EDN: ZAIBLC
УДК 004.85; 616.8-089
DOI: 10.56618/2071-2693_2025_17_4_22

# ЦЕЛЕСООБРАЗНОСТЬ И ОСОБЕННОСТИ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В НЕЙРОХИРУРГИИ

**Юрий Владимирович Кивелёв**[1,2]
✉j.v.kivelev@gmail.com, orcid.org/0000-0002-5499-9628, SPIN-код: 2687-9979

**Владислав Юрьевич Черебилло**[3]
cherebillo@mail.ru, orcid.org/0000-0001-6803-9954, SPIN-код: 2887-6943

**Алексей Леонидович Кривошапкин**[1,4]
alkr01@yandex.ru, orcid.org/0000-0003-0789-8039, SPIN-код: 8329-8540

[1] Акционерное общество «Европейский медицинский центр» (ул. Щепкина, д. 35, Москва, Российская Федерация, 129090)

[2] Федеральное государственное бюджетное образовательное учреждение высшего образования «Петрозаводский государственный университет» Министерства образования и науки Российской Федерации (ул. Красноармейская, д. 31, г. Петрозаводск, Российская Федерация, 185035)

[3] Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный медицинский университет имени академика И. П. Павлова» Министерства здравоохранения Российской Федерации (ул. Льва Толстого, д. 6-8, Санкт-Петербург, Российская Федерация, 197022)

[4] Федеральное государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов имени Патриса Лумумбы» Министерства образования и науки Российской Федерации (ул. Миклухо-Маклая, д. 6, Москва, Российская Федерация, 117198)

**Резюме**
Представлены особенности предварительной обработки больших данных, к которым могут применяться технологии искусственного интеллекта в клинической нейрохирургии. На первичном этапе автоматизированного создания крупных регистров пациентов данные характеризуются ошибками («шум»), пропусками и вариабельностью, что требует применения специфических методов предобработки. Очистка данных от шума, устранение пропусков и решение проблемы вариабельности на начальном этапе критически важны для проведения последующего анализа и применения технологий искусственного интеллекта. Предобработка данных включает в себя обнаружение статистических выбросов, нормализацию, извлечение признаков, выбор признаков. Эти этапы осуществляются такими методами, как анализ полных наблюдений, метод импутации, метод повторения результата последнего наблюдения, методы по уменьшению размерности данных. Применение данных методов позволяет повысить точность и эффективность алгоритмов искусственного интеллекта и улучшить интерпретацию результатов. Освещены практические рекомендации для нейрохирургов по выбору и применению методов предобработки данных.
**Ключевые слова:** большие данные, искусственный интеллект в нейрохирургии

***Для цитирования:*** *Кивелёв Ю. В., Черебилло В. Ю., Кривошапкин А. Л. Целесообразность и особенности предварительной обработки больших данных в нейрохирургии // Российский нейрохирургический журнал им. проф. А. Л. Поленова. 2025. Т. XVII, № 4. С. 22–27. DOI: 10.56618/2071-2693_2025_17_4_22.*

## REASONS AND FEATURES OF BIG DATA PREPROCESSING IN NEUROSURGERY. PRACTICAL GUIDELINE

**Juri V. Kivelev**[1,2]
✉j.v.kivelev@gmail.com, orcid.org/0000-0002-5499-9628, SPIN-code: 2687-9979

**Vladislav Yu. Cherebillo**[3]
cherebillo@mail.ru, orcid.org/0000-0001-6803-9954, SPIN-code: 2887-6943

**Alexey L. Krivoshapkin**[1,4]
alkr01@yandex.ru, orcid.org/0000-0003-0789-8039, SPIN-code: 8329-8540

[1] European Medical Center (35 Schepkina street, Moscow, Russian Federation, 129090)

[2] Petrozavodsk State University (31 Krasnoarmeyskaya street, Petrozavodsk, Russian Federation, 185035)

[3] Pavlov University (6-8 L'va Tolstogo street, Saint Petersburg, Russian Federation, 197022)

