Санкт-Петербургский государственный университет

Виль Мария Юрьевна

Выпускная квалификационная работа

Типологизация больных лимфомой по данным феноменологических результатов проточной цитометрии

Уровень образования: бакалавриат
Направление 01.03.02 «Прикладная математика и информатика»
Основная образовательная программа СВ.5005.2016 «Прикладная математика, фундаментальная информатика и программирование»
Профиль «Математическое и программное обеспечение вычислительных машин»

Научный руководитель:
профессор кафедры диагностики функциональных систем,
д. м. н. Шишкин Виктор Иванович

Рецензент:
доцент, к. м. н. Новик Алексей Викторович

Санкт-Петербург
2020

Содержание.

Введение.

Важность исследований в области онкологии в наше время неоспорима. По оценкам ВОЗ онкологические заболевания являются второй по частоте из основных причин смерти в мире, опережая даже ВИЧ-инфекцию и уступая только сердечно-сосудистым заболеваниям. Заболеваемость и смертность от рака быстро растут во всем мире. По статистике на 2018 год от рака погибли 9,6 миллионов человек, а значит, каждая шестая смерть в мире приходится на раковые заболевания [1]. Помимо того, что рак наносит огромный ущерб мировой демографии, его экономический эффект так же значителен, и он возрастает. Общий годовой экономический ущерб от рака в 2010 году оценивается примерно в 1,16 триллионов долларов США. [2]

В современной онкологии, благодаря высокому уровню оптимизации и простоте в эксплуатации, широкое распространение как в исследовательской, так и в клинико-диагностической деятельности получил метод проточной цитофлуориметрии (далее ПЦ) - современный метод исследования одиночных биологических клеток в дисперсных средах, основанный на измерении светорассеяния и специфического свечения частиц при прохождении через лазерный луч [3,7]. Специфическая флуоресценция обусловлена окрашиванием клеток особыми флюорохромами. Это помогает, например, определить жизнеспособность клетки. Светорассеяние определяется размерами клетки (прямое светорассеяние FS) и ее структурой (боковое светорассеяние SS) (рис.1).

Рис.1. Боковое и прямое светорассеяние

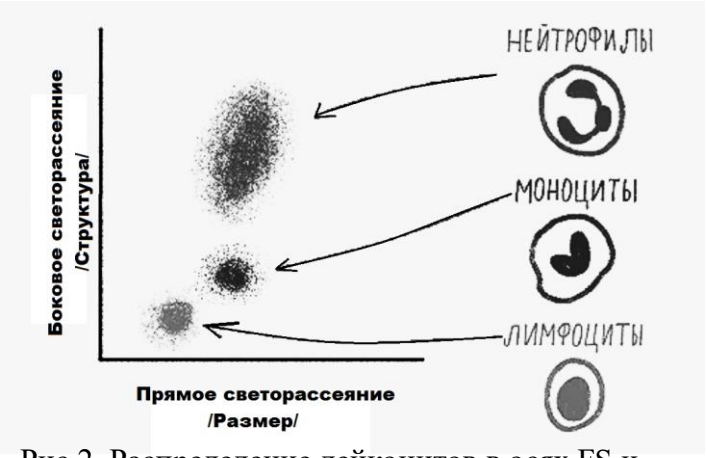

Рис.2. Распределение лейкоцитов в осях FS и SS

Комбинация прямого и бокового светорассеяния позволяет выделять различные субпопуляции белых клеток крови (лейкоцитов) (рис.2) для дальнейшего анализа, так как их можно разделить на три основные группы (лимфоциты, моноциты и гранулоциты (подавляющее большинство гранулоцитов составляют нейтрофилы, которые играют центральную роль в защите организма от инфекционных заболеваний)), отличающиеся между собой размерами и структурой.[5,8]


Рис.3. Стандартное распределение лейкоцитов в осях (слева), и распределение при появлении «пятого кластера» (справа).

При исследовании методов автоматической типологизации белых клеток крови в осях FS и SS А.В. Ореховым была замечена еще одна аномальная группы клеток (рис.3), которая получила условное название: “пятый кластер”.

В данной работе была сделана попытка определения причины появления данного явления в крови пациентов, больных некоторыми онкологическими заболеваниями, и, возможно, определить его медицинскую значимость в эффективности лечения.

Постановка задачи.

Целью данной работы является анализ статистической связи между полом, возрастом, а также некоторыми клиническими параметрами и появлением в крови пациентов, больных одним из трех рассматриваемых онкологических заболеваний, относящихся к гемобластозам[1] (лимфома Ходжкина, неходжкинская лимфома и множественная миелома), обнаруженной субпопуляции лейкоцитов.

Из клинических параметров были выбраны:

- Диагноз;
- Стадия заболевания;
- Исход;
- Предшествующий курс химиотерапии;
- Длительность заболевания;
- Особенности течения;
- Поражение костного мозга;
- Время от первого курса лечения до рецидива;
- Предшествующая лучевая терапия;
- Эффективность мобилизации;
- Стимуляция костного мозга (Применение препаратов, содержащих действующее вещество - филграстим. Провоцирует лейкопоэз[2]);
- Предшествующая DHAP (Особый вид химиотерапии, включающий препараты: дексаметазон, цитарабин, цисплатин. А также его подвид DHAP-R, который включает еще и ритуксимаб).

