Харківський національний університет імені В.Н. Каразіна

Факультет математики і інформатики

Кафедра прикладної математики

Кваліфікаційна робота

магістра

на тему «Множинна імпутація як прийнятний спосіб роботи з відсутніми даними»

Виконав:

студент групи МП-61 6 курсу

спеціальність 113 - прикладна математика

освітньо-наукова програма

«Прикладна математика»

Борзенков А. І.

Наукові керівники:
к. ф.-м. наук Ревіна Т. В.
Principal Statistical Programmer / Analyst
Калініченко В. В.

Рецензент: Senior Statistical Programmer / Analyst Висоцький М. В.

Харків - 2024 рік

Анотація

Борзенков А. І. Множинна імпутація як прийнятний спосіб роботи з відсутніми даними. У даній роботі досліджено метод множинної імпутації, спрямований на обробку відсутніх даних у клінічних випробуваннях. Зосереджено увагу на його застосуванні, зокрема на лінійній регресії. Після аналізу методів імпутації, увага приділяється статистичному аналізу, включаючи метод Каплан-Майера та різні варіанти Т-тесту та ANOVA. Далі показано, як метод множинної імпутації може створити "найгірший" набір даних для оцінки впливу відсутності даних на результати. На заключному етапі ефективність розроблених методів перевіряється на реальних або симульованих даних, що підкреслює їхню застосовність у клінічних дослідженнях.

Abstract

Borzenkov, A. Multiple imputation as a valid way of dealing with missing data. This work examines the multiple imputation method, a powerful tool for handling missing data in clinical trials. Attention is focused on its application, particularly linear regression. Following an analysis of imputation methods, statistical analysis is addressed, including the Kaplan-Meier method and various forms of T-tests and ANOVA. Furthermore, it demonstrates how the multiple imputation method can create a "worst-case" dataset to assess the impact of missing data on results. Finally, the efficacy of the developed methods is tested on real or simulated data, underscoring their applicability in clinical research.

Зміст

1 Вступ

У сучасному дослідницькому середовищі важливість аналізу даних є невід'ємною частиною наукових досліджень. Незалежно від області, в якій проводяться дослідження, відсутність даних може стати суттєвим обмеженням для отримання достовірних результатів. Відсутні дані можуть виникати з різних причин: від відмови пацієнтів від участі в дослідженні до помилок при введенні даних.

Одним з методів роботи з відсутніми даними є багаторазова імпутація. Цей підхід дозволяє створити кілька "повних" наборів даних, в яких відсутні дані замінюються "правдоподібними" значеннями. Потім ці набори даних аналізуються окремо, а результати комбінуються з урахуванням невизначеності (варіативності) різних імпутацій.

У цій роботі ми розглянемо метод багаторазової імпутації, його застосування в клінічних випробуваннях та поточні методи роботи з відсутніми даними. Також ми дослідимо можливість використання багаторазової імпутації для створення "найгіршого" набору даних та перевіримо його на реальних або штучних наближених до реальних даних.

2 Опис методу множинної імпутації

Метод множинної імпутації є потужним інструментом для роботи з відсутніми даними. Як описано в [1-3], основний принцип роботи цього методу полягає в створенні кількох “повних” наборів даних, в яких відсутні значення замінюються “правдоподібними” значеннями.

Давайте розглянемо цей принцип детальніше:

1. **Генерація множини імпутацій**:
    - Спочатку вибирається кілька (зазвичай 5-10) “повних” наборів даних.
    - Відсутні значення в кожному наборі даних заповнюються “правдоподібними” значеннями, використовуючи статистичні методи, такі як лінійна регресія, дерева рішень або байєсівська мережа.
2. **Аналіз окремо для кожного набору даних**:
    - Кожен “повний” набір даних аналізується окремо, використовуючи стандартні статистичні методи.
    - Отримані результати можуть бути різними для кожного набору даних.
3. **Комбінування результатів**:
    - Результати з усіх “повних” наборів даних комбінуються, враховуючи невизначеність (варіативність) різних імпутацій.
    - Це дозволяє отримати більш точні та надійні результати, ніж при аналізі лише одного набору даних.

Метод множинної імпутації дозволяє уникнути втрати статистичної потужності, зберігаючи при цьому достовірність результатів.

2.1 Лінійна регресія

Лінійна регресія шукає лінійну залежність між незалежною змінною X і залежною змінною Y. Цей метод детально описаний в [4-5]. Основна ідея полягає в тому, щоб знайти такі коефіцієнти β_0 (перетин з осі Y) і β_1 (нахил лінії), які найкращим чином апроксимують спостережувані дані.