[4] RUDN University (6 Miklukho-Maklaya street, Moscow, Russian Federation, 117198)

**Abstract**

Current paper presents the features of big data preprocessing, to which artificial intelligence technologies can be applied in clinical neurosurgery. At the initial stage of creating patient registries, the data are characterized by missing values and variability, which requires the use of specific preprocessing methods. Data cleaning, handling missing values, and addressing variability issues at the initial stage are critically important for subsequent analysis and the application of artificial intelligence technologies. Data preprocessing includes detecting statistical outliers, normalization, feature extraction, and feature selection. These steps are performed using methods such as complete case analysis, imputation techniques, the last observation carried forward method, and dimensionality reduction methods. The application of these methods improves the accuracy and efficiency of artificial intelligence algorithms and enhances the interpretation of results. This paper provides practical recommendations for neurosurgeons on the selection and application of data preprocessing methods

**Keywords:** big data, artificial intelligence in neurosurgery

***For citation:*** *Kivelev Ju. V., Cherebillo V. Yu., Krivoshapkin A. L. Reasons and features of big data preprocessing in neurosurgery. Practical guideline. Russian neurosurgical journal named after professor A. L. Polenov. 2025;XVII(4):22–27. (In Russ.). DOI: 10.56618/2071–2693_2025_17_4_22.*

## Введение

В последние годы методы искусственного интеллекта (ИИ) находят широкое применение в клинической нейрохирургии для анализа больших данных. Однако эффективность алгоритмов машинного обучения (МО) зависит от качества исходных данных. Автоматизированное создание крупных регистров на основе медицинских информационных систем (МИС) часто сопряжено с наличием «грязных» данных с неточностями (шумами), пропусками и несоответствиями. Предобработка данных становится критически важным этапом, определяющим успех последующего анализа.

**Целью** работы является описание ключевых этапов и методов предобработки данных для клинических исследований в нейрохирургии, а также предоставление практических рекомендаций для исследователей и врачей.

## Первичный набор данных

Большие данные, полученные из МИС в практической нейрохирургии, зачастую содержат непредвиденные ошибки («шум»), разнообразные несоответствия, а также нередко имеют пропущенные, незаполненные значения [1]. Подобные информационные массивы на первичном этапе формирования можно обозначить как «грязные» данные. Вполне естественно предположить, что запускать «грязные» данные в комплексную статистическую обработку непродуктивно, а проблемные участки необходимо идентифицировать и, по возможности, исправить на самом начальном этапе [2]. Примером шума данных может быть значения возраста более 125 лет, что однозначно указывает на ошибку заполнения [3]. Здесь причиной может быть как человеческий фактор (опечатка регистратора), так и технический сбой функционирования МИС.

В качестве примера несоответствий может выступать запись информации в свободном текстовом формате, содержащая опечатки, а также различные способы описания одних и тех же медицинских характеристик. В силу специфики работы различных подразделений, взаимодействующих с нейрохирургической службой стационаров, полная унификация и стандартизация текстового описания клинических параметров на практике неосуществима, что и объясняет подобные несоответствия в контексте формализации данных. Кроме того, к категории несоответствий можно отнести случаи записи, когда, например, вес пятилетнего ребенка указывается в размере 75 кг. В этом примере 75 кг – показатель, который вполне вписывается в средний вес совершеннолетнего пациента, однако очевидно, не соответствует весу детей дошкольного возраста. «Грязные» данные затрудняют применение методов машинного обучения, так как усложняют выявление значимых признаков [4, 5]. Автоматизированное обнаружение и устранение шума, несоответствий и учиты-

вание пропущенных показателей могут занять довольно большое количество времени, однако это критически важно для получения адекватного и надежного результата при последующей статистической обработке.

*Рекомендации.* Методы устранения шума:

1) удаление выбросов:

– использование пороговых значений (например, возраст в пределах 0–100 лет);

– кластеризация для обнаружения аномальных значений;

2) стандартизация данных: приведение текстовых данных к единому формату (например, через регулярные выражения).

