

Разработка метода поиска клинических исследований по включающим факторам

А.Р. Донская^{1,2}, Н.Д. Сибирный¹, А.В. Зубков^{1,2}, Ю.А. Орлова^{1,2}, М.Ю. Фролов^{2,3}, А.С. Саласюк^{2,3}

¹Волгоградский государственный технический университет, Волгоград

²Волгоградский государственный медицинский университет, Волгоград

³Волгоградский медицинский научный центр, Волгоград

Аннотация: Результаты клинических исследований являются основным источником информации при осуществлении врачебной деятельности в соответствии с принципами доказательной медицины. На данный момент информационные системы, которые позволяли бы врачу подбирать клинические исследования в рамках нозологии, наиболее соответствующие профилю конкретного пациента, с целью дальнейшего анализа их результатов и подбора терапии, отсутствуют. Коллективом авторов был предложен метод поиска клинических исследований по критериям включения, что, в свою очередь, позволит значительно повысить эффективность и сократить время поиска и выбора терапии.

Ключевые слова: клинические исследования, алгоритмы критериального поиска, методы критериального поиска, включающие факторы, поиск ближайшего класса, сервисы.

Введение

Доказательная медицина для врача является одним из основополагающих источников информации, в связи с тем, что при назначении терапии и ведении пациента не существует на сегодня единого шаблона действий, который бы подходил для каждого пациента в отдельности [1]. Так же можно отметить, что ни один врач не обладает всей полнотой знаний в какой-либо области, в результате чего может возникнуть потребность разрешить, например, спорный вопрос относительно наиболее приемлемого варианта лечения. Кроме того, подобная персонифицированная медицина, которая представляет собой подбор лечения на основе индивидуальных особенностей пациента, представляется наиболее перспективным направлением медицины. Для решения подобных вопросов используют, в том числе, подробный анализ и субанализы клинических исследований лекарственных препаратов, участвующие в которых пациенты

наиболее полно соответствовали имеющейся клинической ситуации [2 – 4]. Однако, количество исследований непрерывно растет [5], а средства для качественного и быстрого поиска по всему массиву проведенных исследований на данный момент значительно ограничены либо языковым барьером, либо же недостаточностью функционала и критериев поиска для достаточной оптимизации работы специалиста. Сейчас качественные платформы поиска клинических исследований по критериям включения и с возможностью указания приоритетов на отечественном рынке отсутствуют. Как следствие, врачу приходится самостоятельно сравнивать характеристики пациентов в исследовании, с характеристиками своего пациента для выбора оптимальной терапии.

Целью работы является улучшение существующего процесса поиска клинических исследований, характеристики пациентов в которых максимально соответствуют таковым у конкретного пациента за счет использования метода приоритизации по задаваемым врачом критериям включения в ходе подбора.

Для выполнения поставленной цели необходимо реализовать следующие задачи:

- Изучить процесс подбора и поиска клинических исследований врачами. Выделить ключевые проблемы;
- Разработать метод поиска клинических исследований врачами и выделить необходимые критерии;
- Реализовать сервис поиска клинических исследований.

Постановка задачи

Клинические исследования — это исследования, организованные для оценки различных вмешательств, условия проведения которых направлены на устранение влияния систематических ошибок на результаты [6].

На данный момент, на практике используются методы поиска клинического исследования только по названию (включая ближайшие названия по расстоянию Левенштейна) или ключевым словам.

Каждое клиническое исследование имеет ряд критериев включения, определяющее то, какими характеристиками должны обладать пациенты для включения в данное исследование [6].

Разделить критерии можно на качественные и количественные. Причем все критерии необходимо привести к одному из данных типов. Например, в рамках исследования «SOLOIST-WHF» [7] количественными критериями выступают систолическое и диастолическое артериальное давление, индекс массы тела и другие критерии, которые можно определить при помощи физикального, лабораторного или инструментального обследования. Помимо этого, к количественным критериям относится возраст, длительность заболевания до включения в исследование и другие демографические характеристики пациентов.

К группе качественных критериев можно отнести:

— наличие у пациентов, включенных в исследование, подтвержденных сердечно-сосудистых заболеваний, хронической болезни почек, указаний на острое нарушение мозгового кровообращения в анамнезе и т.д., то есть, характеристики коморбидной патологии;

— факт приема пациентом терапии ключевыми группами лекарственных препаратов, например, статинами или ингибиторами ангиотензинпревращающего фермента;

— гендерная принадлежность, факт курения, и т.д.

Данные критерии носят статистический характер и отражены в виде частот встречаемости в популяции, поэтому логичнее всего данные значения инкапсулировать как качественные бинарные характеристики, с

соответственным отображением в интерфейсе для пользователя, по факту наличия или отсутствия качественного критерия.

