МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Харківський національний університет імені В.Н. Каразіна

Факультет математики і інформатики

Кафедра теоретичної та прикладної інформатики

Кваліфікаційна робота

магістр

на тему Впровадження машинного навчання в клінічне статистичне програмування на основі стандартів CDISC моделі табуляції даних дослідження (SDTM)

Виконав: студент 2 курсу, групи МФ-61
спеціальність 122 «Комп'ютерні науки»
освітньо-наукова програм «Інформатика»

Харюк Д.О.
(прізвище та ініціали)

Керівник
(прізвище та ініціали)

Рецензент
(прізвище та ініціали)

Харків – 2024 року

ЗМІСТ

1.ВСТУП

1.1 Формування мети роботи, задач та обґрунтування актуальності теми

Мета роботи полягає у досліджені можливості використання машинного навчання на етапі створення доменів моделі табуляції даних дослідження або SDTM, у сфері клінічного дослідження.

Досягнення мети наукового дослідження зумовило постанову і вирішення кола питань, що склали його основні задачі:

- Ознайомлення з типами машинного навчання для подальшого використання для досягнення мети.
- Ознайомлення з алгоритмами машинного навчання відповідного типу для подальшого використання для досягнення мети.
- Підготовка даних відповідних стандартам CDISC для машинного навчання.
- Написання програми реалізації обраного алгоритму машинного навчання.
- Отримання, перевірка, візуалізація результатів машинного навчання.

Однією з основних задач у клінічних дослідженнях, від першого опису у «Книзі пророка Даниїла» у Біблії [1] і до сучасності, залишається інтерпретація отриманих даних.

В історії людства помилкові судження призводили, як до великих відкриттів так і жахливих катастроф. У такій галузі науки як клінічні дослідження помилкова інтерпретація може призвести до значних фінансових втрат з боку дослідників, так і до втрати життя досліджуваного, що є не припустимими подіями. Для зменшення вірогідності появи подібних подій існують стандарти компанії CDISC та різноманітні «Форми звіту про випадок».

«Форми звіту про випадок» або «Case Report Form» (надалі CRF) це спеціалізований документ у клінічних дослідженнях призначений для збору даних про пацієнта під час клінічного випробування. Розробка CRF має велике значення для всього клінічного дослідження і безпосередньо впливає на результат [2]. Відповідно до Міжнародної Конференції з Узгоджених Рекомендацій для Належної Клінічної Практики CRF визначається як: друкований, оптичний або електронний документ, призначений для запису необхідної, відповідної до протоколу, інформації про кожного суб’єкта випробування, яка має бути надана спонсору випробування [3].

Стандарти компанії Clinical Data Interchange Standards Consortium або CDICS покликані перетворювати несумісні формати, неузгоджені методології та різноманітні перспективи у потужну структуру для створення даних клінічних досліджень, які відображають чітку картину випробування. Стандарти CDISC поділяються на багато класів, таких як: Study Data Tabulation Model / Модель табуляції даних дослідження, Analisis Data Model / Модель даних аналізу і таке інше.

У цій роботи зроблений акцент на Study Data Tabulation Model / Модель табуляції даних дослідження або SDTM, а саме покликана розглянути можливості прискорення створення SDTM алгоритмами машинного навчання. Етап

інтерпретації табуляції даних може займати від одного дня до двох тижнів на форму CRF. Спроможність майже миттєво ідентифікувати цільовий домен невідомої форми CRF дозволить значно прискорити процес створення SDTM.

1.2 Стислий огляд відомих результатів в області дослідження

На момент написання роботи було знайдено два джерела зі схожою тематикою.

Beyond Traditional Methods: Implementing AI/ML in Clinical Stats Programming, Naga Durga Suravajhula. Робота була презентована на PHUSE 2023 року і представляє собою застосунок для автоматизації рекомендаційного табулювання форми CRF.

Automating SDTM for Clinical Trial Data Submissions Using Artificial Intelligence and Machine Learning, Melanie R. Ciotti, MBA, Dale W. Usner, PhD, Richard B. Abelson, PhD, Statistics & Data Corporation, Tempe, AZ, USA. У статті описано варіацію використання машинного навчання для передбачення табуляції форми CRF, генерації SDTM набору даних та анотації CRF.

1.3 Відомості про одержані результати та їх новизна

У роботі Beyond Traditional Methods: Implementing AI/ML in Clinical Stats Programming для презентації результатів використовується базовий домен Демографічних показників, який є унікальним і незмінним. Попри це результат рекомендаційного табулювання складає приблизно 50% співпадіння і потребує подальшого редагування спеціалістом.

У роботі Automating SDTM for Clinical Trial Data Submissions Using Artificial Intelligence and Machine Learning для презентації результатів так само використовується базовий домен Демографічних показників. Представлено

результати тренування машини на одній змінній, але з розширенням на розуміння яка змінна має бути табульована основна чи додаткова. Не представлені результати генерування SDTM набору даних.

Робота складається зі вступу, основної частини, загальних висновків, списку використаної літератури (39). Зміст роботи висвітлено на 34 сторінках основного тексту і містить 2 рисунки.

2.ОСНОВНА ЧАСТИНА

2.1 Постановка задачі

Задачею у даній роботі було обрано ідентифікація цільового домена невідомої CRF форми, яка складається з відомих запитань. Данна задача являє собою класичну задача класифікація. Для вирішення цієї задачі було виділено наступні цілі:

- Огляд типів штучного інтелекту.
- Обрати тип штучного інтелекту, що буде відповідати задачі класифікації.
- Огляд алгоритмів штучного інтелекту обраного типу.
- Обрати алгоритм штучного інтелекту для вирішення задачі класифікації.
- Підготовка CRF форм актуальних досліджень з відомою табуляцією.
- Конвертувати обрані CRF форми у набір даних для подальшого використання для машинного навчання.
- Реалізувати алгоритм машинного навчання з допомогою мови програмування Python.
- Отримані результати перевірити на відсоток співпадіння.
- Візуалізувати результат, для формування висновку.
- Формулювання висновку на основі отриманих результатів.

2.2 Розвинутий огляд сучасного стану справ в області дослідження

У низці статей описується можливі застосування Штучного Інтелекту (ШІ) у сфері клінічних досліджень. Перша за все це застосування ШІ для дослідження областей низької прибутковості (таргетована терапія, рідкісні захворювання).Розглядаються методи підвищення набору пацієнтів і розробки протоколів, що може підвищити шанси на успіх випробування. Також розглядаються методи покращення моніторингу пацієнтів і аналізу отриманих результатів, що може позитивно вплинути на вимірювання та інтерпретацію результатів [11,23,13,24].

У сфері до клінічних досліджень такий напрямок, як ідентифікація нових молекулярних цілей для вирішення актуальних медичних проблем, може бути прискореним за рахунок прогнозування токсичності з допомогою ШІ. ШІ може дати поштовх у розвитку генерації нових стабільних молекул з дійсним лікувальним потенціалом. У випадку надання доступу до великих масивів даних по фармакокінетиці та фармакодинаміці з попередніх клінічних і до клінічних досліджень, не зважаючи на те чи були вони вдалими чи ні, з'являється можливість для розробки та навчання ефективних алгоритмів для генерації нових молекул. Нажаль даних по фармакокінетиці та фармакодинаміці немає у вільному доступі по причині високої конкуренції серед фармацевтичних компаній [13,15].