Отдельно также будут рассмотрены еще половая принадлежность и возраст.

[1] Гемобластоз - опухолевые заболевания кроветворной и лимфатической ткани.
[2] Лейкопоэз - образование лейкоцитов.

Глава 1. Подготовка к работе.

НМИЦ онкологии им. Н. Н. Петрова предоставил для исследования результаты анализа иммунного статуса пациентов, полученного методом ПЦ. Данные изначально представлены в числовом формате в виде таблицы, где одна строка соответствует одной клетке, прошедшей через лазерный пучок, а в столбцах обозначены параметры, такие как время, размер, структура и степень флуоресценции от окрашивания различными красителями. С помощью программы, написанной Ореховым А.В., есть возможность перевести все эти значения в график. Последовательным просмотром таких графиков в осях FS и SS было отобрано 100 пациентов у который в той или иной степени выделялся исследуемый кластер.

Далее, для каждого пациента был выделен краткий анамнез, включающий в том числе и указанные выше параметры. Было решено ограничиться рассмотрением только пациентов, больных лимфомой Ходжкина, неходжкинской лимфомой или множественной миеломой, так как эти три группы оказались наиболее многочисленными. В результате осталось 57 пациентов.

Была выделена контрольная группа - пациенты, больные теми же заболеваниями, но без “пятого кластера”, в количестве 57 человек. Для них также был рассмотрен краткий анамнез, включающий те же параметры.

В дополнение ко всему этому, группа пациентов с “пятым кластером” была разбита на три подгруппы по его размеру: маленький (30 пациентов), большой (8 пациентов), средний (19 пациентов). Стоит отметить субъективность данного разделения, весь отбор и сортировка производились после построения в ранее упомянутой программе «на глаз». Автоматизировать этот процесс не удалось в силу следующих обстоятельств:

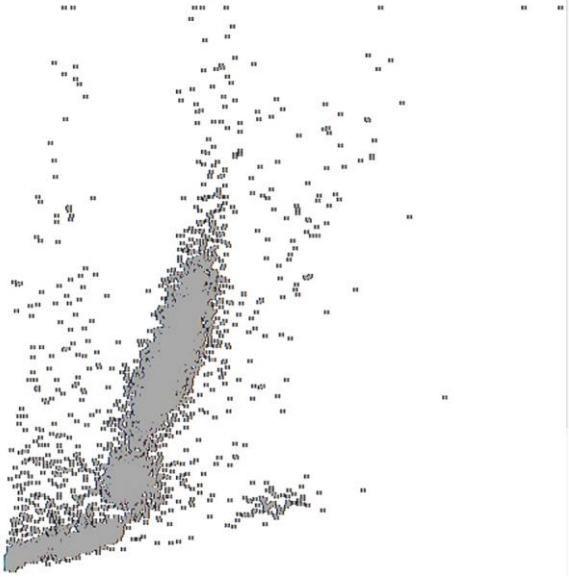

Рис.4. Клетки сосредоточены слева

- размеры и структура клеток крови уникальны для каждого организма, это может вызывать существенное смещение графика от среднестатистического, что усложняет выбор оптимального гейтирования[3]. Пример этому явлению можно увидеть на рисунке 4, клетки крови этого пациента имеют сравнительно малый размер;

- Наличие клеточных агрегатов[4] и дебриса[5] (рис.5) не позволяет точно рассчитать количество клеток в гейте, которые принадлежат непосредственно кластеру (даже если будет решена вышеуказанная проблема).

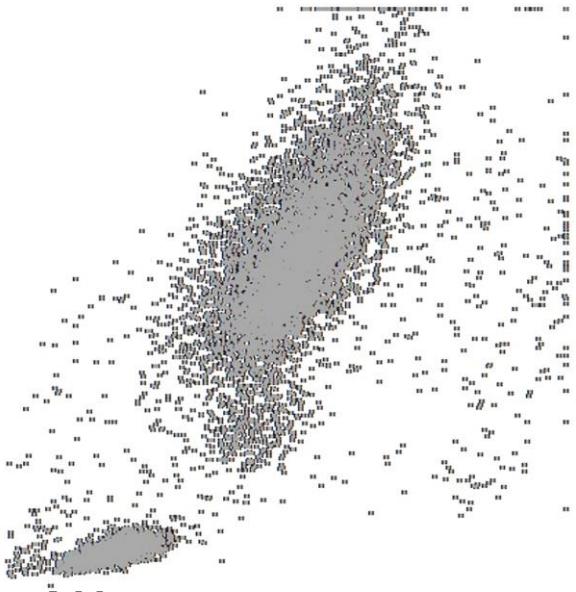

Рис.5. Много клеточных агрегатов справа

[3] Гейтирование - разбиение на области. Соответственно, гейт - область.
[4] Клеточные агрегаты - две или более клеток, “склеенных” между собой.
[5] Клеточный дебрис - остатки клеток после их разрушения.

Глава 2. Метод оценки статистической связи.

2.1. Метод для проверки гипотезы о независимости двух номинальных признаков.

