Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

ДОСЛІДЖЕННЯ НЕВИЗНАЧЕНИХ ЗНАНЬ ТА КІЛЬКІСНА ОЦІНКА НЕВИЗНАЧЕНОСТІ

(тема)

Виконав:
студент 2 курсу, групи ІНФм-19-1
Чугайов О.О.
(прізвище, ініціали)

Спеціальності 122 Комп'ютерні науки
(код і повна назва спеціальності)

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник проф. Кузьомін О.Я.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Кобилін О.А.
(підпис) (прізвище, ініціали)

2020 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра Інформатики
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Освітня програма Інформатика
(повна назва освітньої програми)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Чугайову Олексію Олександровичу
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження невизначених знань та кількісна оцінка невизначеності»

затверджена наказом по університету від «_23_» жовтня 2020 року № 1428Ст.

2. Термін подання студентом роботи до екзаменаційної комісії _2_ ___грудня____ 2020 р.

3. Вихідні дані до роботи статистика з заповненими пропусками,
перелік використовуваних програмних засобів: Intellij IDEA, Delphi.

4. Перелік питань, що потрібно опрацювати в роботі
1. Аналіз предметної області та постановка задачі дослідження
2. Аналіз методів оцінки невизначеності
3. Відновлення даних
4. Алгоритм доповнення даних
5. Програмна реалізація оцінки невизначеності

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) Комп'ютерна презентація

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на атестаційну роботу	23.10.2020	
2	Аналіз завдання, підбір літератури	23.10.20-26.10.20	
3	Аналіз літератури з досліджуваної проблеми	26.10.20-02.11.20	
4	Аналіз технічних засобів для реалізації	02.11.20-06.11.20	
5	Проектування системи	06.11.20-10.11.20	
6	Програмна реалізація	10.11.20-16.11.20	
7	Оформлення пояснювальної записки	16.11.20-30.12.20	
8	Перевірка на плагіат	30.11.20	
9	Рецензування	05.12.20	
10	Підготовка презентації та доповіді	06.12.20	
11	Занесення роботи в електронний архів	08.12.20	
12	Попередній захист атестаційної роботи	09.12.20	

Дата видачі завдання __23_ __жовтня___ 2020 р.

Студент _____
(підпис)

Керівник роботи _____ проф. Кузьомін О. Я.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до атестаційної роботи: 51 с., 6 табл., 10 рис., 33 джерела.

НЕВИЗНАЧЕНІ ДАНІ, АЛГОРИТМ МОНТЕ-КАРЛО, ВІДНОВЛЕННЯ ДАНИХ, МЕТОД ПОМИЛОК, ОЦІНКА НЕВИЗНАЧЕННОСТІ, ФОРМУЛА ПІРСОНА, МЕДИЧНИЙ АГЕНТ.

Об’єктом дослідження є дослідження невизначених даних та порівняння різних алгорітмів пошуку пропусків.

У цьому дослідженні інструментальним засобом для аналізу ступенів впевненості буде теорія імовірностей, яка зв’язана з баєсовим представленням даних і знань про багатофакторний стан проблеми

У даній дипломній роботі було проведено дослідження на тему відновлення пропусків у неповних даних. Також будуть розглянуті механізми породження перепусток і деякі популярні методи роботи з неповнотою даних.

UNCERTAINTY DATA, MONTE CARLO ALGORITHM, DATA RECOVERY, ERROR METHOD, UNCERTAINTY ASSESSMENT, PIRSON’S FORMULA, MEDICAL AGENT.

The object of the study is to study uncertain data and compare different algorithms for finding gaps.

In this study, the tool for analyzing the degrees of confidence will be probability theory, which is associated with the Bayesian representation of data and knowledge about the multifactorial state of the problem.

In this research, a study was conducted on the restoration of gaps in incomplete data. Mechanisms for generating passes and some popular methods of dealing with incomplete data will also be considered.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ISO – Infrared Space Observatory

GFR – Glomerular Filtration Rate

КТ – комп’ютерна томографія

TE – total error

BIPM – Bureau International of Weights and Measures

ВСТУП

У цьому досліджені розглядається простий приклад з області медичної діагностики [1]. Діагностика проводиться під час обстеження пацієнта та являє собою задачу, в якій майже завжди доводиться стикатися з невизначеністю.

Діагностична невизначеність пов’язана з виявленням анамнезу.

По-перше, відбувається фізичний огляд пацієнта, виконуються клінічні дослідження лабораторними методами, у разі необхідності аналізуються результати візуалізації окремих елементів організму пацієнта або цілком його організм та інше [2], зазвичай це все зменшує діагностичну невизначеність. Але треба врахувати ще логіку аналізу даних і знань.

Спроба використовувати логіку першого порядку для подання знань в таких проблемних областях, як медична діагностика [3], закінчується невдачею з трьох основних причин:

– економія зусиль. Для формування повної множини антецедентів або консеквента, необхідного для складання правил, які не мають винятків, потрібно занадто багато роботи і часу, крім того застосування таких правил є сама по собі занадто складною справою;

– відсутність теоретичних знань. Медична наука не має під час діагностування повної теоретичної обґрунтованості для даної проблемної ситуації, яку отримали медичні фахівці під час навчання. Необхідно постійно підвищувати фаховий рівень, мати інтерес до нових технології і результатів досліджень. Тут може використовуватись різноманітні сучасні можливості, зокрема відомості, які в великому обсягу знаходяться в Інтернет;

– відсутність практичних знань. Навіть якщо відомі всі правила логічного виводу, може залишатися невизначеність щодо діагнозу даного конкретного пацієнта, оскільки всі необхідні обстеження були або не могли бути виконані, або мали мале практичне підтвердження в особистій лікувальній практиці лікаря, або не вистачило часу на отримання

достовірного діагнозу чи коштів у пацієнта та інше. Для скорочення часу на діагностування і підвищення якості діагнозу.

Отже при аналізі проблем діагностичною невизначеності необхідно враховувати таку невизначеність, яка пов’язана з вимірюваннями параметрів стосовно стану організму, які поповнюють бази даних (БД) так і баз знань (БЗ) на всіх етапах діагностування [3,4].

Відзначимо, що знання дозволяють саме краще та точніше сформувати висловлювання, які стосуються медичної проблеми діагностування тільки з певним ступенем впевненості (degree of belief).

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

Імовірності надають спосіб оцінки сумарного обліку невизначеності, що виникає з причин економії зусиль і відсутності знань. Не можна знати з усією впевненістю, що турбує даного конкретного пацієнта, які у нього ознаки і симптоми хвороби найбільш значимі для виявлення правильного діагнозу. Ступінь істинності, на відміну від ступенів впевненості, є предметом нечіткої логіки [3, 5].

Проблема оцінки і врахування невизначеності медичних даних,знань щодо медичних аналізів і різноманітних досліджень має дуже важливе значення. Системний аналіз і прийняття рішень при діагностуванні стану організму людини є актуальною та досить складною проблемою [1-15].

Крім зазначеного, зауважимо, що для вирішення поставленої проблеми будимо використовувати мультіагентне представлення медичної системи, яка буде використана для діагностування та лікування [16,17].

1.1 Визначення похибки та невизначеності вимірювань

Сучасні підходи до оцінки помилок при вимірюванні медичних даних, швидше за все, не скоро у повній мірі будуть використані з урахуванням невизначеності, оскільки останні підходи перебувають на ранніх стадіях розвитку їх впровадження. Але підходи щодо невизначеності, швидше за все, отримають все більший розвиток, оскільки фокус переходить від властивості вимірювальної системи до належного використання результату вимірювання при діагностиці та моніторингу ефектів лікування, включаючи всі фактори, що спричиняють невизначеність.

Похибка вимірювання − це відхилення вимірного значення величин від її «істинного» значення. За своєю природою чи характером прояви погрішності можуть бути «випадковими» та «систематичними».

Невизначеність вимірювань − це «отриманих результатів, які мають збіги в істинності». Тобто параметр, пов'язаний з результатами вимірювань, що характеризує розбіжність значень, які могли бути обґрунтовано приєднаними до змінної величини.

Поняття «невизначеність вимірювань» з'явилося понад 50 років тому і пов'язане воно з точністю результатів вимірювань. Необхідність розробки нової концепції оцінки точності результатів вимірювань була викликана відсутністю міжнародної єдності в цих питаннях.

Концепція невизначеності стала результатом розвитку теоретичної метрології та в даний час найбільш повно відповідає сучасним вимогам технічного прогресу і є критерієм оцінки точності на міжнародному рівні.

Історія виникнення терміна «невизначеність вимірювань» пов'язана з заміною основних понять, що, по суті, є в таких обох термінах як «погрішність» і «невизначеність». Це вираз в різних термінах, одночасно поняття − «точності вимірювання».

У світі історично склалось так, що при оцінювані відповідності вимірів, які були пов'язані між собою використовували поняття погрішність чи похибка.

За кордоном застосовувалось поняття − «помилка вимірювання» чи − «помилка вимірювання». Одною з цілей при розробці стандартів якості ISO 9000 було забезпечено безпечне виконання всіх виробничих функцій. У рамках ISO 9000 було розроблено «Керівництво по обчисленню невизначеності у вимірах» − «Посібник із визначення невизначеності у вимірах», у якому описано поняття невизначеності вимірювань та способів її обчислення.

Поки де потрібно вимагати оцінки точності проведення змін у поняттях (наприклад, таке вимагання, що передбачається при акредитації лабораторій) у термінах «невизначеності».

У зв'язку зі вступом країн в ВТО, були прийняти рішення перевести правила проведення та оцінки якості вимірюваних робіт у відповідність до

міжнародних стандартів ІСО [17-19]. Зараз усі вимірювальні лабораторії іноземних членів ВТО повинні оцінювати точність результатів вимірювань у термінах невизначеності.

Поняття «невизначеності» виникло з дослівного переводу з «Посібника для визначення невизначеності у вимірах», ISO-1993. Документ звільнив безліч спорів і розділив громадськість на три табори: прихильники «Керівництва…», противники «Керівництва…» та спеціалісти-практики, що чекають – «чим все це закінчиться».

У підсумку, «все скінчилось тем», що виник документ РМГ 91-2009 «Спільне використання понять «похибка вимірювань» і «невизначеність вимірювання», який дав пояснення щодо використання відповідних термінів [20-24]. Терміни використовуються при розрахунку невизначеності (співвідношення термінів теорії невизначеності з термінами класичної теорії точності (в дужках)):

– невизначеність результату вимірювання (похибка результату вимірювання);

– невизначеність типу А (випадкова похибка) ;

– невизначеність типу Б (систематична похибка) ;

– стандартна невизначеність (стандартне відхилення похибки) результату вимірювання;

– розширена невизначеність (довірчі кордону) результату вимірювання;

– можливість охоплення, ймовірність покриття (довірча ймовірність);

– коефіцієнт охоплення, коефіцієнт покриття (коефіцієнт розподілу похибки).

