УДК 519.255

МЕТОДЫ ВИЗУАЛИЗАЦИИ ДАННЫХ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

С. Ю. Огородникова [1], Е. Д. Константинова [2]

[1] *Уральский Федеральный университет им. первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия*
[2] *Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия*

В статье представлены результаты применения методов визуализации многомерных данных к решению прикладной задачи из области медицины. Использованы ретроспективные данные 586 пациентов с раком мочевого пузыря, перенесших радикальную цистэктомию с декабря 2001 г. по май 2021 г. в онкоурологическом отделении Свердловского областного онкологического диспансера (СООД). В качестве методов визуализации многомерных данных реализованы такие подходы, как кривые выживаемости (по методу Каплана – Мейера), тепловые карты и деревья классификации. Для анализа данных использовались пакеты Statistica for Windows v.10. и MATLAB R2021b. На базе трех подходов показана возможность выявления отношений между численными данными на основе визуального анализа полученной совокупности графических образов. Продемонстрированы примеры построения наглядных, интуитивно понятных решающих правил для каждого из методов при решении задачи оценки 5- и 10-летней выживаемости пациентов Свердловского областного онкологического диспансера, перенесших радикальную цистэктомию тремя различными способами уродеривации. Рассмотрены особенности, преимущества и недостатки каждого из подходов применительно к работе с данными в медико-биологических исследованиях. Показано, что методы визуализации многомерных данных и подходы на их основе перспективны для анализа и наглядного представления экспериментальных численных результатов из области медицины.

Ключевые слова: медико-биологические исследования; визуализация данных; тепловые карты; кривые выживаемости, деревья классификации.

1. Введение

Примерно с конца XVII столетия, с того момента, когда Эдмунд Галлей вывел формулы для расчета страховых платежей, статистика имеет дело с алгоритмами, помогающими в принятии решений [1]. За такой длительный срок масштабы собираемых данных претерпели существенные изменения – появились так называемые большие данные. Причем «большими» данные могут быть как по числу случаев (кейсов, строк), так и по количеству переменных (измеряемых параметров, характеристик, столбцов) в базе данных. Появление «больших данных» в медицине послужило стимулом для разработки алгоритмов и систем поддержки принятия ре-

шений в различных областях врачебной практики, включая такие ответственные, как оценка исходов оперативного вмешательства в онкологии.

Анализ медико-биологических данных (МБД) имеет ряд особенностей, которые переводят его из разряда рутинных задач в творческие. Выделим основные:

– разнообразие типов данных: количественные (непрерывные и дискретные), категориальные (номинальные, могут быть в бинарной форме), ранговые;

–загрязненность (или зашумленность) реальных МБД, часто требующих предварительной обработки;

– достаточно высокий уровень коррелированности реальных МБД;

– невозможность получения 100 %-й точности результатов моделирования, поскольку мы имеем дело с «паутиной причинности» возникновения патологии. Ни один фактор риска возникновения болезни сам по себе не является непосредственной причиной заболевания, он может лишь увеличить или изменить вероятность его появления;

– нежелательность применения моделей типа «черного ящика» с непрозрачной работой алгоритма, поскольку необходима контролируемость алгоритмов, потенциально влияющих на жизнь людей;

– желательность получения интуитивно понятных решающих правил модели для использования специалистами – медиками и биологами.

Две последние (по порядку, но не по значимости) особенности диктуют необходимость широкого внедрения в статистический анализ МБД методов *визуализации*.

Визуализацию данных можно определить как системное, основанное на правилах графическое представление информации, помогающее разобраться в сложных понятиях, нацеленное на обобщение, синтез теории и опыта [2].

Одна из целей визуализации – выделение закономерностей или аномалий в численных данных, а также первичная оценка набора данных для возможности применения в дальнейшем более сложных инструментов анализа. Применяя методы визуализации данных, исследователь нацелен на поиск наиболее выразительных изображений изучаемых объектов и связей между ними.

Хотелось бы особо подчеркнуть, что для решения одной прикладной задачи зачастую не существует единого подходящего метода визуализации. Это можно считать частным случаем известного принципа множественности моделей В. В. Налимова [3], который постулирует следующее: для объяснения и предсказания структуры и (или) поведения сложной системы возможно построение нескольких моделей, имеющих одинаковое право на существование. Применение комбинации нескольких методов для обеспечения высокого качества результатов является в данном случае необходимостью.

Цель работы – визуализировать численные данные из области медицины и показать возможность выявления отношений между ними на основе зрительного анализа полученной совокупности графических изображений (образов).

2. Материалы и методы

Клиническая часть работы по сбору данных выполнена в онкоурологическом отделении СООД. Были ретроспективно проанализированы данные 586 пациентов с раком мочевого пузыря, перенесших радикальную цистэктомию (РЦ) с декабря 2001 г. по май 2021 г. База данных включала в себя такие характеристики пациен-

тов, как пол, возраст, наличие хронических заболеваний, сведения о проведенной операции, наличие и тяжесть послеоперационных осложнений, а также данные о нескольких видах выживаемости.

В исследование были включены 523 мужчины (89,3 %) и 63 женщины (10,7 %) возрастом от 28 до 81 г. Средний возраст пациентов составил 59,8 ± 8,6 лет.

Пациенты были разделены на 3 группы в зависимости от способа уродеривации, осуществленного в ходе операции:

– пациенты с наружным отведением мочи (группа 1, $n = 82$ человека);

– пациенты с отведением мочи в изолированный сегмент подвздошной кишки (группа 2, $n = 373$ человека);

– пациенты с ортоптическими кишечными резервуарами (группа 3, $n = 131$ человек).

В статье рассмотрены и реализованы такие подходы визуализации многомерных данных, как кривые выживаемости, тепловые карты и деревья классификации.

Представленные в работе методы были выполнены в пакетах Statistica for Windows 10 version и MatLAB2021b.

На протокол исследования было получено одобрение Локального этического комитета ИПЭ УрО РАН.

3. Результаты

Рассмотрим результаты решения прикладной задачи по оценке выживаемости пациентов СООД, перенесших РЦ тремя различными способами уродеривации.

3.1. Кривые выживаемости

Когда говорят об исследованиях в области визуализации информации, первым обычно вспоминают исследование «элементарных перцептивных задач» Уильяма Кливленда и Роберта МакГилла, которые еще в 1984 г. доказали, что сравнение объектов в одной шкале, например по оси, является самым простым визуальным действием [4].

