Федеральное государственное бюджетное учреждение
«Национальный медицинский исследовательский центр онкологии
имени Н.Н. Петрова»
Министерства здравоохранения Российской Федерации
(ФГБУ «НМИЦ онкологии им. Н.Н. Петрова» Минздрава России)
Отдел учебно-методической работы

Федеральное государственное бюджетное образовательное учреждение
высшего образования «Северо-Западный государственный
медицинский университет имени И.И. Мечникова»
Министерства здравоохранения Российской Федерации
(ФГБОУ ВО СЗГМУ им. И.И. Мечникова Минздрава России)
Отдел дополнительного профессионального образования
Кафедра онкологии

Беляев А. М., Михнин А. Е., Рогачев М. В.

Подготовка данных и анализ выживаемости в пакетах статистических программ MedCalc и Statistica

Учебное пособие

Санкт-Петербург
2022

УДК:614.1:681.3.06(07)
ББК:51.1(2)я7

Беляев А. М., Михнин А. Е., Рогачев М. В. Подготовка данных и анализ выживаемости в пакетах статистических программ MedCalc и Statistica: учебное пособие для обучающихся в системе высшего и дополнительного профессионального образования. – Санкт-Петербург: НМИЦ онкологии им. Н.Н. Петрова, 2022. – 56 с.

ISBN 978-5-6046979-6-2

Рецензент: Мерабишвили Вахтанг Михайлович, д. м. н., профессор, председатель научно-медицинского Совета по развитию информационных систем онкологической службы Северо-Западного Федерального округа Российской Федерации, заведующий научной лабораторией онкологической статистики федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр онкологии имени Н.Н. Петрова» Министерства здравоохранения Российской Федерации

В учебном пособии подробно освещены вопросы, касающиеся анализу выживаемости. Анализ выживаемости является одной из стандартных задач клинических исследований в области клинической онкологии. Несмотря на обширную литературу по математической статистике, у многих исследователей возникают проблемы самостоятельного практического применения методов анализа выживаемости в доступных статистических пакетах. Настоящее учебно-методическое пособие ориентировано на аспирантов, клинических ординаторов и врачей, начинающих свою научную деятельность в клинической онкологии. Пособие содержит пошаговые инструкции по использованию популярных статистических пакетов MedCalc и Statistica для анализа выживаемости. Выполнение практических рекомендаций, изложенных в пособии, позволит избежать типичных ошибок, встречающихся в публикациях и диссертационных работах.

Учебное пособие предназначено для врачей-исследователей, врачей-онкологов, для врачей, работающих с онкологическими больными, а также для обучающихся в системе высшего образования (аспирантура, ординатура, специалитет) и дополнительного профессионального образования (повышение квалификации и профессиональная переподготовка).

Утверждено
в качестве учебного пособия
Ученым советом ФГБУ «НМИЦ онкологии
им. Н.Н. Петрова» Минздрава России
протокол № 11 от 20 сентября 2022 г.

ISBN 978-5-6046979-6-2

СОДЕРЖАНИЕ

Введение

«В каждом отделе естествознания есть
лишь столько настоящей науки, сколько в нем математики»
И. Кант. Метафизические основы естествознания, 1786 г.

Современные естественные науки, включая медицину, невозможно представить без математической статистики. Целый ряд статистических методов, первоначально разработанных для решения технических и финансовых задач, нашли широкое применение в медицинских исследованиях и сегодня являются «золотым стандартом» доказательной медицины. Возросли требования к статистическому обоснованию работ, направляемых в медицинские журналы и на рассмотрение диссертационных советов. Вместе с тем, появились специализированные программные пакеты, охватывающие широкий спектр статистических методов, адекватное применение которых требует достаточного понимания их возможностей и ограничений.

Общение с аспирантами и клиническими ординаторами при подготовке научных публикаций и рецензирование диссертаций, поступающих в диссертационный совет, показали, что анализ выживаемости онкологических пациентов является одной из типичных и наиболее сложных задач в клинических исследованиях.

Целью настоящего пособия является помощь врачам-исследователям, имеющим лишь начальную подготовку в области математической статистики, в освоении и адекватном практическом применении модулей анализа выживаемости в популярных статистических пакетах MedCalc и Statistica. Соблюдение рекомендаций, изложенных в пособии, позволит избежать типичных ошибок, встречающихся в публикациях и диссертационных работах.

Следует подчеркнуть, что в задачи настоящего пособия не входит описание математических аспектов применяемых статистических методов, которые настоятельно рекомендуется изучить самостоятельно.

Глава 1.
Популярные статистические пакеты

Среди множества статистических программ и пакетов, в медицинских исследованиях наиболее часто используются MedCalc, Statistica, SPSS, R.

MedCalc – простой для освоения и использования пакет специализированных программ для современного анализа биомедицинских данных. Достоинством MedCalc является удобство проведения логистической регрессии, построения и сравнения ROC-кривых. В программе реализованы наиболее удобные инструменты анализа выживаемости: метод Каплана-Мейера и регрессия Кокса. Основным недостатком является сложность экспорта графической информации в текстовый редактор. Также затруднен механизм изменения получаемых графиков. Статистический пакет доступен в интернете, нетребователен к аппаратным ресурсам.

Statistica – универсальный пакет программ для статистического анализа в самых разных сферах деятельности. Имеет удобную систему ввода, редактирования и вывода результатов. Главным недостатком пакета является сложность проведения ROC-анализа, для которого необходимо предварительное создание виртуальной нейронной сети. Это значительно повышает точность расчетов, но требует специальной подготовки исследователя.

SPSS (Statistical Package for the Social Sciences) – мощный инструмент для социологических исследований, который широко используется и для решения медико-биологических задач. SPSS обладает развитой системой анализа количественных данных и проведения ROC-анализа, мощной системой ввода и редактирования исходных данных, а также экспорта полученных текстовых, табличных и графических результатов в текстовые и табличные редакторы. Основными недостатками являются сложность программы, требующая специального

обучения и ограниченная доступность бесплатных версий.

Среда статистических вычислений R обладает очень широкими возможностями, работает на платформах Windows и Linux. Имеет собственный язык программирования. Исходные тексты и бинарные модули доступны в сети репозитариев CRAN. Требует серьезных знаний в области математической статистики и владения языком программирования R.

Учитывая достоинства и недостатки перечисленных статистических пакетов, а также интуитивную понятность и простоту, начинающему исследователю мы рекомендуем для анализа выживаемости использовать популярные пакеты статистических программ MedCalc и Statistica, которым и будет уделено внимание в настоящем пособии.

Глава 2.
Общие рекомендации по созданию базы данных и её подготовке к экспортированию в статистические пакеты

Для создания баз данных существует множество программных продуктов.

Мы остановим свое внимание на Microsoft Office Excel, поскольку большинство начинающих исследователей используют именно этот табличный процессор для формирования базы в виде одной, иногда обширной и труднообозримой таблицы.

В простейшем виде база данных в формате Excel представляет собой таблицу, в которой каждая запись (строка) соответствует одному больному, а столбцы (колонки) содержат значения переменных. Максимальный размер таблицы Microsoft Office Excel 2019 составляет 1048576 (220) строк и 16384 (214) колонки.

Шапка таблицы создается в первой строке и содержит имена переменных, индивидуальные значения которых заносятся в столбцы. Рекомендуется использовать с этой целью только первую строку, поскольку использование второй и третьей строк, а также слияние ячеек в шапке таблицы может затруднить сортировку и экспорт данных в статистические пакеты.

Переменным следует присваивать краткие имена на английском или латинском языке, так как не все статистические программы, которые могут быть использованы в дальнейшем, поддерживают кириллицу.

Список переменных и их подробную расшифровку рекомендуется создать в той же книге на отдельном листе Excel.

В первом столбце таблицы указывается уникальный идентификационный номер (ID) пациента: это может быть порядковый номер, номер истории болезни, а лучше – номер амбулаторной карты, который позволяет быстро найти больного в корпоративной базе учреждения («Виста»).

Второй столбец таблицы должен содержать фамилию, имя и отчество пациента в точном паспортном написании, что позволит в последствии при необходимости сличить записи создаваемой базы данных с базами других организаций (канцеррегистр, ЗАГС и т.п.).

В третьем столбце следует указать год и дату рождения. Еще одним обязательным столбцом является половая принадлежность пациента, которая должна быть категориальной переменной и обозначена цифрой.

Использование букв (текстовая переменная) затрудняет последующую обработку данных в статистических пакетах.

В четвертом столбце таблицы можно указать возраст больного к интересующей нас дате: это может быть дата начала заболевания, обращения к врачу, дата операции или дата начала лечения. Столбец не является обязательным, поскольку возраст может быть легко вычислен по датам рождения и интересующей нас дате, но удобен для анализа.

Следующей категориальной переменной может быть половая принадлежность, если в исследование включены оба пола.

Прочие необходимые сведения о больном, включая адрес регистрации, контактные телефоны и электронные адреса также могут быть внесены в основную таблицу или сохранены в дополнительной таблице на отдельном листе, что избавит в дальнейшем от необходимости обращаться в «Висту» за дополнительными сведениями.

Полный набор переменных определяется задачами исследования, однако следует предусмотреть наличие свободно заполняемых текстовых полей, куда можно заносить любые неструктурированные записи, которые могут потребоваться в дальнейшем.

При заполнении полей в таблице следует придерживаться следующих правил:

- Не использовать одинаковые имена для различных переменных.

- Не ставить нули вместо отсутствующих данных.
- Не использовать для одной переменной (одного столбца) различные форматы ячеек: например, формат даты и цифровой или текстовый формат.

- Не использовать числовой формат ячейки при введении ранговых или категориальных переменных (например, степени дифференцировки опухоли, категории Т, цифрового обозначения локализации опухоли, кода диагноза или вида лечения). В случае цифрового формата в некоторых статистических пакетах код лечения 5 будет учитываться при расчетах, как дающий пятикратный вклад по сравнению с кодом лечения 1.

- В ряде случаев полезно разукрупнение категориальных переменных в бинарные, например, категориальная переменная «гистологический тип опухоли» разбивается на несколько бинарных переменных: «аденокарцинома» 1/0, «плоскоклеточный рак» 1/0, «мелкоклеточный рак» 1/0 и «прочие раки» 1/0. Это увеличивает таблицу, но упрощает сортировку и анализ данных, а также построение регрессионных моделей.

После завершения сбора данных рекомендуется защитить от редактирования лист, содержащий базу, и создать резервную копию файла книги Excel для хранения вне компьютера.

Для дальнейшей работы рекомендуется скопировать защищенный лист с базой данных на новый лист в той же книге Excel и дальнейшее редактирование (добавление или удаление данных, строк и столбцов) и сортировку проводить в копии базы.

Следует иметь в виду, что в старых версиях Excel сортировка

строк таблицы осуществлялась до пустого (не имеющего данных) столбца. Если это оставалось незамеченным, дальнейшая работа с таблицей приводила к неверным результатам.

Следует также иметь в виду, что при экспорте электронных таблиц в некоторые статистические программы пустые строки таблицы могут заполняться нулями.

Для подготовки данных к экспорту в статистические программы рекомендуется удалить текстовые поля (не является обязательным для последних версий программы Statistica) и еще раз проверить правильность формата переменных.

Глава 3.
Выживаемость

Выживаемость (Survival) – это доля выживших к моменту времени t от начала наблюдения t0.

Выживаемость связана со смертностью (Mortality):

Survival = 1 - Mortality

В медицинских исследованиях сложность оценки выживаемости состоит в наличии цензурированных наблюдений (т.е. незавершенных из-за ограничения продолжительности наблюдения или потери контакта с больным). Современные методы анализа выживаемости позволяют с достаточной точностью учитывать эффект цензурирования.

