

Формирование набора больших данных для клинических исследований на примере аневризм сосудов головного мозга

Ю.В. Кивелёв^{1,2}, И. Сааренпя¹, А.Л. Кривошапкин^{2,3,4}

¹ Университет Турку

20520, Финляндия, Хямеентие, 11

² Европейский медицинский центр

129090, г. Москва, ул. Щепкина, 25

³ Российский университет дружбы народов

117198, г. Москва, ул. Миклухо-Маклая, 6

⁴ НМИЦ имени академика Е.Н. Мешалкина Минздрава России

630055, г. Новосибирск, ул. Речкуновская, 15

Резюме

Изменчивость и неоднородность цифровой медицинской информации требует разработки современных алгоритмов по структурированию массивов данных с целью их дальнейшей статистической обработки. Цель исследования – определить ход работы по созданию набора данных (НД) при исследовании аневризм сосудов головного мозга от этапа формирования технического задания до получения финального НД. **Материал и методы.** Процесс создания, обработки и анализа НД пациентов с аневризмами проводился на базе университетской клиники города Турку, Финляндия. В течение последних 20 лет клиника осуществляет медицинский документооборот в цифровом формате, что позволило создать на ее базе отдел хранения цифровых данных с целью максимального сохранения любой доступной цифровой информации. Автоматизированное получение данных пациентов проводилось дата-инженером с использованием языка программирования «R» на основании кодов Международной классификации болезней (МКБ-10). **Результаты и их обсуждение.** В период с января 2000 г. по май 2018 г. в ходе первичного получения данных выявлено 3850 пациентов. После независимой перекрестной проверки электронных историй болезни отсеяно 1218 (32 %) ложноположительных случаев. Данные по оставшимся пациентам были разделены на клинический и реанимационный блоки. Каждое событие, относящееся к конкретной временной дате в НД, определено как инфо-единица. Вся информация в обоих блоках структурирована в формате Excel и представлена в хронологическом порядке для каждого отдельного больного. В целом весь набор данных состоял из более чем 70 000 000 рядов инфо-единиц, выявленных у 2632 пациентов. **Заключение.** Автоматизированный поиск данных позволил создать многокомпонентный структурированный набор данных пациентов с аневризмами сосудов головного мозга. Выработанный алгоритм автоматизированного получения данных имел ограничение в отношении ложнопозитивных случаев, выявленных в 32 % случаев. Таким образом, анализ клинического материала, полученного с помощью цифровых алгоритмов, требует тщательной перекрестной проверки членами исследовательской группы.

Ключевые слова: цифровизация, медицинские данные, набор данных, автоматизированное получение данных, перекрестная проверка данных.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Автор для переписки: Кривошапкин А.Л., e-mail: alkr01@yandex.ru

Для цитирования: Кивелёв Ю.В., Сааренпя И., Кривошапкин А.Л. Формирование набора больших данных для клинических исследований на примере аневризм сосудов головного мозга. *Сибирский научный медицинский журнал.* 2023;43(3):86–94. doi: 10.18699/SSMJ20230311

Establishing of big data clinical dataset in brain vessel aneurysm research

Ju.V. Kivelev^{1,2}, I. Saarenpää¹, A.L. Krivoschapkin^{2,3,4}

¹ Turku University Hospital
20520, Finland, Hämeentie, 11

² European Medical Center
129090, Moscow, Shchepkina str., 25

³ Peoples' Friendship University of Russia (RUDN University)
117198, Moscow, Miklukho-Maklaya str., 6

⁴ Meshalkin National Medical Research Center of Minzdrav of Russia
630055, Novosibirsk, Rechkunovskaya str., 15

Abstract

Variability and heterogeneity of digital medical data requires establishing of modern algorithms which provide appropriate data processing. The aim of the study was to delineate the main steps in formation of a clinical dataset of patients with brain aneurysms from the stage of producing primary mining specifications to formation of a final version. **Material and methods.** Data collection, crosschecking of the cases and analyses of dataset has been carried out in Turku University Hospital. Within last two decades available medical data at our hospital have been stored in digital data lake thus allowing automatized data mining. In frame of our study, data mining was performed by a data scientist utilizing R software. Inclusion criteria were based on a set of diagnosis which were coded in medical charts according to international classification of diseases (ICD 10). **Results and Discussion.** Primary data mining identified 3850 patients with brain aneurysms treated at our hospital from January 2000 till May 2018. After independent manual crosschecking of medical charts of these patients, we found 1218 (32 %) cases, which had no aneurysm (false-positive). Data of remaining true aneurysm-cases were divided into clinical and intensive care unit subsets where every event linked to particular date of treatment was defined as an info-unit. All the data in both subsets were structured into separate Excel files and presented in chronological order for each particular patient. Altogether, dataset included 70 000 000 rows of info-units found in 2632 patients. **Conclusions.** Data mining allowed establishment of detailed clinical dataset of patients with brain aneurysms. Produced mining algorithm had limitation regarding false-positive cases (32 % patients). Based on that, we recommend manual crosschecking of automatically collected dataset before statistical analysis.