Также в рамках качественных характеристик модель должна обрабатывать обратные случаи (например, если пациент мужчина). Все бинарные характеристики должны учитываться при выборе наиболее подходящего исследования по принципу наибольшей встречаемости соответствующего значения качественного критерия в популяции исследования.

Все качественные характеристики для валидации их в интерфейсе должны иметь правила, по которым можно было бы проверить правильность введённых значений. Для этого планируется ввести константные пределы для каждой из типов характеристик.

Описание хранимых данных о клиническом исследовании

Процесс подготовки базы клинических исследований на данный момент носит ручной характер, так как решение об отношении критерия к типу качественных или количественных должен производить медицинский специалист.

Формат данных представляет собой набор табличных файлов, каждый из которой относится к одному из рубрикаторов МКБ. Пример такого файла представлен на Рис. 1.

Столбец 1 отвечает за названия вводимых характеристик, каждые последующие столбцы отвечают за величину каждой из введённых характеристик в рамках данной области.

На выходе модель должна выдавать отсортированные по сходству найденные клинические исследования и их результаты.

Столбец1	EMPEROR Reduced				EMPHASIS		Столбец2
	Дара-HF		SOLOIST-WHF	PARADIGM			
РКИ	Дара-HF		SOLOIST-WHF		EMPHASIS	EMPEROR-Preserved	
Возраст пациентов, годы		66,2 67,2 ± 10,8		69 63,8 ± 11,5	68,7±7,7	71,8	
САД		122 122,6		122	124	131,8	
ДАД		73,5		72	75		
ИМТ, кг/м2		28,2 28		30,4	28,1 27,5	29,77	
HbA1c, %	6,50			7,1			
Длительность СД2, годы	-						
рСКФ		66 61,8		49,2	71,2	60,6	
% пациентов с подтвержденными ССЗ	-						
% пациентов со сниж. почечной ф-й (СКФ <60 мл/мин)		40,6 48		69,9	32,2	50,2	
% пациентов с альбинурией > 30 мг/г	-						
% пациентов, получающих инсулин		27,6		35,7			
% пациентов, получающих статины		67,1		-	62,8	68,1	
% женщины		23,8 23,5		32,6	21 22,7	44,6	
Микроваскулярные заболевания							
Ретинопатия	-						
Нефропатия	-						
Нейропатия	-						
Макроваскулярные заболевания							
ИБС		55,5 52,8			59,9 69,7		
ИМ в анамнезе		44,5			43,4 50,3		
ОИМК в анамнезе	-				8,5 10		
ПАД	-						
АГ		74,8 72,4			70,9 66,7	90,8	
СД2 типа		42 49,8		100	34,7 33,7	48,9	
Дислипидемия	-						
ХСН		100 100		100	100 100	100	
HFrEF (≤40%)		100 100		100	100		
HFrEF (>40%)		0 0		0	0	100	
ФВЛЖ		31,1 27,7		35	29,6 26,2	54,3	
Ампутация н/к	-					0,5	
ФП		38,6 35,6		-	36,2 30	51,5	

Рис. 1. – Пример табличного файла для одной из нозологий

Для более удобного использования в требованиях для отображения были выбраны наиболее важные характеристики клинического исследования:

- Название клинического исследования;
- Конечные точки (например, в исследованиях в области сахарного

диабета 2 типа таковыми являются:

- MACE (тяжелое сердечно-сосудистое нежелательное явление);
- CV death (смерть в связи с сердечно-сосудистыми заболеваниями);
- MI (инфаркт миокарда);
- Stroke (инсульт);
- HF (Госпитализация по поводу сердечной недостаточности или сочетание госпитализаций по поводу сердечной недостаточности или смерти от сердечно-сосудистых заболеваний);
- All death (смерть от любых причин);
- Kidney endpoints (почечная конечная точка, включающая в себя терминальную почечную недостаточность, значимое снижение скорости

клубочковой фильтрации или смерть в связи с прогрессированием хронической болезни почек).

Метод поиска клинических исследований по включающим факторам клинических исследований

Изучив существующие модели и методы (модели машинного обучения для задач классификации [8-10], метод поиска на базе крайних значений [8, 11]), можно сделать вывод, что сегодня данная задача является открытой и относится к ряду задач поисковых оптимизаций [12, 13].

Поэтому в рамках данной статьи представлен метод для поиска ближайшего класса клинических исследований к входному объекту, состоящий из следующих элементов:

- пользовательская оценка критериев важности каждой характеристики входного объекта;
- расчет статистической близости коэффициента;
- учет множества количественных и качественных характеристик.

