VI Национальный Суперкомпьютерный ФОРУМ 2017

Переславль-Залесский, Россия
Институт программных систем имени А.К. Айламазяна РАН
27 ноября – 01 декабря 2017 года

О.Ю. Колесниченко[1], Ю.Ю. Колесниченко[2], Л.О. Минушкина[3], Л.С. Мазелис[4], А.Л. Мазелис[4], А.Э. Николаев[4], К.И. Шахгельдян[4,5], В.Л. Авербух[6,7], И.О. Михайлов[7], А.В. Мартынов[8], В.В. Пулит[8], А.Н. Долженков[8], И.Н. Григоревский[9,10], Г.Н. Смородин[11]

[1]Security Analysis Bulletin, Москва;
[2]Uzgraph, Москва;
[3]ФГБУ ДПО «Центральная государственная медицинская академия УД ПРФ», Москва;
[4]ФГБОУ ВО «Владивостокский государственный университет экономики и сервиса», Владивосток;
[5]Дальневосточный федеральный университет, Владивосток;
[6]Институт математики и механики им. Н.Н. Красовского УрО РАН, Екатеринбург;
[7]Уральский федеральный университет, Екатеринбург;
[8]СП.АРМ, Санкт-Петербург;
[9]Национальная Суперкомпьютерная Технологическая Платформа, Переславль-Залесский;
[10]Институт программных систем им. А.К. Айламазяна РАН, Переславль-Залесский;
[11]Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (ИТМО), Санкт-Петербург

Аналитика Больших данных МИС

Аннотация. В статье представлены результаты аналитики Больших данных, регистрируемых медицинской информационной системой qMS за период с 2013 года по 2017 год – установленные во время госпитализации (несколько разных больниц) коды клинических диагнозов по МКБ-10, отметки о количестве и видах проведенных обследований, процедур и операций. Многоцентровое исследование «Аналитика Больших данных в медицине» выполняется под эгидой Национального Суперкомпьютерного Форума – Национальной Суперкомпьютерной Технологической Платформы России. Осуществлен разный математический анализ: кластерный анализ с использованием языка программирования Python; Булева алгебра логики, бинарный рефлексивный код Грея, реализованный на языке Java; построение графов; 3D-визуализация многомерных дискретных данных. Проанализированы выборки пациентов двух нозологических групп – артериальная гипертензия и сахарный диабет 1 типа. Планируется продолжение исследования с расширением спектра анализируемых Больших данных и методов анализа.

Ключевые слова: Большие данные, медицинская информационная система, qMS, кластерный анализ, код Грея, 3D-визуализация, Дополненная реальность, графовый анализ, сахарный диабет 1 типа, артериальная гипертензия.

Введение

Медицина сегодня подвергается информационной трансформации, становясь Data Driven-медициной. Облачные технологии, информационные системы, аналитика Больших данных – это новый облик медицинского учреждения. Ассоциация HIMSS (The Health Care Information and Management Systems Society) формирует мировые стандарты к медицинским информационным системам (МИС). Высшие ступени «зрелости» МИС по стандартам HIMSS содержат обязательные требования по готовности к аналитике Больших данных. В России необходимо интенсивно осваивать новую область – Большие данные, создавая многоцентровые команды, применяя разные методы аналитики, делая первые шаги к описанию новых практик. Так, многоцентровое исследование «Аналитика Больших данных в медицине» выполняется под эгидой Национального Суперкомпьютерного Форума – Национальной Суперкомпьютерной Технологической Платформы России. Главной целью этого многоцентрового исследования является формирование интероперабельных связей между врачами и специалистами в области математического моделирования для анализа данных МИС.

Используемые на современном этапе математические методы обработки больших массивов данных позволяют пересмотреть подходы к формированию баз знаний, содержащих информацию по закономерностям, зависимостям и связям, описывающим непрерывный лечебный процесс. То, что в рамках технологий бумажного документооборота было вовсе невозможно, при переходе на электронные варианты хранения данных требует не более месяца для разработки методов и несколько минут для работы алгоритмов при их внедрении в программное обеспечение МИС. Современные МИС позволяют существенно расширить и детализировать состав и структуру хранения данных, методы их обобщения и обработки. А выход за рамки форматов и описаний, утвержденных Минздравом, в результате дает возможность повышать качество и точность создаваемых прогностических моделей медицинского обслуживания населения в режиме «умного здравоохранения» или Smart Health.

Методы

Данные, регистрируемые МИС qMS (СП.АРМ) за период с 2013 года по 2017 год, представляют собой установленные во время госпитализации (несколько разных больниц) коды клинических диагнозов по МКБ-10, отметки о количестве и видах проведенных обследований, процедур и операций. Анализируемые выборки включают 861 пациента с диагнозом: сахарный диабет 1 типа; и 685 пациентов с диагнозом: гипертензивная (гипертоническая) болезнь с преимущественным поражением сердца без сердечной недостаточности (гипертоническая болезнь 2 стадии). Исследовательской группе не передавались персональные данные пациентов, информация о каждом из них кодировалась отдельным кодом по каждому эпизоду госпитализации.

В работе применена Булева алгебра логики с целью поиска уникальных совпадений в матрице данных. Этот метод позволяет гибко оценивать большой объем данных с учетом всех индивидуальных характеристик, что соответствует идеологии аналитики Больших данных, в отличие от усредняющего статистического анализа. Алгоритм бинарного рефлексивного кода Грея, реализованный на языке Java, создан Ю.Ю. Колесниченко. Метод математического анализа был применен для выборки пациентов с сахарным диабетом 1 типа.

В качестве инструмента для проведения кластерного анализа использован язык программирования Python. Для автоматизации расчетов была написана программа с использованием интерактивной среды iPython и библиотек NumPy, Pandas и Sklearn, предназначенных специально для анализа и обработки данных. Авторы программы: Л.С. Мазелис и А.Л. Мазелис. Метод математического анализа был применен для выборки пациентов с артериальной гипертензией.

Поиск связей между диагнозами при поступлении и выписке был осуществлен с помощью построения графов, автор программы: К.И. Шахгельдян. Метод математического анализа был применен для расширенной выборки пациентов с сахарным диабетом 1 и 2 типов.

Разработана интерактивная среда 3D-визуализации данных МИС. Разработчики программного обеспечения В.Л. Авербух и И.О. Михайлов. Метод аналитики был применен к выборке пациентов с сахарным диабетом 1 типа. Рассматривается многомерное пространство данных, где в качестве измерений можно использовать характеристики больных и результаты их обследования и лечения (столбцы таблицы метаданных). Разработанный прототип системы позволяет комбинировать сразу несколько типов данных в едином трехмерном поле. Существует возможность масштабирования и гиперактивной детализации информации о каждом конкретном пациенте. Возможна смена набора измерений по ходу анализа данных. Визуальное пространство можно вращать. В дальнейшем предусматривается возможность использования в рамках данной системы средств виртуальной и расширенной реальности (Дополненной реальности).

Отрабатываются подходы: описания социально-медицинского портрета пациента конкретной нозологической группы, выявления прогностических моделей и реперных комбинированных точек лечебно-диагностического процесса, подсчета относительного риска, обнаружения новых акцентов в ведении пациентов, оценки процесса уточнения основного диагноза пациента. Все предложенные направления требуют дальнейшего развития до уровня методических рекомендаций по аналитике медицинских Больших данных.

Результаты и обсуждение

МИС qMS осуществляет запись данных в условиях реального лечебно-диагностического процесса, со всеми преимуществами электронных записей и со всеми недостатками, связанными с внедрением электронных технологий в здравоохранение, включая не 100%-ю регистрацию информации в базы данных со стороны медицинских работников. К анализу данных МИС необходимо искать подход из области аналитики Больших данных, так как они характеризуются большим объемом и постоянным накоплением, разнообразием, всеобъемлющим охватом какого-то процесса с допустимостью единичных неточностей при регистрации данных. Для аналитики Больших данных важно, чтобы анализируемый процесс был целостным, масштабным и непрерывным во времени. В этом аспекте регистрация данных с потока пациентов в больницах является идеальным источником для применения подходов аналитики Больших данных.

Представлены результаты анализа данных пациентов с ***инсулинзависимым сахарным диабетом (СД) 1 типа***. Хранение регистрируемых данных в едином дата-центре СП.АРМ позволило взглянуть на поток пациентов с СД в масштабе всего лечебно-диагностического процесса больниц, а не отдельных специализированных отделений. Если пациент поступал в больницу и имел запись в МИС о диагнозе СД 1 типа, то, независимо в каком отделении он лечился и по какому поводу, он был выбран в общий список для аналитики. Это открывает новые возможности для того, чтобы взглянуть на конкретное заболевание не с узкой позиции заранее сконструированного дизайна для клинического исследования, а в целом, как реально та или иная нозологическая категория влияет на всю заболеваемость пациентов и на организацию лечебно-диагностического процесса.

