

Одеський національний університет імені І. І. Мечникова
Факультет математики, фізики та інформаційних технологій
Кафедра оптимального керування і економічної кібернетики

Кваліфікаційна робота

на здобуття ступеня вищої освіти «магістр»

**«Дослідження та прогнозування ризиків
вітамін-D-дефіцитних станів на основі моделей
машинного навчання»**

**«Research and forecasting of the risk of vitamin D
deficiency conditions using machine learning models»**

Виконав(ла): здобувач(ка) денної форми навчання
спеціальності 113 Прикладна математика
Освітня програма «Прикладна математика»
Корхова Аріна Сергіївна

Керівник: канд. фіз.-мат. наук, доц. Страхов Є. М. _____
Рецензент: канд. фіз.-мат. наук, доц. Яровий А. Т. _____

Рекомендовано до захисту:
Протокол засідання кафедри
№ ____ від _____ 2025 р.
Завідувач кафедри

Захищено на засіданні ЕК № _____
Протокол № ____ від _____ 2025 р.
Оцінка _____ / _____ / _____
Голова ЕК

ЗМІСТ

Вступ	3
1 МАТЕРІАЛИ ТА МЕТОДИ	5
2 ОГЛЯД ПОПЕРЕДНІХ РЕЗУЛЬТАТІВ	7
2.1 Характеристика набору даних та розвідувальний аналіз . . .	7
2.2 Результати математичного моделювання методами регресійного аналізу	7
2.3 Ефективність застосування методів машинного навчання . .	8
2.4 Висновки за результатами попереднього етапу	9
3 ОЦІНКА ВУГЛЕВОДНОГО ОБМІНУ ТА ЙОГО ВЗАЄМОЗВ'ЯЗОК ІЗ РИЗИКОМ ВІТАМІН-D-ДЕФІЦИТНИХ СТАНІВ	10
3.1 Теоретичні основи та математичний вираз індексу НОМА-IR	10
3.2 Математичні методи аналізу та моделювання взаємозв'язку НОМА-IR та вітаміну D	11
3.3 Розрахунок індексу НОМА-IR для вибірки дослідження та аналіз розподілу	13
4 ВИЯВЛЕННЯ ПРИЧИННО-НАСЛІДКОВИХ ЗВ'ЯЗКІВ ЗА ДОПОМОГОЮ МЕТОДУ DOUBLE MACHINE LEARNING	18
4.1 Математичні основи методу Double Machine Learning	18
4.1.1 Постановка задачі каузального виведення	18
4.1.2 Проблема конфаундингу та регуляризаційного зміщення	18
4.1.3 Частково лінійна модель	19
4.1.4 Неуман-ортогональність	19
4.1.5 Алгоритм Double Machine Learning	19
4.1.6 Асимптотичні властивості	20
4.1.7 Оцінювання стандартних похибок	21
4.1.8 Узагальнення: Interactive Model	21
4.2 Специфікація каузальних задач дослідження	21
4.2.1 Причинний вплив інсулінорезистентності на рівень вітаміну D	22
4.2.2 Причинний вплив ожиріння на рівень вітаміну D . . .	25
4.2.3 Гетерогенність ефекту за ІМТ	28
Висновки	31
Список літератури	33

ВСТУП

Актуальність. Вітамін-D-дефіцитні стани є однією з найпоширеніших форм гіповітамінозу у світовому масштабі, що становить серйозну загрозу для громадського здоров'я. За даними глобального аналізу 7.9 мільйонів учасників, проведеного у період з 2000 по 2022 рік, 15.7% населення світу мають рівень 25-гідроксिवітаміну D нижче 30 нмоль/л, 47.9% – нижче 50 нмоль/л, а 76.6% – нижче 75 нмоль/л [1]. Це свідчить про те, що більшість населення планети перебуває в стані субоптимального забезпечення вітаміном D, що має далекосяжні наслідки для здоров'я.

Вітамін D відіграє критичну роль не лише у підтримці мінералізації кісткової тканини та гомеостазу кальцію і фосфору, але й у регуляції імунної системи, серцево-судинного здоров'я та метаболічних процесів. Дефіцит цього вітаміну асоціюється з підвищеним ризиком розвитку рахіту та остеомалії, а також численних позаскелетних патологій, включаючи інфекційні захворювання, онкологічні процеси, цукровий діабет, серцево-судинні захворювання та автоімунні розлади.

Серед американського населення поширеність важкого та помірного дефіциту вітаміну D становить 2.6% та 22.0% відповідно, а недостатності – 40.9% [2]. Особливо вразливими групами є жінки, представники неіспанської чорношкірої популяції, особи віком 20-29 років, а також люди в зимовий період. У пацієнтів з хронічною нирковою недостатністю, реципієнтів після трансплантації органів та осіб з печінковими захворюваннями поширеність дефіциту вітаміну D може сягати 85-99% [3].

Епідеміологічні дослідження демонструють, що проблема вітамін-D-дефіцитних станів має виражену географічну, етнічну та соціально-економічну нерівномірність. Жителі високих широт, країн з низьким та середнім рівнем доходу, а також жінки репродуктивного віку потребують особливої уваги. Існує значна варіабельність показників залежно від пори року, способу життя, дієтичних звичок, використання сонцезахисних засобів та культурних особливостей одягу. Доведено зв'язок рівнів вітаміну D з тригліцеридами (ТГ), коефіцієнтом атерогенності (КА), ліпопротеїдами дуже низької щільності (ЛПДНЩ), ліпопротеїдами низької щільності (ЛПНЩ) та ліпопротеїдами високої щільності (ЛПВЩ).

Традиційні підходи до виявлення дефіциту вітаміну D базуються на визначенні рівня 25-гідроксिवітаміну D у сироватці крові, що вважається найбільш надійним біомаркером вітамінного статусу. Проте масове скринування населення є економічно недоцільним через високу вартість лабораторних досліджень та обмежену доступність тестування в ресурсо-обмежених регіонах. Це створює потребу у розробці альтернативних методів ідентифікації осіб з високим ризиком розвитку дефіциту вітаміну D.

У контексті цих викликів стрімко зростає інтерес до застосування методів машинного навчання для прогнозування ризиків вітамін-D-дефіцитних станів. Машинне навчання базується на великих масивах даних для розробки надійних моделей ризику та прогнозування стану здоров'я людини [4]. Сучасні дослідження демонструють, що моделі машинного навчання можуть з високою точністю передбачати вітамінний статус на основі легкодоступних демографічних, антропометричних, поведінкових та клінічних параметрів, які можуть бути отримані під час простого опитування чи огляду.

Переваги застосування машинного навчання у прогнозуванні вітамін-D-дефіцитних станів включають: можливість обробки великих обсягів різнорідних даних; виявлення складних нелінійних взаємозв'язків між факторами ризику; автоматичний відбір найбільш інформативних предикторів; здатність адаптуватися до специфіки різних популяцій; можливість інтеграції у системи клінічної підтримки прийняття рішень.

Розробка точних прогностичних моделей на основі машинного навчання дозволить: оптимізувати розподіл ресурсів охорони здоров'я шляхом таргетного скринування груп ризику; знизити витрати на непотрібне масове тестування; забезпечити ранню ідентифікацію осіб, які потребують профілактичних або терапевтичних втручань; персоналізувати підходи до попередження та корекції дефіциту вітаміну D; створити інструменти для використання у первинній ланці охорони здоров'я та громадських скринінгових програмах.

Таким чином, дослідження та прогнозування ризиків вітамін-D-дефіцитних станів на основі моделей машинного навчання є актуальним напрямком на перетині прикладної математики, медичної інформатики та громадського здоров'я. Це дослідження має потенціал внести значний вклад у вирішення глобальної проблеми дефіциту вітаміну D шляхом створення ефективних, доступних та масштабованих інструментів для ідентифікації груп ризику та оптимізації профілактичних стратегій.

Мета. Створення математичної моделі, яка дозволяє спрогнозувати вплив кожного із лабораторних і антропометричних показників на рівень 25(OH)D сироватки крові та у подальшому дозволить розробити систему ранньої діагностики та профілактики вітамін D дефіцитних станів.

Предмет дослідження. Рівень вітаміну D серед дорослого населення жителів Півдня України.

Об'єкт дослідження. Моделі машинного навчання для визначення рівня 25(OH)D.

Методи дослідження. Методи розвідувального та кореляційного аналізу, методи машинного навчання та штучного інтелекту.

РОЗДІЛ 1

МАТЕРІАЛИ ТА МЕТОДИ

Вибіркову сукупність дослідження склали 928 мешканців півдня України (Одеська, Миколаївська та Херсонська області). Серед обстежених було 507 жінок (середній вік 47.7 ± 15.3 років) та 421 чоловіків (46.7 ± 15.5 років). Віковий діапазон респондентів становив від 19 до 82 років.

Антропометричне обстеження включало розрахунок індексу маси тіла (ІМТ) за формулою: маса тіла (кг) / зріст (м^2) відповідно до рекомендацій Міжнародної групи з питань ожиріння ВООЗ (1997). Наявність абдомінального ожиріння верифікували при співвідношенні обводу талії до обводу стегон $> 0,8$.

Лабораторний етап дослідження передбачав визначення концентрації сумарного вітаміну 25(OH)D (фракції D2 та D3) на автоматичному імунохімічному аналізаторі Architect i2000sr (лабораторія ТОВ «СМАРТЛАБ»). Показники ліпідного профілю — загальний холестерин (ЗХС), тригліцериди (ТГ), холестерин ліпопротеїдів високої (ЛПВЩ), низької (ЛПНЩ) та дуже низької щільності (ЛПДНЩ), а також коефіцієнт атерогенності (КА) — визначали ферментативно-колориметричним методом за допомогою біохімічного аналізатора Cobas 6000 (Roche Diagnostics) [5].

