Т. 8, № 4. С. 118–124. **Медицинские науки** 2023
Научная статья
УДК 618

СОКОЛОВА
Бажена Евгеньевна
ординатура, Петрозаводский государственный университет (Петрозаводск, Россия),
bazhena98@bk.ru

РАЗМЕТКА МЕДИЦИНСКИХ ТЕКСТОВ ДЛЯ РАЗРАБОТКИ СКРИНИНГОВОЙ СИСТЕМЫ ПРОГНОЗИРОВАНИЯ РИСКА ХРОМОСОМНЫХ АНОМАЛИЙ У ПЛОДА

Научный руководитель:
Ившин Александр Анатольевич
Рецензент:
Корнева Виктория Алексеевна
Статья поступила: 09.11.2023;
Принята к публикации: 28.11.2023;
Размещена в сети: 01.12.2023.

Аннотация. В статье рассматриваются способы выполнения разметки обезличенных клинических данных. Представлены признаки, полученные в результате анализа медицинской литературы, на основании которых возможен расчет риска хромосомных аномалий у плода. Подробно описана характеристика бинарных и теговых признаков. Представлены алгоритмы выполнения разметки в программе Microsoft Excel и на сервисе Label Studio.

Ключевые слова: разметка данных, теговые признаки, бинарные признаки, расчет риска, прогнозирование, хромосомные аномалии

Благодарности. Исследования, описанные в данной работе, были проведены в рамках проекта «Разработка скрининговой системы прогнозирования индивидуального риска хромосомных аномалий у плода в Республике Карелия на основе алгоритмов машинного обучения», поддержанного в рамках Программы поддержки НИОКР студентов, аспирантов и лиц, имеющих ученую степень, обеспечивающих значительный вклад в инновационное развитие отраслей экономики и социальной сферы Республики Карелия, в 2023 году, финансируемой Правительством Республики Карелия (Договор № 3-Г22 от 29.12.2022 между ФГБОУ ВО «Петрозаводский государственный университет» и Фондом венчурных инвестиций Республики Карелия, Соглашение № КГРК-23/18)

Для цитирования: Соколова Б. Е. Разметка медицинских текстов для разработки скрининговой системы прогнозирования риска хромосомных аномалий у плода // StudArctic Forum. 2023. Т. 8, № 4. С. 118–124.

Разметка медицинских данных представляет собой процесс поиска и извлечения определенных признаков (их наличия/отсутствия в тексте, числового значения признака) из неструктурированных текстов (протоколов врачебных осмотров, дневниковых записей, результатов лабораторных исследований и т. д.). Процесс разметки медицинских текстов сводится к поиску тех или иных признаков (например, факторов риска развития хромосомных аномалий у плода) в текстовых документах и их обозначению в понятной для машинной обработке форме.

Для этого оператором производится тщательный анализ текстового документа, выполняется поиск определенных слов или словосочетаний, обозначающих искомый

признак. Например, наличие у плода синдрома Дауна. В рамках медицинского текста данный признак может быть обозначен по-разному. В одной записи, это «синдром Дауна», в другой «трисомия по 21 хромосоме», в третьей и вовсе искомый признак будет обозначен как «47 XX (XY) + 21» и т. д. Во всех перечисленных выше примерах разные по написанию фразы имеют одинаковый смысл, все они содержат информацию о том, что у плода имеется определенная хромосомная патология.

Основная задача в процессе разметки заключается в том, чтобы максимально точно и обширно, учитывая различные варианты обозначения в тексте того или иного состояния, выявить наличие или отсутствие определенного признака.

Все признаки можно условно разделить на две категории: бинарные и теговые. Так, бинарный признак требует только указания на то, есть он в тексте или нет, никаких дополнительных манипуляций с такого рода признаками не проводится. Разметка таких признаков позволяет обучить машину распознаванию наличия или отсутствия определенного фактора риска в медицинской записи[1].

