НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО”
Фізико-математичний факультет
Кафедра математичного аналізу та теорії ймовірностей

“На правах рукопису”
УДК 519.21

До захисту допущено
Завідувач кафедри
_____ Олег КЛЕСОВ
_____

Магістерська дисертація

на здобуття ступеня магістра
за освітньо-науковою програмою
“Страхова та фінансова математика”
зі спеціальності 111 “Математика”
на тему: “Актуарне дослідження епідемії ВІЛ/СНІДу в Україні (2014-2021)”

Виконав:
студент II курсу магістратури, групи ОМ-41мн
Окунєв Єгор Максимович _____

Керівник:
доктор фізико-математичних наук, доцент
Василик Ольга Іванівна _____

Рецензент:
канд. фіз.-мат. наук, заступник декана з наукової роботи
механіко-математичного факультету
Київського національного університету
імені Тараса Шевченка
Яневич Тетяна Олександрівна

_____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань
Студент _____

Київ – 2026 року

Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-математичний факультет
Кафедра математичного аналізу та теорії ймовірностей

Рівень вищої освіти – другий (магістерський)
Спеціальність – 111 "Математика"
Освітньо-наукова програма "Страхова та фінансова математика"

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Олег КЛЕСОВ
_____

ЗАВДАННЯ
на магістерську дисертацію студенту Окунєву Єгору Максимовичу

1. **Тема дисертації** "Актуарне дослідження епідемії ВІЛ/СНІДу в Україні (2014-2021)", науковий керівник дисертації Василик Ольга Іванівна, доктор фізико-математичних наук, доцент, затверджені наказом по університету від "31" березня 2026 р. №1340-с.
2. **Термін подання** студентом дисертації "15" травня 2026 року.
3. **Об'єкт дослідження** епідемія ВІЛ/СНІДу в Україні, як страховий ризик.
4. **Предмет дослідження** оцінка розподілу часу між діагнозами ВІЛ/СНІД та оцінка ймовірності банкрутства відповідного однорідного страхового портфелю.
5. **Перелік завдань, які потрібно розробити**
   (а) Ознайомитись з літературою та методологією на тему епідеміологічних досліджень.
   (б) Користуючись загально доступними даними, знайти та привести до прийнятного для подальшої роботи стану статистичну інформацію про епідемію ВІЛ/СНІДу в Україні за період часу 2014-2021 рр. та загальну смертність за той самий період.

(в) Кластеризувати регіони України за схожими епідеміологічними трендами, для підвищення точності моделі.

(г) Оцінити функції розподілу часу для всіх релевантних до епідемії подій.

(д) Користуючись методологією Крамера-Лундберга вивести інтегральне рівняння ймовірності банкрутства відповідного однорідного страхового портфелю та знайти замкнені вирази для константи Крамера-Лундберга.

6. **Орієнтовний перелік графічного (ілюстративного) матеріалу** 22 слайди.

7. **Дата видачі завдання** "03" лютого 2026 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання	Примітка
1.	Ознайомлення з літературою та методологією на тему епідеміологічних досліджень.	04.02.2026 – 10.02.2026	Виконано
2.	Опрацювання п'ятого та шостого розділів монографії "Risk Theory" Hanspeter Schmidli.	11.02.2026 – 24.02.2026	Виконано
3.	Аналіз літератури та уточнення постановки задачі.	25.02.2026 – 10.03.2026	Виконано
4.	Знаходження та приведення до ладу необхідних статистичних даних із порталу громадського здоров'я.	11.03.2026 – 15.03.2026	Виконано
5.	Кластеризація регіонів за подібністю епідеміологічних трендів (кількість діагнозів ВІЛ/СНІД та смертність від СНІД).	16.03.2026 – 20.03.2026	Виконано
6.	Знаходження непараметричної оцінки емпіричної функції розподілу часу до переходу між станами моделі.	16.03.2026 – 11.04.2026	Виконано
7.	Підбір параметричної моделі, що найкраще описує отриману непараметричну оцінку та дозволяє відновити інформацію, втрачену через інтервально-цензуровану природу наявних емпіричних даних.	12.04.2026 – 22.04.2026	Виконано
8.	Виведення інтегрального рівняння ймовірності банкрутства однорідного страхового портфелю в межах побудованої моделі для кожного однорідного кластеру регіонів.	23.04.2026 – 14.05.2026	Виконано
9.	Знаходження константи Крамера–Лундберга для асимптотичних ймовірностей банкрутства.	23.04.2026 – 14.05.2026	Виконано
10.	Оформлення дипломної роботи	23.04.2026 – 14.05.2026	Виконано

Студент Єгор ОКУНЄВ

Науковий керівник Ольга ВАСИЛИК

РЕФЕРАТ

Магістерська дисертація: 58 сторінок, 22 слайди для проектора, 24 першоджерела.

Актуальність: Актуальність дослідження полягає в тому, що Україна є однією з країн із високими масштабами епідемії ВІЛ/СНІДу [1, 2]. Це створює значний ризик для населення та страхових компаній, який потребує детального аналізу та врахування при формуванні страхових стратегій.

Мета: Мета дослідження полягає у зборі та приведенні до робочого стану інформації про перебіг епідемії ВІЛ/СНІДу в Україні у 2014–2021 роках, кластеризації регіонів за схожими епідеміологічними тенденціями, побудові та оцінці параметрів страхових напівмарківських моделей на індивідуальному рівні, а також у знаходженні інтегрального рівняння, що пов'язує стартовий капітал страхового портфелю з ймовірністю банкрутства.

Завдання:

- Збір та відновлення даних про перебіг епідемії ВІЛ/СНІДу в Україні.
- Кластеризація регіонів за схожими епідеміологічними тенденціями.
- Побудова та оцінка параметрів напівмарківських моделей на індивідуальному рівні.
- Формулювання інтегрального рівняння для ймовірності банкрутства страхового портфелю.
- Дослідження асимптотичної поведінки ймовірності банкрутства за методологією Крамера–Лундберга.

Методи: Для виконання дослідження використано такі методології та технології: OCR-технології для збору даних зі сканованих джерел; лінійні змішані моделі для кластеризації регіонів; теорія напівмарківських ланцюгів; методи аналізу виживання, зокрема оцінки Нельсона–Аалена, Каплана–Мейєра та Тьорнбула; медичні знання про ВІЛ/СНІД, методи лікування та профілактики; методологія Крамера–Лундберга для побудови та аналізу інтегрального рівняння ймовірності банкрутства.

Об'єкт і предмет дослідження: Об'єктом є епідемія ВІЛ/СНІДу в Україні за період із повною звітністю щодо діагнозів і смертності. Предметом є побудова моделі часу та кількості релевантних діагнозів. Через особливості даних аналіз обмежено часом до діагнозу, однак методологія дозволяє покращити результати при надходженні якісніших даних.

Наукова новизна: Новизна роботи полягає у побудові моделі епідемії, відновленні втрачених даних із сканованих джерел за допомогою OCR, а також у виборі оптимальних методологічних підходів.

Практичне значення: Результати дослідження мають практичне значення для України, сектору громадського здоров'я та медичного страхування, оскільки надають інформацію для ухвалення рішень і управління ризиками.

Ключові слова: ВІЛ/СНІД, Україна, напівмарківська модель, OCR, модель Крамера–Лундберга, аналіз виживання, кластеризація.

ABSTRACT

Master's thesis: 58 pages, 22 slides for a projector, 24 primary sources.

Relevance: The relevance of the study lies in the fact that Ukraine is one of the countries most affected by the HIV/AIDS epidemic. This creates a significant risk for the population and insurance companies, requiring detailed analysis and consideration in the formation of insurance strategies.

Aim: The aim of the study is to collect, process, and restore information on the course of the HIV/AIDS epidemic in Ukraine from 2014 to 2021, to cluster regions according to similar epidemiological trends, to construct and estimate parameters of individual-level semi-Markov insurance models, and to derive an integro-differential equation linking the initial capital of an insurance portfolio with the probability of bankruptcy.

Objectives:

- Collecting and restoring data on the course of the HIV/AIDS epidemic in Ukraine.
- Clustering regions based on similar epidemiological trends.
- Constructing and estimating parameters of individual-level semi-Markov models.
- Formulating an integro-differential equation to assess the bankruptcy probability of an insurance portfolio.
- Investigating the asymptotic behavior of the bankruptcy probability using the Cramér–Lundberg methodology.

Methods: The study employs the following methodologies and technologies: OCR technologies for data collection from scanned sources; linear mixed models for clustering regions; semi-Markov chain theory; survival analysis methods, including Nelson–Aalen, Kaplan–Meier, and Turnbull estimators; medical knowledge about HIV/AIDS, prevention and treatment methods; and the Cramér–Lundberg methodology for constructing and analyzing the differential equation of bankruptcy probability.

Object and Subject of Research: The object of the study is the HIV/AIDS epidemic in Ukraine during the period with complete reporting of diagnoses and mortality. The subject of the study is the construction of an adequate model of the timing and number of relevant diagnoses. Due to the nature of available data, the analysis is limited to time until diagnosis; however, the proposed methodology allows improvement if higher-quality data and societal changes in attitudes toward regular medical examinations become available.

Scientific Novelty: The novelty of the work lies in the construction of an up-to-date epidemic model, the restoration of lost data from scanned sources using OCR, and the selection of optimal methodological approaches for completing the study's tasks.

Practical Significance: The results of the study have practical significance for Ukraine, the public health sector, and medical insurance, as they provide information for decision-making and risk management regarding the HIV/AIDS epidemic.

Keywords: HIV/AIDS, Ukraine, semi-Markov model, OCR, Cramér–Lundberg model, survival analysis, clustering.

Зміст

Вступ 10

1 Збір даних 16

2 Побудова моделі 21

3 Регіональна кластеризація 23

4 Побудова епідеміологічного марківського ланцюга 32

5 Оцінка параметрів моделі 39

6 Дослідження ймовірності банкрутства 46

Висновки 53

Список використаних джерел 55

Вступ

Актуальність дослідження полягає в тому, що Україна залишається однією з країн Європи з високими масштабами та значним впливом епідемії ВІЛ/СНІДу. Поширення цієї інфекції створює суттєві виклики для системи громадського здоров'я, а також формує додаткові ризики для страхового сектору, зокрема для систем медичного та актуарного страхування. Наявність довготривалого перебігу захворювання, складної динаміки діагностування та значної регіональної неоднорідності епідеміологічної ситуації вимагає використання сучасних статистичних та математичних методів для адекватного аналізу і моделювання відповідних процесів.

Особливістю дослідження епідемії ВІЛ/СНІДу в Україні є також обмеженість та фрагментарність доступних статистичних даних, значна частина яких представлена у вигляді сканованих епідеміологічних звітів. Це створює додаткові труднощі для подальшого статистичного аналізу та побудови математичних моделей, що робить актуальним застосування сучасних технологій обробки даних, зокрема OCR-технологій, для відновлення та систематизації наявної інформації.

Водночас для дослідження динаміки розвитку епідемії та її впливу на страхові ризики доцільним є застосування сучасних методів статистичного аналізу, зокрема методів кластеризації, аналізу виживання та стохастичного моделювання. Такі підходи дозволяють враховувати регіональні відмінності епідеміологічних процесів, оцінювати індивідуальні траєкторії розвитку захворювання та формувати адекватні моделі для актуарного аналізу.