Рассмотрим реализацию подхода, предложенного Капланом и Мейером в 1958 г. [5–7], позволяющего отображать многомерные объекты в виде кривых линий, достаточно простых и понятных для восприятия (рис. 1).

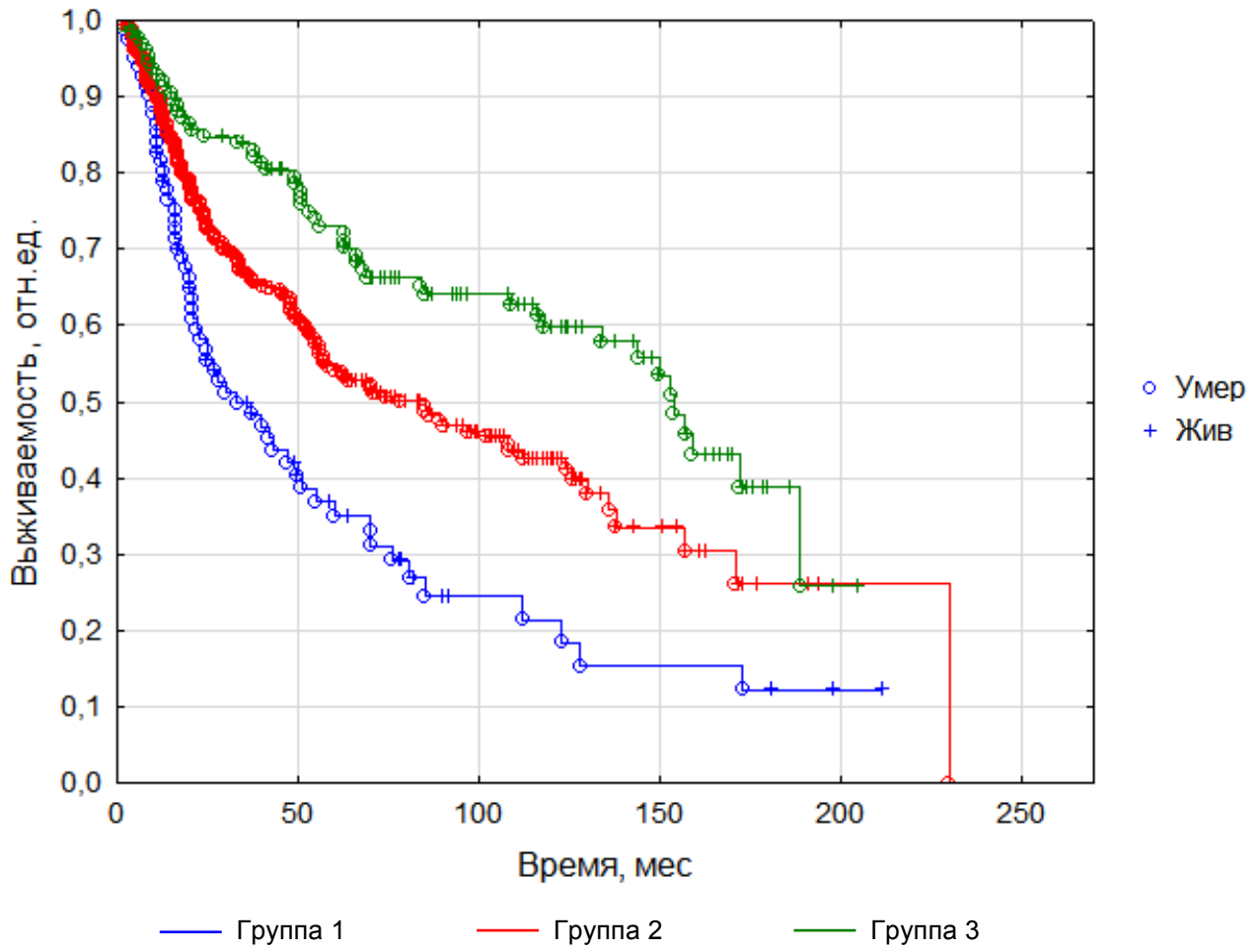


Рис. 1. Кривые общей выживаемости пациентов СООД, перенесших РЦ тремя способами уродеривации

Как следует читать график кривых выживаемости? По оси абсцисс отмечается время в месяцах. Время наблюдения для живых пациентов определяют как время от момента постановки диагноза до даты последнего наблюдения за пациентом, а для умерших пациентов – как время от момента постановки диагноза до даты смерти пациента. По оси ординат – выживаемость в относительных единицах, где «1,0 = 100 % наблюдаемых пациентов живы», а «0,0 = 0 % все наблюдаемые пациенты умерли». Точка на кривой, отмеченная знаком «+» и «○», соответствуют пациенту, который на момент анализа данных был жив или мертв.

Визуальный анализ графика (рис. 1) позволяет сделать следующий вывод: выживаемость у пациентов трех групп различается (кривые достаточно удалены друг от друга). Лучшая выживаемость прослеживается у третьей группы больных, а худшая – у пациентов первой группы, по методу уродеривации.

Существует такой расчетный показатель, как медиана выживаемости (Me) – значение времени, до которого выживаемость для выборки не опускалась ниже 0,5 (50 %). Стоит отдельно подчеркнуть, что речь идет именно о выживаемости, а не о доле выживших пациентов. Основное различие данных терминов заключается в том, что выживаемость представляет собой вероятность пережить определенный промежуток времени и учитывает выбывших пациентов. Доля выживших пациентов на момент времени *t* в свою очередь является отношением переживших момент *t* к объему выборки. В условиях отсутствия выбывших пациентов данные значения равны, однако чем больше будет выбывших, тем больше будет расхождение между этими значениями. Мы бы хотели сделать акцент на том, что в условиях реальных

