

УДК 519.2

DOI: <https://doi.org/10.17721/1812-5409.2026/1.25>

Олександр ГАЛУШКО, асп.
ORCID ID: 0009-0009-0275-6542
e-mail: health24@ukr.net

Київський національний університет імені Тараса Шевченка, Київ, Україна

ІНТЕГРОВАНА МОДЕЛЬ ФАКТОРНОГО ВІДБОРУ ТА КАТЕГОРІАЛЬНОЇ РЕГРЕСІЇ ДЛЯ ПРОГНОЗУВАННЯ РІВНЯ РИЗИКУ ІМПЛАНТАЦІЇ ДІАГНОСТИЧНО-ТЕРАПЕВТИЧНИХ ПРИСТРОЇВ ПРИ ПАТОЛОГІЇ ШЛУНКОВО-СТРАВОХІДНОГО З'ЄДНАННЯ

Представлено інтегрований метод прогнозування рівня ризику імплантації діагностично-терапевтичних пристроїв у ділянці шлунково-стравохідного з'єднання на основі факторного аналізу та категоріальної регресії. На вибірці із 558 пацієнтів було виділено 15 клінічних та ендоскопічних показників, що характеризують стан слизової, анатомічні особливості та симптоматику рефлюксу. Факторний аналіз методом головних осей з обертанням Oblimin дає можливість звести початковий набір змінних до трьох латентних факторів: F_1 – "Клінічні та морфологічні ознаки гастроезофагеальної рефлюксної хвороби", F_2 – "Анатомо-функціональні показники стравоходу" та F_3 – "Ерозивні ураження слизової". Ці фактори були використані як предиктори в категоріальній регресії для прогнозування п'яти рівнів ризику імплантації. Модель продемонструвала високу якість прогнозу ($R^2 = 0,858$), усі фактори були статистично значущими ($p < 0,001$), а показники мультиколінеарності предикторів $> 0,8$ підтвердили відсутність критичної мультиколінеарності. Отримана модель забезпечує стандартизовану стратифікацію ризику, допомагаючи клініцистам об'єктивніше оцінювати доцільність імплантації та визначати пацієнтів, які потребують додаткової оптимізації стану перед процедурою.

Ключові слова: імплантація пристроїв, факторний аналіз, категоріальна регресія, ризик, шлунково-стравохідне з'єднання, стратифікація ризику.

Класифікація відповідно до AMS 2020: 62J12, 62P10.

Вступ

Імплантацію діагностичних і терапевтичних пристроїв у ділянці шлунково-стравохідного з'єднання (зокрема систем бездротового рН-моніторингу, стимуляторів нижнього стравохідного сфінктера, капсульних сенсорів) дедалі частіше застосовують у клінічній практиці для діагностики та лікування гастроезофагеальної рефлюксної хвороби (ГЕРХ) та пов'язаних із нею патологій. Водночас ці втручання пов'язані з ризиком розвитку ускладнень, що зумовлює необхідність попередньої об'єктивної оцінки локального стану слизової оболонки, анатомо-функціональних особливостей стравоходу та клінічної симптоматики пацієнта.

На практиці рішення щодо доцільності та безпечності імплантації часто ґрунтується на експертній оцінці лікаря з урахуванням великої кількості клінічних та ендоскопічних показників різного типу (бінарних, порядкових і кількісних). Це ускладнює стандартизацію підходу, підвищує суб'єктивність рішень й обмежує можливості формалізованої стратифікації ризику. Крім того, велика кількість взаємопов'язаних змінних і наявність рідкісних категорій створюють проблеми для класичних багатовимірних моделей через мультиколінеарність і нестійкість оцінок.

У зв'язку із цим актуальним є розроблення математично обґрунтованих моделей, що дають змогу інтегрувати різномірні клінічні показники, зменшувати розмірність даних без втрати інформації та забезпечувати об'єктивне прогнозування рівня ризику імплантації.

Латентні фактори у медичних дослідженнях застосовують для виявлення прихованих (небезпосередньо невимірюваних) структур у складних переліках клінічних, лабораторних або анкетних даних. Вони допомагають зменшити розмірність даних й об'єднати кілька взаємопов'язаних показників у спільні концептуальні компоненти (напр., "метаболічний синдром", "психоемоційний стан", "анатомо-функціональні порушення"). Подальше використання цих факторів у регресійному аналізі дає змогу будувати прогностичні моделі, які є більш стабільними, інтерпретованими та менш чутливими до мультиколінеарності між первинними змінними. Такий підхід широко застосовують для: оцінки ризику розвитку захворювань (Smith et al., 2017); побудови моделей латентних класів або підтипів пацієнтів (Guo, Wall, & Amemiya, 2006); аналізу психометричних і клінічних шкал (Viana, Gonçalves, & Paez, 2024); розробки методів факторної регресії для інтеграції змінних і латентних показників у єдину модель (Devlieger, 2019) та ін.

Завдяки цьому факторно-регресійний підхід вважають ефективним інструментом для побудови персоналізованих прогностичних систем у клінічній медицині.

Об'єктом пропонованого дослідження є клінічні та ендоскопічні характеристики пацієнтів із патологією шлунково-стравохідного з'єднання, що використовуються для оцінки ризику імплантації діагностичних і терапевтичних пристроїв.

Мета роботи – розроблення й обґрунтування інтегрованої математичної моделі прогнозування рівня ризику імплантації пристроїв у ділянці шлунково-стравохідного з'єднання на основі поєднання факторного аналізу та категоріальної регресії.

Для досягнення цієї мети поставлено такі завдання:

- відібрати релевантні клінічні й ендоскопічні показники, що характеризують стан пацієнтів;
- здійснити факторний аналіз для зменшення розмірності та виділення латентних клінічно інтерпретованих факторів;
- побудувати модель категоріальної регресії з оптимальним шкалюванням для прогнозування порядкової змінної "рівень ризику";
- оцінити якість, стабільність та інтерпретованість отриманої моделі.

У дослідженні спочатку виконано факторний аналіз за метод головних осей із обертанням Oblimin для виділення клінічно інтерпретованих латентних факторів із 15 ендоскопічних і клінічних показників, а потім отримані фактори постають як предиктори в категоріальній регресії для прогнозування п'яти рівнів ризику імплантації.

