МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М.В.ЛОМОНОСОВА
ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ

УТВЕРЖДАЮ
Декан факультета ВМК МГУ,
академик РАН

И.А. Соколов

« » _____ 2023 г.

ОТЧЕТ

О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ ПО ДОГОВОРУ №6.2.18 (ГОСБЮДЖЕТ, № ЦИТИС АААА-А18-118011590152-8) ИССЛЕДОВАНИЕ, РАЗРАБОТКА И ПРИМЕНЕНИЕ ИННОВАЦИОННЫХ ТЕХНОЛОГИЙ ПОСТРОЕНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ ПРОГРАММНЫХ СИСТЕМ

Руководитель проекта,
д.ф.-м.н., профессор, заведующий кафедрой И.В. Машечкин

Москва 2023

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель проекта, д.ф.-м.н., профессор, заведующий кафедрой	И.В. Машечкин
Исполнители темы	
к.ф.-м.н., доцент	М.И. Петровский
д.т.н., профессор	А.П. Рыжов
м.н.с	И.С. Попов
к.ф.-м.н., доцент	М.А. Казачук
математик	Ю.А. Васильев
инженер	О.Е. Горохов
математик	И.С. Лазухин
инженер	С.В. Герасимов
к.ф.-м.н., математик	Д.В. Царев

РЕФЕРАТ

Отчет 98 с., 1 ч., 83 рис., 10 табл., 32 источника.

COVID-19, ПРОГНОЗИРОВАНИЕ, ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, МОДЕЛИ ВЫЖИВАЕМОСТИ, МОДЕЛИ ПРОГНОЗИРОВАНИЯ ЛЕТАЛЬНОСТИ.

В отчете содержится информация о проведенных работах в части:

- Решения задачи анализа и обработки предоставленных исторических данных о развитии пандемии Covid-19 в г. Москве за 2020г. (очистка и приведение значений показателей к общим шкалам и словарям; поиск, удаление или исправление артефактов, выбросов и противоречивых данных; анализ и применение вероятностных методов множественной подстановки пропущенных значений для расчета ключевых показателей, таких как день течения заболевания при проведении анализа для случаев, когда эти данные не указаны или противоречивы);
- Решения задачи разработки на основе подготовленных данных моделей выживаемости для прогнозирования тяжести течения и исхода заболевания у пациентов (моделей выживаемости с использованием методов Каплана-Мейера и пропорциональных рисков Кокса с выявлением ключевых признаков, влияющих на выживаемость, а также выявлением стратифицирующих признаков; моделей выживаемости с учетом стратифицирующих признаков на основе использования методов Каплана-Мейера с выявлением важных предикторов внутри каждой из страт)

в рамках задачи исследования и разработки методов искусственного интеллекта и анализа больших данных в сфере здравоохранения.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящем отчете о НИР применяют следующие термины с соответствующими определениями.

ТЗ	Техническое задание
СПО	Системное программное обеспечение
ПАО	Программно-аппаратное оборудование
ИС	Информационная система
ПЦР	Полимеразная цепная реакция – метод молекулярной биологии, позволяющий добиться значительного увеличения малых концентраций определённых фрагментов нуклеиновой кислоты (ДНК или РНК) в биологическом материале (пробе)
ИФА	Иммуноферментный анализ – лабораторный иммунологический метод качественного или количественного определения различных низкомолекулярных соединений, макромолекул, вирусов и пр., в основе которого лежит специфическая реакция антиген-антитело
ЛИ	Лабораторные исследования
КТ	Компьютерная томография
NN	Neural Network (нейронная сеть)
RF	Random Forest (случайный лес)

ВВЕДЕНИЕ

Всемирная организация здравоохранения 11 марта 2020 года объявила пандемию по заболеванию COVID-19, вызываемому вирусом SARS-CoV-2. Пандемия резко ускорила внедрение цифровых сервисов в работу московского здравоохранения. Начиная с марта за полгода московские врачи вылечили свыше 500 тысяч человек – полмиллиона электронных историй болезни и выздоровления, в которых накоплен огромный объем данных. Вследствие этого актуальным является исследование возможности применения технологий искусственного интеллекта и машинного обучения для создания полезных инноваций для спасения жизни людей.

В рамках данного этапа научно-исследовательских работ проведены:

- Анализ и обработка предоставленных исторических данных о развитии пандемии Covid-19 в г. Москве за 2020г, в частности:
    - очистка и приведение значений показателей к общим шкалам и словарям;
    - поиск, удаление или исправление артефактов, выбросов и противоречивых данных;
    - анализ и применение вероятностных методов множественной подстановки пропущенных значений для расчета ключевых показателей, таких как день течения заболевания при проведении анализа для случаев, когда эти данные не указаны или противоречивы.
- Разработка на основе подготовленных данных моделей выживаемости для прогнозирования тяжести течения и исхода заболевания у пациентов, в частности:
    - моделей выживаемости с использованием методов Каплана-Мейера и пропорциональных рисков Кокса с выявлением ключевых признаков, влияющих на выживаемость, а также выявлением стратифицирующих признаков;
    - моделей выживаемости с учетом стратифицирующих признаков на основе использования методов Каплана-Мейера с выявлением важных предикторов внутри каждой из страт.

1 Аннотация

В рамках данного этапа работ были проведены научно-исследовательские работы в области создания прогностических моделей течения и исхода заболевания у пациентов с COVID-19. На основе данных о клинических особенностях, факторах коморбидности, клинико-лабораторного анализа и других факторов, потенциально связанных с тяжестью течения заболевания и вероятностью смерти пациентов с COVID-19, был разработан комплекс моделей, построенных с использованием передовых методов машинного обучения и прикладного статистического анализа, для прогнозирования тяжести течения и исхода заболевания у пациентов, получающих лечение в амбулаторных и стационарных условиях.

Построенные в рамках данной научно-исследовательской работы математические модели дают возможность судить о степени риска тяжелого течения заболевания, исследовать особенности распределения заболеваемости в зависимости от различных категорий факторов. Перспективной важной функцией данных моделей является описание долгосрочной динамики заболеваемости, включая сезонные циклы, что открывает перспективы прогнозирования тенденций и уровней развития основных показателей эпидемического процесса. Использование методов математического моделирования эпидемического процесса может быть чрезвычайно полезно также при планировании профилактических и противоэпидемических мероприятий, для выбора оптимальных путей борьбы с эпидемическим распространением COVID-19.

2Постановка задачи

Задачами данного этапа научно-исследовательских работ являются:

- Анализ и обработка предоставленных исторических данных о развитии пандемии Covid-19 в г. Москве за 2020г, в частности:
    - очистка и приведение значений показателей к общим шкалам и словарям;
    - поиск, удаление или исправление артефактов, выбросов и противоречивых данных;
    - анализ и применение вероятностных методов множественной подстановки пропущенных значений для расчета ключевых показателей, таких как день течения заболевания при проведении анализа для случаев, когда эти данные не указаны или противоречивы.
- Разработка на основе подготовленных данных моделей выживаемости для прогнозирования тяжести течения и исхода заболевания у пациентов, в частности:
    - моделей выживаемости с использованием методов Каплана-Мейера и пропорциональных рисков Кокса с выявлением ключевых признаков, влияющих на выживаемость, а также выявлением стратифицирующих признаков;
    - моделей выживаемости с учетом стратифицирующих признаков на основе использования методов Каплана-Мейера с выявлением важных предикторов внутри каждой из страт.

3 Исследование и построение решения

3.1 Анализ данных о весеннем периоде развития пандемии Covid-19 в г.Москве

3.1.1 Описание наборов данных

Суммарно за рассматриваемый период было получено и проанализировано более 2 млн. анализов (примерно 1.3 млн. пациентов), из них:

- 60% анализов отрицательный ПЦР;
- 10% анализов положительный ПЦР;
- 27% анализов ИФА.

Дополнительные данные:

- Использованные: КТ, смертность, лабораторные исследования (анализ крови на лейкоциты, лимфоциты, тромбоциты, нейтрофилы, гемоглобин, с-реактивный белок, ферритин, D-димер);
- Не использованные: опросники, диагнозы, рецепты, остальные лабораторные исследования.

3.1.2 Подготовка данных

На вход подаются данные медицинских анализов пациентов (более 1.3 млн человек), больных Covid-19.

Необходимо подготовить эти данные к построению над ними математических моделей:

- *Очистить данные и привести значения показателей к общим шкалам и словарям;*
- *Найти, удалить и исправить артефакты, выбросы и противоречивые данные;*
- *Предложить вероятностные методы множественной подстановки пропущенных значений для расчета ключевых показателей, таких как день течения заболевания при проведении анализа для случаев, когда эти данные не указаны или противоречивы.*

3.1.2.1 Очистка и приведение значений показателей к общим шкалам и словарям

Поскольку по различным пациентам показатели измерялись в разных шкалах, данные показатели были унифицированы. Это необходимо для дальнейшей корректной работы методов машинного обучения. Подробно данный процесс описан в разделе 3.2.2.1 данного документа.

3.1.2.2 Поиск, удаление или исправление артефактов, выбросов и противоречивых данных

В ходе анализа исходных наборов были найдены противоречия и ошибки ввода данных. Соответствующие показания были удалены. Подробно данный процесс описан в разделе 3.2.2.2 данного документа.

Также была проведена предварительная фильтрация: для дальнейшего рассмотрения были оставлены показания пациентов, по которым имелся хотя бы один результат Ig. В итоге, были проанализированы показания порядка 40 тысяч пациентов.

Было получено, что в данных по измерениям КТ и лабораторных исследований имеется большой процент пропусков:

- Показатель:
  - Д-димер 98% пропусков;
  - Ферритин 95% пропусков;
  - Нейтрофилы 75% пропусков;
  - Лимфоциты 70% пропусков;
  - Тромбоциты 60% пропусков;
  - Лейкоциты и CRP 35% пропусков;
  - Гемоглобин 32% пропусков;
  - КТ 10 % пропусков;
- Наблюдений, где заполнены все показатели всего около 100 (0.25%).

Было предложено два способа заполнения пропущенных значений:

- «Простой» подход – заполнение пропусков «средним» при анализе влияния числовых признаков, введение специальной категории «неизвестно» при анализе влияния категориальных предикторов (таких как выход за границы референсных значений);

- «Сложный» подход – на основе кластеризации, заполнение пропусков от прототипа ближайшего кластера.

3.1.2.3 Использование вероятностных методов множественной подстановки пропущенных значений для расчета ключевых показателей, таких как день течения заболевания при проведении анализа для случаев, когда эти данные не указаны или противоречивы

Также было замечено, что по большинству пациентов даты начала заболевания являются некорректными (корректные меньше чем у 30 000 пациентов). Для решения данной проблемы было произведено прогнозирование даты заболевания (см. Рисунки 1, 2):

- На основе подхода k-ближайших соседей [1];
- С использованием Байесовского бутстрепинга [2];
- С учетом возраста, КТ, ПЦР и начальных значений IgM и IgG.

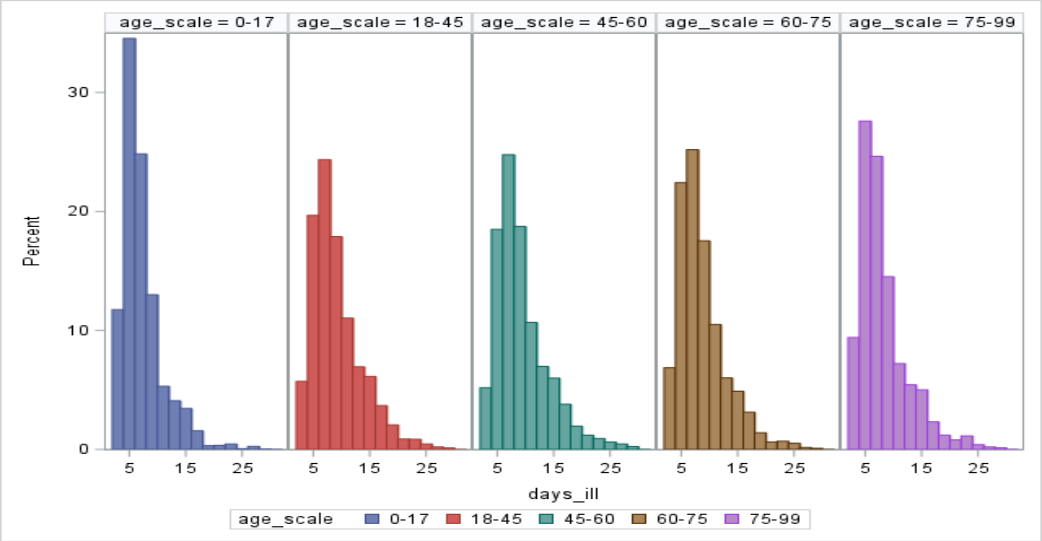

Рисунок 1 — Прогноз даты заболевания

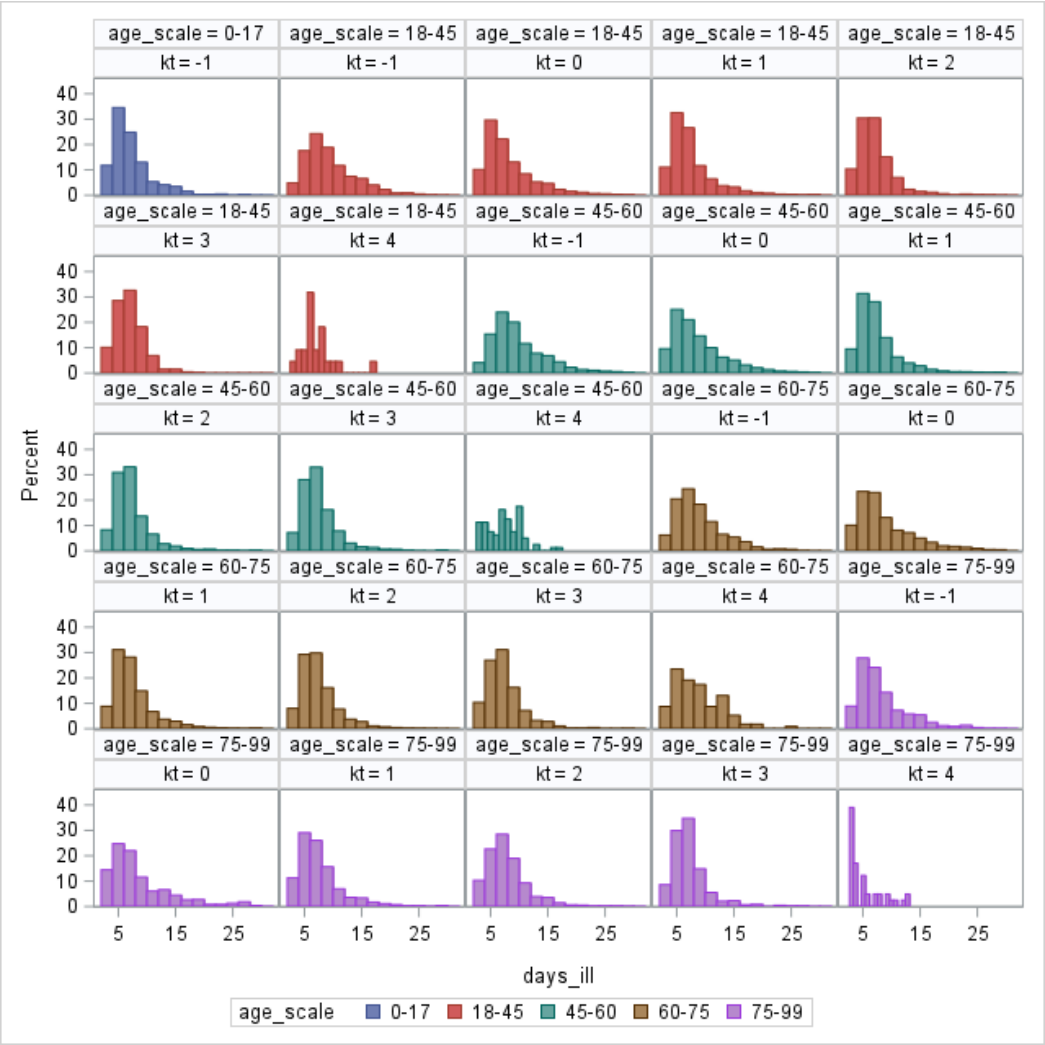

Рисунок 2 — Прогноз даты заболевания

3.1.3 Разработка моделей выживаемости

На вход подаются обработанные данные медицинских анализов пациентов.

Необходимо разработать модели выживаемости с использованием методов Каплана-Мейера и пропорциональных рисков Кокса, выявить ключевые и стратифицирующие признаки.

Целевой переменной является факт летальности.

3.1.3.1 Разработка моделей выживаемости с использованием методов Каплана-Мейера и пропорциональных рисков Кокса, выявление ключевых признаков, влияющих на выживаемость, выявление стратифицирующих признаков

В значительной мере клинические проявления COVID-19, варьирующиеся от бессимптомного течения до полиорганной недостаточности, связаны с особенностями

иммунного ответа у пациента. На Рисунках 3–5 представлена темпоральная (временная) модель данных, отражающая произвольные показатели (положительный результат ПЦР, рост/снижение антител класса IgM и рост антител класса IgG), связанные с определенными промежутками времени.

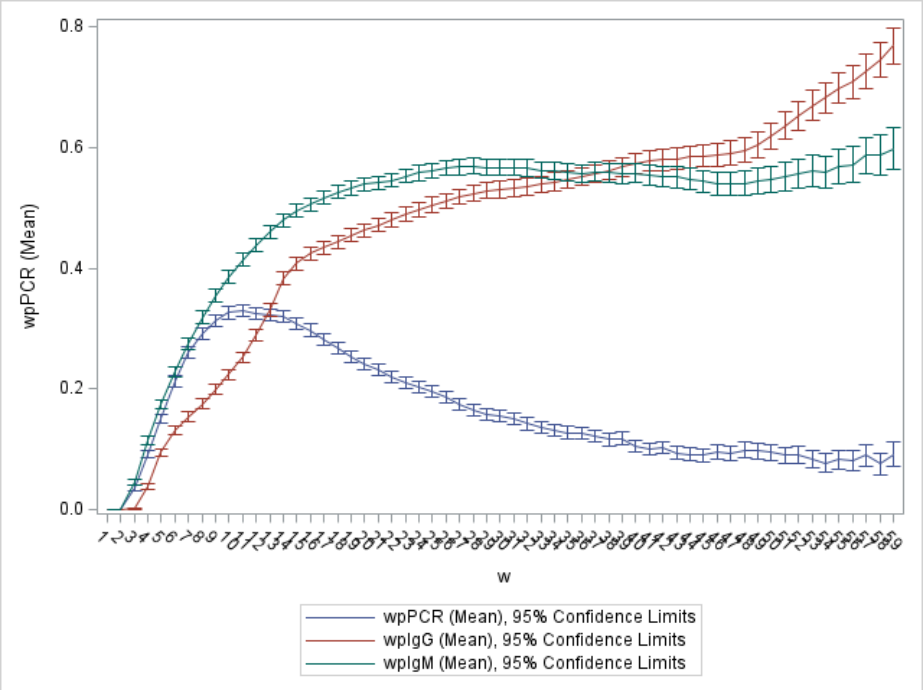

Рисунок 3 — График среднего числа положительных ПЦР и ИФА в выборке по времени

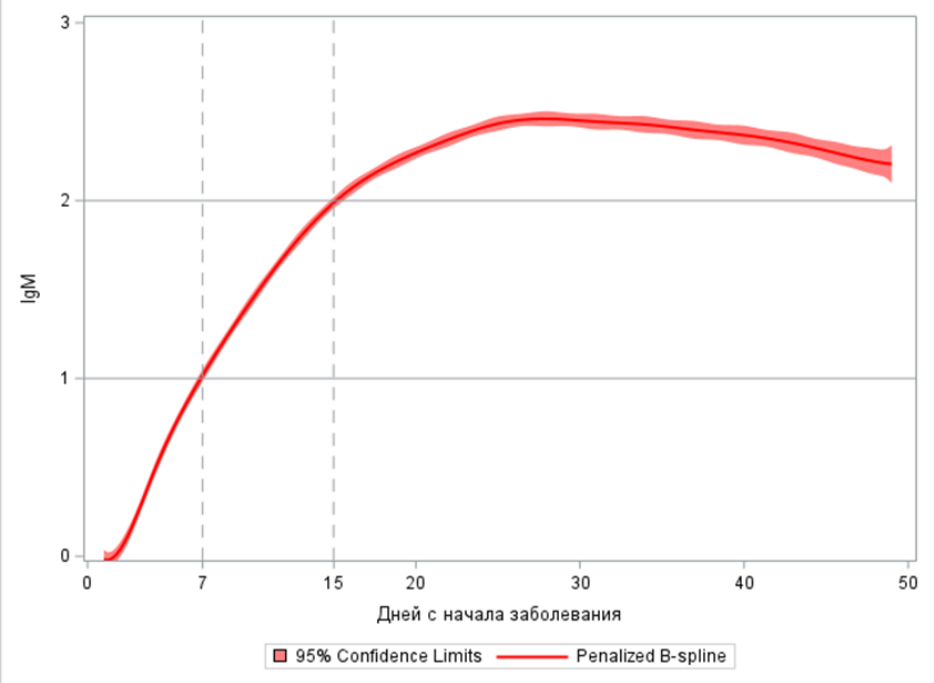

Рисунок 4 — Анализ значений антител класса IgM от дня заболевания

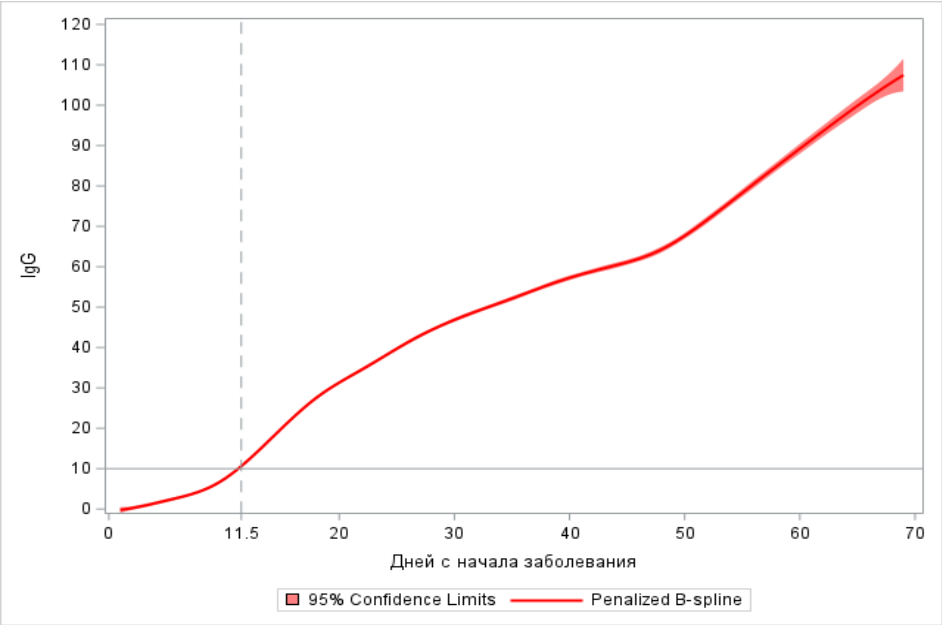

Рисунок 5 — Анализ значений антител класса IgG от дня заболевания

Были построены модели выживаемости Каплана-Мейера [3–8] по возрасту и полу (данные модели не зависят от метода постановки, т.к. пол и возраст всегда заполнены).

В ходе анализа модели выживаемости по возрасту (см. Рисунки 6, 7) было получено, что:

- **Чем старше пациент, тем выше вероятность умереть.**

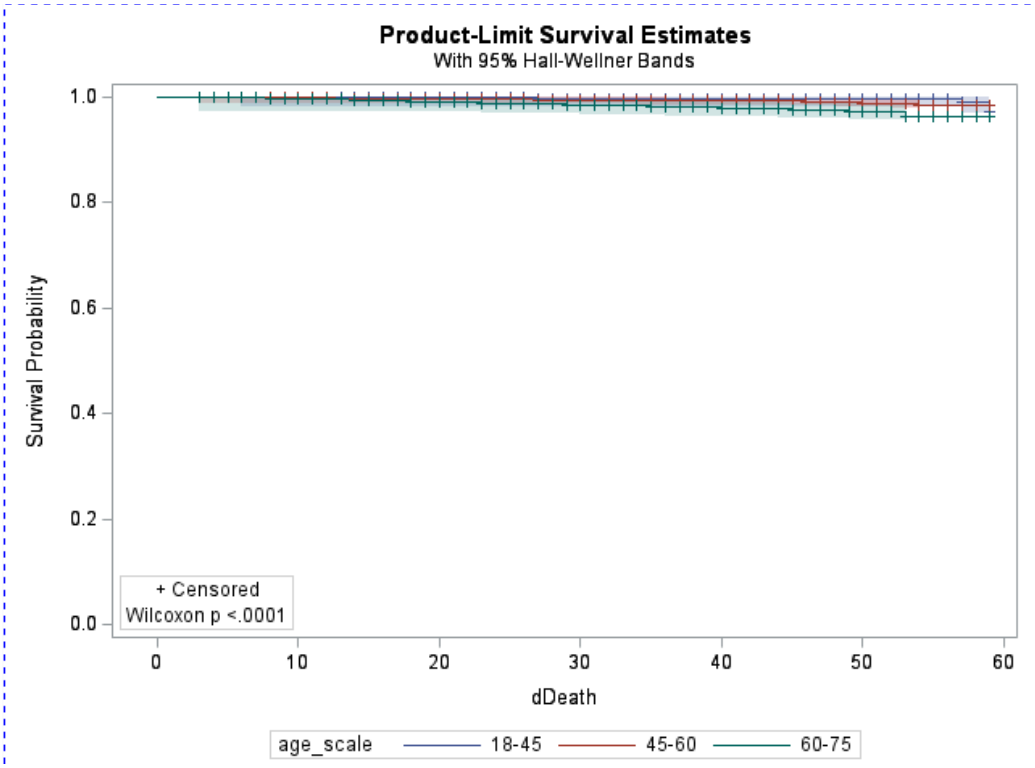

Рисунок 6 — Модель выживаемости Каплана-Мейера по возрасту, график

Adjustment for Multiple Comparisons for the Wilcoxon Test				
Strata Comparison			p-Values	
age_scale	age_scale	Chi-Square	Raw	Scheffe
18-45	45-60	17.1179	<.0001	0.0002
18-45	60-75	91.3644	<.0001	<.0001
45-60	60-75	17.3888	<.0001	0.0002

Рисунок 7 — Модель выживаемости Каплана-Мейера по возрасту, оценки теста Вилкоксона

В ходе анализа модели выживаемости по полу (см. Рисунки 8, 9) было получено, что:

- **Смертность выше у мужчин.**

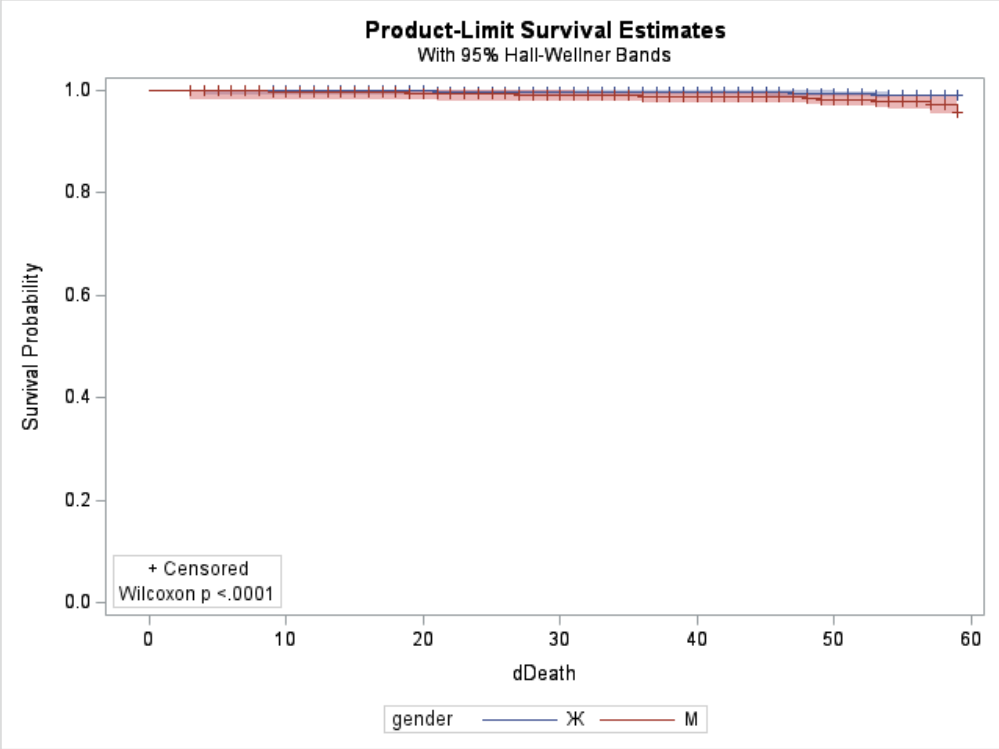

Рисунок 8 — Модель выживаемости Каплана-Мейера по полу, график

Adjustment for Multiple Comparisons for the Wilcoxon Test				
Strata Comparison			p-Values	
gender	gender	Chi-Square	Raw	Scheffe
Ж	M	55.2697	<.0001	<.0001

Рисунок 9 — Модель выживаемости Каплана-Мейера по полу, оценки теста Вилкоксона

В ходе анализа построенных моделей выживаемости Каплана-Мейера по Ig и PCR (не зависят от метода постановки, т.к. Ig и PCR всегда заполнены) (см. Рисунки 10–12), были получены следующие выводы:

- **Нет зависимости от положительного ПЦР;**
- **При IgG и IgM выше клинических порогов вероятность умереть ниже.**