Для прогнозування безпеки описано декілька методологій ШІ. Фактично програмне забезпечення вже має можливість прогнозувати токсичність базуючись на таргетований інформації. Привнесення зміни у традиційне до клінічне дослідження in vitro та дослідження на тваринах може відбутися за рахунок підвищення ефективності прогнозування токсичності [12]. Також моделі можуть знайти своє застосування як інструменти керування ризиками та визначення пріоритетів під час розробки, забезпечуючи виявлення сполук високого ризику, які можуть сильно вплинути на безпечність препарату [13].

Однією з перших по складності задача, для всіх напрямків ШІ, є інтерпретація моделі, особливо враховуючи великий рівень невизначеності на ранніх стадіях розробки. Ключовим є розуміння особливостей моделі та біохімічних механізмів, що лежать в її основі, для інтерпретації та достовірності прогнозів [13].

У напрямку дизайну клінічних досліджень ШІ можуть бути використані для підбору та оптимізації дизайну випробування шляхом усунення явно не підходящих варіантів та для прогнозування клінічних результатів обраного дизайну, що може напряму вплинути на розвиток точної медицини. Фактично, ШІ має потенціал для моделювання даних для виявлення більш ефективного дизайну з кращими статистичними показниками [16]. В одній роботі навіть припускають можливість використання алгоритмів ШІ для розрахунку прогнозу результатів учасників дослідження, що дасть змогу завчасно виявити перспективних учасників, що може повпливати на тривалість випробування [17]. На додачу, аналіз електронних медичних записів за допомогою ШІ може зпрогнозувати вірогідність покидання випробування пацієнтом. Такі знання про пацієнта можна використовувати по різному. Замість того, щоб заздалегідь виключати учасників, які мають вірогідність покинути дослідження, ці інформацію можна використати для заохочення пацієнтів прийняти участь у іншому досліджені. Таким чином можна зменшити необхідну кількість учасників дослідження [10].

В одній з статей продемонстрували можливість зменшення смертності учасників клінічного дослідження від раку на 15-25% за допомогою алгоритмів машинного навчання [18]. Таким чином за допомогою таких алгоритмів машинного навчання можна прогнозувати клінічні результати, які стратифіковані за економічними та генетичними ознаками. Створення таких алгоритмів базується на великих масивах біологічних даних які були зібрані з інтервенційних досліджень, що мають кореляцію біомаркерів, які відносяться до препарату, з даними про відсутність погіршення стану захворювання та рівня смертності, за профілями пухлин [18]. У наступній статті за допомогою моделі машинного

навчання росту пухлин аналізували дані дослідження недрібноклітинного раку легень, з метою перевірки достовірності статусі біомаркерів, і не тільки, у прогнозуванні пухлинної реакції та рівня виживання [14].

Перелічені застосунки штучного інтелекту можуть бути використані для покращення вибору ліків і підготовку досліджуваних препаратів до гістології конкретного типу раку, що призведе до зниження рівня смертності. Звісно це не ідеальні застосунки і вони потребують вдосконалення, однак такі моделі машинного навчання, що використовують комплексні дані з мультиомічних особливостей, мають потенціал до зміни парадигми лікування та методів набору учасників і привнесення нових способів розробки точних клінічних випробувань.

Неможливо звернути увагу на застосування машинного навчання для прогнозування ймовірності успіху випробування. Так у статті [16], приводять приклади того, що алгоритми машинного навчання сприяють ранньому виявленню та прогнозуванню перебігу захворювання, підвищуючи таким чином вірогідність успіху клінічного дослідження. За межами успіху клінічного дослідження, ШІ можна використовувати на ранніх стадіях клінічного дослідження для прогнозування молекулярних особливостей, чутливості цілі, біодоступності та токсичності [15,7]. На пізніх стадіях дослідження ШІ можуть підвищити вірогідність успіху, тим самим збільшити кількість досліджень 2 та 3 фазу, які отримають схвалення регуляторних органів. Як наслідок маємо більш раціональне використання людських та фінансових ресурсів, підвищення забезпечення безпеки учасників і покращення сприйняття громадськістю клінічних досліджень [23,15].

Інформація про дизайн клінічного дослідження та характеристики учасників можуть бути використані для тренування моделі МН, яку можна буде використати не тільки для прогнозування схвалення регуляторних органів, але й оцінити ймовірність переходу між фазами дослідження [19]. Розуміння таких факторів як складність протоколу, вибір кінцевих клінічних точок, груп дослідження та критеріїв включення/виключення дає можливість впливати не

тільки на успіх та невдачу окремої фази дослідження, але і на дизайн випробування на подальших фазах [23,14]. Розробка подібних моделей на, доступних у відповідній літературі, вірогідностях побічних ефектів або відсутності ефективності може допомогти в розробці клінічних досліджень.

У сфері онкології штучний інтелект використовується для створення випробувань In-Silico, в яких використовують клінічні дані для симуляції когорт, які моделюють ефективність лікування [26]. Також розглядають можливість відбору респондентів для покращення результативності на пізніх стадіях розвитку [26]. Проте залишається проблема браку якісних, підібраних і завершених наборів даних, яка безпосередньо впливає на потенціал застосування штучного інтелекту [26].

Розглянемо застосування штучного інтелекту у зміні дизайну клінічного дослідження. Згенеровані штучним інтелектом рішення дозволяють швидше і більш точно генерувати та аналізувати гіпотези, щоб поглибити розуміння еволюції захворювання, а також збільшити вірогідність відкриття ліків, покращити підбір когорт, моніторинг, забезпечення та вибір кінцевої мети [11,23]. Загалом спостерігається покращення результатів при запроваджені методології штучного інтелекту в дизайн. Наприклад, вдосконалення протоколу та верифікація біомаркерів підвищує придатність складу досліджуваних груп. Але покращення та зменшення кількості помилок моделі штучного інтелекту потребують об’єднання знань та зусиль для створення загального виду протоколу зборів, методів архівування та організації великих наборів даних [11].

Також розглядаються можливості симуляції контрольної групи досліджуваних, для прогнозування прогресування захворювання з допомогою добре розробленої моделі штучного інтелекту яка була натренована на достатній кількості якісних даних. Такий підхід дасть можливість вилучити з досліджень групу плацебо і замінити її на синтетичну групу досліджуваних [17]. Подібні зміни можуть покращити результат та пришвидшити дослідження, скоротити

бюджет, зменшити кількості учасників, що призведе до зменшення навантаження на сайти. Однак подібні генерації синтетичних груп дослідження, на основі тренувальних наборів даних з існуючих клінічних досліджень, потребують значних ресурсів і витрат часу. Не менш важливим є можливість усунення етичних проблем щодо рандомізації на контрольну групу плацебо та групу що приймає препарат, за рахунок синтетичної контрольної групи плацебо [11].