Теговый признак всегда имеет числовое значение (например, возраст, рост, вес). В данном случае требуется прямое выделение числового параметра соответствующего признака со строгим соблюдением ряда правил. Например, для разметки признака «возраст матери» требуется выделить текст таким образом, чтобы первым значением было наименование искомого признака (в данном случае это возраст пациентки), а последним обязательно числовое значение, соответствующее параметру (в обязательном порядке - без единиц измерения). Например, словосочетание «женщина в возрасте 34 лет» в пригодной для машинной обработки форме будет выглядеть как «возраст 34». Разметка такого рода признаков позволяет обучить машину распознаванию разного рода параметров, требующих числового/количественного значения (антропометрические данные, индексы диагностики, концентрация определенных лабораторных маркеров и т. д.).

Для выбора необходимой информации для разметки были получены данные о врожденных пороках развития, хромосомных аномалиях, особенностях проведения пренатального скрининга, которые позволили сформировать перечень признаков, необходимых для расчета риска анеуплоидий[2] [Астраханцева], [Беременность], [Джаманкулова], [Alldred].

Все признаки, полученные для разметки, были разделены на две основные категории в зависимости от способа разметки. Получилось 22 бинарных и 16 теговых признаков. Все признаки представлены в таблице 1.

Таблица 1

Перечень признаков с необходимыми значениями для расчета риска анеуплоидий

Теговые признаки	Бинарные признаки

1. Копчико-теменной размер плода 2. ЧСС 3. Толщина воротникового пространства 4. Пульсационный индекс в венозном протоке 5. Трикуспидальная ренургитация 6. ХГЧ 7. PAPP-A 8. АФП 9. Ингибин А 10. Неконъюгированный эстрадиол 11. Материнский возраст 12. Расовая принадлежность 13. Вес матери 14. Рост матери 15. ИМТ 16. Уровень фолиевой кислоты	1. Хромосомные аномалии 2. Синдром Дауна 3. Синдром Эдвардса 4. Синдром Патау 5. Отягощенная наследственность 6. Предыдущий ребенок/плод с Трисомией 21/18/13 7. Гипоплазия НК 8. ВПР 9. НИПТ 10. ИПД 11. Многоплодная беременность 12. Одноплодная беременность 13. ЭКО/ИКСИ/ВРТ 14. Прерывание беременности в анамнезе 15. Замершие беременности в анамнезе 16. Социально-экономический статус матери 17. Табакокурение 18. Наркомания 19. Алкоголизм 20. Ожирение 21. Дефицит фолиевой кислоты 22. Прегравидарная подготовка

Разметка неструктурированных клинических данных пациенток, которым выполнялся пренатальный скининг, выполнялась в двумя способами: в программе Microsoft Excel и на сервисе Label Studio.

Для выполнения разметки в программе Microsoft Excel было создано два алгоритма.

Алгоритм № 1. Подготовка набора данных к разметке:

1) удалить ненужные столбцы таблицы;

2) выбрать записи для анализа (осмотры, результаты исследований);

3) получить необходимую информацию из полей таблицы по признакам;

4) сформировать «словарик», состоящий из набора слов/словосочетаний для каждого признака (эти признаки должны входить в строку в соответствии с выполнением условия нахождения признака (близость нахождения слов));

5) обнаружить попадание признака в строку с помощью регулярных выражений;

6) добавить строки, содержащие признак, в новые выборки для данной категории признаков с ограничением по объему.

В результате выполнения работы с данными по алгоритму были сформированы обучающие выборки, содержащие записи врачебных осмотров и протоколы исследований.

Далее выполнялась разметка медицинских текстов по следующему алгоритму.

Алгоритм № 2. Разметка признаков в программе Microsoft Excel:

1) применить для столбца с медицинскими данными (текстом осмотра) текстовый фильтр «содержит»;

2) в фильтр ввести одно из значений признака, в результате будет сформирована выборка записей, содержащих этот конкретный признак;

3) отметить цифрой «1» ячейку в соответствующем данному признаку столбце (на пересечении строки и нужного признака);

4) в фильтр данных необходимо поочередно вводить все возможные варианты написания признака, повторяя при этом пункт 2, 3;

Далее необходимо из сформированной выборки записей, исключить признак,

содержащий отрицание:

5) применить для столбца с медицинскими данными текстовый фильтр «содержит»;

6) в фильтр данных необходимо поочередно ввести все возможные варианты написания отрицания признака;

7) отметить цифрой «0» ячейку в соответствующем данному признаку столбце.