На рис. 1 показан ***граф***, отражающий связь между «вершинами» – установленными диагнозами СД на этапах поступления в больницу, во время пребывания в стационаре и при выписке, а также между часто встречающимися сопутствующими заболеваниями (коды диагнозов согласно МКБ-10). Графовый анализ визуализирует процесс уточнения

диагноза СД 1 типа в период госпитализации. Данный математический анализ показал, что самый частый диагноз при поступлении пациентов с СД 1 типа – E10.7, с множественными осложнениями; этот диагноз как основной в преобладающем большинстве случаев остается неизмененным при выписке пациентов из стационара. Обнаружено, что в 10% случаев этиология СД уточняется в период госпитализации, при этом в 9% случаев диагноз меняется с СД 1 типа на СД 2 типа. Таким образом, можно говорить о гипердиагностике СД 1 типа в объеме 9% на догоспитальном уровне. Из зафиксированных в МИС сопутствующих диагнозу E10.7 частых заболеваний выявлены: артериальная гипертензия, гастрит и почечная недостаточность в равных долях.

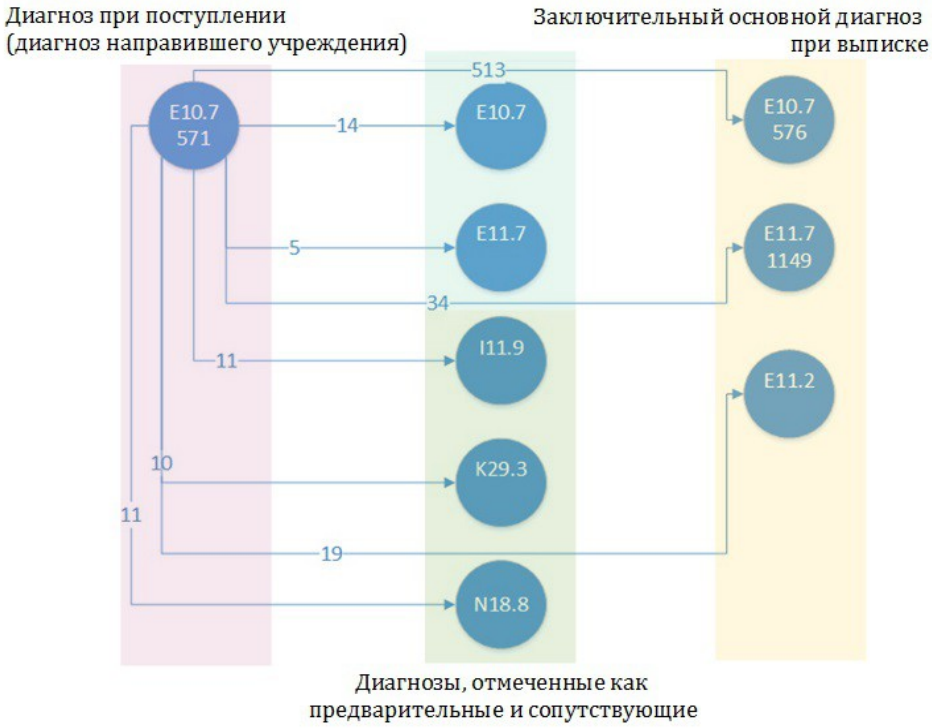


Рис. 1. Графовый анализ: связь между диагнозом E10.7 и наиболее частыми диагнозами за период госпитализации, вершины – диагнозы по МКБ-10; ребра – количество одинаковых связей (или число пациентов). E10.7 – инсулинзависимый сахарный диабет 1 типа с множественными осложнениями; E11.7 – инсулиннезависимый сахарный диабет 2 типа с множественными осложнениями; E11.2 – инсулиннезависимый сахарный диабет 2 типа с поражением почек; I11.9 – гипертензивная (гипертоническая) болезнь с преимущественным поражением сердца без сердечной недостаточности; K29.3 – хронический поверхностный гастрит; N18.8 – другие проявления хронической почечной недостаточности. Анализ выполнен К.И. Шахгельдян.

Ячейки с данными о выполненных диагностических обследованиях, процедурах и операциях пациентам с СД 1 типа были объединены по группам в «логические ячейки» или ***реперные точки***, которые затем подвергались математической обработке методом бинарного рефлексивного кода Грея. Дублирование пациентов в рамках одной реперной точки исключено. Реперные точки представлены в таблице 1.

Таблица 1

Объединение ячеек данных в реперные точки

Логическая ячейка	**n, %***
Реперная точка: Диабетическая нефропатия, искусственное очищение крови (диализ). **Д** (*диализ*)	120, 14%
Реперная точка: биопсия, пересадка почки, трансплантированная почка. **ТП** (*трансплантированная почка*)	36, 4,2%
Реперная точка: макроангиопатия, ИБС, коронарные артерии, оперативное лечение. **ОКА** (*операции на коронарных артериях*)	11, 1,3%
Реперная точка: макроангиопатия, атеросклероз, оперативное лечение. **ОМА** (*операции при макроангиопатиях*)	24, 2,8%
Реперная точка: остеопороз, диагностика. **ДОП** (*диагностика остеопороза*)	125, 14,5%
Реперная точка: поражение ЖКТ, диагностика. **ДПЖКТ** (*диагностика поражений желудочно-кишечного тракта*)	186, 21,6%
Реперная точка: оперативное лечение ЖКТ. **ОЖКТ** (*операции на органах желудочно-кишечного тракта*)	15, 1,7%
Реперная точка: поражение бронхов / легких, диагностика, операции. **ДПБЛ** (*диагностика поражений бронхов и легких, операции*)	233** 27%

* % от всей выборки (861 пациент).

**Из них 181 – рентгенография легких, 52 – остальные обследования, процедуры, операции. По Стандарту специализированной медицинской помощи при инсулинзависимом сахарном диабете (Приказ Министерства здравоохранения РФ от 24 декабря 2012 года № 1552н) рентгенография легких входит в перечень обследований с коэффициентом 1.

На рис. 2 процентное соотношение реперных точек представлено в графическом виде. Показано, что из анализируемого объема ячеек наибольшее число обследований относилось к бронхам и легким. Вторая по частоте встречаемости реперная точка относится к обследованию желудочно-кишечного тракта (ЖКТ). Третья по частоте встречаемости – реперная точка ДОП, рентгеновская денситометрия костей скелета. Можно отметить еще одну реперную точку – Д, диализ, которая отражает самый неблагоприятный исход патогенеза заболевания. В реперную точку Д сгруппированы разные указания в МИС на то, что пациент перешел на диализ (включая не только саму ячейку «Гемодиализ», но и ячейки с указанием на хирургическую подготовку сосудистого доступа и т.д.). Остальные выбранные реперные точки не имеют большой частоты встречаемости.

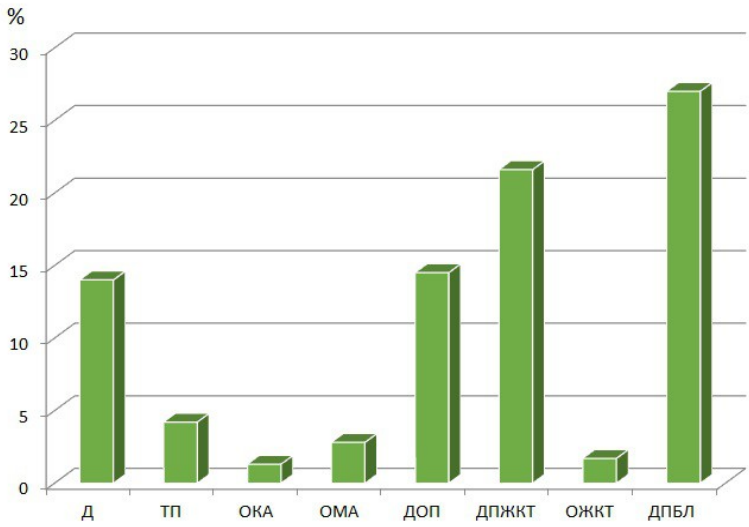


Рис. 2. Процентное распределение реперных точек (см. табл. 1).