В данном случае вход и выход можно формализовать следующим образом: на вход подается объект с набором характеристик, а также пользовательская оценка важности каждой характеристики объекта. Также в качестве индексируемой базы данных используется СУБД с размеченными клиническими исследованиями. На выходе модель выдает найденные клинические исследования. Для удобства оценки характеристики была использована пятибалльная шкала оценки из целочисленных значений.

Формула (1) демонстрирует полную модель, которая состоит из трех основных элементов (слагаемых), которые отвечают за разные типы характеристик. Первое и второе слагаемое отвечает за работу с качественными характеристиками, а третье отвечает за работу с количественными характеристиками.

$$P(A) = \sum_{i=1}^a 0.2 * prior_i * \frac{value_i * relProbability_i}{100} + \sum_{i=1}^a 0.2 * prior_i * \frac{value_i * (100 - relProbability_i)}{100} + \sum_{j=1}^b 0.2 * prior_j * \left(1 - \left| \frac{relVal_j - value_j}{relVal_j} \right| \right) \quad (1),$$

где:

- a - число качественных характеристик;
- $prior_i$ - важность i -ой качественных характеристики;
- $value_i$ - значение i -ой качественной характеристики (0...1), указанное врачом;
- $relProbability_i$ - значение i -ой вероятности, указанной в базе данных в значении класса;
- b - число числовых характеристик;
- $prior_j$ - важность j -ой числовой характеристики (цел. число от 1 до 5);
- $relVal_j$ - среднее значение j -ой относительной характеристики;
- $value_j$ - значение j -ой относительной характеристики, указанное врачом.

Первым слагаемым просчитывается близость только по прямым качественным характеристикам и данное значение лежит в промежутке от 0 до 1. Переменная *value* выступает в рамках данного множества неким управляющим значением, так как оно бинарное и может быть или 0, или 1. При нуле все слагаемое становится равно нулю. Например, в исследовании А было 30% женщин, запрашиваемый объект В имеет данную характеристику равную 1 с приоритетом 4, тогда значение данного слагаемого равно 0,24 (но если бы это был мужчина, то данное слагаемое было бы равно 0, а второе равно 0,56, далее объясняется, почему).

Второе слагаемое является обратной версией первого. Параметр *value* стоит с оператором “Инверсия”, а вероятность из базы данных берется обратная (т.е. если в исследовании принимали участие 60% людей с

подтверждёнными сердечно-сосудистыми заболеваниями, то обратная вероятность ей будет $100 - 60 = 40\%$ процентов людей без подтвержденных сердечно-сосудистых заболеваний).

В формуле (2) представлено третье слагаемое, используемое для подсчета близости, численного значения *value* характеристики-*j* объекта к значению *relVal*. Характеристики *j* класса *i*. Значение данной операции нормировано в промежутке от 0 до 1.

$$1 - \left| \frac{relVal_j - value_j}{relVal_j} \right| \quad (2)$$

На Рис. 2 представлен алгоритм вычисления ближайших классов клинических исследований с использованием формулы (1), описанной выше.



Рис. 2. – Алгоритм вычисления ближайших классов клинических исследований с использованием формулы (1)

Первым шагом алгоритма является инициализация пустого хранилища (КЭШа) пар значений класса и вычисленной оценки близости этого класса к объекту A . Далее осуществляется проход по каждому классу клинических исследований, размеченных в базе данных, где вычисляется значение $P(A)$ и добавляется в кэш. После окончания цикла происходит сортировка элементов кэша по критерию близости, рассчитанной в рамках заложенной итерации. На выходе пользователь получает N -количество классов, которые наиболее близки к данному входному объекту.

Стоит отметить, что при таком подходе при наращивании размеченной базы данных будет увеличиваться точность самого поиска.

Все вышеописанное является методом поиска клинических исследований по включающим факторам, основными шагами которого можно обозначить:

1. Обработка клинических исследований и выделение двух типов критериев;
2. Преобразование выделенных критериев к виду реляционной базы данных;
3. Обращение к алгоритму, описанному выше, в результате работы которого пользователь наблюдает результат, а именно набор ближайших клинических исследований.

Заключение

В результате работы коллективом авторов был проведен анализ основных критериев включения для клинических исследований, на базе которого были выделены основные характеристики пациентов, для чего в дальнейшем была разработана модель. Был предложен алгоритм вычисления ближайших классов клинических исследований на базе разработанной формулы. Был создан инструмент для удобного поиска и навигации по входным объектам в клинических исследованиях.



Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-07-00502, при финансовой поддержке ВолгГТУ в рамках научного проекта №60/478-22, №60/473-22 в рамках программы приоритета 2030 ВолгГТУ.