В клинических исследованиях обычно используют следующие виды выживаемости:

Overall Survival (OS) – наблюдаемая (общая) выживаемость, при расчете которой учитываются летальные исходы от всех причин.

Disease-dependent Survival (DDS) – скорректированная выживаемость, при расчете которой учитываются летальные исходы только от изучаемого заболевания. Частный случай – Cancer-dependent Survival.

Event Free Survival (EFS) – бессобытийная выживаемость. Законченным наблюдением считают одно или несколько предусмотренных в исследовании событий (прогрессирование, рецидив, появление отдаленных метастазов, смерть от лечения и т.д.).

Relapse Free Survival (RFS) – безрецидивная выживаемость (законченным наблюдением считают выявление рецидива), оценивается для пациентов, достигших полной ремиссии.

Disease Free Survival (DFS, DFI) – период, свободный от болезни (законченным наблюдением считают прогрессирование).

Глава 4.
Подготовка данных для анализа выживаемости в Excel

Для анализа выживаемости в экспортируемом файле формата *xlsx должны быть переменные, одна из которых содержит данные о длительности наблюдения (например, SURV), и вторая цензурирующая переменная (CENS), информирующая о событии (завершенном либо незавершенном наблюдении, например, о смерти больного или развитии рецидива). Если пациент к дате последнего контакта был жив, а далее выбыл из прослеживания, наблюдение считается цензурированным.

Для формирования столбца длительности наблюдения (SURV) можно использовать дату начала наблюдения (например, дату операции OP_DATE) и дату окончания наблюдения (например, дату последнего контакта с больным LAST_DATE).

Чаще всего интервал между этими датами вычисляют в месяцах, тогда SURV = (LAST_DATE – OP_DATE)/30,42.

Знаменатель формулы вытекает из деления количества дней в году (365) на 12 месяцев. Если необходима выживаемость в неделях, знаменатель формулы будет равен 7.

В Excel эти вычисления можно провести автоматически, записав формулу в ячейки столбца SURV, однако следует тщательно проверить получившиеся длительности наблюдения, поскольку при отсутствии даты начала наблюдения в соответствующей ячейке (например, даты операции) разница дат может вычисляться как SURV = (LAST_DATE – 01.01.1900). Перед экспортированием файла рекомендуется сохранить столбец SURV в виде значений (меню Excel: копировать, вставка, специальная вставка, значения), хотя в последних версиях статистических пакетов это не является обязательным.

Для сравнения выживаемости в двух или нескольких группах экспортируемая таблица должна содержать группирующую переменную GROUP (столбец, указывающий на принадлежность наблюдения к определенной группе).

Глава 5.
Анализ выживаемости:
метод Каплана-Мейера

Существуют различные методы оценки выживаемости.

Оценка функции выживания – т.е. определение вероятности того, что пациент проживет определенный срок после начала лечения, с учетом цензурирования может быть достигнута без создания таблиц дожития методом Каплана-Мейера[1].

Отметим, что для получения информативных показателей выживаемости желательно иметь не менее 30 наблюдений.

Следует учитывать, что метод Каплана-Мейера использует следующие допущения:

- цензурированные объекты (выбывшие) имеют ту же выживаемость, что и объекты, которые продолжают наблюдаться;
- оценки выживаемости одинаковы для объектов, включенных в исследование на ранних и более поздних сроках;
- событие происходит именно в анализируемый момент времени.

Последнее предположение может искусственно завысить оценку выживаемости, если измерения производятся редко, так как определение момента наступления события откладывается до следующего обследования.

[1] Kaplan E. L. & Meier P. Nonparametric estimation from incomplete observations //J. Amer. Statist. Assn. – 1958. – Vol. 53. - P. 457-481 (Univ. California Radiation Laboratory, CA and University of Chicago, IL)

Глава 6.
Анализ выживаемости методом Каплана-Мейера в MedCalc

На рисунках 1-3 проиллюстрирована оценка выживаемости методом Каплана-Мейера в MedCalc.

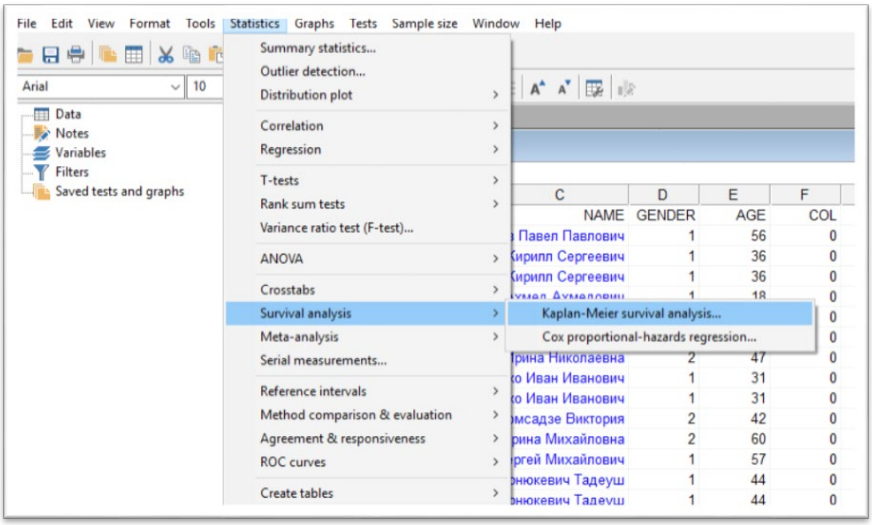

Рис. 1. Оценка выживаемости методом Каплана-Мейера в MedCalc. Путь к модулю: Statistics/Survival analysis/Kaplan-Meier survival analysis.

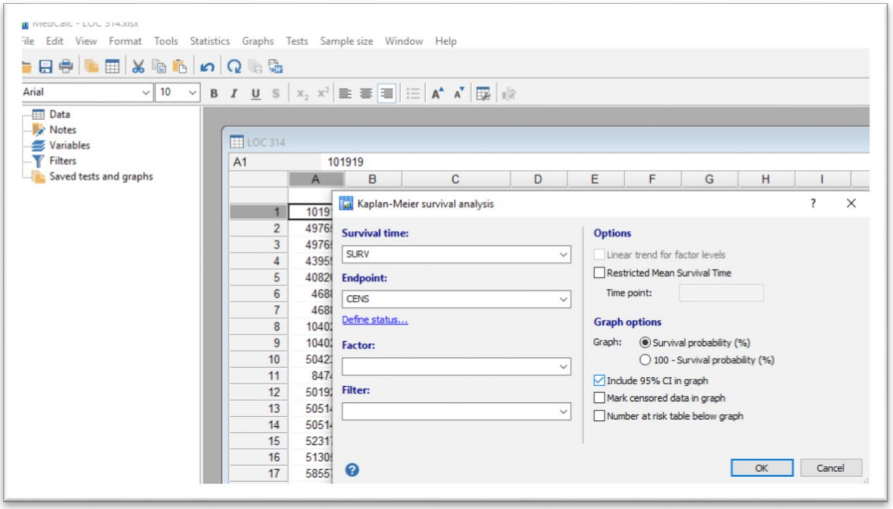

Рис. 2. Оценка выживаемости методом Каплана-Мейера в MedCalc. Окно выбора переменных. Endpoint – выбор цензурирующей переменной.

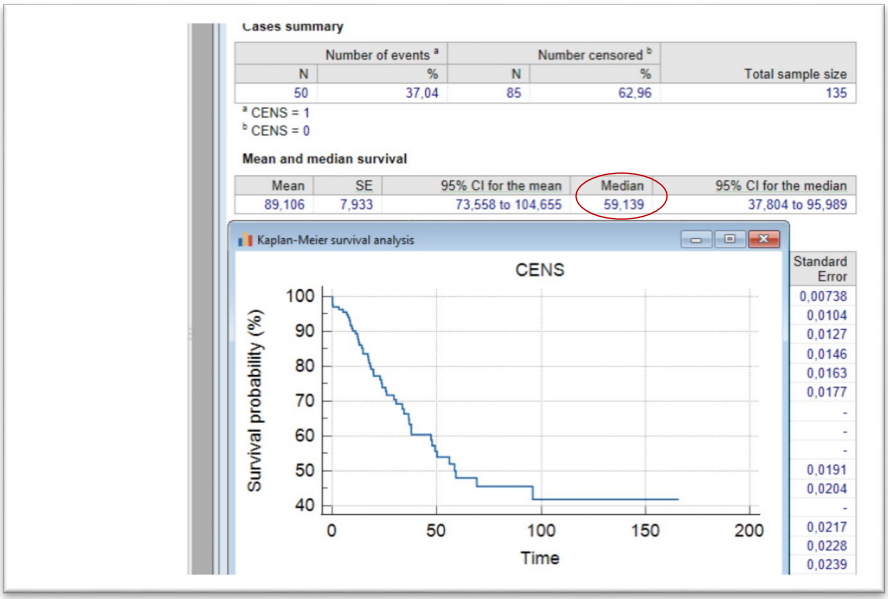

Рис. 3. Оценка выживаемости методом Каплана-Мейера в MedCalc. График кривой кумулятивной выживаемости. Медиана выживаемости и её доверительный интервал представлены в разделе отчета Mean and median survival.

Глава 7.
Анализ выживаемости
методом Каплана-Мейера в Statistica

Пакет Statistica представляет более широкие возможности для исследования выживаемости (рис. 4-10).

Модуль Survival (Survival and Failure Time Analysis) предоставляет выбор: таблицы выживаемости, метод Каплана-Мейера, сравнение выживаемости в двух и нескольких выборках, построение регрессионных моделей.

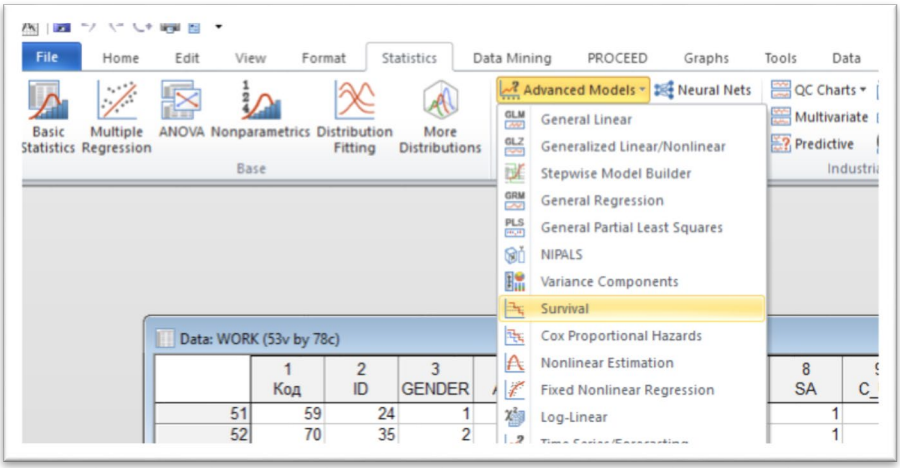

Рис. 4. Оценка выживаемости в Statistica. Путь к модулю: Statistics/Advanced Models/Survival.