Реалізація подібних ідей на практиці і побудова відповідної інфраструктури займе ще багато часу і зустріне багато проблем на своєму шляху, серед яких є етичні міркування та неправомірне використання медичної інформації [23].

Так само проблеми рекрутинга, які обов’язково з’являються при організації та проведенню клінічних досліджень, мають критичний вплив на результат. Погано набрані учасники безпосередньо впливають на час проведення дослідження, що очевидно тягне за собою збільшення фінансових витрат. Основними факторами які впливають на якість є складність протоколу дослідження, погана поінформованість учасників випробування, емоційній страх перед випробуванням та банальна не зацікавленість у досліджені [27,20].

Складність дослідження безпосередньо впливає на складність критеріїв включення та виключення. Це тягне за собою укладення набору пацієнтів, що можуть відповідати необхідним критеріям відбору, щоб уникнути неправдоподібності отриманої інформації або помилкової класифікації [28,21].

У різних терапевтичних областях було підкреслено, що моделі ШІ можуть допомогти у рекрутингу учасників, за складними критеріями включення, поєднуючи такі дані, як демографічні, лабораторні, томографічні та інші омічні дані [17,27,20,22].

Також розглядається можливість для автоматизованого поширення інформації про випробування потенційним учасникам, які відповідають

необхідним критеріям, тобто ШІ може покращити відбір пацієнтів, надаючи інформацію ширшому колу потенційних учасників випробувань через загальнодоступні платформи клінічних досліджень [20]. На приклад, використовуючи великий масив даних з області метаболізму було показано, що інструменти ШІ сприяють більш справедливому відбору учасників, які отримають права на участь у досліджені, що можна вважати гарним результатом[29]. Подібні підходи можуть бути використані для покращення обізнаності потенційних учасників у пошуку підходящого клінічного дослідження і сприяти створенню механізмів підбору, як це було продемонстровано в дослідженні ВІЛ [5].

Ефективне використання подібних алгоритмів штучного інтелекту у рекрутингу може буде засновано на впровадженні стандартизованих критеріїв відбору, що дасть можливість взаємодії цих систем. Алгоритм повинен мати можливість читати і розуміти вхідні дані, щоб його можна було використовувати за призначенням [8,9]. Однією з ідей реалізації подібного є поєднання структурованих даних з інформацією, отриманою в результаті обробки звітів пацієнтів написаних вільним стилем, щоб доповнити інформацію для відповідності критеріїв скринінгу [4]. Відносно нещодавно було продемонстровано гарні результатів у системному відборі учасників дослідження використовуючи алгоритмів штучного інтелекту на основі баз даних сайту clinicaltrials.gov до якої увійшло понад 350 тисяч досліджень [6,30].

Інша проблема, яку штучний інтелект зможе вирішити, пов’язана з навантаженням на ключові теми клінічних досліджень, що змушує спонсорів працювати тільки з коротким списком відомих дослідників, базуючись на рівні їх досвіду і результатів попередньо проведених клінічних досліджень. Штучний інтелект зможе допомогти спонсорам швидко і просто підібрати підходящого дослідника, що прискорить початок дослідження і процес набору персоналу [31].

Декілька вказують на можливість використання штучного інтелекту у проведенні випробування. Один з таких варіантів базується на використанні

Цифрових Технологія Охорони Здоров'я. Інформація зібрана з допомогою автоматизованих інструментів збору даних, таких як натільні пристрої і датчики, та розробка нових цифрових біомаркерів, створених та проаналізованих з допомогою штучного інтелекту, дозволяють отримувати інформацію про стан учасник у режимі близькому до реального часу [32]. Такий підхід дозволить покращити нагляд за учасниками та проводити випробування на учасниках у близькому до критичного стані, що стає очевидним джерелом важливої критичної інформації.

Інший приклад застосування, це контроль дотримання прийому досліджуваного лікарського засобу, що є основною проблемою на клінічних дослідження пов'язаних з психологічними та неврологічними розладами. Вже існуючі технології для відстежування факту прийому досліджуваного лікарського засобу, такі як фіксація відкриття баночки з ліками, не надають точного результату [33]. Саме у цьому випадку можна застосувати альтернативні методи підтвердження споживання досліджуваного лікарського засобу на основі штучного інтелекту. На приклад, пристрій відеоспостереження з вбудованим штучним інтелектом може бути використаним для більш чіткої фіксації підтвердження того, що учасник прийняв ліки. Такий метод вважається єдиним способом підтвердження споживання досліджуваного лікарського засобу без безпосереднього нагляду за учасниками клінічного дослідження з боку персоналу [33,34].

Розглядаючи напрямок медичних зображень, а саме аналізу та керування робочим процесом, у якому штучний інтелект може бути використаним для оптимізації та вдосконалення аналізу медичного зображення, можна помітити можливість покращення клінічного дослідження. Алгоритми штучного інтелекту можуть спростити задачу пошуку важливих маркерів, які зазвичай відмічаються експертами вручну [17,35]. Штучний інтелект також може покращити робочий процес аналізу зображень, беручи на себе роботу класифікації зображень, що дають більше часу експертам на інші задачі [32]. Суттєвою проблемою для

реалізації подібних ідей залишається значні витрати часу та трудомісткість процесу створення репозиторію стандартизованих зображень необхідних для навчання штучного інтелекту [35].

Кажучи про аналіз не можна не відмітити визначення ефекту гетерогенності. Загалом дослідження показують, що ідеально гомогенні ефекти лікування майже не зустрічаються, тому виявлення ефекту гетерогенності є добре відомою задачею для вчених статистиків, що проводять клінічні дослідження [36]. Були створені моделі штучного інтелекту, які були навчені на наборах даних серцево-судинних захворювань, для перевірки даних зібраних під час клінічного дослідження та ідентифікації підгруп з різними ефектами лікування, так само як і для ідентифікації ключових факторів ризику та швидкого допомоги у субпопуляції [27,36]. Такі моделі дозволяють проводити більш комплексний аналіз та покращити розуміння ситуації на досліджені для розробників лікарських засобів. Однак, залишається проблема визнання результатів, отриманих за допомогою цих моделей, регуляторними органами [36].

Немало важливою є проблема імпутації пропущених даних та обробка пропущених візитів у досліджені. Під час пандемії COVID-19 багато клінічних досліджень постраждали через перерозподіл ресурсів між клінічними центрами в яких проводилось дослідження. У пацієнтів виникали труднощі з проїздом до місця проведення дослідження, та через те що учасники заразились вірусом і більше не могли приймати участь у дослідженні. Подібні проблеми призводили до появи пропущених даних та затримки у запланованих візитах, що впливало на статистичний аналіз [37]. У таких випадках ідеально підійшли б алгоритми машинного навчання для вирішення питання пропущених даних та визначення стану учасників, коли візити були відкладені по за межі визначених термінів [37,38].