Збір даних тривав упродовж календарного року, що дозволило нівелювати вплив сезонних коливань інсоляції на рівень вітаміну D. Формування групи спостереження здійснювалося з урахуванням критеріїв виключення. На основі попереднього анкетування з дослідження були вилучені особи з аутоімунними та онкологічними захворюваннями, хронічною патологією печінки та нирок, вагітні та жінки в період лактації. Також виключалися респонденти, які приймали вітамін D або препарати, що впливають на метаболічні процеси (глюкокортикоїди, замісна гормональна терапія, антиконвульсанти тощо).

Інтерпретацію отриманих показників 25(OH)D проводили відповідно до клінічних настанов Endocrine Practice Guidelines Committee та Institute of Medicine (Таблиця 1.1).

Табл. 1.1. Рівні 25(ОН)D сироватки крові відповідно до клінічних настанов Endocrine Practice Guidelines Committee та Institute of Medicine

Значення	Рівень 25(ОН)D сироватки крові
Дефіцит	<20 нг/мл (<50 нмоль/л)
Недостатність	21-29 нг/мл (51-74 нмоль/л)
Достатній рівень	≥30 нг/мл (≥75 нмоль/л)
Токсичний рівень	>150 нг/мл (>375 нмоль/л)

Математична обробка даних та побудова прогностичних моделей виконувалися мовою програмування Python із використанням бібліотек для наукових обчислень та машинного навчання: NumPy [6], Pandas [7], SciPy [8], Scikit-learn [9], SHAP [10], а також інструментів візуалізації Matplotlib [11] та Seaborn [12].

Статистичний аналіз включав розрахунок описових статистик та перевірку розподілу на нормальність. Для виявлення взаємозв'язків застосовували кореляційний та регресійний аналізи. Прогнозування здійснювалося за допомогою моделей багатофакторної лінійної та логістичної регресії. Якість лінійних моделей оцінювали за скоригованим коефіцієнтом детермінації (R^2), а силу впливу предикторів — за стандартизованими коефіцієнтами регресії та еластичності. Для валідації моделей логістичної регресії використовували ROC-аналіз та метрики матриці помилок. Оцінка моделей на основі дерев рішень проводилася за допомогою метрики асигуру та коефіцієнта Джині. Різницю вважали статистично значущою при $p < 0,001$. Для виявлення причинно-наслідкових зв'язків та оцінки каузальних ефектів застосовано метод Double Machine Learning (DML), що базується на принципі Ньюман-ортогональності. В рамках DML оцінювання «заважаючих» параметрів (nuisance parameters) здійснювалося з використанням ансамблевих методів машинного навчання, зокрема Random Forest та Gradient Boosting, а для запобігання перенавчанню використовувалася процедура перехресної підгонки (cross-fitting). Аналіз гетерогенності ефектів впливу проводився шляхом оцінки умовного середнього причинного ефекту (CATE). Інтерпретація моделей та оцінка внеску окремих ознак у прогноз здійснювалася за допомогою методу адитивних пояснень Шеплі (SHAP values).

РОЗДІЛ 2

ОГЛЯД ПОПЕРЕДНІХ РЕЗУЛЬТАТІВ

2.1 Характеристика набору даних та розвідувальний аналіз

На попередньому етапі роботи було проведено комплексне дослідження факторів, що впливають на рівень 25-гідроксивітаміну D (25(OH)D) у сироватці крові дорослого населення Півдня України. Вибірка складалася з 928 осіб (507 жінок та 421 чоловік) віком від 19 до 82 років, які проживають у Херсонській, Миколаївській та Одеській областях. У ході первинного розвідувального аналізу було встановлено високу поширеність дефіцитних станів: лише 33.4% досліджуваних мали достатній рівень вітаміну D, тоді як дефіцит та недостатність спостерігалися у 33.6% та 33% випадків відповідно [5].

Ключовим етапом попереднього дослідження став кореляційний аналіз антропометричних показників. Було виявлено статистично значущий зворотний зв'язок між індексом маси тіла (ІМТ) та рівнем 25(OH)D (коефіцієнт Спірмена $r = -0.181$, $p < 0.001$). Аналіз середніх значень підтвердив, що пацієнти з ожирінням мають достовірно нижчий рівень вітаміну D порівняно з особами, що мають нормальну вагу. Також оцінювався вплив абдомінального ожиріння через співвідношення окружності талії до окружності стегон (ОТ/ОС). Хоча кореляція виявилася слабшою ($r = -0.09$), відмінності середніх значень у групах з різним статусом вітаміну D залишалися статистично значущими.

Додатково було проаналізовано ліпідний профіль. Встановлено наявність кореляційних зв'язків між рівнем вітаміну D та показниками ліпідного обміну, зокрема прямий зв'язок з ліпопротеїдами високої щільності (ЛПВЩ) та зворотний — з коефіцієнтом атерогенності (КА) і тригліцеридами.

2.2 Результати математичного моделювання методами регресійного аналізу

На основі виявлених кореляцій було здійснено спробу побудови прогностичних моделей для визначення точного рівня 25(OH)D за допомогою багатофакторної лінійної регресії. Було протестовано 7 моделей з різними комбінаціями предикторів.

Найкращі показники адекватності продемонструвала чотирифакторна модель (Модель №4), яка включала такі змінні: вік, ІМТ, коефіцієнт атерогенності (КА) та рівень ЛПВЩ. Рівняння регресії мало вигляд:

$$25(OH)D = 53.622 - 0.342 \cdot \text{Вік} - 0.407 \cdot \text{ІМТ} - 0.631 \cdot \text{КА} + 1.004 \cdot \text{ЛПВЩ} \quad (2.1)$$

Незважаючи на статистичну значущість моделі (F -статистика = 55.39), коефіцієнт детермінації залишався низьким ($R_T^2 = 0.224$), що свідчило про те, що лінійна модель пояснює лише близько 22% дисперсії залежної змінної. Це вказало на обмеженість лінійних методів для точного прогнозування кількісного рівня вітаміну D та необхідність переходу до задач класифікації ризиків.

Наступним кроком стало застосування логістичної регресії для прогнозування бінарного результату: наявності або відсутності дефіциту вітаміну D (<30 нг/мл). Для оцінки якості класифікації використовувалися метрики Ассурасу, Precision, Recall та AUC. Найбільш оптимальною з точки зору балансу складності та точності знову виявилася модель, що включає чотири предиктори (вік, ІМТ, КА, ЛПВЩ). Площа під ROC-кривою (AUC) для цієї моделі склала 0.735, що вказує на прийнятну, але не високу якість класифікації. Чутливість моделі склала 87.8%, проте специфічність залишалася низькою (42.6%).

2.3 Ефективність застосування методів машинного навчання

З метою подолання обмежень класичних статистичних методів та врахування нелінійних взаємозв'язків у даних, у попередньому дослідженні було застосовано алгоритм дерев рішень (Decision Trees).

Оптимізація гіперпараметрів (зокрема глибини дерева) дозволила побудувати модель, яка значно перевершила результати логістичної регресії. Аналіз важливості ознак (Feature Importance) підтвердив, що найбільш впливовими предикторами є вік, ІМТ та рівень ЛПВЩ.

Результати тестування моделі дерева рішень показали суттєве покращення всіх метрик якості:

- Точність (Ассурасу) зросла до 0.914 (порівняно з 0.746 у логістичної регресії);
- Специфічність підвищилася до 0.946;
- Інтегральний показник AUC досяг значення 0.92, що на 17% вище за результати лінійних методів.

Візуалізація дерева рішень дозволила інтерпретувати правила класифікації, виділивши критичні порогові значення для віку (41.5 років) та ІМТ, які є маркерами підвищеного ризику.

За результатами проведеного дослідження були опубліковані статті «Mathematical modeling of the dependence of the risk of vitamin D deficiency on anthropometric and laboratory parameters» [13] та «Features of the use of artificial intelligence for the purpose of screening vitamin D deficiency in the adult population» [14]. Також результати роботи були апробовані на науково-практичних конференціях “Інформаційні технології і автоматизація” (Одеса, ОНТУ, 2023) [15], “Стан, досягнення та перспективи інформаційних систем і технологій” (Одеса, ОНТУ, 2024) [14] та опубліковані у вигляді тез. Результати доповідалися на науково-практичній конференції з міжнародною участю “Сучасні теоретичні та практичні аспекти клінічної медицини” (Одеса, ОНМедУ, 2023) [16]

2.4 Висновки за результатами попереднього етапу

Проведене попереднє дослідження дозволило зробити такі ключові висновки, які формують підґрунтя для даної роботи:

- Існує статистично підтверджений зв'язок між дефіцитом вітаміну D та антропометричними показниками (вік, ІМТ) і ліпідним профілем (ЛПВЩ, КА).
- Класичні методи регресійного аналізу (лінійна та логістична регресія) мають обмежену прогностичну здатність ($AUC = 0.73$) і не здатні повною мірою описати складні залежності у медичних даних.
- Застосування алгоритмів машинного навчання, зокрема дерев рішень, дозволило значно підвищити точність прогнозування ризиків ($AUC = 0.92$).

РОЗДІЛ 3

ОЦІНКА ВУГЛЕВОДНОГО ОБМІНУ ТА ЙОГО ВЗАЄМОЗВ'ЯЗОК ІЗ РИЗИКОМ ВІТАМІН-D-ДЕФІЦИТНИХ СТАНІВ

3.1 Теоретичні основи та математичний вираз індексу HOMA-IR

Індекс HOMA-IR (Homeostatic Model Assessment for Insulin Resistance) є одним із найпоширеніших непрямих методів оцінки інсулінорезистентності в клінічній практиці та наукових дослідженнях. Цей показник був запропонований Matthews et al. у 1985 році як математична модель для оцінки функції β -клітин підшлункової залози та чутливості до інсуліну на основі концентрацій глюкози та інсуліну натще [17].

HOMA-IR розраховується за формулою:

$$\text{HOMA-IR} = \frac{\text{Інсулін натще (мкОД/мл)} \times \text{Глюкоза натще (ммоль/л)}}{22.5}. \quad (3.1)$$

Ділення на константу 22.5 забезпечує нормування значень відповідно до умовної «базової» функції β -клітин.

