Московский государственный университет
Факультет вычислительной математики и кибернетики

Особенности обработки медицинских баз данных на примере задачи о профилактике сердечно-сосудистых заболеваний

Кочетов Е. В.,
Буничева А. Я.,
Орлова Я.А.,
Радаева К.В.,
Мухин С. И.

31.10.2024, Математические модели и численные методы в биологии и медицине, ИВМ РАН

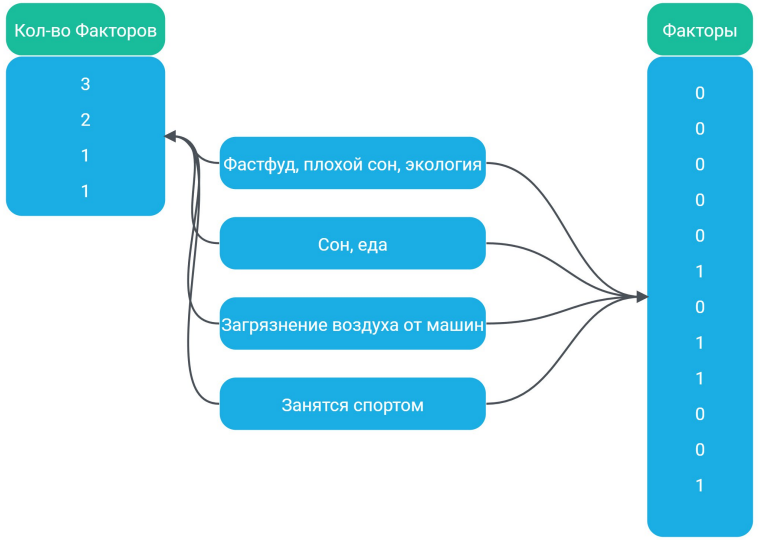
Основные цели

- Обработать базу данных пациентов медицинского научно-образовательного центра
- Исследовать возможность улучшения эффективности профилактики сердечно-сосудистых заболеваний на основе полученных данных
- Формирование «портретов пациентов», требующих разных подходов при коррекции факторов риска

Анкетирование по факторам риска

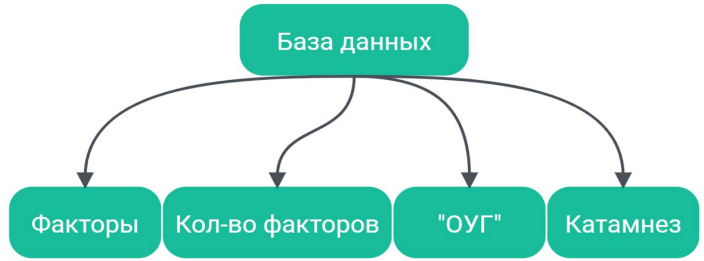
1. Как Вы считаете, какие факторы могут увеличивать риск развития сердечно-сосудистых заболеваний?
2. Какие из этих факторов человек может контролировать?
3. Какие факторы влияют или могут повлиять неблагоприятно на Ваше здоровье?
4. Какие из этих факторов Вы планируете в ближайшее время изменить для сохранения Вашего здоровья?

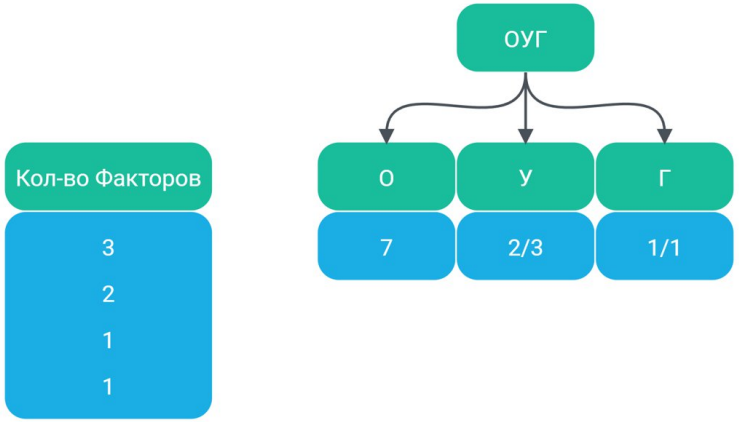
- Доступ к ответам на вопросы — не имеем!
- Выделено 12 основных групп факторов риска.
- Имеем доступ к «кол-во факторов» — количеству уникальных(относительно 12 групп) ответов на каждый из 4х вопросов
- Имеем доступ к «факторы» — бинарный вектор размерности 12, который показывает названные факторы риска при ответе на 4 вопроса анкеты