Багато статей свідчить про те що інтеграція штучного інтелекту у клінічні дослідження є перспективною галуззю, яка постійно розвивається. Штучний

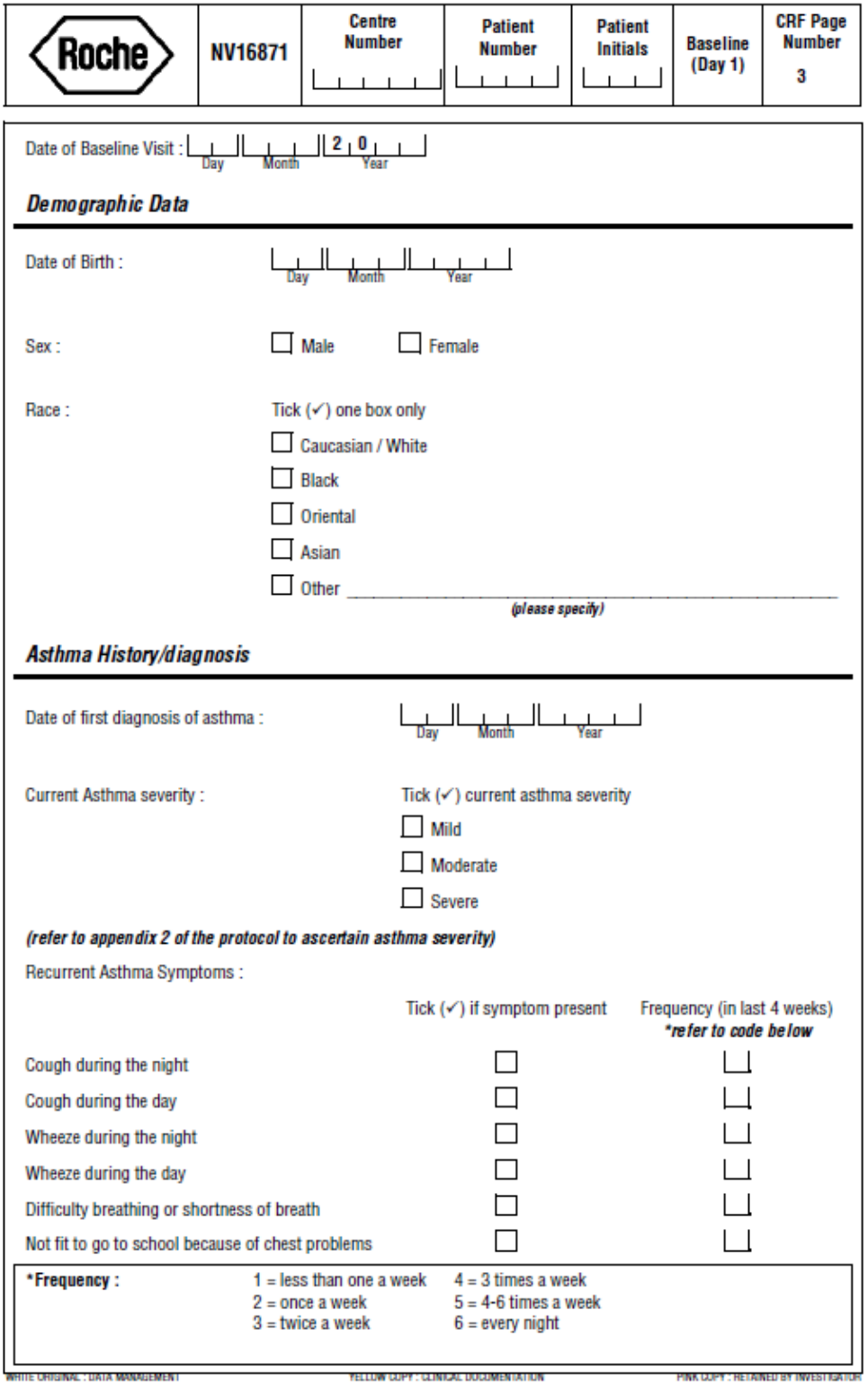
інтелект може стати ключем до подолання поточного застою в розробці ліків і прокласти шлях до нової парадигми розвитку медичних досліджень. Кількість досліджень пов'язаних з оцінку перспектив використання штучного інтелекту, вказують на що в індустрії є зацікавленість цією темою, особливо у практичному застосуванні для проведення більш успішних і економічно вигідних випробувань. Спектр інтеграції штучного інтелекту у процес створення та схвалення ліків є всеосяжним і охоплює всі фази життєвого циклу лікарських засобів.

Однак, все ще існує проблема відсутності конкретних та детальних вказівок щодо використання штучного інтелекту у клінічних дослідження від регуляторних органів, попри наявність публікацій використання штучного інтелекту та стратегічних планів регуляторних онагрів.

Сфера охорони здоров’я, як одна з найбільш регульованих та не схильних до ризику галузей буде повільно і безпечно адаптуватись до світової цифрової трансформації. Тому незважаючи на позитивні звіти, попереду ще довгий шлях на якому майорять питання визначення належних етичних та нормативних рамок використання штучного інтелекту у клінічних дослідженнях і не тільки.

2.3 Методи дослідження

Клінічне дослідження складається з багатьох етапів, один з яких є формування CRF. CRF має бути добре спроектованим, щоб пацієнт або співробітник клініки могли записати правильну інформацію про пацієнта або події пов’язані з пацієнтом.



Roche	NV16871	Centre Number	Patient Number	Patient Initials	Baseline (Day 1)	CRF Page Number 3

Date of Baseline Visit : Day Month 2 0 Year

Demographic Data

Date of Birth : Day Month Year

Sex : ☐ Male ☐ Female

Race : Tick (✓) one box only

☐ Caucasian / White
☐ Black
☐ Oriental
☐ Asian
☐ Other _____ *(please specify)*

Asthma History/diagnosis

Date of first diagnosis of asthma : Day Month Year

Current Asthma severity : Tick (✓) current asthma severity

☐ Mild
☐ Moderate
☐ Severe

(refer to appendix 2 of the protocol to ascertain asthma severity)

Recurrent Asthma Symptoms :

	Tick (✓) if symptom present	Frequency (in last 4 weeks) *refer to code below
Cough during the night	☐	☐☐
Cough during the day	☐	☐☐
Wheeze during the night	☐	☐☐
Wheeze during the day	☐	☐☐
Difficulty breathing or shortness of breath	☐	☐☐
Not fit to go to school because of chest problems	☐	☐☐

***Frequency :** 1 = less than one a week; 2 = once a week; 3 = twice a week; 4 = 3 times a week; 5 = 4-6 times a week; 6 = every night

WHITE ORIGINAL : DATA MANAGEMENT YELLOW COPY : CLINICAL DOCUMENTATION PINK COPY : RETAINED BY INVESTIGATOR

Рис.2.1 - Приклад CRF.

Всі CRF розподіляються по темах відповідно до даних, які будуть зібрані у форму. Однією з умов створення CRF є уникати дублювання інформації, як на

рівні однієї форми так і у декількох формах. Назва форми зазвичай відповідає темі до якої відноситься форма.