Под предобработкой данных понимается процесс очистки и подготовки данных для моделей машинного обучения. Она включает в себя несколько этапов, таких как обнаружение статистических выбросов, нормализация, извлечение признаков, выбор признаков и т. д. [5]. Каждому из этих этапов следует уделить должное внимание, так как зачастую они имеют несколько решений разного уровня сложности. Например, обнаружение выбросов может быть выполнено отдельно для каждого признака, удаляя точки данных за пределами некоторого распределения или за пределами определенного диапазона значений (например, возраст не может быть меньше 0 или больше 125 лет) [5]. Также выбросы могут быть обнаружены с помощью метода кластеризации [4].

Отметим, что в медицинских, и, в частности, нейрохирургических, наборах больших данных особенно актуальна проблема пропущенных, недостающих значений. Например, люмбальная пункция в послеоперационном периоде проводится по определенным показаниям только у части пациентов. Соответственно, в наборе данных эта переменная будет зафиксирована у какой-то части пациентов, но при этом будет отсутствовать у остальных. Соответственно, насыщенность данными мониторинга у одних пациентов может сочетаться с отсутствием соответствующих показателей у других, а это в контексте оценки набора данных рассматривается как пропущенное, недостающее значение [6]. В целом пропущенные данные можно разделить на три группы [7]:

1) полностью случайные пропуски (ПСП) – не зависят от других переменных;

2) случайные пропуски (СП) – связаны с наблюдаемыми переменными;

3) неслучайные пропуски (НП) – обусловлены ненаблюдаемыми факторами.

На практике предобработка наборов больших данных нацелена на коррекцию, в первую очередь, СП или НП [7].

*Рекомендации.* Обнаружение пропусков и ошибок:

1) проверить, есть ли в базе данных отсутствующие значения (например, не указаны возраст, диагноз, результаты обследований);

2) определить, случайны ли пропуски или связаны с конкретными факторами (например, отсутствие данных из-за непроведенного обследования);

3) привести показатели к единому формату (например, показатели внутричерепного давления – только в мм рт. ст.);

4) после исправления убедиться, что данные логичны, последовательны и соответствуют клинической практике

## Работа с отсутствующими данными

Одним из методов обработки пропущенных данных является анализ полных наблюдений, когда в анализ берутся показатели, не имеющие отсутствующих значений [7]. Несмотря на то, что этот метод удобен и используется по умолчанию во многих пакетах статистического программного обеспечения, он может пропустить релевантную информацию исследования, а если пропуски не являются ПСП, он может внести погрешность в конечный результат [7, 8].

Более подходящим вариантом устранения пропусков является метод импутации [3]. Этот способ заменяет недостающие значения в элементах выборки, например, средним значением или медианой выборки соседних известных значений, а также с помощью предиктивных моделей. Использование среднего или медианы для импутации не лишено недостатков, так как, подобно методике полного анализа случаев, оно может внести смещения и погрешность в данные, искусственно создавая дополнительные значения [8]. Более эффективным спосо-

бом является метод интерполяции, прогнозирующий недостающие значения на основе уже доступных переменных [7]. К примеру, при пропусках в выборке, посвященной анализу заболеваемости глиобластомой, будет обоснованным внести в предиктивную модель предположение о том, что эта опухоль у пожилых пациентов будут встречаться чаще, чем у молодых.

С помощью множественной импутации каждый из полученных новых наборов данных анализируется по отдельности и после этого сводится к единому формату вычисления таких параметров, как среднее, стандартная ошибка и доверительные интервалы результатов. Таким образом, множественная импутация снижает искажения, возникающие при заполнении пропусков. Соответственно, использование этого метода видится более предпочтительным по сравнению с однократной импутацией [7].

Коррекция пропусков в выборках медицинских данных с временными рядами может проводиться методом повторения результата последнего наблюдения. При этом делается допущение, что значение не меняется в следующем временном интервале, а сам временной интервал относительно непродолжителен. Например, в выборке данных указан пациент с диагнозом конвекситальной менингиомы у бессимптомного неоперированного пациента, который периодически проходит контрольные обследования. Во временном ряду выборки обнаружились пропуски кода этого диагноза. В соответствии с принципом повторения результата последнего наблюдения, диагноз, поставленный на временном шаге $t$–1, будет сохраняться с большой долей вероятности в момент $t$, и наоборот. Соответственно, заполнение пропуска значением как предыдущего, так и последующего временного шага в этом случае не приведет к искажению выборки.