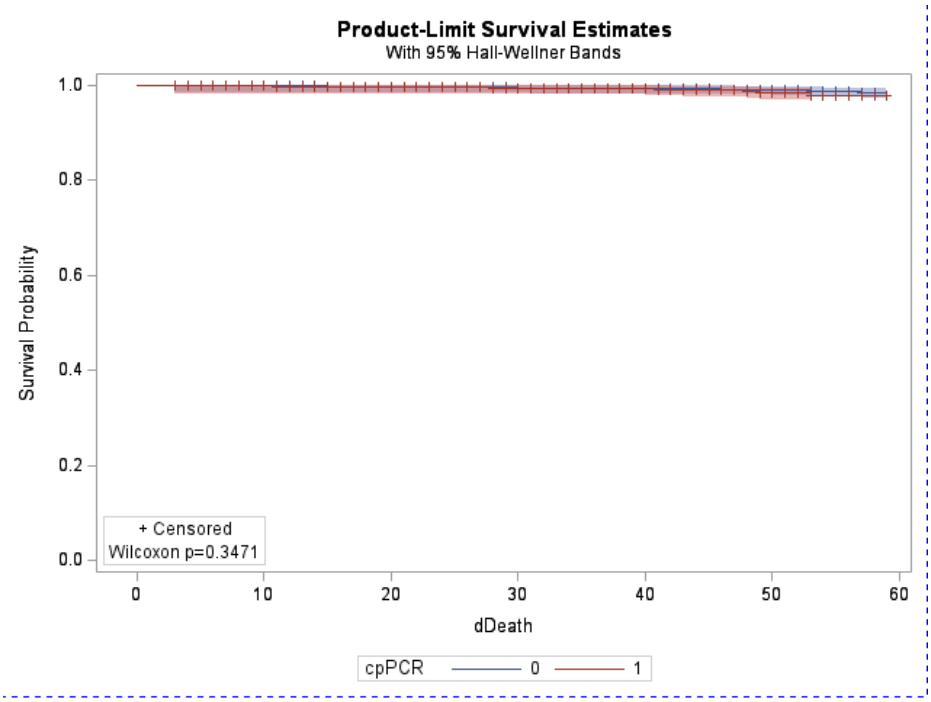

Рисунок 10 — Модель выживаемости Каплана-Мейера по ПЦР, график

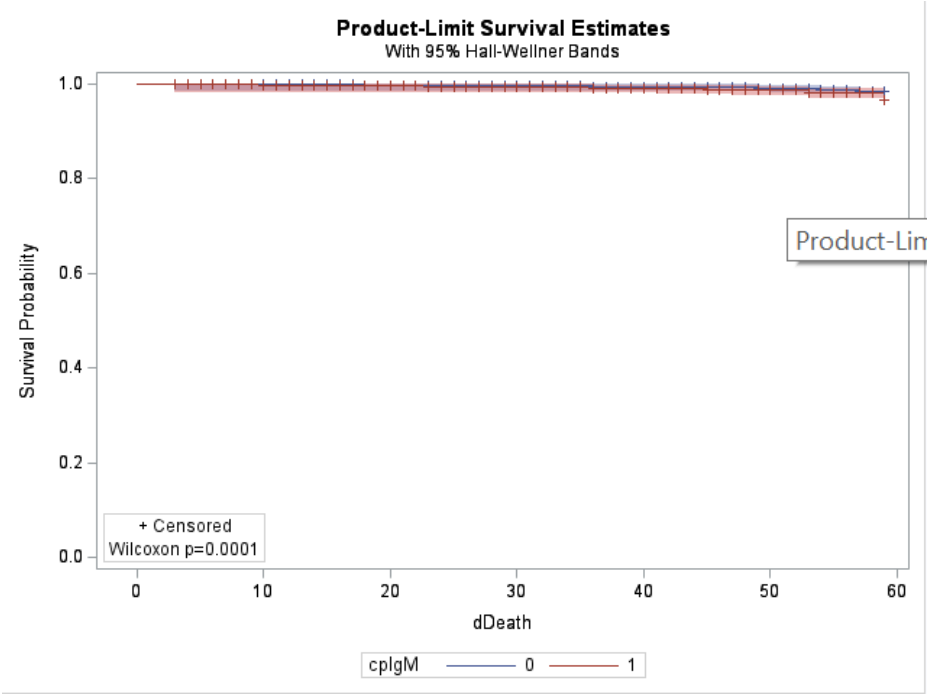

Рисунок 11 — Модель выживаемости Каплана-Мейера по IgM, график

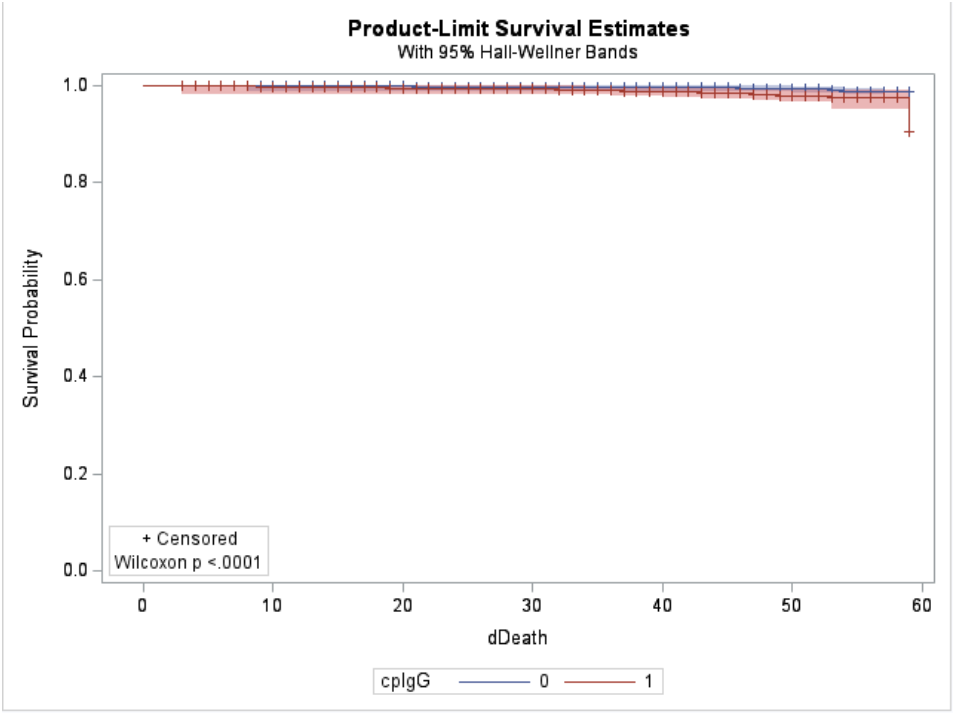

Рисунок 12 — Модель выживаемости Каплана-Мейера по IgG, график

Также были построены модели пропорциональных рисков Кокса [9–14] при «простой» подстановке пропусков и стратификацией по полу (см. Рисунки 13, 14):

- **Порядок добавления значимых переменных в модель (красный цвет – при росте показателя растет отношение риска смерти, зеленый – уменьшается):**
  1. **минимальный CRP**
  2. **возраст**
  3. **выработался ли IgG>10**
  4. **вышел ли за границы CRP**
  5. **минимальный IgG**
  6. **минимальные лейкоциты**
  7. **средний гемоглобин**

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	mnCRP		1	1	344.8686		<.0001
2	age		1	2	84.6085		<.0001
3	cpIgG		1	3	54.6666		<.0001
4	bdCRP		1	4	43.5314		<.0001
5	mnIgG		1	5	17.1760		<.0001
6	mnWBC		1	6	16.7873		<.0001
7	avHGB		1	7	14.8020		0.0001

Рисунок 13 — Анализ модели пропорциональных рисков Кокса при «простой» подстановке пропусков и стратификацией по полу

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
age	1	0.07968	0.00849	88.0428	<.0001	1.083	1.065	1.101
bdCRP	1	-0.35035	0.05433	41.5843	<.0001	0.704	0.633	0.784
mnCRP	1	0.00992	0.0008433	138.3786	<.0001	1.010	1.008	1.012
avHGB	1	-0.02038	0.00480	17.9963	<.0001	0.980	0.971	0.989
mnWBC	1	0.02267	0.00654	12.0057	0.0005	1.023	1.010	1.036
mnIgG	1	-0.01117	0.00265	17.8129	<.0001	0.989	0.984	0.994
cpIgG	1	0.51684	0.19683	6.8947	0.0086	1.677	1.140	2.466

Рисунок 14 — Анализ модели пропорциональных рисков Кокса при «простой» подстановке пропусков и стратификацией по полу

Общая модель выживаемости Каплана-Мейера с выбором значимых признаков при «простой» подстановке представлена на Рисунке 15.

При этом, важными признаками являются (см. Рисунок 16):

- **CRP**
- **Возраст**
- **Средний IgG**
- **Минимальные лейкоциты**
- **Минимальные тромбоциты**
- **Все статистики по нейтрофилам**
- **Средний гемоглобин**

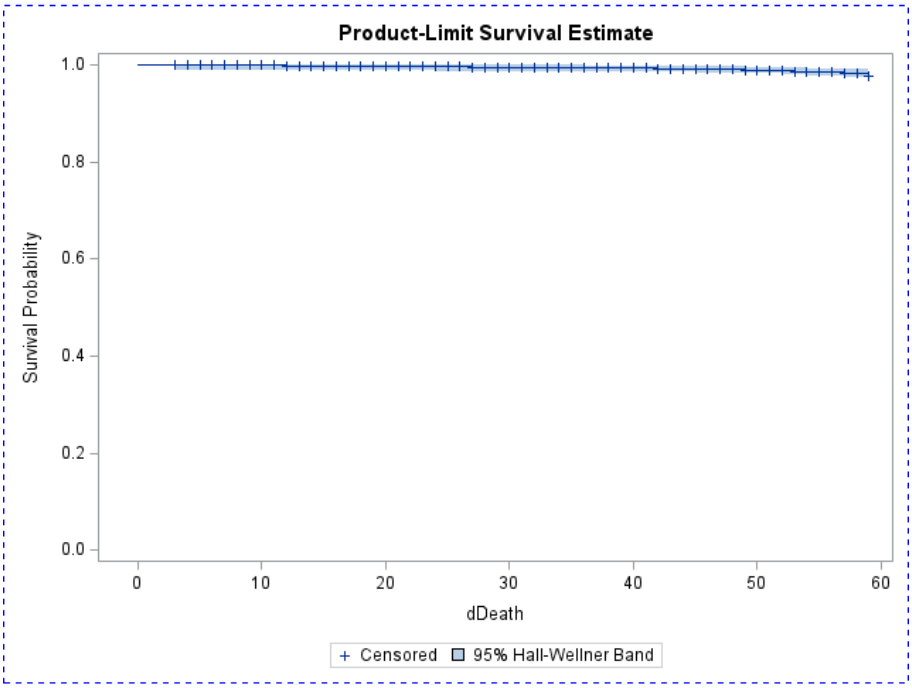

Рисунок 15 — Модель выживаемости Каплана-Мейера с выбором значимых признаков при «простой» подстановке

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
mnCRP	1	417.2	<.0001	417.2	<.0001
age	2	473.3	<.0001	56.1329	<.0001
avIgG	3	517.9	<.0001	44.5931	<.0001
mnWBC	4	558.7	<.0001	40.8072	<.0001
mnPLT	5	565.8	<.0001	7.1338	0.0076
mxNEUT	6	572.4	<.0001	6.6257	0.0101
mnNEUT	7	579.0	<.0001	6.5056	0.0108
avNEUT	8	584.5	<.0001	5.5159	0.0188
avHGB	9	588.7	<.0001	4.2331	0.0396

Рисунок 16 — Модель выживаемости Каплана-Мейера с выбором значимых признаков при «простой» подстановке, оценки теста Вилкоксона

При анализе модели выживаемости Каплана-Мейера по КТ при «простой» подстановке (см. Рисунки 17, 18) были получены следующие результаты:

- **В целом чем выше поражение, тем выше вероятность умереть**
- **Близкие по вероятности выжить категории:**
  - **-1 (не делали КТ), 2 и 3;**
  - **0 и 1**

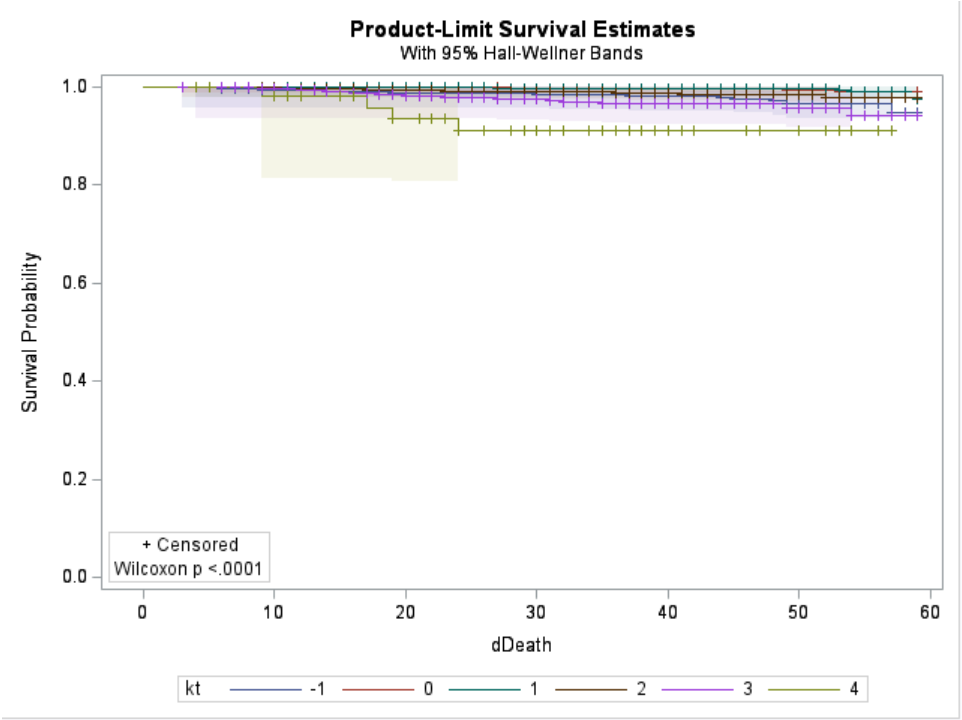

Рисунок 17 — Модель выживаемости Каплана-Мейера по КТ при «простой» подстановке, график

Adjustment for Multiple Comparisons for the Wilcoxon Test				
Strata Comparison			p-Values	
kt	kt	Chi-Square	Raw	Scheffe
-1	0	60.7245	<.0001	<.0001
-1	1	79.1809	<.0001	<.0001
-1	2	4.4909	0.0341	0.4811
-1	3	5.9289	0.0149	0.3132
-1	4	56.7663	<.0001	<.0001
0	1	1.1171	0.2905	0.9526
0	2	31.9426	<.0001	<.0001
0	3	52.3421	<.0001	<.0001
0	4	28.8930	<.0001	<.0001
1	2	45.0065	<.0001	<.0001
1	3	73.1246	<.0001	<.0001
1	4	48.5420	<.0001	<.0001
2	3	0.1551	0.6937	0.9995
2	4	10.6690	0.0011	0.0584
3	4	47.7799	<.0001	<.0001

Рисунок 18 — Модель выживаемости Каплана-Мейера по КТ при «простой» подстановке, оценки теста Вилкоксона

Выводы по выживаемости на основе метода Каплана-Мейера по лабораторным исследованиям при «простой» подстановке:

- **Нет статистической зависимости от:**
  - **Ферритина;**
  - **Д-димера;**
- **Есть зависимость «чем больше тем хуже» от:**
  - **CRP;**
  - **лейкоцитов;**
  - **нейтрофилов;**
  - **лимфоцитов;**
- **Есть зависимость «чем больше, тем лучше» от гемоглобина;**
- **Есть зависимость «при норме смертность выше» от тромбоцитов.**

Были построены модели пропорциональных рисков Кокса при «сложной» подстановке пропусков и стратификацией по полу (см. Рисунки 19, 20):

Порядок добавления значимых переменных в модель (красный цвет – при росте показателя растет отношение риска смерти, зеленый – уменьшается):

- **минимальный CRP;**
- **возраст;**
- **средний IgG;**
- **KT;**
- **максимальный гемоглобин;**
- **минимальные лейкоциты;**
- **минимальные тромбоциты.**

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	mnCRP		1	1	312.7465		<.0001
2	age		1	2	62.1443		<.0001
3	avIgG		1	3	37.0591		<.0001
4	kt		1	4	48.7137		<.0001
5	mxHGB		1	5	14.6755		0.0001
6	mnWBC		1	6	14.2091		0.0002
7	mnPLT		1	7	7.9545		0.0048

Рисунок 19 — Анализ модели пропорциональных рисков Кокса при «сложной» подстановке пропусков и стратификацией по полу

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
age	1	0.07110	0.01062	44.8081	<.0001	1.074	1.052	1.096
avIgG	1	-0.01485	0.00244	37.1973	<.0001	0.985	0.981	0.990
mnCRP	1	0.00804	0.00104	59.8892	<.0001	1.008	1.006	1.010
mxHGB	1	-0.02557	0.00499	26.2878	<.0001	0.975	0.965	0.984
mnPLT	1	-0.00472	0.00162	8.4647	0.0036	0.995	0.992	0.998
mnWBC	1	0.02523	0.00598	17.7728	<.0001	1.026	1.014	1.038
kt	1	0.58220	0.08924	42.5628	<.0001	1.790	1.503	2.132

Рисунок 20 — Анализ модели пропорциональных рисков Кокса при «сложной» подстановке пропусков и стратификацией по полу

В ходе анализа построенной общей модели выживаемости Каплана-Мейера с выбором значимых признаков при «сложной» подстановке (см. Рисунки 21, 22) было получено, что важными признаками являются:

- **CRP;**

- **Возраст;**
- **Средний IgG;**
- **Минимальные лейкоциты;**
- **Минимальные тромбоциты;**
- **Нейтрофилы;**
- **Средний гемоглобин.**

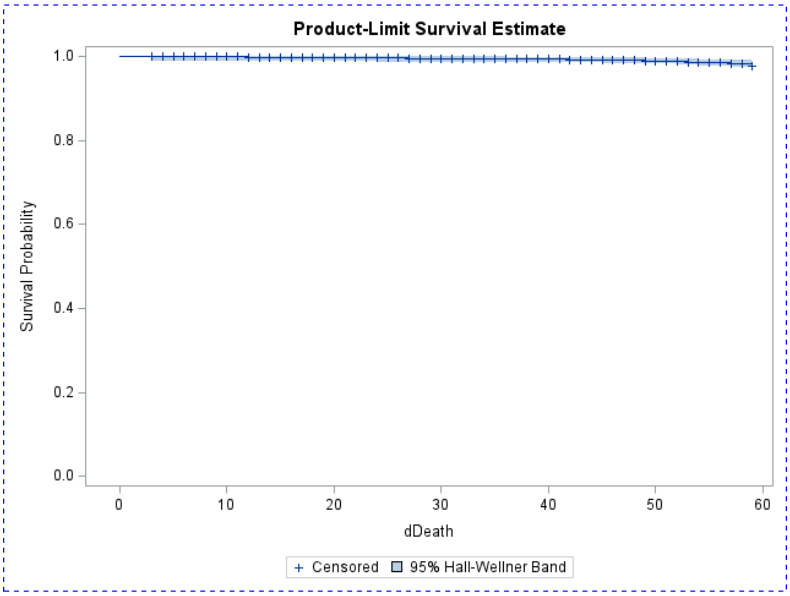

Рисунок 21 — Общая модель выживаемости Каплана-Мейера с выбором значимых признаков при «сложной» подстановке, график

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test

Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
mnCRP	1	417.2	<.0001	417.2	<.0001
age	2	473.3	<.0001	56.1329	<.0001
avIgG	3	517.9	<.0001	44.5931	<.0001
mnWBC	4	558.7	<.0001	40.8072	<.0001
mnPLT	5	565.8	<.0001	7.1338	0.0076
mxNEUT	6	572.4	<.0001	6.6257	0.0101
mnNEUT	7	579.0	<.0001	6.5056	0.0108
avNEUT	8	584.5	<.0001	5.5159	0.0188
avHGB	9	588.7	<.0001	4.2331	0.0396

Рисунок 22 — Общая модель выживаемости Каплана-Мейера с выбором значимых признаков при «сложной» подстановке, оценки теста Вилкоксона

В ходе анализа построенной модели выживаемости Каплана-Мейера по KT при «сложной» подстановке (см. Рисунки 23, 24) были получены следующие выводы:

- **Чем выше поражение, тем выше вероятность умереть;**
- **Похожие категории: 0 и 1, 2 и 3.**

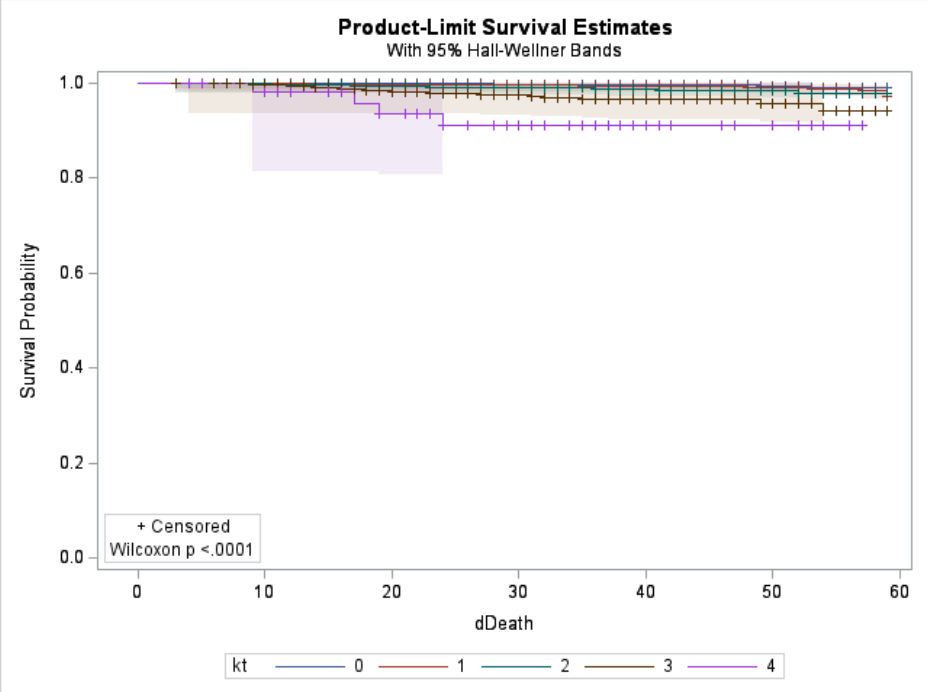

Рисунок 23 — Модель выживаемости Каплана-Мейера по KT при «сложной» подстановке, график

Adjustment for Multiple Comparisons for the Wilcoxon Test

Strata Comparison			p-Values	
kt	kt	Chi-Square	Raw	Scheffe
0	1	2.3650	0.1241	0.6690
0	2	31.9426	<.0001	<.0001
0	3	52.3421	<.0001	<.0001
0	4	28.8930	<.0001	<.0001
1	2	10.4279	0.0012	0.0338
1	3	18.4168	<.0001	0.0010
1	4	5.3449	0.0208	0.2537
2	3	0.1551	0.6937	0.9971
2	4	10.6690	0.0011	0.0305
3	4	47.7799	<.0001	<.0001

Рисунок 24 — Модель выживаемости Каплана-Мейера по KT при «сложной» подстановке, оценки теста Вилкоксона

Общие выводы модели выживаемости Каплана-Мейера по лабораторным исследованиям при «сложной» подстановке выглядят следующим образом:

- **Статистическая значимость есть по всем предикторам!**
- **Есть зависимость «чем больше, тем хуже» от:**
  - **CRP,**
  - **лейкоцитов,**
  - **нейтрофилов,**
  - **лимфоцитов,**
  - **Ферритина,**
  - **Д-димера;**
- **Есть зависимость «чем больше, тем лучше» от гемоглобина;**
- **Есть зависимость «при норме смертность выше» от тромбоцитов.**

При проведении сложной подстановки на основе кластеризации были выделены две группы риска с высокой ранней и высокой поздней смертностью. Таким образом, в качестве стратифицирующей переменной был выбран признак наличия и характера смертности. Подробнее об этом написано в разделе 3.1.3.2 данного отчета.

3.1.3.2 Разработка моделей выживаемости с учетом стратифицирующих признаков на основе использования методов Каплана-Мейера и выявление важных предикторов внутри каждой из страт

В результате «сложной» подстановки на основе кластеризации (см. Рисунок 25):

- были найдены два кластера (9 и 13 из 15), в которых летальность была существенно выше, чем в среднем по выборке;
- причем в 9 кластере умирать начинали существенно раньше (с 10 по 20 день), чем в 13 (с 20 по 30 день).

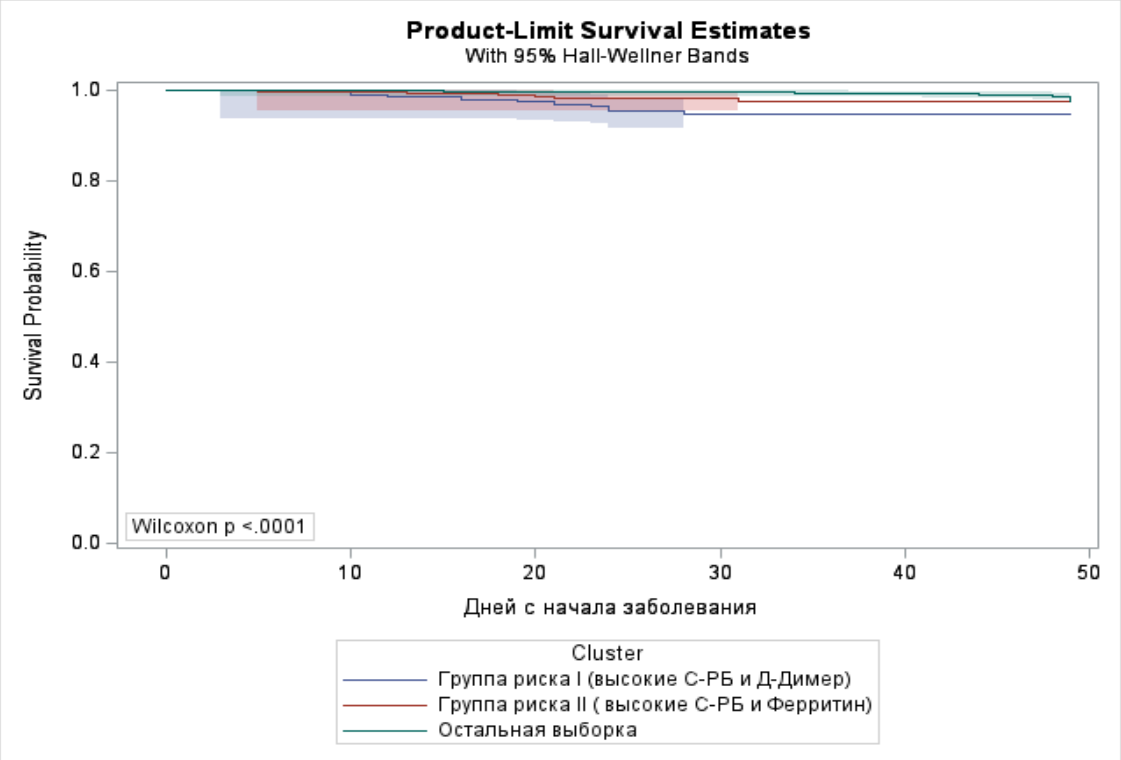

Рисунок 25 — Сложная подстановка, модель выживаемости на основе Каплана-Мейера, стратификация

Факторный анализ кластера 9 с высокой «ранней» (с 10 по 20 день) смертностью (см. Рисунки 26, 27) показал важные признаки, отличающие кластер 9 от всей выборки:

- Минимальный и средний CRP за время болезни у пациентов из кластера 9 значительно выше, чем у пациентов из остальной выборки;
- Минимальный Д-димер за время болезни у пациентов из кластера 9 несколько выше, чем у пациентов из остальной выборки.

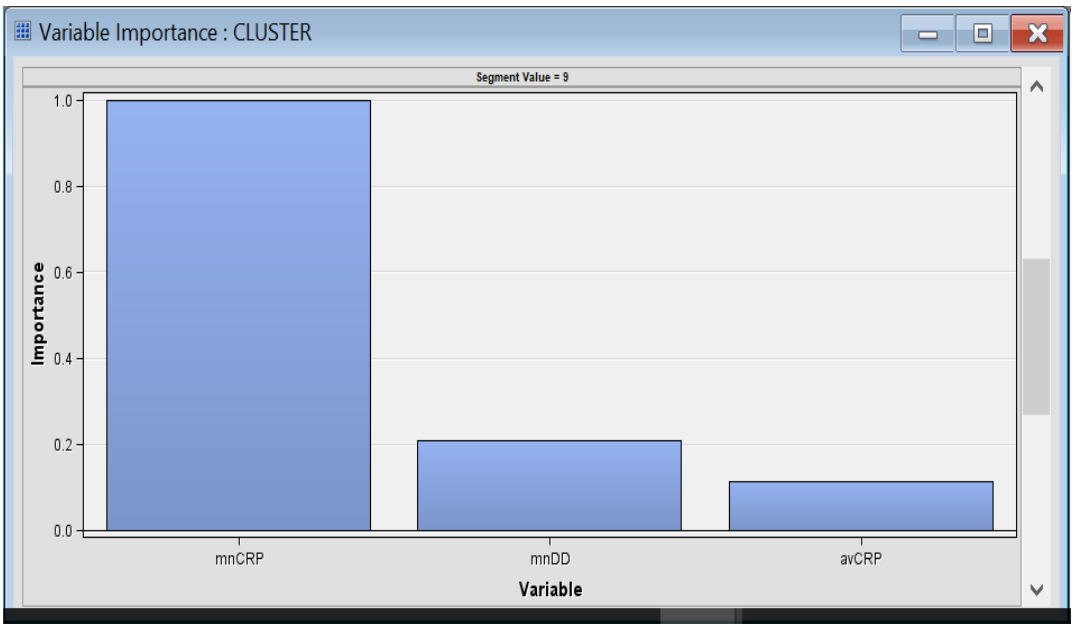

Рисунок 26 — Факторный анализ кластера 9 с высокой «ранней» (с 10 по 20 день) смертностью

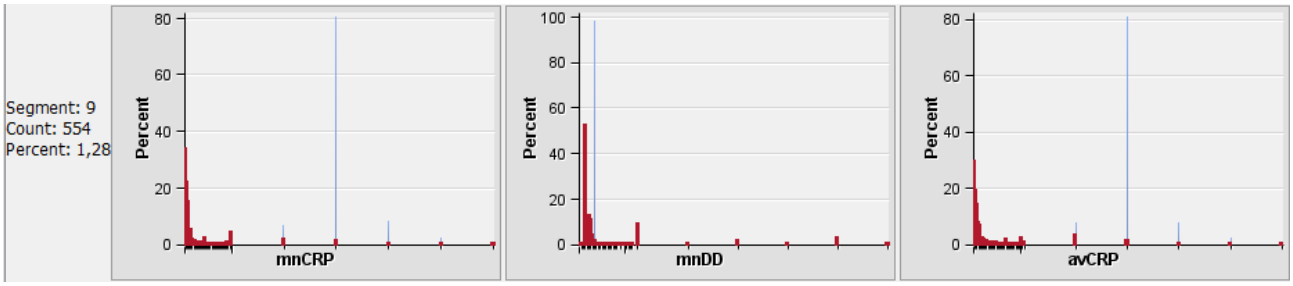

Рисунок 27 — Факторный анализ кластера 9 с высокой «ранней» (с 10 по 20 день) смертностью

Факторный анализ кластера 13 с высокой «поздней» (с 20 по 30 дни) смертностью (см. Рисунки 28, 29) показал важные признаки, отличающие кластер 13 от всей выборки:

- Максимальный ферритин за время болезни у пациентов из кластера 13 значительно выше, чем у пациентов из остальной выборки;
- Максимальный CRP за время болезни у пациентов из кластера 13 несколько выше, чем у пациентов из остальной выборки.