Таким образом, для каждого признака, при его наличии в записи, необходимо отметить в соответствующем данному признаку столбце цифрой «1», а при его отсутствии/отрицании признака – цифрой «0». На рисунке 1 представлен алгоритм разметки признаков в программе Microsoft Excel.

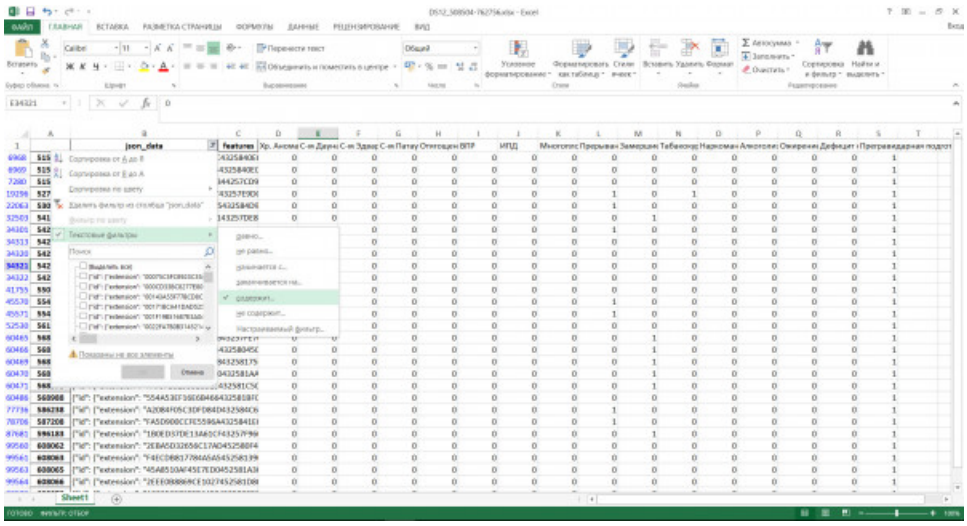

Рис. 1. Разметка признаков в Microsoft Excel. Изображение автора

Разметка массива неструктурированных клинических данных по бинарным признакам выполняется путем указания наличия или отсутствия признака. А разметка по теговым признакам выполняется путем указания признака и его значения, единиц измерения. Далее выполнялось уточнение всех возможных вариантов написания конкретного признака в медицинской документации, включая сокращения, номенклатуру, синонимы.

Например, в текстовом поле осмотра была цель найти признак «табакокурение». Для признака «табакокурение» возможными вариантами написания признака будут: курение; курит; курила; хроническая никотиновая интоксикация; ХНИ; запах табака. Их отмечаем как «1». Возможные варианты отрицания признака: не курит; вредные привычки отрицает; вредных привычек нет. Их отмечаем как «0».

Например, признак «хромосомные аномалии» является бинарным, в медицинской документации может встречаться в виде следующих вариантов написания: хр. аномалии; хром. аномалии; хромосомные болезни; хром. болезни; ХБ; ХА; анеуплоидии. Признак «Альфа-фетопротеин» является теговым, единица измерения Ед/мл, способы написания: α-фетопротеин; AFP; альфа-ФП; АФП. Бинарный признак «многоплодная беременность» имеет кодировку по МКБ-10: О30; О30.1; О30.2; О30.8; О30.9, может быть записан как: многоплодие; многоплодная монохорильная моноамниотическая беременность; ММ б-ть; многоплодная монохориальная диамнотическая беременность; МД бер-ть и т.д.

Другой способ для выполнения разметки медицинских текстов – это использование сервиса Label Studio.

На сервис были выгружены медицинские тексты. Создана сетка бинарных и теговых признаков. Операторами прочитывались все тексты, проводился анализ данных, поиск признаков. При обнаружении бинарного или тегового признака выполнялась отметка в сетке. Например, диагноз «ожирение 1 степени», соответствует одноименному бинарному

признаку, указано присутствие данного признака в тексте. При теговой разметке, выполнялось не только указание наличия признака, но и конкретное нахождение параметра в тексте. Например, лабораторный параметр «свободная бета-субъединица ХГЧ (ХГЧ)» выбран в сетке, а затем выделен в тексте способ написания признака и его числовое значение, единицы измерения. На рисунках 2 и 3 представлен интерфейс системы при выполнении разметки.