1. Формування вибірки, змінних і шкали ризику

У представленій статті побудовано математичну модель, що інтегрує клінічні й ендоскопічні дані для прогнозування рівня ризику до імплантації діагностичних та/або терапевтичних пристроїв, зокрема систем Bravo, JSPH-1, BEST Capsule, EndoStim для стимуляції нижнього стравохідного сфінктера, капсульних датчиків Smart nanowire та інших (Lawenko, & Lee, 2016; Li, Liu, & Tao, 2013), у пацієнтів із патологією шлунково-стравохідного з'єднання, з метою підвищення точності й об'єктивності клінічного ухвалення рішень.

Показники для моделі: гіперемія слизової в субкардіальній ділянці (бінарна змінна: 1 – наявна; 0 – відсутня) – X_1 ; скарги на печію (бінарна змінна: 1 – наявна; 0 – відсутня) – X_2 ; скарги на відчуття регургітації (бінарна змінна: 1 – наявна; 0 – відсутня) – X_3 ; наявність дефектів слизової класифікованих за Лос-Анджелеською класифікацією оцінки рефлюкс-езофагіту (порядкова змінна: дефекти відсутні – 0; тип А – 1; тип В – 2; тип С – 3; тип D – 4) – X_4 ; наявність ерозій, покритих фібрином (бінарна змінна: 1 – наявні; 0 – відсутні) – X_5 ; наявність ерозій, покритих гематином (бінарна змінна: 1 – наявні; 0 – відсутні) – X_6 ; наявність виразок, покритих фібрином (бінарна змінна: 1 – наявні; 0 – відсутні) – X_7 ; наявність виразок, покритих гематином (бінарна змінна: 1 – наявні; 0 – відсутні) – X_8 ; показник розташування Z-лінії (кількісна змінна, см) – X_9 ; висота верхнього стравохідного сфінктера (кількісна змінна, см) – X_{10} ; показник діафрагмального звуження (кількісна змінна, см) – X_{11} ; показник висоти пролабування шлунка у стравохід або кардіальну ділянку шлунка (кількісна змінна, см) – X_{12} ; довжина пролабування (кількісна змінна, см) – X_{13} ; гіперемія в нижній третині стравоходу (бінарна змінна: 1 – наявна; 0 – відсутня) – X_{14} ; куріння (бінарна змінна: 1 – наявне; 0 – відсутнє) – X_{15} .

На основі вибірки із 558 пацієнтів ендоскопічного відділення комунального некомерційного підприємства "Кіровоградська обласна лікарня Кіровоградської обласної ради" та їхніх клінічних характеристик (X_1 – X_{15}) лікарі змогли встановити рівні ризику для лікувальної імплантації. Було виділено 5 рівнів ризику. Проаналізуємо основні характеристики для градації на групи за рівнями ризику.

0 (відсутній ризик). Ймовірність ускладнень мінімальна. Наслідки незначні та легко контролювані. Процедура імплантації вважається безпечною. Базові критерії: може бути гіперемія в нижній третині стравоходу або гіперемія в субкардіальній ділянці шлунка; довжина пролабування менше 2 см; відсутність дефектів слизової.

1 (низький ризик). Ймовірність ускладнень мінімальна. Наслідки незначні, добре контролювані. Процедура імплантації безпечна, потрібне лише стандартне спостереження. Базові критерії: може бути гіперемія в нижній третині стравоходу і/або гіперемія в субкардіальній ділянці шлунка; довжина пролабування не перевищує 2 см; можуть бути дефекти слизової у вигляді ерозій, покритих фібрином.

2 (середній ризик). Ймовірність ускладнень помірна. Наслідки контролювані, однак рекомендоване дотримання дієти, режиму харчування та посилене клінічне спостереження. Базові критерії: може бути гіперемія в нижній третині стравоходу і/або гіперемія в субкардіальній ділянці шлунка; довжина пролабування не перевищує 3 см; можуть бути дефекти слизової у вигляді ерозій та/або виразок, покритих фібрином.

3 (високий ризик). Ймовірність ускладнень підвищена. Наслідки клінічно значущі, потребують посиленого медичного нагляду. Процедура імплантації допускається лише після ретельної оцінки. Базові критерії: довжина пролабування не перевищує 3 см; можуть бути дефекти слизової у вигляді ерозій та/або виразок, покритих гематином. Комбінації "гематин" піднімають ризик до високого незалежно від гіперемії чи набряку.

4 (критичний ризик). Ймовірність ускладнень висока. Наслідки можуть бути небезпечними для життя. Процедура імплантації не рекомендована, поки ризики не будуть усунені або зменшені. Базові критерії: довжина пролабування більше 3 см; можуть бути дефекти слизової у вигляді ерозій та/або виразок, покритих фібрином і гематином.

Зауважимо, що тут наведені ключові маркери щодо градації за групами ризику. Лікарі враховували весь спектр зазначених показників X_1 – X_{15} .

2. Виділення латентних факторів методом головних осей

Оскільки кількість змінних дуже велика, а також деяка частина змінних містить поодинокі значення (є проблема розрізнення даних), то для знаходження моделі, що буде відображати рівні ризику, проведемо факторний аналіз змінних X_1 – X_{15} і визначимо латентні фактори, які пояснюють основні закономірності в даних і зменшують кількість змінних, залишаючи максимально можливу інформацію.

Отже, для спостережуваних p змінних X_1 – X_{15} припускаємо, що існує m ($m < p$) латентних факторів $F_1, F_2, \dots, F_m$, такі, що кожна змінна може бути представлена як:

$$X_j = \lambda_{j1}F_1 + \lambda_{j2}F_2 + \dots + \lambda_{jm}F_m + \varepsilon_j, \quad j = 1, \dots, 15, \quad (1)$$

де λ_{jk} – факторні навантаження, що показують, наскільки X_j пов'язана з фактором F_j ; ε_j – унікальна частина змінної X_j , яку фактори не пояснюють.

Будемо використовувати метод головних осей (Principal Axis Factoring, PAF), що не вимагає припущення про нормальний розподіл даних (Finch, & French, 2019; Yong, & Pearce, 2013). Це робить його привабливим вибором, коли дані не відповідають умовам нормальності. Порівняно з методом максимальної правдоподібності, що потребує нормальності даних, PAF є більш гнучким і стійким до порушень цієї умови.