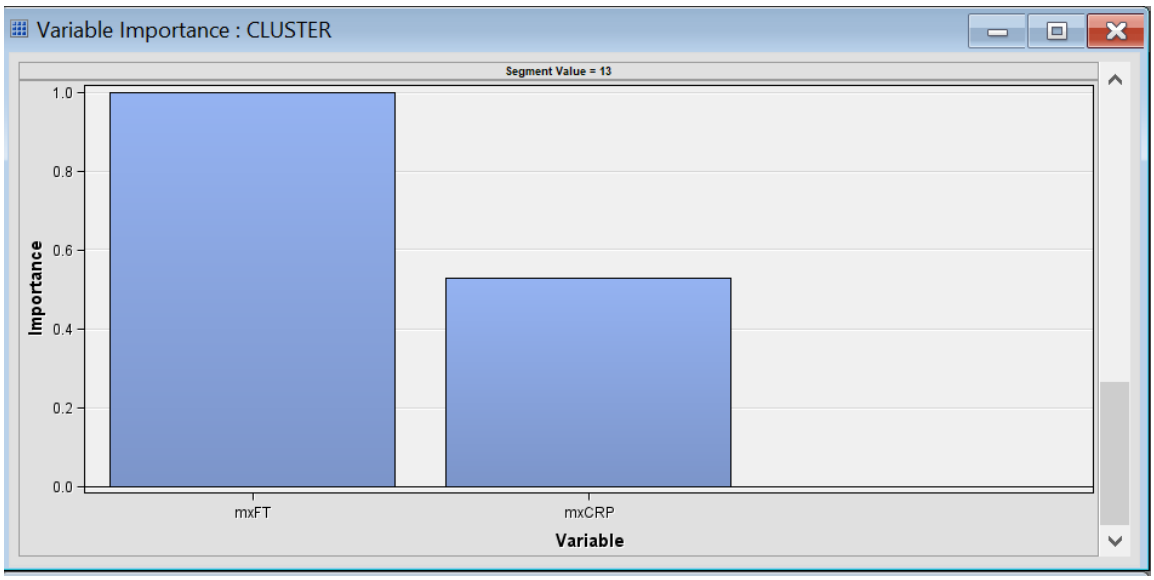

Рисунок 28 — Факторный анализ кластера 13 с высокой «поздней» (с 20 по 30 дни) смертностью

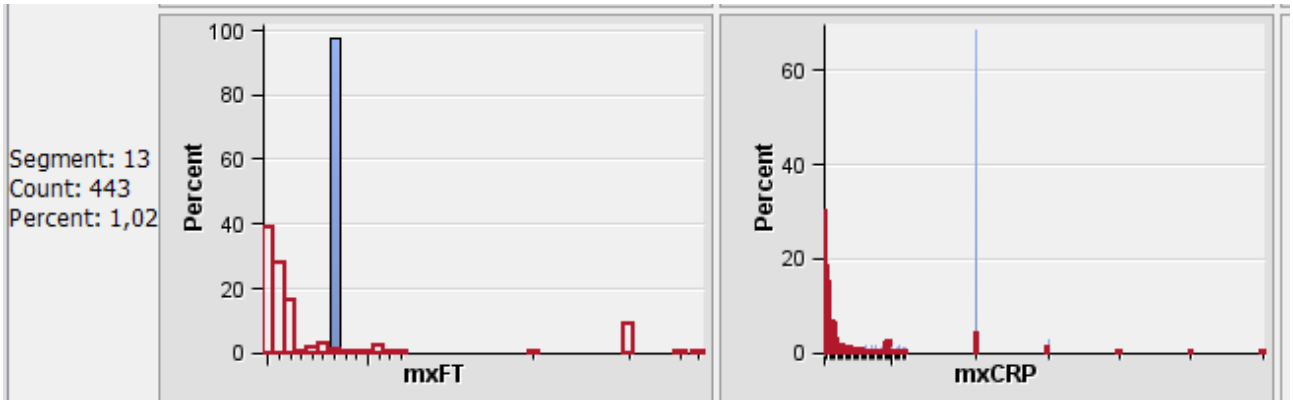

Рисунок 29 — Факторный анализ кластера 13 с высокой «поздней» (с 20 по 30 дни) смертностью

Также, в качестве стратифицирующих переменных можно выбрать возраст, пол, степень тяжести КТ, Ig. Данные исследования, а также соответствующие им графики приведены в предыдущем (3.1.3.1) пункте данного отчета.

3.1.4 Построение описательных моделей прогнозирования летальности с функцией отбора важных предикторов

На вход подаются обработанные данные медицинских анализов пациентов.

Необходимо построить описательные модели прогнозирования летальности с функцией отбора важных предикторов с использованием методов регрессионного анализа и деревьев решений.

Целевой переменной является факт летальности.

Были построены описательные модели прогнозирования летальности с использованием методов регрессионного анализа [15–20] и деревьев решений [21–26], на основе которых были выделены следующие главные особенности и ключевые признаки:

- Чем старше пациент, тем выше вероятность умереть;
- Смертность выше у мужчин;
- Нет зависимости от положительного ПЦР;
- При IgG и IgM выше клинических порогов вероятность умереть ниже.

Было получено, что важными признаками являются:

- Степень тяжести КТ;
- Возраст;
- Средний IgG;
- Минимальные лейкоциты;
- Минимальные тромбоциты;
- Нейтрофилы;
- Средний гемоглобин;
- Уровень С-реактивного белка (СРБ, CRP).

Наибольшую эффективность показали методы на основе деревьев решений (см. Таблица 1).

Таблица 1 — Прогнозирование летальности

Тип модели	**ROC AUC**	**Среднеквадратичная ошибка**
Решающее дерево	0,860	0,074
Регрессия	0,860	0,075

3.1.5 Анализ фактов появления и исчезновения положительного ПЦР

На вход подаются обработанные данные медицинских анализов пациентов.

Необходимо с помощью метода Каплана-Мейера, регрессионных моделей и деревьев решений построить модели прогнозирования факта появления (и исчезновения) во времени положительного ПЦР, выявить ключевые признаки.

Целевой переменной является факт появления положительного ПЦР.

На основе метода Каплана-Мейера, регрессионных моделей и деревьев решений были построены модели прогнозирования факта появления (и исчезновения) во времени положительного ПЦР.

Были выделены следующие ключевые признаки:

- КТ;
- Пол;
- Количество нейтрофилов;
- Количество лейкоцитов (WBC);
- Возраст;
- Значение Гемоглобина (HGB);
- Уровень С-реактивного белка (CRB);
- Количество тромбоцитов (PLT);
- D-димер;
- IgM.

Наилучшее качество работы продемонстрировали решающие деревья (см. Таблица 2).

Таблица 2 — Прогнозирование факта появления (и исчезновения) во времени положительного ПЦР

Название метода	**ROC AUC**	**Среднеквадратичная ошибка**
Решающие деревья	0,844	0,156
Регрессия	0,843	0,160

На основе полученной предиктивной аналитики для практического здравоохранения были сформированы таблицы сочетания признаков, позволяющие прогнозировать характер течения COVID-19 у конкретного пациента с учетом дня заболевания (см. Рисунки 30–32). Временные интервалы представления комбинаций признаков определены следующим образом: интервал до 7 дня от начала заболевания, 7-14 день, 15-50 день. В соответствии с имеющимися данными, можно сделать вывод, что в течение первой недели с момента появления симптомов, ПЦР является положительной практически у всех пациентов. С течением времени число выявляемых случаев заболевания резко снижается при легком и среднетяжелом течении. Персистенция положительных результатов ПЦР при тяжелом

течении заболевания может быть связана с отсутствием или поздней генерацией специфических нейтрализующих антител. Наряду с признанной в мировом сообществе информативностью данных компьютерной томографии легких для прогнозирования течения заболевания, нами была предложена комбинация лабораторных факторов, характеризующихся прогностической значимостью в случае тяжелого течения COVID-19. Так, повышение значений С-реактивного белка и Д-димера в течение первых 7 дней от начала заболевания, нарастание уровня ферритина, нейтрофилеза и тромбоцитопении с 7 по 14 день заболевания ассоциировано с высоким риском тяжелого течения заболевания с развитием острого респираторного дистресс-синдрома. Данный подход позволяет провести адекватную маршрутизацию пациента и стратегию превентивных действий.

ПЦР +/-	IgM +/-	IgG +/-	D-dim ↑ +/-	CRP ↑ +/-	Ферритин ↑ +/-	Трц ↑ +/-	Трц ↓ +/-	Нейтр ↑ +/-	Нейтр ↓ +/-	КТ 1-2 +/-	КТ 3-4 +/-	ТЕЧЕНИЕ
+	-	-	-	+	-	+/-	+/-	+/-	+/-	+	-	Легкое
+	-	-	+/-	+	+	+	-	+/-	+/-	+	-	Средней тяжести
+	-	-	+	++	+	+	-	+	-	-	+	Тяжелое/критическое

Рисунок 30 — Сочетание показателей для интервала до 7 сут от начала заболевания (↑-cut-off верхний для показателя, ↓- cut-off нижний для показателя)

ПЦР +/-	IgM +/-	IgG +/-	D-dim ↑ +/-	CRP ↑ +/-	Ферритин ↑ +/-	Трц ↑ +/-	Трц ↓ +/-	Нейтр ↑ +/-	Нейтр ↓ +/-	КТ 1-2 +/-	КТ 3-4 +/-	ТЕЧЕНИЕ
+/-	+	+/-	-	+	-	+	+/-	+/-	+/-	+	-	Легкое
+/-	+	+/-	+/-	+	-	+	-	+/-	+/-	+	-	Средней тяжести
+	+	-	+	++	+	-	+	+	-	-	+	Тяжелое/критическое

Рисунок 31 — Сочетание показателей для интервала 7-14 сут от начала заболевания (↑-cut-off верхний для показателя, ↓- cut-off нижний для показателя)

ПЦР +/-	IgM +/-	IgG +/-	D-dim ↑ +/-	CRP ↑ +/-	Ферритин ↑ +/-	Трц ↑ +/-	Трц ↓ +/-	Нейтр ↑ +/-	Нейтр ↓ +/-	КТ 1-2 +/-	КТ 3-4 +/-	ТЕЧЕНИЕ
-	+/-	+	-	-	-	+	+/-	+/-	+/-	+	-	Легкое
-	+	++	+	+	-	+	-	+/-	+/-	+	+	Средней тяжести
+	+/-	+/-	+	++	+	-	+	+	-	-	+	Тяжелое/критическое

Рисунок 32 — Сочетание показателей для интервала 15-50 сут от начала заболевания (↑-cut-off верхний для показателя, ↓- cut-off нижний для показателя)

3.1.6 Анализ динамики появления и изменения иммуноглобулинов IgM и IgG

На вход подаются обработанные данные медицинских анализов пациентов.

Необходимо с помощью метода Каплана-Мейера, регрессионных моделей и деревьев решений провести анализ динамики появления и изменения иммуноглобулинов IgM и IgG, выявить ключевые признаки

Целевой переменной является факт появления и изменения иммуноглобулинов IgM и IgG.

Для выполнения задач данного подраздела была проведена фильтрация данных: были отобраны пациенты только с тремя и более ИФА-анализами.

В итоге для дальнейших наблюдений остались данные 8 500 пациентов. Заметим, что со временем число пациентов уменьшается (умирают или выздоравливают и больше не обследуются) (см. Рисунок 33).

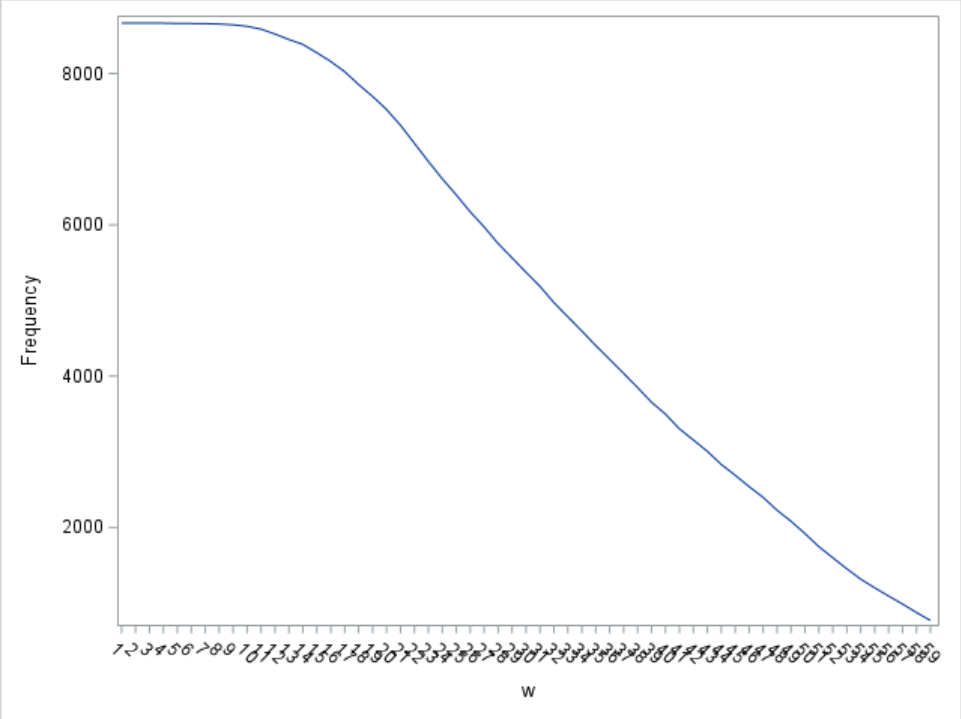

Рисунок 33 — Уменьшение числа контролируемых пациентов во времени

Число положительных ПЦР и ИФА в выборке по времени в процентах представлено на Рисунке 34.

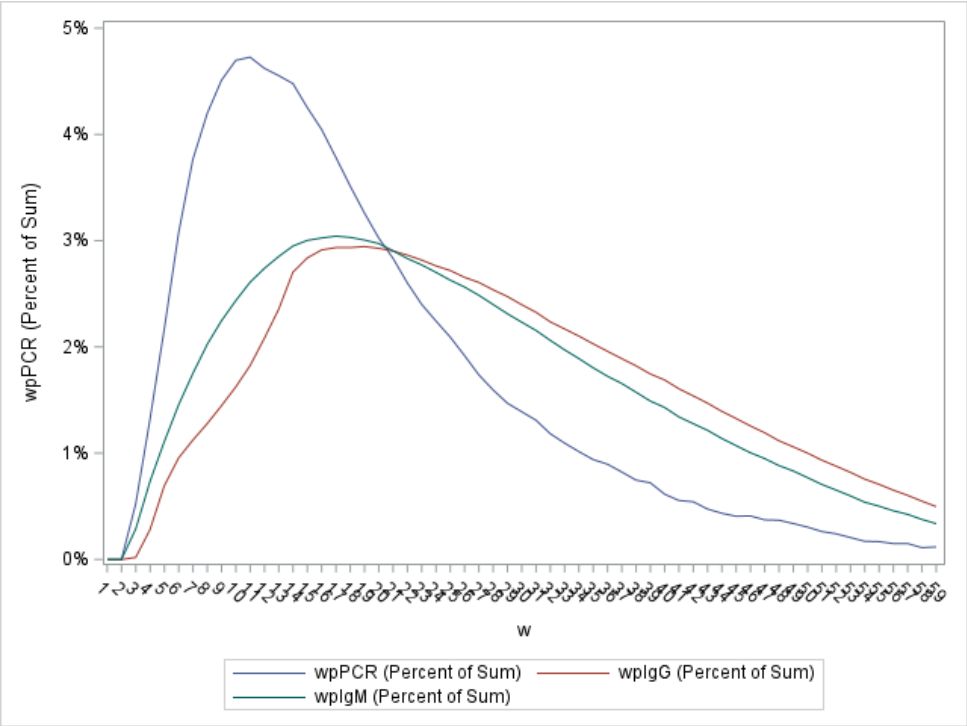

Рисунок 34 — Число положительных ПЦР и ИФА в выборке по времени в процентах

При отдельном рассмотрении положительных ИФА и ПЦР по возрастным группам были получены следующие графики (см. Рисунки 35–37).

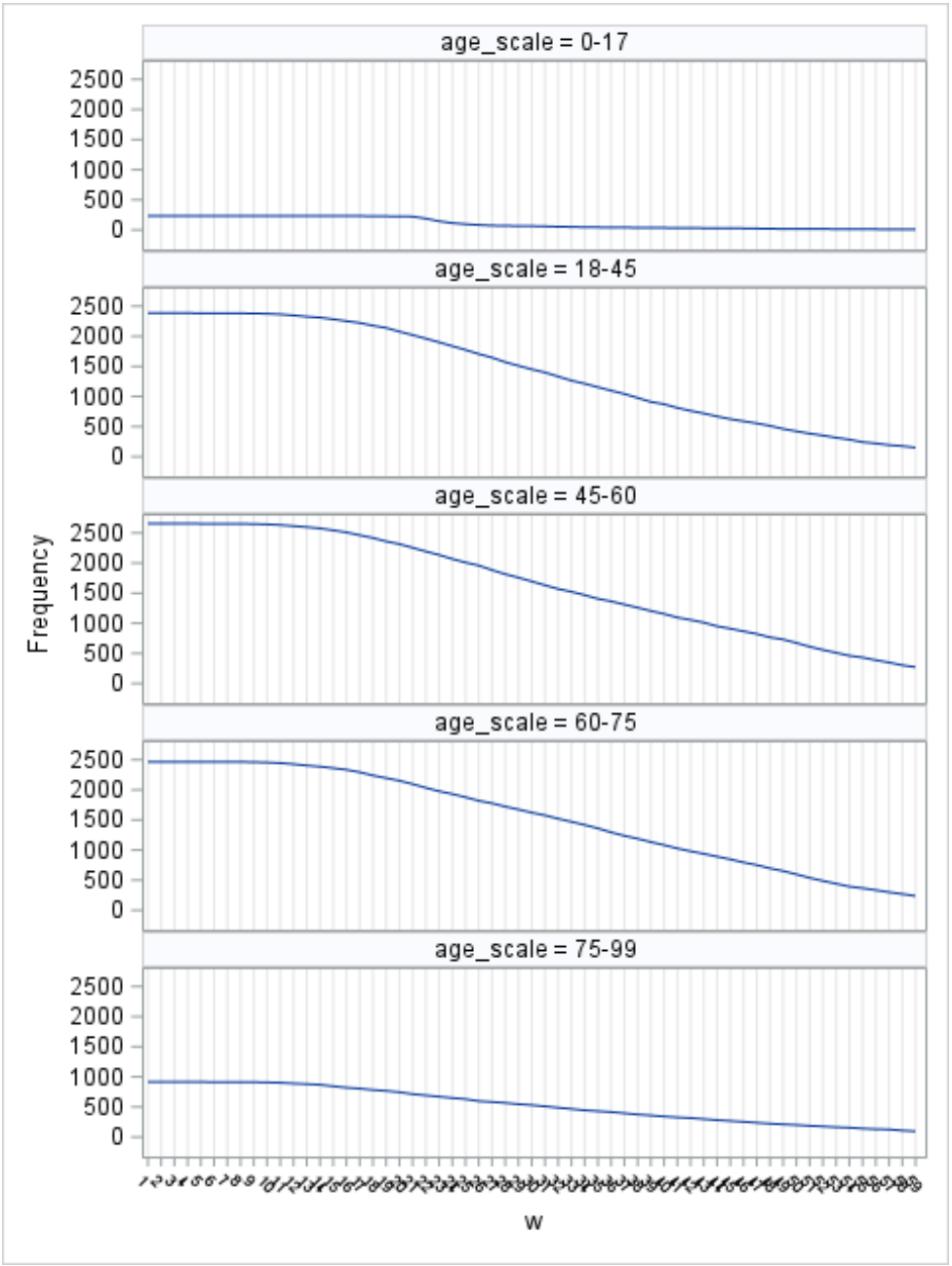

Рисунок 35 — Число пациентов во времени

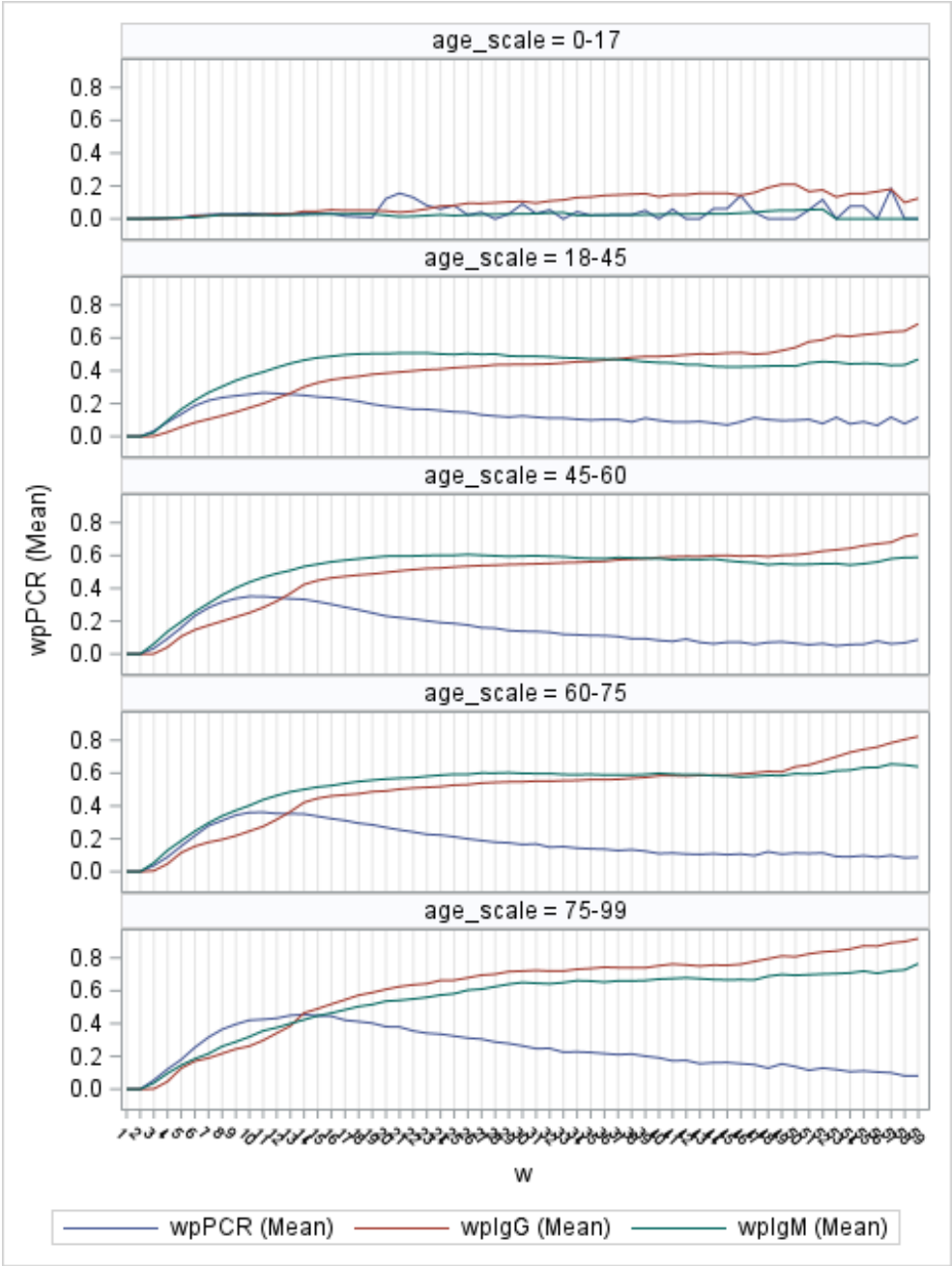

Рисунок 36 — Доля положительных пациентов в выборке по времени

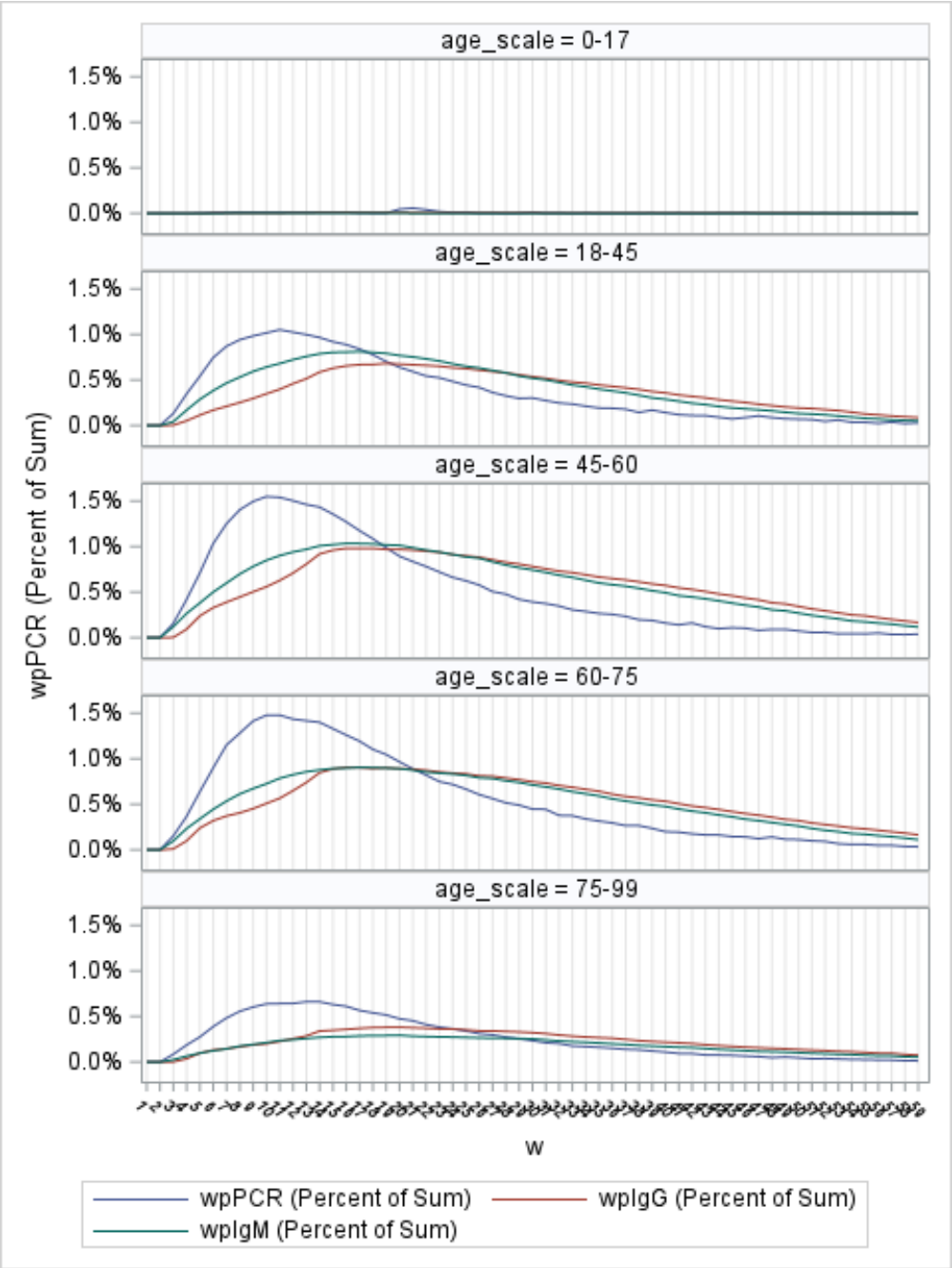

Рисунок 37 — Процент положительных пациентов в выборке

3.1.6.1 Использование метода Каплана-Мейера для построения моделей прогнозирования факта появления пороговых значений IgG и IgM

Был проведен анализ факта появления пороговых значений IgG и IgM с использованием метода Каплана-Мейера. Подробнее данный процесс описан в разделах 3.1.6.2 и 3.1.6.3 данного документа.

3.1.6.2 Выявление ключевых признаков и стратифицирующих признаков, влияющих на факт появления и исчезновения во времени пороговых значений IgG и IgM

В рамках исследования в качестве порогового значения для IgM рассматривалось значение 1, а для IgG – 10.

Для IgM были построены следующие графики распределения:

- Распределение дней до появления IgM>1 (если вообще появляется) (см. Рисунок 38);
- Распределение дней до появления IgM>1 (с учетом возраста) (см. Рисунок 39);
- Распределение дней до появления IgM>1 (с учетом пола) (см. Рисунок 40).