Особливий інтерес у цьому контексті становлять напівмарківські моделі, які дають можливість описувати процеси переходів між різними станами захворювання з урахуванням розподілів часу перебування у кожному стані. Використання таких моделей у поєднанні з методами аналізу виживання дозволяє отримати більш точні оцінки параметрів епідеміологічних процесів. Отримані результати можуть бути використані для побудови актуарних моделей страхових портфелів та оцінювання відповідних фінансових ризиків.

Мета дослідження полягає у зборі, приведенні до робочого стану та відновленні інформації про перебіг епідемії ВІЛ/СНІДу в Україні у період 2014–2021 років, кластеризації регіонів за схожими епідеміологічними тенденціями, побудові та оцінці параметрів страхових напівмарківських моделей на індивідуальному рівні, а також у знаходженні інтегрального рівняння, що пов'язує стартовий капітал страхового портфелю з ймовірністю банкрутства.

Для досягнення поставленої мети у роботі розв'язуються такі завдання:

- збір та відновлення даних про перебіг епідемії ВІЛ/СНІДу в Україні;
- приведення отриманих даних до придатного для статистичного аналізу вигляду;
- кластеризація регіонів України за подібними епідеміологічними тенденціями;
- побудова та оцінка параметрів напівмарківських моделей на індивідуальному рівні;

- формулювання інтегрального рівняння для ймовірності банкрутства страхового портфелю;
- дослідження асимптотичної поведінки ймовірності банкрутства у рамках методології Крамера–Лундберга.

Об'єктом дослідження є епідемія ВІЛ/СНІДу в Україні у період 2014–2021 років за наявною офіційною статистичною звітністю щодо встановлення діагнозів та смертності.

Предметом дослідження є побудова адекватних статистичних і стохастичних моделей часу до настання релевантних медичних станів, а також актуарний аналіз страхових ризиків, пов'язаних із діагностуванням ВІЛ/СНІДу. Через особливості доступних даних аналіз обмежується часом до встановлення діагнозу, однак запропонована методологія може бути розширена у разі появи детальніших даних та підвищення рівня регулярних медичних обстежень населення.

У роботі використано такі методи дослідження: OCR-технології для відновлення статистичних даних зі сканованих джерел; лінійні змішані моделі для кластеризації регіонів; теорію напівмарківських ланцюгів для моделювання індивідуальних переходів між станами; методи аналізу виживання, зокрема оцінки Нельсона–Аалена, Каплана–Мейєра та Тьорнбула; а також підходи актуарної математики, зокрема методологію Крамера–Лундберга для аналізу ймовірності банкрутства страхового портфелю.

Огляд наукових досліджень за темою.

Актуальність дослідження проблеми ВІЛ/СНІДу підтверджується суча-

сними глобальними епідеміологічними оцінками. Зокрема, дані, представлені у [6], відображають значний вплив інфекційних захворювань на показники громадського здоров'я та функціонування систем охорони здоров'я. Аналіз подібних наборів даних дозволяє обґрунтувати необхідність застосування математичних моделей для дослідження динаміки епідемії та оцінювання ефективності медичних і соціальних втручань.

Для побудови адекватної моделі перебігу захворювання важливим є врахування клінічних та біологічних особливостей ВІЛ-інфекції. У роботі [7] досліджуються результати антиретровірусної терапії, зокрема вплив інгібіторів інтегрази на вірусологічні показники пацієнтів. Отримані результати підтверджують суттєву роль терапії у зміні швидкості прогресування захворювання, що є важливим при моделюванні переходів між станами.

Дослідження механізмів передачі інфекції та оцінювання ризиків інфікування мають важливе значення для формування структури моделей. У роботі [8] запропоновано підходи до оцінювання ризику трансмісії інфекцій, що поєднують емпіричні оцінки та математичне моделювання. Особливості окремих шляхів передачі, зокрема перинатальної, розглянуто у [9], що дозволяє враховувати неоднорідність популяції при побудові моделей.

Важливим аспектом є аналіз динаміки вірусного навантаження та переходів між клінічними станами. У роботі [10] показано, що повторні підвищення вірусного навантаження асоціюються з неефективністю терапії, що обґрунтовує використання стохастичних моделей із випадковими моментами переходів між станами.

Окрему увагу у літературі приділено регіональним особливостям епіде-

мії. Зокрема, у [11] досліджено фактори, що впливають на затримку включення ВІЛ-позитивних осіб до медичного нагляду, що є суттєвим чинником поширення інфекції. У роботі [12] показано вплив соціально-політичних факторів, зокрема воєнних дій, на динаміку поширення ВІЛ в Україні. Подальші дослідження ([13]) із використанням філодинамічних підходів дозволяють детальніше описати структуру передачі інфекції у вразливих групах населення. У європейському контексті проблема пізньої діагностики ВІЛ залишається актуальною ([14]), що підтверджується і сучасними оглядовими роботами щодо стану контролю епідемії в регіоні ([15]).

З методологічної точки зору важливим є коректний аналіз даних, які часто мають інтервально-цензуровану природу. У роботі [16] наведено огляд методів аналізу таких даних, зокрема непараметричної оцінки функції розподілу за алгоритмом Тьорнбулла. Застосування цих підходів є доцільним у контексті даного дослідження, оскільки наявні епідеміологічні дані часто містять інформацію про події у вигляді часових інтервалів, а не точних моментів настання.

Таким чином, проведений аналіз літератури свідчить про необхідність комплексного підходу до дослідження епідемії ВІЛ/СНІДу, який поєднує використання сучасних статистичних методів, стохастичного моделювання та актуарного аналізу. Це створює підґрунтя для побудови адекватних моделей, здатних враховувати як індивідуальні особливості перебігу захворювання, так і регіональні відмінності епідеміологічної ситуації.

Наукова новизна отриманих результатів полягає у побудові актуальної статистично-актуарної моделі перебігу епідемії ВІЛ/СНІДу в Україні, від-

новленні та систематизації історичних епідеміологічних даних із використанням OCR-технологій, а також у застосуванні комплексу сучасних статистичних та стохастичних методів для аналізу епідеміологічних процесів і пов'язаних із ними страхових ризиків.

Практичне значення отриманих результатів полягає у можливості використання побудованих моделей для аналізу епідеміологічної ситуації, прогнозування розвитку епідемії та оцінювання страхових ризиків. Результати дослідження можуть бути використані у сфері громадського здоров'я, актуарних досліджень та медичного страхування для підтримки процесів прийняття управлінських рішень.

Магістерська дисертація складається зі вступу, чотирьох розділів, висновків та списку використаної літератури. У першому розділі розглянуто збір і обробку емпіричних даних із епідеміологічних звітів порталу громадського здоров'я. У другому розділі проведено кластеризацію регіонів України за подібними епідеміологічними тенденціями. У третьому розділі побудовано напівмарківську модель перебігу ВІЛ/СНІДу зі станами "недіагностований", "діагностований ВІЛ", "діагностований СНІД" та "смерть" і здійснено оцінювання її параметрів. У четвертому розділі досліджено ймовірність банкрутства однорідного страхового портфелю в моделі Крамера–Лундберга для страхування ризику діагностування ВІЛ/СНІДу.

Усі основні результати дисертаційної роботи отримано автором самостійно.

1 Збір даних

У межах цього розділу ми опишемо, яким чином були зібрані емпіричні дані, що було зроблено для приведення їх до робочого стану та які величини були оцінені.

Основним джерелом статистичних даних у цій роботі є епідеміологічні звіти Порталу громадського здоров'я України [3], у яких зібрано інформацію про перебіг епідемії ВІЛ/СНІДу в Україні, починаючи із січня 2014 року і дотепер. Звіти містять щомісячно оновлювану інформацію по областях України у вигляді таких змінних:

- кількість нових діагнозів ВІЛ;
- кількість нових діагнозів СНІД;
- кількість смертей, спричинених СНІДом;
- загальна кількість людей, що перебувають на медичному обліку ЛЖВ на кінець звітного місяця;
- відповідні величини, перераховані на 100000 осіб популяції.

З огляду на природу звітів наявні дані є ліво- та інтервально цензурованими. Наявність змінних, перерахованих на 100000 осіб популяції, імплікує те, що була проведена загальна оцінка популяції для кожного звітного місяця. Додатково в цій роботі використовувалися дані про загальну смертність, запозичені із сайту Державної служби статистики України [4]. Дані про смертність через повномасштабне вторгнення Росії доступні лише по-

чинаючи із січня 2015 року до січня 2022 року.

Емпіричні дані періодично зберігалися як скани Excel-таблиць, а не безпосередньо як Excel-файли, тож у їхньому вихідному стані вони не були придатними для подальшої роботи. Для вирішення цієї проблеми була використана технологія OCR (оптичне розпізнавання символів). Зокрема, був задіяний сервіс ABBYY. Використання технології оптичного розпізнавання символів здебільшого перетворило дані на придатні до роботи, але також призвело до появи додаткових проблем. Зокрема, програмне забезпечення періодично помилково інтерпретувало символи. Найчастіше це відбувалося з цифрами 0 та 1, які помилково зчитувалися як літери "o" та "i" відповідно. Також після OCR-зчитування з'являлося спорадичне спотворення даних, а вирівнювання вмісту комірок таблиці інтерпретувалося як наявність пробілів і змушувало MS Excel сприймати вміст комірок як текст.

Для того щоб усунути деякі проблеми, які з'явилися внаслідок використання OCR, було вручну видалено зайві пробіли та застосовано глобальні заміни типу $i \to 1$.

Після цього було оцінено кількість смертей з причин, відмінних від СНІДу (англ. Causes different from AIDS = CDA). Далі ми зробили вмотивоване припущення, що смертність від CDA серед певної демографічної групи (наприклад, ЛЖВ) пропорційна кількості смертей від CDA серед людей із загальної популяції (англ. Population = P).

Це дозволило нам оцінити кількість людей, що померли від CDA серед таких демографічних груп: люди, у яких недіагностовано ВІЛ/СНІД (англ. undiagnosed = U), люди, у яких діагностовано ВІЛ (англ. HIV = H), люди,

у яких діагностовано СНІД (англ. AIDS = A). Слід наголосити, що категорія U містить ЛЖВ, які не знають про свій стан. Це зумовлено тим, що дуже часто ВІЛ може бути неможливо діагностувати стандартними тестами впродовж періоду близько пів року після зараження, а також тим, що хвороба може протікати безсимптомно впродовж років після зараження.

Далі використовуватимемо наведені вище латинські скорочення для досліджуваних демографічних груп і такі праві підписи:

- `_` — відсутність підпису означає загальну кількість осіб у відповідній демографічній групі;
- N — приріст відповідної демографічної групи протягом звітного місяця;
- D — кількість осіб відповідної демографічної групи, які померли протягом звітного місяця;
- `A` — смерть, спричинена СНІДом;
- `CDA` — смерть, спричинена CDA;
- `100k` — значення, обчислене на 100 000 осіб загальної популяції.

Додатково ми використовували лівий підпис "t" для уточнення моменту часу спостереження (номер спостережуваного місяця). Наприклад, загальна кількість людей, що померли через CDA впродовж четвертого спостережуваного місяця, позначатиметься як ${}_4P_{D.CDA}$.

