинический вопрос.

Все рассмотренные системы (исключая OCEBM) при вынесении решения об убедительности рекомендаций опираются в первую очередь на оценку достоверности доказательств и на ряд одинаковых дополнительных критериев: соотношение между пользой и вредом от вмешательства, обобщаемость доказательств и стоимость лечения (последнее не относится к NHMRC). Авторы системы NHMRC не включили стоимость лечения в качестве критерия оценки убедительности рекомендаций, аргументируя это тем, что «готовность платить» у лиц, использующих рекомендации, будет сильно зависеть от конкретных обстоятельств [47]. Кроме одинаковых критериев оценки убедительности рекомендаций, авторы разных систем включают и некоторые дополнительные критерии, такие, как применимость рекомендаций в условиях национальной системы здравоохранения (SIGN, NHMRC), жизненные ценности и предпочтения пациентов (GRADE).

Следует отметить, что несмотря на сложность и многоэтапность процесса оценки, ни одна из рассмотренных систем не может полностью исключить необходимости в вынесении суждений (зачастую субъективных) экспертами рабочей группы. Для того чтобы обеспечить максимальную прозрачность процесса оценки, все рассмотренные системы акцентируют внимание на подробном обосновании любого принятого решения, касающегося достоверности доказательств или убедительности рекомендаций.

ПЕРСПЕКТИВЫ ВВЕДЕНИЯ ЕДИНОЙ СИСТЕМЫ ОЦЕНКИ ДОСТОВЕРНОСТИ ДОКАЗАТЕЛЬСТВ И УБЕДИТЕЛЬНОСТИ РЕКОМЕНДАЦИЙ ЗА РУБЕЖОМ

На данный момент в связи с постоянно нарастающим количеством различных клинических рекомендаций и увеличивающейся путаницей с их интерпретацией и применением на практике, назрела очевидная необходимость в унифицированном подходе к оценке достоверности доказательств и убедительности рекомендаций. Однако до сих пор исследователи и врачи не пришли к единому мнению о том, какой должна быть единая система оценки и нужна ли она. Наиболее важными причинами для введения единой системы оценки считаются:

1) уменьшение путаницы в понимании и использовании клинических руководств; 2) исключение возможности выбрать такую систему оценки, которая позволит присвоить наиболее высокий уровень достоверности доказательств и убедительности рекомендаций какому-то конкретному вмешательству [12]. Единственный существенный аргумент против введения единой системы оценки − это сомнения в том, что в рамках одной системы можно справиться с адекватной оценкой всего спектра рассматриваемых медицинских проблем, начиная с клинической эффективности вмешательств и заканчивая эффективностью функционирования системы здравоохранения. Предполагается, что если единая система оценки и будет создана, то, скорее всего, она окажется очень сложной [12].

Несмотря на существующие противоречия, все большее число международных организаций приходят к выводу о необходимости введения единой системы оценки. В настоящее время, пожалуй, самой распространенной и обсуждаемой становится система оценки GRADE, которая уже была адаптирована более чем 50 международными организациями, включая ВОЗ, Кокрановское сотрудничество, SIGN, NICE, AHRQ, Центры по контролю и профилактике заболеваний США (Centers for Disease Control and Prevention, CDC), многие европейские, американские и канадские профессиональные медицинские ассоциации (http://www.gradeworkinggroup.org/). Система GRADE в сравнении с другими системами оценки имеет ряд преимуществ. К ним относятся: классификация исходов лечения по значимости для пациентов; подробно разработанные критерии определения достоверности доказательств как для отдельных исходов, так и для совокупности исходов; прозрачность процесса вынесения суждений; простая классификация уровней достоверности доказательств и убедительности рекомендаций; использование формализованных таблиц для представления результатов оценки; доступность программного обеспечения GRADEpro.

В 2009 г. Cuello-Garcia с соавт. [15] обратились к международным экспертам по составлению клинических рекомендаций с предложением оценить 7 различных систем (GRADE, NICE, SIGN, OCEBM, SORT, CTFPH, USTFPS) по 4-балльной шкале Ликерта с точки зрения простоты использования, затрат рабочего времени и ресурсов, однозначности заключительных формулировок, полноты критериев определения достоверности доказательств (качественные и количественные характеристики, согласованность). По результатам опроса система GRADE наряду с системой NICE набрала наибольшее число баллов и была выбрана в качестве предпочтительной большинством экспертов [15]. В пользу системы GRADE говорят и результаты РКИ, в котором изучалось влияние клинических рекомендаций на принятие врачами решения о применении медицинского вмешательства на практике [16]. Данное РКИ было проведено на выборке педиатров из Мексики, которые были рандомизированы в четыре параллельные группы и должны были сделать выбор, касающийся применения рацекадотрила у детей с диареей, до и после ознакомления с клиническими рекомендациями. Врачам в разных группах давали для изучения клинические рекомендации, основанные на одной из четырех систем оценки (NICE, SIGN, GRADE, OCEBM). Результаты РКИ показали, что чаще всего отношение врачей к использованию рацекадотрила изменялось после прочтения клинических рекомендаций системы GRADE. Среди причин, по которым именно эти рекомендации оказали наибольшее влияние на мнение практикующих врачей, были выделены однозначность рекомендаций (сильная или условная рекомендации) и вызывающий доверие комплексный подход к процессу разработки рекомендаций.

