РАЗРАБОТКА МОДУЛЯ АНАЛИЗА ДАННЫХ О ВЫЖИВАЕМОСТИ ДЛЯ ИНФОРМАЦИОННОЙ СИСТЕМЫ УПРАВЛЕНИЯ КЛИНИЧЕСКИМИ ДАННЫМИ ТРАНСПЛАНТАЦИЙ ГЕМОПОЭТИЧЕСКИХ СТВОЛОВЫХ КЛЕТОК

Кулешов Егор Алексеевич
магистрант,
МИРЭА – Российский технологический университет.
Россия, г. Москва
Научный руководитель:
Старичкова Юлия Викторовна
кандидат технических наук, профессор.
МИРЭА – Российский технологический университет
Россия, г. Москва

АННОТАЦИЯ

Информационная система управления клиническими данными трансплантаций гемопоэтических стволовых клеток является узкоспециализированным программным обеспечением. Оно разработано для ФГБУ «НМИЦ ДГОИ им. Дмитрия Рогачева» Минздрава России и является частью комплексной информационной системы медицинского центра. Система управления клиническими данными трансплантаций гемопоэтических стволовых клеток представляет собой комплекс программ, занимающийся сбором и хранением клинических данных пациентов. С помощью этой информационной системы можно управлять клиническими данными, полученными из лабораторных информационных систем, а также из медицинской информационной системы медицинского центра. Перед проведением ТГСК комиссия изучает данные пациента для выноса решения о проведении операции. Однако, в этой информационной системе отсутствует возможность провести анализ выживаемости пациентов. Для этого был разработан внешний модуль анализа выживаемости пациентов. Благодаря разработке этого модуля, персонал медицинского центра получит удобный инструмент для проведения анализа данных.

ANNOTATION

The information system for managing clinical data of hematopoietic stem cell transplantation is a highly specialized software. It was developed for Dmitry Rogachev National Research Center and is part of the integrated information system of the medical center. The Hematopoietic stem cell Transplantation Clinical Data Management system is a set of programs that collect and store clinical data of patients. With this information system, you can manage clinical data obtained from laboratory information systems, as well as from the medical information system of the medical center. Before performing the HSCT, the commission examines the patient's data in order to make a decision about the operation. However, it is not possible to conduct a survival analysis in this information system. For this purpose, an external module for analyzing patient survival has been developed. Thanks to the development of this module, the staff of the medical center will receive a convenient tool for data analysis.

Ключевые слова: Информационная система; Анализ данных; Базы данных.

Keywords: Data analysis; Information system; Database.

В ходе статьи будут рассмотрены основные методы, необходимые для проведения анализа выживаемости и их применимость, после чего будет приведено информационное обеспечение разработки и приведены результаты.

Для описания средних времен жизни и сравнения нового метода лечения со старыми, можно было бы использовать стандартные параметрические и непараметрические методы.

Ниже будут рассмотрены методы, связанные с анализом выживаемости, а также те, которые участвуют в ходе обработки медицинских данных. Существуют методы вычисления и оценки функции выживаемости, такие как оценки Каплана-Мейера и метод линейной логистической регрессии Кокса.

Оценка Каплана-Мейера - один из предпочтительных вариантов, который можно использовать для измерения доли объектов анализа, живущих в течение определенного времени после лечения. Безусловным преимуществом данного метода является то, что в данном методе можно использовать цензурированные данные. В клинических испытаниях оценивается путем измерения количества объектов, выживших после этого в течение определенного периода времени. Время от определенной точки до наступления данного события, например смерти, называется временем выживания, а анализ групповых данных - анализом выживаемости. На это могут повлиять объекты исследования, которые выбили из исследования или данные о которых частичны, то есть мы теряем с ними связь на полпути в исследовании. Мы называем эти ситуации цензурированными наблюдениями. Оценка Каплана-Мейера — это простейший способ вычисления выживаемости во времени. Вероятность выживания в любой конкретный момент времени рассчитывается по формуле, приведенной ниже:

$$S_t = \frac{\text{Количество объектов на старте} - \text{количество умерших объектов}}{\text{Количество объектов на старте}} \quad (1)$$

Кривая выживаемости может быть построена для различных ситуаций. Метод включает в себя вычисление вероятностей возникновения события в определенный момент времени и умножение этих последовательных вероятностей на любые ранее вычисленные вероятности для получения окончательной оценки. Это можно вычислить для двух групп испытуемых, а также их статистическую разницу в выживаемости. Это можно использовать, когда они сравнивают два препарата и ищут выживаемость субъектов или как величину показывающую эффективность лечения.