Першим кроком зазначеного методу є стандартизація змінних через середні значення та стандартні відхилення змінних

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j}, \quad i = 1, \dots, 558, \quad j = 1, \dots, 15,$$

де Z_{ij} – стандартизоване значення змінної j для i -го пацієнта; $\bar{X}_j$ – середнє значення змінної X_j ; s_j – стандартне відхилення змінної X_j .

Це забезпечує, що всі змінні мають $E(Z_j) = 0$, $Var(Z_j) = 1$.

Далі будемо матрицю кореляцій $R = [r_{jk}]_{p \times p}$, елементи якої визначаємо так:

$$r_{jk} = \frac{\text{Cov}(X_j, X_k)}{s_j s_k}, \quad j, k = 1, \dots, 15.$$

Для бінарних змінних коефіцієнт кореляції Пірсона чисельно збігається з коефіцієнтом контингенції ϕ , який використовуємо для оцінювання зв'язку між бінарними змінними (Akoglu, 2018), тому застосування стандартної матриці кореляцій у SPSS (Statistical Package for the Social Sciences – статистичне програмне забезпечення для соціальних наук) коректне навіть за змішаних типів змінних.

У матричній формі основну факторну модель для всієї вибірки записуємо як:

$$Z = F\Lambda' + E,$$

де Z – матриця стандартизованих змінних розмірності $n \times p$; F – матриця факторних балів (латентних факторів) розмірності $n \times m$; Λ' – транспонована матриця факторних навантажень розмірності $m \times p$; E – матриця унікальних залишків розмірності $n \times p$.

Для кожного окремого спостереження (пацієнта) це рівняння має вигляд:

$$z_i = \Lambda f_i + \varepsilon_i, \quad i = 1, \dots, n,$$

де z_i – вектор стандартизованих змінних для i -го пацієнта; f_i – вектор значень латентних факторів; ε_i – вектор залишків похибок.

PAF розраховує такі факторні навантаження $\Lambda = [\lambda_{jk}]$, щоб коваріаційна структура змінних задовольняла рівність:

$$R = \Lambda\Lambda' + \Psi,$$

де $\Lambda\Lambda'$ – дисперсія, пояснена латентними факторами; $\Psi = \text{diag}(\psi_1, \dots, \psi_p)$ – діагональна матриця унікальних дисперсій, тобто залишків, що не пояснюються факторами.

Для кожної змінної X_j обчислюємо загальну дисперсію змінної, пояснену факторами h_j^2 , тобто частку дисперсії змінної, поясненої виділеними факторами:

$$h_j^2 = \sum_{k=1}^m \lambda_{jk}^2.$$

У пакеті SPSS розраховуємо Initial (початкові) і Extraction (виділені) h_j^2 . Високі значення h_j^2 ($h_j^2 \rightarrow 1$) означають, що фактори добре пояснюють змінну X_j .

Метод PAF відшукує такі факторні навантаження Λ , щоб звести до мінімуму різницю між емпіричною матрицею кореляцій та її факторною апроксимацією ($\Lambda\Lambda' + \Psi$). На практиці це досягається ітераційно. Початкові значення ψ_j задають як $1 - h_j^2$. З матриці $R - \Psi$ обчислюють власні значення та власні вектори. Формуються нові навантаження Λ із власних векторів, що відповідають найбільшим власним значенням. Процес повторюється доти, доки зміни h_j^2 стануть незначними.

Для полегшення інтерпретації результатів скористаємося методом обертання Oblimin із нормалізацією Кайзера. Цей метод дає факторам можливість бути корельованими, що важливо для клінічних ознак, які можуть бути взаємопов'язані. Також факторні навантаження після обертання легше інтерпретувати, вони чітко показують, які змінні входять до кожного латентного фактору.

Після обертання

$$\Lambda^* = \Lambda T,$$

де T – матриця обертання, що спрощує структуру факторних навантажень таким чином, що кожна змінна має суттєве навантаження на один фактор.

Критерієм придатності даних для факторного аналізу є тест Кайзера – Мейєра – Олкіна (KMO), що вимірює пропорцію дисперсії змінних, яку можна пояснити спільними факторами. KMO обчислюємо за формулою:

$$KMO = \frac{\sum_{j \neq k} r_{jk}^2}{\sum_{j \neq k} r_{jk}^2 + \sum_{j \neq k} p_{jk}^2}, \quad p_{jk} = -\frac{r^{jk}}{\sqrt{r^{jj} \cdot r^{kk}}},$$

де r^{jk} – елемент оберненої матриці кореляцій R^{-1} ; r^{jj} , r^{kk} – діагональні елементи R^{-1} , що відповідають змінним X_j , X_k .

Значення KMO близьке до 1 вказує на високу придатність даних для факторного аналізу.

Тест сферичності Бартлетта перевіряє гіпотезу про те, що кореляційна матриця змінних є одиничною, тобто змінні не корельовані між собою. Тест є важливим перед застосуванням факторного аналізу, оскільки значна кореляція між змінними – необхідна умова для виділення латентних факторів. Статистику критерію Бартлетта розраховуємо так:

$$\chi^2 = -\left(n - 1 - \frac{2p + 5}{6}\right) \ln(\det(R)).$$

Якщо в SPSS p -значення тесту Бартлетта менше обраного рівня значущості, то гіпотезу про одиничність кореляційної матриці відхиляють, що свідчить про наявність значущих кореляцій між ними.

Проаналізуємо детально результати факторного аналізу для нашої моделі. З табл. 1 видно, що оптимальна кількість факторів дорівнює 3. Згідно з критерієм Кайзера, у моделі залишають лише ті фактори, власні значення яких перевищують 1 (Statistics Solutions, n. d.).

З табл. 1 видно, що перші три фактори мають власні значення: $\lambda_1 = 6,143$; $\lambda_2 = 2,862$; $\lambda_3 = 1,516$, що разом пояснюють приблизно 70,1 % загальної дисперсії вихідних змінних до обертання. Після обертання Oblimin із нормалізацією Кайзера сумарна пояснена дисперсія становить 64,6 %, що є прийнятним результатом для медико-біологічних досліджень, де змінні мають складну багатовимірну природу.