База данных

1) Факторы — 12-мерный вектор, полученный из анкеты
2) Кол-во факторов — 4-мерный вектор, полученный из анкеты
3) «ОУГ» — осведомленность респондентов о ФР, потенциальная управляемость ФР, готовность респондентов к изменениям
4) Катамнез — всевозможные данные о пациентах, которые включают в себя как данные опроса (курит ли пациент, имеет ли домашнее животное и т. д.), так и медицинские данные (глюкоза (Glu), ширина распред-я эритроцитов (RDW-CV), и т. д.).





Данные

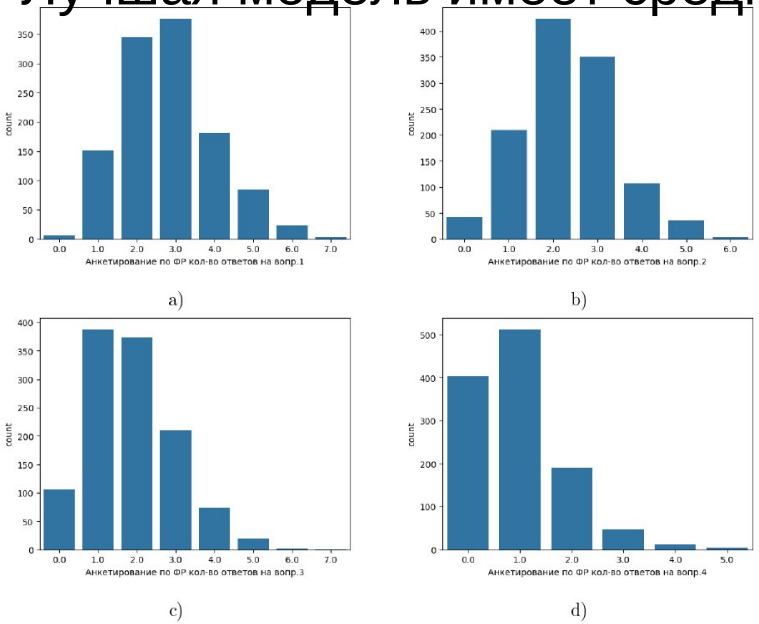
ФАКТОРЫ
Курение
Сахарный диабет
Артериальная гипертензия
ССЗ (кроме АГ)
Холестерин
Неправильное питание
Избыточный вес
Низкая физическая активность
Нарушение сна
Психологические
Межличностные отношения
Социально-экономические

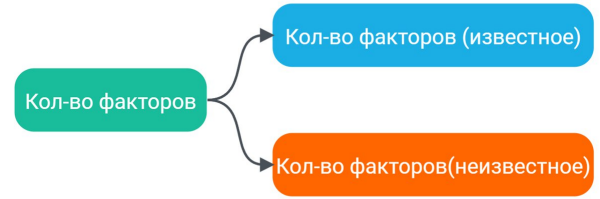
КАТАМНЕЗ
Питомцы
Образование
Тип факультета
Лучшая работоспособность
Сон в течение суток (часы)
Употребление алкоголя
Отношение к здоровью в период пандемии
Пол
Проживание
Наличие АГ
Наличие повышенного холестерина
Наличие ИМ и Инсульт у ближайших родственников

