Санкт-Петербургский государственный университет

**Факультет прикладной математики – процессов управления**

**Кафедра диагностики функциональных систем**

**Бурлакова Людмила Сергеевна**

**Выпускная квалификационная работа бакалавра**

**Сравнительная характеристика четырех групп больных реактивным артритом с помощью методов математической статистики**

Направление 010400

Прикладная математика и информатика

Научный руководитель,  
доктор медицинских наук,

профессор

Шишкин В. И.

Санкт-Петербург

2017

Содержание

[Введение 3](#_Toc482296343)

[Постановка задачи 5](#_Toc482296344)

[Обработка материала 7](#_Toc482296345)

[**Глава 1. Дискриминантный анализ** 9](#_Toc482296346)

[**1.1** Принцип дискриминации 9](#_Toc482296348)

[**1.2** Канонический дискриминантный анализ 10](#_Toc482296349)

[1.2.1 Дискриминантные функции 10](#_Toc482296350)

[1.2.2Коэффициенты канонической дискриминантной функции 11](#_Toc482296351)

[1.2.3 Ограничения 16](#_Toc482296352)

[**1.3** Пошаговый дискриминантный анализ 17](#_Toc482296353)

[1.3.1 Отбор информативных признаков 17](#_Toc482296354)

[1.3.2 Критерии отбора 18](#_Toc482296355)

[1.3.3 Алгоритм Forward Stepwise 21](#_Toc482296356)

[**Глава 2. Результаты исследования** 23](#_Toc482296357)

[Пошаговый дискриминантный анализ 23](#_Toc482296359)

[Канонический анализ 25](#_Toc482296360)

[Визуализация 27](#_Toc482296361)

[Классификация наблюдений 29](#_Toc482296362)

[Выводы 32](#_Toc482296363)

[Заключение 34](#_Toc482296364)

[Список литературы 35](#_Toc482296365)

[Приложения 36](#_Toc482296366)

Введение

В настоящее время медицина в своих исследованиях все чаще обращается к достижениям математических наук, которые позволяют точнее выявлять и трактовать многообразные скрытые связи основных клинико-лабораторных параметров, а также их комбинаций.

Реактивный артрит - воспалительное негнойное заболевание суставов, причины возникновения которого доподлинно неизвестны. Замечено, что это заболевание развивается одновременно или в течение 2-4 недель после перенесенной внесуставной инфекции [1]. Данное заболевание не связано с распространением инфекции по организму и её попаданием в сустав, инфекция здесь играет роль только пускового фактора. То есть артрит - это реакция организма на микроб. Существует предположение, что в основе реактивного артрита лежит генетически детерминированная аномалия иммунной системы (заболевание в 50 раз чаще диагностируется у носителей антигена гистосовместимости HLA-B27 [2]), которая реализуется при инфицировании некоторыми микроорганизмами. Инфекция, вызывающая развитие реактивного артрита, обычно поражает носоглотку, мочевыводящие пути и половые органы, желудочно-кишечного тракт. Таким образом, можно сделать предположение о природе возникновения реактивного артрита и выделить четыре «формы» болезни, разделенные по этиологическому принципу (типу инфекции-возбудителя): носоглоточная, кишечная, урогенитальная и группа с не выявленным возбудителем.

В нашем распоряжении имеются результаты клинического исследования больных реактивным артритом, представленные в виде табличных данных. В таблице отражена информация по 104 индивидуумам, каждый из которых характеризуется набором из 58 признаков. Признаками здесь являются клинические, лабораторные и категориальные показатели, зафиксированные при обращении в медицинское учреждение.

С математической точки зрения, основная цель исследования состоит в нахождении способа, позволяющего на основании этих характеристик определить группу, к которой относится каждый из объектов в имеющейся совокупности. Сопутствующей задачей является задача снижения размерности исходного пространства признаков, или, иначе говоря, формирование оптимального множества признаков, по которым различение по группам внутри имеющегося набора данных осуществлялась бы наиболее верно.

С медицинской точки зрения, целью данной работы является выявление наиболее значимых признаков, на основании значений которых можно было бы с минимальной вероятностью ошибки определить форму реактивного артрита у любого нового пациента.

Постановка задачи

Для однозначного определения информативных признаков, характеризующих тот или иной диагностический процесс на основе законов распределения и значений вероятностей присутствия, могут использоваться следующие методы многомерного статистического анализа:

1. регрессионный;
2. кластерный;
3. дискриминантный.

Регрессионный анализ позволяет выявлять признаки, которые оказывают наибольшее влияние на итоговый показатель, а также определять мощность и направление этого влияния. Однако синтезированные модели могут оказаться неточными или, в случае сложных моделей, переобученными, что снижает эффективность их использования.

Методы кластерного анализа дают возможность сравнивать объекты моделирования по их качественным характеристикам, агрегировать экспертные оценки текущего и прогнозируемого уровней развития объектов, выделять закономерности в структуре анализируемых данных [3]. Несмотря на многочисленные достоинства методов данного класса, их использование зачастую характеризуется значительными временными затратами на проведение достаточно сложных расчетов, в то время как адекватно интерпретировать полученные в результате кластеризации группы объектов не всегда возможно, особенно если информативные признаки, характеризующие объекты исследования, были изначально достаточно неоднородны.

Перечисленные выше недостатки учитывает дискриминантный анализ, предоставляющий возможность изучать различия между двумя и более группами объектов по нескольким переменным одновременно. Использование методов этого вида многомерного анализа имеет ряд преимуществ, например, позволяет учитывать вариабельность параметра, рассматривать совокупность различных клинико-лабораторных показателей, взятых со своими весовыми коэффициентами, отражающими удельный вес влияния каждого показателя на итоговое решение о принадлежности пациента к одной из групп.

Дискриминантный анализ – это общий термин, относящийся к нескольким тесно связанным статистическим процедурам. Эти процедуры можно разделить на [4]:

1. Методы интерпретации межгрупповых различий (дискриминации);
2. Методы классификации наблюдений по группам.

В данной дипломной работе будут рассмотрены процедуры первого типа, которые позволят ответить на следующие вопросы:

* Возможно ли, используя данный набор переменных, отличить одну группу от другой;
* Насколько хорошо эти переменные позволяют провести дискриминацию;
* Какие из них наиболее информативны, а какие могут быть удалены из пространства признаков в связи с избыточностью.

Процедуры второго типа будут рассмотрены только в контексте проверки построенной дискриминантной модели и оценке ее дискриминирующих способностей.

Обработка материала

В работе используются данные о пациентах, перенесших заболевание реактивным артритом, представленные Клинической ревматологической больницей №25 города Санкт-Петербурга.

104 пациента из опытной группы подверглись детальному исследованию и описанию при обращении в данное медицинское учреждение. В последствии обследованные пациенты были разделены на четыре группы в зависимости от возбудителя заболевания (Табл.1). В первую группу вошли 50 пациентов с урогенитальным возбудителем. Во вторую – 19 с не выявленным. К третьей группе были отнесены 19 человек, причиной развития болезни которых стали различные носоглоточные инфекции, такие как ОРВИ, хронический тонзиллит и другие. Четвертая группа включает 16 пациентов с кишечными и гастроэнтерологическим причинами.

Табл.1: Формы реактивного артрита

|  |  |  |
| --- | --- | --- |
| № | Название | Количество наблюдений |
| 1 | Урогенитальная | 50 |
| 2 | Неизвестная | 19 |
| 3 | Носоглоточная | 19 |
| 4 | Кишечная | 16 |

Клинико-лабораторные показатели больных реактивных артритом, на основании которых в дальнейшем будет проведен статистический анализ, для удобства можно разделить на следующие тематические группы [Приложение 1]:

* Признаки клинической активности, характеризующие клиническую интенсивность воспалительного процесса;
* Лабораторные признаки воспаления;
* Иммунологические показатели;
* Категориальные показатели;
* Показатели общего анализа крови.

**Глава 1**

Дискриминантный анализ

1.1 Принцип дискриминации

В общем случае задача дискриминации формулируется следующим образом. Пусть задано пространство признаков X размерностью p>1, точками которого являются наблюдения над объектом, представленные в виде p-мерного случайного вектора *xj = (x(1), x(2), …, x(p))∈ X*, где *x(1), x(2), …, x(p)* – признаки объекта, в дальнейшем именуемые дискриминантными переменными; *j=1,2,…,n*; *n* – общее число наблюдений. Исходная таблица наблюдений разбита на g непересекающихся классов, где каждой строке *xj* поставлен в соответствие некоторый класс категориальной природы *Gi*, *i=1,2,…,g***:**

Если априорно заданные классы действительно отличаются друг от друга по наблюдаемым переменным, их можно представить как скопления точек в некоторых областях рассматриваемого p-мерного пространства.

Требуется установить решающее правило, согласно которому по значениям координат произвольного вектора xj∈X рассматриваемый объект можно было бы отнести к одному из существующих классов Gi:

Таким правилом является задание классификационных функций, вычисление которых может осуществляться различными способами. Рассмотрим методы канонического и пошагового дискриминантного анализа.

1.2 Канонический дискриминантный анализ

**1.2.1** Дискриминантные функции

Каноническая дискриминантная функция является линейной комбинацией дискриминантных переменных и должна удовлетворять определенным условиям. Ее математическое представление имеет следующий вид:

*fkm* – значение канонической дискриминантной функции для *m*-го объекта в классе *k*, где *k=1,..,g*; *m=1,..,nk*;*nk* – количество объектов в *k*-ом классе;

(1)

Xikm – значение дискриминантной переменной *x(i)* для *m*-го объекта в группе *k*;

*Ui* – коэффициенты, обеспечивающие выполнение требуемых условий, где *i=1,…,p*.