До прикладу:

- форма “Демографія” відповідно зберігає демографічну інформацію про пацієнта таку як день народження, вік, стать, раса, етнічність і т.д.;
- форма “Життєво-важливі ознаки” відповідно зберігає інформацію про життєво-важливі показники пацієнта такі як систолічний кров’яний тиск, діастолічний кров’яний тиск, пульс, температура, частота дихання і т.д.;
- форма “Місцевий аналіз сечі” відповідно зберігає інформацію про результати аналізу сечі пацієнта такі як кислотність, глюкоза, протеїн, кетони, кров, білірубін, лейкоцитарна естераза, нітрити, уробіліноген і т.д.;
- форма “Побічна подія або супутнє захворювання” відповідно зберігає інформацію про побічні події та захворювання які були у пацієнта або проявились під час дослідження, а саме назва побічної події або захворювання, дата початку побічної події або захворювання, дата кінця побічної події або захворювання, складність побічної події або захворювання за градацією NCI CTCAE, яка розшифровується як National Cancer Institute Common Terminology Criteria for Adverse Events. Також ця форма зберігає інформацію про можливий препарат який міг викликати побічну реакцію або захворювання та інформацію про потужність складових препарата.

Якщо є форми назви яких відповідають темі до якої відноситься, то мають бути і винятки. Зустрічаються форми у яких назва не може повністю описати, яку саме інформацію така форма зберігає. До прикладу форма “Побічна подія - внутрішньо очне запалення”, ґрунтуючись на назві форми, має зберігати інформацію про побічні події пов’язані з внутрішньо очним запаленням, але вже існує форма “Побічна подія або супутнє захворювання” у якій мають зібрані всі

можливі побічні події або супутні захворювання, що сталися з пацієнтом під час дослідження, також має зберігатись умова не дублювати інформацію серед форм. Відповідно до змінних у формі “Побічна подія - внутрішньо очне запалення” можна зрозуміти що ця форма зберігає інформацію не про факт того що ця побічна подія відбулась, а про опис побічної події типу внутрішньо очне запалення. Наприклад є питання “пацієнт скаржиться на симптоми внутрішньо очного запалення (Оберіть усе, що підходить)” і список відомих симптомів та строчка для вільного запису. Загалом вся форма побудована по принципу, якщо було знайдено якийсь симптом його треба описати і перевірити за допомогою підготовлених можливих варіантів побічних подій та тестів спрямованих на виявлення внутрішньо очне запалення.

Існування подібних форм може бути пояснена особливими умовами такими як неочевидний зв’язок між отриманими даними, визначений зв’язок між отриманими даними, описання процедури та отримані дані і т.д. Також така форма може бути пов’язана з чимось новим, з чим розробники CRF ще не працювали і не можуть точно визначити на які форми поділити новий метод тестування то все записують в одну єдину форму, яка подекуди має монструозні масштаби.

Для спрощення роботи кожна назва форми має свій унікальний короткий запис. Наприклад форма “Демографія” скорочується до “DEM”, форма “Життєво-важливі ознаки” скорочується до “VTLS”, форма “Місцевий аналіз сечі” має скорочення “URNL”, а форма “Побічна подія або супутнє захворювання” скорочується до “AE”.

Аналогічно до назв форм кожен пункт, тест, питання у формі мають свої короткий запис або з додатковою інформацією. Так пункти з форми “Демографія” кодуються наступним чином: день народження має короткий запис “BRTHD”. Вік має запис “AGEIC”. Стать відповідно записана як “SEX” і т.д. Пункти з форми “Життєво-важливі ознаки” маю наступні назви: систолічний кров’яний тиск має запис такий як “SYSBP”, діастолічний кров’яний тиск має скорочення у вигляді

“DIABP”, пульс записується як “PULSE”, температура як “TEMP”, а частота дихання як “RESP”. Можна прослідкувати логіку подібних скорочень, яка базується на подібному методі кодування змінних у доменах CDISC.

Узагальнюючи інформацію про форму та пункти у формі можна прийти до висновку, що форму можна представити як рядок, або декілька рядків, унікальних записів. Приведення форм CRF до такого формату запису підготує форми до використання для навчання штучного інтелекту.

Загального визначення, яке б описувало всі аспекти, що таке штучний інтелекту певно не існує, тому буде наведено три варіанти які покривають більш розповсюджені напрямки.

Штучний інтелект — це галузь науки, яка займається створенням комп’ютерів і машин, здатних міркувати, вчитися та діяти таким чином, що зазвичай потребує людського інтелекту або включає дані, масштаб яких перевищує те, що людина може проаналізувати.[39]

Штучний інтелект — це широка сфера, яка охоплює багато різних дисциплін, включаючи інформатику, аналітику даних і статистику, розробку апаратного та програмного забезпечення, лінгвістику, нейронауку та навіть філософію та психологію.[39]

На операційному рівні для використання в бізнесі штучний інтелект - це набір технологій, які базуються переважно на машинному та глибинному навчанні, що використовуються для аналізу даних, прогнозування та передбачення, категоризації об'єктів, обробки природної мови, рекомендацій, інтелектуального пошуку даних та багато іншого. [39]

Розглядаючи напрямки розвитку штучного інтелекту через призму мети даної роботи виділяється такий напрямок як штучний інтелект з обмеженою пам'яттю. Подібний напрямок розвитку вдосконалюється з часом, навчаючись на даних що оновлюються, за допомогою навчальної моделі.

До цього напрямку розвитку відноситься такий тип штучного інтелекту, як машинне навчання. Машинне навчання - це тип штучного інтелекту, що використовуючи алгоритм навчання на даних для отримання результату.

Машинне навчання розділяють на три моделі навчання:

Поглиблене навчання (Reinforced Learning) - це модель машинного навчання, принцип роботи якої можна описати як «навчання на практиці». Принцип поглибленого навчання називається цикл зворотного зв'язку, який полягає у тому, що модель вчиться виконувати певне завдання методом спроб і помилок, поки її продуктивність не буде в межах бажаного діапазону. Процес визначення правильних та неправильних рішень відбувається за допомогою відгуків, тобто якщо модель виконує завдання правильно то отримує позитивний відгук і негативний коли робить неправильно.

Контрольоване навчання (Supervised Learning) - це модель машинного навчання, яка використовує підготовлені навчальні дані для того щоб спрогнозувати результат на тестових даних з подальшим використанням цієї моделі на актуальних даних. Простіше кажучи, щоб навчити алгоритм розпізнавати щось зображення, текст, числа, потрібно подавати йому відповідно підготовлені зображення, тести, числа з поясненням що це має бути.

Навчання без нагляду (Unsupervised Learning) - це модель машинного навчання, яка вивчає закономірності на основі непідготовлених даних. На відміну від контрольованого навчання, результат такого навчання може бути непередбачуваним аж до моменту реалізації алгоритму. Замість цього алгоритм навчається на основі даних, класифікуючи їх на групи на основі атрибутів.