В силу того, что большинство параметров являются номинальными, в качестве метода анализа будет использоваться критерий для проверки гипотезы о независимость двух номинальных признаков, который является практически полным аналогом критерия хи-квадрат (критерия Пирсона).[4]

Целесообразнее описывать теоретические положения данного критерия на примере, сразу в процессе анализа. Итак, у нас есть 2 признака: признак A – факт наличия кластера и признак B – диагноз. Признак A имеет категории: A_1 – есть кластер, A_2 – нет кластера; признак B имеет категории: B_1 – лимфома Ходжкина, B_2 – неходжкинская лимфома, B_3 – множественная миелома.

Введем случайные события:

A_i = {признак A у случайно выбранного объекта (пациента) имеет i-ую категорию (кластер либо есть, либо нет)}, $i = 1,\dots,m$; (в нашем случае $m = 2$);

B_j = {признак B у случайно выбранного пациента имеет j-ую категорию (пациент болен одним из трех заболеваний)}, $j = 1,\dots,k$; (в нашем случае $k = 3$).

Введем обозначения: $p_{i\bullet} = P(A_i)$, $p_{\bullet j} = P(B_j)$, $p_{ij} = P(A_i \cdot B_j)$, $i = 1,\dots,m, j = 1,\dots,k$.

Если $p_{ij} = p_{i\bullet} \cdot p_{\bullet j}, \forall\, i = 1,\dots,m, \forall\, j = 1,\dots,k$, то номинальные признаки A и B являются независимыми.

Были отобраны 114 (обозначается n) пациентов с онкологией (57 из них имеют на графике анализа иммунного статуса “пятый кластер”, а 57 - нет (контрольная группа)). Путем анализа анамнезов для каждого пациента были выделены категории признака B (в данном случае, диагноз; далее будут

рассмотрены остальные вышеупомянутые признаки) и их значения. Эти данные удобно представлять в виде таблицы сопряженности признаков A и B размера $m \times k$ (Табл.1).

	B_1	…	B_k	
A_1	n_{11}	…	n_{1k}	$n_{1\bullet}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
A_m	n_{m1}	…	n_{mk}	$n_{m\bullet}$
	$n_{\bullet 1}$	…	$n_{\bullet k}$	n

Таб.1. Таблица сопряженности признаков A и B

В таблице 1: n_{ij} – количество пациентов, у которых признак A имеет категорию A_i , а признак B имеет категорию B_j; $n_{i\bullet} = \sum_{j=1}^{k} n_{ij}$ – количество пациентов, у которых признак A имеет категорию A_i; $n_{\bullet j} = \sum_{i=1}^{m} n_{ij}$ – количество пациентов, у которых признак B имеет категорию B_j.

В рассматриваемом примере таблица получилась такого вида:

	Лимфома Ходжкина	Неходжкинская лимфома	Множественная миелома	
Группа с кластером	12	35	10	57
Контрольная группа	11	36	10	57
	23	71	20	114

По таблице сопряженности признаков мы можем оценить неизвестные вероятности $p_{ij}, p_{i\bullet}, p_{\bullet j}, i = 1,\dots,m, j = 1,\dots,k$. Пусть $p_{ij}^{*} = \frac{n_{ij}}{n}$ – частота случайного события $A_i \cdot B_j$, $i = 1,\dots,m$, $j = 1,\dots,k$; $p_{i\bullet}^{*} = \frac{n_{i\bullet}}{n}$ – частота случайного события A_i; $p_{\bullet j}^{*} = \frac{n_{\bullet j}}{n}$ – частота случайного события B_j. Эти частоты – сильно состоятельные, несмещенные оценки вероятностей, соответствующих им.

Зададим нулевую гипотезу в виде:

$$H_0: p_{ij} = p_{i\bullet} \cdot p_{\bullet j}, \forall\, i = 1,\dots,m, \forall\, j = 1,\dots,k.$$

То есть положим, что признаки A и B являются независимыми. Для

проверки этой гипотезы применяется критерий Пирсона хи-квадрат со статистикой вида:

$$\chi_n^2 = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(p_{ij}^* - p_{i\bullet}^* p_{\bullet j}^*)^2}{p_{i\bullet}^* p_{\bullet j}^*} = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij} - n_{i\bullet} n_{\bullet j})^2}{n_{i\bullet} n_{\bullet j}} \qquad (1)$$

При справедливости гипотезы H_0 и $n \to \infty$ статистика (1) имеет распределение хи-квадрат с $r = (m-1)(k-1)$ степенями свободы, а критическая область уровня значимости α имеет вид $(k_{1-\alpha}(r); +\infty)$, где $k_{1-\alpha}(r)$ - квантиль уровня 1-α распределения хи-квадрат с r степенями свободы (Приложение 1).

Рассмотренный критерий состоятелен против альтернативы общего вида:

$$H_A: \exists\, i, j \text{ такие, что } p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j}.$$

2.2. Меры связи, основанные на статистике χ_n^2.

С учетом вида критической области, очевиден тот факт, что большие значения статистики (1) свидетельствуют о наличии зависимости между признаками A и B, но непосредственно основываясь на величине χ_n^2 нельзя судить о степени связи между ними, т.к. $\chi_n^2 \to \infty$ при $n \to \infty$, если признаки А и В зависимы.