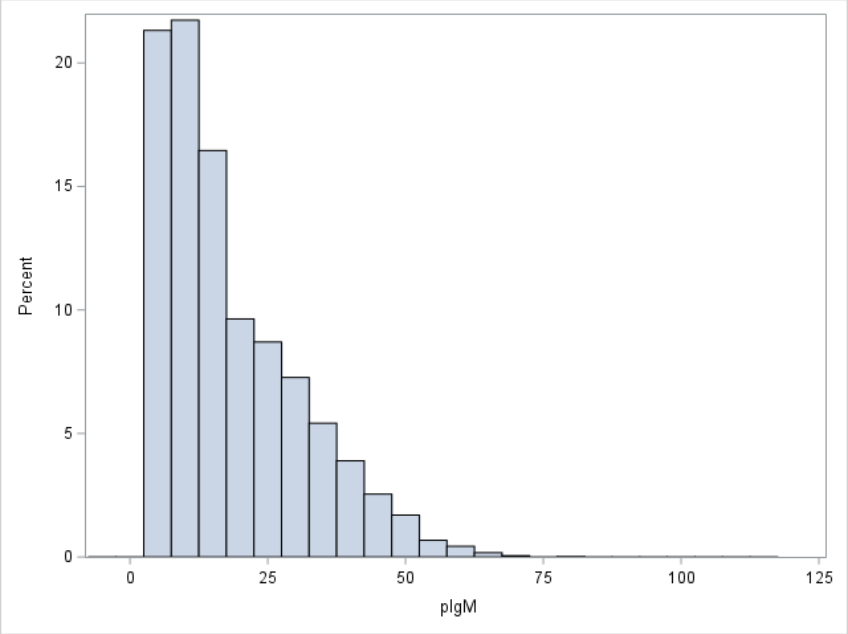

Рисунок 38 — Распределение дней до появления IgM>1 (если вообще появляется)

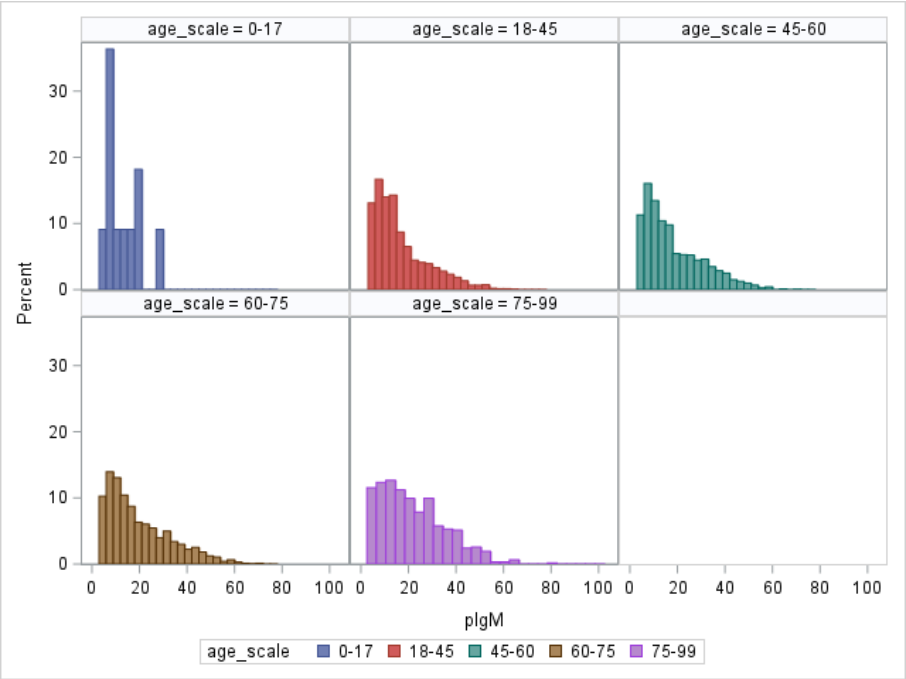

Рисунок 39 — Распределение дней до появления IgM>1 (с учетом возраста)

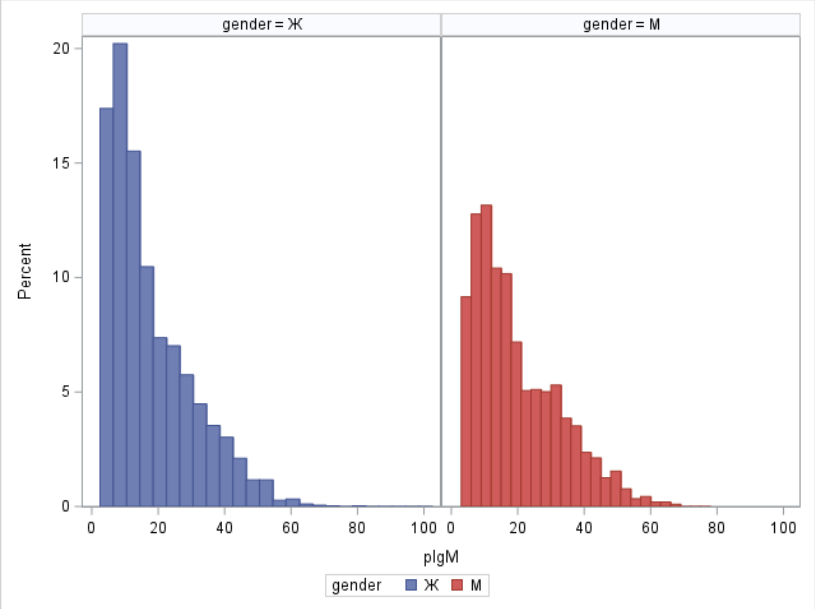

Рисунок 40 — Распределение дней до появления IgM>1 (с учетом пола)

Был построен график вероятности (см. Рисунок 41) не получить IgM>1 от времени (у 80% он появляется):

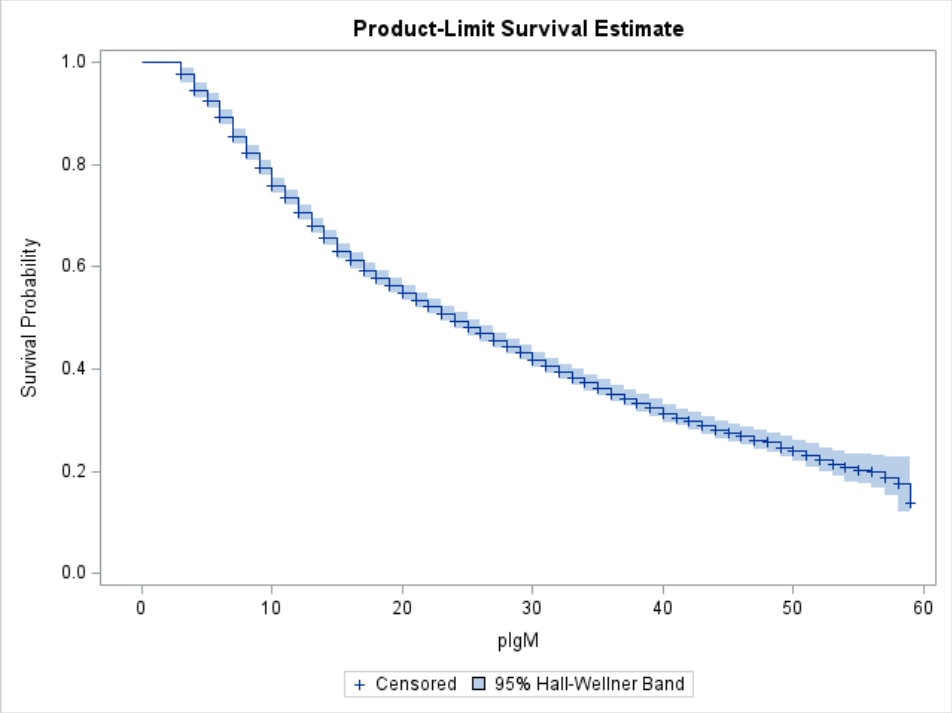

Рисунок 41 — График вероятности не получить IgM>1 от времени

Соответствующие оценки теста Вилкоксона приведены на Рисунке 42:

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
kt	1	294.9	<.0001	294.9	<.0001
mxIgG	2	372.8	<.0001	77.8766	<.0001
mnIgG	3	468.3	<.0001	95.4825	<.0001
mxNEUT	4	526.5	<.0001	58.2521	<.0001
avIgG	5	561.2	<.0001	34.6994	<.0001
mnHGB	6	590.6	<.0001	29.3818	<.0001
mxHGB	7	600.9	<.0001	10.3368	0.0013
mnCRP	8	611.3	<.0001	10.3649	0.0013
mnPLT	9	618.7	<.0001	7.4234	0.0064
cpPCR	10	623.6	<.0001	4.8322	0.0279

Рисунок 42 — Оценки теста Вилкоксона

На основе полученных оценок были выявлены следующие ключевые признаки, влияющие на появление в крови пороговых значений IgM:

- **КТ;**
- **IgG;**
- **Нейтрофилы,**

- **Гемоглобин,**
- **С-реактивный белок,**
- **Тромбоциты,**
- **Положительный ПЦР**
- **Нет возраста!**

По результатам предварительного анализа, в качестве стратифицирующих признаков были выделены следующие:

- Возраст (и пол);
- КТ;
- Положительный ПЦР;
- Среднее значение IgG.

Для IgG были построены следующие графики распределения:

- Распределение дней до появления IgG>10 (если вообще появляется) (см. Рисунок 43);
- Распределение дней до появления IgG>10 (с учетом возраста) (см. Рисунок 44);
- Распределение дней до появления IgG>10 (с учетом пола) (см. Рисунок 45).

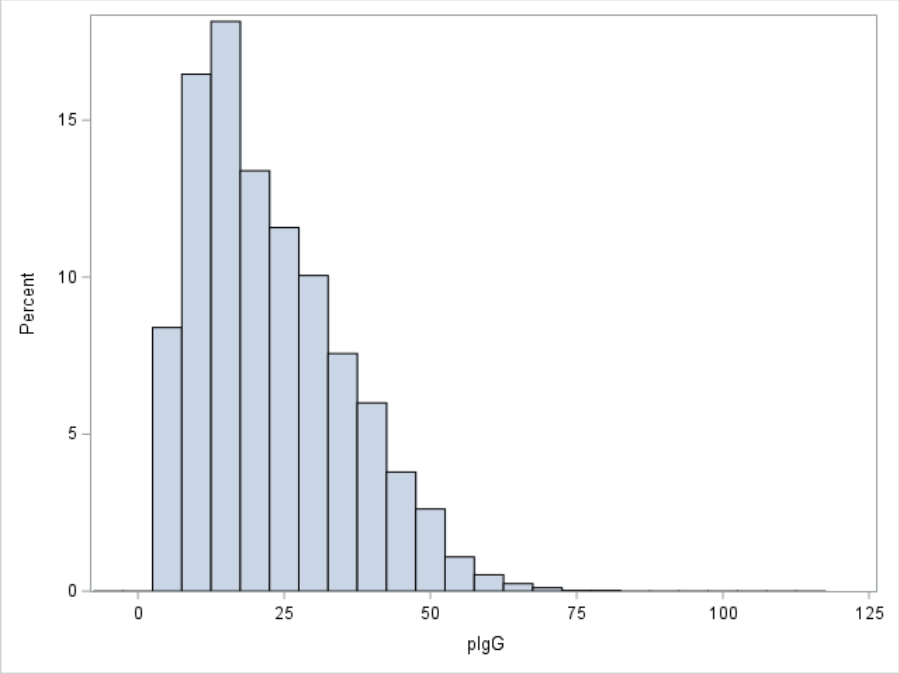

Рисунок 43 — Распределение дней до появления IgG>10 (если вообще появляется)

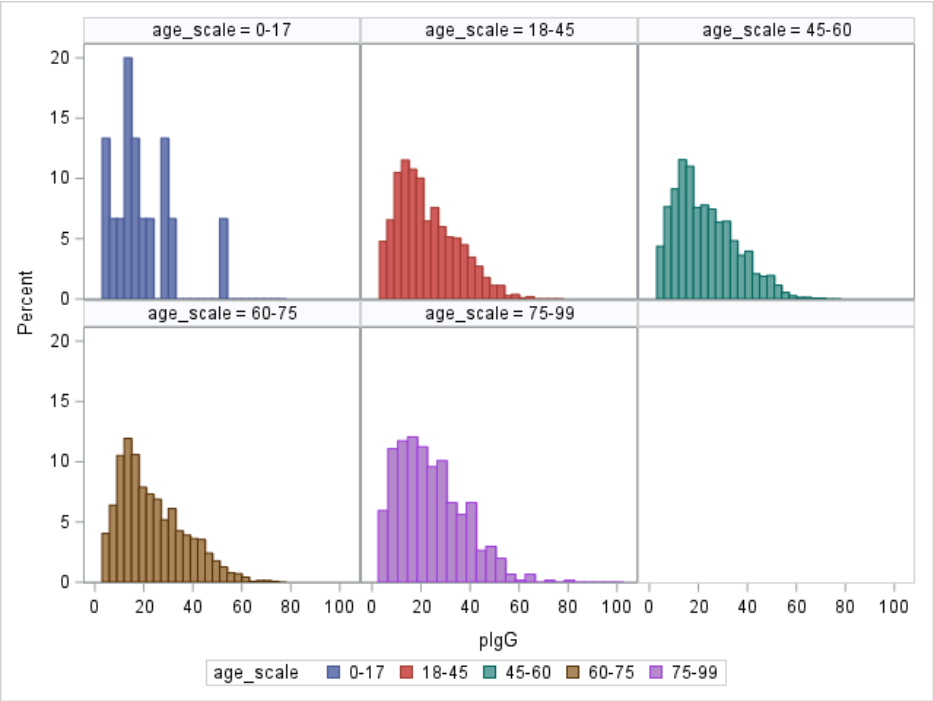

Рисунок 44 — Распределение дней до появления IgG>10 (с учетом возраста)

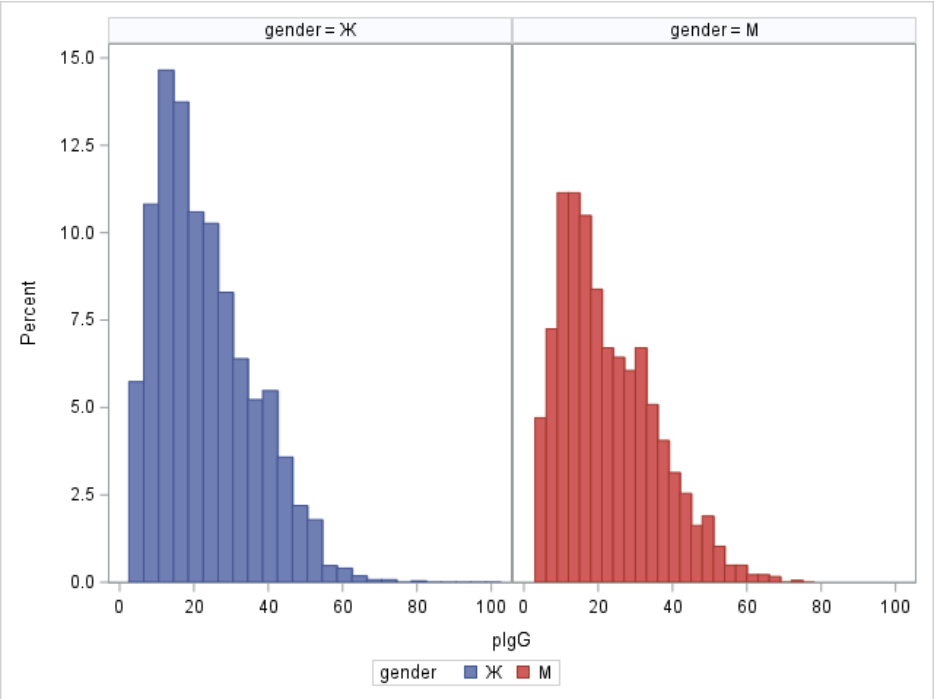

Рисунок 45 — Распределение дней до появления IgG>10 (с учетом пола)

Был построен график вероятности не получить IgG>10 от времени (у 80% он появляется), см. Рисунок 46:

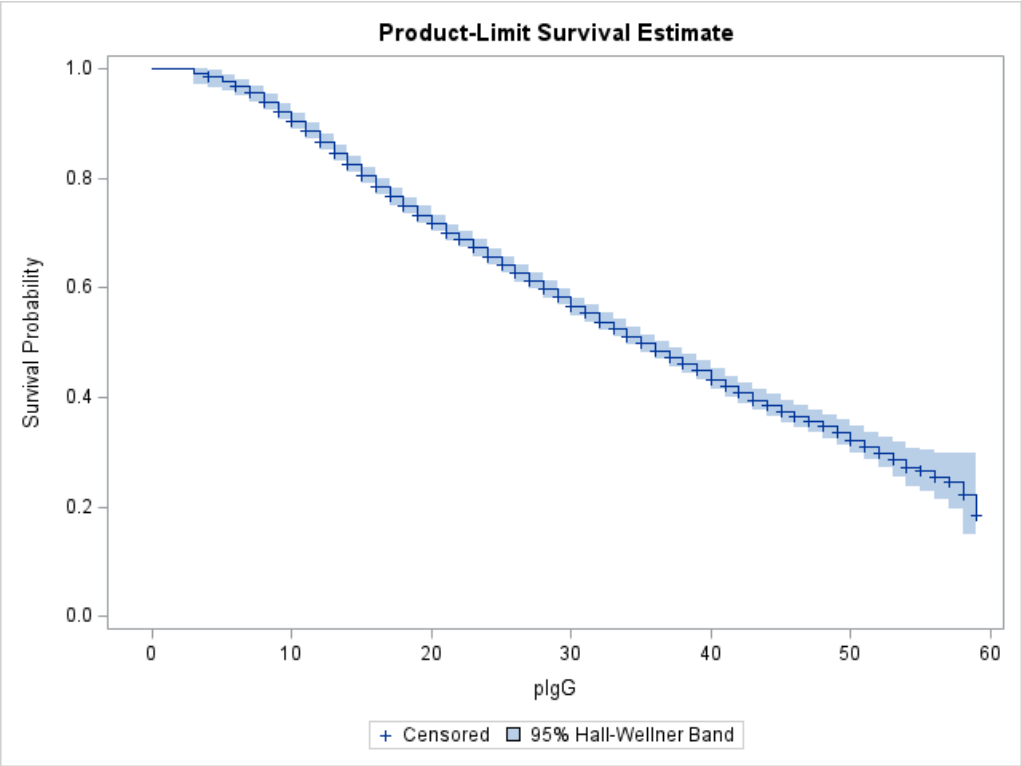

Рисунок 46 — График вероятности не получить IgG>10 от времени

Соответствующие оценки теста Вилкоксона приведены на Рисунке 47:

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
cpPCR	1	2399.8	<.0001	2399.8	<.0001
mxIgM	2	3070.4	<.0001	670.7	<.0001
kt	3	3390.5	<.0001	320.1	<.0001
mnIgM	4	3448.4	<.0001	57.8672	<.0001
mnHGB	5	3470.5	<.0001	22.0782	<.0001
avIgM	6	3481.6	<.0001	11.1513	0.0008
mxNEUT	7	3492.1	<.0001	10.4219	0.0012
mxPLT	8	3499.0	<.0001	6.9205	0.0085
age	9	3504.5	<.0001	5.4800	0.0192

Рисунок 47 — Оценки теста Вилкоксона

На основе полученных оценок были выявлены следующие ключевые признаки, влияющие на появление в крови пороговых значений IgG:

- **КТ;**
- **IgM;**

- **нейтрофилы;**
- **гемоглобин;**
- **возраст;**
- **тромбоциты;**
- **положительный ПЦР.**

По результатам предварительного анализа, в качестве стратифицирующих признаков были выделены следующие:

- Возраст (и пол);
- КТ;
- Положительный ПЦР;
- Среднее значение IgM.

3.1.6.3 Построение аналогичных моделей с выявлением ключевых признаков внутри страт для прогнозирования появления и исчезновения во времени пороговых значений IgG и IgM

Были построены графики формирования IgM (порог >1) во времени по возрасту, полу, КТ, ПЦР, IgG (см. Рисунки 48–52).

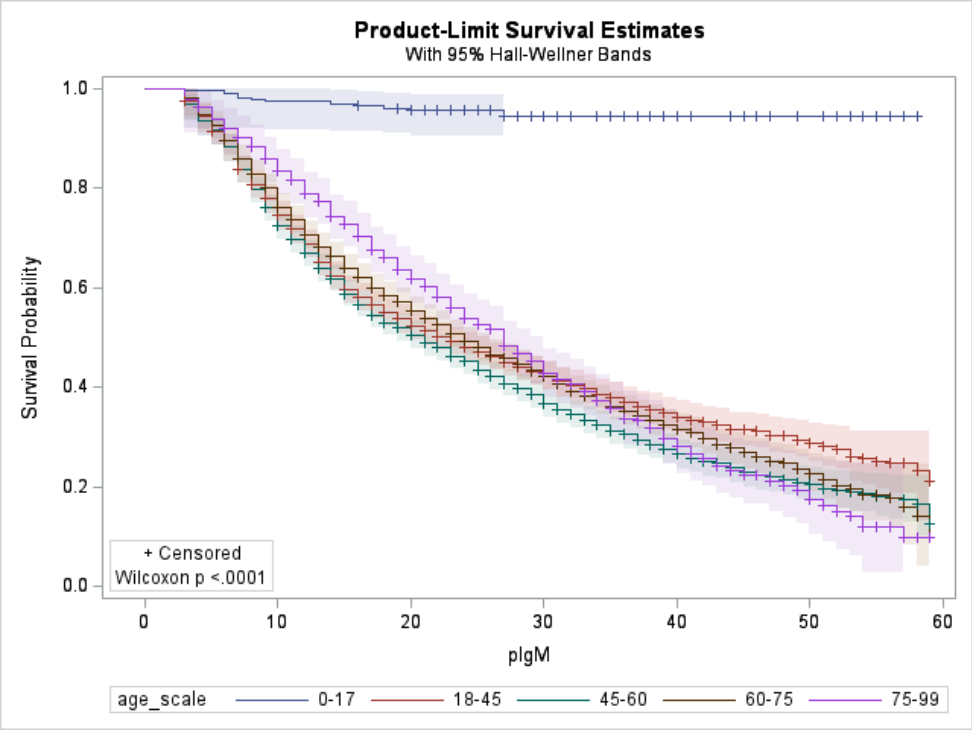

Рисунок 48 — График формирования IgM (порог >1) во времени по возрасту

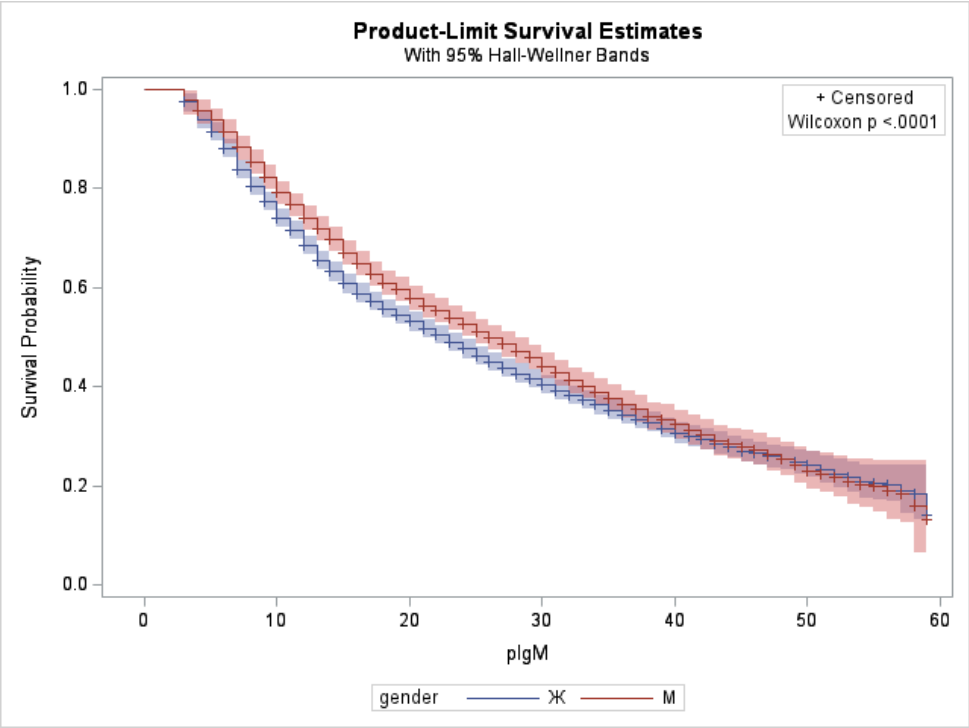

Рисунок 49 — График формирования IgM (порог >1) во времени по полу

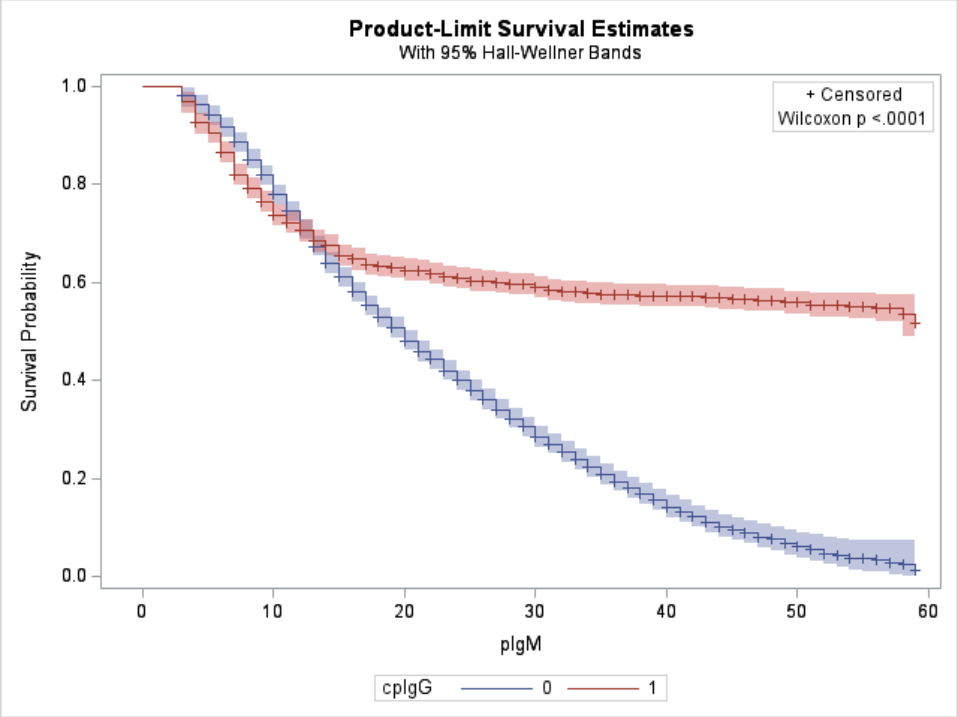

Рисунок 50 — График формирования IgM (порог >1) во времени по IgG

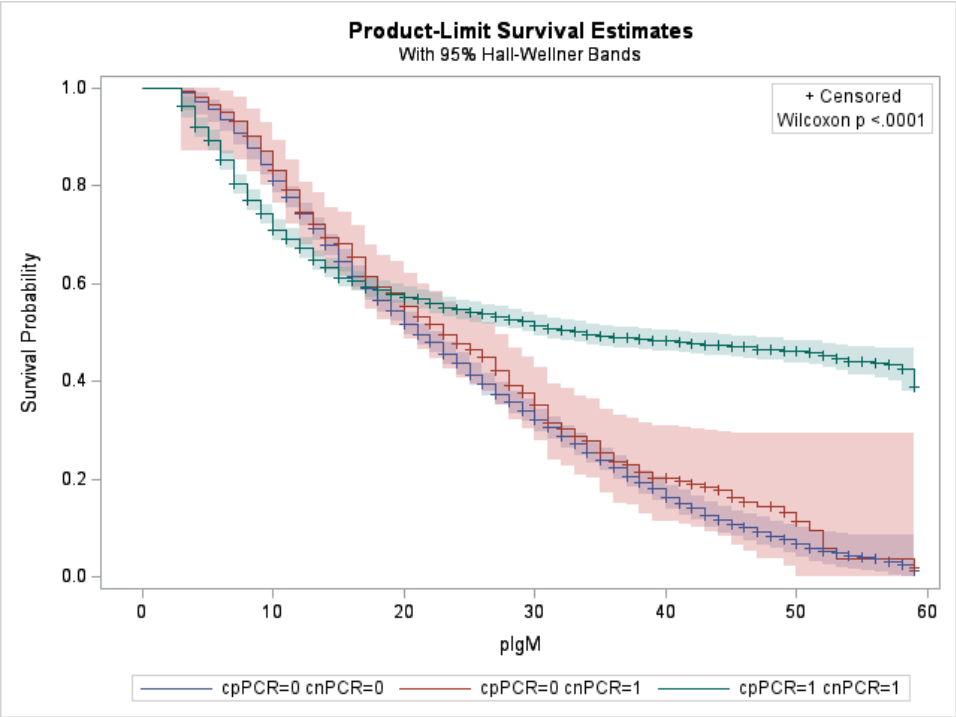

Рисунок 51 — График формирования IgM (порог >1) во времени по ПЦР

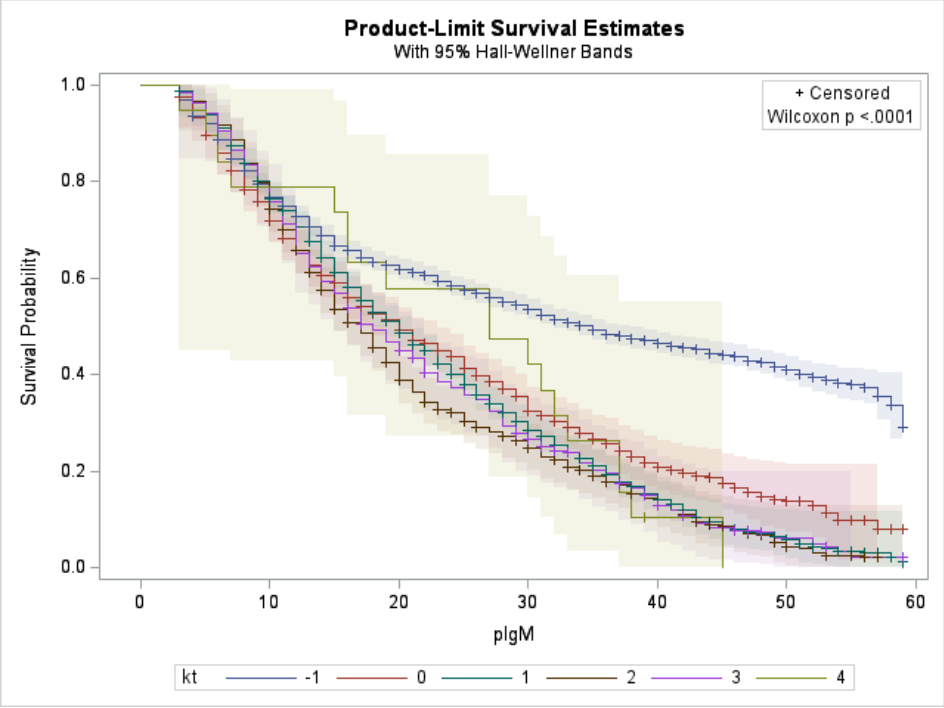

Рисунок 52 — График формирования IgM (порог >1) во времени по КТ

В результате анализа полученных графиков были получены следующие выводы:

- **У детей IgM почти не формируется, у остальных возрастов почти одинаково;**

- **У женщин IgM формируется быстрее, чем у мужчин;**
- **IgM коррелирует с положительным ПЦР, степенью КТ и IgG>10.**

Также были построены графики формирования IgM (порог >1) во времени по результатам значимых лабораторных исследований (см. Рисунки 53–57)

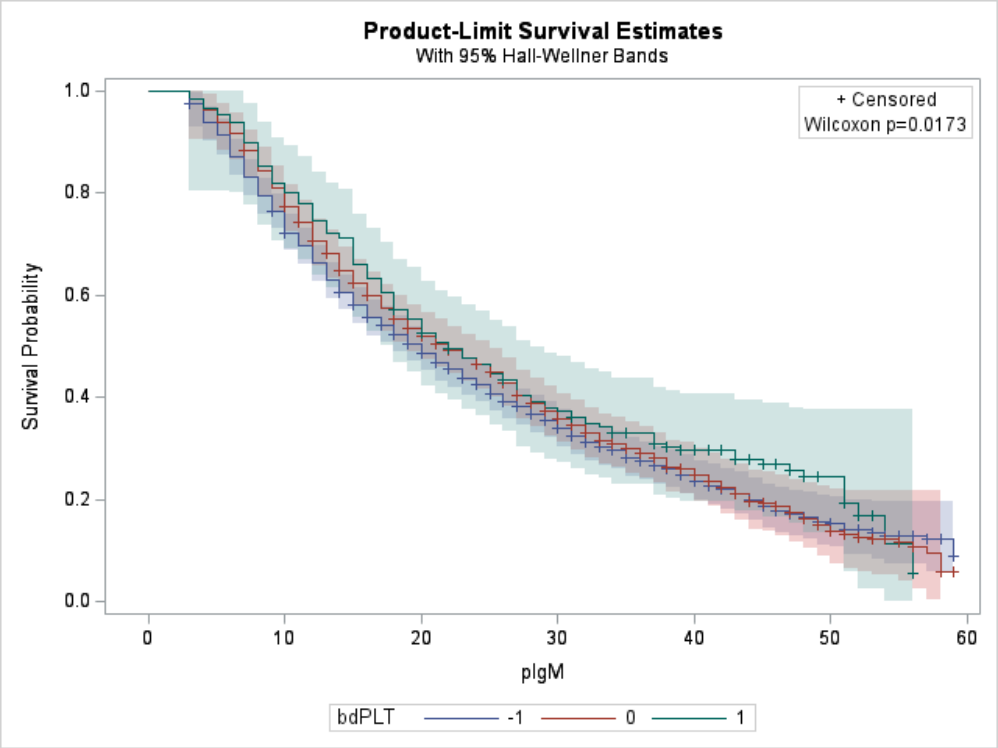

Рисунок 53 — График формирования IgM (порог >1) во времени по результатам значимых лабораторных исследований (тромбоциты)