Вибір саме трьох факторів додатково підтверджується графіком кам'янистого осипу (scree plot) (рис. 1), на якому чітко спостерігаємо "перелом" між третім і четвертим факторами. Після третього фактору власні значення різко зменшуються та наближаються до нуля, що свідчить про незначний внесок наступних факторів у пояснення дисперсії. Отже, рішення про виокремлення трьох латентних факторів є теоретично та статистично обґрунтованим.

Таблиця 1

Пояснена дисперсія факторів за результатами факторного аналізу методом головних осей

Фактор	Початкові власні числа			Після виділення: сума квадратів навантажень			Після обертання: сума квадратів навантажень*
	Загальне	% дисперсії	Кумулятивно, %	Загальне	% дисперсії	Кумулятивно, %	Загальне
1	6,143	40,952	40,952	5,864	39,093	39,093	5,390
2	2,862	19,079	60,030	2,629	17,527	56,619	2,833
3	1,516	10,106	70,136	1,196	7,972	64,592	2,663
4	0,900	6,001	76,138				
5	0,760	5,064	81,202				
6	0,711	4,741	85,944				
7	0,641	4,274	90,218				
8	0,427	2,849	93,067				
9	0,338	2,253	95,320				
10	0,240	1,602	96,921				
11	0,183	1,221	98,142				
12	0,129	0,860	99,003				
13	0,077	0,516	99,519				
14	0,067	0,449	99,969				
15	0,005	0,031	100,000				

Примітки: 1. **Метод виділення** – метод головних осей (РАФ). 2. Позначка (*) означає, що коли фактори корельовані, то суми квадратів навантажень не можна додавати для отримання загальної дисперсії.

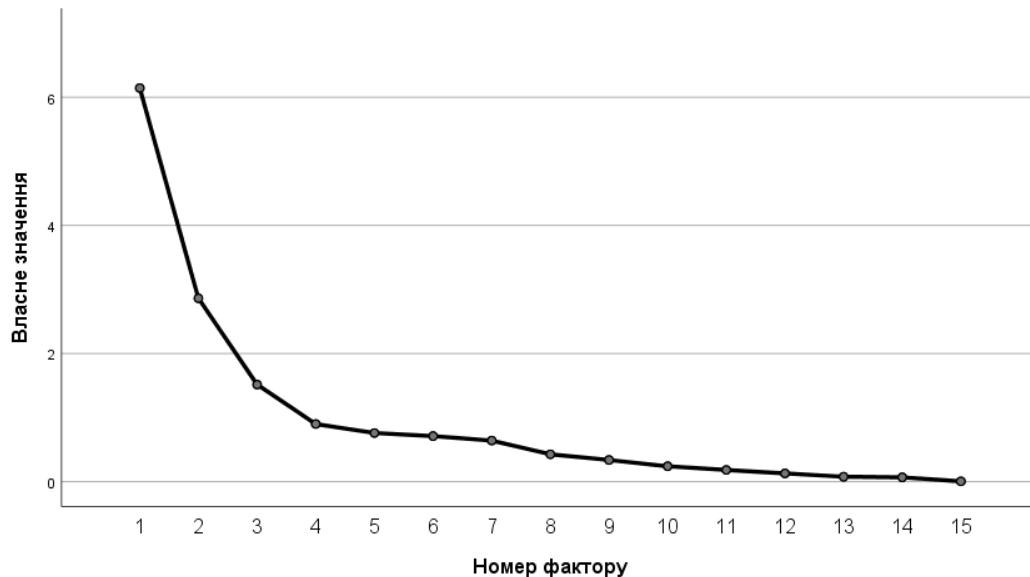


Рис. 1. Графік кам'янистого осипу (scree plot)

На основі кореляційної матриці та таблиці h_j^2 (табл. 2) можемо здійснити клінічну інтерпретацію важливості початкових змінних X_1-X_{15} на визначення факторних змінних F_1-F_3 .

Таблиця 2

Загальні дисперсії змінних, пояснені факторами, та клінічна інтерпретація змінних у факторному аналізі

Предиктори	Початкові h_j^2	Виділені h_j^2	Клінічна інтерпретація	Рівень впливу
X_1	0,732	0,454	Локальне запалення, пов'язане з анатомічними змінами (висота сфінктера, довжина пролабування)	Середній
X_2	0,712	0,678	Класичний симптом ГЕРХ, корелює з дефектами слизової	Високий
X_3	0,787	0,649	Ключовий симптом ГЕРХ, пов'язаний із тяжкими ураженнями	Високий
X_4	0,910	0,922	Ключовий ендоскопічний маркер, корелює із симптомами	Високий
X_5	0,763	0,915	Активне ураження, пов'язане із симптомами, анатомією та курінням	Високий
X_6	0,175	0,083	Рідкісне ураження, клінічно малозначуще	Низький
X_7	0,715	0,639	Тяжкі ураження, пов'язані із симптомами та структурними змінами	Високий
X_8	0,255	0,199	Малоспецифічне, рідкісне ураження	Низький
X_9	0,795	0,711	Анатомічний маркер ГЕРХ, відображає тяжкість рефлюксу	Високий
X_{10}	0,878	0,897	Механічний бар'єр проти рефлюксу, пов'язаний із довжиною пролабування	Високий
X_{11}	0,987	0,858	Основний анатомічний маркер тяжкого рефлюксу	Високий
X_{12}	0,991	0,977	Маркер тяжкого рефлюксу, пов'язаний із діафрагмальним звуженням	Високий
X_{13}	0,966	0,623	Механічне значення, вторинний маркер	Високий
X_{14}	0,533	0,519	Показник запалення, асоційований із симптомами і структурою	Середній
X_{15}	0,608	0,564	Фактор ризику, посилює симптоми й ураження слизової	Середній

Згідно з нашим результатом, значення $KMO = 0,740$ (табл. 3). Це свідчить про значне перекриття інформації між змінними та наявність середньої часткової кореляції. Отже, проведення факторного аналізу є цілком обґрунтованим. Іншими словами, вибірка і структура даних дають змогу виокремити латентні фактори. За тестом сферичності Бартлетта $p = 0,000 < 0,05$, тож гіпотезу про одиничну кореляційну матрицю відкидаємо. Це означає, що змінні взаємопов'язані, і факторний аналіз доцільний.