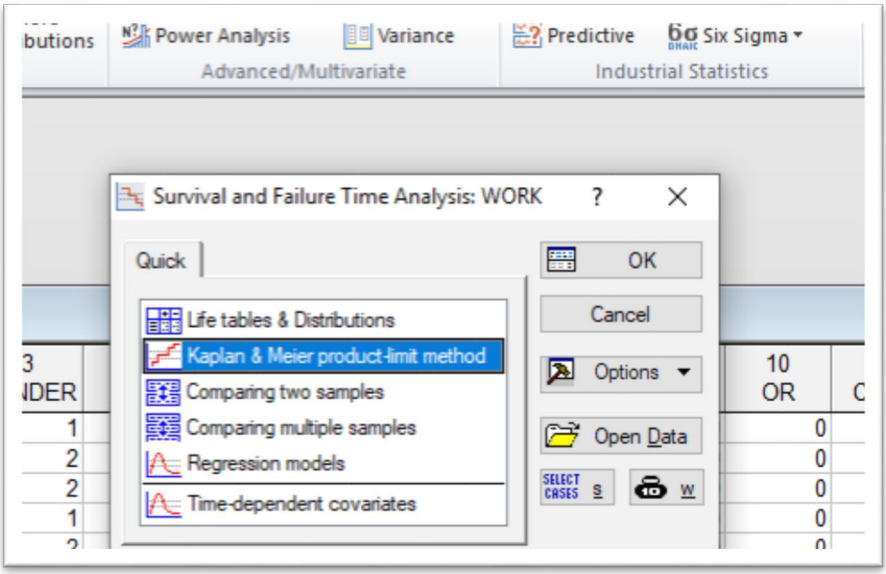

Рис. 5. Оценка выживаемости методом Каплана-Мейера в MedCalc. Модуль Survival and Failure Time Analysis – анализ выживаемости и времени наработки на отказ.

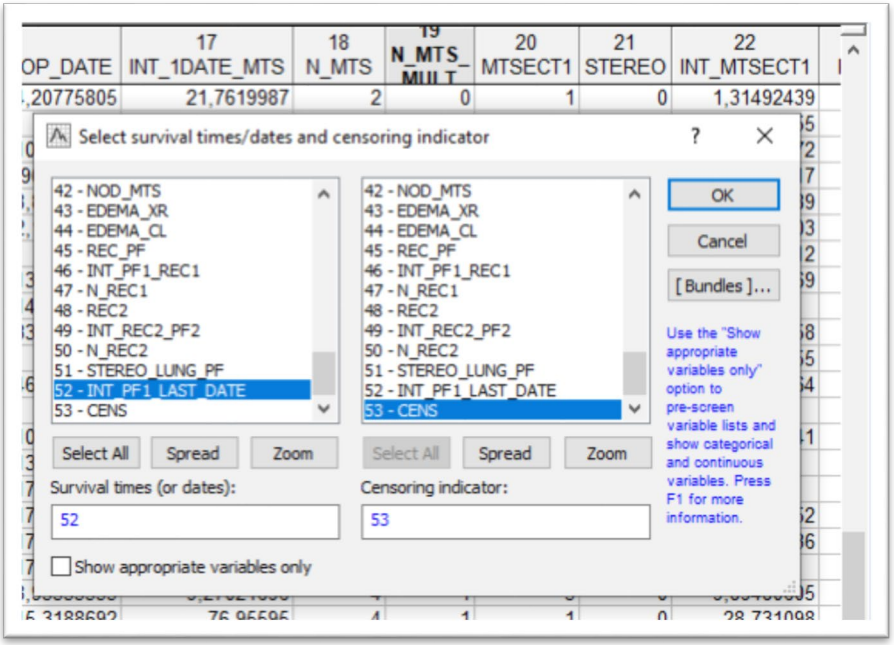

Рис. 6. Оценка выживаемости методом Каплана-Мейера в Statistica. Окно ввода переменных.

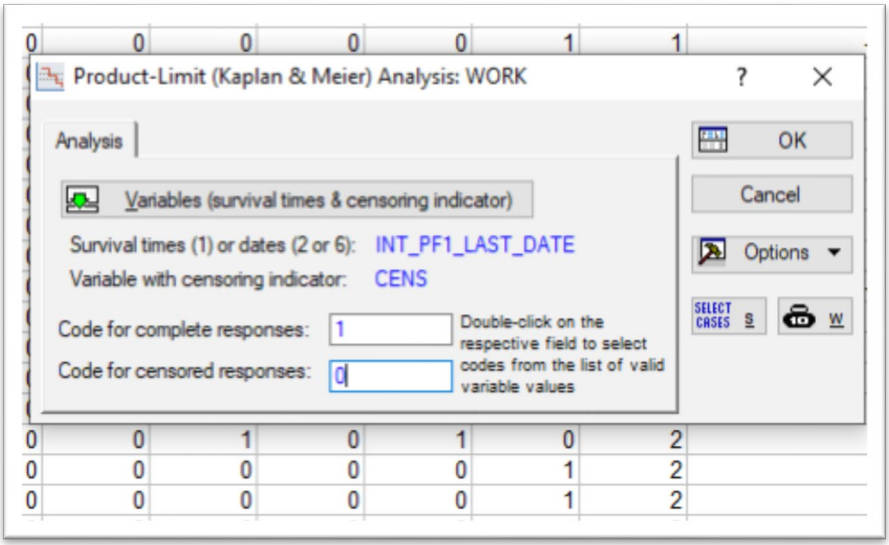

Рис. 7. Оценка выживаемости методом Каплана-Мейера в Statistica. Окно ввода цензурирующей переменной.

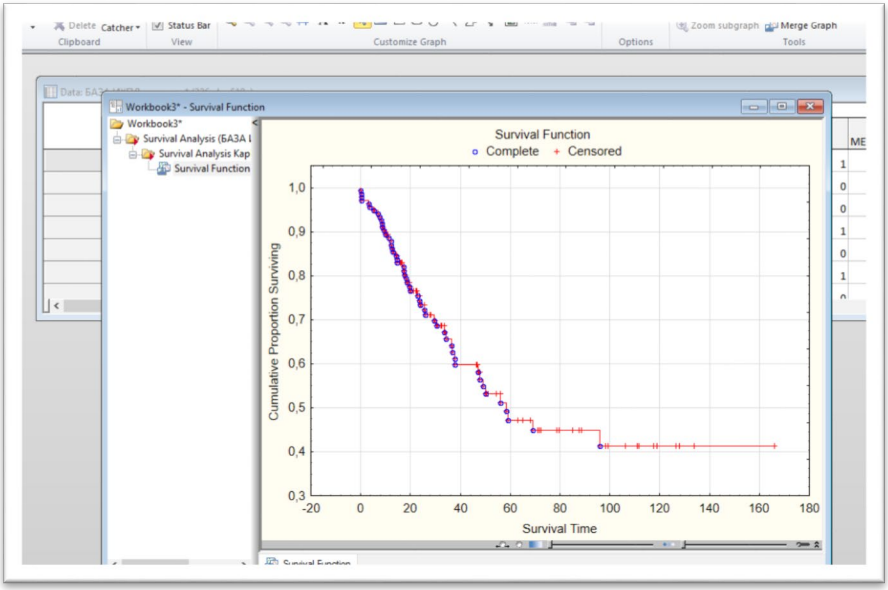

Рис. 8. Оценка выживаемости методом Каплана-Мейера в Statistica.

График кривой кумулятивной выживаемости. Медиана выживаемости и её стандартная ошибка представлены в окне отчета о кумулятивной выживаемости (рис. 9) и разделе отчета Percentiles (рис. 10).

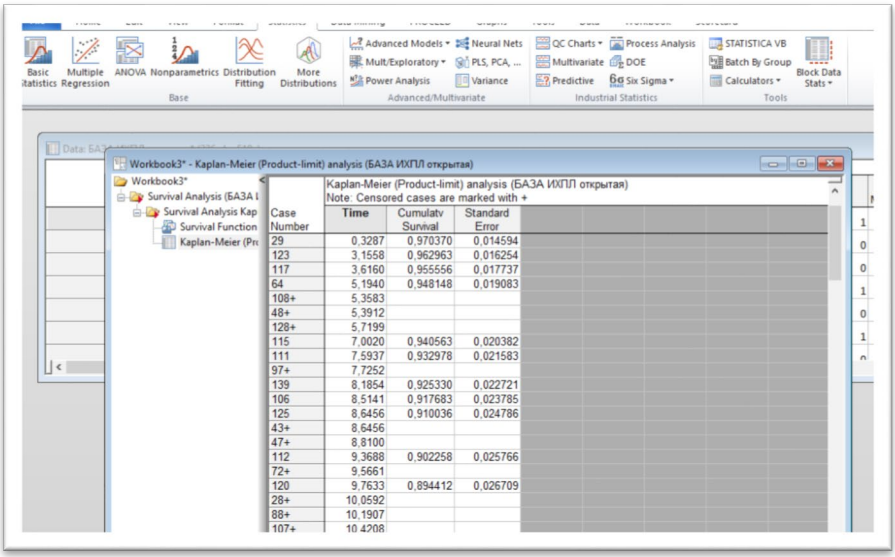

Рис. 9. Оценка выживаемости методом Каплана-Мейера в Statistica. Окно отчета о кумулятивной выживаемости, в котором можно найти медиану выживаемости (Cumulative Survival = 0,5) и её стандартную ошибку.

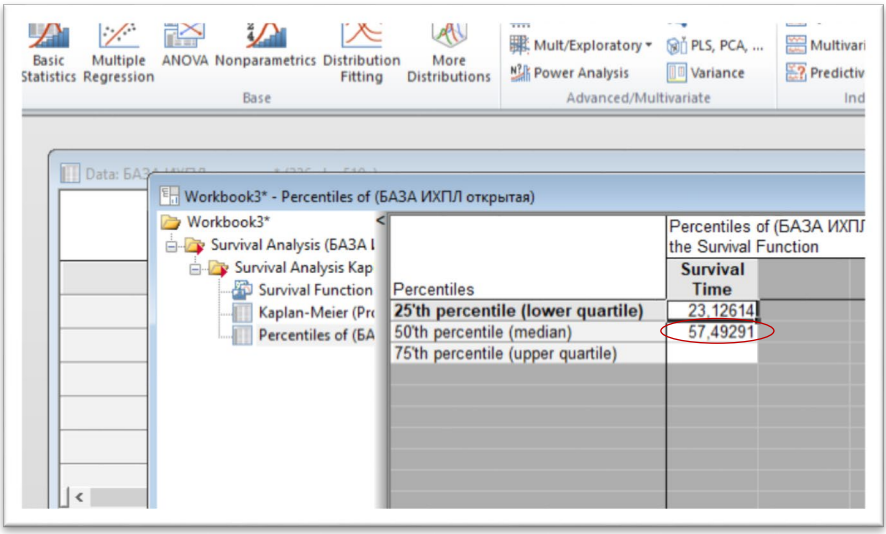

Рис. 10. Оценка выживаемости методом Каплана-Мейера в Statistica. Окно отчета о медиане и квартилях кумулятивной выживаемости.

Глава 8.
Сравнение выживаемости двух или нескольких групп в Statistica

Сразу же следует подчеркнуть, что MedCalc не предоставляет возможности сравнения выживаемости в двух или нескольких группах. В пакете Statistica это легко реализуется (рис. 11-17).