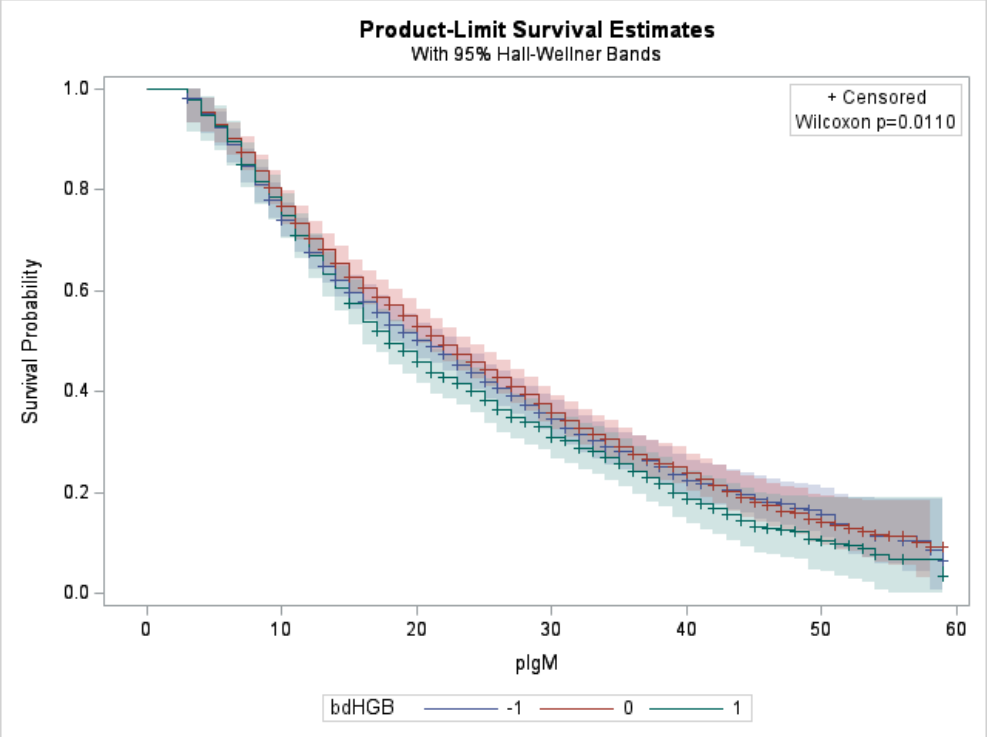

Рисунок 54 — График формирования IgM (порог >1) во времени по результатам значимых лабораторных исследований (гемоглобин)

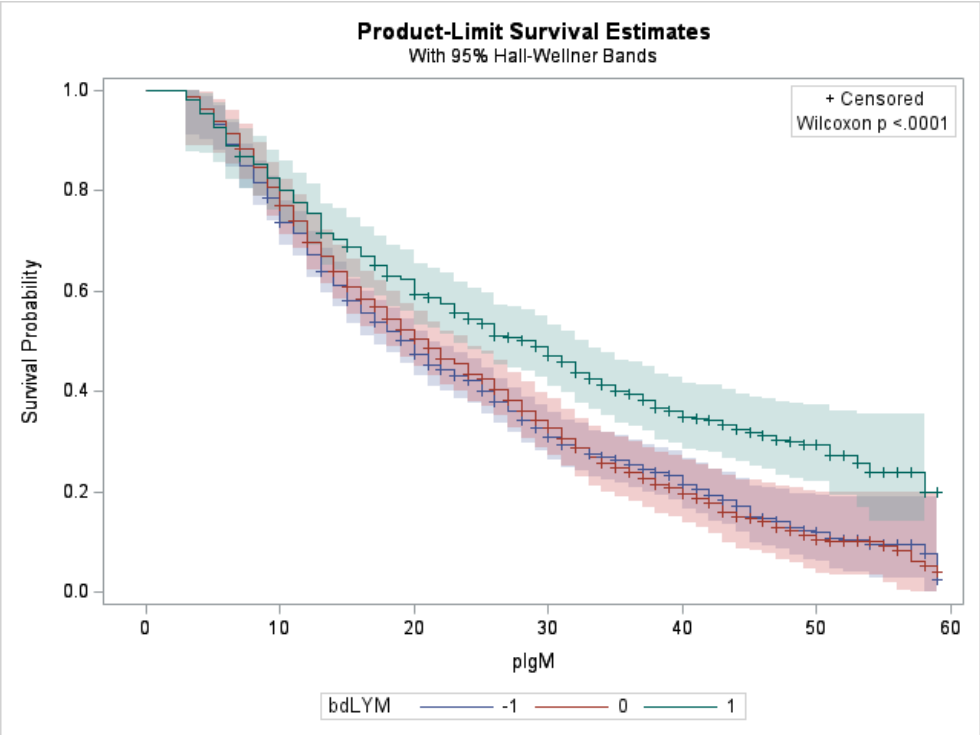

Рисунок 55 — График формирования IgM (порог >1) во времени по результатам значимых лабораторных исследований (лимфоциты)

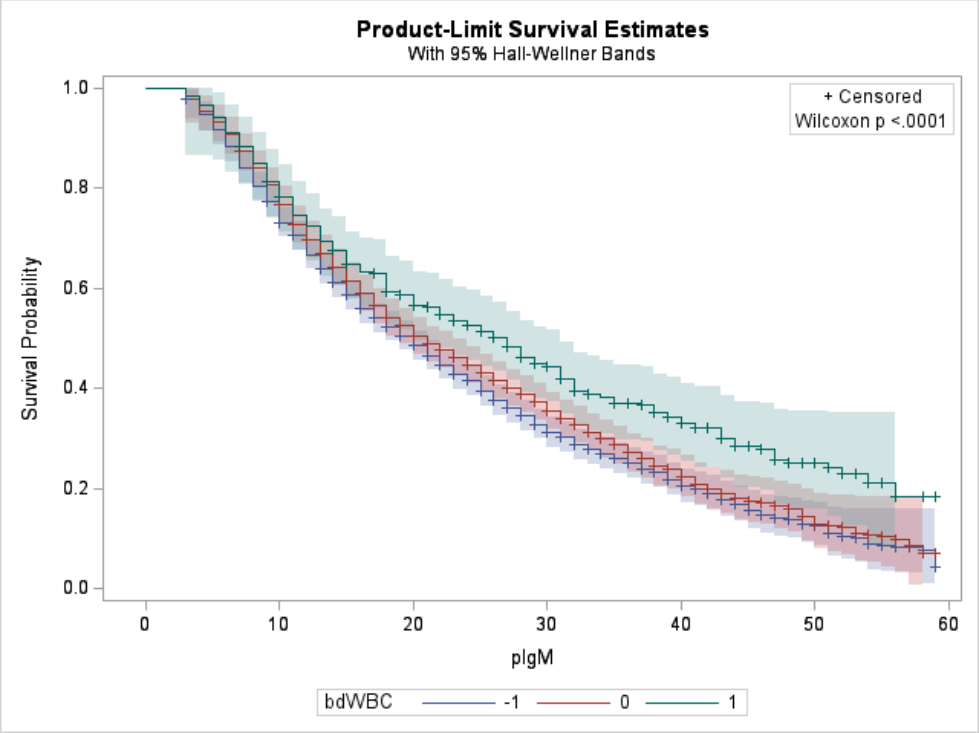

Рисунок 56 — График формирования IgM (порог >1) во времени по результатам значимых лабораторных исследований (лейкоциты)

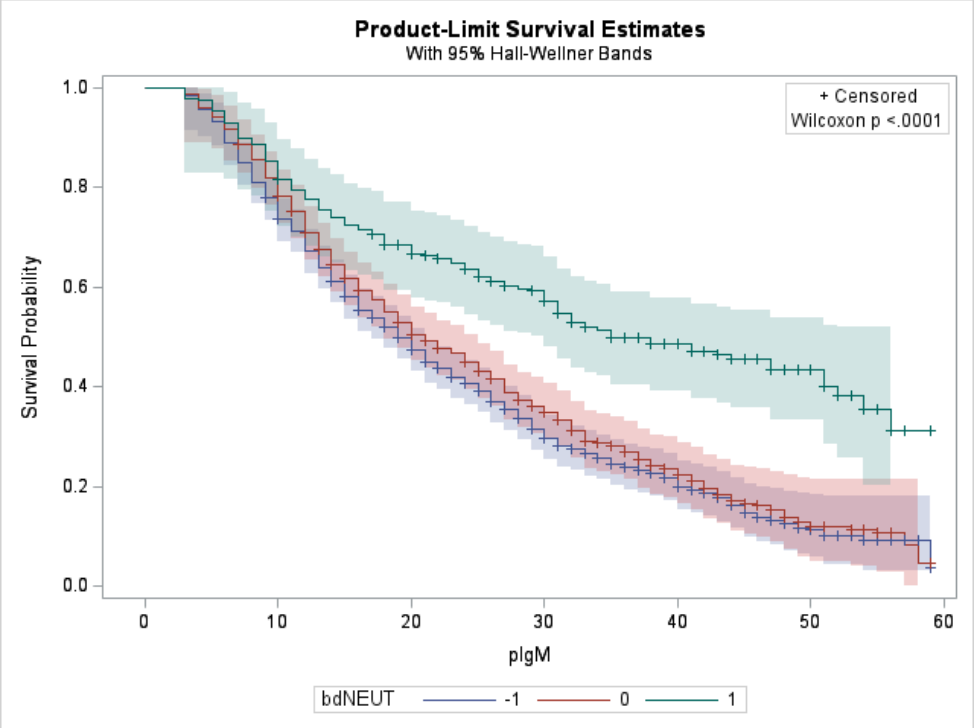

Рисунок 57 — График формирования IgM (порог >1) во времени по результатам значимых лабораторных исследований (нейтрофилы)

В результате анализа полученных графиков были получены следующие выводы:

- **Пониженные тромбоциты или повышенный гемоглобин – раньше IgM>1;**
- **Повышенные нейтрофилы или лейкоциты или лимфоциты – позже и менее вероятно IgM>1.**

Были построены графики формирования IgG (порог >10) во времени по возрасту, полу, КТ, ПЦР, IgM (см. Рисунки 58–62).

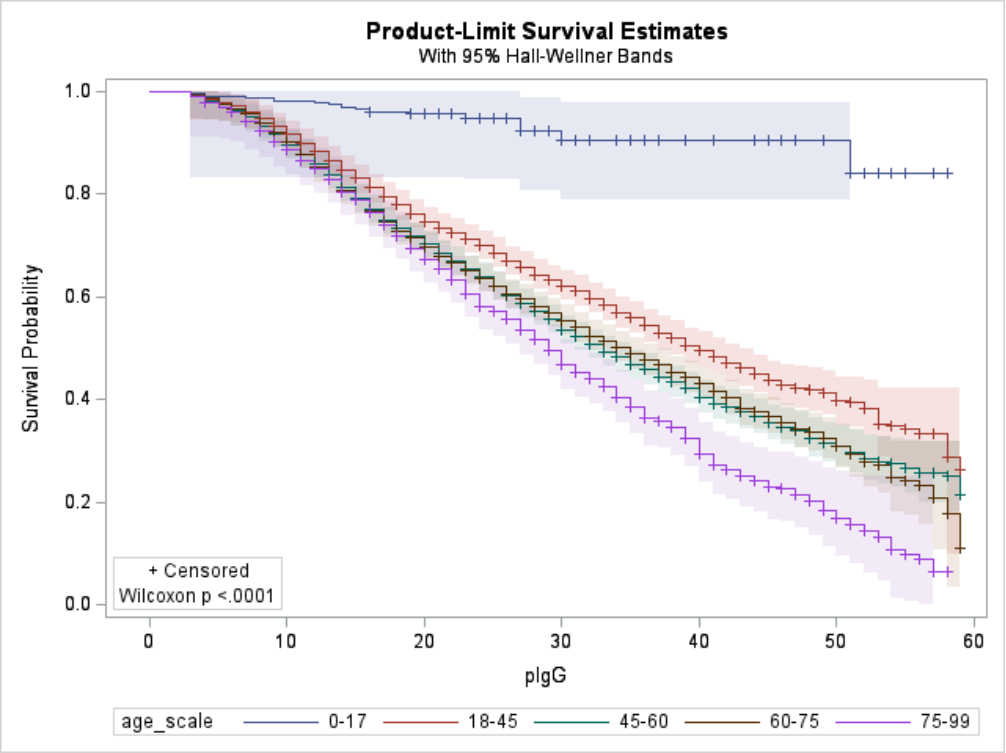

Рисунок 58 — График формирования IgG (порог >10) во времени по возрасту

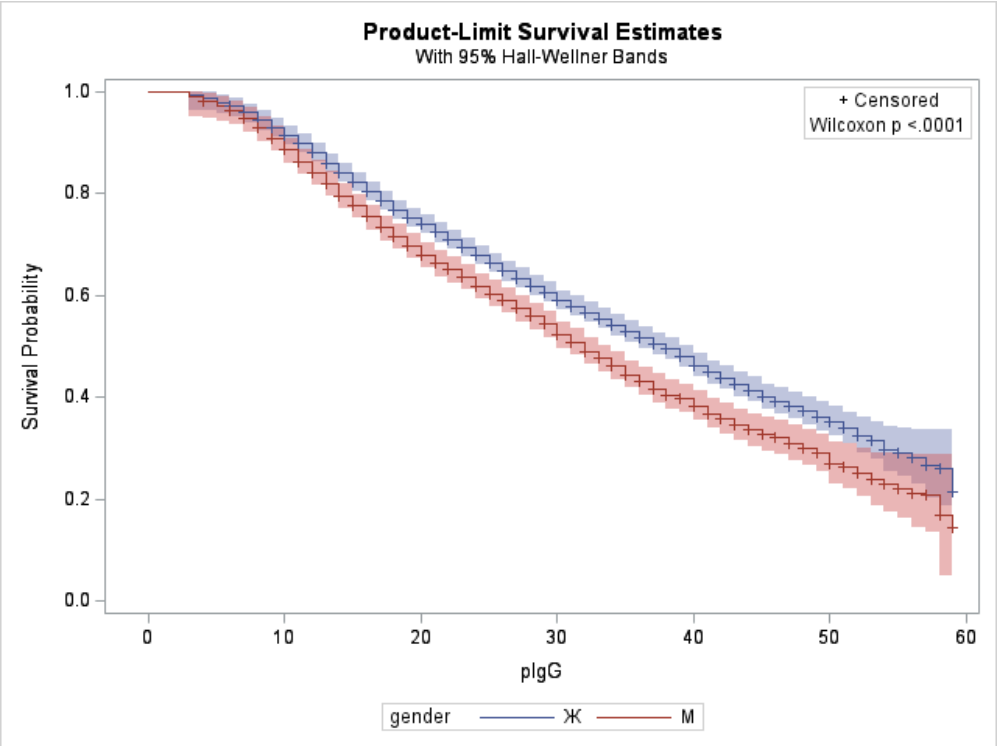

Рисунок 59 — График формирования IgG (порог >10) во времени по полу

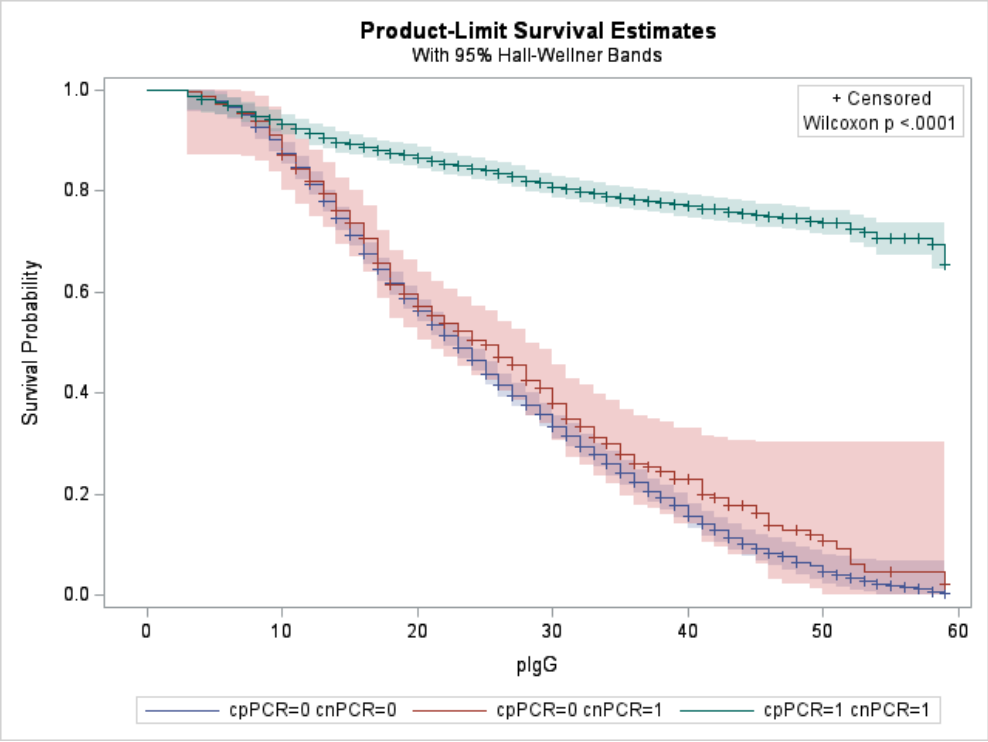

Рисунок 60 — График формирования IgG (порог >10) во времени по ПЦР

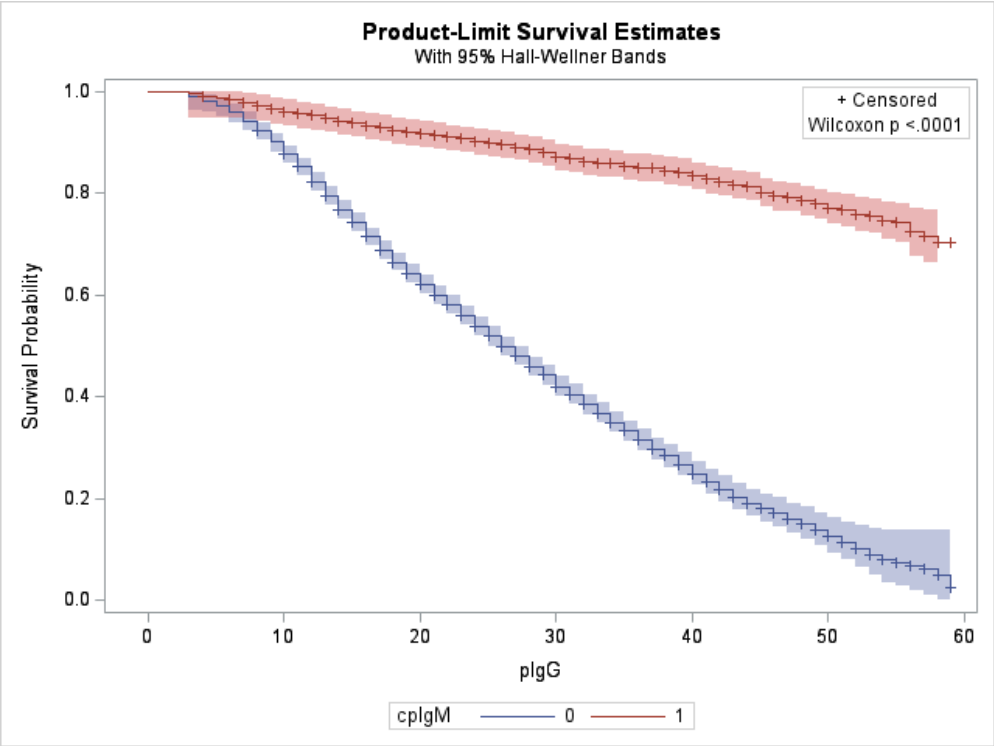

Рисунок 61 — График формирования IgG (порог >10) во времени по IgM

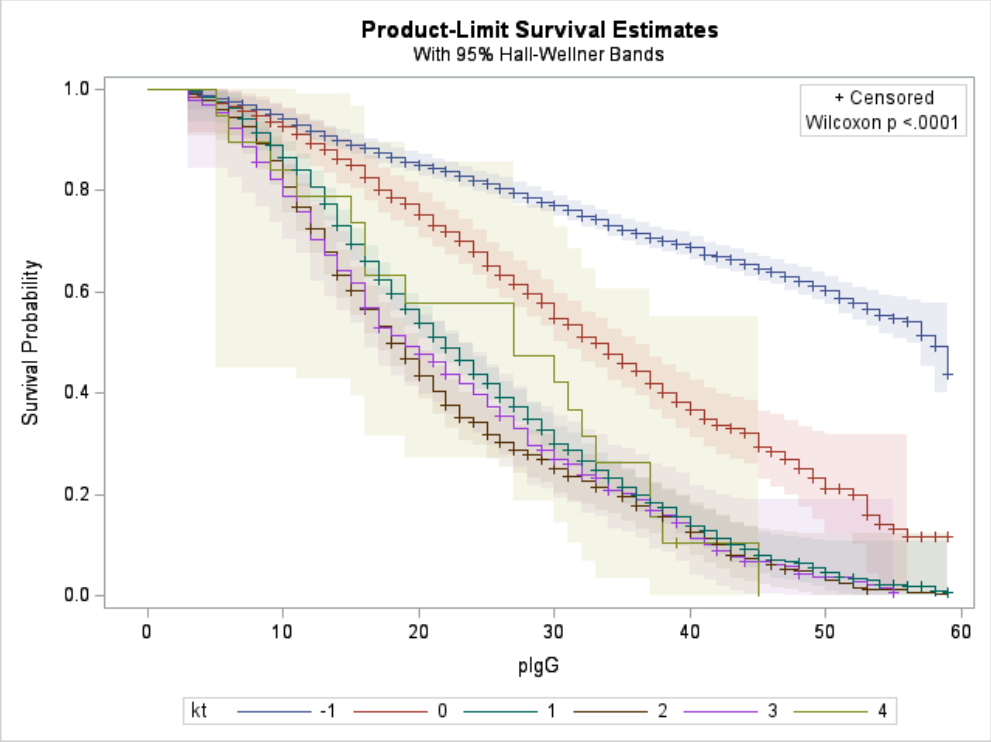

Рисунок 62 — График формирования IgG (порог >10) во времени по КТ

В результате анализа полученных графиков были получены следующие выводы:

- **У детей IgG слабо формируется, дальше чем старше, тем сильнее и быстрее;**
- **У мужчин IgG формируется быстрее и сильнее, чем у женщин;**
- **IgG коррелирует с положительным ПЦР, IgM>1 и степенью поражения по КТ больше 0.**

Также были построены графики формирования IgG (порог >10) во времени по результатам значимых лабораторных исследований (см. Рисунки 63–68).

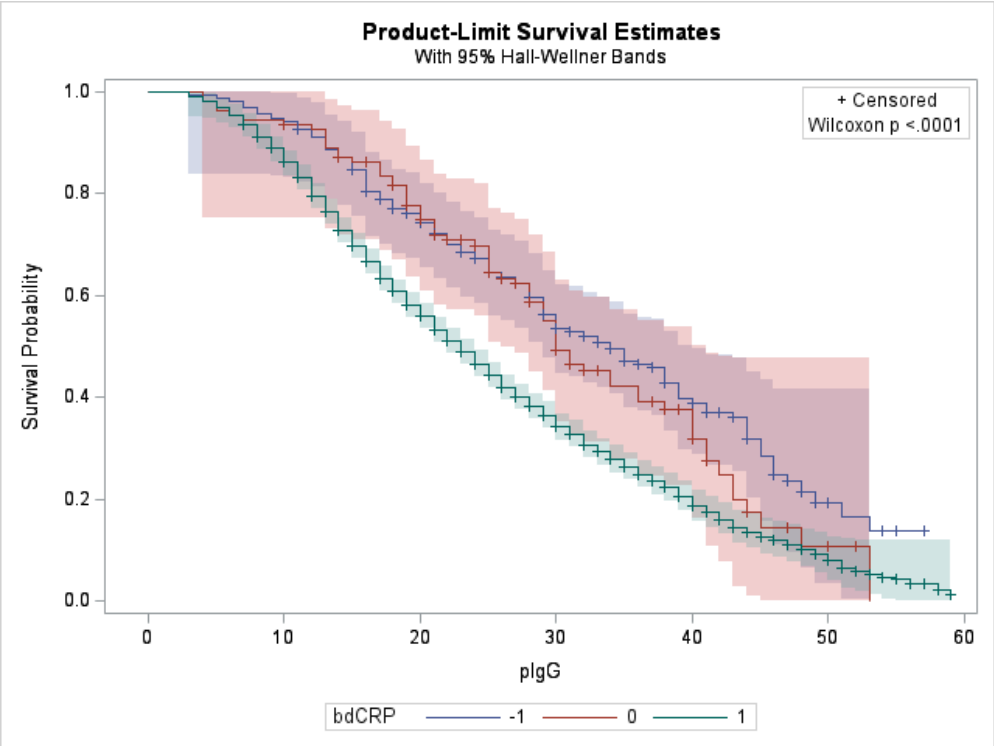

Рисунок 63 — График формирования IgG (порог >10) во времени по результатам значимых лабораторных исследований (с-реактивный белок)

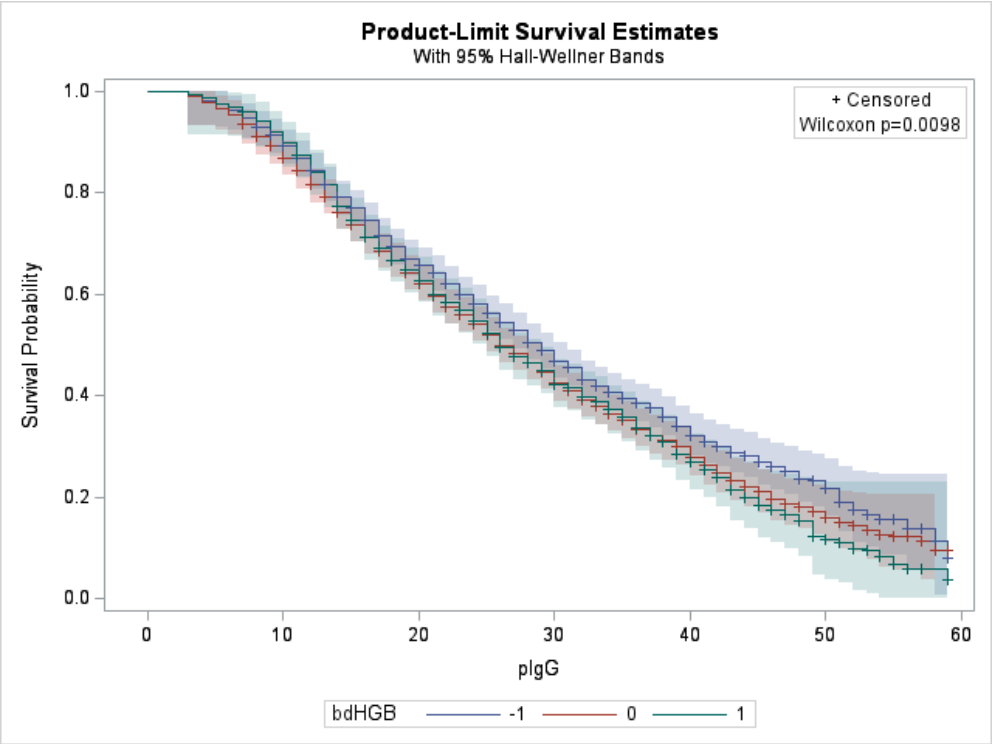

Рисунок 64 — График формирования IgG (порог >10) во времени по результатам значимых лабораторных исследований (гемоглобин)

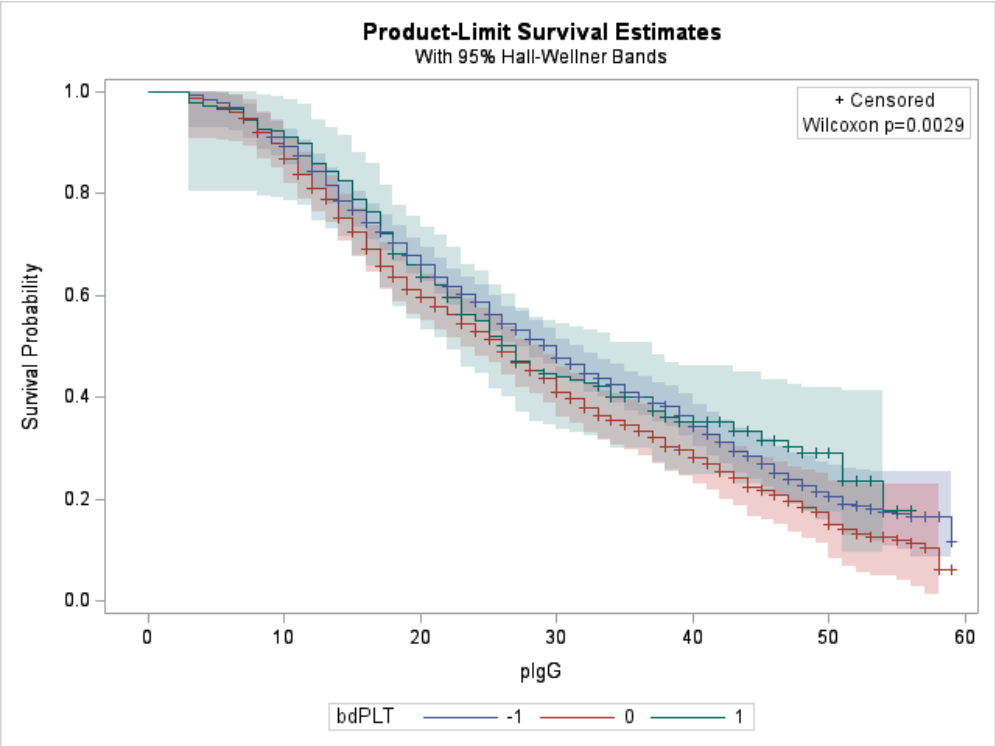

Рисунок 65 — График формирования IgG (порог >10) во времени по результатам значимых лабораторных исследований (тромбоциты)