Тож загальна кількість людей, що померли через CDA у певній неуто-

чненій демографічній групі G, задається формулою:

$$ {}_tG_{D.CDA} = {}_tP_{D.CDA} \cdot \frac{{}_tG}{{}_tP}. $$

Як було зазначено раніше, Порталом громадського здоров'я також були оцінені статистики на 100000 осіб загальної популяції і таким чином вони неявно містять інформацію про оцінки розміру загальної популяції. Це дозволяє нам оцінювати загальну чисельність певної демографічної групи G, якщо відома лише її кількість, оцінена на 100000 осіб загальної популяції:

$$ {}_tG_{\_.100k} = \frac{{}_tG_{\_.100k}}{{}_tP} \cdot 100000, $$

де _ — це довільна статистика, обчислена для 100000 людей загальної популяції.

Остаточно, щоб адресувати спорадичне спотворення (неможливість зчитування) даних, були використані такі формули:

$$ {}_tH = {}_{t-1}H + {}_tH_N - {}_tH_{D.CDA}, $$
$$ {}_tA = {}_{t-1}A + {}_tA_N - {}_tA_{D.CDA} - {}_tA_{D.A}. $$

Ці формули мають інтуїтивну інтерпретацію: для отримання загальної кількості людей з діагностованим ВІЛ впродовж спостережуваного місяця t до загальної кількості людей з діагностованим ВІЛ за попередній місяць слід додати приріст нових діагнозів ВІЛ та відняти оцінку людей, що померли від CDA серед людей, у яких діагностовано ВІЛ.

Аналогічно, для людей з діагностованим СНІДом слід врахувати нові

діагнози СНІДу та смертність від СНІДу і CDA.

Слід зауважити, що кількість прогресій ВІЛ у СНІД серед ЛЖВ враховувати не треба, оскільки абсолютна більшість ЛЖВ отримує антиретровірусну терапію (АРТ), яка інгібує прогресію ВІЛ, не даючи йому досягти термінальної стадії СНІДу, тобто такі переходи майже не трапляються, а ті, що трапляються, неможливо оцінити з наявних даних.

Наступне важливе зауваження стосується переходів між стадіями СНІДу та ВІЛ: у пацієнтів, які отримують антиретровірусну терапію (АРТ), з часом спостерігаються вірусологічна супресія та імунологічне відновлення, що супроводжується зниженням вірусного навантаження до невизначуваного рівня [7]. Унаслідок цього клінічний стан таких пацієнтів більше не відповідає критеріям СНІДу і відповідає хронічно контрольованій ВІЛ-інфекції.

Час, необхідний для такого переходу, є дуже варіативним, тому оцінити кількість таких регресій на основі наявних даних неможливо, але, як і у випадку з прогресіями ВІЛ у СНІД, кількість таких переходів є нехтовно малою, оскільки абсолютна більшість діагнозів ВІЛ/СНІДу трапляється саме на стадії ВІЛ.

Фрагменти цієї роботи висвітлені у тезах доповіді [5].

2 Побудова моделі

Як було сказано раніше, природнім вибором станів для моделі є "Недіагностований" (U), живе з ВІЛ (H), живе зі СНІД (A), та мертвий (D). Інтуїтивно, в такій моделі хочеться розглядати переходи між станами аналогічні природному розвитку ВІЛ, себто, $U \to H \to A$, та всі ці стани сполучені із D. Але така модель не є репрезентативною з огляду на наступні причини:

- Легкий доступ до АРТ дозволяє переходи $A \to H$,
- ЛЖВ, що перебувають на медичному обліку, майже ніколи не переходять від H до A [10],
- Більшість людей у стані A є "пізньо-діагностованими" тобто, в межах нашої моделі таким людям притаманний перехід $U \to A$ [14], [15].

Тож доречно розглядати таку модель станів та переходів між ними: як було зазначено в попередньому розділі, оцінити кількість переходів $A \to H$ неможливо на основі наявних даних. Тож, для подальших досліджень, розглядатимемо стан H/A, що описує діагностованих ЛЖВ на всіх стадіях розвитку хвороби, для якого можливо оцінити всі ймовірності переходів. Тоді модель станів прийме наступний вигляд:

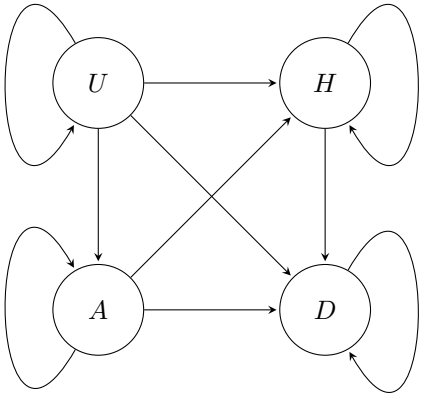

Рис. 1: Деталізована модель

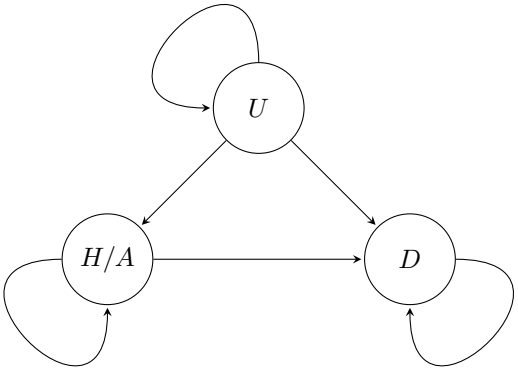

Рис. 2: Спрощена модель

Наостанок зазначимо, що хоча група U включає недіагностованих ЛЖВ, її внесок у смертність, пов'язану зі СНІД, є статистично незначущим, оскільки більшість таких смертей припадає на випадки пізньої діагностики.

3 Регіональна кластеризація

Кластеризація відіграє дуже важливу роль під час актуарних досліджень, оскільки спостереження з одного кластеру найчастіше мають спільні ознаки та властивості, наприклад поведінкові особливості чи географічне розташування. Пропуск кроку кластеризації матиме такі негативні наслідки як недооцінення популяційної дисперсії, нерепрезентативність моделі. Що, в свою чергу, може призвести до неправильних висновків та неефективних рішень щодо ціноутворення для страхового продукту. Врахування гетерогенної структури даних і подальша кластеризація дозволяє актуарію побудувати адекватну модель із вищою точністю, яка якісніше описує досліджувані процеси ризику.

В межах цієї роботи для покращення точності результатів ми кластеризуємо регіони за їх тенденціями епідемії ВІЛ/СНІД.

В нашому дослідженні ми оцінимо параметри трьох лінійних змішаних моделей: перша - для стандартизованих лінійних трендів кількості нових діагнозів ВІЛ в регіоні, друга - для стандартизованих лінійних трендів кількості нових діагнозів СНІД в регіоні, та остання - для стандартизованих лінійних трендів кількості смертей спричинених СНІД в регіоні.

Кожна з цих моделей описує важливий аспект перебігу епідемії ВІЛ/СНІД в регіоні:

- кількість діагнозів ВІЛ описує одночасно два явища: наскільки масштабною є епідемія в регіоні, та наскільки ефективними є превентивні

заходи у боротьбі з епідемією (Якість та кількість освітніх заходів із питань ВІЛ/СНІД, доступ до засобів до- та післяконтактної профілактики ВІЛ, доступ до АРТ, культурне ставлення до регулярного медичного тестування, рівень стигматизації ВІЛ);

- кількість діагнозів СНІД першочергово описує наскільки частим діагностування відбувається на пізніх стадіях;
- остаточно, кількість смертей від причин пов'язаних із СНІД, описує те наскільки часто трапляються настільки запущені випадки, що медичне втручання не має достатньо часу для порятунку життя людини.

Зауважимо, що через те, що в різних регіонах може бути різна культура догляду за здоров'ям та різна ефективність превентивних кампаній, висока або низька кількість діагнозів ВІЛ сама по собі не є індикатором важкості стану епідемії в певному регіоні.

Наприклад, регіони із помірною кількістю діагнозів ВІЛ, але низькими кількостями діагнозів СНІД та пов'язаними із СНІД смертями можна вважати 'низько-ризиковими'. В той час як регіони із низькою кількістю діагнозів ВІЛ та помірною кількістю діагнозів СНІД або смертей пов'язаних із СНІД можуть означати, що в регіоні відбувається систематичне недообстеження.

Для подальшого аналізу, було прийнято рішення про стандартизацію даних. А саме до даних було застосовано z-перетворення, для того, щоб в кожен момент часу, дані мали середнє нуль та одиничне стандартне відхиле-

ння. Нормалізацію було виконано по змінних HIV.N.100k, AIDS.N.100k так AIDS.D.100k, щоб уникнути помилкових подібностей спричинених популяційно-кількісною подібністю регіонів.

Для виділення прихованих лінійних класів в спостережуваних даних, було використано пакет R `lcmm` . Методологічно, задача найкращого виділення прихованих класів є задачею оптимізації функції правдоподібності серед всіх можливих розбиттів даних на наперед задану кількість класів. В нашій роботі, враховуючи велику кількість даних, було знайдено найкращі моделі для одного, двох та трьох станів. Дестандартизовані траєкторії середніх показано на наступних зображеннях.

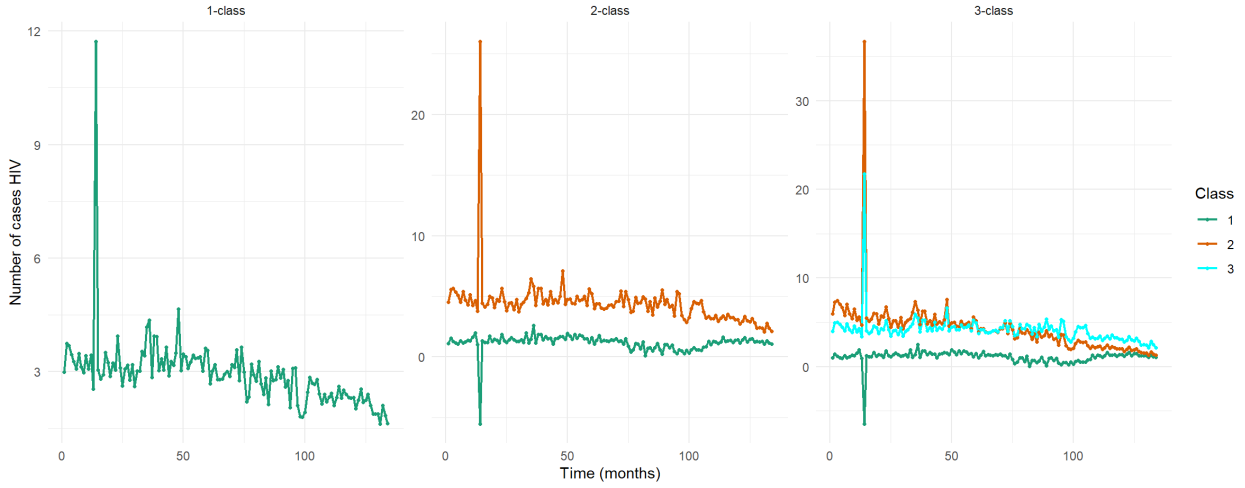

Рис. 3: Траєкторії середнього значення HIV.N.