Як вже згадувалося вище, термін «похибка» прив’язаний до істинного значення вимірюваної величини. Однак, при цьому вихідне «справжнє значення» невідомо. І при проведенні вимірювань вказують інтервал, в якому це «справжнє значення» знаходиться з певним рівнем імовірності:

$$X = A \pm \Delta, P = 0{,}95\ ,$$

де P - довірча ймовірність.

Інтервал від $A - \Delta$ до $A + \Delta$ з імовірністю P містить в собі (рис. 1.1) «справжнє» значення вимірюваної величини:

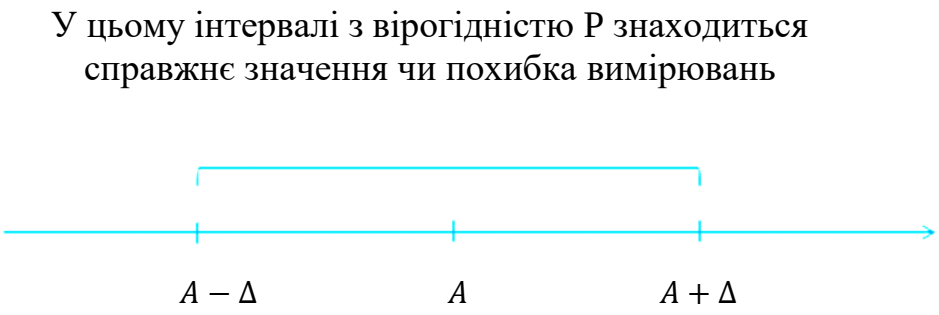

Рисунок 1.1– Діапазон можливих значень щодо похибки

Обчислення невизначеності можливо виконати за наступними кроками:

Крок 1. Обчислюємо середнє арифметичне значення освітленості з усіх вимірів в даній точці:

$$E = \frac{1}{n}\sum_{i=1}^{n} E_I\ ,$$

де $E_i, i = 1 \ldots n$ – виміри в даній точці.

Крок 2. Для джерел невизначеності випадкового характеру обчислюємо невизначеність за типом А:

$$u_A(E) = \sqrt{\frac{\sum_{i=1}^{n}(E_i - E)^2}{n(n-1)}}\ .$$

Крок 3. Для джерел невизначеності систематичного характеру (приладова похибка) обчислюємо невизначеність за типом Б:

$$u_{\text{Б}}(E) = \sqrt{\frac{\Delta E}{n\sqrt{3}}}\,.$$

Крок 4. Обчислюємо сумарну стандартну невизначеність:

$$u_C(E) = \sqrt{u_A^2(E) + u_{\text{Б}}^2(E)}\,.$$

Крок 5. Для довірчої ймовірності (ймовірності охоплення) P = 0,95 (рекомендується у Керівництві щодо розрахунку невизначеності) задаємо коефіцієнт охоплення k = 2 і обчислюємо розширену невизначеність вимірювань:

$$u \;\; = k u_c\,.$$

Зараз у Бюро International des Poids et Mesures (BIPM) існує 5 теорій вимірювання:

– математичні моделі вимірювання розглядають вимірювання як відображення якісних емпіричних відношень до відношень між числами;

– операціоналістські моделі розглядають вимірювання як сукупність операцій, що формують значення та / або регламентують використання кількісних термінів;

– реалістичні моделі розглядають вимірювання як оцінку незалежних від розуму властивостей та / або відношень;

– інформаційні моделі розглядають вимірювання як збір та інтерпретацію інформації про систему;

– теорія моделей у теоретичній / математичній / статистичній моделі розглядає вимірювання як когерентне присвоєння значень параметрам процесу.

Розглядаються реалістичні моделі, представлені в лабораторній медицині методами помилок вимірювання як оцінка міри та вимірювальної системи як «незалежних від розуму властивостей».

Теорії моделей, представлені методами невизначеності вимірювань, стверджують, що на додаток до, є інша відповідна та доступна інформація. Самі результати вимірювань повинні враховуватися як допомога у правильній інтерпретації результатів вимірювань. Методи помилок в даний час домінують у лабораторній медицині незважаючи на впровадження методів невизначеності вимірювань на початку 2000-х років.

Методи помилок фокусуються на практичному процесі вимірювання та його результатах. Помилка властивісті одиничного результату вимірювання розглядається відповідно справжнього значення. Упередженість оцінюється як різниця між середнім значенням повторних результатів вимірювань та дійсним значення. Повторюваність, проміжна точність та неточність відтворюваності оцінюється як міра випадкової помилки. Поєднання упередженості та неточності (точності), виражене як загальна похибка (TE), набуло популярності, оскільки це може бути рентабельно оцінюється одиночними вимірами контрольних зразків.

Методи помилок, що застосовуються в лабораторній медицині, зосереджуються на властивостях вимірювальних систем, а також розробляються для інших результатів вимірювань про стан здоров’я і догляду.

Методи невизначеності також засновані на результатах вимірювань, але їх основна увага приділяється їх використанню для діагностики та контролю результатів лікування. Всі фактори, що впливають на торгівлю та лабораторну медицину, започатковані найвищими міжнародними органами влади метрології (BIPM) становлять важливий розвиток, який все ще триває. Наприклад, у послідовних версіях GUM та VIM. Зміна парадигми в хімії та інших науках про вимірювання також спостерігається.

VIM визначає ключові поняття та терміни в метрології. Оригінальна версія GUM у від 1993/1995 рр. досі діє, але була розширена додатками. Тип помилки ТЕ (загальна помилка) у термінології VIM є комбінацією випадкових та систематичних помилок, виражені за номінальною шкалою. На відміну від ТЕ, це абсолютне значення виміряного зміщення плюс стандартні відхилення, виміряні на шкалі співвідношень у термінології. Точність − це також поєднання упередженості та випадкової помилки, але вимірюється за порядковою шкалою згідно з VIM. Таким чином, ця система вимірювань є більш-менш точною, ніж інша система вимірювань, але точність не вказує, наскільки маємо більш-менш точні значення.

У медичній лабораторії невизначеність вимірювань відома лише опосередковано для даних, які отримані в результаті повторного вимірювання стабілізованих контрольних зразків або для зразків пацієнта. Тому важливо проводити повторні вимірювання контролів для оцінки невизначеності на аналітичній фазі.

GUM стверджує, що невизначеність вимірювання «відображає відсутність точних знань про значення вимірюваної величини». Розподіл результатів, що описують невизначеність вимірювань, має передавати силу обґрунтованої віри в те, де справжнє значення вимірюваної величини полягає. Інтервал повинен містити справжнє значення виміряна величина із заданою імовірністю.

Останні події у філософії вимірювання підкреслюють взаємозв’язок між вимірюванням та теоретичними моделями, на яких базується використання результатів вимірювань.

Версія GUM у 1993/1995 рр. використовувала як загальні підходи до невизначеності, так і помилки, використовуючи відповідно байєсову та частотну статистику [21-24]. Отже розглядається стан знань або ступінь переконань, на відміну від частотних методів, які розглядають імовірність як частоту виникнення. Баєсові методи призначають імовірності гіпотезам, тоді як частотистські методи перевіряють гіпотези, не призначаючи їх попередні

імовірності. Останні доповнення до GUM та остання версія VIM (VIM3), яка була опублікована у 2008 році, повністю перейшла до врахування спільно невизначеності і баєсової статистики.

1.2 Методи оцінки невизначеності

Оцінка невизначеності враховує методи хіміхної метрології стосовно фізичної. Отже, на відміну від методів невизначеності, методи помилок вимагають, щоб справжнє значення змінної було відомим, оскільки похибка є властивістю одного виміру. Методи оцінки щодо невизначеністі не стверджують відсутність справжнього значення, але стверджують відсутність точного значення знання про справжню цінність. Концептуальні відмінності між діагностичною невизначеністю та підходами до похибки вимірювання при оцінці (A) діагностичної невизначеності, (B) упередженість та (C) ТЕ в лабораторній медицині. Підходи до діагностичної невизначеності (A) забезпечують оцінку діагностичної невизначеності; врахуємо сукупну невизначеність (рис. 1.2).

Однією з найбільш відомих і широко використовуваних в даний час є шкала Медичної дослідницької ради (Medical Research Council Scale, MRC). Ця 5-бальна шкала була вперше опублікована в 1952 р. і, в подальшому були деякі зміни. Вона проста у використанні, тому що всього 5 питань позволяють визначити, в якій мірі задишка обмежує активність пацієнта. Вища оцінка за шкалою MRC (4 бали) відповідає максимально вираженій задишці.

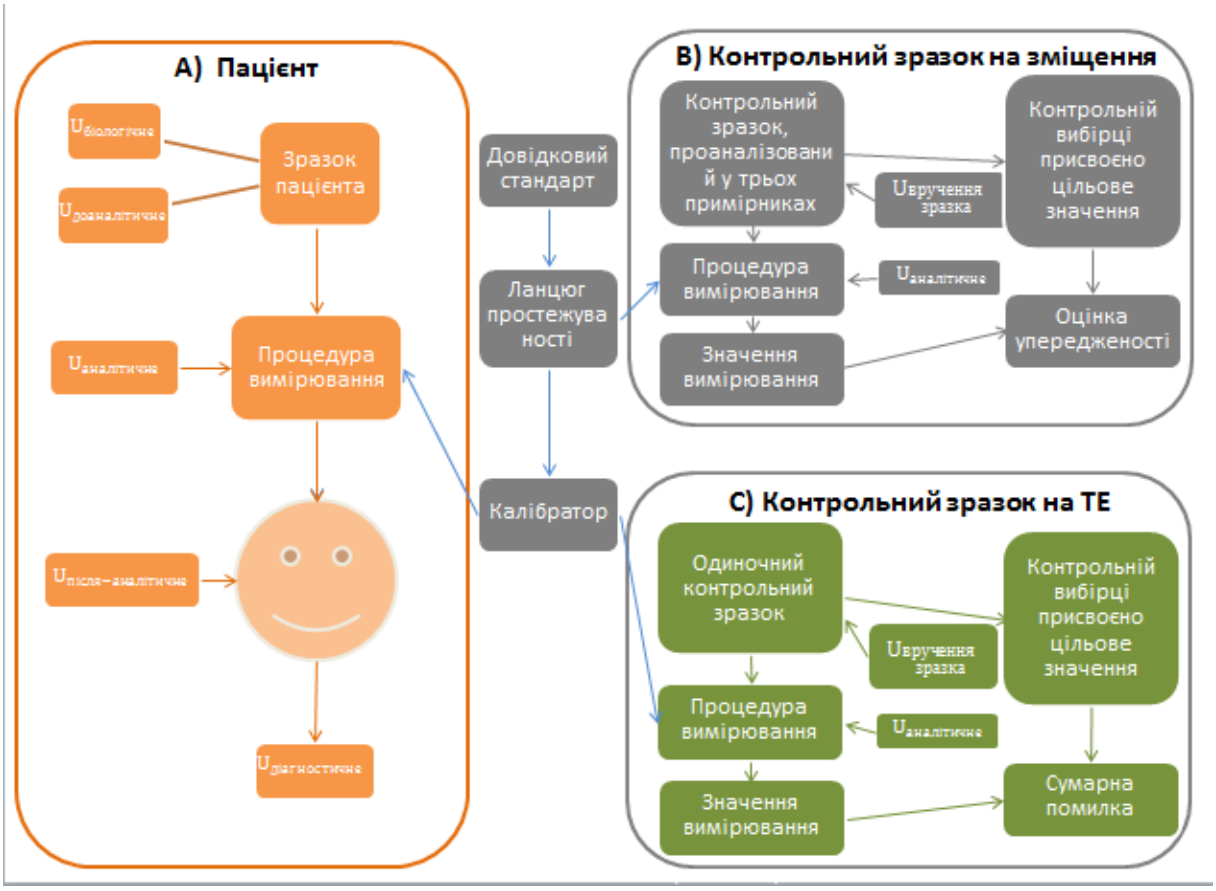

Рисунок 1.2 − Концептуальні відмінності між діагностичною невизначеністю та підходами до похибки вимірювання

Загальний ланцюжок випробувань у лабораторній медицині включає кілька можливих джерел невизначеності від клінічного рішення про замовлення тесту через біологічні варіації, преаналітичну, аналітичну та постаналітичну фази до значення результату тесту у поточні клінічні рішення (рис. 1.3).