Key words: digitalization, medical data, dataset, mining, crosschecking.

Conflict of interest. The authors declare no conflict of interest.

Correspondence author: Krivoschapkin A.L., e-mail: alkr01@yandex.ru

Citation: Kivelev Ju.V., Saarenpää I., Krivoschapkin A.L. Establishing of big data clinical dataset in brain vessel aneurysm research. *Sibirskij nauchnyj medicinskij zhurnal* = *Siberian Scientific Medical Journal*. 2023;43(3):86–94. [In Russian]. doi: 10.18699/SSMJ20230311

Введение

За последние два десятилетия объем цифровых медицинских данных, доступных для научного анализа, стремительно вырос [1]. В современном мире медицина и связанные с ней вспомогательные службы являются одним из главных источников больших данных (БД). Так, около 30 % от всего объема цифровых данных человечества приходится именно на медицинский сектор, составляя 10^{24} байт [2]. Согласно литературным данным, объем цифровой информации, который ежегодно приходится на одного пациента, в среднем составляет 80 мегабайт [3]. Вместе

с тем большие объемы данных *per se* не представляют технических трудностей для обработки, учитывая постоянно нарастающие современные вычислительные мощности компьютеров. Основная трудность в анализе и интерпретации БД в медицине кроется в их ярко выраженной изменчивости и неоднородности [4]. Медицинские БД включают демографические, диагностические, лабораторные, морфологические показатели, данные проведенных лечебных мероприятий и процедур, лекарственной терапии, регистрационные и клинические записи в медицинской документации, операционно-финансовые отчетности, страховые и организационно-хозяйствен-

ные характеристики. Такая ярко выраженная неоднородность информации в медицине, а также наличие качественной и количественной изменчивости при вводе данных медицинским персоналом и регистраторами в электронные системы болезней существенно отличает медицинские БД от многих других сфер деятельности человека.

На данный момент основным источником БД в медицине является область лучевой диагностики, включая магнитно-резонансную (МРТ) и компьютерную томографию (КТ), рентгеновские обследования и т.п. Выраженный рост количества данных демонстрируют также так называемые омиксные биомедицинские технологии, включая геномику, транскриптомику, протеомику, метабономику и т.п. Вместе с тем возможность сбора БД в клинической медицине дала толчок к внедрению новых технологий искусственного интеллекта и машинного обучения при научном анализе и решении конкретных клинических задач.

В нейрохирургической практике одной из потенциальных точек приложения технологии обработки БД и применения искусственного интеллекта и машинного обучения является проблематика ведения пациентов с аневризмами сосудов головного мозга (АГ). Неразорвавшиеся АГ встречаются у 3 % населения и чаще всего выявляются случайно на МРТ или КТ [5, 6]. При вынесении показаний к профилактическому оперативному лечению в случае бессимптомных АГ необходимо взвесить риски естественного течения самого заболевания и риски хирургического вмешательства (как микрохирургического, так и эндоваскулярного). Большую роль при этом играют такие индивидуальные особенности, как состояние соматического статуса пациентов, наличие вредных привычек и других сопутствующих факторов. При выборе в пользу динамического наблюдения для своевременного выявления изменений размеров АГ обычно проводится серия контрольных снимков на протяжении нескольких лет, однако эффективность такой стратегии при небольших АГ (3 мм и менее) не является доказанной [7], хотя именно они встречаются на практике чаще всего. Исходя из этого, при обсуждении стратегии лечения лиц с неразорвавшимися АГ необходимо применять строго индивидуальный подход, не только опираясь на литературные данные, но и учитывая максимально широкий спектр доступных медицинских характеристик пациента в каждом конкретном случае. В этом контексте появление в арсенале клинициста возможности использовать автоматизированные алгоритмы и модели, построенные на индивидуальных реальных данных пациентов с АГ, могут помочь оптимизировать план ведения этой группы пациентов.

Данная работа посвящена описанию основных шагов на пути формирования набора клинических данных, способов его структурирования, проверки и подготовки к дальнейшему анализу на примере группы пациентов с аневризмами головного мозга.

Материал и методы

Работа по созданию, верификации, обработке и анализу набора данных (НД) пациентов с АГ выполнена на базе университетской клиники города Турку, Финляндия. Данная клиника является третичным центром оказания медицинской помощи юго-западного региона страны с населением около 1 млн человек. Все пациенты с нейрохирургической патологией, встречающейся в указанном регионе, поступают в нейроцентр университетской клиники. В течение последних 20 лет клиника осуществляет медицинский документооборот в цифровом формате, что позволило создать на базе клиники отдел хранения цифровых данных (ХЦД) с целью максимального сохранения любой доступной цифровой информации. Научное взаимодействие с отделом ХЦД как отдельной административной единицей осуществлялось в несколько этапов: формирование исследовательского проекта, научной группы и подготовка разрешительной документации; создание технического задания с перечнем критериев поиска нужных случаев для внесения в НД, образование совместной рабочей группы, включающей исследователя-клинициста и инженера, осуществляющего автоматизированное получение данных (АПД) и получение данных из отдела ХЦД; формирование НД, его структурирование, перекрестная проверка и внесение необходимых корректировок и поправок.