Значення HOMA-IR $\geq 2,5$ зазвичай вказує на наявність інсулінорезистентності, хоча порогові значення можуть варіювати залежно від популяції та методів визначення інсуліну [18].

Численні епідеміологічні дослідження демонструють обернену залежність між рівнем 25-гідроксівітаміну D у сироватці крові та показниками інсулінорезистентності. Мета-аналіз, проведений Gagnon et al., який включав понад 11 000 учасників, показав, що низькі рівні вітаміну D асоційовані з підвищеним ризиком розвитку метаболічного синдрому та цукрового діабету 2 типу [19].

Дослідження показують, що при підвищенні концентрації 25(OH)D на кожні 10 нг/мл ризик розвитку цукрового діабету 2 типу знижується приблизно на 4-8% [20]. При цьому особи з рівнем вітаміну D < 20 нг/мл мають значно вищі показники HOMA-IR порівняно з тими, у кого рівень вітаміну D перевищує 30 нг/мл [21].

Важливо відзначити, що зв'язок між дефіцитом вітаміну D та інсулінорезистентністю може бути двоспрямованим. З одного боку, недостатність

вітаміну D може призводити до розвитку інсулінорезистентності через вищезгадані механізми. З іншого боку, метаболічні порушення, пов'язані з інсулінорезистентністю та ожирінням, можуть сприяти зниженню біодоступності вітаміну D через його секвестрацію в жировій тканині [22].

Включення НОМА-IR до моделей прогнозування вітамін D-дефіцитних станів обґрунтовується кількома факторами:

Патофізіологічна обґрунтованість. Як було показано вище, існує потужний біологічний зв'язок між інсуліновою чутливістю та статусом вітаміну D, що робить НОМА-IR релевантним біомаркером для прогнозування ризику дефіциту.

Висока поширеність коморбідності. Інсулінорезистентність та дефіцит вітаміну D часто співіснують, особливо в популяціях з високим ризиком метаболічних порушень, що підвищує прогностичну цінність НОМА-IR [23].

Доступність та стандартизованість. НОМА-IR є легкодоступним показником, що розраховується на основі рутинних лабораторних параметрів (глюкоза та інсулін натще), які часто визначаються в рамках стандартних біохімічних обстежень.

Незалежна прогностична цінність. Дослідження показують, що НОМА-IR може мати незалежну прогностичну цінність щодо ризику дефіциту вітаміну D навіть після поправки на інші фактори ризику, такі як індекс маси тіла, вік та рівень фізичної активності [24].

Таким чином, НОМА-IR є важливим параметром, який доцільно включати до багатофакторних моделей машинного навчання для прогнозування ризику вітамін D-дефіцитних станів, що дозволяє підвищити точність прогнозування та виявити групи високого ризику для своєчасного профілактичного втручання.

3.2 Математичні методи аналізу та моделювання взаємозв'язку НОМА-IR та вітаміну D

Для дослідження взаємозв'язку між НОМА-IR та рівнем вітаміну D використовуються різні статистичні методи залежно від розподілу даних та характеру зв'язків.

Кореляційний аналіз. Для оцінки лінійного зв'язку між НОМА-IR та концентрацією 25(OH)D застосовується коефіцієнт кореляції Пірсона (r), який визначається за формулою:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.2)$$

де x_i – значення НОМА-IR, y_i – рівень вітаміну D, $\bar{x}$ та $\bar{y}$ – середні значення відповідних показників.

У випадку ненормального розподілу даних застосовується непараметричний коефіцієнт кореляції Спірмена (ρ), який обчислюється на основі рангів значень [25].

Регресійний аналіз. Для кількісної оцінки впливу НОМА-IR на рівень вітаміну D використовується лінійна регресія:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \quad (3.3)$$

де y – залежна змінна (рівень вітаміну D), $x_1, x_2, \dots, x_p$ – незалежні змінні (включаючи НОМА-IR), $\beta_0, \beta_1, \dots, \beta_p$ – коефіцієнти регресії, ε – випадкова помилка [26].

Логістична регресія. Для прогнозування бінарного результату (наявність або відсутність дефіциту вітаміну D) використовується логістична регресія. Модель логістичної регресії має вигляд:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}} \quad (3.4)$$

де $P(Y = 1|X)$ – ймовірність наявності дефіциту вітаміну D при заданих значеннях предикторів $X = (x_1, x_2, \dots, x_p)$.

Відношення шансів (odds ratio, OR) для кожного предиктора обчислюється як:

$$OR_i = e^{\beta_i} \quad (3.5)$$

що дозволяє інтерпретувати вплив зміни кожного предиктора на ймовірність результату [27].

3.3 Розрахунок індексу HOMA-IR для вибірки дослідження та аналіз розподілу

На Рисунку 3.1 зображено розподіли індексу HOMA-IR у вибірці залежно від метаболічного статусу учасників. Для груп Insulin Resistant та Normal побудовано гістограми з накладеними оцінками густини ймовірності (KDE). Видно, що значення HOMA-IR у групі інсулінорезистентних респондентів мають більший середній рівень та ширший розкид, зміщуючись у бік вищих значень (переважно вище 3). Натомість у групі метаболічно нормальних учасників розподіл компактніший, зі значеннями, що концентруються в діапазоні приблизно 1.5–2.5. Це підтверджує, що індекс HOMA-IR добре відображає відмінності між особами з нормальною чутливістю до інсуліну та тими, хто має прояви інсулінорезистентності.

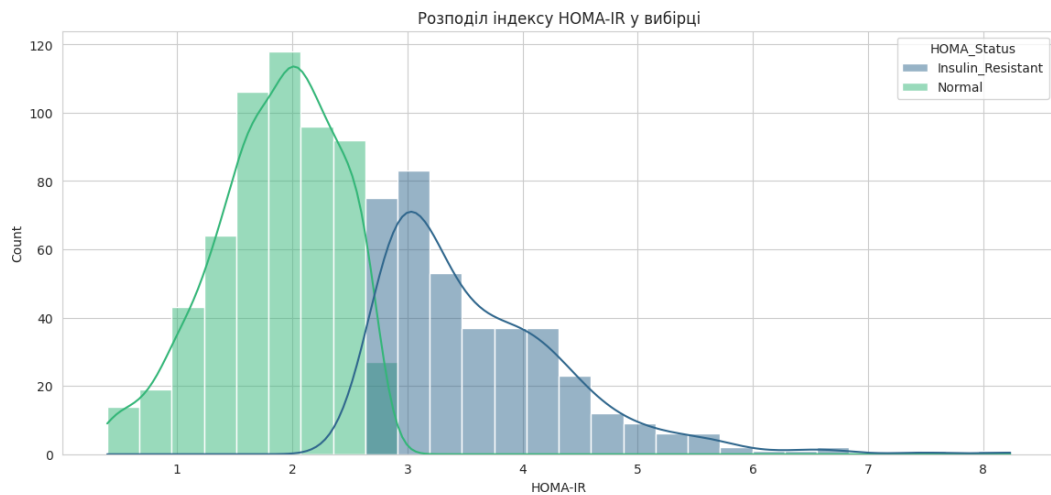


Рис. 3.1. Розподіл індексу HOMA-IR

На Рисунку 3.2 подано порівняння рівнів HOMA-IR залежно від статусу вітаміну D, представлене у форматі boxplot. Розглянуто три категорії: Deficiency, Insufficiency та Optimal. Спостерігається тенденція до підвищених медіанних значень HOMA-IR у групі з дефіцитом вітаміну D, тоді як у групах недостатності та оптимального рівня розподіли зміщені до нижчих значень. Хоча міжгрупові перекриття присутні, діаграма вказує на можливий зв'язок між нижчими концентраціями вітаміну D та більш високими показниками інсулінорезистентності. Це узгоджується з гіпотезою про потенційну роль вітаміну D у регуляції метаболічних процесів.

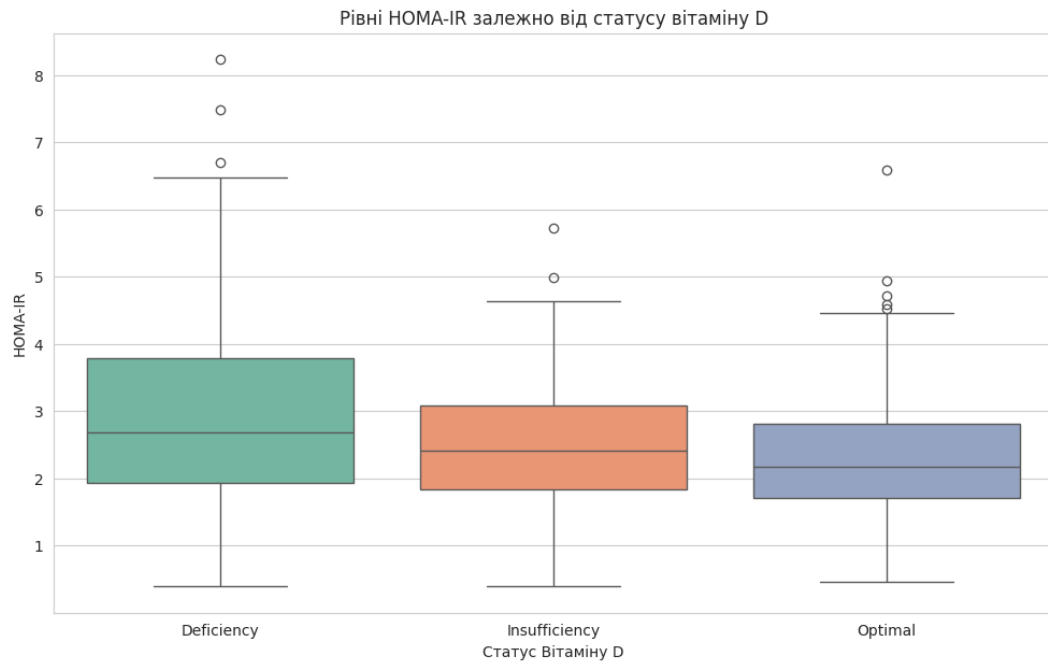


Рис. 3.2. Рівні HOMA-IR залежно від статусу вітаміну D

За результатами кореляційного аналізу Спірмена встановлено статистично значущий обернений (від'ємний) зв'язок між концентрацією вітаміну D та індексом HOMA-IR ($r = -0.187, p < 0.001$). Від'ємний знак коефіцієнта кореляції свідчить про те, що нижчі рівні вітаміну D асоціюються з вищими показниками інсулінорезистентності, і навпаки. Хоча сила зв'язку є слабкою, висока статистична значущість ($p = 4.64 \times 10^{-9}$) підтверджує не випадковий характер цієї закономірності у досліджуваній вибірці.

