УДК 614.2, 616.65-002

АНАЛИЗ ГЛАВНЫХ КОМПОНЕНТ И КЛАССИФИКАЦИЯ БОЛЬНЫХ С ДОБРОКАЧЕСТВЕННОЙ ГИПЕРПЛАЗИЕЙ ПРЕДСТАТЕЛЬНОЙ ЖЕЛЕЗЫ ПРИ НАЛИЧИИ СОПУТСТВУЮЩИХ ЗАБОЛЕВАНИЙ

Клименко Г.Я., С.Ю. Пехов, К.А. Разинкин

Воронежская государственная медицинская академия им. Н.Н.Бурденко
Воронежский государственный технический университет

Рассмотрены возможности анализа главных компонент и классификации больных с доброкачественной гиперплазией предстательной железы при наличии или отсутствии сердечно-сосудистых заболеваний на основе создания векторного пространства латентных факторов с меньшей размерностью

В конце XX – начале XXI в. наблюдается заметный рост урологической заболеваемости, особенностью которой является то, что она возрастает в группе лиц пожилого возраста. Часто диагностируются запущенные, принявшие хронические формы заболевания, которые протекают на фоне тяжёлой соматической патологии, что приводит к высокой летальности [1].

Доброкачественная гиперплазия предстательной железы (ДГПЖ) является самым распространённым из урологических заболеваний мужчин пожилого возраста [2]. В последние 15-20 лет появилось много лекарственных препаратов для терапии ДГПЖ, которые в настоящее время занимают важнейшее место в структуре лечения таких больных. По мнению Е.Б. Мазо и М.И. Белковской (2001), в XXI в. больные будут лечиться, в основном, медикаментозно, с обеспечением высокого качества жизни.

Так как в структуре этиопатогенетических факторов при ДГПЖ существенную роль играют социально-гигиенические факторы риска целесообразно организовать комплексный (наряду с медико-биологическими характеристиками) анализ причин возникновения, течения и особенностей реабилитационных мероприятий при ДГПЖ с целью научно-обоснованного принятия врачебных решений [1].

В связи с высокой медико-социальной значимостью проблемы ДГПЖ, с учетом мультифакторного характера особенностей развития и течения предпринята попытка проведения анализа главных компонент и классификации больных с доброкачественной гиперплазией предстательной железы при наличии или отсутствии сердечно-сосудистых заболеваний на основе создания векторного пространства латентных факторов с меньшей размерностью.

Задача анализа главных компонент, имеет, как минимум, четыре базовых версии: аппроксимировать данные линейными многообразиями меньшей размерности; найти подпространства меньшей размерности, в ортогональной проекции на которые разброс данных (то есть среднеквадратичное отклонение от среднего значения) максимален; найти подпространства меньшей размерности, в ортогональной проекции на которые среднеквадратичное расстояние между точками максимально; для данной многомерной случайной величины построить такое ортогональное преобразование координат, что в результате корреляции между отдельными координатами обратятся в ноль [2].

В соответствии с общепринятой схемой анализ главных компонент и классификация пациентов в зависимости от наличия заболеваний сердечнососудистой системы (ССС) были реализованы в ППП Statistica 8.0 Copyright© StatSoft, Inc (2007). В качестве активных переменных использовались семь видов социально-гигиенических характеристик, а именно: Age (возраст), Formation (образование), Family (семья), Material maintenance (материальное положение), Smoking (курение), Food (пищевые предпочтения), Trade (общественно-профессиональная группа). В качестве вспомогательных были выбраны такие переменные как Dream (характер сна), Life (жилищно-бытовые условия), Gymnastics (занятия утренней гимнастикой). Так же были выделены две группирующие переменные - активная Heart (наличие заболеваний сердечно-сосудистой системы) и Diabetes (наличие сахарного диабета в анамнезе).

Цель данного анализа — изучение взаимосвязей между различными социально-гигиеническими показателями, чтобы выявить скрытые факторы, которые упростили бы процесс классификации изучаемых групп населения. Для достижения этой цели необходимо определить факторные оси в пространстве меньшей размерности, на которые можно спроектировать пространство переменных анализа, а также сделать возможной визуализацию этих групп, т.е. нанести результаты на карту полученного пространства.

На начальном этапе с целью выделения скрытых социально-гигиенических факторов характеризующих состояние больного ДГПЖ был реализован факторный анализ, который позволили выделить два фактора (рис.1). Фактор_1 имеет значимые корреляции с такими переменными как возраст, образование, семейное положение, материальное положение, заболеваемостью сердечно-сосудистой системы. Фактор_2 с общественно-профессиональной группой и заболеваемостью сахарным диабетом.