Відповідно до мети роботи та даних що були підготовлені для реалізації мети, був обраний тип машинного навчання контрольоване навчання.

Контрольоване навчання поділяється на дві групи моделей: лінійні та деревоподібні.

Алгоритми лінійної моделі контрольованого навчання очікують, що вхідні дані є лінійною комбінацією. Тобто, лінійні моделі створюють прогноз для даних в яких є лінійна залежність. З найрозповсюджених лінійних моделей виділяють чотири алгоритми.

- Лінійна регресія. Лінійна регресія це простий алгоритм, який моделює лінійну залежність між вхідними даними та нескінченно можливими вихідними даними. Переваги такого алгоритму полягають у тому що метод простий у розумінні, результат є зрозумілим відповідно до вихідного коефіцієнту, один з найшвидших алгоритмів для навчання. Недоліками такого алгоритму є очікування лінійної залежності між вхідними даними та вихідними, чутливість до викидів, погано працюють з малими даними високої розмірності.

- Логістична регресія. Логістична регресія це простий алгоритм, який моделює лінійну залежність між вхідними та категоріальними вихідними даними. Перевагами такого алгоритму є можливість інтерпретації, результат можна пояснити, легко регулюється, застосовується для багато-класових прогнозів. Недоліками такого алгоритму є очікування лінійної залежності між вхідними даними та вихідними, може працюють з малими даними високої розмірності.

- Хребтова регресія. Хребтова регресія полягає у зменшені коефіцієнту, по квадрату коефіцієнтів, ближче до нуля для ознак які мають низькі прогностичні результати. Перевагами такого алгоритму є невелика схильність до надмірного підналаштування, добре підходить для даних з мультиколінеарністю, можливість інтерпретації, результат можна пояснити. Недоліками даного алгоритму є невибірковість ознак та всі передбачення зберігаються у фінальній моделі.

- Регресія Лассо. Регресія Лассо полягає у зменшенні коефіцієнту, по величині коефіцієнтів, ближче до нуля для ознак які мають низькі

прогностичні результати. Перевагами такого алгоритму є невелика схильність до надмірного підналаштування, може обробляти дані високої розмірності. Недоліками даного алгоритму є можливість до поганої інтерпретованості, оскільки може зберігати сильно корельовані змінні.

Деревоподібні моделі алгоритмів контрольованого навчання полягає у прогнозуванні на основі дерев рішень, тобто моделі використовують серію правил “якщо-тоді” для прогнозу результату. З найрозповсюджених деревоподібних моделей виділяють п'ять алгоритми.

- Дерево рішень. Моделі дерева рішень створюють правила прийняття рішень на основі ознак для отримання прогнозів. Перевагами такого алгоритму є можливість інтерпретації, результат можна пояснити, має можливість працювати з відсутніми значеннями. Недоліками такого алгоритму є чутливість до відхилень та схильність до перенавантаження.

- Випадкові ліси. Фактично, алгоритм випадкові ліси є методом ансамблевого навчання, який побудований на багатьох деревах рішень результати прогнози яких об’єднуються для виведення загального прогнозу. Перевагами такого алгоритму є вища точність у порівнянні з іншими моделями та зменшення ефекту перенавчання. Недоліками такого алгоритму є можлива відсутність інтерпретації результату та можлива висока складність навчання.

Для вирішення поставлених задач був обраний тип машинного навчання контрольоване навчання, а саме алгоритм випадкового лісу. Алгоритм випадкового ліса та адаптовані метадані CRF у форматі даних підходящих для аналізу дають можливість вирішити проблему ідентифікації домену невідомої форми CRF.

2.4. Описання та обґрунтування алгоритмів та результатів дослідження

Адаптована інформація CRF була оформлена у вигляді набору даних з чотирнадцятьма змінними. Перша зміна під назвою pakname зберігає у собі інформацію про назву форми у скороченому варіанті.

Змінні x1-x12 зберігають інформацію про пункти з CRF у скороченому варіанті. Кількість змінних була обрана, як оптимальна між формами з малою кількістю пунктів, а саме три, та формами найбільшою кількістю пунктів, більше ста. Остання змінна, під назвою Domain, зберігає інформацію про цільовий домен який був визначений спеціалістами.

Було прийнято рішення форми з великою кількістю пунктів, більше 12, поділити на декілька строчок. Тобто форма з 100 пунктів буде займати 108 / 12 = 9 строчок. Форми з кількістю пунктів менше 3, не розглядаються у даній роботі про причині унікальності. Порожні записи у строчка будуть замінені на назву цільового домена.

Мовою реалізації роботи був обраний Python. Python був обраний через можливість підтримки модулів та пакетів модулів, що сприяє модульності та повторному використанню коду.

Серед основних переваг мови можна назвати такі:

- чистий синтаксис (для виділення блоків слід використовувати відступи);
- переносимість програм (що властиве більшості інтерпретованих мов);
- стандартний дистрибутив має велику кількість корисних модулів (включно з модулем для розробки графічного інтерфейсу);
- можливість використання Python в діалоговому режимі (дуже корисне для експериментування та розв'язання простих задач);
- стандартний дистрибутив має просте, але разом із тим досить потужне середовище розробки, яке називається IDLE і яке написане мовою Python;

- зручний для розв'язання математичних проблем (має засоби роботи з комплексними числами, може оперувати з цілими числами довільної величини, у діалоговому режимі може використовуватися як потужний калькулятор);
- відкритий код (можливість редагувати його іншими користувачами).

Алгоритм випадкових лісів був реалізований за допомогою бібліотеки Python sklearn яка містить в собі готовий варіант реалізації алгоритму під назвою RandomForestClassifier.

Набір даних був розділений на набір даних для тренування - 80% та набір даних для тестування - 20%.

Відсоток співпадіння 98,3%, що є гарним рівнем.

Результат роботи представлений у вигляді теплової мапи.