Для визначення відмінностей у показниках інсулінорезистентності залежно від статусу вітаміну D вибірку було розподілено на три групи: «Дефіцит» (Deficiency) (< 20 нг/мл), «Недостатність» (Insufficiency) ($20-30$ нг/мл) та «Оптимальний рівень» (Optimal) (≥ 30 нг/мл).

Застосування критерію Крускала-Уолліса виявило високозначущу різницю рівнів HOMA-IR між трьома групами ($H = 32.30, p < 0.001$). Це вказує на те, що статус вітаміну D суттєво впливає на показники вуглеводного обміну.

Для деталізації виявлених відмінностей було проведено попарні порівняння груп (post-hoc аналіз) за допомогою U-критерію Манна-Вітні. Результати показали статистично значущі відмінності у всіх парах порівняння:

- Група дефіциту демонструвала значуще вищі показники HOMA-IR порівняно з групою недостатності ($p = 0.0003$).
- Найбільш виражена різниця спостерігалася між полярними групами: пацієнти з дефіцитом мали достовірно гірші показники інсулінорезистентності порівняно з пацієнтами з оптимальним рівнем вітаміну D

($p < 0.001$).

- Різниця між групами недостатності та оптимального рівня також була статистично значущою ($p = 0.0145$), хоча і менш вираженою, ніж у попередніх парах.

Отримані дані підтверджують гіпотезу про протективну роль вітаміну D щодо розвитку інсулінорезистентності. Спостерігається чіткий градієнт: покращення статусу вітаміну D від дефіциту до оптимального рівня супроводжується статистично значущим зниженням індексу HOMA-IR.

На Рисунку 3.3 візуально простежується негативна залежність між змінними: зі зростанням HOMA-IR значення 25(OH)D мають тенденцію зменшуватися. Хоча розсіювання точок є значним, загальна траєкторія регресійної лінії свідчить про слабку, але стабільну спадну тенденцію. Це може вказувати на потенційний зворотний зв'язок між інсулінорезистентністю та рівнем вітаміну D — учасники з вищими значеннями HOMA-IR, як правило, характеризуються нижчим рівнем 25(OH)D. Такий результат відповідає гіпотезі про можливий взаємозв'язок між метаболічними порушеннями та станом забезпеченості вітаміном D, що потребує подальшого підтвердження на основі статистичних моделей та тестування.

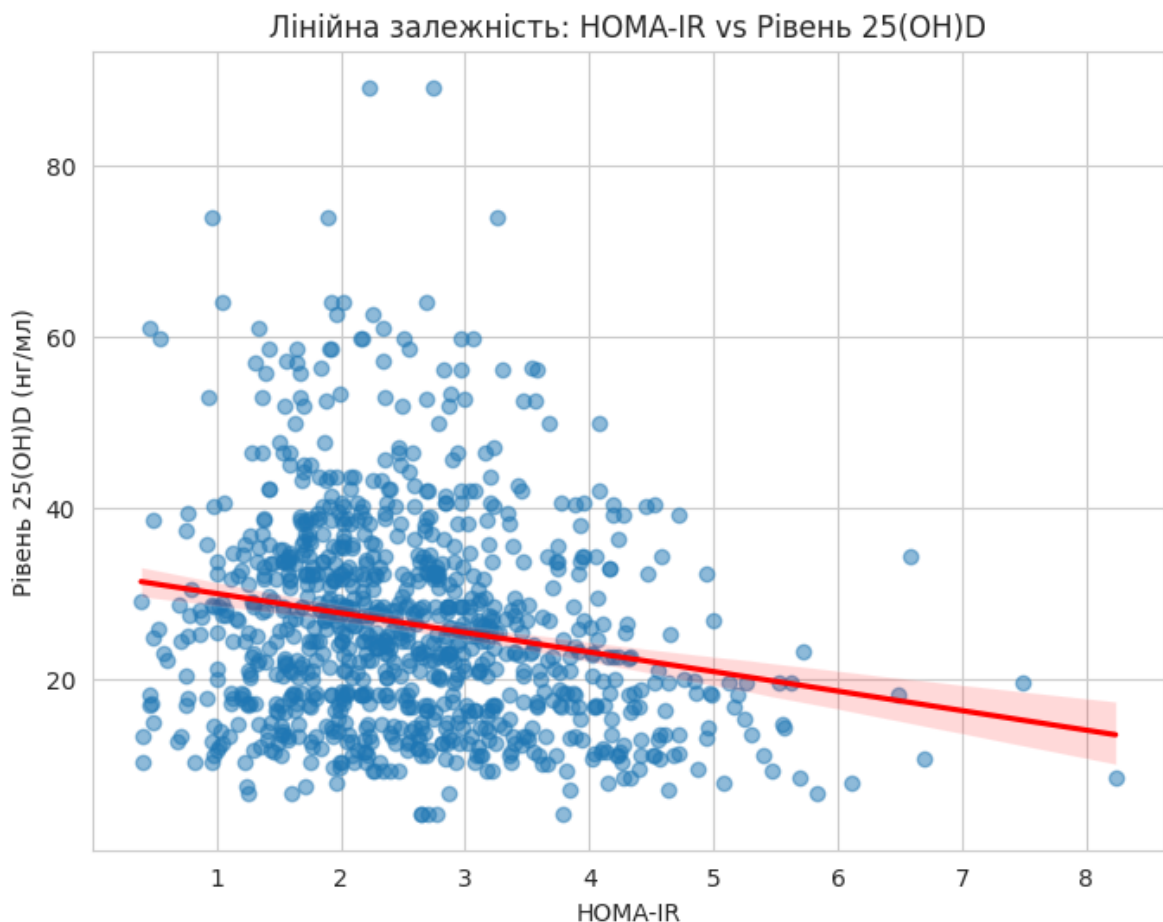


Рис. 3.3. Лінійна залежність: HOMA-IR vs Рівень 25(OH)D

На Рисунку 3.4 представлено SHAP summary plot, який демонструє внесок кожної ознаки у прогноз моделі щодо ризику дефіциту вітаміну D. Значення SHAP відображають, наскільки та в якому напрямку певна змінна впливає на вихід моделі: позитивні значення підвищують імовірність дефіциту, негативні — зменшують. Кожна точка відповідає окремому спостереженню, а її колір вказує на значення відповідної ознаки (від низьких — синій, до високих — рожевий/червоний).

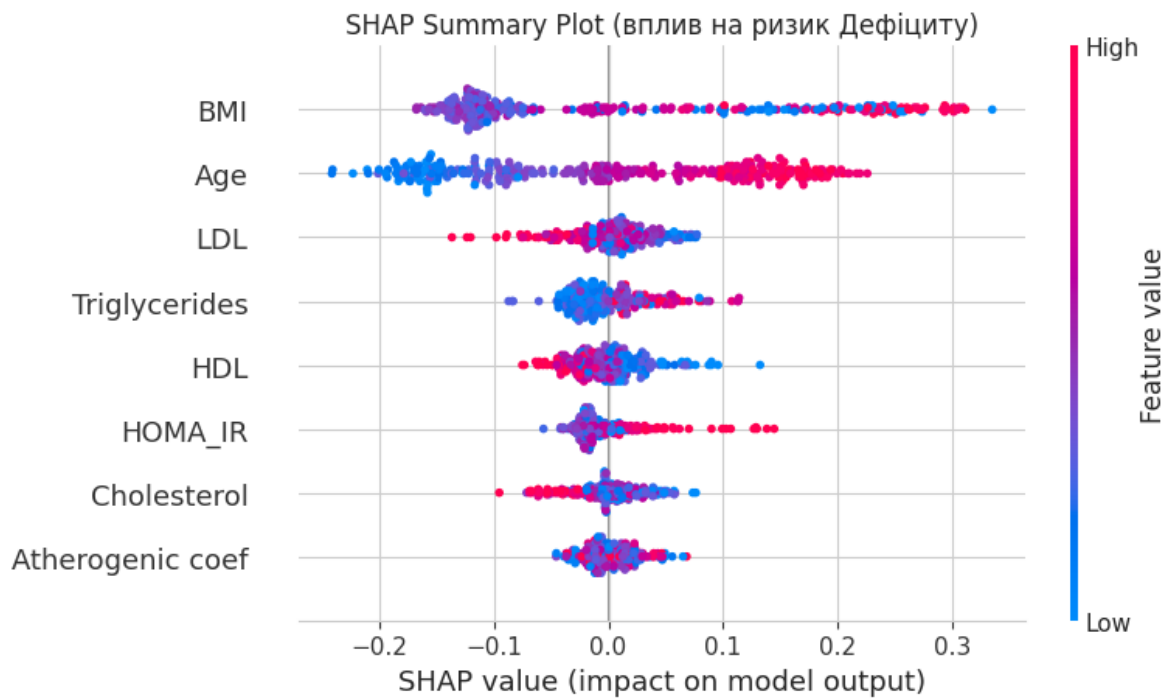


Рис. 3.4. SHAP Summary Plot (вплив на ризик Дефіциту)

Графік дозволяє оцінити як важливість ознак, так і напрям їхнього впливу:

- ІМТ та Вік мають найбільший позитивний вплив на прогноз дефіциту: високі значення цих ознак (червоні точки) розташовані праворуч, що свідчить про підвищення ризику.
- LDL (ЛПНП) та Cholesterol також демонструють тенденцію до підвищення ризику при високих значеннях.
- Для HOMA-IR спостерігається подібний патерн: вищі значення інсулінорезистентності асоціюються з більшими SHAP-значеннями, тобто більшим внеском у прогноз дефіциту.
- Ознаки HDL (ЛПВП) та Triglycerides мають змішаний вплив, однак низькі значення HDL (ЛПВП) частіше асоціюються з підвищеним ризиком (точки синього кольору з позитивними SHAP-значеннями).
- Коефіцієнт атерогенності також демонструє незначний, але помітний позитивний вплив при високих значеннях коефіцієнта.

