УДК 004.852

ПРОГНОЗИРОВАНИЕ РАСХОДОВАНИЯ МЕДИКАМЕНТОВ ДЛЯ ЛЕЧЕНИЯ ОНКОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ НА РАЗРОЗНЕННЫХ ДАТАСЕТАХ

Береснев А.Д. (Университет ИТМО), **Рыжков Н.М.** (Университет ИТМО)
Научный руководитель – к.т.н., ст.тех.сотр. Гусарова Н.Ф.
(Университет ИТМО)

Введение. Отдел по организационно-методической работе с регионами НМИЦ им. Петрова собирает данные от лечебных учреждений, страховых медицинских организаций, комитетов по здравоохранению и министерств здравоохранений регионов северо-западного федерального округа. К округу относится 11 субъектов РФ. Актуальной задачей является прогнозирование расхода лекарств для лечение онкологических заболеваний на основе собираемых данных. Прогноз необходимо построить в условиях разных форматов и источников данных, их неполноты и слабой связности.

Основная часть. Для построения комплексного датасета статистики лечения онкологических заболеваний была разработана информационная система, позволяющая импортировать данные из разных источников, объединять и анализировать их, а так же строить прогностические модели.

С целью проведения анализа был подготовлен датасет содержащий сведения о случаях лечения с помощью лекарственных препаратов. В нем содержатся такие данные, регистрационный номер препарата, дата его введения, пол пациента, его вес и площадь поверхности тела и дозировка. При составлении датасета возникли следующие проблемы:

- расхода лекарств в непосредственном представлении в имеющихся источниках не имелось,
- имеющихся на момент начала анализа данных не хватало, из-за чего понадобился поиск дополнительной информации,
- для формирования датасета необходимо было использовать данные из различных источников.

Во всех колонках, кроме номера и даты, содержится огромное количество пропусков, которые были по необходимости заполнены по определенным правилам. Исходя из дозировок был посчитан расход лекарств и агрегирован помесячно. В результате получился одномерный временной ряд, отображающий зависимость расхода лекарства в миллиграммах в месяц от месяца с конца 2019 до конца 2021 года.

В процессе исследования требовалось подобрать наиболее эффективную модель прогноза расходования лекарственных средств, не предполагающую необходимости построение и обучения сложных моделей требующих полных датасетов. Для прогнозирования временных рядов существует несколько классов моделей, имеющих свои особенности: наивные, экспоненциальное сглаживание, SARIMA, Garch, динамические линейные модели, NNETAR, LSTM и другие. Все они разного уровня сложности для прогнозирования рядов в зависимости от их длины, формы и наличия каких-либо свойств (сезонности, тренда, цикличности).

Перед непосредственным построением моделей были внесены календарные поправки для определения среднего ежедневного расхода в месяц, удалены выбросы в данных (использовался IQR критерий).

В пилотном исследовании применялись методы прогнозирования наивный, скользящее среднее, модель случайного блуждания. Анализ результатов показал, что модели предсказание простых моделей значительно теряет точность при увеличении периода анализа.

Для получения более точного прогноза была использована модель ARIMA, которая представляет собой еще один метод для прогнозирования временных рядов, применялось экспоненциальное сглаживание.
Для нескольких лекарственных препаратов (Капецитабина, Паклитаксела) были получены данные ежедневного расхода по месяцам, автокорреляционной функции данных о расходе. Эти данные были построены для моделей ARIMA с параметрами, полученными автоматически и вручную. Анализ показал, что:
- модель с подобранными вручную параметрами показывает несколько большую точность,
- модель обладает достаточной точностью при прогнозе на небольшие периоды,
- при прогнозе на временные периоды от квартала и выше, падает доверительный интервал прогноза.

Выводы. Проведенное исследование показало, что в настоящий момент построение полного статистического датасета по лечению онкологических заболеваний затруднено. На имеющихся ограниченных данных прогноз лекарственных средств методами анализа временных рядов показывает достаточную точность только на сравнительно небольших временных промежутках (квартал), для более точных оценок на временных промежутках от полугода планируется построить модель применения схем лечения с учетом частоты использования для пациентов разных возрастных групп и провести анализ на ее основе.

Использованные источники:

1. Forecasting: Principles and Practices: https://otexts.com/fpp2/. Дата обращения - 28.10.2022
2. R documentation: https://www.rdocumentation.org/. Дата обращения - 05.11.2022
3. Методы обработки результатов измерений и оценки погрешностей в учебном лабораторном практикуме: учебное пособие; издание второе / Н.С. Кравченко, О.Г. Ревинская; Национальный исследовательский Томский политехнический университет. – Томск: Изд-во Томского политехнического университета, 2017. – 121 с

Береснев А.Д. (автор)

Гусарова Н.Ф. (научный руководитель)