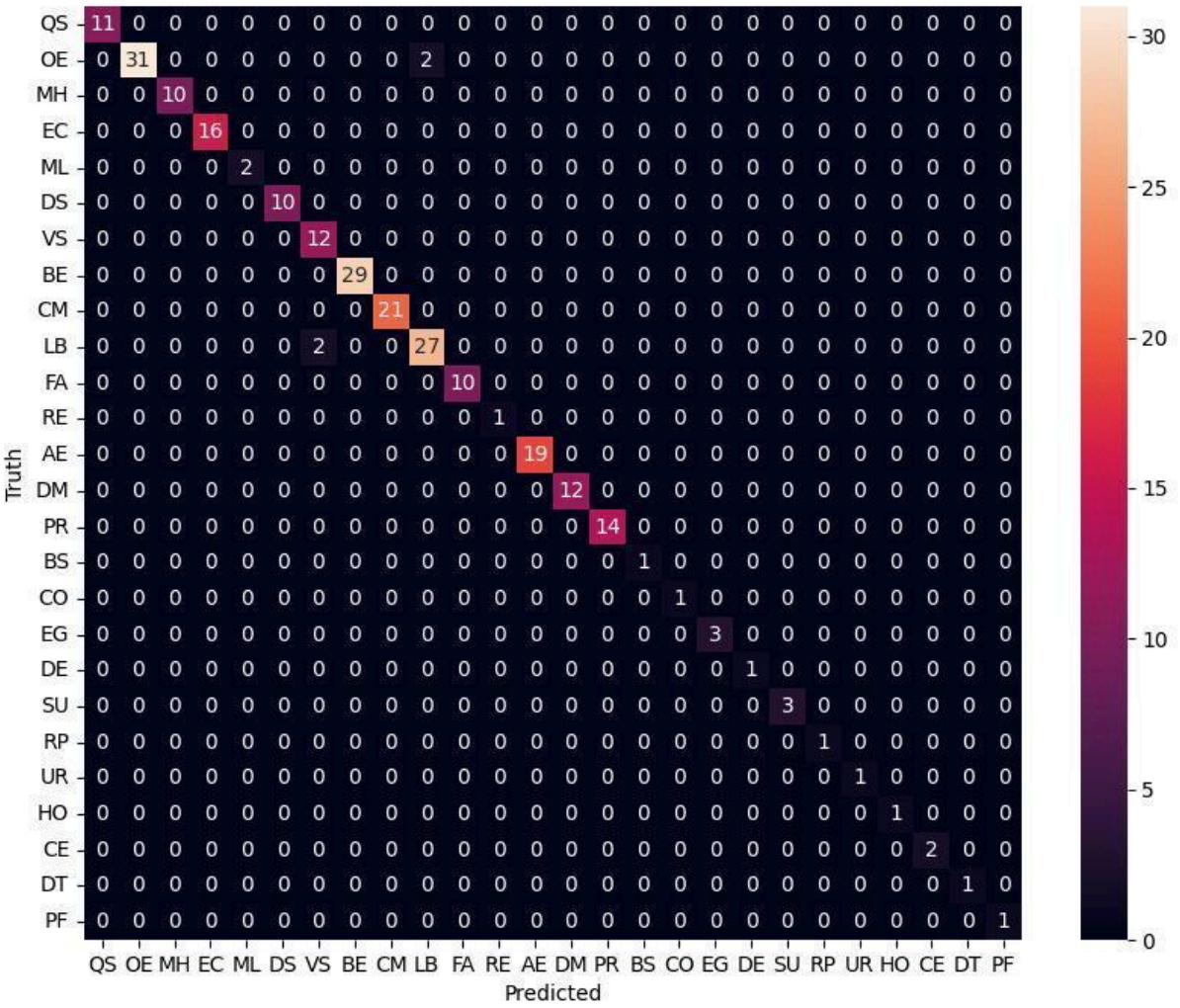


Рис.2.2 - Графічне відображення результатів.

На графічному відображенні результату дослідження можна побачити, що декілька доменів були спрогнозовано помилково. Це може бути викликано тим, що є зміні в яких не змінюється опис, але змінюється табулювання. Наприклад змінна “Not done” в усіх формах буде записана саме так, але табулювання буде залежати від того в які того в який домен йдуть пов’язані з ним зміні.

2.5. Аналіз результатів

Отримані результати свідчать про те що задача дослідження, а саме ідентифікація цільового домену невідомої форми, була виконана.

Машинне навчання має місце для застосування у створенні моделей табулювання даних дослідження. Можна чітко бачити, що відсоток співпадіння має високий рівень відповідності, а час на розуміння до якого домену відноситься форма скоротився від днів до секунд.

Система має потенціал до розвитку та покращення до стану коли 80% роботи над створенням моделі табулювання даних дослідження буде переведено на потужності штучного інтелекту, а експерту залишиться тільки перевірка і складні питання.

ВИСНОВКИ

1. Впровадження машинного навчання в клінічне статистичне програмування на основі стандартів CDISC моделі табуляції даних дослідження (SDTM) можливе і має потенціал до розвитку.

2. Алгоритм випадково лісу показав прекрасні результати при застосуванні на практиці для вирішення питання визначення цільового домену форми.

3. Методи штучного інтелекту мають потенціал на покращення процесу створення SDTM

СПИСОК ЛІТЕРАТУРИ

1. Collier R. Legumes, lemons and streptomycin: A short history of the clinical trial. CMAJ. 2009;180: 23–24.
2. Shantala Bellary, Binny Krishnankutty and M.S.Latha. Basics of case report form designing in clinical research. Perspectives in Clinical Research. 2014 Oct-Dec; 5(4): 159-166.
3. ICH Harmonised Guideline. Integrated Addendum to ICH E6(R1): Guideline for Good Clonical practice E6(R2). Step 4 version. 9 Nov 2016: 3
4. Chaitanya Shivade, Courtney Hebert, Kelly Regan, Eric Fosler-Lussier, Albert M. Lai. Automatic data source identification for clinical trial eligibility criteria resolution. AMIA Annual Symposium Proceedings. 2016; 2016:1149-1158. PMCID: PMC5333255
5. Kevin Zhang, Dina Demner-Fushman. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. Journal of the American Medical Informatics Association: JAMIA. 2017;24(4):781–787. doi: 10.1093/JAMIA/OCW176.
6. Tian Kang, Shaodian Zhang, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association: JAMIA. 2017;24(6): 1062-1071. doi: 10.1093/JAMIA/OCX019.
7. Feisheng Zhong, Jing Xing, Xutong Li and other. Artificial intelligence in drug design. Science China Life Sciences. 2018;61(10):1191–1204. doi: 10.1007/S11427-018-9342-2.
8. Long Chen, Yu Gu, Xin Ji, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. Journal of the

American Medical Informatics Association: JAMIA. 2019;26(11):1218–1226. doi: 10.1093/JAMIA/OCZ109

9. V G Vinod Vydiswaran, et al. Hybrid bag of approaches to characterize selection criteria for cohort identification. Journal of the American Medical Informatics Association: JAMIA. 2019;26(11):1172–1180. doi: 10.1093/JAMIA/OCZ079.

10. Chayakrit Krittanawong, Kipp W Johnson, WH Wilson Tang. How artificial intelligence could redefine clinical trials in cardiovascular medicine: lessons learned from oncology. Personalized Medicine. 2019;16(2):87–92. doi: 10.2217/PME-2018-0130

11. Stefan Harrer, Pratik Shah, Bhavna Antony, Jianying Hu. Artificial Intelligence for Clinical Trial Design. Trends in Pharmacological Sciences. 2019;40(8):577–591. doi: 10.1016/J.TIPS.2019.05.005

12. Anna O Basile, Alexandre Yahi, Nicholas P Tatonetti. Artificial Intelligence for Drug Toxicity and Safety. Trends in Pharmacological Sciences. 2019;40(9):624–635. doi: 10.1016/J.TIPS.2019.07.005.

13. Coryandra Gilvary, Neel Madhukar, Jamal Elkhader, Olivier Elemento. The missing pieces of Artificial Intelligence in Medicine. Trends in Pharmacological Sciences. 2019;40(8):555–564. doi: 10.1016/J.TIPS.2019.06.001.