Результатом проведения анализа выживаемости методом Каплана-Майера будет функция выживаемости пациентов. В разрабатываемом модуле она будет отображаться в виде графика.

Вторым методом анализа, который будет использоваться в разрабатываемом модуле будет линейная логистической регрессии Кокса.

Согласно Дэвиду Коксу, если предположение о пропорциональной опасности верное, можно оценить размер эффекта без учета функции базовой опасности. Поэтому, пропорциональность рисков – это главный фактор, определяющий возможность применения этого метода.

Если допущение пропорциональности рисков не выполняется, то применяется регрессия Кокса с ковариатами, которые зависят от времени. Модель пропорциональных рисков Кокса полупараметрическая, соответственно она не предполагает наличия какой-либо информации о базовой функции отказов, однако при таком подходе определен вид регрессионной функции.

Отличие от оценки Каплана – Мейера заключается в том, что регрессия Кокса моделирует функцию риска, а не выживаемость. Несмотря на это с помощью этой модели можно получить оценку рисков. Такая модель не может рассчитать конкретное значение абсолютного риска, возможно оценивать только относительный риск. В качестве меры относительного риска выступают коэффициенты опасности.

Для наблюдаемых объектов исследования, которые подвергаются цензурированию в момент времени t, надо рассчитывать риск наступления терминального события точно так же, как и для “обычных” объектов наблюдения, которые все еще наблюдаются в момент t. В реальности такое случается довольно редко. Больные пациенты, находящиеся в листе ожидания, подвергаются не только риску смерти, но и могут быть исключенными из исследования из-за различных факторов.

Метод линейной логистической регрессии Кокса - наиболее общая регрессионная модель, по причине того, что в ней не используются какие-либо предположения касательно распределения времени выживания. Такая модель предполагает, что у функции интенсивности имеется уровень у, который является функцией независимых переменных. Не делается предположений о том, как выглядит функции интенсивности [4, с. 35]. Именно поэтому модель Кокса может рассматриваться как в некотором смысле непараметрическая. В

этой модели переменной отклика является «опасность». Согласно модели Кокса, риск наступления события выражается следующей формулой:

$$h(T) = h_0(T)e^{\sum_{i=1}^{p} x_i\beta_i} \quad (2)$$

где: T – время;

$x_1, \ldots, x_p$ — независимые переменные (предикторы);

$h_0(T)$ – базовый риск наступления события, являющийся одинаковым для всех пациентов (только при условии, что все независимые переменные равны 0, соответственно никак не влияют на исход);

$\beta_i, \ldots, \beta_p$ – коэффициенты.

Вычислив коэффициенты, получится измерить насколько различные факторы влияют на опасность в единицу времени относительно друг друга. Важным замечанием будет, что необходимо делать только сравнительные утверждения об опасности, можно сказать, что опасность для одной группы в три раза выше, чем у другой, но неверно будет сказать, насколько высока или низкая, любая функция является компромиссом, связанным с регрессией Кокса. [4, с. 13]. Результатом такой модели будут являться относительные коэффициенты предикатов, которые также будут изображаться на графике.

Итак, выше были описаны математические методы, которые будут использоваться для проведения анализа выживаемости. Далее, для реализации разработки модуля необходимо провести информационное обеспечение проекта.

Для того чтобы понять, как хранятся данные в базе данных, был проведен ее анализ. Так как используемая документоориентированая база данных не обладает связями между сущностями, в базе данных используются коллекции, а не таблицы сущностей. Итак, приступим к анализу модели информационной базы. В базе данных присутствуют 3 коллекции содержащую информацию, характеризующую пациентов:

Entities – сущности предметной области и их структура: поля, связи;

Records – непосредственно клинические данные;

Dictionaries – коллекция содержащая данные для расшифровки некоторых значений из коллекции records. В коллекции records некоторые данные представлены в виде массивов, которые соответствуют данным в коллекции dictionaries. Например, в коллекции records значение диагноза у пациента равно ["1", "0", "0"], это значит, что для того, чтобы получить реальное значение, необходимо перейти в коллекцию dictionaries, в поле diagnoses, в нем будет значение – объект json. В этом объекте необходимо перейти по пути ключей с индексами 1, 0, 0. После перехода по всем ключам, можно получить значение диагноза.

Поскольку в медицинских данных, хранящихся в информационной системе управления клиническими данными трансплантаций гемопоэтических стволовых клеток, присутствует огромная вложенность и классификация данных, была выбрана документоориентированная база данных. Если бы данные хранились в реляционной СУБД пришлось бы создавать огромное количество таблиц и связей для классификации данных. Было бы огромное количество колонок с нулевыми значениями и спутанность данных. Все эти проблемы решила документоориентированная база данных, однако такой подход требует условной логики и сложности с получением данных.