При применении бинарного рефлексивного кода Грея идет поиск по матрице данных всех возможных комбинаций, число которых растет с увеличением числа столбцов переменных. Например, выбранные 8 реперных точек представляют собой 8 столбцов таблицы с данными по каждому пациенту. Итоговый результат математического анализа содержит $2^8 = 256$ вариантов совпадений (или не совпадений) данных в столбцах (все возможные варианты сопоставления данных в 8-ми столбцах). Сделать такой анализ вручную без компьютерной программы не представляется возможным. Для оптимизации количества обрабатываемых вариантов число реперных точек (столбцов данных) решено было ограничить до 8-ми, в то же время можно продолжить исследование, и на следующих этапах выбирать другие реперные точки для анализа с применением кода Грея.

В таблице 2 представлены выбранные комбинации кода Грея, имеющие значение для анализа. В таблице 3 показаны все обнаруженные комбинации. Разработчики МИС могут установить в системе конкретные параметры и применять уникальный код комбинации для моментального поиска по базе данных нужной конкретной комбинации, с целью оценки потока пациентов в больнице за определенный период времени. Это позволяет не только сделать выводы по заболеваемости контингента, поступающего в конкретное лечебное учреждение, но и дает возможность для более точного прогностического планирования медпомощи (закупки препаратов, расходных материалов, медтехники, укомплектованность медицинскими работниками).

На рис. 3 представлена ***схема (граф),*** включающая все обнаруженные варианты при комбинации 8-ми реперных точек (не учтены только комбинации, встречающиеся один раз). Можно отметить, что выявлен треугольник часто встречающихся и связанных для одного и того же пациента реперных точек – ДПБЛ (диагностика поражений бронхов и легких, операции), ДПЖКТ (диагностика поражений желудочно-кишечного тракта), ДОП (диагностика остеопороза).

Таблица 2

Булева алгебра логики, выбранные комбинации 8-ми реперных точек, определенные методом бинарного рефлексивного кода Грея

n – количество пациентов с конкретными уникальными комбинациями реперных точек

ДПЖКТ	ДОП	Д	ОКА	ОМА	ОЖКТ	ТП	ДПБЛ	Код	Комбинация (n)
Часто встречающиеся комбинации с совпадением 2-х реперных точек (n>10)									
1	0	0	0	0	0	0	1	10000001	ДПЖКТ, ДПБЛ (71)
1	1	0	0	0	0	0	0	11000000	ДПЖКТ, ДОП (59)
0	0	1	0	0	0	0	1	00100001	Д, ДПБЛ (32)
0	1	0	0	0	0	0	1	01000001	ДОП, ДПБЛ (28)
0	0	0	0	0	0	1	1	00000011	ТП, ДПБЛ (18)
0	0	1	0	1	0	0	0	00101000	Д, ОМА (14)
1	0	0	0	0	0	1	0	10000010	ДПЖКТ, ТП (15)
1	0	1	0	0	0	0	0	10100000	ДПЖКТ, Д (13)
0	1	0	0	0	0	1	0	01000010	ДОП, ТП (12)
Часто встречающаяся комбинация с совпадением 3-х реперных точек (n>10)									
1	1	0	0	0	0	0	1	11000001	ДПЖКТ, ДОП, ДПБЛ (20)
Комбинации с совпадением 4-х реперных точек (n>2)									
1	1	0	0	0	0	1	1	11000011	ДПЖКТ, ДОП, ТП, ДПБЛ (7)
1	1	1	0	0	0	0	1	11100001	ДПЖКТ, ДОП, Д, ДПБЛ (3)
Единственная комбинация с совпадением 5-ти реперных точек*									
1	1	1	0	0	0	1	1	11100011	ДПЖКТ, ДОП, Д, ТП, ДПБЛ (1)

*Пациентка, 41 год, основной диагноз E10.7, продолжительность госпитализации 32 дня.

Таблица 3

Булева алгебра логики, все обнаруженные комбинации 8-ми реперных точек, определенные методом бинарного рефлексивного кода Грея

n – количество пациентов с конкретными уникальными комбинациями реперных точек

ДПЖКТ (186)	ДОП (125)	Д (120)	ОКА (11)	ОМА (24)	ОЖКТ (15)	ТП (36)	ДПБЛ (233)	Код	Комбинация (n)
0	0	0	0	0	0	1	1	00000011	ТП, ДПБЛ (18)
0	0	0	0	0	1	0	1	00000101	ОЖКТ, ДПБЛ (7)
0	0	0	0	1	0	0	1	00001001	ОМА, ДПБЛ (6)
0	0	0	0	1	0	1	0	00001010	ОМА, ТП (1)
0	0	0	1	0	0	0	1	00010001	ОКА, ДПБЛ (3)
0	0	1	0	0	0	0	1	00100001	Д, ДПБЛ (32)
0	0	1	0	0	0	1	0	00100010	Д, ТП (7)
0	1	0	0	0	0	0	1	01000001	ДОП, ДПБЛ (28)

0	1	0	0	0	0	1	0	01000010	ДОП, ТП (12)
0	1	0	0	0	1	0	0	01000100	ДОП, ОЖКТ (1)
0	1	1	0	0	0	0	0	01100000	ДОП, Д (4)
1	0	0	0	0	0	0	1	10000001	ДПЖКТ, ДПБЛ (71)
1	0	0	0	0	0	1	0	10000010	ДПЖКТ, ТП (15)
1	0	0	0	0	1	0	0	10000100	ДПЖКТ, ОЖКТ (7)
1	0	0	0	1	0	0	0	10001000	ДПЖКТ, ОМА (5)
0	0	1	0	0	1	0	0	00100100	Д, ОЖКТ (2)
0	0	0	1	1	0	0	0	00011000	ОКА, ОМА (1)
0	0	1	0	1	0	0	0	00101000	Д, ОМА (14)
0	0	1	1	0	0	0	0	00110000	Д, ОКА (1)
1	0	1	0	0	0	0	0	10100000	ДПЖКТ, Д (13)
1	1	0	0	0	0	0	0	11000000	ДПЖКТ, ДОП (59)
1	1	1	0	0	0	0	0	11100000	ДПЖКТ, ДОП, Д (4)
0	0	0	0	1	0	1	1	00001011	ОМА, ТП, ДПБЛ (1)
0	0	1	0	0	0	1	1	00100011	Д, ТП, ДПБЛ (6)
0	0	1	0	0	1	0	1	00100101	Д, ОЖКТ, ДПБЛ (1)
0	0	1	0	1	0	0	1	00101001	Д, ОМА, ДПБЛ (4)
0	0	1	1	1	0	0	0	00111000	Д, ОКА, ОМА (1)
0	1	0	0	0	0	1	1	01000011	ДОП, ТП, ДПБЛ (8)
0	1	0	0	0	1	0	1	01000101	ДОП, ОЖКТ, ДПБЛ (1)
0	0	1	0	1	0	1	0	00101010	Д, ОМА, ТП (1)
0	1	1	0	0	0	0	1	01100001	ДОП, Д, ДПБЛ (3)
0	1	1	0	0	0	1	0	01100010	ДОП, Д, ТП (1)
1	0	0	0	0	1	0	1	10000101	ДПЖКТ, ОЖКТ, ДПБЛ (2)
1	0	0	0	1	0	0	1	10001001	ДПЖКТ, ОМА, ДПБЛ (3)
1	0	1	0	0	0	0	1	10100001	ДПЖКТ, Д, ДПБЛ (9)
1	0	1	0	0	0	1	0	10100010	ДПЖКТ, Д, ТП (2)
1	0	1	0	1	0	0	0	10101000	ДПЖКТ, Д, ОМА (2)
1	0	1	0	0	1	0	0	10100100	ДПЖКТ, Д, ОЖКТ (1)
1	1	0	0	0	1	0	0	11000100	ДПЖКТ, ДОП, ОЖКТ (1)
1	1	0	0	0	0	0	1	11000001	ДПЖКТ, ДОП, ДПБЛ (20)
0	0	1	0	1	0	1	1	00101011	Д, ОМА, ТП, ДПБЛ (1)
0	1	1	0	0	0	1	1	01100011	ДОП, Д, ТП, ДПБЛ (1)
1	0	1	0	0	0	1	1	10100011	ДПЖКТ, Д, ТП, ДПБЛ (2)
1	0	1	0	1	0	0	1	10101001	ДПЖКТ, Д, ОМА, ДПБЛ (1)
1	1	0	0	0	0	1	1	11000011	ДПЖКТ, ДОП, ТП, ДПБЛ (7)
1	1	0	0	0	1	0	1	11000101	ДПЖКТ, ДОП, ОЖКТ, ДПБЛ (1)

1	1	1	0	0	0	0	1	11100001	ДПЖКТ, ДОП, Д, ДПБЛ (3)
1	1	1	0	0	0	1	0	11100010	ДПЖКТ, ДОП, Д, ТП (1)
1	1	1	0	0	0	1	1	11100011	ДПЖКТ, ДОП, Д, ТП, ДПБЛ (1)