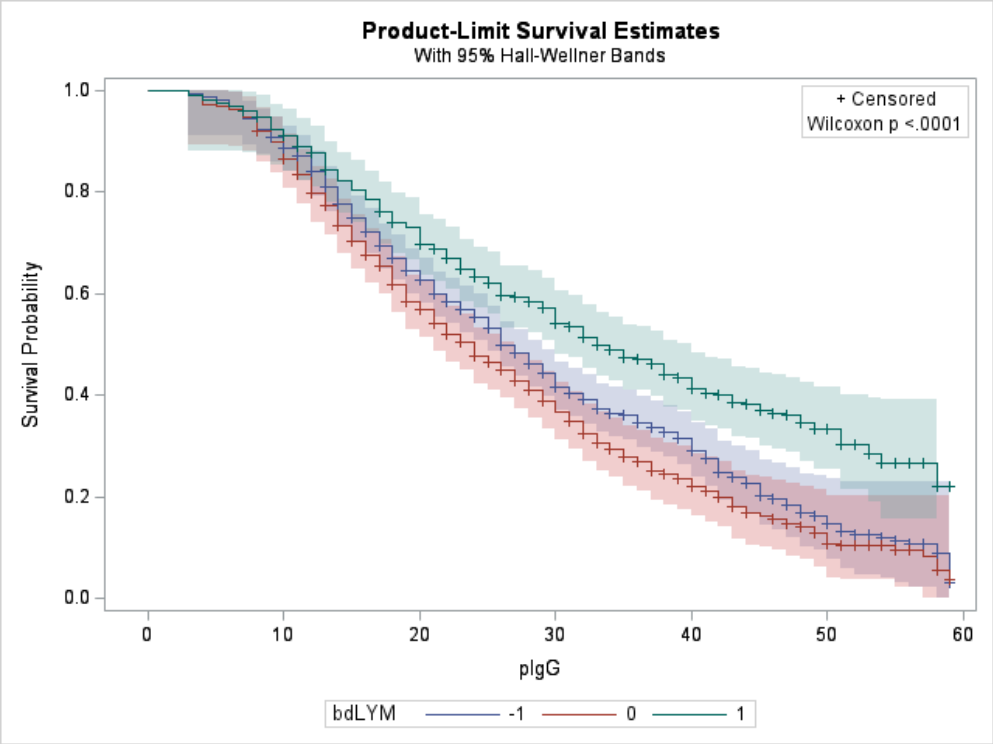

Рисунок 66 — График формирования IgG (порог >10) во времени по результатам значимых лабораторных исследований (лимфоциты)

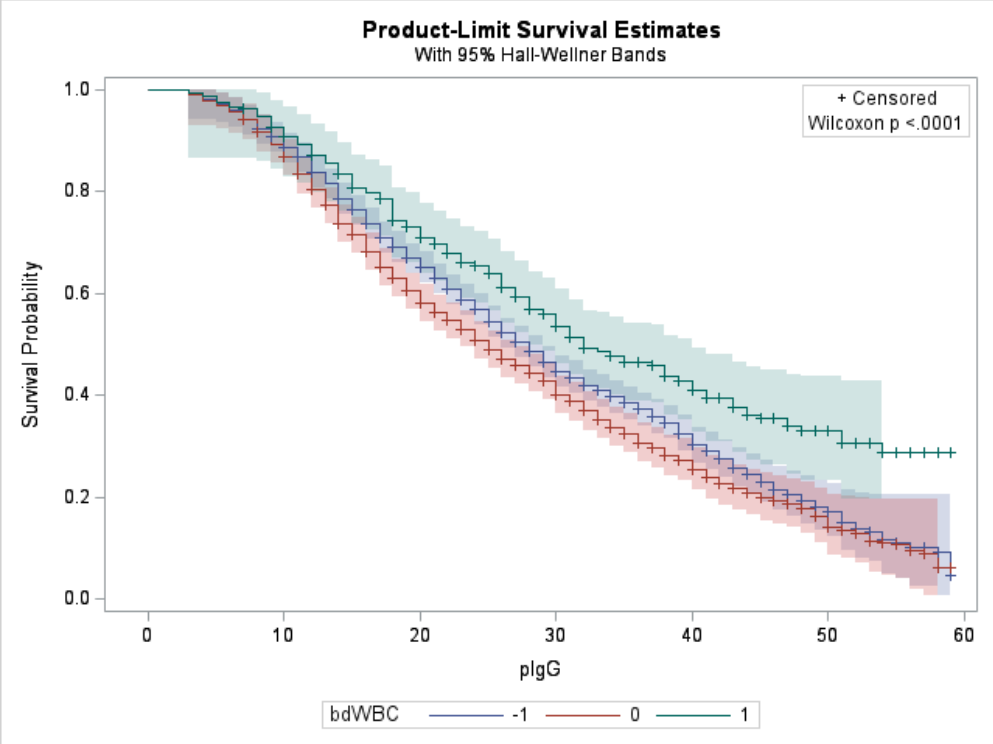

Рисунок 67 — График формирования IgG (порог >10) во времени по результатам значимых лабораторных исследований (лейкоциты)

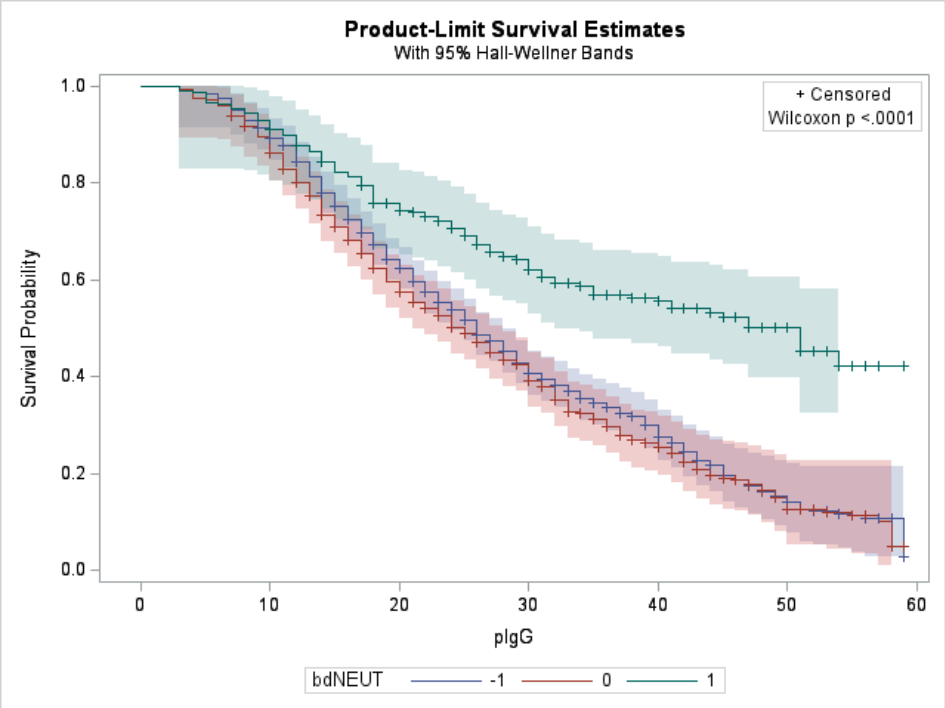

Рисунок 68 — График формирования IgG (порог >10) во времени по результатам значимых лабораторных исследований (нейтрофилы)

В результате анализа полученных графиков были получены следующие выводы:

- **Высокий с-реактивный белок – быстрее и вероятнее появляется IgG>10;**
- **Низкий гемоглобин – медленнее и менее вероятно IgG>10;**
- **Тромбоциты в норме – быстрее и вероятнее появляется IgG>10;**
- **Повышенные лимфоциты – медленнее и менее вероятно появляется IgG>10;**
- **Повышенные лейкоциты – медленнее и менее вероятно появляется IgG>10;**
- **Повышенные нейтрофилы – медленнее и менее вероятно появляется IgG>10.**

3.1.6.4 Построение описательных моделей прогнозирования уровня IgG и IgM (без учета пороговых значений) с функцией отбора важных предикторов на основе регрессионных моделей и деревьев решений

Были построены описательные модели прогнозирования уровней IgG и IgM с функцией отбора важных предикторов на основе регрессионных моделей, деревьев решений, а также нейронных сетей. Наилучшее качество продемонстрировали нейронные сети и регрессионные модели (см. Таблицы 3, 4):

Таблица 3 — Прогнозирование уровня IgG

Тип модели	**ROC AUC**	**Среднеквадратичная ошибка**
Нейронная сеть	0,637	0,084
Регрессия	0,633	0,084
Решающее дерево	0,632	0,084

Таблица 4 — Прогнозирование уровня IgM

Тип модели	**ROC AUC**	**Среднеквадратичная ошибка**
Нейронная сеть	0,666	0,056
Регрессия	0,664	0,056
Решающее дерево	0,663	0,056

В ходе анализа прогнозирования уровня IgG были получены следующие выводы:

- Наиболее значимые переменные: возраст, пол;
- У детей формируется слабо;
- У мужчин формируется быстрее и больше, чем у женщин.

В ходе анализа прогнозирования уровня IgM (см. Рисунок 69) были получены следующие выводы:

- Наиболее значимые переменные: пол;
- У детей формируется слабо;
- У женщин формируется больше и быстрее, чем у мужчин.

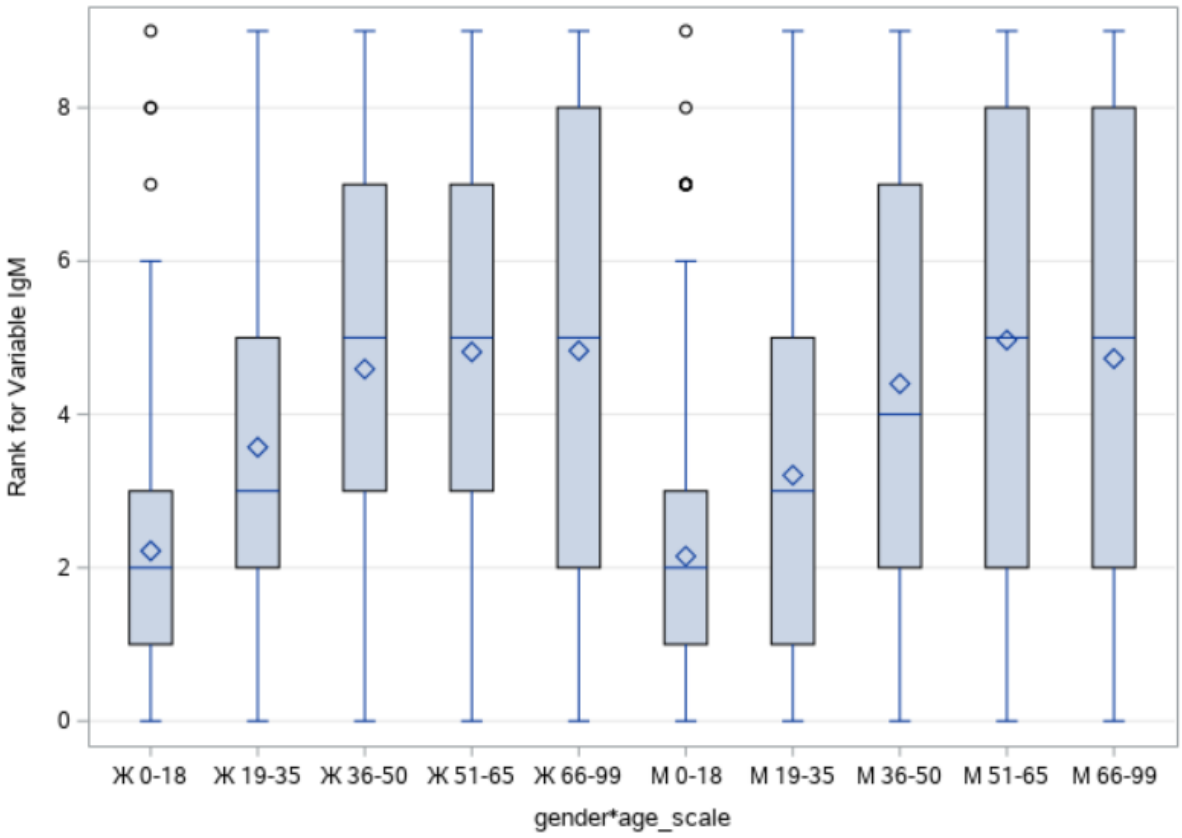

Рисунок 69 — Исследование уровня IgM пациентов

3.1.6.5 Построение модели оценки динамики уровня IgG и IgM положительных больных в выборке во времени с отсчетом от даты начала заболевания

Модель выживаемости по формированию гуморального иммунного ответа показала, что выработка антител класса IgM и IgG у пациентов с симптоматическим течением COVID-19 является прогностически благоприятным признаком (см. Рисунки 70, 71).

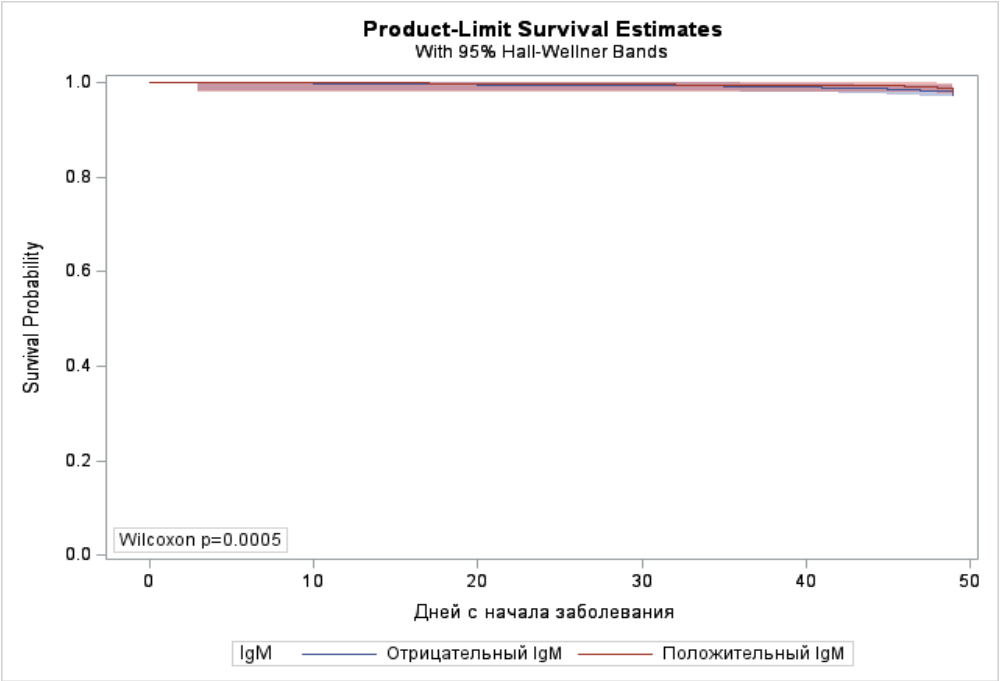

Рисунок 70 — Модель выживаемости Каплана-Мейера по выработке антител IgM

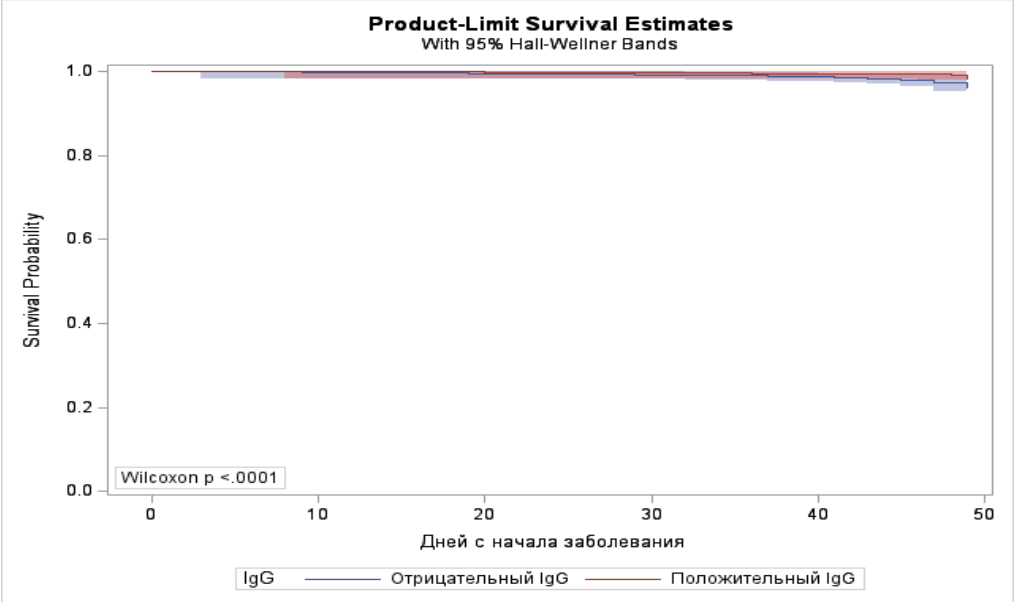

Рисунок 71 — Модель выживаемости Каплана-Мейера по выработке антител IgG

Статистическая значимость данной модели подтверждает гипотезу о том, что у пациентов с тяжелой коронавирусной инфекцией не наблюдался переход от гиперактивного врожденного иммунного ответа к адаптивному иммунному ответу. При этом пациенты с более легким течением заболевания демонстрировали выраженный иммунный ответ. К

седьмому дню болезни было отмечено повышение IgM со стойкой тенденцией к росту до 20-го дня.

3.1.6.6 Построение модели оценки динамики уровня усредненных значений IgG и IgM в выборке во времени с отсчетом от даты начала заболевания, выявление ожидаемых дней появления пороговых значений IgG и IgM

Анализ абсолютных значений IgM от времени болезни показал следующие результаты (см. Рисунки 72, 73):

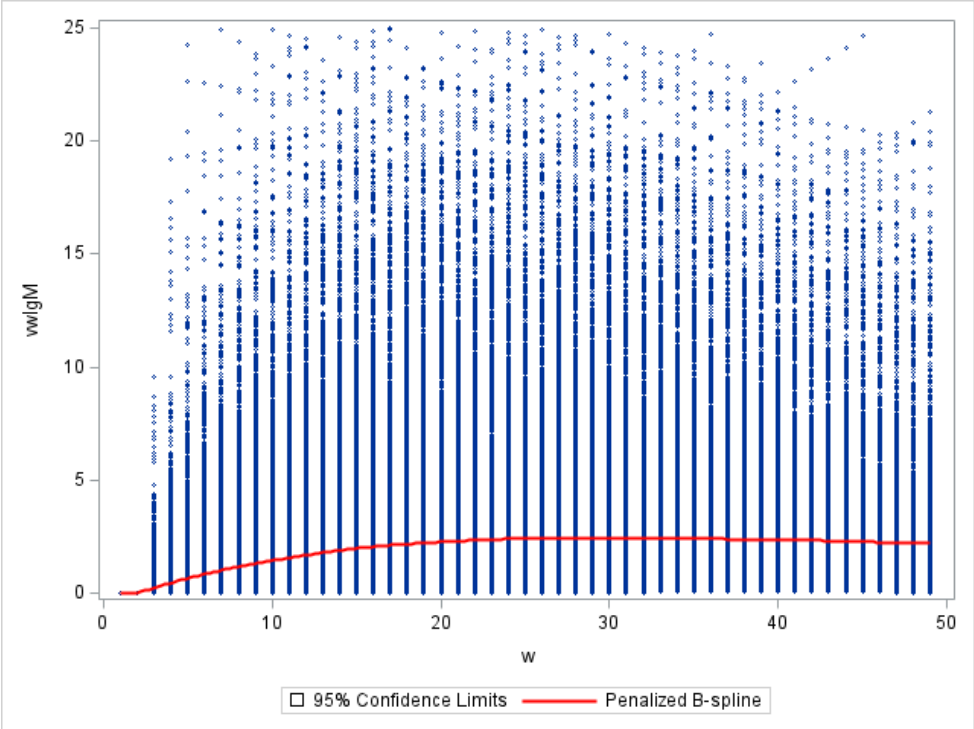

Рисунок 72 — Анализ абсолютных значений IgM от времени болезни

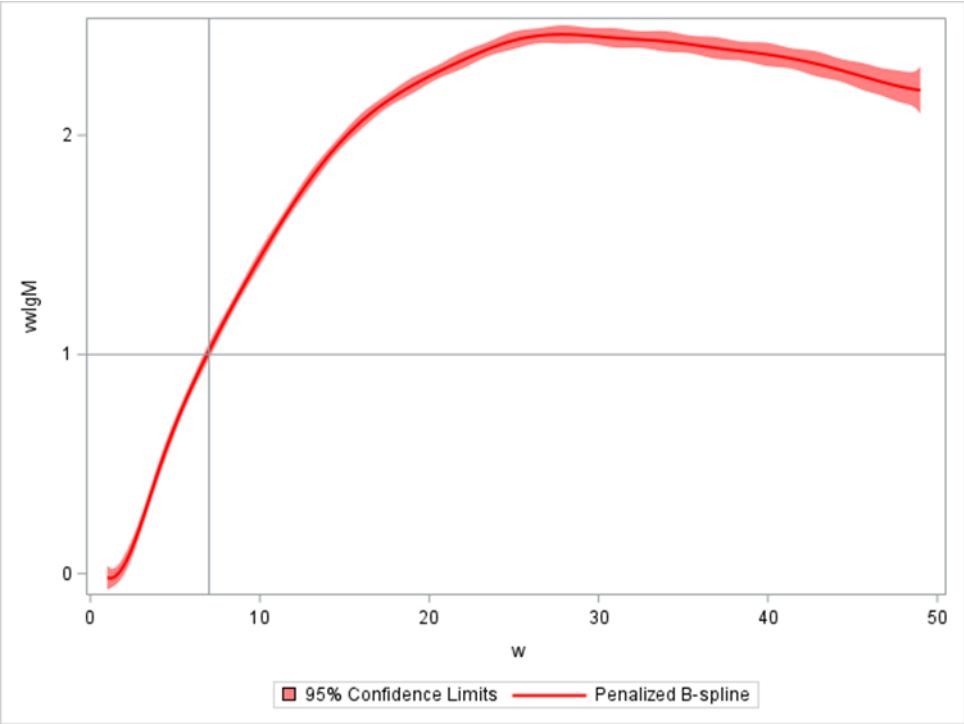

Рисунок 73 — Анализ абсолютных значений IgM от времени болезни

Анализ абсолютных значений IgG от времени болезни показал следующие результаты (см. Рисунки 74, 75):

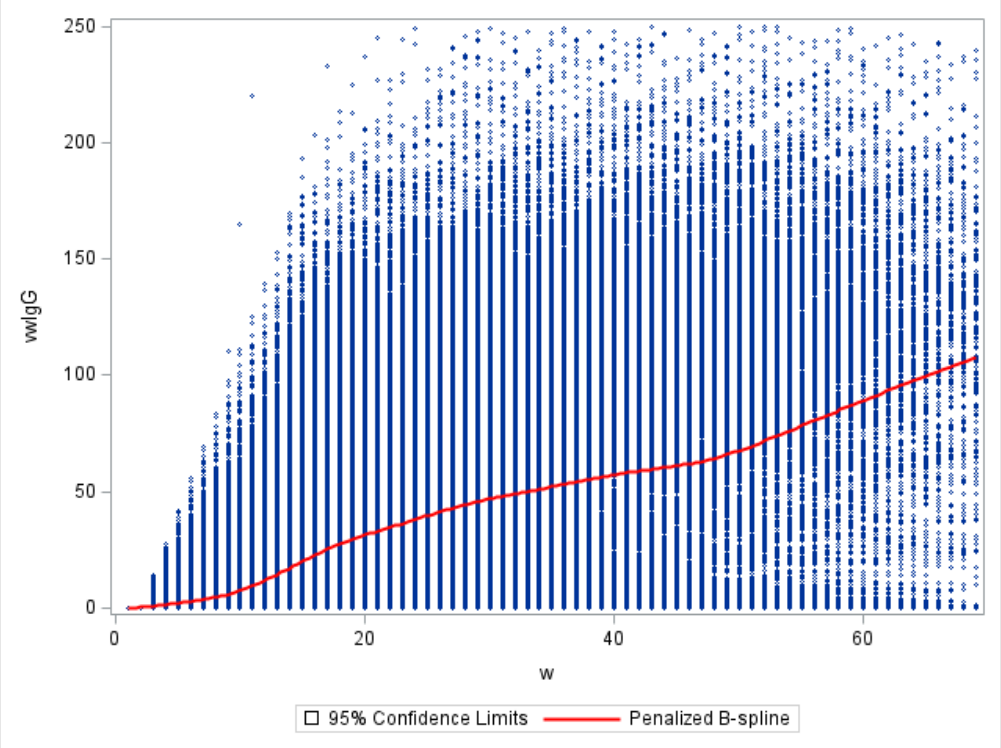

Рисунок 74 — Анализ абсолютных значений IgG от времени болезни

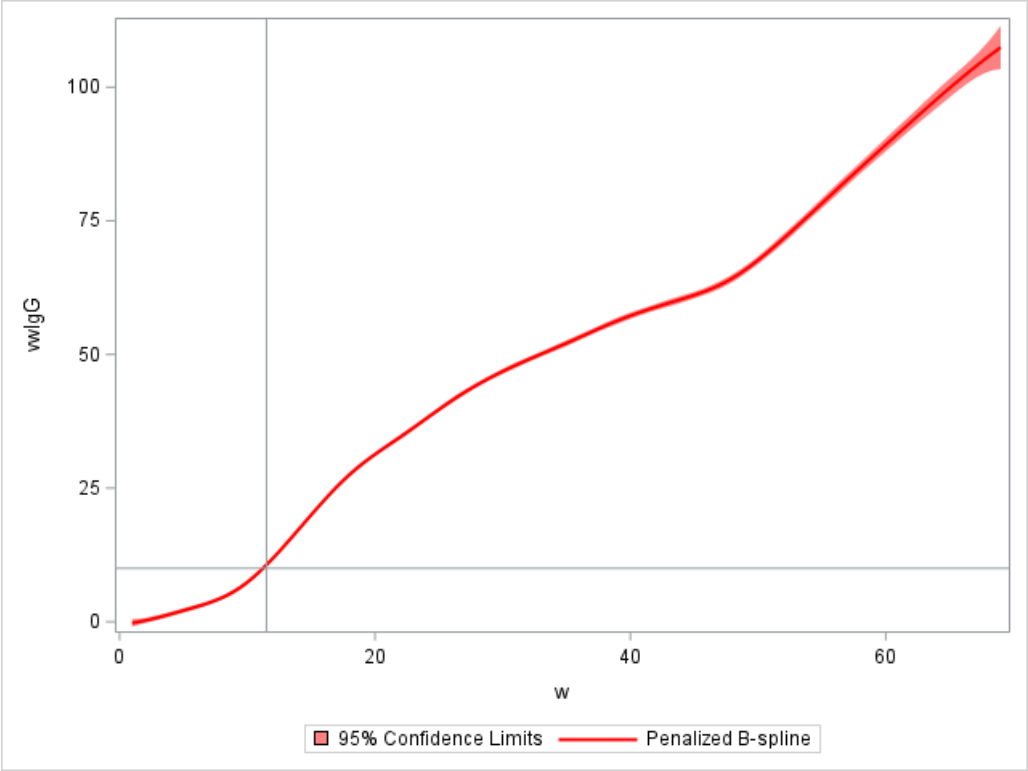

Рисунок 75 — Анализ абсолютных значений IgG от времени болезни

Таким образом, в ходе изучения динамики формирования гуморального иммунитета при COVID-19 с учетом даты начала заболевания выявлено, что антитела класса IgM начинают выявляться у большей части пациентов с 10 суток от начала заболевания. Начало выявления антител класса IgG – 18-21 сутки от начала заболевания. Уровень антител класса IgM снижается к 7-8 неделе от начала заболевания. Антитела класса IgG продолжают определяться более 9 недель.

Таким образом, создана модель противовирусного иммунного ответа, которая может быть использована для построения долгосрочного прогноза эпидемиологической ситуации. Понимание кинетики развития COVID-19 позволяет уточнить сроки тестирования для выявления людей, которые могут нуждаться в лечении, или которые должны быть изолированы, чтобы предотвратить распространение инфекции. Правильная идентификация людей, ранее переболевших COVID-19, важна для оценки распространения инфекции и корректировки мер системы здравоохранения.

3.1.7 Выводы

На основе данных о клинических особенностях, факторах коморбидности, клинико-лабораторного анализа и других факторов, потенциально связанных с тяжестью течения заболевания и вероятностью смерти пациентов с COVID-19, был разработан комплекс моделей, построенных с использованием передовых методов машинного обучения и прикладного статистического анализа, для прогнозирования тяжести течения и исхода заболевания у пациентов, получающих лечение в амбулаторных и стационарных условиях.

В значительной мере клинические проявления COVID-19, варьирующиеся от бессимптомного течения до полиорганной недостаточности, связаны с особенностями иммунного ответа у пациента. Создана модель противовирусного иммунного ответа, которая может быть использована для построения долгосрочного прогноза эпидемиологической ситуации.

Нами была выполнена сложная подстановка пропуска данных и сгенерирована модель пропорциональных рисков. Выделены следующие значимые переменные: минимальное значение СРБ, возраст, среднее значение IgG, стадия по КТ, максимальное значение гемоглобина, минимальное значение лейкоцитов, минимальное значение тромбоцитов. Рост указанных показателей ассоциирован с изменением отношения риска летального исхода. Повышение риска ассоциировано с увеличением минимального значения СРБ, возраста, стадии по КТ, минимального значения лейкоцитов ($p<0,0001$).

Известно, что тяжесть поражения легких на КТ коррелирует с тяжестью заболевания COVID-19. Нами была представлена модель выживаемости по стадиям КТ. Было выявлено, что частота летальных исходов направленно увеличивается от КТ-0 до КТ-4. При этом выделены похожие категории по трендам выживания: КТ – 0 и 1; КТ-2 и 3. Таким образом, вероятность смерти растет сильнее при переходе от КТ-1 к КТ-2 и от КТ-3 к КТ-4. При переходе от КТ-2 к КТ-3 вероятность смерти статистически значимо не возрастала.

При сложной подстановке методом кластеризации выявлены 2 группы риска. 1 группа с пиком летальности с 10 по 20 день от начала заболевания характеризуется значимо более высокими значениями С-РБ (при сравнении минимальных и средних значений показателя за время наблюдения) и Д-димера (при сравнении минимальных значений показателя за время наблюдения) по сравнению с остальной выборкой в течение всего времени заболевания. 2 группа с пиком летальности с 20 по 30 день от начала заболевания отличается значимо более высокими значениями ферритина и С-РБ (при сравнении максимальных значений

показателей за время наблюдения), по сравнению с остальной выборкой в течение всего времени заболевания.

Выявленные закономерности позволяют прогнозировать пиковые даты прогрессивного нарастания клинической симптоматики, что важно для мониторинга пациентов в подостром состоянии.

3.2 Анализ данных о периоде развития пандемии Covid-19 в г.Москве с весны 2020 года до сентября месяца включительно

3.2.1 Описание наборов данных

Набор данных представляет собой выгрузку из системы ЕМИАС с 03.2020 до 10.2020.