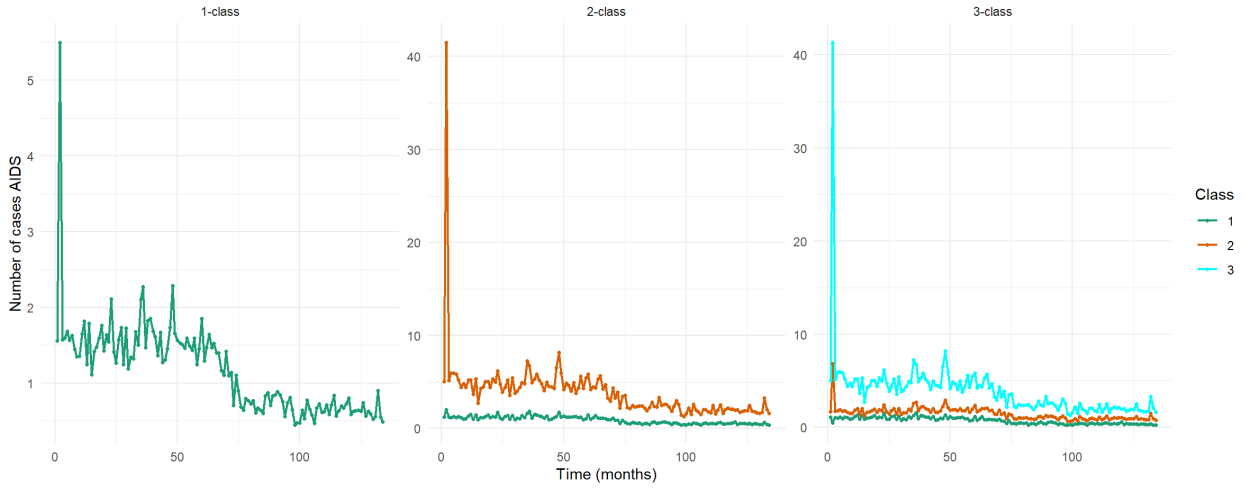

Рис. 4: Траєкторії середнього значення AIDS.N.

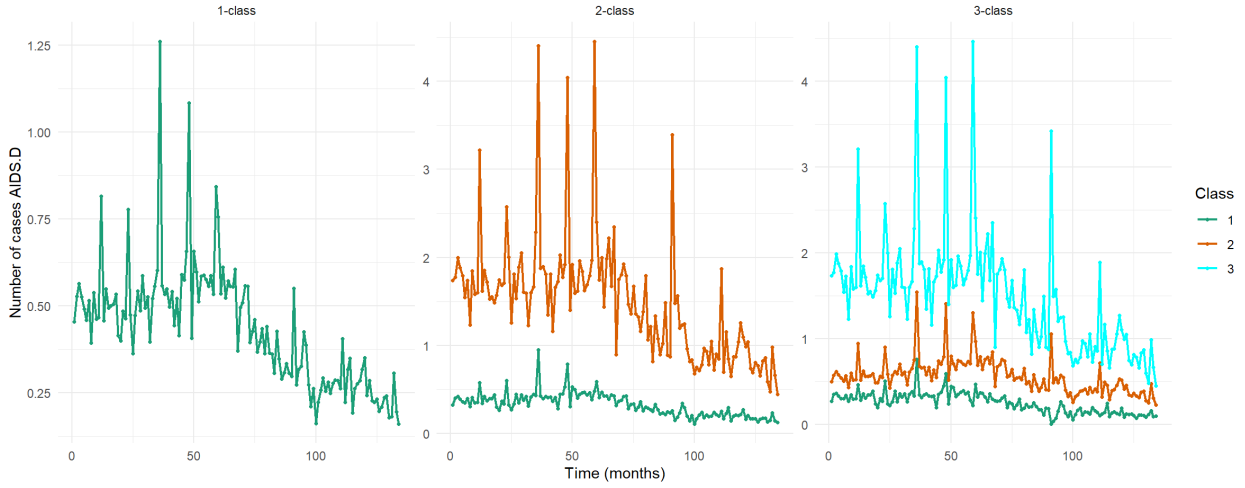

Рис. 5: Траєкторії середнього значення AIDS.D.

Із цих графіків, можна побачити, що незалежно від кількості класів та досліджуваного аспекту епідемії ВІЛ/СНІДу, у даних присутній спадний тренд, який є індикатором успіху превентивних кампаній у боротьбі з ВІЛ/СНІДом в Україні впродовж досліджуваного проміжку часу.

Також, можна побачити, що для кількості діагнозів СНІД та кількості смертей спричинених СНІД, розгляд моделі із трьома класами не породжує істотно нового класу порівняно із моделлю на два класи (додається промі-

жна градація між існуючими). Що спонукає нас на додаткове дослідження проблеми вибору комбінації досліджуваних моделей для найкращих результатів.

Це питання можна розв'язати багатьма способами, але класичним та перевіреним часом рішенням є застосування інформаційного критерію Байєса (BIC). Значення BIC для досліджуваних моделей представлено у таблиці 1:

Табл. 1: Значення Інформаційного критерію Байєса (BIC) для змішаних моделей (HIV.N, AIDS.N, та AIDS.D).

Variable	**1-class**	**2-class**	**3-class**
HIV.N.100k	3435.553	3418.599	3412.978
AIDS.N.100k	4668.3	4660.506	4664.915
AIDS.D.100k	5497.002	5478.181	5477.191

На основі цього, модель із трьома класами була обрана для опису поведінки HIV.N.100k та AIDS.D.100k, та модель із двома класами була обрана для опису поведінки AIDS.N.100k. Відсутність покращення інтерпретабельності моделі з трьома класами для AIDS.N.100K гарно ілюструється дестандартизованими лініями тренду на попередньому графіку. Будучи точними, доданий клас має лінію тренду, яка майже не відрізняється від лінії тренду вже існуючого класу і таким чином не допомагає моделі змістовніше описувати природу наявних даних.

Слід наголосити, що кожна окрема латентна змішана модель описує лише один аспект перебігу епідемії, а саме: динаміку нових діагнозів ВІЛ, нових діагнозів СНІДу або смертності, пов'язаної зі СНІДом. Тому використання лише однієї такої моделі не дозволяє отримати цілісну характе-

ристику регіону.

З цієї причини на наступному етапі було використано конкатеновані вектори ймовірностей належності до класів усіх моделей. Кореляційна структура таких векторів дозволяє виділити регіони, подібні не за одним окремим показником, а за сукупним профілем епідеміологічних тенденцій.

Після оцінювання параметрів та вибору моделей ми можемо отримати вектор ймовірності належності регіону до кожного класу відповідної моделі. Для кластеризації регіону ми розглядатимемо конкатенований вектор ймовірностей належності для кожного регіону.

Щоб виділити регіони із подібними епідеміологічними трендами, було побудовано зважений повний граф, де кожній вершині відповідає певний регіон, а сполучне ребро кожної пари вершин зважене на кореляцію між конкатенованими ймовірнісними векторами що відповідають регіонам із пари.

Для уникання інформаційного засмічення через неістотні зв'язки (наприклад слабко або від'ємно корельовані зв'язки), була застосована жорстка порогова фільтрація. Тобто, було обрано поріг τ, і ребра e з вагою $\omega(e) < \tau$ було вилучено.

Для інформованого вибору τ, ми дослідили частотну гістограму значень із кореляційної матриці конкатенованих ймовірнісних векторів регіонів:

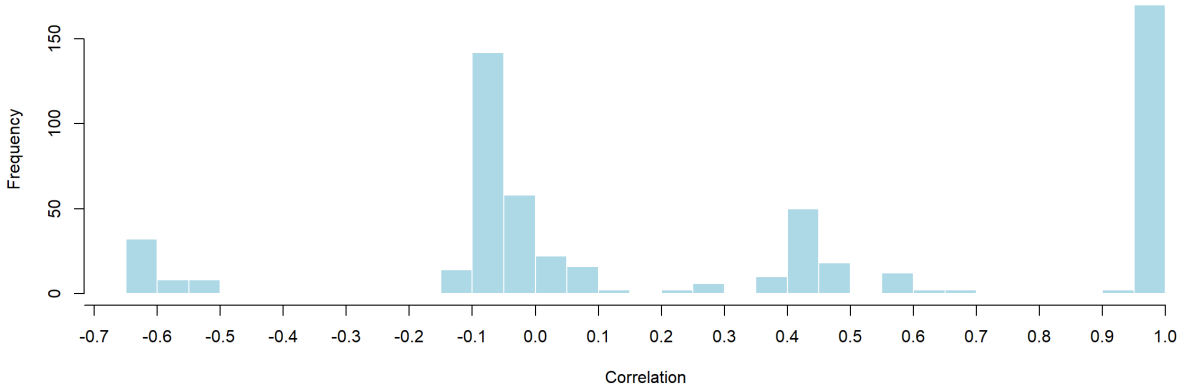

Рис. 6: Гістограма значень в кореляційній матриці

З гістограми видно, що більшість сильних кореляцій перевищують 0.9, тому для подальшого аналізу було обрано поріг $\tau = 0.9$.

Після жорсткої порогової фільтрації застосовано алгоритм Walktrap для кластеризації регіонів із подібними епідеміологічними трендами. Усі обчислення виконано за допомогою пакета `igraph` для мови програмування R.

Маючи кластеризовану мережу, проаналізуємо характеристики отриманих кластерів 2.

Табл. 2: Характеристика кластерів на основі присвоєних класів

Cluster	**Size**	**HIV.N Class**	**AIDS.N Class**	**AIDS.D Class**
1	9	1	1	1
2	9	3	1	2
3	2	2	1	2
4	2	3	2	3
5	1	1	1	2

Візуалізацію результуючої регіональної мережі наведено на рис. 7.

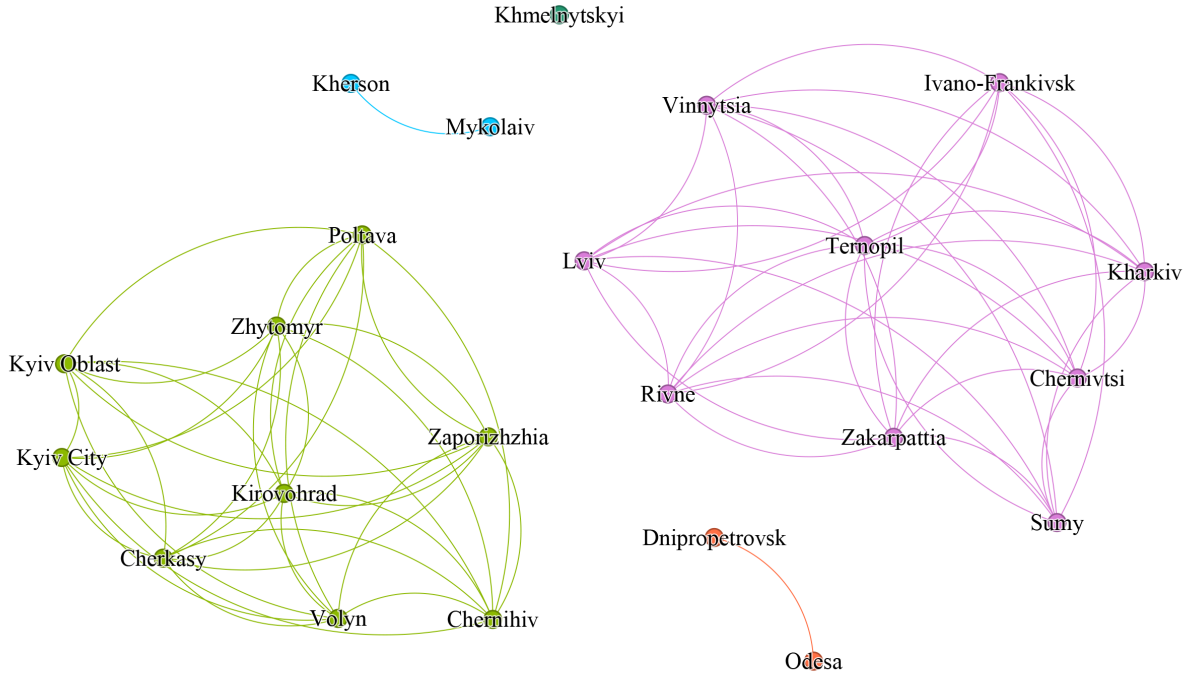

Рис. 7: Епідеміологічні регіональні кластери

Кластер 1 включає регіони з низькою кількістю діагнозів ВІЛ та СНІДу і низькою смертністю, пов'язаною зі СНІДом. Уявна помірність епідемії може бути наслідком недостатнього тестування або неповної звітності, а не справді низької поширеності.