Итак, коллекции entities, dictionaries и records не обладают связями между собой в контексте базы данных, однако связаны между собой логически. Концептуальная модель представлена на рисунке 1:

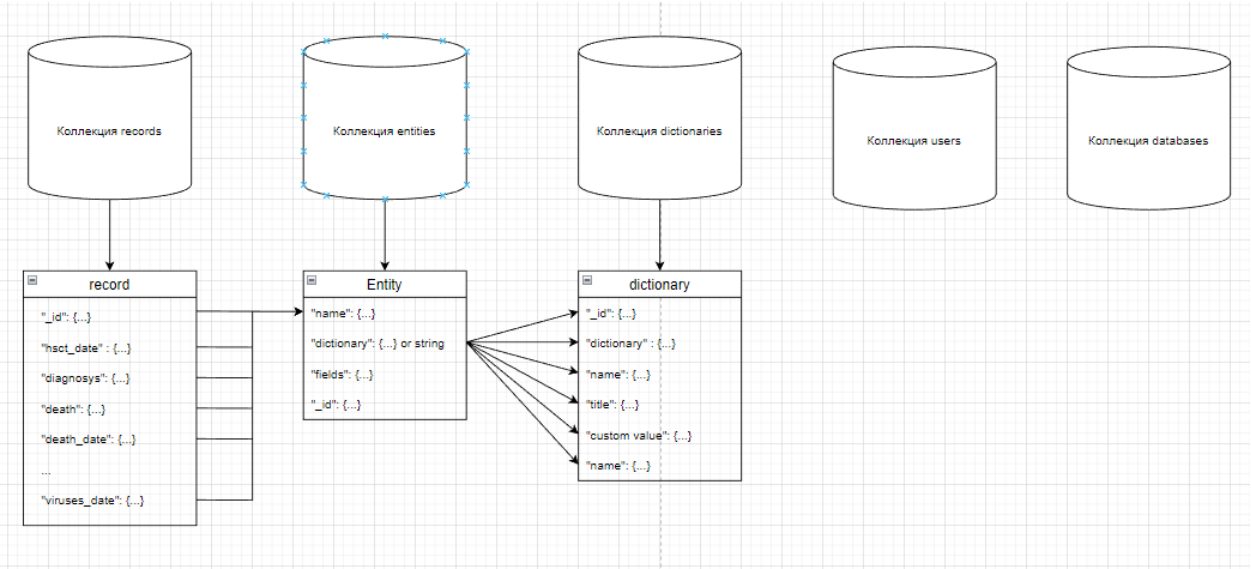

Рисунок 1. Концептуальная схема базы данных

Каждое значение ключа в коллекции records соответствует значению ключа “name” в коллекции entities. Значение ключа “dictionary” в коллекции entities может принимать значения типа словаря или типа строки. В случае принятия значения типа строки, элемент коллекции entities имеет логическую связь с элементом в коллекции dictionaries. Значение ключа “dictionary” в коллекции entities соответствует значению элемента в коллекции dictionaries со значением ключа “name”. Проходя по такому пути, будут получаться данные для программного модуля.

В разработанном приложении пользовательский интерфейс представлен в виде одностраничного веб-сайта. Ниже продемонстрированы варианты взаимодействия пользователя с программным модулем анализа выживаемости.

На рисунке 2 представлен интерфейс модуля при заходе пользователя на сервис.

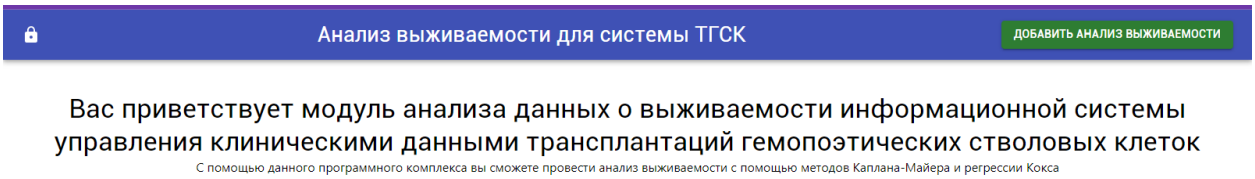

Рисунок 2. Интерфейс модуля сервиса

Далее, при нажатии пользователем кнопки “Добавить анализ выживаемости”, появляется меню, представленное на рисунке 3, в котором пользователь выбирает вид анализа.