У лінійній регресії модель можна виразити наступним чином:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

де:

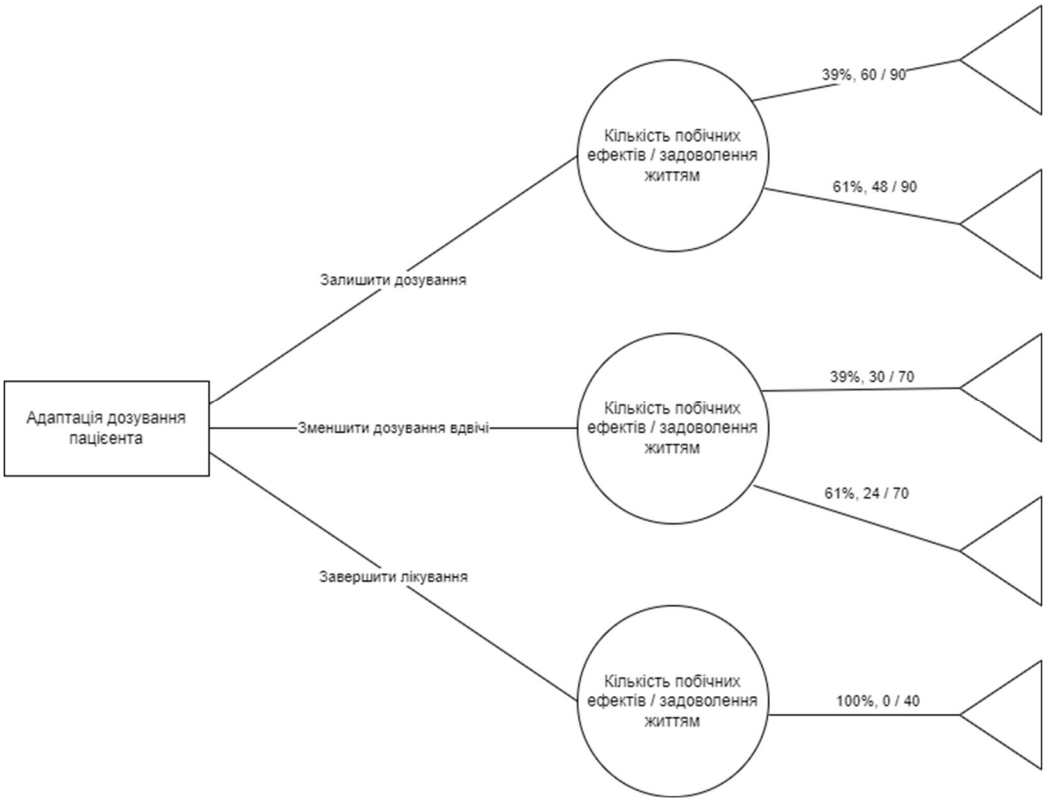
- Y - залежна змінна (вихідний параметр),
- X - незалежна змінна (вхідний параметр),
- β_0 - перетин з осі Y (константа),
- β_1 - нахил лінії (коефіцієнт регресії),
- ε - помилка (різниця між фактичними та прогнозованими значеннями).

2.2 Дерево рішень

Дерево рішень [6] - це діаграма, яка представляє рішення, які необхідно прийняти, різні сценарії, які можуть виникнути, і всі можливі результати. Вона допомагає глобально побачити всі можливі сценарії та наскільки ймовірно кожен з них здійсниться, що дає змогу знати, наскільки ризикованим є кожне рішення. Дерево рішень складається з таких елементів:

- Вузол рішення (□): відповідає рішенню, яке необхідно прийняти. У дереві рішень він представлений квадратом.
- Вузол імовірності (○): символізує те, що може мати місце кілька сценаріїв. Кожна гілка, що виходить із вузла імовірності, представляє інший сценарій. Вузол імовірності малюється з порожнім колом у дереві рішень.

- Кінцевий вузол (△): представляє результат, тому їх легко ідентифікувати, оскільки жодна гілка не залишає їх. У дереві рішень вони представлені трикутниками.



Малюнок 1. Приклад дерева рішень

3 Опис статистичного аналізу в клінічних випробуваннях

Для перевірки ефективності та безпеки ліків в клінічних випробуваннях використовуються різноманітні статистичні методи. Ось деякі з них:

- **Контрольовані рандомізовані дослідження (RCT)**. Це “золотий стандарт” для оцінки ефективності ліків. Учасники випадковим чином розподіляються між групами, які отримують лікування або плацебо.
- **Аналіз виживаності.** Використовується для оцінки часу до настання певної події, наприклад, рецидиву захворювання. Метод Каплана-Майєра та модель пропорційних ризиків Кокса є поширеними техніками.
- **Мета-аналіз**. Цей метод об’єднує результати кількох досліджень для отримання загального висновку про ефективність лікування.
- **Логістична регресія**. Використовується для оцінки ймовірності настання події (наприклад, поліпшення стану) в залежності від лікування.
- **Т-тест або ANOVA**. Ці тести порівнюють середні значення між двома або більше групами для оцінки статистичної значущості різниці.
- **Час до настання події**. Аналізи, які оцінюють час від початку дослідження до настання певної події, такої як погіршення стану.
- **Баєсівський аналіз**. Метод, який включає попередні знання та оновлює ймовірності на основі отриманих даних.

Ці методи дозволяють оцінити, чи є лікування ефективним порівняно з плацебо або іншими стандартними методами лікування, а також визначити, які

побічні ефекти можуть виникати і якою мірою. Важливо, що вибір статистичного методу залежить від типу даних, цілей дослідження та дизайну випробування [7].