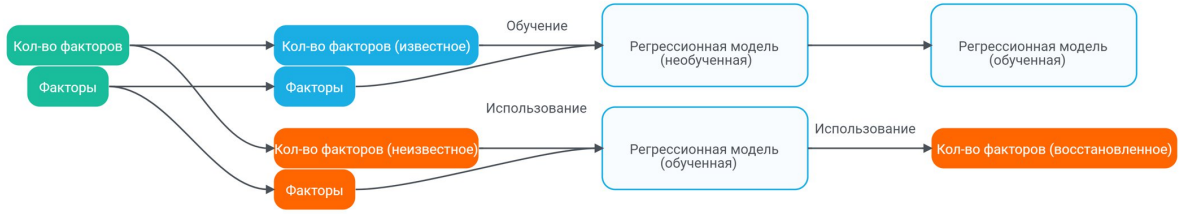
КАТАМНЕЗ
Курение
Ходьба в быстром темпе
Фрукты/овощи в рационе
Подсаливание пищи
Спорт 3 раза в неделю или чаще
Возраст
ИМТ
Обхват талии
Алкоголь общее кол-во баллов

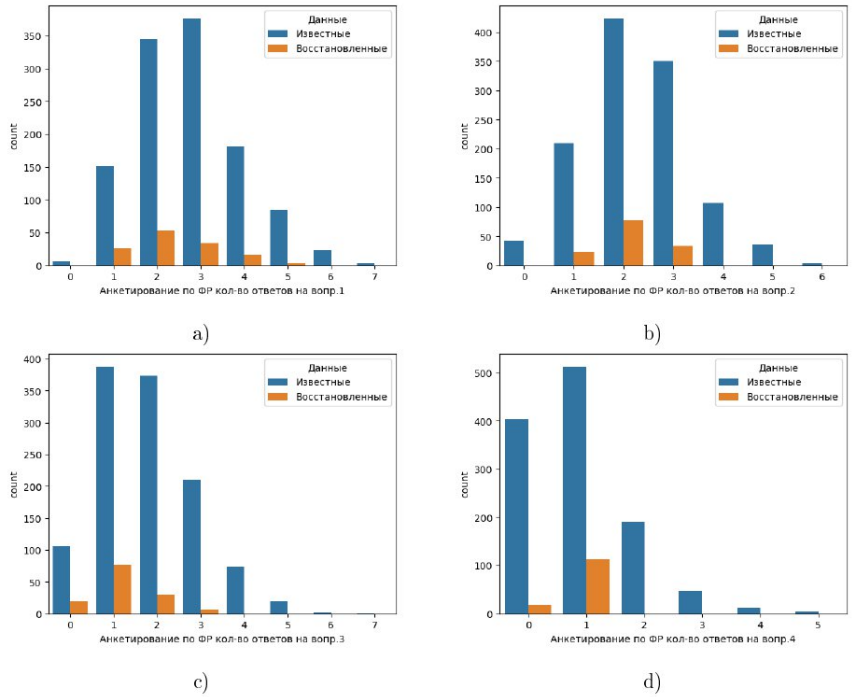
Восстановление данных

- Кол-во факторов содержат пропуски (129), но в этих случаях заполнены факторы
- Восстанавливались значения кол-ва факторов по значению факторов
- Рассматривались различные модели, и выбиралась лучшая($\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^*|$) на кроссвалидации
- Лучшая модель имеет среднее MAE равное 0.6.