Получение НД пациентов с АГ проводилось инженером с использованием языка программирования «R» на основании следующих кодов Международной классификации болезней (МКБ-10) I67.1 (аневризма мозга без разрыва), I60.0-I60.7 (субарахноидальное кровоизлияние (разных локализаций), I60.9 (нетравматическое субарахноидальное кровоизлияние), I69.0 (последствия цереброваскулярных болезней), I69.1 (последствия внутричерепного кровоизлияния), Z82.3 (инсульт в семейном анамнезе).

В исследование включены все пациенты, обследованные и/или получавшие лечение в университетской клинике города Турку с января 2000 г. по май 2018 г., данные которых были доступны в отделе ХЦД. Учитывая, что кодирование диагнозов до 2003 г. проводилось по системе МКБ-9, АПД данных по пациентам с 2000 по 2003 г. про-

водился с автоматической перекодировкой диагнозов и приведением их к стандартам МКБ-10.

Результаты

За указанный период найдено 3850 пациентов, удовлетворяющих согласованным критериям АПД. После перекрестной проверки электронных историй болезни тремя сотрудниками исследовательской группы независимо друг от друга (на каждого было распределено по 1300 пациентов в случайном порядке) обнаружено 1218 (32 %) ложноположительных случаев. Так, например, в медицинской информационной системе (МИС) выявлены опечатки при введении кода диагноза по МКБ; подозрение на аневризму и/или аневризматическое субарахноидальное кровоизлияние указано как окончательный диагноз, однако позже опровергнуто уточняющими обследованиями; введен ошибочный код диагноза при диагностировании аневризмы внутри артериовенозной мальформации, которая в данном случае была не относящейся к изучаемой теме нозологической единицей и т.п.

После удаления ложноположительных случаев в окончательной версии осталось 2632 пациента с диагностированной неразорвавшейся и/или разорвавшейся аневризмой головного мозга. Путем дальнейшего процесса получения данных собрана информация обо всех медицинских контактах, лечебных мероприятиях, любых диагностических и/или лабораторных обследованиях за указанный период, вне зависимости от того, когда в этом промежутке времени была диагностирована аневризма. Например, пациент, поступивший в 2015 г. по поводу разрыва аневризмы и субарахноидального кровоизлияния, был включен в НД таким образом, что все его медицинские данные, предыдущие за 15 лет, а также последующие данные до мая 2018 г. были автоматически внесены в НД и структурированы в хронологическом порядке. Благодаря этой методике мы смогли максимально охватить всю доступную медицинскую историю пациентов на протяжении 18 лет. При этом каждое событие, относящееся к конкретной временной дате в НД, мы определили как инфо-единицу, в качестве которой мы принимали и дневниковую запись лечащего врача, и конкретный числовой показатель, если в обоих случаях события имели конкретную привязку к дате загрузки в МИС (обычно с точностью до минуты). Введение данных в виде инфо-единиц позволило объединить в один ряд текстовые и числовые показатели, которые могли быть упорядочены в хронологическом порядке. Далее данные были

структурированы в виде двух блоков, клинического и реанимационного.

В клинический блок вошли демографическая информация; список всех зафиксированных диагнозов; перечень подразделений больницы и даты всех контактов с ней; список проведенных инструментально-диагностических мероприятий лабораторной, лучевой и патоморфологической диагностики с отображением всех результатов обследования и заключений диагностов в текстовом формате; вмешательства, выполненные в условиях операционной, с данными кодов, названий и времени проведенных операций, основными параметрами анестезиологического пособия; проведенная в стационаре лекарственная терапия с указанием названий и дозировок препаратов, записями базового мониторингирования состояния пациента средним медицинским персоналом и дневниковые и выписные записи лечащих врачей в текстовом формате. Для удобства интерпретации данные в этом блоке в документе Excel были распределены по отдельным закладкам (рис. 1, 2). Общее количество рядов инфо-единиц в клиническом блоке составило чуть более 4 000 000.

Реанимационный блок включал три подраздела: 1) текстовые записи мониторингирования состояния пациента средним медицинским персоналом в текстовом формате (около 2 100 000 рядов инфо-единиц); 2) числовые данные инструментального мониторингирования (артериальное давление, внутричерепное давление, параметры респираторной поддержки и т.п., около 38 000 000 рядов инфо-единиц), в условиях реанимационного отделения данные автоматически вносились в МИС каждые две минуты, независимо от продолжительности мониторингирования; 3) назначения лечащего врача и лекарственная терапия (текст и/или цифры, около 30 000 000 рядов инфо-единиц).