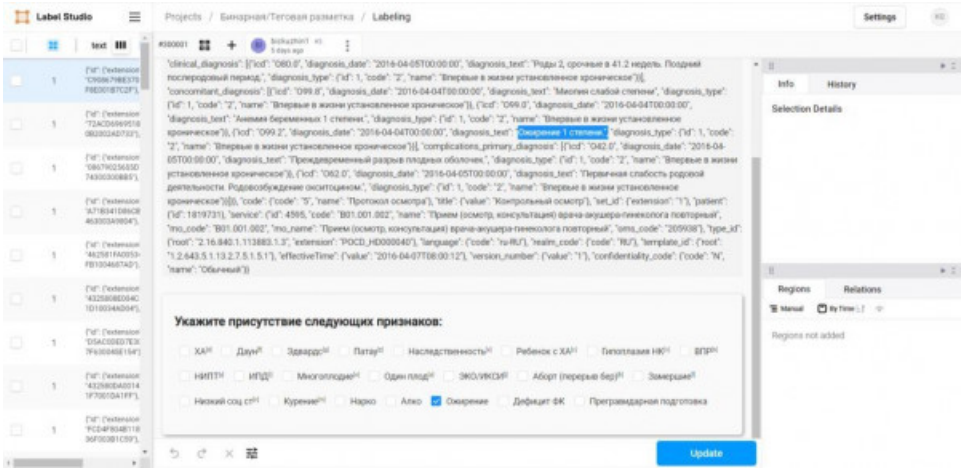

Рис. 2. Интерфейс Label Studio при разметке бинарного признака. Изображение автора

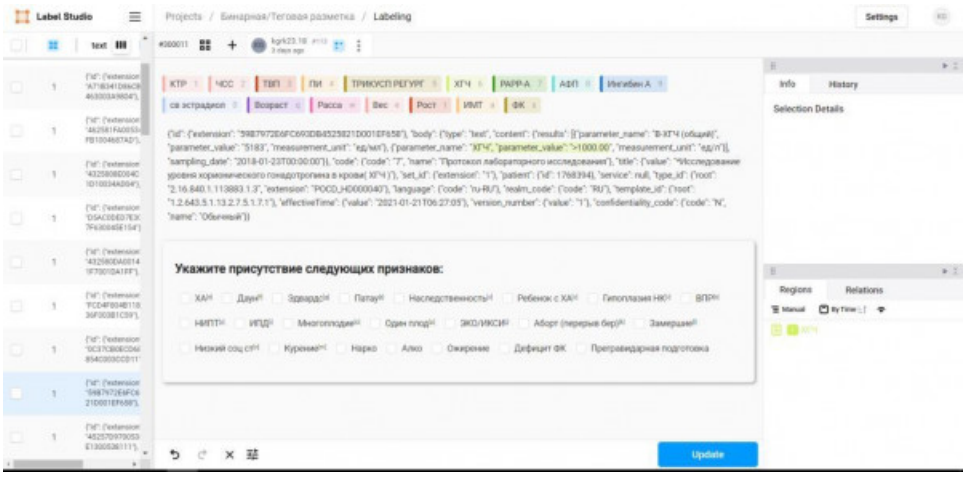

Рис. 3. Интерфейс Label Studio при разметке тегового признака. Изображение автора

В результате проведенных работ по разметке неструктурированных текстов получен набор размеченных медицинских данных. Размечено 30000 медицинских текстов (данные анамнеза, осмотра, заключений врача, рекомендаций, УЗ-маркеров и биохимических маркеров скрининга). Выполненная работа по разметке массива неструктурированных клинических данных используется при разработке модуля извлечения признаков анеуплоидий.

Примечания

[1] Основные метрики задач классификации в машинном обучении // WEBIOMED. URL: https://webiomed.ru/blog/osnovnye-metriki-zadach-klassifikatsii-v-mashinnom-obuchenii (дата обращения: 09.05.2023).