В качестве меры связи признаков А и В можно рассматривать коэффициент взаимной сопряженности Пирсона:

$$P = \sqrt{\frac{\chi_n^2}{\chi_n^2 + n}};$$

Этот коэффициент называется коэффициентом взаимной сопряженности или коэффициентом Пирсона. При чем при возрастании m и k $P^2 \to r^2$, здесь r - коэффициент корреляции. Но, в отличие от r, максимум P равен $\sqrt{\frac{l-1}{l}} < 1$, где $l = min(m-1; k-1)$. Для устранения этого недостатка Крамер ввел другую меру связи:

$$C = \sqrt{\frac{\chi_n^2}{n \cdot min((m-1);(k-1))}}.$$

Она называется коэффициентом Крамера.

$C \in [0; 1]$ и верхний предел $C = 1$ достигается тогда и только тогда, когда каждая строка (при m ≥ k) или каждый столбец (при m ≤ k) в таблице содержит лишь 1 отличный от нуля элемент.

Если гипотеза о независимости признаков отвергнута, то принято считать, что значения P и C:

- $\in [0; 0,3)$ говорят о слабой связи признаков;
- $\in [0,3; 0,7)$- о средней связи;
- $\in [0,7; 1]$- о значительной связи.

Вообще говоря, коэффициенты Крамера и Пирсона принято рассчитывать только в случае, когда гипотеза о независимости признаков отвергнута, но здесь и далее будем рассчитывать эти коэффициенты даже если это не так, для наглядности.

2.3. Выбор уровня значимости α.

При исследовании выборки всегда имеется вероятность того, что полученный вывод может быть ошибочным. Существует два типа статистических ошибок:

- Ошибка первого рода;
- Ошибка второго рода.

Ошибка первого рода – это ошибка принятия решения, в результате которого истинная нулевая гипотеза отклоняется. Это значит, что связь обнаруживается там, где в действительности ее нет. В свою очередь, ошибка второго рода заключается в принятии нулевой гипотезы, которая при этом является ложной, то есть в том, чтобы не обнаружить связь там, где она есть. С этими понятиями напрямую связан вопрос о выборе уровня значимости.

Уровень значимости - это критическая вероятность ошибки второго рода. Иначе говоря, это допустимая с точки зрения исследователя вероятность того, что связь сочтена существенной в то время, как она отсутствует [6].

Однозначно ответить на вопрос о приемлемом уровне значимости нельзя. Чтобы минимизировать ошибку первого рода, следует брать очень малые значения α (обычно это 0,01 или 0,05), так как вероятность ошибки первого рода снижается при уменьшении уровня значимости и соответственно уменьшается размер критической области. Однако в то же время при неизменном объеме выборки вероятность ошибки второго рода увеличивается. То есть ошибки имеют обратную зависимость друг от друга, следовательно, нельзя свести к минимуму обе ошибки. Поэтому нужно выбрать такой уровень значимости, который был бы балансом между ними: более высокие значения α сведут к минимуму вероятность ошибки второго рода, а более низкие снизят вероятность ошибки первого рода.

Наиболее разумным с точки зрения этого исследования является решение не повышать вероятность ошибки второго рода за счет уменьшения вероятности ошибки первого рода, так как гораздо критичнее будет пропустить связь там, где она есть. Поэтому все дальнейшие вычисления будем проводить на уровне значимости $\alpha = 0,2$.

Глава 3. Исследование без учета размера.

3.1. Диагноз.

Продолжая рассуждения, начатые ранее, получаем:

- Наблюдаемое значение критерия: 0,05756;
- Степени свободы: 2;
- Табличное значение квантиля: 3,22;
- Критическая область: (3,22; +∞).
- P = 0,022465;

- C = 0,022471;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера не зависит от диагноза.

Судя по значениям коэффициентов, можно сказать, что зависимость между признаками A и B очень слабая, чего и следовало ожидать, ведь гипотеза о независимости была принята.

3.2. Курс химиотерапии.

Здесь и далее все рассуждения проводятся аналогично, меняется только признак B.

	Не было	Первый	Второй	Третий	
Группа с кластером	34	12	10	1	57
Контрольная группа	38	15	4	0	57
	72	27	14	1	114

- Наблюдаемое значение критерия: 4,12698;
- Степени свободы: 3;
- Табличное значение квантиля: 4,64;
- Критическая область: (4,64; +∞)
- P = 0,186914004;
- C = 0,19026722.

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, мы можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера не зависит от курса химиотерапии.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

3.3. Исход.

	Жив	Мертв	
Группа с кластером	53	4	57
Контрольная группа	54	3	57
	107	7	114

- Наблюдаемое значение критерия: 0,1522;
- Степени свободы: 1;
- Табличное значение квантиля: 1,64;
- Критическая область: (1,64; +∞)
- P = 0,036514837;
- C = 0,036539205;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера не зависит от исхода.

Судя по значениям коэффициентов, можно сказать, что связь очень слабая.

3.4. Особенности течения.

	Прогрессирование	Резистентное к терапии	Ответ на ХТ	
Группа с кластером	20	2	34	56
Контрольная группа	22	3	32	57
	44	5	66	113

- Наблюдаемое значение критерия: 0,34702;
- Степени свободы: 2;
- Табличное значение квантиля: 3,22;

- Критическая область: (3,22; +∞)
- P = 0,05533161;
- C = 0,055416506;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера не зависит от исхода.

Судя по значениям коэффициентов, можно сказать, что связь очень слабая.

3.5. Поражение костного мозга.