Кластер 2 представляє регіони з помірною захворюваністю на ВІЛ, низькою кількістю випадків СНІДу та помірною смертністю (зазвичай менше однієї смерті на місяць). Відносно вища кількість діагнозів ВІЛ свідчить про ширше охоплення тестуванням і більш точне відображення епідемічної ситуації.

Кластер 3 охоплює регіони зі стабільною захворюваністю на ВІЛ, низькою кількістю випадків СНІДу та помірною смертністю. Хоча тенденції епідемії виглядають контрольованими, посилення інформаційних кампаній

могло б додатково знизити передачу ВІЛ, щоб ці регіони відповідали загальній тенденції до зниження.

Кластер 4 включає регіони, які історично мали найвищий тягар ВІЛ/СНІД — зокрема Одеську та Дніпропетровську області. Незважаючи на спадну тенденцію діагнозів ВІЛ, стабільно високі показники СНІДу та смертності свідчать про те, що недостатні профілактичні та інформаційні заходи залишаються ключовими чинниками тяжкості, що узгоджується з попередніми дослідженнями [11].

Кластер 5 включає регіони, ймовірно, з недодіагностованими випадками, хоча епідемія там менш тяжка, ніж у Кластері 4.

На основі результатів отриманих в цьому розділі, ми будемо досліджувати і будувати узагальнені моделі Крамера-Лундберга для кожного кластеру окремо, щоб отримати якісніші оцінки.

Результати отримані в цьому розділі було опубліковано у статті [17].

4 Побудова епідеміологічного марківського ланцюга

Для ґрунтовного дослідження епідемії ВІЛ/СНІДу в Україні дуже важливою є змога мати принаймні короткострокові адекватні прогнози кількості нових діагнозів та смертей.

Для розв'язання цієї задачі доцільною відправною точкою є побудова марківського ланцюга на популяційному рівні. Марковська властивість хоч і не узгоджується із реальною природою даних, не має значущого впливу на якість прогнозів у короткостроковій перспективі.

Розглядаючи марківську модель з дискретним часом на індивідуальному рівні, ми вважаємо, що час до переходу між станами геометрично розподілений, а загальна кількість переходів має біноміальний розподіл (з однаковим параметром ймовірності p та різним параметром кількості випробувань n для кожного місяця). Оцінка параметра p виводиться безпосередньо:

$$
\begin{aligned}
L(p) &= \prod_{i=1}^{n} C_{n_i}^{x_i} p^{x_i} (1-p)^{n_i - x_i}, \\
\mathcal{L}(p) &= \ln L(p) \\
&= \sum_{i=1}^{n} \ln C_{n_i}^{x_i} + x_i \ln p + (n_i - x_i) \ln(1-p), \\
\mathcal{L}'(p) &= \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \sum_{i=1}^{n} (n_i - x_i)
\end{aligned}
$$

$$= \frac{1}{p(1-p)} \left(\sum_{i=1}^{n} x_i - p \sum_{i=1}^{n} n_i \right) = 0,$$

$$\therefore \quad p = \frac{\sum x_i}{\sum n_i}.$$

Таким чином, оцінки для ймовірностей переходів виглядають наступним чином:

$$p_{U,H/A} = \frac{\sum_t H_N + \sum_t A_N}{\sum_t U},$$

$$p_{U,D} = \frac{\sum_t U_{D.CDA}}{\sum_t U},$$

$$p_{H/A,D} = \frac{\sum_t H_{D.CDA}}{\sum_t H} + \frac{\sum_t A_{D.CDA} + \sum_t A_{D.A}}{\sum_t A}$$

Оцінивши параметри моделі, можна адресувати проблему цензурування у наявних даних. Зокрема, можна знайти марківську модель із неперервним часом на індивідуальному рівні, яка при дискретизації неперервного часу помісячно перетворюється на модель Маркова із дискретним часом на індивідуальному рівні, яку ми вже оцінили на основі емпіричних даних. Для того щоб отримати матрицю переходу марківської моделі із дискретним часом на основі марківської моделі із неперервним часом, яка має інфінітезімальний генератор Q, використовується формула $P = e^{tQ}$; тож для реконструкції інфінітезімального генератора Q треба скористатися формулою $tQ = \ln P$.

Оцінивши параметри Марківської моделі із неперервним часом на індивідуальному рівні, ми маємо змогу перейти до марківської моделі на популя-

ційному рівні, використовуючи методологічні кроки, описані в [18]. Знаючи інфінітезімальний генератор Q та стартові кількості індивідів у кожному стані в популяції, вектор середніх очікуваних кількостей та коваріаційна матриця задаються наступними формулами [19]:

$$x(t) = x(0)e^{tQ}, \tag{1}$$

$$\mathcal{C}(t) = \mathrm{Cov}[x(t)] = \mathrm{Diag}(x(0)e^{tQ}) - e^{tQ^\top}\,\mathrm{Diag}(x(0))e^{tQ}. \tag{2}$$

Використовуючи ці формули, маємо емпіричну оцінку матриці переходу P для моделі із дискретним часом на індивідуальному рівні:

$$\begin{bmatrix} 0.998931 & 3.361732e-05 & 0.001035343 \\ 0 & 9.450082e-01 & 0.054991825 \\ 0 & 0 & 1 \end{bmatrix}$$

Оцінка матриці відповідної моделі із неперервним часом Q виглядає наступним чином:

$$\begin{bmatrix} -0.001069532 & 3.459568e-05 & 0.001034936 \\ 0 & -0.056561701 & 0.056561701 \\ 0 & 0 & 0 \end{bmatrix}$$

Для тестування розглянемо кількісний вектор популяції, який реалістично відображає пропорцію недіагностованих осіб та кількість діагностованих ЛЖВ:

$$\hat{x}(0) = (43192218,\ 171418,\ \_)$$

Покладемо кількість мертвих індивідів рівною нулю, оскільки нам пер-

шочергово цікаве моделювання зміни кількості ЛЖВ.

Матсподівання кількісного популяційного вектора через 1 місяць має вигляд:

$$\hat{x}(1) = (43145390,\ 153817,\ \_)$$

Абсолютна відсоткова різниця має вигляд при $t = 1$:

$$\delta x(1) = (0.02018511,\ -0.1086147,\ \_)$$

Очікуваний вектор після 3-х місяців спостережень виглядає як:

$$\hat{x}(3) = (43051887,\ 124446,\ \_)$$

Абсолютна відсоткова різниця порівняно із справжніми даними виглядає так при $t = 3$:

$$\delta x(3) = (-0.03802216,\ -0.2782137,\ \_)$$

Із цих двох прикладів можна побачити, що розглянута модель має тенденцію недооцінювати кількість ЛЖВ під час прогнозування. Це явище спричинене двома факторами: припущенням про експоненційний розподіл часу для переходів, неможливістю врахувати стартовий вік особи на момент початку спостереження (вік є принциповим фактором, що можна побачити із статево-вікової піраміди ЛЖВ).

Оцінкою коваріаційної матриці процесу на популяційному рівні є матри-

ця $\mathcal{C}(3)$:

$$\begin{bmatrix} 1795865.4 & -119254.9170 & -179489.9249 \\ -119254.9 & 124101.3380 & -518.8347 \\ -179489.9 & -518.8347 & 186522.1861 \end{bmatrix}$$

Для виконання моделювання використовувалась бібліотека `msm` мови програмування R [20] для проведення симуляцій на популяційному рівні. Для досягнення цього ми провели багато симуляцій на індивідуальному рівні і потім агрегували дані по цим симуляціям. Це дозволило нам побудувати оцінки параметрів для моделі на популяційному рівні. Проводячи численні симуляції, ми отримали емпіричні довірчі інтервали для значень процесу на популяційному рівні у різні моменти часу.

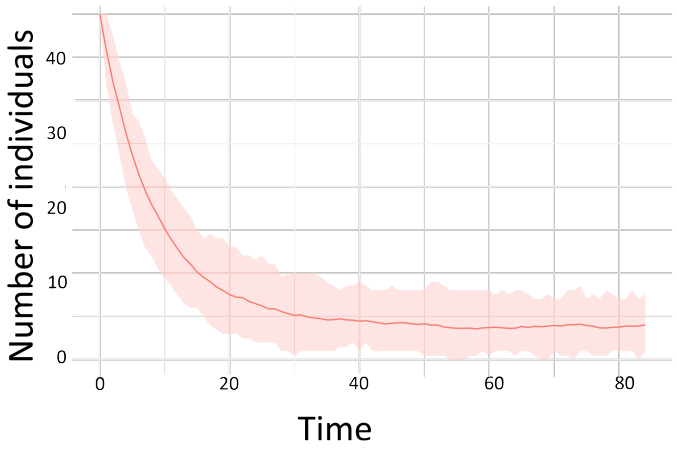

Рис. 8: Довірчий інтервал для кількості людей із ВІЛ/СНІД при населенні 10 000 та пропорційних підрахунках.

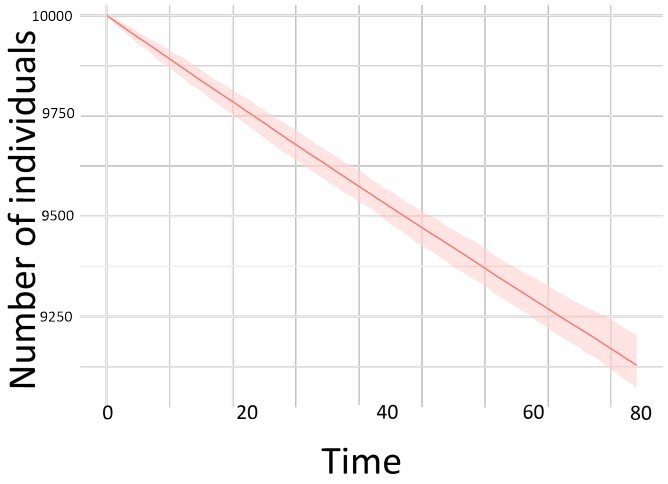

Рис. 9: Довірчий інтервалдля кількості недіагностованих людей із ВІЛ/СНІД або без нього при населенні 10 000 та пропорційних підрахунках.