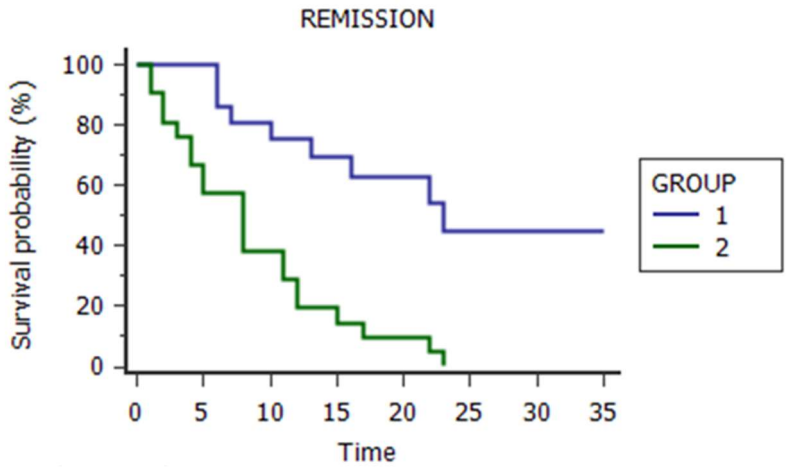
3.1 Метод Каплан-Майера

Метод Каплан-Майера [9] - це статистичний метод для оцінки часу виживання при дослідженні виживання, особливо в медичних дослідженнях або при дослідженні смертності. Використовується для аналізу часу до події, таких як смерть, відмова від лікування або інший важливий кінцевий результат.

Припустимо, що ми маємо набір даних, який складається з часу спостереження і статусу події (подія відбулася або ні) для кожного учасника. Метод Каплан-Майера дозволяє побудувати криву виживання, яка відображає ймовірність того, що учасник виживе протягом певного часу.

Для побудови кривої виживання за методом Каплан-Майера використовуються наступні кроки [10]:

1. Сортування даних: Спочатку дані сортуються в порядку зростання часу спостереження.
2. Підрахунок ризиків імовірностей виживання: Для кожного часового інтервалу, в якому відбувається подія, обчислюються ймовірності виживання (відсоток людей, що вижили до цього часу).
3. Обчислення кумулятивних ймовірностей виживання: Кумулятивні ймовірності виживання обчислюються, як добуток імовірностей виживання для кожного часового інтервалу.
4. Побудова кривої виживання: Крива виживання будується шляхом поєднання точок, що відповідають часовим інтервалам і кумулятивним ймовірностям виживання.



Малюнок 2. Приклад графіку за методом Каплан-Майера

Формула для обчислення кумулятивних ймовірностей виживання (S(t)) для кожного часового інтервалу може бути виражена наступним чином:

$$S(t) = S(t-1) \times \left(1 - \frac{d}{n}\right)$$

Де:

- S(t) - кумулятивна ймовірність виживання до часу
- S(t−1) - кумулятивна ймовірність виживання до попереднього часу.
- d - кількість подій (смертей або відмов від лікування) в даному часовому інтервалі.
- n - кількість учасників на початку цього часового інтервалу.

Отже, за допомогою цих формул можна побудувати криву виживання за методом Каплан-Майера для набору даних про час виживання. Ця крива допомагає аналізувати виживання та порівнювати групи учасників за їхньою тривалістю виживання.

3.2 Т-тест

Т-тест — це статистичний тест, який використовується для порівняння середніх значень двох груп [11-12]. Існують три основні типи т-тестів:

одновибірковий, незалежний (для двох незалежних вибірок) та парний т-тест. Ось детальний опис кожного з них:

3.2.1 Одновибірковий т-тест

Цей тест порівнює середнє значення вибірки з відомим стандартним значенням. Формула для одновибіркового т-тесту:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

де:

- $\bar{x}$ — середнє значення вибірки,
- μ — відоме середнє значення для порівняння,
- s — стандартне відхилення вибірки,
- n — розмір вибірки.

3.2.2 Незалежний т-тест

Цей тест порівнює середні значення двох незалежних груп. Формула для незалежного т-тесту:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

де:

- $\overline{x_1}$ та $\overline{x_2}$ — середні значення двох вибірок,
- s_p^2 — об’єднана дисперсія вибірок,
- n_1 та n_2 — розміри вибірок.

Об’єднана дисперсія вибірок s_p^2 обчислюється як:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

де s_1^2 та s_2^2 — дисперсії кожної вибірки.

3.2.3 Парний т-тест

Цей тест використовується, коли маємо парні або залежні вибірки, наприклад, коли одна група тестується двічі (до та після експерименту). Формула для парного т-тесту:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

де:

$\bar{d}$ — середнє різниць парних значень,

s_d — стандартне відхилення різниць,

n — кількість пар.

У всіх випадках, значення t порівнюється з критичним значенням t-розподілу з відповідними ступенями свободи для визначення статистичної значущості різниці середніх. Якщо обчислене значення t більше за критичне, то різниця вважається статистично значущою. Частіше за все в клінічних випробуваннях використовується незалежний т-тест.

3.3 ANOVA

Метод ANOVA (аналіз варіативності) - це статистичний метод, який використовується для тестування різниці між двома або більше середніми значеннями [13-14]. Він схожий на t-тест, але t-тест зазвичай використовується для порівняння двох середніх, тоді як ANOVA використовується, коли у вас є більше двох середніх для порівняння.

ANOVA базується на порівнянні варіативності (або варіації) між вибірками даних з варіацією в межах кожної конкретної вибірки. Якщо варіативність між групами висока, а варіативність в межах групи низька, це свідчить про те, що середні значення груп суттєво відрізняються.