Вся информация в обоих блоках структурирована в формате Excel и представлена в хронологическом порядке для каждого отдельного пациента. В целом весь НД по 2632 пациентам состоял из более чем 70 000 000 рядов инфо-единиц.

Обсуждение

В настоящей работе мы описали схему создания НД по типу БД на примере пациентов с аневризмами сосудов головного мозга. Наличие в университетской клинике города Турку отдела по сбору и хранению цифровой информации, создание рабочей группы, включающей клиницистов и инженеров, разработка алгоритмов АПД по четко обозначенным критериям поиска позволили создать многокомпонентный структурированный НД, масштаб и глубина которого (70 000 000 рядов инфо-единиц)

patient_ID	sex	diagnosis_date	age	diagnosis_specification	reception_unit_name	area_of_responsibility	speciality_code	speciality_specific
1	1 female	2007.01.12 13:53:00	70	Z03.9 undefined disease	SKIR	SKIR	200	orthopedics
2	1 female	2007.02.09 13:16:00	70	M51.1 NIKAMAVÄLILEVYSRAUS JA HERMOJUUL	SKIR	SKIR	200	orthopedics
3	1 female	2007.04.25 10:35:01	70	R19.8 MUU RUUANSULATUSELINTEN JA VATSA	227 gastroenterology	GAS	10G	gastroenterology
4	1 female	2007.07.16 09:51:10	70	A04.7 CLOSTRIDIUM DIFFICILEN AIHEUTTAMA	C227 gastroenterology	GAS	10G	gastroenterology
5	1 female	2008.03.05 13:06:42	71	Z01.8 MUU ERITYISTUTKIMUS	SSIS internal medicine outpatient clinic	SSIS	10	internal medicine
6	1 female	2008.12.08 11:01:00	71	L40.0 psoriasis	620 dermatology and venereology	IHO	60	dermatology
7	1 female	2009.02.13 12:19:00	72	L40.0 psoriasis	620 dermatology and venereology	IHO	60	dermatology
8	1 female	2009.09.28 06:52:00	72	R04.0 NENÄVERENVUOTO	525 otology outpatient clinic	KOR	55	otorhinolaryngolog
9	1 female	2009.09.29 12:45:00	72	R04.0 NENÄVERENVUOTO	515 otology	KOR	55	otorhinolaryngolog
10	1 female	2011.11.15 10:31:09	74	Z72.0 Tupakan käyttö	220 emergency room	EMS	10	internal medicine
11	1 female	2011.11.15 10:31:09	74	Z20.09 Epästabiili angina määrittämätön sepelv	220 emergency room	EMS	10	internal medicine
12	1 female	2011.11.15 10:31:09	74	L40.0 psoriasis	220 emergency room	EMS	10	internal medicine
13	1 female	2011.11.15 10:31:09	74	J45.9 Määrittämätön astma	220 emergency room	EMS	10	internal medicine
14	1 female	2011.11.17 15:00:00	74	L40.0 psoriasis	011 internal medicine	YSIS	10	internal medicine
15	1 female	2011.11.17 15:00:00	74	J45.9 Määrittämätön astma	011 internal medicine	YSIS	10	internal medicine
16	1 female	2011.11.17 15:00:00	74	J20.00 Epästabiili angina ei osoitettavaa sepelv	011 internal medicine	YSIS	10	internal medicine
17	1 female	2011.11.17 15:00:00	74	J45.9 Allerginen astma	011 internal medicine	YSIS	10	internal medicine
18	1 female	2011.11.17 15:00:00	74	I10 Essentiaalinen (primaarinen) verenpainet	011 internal medicine	YSIS	10	internal medicine
19	1 female	2011.12.11 09:58:41	74	I10 Essentiaalinen (primaarinen) verenpainet	220 emergency room	EMS	10	internal medicine
20	1 female	2011.12.11 09:58:41	74	L40.9 Määrittämätön psoriaasi	220 emergency room	EMS	10	internal medicine
21	1 female	2011.12.11 09:58:41	74	R07.4 Määrittämätön rintakipu	220 emergency room	EMS	10	internal medicine
22	1 female	2011.12.11 09:58:41	74	I25.9 Määrittämätön iskeeminen sydänsairaus	220 emergency room	EMS	10	internal medicine
23	1 female	2011.12.11 09:58:41	74	J45.9 Määrittämätön astma	220 emergency room	EMS	10	internal medicine
24	1 female	2011.12.15 13:07:00	75	R07.4 Määrittämätön rintakipu	011 internal medicine	YSIS	10	internal medicine
25	1 female	2011.12.15 13:07:00	75	I25.9 Määrittämätön iskeeminen sydänsairaus	011 internal medicine	YSIS	10	internal medicine

Рис. 1. Пример из клинического блока НД. Пациентке присвоен идентификационный номер "1". Первый медицинский контакт зафиксирован в 2007 г. по поводу ортопедического заболевания. Все последующие контакты записаны в хронологическом порядке. Нижний ряд закладок содержит всю доступную медицинскую информацию, структурированную также в хронологическом порядке. Она включает все данные по инструментальным исследованиям, проведенному оперативному и лекарственному лечению и т.п.