Таблиця 3

Критерії придатності даних для факторного аналізу (KMO та тест сферичності Бартлетта)		
Статистика критерію KMO		0,740
Тест сферичності Бартлетта	Хі-квадрат статистика	8597,565
	Ступені свободи	105
	Рівень значущості (p)	0,000

На основі таблиці факторних навантажень для 15 клінічних змінних методом головних осей з обертанням Oblimin виділимо латентні фактори та надамо їх змістовну інтерпретацію (в дужках зазначатимемо факторні навантаження клінічних показників).

Фактор 1 – F_1 "Клінічні та морфологічні ознаки ГЕРХ" охоплює такі показники: показник діафрагмального звуження (0,830); наявність виразок, покритих фібрином (0,780); показник розташування Z-лінії (0,748); скарги на відчуття регургітації (0,730); наявність дефектів слизової за Лос-Анджелеською класифікацією (0,729); куріння (0,598); скарги на печію (0,592); гіперемія в нижній третині стравоходу (0,572); наявність виразок, покритих гематином (0,438).

Фактор 2 – F_2 "Анатомо-функціональні показники стравоходу" охоплює такі показники: висота верхнього стравохідного сфінктера (– 0,870); показник довжини пролабування (– 0,783); показник висоти пролабування шлунка у стравохід або кардіальну ділянку (0,698); наявність гіперемії слизової в субкардіальній ділянці (0,635).

Фактор 3 – F_3 "Ерозивні ураження слизової" охоплює такі показники: наявність ерозій, покритих фібрином (0,973); наявність ерозій, покритих гематином (0,256).

Після обертання розраховуємо факторні бали для кожного спостереження. У нашому випадку це буде:

$$f_{i1}, f_{i2}, f_{i3}, \quad i = 1, \dots, 558,$$

де f_{i1} – факторний бал для фактора 1 "Клінічні та морфологічні ознаки ГЕРХ"; f_{i2} – факторний бал для фактора 2 "Анатомо-функціональні показники стравоходу"; f_{i3} – факторний бал для фактора 3 "Ерозивні ураження слизової".

Факторні бали f_{ik} для кожного спостереження i на факторі k розраховуємо як зважену комбінацію стандартизованих змінних:

$$f_{ik} = \omega_{1k}Z_{i1} + \omega_{2k}Z_{i2} + \dots + \omega_{pk}Z_{ip},$$

де ω_{jk} – коефіцієнт ваги для змінної Z_j на факторі F_k . Вагу в пакеті SPSS обчислюємо методом регресії, де факторні бали – це найкраща оцінка факторів за мінімумом середньоквадратичної похибки.

Далі обчислюємо кореляційну матрицю факторних балів:

$$r_{F_j F_k} = \frac{Cov(F_j, F_k)}{\sigma_{F_j} \sigma_{F_k}}, \quad j, k = 1, 2, 3,$$

де $Cov(F_j, F_k)$ – коваріація між факторами F_j та F_k ; σ_{F_j} , σ_{F_k} – стандартні відхилення факторних балів відповідно.

У табл. 4 наведено матрицю відповідних факторних кореляцій.

Таблиця 4

Матриця факторних кореляцій			
Фактор	1	2	3
1	1,000	– 0,149	0,288
2	– 0,149	1,000	– 0,001
3	0,288	– 0,001	1,000

Позадіагональні елементи в табл. 4 показують взаємну кореляцію між факторами: F_1 і F_2 ($r = -0,149$) – слабка негативна кореляція; майже незалежні; F_1 і F_3 ($r = 0,288$) – слабка позитивна кореляція, деяке перетинання ознак; F_2 і F_3 ($r = -0,001$) – практично відсутня кореляція, фактори незалежні.

Кінцева форма факторної моделі (1), що показує, як кожна змінна розкладається на вплив факторів плюс залишкову частину, має вигляд:

$$X_1 = -0,130 \cdot F_1 + 0,635 \cdot F_2 + 0 \cdot F_3 + \varepsilon_1;$$

$$X_3 = 0,730 \cdot F_1 - 0,217 \cdot F_2 + 0 \cdot F_3 + \varepsilon_3;$$

$$X_5 = 0 \cdot F_1 + 0 \cdot F_2 + 0,973 \cdot F_3 + \varepsilon_5;$$

$$X_7 = 0,780 \cdot F_1 - 0,153 \cdot F_2 - 0,373 \cdot F_3 + \varepsilon_7;$$

$$X_9 = 0,748 \cdot F_1 + 0,303 \cdot F_2 + 0,201 \cdot F_3 + \varepsilon_9;$$

$$X_{11} = 0,830 \cdot F_1 + 0,359 \cdot F_2 + 0,192 \cdot F_3 + \varepsilon_{11};$$

$$X_{13} = 0 \cdot F_1 - 0,783 \cdot F_2 + 0 \cdot F_3 + \varepsilon_{13};$$

$$X_{15} = 0,598 \cdot F_1 + 0 \cdot F_2 + 0,301 \cdot F_3 + \varepsilon_{15}.$$

$$X_2 = 0,592 \cdot F_1 - 0,182 \cdot F_2 + 0,369 \cdot F_3 + \varepsilon_2;$$

$$X_4 = 0,729 \cdot F_1 - 0,258 \cdot F_2 + 0,347 \cdot F_3 + \varepsilon_4;$$

$$X_6 = 0 \cdot F_1 + 0 \cdot F_2 + 0,256 \cdot F_3 + \varepsilon_6;$$

$$X_8 = 0,438 \cdot F_1 + 0 \cdot F_2 - 0,202 \cdot F_3 + \varepsilon_8;$$

$$X_{10} = 0,133 \cdot F_1 - 0,870 \cdot F_2 + 0,260 \cdot F_3 + \varepsilon_{10};$$

$$X_{12} = 0,698 \cdot F_1 + 0,729 \cdot F_2 + 0,188 \cdot F_3 + \varepsilon_{12};$$

$$X_{14} = 0,572 \cdot F_1 - 0,186 \cdot F_2 + 0,226 \cdot F_3 + \varepsilon_{14};$$

Отже, три виділені фактори переважно незалежні і відображають різні аспекти клінічних змін у пацієнтів.