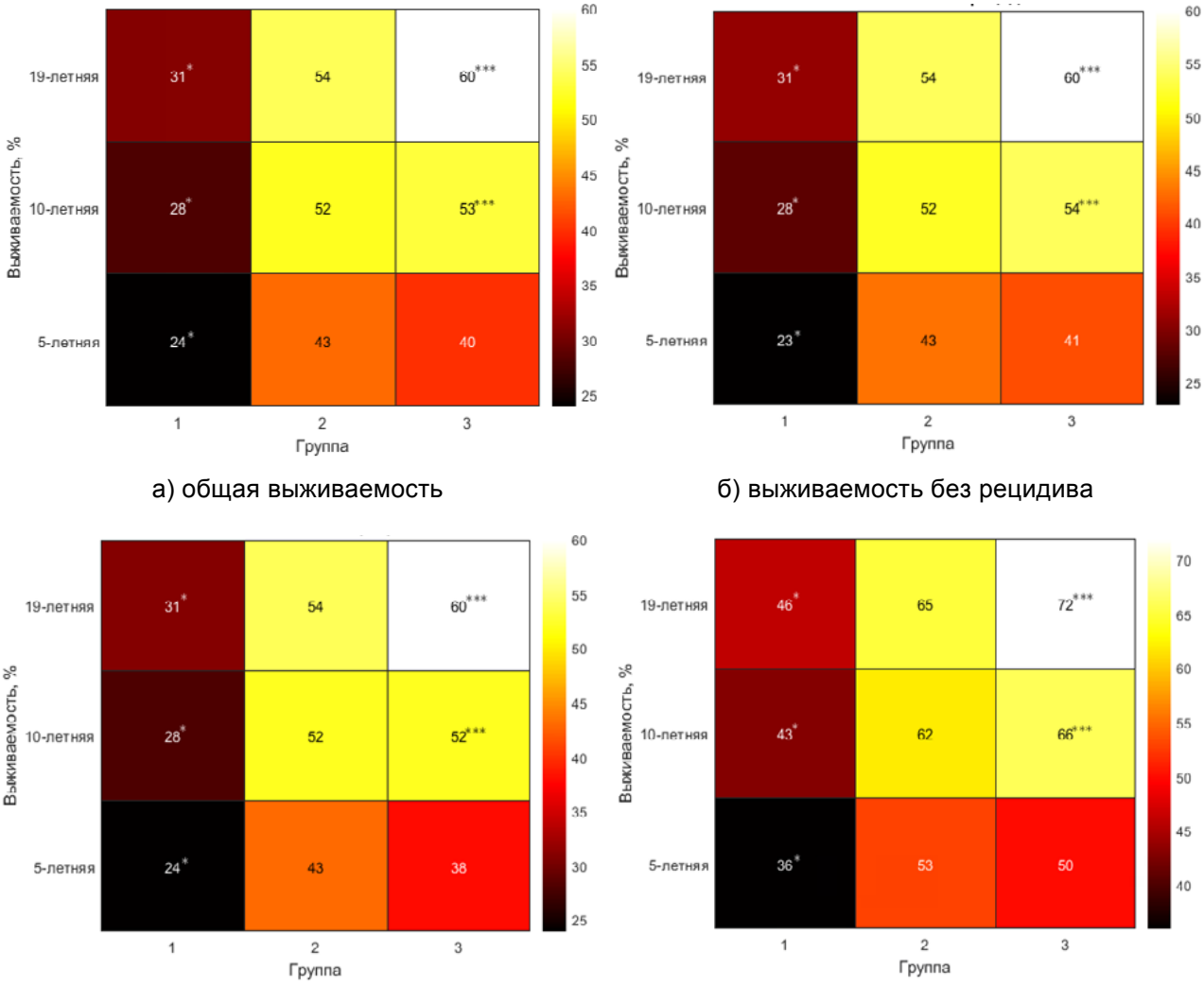
медицинских исследований сложно представить ситуацию отсутствия выбывших из исследования (по разным причинам) пациентов.

Медианы общей выживаемости были рассчитаны для каждой группы: Me в первой группе составила 37 месяцев, во второй – 78 месяцев, в третьей группе данный показатель оказался максимальным – 154 месяца. Рассчитанные показатели медиан выживаемости подтвердили результаты визуального анализа кривых – лучшую выживаемость демонстрируют пациенты третьей группы; худшую – пациенты первой группы, по методу уродеривации. При этом визуальный анализ интуитивно понятен, не требует специальной математической подготовки и проведения дополнительных расчетов.

3.2. Тепловые карты

Карты – один из древнейших способов визуализации, отображающий окружающую реальность. В различных областях научных исследований достаточно популярны тепловые карты – метод визуализации многомерных данных, разработанный и впервые примененный более 100 лет назад [8, 9].

Представленный подход реализован в пакете MatLAB2021b. На рис. 2 представлены тепловые карты различных видов выживаемости для трех групп пациентов СООД, перенесших РЦ.



а) общая выживаемость

б) выживаемость без рецидива

в) выживаемость без прогрессии

г) раковоспецифичная выживаемость

Рис. 2. Тепловые карты 5 и 10-летней выживаемости пациентов СООД

Интерпретацию тепловых карт проводят, опираясь на цветовую шкалу, расположенную справа от основного рисунка. В нашем исследовании использовалась наиболее удобная для восприятия последовательная цветовая шкала оттенков: от черного цвета (наименьшее значение выживаемости) к белому (наибольший процент выживших пациентов).

Рассмотрим в качестве примера общую 10-летнюю выживаемость пациентов 2 группы: нужная нам ячейка расположена на пересечении строки «10-летняя выживаемость» и столбца «2 группа». Значение выживаемости составляет 52 %; желтый цвет ячейки по последовательной цветовой шкале оттенков близок к белому, что соответствует лучшим показателям выживаемости. Соседняя ячейка (10-летняя выживаемость пациентов 1 группы) окрашена в более темный оттенок, что говорит о худших (относительно 2 группы) показателях выживаемости (28 %).

Анализируя полученные карты (рис. 2), можем говорить об одинаковом характере поведения всех рассмотренных видов выживаемости для трех групп пациентов: наименьшая выживаемость у пациентов первой группы на всех временных промежутках (левый столбец), для пациентов второй и третьей групп показатели выживаемости практически не различаются, наибольшее значение (60 и 72 месяца) соответствует выживаемости пациентов третьей группы за все время исследования (19 лет).