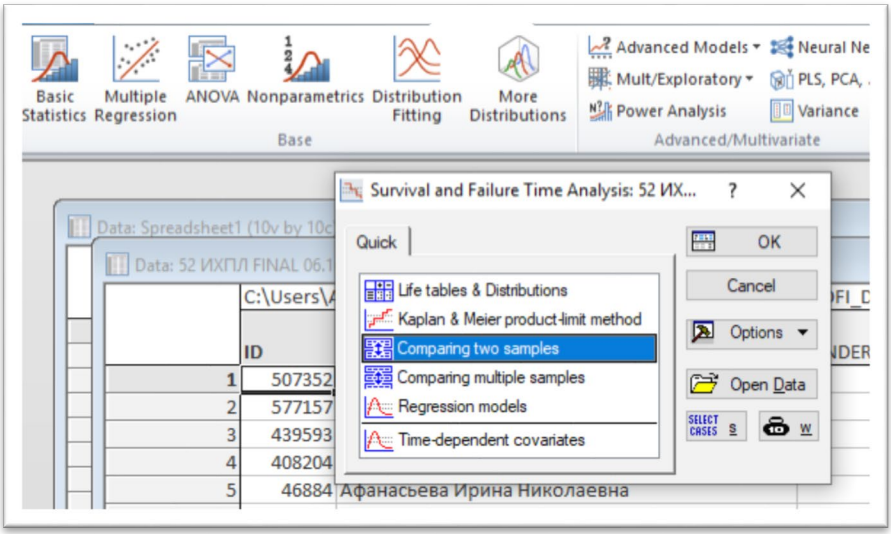

Рис. 11. Сравнение выживаемости двух групп в Statistica. Путь к модулю: Statistics/Advanced Models/Survival/Comparing two samples.

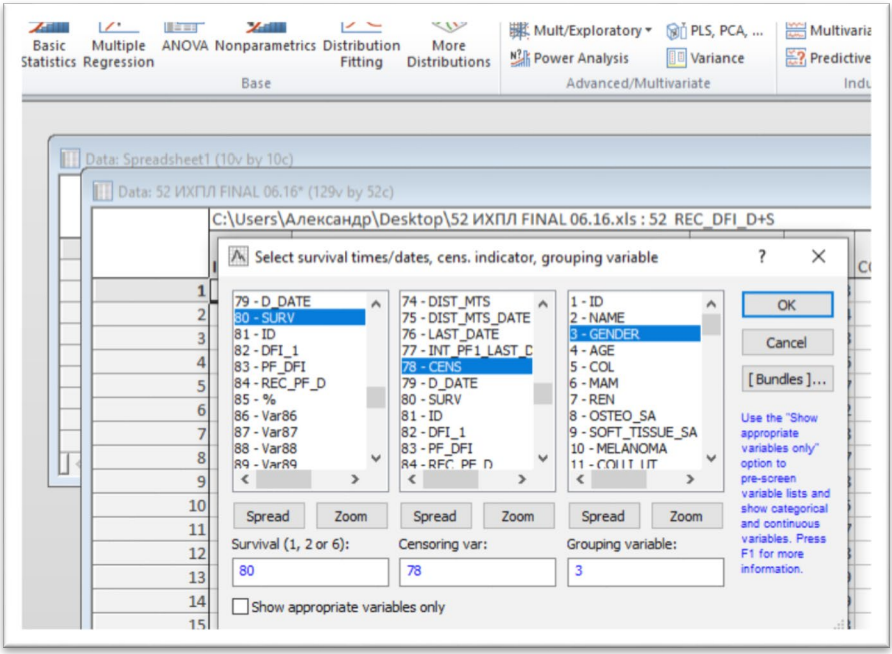

Рис. 12. Сравнение выживаемости двух групп в Statistica. Окно выбора переменных.

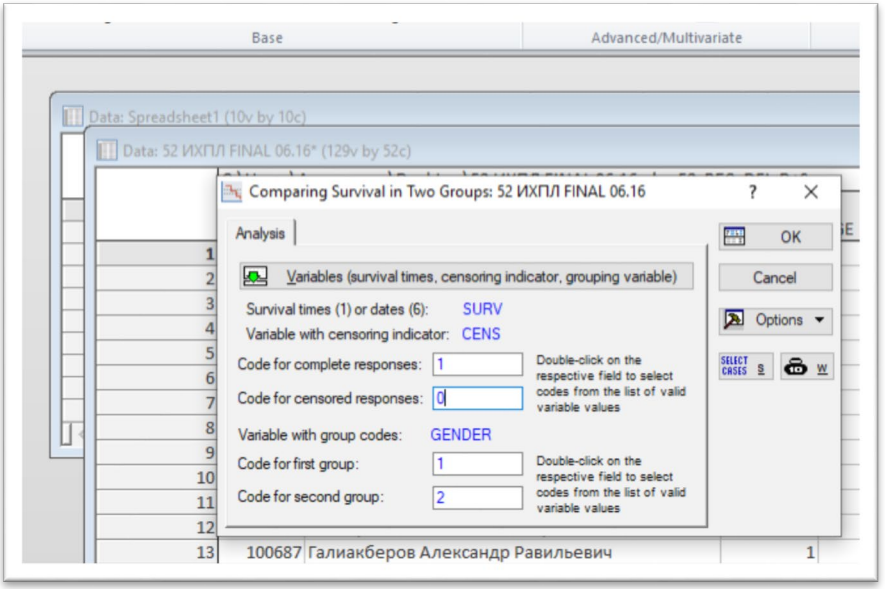

Рис. 13. Сравнение выживаемости двух групп в Statistica. Окно выбора кодов цензурирующей и группирующей переменных.

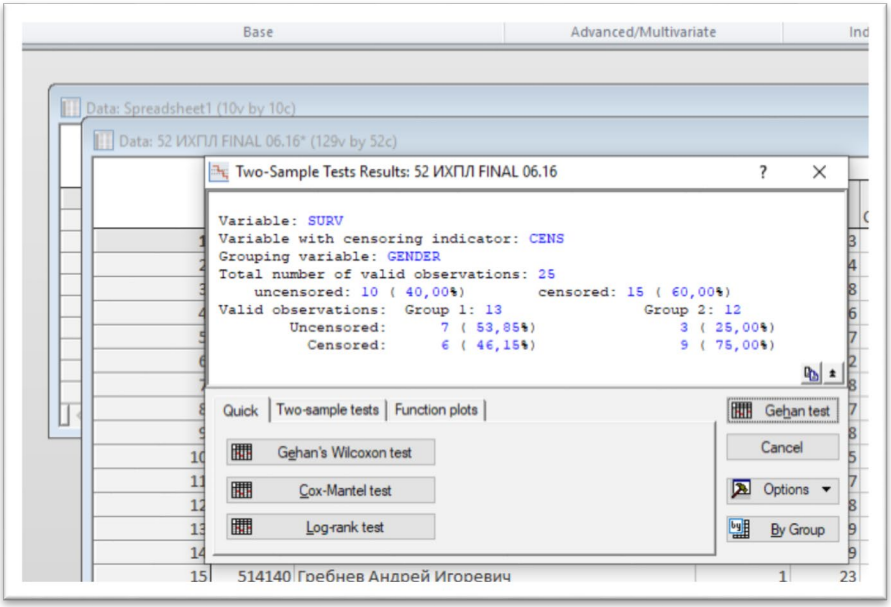

Рис. 14. Сравнение выживаемости двух групп в Statistica. Окно выбора теста для выявления различий выживаемости.

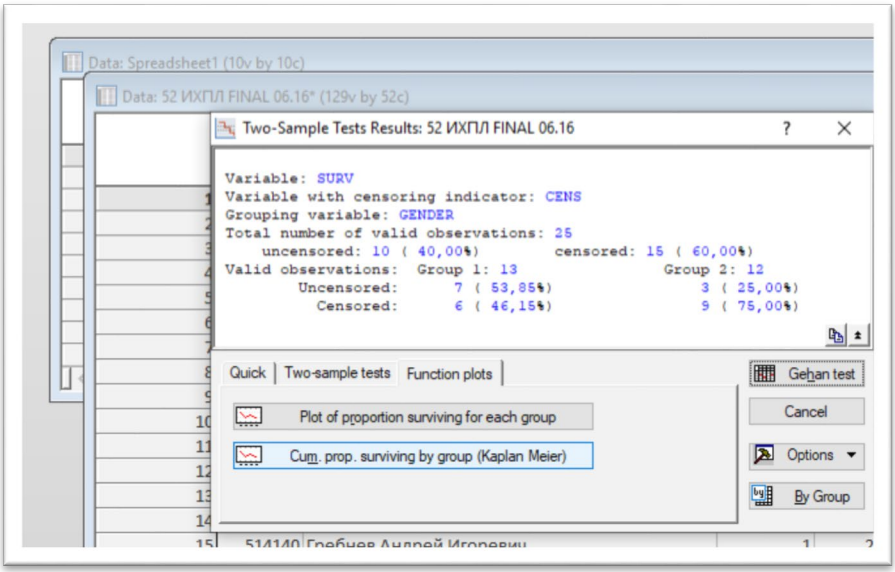

Рис. 15. Сравнение выживаемости двух групп в Statistica. Окно выбора вывода графического представления результатов.

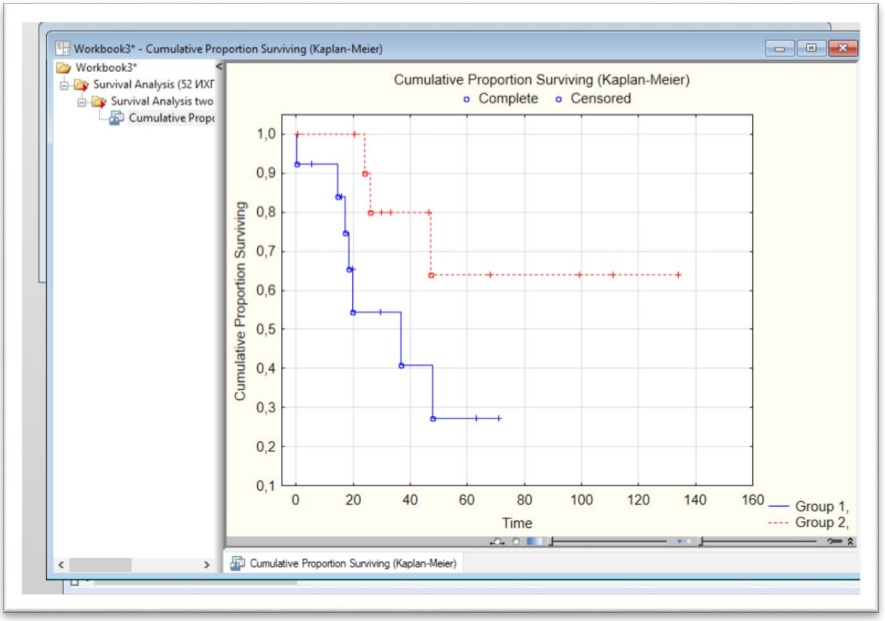

Рис. 16. Сравнение выживаемости двух групп в Statistica. Графики кривых кумулятивной выживаемости Каплана-Мейера в группах.

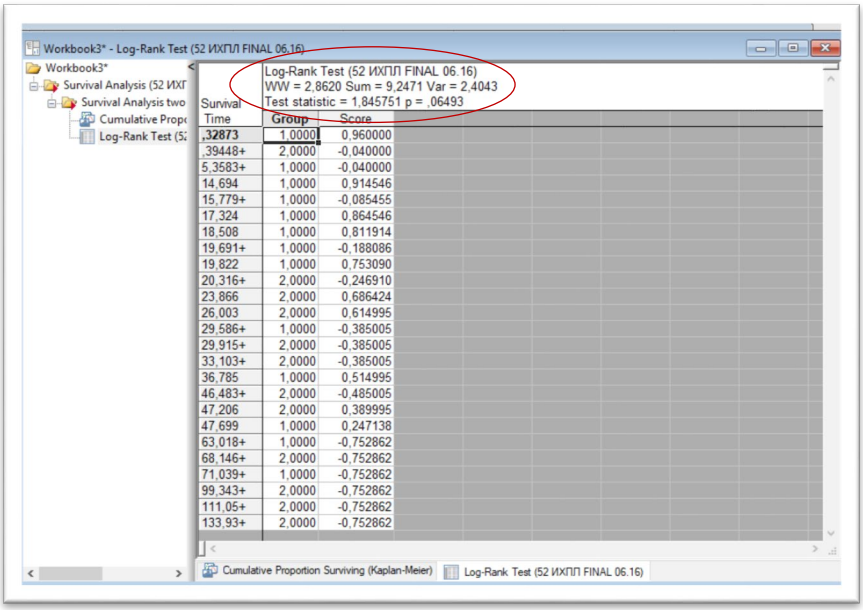

Рис. 17. Сравнение выживаемости двух групп в Statistica с использованием Log-Rank теста. Окно вывода результатов.