Литература

1. Основы доказательной медицины. Учебное пособие для системы послевузовского и дополнительного профессионального образования врачей. М.: Силицея-Полиграф. 2010. 136 с.
 2. Шульман В. А. Доказательная медицина. Насколько она доказательна? // Сибирское медицинское обозрение. 2007. №2. С. 92-96.
 3. Бабанов С.А. Испытания лекарственных средств с позиции доказательной медицины: проблемы и решения // Проблемы стандартизации в здравоохранении. 2007. №2. 15 с.
 4. Абаев Ю.К. История развития доказательной медицины // Медицинские новости. 2007. №6. С. 7-14.
 5. Ассоциация организаций по клиническим исследованиям Информационно-аналитический бюллетень № 24 URL: acto-russia.org/files/bulletin_24.pdf (дата обращения: 10.05.2022).
 6. Мелиов О.Г. Клинические исследования. 3 - е изд., доп. М.: Издательство «Атмосфера». 2013. 200 с.
 7. Nidhra, S. Black Box and White Box Testing Techniques - A Literature Review // International Journal of Embedded Systems and Applications. 2012. №2. P.29-50. URL: [10.5121/ijesa.2012.2204](https://doi.org/10.5121/ijesa.2012.2204).
 8. Docherty K., McMurray J. SOLOIST-WHF: Sodium-glucose cotransporter 2 (SGLT2) inhibitors should be initiated in patients hospitalised with worsening heart failure // European Journal of Heart Failure. 2020. №23. URL: [10.1002/ejhf.2075](https://doi.org/10.1002/ejhf.2075).
-



9. Микони С.В. О классе, классификации и систематизации // Онтология проектирования. 2016. №1 (19). С. 67-80.
10. Стебаков И.Н., Шутин Д.В., Марахин Н.А. Машинное обучение в реабилитационной медицине и пример классификатора движений пальцев для кистевого тренажера // Инженерный вестник Дона, 2020, №6. URL: ivdon.ru/ru/magazine/archive/N6y2020/6514.
11. Воронин В.В., Сизякин Р.А., Гапон Н.В., Франц В.А., Колосов А.Ю. Алгоритм реконструкции изображений на основе анализа локальных бинарных окрестностей // Инженерный вестник Дона, 2013, №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1857.
12. Bishop C. M. Pattern recognition. // Machine Learning. 2006. №128. pp. 1-58.
13. Salvatore F. In Search of HR Intelligence: Evidence-Based HR Analytics Practices in High Performing Companies // People + Strategy. 2014. p.12.

References

1. Osnovy dokazatel'noy meditsiny. Uchebnoye posobiye dlya sistemy poslevuzovskogo i dopolnitel'nogo professional'nogo obrazovaniya vrachey. M.: Silitseya-Poligraf [Fundamentals of evidence-based medicine. Textbook for the system of postgraduate and additional professional education of doctors]. 2010. p. 136.
2. Shul'man V. A. Dokazatel'naya meditsina. Naskol'ko ona dokazatel'na? Sibirskoye meditsinskoye obozreniye. 2007. №2. pp. 92-96.
3. Babanov S.A. Problemy standartizatsii v zdravookhranении. 2007. №2. p. 15.
4. Abayev Yu.K. Meditsinskiye novosti. 2007. №6. pp. 7-14.
5. Assotsiatsiya organizatsiy po klinicheskim issledovaniyam Informatcionno-analiticheskiy byulleten' № 24 URL: acto-russia.org/files/bulletin_24.pdf.



6. Meliov O.G. Klinicheskiye issledovaniya [Clinical studies]. 3-e izd., dop. M.: Izdatel'stvo «Atmosfera». 2013. p. 200.
7. Nidhra, S. International Journal of Embedded Systems and Applications. 2012. №2. pp.29-50. URL: 10.5121/ijesa.2012.2204.
8. Docherty K., McMurray J. European Journal of Heart Failure. 2020. №23. URL: 10.1002/ejhf.2075.
9. Mikoni S.V. Ontologiya proyektirovaniya. 2016. №1 (19). pp. 67-80.
10. Stebakov I.N., Shutin D.V., Marakhin N.A. Inzhenernyj vestnik Dona, 2020, №6. URL: ivdon.ru/ru/magazine/archive/N6y2020/6514.
11. Voronin V.V., Sizyakin R.A., Gapon N.V., Frants V.A., Kolosov A.YU. Inzhenernyj vestnik Dona, 2013, №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1857.
12. Bishop C. M. Machine Learning. 2006. №128. pp. 1-58.
13. Salvatore F. People + Strategy. 2014. p.12.