Количество функций определяется количеством значимых корней канонической корреляции. Они задают новое пространство, в котором каждой оси соответствует одна дискриминантная функция и в котором определяется "центр тяжести" каждой группировки – центроид. Центроид класса – это точка, координаты которой есть усредненные значения переменных по всем объектам в данном классе. В связи с тем, что он занимает положение типичных наблюдений, центроиды могут быть использованы для изучения различий между классами. Рассмотрение объектов в исходном p-мерном признаковом пространстве, как правило, характеризуется трудностью интерпретации информации по всем переменным в совокупности. Однако У.Клекка в своей работе [4] продемонстрировал, что для различения положения центроидов достаточно ограничиться размерностью на единицу меньшей числа классов, что существенно упрощает представление итоговых результатов.

Соотношение (1) задает математическое преобразование p-мерного пространства дискриминантных переменных в q-мерное пространство канонических функций (где q=g-1 – максимальное число функций). Каждой оси соответствует свое соотношение вида (1). Для текущего наблюдения значение fkm интерпретируется как координата объекта в пространстве канонических дискриминантных функций [4].

Объект, определенный в этом пространстве как точка, относится к той группировке, к центроиду которой он расположен ближе. В связи с этим, при проверке модели могут возникнуть так называемы ошибки классификации: когда объект, который эксперт априорно отнес к одному классу, модель может отнести к другому. Доля правильно классированных объектов способна косвенно подтвердить правильность разделения групп. В качестве меры расстояния между объектами в p-мерном пространстве будет использовано выборочное расстояние Махаланобиса, преимущества которого по сравнению с другими метриками описаны в параграфе 1.3.2.

**1.2.2** Коэффициенты канонической дискриминантной функции

Для вычисления коэффициентов Ui канонической дискриминантной функции необходим некий статистический критерий, который позволил бы оценить степень различия между объектами и классами. Результаты дискриминации будут тем лучше, чем меньше разброс точек относительно центроида внутри класса и чем больше расстояние между самими центроидами классов.

Один из методов поиска наилучшей дискриминации состоит в нахождении такой канонической дискриминантной функции f, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой:

(2)

Где B - межгрупповая и W – внутригрупповая матрицы рассеяния наблюдаемых переменных от средних.

Введем следующие обозначения для пояснения происхождения названных матриц:

nk – число наблюдений в k-ом классе, k=1,..,g

n – общее число наблюдений по всем классам

Xikm – величина переменной i для m-го наблюдения в k-ом классе

ik – средняя величина переменной i в k-ом классе

i – среднее значение переменной i по всем классам (общее среднее)

Рассмотрим задачу максимизации отношения (2) при наличии g классов. Для того, чтобы оценить информацию, характеризующую степень различия между объектами по всему пространству точек, необходимо вычислить матрицу рассеяния T.

Элементы матрицы межгрупповых различий T задаются соотношением:

где , *i=1,..,p*

Данная матрица дает полное представление о распределении точек по пространству, определяемому переменными. Элементы, стоящие на диагонали, представляют собой сумму квадратов отклонений от общего среднего и показывают, как ведут себя наблюдения по отдельно взятой переменной. Внедиагональные элементы равны сумме произведений отклонений по одной переменной на отклонения по другой.

Для измерения разброса внутри классов используется матрица W, отличающаяся от T только тем, что ее элементы определяются векторами средних для отдельных классов, а не векторами средних для общих данных:

Также введем матрицу межгрупповых сумм квадратов отклонений и попарных произведений B, которая будет нулевой только в том случае, когда центроиды различных классов совпадают. Если положение групп в пространстве различается (то есть их центроиды не совпадают), то степень разброса наблюдений внутри классов будет меньше межгруппового разброса. Элементы данной матрицы задаются следующим образом:

(3)

Матрицы B и W содержат всю необходимую информацию о зависимости внутри и между классами, а их отношение дает меру различия между ними.

Для улучшения результатов разделения наблюдений на классы необходимо подобрать коэффициенты дискриминантной функции из условия максимизации отношения (2) при условии ортогональности дискриминантных плоскостей. Условие ортогональности означает, что вклады дискриминантных функций в различение групп не будут перекрываться. Тогда задача нахождение коэффициентов сведется к решению задачи о собственных значениях и векторах [4].

*Утверждение*. Если спроектировать g групп p-мерных выборок на q-мерное пространство, порожденное собственными векторами (V1q ,..., Vpq), каждому из которых соответствует свое собственное число λ, то отношение (2) будет максимальным. То есть рассеивание между группами будет максимальным при заданном внутригрупповом рассеивании.

Следовательно, требуется найти решения следующей системы уравнений относительно λi и *Vi* :

Где λi – собственное число; *Vi* – собственный вектор, последовательность p коэффициентов;

bji и wji – элементы матриц B и W соответственно.

Для получения единственного правильного решения налагают дополнительное условие:

Максимально возможное количество нетривиальных решений этих уравнений равно q.

Каждое решение, имеющее свое собственное значение λi и собственный вектор Vi, соответствует одной канонической дискриминантной функции. При этом функция с наибольшим собственным значением является самым мощным дискриминатором, и, соответственно, наибольшее рассеяние между классами будет достигаться при проектировании g выборок на ось, соответствующую этому максимальному собственному числу λi.

Далее необходимо определить коэффициенты функции вида (1). Дискриминантную функцию можно получить по нестандартизированным и стандартизированным коэффициентам. Компоненты собственного вектора *Vi*, полученного в результате решения системы, также можно использовать в качестве коэффициентов, но в таком случае начало координат не будет совпадать с «главным центроидом» – точкой, в которой каждая ось равна нулю. Для того, чтобы начало координат совпало с главным центроидом, необходимо провести нормирование компонент собственного вектора *Vi* по следующим формулам:

Нормированные коэффициенты (4) называются нестандартизированными, так как получены по нестандартизированным исходным данным.  Эти коэффициенты приводят к дискриминантным значениям, единицей измерения которых является стандартное квадратичное отклонение. При таком подходе каждая ось в преобразованном пространстве сжимается или растягивается таким образом, что соответствующее дискриминантное значение для данного объекта представляет собой число стандартных отклонений точки от главного центроида.

(4)

Применение стандартизированных коэффициентов позволяет привести значения дискриминантной функции к стандартной форме, т.е соответствующие дискриминантные значения по совокупности наблюдений (объектов) будут иметь нулевое среднее и единичное внутригрупповое стандартное отклонение. Преобразование нестандартизированных коэффициентов к стандартизированной форме осуществляется по формуле:

(5)

Где *wij* – сумма внутригрупповых квадратов i-ой переменной, определяемая по формуле (3).

Использование именно стандартизированных коэффициентов удобно для решения поставленной нами задачи о снижения размерности исходного признакового пространства переменных путем удаления малоинформативных признаков. Эти коэффициенты основаны на стандартизированных переменных и, соответственно, принадлежат к одной шкале абсолютной измерений, поэтому их можно сравнивать, чтобы определить величины и направления вкладов переменных в каждую каноническую функцию. Абсолютная величина коэффициента в данном случае анализируется в стандартной форме: чем она больше, тем больше вклад этой переменной в дискриминацию. Таким образом, стандартизированные коэффициенты позволяют оценить содержательность каждой переменной в наборе: если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, ее можно исключить [5].

Как уже было сказано, степень разделения выборочных групп зависит от величины собственных чисел. Наибольшей разделительной способностью обладает первая дискриминантная функция, соответствующая наибольшему собственному числу λ1, вторая обеспечивает максимальное разделение после первой и т.д. Оценить различительную способность i-ой функции в сравнение с остальными можно по ее проценту объясненной дисперсии, который рассчитывается с помощью собственных чисел дискриминантных функций по формуле:

(6)

**1.2.3** Ограничения

Прежде чем использовать линейный дискриминантный анализ, требуется проверить на фиксированном уровне значимости следующие гипотезы:

1. Исходная выборка должна быть извлечена из нормальной генеральной совокупности;
2. Ковариационные матрицы групп должны быть равны между собой.

Но в действительности линейный дискриминантный анализ способен давать результаты и в случае нарушения рассматриваемых гипотез. Лахенбрук в своей работе [6] говорит о том, что линейный дискриминантный анализ показывает хорошие результаты, имея дело с дискретными признаками. Также использовать рассматриваемое решающее правило можно и в случаях, когда гипотеза о равенстве ковариационных матриц или гипотеза о нормальности распределения генеральной совокупности оказывается не принятой. Однако это может привести к большой ошибке при классификации.

1.3 Пошаговый дискриминантный анализ

**1.3.1** Отбор информативных признаков

Как правило, не все признаки, имеющиеся в распоряжении исследователей, вносят одинаковый вклад в разделение объектов на классы. Одним из способов исключения малозначимых переменных и снижения размерности исходного пространства признаков является алгоритм пошагового дискриминантного анализа Forward Stepwise (процедура прямого отбора), основанный на идее поэтапного включения наиболее информативных переменных в модель. Общий принцип отбора состоит в следующем: на каждом шаге с помощью какого-либо статистического критерия выбирается переменная, которая вносит наибольший вклад в различие между совокупностями в сочетании с отобранными ранее переменными. Процесс продолжается до тех пор, пока оставшиеся переменные не перестанут улучшать дискриминацию.

Таким образом порождается оптимальное множество дискриминантных переменных, которое, тем не менее, может не быть максимальной (наилучшей) комбинацией. Для получения максимального решения требуется проверить всевозможные сочетания переменных (пар, троек и т.д.). Однако на практике такая проверка может оказаться дорогой и временно затратной, в связи с чем комбинацию, полученную в результате работы процедуры последовательного отбора Forward Stepwise, будем считать лучшей.

**1.3.2** Критерии отбора

В качестве критерия отбора наиболее информативных переменных должны использоваться некоторые статистки, позволяющие оценить вклад конкретной переменной в дискриминацию. Рассмотрим те из них, на основании которых в дальнейшем будет проведен анализ. Также следует отметить, что зачастую для принятия решения о включении того или иного признака в модель, лучше руководствоваться результатами не одного критерия, а рассмотреть их в совокупности [7].

1. *Коэффициент канонической корреляции Ri* – мера связи, позволяющая оценить степень зависимости между группами и дискриминантной функцией. Максимальная величина этого коэффициента равна 1. Каноническая корреляция связана с собственным значением i-ой ДФ следующий образом:

Где λi – значение, вычисляемое по формуле (2). Чем больше величина *Ri*, тем лучше разделительная способность функции.