Загалом графік ілюструє, що модель виявляє кілька метаболічних показників — зокрема BMI (ІМТ), Age (вік), LDL (ЛПНП), HOMA-IR — як суттєві фактори ризику дефіциту 25(OH)D. Таким чином, SHAP summary plot підтверджує важливість комплексної оцінки метаболічного профілю під час моделювання ризику дефіцитних станів.

РОЗДІЛ 4

ВИЯВЛЕННЯ ПРИЧИННО-НАСЛІДКОВИХ ЗВ'ЯЗКІВ ЗА ДОПОМОГОЮ МЕТОДУ DOUBLE MACHINE LEARNING

4.1 Математичні основи методу Double Machine Learning

Метод Double Machine Learning (DML), запропонований Chernozhukov et al. (2018) [28], представляє собою сучасний підхід до каузального виведення в умовах високорозмірних даних, який поєднує переваги методів машинного навчання з строгістю економетричного аналізу.

4.1.1 Постановка задачі каузального виведення

Розглянемо стандартну задачу оцінювання причинно-наслідкового ефекту. Нехай $Y \in \mathbb{R}$ — змінна відгуку (наприклад, рівень вітаміну D), $D \in \mathbb{R}$ — змінна впливу або treatment (наприклад, експозиція до сонячного світла), і $X \in \mathbb{R}^p$ — вектор коваріат високої розмірності (демографічні характеристики, дієтичні фактори, клінічні показники). Базова структурна модель має вигляд:

$$Y = \theta_0 D + g_0(X) + U, \quad \mathbb{E}[U|X, D] = 0 \quad (4.1)$$

де θ_0 — параметр каузального інтересу (істинний ефект впливу), $g_0(X)$ — невідома функція базового рівня, а U — незалежний шум [29].

4.1.2 Проблема конфаундингу та регуляризаційного зміщення

Класичні методи оцінювання, такі як звичайна регресія найменших квадратів (OLS), можуть давати зміщені оцінки θ_0 з двох основних причин:

- 1) Проблема конфаундингу (confounding bias): Якщо існують спільні причини X , які впливають як на D , так і на Y , то пряма регресія дає зміщену оцінку каузального ефекту [30].
- 2) Регуляризаційне зміщення (regularization bias): При використанні методів машинного навчання з регуляризацією (Lasso, Ridge, Random

Forest) для оцінювання $g_0(X)$ виникає систематичне зміщення параметра θ_0 , навіть якщо предиктивна якість моделі висока [28].

4.1.3 Частково лінійна модель

DML працює в рамках частково лінійної моделі (Partially Linear Model, PLM), яка має вигляд:

$$Y = D\theta_0 + g_0(X) + U \quad (4.2)$$

$$D = m_0(X) + V \quad (4.3)$$

де: $\mathbb{E}[U|X] = 0$, $\mathbb{E}[V|X] = 0$ — умови ортогональності

$m_0(X) = \mathbb{E}[D|X]$ — функція схильності (propensity function)

$g_0(X) = \mathbb{E}[Y|X]$ — функція базового результату

Ключова ідея полягає в тому, що як $g_0(X)$, так і $m_0(X)$ можуть бути високорозмірними нелінійними функціями, які ми оцінюємо за допомогою методів машинного навчання [31].

4.1.4 Неуман-ортогональність

Центральною концепцією DML є принцип Неуман-ортогональності, який захищає оцінку θ_0 від помилок специфікації допоміжних функцій g_0 та m_0 . Нехай $\psi(W; \theta, \eta)$ — це момент-функція, де $W = (Y, D, X)$, θ — параметр інтересу, а $\eta = (g, m)$ — nuisance параметри. Умова ортогональності має вигляд:

$$\partial_\eta \mathbb{E}[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0 \quad (4.4)$$

для всіх η в околі η_0 [28].

Для частково лінійної моделі ортогональний момент має вигляд:

$$\psi(W; \theta, \eta) = (Y - D\theta - g(X))(D - m(X)) \quad (4.5)$$

Ця форма забезпечує, що похибки першого порядку в оцінках $\hat{g}$ і $\hat{m}$ не впливають на асимптотичний розподіл $\hat{\theta}$ [32].

4.1.5 Алгоритм Double Machine Learning

Алгоритм DML складається з наступних кроків [28]:

Крок 1: Sample Splitting Розділимо вибірку $(Y_i, D_i, X_i)_{i=1}^n$ на K взаємовияткових підмножин (фолдів) I_k , $k = 1, \dots, K$.

Крок 2: Cross-fitting Для кожного фолду k :

I_k : а) Навчаємо моделі $\hat{g}^{(-k)}$ та $\hat{m}^{(-k)}$ на допоміжних даних $I_k^c = 1, \dots, n \setminus I_k$:

$$\hat{g}^{(-k)} = \arg \min_{g \in \mathcal{G}} \sum_{i \in I_k^c} (Y_i - g(X_i))^2 \quad (4.6)$$

$$\hat{m}^{(-k)} = \arg \min_{m \in \mathcal{M}} \sum_{i \in I_k^c} (D_i - m(X_i))^2 \quad (4.7)$$

б) Обчислюємо залишки на тестовій підмножині I_k :

$$\tilde{Y}_i = Y_i - \hat{g}^{(-k)}(X_i), \quad \tilde{D}_i = D_i - \hat{m}^{(-k)}(X_i), \quad i \in I_k \quad (4.8)$$

Крок 3: Агрегація Оцінюємо θ_0 за допомогою звичайного методу найменших квадратів на ортогоналізованих даних:

$$\hat{\theta} = \left(\sum_{k=1}^K \sum_{i \in I_k} \tilde{D}_i^2 \right)^{-1} \left(\sum_{k=1}^K \sum_{i \in I_k} \tilde{D}_i \tilde{Y}_i \right) \quad (4.9)$$

або еквівалентно:

$$\hat{\theta} = \frac{\sum_{i=1}^n \tilde{D}_i \tilde{Y}_i}{\sum_{i=1}^n \tilde{D}_i^2} \quad (4.10)$$

4.1.6 Асимптотичні властивості

За умов регулярності, оцінка DML має наступні асимптотичні властивості [28]:

Теорема 4.1 (Узгодженість та асимптотична нормальність). Нехай виконуються наступні умови:

$$|\hat{g} - g_0|_{L_2(P)} = o_P(n^{-1/4})$$

$$|\hat{m} - m_0|_{L_2(P)} = o_P(n^{-1/4})$$

$$\mathbb{E}[(D - m_0(X))^2] > c > 0$$

Тоді:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2) \quad (4.11)$$

де

$$\sigma^2 = \mathbb{E}[(Y - D\theta_0 - g_0(X))(D - m_0(X))]^2 / (\mathbb{E}[(D - m_0(X))^2])^2$$

Важливо відзначити, що умови на швидкість збіжності $\hat{g}$ та $\hat{m}$ є

$o_P(n^{-1/4})$, що значно слабше, ніж класична параметрична швидкість $o_P(n^{-1/2})$. Це дозволяє використовувати складні моделі машинного навчання навіть у високорозмірних задачах [29].

4.1.7 Оцінювання стандартних похибок

Стандартні похибки для $\hat{\theta}$ можуть бути оцінені за допомогою робастної формули:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{\tilde{D}_i \tilde{Y}_i}{\bar{D}^2} - \frac{\hat{\theta} \tilde{D}_i^2}{\bar{D}^2} \right)^2 \quad (4.12)$$

де $\bar{D}^2 = n^{-1} \sum_{i=1}^n \tilde{D}_i^2$ [28].

Довірчий інтервал $(1 - \alpha)$ рівня для θ_0 має вигляд:

$$\hat{\theta} \pm z_{1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \quad (4.13)$$

де $z_{1-\alpha/2}$ — квантиль стандартного нормального розподілу.

4.1.8 Узагальнення: Interactive Model

DML також може бути застосований до інтерактивної моделі (Interactive Regression Model), яка дозволяє гетерогенні ефекти впливу [33]:

$$Y = g_0(D, X) + U, \quad \mathbb{E}[U|D, X] = 0 \quad (4.14)$$

Локальний каузальний ефект визначається як:

$$\theta_0(x) = \frac{\partial}{\partial d} g_0(d, x) \Big|_{d=d_0} \quad (4.15)$$

Для оцінювання $\theta_0(x)$ використовується модифікований алгоритм DML з локальною лінеаризацією та kernel-методами [34].

4.2 Специфікація каузальних задач дослідження

У цьому розділі представлено математичну формалізацію каузальних задач, що досліджуються в роботі з використанням методу Double Machine

Learning. Кожна задача спрямована на виявлення специфічних причинно-наслідкових механізмів розвитку вітамін-D-дефіцитних станів.

4.2.1 Причинний вплив інсулінорезистентності на рівень вітаміну D

Теоретичне обґрунтування

Інсулінорезистентність, що визначається одним з таких показників як індекс НОМА-IR (Homeostatic Model Assessment for Insulin Resistance), представляє собою стан зниженої чутливості тканин до дії інсуліну [17]. Епідеміологічні дослідження демонструють стійку негативну асоціацію між рівнями інсуліну та концентрацією 25-гідроксिवітаміну D [25(OH)D] у сироватці крові [21].