Сравнение выживаемости нескольких групп в пакете Statistica проводится в модуле Statistics/Advanced Models/Survival/Comparing multiple samples (рис. 11-17). Вводятся группирующая и цензурирующая переменные, программа строит на одном графике кривые выживаемости каждой из групп. Однако для получения оценок межгрупповых различий необходимо дополнительно выполнить попарное сравнение и оценку кривых выживаемости в модуле Statistics/Advanced Models/Survival/Comparing two samples.

Глава 9.
Анализ выживаемости: модель пропорциональных рисков Кокса

Модель Кокса (Cox proportional hazards model) является наиболее часто используемым современным инструментом анализа выживаемости. Регрессионная модель Кокса применяется для изучения влияния независимых переменных (predictor variables), рассматриваемых как факторы риска, на время наступления события (survival time). Модель прогнозирует риск наступления события (Hazard Risk) и оценивает влияние изучаемых переменных на величину риска. Hazard Risk является функцией, зависимой от времени. Никаких предположений о виде функции риска не делается, однако предполагается, что все переменные линейно влияют на логарифм функции риска: по указанным причинам метод является полупараметрическим.

9.1. Условия применения модели Кокса

1. Независимость предикторов, которая проверяется вычислением коэффициентов корреляции. Зависимость некоторых предикторов очевидна, например, размер опухоли и категория Т, линейный размер опухоли и её объем и т.д. Четких правил выбора одного из взаимозависимых предикторов не существует. Может быть построен целый ряд моделей с каждой из взаимозависимых переменных и затем выбрана наиболее точная регрессионная модель.

2. Отношение рисков не изменяется во времени (пропорциональность рисков). Например, если риск инфаркта миокарда в возрасте 50 лет у мужчин в два раза выше, чем у женщин, то он должен оставаться таким же в 60 лет или любом другом возрасте. Условие является главным для построения регрессионной модели Кокса *без зависимых от времени переменных*. Очевидно, что данное условие трудно проверяемо и выполняется далеко не всегда: даже в представленном примере риск развития инфаркта у мужчин и женщин выравнивается к возрасту 60-70 лет.

3. Моменты начала и окончания исследования (возникновения

события или окончания периода наблюдения) должны быть точно определены для каждого наблюдения.

4. Цензурированные и законченные наблюдения не должны отличаться по выживаемости друг от друга. Данное условие также труднопроверяемо. Во всяком случае, очевидные причины такого рода различий должны быть исключены.

5. Количество переменных, используемых для построения регрессионной модели, не должно превышать 1/10 числа наблюдений.

9.2. Выбор переменных для включения в модель Кокса

Существуют различные способы отбора переменных (предикторов) для включения в модель.

Рекомендуется для каждой числовой переменной сравнить кривые выживаемости в первых двух квартилях и в двух последних квартилях с помощью лог-ранк-критерия, т.е. вариационный ряд значений переменной разделяется на две равные части по медиане и сравнивается выживаемость в двух его половинах. Для бинарных переменных (например, пол) можно сравнить выживаемость в двух неравных частях.

Допускается также раздельно исследовать каждую переменную непосредственно в модели Кокса (унивариантный анализ).

В результате получаем список переменных с различными уровнями статистической значимости.

Для включения в мультивариантные модели Кокса отбираются переменные, получившие при унивариантном анализе уровни значимости $p < 0{,}1$.

Существует два способа построения моделей Кокса: способ пошагового добавления переменных и способ пошагового удаления переменных.

В первом случае в качестве основной берется переменная с наиболее высоким уровнем статистической значимости и к ней последовательно добавляются переменные с более низким уровнем значимости с оценкой на каждом шаге качества получаемых моделей.

Во втором случае в модель включаются сразу все переменные с p <0,1, оценивается качество модели и далее последовательно удаляются предикторы, получившие в модели p >0,05. На каждом следующем шаге вновь оценивается качество модели. В результате создается целый набор моделей, состоящих из различных комбинаций предикторов, из которого окончательно выбирается модель наилучшего качества.

Модель Кокса на выходе возвращает отношение рисков (Hazard Ratio, HR) и его доверительные интервалы для каждого из предикторов. Интерпретация Hazard Ratio: при HR = 2 пациент, доживший до некоторого момента времени имеет в 2 раза больший шанс умереть к следующему моменту времени.

Вероятность того, что событие наступит раньше, может быть рассчитана из показателя HR по формуле: p = HR/ (1 + HR). Таким образом, HR = 2 соответствует 67% шансу более раннего наступления события.

9.3. Модель Кокса в MedCalc

Рассмотрим модель Кокса в MedCalc (рис.18-24).

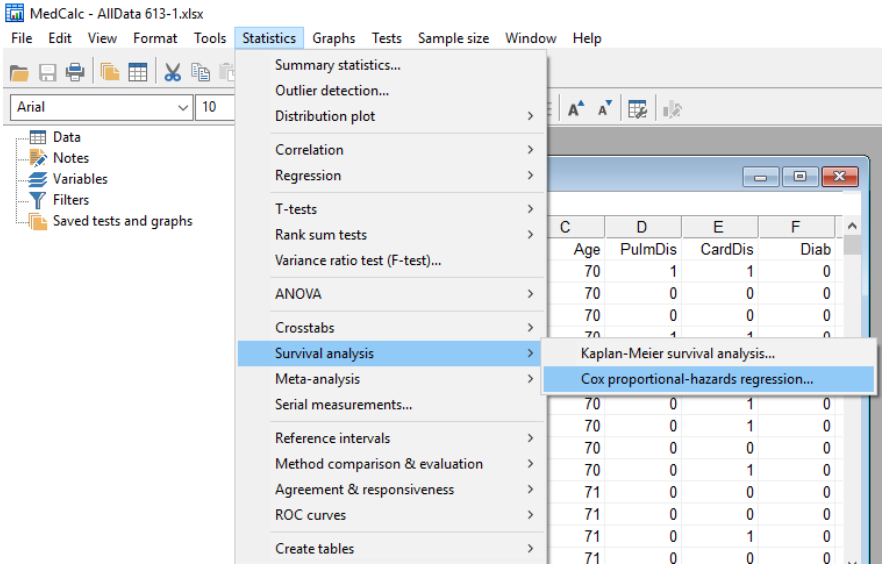

Рис. 18. Модель Кокса в MedCalc. Путь: Statistics/Survival analysis/Cox proportional-hazard regression…

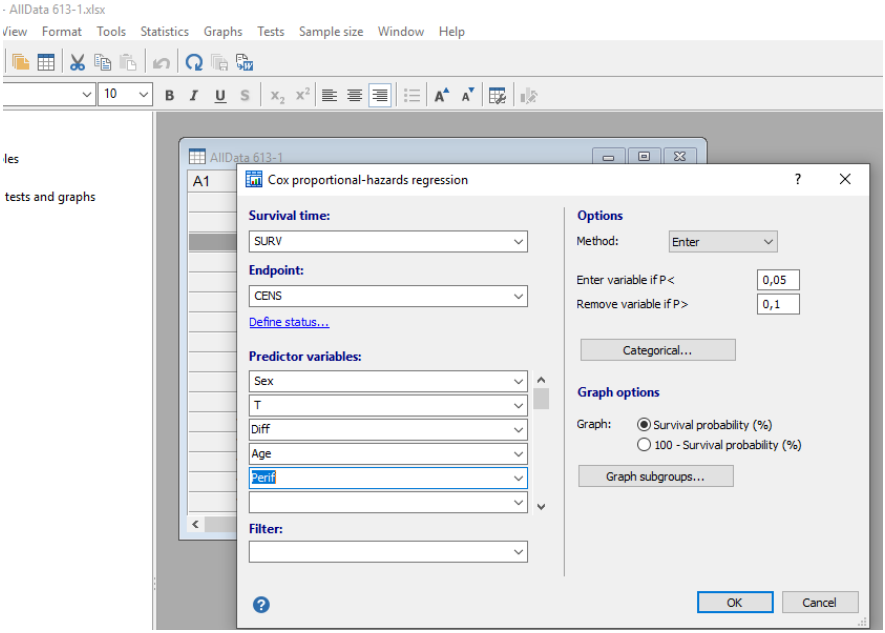

Рис. 19. Модель Кокса в MedCalc. Окно ввода переменных. Вводятся предикторы, имеющие p <0,1 при унивариантном анализе.

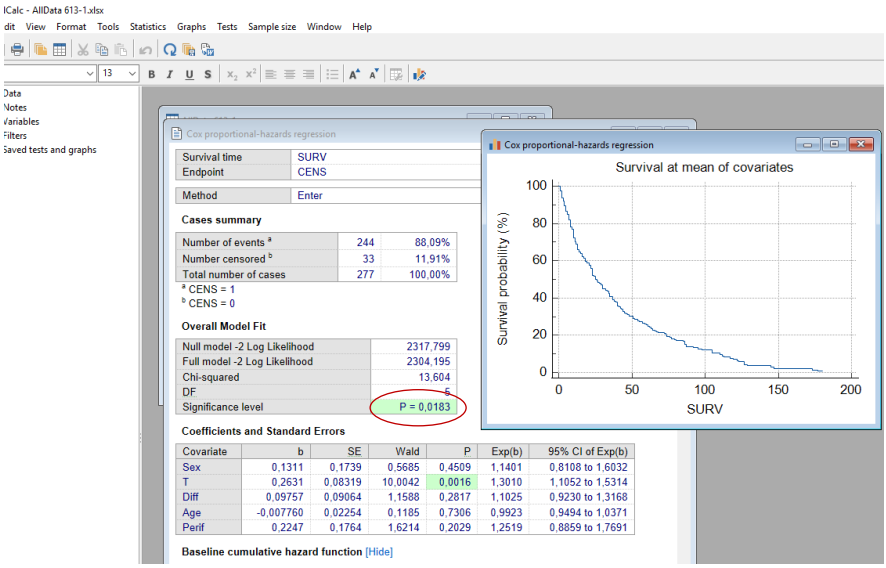

Рис. 20. Модель Кокса в MedCalc. Результат построения регрессионной модели (1) методом пошагового удаления предикторов, шаг 1: включены все значимые переменные. Качество модели: таблица Overall Model Fit, параметры модели: таблица Coefficients and Standard Errors. Уровень статистической значимости модели p = 0,0183.