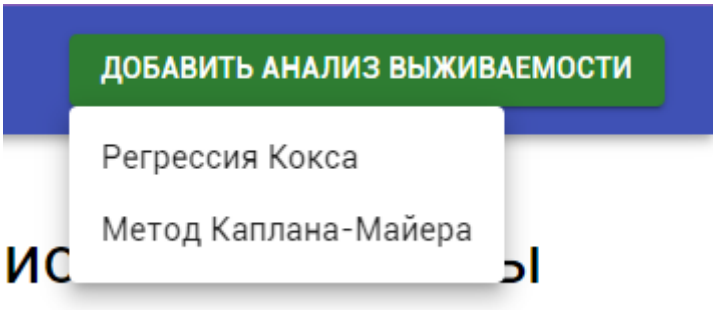

Рисунок 3. Окно выбора вида анализа

При выборе пользователем одного из видов анализа, меню скрывается и появляются соответствующие окна. Для структурированности повествования предлагается сначала рассмотреть формы метода Каплана-Майера, а затем регрессии Кокса.

Итак, при выборе метода анализа Каплана-Майера открывается следующее окно.

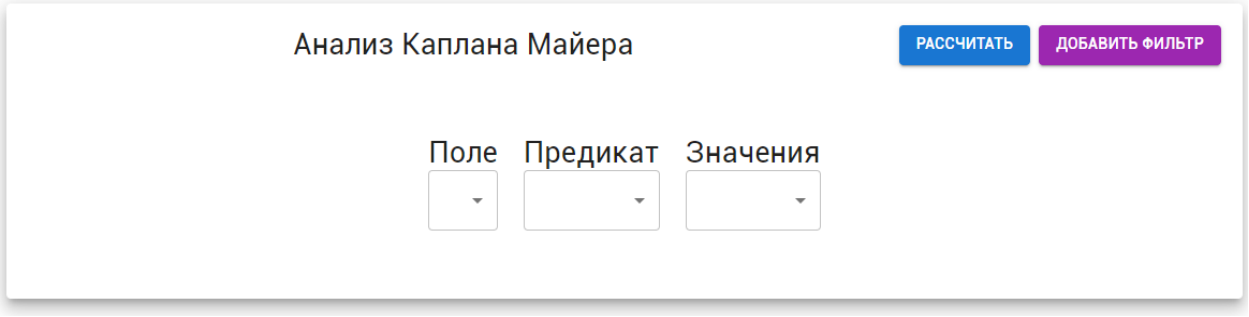

Рисунок 4. Окно выбора вида анализа

Далее пользователь может выбрать поле, предикат и значение, которому должно соответствовать поле. Исходя из этих фильтров будет формироваться выборка пациентов, которые будут участвовать в анализе. В списке полей находятся поля для фильтрации, а в поле “Значения” – значения, которые могут принимать эти поля. Это продемонстрировано на рисунках 5 и 6.