Factor Loadings (Unrotated) (Act
Extraction: Principal component
(Marked loadings are >,700000)

Variable	Factor 1	Factor 2
AGE	**-0,976956**	0,014098
FORMATION	**-0,977317**	-0,113595
FAMILY	**0,913825**	-0,308599
MATERIAL MAIN	**0,825805**	-0,191910
SMOKING	0,605547	-0,306484
FOOD	0,094597	-0,047153
TRADE	0,498674	**0,760391**
DREAM	0,690115	0,540797
LIFE	-0,193677	0,699589
GYMNASTICS	0,399327	0,327339
HEART	**0,779552**	-0,370241
DIABETES	-0,088866	**-0,836007**
Expl.Var	5,339765	2,544538
Prp.Totl	0,444980	0,212045

Рис.1. Расчет факторных нагрузок

Таким образом, с учетом рассчитанных значений факторных нагрузок Фактор_1 можно назвать сердечно-сосудистым фактором, а Фактор_2 – диабетическим. Далее построим график каменистой осыпи, на котором в виде кусочно-линейной функции изображены собственные значения (рис.2).

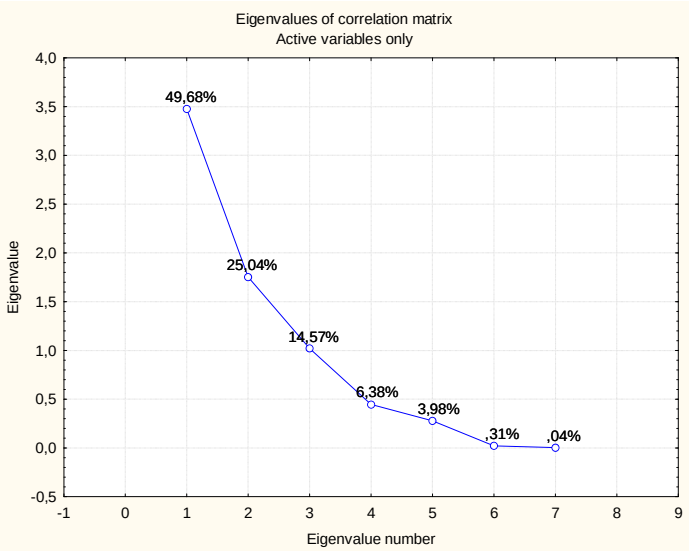


Рис.2. График каменистой осыпи

По критерию Кэттеля определяем собственное значение, начиная с которого «горка» теряет свою кривизну, т.е. убывание собственных значений максимально замедляется. Число выделяемых факторов должно быть равно номеру этого собственного значения. Из графика видно, что такими собственными значениями являются значения 2 или 3. Поэтому число выделяемых факторов может быть равно 2 или 3. На рис. 3 представлена таблица собственных значений, в которой приведен процент объясненной дисперсии (*Total variance*), кумулятивное собственное значение (*Cumulative Eigenvalue*) и кумулятивный процент (*Cumulative %*) объясненной дисперсии. Собственные значения представлены в порядке убывания, отражая тем самым степень важности соответствующих выделенных факторов для объяснения вариации исходных данных. Так, фактор, соответствующий максимальному собственному значению (3,447410), описывает приблизительно 49,6% общей вариации. Второй фактор для значения (1,752522) описывает 25,0 % общей вариации и т.д.

Value number	Eigenvalues of correlation matrix, and related statisti Active variables only			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,477410	49,67729	3,477410	49,6773
2	1,752522	25,03603	5,229932	74,7133
3	1,019694	14,56706	6,249626	89,2804
4	0,446733	6,38190	6,696359	95,6623
5	0,278795	3,98279	6,975155	99,6451
6	0,021777	0,31110	6,996931	99,9562
7	0,003069	0,04384	7,000000	100,0000

Рис.3. Таблица собственных значений

Далее вычислим факторные координаты переменных. Для этого построим таблицу координат исходных факторов в пространстве новых выделенных факторов (рис. 4).