Нижче наведено основні формули, що використовуються в ANOVA:

1. **Сума квадратів обробки (SS обробки)**. Це сума квадратних відмінностей між середніми значеннями груп і загальним середнім. Формула для обчислення SS обробки: $SS_{\text{обробки}} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$, де k - кількість груп, n_i - кількість спостережень в і-й групі, $\bar{X}_i$ - середнє значення і-ї групи, $\bar{X}$ - загальне середнє значення
2. **Сума квадратів помилок (SS помилок)**. Це сума квадратних відмінностей між кожним спостереженням і його середнім значенням групи. Формула для обчислення SS помилок: $SS_{\text{помилок}} = \sum_{i=1}^{k} \quad \sum_{j=1}^{n_i} \quad (X_{ij} - \bar{X}_i)^2$,

   де X_{ij} - j-те спостереження в і-й групі
3. **Загальна сума квадратів (SS загальна).** Це сума квадратних відмінностей між кожним спостереженням і загальним середнім значенням. Формула для обчислення SS загальна:

$$SS_{\text{загальна}} = \sum_{i=1}^{k} \quad \sum_{j=1}^{n_i} \quad (X_{ij} - \bar{X})^2$$

4. **Обчислення ступенів свободи.**

   Ступені свободи для обробки: $df_{\text{обробки}} = k - 1$.

   Ступені свободи для помилок: $df_{\text{помилок}} = N - k$.

   Загальні ступені свободи: $df_{\text{загальна}} = N - 1$.

   Де N - загальна кількість спостережень.
5. **Обчислення середніх квадратів.**

   Середній квадрат обробки: $MS_{\text{обробки}} = \frac{df_{\text{обробки}}}{SS_{\text{обробки}}}$

   Середній квадрат помилок: $MS_{\text{помилок}} = \frac{df_{\text{помилок}}}{SS_{\text{помилок}}}$
6. **Обчислення F-статистики.** $F = \frac{MS_{\text{помилок}}}{MS_{\text{обробки}}}$

Ця F-статистика потім використовується для визначення p-значення, яке використовується для відхилення нульової гіпотези, якщо p-значення менше встановленого рівня значущості (зазвичай 0.05).

4 Використання множинної імпутації для створення "найгіршого" набору даних

У поточному контексті клінічних випробувань використовуються різноманітні методики обробки пропущених даних, які викладені у Додатку 1. Хоча їх існує значна кількість, однак у практиці зазвичай застосовують лише кілька з них, а у більшості випадків пропущені дані лишаються незаповненими.

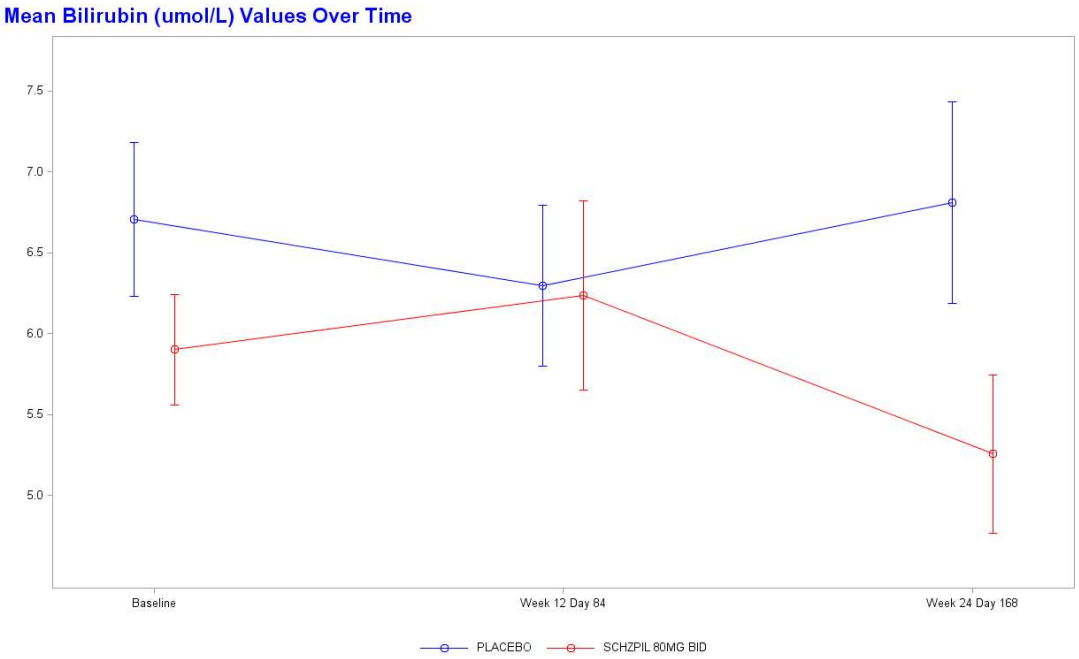
Неналежне урахування пропущених даних під час аналізу може призвести до спотворення результатів та неправильних висновків щодо ефективності або безпеки лікування. Зокрема, пропущені дані можуть впливати на розподіл вибірки і, отже, на отримані результати стосовно ефективності та безпеки досліджуваного препарату.

З метою вирішення цієї проблеми рекомендується використовувати метод множинної імпутації, а з серед відновлених та оригінальних вибірок в незалежних групах вибирати для остаточного аналізу ті, що показують найгірші результати. Наприклад, можна віддати перевагу тим, для яких аналіз варіації між двома групами виявить найменшу відмінність. Врахування ефективності та безпеки ліків в найгіршому сценарії дозволить більш впевнено зробити висновки на основі статистичного аналізу.