2) *Статистика лямбда Уилкса (Wilk’s Lambda)* – отношение меры внутриклассовой изменчивости к общей изменчивости. Принимает значение в интервале [0;1], где «0» означает, что классы полностью однородны, а «1» – что разделение объектов на классы не верно. То есть, чем меньше значение данной статистики, тем лучше разделение на группы.

3) *Λ-статистика Уилкса (partial Lambda)* – мера различия между классами по нескольким переменным. Позволяет оценить одиночный вклад соответствующей переменной в дискриминацию между совокупностями. Эта статистика показывает отношения лямбды Уилкса после добавления данной переменной к лямбде Уилкса до добавления переменной. Позволяет оценить остаточную дискриминантную способность – способность различать группы с помощью переменных, не включенных в модель. Если она мала, то выполненный анализ достиг своей цели.

Где k – число уже вычисленных функций.

1. *Статистика F-включения* – это частная F-статистика, оценивающая улучшение качества дискриминации от использования рассматриваемой переменной по сравнению с различением, достигнутым с помощью отобранных ранее переменных. Является основным критерием отбора. Перед проведением анализа задается критическое значение F-включения: если на некотором шаге значения данной статистики для всех переменных вне модели будут меньше ее критической величины, процедура отбора должна завершиться. Чтобы убедиться в статистической значимости улучшения качества дискриминации после включения проверяемой переменной в модель, значение данной статистики рекомендуется рассматривать вместе со значением толерантности.
2. *Статистика F-удаления* – также является частной F-статистикой, оценивает значимость ухудшения качества дискриминации после удаления переменной из списка уже отобранных переменных. Данная процедура проводится в начале каждого шага, чтобы проверить, имеется ли какая-либо переменная, уже не вносящая достаточно большого вклада в дискриминацию.

На заключительном шаге статистика F-удаления используется в качестве критерия для ранжирования дискриминантных способностей отобранных переменных. Чем больше ее значение, тем больше вклад этой переменной в различение.

1. *Толерантность* – мера избыточности признака, оценивает степень нескоррелированности данной переменной со всеми остальными. Толерантность еще не отобранной переменной вычисляется по следующей формуле:

где *R–* коэффициент множественной корреляции текущего признака со всеми остальными, использованными в анализе.

Чем меньше ее значение, тем сильнее текущий признак связан с остальными переменными, и, соответственно, тем избыточнее, так как несет в себе малую дополнительную информацию. Если толерантность признака близка к нулю, то его лучше не включать в модель.

1. *Выборочное расстояния Махаланобиса* – мера, позволяющая рассчитать расстояние от каждого объекта до центроида любого из классов, чтобы затем отнести объект в ближайший класс. В отличие от евклидовой метрики, может применяться даже в тех случаях, когда переменные коррелированы, измеряются в разных единицах и имеют различные стандартные отклонения. Будет использована для проверки классифицирующих свойств построенной модели в новом оптимальном пространстве признаков.

Квадрат расстояния от точки X (текущий объект) до центроида класса k вычисляется по формуле:

1. *τ-статистика ошибок* – стандартизированная мера эффективности для любого количества классов. Позволяет оценить адекватность полученной модели апостериорно, вычисляя процентное содержание правильно классифицированных объектов.

Где nc – число правильно классифицированных объектов,

pi – априорная вероятность принадлежности классу (вероятности того, что наблюдение принадлежит соответствующему классу без использования какой-либо информации о значениях переменных в модели):

**1.3.3** Алгоритм Forward Stepwise

Пусть произвольное наблюдение X описывается p-мерным вектором признаков: 𝑥 = (*𝑥(1), 𝑥(2), ..., 𝑥(𝑝)*). Алгоритм Forward Stepwise выглядит следующим образом:

1. Вначале из модели удаляются все рассматриваемые признаки.

2. Задается критическое значение статистики Fкр-включения.

3. Для каждого *𝑥(i)* высчитывается Fвкл(𝑥(i)), *i=1,..,*𝑝 и затем в модель отбирается та переменная, значение соответствующей статистики которой наибольшее среди рассматриваемых. Включенный в модель признак обозначается за 𝑡1.

4. Среди остальных 𝑝 − 1 переменных ищется признак с наибольшим значением статистики F-включения:

Вместе с тем, статистика должна удовлетворять условию: Fвкл > Fкр-включения. Если условие не выполняется ни для одной статистики из рассматриваемых, то пошаговый процесс останавливается.

5. Процесс будет продолжается аналогичным образом до тех пор, пока ни у одной из переменных соответствующая статистика не будет удовлетворять условию включения или все переменные не войдут в модель.

В результате работы данного алгоритма будет получен набор переменных 𝑡1,...,𝑡𝑟 (r – число шагов в алгоритме), который будет являться оптимальным множеством – новым признаковым пространством. Данный набор может дать лучший результат дискриминации, чем изначальный набор признаков.

**Глава 2**

Результаты исследования

Расчеты производились при помощи статистического пакета SPSS.

Пошаговый дискриминантный анализ

Уровень значимости, или порог, при котором результаты объявляются статистически значимыми, p-level установлен равным 0.05.

В результате процедуры прямого отбора наиболее информативных признаков при критическом значении статистики Fкр-включения = 2.5 и критическом значении толерантности T = 0.01, в оптимальное множество были включены 9 дискриминантных переменных из 58 при статистически незначимой остаточной дискриминации (Рис.1). На первый взгляд, дискриминация между группами высоко значима (общая статистика лямбда Уилкса=0.06; F=15.872; p-level<0.00001). Проанализируем независимые вклады каждой переменной в модели в предсказание.

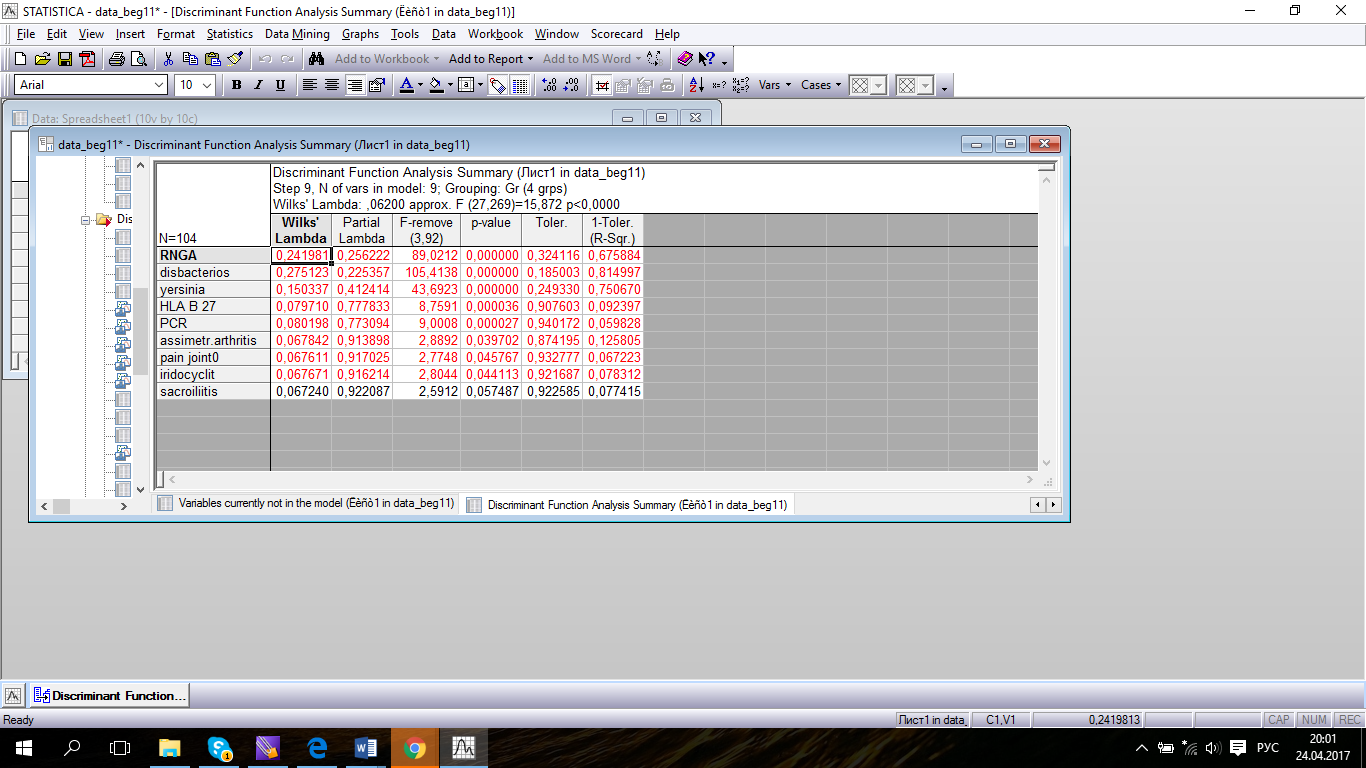


Рис.1: Дискриминантная модель с Fкр-вкл=2.5

Статистика F-удаления и частная статистика лямбда Уилкса показывают, что наибольший вклад в общую дискриминацию вносит переменная disbacterios, следующая по значимости – RNGA. Еще три признака – yersenia, PCR, HLA B 27 - также информативны, это подтверждают несколько статистических критериев в совокупности: высокий уровень статистической значимости p-value, относительно высокие показатели F-критерия удаления, на основании которого происходит оценка дискриминантных способностей переменной, и «хорошие» значения частной лямбды Уилкса.

Последние четыре признака в построенной модели вносят примерно одинаковый вклад в различение классов и, следовательно, дублируют друг друга, то есть три из них являются избыточными. Проанализировав значения статистик для соответствующих переменных, было принято решение оставить в модели показатель «асимметричного артрита» и исключить остальные переменные, после чего построить новую дискриминантную модель без учета удаленных переменных. Для этого необходимо повысить критическое значение статистики F-включения. Пусть Fкр-включения = 3.