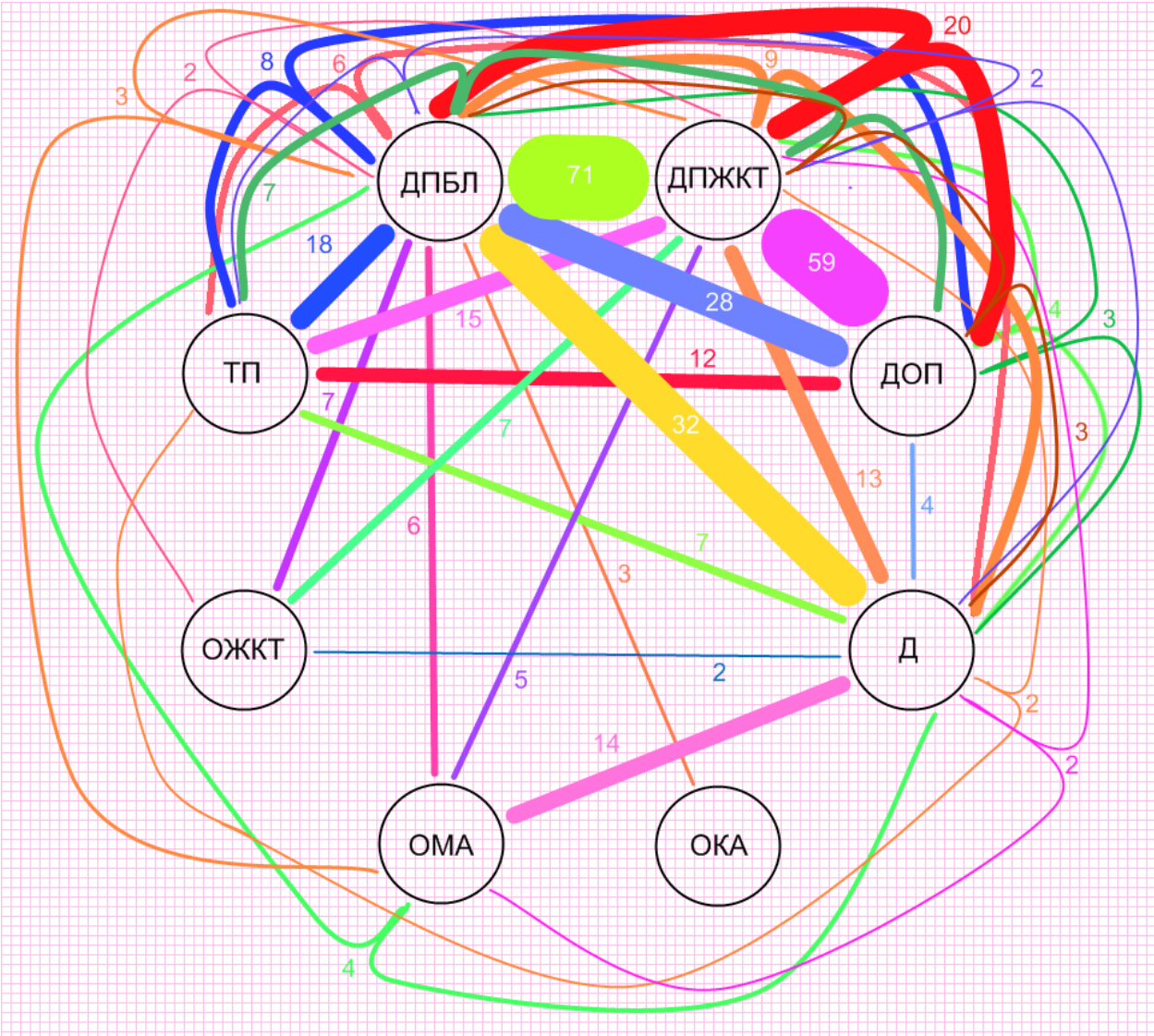


Рис. 3. Взвешенный ненаправленный граф (послойное нанесение в Paint.net) по результатам математической обработки реперных точек бинарным рефлексивным кодом Грея. Граф выполнен Ю.Ю. Колесниченко.

Представляется интересной связь двух реперных точек Д (диализ) и ОМА (операции при макроангиопатиях). Данная связь очень наглядно отражает общий патогенетический негативный исход при СД – поражение почек с развитием ХПН (хроническая почечная недостаточность) и необходимостью применять гемодиализ (или перитонеальный диализ), и атеросклеротическое поражение крупных артерий, требующее хирургического вмешательства.

В нашем исследовании записей МИС мы показали пример оценки шансов совпадения осложнений, требующих дорогостоящего медицинского обеспечения и

соответствующего планирования медпомощи. Реперная точка ОМА не включает операции на коронарных артериях, они специально были определены в отдельную реперную точку ОКА. При рассмотрении связей между реперными точками было замечено, что можно подсчитать ***отношение шансов (Odds Ratio, OR)*** обнаружения атеросклеротического поражения периферических и брахиоцефальных артерий, требующего оперативного вмешательства при ХПН с диализом у пациентов с СД 1 типа. Получился статистически достоверный показатель OR=9,6 при $p<0,05$. То есть, у пациентов с ХПН при СД 1 типа, подвергающихся диализу, в 9,6 раз выше шанс обнаружить атеросклеротическое поражение периферических и брахиоцефальных артерий, требующее оперативного вмешательства. Также был рассчитан ***относительный риск (Relative Risk, RR)*** для тех же двух реперных точек: RR=8,6 при $p<0,05$ (чувствительность Se=0,58, специфичность Sp=0,87). ХПН при СД 1 типа с диализом повышает риск атеросклеротического поражения периферических и брахиоцефальных артерий, требующего оперативного вмешательства, в 8,6 раз.

Можно выделить группу пациентов, для которых совпали все три ведущие реперные точки (комбинированная реперная точка), их 20 человек (2,3% от всей выборки 861 пациента). По этой доле ***комбинированного «тройного совпадения»*** можно судить о степени нагрузки на стационар, обусловленной поступающим контингентом больных за определенный промежуток времени, а также прогнозировать эту нагрузку при изменении процента «тройного совпадения». С помощью бинарного рефлексивного кода Грея были выявлены связи определенных диагнозов с комбинированной точкой «тройного совпадения»: сочетание с диагнозом E10.7, с множественными осложнениями – n=11; E10.2, с поражением почек – n=8.

При анализе группы пациентов с отметкой о диагнозе E10.0, кома (n=44), было выявлено следующее распределение сопутствующих диагнозов: злокачественные новообразования легких и бронхов (n=9); другие злокачественные новообразования (n=3); доброкачественные и неуточненные новообразования (n=5); панкреатиты (n=8); заболевания ЖКТ (n=9); заболевания сердца и сосудов (n=7); зоб (n=3). Эти данные указывают на преобладание рака легких и бронхов в группе декомпенсированного СД 1 типа, а также ***на связь рака легких и бронхов с утяжелением клинического течения СД 1 типа***. Для уточнения выявленной тенденции был проведен анализ всех отметок о раке в анализируемой выборке пациентов. Было обнаружено, что 11 пациентов имеют злокачественные опухоли легких и бронхов (3 мужчин и 8 женщин, возраст 70,6±4,2 лет (от 52 лет до 87 лет). Также в выборке пациентов отмечено 10 случаев с диагнозами рубрики МКБ-10 D: доброкачественные образования паращитовидной железы, молочной железы, восходящей ободочной кишки, трахеи, (n=6); новообразования неопределенного или неизвестного характера щитовидной железы и брюшины, забрюшинного пространства (n=4).

Если сравнивать с показателем заболеваемости раком легких по России на 100 тыс. человек, то этот показатель не превышает 100 (колеблется от 66 до 96 на 100 тыс. чел.), при доминировании мужской заболеваемости (для женщин этот показатель колеблется от 6,9 до 12,5 на 100 тыс. чел.) [1]. Переведя данные нашего исследования на 100 тыс. чел., получится, что для больных СД 1 типа показатель заболеваемости раком легких достигает 1277 на 100 тыс. пациентов (при n=11, диагнозы МКБ-10 C34 и C78) или 929 на 100 тыс. пациентов (при n=8, без учета МКБ-10 C78, или только для женщин).