3. Побудова й оцінка моделі категоріальної регресії ризику

Перейдемо від 15 клінічних показників різного типу і природи до 3 факторних змінних, що повно та значущо охоплюють дані показники. На основі отриманих трьох факторних змінних побудуємо модель категоріальної регресії, яка дасть можливість класифікувати пацієнтів за рівнями ризику для лікувальної імплантації.

Категоріальна регресія (в SPSS: Optimal Scaling Regression / CATREG) – це узагальнення лінійної регресії, що дає змогу використовувати змішані типи змінних: кількісні, порядкові, номінальні. Перед побудовою регресійної моделі виконується оптимальне шкалювання: кожній номінальній або порядковій змінній присвоюють оптимальні числові значення, які максимізують множинну кореляцію між предикторами та залежною змінною.

Зауважимо, що залежна змінна у нас порядкова (можливо її також визначати як номінальну) й у неї наявні такі категорії ризику: 0 (відсутній ризик); 1 (низький ризик); 2 (середній ризик); 3 (високий ризик); 4 (критичний ризик). Інші змінні – кількісні – це факторні навантаження для трьох факторів.

Підсумуємо: $Y \in \{0,1,2,3,4\}$ – це порядкова змінна; незалежні змінні – кількісні факторні бали (латентні фактори), які ми отримали із факторного аналізу: F_1, F_2, F_3 .

Тоді категоріальна регресійна модель у нашому випадку матиме вигляд:

$$\psi(Y_i) = \beta_0 + \beta_1 \phi_1(f_{1i}) + \beta_2 \phi_2(f_{2i}) + \beta_3 \phi_3(f_{3i}) + \varepsilon_i, \quad i = 1, \dots, n,$$

де $\psi: \{0,1,2,3,4\} \mapsto \mathbb{R}$ – монотонна функція оптимального шкалювання залежної змінної (зберігає порядок): $\psi(0) < \psi(1) < \psi(2) < \psi(3) < \psi(4)$; $\phi_j: \mathbb{R} \mapsto \mathbb{R}$ – функція оптимального шкалювання j -го предиктора ($j = 1,2,3$); $\beta_0, \beta_1, \beta_2, \beta_3$ – коефіцієнти регресії; ε_i – випадкова похибка з $E(\varepsilon_i) = 0$.

Коефіцієнти $\beta_0, \beta_1, \beta_2, \beta_3$ та трансформації $\psi, \phi_1, \phi_2, \phi_3$ визначаються шляхом мінімізації суми квадратів залишків:

$$\min_{\beta_0, \beta_1, \beta_2, \beta_3, \psi, \phi_1, \phi_2, \phi_3} \sum_{i=1}^n \left[\psi(Y_i) - \beta_0 - \sum_{j=1}^3 \beta_j \phi_j(f_{ji}) \right]^2.$$

В алгоритмі використовуємо наближення через дискретизацію (IBM, n. d.):

$$\phi_j(f_{ji}) = \sum_{k=1}^{K_j} c_{jk} \mathbf{1}(f_{ji} \in I_{jk}),$$

де I_{jk} – інтервали (градації) значень фактора F_j ; c_{jk} – оптимальні числові коефіцієнти шкали для кожної градації; $\mathbf{1}(f_{ji} \in I_{jk})$ – індикаторна функція (1, якщо f_{ji} належить інтервалу I_{jk}).

Зрештою формуються нові змінні:

$$Y_i^* = \psi(Y_i), \quad f_{ji}^* = \phi_j(f_{ji}), \quad j = 1,2,3.$$

Модель у просторі шкалованих змінних має вигляд лінійної регресії:

$$Y_i^* = \beta_0 + \beta_1 f_{1i}^* + \beta_2 f_{2i}^* + \beta_3 f_{3i}^* + \varepsilon_i. \quad (2)$$

Відповідно, β_j – приріст очікуваного шкалованого індексу ризику Y^* за одиничного приросту F_j^* в умовах фіксованих інших предикторів. Знак β_j визначає напрям впливу фактора на шкалований ризик, а модуль – відносну силу впливу.

Показники якості моделі розраховуємо аналогічно, як і у випадку лінійної регресії, їх наведено у табл. 5 та 6.

Таблиця 5

Параметри якості моделі категоріальної регресії для залежної змінної "Група ризику"

Множинний R	R ²	Скоригований R ²	Середня квадратична помилка прогнозу
0,926	0,858	0,857	0,142

Множинний коефіцієнт кореляції між фактичними і прогнозованими значеннями залежної змінної "Група ризику" $R = 0,926$ свідчить про дуже сильний зв'язок. Коефіцієнт детермінації $R^2 = 0,858$ означає, що 85,8 % варіації "Група ризику" пояснюється моделлю. Скоригований $R^2 = 0,857$ з поправкою на кількість предикторів і розмір вибірки майже такий самий, як R^2 , тобто модель не надто ускладнена.

Таблиця 6

Результати аналізу дисперсії (ANOVA) для моделі категоріальної регресії залежної змінної "Група ризику"

Джерело	Сума квадратів	Ступені свободи	Середній квадрат	F-критерій	Рівень значущості (p – value)
Регресія	476,376	5	95,275	665,264	0,000
Залишки	78,624	549	0,143		
Загальна	555,000	554			

З табл. 6 видно, що модель значущо пояснює змінну Y – "Група ризику", оскільки p – value = 0,000. Отже, ймовірність того, що модель випадкова, менша 0,001.

Стандартизовані коефіцієнти (табл. 7) показують силу та напрямок впливу предикторів: F_1 має невеликий позитивний вплив (+ 0,126); F_2 – дуже сильний негативний вплив (– 0,857); F_3 – невеликий позитивний вплив (+ 0,164). Кожен коефіцієнт статистично значущий (p-value < 0,001). Відповідно найсильніший предиктор F_2 , що свідчить про те, що змінні, які початково сформували зазначений фактор, найсильніше впливають на виявлення рівня ризику для лікувальної імплантації. Проте зауважимо, що всі фактори здійснюють суттєвий вплив на залежну змінну навіть на рівні значущості 0,001.