Fig 1. An example from clinical subset. A patient encrypted with ID 1. Her first medical contact was in 2007 due to orthopedic disease. All further medical contacts are presented in chronological order. The row of sheets contains all available medical data structured chronologically. They include data on imaging, surgical and/or medical treatment etc.

A	B	C	D	E	F	G	H	I	J	K
83	1 female	2015.11.23 08:19:55	78 I49.3		Ennenalkainen kammiodepolarisaatio	NEU	general neurology	NEU	77Y	general neurology
84	1 female	2015.11.23 08:19:55	78 I25.9		Määrättämätön iskeeminen sydänsairaus	NEU	general neurology	NEU	77Y	general neurology
85	1 female	2015.11.23 08:19:55	78 I48		Eteisvärinä tai eteislepatus	NEU	general neurology	NEU	77Y	general neurology
86	1 female	2015.11.23 08:19:55	78 I10		Essentiaalinen (primaarinen) verenpainet	NEU	general neurology	NEU	77Y	general neurology
87	1 female	2015.11.23 08:19:55	78 I44.8		Pitkäaikainen aiheuttava keuhkosairaus i	NEU	general neurology	NEU	77Y	general neurology
88	1 female	2015.11.23 08:20:13	78 I67.1		Repeytymätön aivoaneurysma	NEU	general neurology	NEU	77Y	general neurology
89	1 female	2015.11.23 08:20:30	78 G46.7*167		Muu lakunaarinen oireyhtymä	NEU	general neurology	NEU	77Y	general neurology
90	1 female	2015.01.26 12:42:48	79 K21.9		Ruokatorven refluksisairaus ilman ruokat	ÄSIS	internal medicine outpatient clinic	ÄSIS	10G	gastroenterology
91	1 female	2015.01.26 12:42:57	79 K44.9		Palleatyrä ilman kureutumista tai kuoliot	ÄSIS	internal medicine outpatient clinic	ÄSIS	10G	gastroenterology
92	1 female	2015.02.02 12:49:44	79 K57.3		Paksusuolen umpipussitauti ilman suoler	ÄSIS	internal medicine outpatient clinic	ÄSIS	10G	gastroenterology
93	1 female	2015.02.12 12:38:12	79 L40.0		psoriasis	IHO	dermatology	IHO	60Y1	Ihotaudit
94	1 female	2015.02.12 12:38:12	79 L30.9		Määrättämätön dermatiitti	IHO	dermatology	IHO	60Y1	Ihotaudit
95	1 female	2015.02.25 14:55:45	79 K44.9		Palleatyrä ilman kureutumista tai kuoliot	ÄSIS	internal medicine outpatient clinic	ÄSIS	10G	gastroenterology
96	1 female	2015.02.25 14:55:45	79 K21.9		Ruokatorven refluksisairaus ilman ruokat	ÄSIS	internal medicine outpatient clinic	ÄSIS	10G	gastroenterology
97	1 female	2015.09.20 08:15:50	79 M43.1		Nikamansiirtymä (spondylolisteesi)	SELKA	spinal care	SELKA	20C	spinal care
98	1 female	2015.09.20 08:16:06	79 M48.0		Selkäranganavan ahtaus	SELKA	spinal care	SELKA	20C	spinal care
99	1 female	2017.02.03 00:41:35	80 R10.4		Muu tai määrättämätön vatsakipu	ECU	emergency room	ECU	99PN	emergency
100	1 female	2017.09.22 13:14:46	80 I48		Eteisvärinä tai eteislepatus	COR	cardiac center	COR	10C	heart center
101	1 female	2017.09.30 12:49:59	80 I48		Eteisvärinä tai eteislepatus	ECU	emergency room	ECU	99PT3	emergency
102	1 female	2017.10.11 14:04:09	80 I48		Eteisvärinä tai eteislepatus	COR	cardiac center	COR	10C	heart center
103	1 female	2017.11.02 12:30:06	80 I48		Eteisvärinä tai eteislepatus	COR	cardiac center	COR	10C	heart center
104	1 female	2017.11.22 22:38:14	80 I48		Eteisvärinä tai eteislepatus	ECU	emergency room	ECU	99PT3	emergency
105	1 female	2017.12.18 11:11:52	81 I48		Eteisvärinä tai eteislepatus	COR	cardiac center	COR	10C3	heart center
106	1 female	2018.04.09 14:20:35	81 I25.1		Ateroskleroottinen sydänsairaus	COR	cardiac center	COR	10C	heart center
107	2 male	2004.12.25 12:31:00	61 S72.4		REISILUUN ALAOSAN MURTUMA	220	emergency room	EMS	20Z	Traumatology
patients diagnosis diagnosis_old service interventions radiology drugs_given drugs prescription oncology pathological_diag pathological_main pathological_sheet pathological_exams										