3.3. Деревья классификации

Одна из ответственных областей применения статистической обработки медицинских данных – это оценка исходов оперативного вмешательства. Большое количество данных, влияющих на выбор лечения пациента, вызывает необходимость разработки алгоритмов и систем поддержки принятия врачебных решений. При этом главными требованиями к подобным алгоритмам остаются прозрачность, наглядность и простота интерпретации решающих правил. Наиболее ярко отражает черты многомерного анализа данных, отвечая при этом всем вышеупомянутым требованиям, алгоритм построения деревьев классификации (деревьев решений).

Были построены деревья решений для оценки 5- и 10-летней выживаемости пациентов СООД, перенесших РЦ (рис. 3, 4).

Представленный подход реализован в пакете Statistica for Windows v.10., по описанной в [10–12] методике.

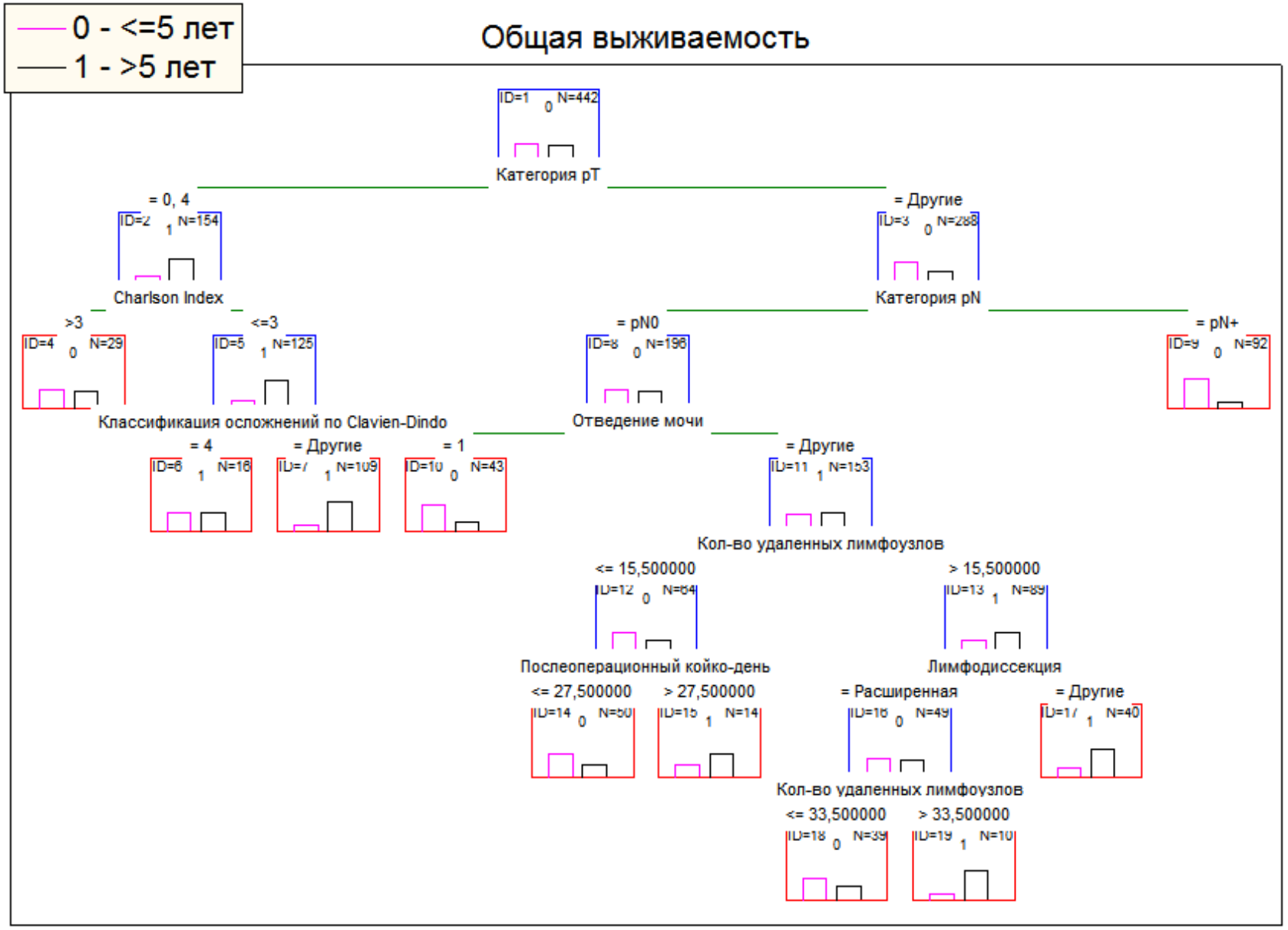


Рис. 3. Дерево решений для оценки 5-летней выживаемости

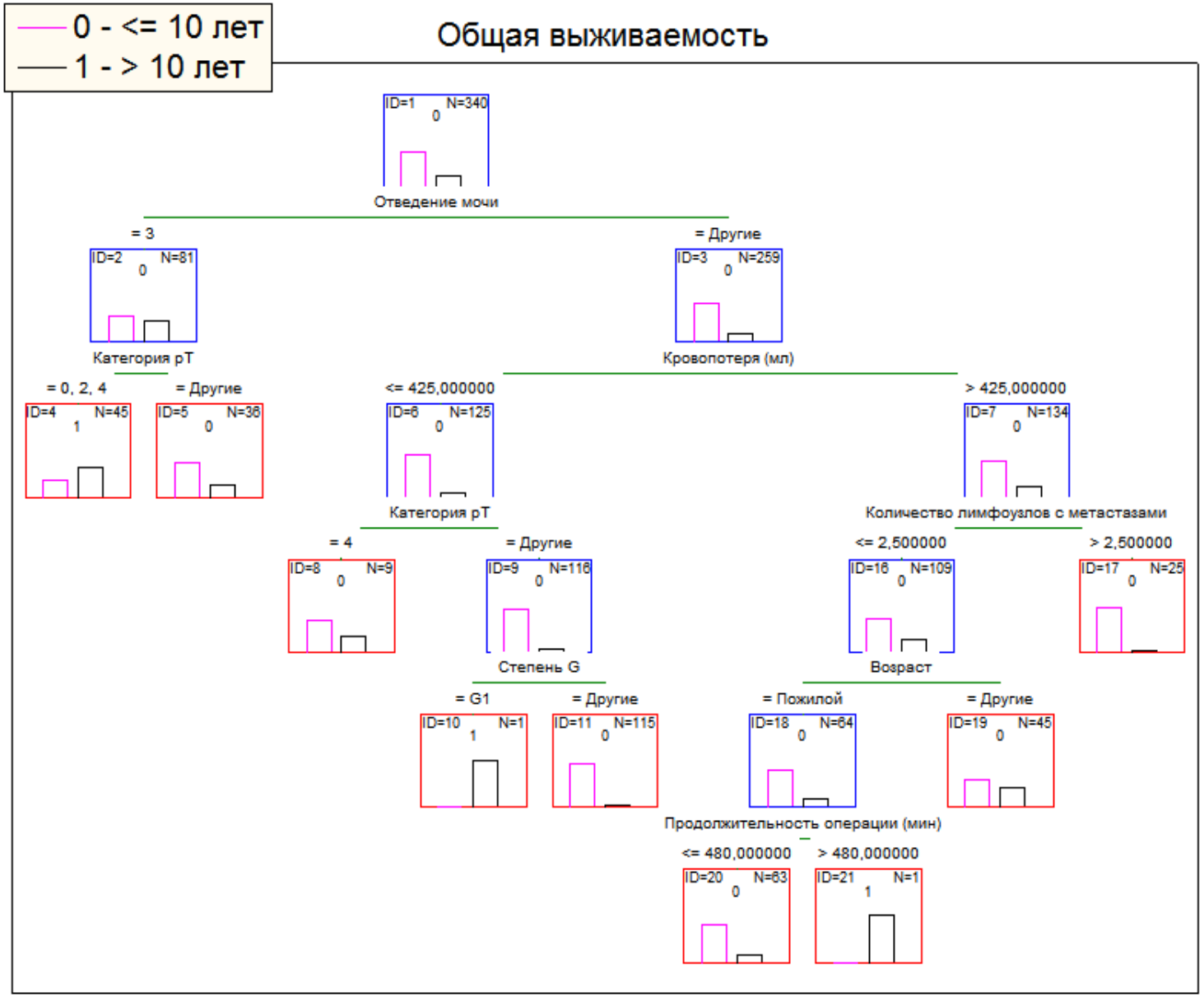


Рис. 4. Дерево решений для оценки 10-летней выживаемости

На рис. 3 представлено дерево решений для оценки 5-летней выживаемости пациентов СООД, перенесших РЦ, построенное методом ветвления CART с оцениваемыми априорными вероятностями, равными ценами ошибок, и фактической остановкой алгоритма с долей неклассифицированных объектов 0,1. Представленная модель содержит 9 терминальных вершин. Чувствительность такой классификации составляет 80,3 %, специфичность – 71,9 %, а точность – 76,0 %.

Дерево решений для оценки 10-летней выживаемости пациентов СООД, перенесших РЦ (рис. 4), построено методом ветвления CART с оцениваемыми априорными вероятностями, равными ценами ошибок, и остановкой алгоритма после первой ошибки классификации. Представленная модель содержит 8 терминальных вершин, чувствительность такой классификации составляет 37,5 %, специфичность – 93,5 %, а точность – 80,3 %.

На примере рис. 4 опишем, как трактовать представленное на нем дерево классификации.

Самая верхняя вершина носит название корневой и содержит всю исследуемую выборку. Алгоритм делит корневую вершину (исходную выборку) на две промежуточные. Название переменной, по которой происходит деление (ветвление), указано под разделяемой вершиной. Последовательно задаются вопросы, ответы на которые могут принимать только два значения. Окончательное решение зависит от ответов на все предыдущие вопросы. Процесс ветвления продолжается до остановки по выбранному критерию. Вершины, после которых не производится дальнейшее ветвление, называются терминальными.