Метод помилок зосереджуються насамперед на аналітичній фазі. А методи невизначеності фокусуються на підрахунку всіх джерел невизначеностей, включаючи біологічні зміни, доаналітичні варіації, аналітичну варіацію та постаналітичну варіацію як допоміжний засіб у діагностиці лікування.

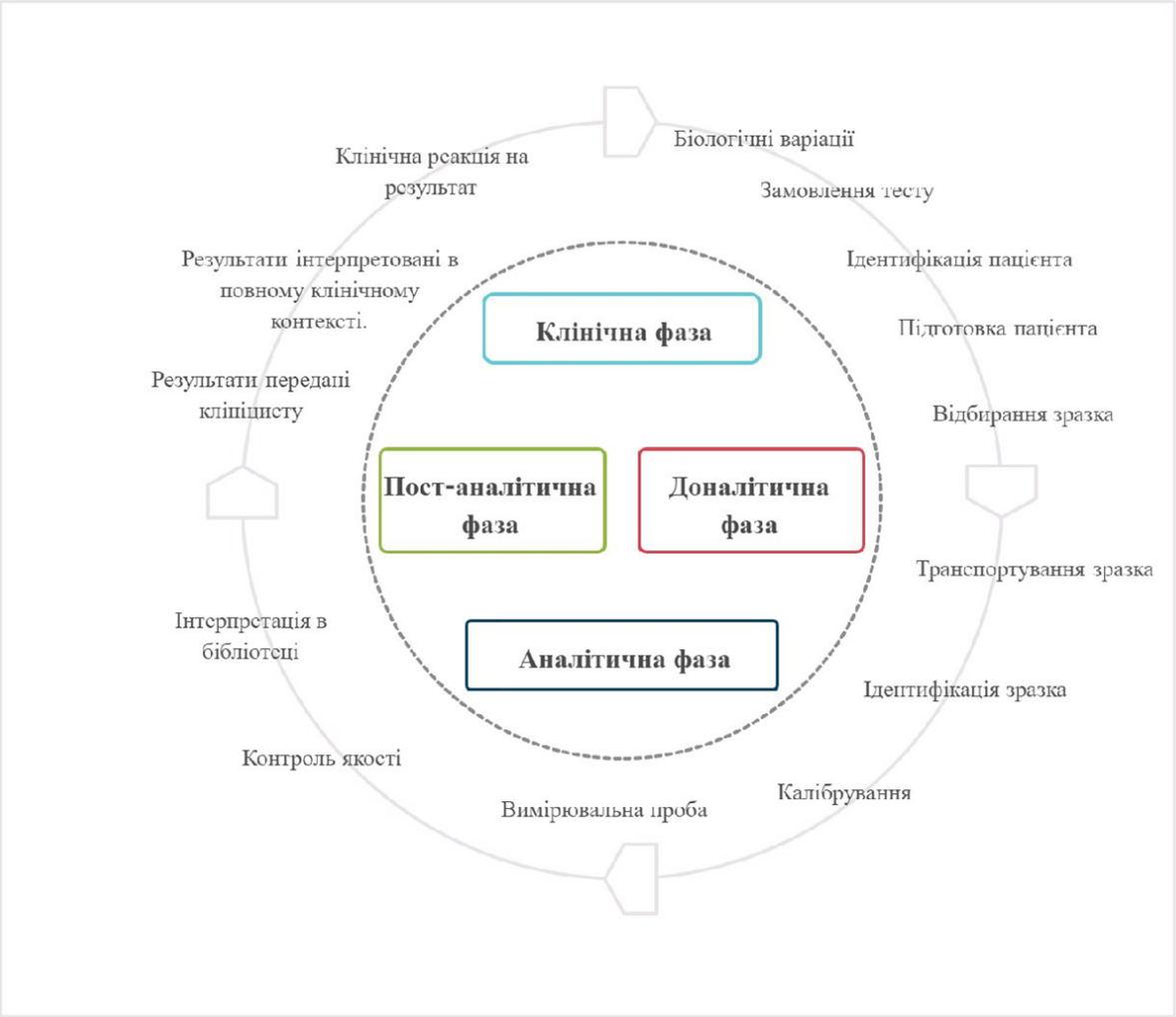

Рисунок 1.3 – Загальний ланцюжок випробувань у лабораторній медицині

Ефекти усіх причин невизначеності у загальному ланцюжку випробувань при використанні лабораторних результатів для діагностики пацієнтів, включаючи біологічні варіації та аналітичні, аналітичні та постаналітичні варіації. Як LPU, наданий GUMом, так і просте додавання дисперсій відповідно до теорему Піфагора можна використовувати для оцінки діагностичної невизначеності.

Методи LPU мають свою основну силу в здатності боротися з численними та складними причинами невизначеності, але їх основною слабкістю є їх теоретична та практична складність та відсутність практичні впровадження в лабораторній медицині.

Підходи до похибки вимірювання (B, C) надати оцінки невизначеності методів вимірювання; аналітична фаза всього ланцюга випробувань. Якщо основною метою програми зовнішнього контролю якості є те, щоб окремо визначити упередженість та неточність, підхід (B) є кращим. Зовнішня якість контрольні програми, спрямовані на оцінку ТЕ, використовують одномісні вимірювання (C). Їх підтримують вичерпні теоретичні моделі, прийняті контролюючими органами та які мають широке практичне використання в лабораторній медицині. Біологічні, доаналітичні та постаналітичні варіації не є актуальними щодо моніторингу змінних у контрольних зразках, але ці варіації актуальні при роботі із зразками пацієнтів.

Термінологія, що використовується у VIM3 для опису компонентів похибки та невизначеності вимірювання відображена нижче (рис. 1.4).

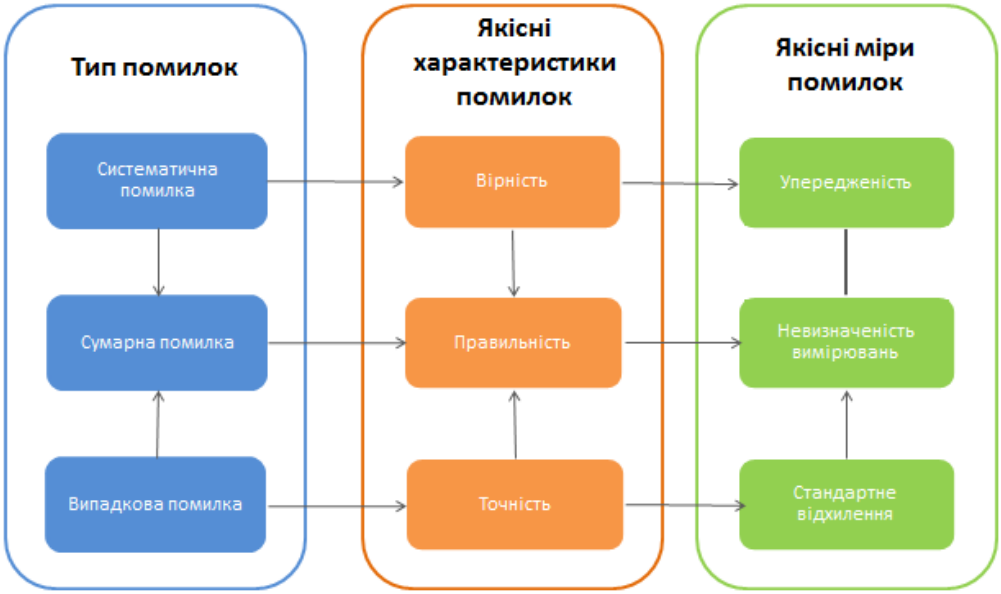

Рисунок 1.4 – Термінологія VIM3

Похибка вимірювання або просто похибка, є властивістю одного вимірювання – значення кількості мінус еталонне значення величини.

Значення еталонної величини служить для змінної а як сурогатне справжнє значення «в системі (або моделі)». Справжнє значення або контрольне значення вимірюваної величини в конкретному зразку пацієнта

не може бути відомим. Отже, справжнє значення результату одного пацієнта не може бути відомим, і, тому немає впевненості у результаті, який повиннен виражатися в імовірнісні термінів на основі частотних статистичних даних (довірчих інтервалів) для моделей помилок та баєсової статистики (щільності імовірності) для моделей невизначеності.

Серед міфів про методи вимірювань невизначеності є те, що вони вимагають щоб упередження було усунуте до того, як можуть бути зроблені розрахунки для оцінки невизначеності. Підходи до обчислення невизначеності стверджують, що упередження слід усунути при їх виявленні. Однак, якщо упередженості неможливо усунути, або коли усунення упередженості має збільшити ризик, то загальна невизначеність вимірювання може бути оброблена як будь-який інший тип В невизначеності.

Підходи до вимірювань невизначеності у лабораторній медицині стикаються з декількома викликами:

– рівень знань з математики та передової статистики в лабораторній медицині, як правило, занадто низький, щоб повністю зрозуміти і застосувати закон поширення невизначеності (LPU), включаючи рівнянні вимірювання, матриці коваріації, часткові похідної, розкладання Тейлора та байєсівські статистичні дані, які необхідні для повної реалізації підходів невизначеності;

– методи помилок / частоти використовуються у всіх лабораторіях і, як правило, добре зрозуміли.