В наборах медицинских данных пропуски могут отражать решения медицинских работников, что позволяет отнести их к категории НП. Важно отметить, что отсутствие данных в элементах выборки само по себе может быть информативным признаком. Об этом свидетельствуют исследование [6] – работа была посвящена сравнению различных методов обработки пропущенных значений в данных педиатрического отделения интенсивной терапии (ОРИТ). Исследователи создали несколько моделей для прогнозирования диагнозов после лечения в ОРИТ и сравнили различные методы обработки пропущенных значений. Этими методами были нулевая импутация, интерполяция с переносом вперед и обнаружение индикаторов пропусков в сочетании с нулевой интерполяцией и переносом вперед. Наилучшие результаты по всем показателям были достигнуты при нулевой интерполяции и использовании индикаторов пропусков с помощью длинной нейронной сети с кратковременной памятью (long short-term memory neural network, LSTM). Кроме того, исследователи обучили модели только на индикаторах пропусков, показав, что для нескольких заболеваний информация о том, какие тесты были применимы, а какие нет, оказалась столь же предсказуемой, как и фактические измерения.

*Рекомендации.* Методы обработки пропусков:

1) анализ полных наблюдений – прост в применении, но может привести к потере информации и смещению данных;

2) импутация – замена пропущенных значений средним, медианой или использованием предиктивных моделей;

3) множественная импутация – создание нескольких наборов данных и объединение результатов для повышения точности;

4) метод повторения последнего наблюдения – используется для временных рядов, когда значения переносятся на последующие временные точки.

## Многоразмерность

Электронная история болезни пациента содержит стандартизованный набор информации, включая диагнозы, справочную информацию, текстовые данные, данные о жизненных показателях и различные другие типы данных. Все это делает данные пациента многоразмерными и создает проблему, называемую «проклятием размерности» [9]. При увеличении числа переменных растут вычислительная сложность и риск переобучения моделей. Кроме того, возникает и проблема обобщаемости модели. Что-

бы справиться с «проклятием размерности», необходимо или отфильтровать признаки, или уменьшить размерность, либо сделать и то и другое. Методы уменьшения размерности призваны преобразовать данные в меньший набор новых признаков, которые, по сути, содержат ту же информацию, что и исходные признаки. Способы уменьшения размерности представлены методами отбора признаков, которые направлены на удаление избыточных признаков из набора данных. Эти методы можно разделить на несколько групп, включая фильтр, обертку и гибридные методы [10–12]. Методы фильтрации обычно являются наименее затратными с вычислительной точки зрения [12]. Они не зависят от алгоритмов машинного обучения, используемых для прогнозирования, и основаны на различных статистических тестах. Хорошими примерами методов, основанных на фильтрах, являются коэффициент корреляции Пирсона, использующийся для отбора только тех признаков, которые коррелируют с переменной отклика [13], и коэффициент «вздутия» дисперсии, который направлен на устранение мультиколлинеарности – тесной взаимосвязи между признаками [14]. Поскольку методы, основанные на фильтрах, не зависят от применяемого далее алгоритма машинного обучения, они обладают более широкими возможностями для предобработки [12].

Оберточные методы оценивают подмножества признаков путем обучения модели МО на этом подмножестве, а затем определяют производительность модели [12]. Этот процесс продолжается до тех пор, пока не будет достигнут критерий остановки. Критерием остановки может быть достижение определенного количества признаков, или факт того, что производительность модели более не увеличивается. Оберточные методы страдают от повышенных потерь при поиске, избыточной погрешности и имеют меньшую обобщаемость, поскольку для выбора признаков используется конкретный алгоритм машинного обучения [12]. Примером оберточного метода является оператор наименьшего абсолютного сокращения и отбора регрессии (least absolute shrinkage and selection operator, LASSO) [15]. Регрессионные модели по типу LASSO, как правило, легко интерпретируются, и считается, что они устойчивы к мультиколлинеарности [16]. Гибридные методы подразумевают использование как фильтров, так и оберточных методов [12, 17].