Рис. 2. Пациентка под номером «1». Последняя доступная медицинская запись в 2018 г. (выделена серым, строка 106). С 2007 по 2018 г. больная имела 106 медицинских контактов. Красным отмечена инфо-единица, которой соответствует момент диагноза неразрывающейся аневризмы (МКБ 10 код I67.1).

Данная инфо-единица датирована 23.11.2015

Fig. 2. Patient 1. The last available contact is dated to 2018 (colored with grey, line 106). From 2007 to 2018 the patient had 106 medical events. A bolded red row shows an event when an unruptured aneurysm has been diagnosed (ICD 10 code I67.1). This event happened at 23.11.2015

фактически несопоставимы с НД, созданными вручную. НД таких размеров, очевидно, открывает новые возможности для научного анализа в клинических исследованиях, однако сам процесс создания НД требует новых подходов к решению задач по структурированию полученной информации. Наши результаты подтверждают, что именно четкое и структурированное техническое задание, предоставляемое клиницистами исследовательской группы инженеру, является важнейшей частью процесса формирования качественного НД для дальнейшего анализа. Это связано в первую очередь с тем, что цифровая информация в отделе ХЦД находится в неупорядоченном виде и характеризуется крайней степенью неоднородности. Не имея четко обозначенных критериев поиска и схемы структурирования конечного продукта АПД, клиницист может получить НД, практически непригодный для дальнейшего научного анализа. Наш опыт говорит о том, что требуется как минимум 3–4 совместных заседаний рабочей группы, в ходе которых следует систематически прорабатывать основные нюансы технического задания, внося максимальное количество корректировок для улучшения качества окончательного продукта. При этом необходимо создавать промежуточные пробные версии НД и проводить их глубокий анализ, который, как правило, позволяет быстро выявить технические недочеты АПД с клинической точки зрения.

Мы также убедились, что и после создания инженером финальной с технической точки зрения версии НД клиницистам исследовательской группы следует провести перекрестную проверку историй болезни вручную. Так, в нашем случае 32 % пациентов, внесенных в НД с помощью АПД, не имели клинически подтвержденного диагноза аневризмы головного мозга. Причиной ложноположительных результатов являлось чаще всего то обстоятельство, что алгоритм АПД фиксировал любое упоминание кода диагноза аневризмы в ХЦД и автоматически вносил пациента в наш НД, включая и те случаи, когда диагноз аневризмы указывался предварительно и позже не подтверждался или, например, был результатом опечатки при оформлении пациента регистратором. Подобные случаи «человеческого фактора» являются неотъемлемой частью реальной клинической работы, особенно в медицинских организациях с большими объемами пациентов, и именно это обстоятельство требует тщательной перепроверки НД, созданных автоматизированными алгоритмами. Только после проверки данных вручную и удаления неправильно внесенных случаев исследовательская группа может переходить к дальнейшему этапу работы с НД.

Клинический НД для каждого пациента может быть собран в единые цифровые образования и поля, которые фактически и определяют расширенную картину статуса его здоровья. Наличие индивидуализированных наборов информации обуславливает необходимость организации цифрового пространства в пределах как отдельной медицинской организации, так и на более высоких уровнях управления медициной. Исходя из реальной клинической практики, по степени цифровизации можно выделить четыре уровня обработки крупных массивов медицинских данных [8–10].

Уровень 1. Данные заносятся и хранятся только в бумажном виде. На этом уровне продуктивный и эффективный анализ медицинской информации крайне ограничен и фактически не позволяет применять современные методы обработки данных.

Уровень 2. Данные заносятся и хранятся в цифровом виде в МИС лечебной организации. На этом уровне появляется возможность применения расширенного анализа. Его недостатком является структурная ограниченность МИС, которая создается как цифровой аналог бумажной медицинской документации и, соответственно, не позволяет применять технологии получения данных и их глубокого анализа.

Уровень 3. Тотальная цифровизация всей медицинской информации с организацией цифрового пространства, которая позволяет использовать технологии искусственного интеллекта и машинного обучения для создания моделей, применяемых в улучшении продуктивности организации.

Уровень 4. Создание интеллектуальной цифровой экосистемы, где полученные в ходе автоматического анализа модели позволяют применять персонализированные стратегии лечения и улучшать финансово-организационные показатели работы медицинского учреждения.