– справжнє значення або контрольне значення вимірюваної величини в конкретному зразку пацієнта не може бути відомим. Отже, справжнє значення результату одного пацієнта не може бути відомим, і, отже, впевненість у результаті повинна виражатися в імовірнісні терміни на основі частотних статистичних даних (довірчих інтервалів) для моделей помилок та баєсова статистика (щільність ймовірності) для моделей невизначеності;

– рівень знань щодо біологічних, доаналітичних та постаналітичних змін у лабораторній медицині все ще значно відстає від знань про причини

аналітичних варіацій, і це зменшує надію на те, що правильне використання невизначеності методи покращать клінічне використання методів вимірювання.

Незважаючи на перешкоди, методи невизначеності були належним чином застосовані в лабораторній медицині.

1.3 Методи розрахунку помилок

Хоча методи помилок добре розроблені і все ще домінують у забезпеченні якості вимірювальні системи в лабораторній медицині, перехід до використання вимірювань та методи невизначеності вже мали місце в інших галузях метрології.

Відмінності методів між розрахунками помилок та невизначеністю потрібно розуміти, але не слід надмірно їх підкреслювати.

Для одного результату вимірювання неможливо дізнатися окремий внесок упередженості та неточності в ТЕ цього результату. Методи ТЕ надають перевагу і є особливо доречно, коли для контролю якості використовуються одиночні зразки. Коли ТЕ використовуються методи, 21 зміщення та неточність (помножені на *z*-коефіцієнт) додаються лінійно, що призводить до значення для ТЕ: ТЕ 5 зміщення 1 zCVa. ТЕ використовується для оцінки меж інтервал навколо дійсного значення, де можна знайти виміряні результати аналітики за допомогою а певна ймовірність, зазвичай 95% ймовірність.

Серед основних проблем, пов’язаних з підходом до помилок, є:

– випадок, коли справжнє значення неможливо дізнатись, тобто ТЕ не можна оцінити;

– різні вирази ТЕ або гранично допустиме відхилення мають спільне злиття величин, які за своєю суттю несумісні; упередження має а знак,

позитивний чи негативний, тоді як стандартне відхилення представляє інтервал (2 стандартних відхилення) величин;

– випадок, коли існує кілька варіантів підрахунку ТЕ.

Можна сподіватися, що дебати між прихильниками невизначеності та помилок підходять в лабораторній медицині сприятиме зростанню розуміння та розвитку в рамках обох підходів у лабораторній медицині.

1.4 Постановка задачі дослідження

Метою дослідження є реалізація програмного продукту, який застосовує методи відновлення пропущених даних. Також зробити порівняльний аналіз ефективності застосування багаторазової вставки за правилом Рубіна і одноразових методів заповнення пропусків

Об’єктом дослідження є медичні дані з пропусками. Необхідно оцінити ефективність різних підходів відновлення даних.

Необхідно розробити адаптивний метод проектування спеціалізованого медичного комплексу для діагностування захворювань з мультіагентним представленням компонентів комплексу (ММКД) [1,3,17,18].

У дослідженні розглядаються можливості оцінити та врахувати невизначеність даних на прикладі розповсюдженого засобу контролю температури тіла, при використанні комп’ютерної томографії і вибору методу аналізу даних і знань під час пандемії COVID-19 [16].

Агентні компоненти структури ММКД визначаються на основі багатопрофільних досліджень стану організму пацієнта (опитування пацієнта, клінічних аналізів, знімків після призначених лікарем). Простір агентних структур, які розташовані у базі прецедентів, відповідають раціональним аналогічним рішенням для прийняття медичних рішень відносно діагностуванню чи стратегії лікування.

Для реалізації програмних агентів необхідно відзначити такі поняття, як адаптація та навчання. Адаптація − це процес зміни параметрів і структури у системі ММКД, а можливо, і керуючих рішень щодо впливів на діагностичний процес стосовно поточної інформації з метою досягнення певного стану системи при початковій невизначеності і умовах отримання медичних даних і знань. Навчання − це процес, в результаті якого система ММКД поступово набуває здатність відповідати потрібним реакціям на певні впливи.

2 ПОШУК НЕВИЗНАЧЕНОСТІ ТА ВІДНОВЛЕННЯ ДАНИХ

В медичній сфері поняття невизначеності тісно пов’язане з проблемою повноти знання про середовище функціонування організму людини, зокрема обмеженістю інформації (як одним з видів ресурсів), яка необхідна для прийняття діагностичних рішень [19-21]. У даному випадку обмеженість інформаційного ресурсу викликана невідповідністю його основним якісним характеристикам, зокрема таких:

– повнота – відображення всіх медичних процесів та явищ, які є суттєвими та важливими для прийняття раціонального діагностичного рішення;

– зрозумілість – доступність її усвідомлення тим користувачам (лікарям), для якого вона призначена;

– достовірність – об’єктивність та відповідність реальному стану частин та цілком організму людини;

– актуальність – відповідність поточному моменту часу.

Виходячи з цього, невизначеністю є недостатність інформації про умови, в яких буде здійснюватися діагностування чи моніторинг стані пацієнта під час лікування, що обумовлює складність визначення кінцевих результатів діагностування на всіх рівнях медичної діяльності.

2.1 Просте додавання варіацій за допомогою теореми Піфагора

Методи помилок використовують теорему Піфагора для обчислення повної дисперсії як корінь квадратний суми дисперсій для всіх складових дисперсії, що складають загальну невизначеність (рис. 2.1). Теорема Піфагора у тригонометрії стосується правила трикутника. Таким чином можна додавати лише незалежні випадкові величини. Геометрично випадкові величини представлені у вигляді векторів, довжини яких відповідають з їх

стандартними відхиленнями. Коли змінні незалежні, їх вектори є ортогональними [22, 23]. У цьому випадку стандартне відхилення суми або різниці змінних – це лише гіпотенуза прямокутного трикутника. Теорема Піфагора також не є доцільним за наявності значної упередженості / систематичної помилки в будь-якій з дисперсій компоненти.

В рузультаті є компоненти діагностичної варіації в LPU. Теорема Піфагора стверджує, що квадрат гіпотенузи дорівнює сумі квадратів інших 2 сторін (рис. 2.1). Це означає що сума площі суміжного до гіпотенузи квадрата дорівнює сумі площ суміжних з катетами квадратів.

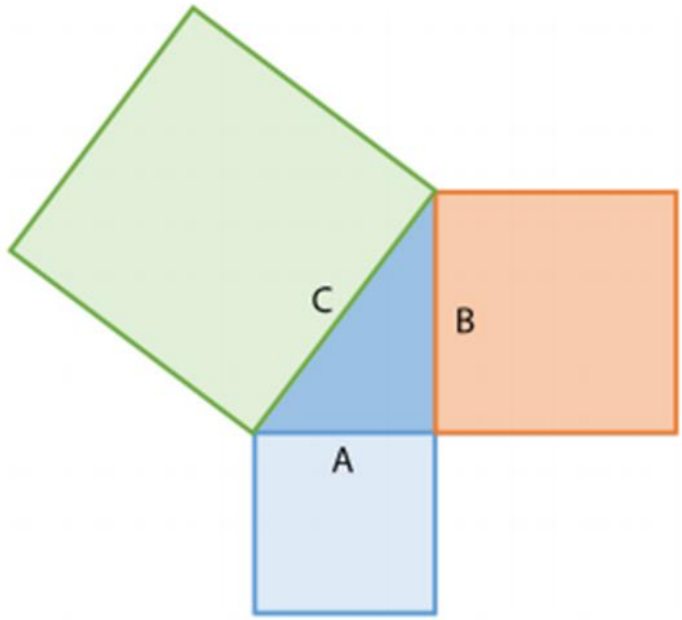

Рисунок 2.1 − Теорема Піфагора

Стандарти ISO 17025 та ISO 15189 вимагають від лабораторій оцінки невизначеності вимірювань. Однак методи розрахунку похибки вимірювання не є відповідно до стандартів, тому органи акредитації приймають як підходи до похибки, так і невизначеності [24, 25]. Методи помилок зверху вниз для розрахунку похибки вимірювань із зразків внутрішнього контролю якості є, мабуть, найбільш часто використовуваними підходами для розрахунку невизначеності вимірювань в акредитованих ISO медичних лабораторіях.

2.2 Використання закону поширення невизначеності

Спочатку GUM рекомендував переважно LPU, тоді як додаток GUM1 / JCGM 101 використовує метод Монте-Карло. Обидва методи засновані на прямому оцінюванні невизначеності, яке створює математичну модель введення-виведення (вимірювання рівняння), опис варіації виходу, спричиненої варіаціями вхідних даних.

Невизначеність вимірюваної величини, якщо це можливо, виражається як функція набору впливання на величини, включаючи їх коваріації. LPU може враховувати неортогональність та упередженість за допомогою моделей вимірювання, матриці коваріації, часткові похідні та розкладання і тому теоретично переважно при розрахунку похибки вимірювань. Перевага LPU полягає в тому, що всі відповідні компоненти невизначеності враховуються як коли їх можна оцінити безпосередньо статистичними методами (тип A) і коли оцінюється іншими способами (тип B), включаючи освічену оцінку на основі досвід.

Існує також рух щодо застосування байєсівської статистики для оцінки невизначеності вимірювань, в якому стан розподілу знань про кількість, що цікавить, випливає з попередньої інформації про кількість та інших впливати на величини; та виміряні дані, використовуючи імовірнісну інверсію або обернену оцінку невизначеності.

2.3 Використання методу Монте-Карло

Моделювання Монте-Карло легше зрозуміти, ніж традиційні методи LPU (рис. 2.2). Це вимагає інформації про розподіл ймовірностей усіх факторів, що впливають на невизначеність.