	Не было	Опухолевое	
Группа с кластером	46	11	57
Контрольная группа	47	10	57
	93	21	114

- Наблюдаемое значение критерия: 0,05837;
- Степени свободы: 1;
- Табличное значение квантиля: 1,64;
- Критическая область: (1,64; +∞)
- P = 0,02262235;
- C = 0,022628141;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера не зависит от поражения костного мозга.

Судя по значениям коэффициентов, можно сказать, что связь очень слабая.

3.6. Время до рецидива от первого лечения.

	Не было	До года	1-2 года	2-3 года	3-4 года	4-5 лет	7-8 лет	9-10 лет	
Группа с кластером	43	9	1	3	0	0	1	0	57
Контрольная группа	47	2	1	2	2	2	0	1	57
	90	11	2	5	2	2	1	1	114

- Наблюдаемое значение критерия: 10,83232323;
- Степени свободы: 7;
- Табличное значение квантиля: 9,80;
- Критическая область: (9,8; +∞)
- P = 0,294575944;
- C = 0,308253758;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и временем до рецидива от первого лечения.

Если судить по коэффициенту Пирсона, то связь принадлежит к разряду слабых. Коэффициент Крамера говорит о средней связи.

3.7. Предшествующая лучевая терапия.

	Да	Нет	
Группа с кластером	8	49	57
Контрольная группа	3	54	57
	11	103	114

- Наблюдаемое значение критерия: 2,51544;
- Степени свободы: 1;

- Табличное значение квантиля: 1,64;
- Критическая область: (1,64; +∞)
- P = 0,1469317748;
- C = 0,1485439778;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и предшествующей лучевой терапией.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

3.8. Длительность заболевания.

	Не было	До года	1-2 года	2-3 года	3-4 года	4-5 лет	7-8 лет	10-11 лет	
Группа с кластером	2	39	11	2	1	1	0	1	57
Контрольная группа	0	34	7	6	4	4	1	1	57
	2	73	18	8	5	5	1	2	114

- Наблюдаемое значение критерия: 9,83136;
- Степени свободы: 7;
- Табличное значение квантиля: 9,80;
- Критическая область: (9,8; +∞)
- P = 0,2817678053;
- C = 0,2936663975;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть

основания полагать, что существует связь между появлением кластера и длительностью заболевания.

Судя по коэффициентам сопряженности связь слабая, хоть значения и достаточно близки к нижней границе средней связи.

3.9. Эффективность мобилизации.

	Недостаточная	Менее 1,00	1,00-5,00	5,00-10,00	10,00-15,00	15,00-20,00	20,00-25,00	25,00-30,00	
Группа с кластером	15	3	13	17	6	3	0	0	57
Контрольная группа	10	1	22	13	4	3	2	2	57
	25	4	35	30	10	6	2	2	114

- Наблюдаемое значение критерия: 9,247619048;
- Степени свободы: 7;
- Табличное значение квантиля: 9,80;
- Критическая область: (9,8; +∞)
- P = 0,273921232;
- C = 0,2848147913;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, т.е. нулевую гипотезу следует принять и возможность появления кластера не зависит от эффективности мобилизации.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

3.10. Стимуляция костного мозга.

	Да	нет	
Группа с кластером	54	3	57
Контрольная группа	57	0	57
	111	3	114

- Наблюдаемое значение критерия: 3,08108;
- Степени свободы: 1;
- Табличное значение квантиля: 1,64;
- Критическая область: (1,64; +∞)
- P = 0,1622214211;
- C = 0,1643989873;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и стимуляцией костного мозга.

Судя по коэффициентам сопряженности связь слабая.

3.11. DHAP.

	Нет	Да	DHAP-R	
Группа с кластером	34	17	6	57
Контрольная группа	38	14	5	57
	72	31	11	114

- Наблюдаемое значение критерия: 0,60345;
- Степени свободы: 2;
- Табличное значение квантиля: 3,22;
- Критическая область: (3,22; +∞)
- P = 0,07256433067;
- C = 0,0727561352;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что

нулевую гипотезу следует принять, т.е. возможность появления кластера не зависит от применения DHAP.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

3.12. Стадия.

	I	II	III	IV	
Группа с кластером	2	12	10	30	54
Контрольная группа	0	12	18	22	52
	2	24	28	52	106

- Наблюдаемое значение критерия: 5,480699;
- Степени свободы: 3;
- Табличное значение квантиля: 4,64;
- Критическая область: (4,64; +∞)
- P = 0,2217267678;
- C = 0,2273866873;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и стадией заболевания.

Судя по коэффициентам сопряженности связь слабая.

3.13. Пол.

	Мужской	Женский	
Группа с кластером	33	24	57
Контрольная группа	25	32	57
	58	56	114

- Наблюдаемое значение критерия: 2,2463;
- Степени свободы: 1;
- Табличное значение квантиля: 1,64;
- Критическая область: (1,64; +∞)
- P = 0,1390096094;
- C = 0,1403724813;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и полом.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

3.14. Возраст.

	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-65	
группа с кластером	4	4	9	6	5	4	9	11	5	57
Контрольная группа	1	3	11	10	8	5	8	5	6	57
	5	7	20	16	13	9	17	16	11	114

- Наблюдаемое значение критерия: 6,346;
- Степени свободы: 8;
- Табличное значение квантиля: 11,03;
- Критическая область: (11,03; +∞)
- P = 0,229633096;
- C = 0,235938004;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что

нулевую гипотезу следует принять, т.е. возможность появления кластера не зависит от возраста.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

Глава 4. Исследование с учетом размера.