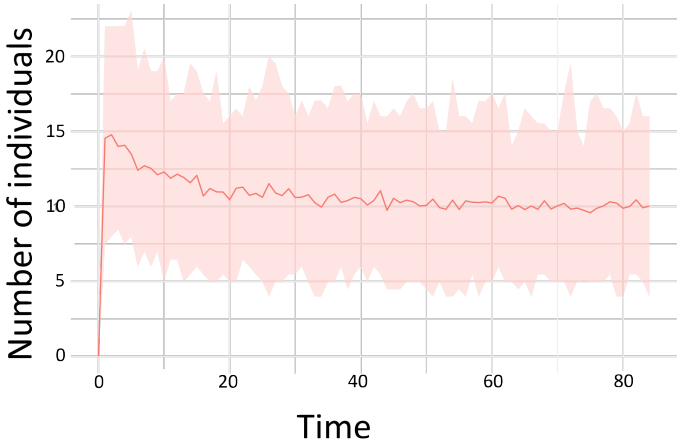

Рис. 10: Довірчий інтервал для кількості померлих людей із ВІЛ/СНІД або без нього при населенні 10 000 та пропорційних підрахунках.

Під час моделювання ми не розглядали зростання популяції людей в межах дослідження, оскільки зібрані дані стосуються осіб віком 15 і старше. Вплив кількості парентеральних інфікувань, як можна побачити в [3], є нехтовним. Також, згідно з [9], діти, народжені ЛЖВ, що мають змогу народжувати, проходять спеціальні процедури профілактики, і менше ніж 1% таких дітей заражаються на ВІЛ; таким чином, врахування інфікувань новонароджених не є необхідним для такого часового проміжку спостере-

жень [8]. А врахування малої кількості дітей та осіб підліткового віку, що досягли віку 15 і старше в межах дослідження, не можливо врахувати на основі наявних даних. Керуючись кредо GBD (Global Burden of Disease), перевага надається моделі з можливістю методологічного вдосконалення над відсутністю моделі як такої.

Результати отримані в цьому розділі опубліковано у роботі [21]

5 Оцінка параметрів моделі

У цьому розділі розглядається задача оцінки розподілу часу переходів між станами моделі на рівні окремого індивіда: розподілу переходу зі стану U у стан H/A, з U у D та з H/A у D.

Природним методом оцінки для наших даних є алгоритм Тьорнбула [22], з огляду на інтервально-цензурований характер спостережень (кожен випадок фіксується в межах звітного місяця, проте точний час події невідомий).

У класичній постановці алгоритму Тьорнбула для кожного індивіда відомий проміжок часу, протягом якого могла статися страхова подія. Натомість наші дані мають агрегований формат із помісячними кількостями діагнозів ВІЛ/СНІДу. У межах оцінки розподілу часу до діагностування смерть індивіда трактується як вихід із дослідження, тобто відповідне спостереження є правосторонньо цензурованим. Слід зазначити, що з поточних робочих даних неможливо оцінити розподіл часу до інфікування, тому дослідження обмежується розподілом часу до діагностування.

У загальній постановці алгоритму Тьорнбула для кожного індивіда i задано інтервал спостережуваної події $A_i = [L_i, R_i]$ та інтервал усічення $B_i \supset A_i$. Позначимо через $[p_j, q_j]$, $j = 1, \ldots, m$, усі максимальні перетини інтервалів A_i — так звані інтервали Тьорнбула. Алгоритм зводить непараметричну максимізацію функції правдоподібності до пошуку ймовірностей $s_1, \ldots, s_m$ цих інтервалів через ітеративну процедуру EM-типу з кроками:

$$\mu_{ij}(\mathbf{s}) = \frac{\alpha_{ij} s_j}{\sum_k \alpha_{ik} s_k}, \qquad \nu_{ij}(\mathbf{s}) = \frac{(1 - \beta_{ij}) s_j}{\sum_k \beta_{ik} s_k},$$

$$s_j^{\text{new}} = \frac{\sum_{i=1}^{N} [\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})]}{\sum_{j=1}^{m} \sum_{i=1}^{N} [\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})]},$$

де $\alpha_{ij} = \mathbf{1}\{[p_j, q_j] \subset A_i\}$ та $\beta_{ij} = \mathbf{1}\{[p_j, q_j] \subset B_i\}$.

Структура наявних даних є суттєво простішою порівняно із загальними припущеннями алгоритму, що дозволяє отримати значно компактніший варіант процедури переоцінювання. Зокрема, виконуються такі умови:

- інтервали цензурування збігаються з інтервалами Тьорнбула, тобто $A_i = [p_{j_i}, q_{j_i}]$ для деякого j_i;
- кожен індивід належить рівно до одного інтервалу Тьорнбула;
- усічення відсутнє: $\beta_{ij} \equiv 1$, тож $\nu_{ij} \equiv 0$;
- деякі індивіди випадають зі спостереження з причин, не пов'язаних із досліджуваною подією (правостороннє цензурування).

Розіб'ємо індивідів на два класи: $\mathcal{O}$ — ті, для яких подія спостережена в інтервалі j_i, та $\mathcal{C}$ — правосторонньо цензуровані у момент $C_i \in [p_{\ell_i}, q_{\ell_i})$. Введемо агреговані величини:

$$n_j = |\{i \in \mathcal{O} : j_i = j\}|, \qquad m_\ell = |\{i \in \mathcal{C} : C_i \in [p_\ell, q_\ell)\}|,$$

тобто n_j — кількість спостережених подій в інтервалі j, а m_ℓ — кількість індивідів, що вибули зі спостереження в інтервалі ℓ.

Оскільки всі m_ℓ цензурованих з інтервалу ℓ мають однаковий знаменник $S_\ell = \sum_{k>\ell} s_k$ (хвостова ймовірність після інтервалу ℓ), крок E спрощується

до:

$$\mu_{ij} = \begin{cases} 1, & i \in \mathcal{O},\ j = j_i, \\ 0, & i \in \mathcal{O},\ j \neq j_i, \\ \dfrac{s_j}{S_{\ell_i}}, & i \in \mathcal{C},\ j > \ell_i, \\ 0, & i \in \mathcal{C},\ j \leq \ell_i. \end{cases}$$

Підставляючи у крок M та агрегуючи по інтервалах, отримуємо спрощену формулу переоцінювання:

$$\boxed{s_j^{\text{new}} = \frac{1}{N}\left(n_j + s_j \sum_{\ell<j} \frac{m_\ell}{S_\ell}\right)}, \tag{3}$$

де $S_\ell = \sum_{k>\ell} s_k$. Умова самоузгодженості $s_j = s_j^{\text{new}}$ дає явний вираз:

$$s_j = \frac{n_j/N}{1 - \sum_{\ell<j} \dfrac{m_\ell}{S_\ell}}, \tag{4}$$

що є прямим аналогом формули Каплана–Мейєра, адаптованим до інтервально цензурованих даних із структурою, що збігається з інтервалами Тьорнбула. Змістовна інтерпретація формули (3) є такою: перший доданок n_j відповідає прямо спостереженим подіям в інтервалі j, тоді як другий доданок акумулює очікуваний внесок від цензурованих індивідів — кожна група m_ℓ рівномірно перерозподіляє свою масу між інтервалами $j > \ell$ пропорційно до поточних ймовірностей s_j.

Таким чином, алгоритм Тьорнбула у спрощеній формі (3) дозволяє отримати непараметричну оцінку функції розподілу часу переходу між станами моделі. Однак така оцінка є усіченою на максимальному спостережуваному

значенні, і для більш інформативної інтерпретації доцільно застосовувати параметричний підхід. Параметричні функції розподілу дискретизуються на інтервали Тьорнбула та порівнюються з непараметричною оцінкою для вибору класу розподілу, який найкраще описує явище.

Для оцінки близькості функцій розподілу застосовуються метрики Колмогорова та Крамера–фон Мізеса. Метрика Колмогорова чутлива до локальних відхилень навіть на множині малої міри, але менш інформативна щодо накопичених помилок на всій області. Метрика Крамера–фон Мізеса, навпаки, добре оцінює сумарну близькість функцій, але менш чутлива до локальних аномалій. У нашому випадку, оскільки функції розподілу за своєю природою не мають аномальних локальних стрибків, доцільним є використання метрики Крамера–фон Мізеса для порівняння параметричних функцій із непараметричною оцінкою (рис. 11, 12).

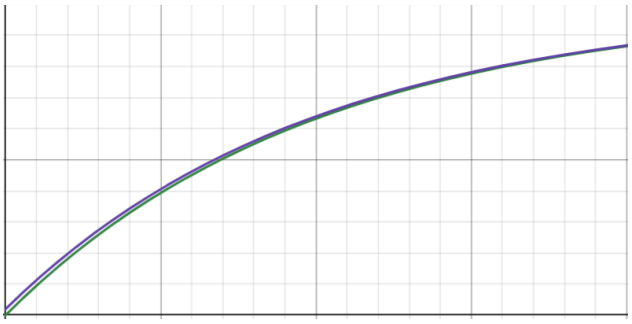

Рис. 11: Ситуація, у якій підхід Колмогорова не є ефективним.

На основі виділених раніше кластерів було визначено найкращі розподіли серед класів Вейбула, логнормальних та гамма-розподілів за критерієм мінімуму статистики Крамера–фон Мізеса, результати представлено на рис. 13–17.

Із представлених результатів видно, що систематично найкращу відпо-

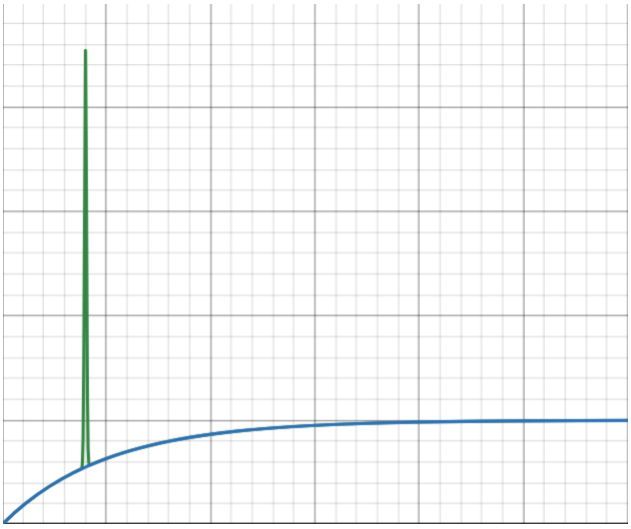

Рис. 12: Ситуація, у якій підхід Крамера–фон Мізеса не є ефективним.

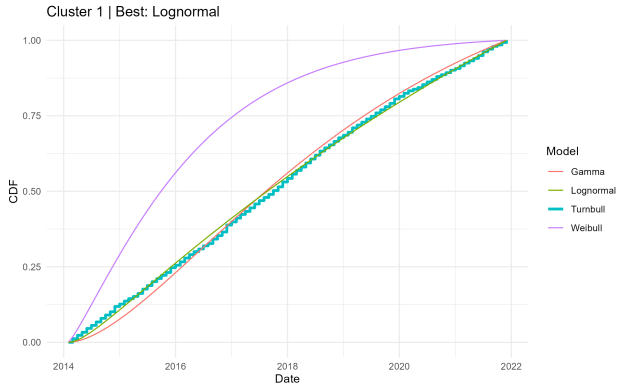

Рис. 13: Точність найкращих моделей для першого епідеміологічного кластеру.