Проследим, как формируется класс пациентов, которые смогут прожить 10 и более лет после проведения оперативного вмешательства (класс «1»). Для этого пройдем путь от корневой вершины до терминальных вершин, помеченных «1». Решающее правило попадания в класс «1» формируется, по сути, описанием этого маршрута. Пациент сможет прожить 10 и более лет после проведения оперативного вмешательства, если он из 1-й или 2-й группы по способу уродеривации; с кровопотерей в ходе операции 425 мл или меньше; при любой категории pT, кроме 4-й; и с низкой степенью злокачественности опухоли (G1). Описание решающего правила метода ДК похоже на то, как положение листа на дереве можно описать, указав ведущую к нему последовательность ветвей (начиная со ствола и заканчивая самой последней веткой, на которой располагается лист).

4. Обсуждение

Основная цель любой информационной технологии – это получение исследователем адекватной информации для ее дальнейшего анализа и принятия на его основе какого-либо решения. Мы показали, как методы визуализации позволяют буквально «одним взглядом» обнаружить особенности и выявить закономерности в больших объемах экспериментальных численных данных.

Перейдем к подробному обсуждению каждого из представленных подходов.

Такой подход визуализации многомерных данных, как *кривые выживаемости по методу Каплана – Мейера*, широко применяется в медицинских исследованиях и является золотым стандартом при оценке эффективности лечения в онкологии. Метод реализуем во многих современных пакетах статистической обработки данных (Statistica, SPSS, SAS и др.).

Интерпретируя *кривые выживаемости*, стоит помнить, что для оценки статистической значимости различий одного визуального анализа недостаточно. Возни-

кает необходимость в применении математического алгоритма сравнения кривых выживаемости, результатом которого является вычисленное *p*-значение, по близости к уровню значимости α которого возможно судить о статистически значимых различиях между полученными кривыми. Существует несколько математических алгоритмов, успешно зарекомендовавших себя при определенных условиях поставленной задачи. В медицинских исследованиях сравнение кривых выживаемости проводят с помощью Log-Rank Test или логрангового критерия [13–16].

Как было показано в разделе «Результаты», для того чтобы сделать вывод о статистической значимости различий, можно сравнивать и медианы выживаемости. Сравнение происходит по аналогии с логранговым критерием, с использованием так называемого медианного теста, являющегося частным случаем критерия Краскела – Уоллиса и непараметрической альтернативой одномерному дисперсионному анализу.

Широкое распространение и признание метод Каплана – Мейера получил благодаря прозрачному и простому для понимания алгоритму построения кривых выживаемости.

Относительным ограничением распространенности метода, особенностью применения подхода во врачебной практике, можно считать необходимость в дополнительной процедуре сравнения кривых выживаемости с помощью статистических критериев, что требует определенной математической подготовки.