Variable	Factor coordinates of th Active and Supplement *Supplementary variabl	
	Factor 1	Factor 2
AGE	0,937681	0,236268
FORMATION	-0,359810	0,812852
FAMILY	0,693071	0,616908
MATERIAL MAIN	-0,497055	0,106482
SMOKING	-0,941613	0,005143
FOOD	-0,687168	-0,251065
TRADE	0,618430	-0,762232
*DREAM	0,582319	-0,711791
*LIFE	-0,137734	-0,457884
*GYMNASTICS	-0,468103	0,160597

Рис.4. Факторные координаты выделенных факторов

В терминологии факторного анализа факторные координаты в методе главных компонент также называют «факторными нагрузками. С точки зрения математики, главный компонент — это линейная комбинация переменных, которые сильно коррелируют с ним. В дальнейшем подразумевается, что факторные координаты переменной — это корреляции между переменной и факторными осями.

Следовательно, интерпретация главных компонент должна быть сделана в терминах корреляции, т.е. нужно выделить те переменные (наблюдения), которые имеют наибольшие (абсолютные) значения факторных координат для данных факторов. Большее абсолютное значение факторной нагрузки переменной с каким-либо фактором говорит о том, что переменная сильнее связана с этим фактором. Другими словами, чем больше величина факторной координаты переменной, тем лучше переменные показывают структуру, представленную этим фактором. Например, фактор с высокими факторными нагрузками для трех измерений размеров человека, таких, как вес, рост и окружность грудной клетки, может рассматриваться как представляющий «размер» (т.е. абстракция трех переменных) человека.

Координаты отображаются как для активных переменных, так и для вспомогательных. Как видно из таблицы, первая факторная ось, соответствующая собственному значению 3,44, имеет сильные обратные корреляции с переменными возраст и образование, сильные прямые корреляции с семейным и материальным положением и умеренную корреляцию с заболеваемостью сердечно-сосудистой системы. Аналогично интерпретируются корреляции по второму фактору.

Процессу интерпретации факторов помогают графики факторных координат переменных и наблюдений (рис. 5). Как видно из рисунка, все переменные изображены в виде точек на единичном круге, так как корреляции (координаты точек) наблюдений с факторными осями принимают значения из интервала [0,1].

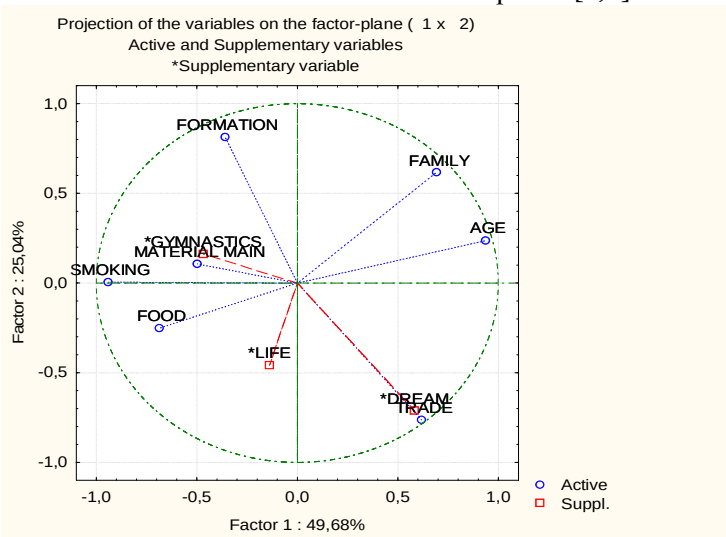


Рис.5. График факторных координат

Горизонтальная ось системы координат соответствует фактору 1, а вертикальная — фактору 2. В зависимости от знаков координат толчки расположены в соответствующих квадрантах плоскости. Основные и вспомогательные переменные изображены (на мониторе) соответственно кружочком синего цвета и прямоугольником красного цвета. Этот круг является визуальным индикатором того, насколько хорошо каждая переменная воспроизводится текущим набором выделенных факторов — чем ближе переменная к единичной окружности, тем лучше она воспроизведена в найденной системе координат.