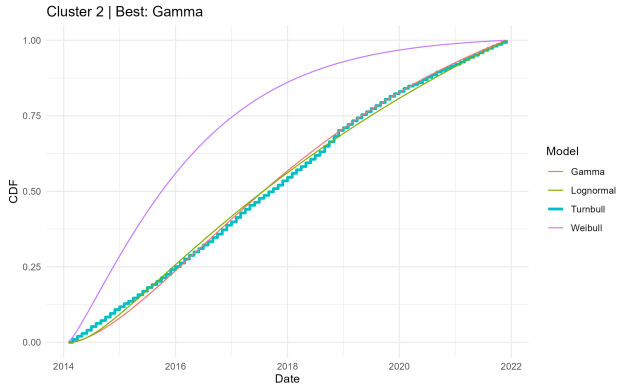

Рис. 14: Точність найкращих моделей для другого епідеміологічного кластеру.

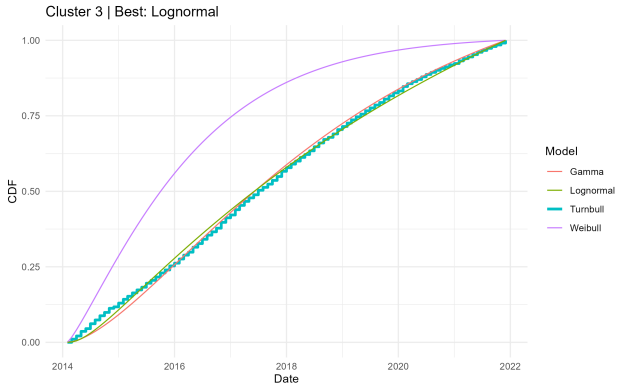

Рис. 15: Точність найкращих моделей для третього епідеміологічного кластеру.

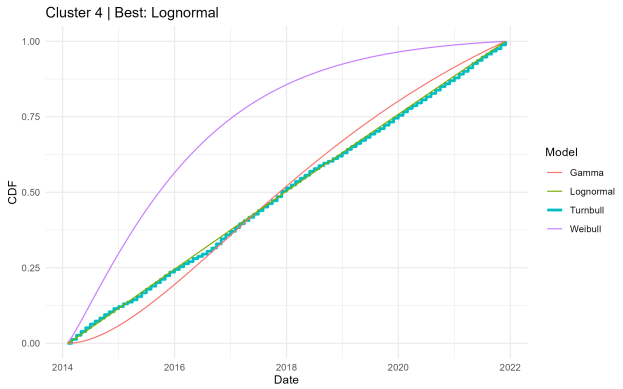

Рис. 16: Точність найкращих моделей для четвертого епідеміологічного кластеру.

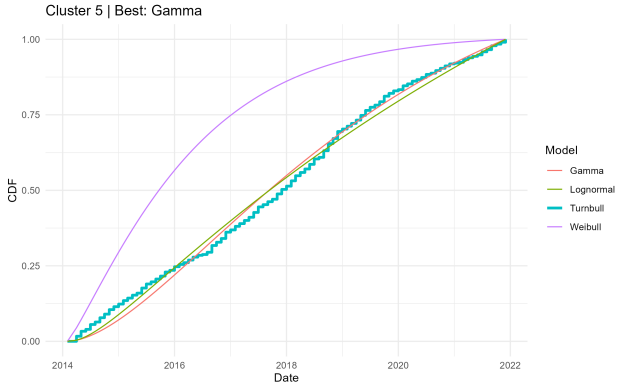

Рис. 17: Точність найкращих моделей для п’ятого епідеміологічного кластеру.

відність — найменше значення статистики Крамера–фон Мізеса — демонструє клас логнормальних розподілів.

За результатами цього розділу підготовано статтю, яку прийнято до друку в журнал "Вісник Київського національного університету імені Тараса Шевченка. Фізико-математичні науки"

6 Дослідження ймовірності банкрутства

Отримавши оцінки розподілу часів між діагнозами ВІЛ/СНІД, природним є побудова моделі Крамера–Лундберга для дослідження ймовірності банкрутства однорідного страхового портфелю, де страхова виплата здійснюється у разі встановлення діагнозу ВІЛ/СНІД.

У класичній моделі Крамера–Лундберга лічильний процес страхових випадків є пуасонівським, тобто інтервали часу між страховими подіями мають експоненційний розподіл. У нашому випадку доцільно розглядати узагальнену модель Спарре–Андерсена, де лічильний процес є процесом відновлення із логнормально розподіленими інтервалами часу між подіями. Керуючись термінологією з [23], покладемо розмір збитку фіксованою константою $y > 0$, а час між страховими подіями — $T \sim \text{LogNormal}(\mu, \sigma^2)$.

Процес капіталу страхової компанії в момент часу t має вигляд:

$$C_t = u + ct - yN(t),$$

де $u \geq 0$ — початковий капітал, $c > 0$ — ставка надходження премій, $N(t)$ — процес відновлення із T-розподіленими міжподієвими часами.

Інтегральне рівняння ймовірності банкрутства

Позначимо через $\varphi(u)$ ймовірність банкрутства, тобто ймовірність того, що існує скінченний момент часу τ такий, що $C_\tau < 0$. Для виведення рівняння для $\varphi(u)$ скористаємось умовою на час τ_1 настання першої страхової події.

У момент τ_1 капітал компанії безпосередньо після виплати становить $C_{\tau_1} = u + c\tau_1 - y$. Тут природно виділити два взаємовиключних випадки:

- якщо $u + c\tau_1 - y < 0$, тобто $\tau_1 < \frac{y-u}{c}$, настає **негайне банкрутство**; цей сценарій можливий лише при $u < y$, коли стартовий капітал не покриває першої виплати;
- якщо $u + c\tau_1 - y \geq 0$, тобто $\tau_1 \geq \frac{y-u}{c}$, банкрутства не сталось, і завдяки марківській властивості процесу відновлення він **перезапускається** з нового капіталу $u + c\tau_1 - y$.

Інтегруючи по всіх можливих значеннях τ_1 із щільністю $f_T(t)$, отримуємо інтегральне рівняння відновлення для $\varphi(u)$:

$$\varphi(u) = \int_0^{\max(0,\, y-u)} f_T(t)\, dt + \int_{\max(0,\, y-u)}^{\infty} \varphi(u + ct - y)\, f_T(t)\, dt. \tag{5}$$

Підставляючи логнормальну щільність $f_T(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\!\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right)$, рівняння (5) набуває вигляду:

$$\varphi(u) = \Phi\!\left(\frac{\ln(\max(y-u,\,0)/c) - \mu}{\sigma}\right) + \frac{1}{\sigma\sqrt{2\pi}} \int_{\max(0,\, y-u)}^{\infty} \frac{\exp\!\left(-\dfrac{(\ln t - \mu)^2}{2\sigma^2}\right)}{t}\, \varphi(u + ct - y)\, dt. \tag{6}$$

де Φ — функція розподілу стандартної нормальної випадкової величини. Рівняння (6) є точним інтегральним рівнянням відновлення по капіталу u, придатним як для якісного аналізу, так і для числового розв'язання методом послідовних наближень.

Зауважимо, що структура рівняння (6) якісно різниться залежно від співвідношення між u та y:

- при $u \geq y$ перший доданок обертається на нуль, оскільки жоден перший збиток не може спричинити негайного банкрутства, і рівняння спрощується до:

$$\varphi(u) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty \frac{\exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right)}{t} \varphi(u + ct - y)\, dt;$$

- при $u < y$ існує ненульова ймовірність негайного банкрутства від першої виплати, що відображається у ненульовому першому доданку.

Константа Крамера–Лундберга для гамма-розподілених часів

Іншим важливим питанням є асимптотична поведінка $\varphi(u)$ при $u \to \infty$. Класична теорія передбачає експоненційне згасання:

$$\varphi(u) \sim Ce^{-Ru}, \quad u \to \infty,$$

де $R > 0$ — константа Крамера–Лундберга. Її знаходження вимагає існування твірної функції моментів розподілу інтервалів часу між подіями. Оскільки логнормальний розподіл не має скінченної твірної функції моментів на всій додатній півосі, розглянемо гамма-розподіл із параметрами $k > 0$ та $\theta > 0$, який також добре описує емпіричні розподіли інтервалів часу між діагнозами (див. Розділ 5) і для якого відповідна твірна функція моментів існує:

$$M_T(r) = (1 - r\theta)^{-k}, \quad r < \frac{1}{\theta}.$$

Згідно з лемою 6.1 з [23], якщо існує $\widehat{\theta}(r)$ така, що виконується рівність

$$M_Y(r)\, M_T\big(-\widehat{\theta}(r) - cr\big) = 1,$$

то процес $e^{-rC_t - \widehat{\theta}(r)T_t}$ є мартингалом, де T_t — сумарний час до моменту t. Константа Крамера–Лундберга R відповідає додатному розв'язку рівняння $\widehat{\theta}(r) = 0$.

Оскільки збиток є фіксованим $Y = y$, то $M_Y(r) = e^{ry}$, а інтервали часу між вимогами мають гамма-розподіл із параметрами форми k та швидкості β, тому $M_T(r) = \big(1 - r/\beta\big)^{-k}$. Підставляючи у рівняння леми, отримуємо:

$$e^{yr} \cdot \left(1 + \frac{\theta(r) + cr}{\beta}\right)^{-k} = 1. \tag{7}$$

Розділимо обидві частини на e^{yr} та піднесемо до степеня $-1/k$:

$$1 + \frac{\theta(r) + cr}{\beta} = e^{yr/k}. \tag{8}$$

Звідси одразу знаходимо $\widehat{\theta}(r)$:

$$\theta(r) = \beta\Big(e^{yr/k} - 1\Big) - cr. \tag{9}$$

Знаходження константи Крамера–Лундберга

Константа R є додатним розв'язком рівняння $\theta(r) = 0$, тобто

$$\beta\Big(e^{yr/k} - 1\Big) = cr. \tag{10}$$

Існування такого додатного кореня обґрунтуємо аналізом функції $f(r) = \beta(e^{yr/k} - 1) - cr$. Маємо $f(0) = 0$ та

$$f'(0) = \frac{\beta y}{k} - c. \tag{11}$$

За умовою чистого прибутку (net profit condition) середній прибуток за один цикл є додатним, що означає $c > \beta y/k$, тобто $f'(0) < 0$. Оскільки $f(r) \to +\infty$ при $r \to +\infty$, а $f(0) = 0$ і $f'(0) < 0$, функція f спочатку спадає, а потім зростає до $+\infty$, тому існує рівно один додатний корінь $R > 0$.

Для знаходження R у явному вигляді зробимо заміну $u = yr/k$, тоді $r = ku/y$, і позначимо $\lambda = ck/(\beta y)$. За умовою чистого прибутку $\lambda > 1$. Рівняння (10) набуває вигляду

$$e^u = \lambda u + 1. \tag{12}$$

Шуканий корінь відповідає $u^* > 0$. Зробимо підстановку $v = \lambda u + 1$, тобто $u = (v - 1)/\lambda$:

$$\begin{aligned} e^{(v-1)/\lambda} &= v, \\ e^{v/\lambda} \cdot e^{-1/\lambda} &= v, \\ e^{-1/\lambda} &= v\, e^{-v/\lambda}. \end{aligned} \tag{13}$$

Помножимо обидві частини на $(-1/\lambda)$:

$$-\frac{1}{\lambda}\, e^{-1/\lambda} = -\frac{v}{\lambda}\, e^{-v/\lambda}. \tag{14}$$

Рівняння (14) має вигляд $Xe^X = A$, де $X = -v/\lambda$ та $A = -(1/\lambda)e^{-1/\lambda}$.