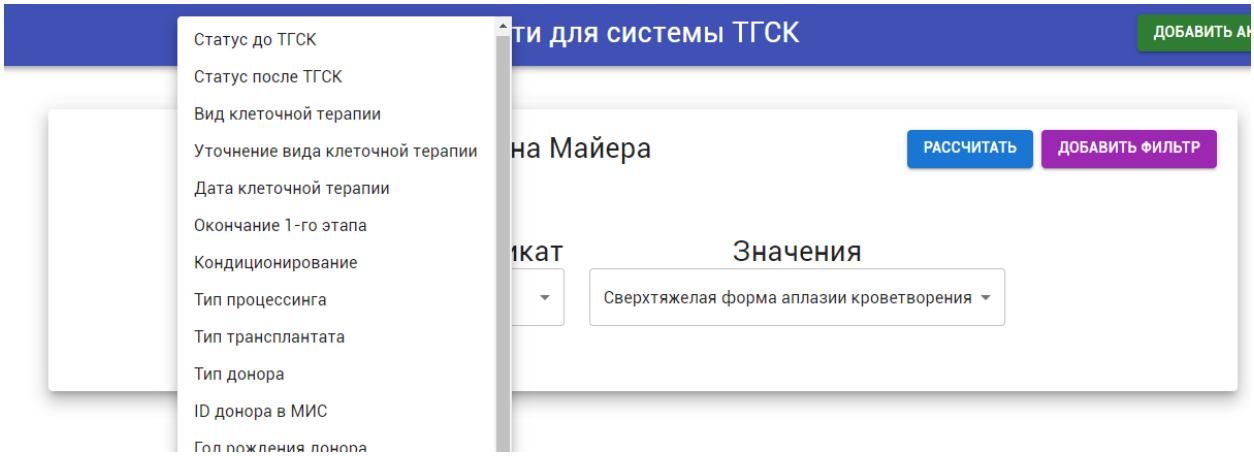

Рисунок 5. Демонстрация выбора значения поля

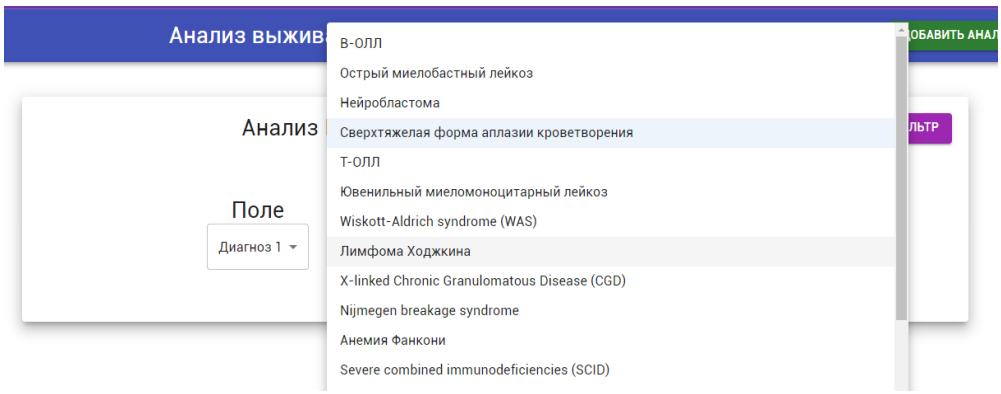

Рисунок 6. Демонстрация выбора значения поля “Значение”

Пользователь может выбирать несколько фильтров. На рисунке 7 приведен пример фильтрации выборки пациентов по значению диагноза и типу донора.

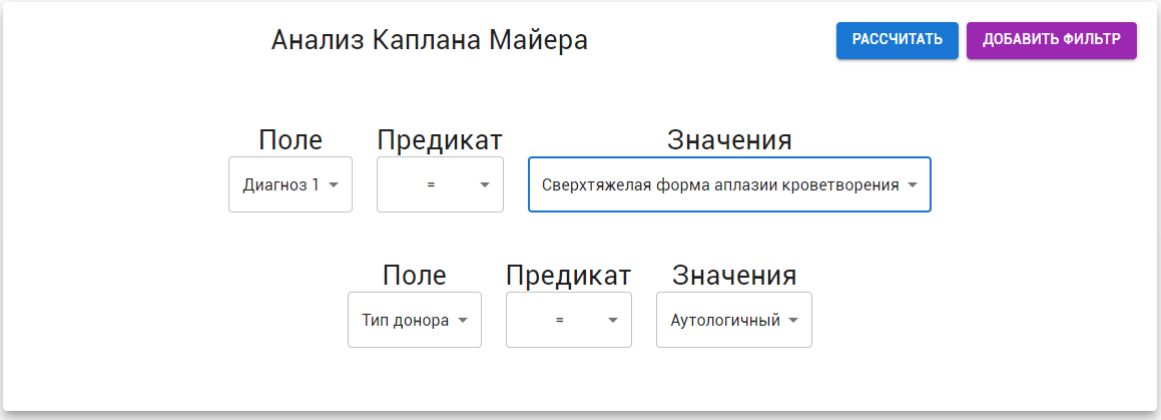

Рисунок 7. Пример фильтрации выборки пациентов по значению диагноза и типу трансплантанта

Пользователь может добавлять неограниченное количество фильтров, однако, если не останется записей, либо фильтры будут некорректны (например, если бы в примере выше предикат был отличный от “=”), то пользователю выведется alert-уведомление.

После выбора фильтров пользователь может нажать на кнопку “Рассчитать” и произойдет расчет соответствующего анализа. Результаты приведены на рисунке 8.