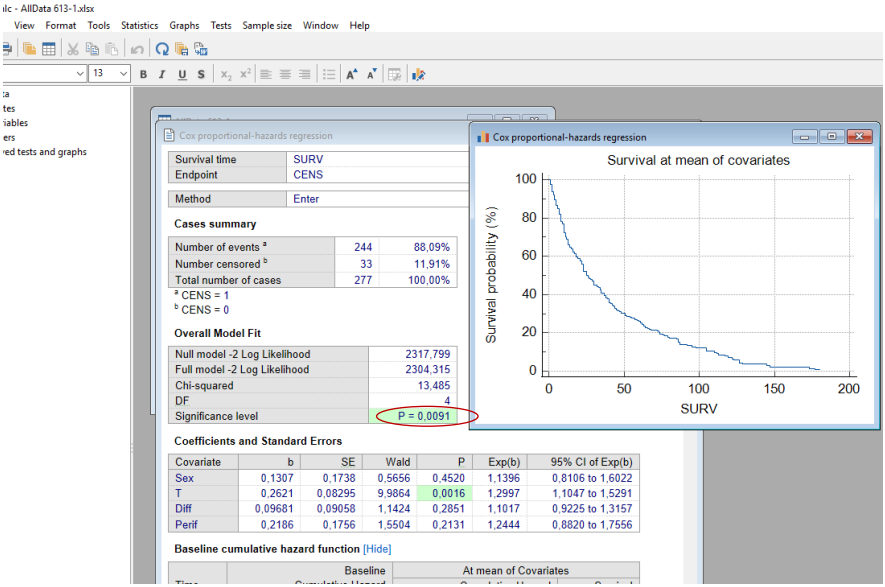

Рис. 21. Модель Кокса в MedCalc. Результат построения регрессионной модели (2) методом пошагового удаления предикторов, шаг 2: удаление переменной AGE, проявившей самый низкий уровень значимости. Уровень статистической значимости модели (2) p = 0,0091.

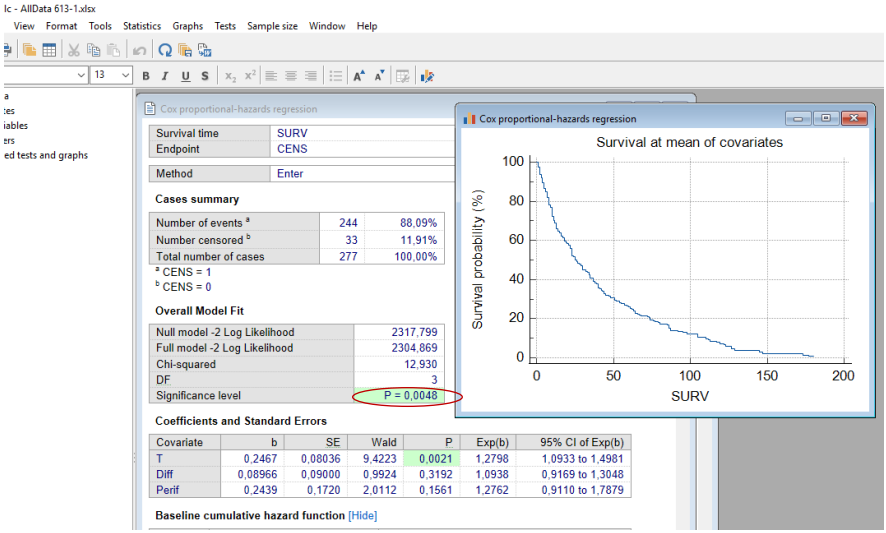

Рис. 22. Модель Кокса в MedCalc. Результат построения регрессионной модели (3) методом пошагового удаления предикторов, шаг 3: удаление переменной SEX. Уровень статистической значимости модели p = 0,0048.

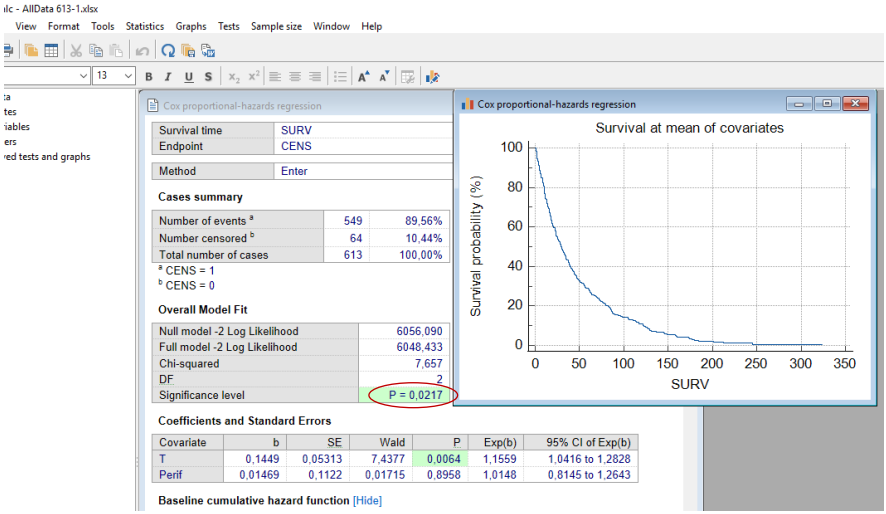

Рис. 23. Модель Кокса в MedCalc. Результат построения регрессионной модели (4) методом пошагового удаления предикторов, шаг 4: удаление переменной DIFF. Результат построения регрессионной модели (4). Уровень статистической значимости модели p = 0,0217.

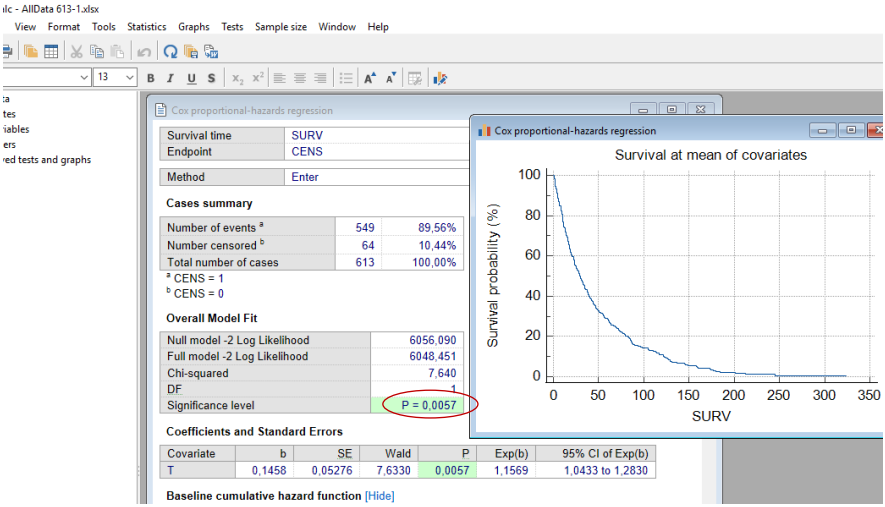

Рис. 24. Модель Кокса в MedCalc. Результат построения регрессионной модели (5) методом пошагового удаления предикторов, шаг 5: удаление переменной PERIF. Уровень статистической значимости модели p = 0,0057.

9.4. Выбор наилучшей из созданных моделей Кокса

Методом пошагового удаления переменных было построено пять моделей, объясняющих влияние предикторов на выживаемость.

Соответствие полученных моделей фактическим данным оценивается критерием хи-квадрат (Chi-squared) с учетом степеней свободы (DF).

Итоговым показателем качества модели является её уровень статистической значимости (Significance level).

Модель (3), изображенная на рис. 22, имеет наивысший уровень статистической значимости из пяти созданных моделей Кокса p = 0,0048.

9.5. Интерпретация модели (3)

Из модели (3) следует, что единственным значимым предиктором выживаемости является переменная T (p=0,0021), имеющая Exp(b) (Hazard Ratio) = 1,2798.

Включение в модель двух дополнительных предикторов Diff и Perif, оказывающих недостоверное влияние, улучшают качество модели.

9.6. Модель Кокса в Statistica

Построение моделей Кокса в пакете статистических программ Statistica выполняется аналогичным образом (рис. 25-38).

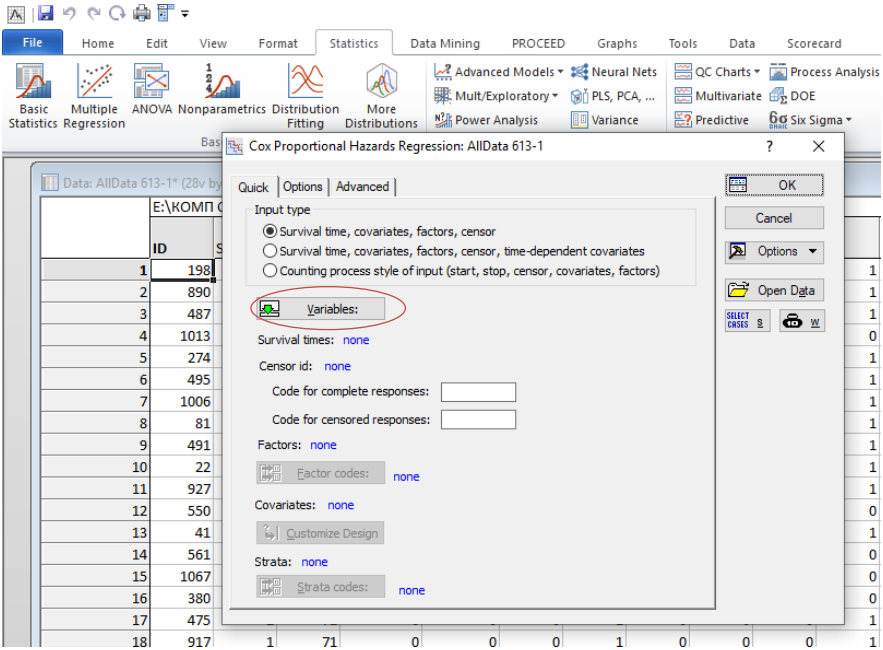

Рис. 25. Построение модели Кокса в Statistica. Путь к модулю: Statistics/Advanced models/Cox Proportional Hazards/ Окно введения переменных.

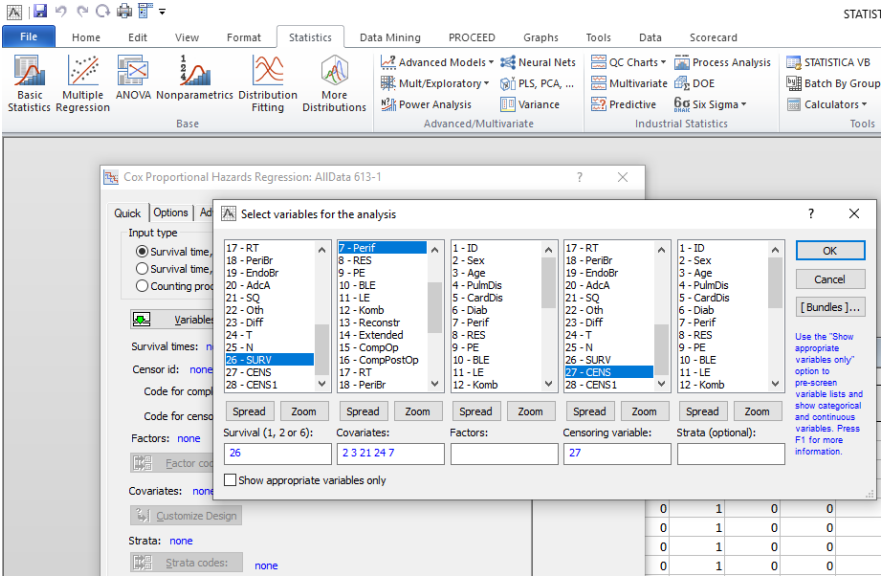

Рис. 26. Построение модели Кокса в пакете Statistica. Выбор переменных.