Следующим рассмотренным подходом были *тепловые карты*, которые относятся к методам визуализации многомерных данных с помощью цветных индикаторов. При этом табличное представление информации все же сохраняется, но числа в ячейках заменяются на заливку цветом согласно выбранной шкале.

Важным фактором для правильной и интуитивно понятной интерпретации тепловой карты является хорошо подобранная цветовая схема. Существуют расходящиеся и последовательные цветовые схемы [17]. Расходящиеся палитры фиксируют цвета в нижнем и верхнем конце, а также в середине диапазона данных. Такое решение лучше подходит для данных, которые варьируются как в отрицательном, так и в положительном направлениях. Последовательные же палитры фиксируют наименьшее и наибольшее значения всего диапазона данных, оптимальны для неотрицательных данных (например, процент от 0 до 100). В нашем исследовании использована последовательная палитра, удобная для восприятия при отсутствии в анализе отрицательных численных данных. Этот метод представления данных активно применяется в исследованиях в области медицины, экономики и социологии [18–22].

В контексте применения во врачебной практике к преимуществам *тепловых карт* как метода визуализации многомерных данных можно отнести прозрачность алгоритма, наглядность результатов анализа, а также простоту интерпретации полученных графических образов.

Из недостатков тепловых карт можно выделить сложность обработки слишком большого объема данных, так как при этом метод теряет свое основное преимущество – простоту интерпретации. Общая схема такой карты слишком массивна, а отдельные ячейки в ней плохо различимы, что приводит к возможности оценить лишь «общую картину» происходящего.

Последним по порядку изложения, но отнюдь не по значению, является метод *деревьев классификации* (*ДК*). Применение метода ДК, на наш взгляд, способствует быстрому выявлению скрытых связей между данными в самой очевидной форме.

К преимуществам метода можно отнести высокую интерпретируемость модели, поддержку и числовых, и категориальных признаков, малое число входящих параметров. Легко визуализируется как само дерево, так и конкретное решающее правило (путь в дереве от корневой вершины к терминальной). К тому же алгоритм ДК способен учитывать опыт группы экспертов, рассматривающих конкретную проблему. Последнее обстоятельство повышает актуальность предложенного подхода.

К недостаткам метода ДК, построенных способом CART, можно отнести нестабильность. Небольшие изменения в данных могут существенно изменять построенное дерево решений. Эту проблему решают с помощью построения ансамблей деревьев классификации [3, 10–12].

Анализируя показатели качества построенных нами деревьев решений (см. раздел «Результаты»), нетрудно заметить, что модели качественно классифицируют лишь отрицательные случаи (пациенты СООД, не доживающие до 5- и 10-летней временной отметки). В ряде случаев такие модели допускаются к практическому применению, так как задача подразумевает точное определение только одной группы классифицированных объектов. Например, при построении прогностической модели оценки платежеспособности потенциальных заемщиков более важной задачей является выявление и классификация кредиторов с потенциальными задолженностями, нежели порядочных заемщиков, исправно вносящих платежи.

В случае же медицинских исследований такую модель не всегда можно назвать вполне удовлетворительной, так как ошибочная классификация «здоровых» пациентов в группу риска может спровоцировать ухудшение качества их жизни. Однако это не повод усложнять дерево, добавляя новые вершины в погоне за высоким процентом правильной классификации. Стоит признать, что при работе с МБД количественная эффективность не может быть принята как единственный критерий качества алгоритма, и ради сохранения простоты алгоритма бывает разумно отказаться от попыток достичь более высоких показателей его эффективности. Это обстоятельство еще раз подчеркивает необходимость совместной работы с экспертами в данной области.

5. Заключение

На базе трех подходов показана возможность выявления отношений между численными данными на основе визуального анализа полученной совокупности графических образов. Тот факт, что результаты анализа доступны для понимания медикам, а не только специалистам в математике и статистике, является большим преимуществом использования методов визуализации.

Определенно, за применением графики в исследованиях большое будущее, и не только потому, что увеличивается скорость передачи информации и повышается уровень ее понимания. Возможно, гораздо важнее тот факт, что применение методов визуализации способствует развитию таких важных для любого исследователя качеств, как интуиция и образное мышление.

6. Выводы

1. Показано, что методы визуализации многомерных данных и подходы на их основе перспективны для анализа и наглядного представления экспериментальных численных результатов из области медицины.
2. Для трех групп пациентов СООД, перенесших РЦ, построены кривые выживаемости методом Каплана – Мейера. Визуальный анализ кривых подтвердил

результат, полученный путем расчета медиан выживаемости. При этом визуальный анализ прост для понимания, не требует специальной математической подготовки и проведения дополнительных расчетов.

3. Построены тепловые карты для оценки 5- и 10-летней выживаемости пациентов СООД, перенесших РЦ тремя различными способами уродеривации. Продемонстрированы очевидные преимущества метода: максимально простой и понятный для восприятия результат в виде таблицы, ячейки которой окрашены согласно подобранной цветовой шкале. Данный подход оптимален в решении задач, которые требуют общего заключения, а также как первичная оценка набора данных для возможности применения в дальнейшем более сложных инструментов анализа.
4. Построены деревья решений для оценки 5- и 10-летней выживаемости пациентов, перенесших РЦ тремя различными способами уродеривации. Продемонстрировано, что алгоритмы на основе ДК способны не только наглядно визуализировать сложные связи между набором переменных, но и послужить основой для построения прогностических моделей оценки эффективности лечения.
5. Показано, что для объяснения поведения сложной системы возможно построение нескольких моделей, имеющих одинаковое право на существование.