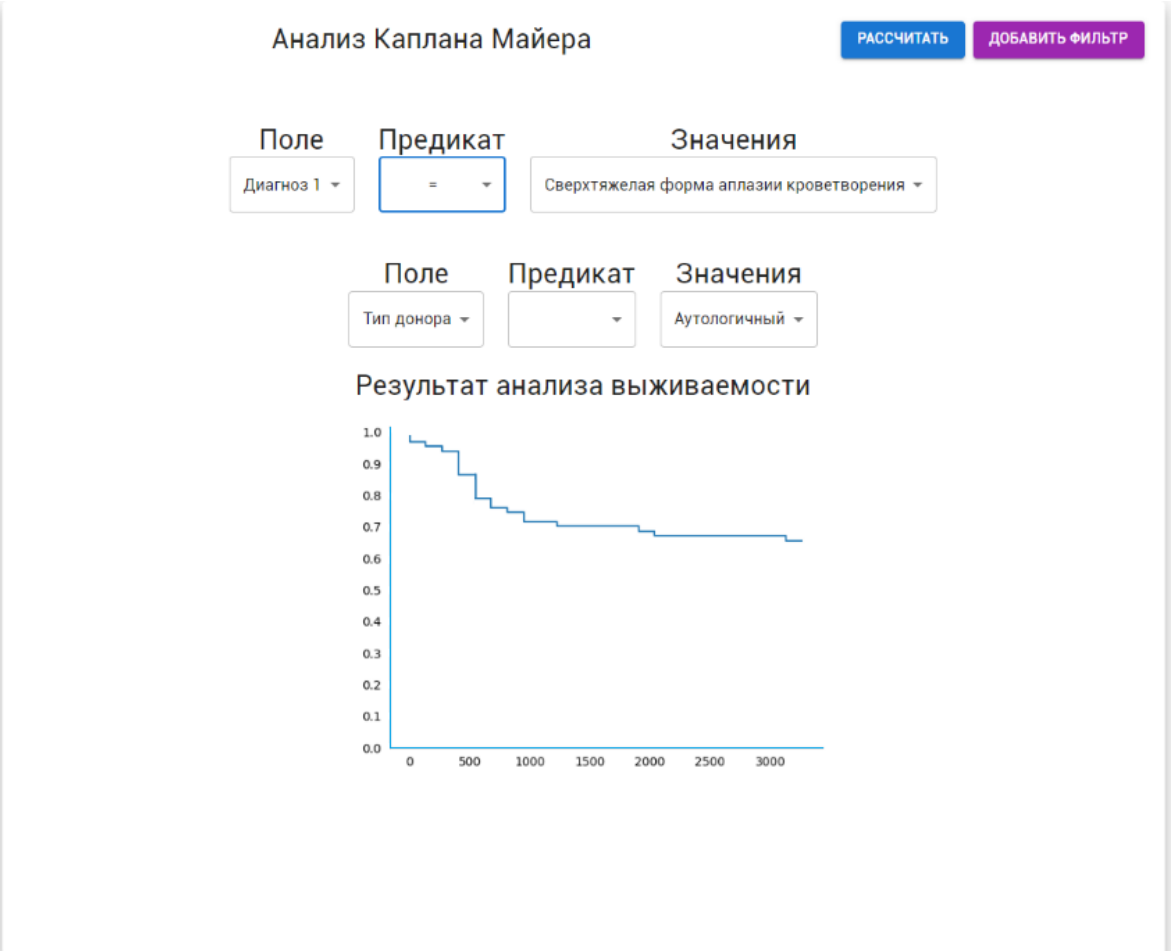

Рисунок 8. Результаты анализа методом Каплана-Майера

Далее, если пользователь хочет продолжить и провести еще один анализ, он может повторно нажать на кнопку «Добавить анализ выживаемости» и выбрать нужный. В результате следующий анализ отобразится под предыдущим. Это продемонстрировано на рисунке 9.