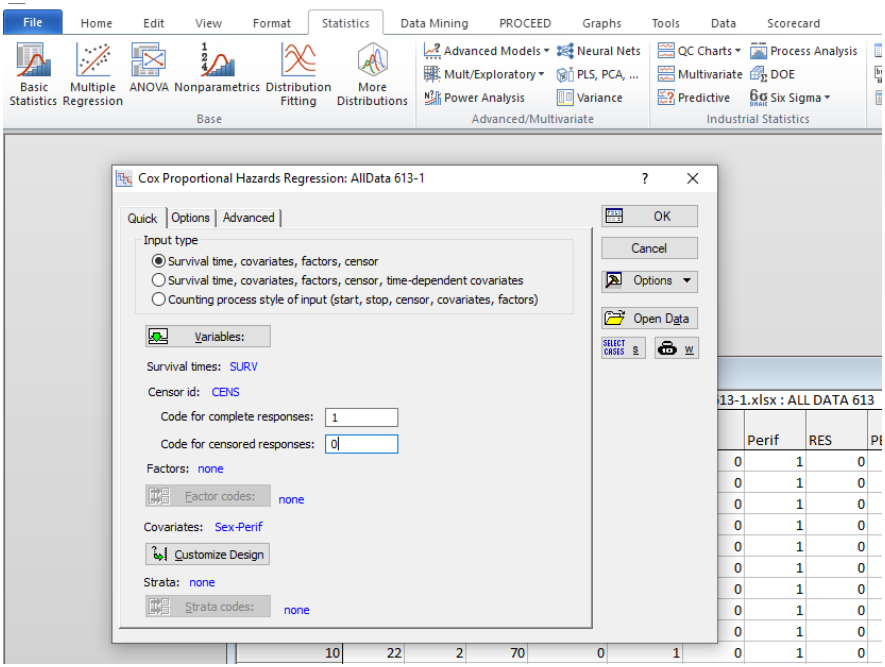

Рис. 27. Построение модели Кокса в Statistica. Введение цензурирующей переменной.

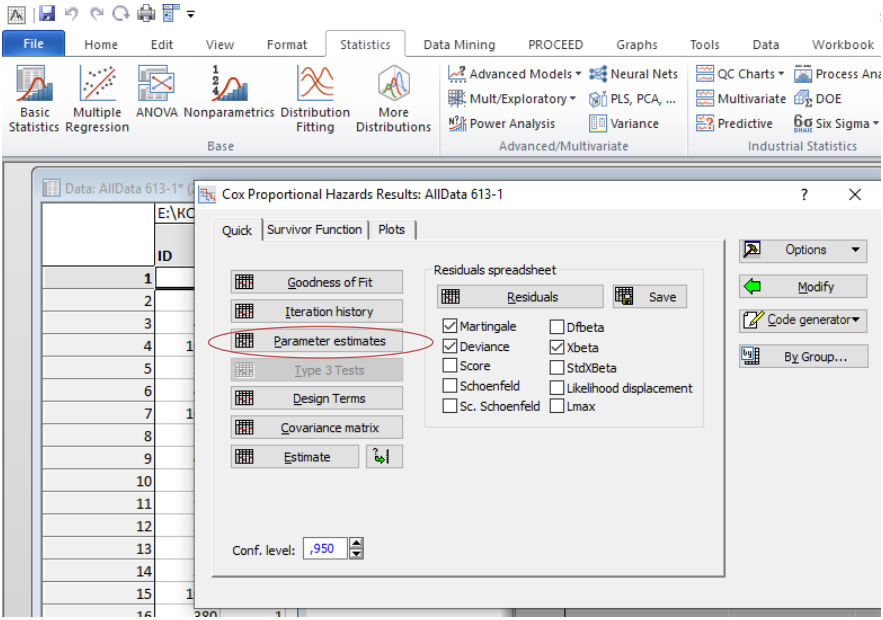

Рис. 28. Построение модели Кокса в Statistica. Выбор параметров оценки. Xbeta выводит в отчёте Hazard Ratio.

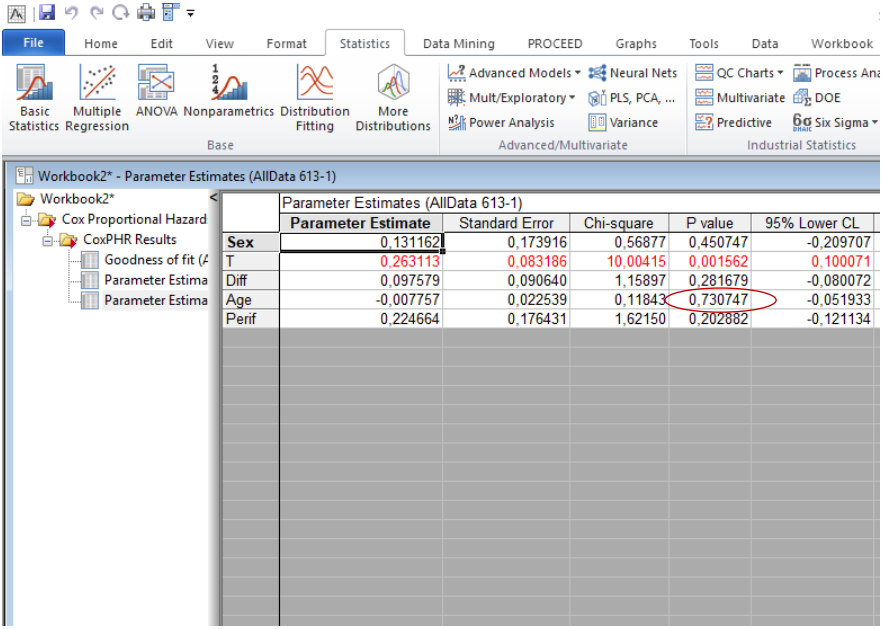

Рис. 29. Построение модели Кокса методом пошагового удаления предикторов в Statistica. Шаг 1. Регрессионная модель (1), включающая 5 предикторов. На втором шаге будет удален предиктор Age с наибольшим значением P value.

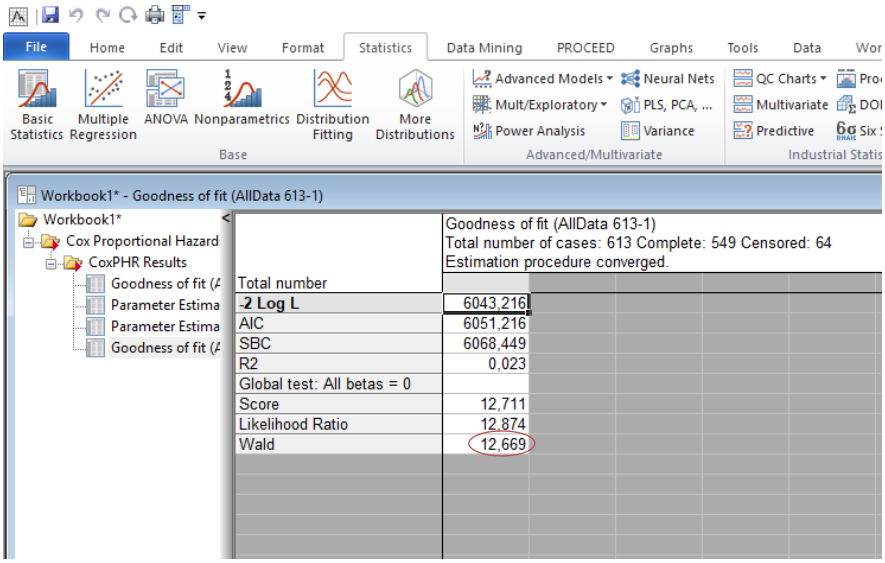

Рис. 30. Построение модели Кокса методом пошагового удаления предикторов в Statistica. Оценка качества регрессионной модели (1).

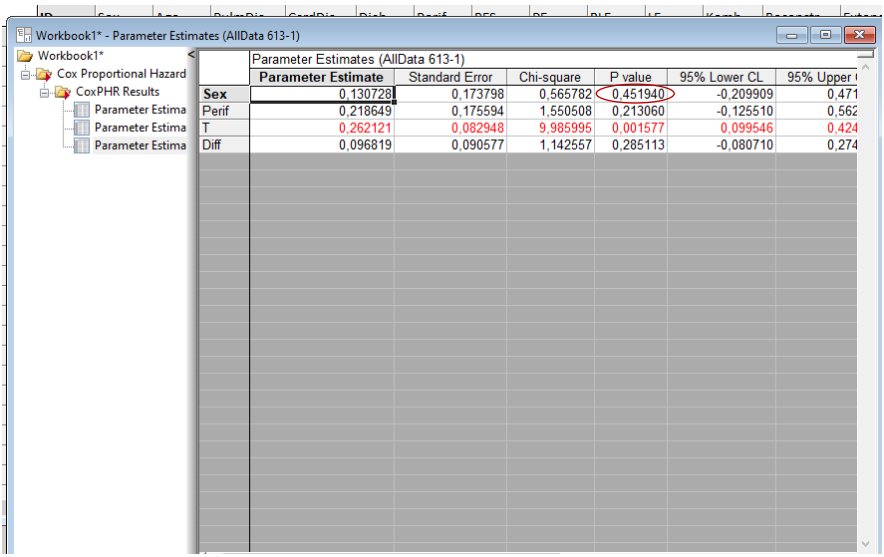

Рис. 31. Построение модели Кокса методом пошагового удаления предикторов в Statistica.

Шаг 2. Регрессионная модель (2). Оценка качества регрессионной модели (2). На третьем шаге будет удален предиктор Sex с наибольшим значением P value.

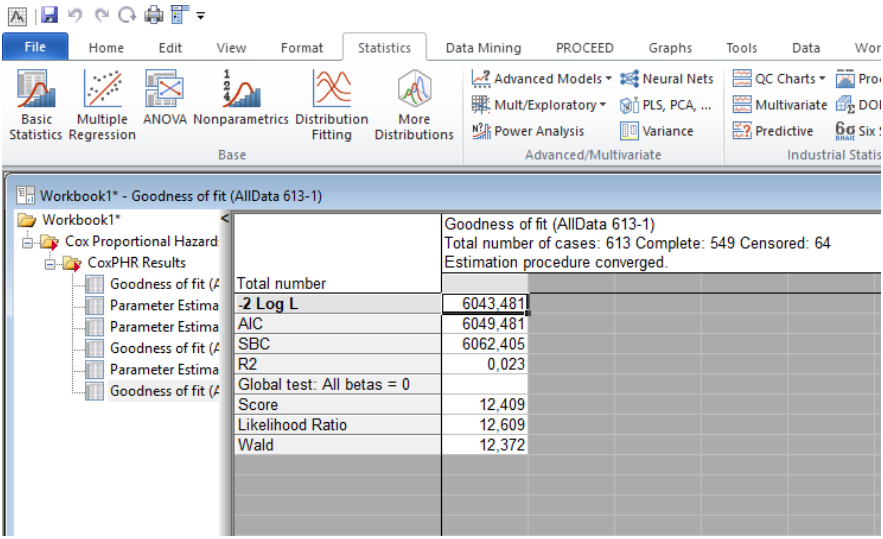

Рис. 32. Построение модели Кокса в Statistica. Оценка качества регрессионной модели (2).