Результаты работы программы с измененным порогом включения в модель представлены на Рис.2. Основываясь на общей статистике Уилкса = 0.08019, значению которой соответствует аппроксимация критерия Фишера F(18.269) = 21.539 и вероятностью ошибки p ≈ 0.0, гипотеза о равенстве групповых средних для различных классов пациентов, разделенных по типу возбудителя реактивного артрита, должна быть отвергнута, т.е исходное деление на классы статистически значимо.

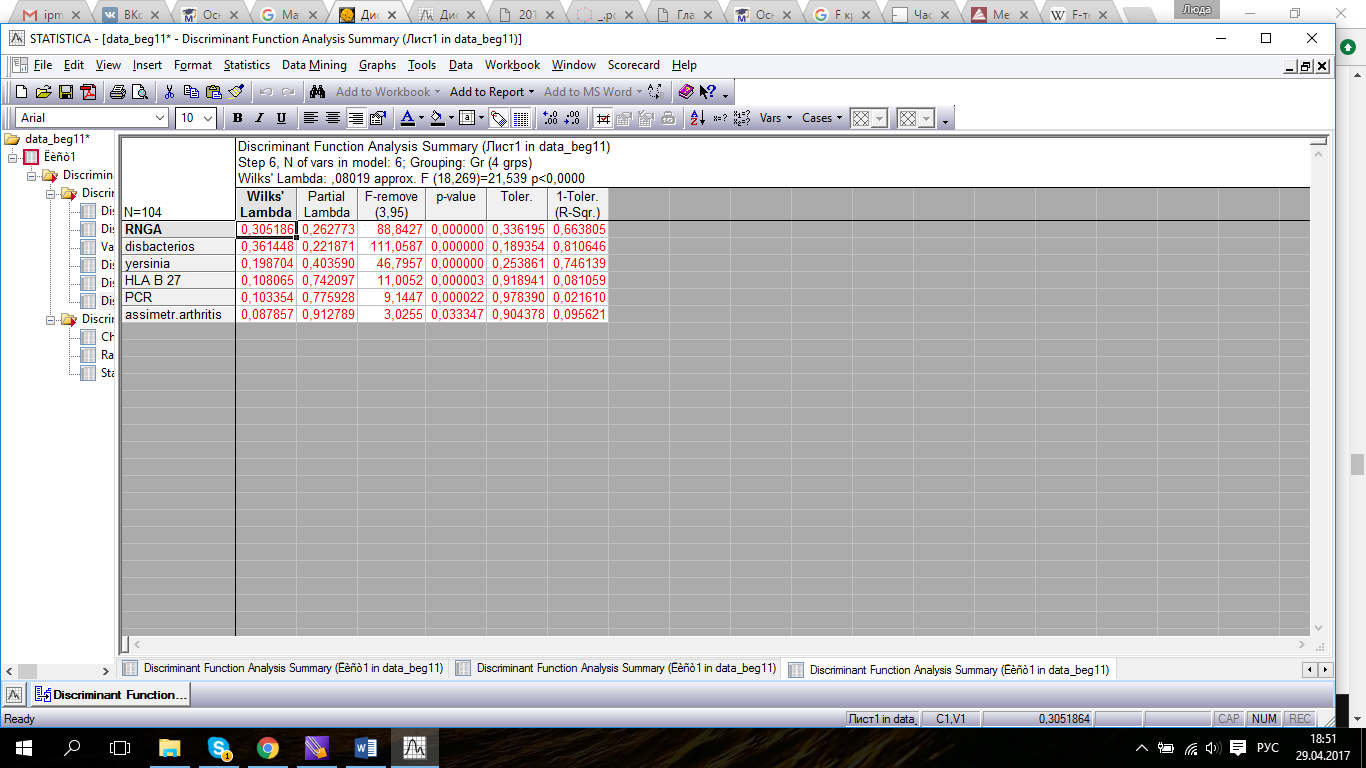


Рис.2: Дискриминантная модель с Fкр-вкл=3

Критерий F-удаления позволяет проранжировать вошедшие в модель переменные по доле их вклада в различение классов. На этой стадии исследования можно заключить, что показатели RNGA (реакция непрямой гемагглютинации) и дисбактериоза являются главными переменными, которые позволяют различать группы пациентов. Для получения дальнейших результатов о природе дискриминации следует провести канонический анализ.

Канонический анализ

Для того, чтобы увидеть, как шесть отобранных переменных разделяют совокупность пациентов на группы, необходимо вычислить дискриминантные функции. В соответствии с формулой (1), программа вычислит три различные независимые (ортогональные) функции.

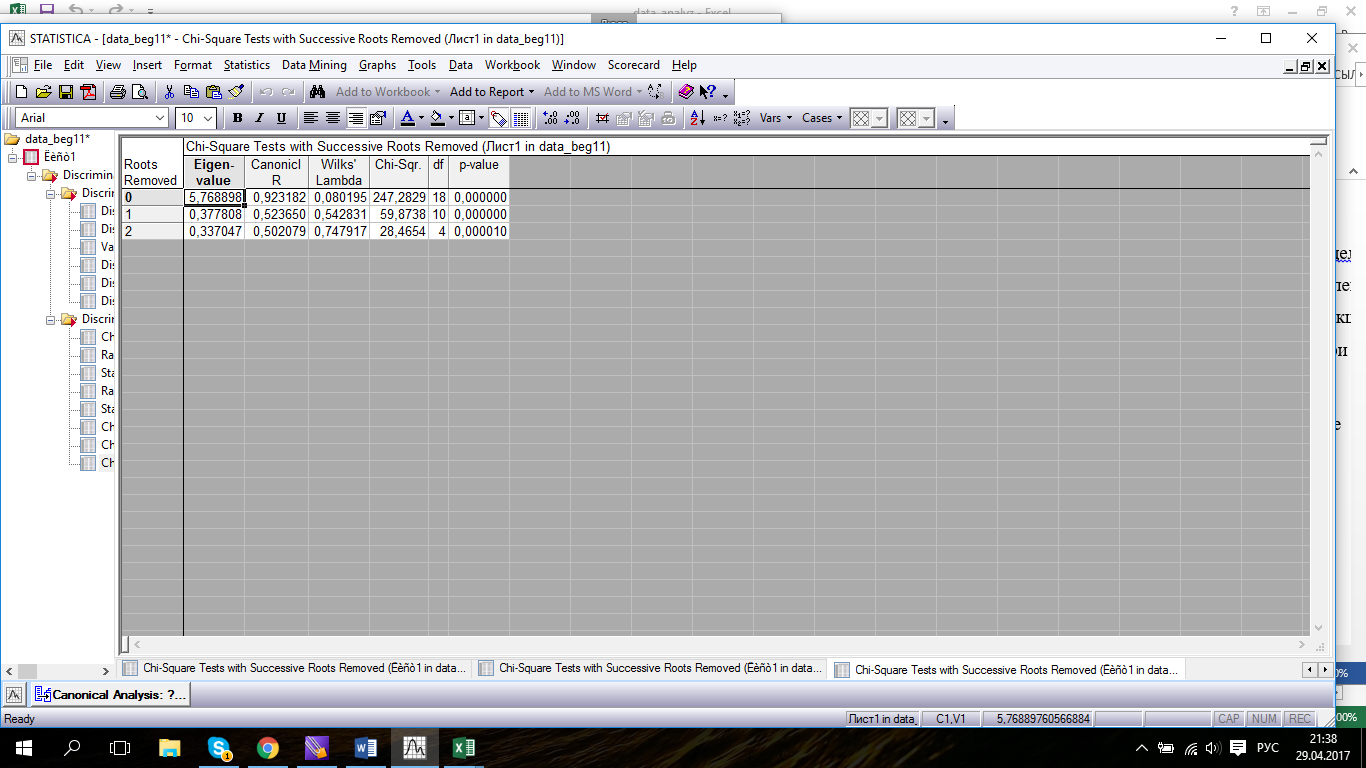
Прежде чем начать интерпретацию результатов построения, необходимо определить, являются ли все три дискриминантные функции (корни) статистически значимыми.

Рис.3:Результаты пошагового отбора для канонических корней

Данные таблицы (Рис.3) указывают на хорошую дискриминацию групп: наибольшей разделительной способностью обладает первая дискриминантная функция, ее кумулятивная доля объясненной дисперсии (дискриминирующей мощности), в соответствие с формулой (6), составляет 88.9%. Большая величина ее коэффициента канонической корреляции R соответствует высокому показателю связи между этой функцией и делением на группы; малое значение Λ-статистики Уилкса означает, что все отобранные переменные эффективно участвуют в различении групп; статистика Хи-квадрат значима с уровнем p-value < 0.0000001.

Вторая и третья дискриминантные функции также являются статистически значимыми, но обладают меньшей мощностью различения классов, равной примерно 5.8%.