Дальнейшее изучение взаимосвязи признаков будем проводить аналогично приведенному выше, но более детализировано. А именно, рассмотрим в качестве категорий признака А не только факт наличия или отсутствия, но и еще разделение кластера по размеру.

Итак, введем новые категории признака A: A_1 – кластер отсутствует, A_2 – кластер маленького размера, A_3 – кластер среднего размера, A_4 – большой кластер. Далее проведем аналогичный анализ тех же признаков с учетом данных изменений.

4.1. Диагноз.

	Лимфома Ходжкина	Неходжкинская лимфома	Множественная миелома	
Контрольная группа	11	36	10	57
Маленький кластер	5	18	7	30
Средний кластер	3	14	2	19
Большой кластер	4	3	1	8
	23	71	20	114

- Наблюдаемое значение критерия: 6,35315;
- Степени свободы: 8;
- Табличное значение квантиля: 11,03;
- Критическая область: (11,03; +∞)
- P = 0,2297554358;

- C = 0,166927197;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера и его размер не зависят от диагноза.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

4.2. Курс химиотерапии.

	Не было	Первый	Второй	Третий	
Контрольная группа	38	15	4	0	57
Маленький кластер	16	8	5	1	30
Средний кластер	13	2	4	0	19
Маленький кластер	5	2	1	0	8
	72	27	14	1	114

- Наблюдаемое значение критерия: 8,07398;
- Степени свободы: 9;
- Табличное значение квантиля: 12,24;
- Критическая область: (12,24; +∞)
- P = 0,257177016;
- C = 0,2661248429;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера и его размер не зависят от курса химиотерапии.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

4.3. Исход.

	Мертв	Жив	
Контрольная группа	3	54	57
Маленький кластер	4	26	30
Средний кластер	0	19	19
Большой кластер	0	8	8
	7	107	114

- Наблюдаемое значение критерия: 4,53564753;
- Степени свободы: 3;
- Табличное значение квантиля: 4,64;
- Критическая область: (4,64; +∞)
- P = 0,19561185;
- C = 0,19946524;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, т.е. нулевую гипотезу следует принять, а возможность появления кластера и его размер не зависят от исхода.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

4.4. Особенности течения.

	Прогрессирование	Резистентное к терапии	Ответ на ХТ	
Контрольная группа	22	3	32	57
Маленький кластер	12	1	17	30
Средний кластер	4	1	13	18
Большой кластер	4	0	4	8
	42	5	66	113

- Наблюдаемое значение критерия: 2,86023;
- Степени свободы: 6;
- Табличное значение квантиля: 8,56;
- Критическая область: (8,56; +∞)
- P = 0,1571206396;
- C = 0,112498365;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера и его размер не зависят от особенностей течения.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

4.5. Поражение костного мозга.

	Нет	Опухолевое	
Контрольная группа	47	10	57
Маленький кластер	24	6	30
Средний кластер	17	2	19
Большой кластер	5	3	8
	93	21	114

- Наблюдаемое значение критерия: 2,80476;
- Степени свободы: 3;
- Табличное значение квантиля: 4,64;
- Критическая область: (4,64; +∞)
- P = 0,1549593285;
- C = 0,1568539914;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что

нулевую гипотезу следует принять, т.е. возможность появления кластера и его размер не зависят от поражения костного мозга.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

4.6. Время до рецидива от первого лечения.

	Не было	До года	1-2 года	2-3 года	3-4 года	4-5 лет	7-8 лет	9-10 лет	
Контрольная	47	2	1	2	2	2	0	1	57
Маленький	23	4	1	2	0	0	0	0	30
Средний	15	4	0	0	0	0	0	0	19
Большой	5	1	0	1	0	0	1	0	8
	90	11	2	5	2	2	1	1	114

- Наблюдаемое значение критерия: 27,30005;
- Степени свободы: 21;
- Табличное значение квантиля: 26,17;
- Критическая область: (26,17; +∞)
- P = 0,4395523099;
- C = 0,282532669;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и временем до рецидива от первого лечения.

Судя по коэффициенту сопряженности Крамера связь слабая, по коэффициенту Пирсона связь принадлежит к разряду средних.

4.7. Предшествующая лучевая терапия.

	Да	Нет	
Контрольная группа	3	54	57
Маленький кластер	5	25	30
Средний кластер	2	17	19
Большой кластер	1	7	8
	11	103	114

- Наблюдаемое значение критерия: 3,04369;
- Степени свободы: 3;
- Табличное значение квантиля: 4,64;
- Критическая область: (4,64; +∞)
- P = 0,1612598156;
- C = 0,1633983755;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. появление кластера и его размер не зависят от предшествующей лучевой терапии.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

4.8. Длительность заболевания.

	Не было	До года	1-2 года	2-3 года	3-4 года	4-5 лет	7-8 лет	10-11 лет	
Контрольная	0	34	7	6	4	4	1	1	57
Маленький	0	21	6	1	1	1	0	0	30
Средний	2	12	5	0	0	0	0	0	19
Большой	0	6	0	1	0	0	0	1	8
	2	73	18	8	5	5	1	2	114

- Наблюдаемое значение критерия: 28,56944;
- Степени свободы: 21;
- Табличное значение квантиля: 26,17;
- Критическая область: (26,17; +∞)
- P = 0,4476490161;
- C = 0,2890265915;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и длительности заболевания.