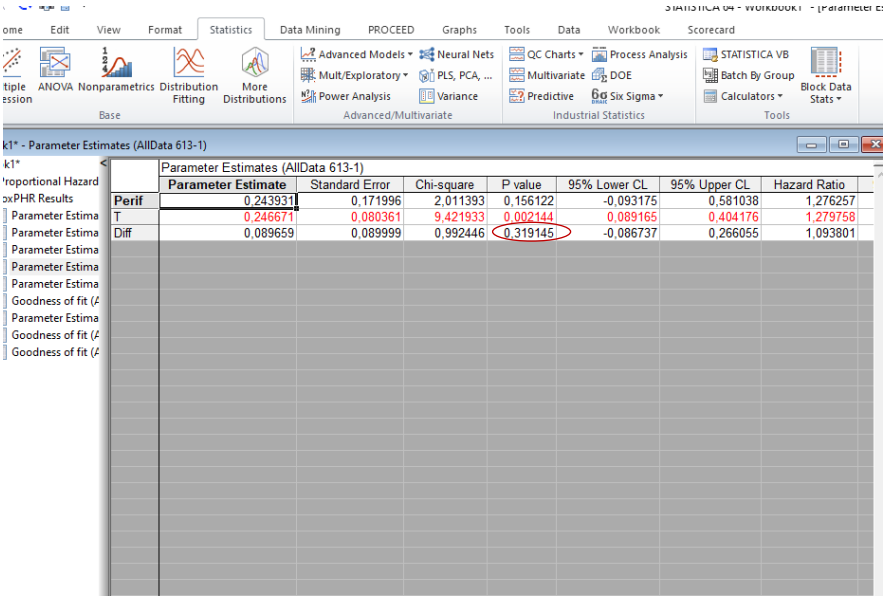

Рис. 33. Построение модели Кокса методом пошагового удаления предикторов в Statistica. Шаг 3. Регрессионная модель (3). На четвёртом шаге будет удален предиктор Diff с наибольшим значением P value.

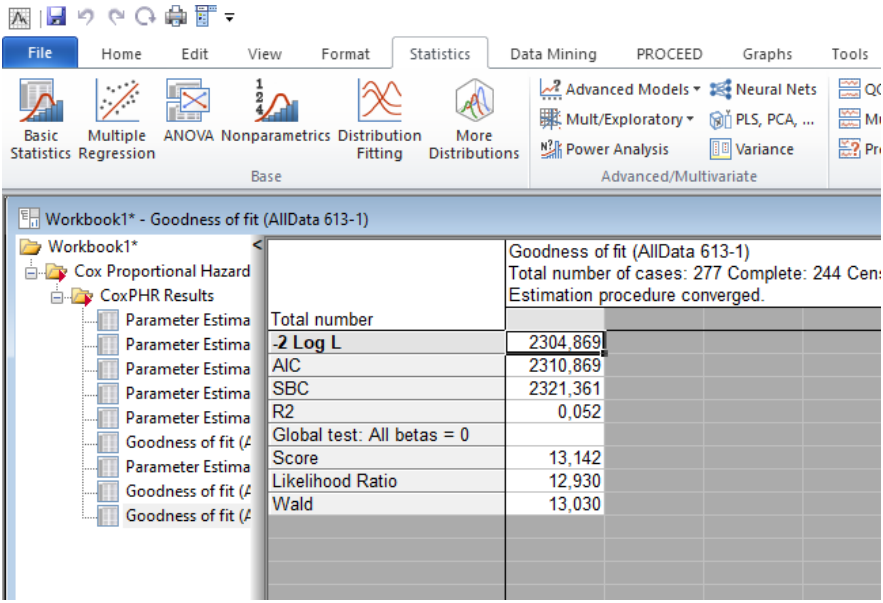

Рис. 34. Построение модели Кокса в Statistica. Оценка качества регрессионной модели (3).

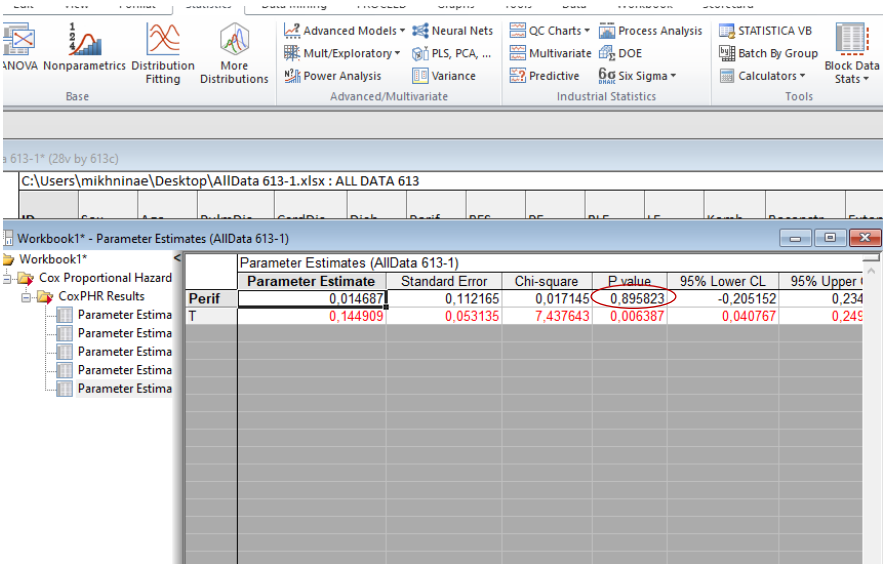

Рис. 35. Построение модели Кокса методом пошагового удаления предикторов в Statistica. Шаг 4. Регрессионная модель (4). На пятом шаге будет удален предиктор Perif с наибольшим значением P value.

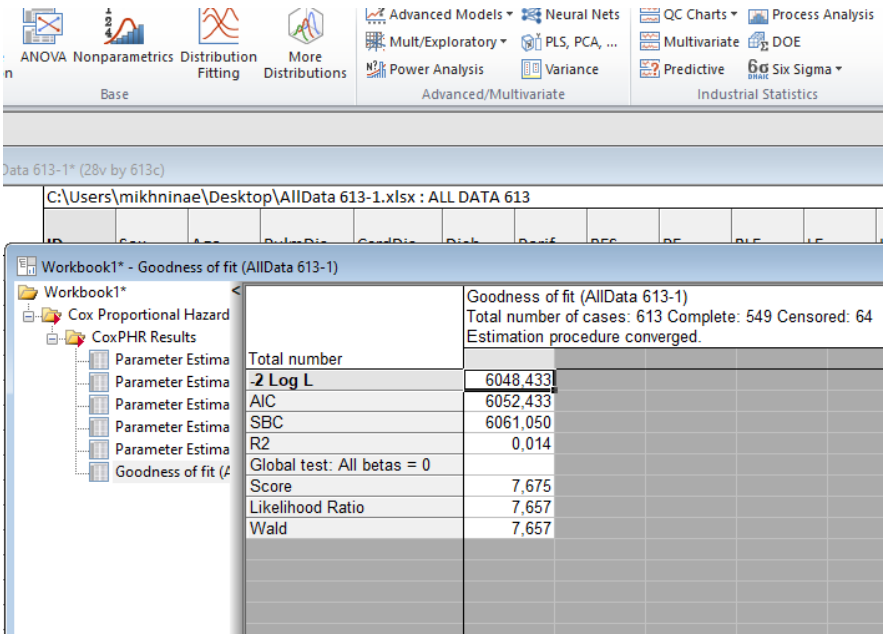

Рис. 36. Построение модели Кокса в Statistica. Оценка качества регрессионной модели (4).

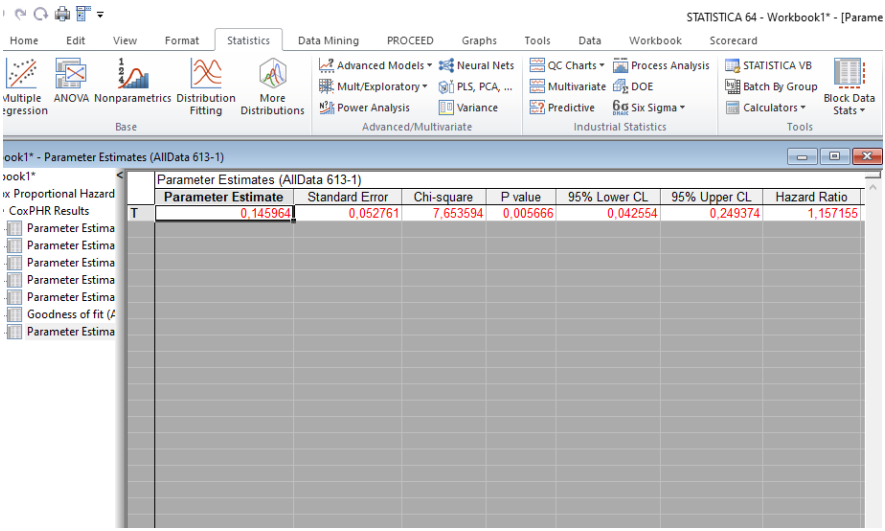

Рис. 37. Построение модели Кокса методом пошагового удаления предикторов в Statistica. Шаг 5. Регрессионная модель (5).

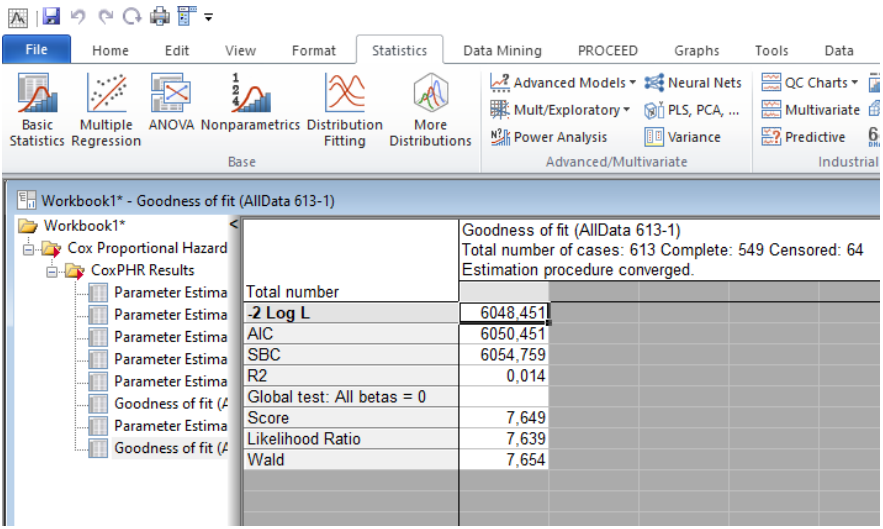

Рис. 38. Построение модели Кокса в Statistica. Оценка качества регрессионной модели (5).

Из пяти построенных моделей по критерию Вальда (13,030) наилучшей является также модель (3), изображенная на рис. 33. Легко видеть, что уровни статистической значимости предикторов и Hazard Ratio в точности соответствуют таковым, полученным в программе MedCalc.

Список литературы

1. Боровиков В. П. Популярное введение в современный анализ данных в системе STATISTICA. – Москва: Горячая линия - Телеком, 2013. – 288 с.

2. Гланц С. Медико-биологическая статистика. Электронная книга = Primer of BIOSTATISTICS. – 4-е изд. – Москва: Практика, 1999. –459 с.

3. Наследов А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках. – 2-е изд. – Санкт-Петербург: Питер, 2006.

4. Lukic I. K. MedCalc Version 7.0.0.2. Software Review // Croatian Medical Journal. – 2003. – Vol. 44. – P. 120-121.

ISBN 978-5-6046979-6-2

9 785604 697962 >

Отпечатано в ООО «АРТЕК»,
СПб, 6-я линия В.О., д.3/10
E-mail: artek-1@mail.ru, т. +7(911) 239-25-32
Подписано в печать 06.09.22
Формат 60x90/16. Печ. л. 3,5.
Тираж 50 экз.