7. Список литературы

1. *Шпигельхалтер, Д.* Искусство статистики. Как находить ответы в данных / Д. Шпигельхалтер. – М. : Манн, 2021. – 449 с. – ISBN 978-5-00-169250-8.
2. Применение методов визуализации при исследовании структуры экспериментальных многомерных данных / В. А. Воловоденко, О. Г. Берестнева, Е. В. Немеров, И. А. Осадчая // Известия Томского политехнического университета, 2012. – Т. 320, № 5. – С. 125–130. – EDN OZQTGT.
3. *Налимов, В. В.* Теория эксперимента / В. В. Налимов. – М. : Наука, 1971. – 206 с.
4. *Cleveland, W. S.* Graphical perception: Theory, experimentation, and application to the development of graphical methods / W. S. Cleveland, R. McGill // J. of the American statistical association. – 1984. – Т. 79, No. 387. – P. 531–554. – DOI 10.1080/01621459.1984.10478080.
5. *Гланц, С.* Медико-биологическая статистика / С. Гланц. – М. : Практика, 1999. – 459 с.
6. Estimating hazard ratios from published Kaplan – Meier survival curves: A methods validation study / R. Saluja, S. Cheng, K. A. delos Santos, K. K. Chan // Research Synthesis Methods. – 2019. – Т. 10, No. 3. – P. 465–475. – DOI: 10.1002/jrsm.1362.
7. *Messori, A.* Synthetizing published evidence on survival by reconstruction of patient-level data and generation of a multi-trial Kaplan – Meier curve / A. Messori // Cureus. – 2021. – Т. 13, No. 11. – DOI: 10.7759/cureus.19422.
8. *Wilkinson, L.* The history of the cluster heat map / L. Wilkinson, M. Friendly // The American Statistician. – 2009. – Т. 63, No. 2. – P. 179–184. – DOI: 10.1198/tas.2009.0033.
9. *Романова, И. К.* Современные методы визуализации многомерных данных: анализ, классификация, реализация, приложения в технических системах / И. К. Романова // Наука и образование: научное издание МГТУ им. Н. Э. Баумана. – 2016. – № 3. – С. 133–167. – DOI 10.7463/0316.0834876. – EDN VTKQIZ.
10. *Buntine, W.* Learning classification trees / W. Buntine // Artificial Intelligence frontiers in statistics. – 2020. – P. 182–201.

11. Benchmarking cesarean delivery rates using machine learning-derived optimal classification trees / A. C. Gimovsky, D. Zhuo, J. T. Levine [et al.] // Health Services Research. – 2022. – Т. 57, No. 4. – P. 796–805. – DOI: 10.1111/1475-6773.13921.
12. *Нессонова, М. Н.* Математические модели и методы построения классификаторов в медицине / М. Н. Нессонова. – LAP LAMBERT Academic Publishing, 2018. – 213 с. – ISBN 987-613-9-58671-4.
13. Does radical cystectomy improve overall survival in octogenarians with muscle-invasive bladder cancer? / S. Yoo, D. You, I. G. Jeong [et al.] // Korean J. of Urology. – 2011. – Т. 52, No. 7. – P. 446–451. – DOI: 10.4111/kju.2011.52.7.446.
14. Long-term oncologic outcomes after radical cystectomy for bladder cancer at a single institution / T. Kwon, I. G. Jeong, D. You [et al.] // J. of Korean medical science. – 2014. – Т. 29, No. 5. – P. 669–675. – DOI: 10.3346/jkms.2014.29.5.669.
15. *Schober, P.* Survival analysis and interpretation of time-to-event data: the tortoise and the hare / P. Schober, T. R. Vetter // Anesthesia and analgesia. – 2018. – Т. 127, No. 3. – P. 792. – DOI: 10.1213/ANE.0000000000003653.
16. Opencrimemapping. org: An Online Tool for Visualizing Crime / M. Crowder, L. Darr, G. Garza, B. Allen // SMU Data Science Review. – 2018. – Т. 1, No. 3. – P. 11.
17. *Harrower*, *M.* Colorbrewer.org: an online tool for selecting colour schemes for maps / M. Harrower, C. A. Brewer // The Cartographic J. – 2003. – Т. 40, No. 1. – P. 27–37.
18. *Жгун*, *Т. В.* Визуализация данных при оценке изменения качества социально-экономических систем / Т. В. Жгун, П. Д. Кристофер // Системный анализ в проектировании и управлении. – 2020. – Т. 24, № 1. – С. 257–267. – DOI: 10.18720/SPBPU/2/id20-130.
19. *DeBold*, *T.* Battling infectious diseases in the 20th century: the impact of vaccines / T. DeBold, D. Friedman // Wall Street J. – 2015. – URL: http://graphics.wsj.com/infectious-diseases-and-vaccines/.
20. Heat map visualization for electrocardiogram data analysis / H. Guo, W. Zhang, C. Ni [et al.] // BMC cardiovascular disorders. – 2020. – Т. 20, No. 1. – P. 1–8. – DOI: 10.1186/s12872-020-01560-8.
21. *Metsalu*, *T.* ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap / T. Metsalu, J. Vilo // Nucleic acids research. – 2015. – Т. 43, No. W1. – P. 566–570. – DOI: 10.1093/nar/gkv468.
22. *Lee*, *S. Y.* High-dose drug heat map analysis for drug safety and efficacy in multi-spheroid brain normal cells and GBM patient-derived cells / S. Y. Lee, Y. Teng, M. Son [et al.] // Plos one. – 2021. – Т. 16, No. 12. – e0251998. – DOI: 10.1371/journal.pone.0251998.

Сведения об авторах:

Огородникова Светлана Юрьевна, студентка Уральского федерального университета им. первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия. Эл. почта: ogorodnikova.sveta2507@gmail.com.

Константинова Екатерина Даниловна, канд. ф.-м. н., старший научный сотрудник, и. о. зав. лаборатории Математического моделирования в экологии и медицине Института промышленной экологии УрО РАН, г. Екатеринбург, Россия. Эл. почта: K_Konst@ecko.uran.ru

DATA VISUALIZATION METHODS IN BIOMEDICAL RESEARCH

S. Y. Ogorodnikova [1], E. D. Konstantinova [2]

[1] *Ural Federal University named after the first President of Russia B. N. Yeltsin, Ekaterinburg, Russia*
[2] *Institute of Industrial Ecology, Ural Branch, Russian Academy of Sciences, Ekaterinburg, Russia*

The article presents the results of applying multidimensional data visualization methods to solving an applied problem from the field of medicine. We used retrospective data of 586 patients with bladder cancer who underwent radical cystectomy from December 2001 to May 2021 in the urological oncology department of the Sverdlovsk Regional Oncology Center. Such approaches as survival curves (according to the Kaplan – Meier method), heat maps and classification trees are implemented as methods for visualizing multidimensional data. The Statistica for Windows 10 version and MatLAB2021b packages were used for data analysis. On the basis of three approaches, the possibility of identifying relationships between numerical data based on a visual analysis of the resulting set of graphic images is shown. Examples of constructing visual, intuitive decision rules for each of the methods in solving the problem of assessing the 5- and 10-year survival of Sverdlovsk Regional Oncology Center patients who underwent radical cystectomy with three different methods of uroderivation are demonstrated. The features, advantages and disadvantages of each of the approaches in relation to working with data in biomedical research are considered. It is shown that the methods of visualization of multidimensional data and approaches based on them are promising for the analysis and visual presentation of experimental numerical results from the field of medicine.