*Рекомендации.* Методы уменьшения размерности:

1) фильтрующие методы:

– коэффициент корреляции Пирсона;

– коэффициент «вздутия» дисперсии;

2) оберточные методы:

– рекурсивное исключение признаков;

– метод LASSO (устойчив к мультиколлинеарности);

3) гибридные методы – сочетание фильтров и оберточных подходов для повышения точности и снижения вычислительных затрат.

## Заключение

Применение методов искусственного интеллекта в клинической нейрохирургии может потребовать предварительной обработки данных. Большие данные, полученные из медицинских информационных систем в практической нейрохирургии, зачастую содержат непредвиденные ошибки, разнообразные несоответствия, а также нередко имеют пропущенные, незаполненные значения. Методы устранения недостатков в наборах данных направлены на повышение эффективности последующего анализа алгоритмами искусственного интеллекта. Для достижения этих задач мы рекомендуем учитывать следующие аспекты:

1) при наличии шума использовать многоканальные методы обнаружения и устранения выбросов;

2) при работе с пропусками отдавать предпочтение множественной импутации для повышения точности;

3) для уменьшения размерности использовать комбинацию фильтров и методов рекурсивного исключения признаков;

4) интегрировать этапы предобработки на ранних стадиях разработки модели для снижения риска систематических ошибок.

**Конфликт интересов.** Авторы заявляют об отсутствии конфликта интересов. **Conflict of interest.** The authors declare no conflict of interest.

**Финансирование.** Исследование проведено без спонсорской поддержки. **Financing.** The study was performed without external funding.

## Литература / References


1. *Chu X. et al.* Data cleaning: Overview and emerging challenges. Proceedings of the 2016 International Conference on Management of Data, SIGMOD '16. 2016, Association for Computing Machinery. New York, 2016, pp. 2201–2206. Doi: 10.18699/SSMJ20230311.
2. *Кивелев Ю. В., Сааренпя И., Кривошапкин А. Л.* Формирование набора больших данных для клинических исследований на примере аневризм сосудов головного мозга // Сибир. науч. мед. журн. 2023. Т. 453, № 3. С. 86–94. [Kivelev J. V., Saarenpää I., Krivoshapkin A. L. Establishing of big data clinical dataset in brain vessel aneurysm research. Sibirskij nauchnyi medicinski jurnal. 2023;43(3):86–94. (In Russ.)]. Doi: 10.18699/SSMJ20230311.
3. *Suvanto K.*-O. Machine learning based prediction of intensive care unit motrality for patients with aneurysmal subarachnoid hemorrhage . Faculty of information technology and electrical engeneering; University of Oulu. 2022, pp. 84.
4. *Bhaya W.,* Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences. 2017;(12):4102–4107. Doi: 10.3923/jeasci.2017.4102.4107.
5. *Kotsiantis S. B., Kanellopoulos D., Pintelas P. E.* Data preprocessing for supervised leaning. International Journal of Computer and Information Engineering. 20;(1):111–117. Doi: 10.5281/zenodo.1082415.
6. *Lipton Z. C., Kale D. C., Wetze R.* Modeling missing data in clinical time series with RNNs. 2016. Available from: https://arxiv.org/abs/1606.0413 [Accessed 15 September 2025].
7. *Mack C., Su Z., Westreic D.* Managing missing data in patient registries: Addendum to registries for evaluating patient outcomes: A user's guide. 3rd ed. 2018.
8. *Zhang Z.,* Missing data imputation: focusing on single imputation. Ann Transl Med. 2016;4(1):9. Doi: 10.3978/j.issn.2305-5839.2015.12.38.
9. *Berisha V. et al.* Digital medicine and the curse of dimensionality. NPJ Digit Med. 2021;4(1):153. Doi: 10.1038/s41746-021-00521-5.
10. *Sorzano C. O. S., Vargas J., Montano A. P.* A survey of dimensionality reduction techniques. 2014. Available from: https://arxiv.org/abs/1403.2877 [Accessed 15 September 2025].
11. *van Der Maaten L. J. P., Postma E. O., van Den Herik H. J.* Dimensionality reduction: a comparative review. Journal of Machine Learning Research. 2009;(10):1–41.
12. *Gnana D. A. A., Balamurugan S. A. A., Leavline E. J.* Literature review on feature selection methods for high-dimensional data. International Journal of Computer Applications. 2016;(136):9–17. Doi: 10.5120/ijca2016908317.
13. *Benesty J. et al.* Pearson Correlation Coefficient, in Noise Reduction in Speech Processing. Springer: Berlin, Heidelberg; 2009, pp. 1–4. Doi: 10.1007/978-3-642-00296-0_5.
14. *Craney T. A., Surles J. G.* Model-dependent variance inflation factor cutoff values Quality Engineering. 2002;(14):391–403. Doi: 10.1081/QEN-120001878.
15. *Tibshirani R.* Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;(58):267–288. Doi: 10.1111/j.2517-6161.1996.tb02080.x.
16. *Dormann C. F. et al.* Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 2013;(36):27–46. Doi: 10.1111/j.1600-0587.2012.07348.x.
17. *Aziz R., Verma C., Srivastava N.* Dimension reduction methods for microarray data: a review. AIMS Bioengineering. 2017, pp. 179–197. Doi: 10.3934/bioeng.2017.1.179.