Вибір гілки функції Ламберта. Аргумент функції Ламберта дорівнює $A = -(1/\lambda)\, e^{-1/\lambda}$. Оскільки $\lambda > 1$, маємо $1/\lambda \in (0, 1)$, тому $e^{-1/\lambda} \in (e^{-1}, 1)$, і отже

$$A = -\frac{1}{\lambda}\, e^{-1/\lambda} \in \left(-\frac{1}{e},\, 0\right). \tag{15}$$

На проміжку $(-1/e,\, 0)$ рівняння $We^W = A$ має *два* дійсні розв'язки. Гілка W_{-1} дає $W_{-1}(A) \in (-\infty, -1)$ і відповідає $v > \lambda$, тобто $u > 0$, що відповідає шуканому додатному кореню $R > 0$. Гілка W_0 дає $W_0(A) \in (-1, 0)$ і відповідає тривіальному кореню $r = 0$.

Отже, застосовуємо W_{-1} до обох частин (14):

$$\begin{aligned} -\frac{v}{\lambda} &= W_{-1}\left(-\frac{1}{\lambda}\, e^{-1/\lambda}\right), \\ v &= -\lambda\, W_{-1}\left(-\frac{1}{\lambda}\, e^{-1/\lambda}\right). \end{aligned} \tag{16}$$

З означення $v = \lambda u + 1$ та $u = yR/k$ маємо $v = cR/\beta + 1$, тому

$$\begin{aligned} \frac{cR}{\beta} + 1 &= -\lambda\, W_{-1}\left(-\frac{1}{\lambda}\, e^{-1/\lambda}\right), \\ \frac{cR}{\beta} &= -1 - \lambda\, W_{-1}\left(-\frac{1}{\lambda}\, e^{-1/\lambda}\right). \end{aligned} \tag{17}$$

Повертаючись до $\lambda = ck/(\beta y)$, остаточно отримуємо:

$$R = -\frac{\beta}{c}\left[1 + \frac{ck}{\beta y} W_{-1}\left(-\frac{\beta y}{ck} e^{-\beta y/(ck)}\right)\right]. \tag{18}$$

Зазначимо, що $W_{-1}(A) < -1$ при $A \in (-1/e, 0)$, тому $\lambda W_{-1}(A) < -\lambda < -1$, і отже $-1 - \lambda W_{-1}(A) > 0$, що підтверджує $R > 0$.

Таким чином ми показали, що і логнормальна і гамма модель мають теоретичні переваги в аналізі поведінки банкрутства страхового портфелю,

де страховою подією вважається діагностування страхувальником ВІЛ або СНІД. Зокрема логнормальна модель краще описує наявні дані, але не дає змоги досліджувати асимптотичну поведінку банкрутства класичними методами, а гамма модель, хоча і дає незначне погіршення в якості наближення емпіричних даних, дає змогу досліджувати асимптотичну поведінку ймовірності банкрутства використовуючи стандарнту методологію.

Результати цього розділу було представлено на XIV Всеукраїнській науковій конференції молодих математиків [24].

Висновки

У роботі розроблено багаторівневий підхід до моделювання епідеміологічного процесу ВІЛ/СНІД в Україні з використанням методів стохастичного моделювання та актуарної математики.

Основні наукові результати роботи можна узагальнити таким чином:

- розроблено та реалізовано повну процедуру підготовки, очищення та структуризації епідеміологічних даних, включаючи обробку OCR-похідних похибок та відновлення пропущеної інформації;
- побудовано індивідуальну стохастичну модель епідеміологічного процесу ВІЛ/СНІД, що описує переходи між станами U, H/A та D;
- оцінено параметри марківського ланцюга як у дискретному, так і в неперервному часі, а також побудовано відповідну популяційну модель з аналітичними виразами для математичного сподівання та коваріаційної структури;
- розроблено методологію короткострокового прогнозування кількості діагнозів та смертей на основі стохастичного моделювання та агрегування індивідуальних траєкторій;
- здійснено кластеризацію регіонів України за спільними епідеміологічними трендами із використанням латентних класових моделей та мережевого підходу на основі кореляцій структурних векторів;
- виявлено, що логнормальний розподіл найкраще описує емпіричний розподіл часу між діагнозами ВІЛ/СНІД;

- виведено узагальнену модель Крамера–Лундберга для процесу відновлення з гамма-розподіленими інтервалами часу між подіями та отримано аналітичний вираз для критичного параметра - коефіцієнта Крамера-Лундберга.

Подальші напрями дослідження включають якісне дослідження інтегрального рівняння для ймовірності банкрутства, оцінювання частки ЛЖВ серед недіагностованої популяції, побудову напівмарківських моделей, що дозволять здійснювати більш точне довгострокове прогнозування кількості ЛЖВ та нових діагнозів, а також порівняльний аналіз ефективності заходів з протидії поширенню ВІЛ/СНІДу в Україні на основі отриманих кластерів та їх характеристик.

Додатково перспективним є врахування коваріатів віку та статі для покращення точності моделювання, однак поточна структура доступних даних не дозволяє коректно інтегрувати віково-статеву стратифікацію у межах цього дослідження.

Результати дослідження можуть бути використані фахівцями у сфері громадського здоров'я, діяльність яких пов'язана з протидією поширенню ВІЛ/СНІДу в Україні; дослідниками, що працюють над удосконаленням статистичних оцінок у межах проєкту Global Burden of Disease, а також актуаріями при моделюванні ВІЛ/СНІДу як страхового ризику.

Список використаних джерел

[1] Kruglov Y., Kobyshcha Y., Salyuk T., Varetska O., Shakarishvili A., Saldanha V. The most severe HIV epidemic in Europe: Ukraine's national HIV prevalence estimates for 2007. *Sexually Transmitted Infections.* 2008. Vol. 84, No. 1. P. 37–41. DOI: https://doi.org/10.1136/sti.2008.031195.

[2] Barnett T., Whiteside A., Khodakevich L., Kruglov Y., Steshenko V. The HIV/AIDS epidemic in Ukraine: Its potential social and economic impact. *Social Science Medicine.* 2000. Vol. 51, No. 9. P. 1387–1403. DOI: https://doi.org/10.1016/s0277-9536(00)00104-0.

[3] State Enterprise "Center of Public Health" of the Ministry of Health of Ukraine. *Statistics on HIV/AIDS in Ukraine*, 2025. Available at: https://phc.org.ua/kontrol-zakhvoryuvan/vilsnid/statistika-z-vilsnidu.

[4] State Statistics Service of Ukraine. *Mortality Rates by Regions of Ukraine, 2015–2022* [Data set], 2022. Available at: https://www.ukrstat.gov.ua/operativ/operativ2022/ds/kjp_reg/arh_kjp_reg2022_u.html.

[5] Є. М. Окунєв, О. І. Василик, *Підготовка даних щодо епідемії ВІЛ/СНІД для актуарного дослідження*, XIII Всеукраїнська наукова конференція молодих математиків: тези доповідей, Київ, 2025, с. 41–42.

[6] Institute for Health Metrics and Evaluation (IHME), Bill and Melinda Gates Foundation (BMGF), Premise Data Corporation. *Premise General Population COVID-19 Health Services Disruption Survey 2020.* Seattle,

United States of America: Institute for Health Metrics and Evaluation (IHME), 2021. https://doi.org/10.6069/hvft19v3

[7] Lu, H., Cole, S. R., Westreich, D., Hudgens, M. G., Adimora, A. A., Althoff, K. N., et al. (2022). Virologic outcomes among adults with HIV using integrase inhibitor-based antiretroviral therapy. *AIDS*, 36(2), 277–286.

[8] Kleinman S., Busch M. The risks of transfusion-transmitted infection: direct estimation and mathematical modelling. *Best Practice Research Clinical Haematology.* 2000. Vol. 13, No. 4. P. 631–649. DOI: https://doi.org/10.1053/beha.2000.0104.

[9] Fowler M., Flynn P., Aizire J. What is new in perinatal HIV prevention? *Current Opinion in Pediatrics.* 2018. Vol. 30, No. 1. P. 144–151. DOI: https://doi.org/10.1097/mop.0000000000000579.

[10] Raboud J. et al. Consecutive rebounds in plasma viral load are associated with virological failure at 52 weeks among HIV-infected patients. *AIDS.* 2002. Vol. 16, No. 12. P. 1627–1632. DOI: https://doi.org/10.1097/00002030-200208160-00008.

[11] Neduzhko O. et al. Factors associated with delayed enrollment in HIV medical care among HIV-positive individuals in Odessa Region, Ukraine. *Journal of the International Association of Providers of AIDS Care.* 2017. Vol. 16, No. 2. P. 168–173. DOI: https://doi.org/10.1177/2325957416686194.

[12] Vasylyeva T. I. et al. Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine. *PNAS.* 2018. Vol. 115, No. 5. P. 1051–1056. DOI: https://doi.org/10.1073/pnas.1701447115.

[13] Kovalenko G. et al. Phylodynamics and migration data help describe HIV transmission dynamics in internally displaced people who inject drugs in Ukraine. *PNAS Nexus.* 2023. Vol. 2, No. 3. p. gad008. DOI: https://doi.org/10.1093/pnasnexus/pgad008.

[14] Miranda M. N. S. et al. Determinants of HIV-1 late presentation in patients followed in Europe. *Pathogens.* 2021. Vol. 10, No. 7. P. 835. DOI: https://doi.org/10.3390/pathogens10070835.

[15] Parczewski M. et al. Control of HIV across the WHO European region: progress and remaining challenges. *The Lancet Regional Health–Europe.* 2025. Vol. 52. P. 1–14. DOI: https://doi.org/10.1016/j.lanepe.2025.101243.

[16] G'omez G., Calle M. L., Oller R., Langohr K. Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling.* 2009. Vol. 9, No. 4. P. 259–297.

[17] Окунєв, Є. М. (2025). Кластеризація регіонів України на основі епідеміологічної ситуації ВІЛ/СНІД. Науковий вісник Ужгородського університету. Серія «Математика і інформатика», 47(2), 199–206. https://doi.org/10.24144/2616-7700.2025.47(2).199-206

[18] Johnson N., Kemp A., Kotz S. *Univariate discrete distributions.* Wiley, 2005. DOI: https://doi.org/10.1002/0471715816.

[19] Cox D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B.* 1972. Vol. 34, No. 2. P. 187–202.

[20] Jackson C. *Multi-state modelling with R: the msm package.* Cambridge, 2007. DOI: `https://doi.org/10.32614/CRAN.package.msm`.

[21] O. Vasylyk, Y. Okuniev, *Modelling HIV/AIDS Epidemic in Ukraine as Continuous Time Markov Process*, Bulletin of Taras Shevchenko National University of Kyiv. Physics and Mathematics, accepted for publication, 2025.

[22] Turnbull B. W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B.* 1976. Vol. 38, No. 3. P. 290–295.

[23] Schmidli H. *Risk theory.* Springer, 2017.

[24] Y. Okuniev, *Sparre Andersen Risk Modeling with Gamma-Fitted Interarrival Times under Interval Censoring*, XIV Всеукраїнська наукова конференція молодих математиків: тези доповідей, 2026, pp. 150–151.