[2] Клинические рекомендации по нормальной беременности. Москва, 2020. 87 с. // Рубрикатор клинических рекомендаций [сайт]. URL: https://cr.minzdrav.gov.ru/schema/288_1 (дата обращения 20.04.23); Об утверждении

Порядка оказания медицинской помощи по профилю "акушерство и гинекология": Приказ Министерства здравоохранения РФ от 20 октября 2020 г. № 1130н // Гарант [сайт]. URL: https://base.garant.ru/74840123 (дата обращения: 04.04.2023).

СПИСОК ЛИТЕРАТУРЫ

Астраханцева М.А. Профилактика и диагностика врождённых пороков развития / М.А. Астраханцева, Кику П.Ф., Воронин С.В., Сухова А.В. // Здравоохранение Российской Федерации. 2021. Т. 65. № 3. С. 230-236.

Беременность ранних сроков. От прегравидарной подготовки к здоровой гестации / Под ред. В.Е. Радзинского, А.А. Оразмурадова. Москва: StatusPraesens, 2020. 800 с.

Джаманкулова Ф.С. Оценка факторов риска у беременных женщин и прогнозирование развития врожденных пороков плода / Ф.С. Джаманкулова, М.С. Мусуралиев, А.А. Сорокин // Казанский медицинский журнал. 2018. № 5. С. 748–753.

Alldred S.K. First and second trimester serum tests with and without first trimester ultrasound tests for Down’s syndrome screening / S.K. Alldred, Y. Takwoingi, B. Guo, M. Pennant, J.J. Deeks, J.P. Neilson // Cochrane database Syst Rev. 2017. Mar 15.

García-Pérez L. Cost-effectiveness of cell-free DNA in maternal blood testing for prenatal detection of trisomy 21, 18 and 13: a systematic review / L. García-Pérez, R. Linertová, M. Álvarez-de-la-Rosa, J.C. Bayón, I. Imaz-Iglesia, J. Ferrer-Rodríguez // Eur J Health Econ. 2018. Sep. 19(7). P. 979–991.

Bazhena E. SOKOLOVA — medical residency, Petrozavodsk State University (Petrozavodsk, Russia), bazhena98@bk.ru

MARKUP OF MEDICAL TEXTS FOR DEVELOPING A SCREENING SYSTEM TO PREDICT THE RISK OF FETAL CHROMOSOMAL ABNORMALITIES

Scientific adviser:
Aleksandr A. Ivshin
Reviewer:
Viktoria Alekseevna Korneva
Paper submitted on: 11/09/2023;
Accepted on: 11/28/2023;
Published online on: 12/01/2023.

Abstract. This article examines various methods for the markup of anonymized clinical data and presents features derived from a medical literature review that can be used to calculate the risk of fetal chromosomal abnormalities. Additionally, the article gives a detailed description of the binary and tagged features and proposes algorithms for performing the markup using Microsoft Excel and Label Studio tool.

Keywords: data markup, tag features, binary features, risk calculation, prognosis, chromosomal abnormalities

For citation: Sokolova, B. E. Markup of Medical Texts for Developing a Screening System to Predict the Risk of Fetal Chromosomal Abnormalities. StudArctic Forum. 2023, 8 (4): 118–124.

REFERENCES

Astrakhantseva M.A., Kiku P.F., et al. Prevention and diagnosis of congenital malformations. *Health Care of the Russian Federation*, 2021, Vol. 65, No. 3, pp. 230-236. (In Russ.)

Radzinsky V.E., Orazmuradov A.A., eds. *Early pregnancy. From pregravidar preparation to healthy gestation.* Moscow, StatusPraesens, 2020, 800 p.

Jamankulova F.S., Musuraliev M.S., et al. Estimation of the risk factors in pregnant women and prediction of congenital fetal anomalies. *Kazan Medical Journal*, 2018, No. 5, pp. 748-753. (In Russ.)

Alldred S.K., Takwoingi Y., et al. First and second trimester serum tests with and without first trimester ultrasound tests for Down's syndrome screening. *Cochrane Database Syst Rev*, 2017, No. 3.

García-Pérez L., Linertová R., et al. Cost-effectiveness of cell-free DNA in maternal blood testing for prenatal detection of trisomy 21, 18 and 13: a systematic review. *Eur J Health Econ*, 2018, No. 7, P. 979-91.