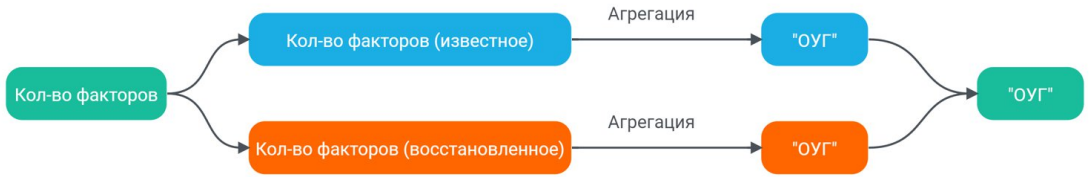


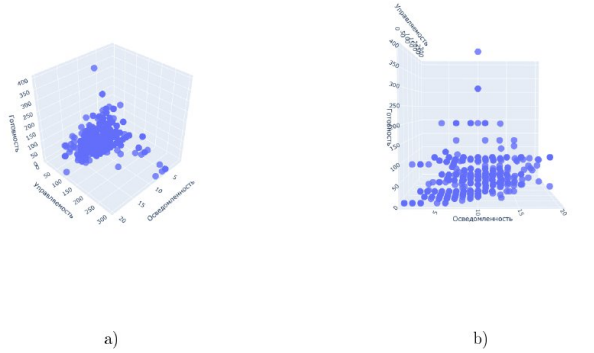


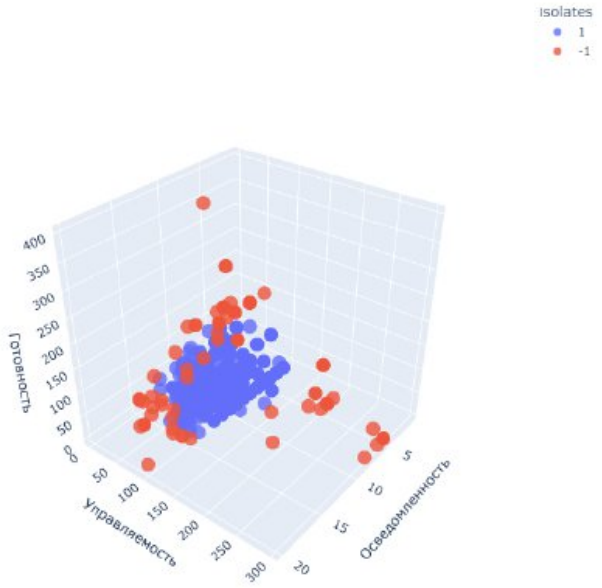


Переход к «ОУГ»

- С учетом восстановленных данных получено 1291 заполненных «кол-во факторов»
- По ним перешли в пространство «ОУГ»
- Устранили аномальные значения(91 объект) с помощью изоляционного леса