Проте механізми цього зв'язку залишаються предметом дискусій. Chiu et al. (2004) запропонували гіпотезу про те, що гіперінсулінемія може пригнічувати α -гідроксилазу в нирках, знижуючи синтез активної форми вітаміну D [21].

Альтернативна гіпотеза полягає в тому, що дефіцит вітаміну D може погіршувати функцію β -клітин підшлункової залози, тим самим сприяючи розвитку інсулінорезистентності [35].

Математична постановка задачі

Нехай $Y_i \in \mathbb{R}$ — концентрація 25(OH)D (нмоль/л) для індивіда i , $D_i \in \mathbb{R}^+$ — значення індексу НОМА-IR, і $X_i \in \mathbb{R}^p$ — вектор коваріат, що включає демографічні характеристики, антропометричні показники, дієтичні фактори та інші біохімічні маркери. Структурна каузальна модель має вигляд:

$$Y_i = \theta_0 D_i + g_0(X_i) + U_i, \quad \mathbb{E}[U_i | X_i, D_i] = 0 \quad (4.16)$$

де: θ_0 — середній каузальний ефект (Average Treatment Effect, АТЕ) одного збільшення НОМА-IR на рівень 25(OH)D $g_0(X_i) = \mathbb{E}[Y_i | X_i]$ — базова функція відгуку U_i — ідіосинкратичний шок

Інтерпретація параметра: Якщо $\theta_0 < 0$, це вказує на те, що підвищення інсулінорезистентності причинно знижує концентрацію вітаміну D. Величина $|\theta_0|$ показує, на скільки нмоль/л зменшується 25(OH)D при збільшенні НОМА-IR на одну одиницю, за інших рівних умов.

Проблема конфаундингу

Критичною проблемою є наявність спільних причин X_i , що впливають одночасно на НОМА-IR і на рівень вітаміну D. Основні конфаундери включають:

Індекс маси тіла (ІМТ): Ожиріння асоціюється з інсулінорезистентністю [36].

Вік: Інсулінорезистентність збільшується з віком, а синтез вітаміну D в шкірі знижується у літніх осіб [37].

Формально, конфаундинг виникає, якщо:

$$\mathbb{E}[U_i|D_i] \neq 0 \quad (4.17)$$

що порушує умову ідентифікації каузального ефекту [30].

Умова незміщеності та ідентифікація

За умови, що набір коваріат X_i задовольняє умову умовної незалежності (conditional independence assumption):

$$(Y_i(d), D_i) \perp D_i \mid X_i \quad \forall d \in \mathcal{D} \quad (4.18)$$

де $Y_i(d)$ — потенційний результат при treatment рівні d , каузальний ефект θ_0 ідентифікується [38].

Ця умова також відома як умова неконфаундингу або відсутності неспостережуваних конфаундерів. Вона стверджує, що після контролю за X_i призначення treatment D_i є умовно випадковим.

Оцінювання за допомогою DML

Застосовуємо алгоритм DML для оцінювання θ_0 :

Етап 1: Оцінюємо nuisance функції за допомогою методів машинного навчання (Random Forest, Gradient Boosting, Elastic Net):

$$\hat{g}(x) = \mathbb{E}[Y|X = x], \quad \hat{m}(x) = \mathbb{E}[D|X = x] \quad (4.19)$$

Етап 2: Обчислюємо ортогоналізовані залишки:

$$\tilde{Y}_i = Y_i - \hat{g}(X_i), \quad \tilde{D}_i = D_i - \hat{m}(X_i) \quad (4.20)$$

Етап 3: Регресія залишків:

$$\hat{\theta} = \left(\sum_{i=1}^n \tilde{D}_i^2 \right)^{-1} \sum_{i=1}^n \tilde{D}_i \tilde{Y}_i \quad (4.21)$$

Статистичний висновок

Стандартна похибка $\hat{\theta}$ оцінюється як:

$$SE(\hat{\theta}) = \sqrt{\frac{1}{n} \cdot \frac{\sum_{i=1}^n (\tilde{Y}_i - \hat{\theta} \tilde{D}_i)^2 \tilde{D}_i^2}{\left(\sum_{i=1}^n \tilde{D}_i^2 \right)^2}} \quad (4.22)$$

95% довірчий інтервал:

$$CI_{95\%} = \hat{\theta} \pm 1.96 \cdot SE(\hat{\theta}) \quad (4.23)$$

Статистична значущість перевіряється за допомогою t-тесту:

$$t = \frac{\hat{\theta}}{SE(\hat{\theta})} \sim t_{n-1} \quad (4.24)$$

Нульова гіпотеза $H_0 : \theta_0 = 0$ відхиляється при $p < 0.05$.

Представлений на рисунку 4.1 схил (ATE = 0.359) є оцінкою середнього причинного ефекту НОМА-ІР на рівень вітаміну D, ізольований від впливу інших предикторів. Графік демонструє незначну позитивну тенденцію, що свідчить: після корекції на інші фактори вищі значення НОМА-ІР асоціюються з дещо вищими залишковими значеннями 25(ОН)D. Однак розсіювання точок широке, а ефект — малий, що узгоджується з тим, що прямий причинний зв'язок між цими змінними є слабким порівняно з впливом інших факторів, включених до моделі.



Рис. 4.1. Причинний вплив HOMA-IR на рівень 25(OH)D

4.2.2 Причинний вплив ожиріння на рівень вітаміну D

Теоретичне обґрунтування

Ожиріння, визначене як індекс маси тіла (ІМТ) ≥ 30 кг/м², є одним з найбільш стійких предикторів дефіциту вітаміну D у популяційних дослідженнях [39]. У попередній роботі «Система оцінки ризиків розвитку вітамін-D-дефіцитних станів. Сучасні теоретичні та практичні аспекти клінічної медицини для здобувачів освіти другого (магістерського) рівня» статистичний аналіз з використанням коефіцієнтів еластичності дозволив виконати кількісну оцінку залежності рівня 25(OH)D сироватки крові від ІМТ та ОТ/ОС. Встановлено, що збільшенням ІМТ на 1 кг/м² відповідає зниженню рівня вітаміну D у середньому на 0.603 нг/мл ($p < 0.001$), а при збільшенні окружності талії на 1 см — на 0.163 нг/мл ($p < 0.001$). Отримані результати дозволяють віднести ожиріння до стану, асоційованого із дефіцитом вітаміну D [16].

Математична постановка: Binary Treatment

Нехай treatment $D_i \in 0, 1$ визначається як:

$$D_i = \begin{cases} 1, & \text{якщо BMI}_i \geq 30, \text{ (ожиріння)} \\ 0, & \text{якщо BMI}_i < 30, \text{ (без ожиріння)} \end{cases} \quad (4.25)$$

Структурна каузальна модель:

$$Y_i = Y_i(0) + D_i \cdot (Y_i(1) - Y_i(0)) = Y_i(0) + D_i \cdot \theta_0 + \eta_i \quad (4.26)$$

де: $Y_i(0)$ — потенційний результат (25(OH)D) без ожиріння $Y_i(1)$ — потенційний результат з ожирінням $\theta_0 = \mathbb{E}[Y_i(1) - Y_i(0)]$ — середній каузальний ефект (ATE)

Ідентифікаційна стратегія

Для ідентифікації ATE потрібні дві ключові умови [38]:

1. Unconfoundedness (Незміщеність):

$$(Y_i(0), Y_i(1)) \perp D_i \mid X_i \quad (4.27)$$

2. Overlap (Перекриття):

$$0 < \mathbb{P}(D_i = 1 \mid X_i = x) < 1 \quad \forall x \in \mathcal{X} \quad (4.28)$$

Умова перекриття гарантує наявність як treated, так і control одиниць в усьому діапазоні коваріат, що дозволяє проводити контрфактуальні порівняння.

Propensity Score та балансування

Propensity score визначається як:

$$e(x) = \mathbb{P}(D_i = 1 \mid X_i = x) \quad (4.29)$$

За теоремою Rosenbaum & Rubin (1983) [40], якщо unconfoundedness виконується для X_i , вона також виконується для $e(X_i)$:

$$(Y_i(0), Y_i(1)) \perp D_i \mid e(X_i) \quad (4.30)$$

Це дозволяє зредукувати багатовимірну проблему балансування до одновимірної.

Оцінювання АТЕ для бінарного treatment

Застосовуємо DML з бінарним treatment:

Етап 1: Оцінюємо propensity score та conditional mean outcome:

$$\hat{e}(x) = \mathbb{P}(D = 1|X = x), \quad \hat{\mu}(x) = \mathbb{E}[Y|X = x] \quad (4.31)$$

Етап 2: Обчислюємо ортогоналізовані залишки:

$$\tilde{D}_i = D_i - \hat{e}(X_i), \quad \tilde{Y}_i = Y_i - \hat{\mu}(X_i) \quad (4.32)$$

Етап 3: Оцінюємо АТЕ:

$$\hat{\theta}_{\text{ATE}} = \frac{1}{n} \sum_{i=1}^n \frac{\tilde{D}_i \tilde{Y}_i}{\hat{e}(X_i)(1 - \hat{e}(X_i))} \quad (4.33)$$

або еквівалентно через Augmented Inverse Propensity Weighting (AIPW) estimator [41]:

$$\begin{aligned} \hat{\theta}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^n & \left[\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{(D_i - \hat{e}(X_i)) \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - \right. \\ & \left. - \frac{(1 - D_i) Y_i}{1 - \hat{e}(X_i)} + \frac{(D_i - \hat{e}(X_i)) \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)} \right] \end{aligned} \quad (4.34)$$

де $\hat{\mu}_d(x) = \mathbb{E}[Y|D = d, X = x]$.