Судя по коэффициенту сопряженности Крамера связь слабая, по коэффициенту Пирсона связь принадлежит к разряду средних.

4.9. Эффективность мобилизации.

	Недостаточная	Менее 1,00	1,00-5,00	5,00-10,00	10,00-15,00	15,00-20,00	20,00-25,00	25,00-30,00	
Контрольная	10	1	22	13	4	3	2	2	57
Маленький	7	1	6	11	4	1	0	0	30
Средний	7	0	4	5	2	1	0	0	19
Большой	1	2	3	1	0	1	0	0	8
	25	4	35	30	10	6	2	2	114

- Наблюдаемое значение критерия: 26,20752381;
- Степени свободы: 21;
- Табличное значение квантиля: 26,17;
- Критическая область: (26,17; +∞)
- P = 0,432341908;
- C = 0,276821579;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и эффективностью мобилизации.

Судя по коэффициенту сопряженности Крамера связь слабая, по коэффициенту Пирсона связь принадлежит к разряду средних.

4.10. Стимуляция костного мозга.

	Нет	Да	
Контрольная группа	0	57	57
Маленький кластер	1	29	30
Средний кластер	1	18	19
Большой кластер	1	7	8
	3	111	114

- Наблюдаемое значение критерия: 5,15225;
- Степени свободы: 3;
- Табличное значение квантиля: 4,64;
- Критическая область: (4,64; +∞)
- P = 0,2079444966;
- C = 0,212591616;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и стимуляцией костного мозга.

Судя по коэффициентам сопряженности связь слабая.

4.11. DHAP.

	Нет	Да	DHAP-R	
Контрольная группа	38	14	5	57
Маленький кластер	16	9	5	30
Средний кластер	13	5	1	19
Большой кластер	5	3	0	8
	72	31	11	114

- Наблюдаемое значение критерия: 3,93075;
- Степени свободы: 6;
- Табличное значение квантиля: 8,56;
- Критическая область: (8,56; +∞)
- P = 0,1825676862;
- C = 0,1313015972;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера и его размер не зависят от применения DHAP.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

4.12. Стадия.

	I	II	III	IV	
Контрольная группа	0	12	18	22	52
Маленький кластер	2	7	7	12	28
Средний кластер	0	3	2	13	18
Большой кластер	0	2	1	5	8
	2	24	28	52	106

- Наблюдаемое значение критерия: 12,441399;
- Степени свободы: 9;
- Табличное значение квантиля: 12,24;
- Критическая область: (12,24; +∞)
- P = 0,3241028323;
- C = 0,1977976049;

Наблюдаемое значение критерия принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует отвергнуть в пользу альтернативной, т.е. есть основания полагать, что существует связь между появлением кластера и стадией заболевания.

Судя по коэффициенту сопряженности Крамера связь слабая, по коэффициенту Пирсона связь средняя.

4.13. Пол.

	Женский	Мужской	
Контрольная группа	32	25	57
Маленький кластер	12	18	30
Средний кластер	9	10	19
Большой кластер	3	5	8
	56	58	114

- Наблюдаемое значение критерия: 2,57799;
- Степени свободы: 3;
- Табличное значение квантиля: 4,64;
- Критическая область: (4,64; +∞)
- P = 0,1487072151;
- C = 0,1503792414;

Наблюдаемое значение критерия не принадлежит критической области при выбранном α, следовательно, нулевую гипотезу следует принять, т.е. возможность появления кластера и его размер не зависят от пола.

Судя по значениям коэффициентов, можно сказать, что связь слабая.

4.14. Возраст.

	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-65	
Контроль	1	3	11	10	8	5	8	5	6	57
Маленький	2	2	5	2	4	2	4	6	3	30
Средний	0	2	3	2	1	0	5	4	2	19
Большой	2	0	1	2	0	2	0	1	0	8
	5	7	20	16	13	9	17	16	11	114

- Наблюдаемое значение критерия: 26,68022213;
- Степени свободы: 24;
- Табличное значение квантиля: 29,55;
- Критическая область: (29,55; +∞)
- P = 0,435490011;
- C = 0,2793069019;

Наблюдаемое значение критерия не принадлежит критической области при выбранном уровне значимости, следовательно, можем сделать вывод, что нулевую гипотезу следует принять, т.е. возможность появления кластера и его размер не зависят от возраста.

Судя по коэффициенту сопряженности Крамера связь слабая, по коэффициенту Пирсона связь средняя.

Выводы.

Как видно из глав 3-4, значимой связи между появлением такого явления, как “пятый кластер” и рассмотренными параметрами найти не удалось. Более того, даже если мы и можем причислить какую-либо связь к разряду средних, то сделать мы это можем лишь по одному из рассматриваемых коэффициентов сопряженности (в случае рассмотрения без учета размера — это коэффициент Крамера, с учетом — коэффициент Пирсона). При этом детализация с добавлением разбиения пациентов исследуемой группы по размеру кластера хоть и увеличивает число связей, отнесенных к разряду средних, но только за счет увеличения коэффициента Пирсона, в то время как коэффициент Крамера, даже если и увеличивается, то незначительно (той точки зрения, что по введенной классификации связей этого увеличения недостаточно для перехода от слабой связи к средней), а в некоторых случаях и уменьшается. Последнее связано с тем, что при рассмотрении без учета размера $min((m-1);(k-1))$, расположенный в знаменателе формулы коэффициента Крамера, всегда равен 1, а в случае учета размера ≥ 1.