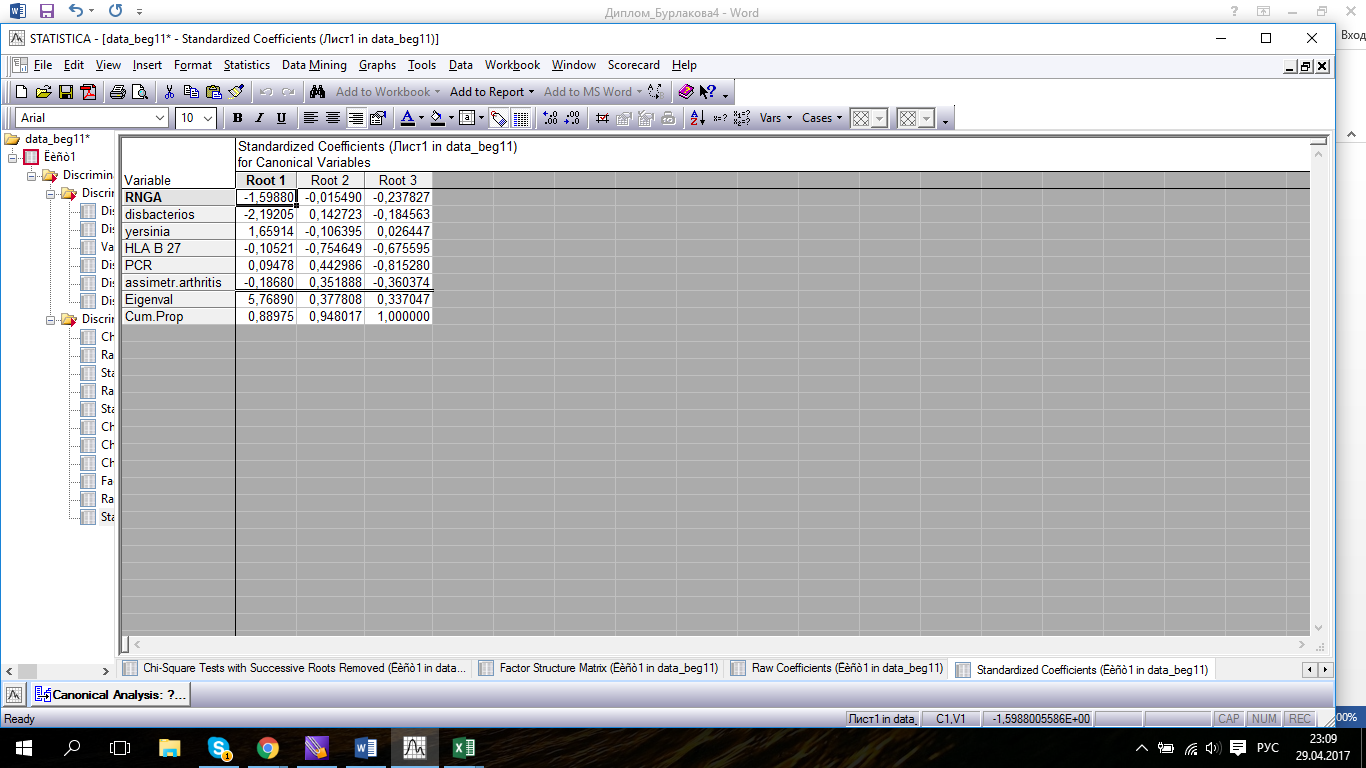
 Для оценки относительного вклада каждой переменной различение групп (в значение дискриминантной функции) были вычислены стандартизованные дискриминантные коэффициенты.

Рис.4: Стандартизированные коэффициенты для дискриминантных функций

Как видно (рис.4), наиболее значимыми переменными по абсолютной величине являются:

* показатели RNGA, дизбактериоза и иерсинии – для первой дискриминантной функции;
* HLA B-27 и (в значительно меньшей мере) PCR – для второй;
* PCR и HLA B-27 – для третьей.

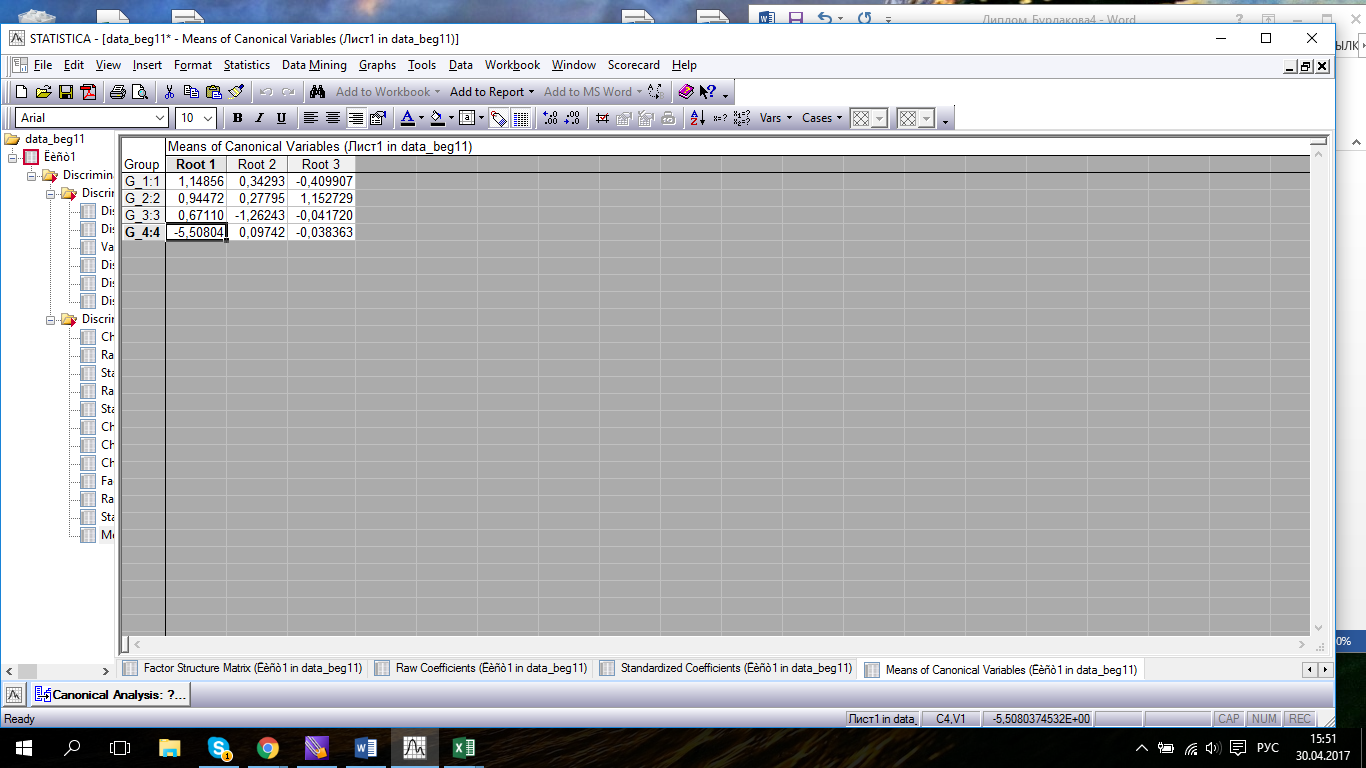
Зная переменные, которые наиболее эффективно различают формы реактивного артрита, можно определить природу дискриминации для каждого канонического корня. Рассмотрим таблицу средних значений канонических переменных, которые позволяют определить группы, лучше всего идентифицируемые конкретной дискриминантной функцией.

Рис.5: Таблица средних канонических переменных

Значения в таблице (рис.5) говорят о том, что первая ДФ наилучшим образом отделяет 4-ую группу реактивного артрита от всех остальных. Каноническое среднее «кишечной» формы значительно отличается от канонических средних других групп. Вторая ДФ выделяет главным образом третью – «носоглоточную» - группу, однако качество разделения по этой оси, как и следовало ожидать, стало меньше. Третья ДФ обладает такой же дискриминантной мощностью, что и вторая, и отделяет вторую группу, в которую вошли пациенты с неизвестным возбудителем заболевания реактивным артритом.

Визуализация

Для наглядного представления результатов разделения совокупности пациентов на классы следует построить диаграмму рассеяния канонических значений в пространстве канонических дискриминантных функций.

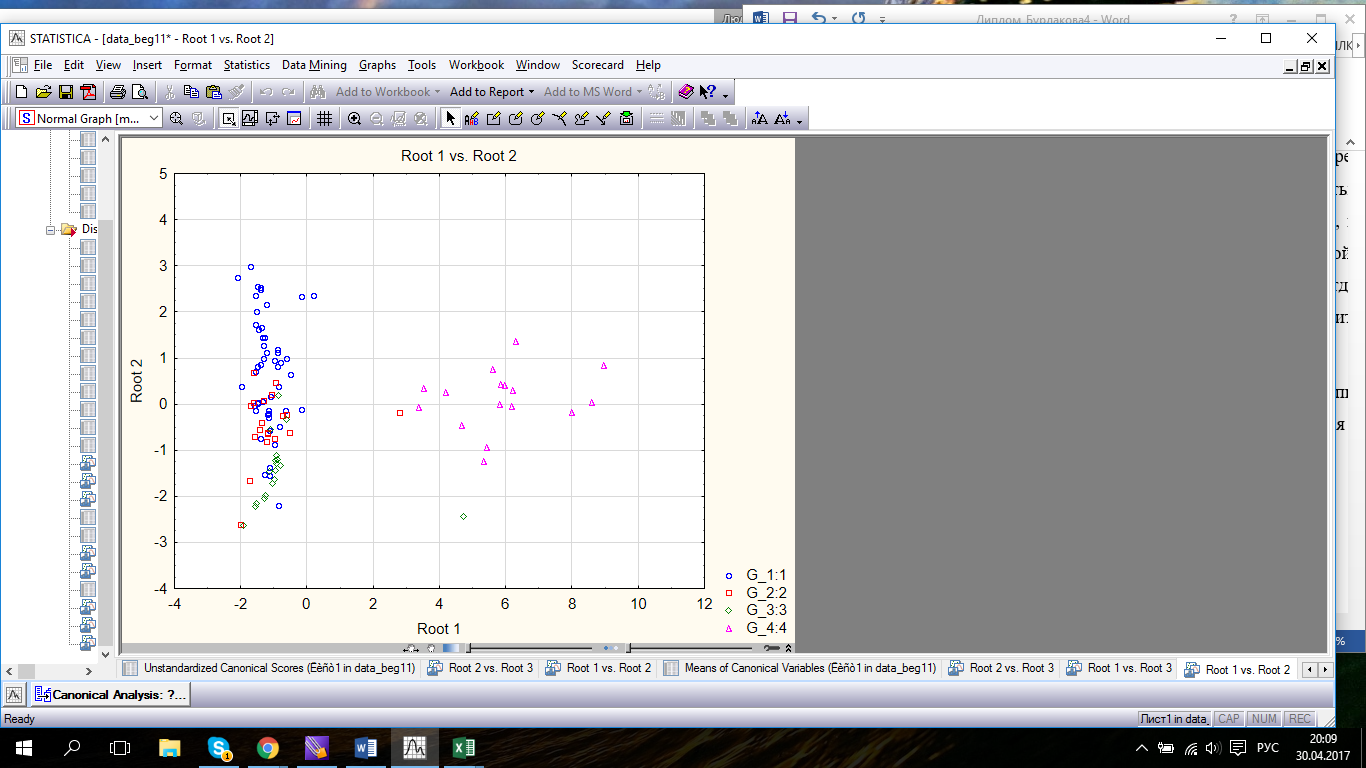


Рис.6 Диаграмма рассеяния 4-х групп пациентов, больных РА, в каноническом пространстве 1-ой и 2-ой ДФ