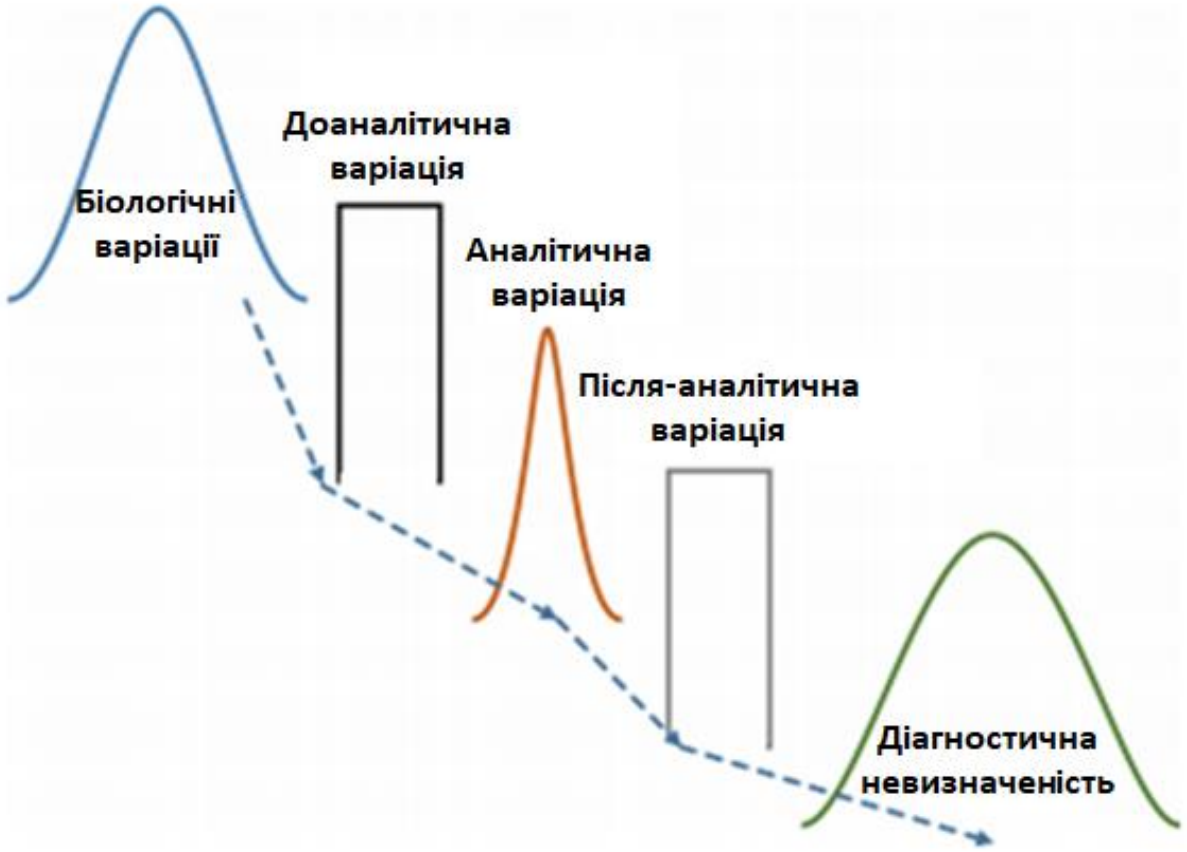

Рисунок 2.2 − Спрощення принципів методів Монте-Карло для оцінки діагностики

Дисперсія в дослідженні, що використовується для оцінки біологічних змін, звикла у моделі, що використовується у щодо внесоку біологічної варіаціїт та оцінці діагностичної невизначеності.

Преаналітична варіація оцінюється на основі квадратичної функції імовірності, властивості якої оцінювання як невизначеність типу B. Аналітична варіація оцінюється як дисперсія повторюваності; всі результати вимірювань в лабораторній організації при вимірюванні однакових зразків для певної вимірюваної величини з використанням усіх систем вимірювання в різних точках вчасно.

Постана алітична варіація оцінюється на основі квадратичної функції імовірності, тобто властивості оцінюються як невизначеністі типу B. Для створення діагностичної невизначеності одночасно з усіх функцій

імовірності беруть щонайменше 100 000 повторних зразків для оцінки розподілу імовірностей.

Змінна первинного відсотка; у цьому випадку діагностичної невизначеністі вимірюваної величини. Діагностична невизначеність включає біологічну варіацію, доаналітичної варіації, аналітичну варіацію та поаналітичної варіації. Біологічна варіація та аналітична варіація можуть бути виражені як коефіцієнти варіації гауссового розподілу, а доаналітичні та постаналітичні розподіли ймовірності можуть бути виражені як прямокутні розподілу (або як помилку, або як її не має).

Існує кілька програмних засобів для застосування методів Монте-Карло, включаючи додаткові модулі для Microsoft Excel. Серед інших переваг методів Монте-Карло є в тому що:

– для оцінки діагностики не потрібна жодна математична функція (вихідна функція) щодо невизначеністі;

– на додаток до припущення не потрібні ніякі припущення щодо вхідних величин які відповідають гауссовому розподілу;

– немає необхідності обчислювати часткові похідні;

– на них не впливають часткові похідні, які зникають при оцінці вхідних даних кількостьна оцінка методів передискретизації мала широку доступність до недорогих обчислювальних потужностей у 1980-х рр. Суттєво покращилася варіанти та вартість роботи з даними, незалежно від теоретичного розподілу А.

Постаналітична варіація оцінюється на основі квадратичної функції імовірності, яка має властивості, що оцінюються як невизначеність типу В. Для створення діагностичної невизначеності одночасно для усіх функцій імовірності беруть щонайменше 100 000 повторних зразків, які мають відповідні розподіли імовірностей.

Невизначеність вимірювання має повну помилку щодо порівнянно до методів передискретизації / завантаження, які методи для основного потіку аналізу даних. Передискретизація щодо заміни означає, що мається поряд із

100 000 до 1 мільйона зразків із заміною, які відбираються стосовно вихідної вибірки, а статистичні дані, що цікавлять досліджувача щодо обчислювання з цієї оцінкою псевдонаселення як оцінка відповідного відповідного параметра.

Методи передискретизації не передбачають припущення для спостереження, яке розподіляються згідно певного теоретичного розподілу, але необхідно припустити що важливо для основного розподілу населення практично такий самий, як і в конкретній вибірці для аналіза населення. Це припущення означає, що достатня кількість спостереження потрібна у вибірці, щоб переконатися, що вона представляє необхідну сукупність.

Приблизно від 100 000 до 1 мільйона вибірок із розподілу, що впливає переважно, коли виміряне значення супроводжується його похибкою вимірювання, це поперше стає результатом вимірювання, яке виражається інтервалом охоплення. У цьому типу інтервалу, слід взяти до уваги 3 цифри: виміряне значення, нижню і верхню межі цього інтервалу. Цей інтервал охоплення зазвичай отримують після оцінювання.

Спрощення принципів передискретизації з використанням методів заміни для оцінки діагностичної невизначеності ілюструється як гауссовий розподіл (рис. 2.3).

Усі вихідні дані дослідження, що використовуються для оцінки біологічних змін, використовуються для повторного відбору проб (що тут і проілюстровано як томбола). Преаналітична варіація, що оцінюється за допомогою квадратної функції імовірності та властивості, оцінюються як невизначеність типу B. Аналітична варіація оцінюється як варіація відтворюваності; всі вимірювання мають результати в лабораторній організації при вимірюванні однієї і тієї ж проби для певного вимірювання величини, які використовують всі вимірювальні системи у різні аналізуючи моменти.

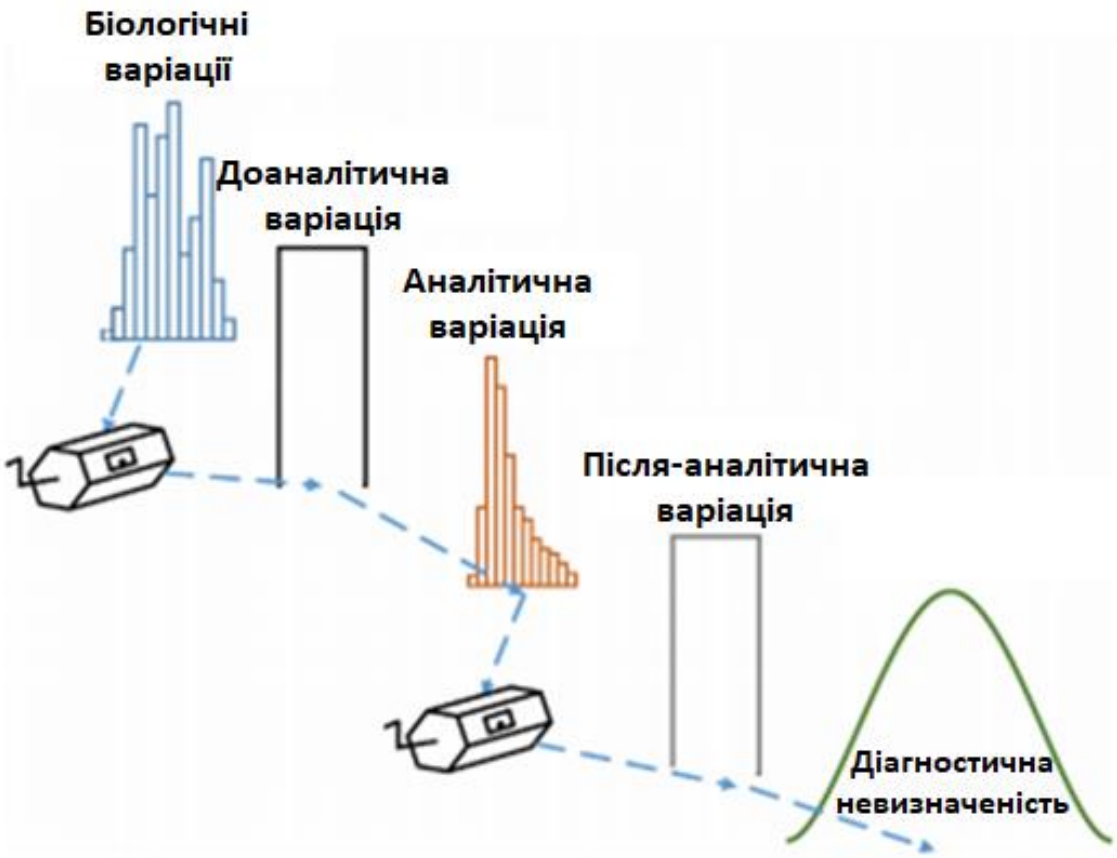

Рисунок 2.3 – Спрощення принципів передискретизації з використанням методів заміни для оцінки діагностичної невизначеності

Всі дані використовуються для передискретизації (тут ілюструється як томбола) мають внесок аналітичної варіації до оцінки діагностичної невизначеності. Постаналітична варіація оцінюється для змінної як квадратична функція імовірності [25-27]. Властивості цієї функції оцінюються як невизначеність типу В.

На рисунку 2.3 продемонстровано збільшення принципів передискретизації з використанням методів заміни діагностичної невизначеності для оцінки (тут ілюструється як гаусовий розподіл та приклад будь-якого можливого розподілу). Усі вихідні дані дослідження, що використовуються для оцінки біологічних змінних, використовуються повторний відбор проб (тут проілюстровано як томбола), біологічне відхилення до оцінки діагностичної невизначеності. Преаналітична варіація оцінюється за допомогою квадратної функції імовірності, властивості, що

має оцінку як невизначеність типу В. Аналітична варіація оцінюється як варіація зміни та всі вимірювання як результати отримані у лабораторній організації при вимірюванні однієї і тойже пробі для відповідного вимірювання величини, що використовують всі вимірювальні системи в різни моменти часу. Всі згадані дані використовуються для передискретизації (тут ілюструється як томбола) як внесок аналітичної варіації для оцінки діагностичної невизначеності. Посталітична варіація оцінюється для змінної як квадратична функція імовірності, властивості якої оцінюються як невідомість типу В.

Використання свіжих різноманітних зразків для пацієнтів щодо контролю якості має сенс стосовно кількох причин:

– матеріал має оптимальні матричні властивості (є змінним);

– матеріал доступний безкоштовно для всіх лабораторій, що приймають рутинні зразки пацієнтів;

– існує загальна домовленість, щодо вимірювальних систем та реагентів що повинні привести оптимально до однакових результатів при аналізі одних і тих же зразків від пацієнтів;

– методи є оптимальними для ідентифікації вимірювальних систем в данній організації, яка вносять найбільшу частину загальної невизначеності вимірювань, спричиненої упередженість.