5 Перевірка на штучних наближених до реальних даних

У рамках проведення клінічних досліджень з ліками для лікування шизофренії, було сформовано штучні дані, які апроксимують реальні параметри. Особлива увага приділялася змінам у рівнях білірубіну, який є важливим показником функціонування печінки, і його динаміці в контексті застосування як ліків, так і плацебо. Збільшення рівня білірубіну в організмі пацієнтів є важливим моментом для визначення безпеки лікування, оскільки воно може свідчити про можливі побічні ефекти, пов'язані з дією ліків на функцію печінки. Таким чином, аналіз зміни цього показника, серед інших лабораторних даних, виступає як важливий індикатор безпеки та толерантності ліків пацієнтами з шизофренією.

Для подальшого аналізу даних було випадковим чином згенеровано пусті дані з ймовірністю 7%, що дозволило врахувати можливість втрати деяких спостережень у наборі даних. Після цього було побудовано графік (Малюнок 3), що відображає зміну середнього показника в залежності від часу. Аналізуючи цей графік, ми зауважили, що на останньому візиті спостерігається найбільша різниця у показниках між групами, які отримували плацебо та ліки від шизофренії. Це вказує на можливу значимість впливу лікування на динаміку показників і може вказувати на ефективність ліків у порівнянні з контрольною групою.



Малюнок 3. Середній показник білірубіну протягом дослідження

Після отримання графічних даних було обчислено показник chg9, який розраховується як різниця між значеннями показника на останньому візиті та значеннями на початковому візиті, останній який вважається базовим для порівнянь. Далі для аналізу використовувався метод дисперсійного аналізу (ANOVA), щоб визначити, чи існують статистично значущі відмінності між групами, які отримували плацебо та ліки від шизофренії. Такий підхід дозволив оцінити вплив лікування на динаміку показників та визначити, чи є ці різниці статистично значущими. Значення, що ми отримали дорівнює 0,7164.

TRT01AN	TRT01A	USUBJID	NAME OF FORMER VARIABLE	vis0	vis6	vis9	chg9
2	SCHZPIL 80MG BID	297662-1020	AVAL	6.84	.	.	.
2	SCHZPIL 80MG BID	297662-1037	AVAL	5.13	.	.	.
2	SCHZPIL 80MG BID	297662-1043	AVAL	3.42	3.42	5.13	1.71
2	SCHZPIL 80MG BID	297662-1060	AVAL	3.42	.	.	.
2	SCHZPIL 80MG BID	297662-1102	AVAL	6.84	.	.	.
2	SCHZPIL 80MG BID	297665-1030	AVAL	8.55	.	.	.
2	SCHZPIL 80MG BID	297676-1009	AVAL	3.42	.	.	.
2	SCHZPIL 80MG BID	297677-1017	AVAL	5.13	17.1	3.42	-1.71
2	SCHZPIL 80MG BID	297677-1049	AVAL	3.42	.	.	.
2	SCHZPIL 80MG BID	297677-1063	AVAL	6.84	5.13	6.84	0
2	SCHZPIL 80MG BID	297677-1079	AVAL	8.55	.	.	.
2	SCHZPIL 80MG BID	297678-1017	AVAL	8.55	.	8.55	0
2	SCHZPIL 80MG BID	297678-1042	AVAL	3.42	3.42	3.42	0
2	SCHZPIL 80MG BID	297678-1066	AVAL	5.13	3.42	5.13	0
2	SCHZPIL 80MG BID	297678-1068	AVAL	3.42	1.71	5.13	1.71
2	SCHZPIL 80MG BID	297678-1091	AVAL	8.55	8.55	3.42	-5.13
2	SCHZPIL 80MG BID	297678-1092	AVAL	5.13	5.13	.	.
2	SCHZPIL 80MG BID	297678-1116	AVAL	11.97	8.55	6.84	-5.13
2	SCHZPIL 80MG BID	297679-1002	AVAL	6.84	10.26	.	.
2	SCHZPIL 80MG BID	297679-1003	AVAL	6.84	5.13	3.42	-3.42
2	SCHZPIL 80MG BID	297679-1005	AVAL	10.26	.	.	.
2	SCHZPIL 80MG BID	297679-1009	AVAL	3.42	10.26	.	.
2	SCHZPIL 80MG BID	297679-1019	AVAL	8.55	.	.	.
2	SCHZPIL 80MG BID	297679-1052	AVAL	5.13	6.84	6.84	1.71
2	SCHZPIL 80MG BID	297679-1056	AVAL	3.42	3.42	3.42	0
2	SCHZPIL 80MG BID	297679-1058	AVAL	5.13	3.42	5.13	0
2	SCHZPIL 80MG BID	297679-1063	AVAL	1.71	3.42	1.71	0
2	SCHZPIL 80MG BID	297679-1069	AVAL	8.55	.	.	.
2	SCHZPIL 80MG BID	297679-1074	AVAL	6.84	13.68	.	.
2	SCHZPIL 80MG BID	297679-1105	AVAL	6.84	.	.	.
2	SCHZPIL 80MG BID	297679-1118	AVAL	8.55	10.26	5.13	-3.42
2	SCHZPIL 80MG BID	297679-1122	AVAL	3.42	3.42	1.71	-1.71
2	SCHZPIL 80MG BID	297679-1123	AVAL	1.71	5.13	1.7083	-0.0017
2	SCHZPIL 80MG BID	297680-1008	AVAL	5.13	.	.	.
2	SCHZPIL 80MG BID	297680-1070	AVAL	6.84	1.71	3.42	-3.42
2	SCHZPIL 80MG BID	297680-1089	AVAL	6.84	6.84	6.84	0