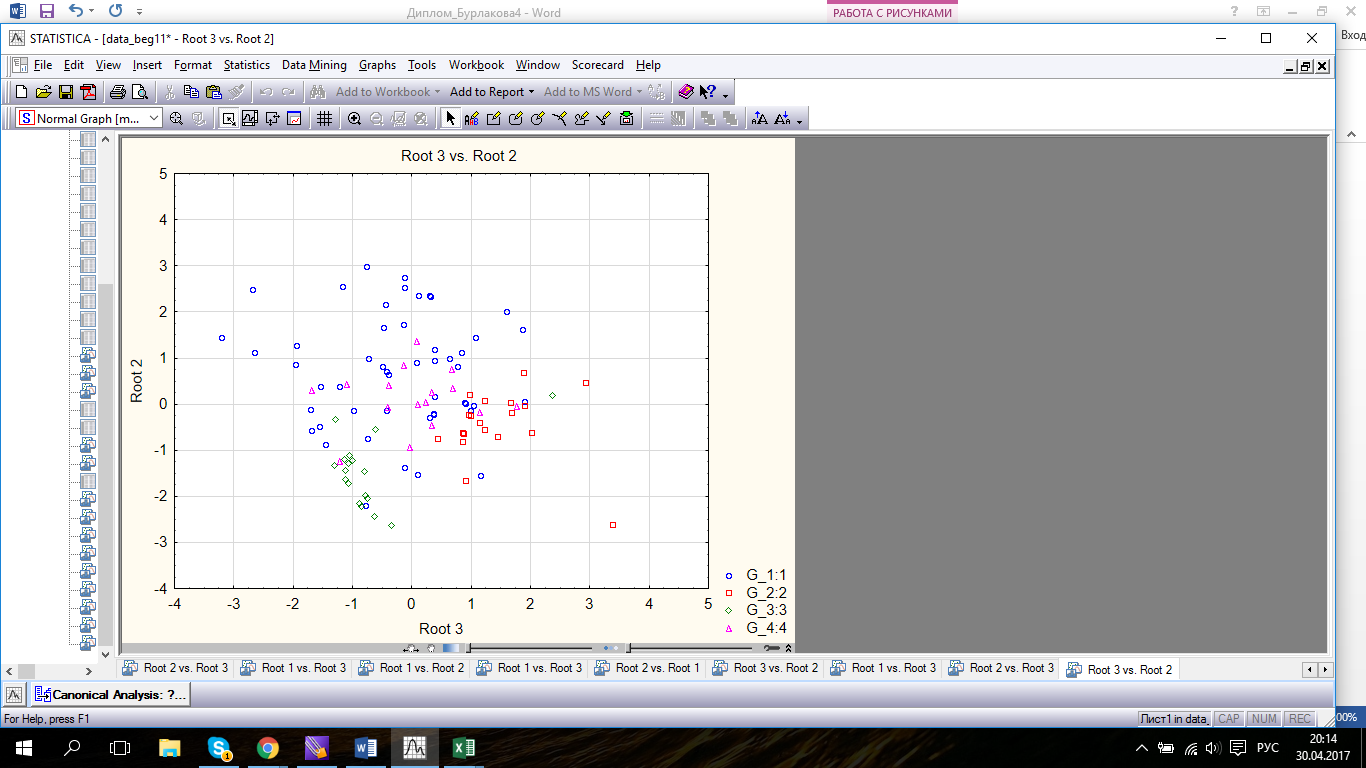
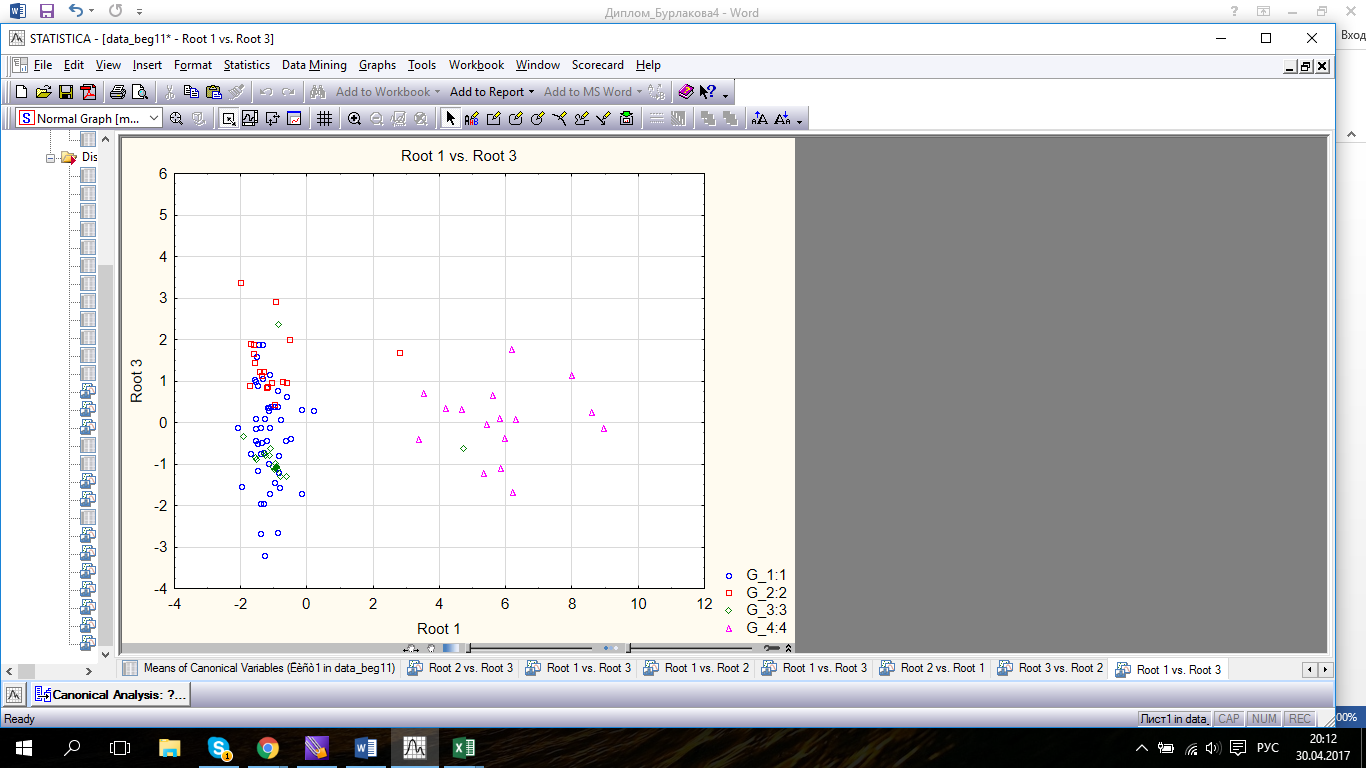
Координатная сеть образуется двумя каноническими корнями, где каждый корень дает набор множителей для переменных, включенных в модель, - канонические коэффициенты. Чтобы получить координату точки по данному корню, необходимо просуммировать произведения значений этих переменных на соответствующие коэффициенты [8].

Диаграмма рассеяния для канонических значений в координатном пространстве первой\второй дискриминантных функций (рис.6) подтверждает интерпретацию полученных выше результатов. Видно, что объекты из 4-ой (кишечной) группы локализованы справа, разрыв между этой группой и остальными определяется исключительно первым корнем (ось Root 1). При этом некая сгруппированность канонических значений наблюдений, относящихся ко 2-ой и 3-ей группам (неопределенной и носоглоточной), вокруг соответствующих центроидов также присутствует, но их значения перекрываются. Наблюдения из 1-ой (урогенитальной) группы нельзя визуально выделить из множества, образованного значениями канонических переменных 1-ой, 2-ой и 3-ей форм реактивного артрита. Таким образом, пациенты оказываются фактически разорваны на 2 группировки:

1. С «кишечной» причиной заболевания;
2. С урогенитальной, носоглоточной и неясной причинами.

Рис.6б: Диаграмма рассеяния в пространстве 2-ой и 3-ей ДФ

Рис.6а: Диаграмма рассеяния в пространстве 1-ой и 3-ей ДФ



На рис.6а и рис.6б представлены диаграммы рассеяния канонических значений в пространствах первой\третьей и второй\третьей дискриминантных функций. Так как дискриминирующая мощность второй и третьей функций одинакова, их диаграммы рассеяния практически совпадают (рис.6 и рис.6а). Разделения объектов в пространстве второй\третьей дискриминантных функций (рис.6б), как и следовало ожидать, не происходит вовсе, так как доля объясненной дисперсии, приходящаяся на эти функции, в сумме равна 11.6%.