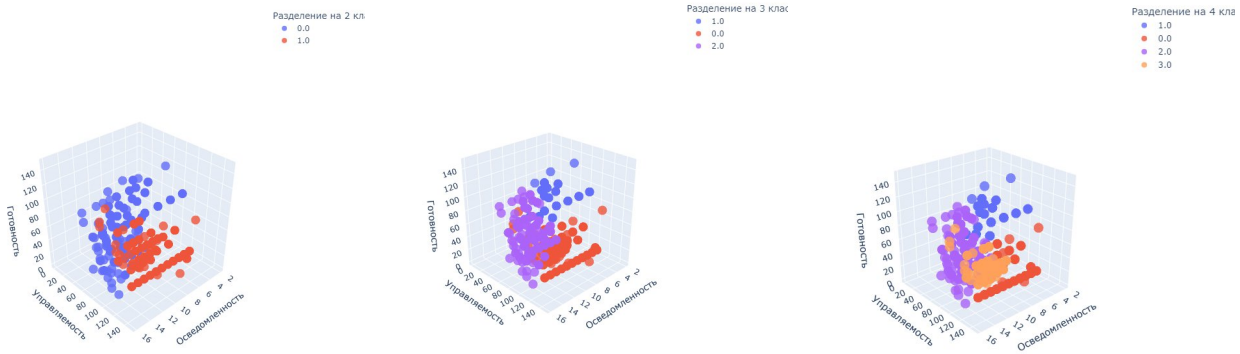


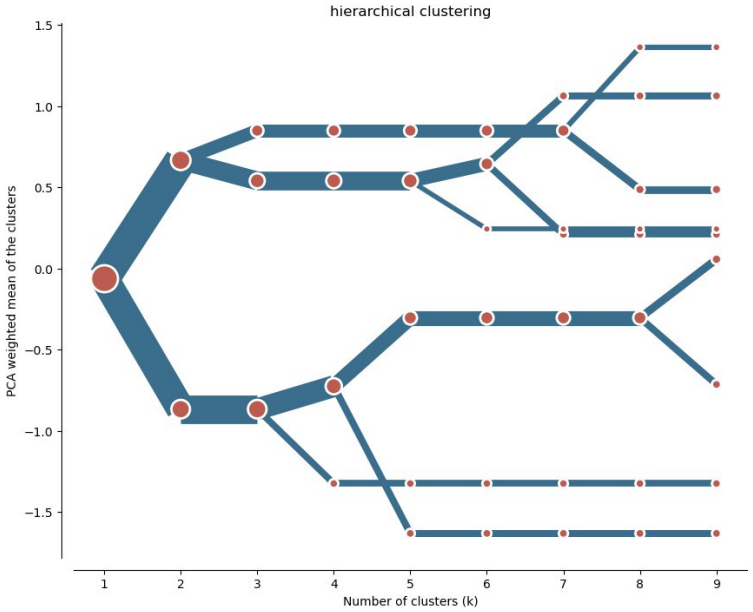


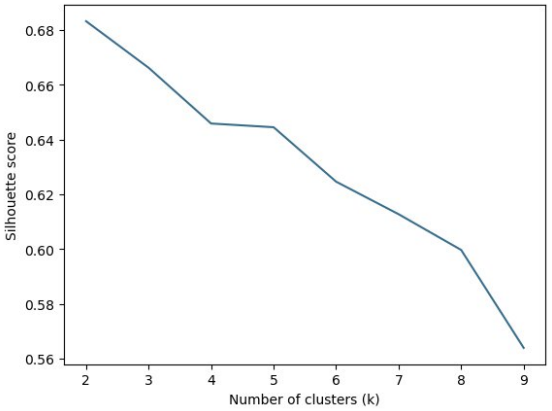


Кластеризация в пространстве «ОУГ»

- Производилась кластеризация в пространстве «ОУГ»
- Наиболее интересен случай с четырьмя кластерами
- Чтобы понять, насколько статистически значимы различия в полученных кластерах использовался t-критерий Уэлча



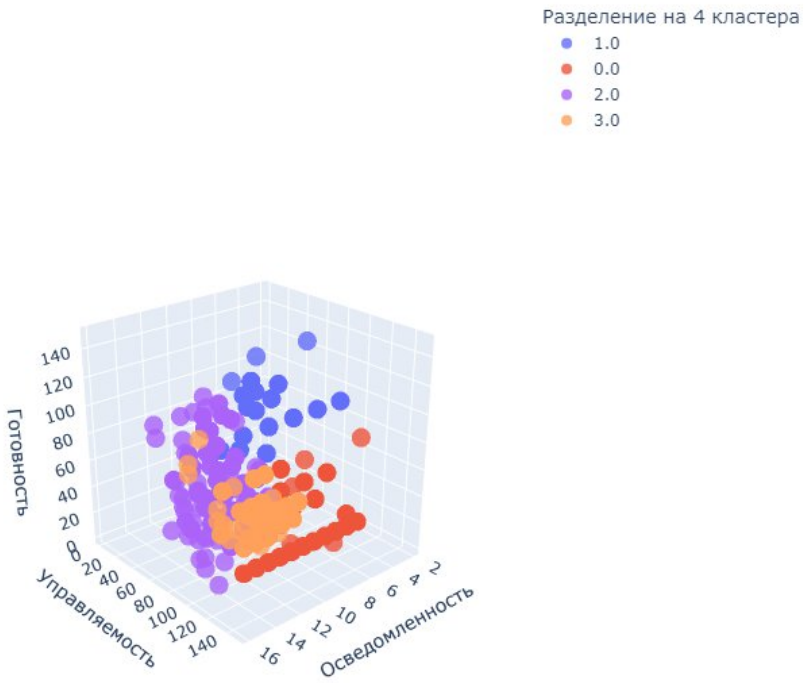




Результаты кластеризации в пространстве «ОУГ»