Применение режима ***интерактивной 3D-визуализации*** данных МИС для выборки пациентов с СД 1 типа позволило создать гиперактивную среду, в которой возникло нелимитированное множество вариантов для проведения анализа лечебно-диагностического процесса в период госпитализации. На рис. 4, 5 и 6 показаны примеры 3D-визуализации пространства данных с разными опциями. На первом этапе создан прототип системы интерактивной визуализации, позволяющий комбинировать сразу несколько типов данных в едином трехмерном поле, с возможностью масштабирования и

гиперактивной детализации информации о каждом конкретном пациенте. Показана проекция многомерного пространства на трехмерное пространство визуализации – то есть отображение выбранных столбцов таблицы метаданных в нескольких визуальных измерениях: три пространственные координаты, размер, цвет и форма маркеров. Также дополнительной опцией может быть анимация маркеров. При использовании средств виртуальной и расширенной реальности пользователь (врач) получит возможность анализа данных изнутри визуального поля, а также дополнительные средства навигации по данным.

На рис. 4 представлена вся выборка пациентов с СД 1 типа, в фокусе анализа – частота проведения этим пациентам УЗИ ***щитовидной железы****. Актуальность анализа проведения обследований щитовидной железы у пациентов с СД 1 типа обусловлена встречающимся единым аутоиммунным генезом поражения поджелудочной и щитовидной желез – аутоиммунный полигландулярный синдром. Цветом показана длительность ожидания УЗИ щитовидной железы: желтый цвет соответствует среднему ожиданию УЗИ, зеленый – длительному ожиданию УЗИ (красный цвет соответствует отсутствию проведения УЗИ). Данная конфигурация опций позволяет врачу сразу охватить взглядом картину распределения пациентов по длительности госпитализации и по ожидаю УЗИ щитовидной железы, во взаимосвязи с возрастом, полом и кодами / подразделами по МКБ-10. Детальное описание этого направления аналитики запланировано для следующих публикаций. Кратко можно отметить, что в основном не было задержек в проведении этого исследования, а конкретные случаи длительного ожидания можно гиперактивно рассмотреть, изменив масштаб. Исследование пациенты проходили в первую половину своего срока госпитализации. Видна корреляция – чем больше возраст пациентов, тем чаще им назначают УЗИ щитовидной железы. На рис. 5 представлено 3D-распределение по длительности эпизодов госпитализации. Длительность эпизода показана как цветом, так и координатой по Z. Видна доминирующая в выборке группа в определенном диапазоне возраста и с определенной длительностью нахождения в стационаре. При этом в целом можно заместить корреляцию между возрастом и длительностью эпизода госпитализации. На рис. 6 представлена 3D-картина взаимосвязи количества проведенных в выборке пациентов УЗИ щитовидной железы и времени ожидания этого исследования. Можно отметить доминирующую возрастную группу и неоднородное распределение маркеров по ожиданию УЗИ. Более детальное описание запланировано на следующем этапе исследования в процессе совершенствования версии программного обеспечения.

*По Стандарту специализированной медицинской помощи при инсулинзависимом сахарном диабете (Приказ Министерства здравоохранения РФ от 24 декабря 2012 года № 1552н) ультразвуковое исследование щитовидной железы входит в перечень обследований с коэффициентом 0,1.

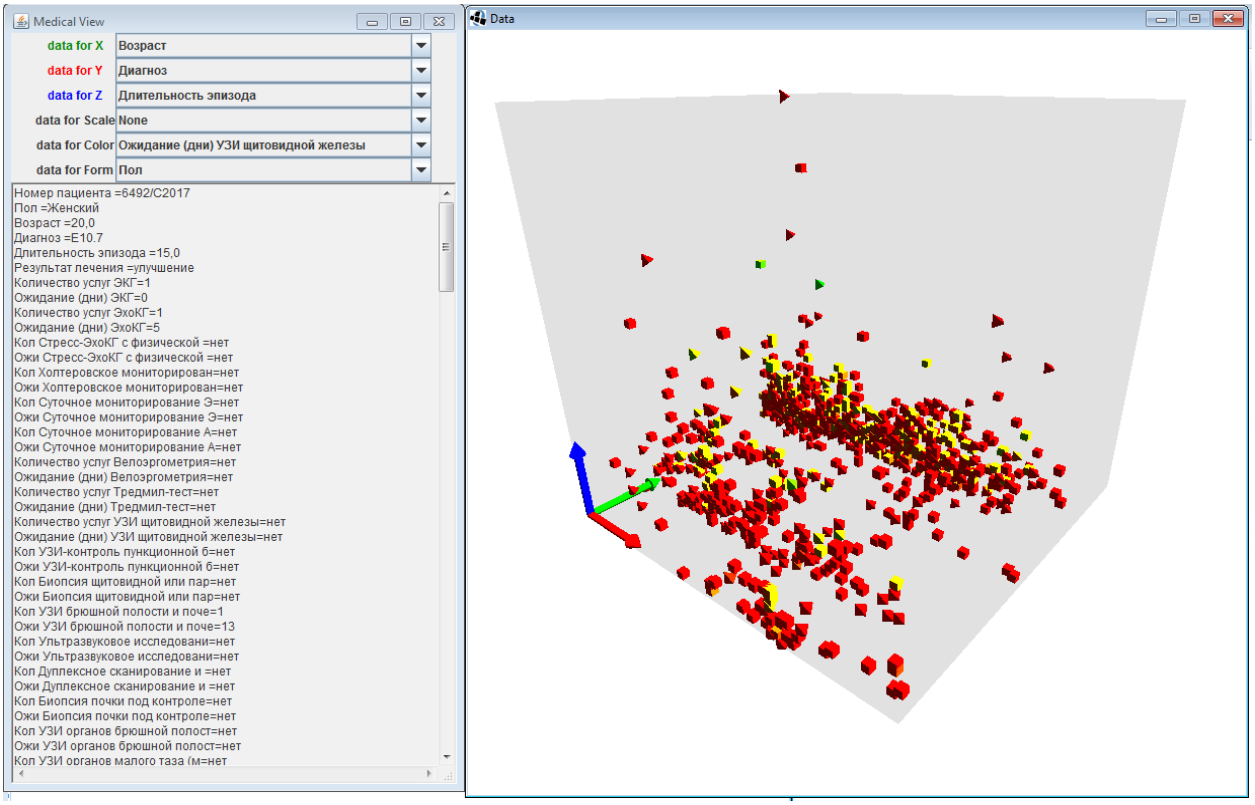


Рис. 4. Окно интерфейса разработанной системы 3D-визуализации данных МИС. Разработчики программного обеспечения В.Л. Авербух и И.О. Михайлов. Выборка пациентов с СД 1 типа. Показана возможность визуализировать в виртуальном трехмерном пространстве данные о пациентах (виртуальное пространство можно вращать, а данные по конкретному пациенту можно посмотреть отдельно, сделав выбор из множества фигур). По оси Z – длительность эпизода госпитализации, цветом показана длительность ожидания УЗИ щитовидной железы: желтый цвет соответствует среднему ожиданию УЗИ, зеленый – длительному ожиданию УЗИ; красный цвет соответствует отсутствию проведения УЗИ. Данная конфигурация опций позволяет врачу сразу охватить взглядом картину распределения пациентов по длительности госпитализации и по ожидаю УЗИ щитовидной железы, во взаимосвязи с возрастом, полом и кодами / подразделами по МКБ-10.