Выгрузка по амбулаторным пациентам имеет следующую файловую структуру:

- Данные из КТ-центров («amb_kt_result_400k.csv»)
  - Объем данных: 45 Мб
  - Количество наблюдений: 173’535
  - Содержит следующие столбцы:
    - **uuid** – уникальный идентификатор
    - **event_start_time –** время проведения КТ
    - **63emperature_tela_time –** время измерения температуры
    - **63emperature_tela_value –** значение температуры
    - **osnovnoy_diagnoz, soputstvujuschij_diagnoz –** диагнозы пациента в виде кода МКБ-10
    - **has_kashel, kashel_type –** наличие и тип кашля (сухой, с мокротой)
    - Бинарные признаки одышки, заложенности, слабости
    - **chdd –** частота дыхательных движений
    - **osmotr_tyazhest –** тяжесть пациента при осмотре (без симптомов, легкая, средняя, тяжелая)
    - **KT_nalichie_pnevmonii –** бинарный признак наличия пневмонии
    - **KT_stepen_tjazhesti –** степень тяжести КТ (нулевая, легкая, средне-тяжелая, тяжелая, критическая)
    - **resultat_KT –** результат КТ (КТ-0, КТ-1, КТ-2, КТ-3, КТ-4)
- Данные по проведенным амбулаторным анализам («amb_an_result_400k.csv»)

- Объем данных: 5 Гб
- Количество наблюдений: 14’172’297
- Содержит следующие столбцы:
    - **uuid** – уникальный идентификатор
    - **event_start_time –** время проведения анализа
    - **nazvanie_issledovanija –** название исследования
    - **test_time –** время сбора анализа
    - **nazvanie_testa –** название теста
    - **znachenie_rezultata_ed_izm –** единицы измерения результата
    - **znachenie_rezultata –** результат теста
    - **referensnye_znachenija –** референсные значения
    - **otklonenie_ot_normy, kritichnost_otklonenija –** признаки отклонения результата.

- Данные по проведенным тестам ПЦР и ИФА («amb_pcr_ifa_result_400k.csv»)
    - Объем данных: 1 Гб
    - Количество наблюдений: 1’834’582, в частности:
        - 1’440’403 наблюдения по ПЦР
        - 394’179 наблюдения по ИФА
    - Содержит следующие столбцы:
        - **uuid** – уникальный идентификатор
        - **birth_dt –** дата рождения
        - **gender –** пол
        - **pcr_ifa –** тип теста: ПЦР или ИФА
        - **mu_name –** наименование пункта сбора тестов
        - **department_name –** наименование филиалов
        - **dis_date –** дата поступления
        - **get_date_at, get_time_at, send_date_at, send_time_at –** даты получения и выдачи результатов теста
        - **mkb10_name, mkb10_code –** название и код выявляемого диагноза
        - **samples_type –** тип сбора анализа
        - **samples_result –** результат анализа
        - **laboratory_name –** наименование лаборатории
- Общие данные по пациентам («amb_400K_все_uuid.xlsx»)
    - Объем данных: 47 Мб

- Количество наблюдений: 379’995
- Содержит следующие столбцы:
    - **uuid** – уникальный идентификатор
    - **Причина закрытия ЭС**
    - **Группа риска**
    - **Дата рождения**
    - **Пол**
    - **Дата взятия первого анализа**
    - **Дата получения результата первого анализа**
    - **Результат повторного анализа**
    - **Дата получения повторного анализа**
    - **Дата получения предпоследнего результата ПЦР**
    - **Результат предпоследнего анализа ПЦР**
    - **Результат последнего анализа ПЦР**
    - **Тяжесть заболевания**
    - **Тяжесть заболевания.1**
    - **Тяжесть заболевания.2**
    - **Тяжесть заболевания.3**
    - **Тяжесть заболевания.4**
    - **Откуда пришел в стационар**
    - **Дата госпитализации в стационар**
    - **ОРИТ**
    - **ИВЛ**
    - **ЭКМО**
    - **Дата размещения в обсерваторе**
    - **Характеристика статуса в поликлинике**

Выгрузка по стационарным пациентам имеет следующую файловую структуру:

- Данные по пациентам (*_patient.xlsx)
    - Объем данных: 15 Мб
    - Количество наблюдений: 251’382
    - Содержит следующие столбцы:
        - **id –** стационарный уникальный идентификатор
        - **возраст**

        - **дата поступления**
        - **дата выписки**
        - **даты заболевания**
        - **исход –** выписан или умер
        - **МКБ-10 –** код диагноза
        - **Диагноз –** полное наименование диагноза
        - **emias_id –** выписан
- Данные по пациентам (*_laboratory.xlsx)
    - Объем данных: 80 Мб
    - Количество наблюдений: 1’048’575
    - Содержит следующие столбцы:
        - **id –** стационарный уникальный идентификатор
        - **nt_name –** название исследования
        - **date_time –** время сбора анализа
        - **test_name –** название теста
        - **rate –** референсные значения
        - **value –** результат теста
        - **name_unit –** единицы измерения результата
- Данные по медикаментам (*_medication.xlsx)
    - Объем данных: 6 Мб
    - Количество наблюдений: 138’012
    - Содержит следующие столбцы:
        - **id –** стационарный уникальный идентификатор
        - **Дата –** дата назначения медикамента
        - **мнн –** наименование медикамента
        - **дозировка**
        - **способ введения**
        - **кратность**
        - **длительность**
- Данные сопоставления идентификаторов (simi_uuid.xlsx)
    - Объем данных: 14 Мб
    - Количество наблюдений: 379’995
    - Содержит следующие столбцы:
        - **№ ЭС**

- **uuid –** глобальный уникальный идентификатор
- **kis_uuid –** стационарный уникальный идентификатор

3.2.2 Подготовка данных

На вход подаются данные медицинских анализов пациентов (более 1.3 млн человек), больных Covid-19.

Необходимо подготовить эти данные к построению над ними математических моделей:

- *Очистить данные и привести значения показателей к общим шкалам и словарям;*
- *Найти, удалить и исправить артефакты, выбросы и противоречивые данные.*

На основе исходных наборов данных формируются следующие наборы данных:

- Набор данных прогнозирования степени тяжести КТ:
  1. Данные по стационарным пациентам приводятся к формату амбулаторных пациентов:
     a. Для каждого id сопоставляется uuid
     b. Амбулаторные пациенты дополняются признаками возраста и пола
     c. Объединяются данные стационара и амбулатории:
        - Данные анализов: «amb_an_result» и *_laboratory.xlsx (выделение анализов)
        - Данные тестов: «amb_an_result» и *_laboratory.xlsx (выделение тестов)
  2. Для каждого наблюдения из КТ центра производится поиск ближайших (окно – неделя) тестов пациентов:
     a. Тесты фильтруются по необходимым группам
     b. На основе uuid производится срез тестов пациента
     c. Тесты сортируются по близости даты к дате проведения КТ
     d. Выбирается ближайший тест
  3. В качестве тестов пациентов взяты следующие группы анализов:
     a. Тесты общего клинического анализа крови
     b. Наиболее заполненные тесты биохимического анализа (определение аланинаминотрансферазы (АЛТ), определение альбумина, определение аспартатаминотрансферазы (АСТ), определение билирубина общего,

определение билирубина прямого (конъюгированного) моноглюкоронида и диглюкоронида, определение калия общего, определение креатинина, определение лактатдегидрогеназы, определение мочевины, определение натрия общего, определение общего белка, определение хлора, определение щелочной фосфатазы, относительное количество нормобластов)

4. По всем выделенным тестам производится фильтрация исходного набора данных. Таким образом, т.к. производится поиск теста (а не анализа), возможно получение результатов теста от нескольких источников исследований.
   Например, Гематокрит встречается в нескольких исследованиях.
5. На основе словарей слияния производится объединение нескольких тестов (от разных исследований), в один признак теста.
6. Выделяются наиболее заполненные признаки среди близких тестов
   Например, для тромбоцитов существует множество признаков после объединения исследований:
   - Общий объем тромбоцитов в крови (тромбокрит, PCT)
   - Количество тромбоцитов
   - Средний объем тромбоцитов в крови
   - Ширина распределения тромбоцитов по объему

   Так как данные признаки коррелируют между собой, в ходе формирования признакового пространства выбирается наиболее заполненный признак.
7. К сформированным признакам добавляются тесты ПЦР, ИФА аналогичным образом (выделение ближайших тестов с окном – неделя)
8. Формируются признаки хронических болезней:
   a. Сбор всех диагнозов на основе файла тестов
   b. Сопоставление каждому наблюдению множества диагнозов (объединение всех диагнозов)
   c. Выявление хронических диагнозов по следующим множествам (болезнь – код МКБ-10):
      - ишемическая болезнь сердца: I11 I20 I24 I25 I51
      - артериальная гипертензия: I10 O10-13 G97 I27 K76 P29 I15
      - сахарный диабет: G63 E10-14 H36 M14 G59 E23 N08 O24
      - хронические болезни легких: F53

- ожирение: E66

d. Формирование бинарных признаков

9. Формируются признаки «отсутствие признака N» для каждого «признак N»: 1, если в исходной выборке значение признака пусто, 0 иначе.
10. Заполнение пропущенных значений медианными значениями по выборке.

- Набор данных для оценки эффективности схем лечения:
    1. Изначально используются данные по стационарным пациентам.
    2. В наборе данных по медикаментам наблюдения агрегируются на основе уникального идентификатора. Поля **«мнн», «дозировка», «способ введения», «кратность», «длительность»** объединяются путем конкатенации.
    3. Каждому наблюдению стационарного пациента сопоставляется сформированное наблюдение в пункте 2.
    4. На основе набора данных по проведенным КТ производится обогащение набора данных:
        - Каждому наблюдению сопоставляется список клинических, физикальных признаков в категориальном виде: если признак категориальный или не имеет референсных границ, то входит в набор без изменений. Если признак непрерывный, то на основе референсных значений принимается одна из следующих категорий признака:
            - -1 – значение меньше нижней границы
            - 0 – значение находится в референсном интервале
            - 1 – значение больше верхней границы
        - Каждому наблюдению сопоставляется список хронических заболеваний в бинарном виде (присутствует, отсутствует).
    5. Производится агрегация всех наблюдений на основе уникального идентификатора. Поля **«МКБ-10», «Диагноз»** объединяются путем конкатенации.
    6. Каждому наблюдению сопоставляется значение признака летального исхода (на основе агрегированного поля **«исход»**).
    7. Производится фильтрация пациентов, не имеющих в списке диагнозов COVID-19 (МКБ U07.2).
    8. На основе буклета «Клинический протокол лечения больных новой коронавирусной инфекцией covid-19», а также рекомендованных схем лечения по

Минздраву, производится генерация признаков, отвечающих тем или иным схемам лечения:

- Каждая схема лечения – отдельный признак.
- Каждая схема представляет собой список множеств. Каждая позиция списка отображает вариативность того или иного медикамента.
- Все позиции списка не имеют пересечений между собой и отображают необходимые медикаменты для применимости схемы.
- Схема является примененной для пациента, если его признак **«мнн»** включает хотя бы один медикамент из каждой позиции списка схемы лечения.

9. Все категориальные признаки кодируются в виде числовых констант.

3.2.2.1 Очистка и приведение значений показателей к общим шкалам и словарям, включая учет референсных значений

- Данные из КТ-центров
  - В поле **70emperature_tela_value** имеются значения температуры:3.0, 3.6, 3.8.
  - В поле **chdd** имеются следующие значения ЧДД: 0, 1 и более 150.

  Решение: удаление данных наблюдений.
- Данные по проведенным амбулаторным анализам
  - Неунифицируемость единиц измерения (**znachenie_rezultata_ed_izm**).

    Встречаются как стандартные единицы измерения («фл», «10^12/л», «Ед/мл» и т.д), так и их различные вариации («fL», «усл.ед», «млн», «мг%»).

    Также, встречаются не интерпретируемые размерности:

    «/L», «/Ед», «1/поле зрения высокого увеличения», «%10^9л».

    Решение: предложен следующий алгоритм обработки единиц измерения (далее е.и.) признака **N:**

    1. Выявление наиболее частых е.и. данного признака
    2. Выделение префиксов и постфиксов данной е.и.
    3. Для остальных е.и. вычисляется расстояние до исходной е.и.
       a. Обработка префиксов и постфиксов «м, мк, н, мл, к, М»
       b. Обработка степенных значений: 10^9 и т.д.
  - Неунифицируемость формата результатов (**znachenie_rezultata**)

Помимо вещественных чисел, возможны и категориальные оценки выявления (обнаружено, не обнаружено), но данные категориальные оценки не имеют унифицируемого формата.

Например, смысловое значение «не обнаружено» может иметь следующие вариации: «Не обнаружено», «не обнаружено», «не обнаружены», «нет», «Нет», «0 (Отрицательно)», «Отрицательно», «ОТРИЦАТЕЛЬНО», «не обнаружен», «-», «Отрицательный», «0 (Не обнаружено)», «Не обнаруж,», «не обнаружена», «Отсутствует», «не обн.».

Данные вариации не являются редкими (каждый тип встречается более 5000 раз).

Решение: предложен следующий алгоритм обработки результатов:

1. Для категориальных переменных составлены словари значений, все возможные значения приведены к нижнему регистру и приводятся к единым категориям.
2. Для непрерывных признаков производится удаление незначащих символов, приведение к вещественному виду. В случае множественных значений, производится разбиение по разделителям и выбирается первое значение.
3. Если в ходе обработки возвращена ошибка – данное значение заменяется средним по обучающей выборке.

- Неунифицируемость формата референсных значений (**referensnye_znachenija**).

В качестве референсных значений могут быть указаны следующие категории представлений:

- Повторение результатов теста:
  «отрицательно», «не обнаружено», «не обнаружены\n\n» и т.д.
- Интервалы и промежутки:
  «<10», «10-20», «3,5 – 6,1», «2.04 – 5.80», «меньше 2000»
- Полная сводка возможных интервалов:
  «Мужчины: 0,0-15,0. Женщины: 0,0-20,0.», «отр.<9 пол.>11»,
  «близко к оптимальному уровню 2,6 – 3,3\nоптимальный уровень < 2,6\nпограничный уровень 3,3 – 4,1\nвысокий уровень 4,1 – 4,9\nочень высокий уровень > 4,9»
- Смесь интервалов с размерностями:
  «<34 мкмоль/л»
- Перечень категорий:

«Кислая, слабокислая, нейтральная», «светло–желтый, желтый, соломенно-желтый»

- Некорректные значения:
  «отрицательно, исслед. Не проводилось», «оформленный»
- И т.д.

Решение: предложен следующий алгоритм обработки результатов:

1. Поддерживаются 2 основных типа референсных значений:
   a. интервал вида **x – y**
   b. полуинтервалы <x, >y
2. Для каждого из вышеизложенных типов исходных референсных значений производится приведение к типам a,b.
3. Если в ходе обработки возвращена ошибка – данное значение заменяется средним по обучающей выборке.

- Данные по проведенным тестам ПЦР и ИФА
  - Неунифицируемость формата результатов (**samples_result**):
    - ИФА:
      Выявлены основные 2 формата представления результатов.
      1. Численное представление igg/igm: «nCoV IgM: 0.12\nnCoV IgG: 0»
      2. Категориальное представление igg/igm: «nCoV IgG: Не обнаружено\nnCoV IgM: Не обнаружено» в различных регистровых форматах.

Решение: создание 4х признаков для ИФА: igg_n, igm_n (отражают конкретное значение igg, igm), igg_def, igm_def (отражают бинарный признак обнаружения). На основе igg_n, igm_n также заполняются igg_def, igm_def по следующим правилам:

1. igg_def «Обнаружено», если igg_n > 10.0
2. igm_def «Обнаружено», если igm_n > 2.0

3.2.2.2 Поиск, удаление или исправление артефактов, выбросов и противоречивых данных

- Данные из КТ-центров
  - Несогласованность полей (**72emperature_tela_time** и **event_start_time)**

Между временем проведения КТ и взятием температуры могло пройти много дней. 171466 наблюдений не имеют данную проблему, однако, среди оставшихся наблюдений, наблюдается следующая картина:

- 1 день – 1052 наблюдения
- 2 дня – 141 наблюдения
- 3 дня – 94 наблюдения
- 4 дня – 91 наблюдение
- 5 дней – 64 наблюдения
- 6 дней – 67 наблюдений
- 7 дней – 57 наблюдений
- Более 7 дней – 504 наблюдение (максимальная разница во времени – 104 дня)

Решение: удаление наблюдений с разницей во времени более 7 дней.

- Несогласованность полей (**KT_stepen_tjazhesti** и **resultat_KT)**

Согласно описанию выше, логично сопоставление между категориями: КТ-0 – нулевая, КТ-1 – легкая, КТ-2 – средне-тяжелая, КТ-3 – тяжелая, КТ-4 – критическая.

Однако, в наборе данных такому сопоставлению отвечают лишь 128’408 наблюдений. Для оставшихся наблюдений имеется следующая картина (**KT_stepen_tjazhesti** и **resultat_KT**):

- 1934 наблюдений: легкая – КТ-2
- 1210 наблюдений: средне-тяжелая – КТ-1
- 904 наблюдения: легкая – КТ-0
- 746 наблюдений: средне-тяжелая – КТ-3
- 423 наблюдения: нулевая – КТ-1
- 262 наблюдения: легкая – КТ-3
- 239 наблюдений: тяжелая – КТ-2
- 232 наблюдения: средне-тяжелая – КТ-0
- Остальные случаи: суммарно не более 350 наблюдений

Также, в 38700+ наблюдений **KT_stepen_tjazhesti** не указана.

Решение: будем считать, что пациент имеет степень тяжести КТ **N**, если его поле **resultat_KT** имеет значение КТ-**N**, а **KT_stepen_tjazhesti** имеет степень тяжести **не ниже** КТ-**N.**

- Данные по проведенным амбулаторным анализам
  - Некорректность заполненности полей (**znachenie_rezultata**):

    Возможны следующие вариации значений:

    - «Отправлено в ДЦЛИ»
    - «Исследование временно не выполняется»
    - «Мутная», «Непрозрачная»
    - «Мало», «немного»
    - «норма»
    - И т.д.

    Данные значения встречаются в анализах, результат которых должен являться количественным показателем.

    Также, в результатах могут встречаться интервалы значений и текстовое описание результата.
- Данные по проведенным тестам ПЦР и ИФА
  - Наличие нескольких различных дат (**dis_date,get_date_at,send_date_at)**

    Для исходного набора данных не была дана расшифровка значений дат и корректной привязки тестов на их основе.

    Экспериментальным путем было выяснено, что наиболее точной датой проведения теста следует считать **get_date_at.**

3.2.3 Построение моделей оценки степени поражения по КТ в зависимости от результатов осмотра и анализа крови (клинического, СРБ, Д-димер, на Ферритин и др.) с использованием регрессионных моделей, деревьев решений и их ансамблей, нейросетей

На вход подаются данные медицинских анализов пациентов (более 1.3 млн человек), больных Covid-19.

Необходимо построить модели оценки степени поражения по КТ в зависимости от результатов осмотра и анализа крови (клинического, СРБ, Д-димер, на Ферритин и др.) с использованием регрессионных моделей, деревьев решений и их ансамблей, нейросетей.

Целевой переменной является степень поражения по КТ.

Также по результатам исследования необходимо разработать программный Демонстрационный прототип «калькулятора» степени поражения по КТ.

Сформированный набор данных является дисбалансным относительно классов степеней тяжести КТ:

- КТ 0 – 36064 наблюдений
- КТ 1 – 91779 наблюдений
- КТ 2 – 32414 наблюдений
- КТ 3 – 10266 наблюдений
- КТ 4 – 936 наблюдений

В таком случае должна производиться балансировка классов перед обучением прогнозных моделей. Балансировка производится путем выбора в большем классе подмножества наблюдений, размер которого совпадает с размером меньшего класса. Также, отметим, что наблюдений класса КТ 4 значительно меньше иных классов. Поэтому, далее будем относить наблюдения класса КТ 4 к классу КТ 3.

В качестве моделей прогнозирования степени тяжести использовались модели машинного обучения с учителем, в частности, одноклассовые и многоклассовые классификаторы: *метод случайного леса* и *нейронные сети*.

Метод случайного леса (Random Forest)

Основная идея Случайного леса заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт невысокое качество классификации, но за счёт их большого количества результат получается более точным.

Пусть обучающая выборка состоит из N примеров, размерность пространства признаков равна M, и задан параметр m – количество признаков для обучения. Наиболее распространённый способ построения ансамбля деревьев – бэггинг, заключается в следующем:

1. Генерируется случайная подвыборка с повторениями размером N из примеров обучающей выборки. (Таким образом, некоторые образцы попадут в неё несколько раз, а какие-то не войдут вообще).
2. Построим решающее дерево, классифицирующее образцы сгенерированной подвыборки, причём при добавлении очередного узла дерева будем выбирать набор признаков, на основе которых производится разбиение.
3. Дерево строится до полного исчерпания подвыборки.

Итоговая классификация объектов проводится путём голосования: каждое дерево ансамбля относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Нейронные сети (Neural Networks)

Функционирование нейронной сети [27–32] имитирует функционирование человеческой нейронной системы мозга. Доказано, что с помощью нейронных сетей можно сколь угодно точно аппроксимировать любую непрерывную функцию и имитировать любой непрерывный автомат.

Поскольку модель нейрона реализует функцию от его входов, нейроны можно объединять в соответствии с правилами суперпозиции функций, получая более сложные модели, называемые персептронами или искусственными нейронными сетями прямого распространения.

Для решения задачи прогнозирования был выбран многослойный персептрон, представляющий собой обобщение однослойного персептрона – однослойной *нейронной сети*, все нейроны которой имеют жесткую пороговую функцию активации.

Многослойный персептрон имеет несколько отличительных признаков: каждый нейрон имеет *нелинейную функцию активации*, сеть содержит один или несколько слоев *скрытых нейронов*. Также для многослойного персептрона выделяют два типа сигналов:

1. *Функциональный сигнал* – это входной сигнал сети, передаваемый по всей сети в прямом направлении. В каждом нейроне, через который передается функциональный сигнал, вычисляется функция активации от взвешенной суммы его входов с поправкой в виде порогового элемента – единичного сигнала с весовым коэффициентом.
2. *Сигнал ошибки* – это сигнал выхода сети и распространяющийся в обратном направлении от слоя к слою. Сигнал ошибки вычисляется каждым нейроном на основе заданной функции ошибки.

Обучение многослойного персептрона состоит в подборе значений весов слоев сети, чтобы при заданном входном векторе получить на выходе значения сигналов, которые с требуемой точностью будут совпадать с ожидаемыми значениями.

Для обучения многослойного персептрона используется *метод обратного распространения ошибки* (от англ. Back *propagation*) – алгоритм обучения, основанный на

вычислении градиента *функции ошибок*. В процессе обучения веса нейронов каждого *слоя* нейросети корректируются с учетом сигналов, поступивших с предыдущего *слоя*, и *невязки* (отклонения) каждого *слоя*, которая вычисляется рекурсивно в обратном направлении от последнего *слоя* к первому.

При одноклассовой классификации в качестве функции ошибок использовалась бинарная кросс-энтропия, а в качестве функции активации использовалась логистическая функция. При многоклассовой классификации в качестве выхода нейросети ожидается вектор вероятностей степеней КТ и в качестве функции ошибок используется косинусная похожесть.

Также, для повышения точности моделей, дополнительно может использоваться калибровка моделей. В данной реализации использовалась калибровка Платта, заключающаяся в добавлении к выходам классификатора логистической регрессии с последующим поиском оптимальных параметров методом математического правдоподобия.

Отметим, что построенная нейросеть является регрессионной, поскольку на ее выходе находится вероятность, а не бинарный ответ.

Для оценки качества прогнозирования будем использовать две метрики.

- *Точность (accuracy)*

  Метрика отражает долю верно классифицированных объектов относительно общего количества всех объектов.

- *Площадь под ROC-кривой (ROC-AUC)*

  ROC-кривая – график, отображающий соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущие признак, и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак, при варьировании порога решающего правила (ошибок I рода).

  Площадь под ROC-кривой AUC (англ. Area Under Curve) принимает значение от 0 до 1 и интерпретируема как вероятность того, что классификатор присвоит больший вес случайно выбранному положительному наблюдению, чем случайно выбранному отрицательному наблюдению.

3.2.3.1 Отделение слабой степени поражения (КТ 0 или 1) от всех остальных – для принятия решения о необходимости выполнения КТ

На основе сформированного набора данных прогнозирования степени тяжести КТ наблюдениям назначаются целевые значения:

- 0 – легкая степень тяжести (КТ 0 или 1);
- 1 – иначе (КТ 2 или 3).

После балансировки, набор данных содержит по 40000 наблюдений на каждый из классов. Далее будет описан полный алгоритм обучения моделей:

- Набор данных разделяется на:
  - признаковое пространство пациентов, содержащее клинические данные, данные физикального осмотра и анамнез
  - вектор целевых значений
- Эксперименты проводились с использованием 10-кратной кросс-валидации. Наборы разделяются на обучающую и тестовую выборку в отношении 90/10
- На основе обучающей выборки производится обучение одноклассовых классификаторов
- На основе тестовой выборки производится расчет метрик качества

Экспериментальные результаты приводятся в Таблице 5.

Таблица 5 — Оценка качества моделей прогнозирования легкой степени тяжести

Метод	**ROC AUC**	**Accuracy**
RF	0.918	0.839
RF с калибровкой	0.919	0.844
NN	0.892	0.811
NN с калибровкой	0.892	0.812

3.2.3.2 Отделение тяжелой степени поражения (КТ 3 или 4) от всех остальных – для принятия решения о необходимости выполнения КТ (возможно повторного) для уточнения степени поражения

На основе сформированного набора данных прогнозирования степени тяжести КТ наблюдениям назначаются целевые значения:

- 1 – тяжелая степень тяжести (КТ 3);
- 0 – иначе (КТ 0, 1 или 2).

После балансировки, набор данных содержит по 10000 наблюдений на каждый из классов. Далее будет описан полный алгоритм обучения моделей:

- Набор данных разделяется на:
    - признаковое пространство пациентов, содержащее клинические данные, данные физикального осмотра и анамнез
    - вектор целевых значений
- Эксперименты проводились с использованием 10-кратной кросс-валидации Наборы разделяются на обучающую и тестовую выборку в отношении 90/10
- На основе обучающей выборки производится обучение одноклассовых классификаторов
- На основе тестовой выборки производится расчет метрик качества

Экспериментальные результаты приводятся в Таблице 6.

Таблица 6 — Оценка качества моделей прогнозирования тяжелой степени тяжести

Метод	**ROC AUC**	**Accuracy**
RF	0.938	0.869
RF с калибровкой	0.941	0.872
NN	0.909	0.838
NN с калибровкой	0.909	0.841

3.2.3.3 Прогнозирование непосредственно степени поражения по КТ в категориях: 0, 1, 2, 3 и выше

На основе сформированного набора данных прогнозирования степени тяжести КТ наблюдениям назначаются целевые значения, равные степени тяжести КТ. После балансировки, набор данных содержит по 10000 наблюдений на каждый из классов. Далее будет описан полный алгоритм обучения моделей:

- Набор данных разделяется на:
    - признаковое пространство пациентов, содержащее клинические данные, данные физикального осмотра и анамнез
    - вектор целевых значений
- Полученные наборы разделяются на обучающую и тестовую выборку в отношении 90/10

- На основе обучающей выборки производится обучение многоклассовых классификаторов
- На основе тестовой выборки производится расчет метрик качества

Экспериментальные результаты приводятся в Таблице 7.

Таблица 7 — Оценка качества моделей прогнозирования степени поражения по КТ

Метод	**Accuracy**
RF	0.705
RF с калибровкой	0.712
NN	0.641
NN с калибровкой	0.642

3.2.3.4 Отбор на основе полученных результатов наилучшей модели и необходимого ей набора признаков

На основе проведенных экспериментов, для прогнозирования различных степеней тяжести КТ были выбраны модели: RF с калибровкой, NN с калибровкой. По умолчанию, далее используется RF с калибровкой, а нейронная сеть является опциональной и может быть активирована дополнительно. На основе данных моделей были выявлены наиболее важные признаки для прогноза (взято 30 наиболее важных признаков от каждой модели).

Далее, на основе экспертной медицинской оценки значимости параметров, были оценены все полученные списки важных признаков и сформирован общий список из 35 признаков.

На основе кросс-валидации были выбраны лучшие модели случайного леса и нейронной сети.

3.2.3.5 Разработка программного Демонстрационного прототипа «калькулятора» степени поражения по КТ

Описание раздела содержится в главе **Error! Reference source not found.** данного отчета (**Error! Reference source not found.**).

3.2.4 Построение моделей прогнозирования риска летального исхода пациентов

На вход подаются данные медицинских анализов пациентов (более 1.3 млн человек), больных Covid-19.

Необходимо построить модели прогнозирования риска летального исхода пациентов с использованием регрессионных моделей, деревьев решений и их ансамблей, нейросетей.

Целевой переменной является факт летальности.

На основе регрессионных моделей, деревьев решений и их ансамблей, а также нейросетей были построены модели прогнозирования риска летального исхода пациентов.

С использованием методов латентно-семантического анализа были выделены следующие хронические заболевания, наиболее сильно влияющие на летальность (в порядке убывания значимости):

- Гипертензивная болезнь сердца;
- Хроническая ишемическая болезнь сердца;
- Инсулиннезависимый диабет;
- Ожирение;
- Сердечная недостаточность.

Далее, было рассмотрено два подкласса моделей. Первые строились с учетом хронических болезней пациента, вторые – без.

Для подкласса моделей с учетом хронических болезней пациентов были выделены следующие значимые признаки (в порядке убывания значимости):

- Возраст;
- Наличие вышеупомянутых хронических заболеваний;
- Степень тяжести КТ.

Наилучшие результаты для данного подкласса моделей продемонстрировала нейронная сеть (см. Таблица 8).

Таблица 8 — Прогнозирование летальности, подкласс моделей с учетом хронических заболеваний

Тип модели	**ROC AUC**	**Среднеквадратичная ошибка**
Нейронная сеть	0,868	0,075
Регрессия	0,865	0,075

Тип модели	ROC AUC	Среднеквадратичная ошибка
Решающее дерево	0,848	0,078

Для подкласса моделей без учета хронических болезней были выделены следующие значимые признаки (в порядке убывания значимости):

- Степень тяжести КТ;
- Значение Гемоглобина (HGB);
- Количество нейтрофилов;
- Д-димер;
- IgM;
- IgG;
- Возраст;
- Количество тромбоцитов (PLT);
- Абсолютное число лимфоцитов;
- Количество лейкоцитов (WBC);
- Уровень С-реактивного белка (CRB).

Наилучшие результаты для данного подкласса моделей так же продемонстрировала нейронная сеть (см. Таблица 9).

Таблица 9 — Прогнозирование летальности, подкласс моделей без учета хронических заболеваний

Тип модели	ROC AUC	Среднеквадратичная ошибка
Нейронная сеть	0,868	0,072
Решающее дерево	0,860	0,074
Регрессия	0,860	0,075

3.2.5 Построение моделей для предварительной статистической оценки эффективности схем лечения Covid-19

На вход подаются данные медицинских анализов пациентов (более 1.3 млн человек), больных Covid-19, а также принимаемых ими медикаментов.

Необходимо с помощью метода Каплана-Мейера построить модели для предварительной статистической оценки эффективности схем лечения Covid-19.

Целевой переменной является статистически значимое увеличение выживаемости в анализируемой группе.