Однако не стоит забывать, что несмотря на все возрастающую популярность и ряд преимуществ, система GRADE обладает и рядом недостатков, наиболее серьезные из которых − сложность в использовании и высокая трудозатратность; эти недостатки могут ограничить ее дальнейшее распространение.

Таким образом, в настоящий момент в мировом медицинском сообществе назрела необходимость в единой системе оценки достоверности доказательств и убедительности рекомендаций и даже наметилась тенденция к переходу на такую систему. Но, скорее всего, этот переход займет не один год, и станет ли этой системой именно система оценки GRADE, на данный момент остается неясным.

ПЕРСПЕКТИВЫ РАЗВИТИЯ СИСТЕМЫ ОЦЕНКИ ДОСТОВЕРНОСТИ ДОКАЗАТЕЛЬСТВ И УБЕДИТЕЛЬНОСТИ РЕКОМЕНДАЦИЙ В РОССИИ

В России система оценки достоверности доказательств наиболее активно используется при составлении Перечня жизненно необходимых и важнейших лекарственных препаратов (ПЖНВЛП), а также ряда отечественных клинических руководств. Согласно последней доступной в 2012 г. версии «Положения о порядке формирования проекта ПЖНВЛП», представленного Министерством здравоохранения и социального развития РФ, в процессе формирования перечня проводится клиническая экспертиза вынесенных на рассмотрение лекарственных препаратов. При проведении клинической экспертизы главными внештатными специалистами Министерства анализируется качество каждого отдельного клинического исследования лекарственного препарата, после чего

эффективность препарата оценивается по перечисленным ниже уровням убедительности доказательности (соответствуют уровням достоверности доказательств):

A – высокая достоверность, информация основана на результатах нескольких независимых клинических исследований с совпадением результатов, обобщенных в систематических обзорах;

B – умеренная достоверность, информация основана на результатах нескольких независимых рандомизированных и близких по целям клинических исследований;

C – ограниченная достоверность, информация об эффективности основана на результатах одного клинического исследования;

D – строгие научные доказательства отсутствуют, соответствующие клинические исследования не проводились, сведения об эффективности основаны на мнениях экспертов.

Критериями положительного заключения клинической экспертизы является присвоение уровня А или В. Очевидно, что такой подход к определению достоверности доказательств эффективности лекарственных препаратов при составлении ПЖНВЛП является сильно упрощенным по сравнению с рассмотренными в этом обзоре системами оценки, поскольку в процессе присвоения УДД недостаточно учитываются принятые в международной практике критерии достоверности доказательств, а именно качественная и количественная характеристики и согласованность доказательств.

Значительная часть составленных в России клинических руководств не опирается на систематический анализ научных доказательств [17]. Когда все же предпринимаются попытки использовать принципы доказательной медицины для обоснования рекомендаций, примененный подход часто не соответствует стандартам международной практики. Например, при составлении рекомендаций 2010 г. Российского респираторного общества «Внебольничная пневмония у взрослых: практические рекомендации по диагностике, лечению и профилактике» были использованы «категории доказательств» (табл. 2) [49]. «Категория доказательств» одновременно указывала как на достоверность доказательств, так и на убедительность рекомендаций. При оценке достоверности доказательств были учтены их качественные и количественные характеристики, однако не была учтена согласованность. При подготовке рекомендаций единственным критерием был уровень достоверности доказательств, т.е. ни один из дополнительных критериев (соотношение между пользой и вредом от вмешательства, обобщаемость доказательств, стоимость лечения, ценности и предпочтения пациентов) не рассматривался.