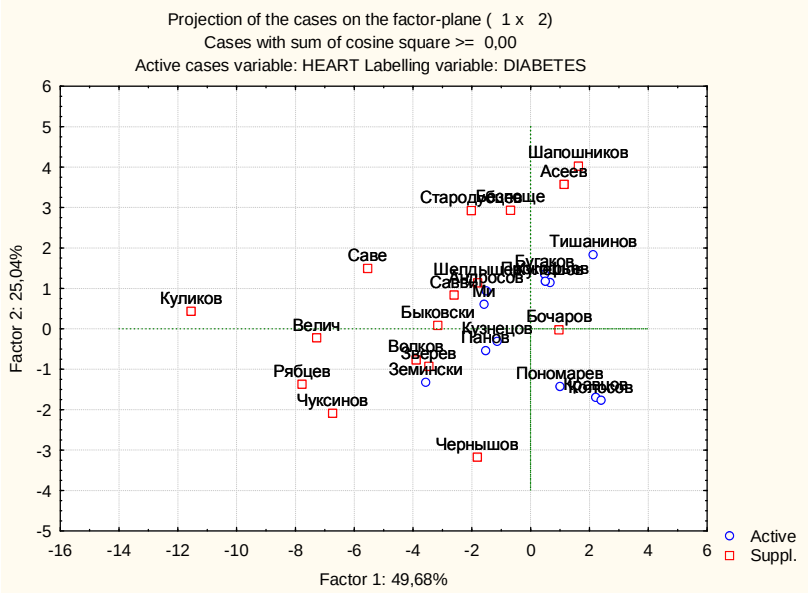
Далее построим таблицу вклада переменных (рис.6). Вклад переменной — это относительный вклад переменной в дисперсию факторной оси.

Значения этой статистики используются для отсеивания переменных, перед тем как они рассматриваются на основе факторных координат, т.е. корреляций для интерпретации факторных осей.

Variable	Variable contributior	
	Factor 1	Factor 2
AGE	0,252845	0,031853
FORMATION	0,037230	0,377015
FAMILY	0,138134	0,217159
MATERIAL MAIN	0,071048	0,006470
SMOKING	0,254970	0,000015
FOOD	0,135791	0,035967
TRADE	0,109983	0,331521

Рис. 6. Таблица вкладов основных переменных

Естественно, те переменные должны быть кандидатами для дальнейшей проверки, вклад (относительный) которых в дисперсию оси фактора больше. Обращает на себя внимание тот факт, что значения вкладов «пропорциональны» факторным нагрузкам. Теперь возможно построения графика наблюдений в факторном пространстве (рис. 7).



На этом графике изображаются как основные наблюдения, интерпретируемые как отсутствие заболеваний сердечно-сосудистой системы), которые использовались при расчете факторов (кружочки синего цвета), так и вспомогательные (имеются заболевания ССС) наблюдения (квадратики красного цвета).

Обратите внимание, что основные и вспомогательные наблюдения сгруппированы в разных областях плоскости, т.е. они объединены в группы однородности — кластеры. При этом кластер с вспомогательными группами отсутствия заболеваний ССС расположен во второй полуплоскости, т.е. имеет отрицательные значения координат по первой, горизонтальной оси и частично положительные значения координат

по второй, вертикальной оси (кроме некоторых наблюдений). Напомним, что горизонтальная факторная ось была нами интерпретирована как сердечно-сосудистый фактор, связанный с возрастом, образованием, материальным положением детьми. При этом отрицательную часть оси определяют переменные жилищно-бытовые условия и материальное положение (рис.5).

Из графика видно, что пациенты с отсутствием сердечно-сосудистой патологии сосредоточены в области переменных занятия гимнастикой и материальное положение. Это означает, что социально-гигиенические

факторы риска, характеризующие заболеваемость ДГПЖ в группах больных с сердечно-сосудистыми заболеваниями тесно связана с физическими упражнениями и материальным положением. Значимость социально-гигиенических характеристик у больных ДГПЖ без отсутствия сопутствующей патологии и в частности заболеваний сердечно-сосудистой системы носят более разносторонний характер — они сгруппированы более или менее равномерно на всей плоскости, что говорит о незначительной зависимости от какого либо конкретного фактора риска развития ДГПЖ.

Литература

1. Гориловский Л. М. Урологические заболевания в гериатрии / Л. М. Гориловский // Клиническая геронтология: Ежеквартальный научно-практический журнал. — 2002. — Том 8, № 11 . — С. 3-13.

2. Лопаткин Н.А. Доброкачественная гиперплазия предстательной железы. Под ред. Н.А. Лопаткина, Москва, 1999.

3. Халафян А.А. Современные методы медицинских исследований. М.: Издательство ЛКИ, 2008. – 320 с.

THE ANALYSIS OF THE PRINCIPAL COMPONENT AND CLASSIFICATION OF PATIENTS WITH THE GOOD-QUALITY HYPERPLASIA OF THE PROSTATE AT PRESENCE OFACCOMPANYING DISEASES

Klimenko G. Ja, S.J.Pehov, K.A.Razinkin

The Voronezh state medical academy of N.N.Burdenko
The Voronezh state technical university

Possibilities of the analysis of the principal component and classifications of patients with a good-qualitative hyperplasia of a prostate are surveyed at presence of accompanying diseases on the basis of creation of vectorial space of latent factors with smaller dimension