Методи розділених зразків трудомісткі за відсутності ефективних комп’ютеризованих систем, але зручні при правильній реалізації. Більшість лабораторних організацій, які впроваджуванні для використання методів розділених зразків і беруть активну участь у продовжені своєї участі у роботі систем зовнішнього контролю якості з метою порівняння їх результатів на національному та транснаціональному рівнях.

Спрощення традиційної зовнішньої схеми контролю якості (ECQ / перевірка кваліфікації)(А) та схема контролю (В) орієнтована на мінімізацію упередженості в конгломераті лабораторій (Лабораторія), куди лабораторії В до Н регулярно надсилають зразки пацієнтів, які вони вже мають

аналізуватися до лабораторії А, яка бере участь у національній або міжнародній ECQ (рис. 2.4).

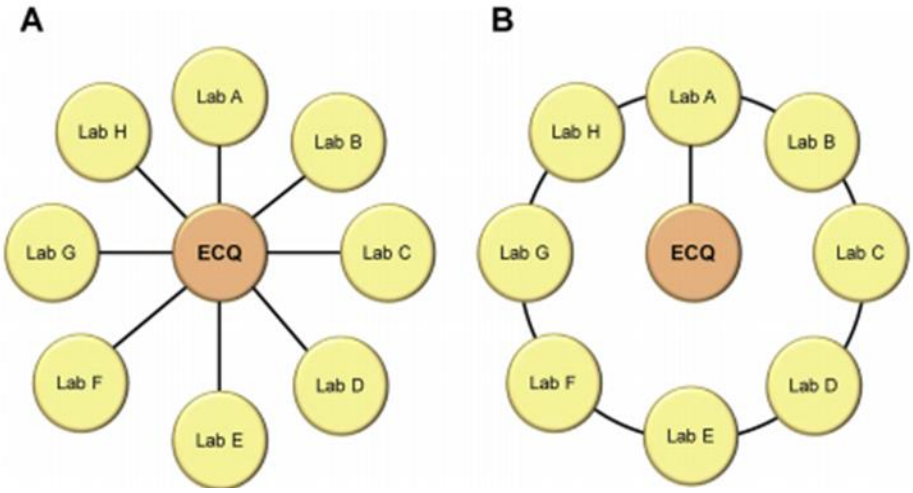

Рисунок 2.4 − Спрощення традиційної зовнішньої схеми контролю якості

Суттєві відмінності між поняттями вимірюваної величини та аналізу повинні бути належним чином відзначені. Вимірювана величина оціненої швидкості клубочкової фільтрації (GFR) – це кількість призначена для вимірювання, тоді як аналітований креатинін або цистатин с є речовиною, що представляє інтерес при вимірюванні в плазмі та сечі аналізованих речовин, концентрація яких входить до рівняння, що використовуються для цієї метих [28-31]. Розглянемо використання як сурогату маркеру швидкості клубочкової фільтрації. Процес звільнення зазвичай оцінюється, наприклад, за допомогою ЕДТА (етилендіамін тетраоцтової кислоти) або кліренсу іогексолу. Наприклад є більш непряма оцінка кліренсу – об’єм плазми крові повністю очищений для креатиніну / цистатину змінної с за одиницю часу. Є кілька варіантів під час спроби мінімізувати вплив невизначеності концентрацій креатиніну на оцінку ЕДТА, наприклад, використовуючи ферментативні методи замість методів Яффе, або виправлення впливу м’язової маси, зазначивши, чи є суб’єкт чоловічої статі, чи жінки, африканського походження чи ні, та з урахуванням віку. Інформацію про оцінки факторів наведено в таблиці 2.1.

Таблиця 2.1 – Оцінка факторів, вимірюваних або констант, у питаннях, що оцінюють кліренс концентрацій креатиніну або кустаніну.

Фактори, що впливають на зв’язок між концентраціями креатиніну та оцінками за GFR	Ким/Чим оцінено
М’язова маса	Чоловік чи жінка Африканське походження чи ні
Вік	Вік
Площа поверхні тіла	Розраховано на основі даних про стать, зріст і вагу
Концентрація альбуміну в плазмі	
Концентрація сечовини в плазмі	
Прийом їжі, що містить креатинін	
Фізичні вправи	
Чинники у різних питаннях eGFR	Оцінено в популяціях, у яких ШКФ вимірювались незалежними методами

Інші фактори, включаючи концентрацію альбуміну або сечовини в плазмі, рідко трапляються враховано в розрахунку, оскільки вони відіграють лише незначну роль у вдосконаленні Оцінка коефіцієнта шуму (КФР) гломерулярної фільтрації. Суттєві невизначеності, пов’язані з тим, коли оцінка математичних факторів у рівняннях РШФР рідко відображається.

При оцінці факторів використовувались різні методи розрахунку невизначеності у рівняннях або які вимірювані величини включались, але рідко згідно з узгодженою методологією; наприклад, LPU. Незважаючи на те,

що eGFR використовується для регулювання дозування діагностичної невизначеності при оцінці РКШ як міри швидкості клубочкової фільтрації рідко, якщо коли-небудь повідомляти про результати.

Незважаючи на теоретичні переваги розширених оцінок невизначеності в лабораторі залишається показати, що звітування про СКЗ разом із належними оцінками невизначеності буде успішно запроваджено та широко використано на практиці, враховуючи відчувається відсутність інтересу до оцінок невизначеності в клінічній медицині. Імовірно що знання даних повної діагностичної оцінки медичних тестів у практичній медичній допоможе надалі віддавати перевагу оцінкам невизначеності, незважаючи на те, що вони тільки доповнюють.

Поняття справжнього значення, якщо воно взагалі використовується та пов’язане із використовуваням у системі вимірювання. Однак система вимірювання є лише частиною загальної реальності лабораторної медицини. Кінцева мета вимірювання концентрації вимірюваної величини для пацієнта повинна покращити розуміння можливого стану захворювання або провести моніторинг ефекта лікування. На виконання цієї мети впливають невизначеності не тільки системи вимірювань (невизначеність вимірювань), але й біологічні зміни, попередньої аналітичної варіації та поаналітична варіація. Слід взяти всі ці компоненти невизначеності та враховувати у всебічній моделі реальності, яка всебічно відображає діагностичну невизначеність результату вимірювання. Прогнозні значення – це діагностична невизначеність, визначена за допомогою баєсівського підходу для поєднання імовірності попереднього тестування за специфікацією продуктивності (що є тією самою сумарною невизначеністю процесу тестування).

Статистика, філософія досліджень та метрологія – це рітельно обговорювані науки в історичному плані в цей час. Сучасна дискусія між методами помилок та невизначеності є, наприклад, відображенням напруженості між статистичним та баєсовим підходами. Один або обидва з

них можуть перемагати в довгостроковій перспективі або, можливо, бути оскарженими новими лініями наукової думки, включаючи причинно-наслідковий аналіз.

Площа поверхні тіла розраховується на основі даних про стать, зріст і вагу. Враховується концентрація альбуміну в плазмі, концентрація сечовини в плазмі, прийом їжі, що містить креатинін, виконання фізичних вправ та фактори в різних рівняннях eGFR. Також оцінюється в популяціях, для яких GFR було виміряно незалежними засобами. Наприклад, за допомогою кліренсу та іогексолу. Можливо, час визнати, що всі наші гносеологічні інструменти є тимчасовими та помилковими елементами, і що шлях до кращих знань обов'язково близький до критичного мислення. Онтологічна однозначність щодо причинно-наслідкових зв'язків та / або статистичних даних не з'являються самі по собі або завдяки будь-якому аналітичному процесу, замість нього, в чесній і критичній діяльності безліч невдач і певних успіхів.

Моделі помилок, включаючи моделі ТЕ для кількісної оцінки якості вимірювальних систем в лабораторній медицині, широко застосовуються і добре слугують лабораторіям з часу їх існування тобто введення в 1970-х роках. Розвиток в інших галузях міжнародної метрології запровадив узагальнені та ще більш комплексні методи (LPU) для кількісного визначення поєднання декількох невизначеностей не тільки вимірювальних систем, але і комбіновані ефекти всіх частин всього ланцюга випробувань (доаналітичного, аналітичного, післяаналітичний та клінічний), коли використовується для діагностики та моніторингу результатів лікування. LPU та баєсові розрахунки ще не повністю показали свою практичну додану вартість в лабораторії та в клінічній медицині. Такі зрушення залежать від витонченості кінцевого споживача результатів вимірювань у практичній медичній допомозі.

Поняття невизначеності все частіше використовуються у всіх сферах людських зусиль, у тому числі комерції, інженерних та екологічних наук, і їх

викладають загалом навчальних програм, включаючи медицину, тому їх практична цінність, швидше за все, буде більше оцінюватися. Підходи LPU, швидше за все, отримають більш широке визнання в лабораторній медицині, оскільки фокус переходить від властивості вимірювальної системи до належного використання результату вимірювання при діагностиці та моніторингу ефектів лікування, включаючи всі фактори, що спричиняють невизначеність. Переглянуті версії GUM та VIM [32] імовірно явно підтримують як підходи до помилок, так і невизначеності, включаючи обидва частотні та баєсівські статистичні методи [33].

2.4 Методи відновлення даних

Під «неповністю» або «некомплектністю» мається на увазі те, що деякі дані з тих чи інших причин пропущені або відсутні у вихідному масиві даних. Традиційними причинами, що призводять до появи пропусків, є:

- неможливість їх отримання або обробки;
- спотворення або приховування інформації;
- всілякі поломки технічного обладнання;
- природні явища;
- економічні причини і т.ін.

У підсумку на вхід програм аналізу зібраних даних надходять неповні відомості. З проблемою обробки пропусків у масивах даних доводиться стикатися при проведенні різноманітних технічних, соціологічних, економічних, астрономічних, біологічних, статистичних та ін. досліджень. Західні дослідники такі дані з пропущеними значеннями називають «missing data» або «incomplete data» [15]. В [15] наведено результати аналізу методів, що застосовується у напрямку врахування невизначеності у даному напрямку (рис. 2.5).