Таким образом, в отношении оценки достоверности доказательств и убедительности рекомендаций Россия отстает от развитых стран мира. Для внедрения таких инструментов доказательной медицины, как шкалы УДД и УУР, отечественные эксперты должны опираться на международный опыт общепризнанных институтов и ассоциаций по оценке медицинских технологий. При выборе системы оценки, которая будет адаптирована для использования в России, необходимо принимать во внимание уровень квалификации экспертов, временные затраты и затраты финансовых ресурсов, которые может себе позволить российская система здравоохранения. Необходим также анализ того, насколько выбранная система оценки будет соответствовать поставленным целям (будет ли это формирование перечней лекарственных препаратов или составление отечественных клинических руководств). Особого внимания при выборе системы оценки заслуживает система GRADE, поскольку на данный момент она является одной из наиболее полных, прозрачных и объективных систем оценки, а также главным кандидатом на место единой международной системы. Возможно, потребуется некоторая модификация системы в процессе ее адаптации к условиям, сложившимся в клинической практике и здравоохранении России. Однако авторы системы GRADE выступают против внесения в нее значительных изменений, поскольку иначе будет утерян смысл создания единой международной системы оценки.

Вне зависимости от того, какая система оценки получит распространение в России, важно осознавать, что целью эффективной и действенной системы оценки должно являться не полное исключение необходимости в использовании суждений, а в обеспечении прозрачности и логичности процесса оценки доказательств и вынесения рекомендаций. Ведь, несмотря на устоявшееся использование уровневого подхода в доказательной медицине, качество и достоверность доказательств, а также убедительность рекомендаций являются непрерывными категориями, и их разделение на уровни может несколько упрощать реальную ситуацию. С одной стороны, введение уровней способствует принятию оптимальных решений в здравоохранении, использованию и распространению клинических рекомендаций на практике. С другой стороны, по психологическим причинам наличие иерархии уровней в какой-то мере снимает с людей ответственность и освобождает от необходимости в размышлении и вынесении собственных суждений [29]. Уровни достоверности доказательств и убедительности рекомендаций ни при каких обстоятельствах, не должны быть использованы без тщательного обдумывания и анализа каждой конкретной ситуации [13, 14].

Таблица 2. **Категории доказательств и их интерпретация, приведенные в клиническом руководстве «Внебольничная пневмония у взрослых: практические рекомендации по диагностике, лечению и профилактике» [49]**

Категория доказательств	Источник доказательств	Определение
A	Рандомизированные контролируемые исследования	Доказательства основаны на хорошо спланированных рандомизированных исследованиях, проведенных на достаточном количестве пациентов, необходимом для получения достоверных результатов. Могут быть обоснованно рекомендованы для широкого применения
B	Рандомизированные контролируемые исследования	Доказательства основаны на рандомизированных контролируемых исследованиях, однако количество включенных пациентов недостаточно для достоверного статистического анализа. Рекомендации могут быть распространены на ограниченную популяцию
C	Нерандомизированные клинические исследования	Доказательства основаны на нерандомизированных клинических исследованиях или исследованиях, проведенных на ограниченном количестве пациентов
D	Мнение экспертов	Доказательства основаны на выработанном группой экспертов консенсусе по определенной проблеме