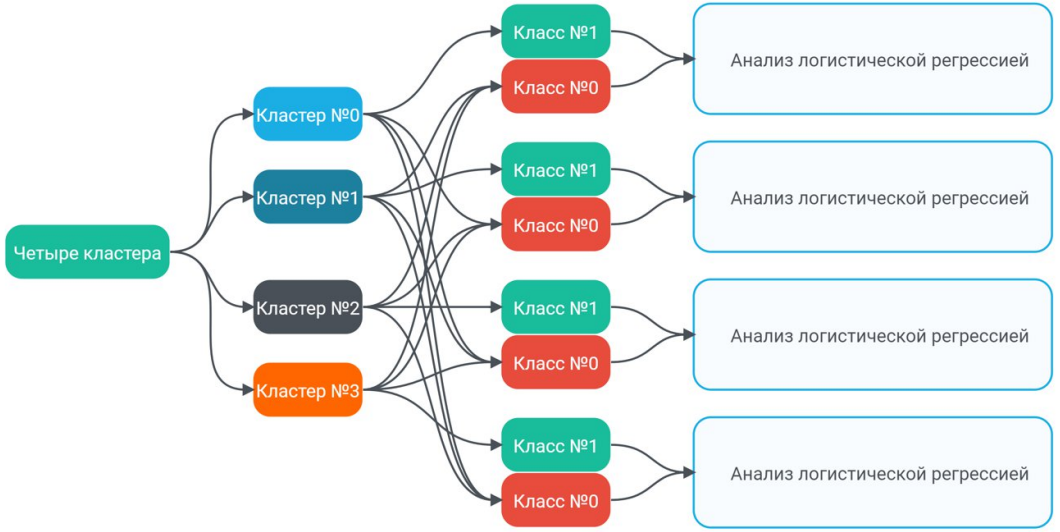
- Выделены четыре кластера с характерными свойствами.
- Кластер с наибольшим значением средней управляемости — группа №0
- Кластер с наибольшим значением средней готовности — группа №1
- Кластер с наибольшим значением средней осведомленности — группа №2
- Кластер с наименьшими средними значениями по всем показателям — группа №3

	Группа №0	Группа №1	Группа №2	Группа №3
Осведомленность	6.53	7.26	11.40	5.74
Управляемость	100.64	88.70	76.01	46.96
Готовность	16.38	98.01	55.74	0.80
Количество пациентов	438	377	261	135



Формирование «портрета пациента»

- Использовались модели логистической регрессии для установки связей в данных.
- Перебирались возможные комбинации признаков для выбора такого набора, который статистически значимо (p-value < 0.05) влияет на принадлежность к определенной группе.
- Для перебора и обучения использовался суперкомпьютер МГУ 270.
- Сравнивались пациенты, которые принадлежат одной группе (класс №1) со всеми остальными (класс №0).



Полученные «портреты пациентов»

Группа №0

Признак	Тенденция
Имеет ли кошку?	Увеличение на 66.34% если есть кошка
Лучшая работоспособность утром?	уменьшение на 19.98% если имеет лучшую работоспособность утром
ИМТ	увеличение на 3.19% с каждым пунктом
Не имеет АГ?	уменьшение на 29.70% если не имеет АГ

Группа №1

Признак	Тенденция
Отношение к здоровью в период пандемии не изменилось	уменьшение на 25.86% если отношение к здоровью в период пандемии не изменилось
Имеет ли вообще животное?	уменьшение на 30.05% если имеет вообще животное
Проживает один?	уменьшение на 26.39% если проживает один
Алкоголь общее кол-во баллов	уменьшение на 10.48% за каждый балл в анкете
Мужчина?	уменьшение на 21.68% если мужчина
Курит?	уменьшение на 35.77% если курит

Полученные «портреты пациентов»

Группа №2

Признак	Тенденция
Имеет повышенный холестерин?	увеличение на 110.61% если имеет повышенный холестерин
Имеет высшее образование?	увеличение на 26.65% если имеет высшее образование
Ест.-науч.?	уменьшение на 32.39% если не является представителем ест.-науч. факультета
Возраст	увеличение на 2.78% за каждый год

Группа №3

Признак	Тенденция
Сон в течение суток вне 7-8	увеличение на 70.72% если сон менее 7 часов или более 8
Имеет кандидата наук?	увеличение на 159.50% если имеет кандидата наук
Медик?	увеличение на 208.88% если медик
Употребление алкоголя < 1 раза в месяц	уменьшение на 34.77% если употребление алкоголя < 1 раза в месяц
Имеет АГ?	увеличение на 217.08% если имеет АГ
Занятие спортом 3 раза в неделю или чаще	уменьшение на 56.27% если занятие спортом менее 3х раз в неделю

Спасибо за внимание!