Малюнок 4. Вхідний набір даних, що має пусті значення

Після застосування методу множинної імпутації, який використовується для заповнення відсутніх значень у наборах даних, було створено п'ять додаткових наборів даних з відновленими пропущеними значеннями. Після цього кожен з цих додаткових наборів було об'єднано з вихідним набором даних, створюючи в результаті по шість наборів даних для кожної досліджуваної групи (плацебо та ліки). Далі, для забезпечення повноти аналізу, були створені комбіновані набори даних, де групи були скомбіновані одна з однією, що призвело до отримання усього 36 наборів даних (6 * 6) для кожної досліджуваної групи. Такий підхід дозволяє врахувати різноманітність можливих варіантів та вплив імпутації на результати аналізу даних.

Imputation Number	TRT01AN	TRT01A	USUBJID	vis0	vis6	vis9	chg9
1	2	SCHZPIL 80MG BID	297662-1020	6.84	10.815258511	4.8026548152	-2.037345185
1	2	SCHZPIL 80MG BID	297662-1037	5.13	2.6879205953	2.4235341881	-2.706465812
1	2	SCHZPIL 80MG BID	297662-1043	3.42	3.42	5.13	1.71
1	2	SCHZPIL 80MG BID	297662-1060	3.42	11.386385231	5.3544325164	1.9344325164
1	2	SCHZPIL 80MG BID	297662-1102	6.84	4.1949411104	6.6961256191	-0.143874381
1	2	SCHZPIL 80MG BID	297665-1030	8.55	8.9436155863	8.1680960235	-0.381903977
1	2	SCHZPIL 80MG BID	297676-1009	3.42	1.8472750575	6.2123233333	2.7923233333
1	2	SCHZPIL 80MG BID	297677-1017	5.13	17.1	3.42	-1.71
1	2	SCHZPIL 80MG BID	297677-1049	3.42	4.2702796112	2.7286102061	-0.691389794
1	2	SCHZPIL 80MG BID	297677-1063	6.84	5.13	6.84	0
1	2	SCHZPIL 80MG BID	297677-1079	8.55	5.4006057063	6.8363014363	-1.713698564
1	2	SCHZPIL 80MG BID	297678-1017	8.55	4.013184667	8.55	0
1	2	SCHZPIL 80MG BID	297678-1042	3.42	3.42	3.42	0
1	2	SCHZPIL 80MG BID	297678-1066	5.13	3.42	5.13	0
1	2	SCHZPIL 80MG BID	297678-1068	3.42	1.71	5.13	1.71
1	2	SCHZPIL 80MG BID	297678-1091	8.55	8.55	3.42	-5.13
1	2	SCHZPIL 80MG BID	297678-1092	5.13	5.13	6.6968222811	1.5668222811
1	2	SCHZPIL 80MG BID	297678-1116	11.97	8.55	6.84	-5.13
1	2	SCHZPIL 80MG BID	297679-1002	6.84	10.26	4.6279514702	-2.21204853
1	2	SCHZPIL 80MG BID	297679-1003	6.84	5.13	3.42	-3.42
1	2	SCHZPIL 80MG BID	297679-1005	10.26	8.624199001	5.9860120851	-4.273987915
1	2	SCHZPIL 80MG BID	297679-1009	3.42	10.26	9.0112781428	5.5912781428
1	2	SCHZPIL 80MG BID	297679-1019	8.55	3.9061719258	7.6216854887	-0.928314511
1	2	SCHZPIL 80MG BID	297679-1052	5.13	6.84	6.84	1.71
1	2	SCHZPIL 80MG BID	297679-1056	3.42	3.42	3.42	0
1	2	SCHZPIL 80MG BID	297679-1058	5.13	3.42	5.13	0
1	2	SCHZPIL 80MG BID	297679-1063	1.71	3.42	1.71	0
1	2	SCHZPIL 80MG BID	297679-1069	8.55	12.043112322	9.4232099424	0.8732099424
1	2	SCHZPIL 80MG BID	297679-1074	6.84	13.68	5.9956304877	-0.844369512
1	2	SCHZPIL 80MG BID	297679-1105	6.84	5.3144523993	10.175210253	3.3352102528
1	2	SCHZPIL 80MG BID	297679-1118	8.55	10.26	5.13	-3.42
1	2	SCHZPIL 80MG BID	297679-1122	3.42	3.42	1.71	-1.71

Малюнок 5. Приклад набору даних з імпутованими показниками

Після отримання 36 різних вибірок, кожна з яких була аналізована окремо за допомогою методу ANOVA, для кожного набору даних було збережено значення p-value. Тепер необхідно обрати серед отриманих 36 p-value те, яке є найбільшим, оскільки воно відповідає найменшій вірогідності відкидання нульової гіпотези. Таке значення може вказувати на найбільшу відмінність між групами та на ймовірність, що ця відмінність не є випадковою. Такий підхід дозволяє визначити найбільш вагомі різниці між групами та зосередитися на них у подальшому аналізі.