Тенденция распространения новых цифровых технологий обработки больших массивов данных в медицине с каждым годом заметно растет. Согласно сведениям международной корпорации данных (International Data Corporation), расходы на проекты по работе с БД с 2017 по 2021 г. выросли с 12 до 57,6 млрд долларов [11]. Основные тенденции применения этих методик можно сгруппировать по нескольким направлениям:

1. Внедрение чат-ботов и цифровых помощников для рутинного взаимодействия с пациентами, что позволит медицинским работникам сосредоточиться на решении более сложных комплексных задач [12].

2. Аналитическая поддержка клинических решений согласно имеющимся стандартам и ру-

ководствам [13], прогнозирование развития клинической ситуации и определение вероятности повторного обращения, длительности госпитализации, летального исхода [14].

3. Адаптация медицинских текстов для информирования общественности [15].

4. Применение технологии сбора данных и последующее создание моделей для конкретного медицинского учреждения [16].

Компьютерный анализ БД позволяет улучшить стратегии лечения при заболеваниях нервной системы [17, 18]. В литературе опубликованы вычислительные модели, разработанные на больших данных МРТ и КТ, которые позволили с 91%-й вероятностью предсказать динамику геометрических показателей неразорвавшихся аневризм на основании параметров тока крови в сосудах головного мозга [18]. Кроме того, усовершенствование технологии глубокого обучения позволило создать классифицирующие модели изображений МРТ при болезни Альцгеймера, опухолях мозга и других патологиях [6, 17, 19]. Фактически данная технология превысила способности человека по выделению и классификации уникальных для определенных неврологических заболеваний характеристик из набора БД, содержащих неоднородную и необработанную информацию. Это позволяет применять БД в клинической работе для уменьшения влияния человеческого фактора и снижения риска ошибок диагностики, а также для ускорения процесса лечения. В литературе имеются наблюдения, подтверждающие эффективность моделей, основанных на БД, в прогнозировании исходов и рисков осложнений в хирургии позвоночника [17, 20, 21], а также в диагностике и определении степени злокачественности глиом головного мозга, в прогнозе выживаемости этих пациентов, а также в локализации эпилептогенного очага у пациентов с эпилепсией [22].

Заключение

В настоящей статье мы проанализировали ход работы по созданию набора БД пациентов с аневризмами головного мозга. Автоматизированный поиск данных позволил создать многокомпонентный структурированный НД, включающий 70 000 000 рядов инфо-единиц. Тесная кооперация с инженерами и четкая формулировка технического задания по АПД были важнейшей частью создания качественного НД. Вместе с тем в ходе перекрестной проверки историй болезней мы обнаружили, что выработанный алгоритм АПД имел ограничения в отношении ложнопозитивных случаев (в нашей клинике 32 %), связанных

с погрешностями оформления электронных историй болезни, включая опечатки регистратора, внесение неподтвердившегося диагноза и т.п. По нашему мнению, описанная методика позволяет получить наиболее качественный набор БД, что в конечном счете определяет надежность результатов последующего статистического анализа.

Список литературы / References

1. Aue G., Biesdorf S., Henke N. How health-care systems can become digital-health leaders. McKinsey and Company Healthcare Systems and Services. Available at: <https://www.mckinsey.com/industries/healthcare/our-insights/how-healthcare-systems-can-become-digital-health-leaders>
2. Faggella D. Where healthcare's big data actually comes from. Available at: <https://www.techemergence.com/where-healthcares-big-dataactually-comes-from>
3. Huesch M., Mosher T. Using it or losing it? The case for data scientists inside health care. *NEJM catalyst*. Available at: <https://catalyst.nejm.org/doi/full/10.1056/CAT.17.0493>
4. Gopal G., Suter-Crazzolara C., Toldo L., Eberhardt W. Digital transformation in healthcare - architectures of present and future information technologies. *Clin. Chem. Lab. Med.* 2019;57(3):328–335. doi: 10.1515/ccclm-2018-0658
5. Backes D., Rinkel G., Greving J., Velthuis B., Murayama Y., Takao H., Ishibashi T., Igase M., ter-Brugge K., Agid R.,... Vergouwen M. ELAPSS score for prediction of risk of growth of unruptured intracranial aneurysms. *Neurology*. 2017;88(17):1600–1606. doi: 10.1212/WNL.0000000000003865
6. Chang K., Bai H., Zhou H., Su C., Bi W., Agbodza E., Kavouridis V., Senders J., Boaro A., Beers A., ... Kalpathy-Cramer J. Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging. *Clin. Cancer Res.* 2018;24(5):1073–1081. doi: 10.1158/1078-0432.CCR-17-2236
7. Malhotra A., Wu X., Forman H., Matouk C., Gandhi D., Sanelli P. Management of tiny unruptured intracranial aneurysms: a comparative effectiveness analysis. *JAMA Neurol.* 2018;75(1):27–34. doi: 10.1001/jamaneurol.2017.3232
8. McKenna B. SAP banks on “intelligent enterprise” at Sapphire 2018. Available at: <https://www.computerweekly.com/news/252442802/SAP-banks-on-intelligent-enterprise-at-Sapphire-2018>
9. Quinn J.B. Strategic outsourcing: leveraging knowledge capabilities. *MIT Sloan Management Review*. Summer 1999. Available at: <https://sloanreview.mit.edu/article/strategic-outsourcing-leveraging-knowledge-capabilities/>