14. Kien Wei Siah, Sean Khozin, Chi Heem Wong, and Andrew W. Lo. Machine-learning and stochastic Tumor Growth Models for Predicting Outcomes in patients with Advanced Non-Small-Cell Lung Cancer. JCO Clinical Cancer Informatics. 2019;3(3):1–11. doi: 10.1200/CCI.19.00046.

15. Alex Zhavoronkov, Quentin Vanhaelen, Tudor I Oprea. Will Artificial Intelligence for Drug Discovery Impact Clinical Pharmacology? Clinical Pharmacology and Therapeutics. 2020;107(4):780–785. doi:10.1002/CPT.1795

16. Neetu Sangari, Yanzhen Qu. A Comparative Study on Machine Learning Algorithms for Predicting Breast Cancer Prognosis in Improving Clinical Trials. Published: 2020 International Conference on Computational Science and Computational Intelligence (CSCI); 2020:813–818. doi:10.1109/CSCI51800.2020.00152

17. Cecilia S. Lee, Aaron Y. Lee. How Artificial Intelligence Can Transform Randomized Controlled Trials. Translational Vision Science & Technology. 2020;9(2). doi:10.1167/TVST.9.2.9

18. Alexander V. Schperberg, Amélie Boichard, Igor F. Tsigelny, Stéphane B. Richard, Razelle Kurzrock. Machine learning model to predict oncologic outcomes for drugs in randomized clinical trials. 2020;147(9):2537–2549. doi: 10.1002/ijc.33240

19. Felipe Feijoo, Michele Palopoli, Jen Bernstein, Sauleh Siddiqui, Tenley E. Albright. Key indicators of phase transition for clinical trials through machine learning. Drug Discovery Today. 2020;25(2):414–421. doi: 10.1016/J.DRUDIS.2019.12.014.

20. Janette Vazquez, Samir Abdelrahman, Loretta M. Byrne, et al. Using supervised machine learning classifiers to estimate likelihood of participating in clinical trials of a de-identified version of ResearchMatch. Journal of Clinical and Translational Science. 2020;5(1). doi: 10.1017/CTS.2020.535.

21. Manon Ansart, Stephane Epelbaum, et al. Reduction of recruitment costs in preclinical AD trials: validation of automatic pre-screening algorithm for brain amyloidosis. Statistical Methods in Medical Research. 2020;29(1):151–164. doi: 10.1177/0962280218823036

22. J. Thaddeus Beck, et al. Artificial Intelligence Tool for optimizing eligibility screening for clinical trials in a large Community Cancer Center. JCO Clinical Cancer Informatics. 2020;4(4):50–59. doi: 10.1200/CCI.19.00079

23. Gaspar Delso, Davide Cirillo, Joshua D Kaggie, Alfonso Valencia, Ur Metser, Patrick Veit-Haibach. How to design AI-Driven clinical trials in Nuclear Medicine. Seminars Nuclear Medicine. 2021;51(2):112–119. doi: 10.1053/J.SEMNUCLMED.2020.09.003.

24. Allison Gates, Michelle Gates, Shannon Sim, Sarah A Elliott, Jennifer, Lisa Hartling. Creating efficiencies in the extraction of data from randomized trials: a prospective evaluation of a machine learning and text mining tool. BMC Medical Research Methodology. 2021;21(1):169. doi: 10.1186/s12874-021-01354-2.

25. Clinical Data Interchange Standards Consortium (CDISC). Therapeutic Areas by Disease Area | CDISC. Accessed October 25., 2021. https://www.cdisc.org/standards/therapeutic-areas/disease-area

26. Likhitha Kolla, Fred K. Gruber, Omar Khalid, Colin Hill, Ravi B. Parikh. The case for AI-driven cancer clinical trials - the efficacy arm in silico. Biochimica et Biophysica Acta - Reviews on Cancer. 2021;1876(1). doi: 10.1016/J.BBCAN.2021.188572.

27. E. Hope Weissler, Tristan Naumann, Tomas Andersson, et al. The role of machine learning in clinical research: transforming the future of evidence generation. Trials. 2021;22(1). doi: 10.1186/S13063-021-05489-X.

28. Danielle Beaulieu, James D. Berry, Sabrina Paganoni, et al. Development and validation of a machine-learning ALS survival model lacking vital capacity (VC-Free) for use in clinical trials during the COVID-19 pandemic. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2021;22(1):22–32. doi: 10.1080/21678421.2021.1924207.

29. Miao Qi, Owen Cahan, et al. Quantifying representativeness in randomized clinical trials using machine learning fairness metrics. JAMIA Open. 2021;4(3). doi: 10.1093/JAMIAOPEN/OOAB077.

30. Hao Liu, Yuan Chi, Alex Butler, Yingcheng Sun, Chenhua Weng. A knowledge base of clinical trial eligibility criteria. Journal of Biomedical Informatics. 2021;117. doi: 10.1016/J.JBI.2021.103771

31. Jelena Gligorijevic and other. Optimizing clinical trials recruitment via deep learning. Journal of the American Medical Informatics Association: JAMIA. 2019;26(11):1195–1202. doi: 10.1093/JAMIA/OCZ064.

32. E. Hope Weissler, Tristan Naumann, Tomas Andersson, et al. The role of machine learning in clinical research: transforming the future of evidence generation. Trials. 2021;22(1). doi: 10.1186/s13063-021-05489-x.

33. Kennath Getz, Zachary Smith, et al. Assessing the scope and predictors of intentional dose non-adherence in clinical trials. Therapeutic Innovation and Regulatory Science. 2020;54(6):1330–1338. doi: 10.1007/S43441-020-00155-X.

34. Vidya Koesmahargyo, Anzar Abbas, Li Zhang, et al. Accuracy of machine learning-based prediction of medication adherence in clinical research. Psychiatry Research. 2020;294. doi: 10.1016/J.PSYCHRES.2020.113558

35. Irene Mayorga-Ruiz, Ana Jiménez-Pastor, Belen Fos-Guarinos, et al. The role of AI in clinical trials. Artificial Intelligence in Medical Imaging. Published online January. 2019; 29:231–243. doi: 10.1007/978-3-319-94878-2_16.

36. Benjamin A. Goldstein; Joseph Rigdon. Using machine learning to identify Heterogeneous Effects in Randomized clinical trials-moving beyond the forest plot and into the forest. JAMA Netw open. 2019;2(3). doi: 10.1001/JAMANETWORKOPEN.2019.0004.

37. William R. Zame, Ioana Bica, et al. Machine learning for clinical trials in the era of COVID-19. Statistics in Biopharmaceutical Research. 2020;12(4):506–517. doi: 10.1080/19466315.2020.1797867.

38. Nina Zhou, Paul Manser. Does including machine learning predictions in ALS clinical trial analysis improve statistical power? Annals of Clinical and Translational Neurology. 2020;7(10):1756–1765. doi: 10.1002/ACN3.51140.

39. What is Artificial Intelligence (AI)? https://cloud.google.com/learn/what-is-artificial-intelligence