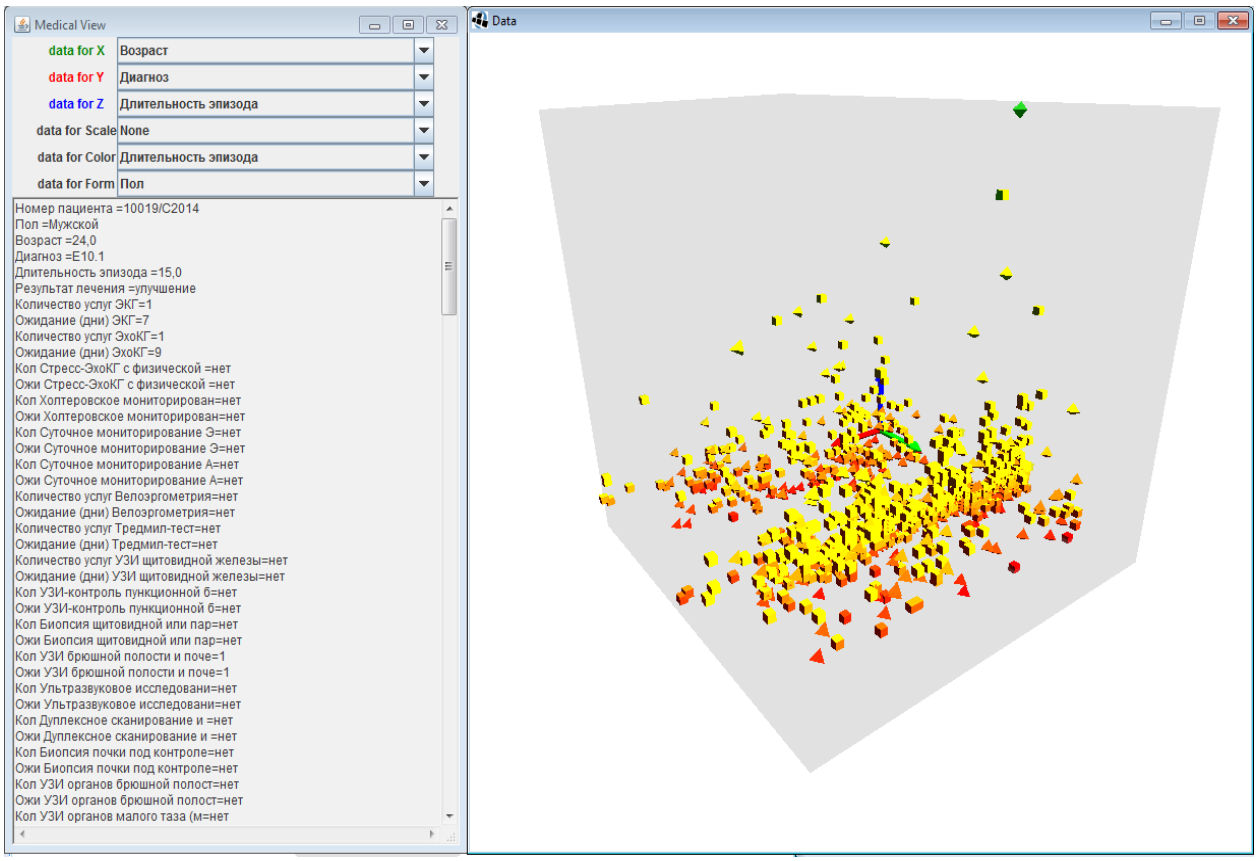


Рис. 5. Окно интерфейса разработанной системы визуализации данных МИС. Разработчики программного обеспечения В.Л. Авербух и И.О. Михайлов. Выборка пациентов с СД 1 типа. Показана возможность визуализировать в виртуальном трехмерном пространстве данные о пациентах (виртуальное пространство можно вращать, а данные по конкретному пациенту можно посмотреть отдельно, сделав выбор из множества фигур). По оси Z – длительность эпизода госпитализации, цветом также показана длительность эпизода госпитализации. Данная конфигурация опций позволяет врачу сразу охватить взглядом картину распределения пациентов по длительности госпитализации, во взаимосвязи с возрастом, полом и кодами / подразделами по МКБ-10.

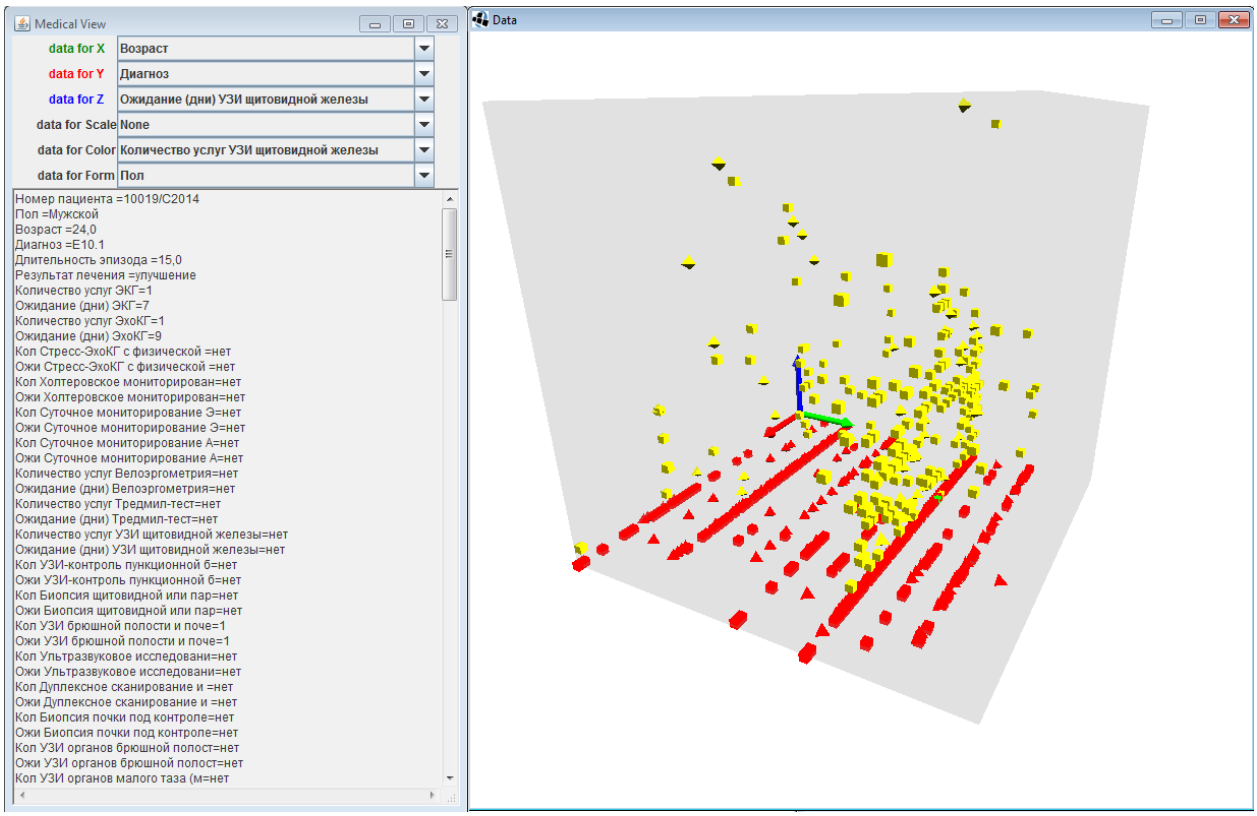
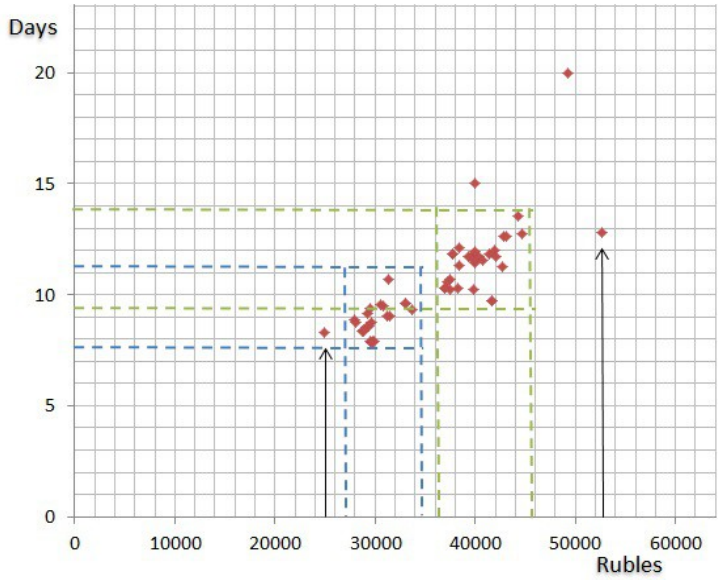
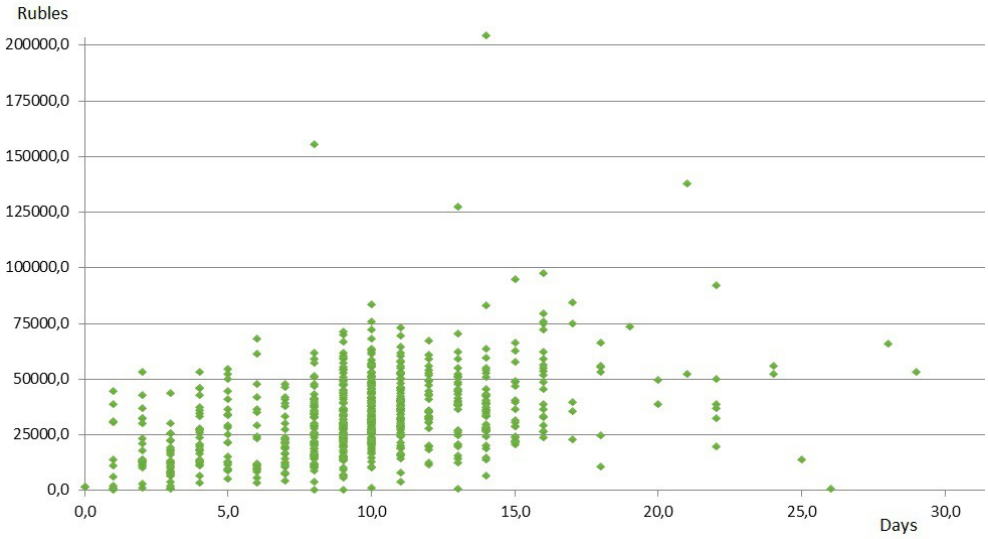


Рис. 6. Окно интерфейса разработанной системы визуализации данных МИС – авторы. Разработчики программного обеспечения В.Л. Авербух и И.О. Михайлов. Выборка пациентов с СД 1 типа. Показана возможность визуализировать в виртуальном трехмерном пространстве данные о пациентах (виртуальное пространство можно вращать, а данные по конкретному пациенту можно посмотреть отдельно, сделав выбор из множества фигур). По оси Z – длительность ожидания УЗИ щитовидной железы. Красный цвет маркера – УЗИ не проводилось, желтый цвет – проведено одно обследование. Данная конфигурация опций позволяет врачу сразу охватить взглядом картину распределения пациентов по получению и ожиданию УЗИ щитовидной железы, во взаимосвязи с возрастом, полом и кодами / подразделами по МКБ.