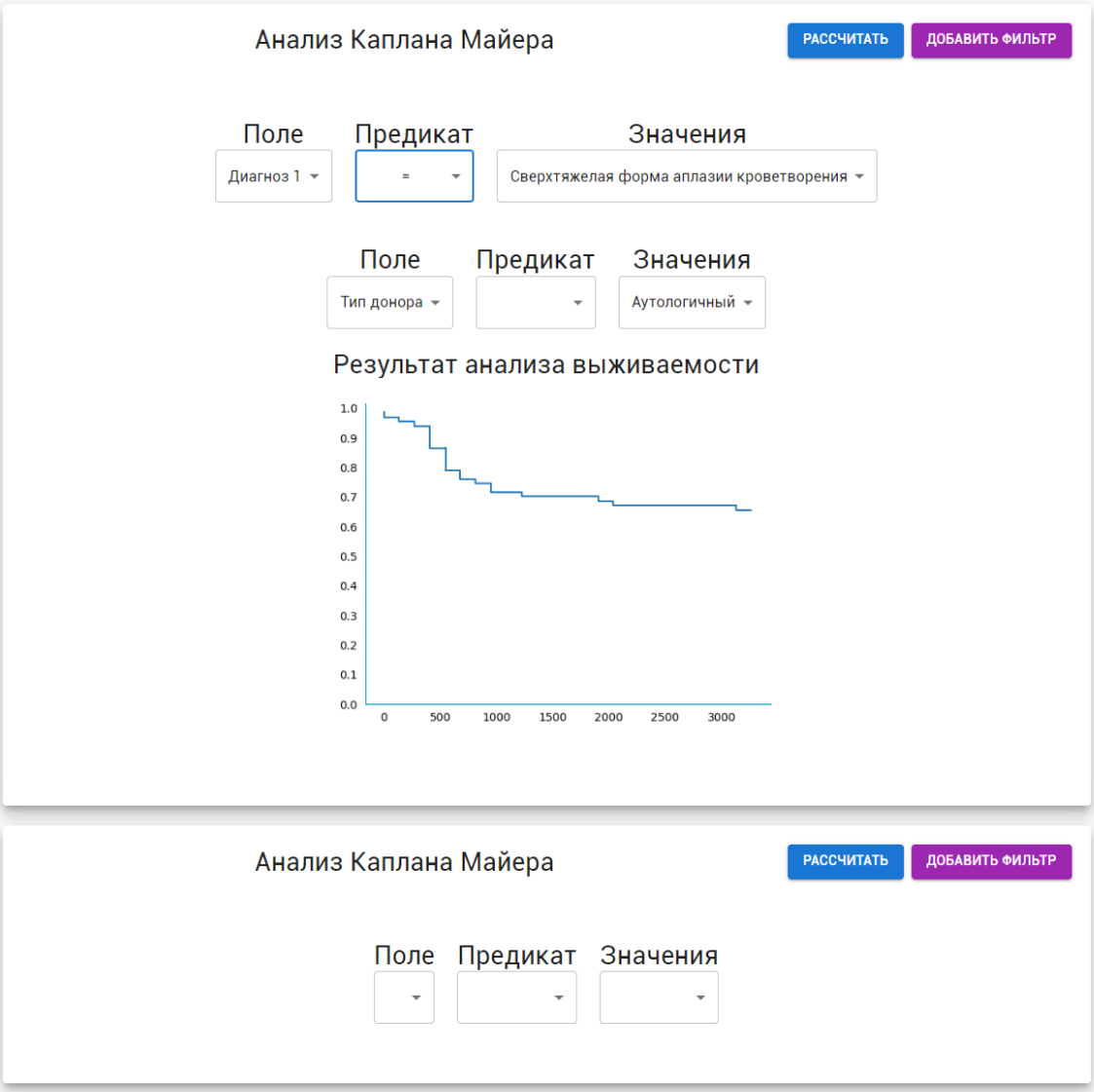

Рисунок 9. Добавление дополнительного анализа выживаемости

Метод оценки Каплана-Майера не способен оценить влияние тех или иных характеристик-предикатов пациентов. Метод позволял только оценить, насколько подвержены риску пациенты по истечении определенного времени. Однако, часто необходимо числено определить, насколько влияют факторы на выживаемость пациента. Далее будет рассмотрена форма проведения анализа с помощью регрессии Кокса. Фильтрация осуществляется аналогичным предыдущим способом, однако для анализа с помощью регрессии Кокса необходимо ввести дополнительные данные. Это продемонстрировано на рисунке 10. Пользователю необходимо ввести данные представляющие

предикаторы-факторы, для того чтобы определить, насколько эти факторы, оказывают влияние на выживаемость пациента. Это представлено в виде удобного инструмента.

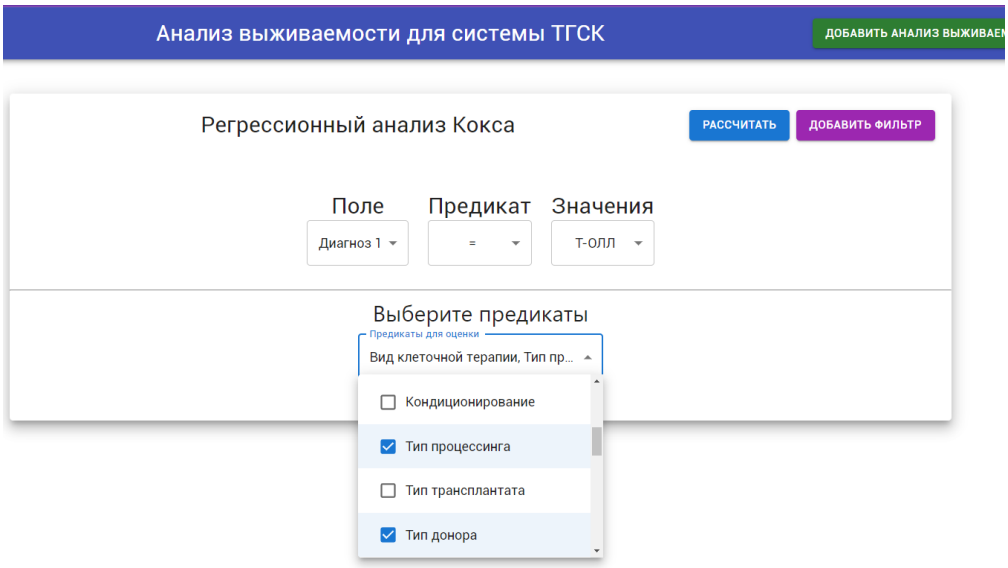

Рисунок 10. Выбор предикатов, которые будут оцениваться в ходе анализа регрессией Кокса

Результат анализа представлен на рисунке 11.