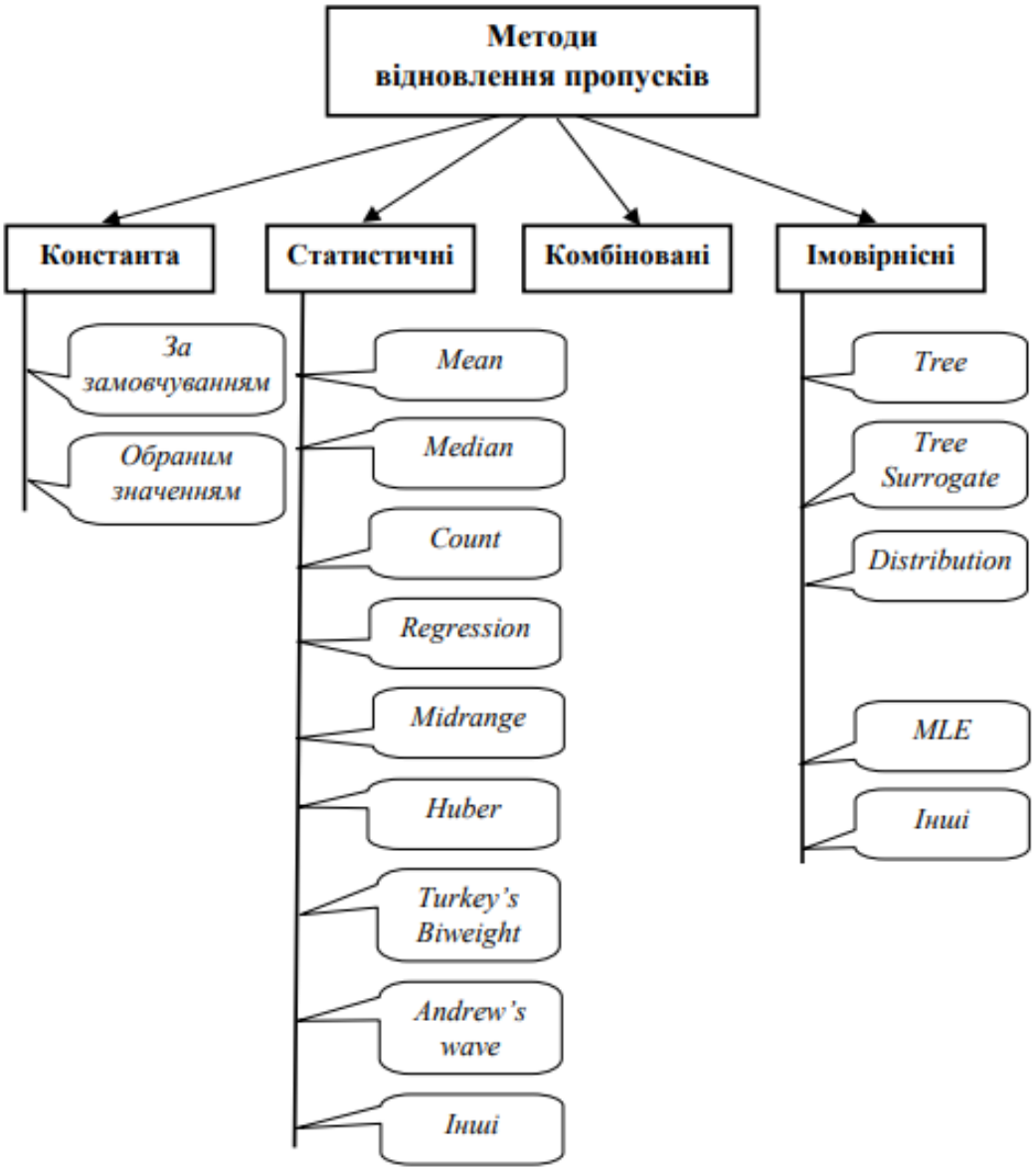

Рисунок 2.5 – Класифікація методів заповнення пропущених даних

2.5 Реалізація алгоритму доповнення даних

За роботу з пропущеними даними в статистичному пакеті SPSS передбачена модель Missing Values, яка містить дві процедури:

1. Процедура «Аналіз пропущених значень» виконує три основні функції:

– Описує структуру пропущених даних. Де розташовані пропущені значення? Наскільки широку область вони охоплюють? Чи є тенденція до пропуску значень в декількох спостереженнях у пар змінних?

– Оцінює середні, середньоквадратичні відхилення, ковариации і кореляції для різних методів обробки пропущених значень: за списками, попарно, регресія або ОМП (максимізація очікувань). Попарний метод виводить також частоти повних пар спостережень.

– Виробляє включення (імпутацію) на місце пропущених значень оціночних значень, використовуючи метод регресії або ОМП (максимізація очікувань).

2. Процедура «Множинної імпутація». Мета «множинної імпутації» – згенерувати можливі значення для пропущених значень, створивши кілька нібито повних наборів даних. Аналітичні процедури, що працюють з наборами даних багаторазового включення, генерують результати для кожного нібито повного набору даних плюс об’єднаний результат, який оцінює, якими були б результати, якби початковий набір даних був без пропущених значень. Ці об’єднані результати зазвичай є більш точнішими, ніж отримані за методами одного включення.

3 РЕАЛІЗАЦІЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Розроблене програмне забезпечення noMissingness засноване на описаних вище методах. Програма дозволяє створювати нові дані або ж відкривати і редагувати дані з файлу формату .txt з поділом знаком табуляції.

Для роботи з даними необхідно вибрати в меню відновлення пропущених значень. У вікні «Вибір методів» відзначити методи одноразових включень, на основі яких будується агрегована оцінка за правилом Рубіна на стадії багаторазового включення. Додатково можна вивести на екран матрицю коефіцієнтів кореляції Пірсона і матрицю коваріацій.

3.1 Опис алгоритму програми

Після того як користувач вніс у програму дані і вибрав методи одноразових включень, програма на початку роботи знайде матрицю коваріації, матрицю коефіцієнтів кореляції за формулою Пірсона і 95% довірчі інтервали по кожному колонку. Матриця коефіцієнтів кореляції використовується для знаходження повнокомплектною вибірки з найбільшим коефіцієнтом кореляції з неполнокомплектной вибірки, над якою проводиться аналіз, в іншому випадку (якщо немає повнокомплектних вибірок) програма знайде неполнокомплектную вибірку і заповнить її середнім за вибіркою для подальшого аналізу. Даний алгоритм, в рамках програми, використовується в усіх процедурах одноразового включення за винятком заповнення середнього по вибірці. Матриця ковариации потрібна для початкових розрахунків в методах, які використовують рівняння регресії. Довірчі інтервали необхідні для подальшого аналізу включень.

При використанні методу програма працює з копією даних для збереження цілісності оригіналу і далі обробляє дані по їх опису.

При цьому дані з включеннями з одноразових методів і з багаторазового методу будуть автоматично збережені в файлі SaveFF.txt в корені програми.

3.2 Обчислювальний експеримент

На основі медичних повнокомплектних даних з дослідження ожиріння дітей з 15 спостережень і 10 ознак побудований обчислювальний експеримент зі згенерували повністю випадковими пропусками (MCAR) як зображено на рисунку 3.1, що становлять 2,8% від усіх даних. Масиви даних оброблялися описаними ранніше методами відновлення пропущених значень: підстановка середнього за вибіркою, підстановка медіани, метод Hot Deck, модель максимальної правдоподібності (EM алгоритм), парна регресія, стохастична регресія і метод multiple imputation Рубіна з використанням Jack Knife.

№	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9
1	5,1	4,3	58,1	2,74	0,140	21,1	75	90	10
2	4,7	4,9	45,1	2,23	0,130	18,5	78	88	12
3	4,2	5,1		5,90	0,270	17,0	66	108	10
4	3,3	3,0	25,5	2,31	0,140		76	92	12
5	4,4	3,6	31,4	2,10	0,150	17,0	80	96	12
6	4,7	4,3	41,0	2,45	0,150	18,6	94	104	8
7	4,6		58,0	3,40	0,200	21,2	80	94	10
8	6,3	4,2	40,3	1,90	0,150	18,3	80	92	8
9	4,8		28,4	2,80	0,180	15,8	80	88	15
10	4,8	5,0	56,7	2,20	0,130		84	90	11
11	5,2	5,0	55,0	2,80	0,150	18,1	88	98	20
12	4,8	4,6	57,8	3,00	0,160	18,8	80	86	23
13	5,1	4,8		3,10	0,110	21,7	80	85	42
14	5,4	5,6	48,0	3,40	0,140	22,0	89	104	14
15	4,9	4,2	42,0	3,60	0,160	21,6	78	90	16

Рисунок 3.1 – Вихідні дані з пропусками

Скористаємося пакетом SPSS щоб знайти довірчий інтервал для кожного з параметрів з пропусками. Результати наведені в таблиці 3.1.

Таблиця 3.1 – Довірчі інтервали при 6 пропусках

	95% довірчий інтервал	
	Нижня межа	Верхня межа
VAR2	4,233	4,733
VAR3	41,503	52,444
VAR6	18,333	20,499

Результати роботи програми для 6 пропусків наведені в таблиці 3.2.

Таблиця 3.2 – Результати заповнення при 6 пропусках

	Підстановка середнього	Підстановка медіани	Hot Deck	EM алгоритм	Парна регресія	Стохастична регресія	Multiple imputation	Jack Knife	Реальне значення
1	4,958	4,450	4,89	4,803	4,142	4,726	4,662	4,662	4,4
2	4,958	4,450	3,63	4,944	4,283	4,867	4,522	4,522	4,4
3	49,695	43,550	56,6	43,786	30,72	45,205	44,935	44,93	47,3
4	49,695	43,550	57,8	52,363	39,30	53,782	49,415	49,41	44,0
5	21,128	18,550	15,8	20,670	17,82	19,834	18,967	18,96	18,6
6	21,128	18,550	19,6	20,406	17,55	19,570	19,483	19,48	16,4

Для значень, відновленими методом багаторазової вставки за правилом Рубіна, що не ввійшли в довірчий інтервал метод коригування Jack Knife підібрав нові, як наведено в таблиці 3.3.

Таблиця 3.3 – Відхилення при 6 пропусках

Відхилення (%)	Підстановка	Підстановка медіани	Hot Deck	EM алгоритм	Парна регресія	Стохастична регресія	Multiple imputation	Jack Knife
0-20	5	6	4	5	5	5	6	6
20-50	1	0	2	1	1	1	0	0
50-100	0	0	0	0	0	0	0	0
100+	0	0	0	0	0	0	0	0

Далі, для порівняння ефективності методів проведено подібний аналіз в таблиці 3.4 з тими ж вхідними даними, але з 15 пропусками:

Таблиця 3.4 – Довірчі інтервали при 15 пропусках

	95% довірчий інтервал	
	Нижня межа	Верхня межа
VAR2	5,15	6,025
VAR3	43,975	56,55
VAR6	18,2	21,775

Результати роботи програми дляя 15 пропусків наведені в таблиці 3.5.