При ***кластерном анализе*** выборки пациентов с ***артериальной гипертензией*** были получены кластеры по трем наборам медуслуг: Min Metadata – минимальный набор (10 необходимых и наиболее часто реализуемых медуслуг), Middle Metadata – укороченный набор (24 медуслуги, соответствующие Стандарту медпомощи при первичной артериальной гипертензии, приказ Минздрава РФ от 9 ноября 2012 г. №708н), Max Metadata – максимальный набор (все услуги из 39 предоставленных). Также кластерный анализ проводился по двум направлениям: количество медуслуг – Series treatment, и время их ожидания – Series time. Для каждой отдельной выборки были построены разбиения на 2, 3, 4 кластера (обозначения: кластеризация полного набора метаданных на два кластера – Max-2, три кластера – Max-3, четыре кластера – Max-4, и т.д.). Кластеры обозначались буквой «k» и номером перед ней (например, Max-2 1k и Max-2 2k означает кластеризацию полного набора метаданных на два кластера 1k и 2k).



7.1



7.2

Рис. 7. 7.1 – Распределение кластеров по двум характеристикам (средние значения для каждого кластера) «стоимость лечения – длительность госпитализации». Стрелками указаны кластер с самой низкой стоимостью лечения и кластер с самой высокой стоимостью лечения. Кластерный анализ выполнили: Л.С. Мазелис, А.Л. Мазелис, Николаев А.Э. 7.2 – Полная выборка пациентов, простое распределение значений длительности госпитализации и стоимости лечения для каждого пациента без математической обработки.

Таблица 4

Кластеры, распределенные на две группы по характеристикам «стоимость-длительность» стационарного лечения

Группы	Кластеры
Группа 1 395 пациентов; стоимость лечения до 35 тыс. руб.; средняя длительность пребывания в стационаре 8,5 суток	**Series treatment** Max-2 1k; Middle-2 1k; Min-2 1k; Max-3 1k; Middle-3 2k; Min-3 1k; Max-4 1k; Min-4 1k
	Series time Max-2 1k; Middle-2 1k; Min-2 1k; Max-3 2k; Max-3 3k; Middle-3 1k; Middle-3 2k; Min-3 1k; Max-4 2k; Max-4 4k; Middle-4 1k; Middle-4 2k; Middle-4 3k; Min-4 3k
Группа 2 290 пациентов; стоимость лечения от 35 тыс. руб.; средняя длительность пребывания в стационаре 11,2 суток	**Series treatment** Max-2 2k; Middle-2 2k; Min-2 2k; Max-3 2k; Max-3 3k; Middle-3 1k; Middle-3 3k; Min-3 2k; Min-3 3k; Max-4 2k; Max-4 3k; Max-4 4k; Middle-4 1k; Middle-4 2k; Middle-4 4k; Min-4 2k; Min-4 4k
	Series time Max-2 2k; Middle-2 2k; Min-2 2k; Max-3 1k; Middle-3 3k; Min-3 2k; Min-3 3k; Max-4 1k; Middle-4 4k; Min-4 1k; Min-4 4k

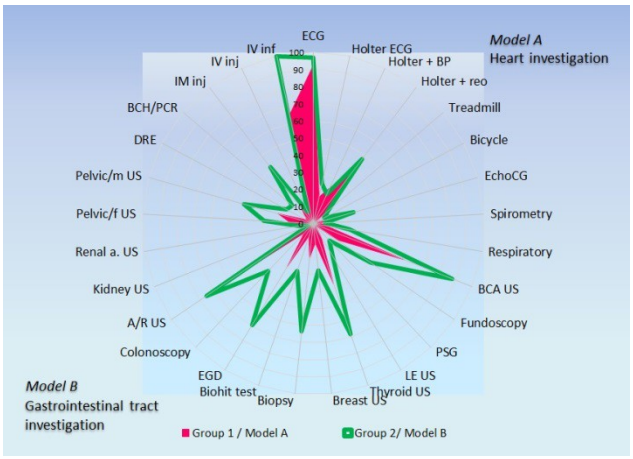
В таблице 4 представлен весь перечень полученных кластеров. Проведена визуализация всех полученных кластеров на графике по критерию «стоимость лечения – длительность госпитализации» (средние показатели в каждом отдельном кластере), см. рис. 7.1. Для сравнения на рис. 7.2 отражено простое распределение без кластеризации всех зарегистрированных случаев лечения, полная выборка 685 пациентов. При визуализации на графике кластеры распределились на ***две группы***, что стало основой для описания двух моделей стационарного ведения пациентов с диагнозом по МКБ-10 I11.9: гипертензивная (гипертоническая) болезнь с преимущественным поражением сердца без сердечной недостаточности (или гипертоническая болезнь 2 стадии).

На рис. 8.1 показаны ***две обнаруженные модели стационарного ведения пациентов с артериальной гипертензией*** – с преимущественным обследованием сердечно-сосудистой системы (Модель А, группа 1; 58% пациентов) и преимущественным обследованием ЖКТ, сопутствующих заболеваний (Модель В, группа 2; 42% пациентов).

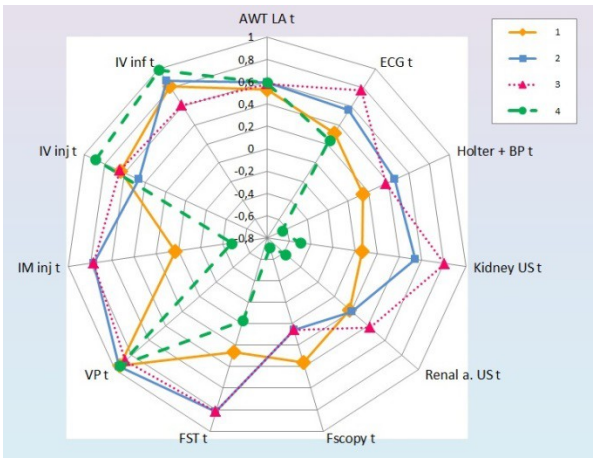
На рис. 8.2 представлена визуализация корреляционного анализа некоторых средних для каждого кластера параметров ожидания с общим временем пребывания в стационаре. Выявлены сильные корреляционные связи, указывающие на то, что более всего влияют на длительность пребывания в стационаре ожидание взятия крови на анализ и ожидание ультразвукового обследования, а также сроки проведения парентеральной терапии.

На основе кластерного анализа выявлен ***социально-медицинский портрет*** пациента (или iПациента), страдающего гипертензивной болезнью сердца: в основном такому пациенту требуется одно обследование ЭКГ; только в половине случаев проводится внутривенное капельное введение препаратов; почти равное внимание уделяется как обследованию сердца, так и обследованию ЖКТ, что оправдано с точки зрения оценки побочных эффектов антигипертензивных препаратов, а также указывает на наличие определенной направленности сопутствующих заболеваний у данной категории больных.

Анализ данных МИС показал, что 80-90% пациентов с диагнозом по МКБ-10 Код: I11.9 (артериальная гипертензия) могут быть обследованы и пролечены в условиях дневного стационара без длительной госпитализации. Для укрепления такой тактики ведения пациентов необходимо развивать телемедицинские службы круглосуточного амбулаторного консультирования [2-4].