Перевірка якості оцінювання

1. Баланс коваріат: Обчислюємо стандартизовану різницю середніх (Standardized Mean Difference):

$$\text{SMD}_j = \frac{\bar{X}_{j,D=1} - \bar{X}_{j,D=0}}{\sqrt{(s_{j,D=1}^2 + s_{j,D=0}^2)/2}} \quad (4.35)$$

Баланс вважається задовільним, якщо $|\text{SMD}_j| < 0.1$ для всіх коваріат j [42].

2. Overlap діагностика: Візуалізуємо розподіли propensity scores для treated та control груп. Області з слабким перекриттям (наприклад, $\hat{e}(x) <$

0.05 або $\hat{e}(x) > 0.95$) можуть бути виключені (trimming) [43].

3. Sensitivity analysis: Оцінюємо чутливість результатів до можливого неспостережуваного конфаундингу за допомогою методу Rosenbaum bounds [44] або E-value [45].

За результатами розрахунків, зворотний ефект дорівнював 0.0037 (P-value: 0.2548). Це призводить до висновку, що **зворотний зв'язок не є статистично значущим, тобто значення Вітаміну D не знижує інсулінорезистентність.**

4.2.3 Гетерогенність ефекту за ІМТ

Теоретичне обґрунтування

Для діагностики та класифікації ступеня ожиріння використовувався індекс маси тіла (ІМТ), також відомий як індекс Кетле. Цей показник розраховується як відношення маси тіла (у кілограмах) до квадрата зросту (у метрах) за формулою:

$$IMT = \frac{m}{h^2} \quad (4.36)$$

де m — маса тіла (кг), h — зріст (м).

Інтерпретація отриманих результатів здійснювалася відповідно до чинної класифікації Всесвітньої організації охорони здоров'я (ВООЗ) для дорослих осіб [46].

Згідно з цими критеріями, нормальна маса тіла відповідає діапазону ІМТ 18.5–24.9 кг/м². Показники, що перевищують ці значення, класифікують як передожиріння (надлишкова маса тіла) або ожиріння відповідного ступеня (Таблиця 4.1).

Табл. 4.1. Класифікація маси тіла за індексом Кетле (ВООЗ)

Класифікація	ІМТ (кг/м ²)
Дефіцит маси тіла	< 18.5
Нормальна маса тіла	18.5 – 24.9
Передожиріння (надлишкова маса тіла)	25.0 – 29.9
Ожиріння	≥ 30.0

Для дослідження гетерогенності причинного впливу індексу НОМА (T) на рівень вітаміну D (Y) в залежності від індексу маси тіла (BMI), було використано підхід, що базується на ортогоналізації (Double Machine

Learning). Нехай залежність описується частково лінійною моделлю:

$$Y = \theta(X)T + g(X) + \varepsilon, \quad E[\varepsilon|X, T] = 0 \quad (4.37)$$

$$T = m(X) + \eta, \quad E[\eta|X] = 0 \quad (4.38)$$

де: Y — цільова змінна (VitD); T — змінна впливу (НОМА); X — вектор коваріат (конфаундерів); $\theta(X)$ — умовний середній причинний ефект (CATE), який ми хочемо оцінити; $g(X)$ та $m(X)$ — «заважаючі» функції (nuisance parameters), що моделюють залежність виходу та впливу від коваріат.

Алгоритм оцінювання складався з наступних кроків:

- 1) Оцінювання «заважаючих» параметрів: За допомогою алгоритму Random Forest були побудовані дві моделі регресії для оцінки умовних математичних сподівань:

$$\hat{q}(X) = \hat{E}[Y|X] \quad (\text{прогноз VitD за коваріатами}) \quad (4.39)$$

$$\hat{p}(X) = \hat{E}[T|X] \quad (\text{прогноз НОМА за коваріатами}) \quad (4.40)$$

- 2) Обчислення ортогоналізованих залишків (Residualization): Згідно з теоремою Робінсона (Robinson transformation), для усунення зміщення, спричиненого конфаундерами X , необхідно перейти до залишків:

$$\tilde{Y} = Y - \hat{q}(X) \quad (4.41)$$

$$\tilde{T} = T - \hat{p}(X) \quad (4.42)$$

Цей крок «очищує» змінні НОМА та VitD від впливу інших факторів (вік, стать тощо), залишаючи лише ту варіацію, яка не пояснюється коваріатами.

- 3) Оцінка CATE в підгрупах (Subgroup Analysis): Вибірка була розділена на чотири групи (G_k) на основі квартилів зміної BMI (Low, Normal, Overweight, Obese). Для кожної групи k оцінювався локальний причинний ефект $\hat{\theta}_k$ шляхом розв'язання задачі мінімізації квадратів залишків без вільного члена (intercept):

$$\hat{\theta}_k = \arg \min_{\theta} \sum_{i \in G_k} (\tilde{Y}_i - \theta \tilde{T}_i)^2 \quad (4.43)$$

Такий підхід дозволяє інтерпретувати коефіцієнт $\hat{\theta}_k$ як чистий причинний ефект НОМА на VitD, специфічний для конкретної категорії індексу маси тіла.

Аналіз гетерогенності ефекту (CATE) демонструє суттєву варіативність впливу інсулінорезистентності на рівень вітаміну D в залежності від індексу маси тіла пацієнтів (Рис. 4.2).

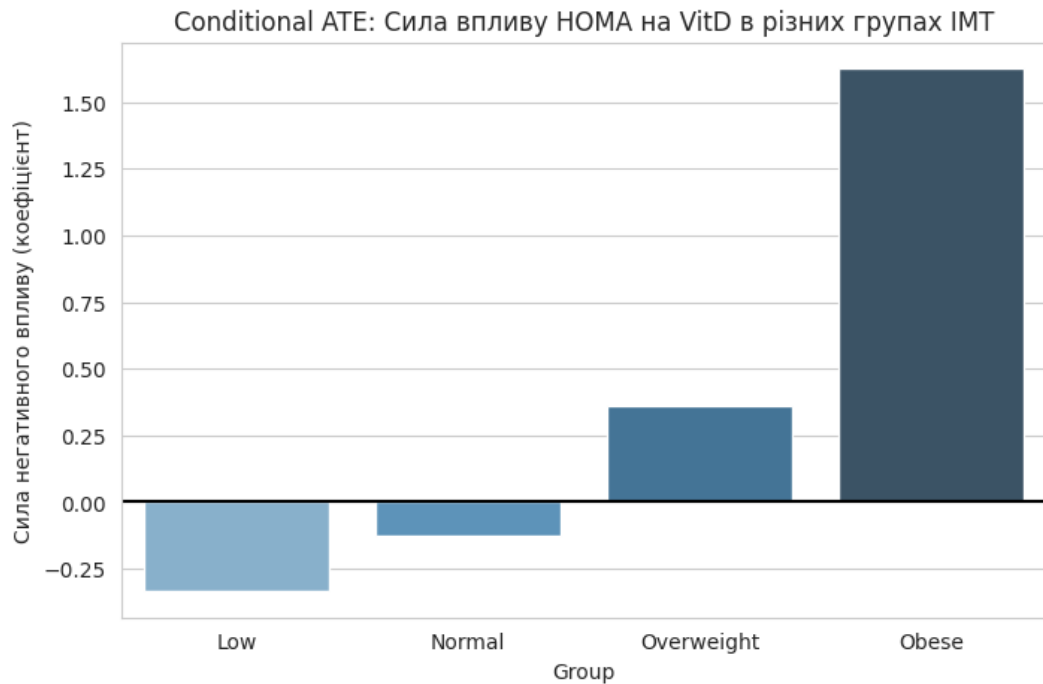


Рис. 4.2. Умовний середній причинний ефект (CATE) впливу індексу НОМА на рівень вітаміну D у різних групах за індексом маси тіла (ІМТ)

Групи «Low» та «Normal»: Для пацієнтів з низьким та нормальним ІМТ спостерігається від’ємний коефіцієнт впливу (близько -0.3 та -0.1 відповідно). Це вказує на класичну обернену залежність: зростання індексу НОМА асоціюється зі зниженням рівня вітаміну D після врахування всіх коваріат.

Групи «Overweight» та «Obese»: У групах з надмірною вагою та ожирінням характер зв’язку змінюється на протилежний. Спостерігаються додатні коефіцієнти, причому в групі «Obese» ефект є найбільш вираженим (коефіцієнт > 1.5).

Результати свідчать про те, що взаємозв’язок між НОМА та вітаміном D не є однорідним. Негативний вплив інсулінорезистентності на вітамін D притаманний переважно людям без надмірної ваги. Натомість у пацієнтів з ожирінням механізм взаємодії цих показників може бути іншим, або ж результати в цій підгрупі зазнають впливу специфічних нелінійних ефектів, які потребують додаткового клінічного трактування.

ВИСНОВКИ

У роботі вирішено актуальну науково-прикладну задачу прогнозування ризиків вітамін-D-дефіцитних станів серед населення Півдня України з використанням методів машинного навчання та каузального виведення. На основі аналізу даних 928 респондентів та побудови математичних моделей отримано такі результати:

1. Статистичний аналіз та обмеженість лінійних методів.

У ході розвідувального аналізу встановлено високу поширеність дефіцитних станів у вибірці (лише 33.4% мають достатній рівень вітаміну D) та підтверджено кореляційні зв'язки між рівнем 25(OH)D та антропометричними показниками (ІМТ, вік) і ліпідним профілем.

Доведено недостатню ефективність класичних регресійних моделей для прогнозування кількісного рівня вітаміну D. Багатофакторна лінійна регресія (Модель №4), незважаючи на статистичну значущість ($F = 55.39$), пояснює лише 22.4% дисперсії залежної змінної ($R^2 = 0.224$), що вказує на наявність складних нелінійних залежностей, які не описуються адитивними моделями. Логістична регресія для класифікації ризику показала посередню якість ($AUC = 0.735$) із низькою специфічністю (42.6%).

2. Ефективність машинного навчання.