Заключение.

К сожалению, несмотря на проделанную работу, действительно значимой связи найти пока не удалось. Однако, даже принимая во внимание этот факт, вполне обоснованно считать, что этот “отрицательный” результат, тоже важен, ведь он дает понять, что, требуется либо искать причины в совсем другом направлении, либо нужно “копать” глубже и рассматривать некоторые аспекты более детально. А может получится так, что даже та детализации, с которой было проведено исследование в этой работе, избыточна и рассматриваемое явление не имеет вообще никакого отношения к раку и это более общее медицинское явление, но в любом случае, необходимо найти его

причины, ведь оно присутствует в крови не у всех людей, даже среди онкобольных. Решить вопрос дальнейших действий, на мой взгляд, поможет существенное увеличение исследуемой выборки. Проблема заключается лишь в отсутствии на данный момент достаточного количества данных, в остальном же подход к работе был организован так, что пересчет всех значений на более крупном массиве данных не вызовет трудностей.

Как бы там ни было, любое научное исследование в области медицины, а тем более в области онкологии, имеет очевидную ценность для общества, ведь чем лучше мы понимаем человеческий организм и природу свойственных ему заболеваний, тем более эффективно мы можем разрабатывать пути лечения и делать жизнь людей лучше.

Список литературы.

1. Bray F., Ferlay J., Soerjomataram I., Siegel R.L., Torre L.A., Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries // CA: A Cancer Journal for Clinicians, 2018, volume 68, issue 6. C. 394-424.
2. Stewart B.W., Wild C.P., editors World cancer report 2014. Lyon: International Agency for Research on Cancer, 2014. 630 с.
3. Балалаева И.В. Проточная цитофлуориметрия: Учебно-методическое пособие. Нижний Новгород: Нижегородский госуниверситет, 2014. 75 с.
4. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: Издательский дом Высшей школы экономики, 2012. 312 с.
5. Дранник Г.М. Клиническая иммунология и аллергология. Одесса: Астро-Принт, 1999. 603 с.
6. Дубина И.Н. Математические основы эмпирических социально-экономических исследований: учебное пособие. Барнаул: Изд-во Алт. ун-та., 2006. 263 с.
7. Зурочка А.В., Хайдуков С.В., Кудрявцев И.В., Черешнев В.А. Проточная цитометрия в медицине и биологии. 2-е изд. Екатеринбург: УрО РАН, 2014. 574 с.
8. Хаитов Р.М., Игнатьева Г.А., Сидорович И.Г. Иммунология: Учебник. М.: Медицина, 2000. 432 с.

Приложения.

r	α								
	0,05	0,1	0,2	0,3	0,5	0,7	0,8	0,9	0,95
1	0,00	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84
2	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99
3	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,82
4	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49
5	1,15	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07
6	1,65	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59
7	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07
8	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51
9	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92
10	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31
11	4,58	5,58	6,99	8,15	10,34	12,90	14,63	17,28	19,68
12	5,23	6,30	7,81	9,03	11,34	14,01	15,81	18,55	21,03
13	5,89	7,04	8,63	9,93	12,34	15,12	16,98	19,81	22,36
14	6,57	7,79	9,47	10,82	13,34	16,22	18,15	21,06	23,69
15	7,26	8,55	10,31	11,72	14,34	17,32	19,31	22,31	25,00
16	7,96	9,31	11,15	12,62	15,34	18,42	20,47	23,54	26,29
17	8,67	10,08	12,00	13,53	16,34	19,51	21,62	24,78	27,60
18	9,39	10,86	12,86	14,44	17,34	20,60	22,76	25,59	28,87
19	10,11	11,65	13,72	15,35	18,34	21,70	23,90	27,20	30,14
20	10,85	12,44	14,58	16,27	19,34	22,80	25,04	28,41	31,41
21	11,59	13,24	15,44	17,18	20,30	23,90	26,17	29,61	32,67
22	12,34	14,04	16,31	18,10	21,30	24,90	27,30	30,81	33,92
23	13,09	14,85	17,19	19,02	22,30	26,00	28,43	32,01	35,17
24	13,85	15,66	18,06	19,94	23,30	27,10	29,55	33,20	36,42
25	14,61	16,47	18,94	20,90	24,30	28,20	30,78	34,38	37,65
26	15,38	17,29	19,82	21,80	25,30	29,20	31,80	35,56	38,89
27	16,15	18,11	20,70	22,70	26,30	30,30	32,91	36,74	40,11
28	16,93	18,94	21,60	23,60	27,30	31,40	34,03	37,92	41,34
29	17,71	19,77	22,50	24,60	28,30	32,50	35,14	39,09	42,56
30	18,49	20,60	23,40	25,50	29,30	33,50	36,25	40,26	43,77

Приложение 1. Таблица квантилей k_αуровня α распределения хи-квадрат.