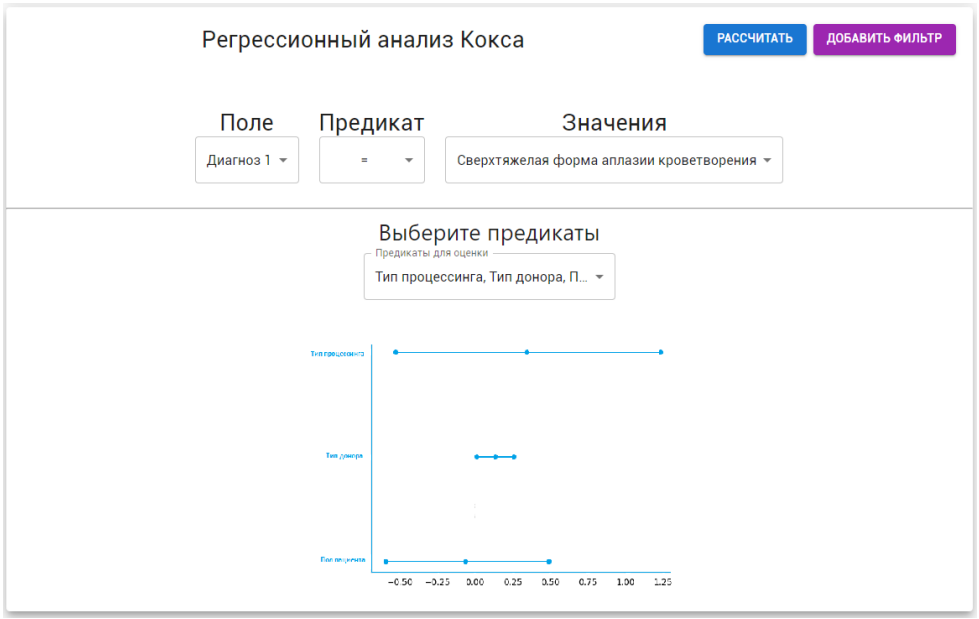

Рисунок 11. **Результат анализа с помощью метода регрессии Кокса**

Итак, в результате описания экранных форм, были показаны виды анализа, которые пользователь сможет использовать, а также продемонстрирована работа с фильтрами и предикатами.

Таким образом, в результате взаимодействия с программным интерфейсом, пользователь сможет получать результаты анализа данных с помощью метода регрессии Кокса или метода Каплана-Майера.

Список литературы:

1. НМИЦ ДГОИ им. Дмитрия Рогачева, официальный сайт [Электронный ресурс]. – Режим доступа: https://fnkc.ru/ (дата обращения: 25.05.2022)
2. Визуализация данных: основные правила, полезные приемы и инструменты [Электронный ресурс]. – Режим доступа: https://www.owox.ru/blog/articles/data-visualization/ (дата обращения: 27.04.2022)
3. Старичкова Ю. В., Воронин К. А., Борисова Н. В., Масчан М. А., Румянцев А. Г. Подходы к интеграции комплекса программных средств управления процессами и сложноструктурированными медицинскими данными с медицинскими и лабораторными информационными системами в учреждениях здравоохранения. 2020
4. Хасти Тревор. - Основы статистического обучения: интеллектуальный анализ данных, логический вывод и прогнозирование. – 2020
5. Стентон Гланц Медико-биологическая статистика. Primer of BIOSTATISTICS. — 4-е изд. — М.: Практика, 2019
6. Мухина А.А., Кузьменко Н.Б., Родина Ю.А, Хорева А.Л, Моисеева А.А, Швец О.В, и др. Эпидемиология первичных иммунодефицитов в Российской Федерации. Педиатрия. Журнал им. Г.Н. Сперанского. 2020; (Дата обращения: 05.01.2022)
7. Highcharts – Краткое руководство [Электронный ресурс]. – Режим доступа: https://coderlessons.com/tutorials/bolshie-dannye-i-analitika/uznaite-highcharts/highcharts-kratkoe-rukovodstvo (дата обращения: 20.05.2021)

8. Лучшие библиотеки визуализации данных JavaScript [Электронный ресурс]. – Режим доступа: https://www.monterail.com/blog/javascript-libraries-data-visualization#first (дата обращения: 27.04.2021)