## Сведения об авторах


*Юрий Владимирович Кивелёв* – кандидат медицинских наук, нейрохирург клиники нейрохирургии Европейского медицинского центра (Москва, Россия); доцент Медицинского института им. А. П. Зильбера Петрозаводского государственного университета (г. Петрозаводск, Россия);

*Владислав Юрьевич Черебилло* – доктор медицинский наук, профессор, врач-нейрохирург, заведующий кафедрой и клиникой нейрохирургии Первого Санкт-Петербургского государственного медицинского университета им. акад. Павлова (Санкт-Петербург, Россия);

*Алексей Леонидович Кривошапкин* – доктор медицинский наук, профессор, член-корреспондент Российской академии наук, врач-нейрохирург, заведующий кафедрой неврологии и нейрохирургии Российского университета дружбы народов им. Патриса Лумумбы (Москва, Россия); врач-нейрохирург, заведующий Отделением нейрохирургии Европейского медицинского центра (Москва, Россия); главный научный сотрудник научно-исследовательского отдела ангионеврологии и нейрохирургии Национального медицинского исследовательского центра им. акад. Е. Н. Мешалкина (г. Новосибирск, Россия).


## Information about the authors


*Juri V. Kivelev* – Cand. of Sc. (Med.), Neurosurgeon at the Neurosurgery Clinic, European Medical Center (Moscow, Russia); Associate Professor at the A. P. Silber Medical Institute, Petrozavodsk State University (Petrozavodsk, Russia);

*Vladislav Yu. Cherebillo* – Dr. of Sci. (Med.), Full Professor, Head at the Department and Clinic of Neurosurgery, Pavlov University (St. Petersburg, Russia);

*Alexey L. Krivoshapkin* – Dr. of Sci. (Med.), Full Professor, Corresponding Member of the Russian Academy of Sciences, Neurosurgeon, Head at the Department of Neurology and Neurosurgery, RUDN University (Moscow, Russia); Neurosurgeon, Head at the Neurosurgery Department, European Medical Center (Moscow, Russia); Chief Researcher at the Scientific Research Department of Angioedema and Neurosurgery, National Medical Research Center named after Academician E. N. Meshalkina (Novosibirsk, Russia).



Поступила в редакцию 31.08.2025
Поступила после рецензирования 10.10.2025
Принята к публикации 12.12.2025

Received 31.08.2025
Revised 10.10.2025
Accepted 12.12.2025