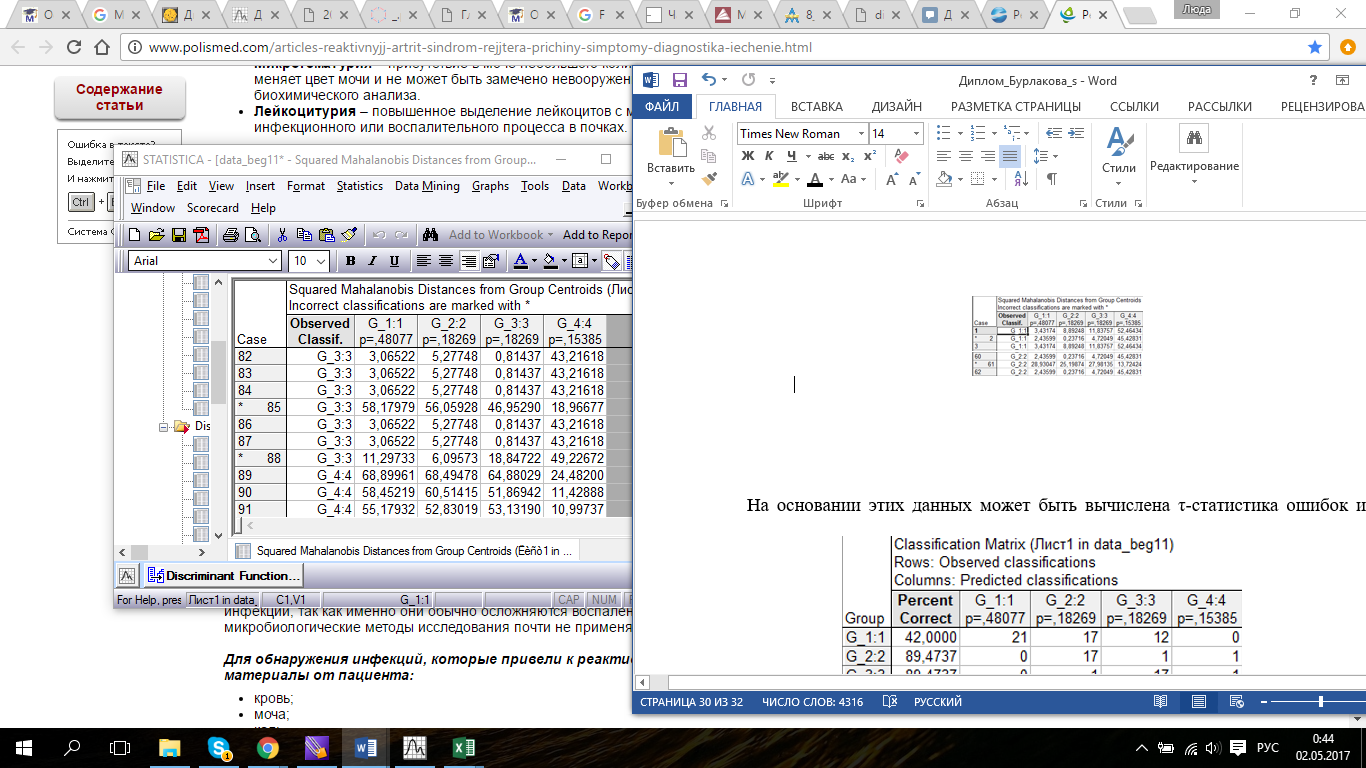
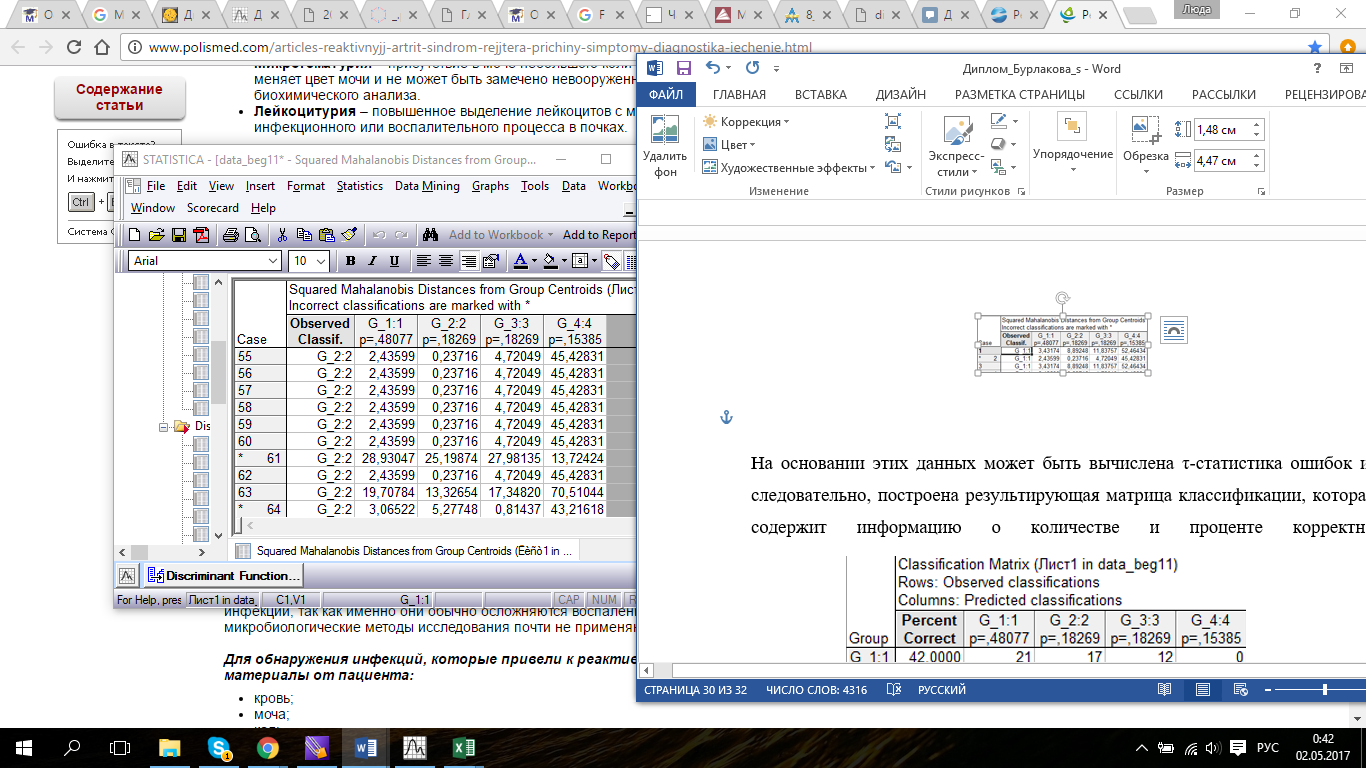
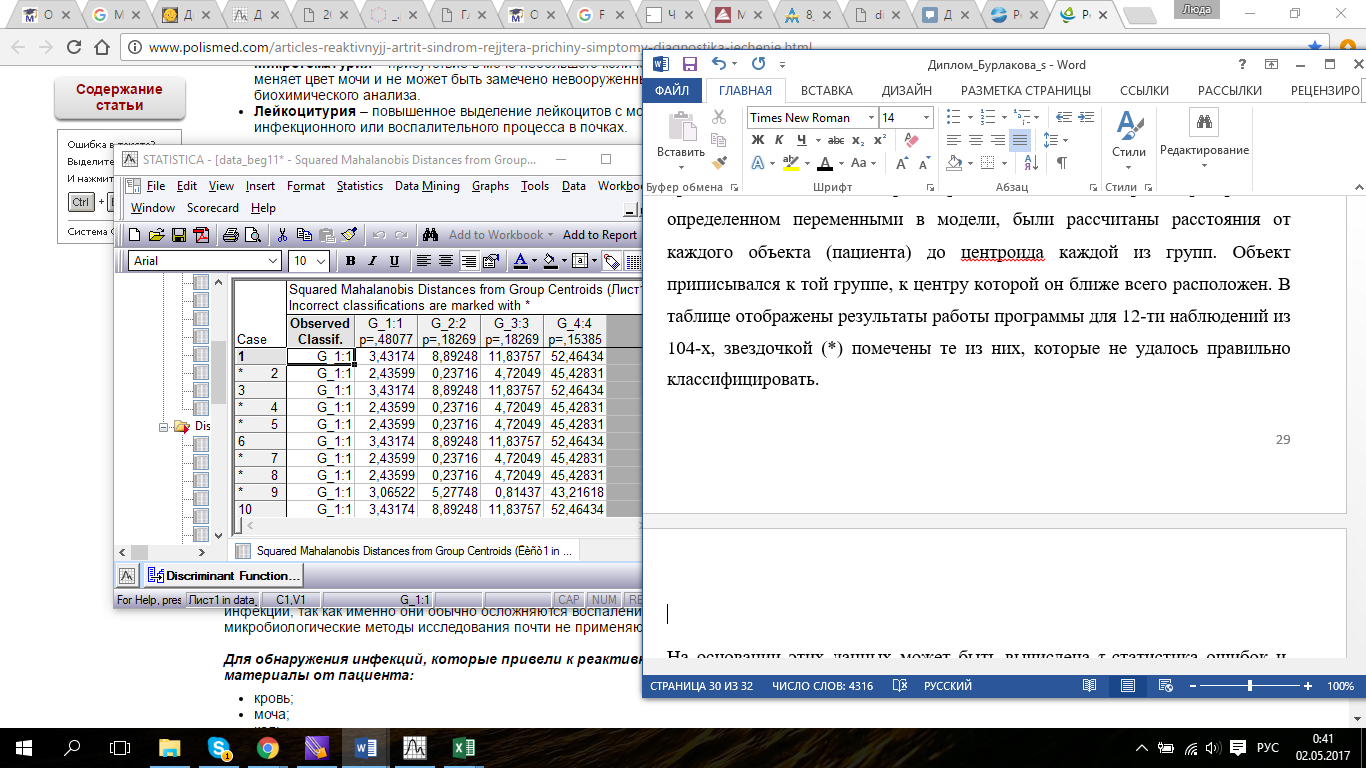
Из чего можно сделать промежуточный вывод о том, что наиболее значимая и ясная дискриминация возможна только для «кишечной» формы реактивного артрита с использованием первой дискриминантной функции.

Классификация наблюдений

Процедура классификации фактически является обратной процедуре дискриминации: по значениям переменных, характеризующих «неизвестное» наблюдение, необходимо определить, к какой группе оно относится. В целях данной дипломной работы не была поставлена задача построения такого решающего правила. Тем не менее, средства классификации также можно использовать для проверки степени адекватности построенной дискриминантной модели, применив определенные правила к «известным» объектам, которые были учтены при построении канонических функций. Доля правильно классифицированных объектов способна косвенно подтвердить степень разделения групп.