10. What is value-based healthcare? *NEJM Catalyst*. Available at: <https://catalyst.nejm.org/doi/full/10.1056/CAT.17.0558>
11. Shirer M., Daquilla M. IDC spending guide forecasts worldwide spending on cognitive and artificial intelligence systems to reach \$57.6 Billion in 2021. IDC. Available at: <https://www.idc.com/getdoc.jsp?containerId=prUS43095417>
12. Laranjo L., Dunn A., Tong H., Kocaballi A., Chen J., Bashir R., Surian D., Gallego B., Magrabi F., Lau A., Coiera E. Conversational agents in healthcare: a systematic review. *J. Am. Med. Inform. Assoc.* 2018;25(9):1248–1258. doi: 10.1093/jamia/ocy072
13. Perrin D. Lessons from the data: applying machine learning for clinical decision support. Vector medicine, Boston Children's Hospital. Available at: <https://vector.childrenshospital.org/2017/04/machine-learning-clinical-decision-support/>
14. Rajkomar A., Dean J., Kohane I. Machine learning in medicine. *N. Engl. J. Med.* 2019;380(14):1347–1358. doi: 10.1056/NEJMra1814259
15. Madhavan R. Machine translation – 14 current applications and services. Available at: <https://emerj.com/ai-sector-overviews/machine-translation-14-current-applications-and-services/>
16. Gillies J.C., Baird A.G., Gillies E.M. Balancing proactive and reactive care. *Occasional Paper in Royal College of General Practitioners*. 1995;71(71):15–28.
17. Aoe J., Fukuma R., Yanagisawa T., Harada T., Tanaka M., Kobayashi M., Inoue Y., Yamamoto S., Ohnishi Y., Kishima H. Automatic diagnosis of neurological diseases using MEG signals with a deep neural network. *Sci. Rep.* 2019;9(1):5057. doi: 10.1038/s41598-019-41500-x
18. Salimi Ashkezari S.F., Mut F., Slawski M., Cheng B., Yu A.K., White T.G., Woo H.H., Koch M.J., Amin-Hanjani S., Charbel F. T., ... Cebal J.R. Prediction of bleb formation in intracranial aneurysms using machine learning models based on aneurysm hemodynamics, geometry, location, and patient population. *J. Neurointerv. Surg.* 2022;14(10):1002–1007. doi: 10.1136/neurintsurg-2021-017976
19. Gulshan V., Peng L., Coram M., Stumpe M., Wu D., Narayanaswamy A., Venugopalan S., Widner K., Madams T., Cuadros J., ... Webster D. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402–2410. doi: 10.1001/jama.2016.17216
20. Hitchcock E., Gibson W. A Review of the genetics of intracranial berry aneurysms and implications for genetic counseling. *J. Genet. Couns.* 2017;26(1):21–31. doi: 10.1007/s10897-016-0029-8
21. Sing D., Metz L., Dudli S. Machine learning-based classification of 38 years of spine-related literature into 100 research topics. *Spine.* 2017;42(11):863–870. doi: 10.1097/BRS.0000000000002079
22. Senders J., Arnaout O., Karhade A., Dasenbrock H., Gormley W., Broekman M., Smith T. Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery.* 2018;83(2):181–192. doi: 10.1093/neuros/nyx384

Сведения об авторах:

Кивелёв Юрий Владимирович, PhD, ORCID: 0000-0002-5499-9628, e-mail: j.v.kivelev@gmail.com

Сааренпя Илкка, PhD, ORCID: 0000-0001-7013-6569, e-mail: ilkka.saarenpaa@tyks.fi

Кривошапкин Алексей Леонидович, д.м.н., проф., ORCID: 0000-0003-0789-8039, e-mail: alkr01@yandex.ru

Information about the authors:

Juri V. Kivelev, PhD, ORCID: 0000-0002-5499-9628, e-mail: j.v.kivelev@gmail.com

Ilkka Saarenpää, PhD, ORCID: 0000-0001-7013-6569, e-mail: ilkka.saarenpaa@tyks.fi

Alexey L. Krivoschapkin, PhD, professor, ORCID: 0000-0003-0789-8039, e-mail: alkr01@yandex.ru

Поступила в редакцию 04.02.2023

После доработки 10.04.2023

Принята к публикации 15.04.2023

Received 04.02.2023

Revision received 10.04.2023

Accepted 15.04.2023