3.2.5.1 Рассмотрение рекомендованных Минздравом базовых схем лечения Covid-19, включая 3 схемы для легкого течения, 8 схем для средней тяжести, 5 для тяжелого течения и 5 для случая цитокинового шторма

Были рассмотрены следующие рекомендованные Минздравом базовые схемы лечения Covid-19 (см. Таблица 10).

Таблица 10 — Рассмотренные схемы лечения Covid-19

Название схемы	**Тяжесть заболевания**	**Лекарства, входящие в схему**
light_1	Легкая	Гидроксихлорохин
light_2	Легкая	Мефлохин
light_3	Легкая	ИФН-а + умифеновир
middle_1	Средняя	Фавипиравир +/- барицитиниб или тофацитиниб
middle_2	Средняя	Гидроксихлорохин+азитромицин +/- барицитиниб или тофацитиниб
middle_3	Средняя	Мефлохин+азитромицин +/- барицитиниб или тофацитиниб
middle_4	Средняя	Лопинавир/ритонавир + ИНФ – b1b +/- барицитиниб или тофацитиниб
middle_5	Средняя	Фавипиравир +/- Олокизумаб
middle_6	Средняя	Гидроксихлорохин+азитромицин +/- олокизумаб
middle_7	Средняя	Мефлохин+азитромицин +/- олокизумаб
middle_8	Средняя	Лопинавир/ритонавир + ИНФ – b1b +/- олокизумаб
hard_1	Тяжелая	Фавипиравир +/- тоцилизумаб или сарилумаб
hard_2	Тяжелая	Гидроксихлорохин + азитромицин +/- тоцилизумаб или сарилумаб
hard_3	Тяжелая	Мефлохин + азитромицин +/- тоцилизумаб или сарилумаб
hard_4	Тяжелая	Лопинавир/ритонавир + ИФН-b1b +/- тоцилизумаб или сарилумаб
hard_5	Тяжелая	Лопинавир/ритонавир + гидроксихлорохин +/- тоцилизумаб или сарилумаб

Название схемы	Тяжесть заболевания	Лекарства, входящие в схему
cyto_1	Цитокиновый шторм	Метилпреднизолон + тоцилизумаб (сарилумаб)
cyto_2	Цитокиновый шторм	Дексаметазон + тоцилизумаб (сарилумаб)
cyto_3	Цитокиновый шторм	Метилпреднизолон + канакинумаб
cyto_4	Цитокиновый шторм	Дексаметазон + канакинумаб
cyto_5	Цитокиновый шторм	Метилпреднизолон или дексаметазон
cyto_6	Цитокиновый шторм	Тоцилизумаб или сарилумаб или канакинумаб
mos_light_1	Легкая	фавипиравир риамиловир
mos_light_2	Легкая	фавипиравир риамиловир гидроксихлорохин
mos_middle_1	Средняя	фавипиравир риамиловир гидроксихлорохин антикоагулянтны
mos_middle_2	Средняя	фавипиравир риамиловир тофацитиниб барицитиниб
mos_middle_3	Средняя	фавипиравир риамиловир ингибриторы цитокинов
mos_middle_4	Средняя	фавипиравир риамиловир ингибриторы цитокинов тофацитиниб барицитиниб
mos_hard_1	Тяжелая	фавипиравир риамиловир гидроксихлорохин глюкокортикостероиды
mos_hard_2	Тяжелая	фавипиравир риамиловир гидроксихлорохин глюкокортикостероиды антикоагулянты
mos_hard_3	Тяжелая	риамиловир ингибриторы цитокинов глюкокортикостероиды
mos_cyto_1	Цитокиновый шторм	антикоагулянты ингибриторы цитокинов
mos_cyto_2	Цитокиновый шторм	антикоагулянты ингибриторы цитокинов глюкокортикостероиды

3.2.5.2 Реализация процедуры на основе метода Каплана-Мейера, позволяющей оценить эффективность схемы лечения в заданной популяции с точки зрения статистически значимого увеличения выживаемости в анализируемой группе

Было проведено сравнение всех схем лечения отдельно по каждой группе тяжести. По каждой группе были найдены важные признаки, влияющие на летальность. Были построены модели выживаемости с учетом этих признаков.

Общие модели выживаемости Каплана-Мейера для каждой из рассматриваемой групп, а также перечень найденных ключевых признаков приведены ниже.

Легкая степень тяжести

Для легкой степени тяжести модель выживаемости Каплана-Мейера выглядит следующим образом (см. Рисунок 76):

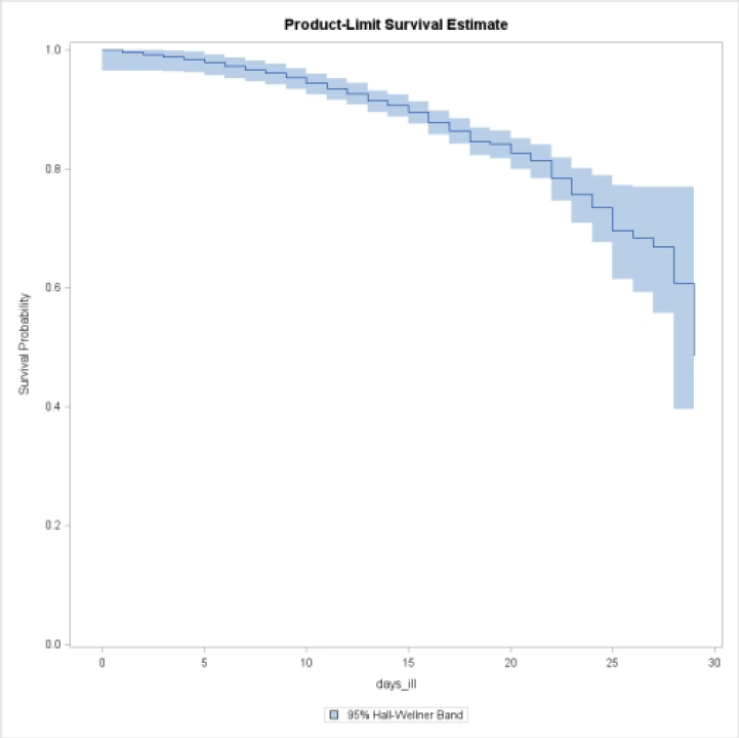

Рисунок 76 — Модель Каплана-Мейера, легкая степень тяжести

При этом, наиболее значимыми признаками являются (см. Рисунок 77):

- Количество лимфоцитов;
- Возраст;
- Количество лейкоцитов;
- Наличие диабета;

- Положительный IGM;
- Количество тромбоцитов;
- Степень поражения по КТ.

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
LYM	1	51.6919	<.0001	51.6919	<.0001
age	2	82.0175	<.0001	30.3256	<.0001
WBC	3	103.6	<.0001	21.5731	<.0001
e10	4	123.8	<.0001	20.1653	<.0001
IGM_N	5	140.9	<.0001	17.1035	<.0001
PLT	6	152.3	<.0001	11.4100	0.0007
resultat_KT	7	157.9	<.0001	5.6700	0.0173

Рисунок 77 — Легкая степень тяжести, значимые признаки

Средняя степень тяжести

Для средней степени тяжести модель выживаемости Каплана-Мейера выглядит следующим образом (см. Рисунок 78):

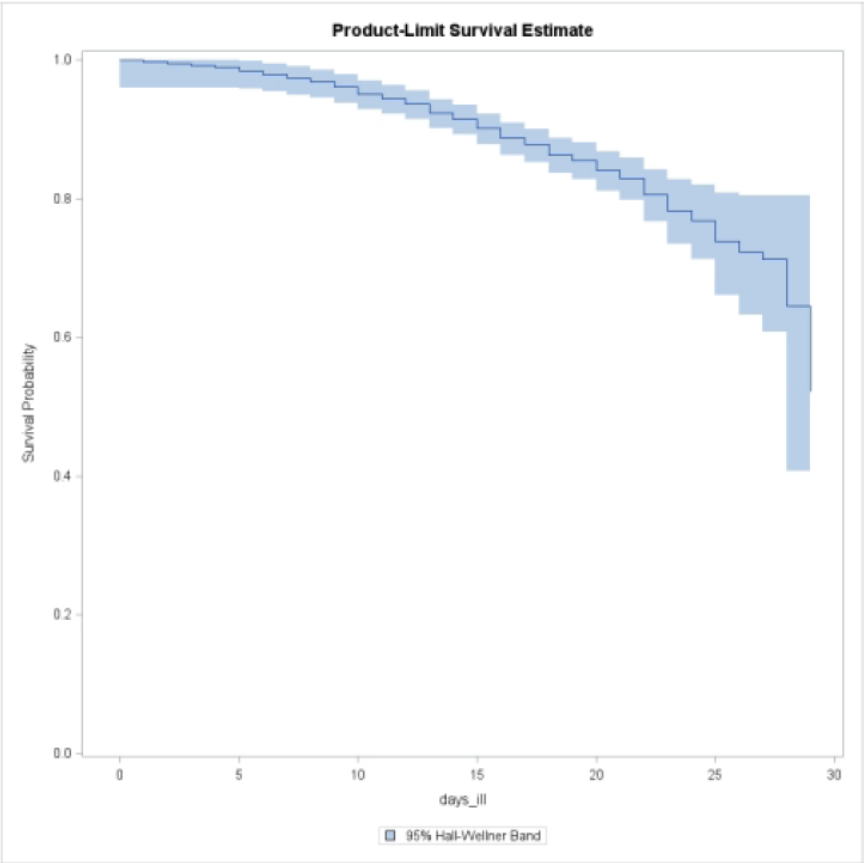

Рисунок 78 — Модель Каплана-Мейера, средняя степень тяжести

При этом, наиболее значимыми признаками являются (см. Рисунок 79):

- Количество лимфоцитов;
- Возраст;
- Количество лейкоцитов;
- Количество тромбоцитов;
- D-димер;
- Положительный IGM;
- Степень поражения по КТ.

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
LYM	1	41.9170	<.0001	41.9170	<.0001
age	2	65.6490	<.0001	23.7320	<.0001
WBC	3	88.4481	<.0001	22.7991	<.0001
PLT	4	107.4	<.0001	18.9298	<.0001
DD	5	113.1	<.0001	5.7323	0.0167
IGM_N	6	117.9	<.0001	4.8136	0.0282
resultat_KT	7	122.7	<.0001	4.7483	0.0293

Рисунок 79 — Средняя степень тяжести, значимые признаки

Тяжелая степень тяжести

Для тяжелой степени тяжести модель выживаемости Каплана-Мейера выглядит следующим образом (см. Рисунок 80):

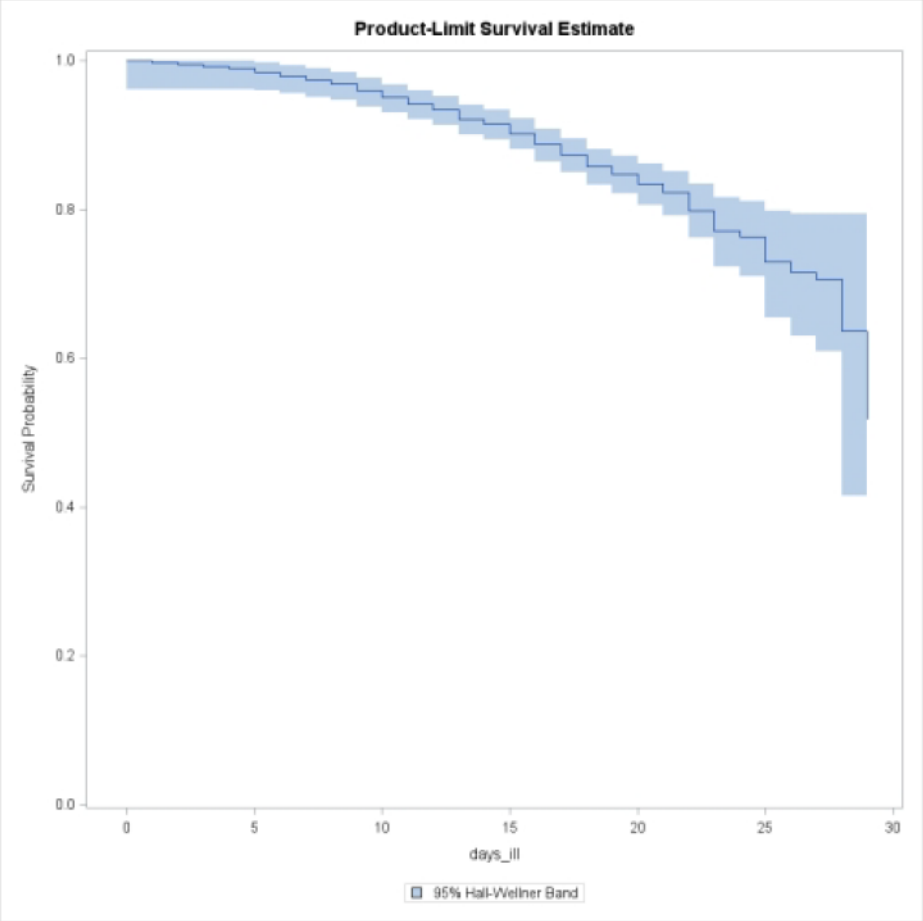

Рисунок 80 — Модель Каплана-Мейера, тяжелая степень тяжести

При этом, наиболее значимыми признаками являются (см. Рисунок 81):

- Количество лимфоцитов;
- Количество лейкоцитов;
- Возраст;
- Количество тромбоцитов;
- Сахарный диабет;
- Положительный IGM;
- Степень поражения по КТ.

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test

Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
LYM	1	68.9837	<.0001	68.9837	<.0001
WBC	2	93.0848	<.0001	24.1011	<.0001
age	3	114.7	<.0001	21.6013	<.0001
PLT	4	124.7	<.0001	10.0441	0.0015
e10	5	135.6	<.0001	10.8450	0.0010
IGM_N	6	144.8	<.0001	9.2417	0.0024
resultat_KT	7	149.6	<.0001	4.7696	0.0290

Рисунок 81 — Тяжелая степень тяжести, значимые признаки

Цитокиновый шторм

Для цитокинового шторма модель выживаемости Каплана-Мейера выглядит следующим образом (см. Рисунок 82):

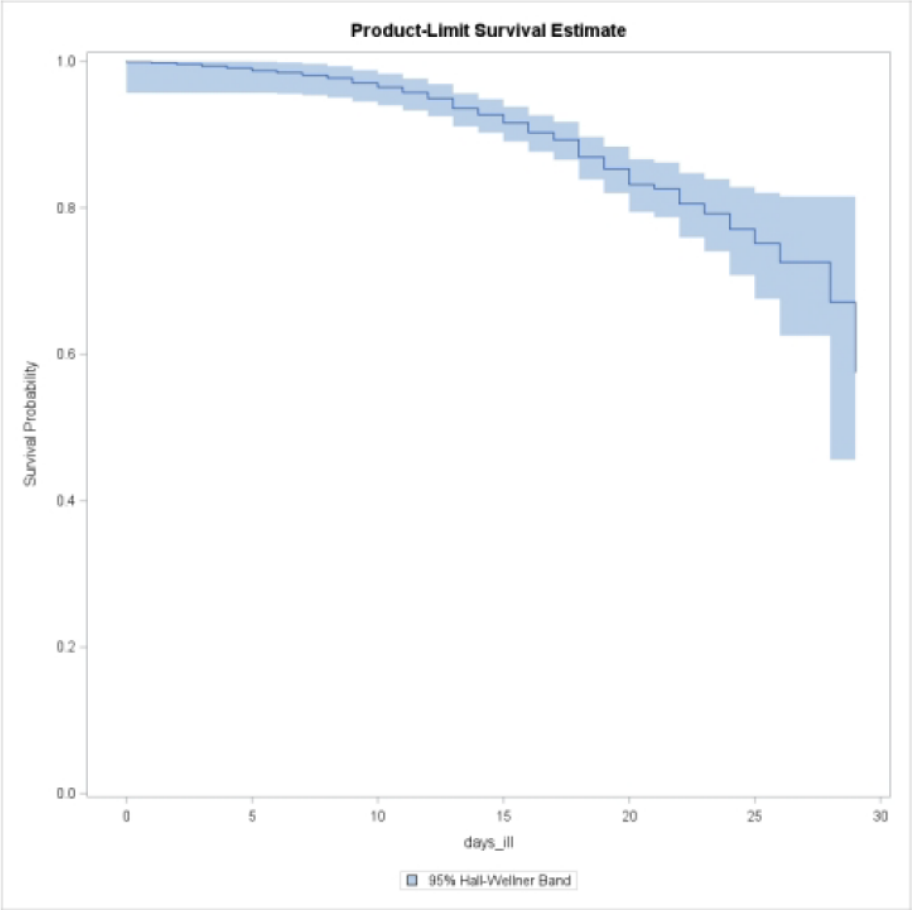

Рисунок 82 — Модель Каплана-Мейера, цитокиновый шторм

При этом, наиболее значимыми признаками являются (см. Рисунок 83):

- Количество лимфоцитов;
- Возраст;
- D-димер;
- Исследование Ферритина;
- Количество нейтрофилов;
- Количество тромбоцитов;
- С-реактивный белок.

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
LYM	1	44.1629	<.0001	44.1629	<.0001
age	2	56.3523	<.0001	12.1893	0.0005
DD	3	63.4731	<.0001	7.1209	0.0076
FRT	4	68.7771	<.0001	5.3039	0.0213
NEU	5	74.6835	<.0001	5.9064	0.0151
PLT	6	77.5787	<.0001	2.8952	0.0888
CRB	7	79.1630	<.0001	1.5843	0.2081

Рисунок 83 — Цитокиновый шторм, значимые признаки

3.2.5.3 Реализация поиска по группам и схемам с целью формулировки гипотезы, когда какая схема более эффективна

В результате анализа было получено, что:

Легкая степень тяжести

Для легкой степени тяжести наименее эффективной во всех случаях является схема лечения light_1. Наиболее эффективными являются схемы лечения light_3 и mos_light_2. Средней эффективностью обладает схема mos_light_1.

При пониженном или повышенном содержании лейкоцитов или тромбоцитов наиболее эффективна схема лечения mos_light_1.

Средняя степень тяжести

Для средней степени тяжести наименьшей эффективностью обладает схема mos_middle_2, наибольшей – mos_middle_1. Низкой эффективностью также обладает схема mos_middle_4. Высокая степень эффективности у схем middle_8, middle_4, middle_5.

При пониженном значении D-димера схемы middle_1, middle_2, middle_5, middle_6, неэффективны. При повышенном значении D-димера наиболее эффективны схемы middle_8, mos_middle_2, mos_middle_4, наименее – middle_2 и middle_6.

При пониженном количестве лейкоцитов наименее эффективна схема middle_4, наиболее эффективны – middle_5, middle_8, mos_middle_3. При повышенном количестве лейкоцитов наименее эффективны схемы middle_2 и middle_6.

При пониженном количестве нейтрофилов наименее эффективны схемы middle_1, middle_2, middle_6. Наиболее эффективны – middle_4, middle_5, middle_8 и mos_middle_3. При повышенном количестве нейтрофилов наименее эффективны схемы middle_2 и middle_6.

Тяжелая степень тяжести

Для тяжелой степени тяжести наименьшей эффективностью обладает схема mos_hard_1, наибольшей – mos_hard_2. Низкая степень эффективности и у схемы лечения hard_5.

При пониженном количестве лейкоцитов наименее эффективна схема hard_5, наиболее эффективна схема лечения hard_4.

При повышенном количестве тромбоцитов наименее эффективна схема hard_1.

При степени поражения легких КТ-3 наиболее эффективна схема hard_1, наименее эффективна схема лечения hard_5.

Для молодых заболевших наименее эффективна схема лечения hard_5, для пожилых пациентов – mos_hard_1.

Для больных сахарным диабетом наибольшей эффективностью обладает схема лечения hard_1, наименьшей – hard_5.

Цитокиновый шторм

Для цитокинового шторма наименьшей эффективностью обладает схема лечения cyto_1. Наибольшей эффективностью – схемы cyto_2 и mos_cyto_2.

При повышенном содержании D-димера наименьшей эффективностью обладает схема cyto_1, наибольшей эффективностью – mos_cyto_1.

При повышенном содержании ферритина наибольшей эффективностью обладают схемы mos_cyto_1 и mos_cyto_2.

При пониженном содержании нейтрофилов наименьшей эффективностью обладает схема cyto_1, наибольшей – cyto_2.

Для молодых пациентов более эффективна схема лечения cyto_2, для пожилых – mos_cyto_2.

3.2.6 Выводы

Были проведены исследование и разработка моделей оценки степени поражения по КТ в зависимости от результатов осмотра и анализа крови (клинического, СРБ, Д-димер, на

Ферритин и др.) с использованием регрессионных моделей, деревьев решений и их ансамблей, а также нейросетей. По результатам проведенных экспериментов наилучшие результаты показали разработанные модели RF с калибровкой, а также NN с калибровкой. Данные модели вошли в демонстрационный прототип «калькулятора» степени поражения по КТ, позволяющего без проведения КТ предсказать степень тяжести пневмонии пациента на основе его анализа крови.

Также были проведены исследование и разработка моделей прогнозирования риска летального исхода пациентов с использованием регрессионных моделей, деревьев решений и их ансамблей, а также нейросетей. Был выделен список хронических заболеваний, наиболее сильно влетающих на летальность, а также набор наиболее значимых признаков. Наилучший результат в данной задаче показала разработанная нейронная сеть.

Дополнительно, были проведены исследование и разработка моделей для предварительной статистической оценки эффективности схем лечения Covid-19. Было получено, что:

- Для легкой степени тяжести наименее эффективной во всех случаях является схема лечения light_1. Наиболее эффективными являются схемы лечения light_3 и mos_light_2. Средней эффективностью обладает схема mos_light_1.
- Для средней степени тяжести наименьшей эффективностью обладает схема mos_middle_2, наибольшей – mos_middle_1. Низкой эффективностью также обладает схема mos_middle_4. Высокая степень эффективности у схем middle_8, middle_4, middle_5.
- Для тяжелой степени тяжести наименьшей эффективностью обладает схема mos_hard_1, наибольшей – mos_hard_2. Низкая степень эффективности и у схемы лечения hard_5.
- Для цитокинового шторма наименьшей эффективностью обладает схема лечения cyto_1. Наибольшей эффективностью – схемы cyto_2 и mos_cyto_2.

3.3 Выводы

В рамках данного раздела были решены следующие задачи:

- Задача разработки моделей выживаемости;
- Задача построения описательных моделей прогнозирования летальности с функцией отбора важных предикторов;

- Задача анализа фактов появления и исчезновения положительного ПЦР;
- Задача анализа динамики появления и изменения иммуноглобулинов IgM и IgG;
- Задача построения моделей оценки степени поражения по КТ в зависимости от результатов осмотра и анализа крови (клинического, СРБ, Д-димер, на Ферритин и др.) с использованием регрессионных моделей, деревьев решений и их ансамблей, нейросетей;
- Задача построения моделей прогнозирования риска летального исхода пациентов;
- Задача построения моделей для предварительной статистической оценки эффективности схем лечения Covid-19.

4 Заключение

В рамках данного этапа работ были проведены научно-исследовательские работы в области исследования и разработки методов искусственного интеллекта и анализа больших данных в сфере здравоохранения.

Решены все поставленные задачи:

- Проведены анализ и обработка предоставленных исторических данных о развитии пандемии Covid-19 в г. Москве за 2020г, в частности:
  - очистка и приведение значений показателей к общим шкалам и словарям;
  - поиск, удаление или исправление артефактов, выбросов и противоречивых данных;
  - анализ и применение вероятностных методов множественной подстановки пропущенных значений для расчета ключевых показателей, таких как день течения заболевания при проведении анализа для случаев, когда эти данные не указаны или противоречивы.
- Проведена разработка на основе подготовленных данных моделей выживаемости для прогнозирования тяжести течения и исхода заболевания у пациентов, в частности:
  - моделей выживаемости с использованием методов Каплана-Мейера и пропорциональных рисков Кокса с выявлением ключевых признаков, влияющих на выживаемость, а также выявлением стратифицирующих признаков;
  - моделей выживаемости с учетом стратифицирующих признаков на основе использования методов Каплана-Мейера с выявлением важных предикторов внутри каждой из страт.

По результатам работы были опубликованы следующие научные статьи в журналах, индексируемых WoS / Scopus:

- 2023 Sensitivity of Survival Analysis Metrics. Vasilev Iulii, Petrovskiy Mikhail, Mashechkin Igor // в журнале Mathematics, издательство MDPI (Basel, Switzerland), том 11, № 20, с. 1-34 **(Q1)**.

- 2023 Adaptive Sampling for Weighted Log-Rank Survival Trees Boosting. Vasilev Iulii, Petrovskiy Mikhail, Mashechkin Igor //в журнале Lecture Notes in Computer Science, том 13822, с. 98-115.

Также за отчетный период были опубликованы следующие научные статьи в журналах, индексируемых WoS / Scopus / RSCI, по результатам, полученным на предыдущих этапах данной научно-исследовательской работы:

- 2023 Fuzzy CNN Autoencoder for Unsupervised Anomaly Detection in Log Data. Gorokhov Oleg, Petrovskiy Mikhail, Mashechkin Igor, Kazachuk Maria // в журнале Mathematics, издательство MDPI (Basel, Switzerland), том 11, № 18 **(Q1)**.
- 2023 SRGZ: Методы машинного обучения и свойства каталога оптических компаньонов точечных рентгеновских источников СРГ/ЕРОЗИТА в области покрытия Desi Legacy Imaging Surveys. Мещеряков А.В., Борисов В.Д., Хорунжев Г.А., Медведев П.А., Гильфанов М.Р., Бельведерский М.И., Сазонов С.Ю., Буренин Р.А., Кривонос Р.А., Бикмаев И.Ф., Хамитов И.М., Герасимов С.В., Машечкин И.В., Сюняев Р.А. // в журнале Письма в Астрономический журнал: Астрономия и космическая астрофизика, издательство ФГБУ "Издательство "Наука" (Москва), том 49, № 7, с. 441-494.
- 2023 Интеллектуальные технологии сегментации и классификации микробиологических фотоизображений. Горохов О.Е., Казачук М.А., Лазухин И.С., Машечкин И.В., Панкратьева Л.Л., Попов И.С. // в журнале Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика, издательство Изд-во Моск. ун-та (М.), № 4, с. 21-32.

Дополнительно, был получен патент на изобретение:

- 2023 Система и способ обнаружения и классификации колоний микроорганизмов на изображениях на основе технологий искусственного интеллекта и компьютерного зрения. Авторы: Казачук М.А., Горохов О.Е., Лазухин И.С., Машечкин И.В., Попов И.С. Номер 2791813.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Kozma L. k Nearest Neighbors algorithm (kNN) //Helsinki University of Technology. – 2008.
2. Rubin D. B. The bayesian bootstrap //The annals of statistics. – 1981. – С. 130-134.
3. Dmitrienko A., Koch G. G. Analysis of clinical trials using SAS: a practical guide. – SAS Institute, 2017.
4. Lenz S. T. Glenn A. Walker, Jack Shostak: Common statistical methods for clinical research with SAS examples 3rd edition. – 2012.
5. Gijbels I. et al. Almost sure asymptotic representation for a class of functionals of the Kaplan-Meier estimator //The Annals of Statistics. – 1991. – Т. 19. – №. 3. – С. 1457-1470.
6. Satten G. A., Datta S. The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average //The American Statistician. – 2001. – Т. 55. – №. 3. – С. 207-210.
7. Kaplan E. L., Meier P. Nonparametric estimation from incomplete observations //Journal of the American statistical association. – 1958. – Т. 53. – №. 282. – С. 457-481.
8. Bose A., Sen A. Asymptotic distribution of the Kaplan–Meier U-statistics //Journal of multivariate analysis. – 2002. – Т. 83. – №. 1. – С. 84-123.
9. Cox D. R. Regression models and life- tables //Journal of the Royal Statistical Society: Series B (Methodological). – 1972. – Т. 34. – №. 2. – С. 187-202.
10. Wang C. Y., Chen H. Y. Augmented inverse probability weighted estimator for Cox missing covariate regression //Biometrics. – 2001. – Т. 57. – №. 2. – С. 414-419.
11. White I. R., Royston P. Imputing missing covariate values for the Cox model //Statistics in medicine. – 2009. – Т. 28. – №. 15. – С. 1982-1998.
12. Cox D. R. Partial likelihood //Biometrika. – 1975. – Т. 62. – №. 2. – С. 269-276.
13. Andersen P. K., Gill R. D. Cox's regression model for counting processes: a large sample study:(preprint) //Stichting Mathematisch Centrum. Mathematische Statistiek. – 1981. – №. SW 73/81.
14. Machin D., Cheung Y. B., Parmar M. Survival analysis: a practical approach. – John Wiley & Sons, 2006.
15. Myers R. H., Myers R. H. Classical and modern regression with applications. – Belmont, CA : Duxbury press, 1990. – Т. 2.
16. Graybill F. A. Theory and application of the linear model. – North Scituate, MA : Duxbury press, 1976. – Т. 183.

17. Montgomery D. C., Peck E. A., Vining G. G. Introduction to linear regression analysis. – John Wiley & Sons, 2012. – Т. 821.
18. Seber G. A. F., Lee A. J. Linear regression analysis. – John Wiley & Sons, 2012. – Т. 329.
19. Yan X., Su X. Linear regression analysis: theory and computing. – World Scientific, 2009.
20. Douglas C.. Montgomery, Peck E. A., Vining G. G. Introduction to linear regression analysis. – Wiley, 2001.
21. Breiman L. et al. Classification and Regression Trees (The Wadsworth Statistics/Probability Series) Chapman and Hall //New York, NY. – 1984. – С. 1-358.
22. Gunluk O. et al. Optimal generalized decision trees via integer programming //arXiv preprint arXiv:1612.03225. – 2016.
23. Strobl C. et al. Conditional variable importance for random forests //BMC bioinformatics. – 2008. – Т. 9. – №. 1. – С. 307.
24. Genuer R., Poggi J. M., Tuleau-Malot C. Variable selection using random forests //Pattern recognition letters. – 2010. – Т. 31. – №. 14. – С. 2225-2236.
25. Biau G. Analysis of a random forests model //The Journal of Machine Learning Research. – 2012. – Т. 13. – №. 1. – С. 1063-1095.
26. Breiman L. Bagging predictors //Machine learning. – 1996. – Т. 24. – №. 2. – С. 123-140.
27. Kak S. C., Chen Y., Wang L. Data Mining Using Surface and Deep Agents Based on Neural Networks //AMCIS. – 2010. – С. 16.
28. Хайкин С. Нейронные сети: полный курс, 2-е издание. – Издательский дом Вильямс, 2008.
29. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain //Psychological review. – 1958. – Т. 65. – №. 6. – С. 386.
30. LeCun Y., Bengio Y., Hinton G. Deep learning //nature. – 2015. – Т. 521. – №. 7553. – С. 436-444.
31. Srivastava N. et al. Dropout: a simple way to prevent neural networks from overfitting //The journal of machine learning research. – 2014. – Т. 15. – №. 1. – С. 1929-1958.
32. Glover F. W., Kochenberger G. A. (ed.). Handbook of metaheuristics. – Springer Science & Business Media, 2006. – Т. 57.