Pr > |t|	p_imp	d_imp
0.7164	0	0
0.6413	0	1
0.4426	0	2
0.8064	0	3
0.6464	0	4
0.6455	0	5
0.9655	1	0
0.9679	1	1
0.7233	1	2
0.8532	1	3
0.9765	1	4
0.9704	1	5
0.9327	2	0
0.9252	2	1
0.6688	2	2
0.8838	2	3
0.9341	2	4
0.9281	2	5
0.8263	3	0
0.7808	3	1
0.5647	3	2
0.9499	3	3
0.7878	3	4
0.7838	3	5
0.9461	4	0
0.9136	4	1
0.8390	4	2
0.7405	4	3
0.9045	4	4
0.9117	4	5
0.9641	5	0
0.9667	5	1
0.7147	5	2
0.8488	5	3
0.9756	5	4
0.9692	5	5

Малюнок 6. Отримані 36 p-value

У результаті аналізу отримано різноманітні значення p-value, де мінімальне значення становить 0,4426, а максимальне - 0,9765. Це свідчить про те, що між групами існує статистично значуща різниця, рівня якої складає 0,5339. Важливо зазначити, що пропущені показники під час дослідження можуть вплинути на остаточний висновок стосовно безпеки ліків.

У подальшому дослідженні доцільним виявляється уточнення параметрів методу множинної імпутації, які буде рекомендовано використовувати під час проведення реальних клінічних випробувань, враховуючи їх моделі. Особливу увагу слід приділити кількості імпутованих вибірок, які слід створити, а також розглянутий діапазон варіативності цих вибірок. Такий підхід дозволить уникнути спотворення результатів та забезпечить надійність та об'єктивність аналізу даних під час клінічних випробувань.

6 Висновки

У рамках цієї роботи було розглянуто метод множинної імпутації, що є одним із потужних інструментів для обробки пустих даних у клінічних випробуваннях. Метод множинної імпутації дозволяє заповнити пропущені значення у наборі даних, забезпечуючи цілісність та достовірність аналізу. Особливу увагу було приділено такій стратегії множинної імпутації як лінійній регресії.

Після того, як було розглянуто методи імпутації, увага була спрямована на статистичний аналіз у клінічних випробуваннях. Були детально описані метод Каплан-Майера для аналізу виживання та різні варіанти Т-тесту (одновибірковий, незалежний та парний), які широко використовуються для порівняння груп та виявлення різниць між ними. Крім того, розглянуто метод ANOVA, який дозволяє аналізувати різниці між трьома або більше групами.

Далі в роботі показано, як метод множинної імпутації може бути використаний для створення "найгіршого" набору даних. Це дозволяє врахувати різноманітність можливих впливів пропущених даних на результати аналізу та забезпечити надійність висновків.

На заключному етапі була проведена перевірка розроблених методів на реальних або штучних, але наближених до реальних даних, що показало їхню застосовність у практичних клінічних дослідженнях.

7 Бібліографія

1. Rubin D. B. Multiple Imputation for Nonresponse in Surveys [Electronic source] / D. B. Rubin. ─ New York [at al.] : John Wiley & Sons, 1987. ─ 258 p. ─ DOI:10.1002/9780470316696.
2. Little R. J. A. Statistical Analysis with Missing Data / R. J. A. Little, D. B. Rubin. ─ Second Edition. ─ [S. l.] : John Wiley & Sons, 2002. ─ 389 p. ─ DOI:10.1002/9781119013563.
3. Enders C. K. Applied Missing Data Analysis [Electronic source] / C. K. Enders ; Series Editor's Note by T. D. Little. ─ New York ; London : THE GUILFORD PRESS, 2010. ─ 401 p. ─ URL: http://hsta559s12.pbworks.com/w/file/fetch/52112520/enders.applied.
4. Montgomery D. C. Introduction to Linear Regression Analysis [Electronic source] / D. C. Montgomery, E. A. Peck, G. G. Vining. ─Sixth Edition. ─ [S. l.] : John Wiley & Sons, 2021. ─ 704 p. ─ URL: https://www.wiley.com/en-br/Introduction+to+Linear+Regression+Analysis%2C+6th+Edition-p-9781119578758.
5. Rencher A. C. Linear Models in Statistics [Electronic source] / A. C. Rencher, G. B. Schaalje. ─ Second Edition. ─ [S. l.] : John Wiley & Sons, 2008. ─ 672 p. ─ DOI:10.1002/9780470192610.
6. Quinlan, J. R. Induction of Decision Trees [Electronic source] / J. R. Quinlan // Machine Learning. ─ 1986. ─ Volume 1, No 1. ─ P. 81─106. ─ DOI: https://doi.org/10.1007/BF00116251.
7. Duolao Wang. Clinical Trials : A Practical Guide to Design, Analysis, and Reporting / Duolao Wang, Ameet Bakhai. – London : Remedica, 2006. – 498 p.
8. Pocock S. J. Clinical Trials: A Practical Approach / S. J. Pocock. – First Edition. – [S. l.] : John Wiley & Sons, 1984. – 266 p.
9. Kaplan, E. L. Nonparametric Estimation from Incomplete Observations [Electronic source] / E. L. Kaplan, P. Meier // Journal of the American

Statistical Association. ─ 1958. ─ Volume 53, No. 282. ─ P. 457─481. ─ URL: https://www.jstor.org/stable/2281868.