8.1



8.2

Рис. 8. 8.1 – Визуализация двух моделей стационарного обследования. Модель А: обследование сердечно-сосудистой системы (Model A, Heart investigation). Модель Б: обследование преимущественно желудочно-кишечного тракта, сопутствующих заболеваний (Model B, gastrointestinal tract investigation). Показано число обследований и процедур, стандартизированных на 100 пациентов. Сокращения и условные обозначения см. ниже. 8.2 – Корреляционный анализ по Пирсону между средним временем пребывания в стационаре и некоторыми средними параметрами ожидания процедур и обследований. 1 – все кластеры; 2 – 8 кластеров по максимальному набору медуслуг; 3 – 8 кластеров по укороченному набору; 4 – 8 кластеров по минимальному набору.

Сокращения для рис. 8.1: ECG – электрокардиография; *Holter ECG* – суточное мониторирование ЭКГ в 12-ти отведениях; *Holter + BP* – суточное мониторирование ЭКГ и артериального давления; *Holter + reo* – комбинированное суточное мониторирование ЭКГ в 12-ти отведениях, АД и реопульмограммы; *Treadmill* – тредмил стресс-тест; *Bicycle* – велоэргометрия; *EchoCG* – стресс-эхокардиография; *Spirometry* – спирометрия с пробой с бронхолитиком; *Respiratory* – комплексное исследование легких; *BCA US* – триплексное исследование брахиоцефальных артерий на интра- и экстракраниальном уровне; *Fundoscopy* – исследование глазного дна с применением линзы Гольдмана; *PSG* – комплексное полисомнологическое обследование; *LE US* – дуплексное исследование артерий и вен нижних конечностей, дуплексное исследование вен нижних конечностей с функциональными пробами; *Thyroid US* – УЗИ щитовидной железы; *Breast US* – УЗИ молочных желез; *Biopsy* – биопсия при эндоскопии; *Biohit* – тест Biohit; *EGD* – эзофагогастродуоденоскопия; *Colonoscopy* – колоноскопия; *A/R US* – УЗИ органов брюшной полости; *Kidney US* – УЗИ почек; *Renal a. US* – УЗИ почечной артерии; *Pelvic/f US* – УЗИ малого таза у женщин; *Pelvic/m US* – УЗИ малого таза у мужчин; *DRE* – ректальное пальцевое исследование предстательной железы; *BCH/PCR* – взятие материала на бактериологическое, цитологическое, гормональное или ПЦР-исследование; *IM inj* – инъекция внутримышечная; *IV inj* – инъекция внутривенная; *IV inf* – внутривенное капельное введение лекарственных средств.

Сокращения для рис. 8.2: AWT LA t – среднее ожидание последнего анализа; *ECG t* – ожидание электрокардиографии; *Holter + BP t* – ожидание суточного мониторирования ЭКГ и артериального давления; *Kidney US t* – ожидание УЗИ почек; *Renal a. US t* – ожидание УЗИ почечной артерии; *Fscopy t* – ожидание исследования глазного дна с применением линзы Гольдмана; *FST t* – ожидание забора крови из пальца; *VP t* – ожидание забора крови из вены; *IM inj t* – ожидание внутримышечной инъекции; *IV inj t* – ожидание внутривенной инъекции; *IV inf t* – ожидание внутривенного капельного введения лекарственных средств.

Выводы

1. Необходимо развивать аналитику Больших данных МИС с целью формирования принципиально нового механизма контроля и анализа лечебно-диагностического процесса, с быстрой адаптацией к изменениям в потоке пациентов. Анализ данных МИС открывает возможности для более широкого изучения отдельных нозологических категорий пациентов, с учетом их вовлеченности в лечебно-диагностический процесс в целом.

2. Информационная «цифровая» концепция «iПациент» (intranet-пациент, отраженный в записях МИС) позволяет анализировать записи МИС с точки зрения социально-медицинской обратной связи и вносить изменения в стандарты лечения в соответствии с обнаруженными реальными требованиями пациентов, формируя гибкие социально-медицинские стандарты «aaS» (as-a-Service, стандарты как услуга).

Список литературы

[1]. Мерабишвили В.М., Дятченко О.Т. Статистика рака легкого: заболеваемость, смертность, выживаемость. Практическая онкология. 2000, 3: 3-7.

[2]. Frolio L. Big Data Insights in Healthcare. Dell EMC Global Services Blog «In Focus». Hopkinton MA, USA, 2015.

[3]. Reyss A., Balandin S. Healthcare, medical support and consultancy applications and services for mobile devices, in Proc. IEEE SIBIRCON Conf., 2010, pp. 300-305.

[4]. Schmarzo B. Big Data MBA: Driving Business Strategies with Data Science. USA: John Wiley & Sons. Inc., 2016.

Big Data analytics of Medical Information System (MIS) records

KOLESNICHENKO Olga, Editor in Chief of Security Analysis Bulletin, Candidate of Medical Sciences, Moscow;
E-mail: oykolesnichenko@list.ru

KOLESNICHENKO Yuriy, Editor in Chief of Uzgraph; Chief Technical Officer of Cybersecurity, Security Analysis Bulletin, Moscow;
E-mail: green-apple_2000@mtu-net.ru

MINUSHKINA Larisa, Doctor of medicine, Professor of the Department of Therapy, Cardiology and Functional Diagnostics with the course of Nephrology, Central State Medical Academy at the Department of Presidential Affairs of the Russian Federation, Moscow;
E-mail: Minushkina@mail.ru

MAZELIS Lev, Doctor of Economic Sciences, Professor, Head of the Department of Mathematics and Modeling, Vladivostok State University of Economics and Service, Vladivostok;
E-mail: Lev.Mazelis@vvsu.ru

MAZELIS Andrey, Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Mathematics and Modeling, Vladivostok State University of Economics and Service, Vladivostok;
E-mail: Andrey.Mazelis@vvsu.ru

NIKOLAEV Alexander, postgraduate, Department of Mathematics and Modeling, Vladivostok State University of Economics and Service, Vladivostok;
E-mail: sa10121992@yandex.ru

SHAHGELDYAN Carina, Doctor of Technical Sciences, Professor, Director of the Institute of Information Technologies of Vladivostok State University of Economics and Service; Head of the Laboratory for Big Data Analytics in Health Care and Biomedicine, School of Biomedicine, Far Eastern Federal University, Vladivostok;
E-mail: carina.shahgeldyan@vvsu.ru

AVERBUKH Vladimir, Candidate of Technical Sciences, Senior Research Fellow, Head of the Computer Visualization Department of the System Software Division of the N.N. Krasovskii Institute of Mathematics and Mechanics of the Urals Branch of the Russian Academy of Sciences; Associate Professor of Ural Federal University, Yekaterinburg;
E-mail: Averbukh@imm.uran.ru

MIKHAYLOV Igor, Assistant Professor of Ural Federal University, Yekaterinburg;
E-mail: igormich88@gmail.com

MARTYNOV Alexander, Chief Executive Officer, SP.ARM, St. Petersburg;
E-mail: Martynov@sparm.com

PULIT Valeriy, Candidate of Chemical Sciences, Senior systems analyst, SP.ARM, St. Petersburg;
E-mail: Valeriy.Pulit@sparm.com

DOLZHENKOV Anatoliy, Science Director, SP.ARM, St. Petersburg;
E-mail: dol@sparm.com

GRIGOREVSKY Ivan, Candidate of Technical Sciences, Deputy Chairman of the Organizing Committee of National Supercomputing Forum, National Supercomputing Technology Platform, Aylamazyan Program Systems Institute of Russian Academy of Sciences, Pereslavl Zalessky;
E-mail: gin@nscf.ru

SMORODIN Gennady, Candidate of Technical Sciences, MBA, Associate Professor of Saint Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University), St. Petersburg; Russia
E-mail: gsmorodin@gmail.com

Abstract. The Study «Big Data analytics in medicine» has been provided by National Supercomputing Forum / National Supercomputing Technology Platform (Russia). The goal of Study is improvement of care management process in a hospital using Big Data analytics of Medical Information System records. The results of Big Data analytics of Medical information system qMS records are presented. The records of MIS qMS collected during period from 2013 to 2017, from Russian hospitals. Patients have Diabetes Mellitus Type 1 and Arterial Hypertension. Patients' data includes: ICD-10 clinical diagnoses, records about implemented investigation procedures, operations, pharmacological treatment. Various mathematical analysis was implemented: Cluster analysis in Python; Gray reflected binary code (Boolean algebra) in Java; Graph analysis; 3D-visualization of multi-degree data. This Study will continue with expansion of set of Big Data and analysis methods.

Key Words and Phrases: Big Data, Medical Information System, qMS, Cluster analysis, Gray code, 3D-visualization, Augmented Reality, Graph analysis, Diabetes mellitus Type 1, Arterial Hypertension.