Застосування алгоритмів машинного навчання, зокрема Дерев рішень (Decision Trees), дозволило якісно покращити точність діагностики. Оптимізована модель досягла точності (Ассурасу) 0.914 та специфічності 0.946, що суттєво перевершує результати логістичної регресії. Інтегральний показник якості AUC склав 0.92. Це підтверджує, що нелінійні методи машинного навчання є більш адекватним інструментом для роботи з медичними даними цього типу.

3. Роль метаболічних маркерів та SHAP-аналіз.

За допомогою методу SHAP (SHapley Additive exPlanations) визначено ієрархію впливу факторів ризику. Найвагомішими предикторами дефіциту вітаміну D визначено індекс маси тіла (BMI), вік (Age), рівень ліпопротеїдів низької щільності (LDL) та індекс інсулінорезистентності (HOMA-IR). Візуалізація SHAP summary plot підтвердила, що високі значення HOMA-IR мають позитивний внесок у прогнозування ризику дефіциту.

4. Каузальний аналіз методом Double Machine Learning (DML).

Для виявлення причинно-наслідкових зв'язків, очищених від впливу конфаундерів, було застосовано метод Double Machine Learning.

- Оцінка середнього каузального ефекту (ATE) показала, що прямий лінійний вплив HOMA-IR на рівень вітаміну D у загальній популяції

є слабким ($ATE = 0.359$) і має незначну позитивну тенденцію після врахування коваріат.

- Аналіз зворотного причинного зв'язку (вплив ожиріння/метаболічного статусу на вітамін D) показав відсутність статистично значущого ефекту в рамках побудованої моделі ($p = 0.2548$).

5. Гетерогенність ефектів (CATE).

Найважливішим результатом застосування DML стало виявлення гетерогенності причинного впливу інсулінорезистентності на рівень вітаміну D залежно від індексу маси тіла.

Встановлено, що характер зв'язку кардинально відрізняється в різних підгрупах:

- Для осіб з низьким та нормальним ІМТ спостерігається класичний негативний вплив: зростання інсулінорезистентності призводить до зниження рівня вітаміну D (коефіцієнти ≈ -0.3 та -0.1).
- У групах з надмірною вагою та ожирінням ефект змінюється на позитивний, досягаючи значень > 1.5 для групи Obese.

Загальний підсумок:

У роботі розроблено та обґрунтовано математичну модель на основі дерев рішень, яка дозволяє з високою точністю ($AUC = 0.92$) прогнозувати ризик дефіциту вітаміну D, використовуючи доступні анамнестичні та лабораторні дані. На початку 2026 року за результатами проведеного дослідження буде опублікована стаття «Оцінка вуглеводного обміну та його взаємозв'язок із ризиком вітамін-D-дефіцитних станів у населення півдня України». Застосування методу Double Machine Learning дозволило виявити приховану нелінійну структуру взаємозв'язку між вуглеводним обміном та вітамінним статусом, довівши, що механізми цього зв'язку є різними для людей з нормальною вагою та ожирінням. Отримані результати можуть бути використані для створення систем підтримки прийняття лікарських рішень та персоналізації стратегій профілактики.

СПИСОК ЛІТЕРАТУРИ

1. Cui A, Zhang T, Xiao P, Fan Z, Wang H, Zhuang Y. Global and regional prevalence of vitamin D deficiency in population-based studies from 2000 to 2022: A pooled analysis of 7.9 million participants. *Front Nutr*. 2023 Mar 17.
2. Cui A, Xiao P, Ma Y, Fan Z, Zhou F, Zheng J, Zhang L. Prevalence, trend, and predictor analyses of vitamin D deficiency in the US population, 2001-2018. *Front Nutr*. 2022 Oct 3.
3. Amrein K, Scherkl M, Hoffmann M, Neuwersch-Sommeregger S, Köstenberger M, Tmava Berisha A, Martucci G, Pilz S, Malle O. Vitamin D deficiency 2.0: an update on the current status worldwide. *Eur J Clin Nutr*. 2020 Nov;74(11):1498-1513.
4. Guo J, He Q, Li Y. Machine learning-based prediction of vitamin D deficiency: NHANES 2001-2018. *Front Endocrinol (Lausanne)*. 2024 Feb 16.
5. Шанигін А. В. (2023) Гігієнічне обґрунтування системи профілактики недостатності та дефіциту вітаміну D у населення півдня України. Одеський Національний Медичний Університет.
6. NumPy — 2025 — <https://numpy.org/>
7. Pandas — 2025 — <https://pandas.pydata.org/>
8. SciPy — 2025 — <https://scipy.org/>
9. Scikit-learn — 2025 — <https://scikit-learn.org/stable/>
10. SHAP — 2025 — <https://shap.readthedocs.io/en/latest/>
11. Matplotlib — 2025 — <https://matplotlib.org/>
12. Seaborn — 2025 — <https://seaborn.pydata.org/>
13. Shanyhin, Anton, Babienko, Volodymyr, Strakhov, Yevhen, Korkhova, Arina. Mathematical modeling of the dependence of the risk of vitamin D deficiency on anthropometric and laboratory parameters. *Journal of Education, Health and Sport* [online]. 28 April 2023, T. 13, nr 4, s. 356–366.
14. Strakhov Ye., Korkhova A. Features of using artificial intelligence for screening vitamin D deficiency in adults // *Стан, досягнення та перспективи інформаційних систем і технологій* — 2024. Матеріали XXIV Всеукраїнської науково-технічної конференції молодих вчених, аспірантів та студентів. Одеса, 18-19 квітня 2024 р. Одеса, Видавництво ОНТУ, 2024 р. С. 474-475.
15. Корхова А. С., Страхов Є. М. Моделювання оцінки ризику виникнення станів дефіциту вітаміну D. // *Інформаційні технології і автоматизація* — 2023. Матеріали XVI міжнародної науково-практичної конференції. Одеса, 19-20 жовтня 2023 р. Одеса, Видавництво ОНТУ, 2023 р. С. 428-429.

16. Шанигін А.В., Страхов Є.М., Корхова А.С., Коломійченко Ю.В. Система оцінки ризиків розвитку вітамін-D-дефіцитних станів. Сучасні теоретичні та практичні аспекти клінічної медицини для здобувачів освіти другого (магістерського) рівня : наук.-практ. конф. з міжнар. участю, присвячена 95-річчю з дня народження Л. В. Прокопової. Одеса, 27–28 квітня 2023 року. Одеса: ОНМедУ, 2023.
17. Matthews DR, Hosker JP, Rudenski AS, et al. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*. 1985;28(7):412-419.
18. Wallace TM, Levy JC, Matthews DR. Use and abuse of HOMA modeling. *Diabetes Care*. 2004;27(6):1487-1495.
19. Gagnon C, Lu ZX, Magliano DJ, et al. Low serum 25-hydroxyvitamin D is associated with increased risk of the development of the metabolic syndrome at five years: results from a national, population-based prospective study. *J Clin Endocrinol Metab*. 2012;97(6):1953-1961.
20. Scragg R, Sowers M, Bell C. Serum 25-hydroxyvitamin D, diabetes, and ethnicity in the Third National Health and Nutrition Examination Survey. *Diabetes Care*. 2004;27(12):2813-2818.
21. Chiu KC, Chu A, Go VL, Saad MF. Hypovitaminosis D is associated with insulin resistance and beta cell dysfunction. *Am J Clin Nutr*. 2004;79(5):820-825.
22. Wortsman J, Matsuoka LY, Chen TC, Lu Z, Holick MF. Decreased bioavailability of vitamin D in obesity. *Am J Clin Nutr*. 2000;72(3):690-693.
23. Strange RC, Shipman KE, Ramachandran S. Metabolic syndrome: A review of the role of vitamin D in mediating susceptibility and outcome. *World J Diabetes*. 2015;6(7):896-911.
24. Kayaniyil S, Vieth R, Retnakaran R, et al. Association of vitamin D with insulin resistance and beta-cell dysfunction in subjects at risk for type 2 diabetes. *Diabetes Care*. 2010;33(6):1379-1381.
25. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24(3):69-71.
26. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. 5th ed. Wiley; 2012.
27. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Wiley; 2013.
28. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
29. Athey, S., & Imbens, G. W. (2019). Machine learning methods that economi-

- sts should know about. *Annual Review of Economics*, 11, 685-725.
30. Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
 31. Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56(4), 931-954.
 32. Chernozhukov, V., Newey, W. K., & Robins, J. (2018). Double/de-biased machine learning using regularized Riesz representers. arXiv:1802.08667.
 33. Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. arXiv preprint arXiv:2004.14497.
 34. Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2), 264-289.
 35. Pittas, A. G., Lau, J., Hu, F. B., & Dawson-Hughes, B. (2007). The role of vitamin D and calcium in type 2 diabetes. A systematic review and meta-analysis. *The Journal of Clinical Endocrinology & Metabolism*, 92(6), 2017-2029.
 36. Hotamisligil, G. S. (2006). Inflammation and metabolic disorders. *Nature*, 444(7121), 860-867.
 37. MacLaughlin, J., & Holick, M. F. (1985). Aging decreases the capacity of human skin to produce vitamin D3. *The Journal of Clinical Investigation*, 76(4), 1536-1538.
 38. Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.
 39. Vimalaewaran, K. S., Berry, D. J., Lu, C., Tikkanen, E., Pilz, S., Hiraki, L. T., ... & Hyppönen, E. (2013). Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts. *PLoS Medicine*, 10(2), e1001383
 40. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
 41. Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.
 42. Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399-424.
 43. Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187-199.
 44. Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer.
 45. VanderWeele, T. J., & Ding, P. (2017). *Sensitivity analysis in observational*

- research: introducing the E-value. *Annals of Internal Medicine*, 167(4), 268-274.
46. Obesity: preventing and managing the global epidemic. Report of a WHO Consultation. World Health Organization technical report series. 2000. Vol. 894. P. 1–253.
 47. Korkhova A.S., Strakhov Ye.M. — 2025 — Resource access mode: <https://github.com/arina-korkhova/vitaminD>