Key words: biomedical research; data visualization; heat maps; survival curves; classification trees.

References

1. *Shpigel'halter, D*. The art of statistics. How to find answers in data / D. Shpigel'halter. – M. : Mann, 2021. – 449 p. – ISBN 978-5-00-169250-8. (In Russian).
2. Application of visualization methods in the study of the structure of experimental multidimensional data / V. A. Volovodenko, O. G. Berestneva, E. V. Nemerov, I. A. Osadchaya // Proceedings of the Tomsk Polytechnic University, 2012. – Vol. 320, No. 5. – P. 125–130. – EDN OZQTGT. (In Russian).
3. *Nalimov, V. V*. Theory of experiment / V. V. Nalimov. – M. : Nauka, 1971. – 206 p. (In Russian).
4. *Cleveland, W. S*. Graphical perception: Theory, experimentation, and application to the development of graphical methods / W. S. Cleveland, R. McGill // J. of the American statistical association. – 1984. – T. 79, No. 387. – P. 531–554. – DOI 10.1080/01621459.1984.10478080.
5. *Glants, S*. Medico-biological statistics / S. Glants. – M. : Praktika, 1999. – 459 p. (In Russian).

6. Estimating hazard ratios from published Kaplan – Meier survival curves: A methods validation study / R. Saluja, S. Cheng, K. A. delos Santos, K. K. Chan // Research Synthesis Methods. – 2019. – T. 10, No. 3. – P. 465–475. – DOI: 10.1002/jrsm.1362.
7. *Messori, A*. Synthetizing published evidence on survival by reconstruction of patient-level data and generation of a multi-trial Kaplan – Meier curve / A. Messori // Cureus. – 2021. – T. 13, No. 11. – DOI: 10.7759/cureus.19422.
8. *Wilkinson, L*. The history of the cluster heat map / L. Wilkinson, M. Friendly // The American Statistician. – 2009. – T. 63, No. 2. – P. 179–184. – DOI: 10.1198/tas.2009.0033.
9. *Romanova, I. K*. Modern methods of visualization of multidimensional data: analysis, classification, implementation, applications in technical systems / I. K. Romanova // Science and Education: scientific edition of Bauman Moscow State Technical University. – 2016, No. 3. – P. 133–167. – DOI 10.7463/0316.0834876. – EDN VTKQIZ. (In Russian).
10. *Buntine, W*. Learning classification trees / W. Buntine // Artificial Intelligence frontiers in statistics. – 2020. – P. 182–201.
11. Benchmarking cesarean delivery rates using machine learning-derived optimal classification trees / A. C. Gimovsky, D. Zhuo, J. T. Levine [et al.] // Health Services Research. – 2022. – T. 57, No. 4. – P. 796–805. – DOI: 10.1111/1475-6773.13921.
12. *Nessonova, M. N*. Mathematical models and methods of constructing classifiers in medicine / M. N. Nessonova. – LAP LAMBERT Academic Publishing, 2018. – 213 p. – ISBN 987-613-9-58671-4. (In Russian).
13. Does radical cystectomy improve overall survival in octogenarians with muscle-invasive bladder cancer? / S. Yoo, D. You, I. G. Jeong [et al.] // Korean J. of Urology. – 2011. – T. 52, No. 7. – P. 446–451. – DOI: 10.4111/kju.2011.52.7.446.
14. Long-term oncologic outcomes after radical cystectomy for bladder cancer at a single institution / T. Kwon, I. G. Jeong, D. You [et al.] // J. of Korean medical science. – 2014. – T. 29, No. 5. – P. 669–675. – DOI: 10.3346/jkms.2014.29.5.669.
15. *Schober*, *P*. Survival analysis and interpretation of time-to-event data: the tortoise and the hare / P. Schober, T. R. Vetter // Anesthesia and analgesia. – 2018. – T. 127, No. 3. – P. 792. – DOI: 10.1213/ANE.0000000000003653.
16. Opencrimemapping. org: An Online Tool for Visualizing Crime / M. Crowder, L. Darr, G. Garza, B. Allen // SMU Data Science Review. – 2018. – T. 1, No. 3. – P. 11.
17. *Harrower*, *M*. Colorbrewer.org: an online tool for selecting colour schemes for maps / M. Harrower, C. A. Brewer // The Cartographic J. – 2003. – T. 40, No. 1. – P. 27–37.
18. *Zhgun*, *T. V*. Data visualization in assessing changes in the quality of socio-economic systems / T. V. Zhgun, P. D. Christopher // System analysis in design and management. – 2020. – Vol. 24, No. 1. – P. 257–267. – DOI: 10.18720/SPBPU/2/id20-130.
19. *DeBold*, *T*. Battling infectious diseases in the 20th century: the impact of vaccines / T. DeBold, D. Friedman // Wall Street J. – 2015. – URL: http://graphics.wsj.com/infectious-diseases-and-vaccines/.
20. Heat map visualization for electrocardiogram data analysis / H. Guo, W. Zhang, C. Ni [et al.] // BMC cardiovascular disorders. – 2020. – T. 20, No. 1. – P. 1–8. – DOI: 10.1186/s12872-020-01560-8.
21. *Metsalu*, *T*. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap / T. Metsalu, J. Vilo // Nucleic acids research. – 2015. – T. 43, No. W1. – P. 566–570. – DOI: 10.1093/nar/gkv468.
22. *Lee*, *S. Y*. High-dose drug heat map analysis for drug safety and efficacy in multi-spheroid brain normal cells and GBM patient-derived cells / S. Y. Lee, Y. Teng, M. Son [et al.] // Plos one. – 2021. – T. 16, No. 12. – e0251998. – DOI: 10.1371/journal.pone.0251998.