10. Allison, P. D. Survival Analysis Using SAS : A Practical Guide / P. D. Allison. ─ Second Edition. ─ [S. l.] : SAS Institute, 2010. – 336 p.
11. Student. The Probable Error of a Mean [Electronic source] / Student // Biometrika. – 1908. ─ Volume 6, No. 1. – P. 1─25. – DOI: https://doi.org/10.2307/2331554.
12. Rosner, B. Fundamentals of Biostatistics / B. Rosner. ─ Eighth Edition. ─ [S. l.] : Cengage Learning, 2015. – 888 p.
13. Fisher R. A. The Design of Experiments [Electronic source] / R. A. Fisher. ─ New York : Hafner press ; London : Collier Macmillan Publishers, 1935. – 252 p. – URL: https://home.iitk.ac.in/~shalab/anova/DOE-RAF.pdf.
14. Hinkelmann, K. Design and Analysis of Experiments. Volume 2: Advanced Experimental Design [Electronic source] / K. Hinkelmann, O. Kempthorne. – [S. l.] : John Wiley & Sons, 2008. – 808 p. – URL: https://web.archive.org/web/20170809115004id_/http://pustaka.unp.ac.id/file/abstrak_kki/EBOOKS/Advanced%20Experimental%20Design..pdf.

ДОДАТОК 1

Поточні методи заповнення пустих даних в клінічних випробуваннях

Average	техніка імпутації, яка встановлює значення аналізу в записі на середнє значення суб'єкта за визначеним набором записів.
Best Case	техніка імпутації, яка встановлює значення аналізу в записі на найкращий можливий результат.
Baseline Observation Carried Forward	техніка імпутації, яка встановлює значення аналізу в записі на базове (початкове) спостереження суб'єкта без відсутності даних.
Best Observed Case	техніка імпутації, яка встановлює значення аналізу в записі на найкращий результат, зафіксований у суб'єкта.
Best Observation Carried Forward	техніка імпутації, яка встановлює значення аналізу в записі на попереднє найкраще значення суб'єкта без відсутності даних.
Copy	техніка, яка використовує логіку призначення для дублювання значення аналізу для подальших оцінок.
Extrapolation	техніка імпутації, яка встановлює значення аналізу в записі на оцінене або припущене значення на основі розширення відомої послідовності значень. Ця категорія також включає умовно розділені питання, де фактична відповідь імпутується.
One Half of Lower Limit of Quantification	техніка імпутації, яка встановлює значення аналізу в записі на половину нижнього обмеження кількісного тесту.
Interpolation	техніка імпутації, яка встановлює значення аналізу в записі на функцію відомих значень для оцінки значення в межах діапазону.
Lower Limit of Detection	техніка імпутації, яка встановлює значення аналізу в записі на нижнє обмеження виявлення тесту.

Lower Limit of Quantification	техніка імпутації, яка встановлює значення аналізу в записі на нижнє обмеження кількісного тесту.
Last Observation Carried Forward	техніка імпутації, яка встановлює значення аналізу в записі на попереднє невідсутнє значення суб'єкта.
Last Observed Value	техніка імпутації, яка встановлює значення аналізу в записі на останнє зареєстроване невідсутнє спостереження суб'єкта.
Last Value Prior to Dosing	техніка імпутації, яка встановлює значення аналізу в записі на останнє спостережене значення суб'єкта до початку лікування в дослідженні.
Maximum	техніка імпутації, яка встановлює значення аналізу в записі на максимальне значення суб'єкта за визначеним набором записів.
Minimum	техніка імпутації, яка встановлює значення аналізу в записі на мінімальне значення суб'єкта за визначеним набором записів.
Maximum Likelihood	техніка імпутації, яка встановлює значення аналізу в записі на оцінку, яка максимізує ймовірність спостереження фактично зафіксованого.
Mean of Other Group	техніка імпутації, яка встановлює значення аналізу в записі на середнє значення з референтної групи.
Mean Observed Value in a Group	техніка імпутації, яка встановлює значення аналізу в записі на середнє значення, спостережене в групі суб'єктів.
Phantom Record	техніка, яка створює запис з відсутнім значенням аналізу, коли відсутній запис для даного візиту аналізу або часової точки аналізу.
Penultimate Observation Carried Forward	техніка імпутації, яка встановлює значення аналізу в записі на передостаннє невідсутнє значення суб'єкта.

Screening Observation Carried Forward	техніка імпутації, яка встановлює значення аналізу в записі на невідсутнє спостереження при скринінгу суб'єкта.
Upper Limit of Detection	техніка імпутації, яка встановлює значення аналізу в записі на верхнє обмеження виявлення тесту.
Upper Limit of Quantification	техніка імпутації, яка встановлює значення аналізу в записі на верхнє обмеження кількісного тесту.
Worst Case	техніка імпутації, яка встановлює значення аналізу в записі на найгірший можливий результат.
Worst Observed Case	техніка імпутації, яка встановлює значення аналізу в записі на найгірший зафіксований результат у суб'єкта.
Worst Observed Value Carried Forward	техніка імпутації, яка встановлює значення аналізу в записі на найгірше спостереження суб'єкта без відсутності даних.
Worst Observed Value in a Group	техніка імпутації, яка встановлює значення аналізу в записі на найгірше значення, спостережене в групі суб'єктів.