Таблиця 7

Стандартизовані коефіцієнти та статистична значущість предикторів моделі категоріальної регресії для залежної змінної "Група ризику"

Предиктор	Стандартизовані коефіцієнти		Ступені свободи	F-критерій	Рівень значущості (p – value)
	Beta	Оцінка стандартної похибки			
F_1	0,126	0,028	2	20,869	0,000
F_2	-0,857	0,021	1	1681,605	0,000
F_3	0,164	0,033	2	25,502	0,000

Відповідно нашу модель категоріальної регресії можна записати у вигляді (2):

$$Y_i^* = \beta_0 + 0,126 \cdot f_{1i}^* - 0,857 \cdot f_{2i}^* + 0,164 \cdot f_{3i}^* + \varepsilon_i.$$

У табл. 8 наведені важливі характеристики категоріальної регресії. Нульова кореляція – це проста кореляція кожного предиктора із залежною змінною, без контролю за іншими предикторами. Часткова кореляція – кореляція предиктора із залежною змінною, урахувавши вплив інших предикторів. Напівчасткова кореляція: у цьому випадку вплив інших предикторів вилучається лише із предиктора, а не із залежної змінної. Внесок ознаки – це показник, що відображає "вагу" предиктора після оптимального шкалювання у категоріальній регресії. Показник мультиколінеарності зі значенням $< 0,1$ свідчить про сильну мультиколінеарність (Liao, 2019).

Таблиця 8

Кореляції, внесок предикторів та показники мультиколінеарності для моделі категоріальної регресії (залежна змінна "Група ризику")

Предиктор	Кореляція			Внесок предиктора	Показник мультиколінеарності	
	Нульова кореляція	Часткова кореляція	Напівчасткова кореляція		Після трансформації	До трансформації
F_1	0,255	0,311	0,123	0,037	0,961	0,823
F_2	-0,906	-0,911	-0,831	0,905	0,939	0,886
F_3	0,300	0,393	0,161	0,057	0,960	0,920

Отже, найвищі абсолютні значення усіх типів кореляцій має фактор F_2 (нульова -0,906, часткова -0,911, напівчасткова -0,831). Це знову підтверджує, що саме F_2 є основним фактором, що визначає зміну залежної змінної "Група ризику". Високе значення внеску предиктора після трансформації (0,905) також підтверджує його провідну роль у моделі. F_1 і F_3 демонструють слабші зв'язки із залежною змінною (кореляції 0,255 та 0,39). Їхній внесок після трансформації становить 0,037 і 0,057 відповідно, що вказує на другорядний внесок у модель.

Значення показників мультиколінеарності до та після трансформації для всіх факторів суттєво не відрізняються. Це означає, що оптимальне шкалювання не спотворило структуру зв'язків, а лише уточнило їх.

Усі значення показника мультиколінеарності $> 0,8$, що свідчить про високу незалежність факторів між собою, тобто про відсутність мультиколінеарності.

Отже, модель стабільна, і жоден фактор не дублює інформацію іншого.

Дискусія і висновки

Методологічний підхід "факторний аналіз у поєднанні з категоріальною регресією" ефективно скоротив багатовимірну початкову інформацію (15 змінних) до трьох латентних факторів (F_1 – F_3) зі збереженням клінічної інтерпретованості і статистичної інформативності.

Індикатори придатності даних (статистика КМО = 0,740; тест сферичності Бартлетта $p < 0,001$) підтверджують доцільність факторного аналізу для цієї вибірки ($n = 558$). Перші три фактори пояснюють приблизно 70 % дисперсії до обертання (після обертання сумарна інтерпретована дисперсія становить 64,6 %), що є прийнятним результатом для медико-біологічних даних.

Коротко проаналізуємо клінічне значення виділених факторів. Фактор F_1 "Клінічні та морфологічні ознаки ГЕРХ" охоплює симптоми та структурні маркери рефлюксу, а також має помірний позитивний вплив на ризик. Фактор F_2 "Анатомо-функціональні показники стравоходу" є найсильнішим предиктором і має значний (негативний у межах моделі шкалювання) вплив на категорію ризику, що підкреслює роль анатомічних характеристик (висота сфінктера, довжина пролабування, висота пролабування) у визначенні придатності до імплантації. Фактор F_3 "Ерозивні ураження слизової" додає інформацію про активність ерозивного процесу і підвищує прогностичну здатність моделі.

Категоріальна регресія забезпечила високу якість прогнозу: множинний $R = 0,926$ і $R^2 = 0,858$ (скоригований $R^2 = 0,857$), що вказує на те, що запропонована комбінація факторів пояснює значну частину варіації рівнів ризику. Усі три фактори статистично значущі ($p < 0,001$). Показники мультиколінеарності предикторів $> 0,8$ свідчать про відсутність критичної мультиколінеарності.

Отримана модель може слугувати допоміжним інструментом для стандартизованої стратифікації ризику перед імплантацією пристроїв у ділянці шлунково-стравохідного з'єднання, сприяючи об'єктивнішому ухваленню рішень і визначенню пацієнтів, яким потрібна додаткова оптимізація стану перед процедурою.

Зауважимо, що дослідження обмежене однорегійною вибіркою та ретроспективною / обсерваційною природою аналізу; модель потребує зовнішньої валідації на незалежних когортах, перевірки робастності для різних підгруп пацієнтів (напр., за віком, супутніми хворобами) й оцінки клінічного впливу за проспективного застосування. Подальші дослідження можуть розглянути інтеграцію додаткових маркерів (функціональні тести, тривалі моніторинги) і використання інших алгоритмів класифікації для порівняння продуктивності.

Джерела фінансування. Фінансування частково забезпечено Київським національним університетом імені Тараса Шевченка.