ЛИТЕРАТУРА

1. Hill A. B. The environment and disease: association or causation? Proc R Soc Med. 1965; 58: 295-300.
2. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment Number 47. Prepared for: Agency for Healthcare Research and Quality. AHRQ Publication No. 02-E016, April 2002, 199.
3. Canfield S. E., Dahm P. Rating the quality of evidence and the strength of recommendations using GRADE. World J Urol. 2011 Jun.; 29 (3): 311-317.
4. Guyatt G. H., et al. Going from evidence to recommendations. BMJ. 2008 May 10; 336 (7652): 1049-1051.
5. Canadian Task Force on the Periodic Health Examination: The periodic health examination. Can Med Assoc J. 1979; 121: 1193-1254.
6. Canadian Optimal Medication Prescribing and Utilization Service. http://www.cadth.ca/media/compus/pdf/COMPUS_Evaluation_Methodology_final_e.pdf.
7. SIGN guideline № 92, Management of hepatitis C. 2006 [www.sign.ac.uk/pdf/sign92.pdf]
8. Ghany M. G., et al. Diagnosis, management, and treatment of hepatitis C: an update. Hepatology. 2009 Apr; 49 (4): 1335-1374.
9. European Association for the Study of the Liver., EASL Clinical Practice Guidelines: management of hepatitis C virus infection. J Hepatol. 2011 Aug.; 55 (2): 245-264.
10. Atkins D., et al. Grading quality of evidence and strength of recommendations. BMJ. 2004 Jun 19; 328 (7454): 1490.
11. Juni P., et al. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999 Sep 15; 282 (11): 1054-1060.
12. Schünemann H. J., Fretheim A., Oxman A. D. Improving the use of research evidence in guideline development: 9. Grading evidence and recommendations. Health Res Policy Syst. 2006 Dec 5; 4: 21.
13. Howick J. The Philosophy of Evidence-Based Medicine. Oxford: Wiley-Blackwell; 2011. 248.
14. Haynes R. B., Devereaux P. J., Guyatt G. H. Clinical expertise in the era of evidence-based medicine and patient choice. ACP J Club. 2002 Mar-Apr; 136 (2): A11-14.
15. Cuello-García C. A., Dávalos-Rodríguez M. L. Perceptions and attitudes towards different grading systems from clinical guidelines developers. J Eval Clin Pract. 2009 Dec; 15 (6): 1074-1076.
16. Cuello-García C. A., Pacheco Alvarado K. P., Pérez Gaxiola G. Grading recommendations in clinical practice guidelines: randomised experimental evaluation of four different systems. Arch Dis Child. 2011 Aug; 96 (8): 723-728.
17. Власов В. В., Шварц Ю. Г. Проблемы составления и использования клинических рекомендаций и формуляров в России. Международный журнал медицинской практики, 2000; 11: 5-13.
18. Kearon C., et al. A comparison of three months of anticoagulation with extended anticoagulation for a first episode of idiopathic venous thromboembolism. N Engl J Med. 1999; 340: 901.
19. Campbell I. A., et al. Anticoagulation for three versus six months in patients with deep vein thrombosis or pulmonary embolism, or both: randomized trial. BMJ. 2007; 334: 674.
20. Kearon C., et al. Comparison of 1 month with 3 months of anticoagulation for a first episode of venous thromboembolism associated with a transient risk factor. J Thromb Haemost. 2004; 2: 743-749.
21. Agnelli G., et al. Three months versus one year of oral anticoagulant therapy for idiopathic deep venous thrombosis. N Engl J Med. 2001; 345: 165-169.
22. Balshem H., et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr; 64 (4): 401-406. Epub 2011 Jan 5.
23. Воробьев П. А., Аксюк З. Н. Стандартизация и оценка качества медицинской помощи. Проблемы стандартизации в здравоохранении. 1999; 1: 8-15.
24. Воробьев П. А. Протоколы ведения больных — подходы к созданию. Проблемы стандартизации в здравоохранении. 1999; 1: 49-55.
25. SIGN 50. A guideline developer's handbook. Scottish Intercollegiate Guidelines Network - Revised Edition, November 2011, 104 p.
26. Applying the GRADE methodology to SIGN guidelines: core principles. http://www.sign.ac.uk/pdf/gradeprinciples.pdf.
27. Dawes M. Putting evidence into practice. BMJ. 2011 Apr 11; 342: d2072.
28. OCEBM Levels of Evidence Working Group «The Oxford 2011 Levels of Evidence». Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653
29. Howick J., et al. The 2011 Oxford CEBM Levels of Evidence (Introductory Document). Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653
30. Howick J., et al., Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document). Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653
31. Canfield S.E., Dahm P. Rating the quality of evidence and the strength of recommendations using GRADE. World J Urol. 2011 29 (3): 311-317.
32. The GRADE working group. Grading quality of evidence and strength of recommendations. BMJ. 2004; 328: 1490-1494.
33. Guyatt G. H., et al. GRADE Working Group. Rating quality of evidence and strength of recommendations: Going from evidence to recommendations. BMJ. 2008 May 10; 336(7652): 1049-1051.
34. Guyatt G. H., et al. GRADE Working Group. Rating quality of evidence and strength of recommendations: What is "quality of evidence" and why is it important to clinicians? BMJ. 2008 May 3; 336 (7651): 995-998.
35. Guyatt G. H., et al., for the GRADE Working Group. Rating quality of evidence and strength of recommendations GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008; 336: 924-926.