С помощью выборочного расстояния Махаланобиса D2, предназначенного для измерения расстояний в многомерном пространстве, определенном переменными в модели, были рассчитаны расстояния от каждого объекта до центроида каждой из групп. Объект приписывался к той группе, к центру которой он ближе всего расположен. В таблице (рис.7) отображены результаты работы программы для 12-ти наблюдений из 104-х, звездочкой (\*) помечены те из них, которые не удалось правильно классифицировать.

Рис.7: Фрагмент таблицы квадратов расстояний Махаланобиса



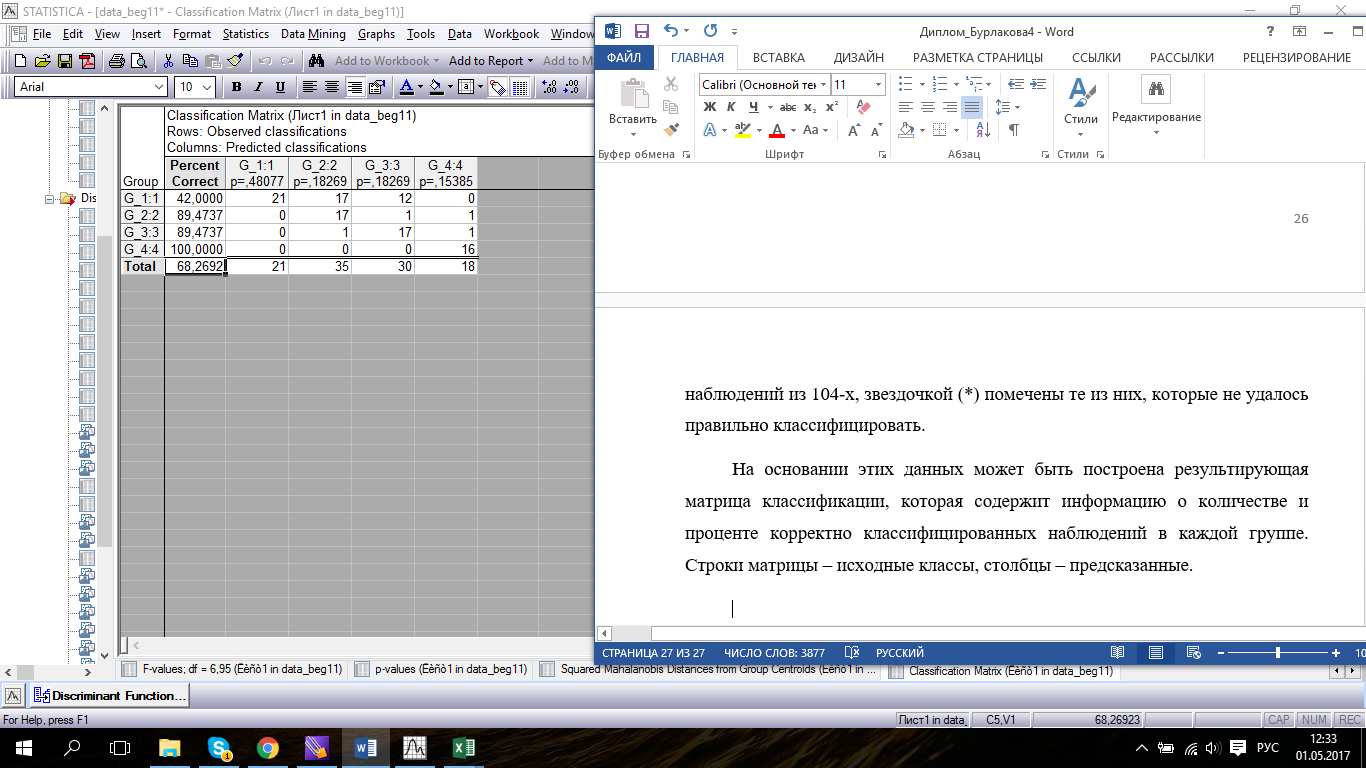
На основании этих данных может быть вычислена τ-статистика ошибок и, следовательно, построена результирующая матрица классификации (рис.8), которая содержит информацию о количестве и проценте корректно классифицированных наблюдений в каждой группе. Строки матрицы – исходные классы, столбцы – предсказанные, p – априорная вероятность принадлежности к группе, пропорциональная количеству объектов в той или иной совокупности.

Рис.8: Матрица классификации

По результатам расчетов, общая точность предсказаний, оцененная с помощью апостериорных вероятностей, получилось равной 68%. Данный процент нельзя считать высоким, но если рассмотреть результаты классификации отдельно по каждой из групп, становится очевидной причина такой не высокой общей точности предсказаний.

Наилучшим образом модель научилась различать 4-ую форму реактивного артрита – точность предсказания для нее равна 100%. Доля корректных предсказаний для 2-ой и 3-ей групп тоже достаточно высока – 89.5%. Наихудший результат классификации у 1-ой наблюдений из группы – всего 42% точности, это связано с морфологической неоднородностью класса и большой площадью пересечения его с другими классами, что и было показано на диаграмме рассеяния (рис.6).

Выводы

В результате применения методов многомерного дискриминантного анализа были выявлены признаки, оказывающие наиболее значимое влияние на «постановку диагноза» – отнесение пациентов к одному из четырех типов реактивного артрита. В оптимальное множество таких переменных вошли следующие показатели:

1. Выявление триггерного инфекта посредством реакции непрямой гемагглютинации (RNGA);
2. Дисбактериоз;
3. Иерсиния;
4. Антиген гистосовместимости HLA-B27;
5. ПЦР (Полимеразная цепная реакция, выявляющая наличие инфекции в организме);
6. Асимметричное поражение суставов.

Таким образом, удалось также косвенно подтвердить предположение о значимости наличия антигена гистосовместимости HLA-B27 в процессе диагностирования заболевания, о котором было сказано во введении.

В ходе канонического дискриминантного анализа в пространстве отобранных переменных был выявлен значимый разрыв между кишечной группой реактивного артрита и тремя другими группами в совокупности. Такого же наглядного разделения между урогенитальной, носоглоточной и группой с невыясненной причиной заболевания не произошло. С медицинской точки зрения это может означать, что изначальное деление на классы не было полностью верным и следует продолжить исследования в данной области.

В результате проверки дискриминирующих способностей переменных на основании стандартизированных коэффициентов первой канонической функции можно заключить, что наибольшее влияние на причисление пациента к «кишечной» форме реактивного артрита оказывают наличие в его организме бактерий дисбактериоза и иерсинии (кишечные инфекции) и показатель RNGA, что не противоречит биологическому смыслу. Сделать аналогичные выводы для остальных типов заболевания не представляется возможным ввиду недостаточной дискриминирующей мощности двух других канонических функций.

Заключение

В данной выпускной квалификационной работе был рассмотрен один из видов многомерного статистического анализа – дискриминантный анализ, с помощью которого производилось исследование медицинских данных на примере больных реактивным артритом, возникающим после некоторых инфекций. Реактивный артрит характеризуется классическими отличиями от других воспалительных заболеваний суставов. При этом, важным вопросом остается поиск возможных клинико-лабораторных различий в группах пациентов в зависимости от инфекции после которой возникает заболевание. В результате применения методов канонического дискриминантного анализа удалось установить отличие группы реактивного артрита, развившегося после достоверной кишечной инфекции, от других форм. Значимого различения в остальных группах, сформированных по этиологическому фактору (после урогенитальной, носоглоточной инфекции или без достоверного этиологического фактора), выявлено не было. С помощью пошагового дискриминантного анализа были отобраны наиболее значимые, с точки зрения различения типов заболевания, характеристики. Результаты исследования могут быть использованы при дальнейшем изучении патогенеза реактивного артрита.

Список литературы

[1] Реактивные артриты. Методические рекомендации / Ставрополь. Изд. СГМА, 2003. 2-10 с

[2] Причины реактивного артрита. https://health.mail.ru/disease/reaktivnyi\_artrit/

[3] Mark S. Aldenderef, Roger K. Blashfield. Cluster Analysis (Second Printing, 1985) p.143-146

[4] William R.Klecka. Discriminant Analysis (Seventh Printing, 1986) p.89-109

[5] Медик В.А., Токмачев М.С. Статистика здоровья населения и здравоохранения: учеб.пособие / М.: Финансы и статистика, 2009. 104 с

[6] Lachenbruch P.A. Some unsolved practical problems in discriminant analysis. (Chapel Hill: University of North Carolina, 1975) p.10

[7] Буреева Н.Н. Многомерный статистический анализ с использованием ППП “STATISTICA” // Учебно-методический материал по программе повышения квалификации «Применение программных средств в научных исследованиях и преподавании математики и механики», 2007. 74-88 с

[8] Ланг Т.А. Как описывать статистику в медицине / М.: Практическая медицина, 2011. 333 с

Приложения

*Приложение №1*

*Клинико-лабораторные характеристики больных реактивным артритом*

Примечание: зеленым помечены признаки, отобранные в результате пошагового дискриминантного анализа.

|  |  |
| --- | --- |
| Признаки клинической активности, характеризующие интенсивность воспалительного процесса | synovit, iridocyclit, entesit, axillobursit, assimetr.arthritis, sacrolitis, character, HLA B-27, leuc, paint joint, swollen, VASp, VAS, VASd, HAQ |
| Лабораторные признаки воспаления | ESR, CRP, Fibrin |
| Иммунологические показатели | CIC, IgG, IgA, IgM, lymf, p-lymf, cd3, p-cd3, cd4, p-cd4, cd8, p-cd8, cd4/cd8, cd19, p-cd19, IL 6, IL2, IF gamma, TNF, IL4, IL1, IL1Ra1 |
| Категориальные показатели | chlamidia, yreaplazma, mycoplasma, yersinia, rez-god, disbacterios |
| Показатели общего анализа крови | Hb, Er, CP, Plate, Leuc, pal+segm, eos, lymf, mon, ALT, RNGA, PCR |