Список використаних джерел

- Akoglu, H. (2018). User's guide to correlation coefficients. *Turk J Emerg Med*, 18(3), 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Devlieger, I. (2019). *Factor score regression: Models for the analysis of variables and factors* [Doctoral dissertation, Ghent University]. Ghent University Academic Repository. <https://backoffice.biblio.ugent.be/download/8655145/8655146>
- Finch, W. H., & French, B. F. (2018). *Educational and Psychological Measurement* (1st ed.). Routledge. <https://doi.org/10.4324/9781315650951>
- Guo, J., Wall, M., & Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, 7(1), 145–163. <https://doi.org/10.1093/biostatistics/kxi046>
- IBM. (n. d.). *CATREG*. IBM SPSS Statistics Algorithms. <https://public.dhe.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/14.0/catreg.pdf>

- Lawenko, R. M. A., & Lee, Y. Y. (2016). Evaluation of gastroesophageal reflux disease using the Bravo capsule pH system. *Journal of Neurogastroenterology and Motility*, 22(1), 25–30. <https://pubmed.ncbi.nlm.nih.gov/26717929/>
- Li, J. N., Liu, C. L., & Tao, X. H. (2013). Clinical utility and tolerability of JSPH-1 wireless esophageal pH monitoring system. *BMC Gastroenterology*, 13(10). <https://doi.org/10.1186/1471-230X-13-10>
- Liao, Y. (2019). *Feature Selection: Removing Multicollinearity – Python Automation and Machine Learning for EM and ICs*. An Online Book, First Edition. https://www.globalsino.com/ICs/page3268.html?utm_source=chatgpt.com
- Smith, L. E., Smith, D. K., Blume, J. D., Siew, E. D., & Billings IV, F. T. (2017). Latent variable modeling improves AKI risk factor identification and AKI prediction compared to traditional methods. *BMC Nephrology*, 18(1), 55. <https://doi.org/10.1186/s12882-017-0465-1>
- Statistics Solutions. (n. d.). *Comprehensive guide to factor analysis*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/?utm_source=chatgpt.com
- Viana, A. T., Gonçalves, K. C. M., & Paez, M. S. (2024). Factor model for ordinal categorical data with latent factors explained by auxiliary variables applied to the Major Depression Inventory. *Journal of Applied Statistics*, 51(14), 2866–2893. <https://doi.org/10.1080/02664763.2024.2321913>
- Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94. <https://doi.org/10.20982/tqmp.09.2.p079>

References

- Akoglu, H. (2018). User's guide to correlation coefficients. *Turk J Emerg Med*, 18(3), 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Devlieger, I. (2019). *Factor score regression: Models for the analysis of variables and factors* [Doctoral dissertation, Ghent University]. Ghent University Academic Repository. <https://backoffice.biblio.ugent.be/download/8655145/8655146>
- Finch, W. H., & French, B. F. (2018). *Educational and Psychological Measurement* (1st ed.). Routledge. <https://doi.org/10.4324/9781315650951>
- Guo, J., Wall, M., & Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, 7(1), 145–163. <https://doi.org/10.1093/biostatistics/kxi046>
- IBM. (n. d.). *CATREG*. IBM SPSS Statistics Algorithms. <https://public.dhe.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/14.0/catreg.pdf>
- Lawenko, R. M. A., & Lee, Y. Y. (2016). Evaluation of gastroesophageal reflux disease using the Bravo capsule pH system. *Journal of Neurogastroenterology and Motility*, 22(1), 25–30. <https://pubmed.ncbi.nlm.nih.gov/26717929/>
- Li, J. N., Liu, C. L., & Tao, X. H. (2013). Clinical utility and tolerability of JSPH-1 wireless esophageal pH monitoring system. *BMC Gastroenterology*, 13(10). <https://doi.org/10.1186/1471-230X-13-10>
- Liao, Y. (2019). *Feature Selection: Removing Multicollinearity – Python Automation and Machine Learning for EM and ICs*. An Online Book, First Edition. https://www.globalsino.com/ICs/page3268.html?utm_source=chatgpt.com
- Smith, L. E., Smith, D. K., Blume, J. D., Siew, E. D., & Billings IV, F. T. (2017). Latent variable modeling improves AKI risk factor identification and AKI prediction compared to traditional methods. *BMC Nephrology*, 18(1), 55. <https://doi.org/10.1186/s12882-017-0465-1>
- Statistics Solutions. (n. d.). *Comprehensive guide to factor analysis*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/?utm_source=chatgpt.com
- Viana, A. T., Gonçalves, K. C. M., & Paez, M. S. (2024). Factor model for ordinal categorical data with latent factors explained by auxiliary variables applied to the Major Depression Inventory. *Journal of Applied Statistics*, 51(14), 2866–2893. <https://doi.org/10.1080/02664763.2024.2321913>
- Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94. <https://doi.org/10.20982/tqmp.09.2.p079>

Отримано редакцією журналу / Received: 30.11.25

Прорецензовано / Revised: 03.02.26

Схвалено до друку / Accepted: 16.04.26

Опубліковано / Published: 05.06.26

Oleksandr HALUSHKO, PhD Student

ORCID ID: 0009-0009-0275-6542

e-mail: health24@ukr.net

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

INTEGRATED MODEL OF FACTOR SELECTION AND CATEGORICAL REGRESSION FOR PREDICTING THE IMPLANTATION RISK OF DIAGNOSTIC-THERAPEUTIC DEVICES IN GASTROESOPHAGEAL JUNCTION PATHOLOGY

The article presents an integrated method for predicting the implantation risk of diagnostic-therapeutic devices in the gastroesophageal junction based on factor analysis and categorical regression. In a sample of 558 patients, 15 clinical and endoscopic indicators were selected, characterizing the mucosal condition, anatomical features, and reflux symptoms. Factor analysis using principal axis factoring with oblimin rotation reduced the initial set of variables to three latent factors: F_1 – "clinical and morphological features of gastroesophageal reflux disease", F_2 – "anatomical and functional indicators of the esophagus", and F_3 – "erosive mucosal lesions". These factors were used as predictors in categorical regression to forecast five levels of implantation risk. The model demonstrated high predictive quality ($R^2 = 0,858$), all factors were statistically significant ($p < 0,001$) and multicollinearity indicators of predictors $> 0,8$ confirmed the absence of critical multicollinearity. The resulting model provides standardized risk stratification, enabling clinicians to more objectively assess the appropriateness of device implantation and identify patients who require additional optimization prior to the procedure.

Key words: device implantation, factor analysis, categorical regression, risk, gastroesophageal junction, risk stratification.

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.