Таблиця 3.5 – результати заповнення при 15 пропусках

	Підстановка середнього	Підстановка медіани	Hot Deck	ЕМ алгоритм	Парна регресія	Стохастична регресія	Multiple imputation	Jack Knife	Реальне значення
1	5,027	4,250	5,3	4,390	2,747	4,368	4,362	4,362	4,9
2	5,027	4,250	3,19	4,463	2,820	4,452	4,034	4,276	3,0
3	5,027	4,250	2,91	5,457	3,814	5,446	4,485	4,485	4,4
4	5,027	4,250	2,53	4,910	3,267	4,899	4,147	4,470	4,4
5	5,027	4,250	3,30	4,362	2,720	4,351	4,002	4,258	5
6	48,850	41,150	55,2	42,225	28,978	39,717	42,662	42,66	47,3
7	48,850	41,150	58,1	47,568	34,321	45,060	45,841	45,84	41
8	48,850	41,150	58,1	47,568	34,321	45,060	45,841	45,84	41,1
9	21,340	17,550	13,1	21,385	14,583	30,647	19,766	19,76	18,5
10	48,850	41,150	31,3	51,842	39,321	50,247	43,793	43,79	44,2
11	21,340	17,550	12,7	21,321	14,520	31,221	19,776	19,77	17,1
12	21,340	17,550	13,0	21,385	14,583	30,647	19,766	19,76	18,5
13	21,340	17,550	14,3	21,385	14,583	31,285	20,083	20,08	15,8
14	21,340	17,550	21,9	21,353	14,552	31,253	21,332	19,34	18,6
15	21,340	17,550	19,5	21,258	14,698	31,158	20,931	18,88	22,1

Дані про відхилення при 15 пропусків наведені в таблиці 3.6.

Таблиця 3.6 – Відхилення при 15 пропусках

Відхилення (%)	Підстановка	Підстановка медіани	Hot Deck	EM алгоритм	Парна регресія	Стохастична регресія	Multiple imputation	Jack Knife
0-20	12	12	6	11	7	7	11	13
20-50	2	3	9	5	8	4	4	2
50-100	1	0	0	0	0	4	0	0
100+	0	0	0	0	0	0	0	0

З проведеного аналізу можна зробити висновок, що методи, в основі яких лежить рівняння регресії (включаючи EM алгоритм), більш схильні до зміщення в результаті збільшення некомплектності даних. Метод Hot Deck так само показав відхилення від реальних значень, але не більше 50% як у випадку зі стохастичною регресією і підстановкою середнього за вибіркою, а метод Jack knife в якійсь мірі покращив оцінки методу багаторазового включення.

ВИСНОВКИ

Розуміння даних і отримання обґрунтованих висновків мають першорядне значення в нинішню епоху великих даних. Для цього широко використовуються методи машинного навчання і теорії імовірностей в різних областях. Один критично важливий, але менш вивчений аспект – це те, як дані і невизначеності моделі фіксуються і аналізуються. Правильна кількісна оцінка невизначеності дає цінну інформацію для прийняття оптимального рішення стосовно діагнозу та лікувального сценарію. У цій роботі були розглянуті відповідні дослідження, проведені за останні 30 по обробці невизначеностей в медичних даних з використанням теорії ймовірностей і методів машинного навчання.

Медичні дані більш схильні до невизначеності через наявність шуму в даних. Тому для точного діагнозу дуже важливо мати чисті медичні дані без зайвого шуму. Для вирішення цієї проблеми необхідно знати джерела шуму в медичних даних. На підставі медичних даних, отриманих лікарем, призначається діагноз захворювання і план лікування.

Є багато знань про оптимальні методи лікування, оскільки в медичній науці є багато джерел невизначеності. Результати показують, що є кілька проблем, які необхідно вирішити при обробці невизначеності в медичних необроблених даних і нових моделях. У цій роботі підсумовувано роботу різних методів, використовуваних для вирішення цієї проблеми.

В даний час застосування нових методів глибокого навчання для вирішення таких невизначеностей значно розширилося. В процесі створення програмного продукту була розглянута реалізація функціоналу пакету SPSS, що працює з пропусками в даних. Був реалізований програмний продукт, який застосовує методи відновлення пропущених даних. Програмний продукт використовує популярні методи відновлення даних для заповнення пропусків в некомплектних даних.

А також, було проведено обчислювальний експеримент, в результаті якого зроблений порівняльний аналіз ефективності застосування багаторазової вставки за правилом Рубіна і одноразових методів заповнення пропусків. Після оцінки ефективності підходів можна зробити висновок, що Multiple imputation дійсно дає хороші результати відновлення. За реальними чисельним значенням вони можуть відрізнятися, але тим не менше вони будуть входити в 95% довірчного інтервалу, а метод Jack Knife на кроці аналізу в якійсь мірі здатний поліпшити результат, але має сенс використовувати при некомплектності від 20% і вище, оскільки при менших втратах ефекту не видно.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Mobile expert system for diagnostic human state in emergency situations./ Kuzomin, O., Dudka, O., Vasylenko, O., Shylo, R., Lyashenko, V.// International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(4), pp. 6484-6489, 334

2. Intellectual Models and Means of the Biometric System Dynamics of Rinosinusite / Oleksii Vasilenko, Oleksandr Kuzomin, Tatyana Khripushina // International Journal "Information Models and Analyses" Volume 7, Number 4. PP.2019. 350 – 361.

3. Литтл Р.Дж.А., Статистический анализ данных с пропусками. Финансы и статистика / Литтл Р.Дж.А., Рубин Д.Б – Москва, 1991, 336 стр.

4. Intelligent geoinformatic expert system for providing emergency help during extreme situations./Kuzomin, O., Stukin, M., Bozhkov, D.//International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, GEM, 2019, 18(2.2), pp. 269-276

5. V.S. Petrovichev, A.V. Melekhov, M.A. Saifullin, I. G. Nikitin / Computed tomography for coronavirus infection: differential diagnosis based on clinical examples // Archives of Internal Medicine. Original articles. No. 5.2020. S. 357-371

6. The patient organism modeling for diagnosis with the usage of a multi agent representation /Kuzomin, O., Dudka, O., Vasylenko, O., Lyashenko, V. // International Journal of Emerging Trends in Engineering Research, 2020, 8(9), pp. 5733-5739

7. Using of ontologies for building databases and knowledge bases for consequences management of emergency/Kuzomin, O., Dudka, O., Vasylenko, O., Radchenko, V., Lyashenko, V.//International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(4), pp. 5040-5045

8. Research of the intellectual system of knowledge search in Databases / Oleksii Vasilenko, Oleksandr Kuzomin, Bohdan Maliar // International Journal "Information Models and Analyses" Volume 7, Number 4. PP.2019. 327 – 338.

9. Research of medical diagnostic data search methods / Oleksii Vasilenko, Oleksandr Kuzomin, Oleksandr Shapoval // International Journal "Information Models and Analyses" Volume 7, Number 4. PP.2019. 339 – 349.

10. Forming Medical Database and Knowledge for Diagnostic Disease / Oleksii Vasilenko, Oleksandr Kuzomin Vladislav Shvets. // International Journal "Information Models and Analyses" Volume 7, Number 4. PP.2019. 362 – 372.

11. Automated Tests for Errors in Computer System ‘Environment’/ Oleksii Vasilenko, Oleksandr Kuzomin. // International Journal "Information Models and Analyses" Volume 7, Number 4. PP.2019. 373 – 384.

12. Developing methods based on text mining technology to Improve the quality and speed of automatic clustering of Documents / Oleksii Vasilenko, Oleksandr Kuzomin, Artem Mertsalov // International Journal "Information Models and Analyses" Volume 7, Number 4. PP.2019. 385 – 397.

13. Розробка багатоагентних структур для вирішення проблем медичної системи діагностування / О.Я. Кузьомін, О.О. Василенко, Свістунов І.О. // Радіоелектроніка та інформатика. 2020. №2. С. 47 – 54.

14. Н.В. Кузнєцова / Виявлення та оброблення невизначеностей у формі неповних даних методами інтелектуального аналізу.// Системні дослідження та інформаційні технології, 2016, № 2. С.104-115.

15. В.С. Петровичев, А.В. Мелехов, М.А. Сайфуллин, И.Г. Никитин / Компьютерная томография при коронавирусной инфекции: дифференциальный диагноз на клинических примерах // Архивъ внутренней медицины. Оригинальные статьи. №5.2020. С. 357-371

16. The patient organism modeling for diagnosis with the usage of a multi agent representation /Kuzomin, O., Dudka, O., Vasylenko, O., Lyashenko, V. // International Journal of Emerging Trends in Engineering Research, 2020, 8(9), pp. 5733-5739

17. Розробка структур медичних агентів для вирішення проблем медичної системи діагностування / О.Я. Кузьомін, О.О. Василенко, А.В. Горшколєпов // Радіоелектроніка та інформатика. 2020. №2. С. 55- 65.

18. Н.В. Кузнєцова / Виявлення та оброблення невизначеностей у формі неповних даних методами інтелектуального аналізу.// Системні дослідження та інформаційні технології, 2016, № 2. С.104-115.

19. JCGM. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Paris: Joint Committee for Guides in Metrology; 2008.

20. Barwick V, Prichard E. Terminology in analytical measurement - introduction to VIM. Brussels: Eurachem; 2011. Available at: https://www.eurachem.org/index.php/ publications/guides/terminology-in-analytical-measurement. Accessed February15, 2016.

21. De Bievre P. The 2007 International Vocabulary of Metrology (VIM), JCGM 200:2008 [ISO/IEC Guide 99]: meeting the need for intercontinentally understood concepts and their associated intercontinentally agreed terms. Clin Biochem 2009.

22. Гаврилов П.А. Обзор методов предобработки, используемых для решения задач классификации в условиях неполноты: Вестник Рязанского государственного радиотехнического университета. № 55, 2016, стр. 141-145.

23. Орлов А.И., Прикладная статистика М.: Издательство «Экзамен», 2004, 656 стр.

24. Попова В.Б. Особенности регрессионного анализа с применением метода Джекнайф. Мичуринский государственный аграрный университет – Мичуринск, 2009, стр. 32-36.

25. Rubin D.B. Statistical Analysis with Missing Data. New York: John Wiley & Sons, 1987, 408 p.

26. Craig K. Enders, APPLIED MISSING DATA ANALYSIS. The Guilford Press. 72 Spring Street – New York, 2010, 401 p.

27. Chen, The comparative efficacy of imputation methods for missing data in structural equation modeling / Chen, & Harlow, European Journal of Operational Research, 2003,pp 53-79.

28. Schafer, J.L., Missing data: Our view of the state of the art. Psychological Methods / Schafer, J.L., & Graham, J.W., 2002, pp 147-177.

29. W.Yung, Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling. Survey Methodology / W.Yung, 1996, pp 23-31.

30. Giordani A, Mari L. Measurement, models, and uncertainty. IEEE T Instrum Meas 2012;61(8):2144–52.

31. Pearl J. Causal inference in statistics: an overview. Stat Surv 2009;3:96–146.

32. Bich W. Revision of the ‘Guide to the Expression of Uncertainty in Measurement’. Why and how. Metrologia 2014;51(4):S155–8.

33. Uncertainty in Measurement and Total Error 33 Модуль Missing Values [Электронный ресурс], сайт компании «IBM». Режим доступа: https://www.ibm.com/support/knowledgecenter/ru/SSLVMB_22.0.0/kc_gen/com.ibm.spss.statistics.help_statistics_mainhelp-gen11.html