36. Guyatt G. et al. GRADE guidelines: 1. Introduction - GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011 Apr; 64 (4): 383-94.
37. Guyatt G. H., et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011 Apr; 64 (4): 395-400.
38. Balshem H., et al. GRADE guidelines: 3. Rating the quality of evidence - introduction. J Clin Epidemiol. 2011 Apr;64(4):401-6.
39. Guyatt G. H., et al. GRADE guidelines: 4. Rating the quality of evidence - risk of bias. J Clin Epidemiol. 2011 Apr; 64 (4): 407-15.
40. Guyatt G. H., et al. GRADE guidelines: 5. Rating the quality of evidence - publication bias. J Clin Epidemiol. 2011 Dec; 64 (12): 1277-82.
41. Guyatt G. H., et al. GRADE guidelines: 6. Rating the quality of evidence - imprecision. J Clin Epidemiol. 2011 Dec; 64 (12): 1283-93.
42. Guyatt G. H., et al. The GRADE Working Group. GRADE guidelines: 7. Rating the quality of evidence - inconsistency. J Clin Epidemiol. 2011 Dec; 64 (12): 1294-302.
43. Guyatt G. H., et al. The GRADE Working Group. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011 Dec; 64 (12): 1311-6.
44. Guyatt G. H., et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol. 2012. Apr 27. [Epub ahead of print]
45. Guyatt G. H., et al. GRADE guidelines: 12. Preparing summary of findings tables - binary outcomes. J Clin Epidemiol. 2012. May 18. [Epub ahead of print]
46. NICE. The guidelines manual 2009. http://www.nice.org.uk/guidelinesmanual
47. NHMRC levels of evidence and grade of recommendations. 2009.
48. Hillier S., et al. FORM: an Australian method for formulating and grading recommendations in evidence-based clinical guidelines. BMC Med Res Methodol. 2011. Feb 28; 11: 23.
49. Australian Government. (2009). NHMRC levels of evidence and grades for recommendations for developers of guidelines. http://www.nhmrc.gov.au
50. Российское респираторное общество, Межрегиональная ассоциация по клинической микробиологии и антимикробной химиотерапии. «Внебольничная пневмония у взрослых: практические рекомендации по диагностике, лечению и профилактике», Москва, 2010. 107 с.

Сведения об авторах:

Андреева Наталья Сергеевна

старший научный сотрудник отдела доказательной медицины, биостатистики и математического моделирования АНО «Национальный центр по оценке технологий здравоохранения», канд. биол. наук

Реброва Ольга Юрьевна

профессор кафедры медицинской кибернетики и информатики ГБОУ ВПО РНИМУ им. Н.И. Пирогова, Москва, Россия, д-р мед. наук

Зорин Никита Александрович

преподаватель курса доказательной медицины отдела аспирантуры ФГБУ НЦЭСМП Минздрава России, канд. мед. наук, доцент

Авксентьева Мария Владимировна

руководитель отдела клинико-экономического анализа АНО «Национальный центр по оценке технологий здравоохранения», профессор кафедры общественного здравоохранения и профилактической медицины Первого Московского государственного медицинского университета им. И.М. Сеченова, д-р мед. наук

Омельяновский Виталий Владимирович

Председатель Совета АНО «Национальный центр по оценке технологий здравоохранения», д-р мед. наук, профессор

Адрес для переписки:

117335, г. Москва, а/я 88
Телефон: +7(495) 545-0927
E-mail: nat.andreyeva@gmail.com

RESEARCH. ANALYSIS. EXPERTISE

Evidence-Based Medicine

Systems for Assessing the Reliability of Scientific Evidence and the Soundness of Guidelines: Comparison and Prospects for Unification

N. S. Andreeva[1], O. Y. Rebrova[2], N. A. Zorin[3], M. V. Avxentyeva[1, 4], V. V. Omelyanovsky[1]

[1] National Center for Health Technology Assessement, 117335, Moscow, post–office box 88, Russia

[2] The Russian National Research Medical University named after N. I. Pirogov, 117997, Moscow, Ostrovityanova St., 1, Russia

[3] Scientific Center for Expertise of Medical Application Products, 127051, Moscow, Petrovskiy boulevard 8, Russia

[4] I. M. Sechenov First Moscow State Medical University, 119991, Moscow, Trubetskaya St. 8, bild. 2, Russia

The decision to implement medical technologies and include them in clinical guidelines should be based on an integrated analysis of all available scientific evidence of their effectiveness and safety. In this review we describe the systems for assessing the reliability of scientific evidence and the soundness of recommendations that are currently endorsed by well-known international agencies for health technology assessment and organizations responsible for the production of clinical guidelines (SIGN, OCEBM, GRADE, NICE, NHMRC). We have also performed a comparative analysis of the criteria used to assess the reliability of evidence (such as qualitative and quantitative variables and consistency of evidence) and the soundness of recommendations (such as generalizability of evidence, benefit-to-harm ratio of an intervention, cost of treatment, values and preferences of the patients, applicability of the recommendations under the conditions of the national healthcare system). Furthermore, we have analyzed the principles of classifying the evidence for the effectiveness of medical technologies according to its level of reliability and for the classification of clinical recommendations according to how sound they are. The final section of the review focuses on the prospects for implementing a unified system for the assessment of the reliability of evidence and soundness of recommendations in Russia and internationally.

KEYWORDS: levels of evidence, soundness of recommendations, GRADE, assessment systems, clinical guidelines.