

Article

Causal Graphical Model of Bacterial Vaginosis in Pregnant Women

Maricela García-Avalos ¹, Juana Canul-Reich ^{1,*}, Lil María Xibai Rodríguez-Henríquez ^{2,3}
and Erick Natividad De la Cruz-Hernández ⁴

¹ Academic Division of Science and Information Technology, Universidad Juárez Autónoma de Tabasco, Cunduacan-Jalpa de Mendez Road Km. 1, Cunduacán 86690, Tabasco, Mexico; maricela.garcia@ujat.mx

² Coordination of Computational Sciences, National Institute of Astrophysics, Optics, and Electronics, Luis Enrique Erro No. 1, Tonantzintla, San Andrés Cholula 72840, Puebla, Mexico; lmrodriguez@inaoep.mx

³ SECIHTI, Av. de los Insurgentes Sur 1582, Crédito Constructor, Benito Juárez, Ciudad de Mexico 03940, Mexico

⁴ Multidisciplinary Academic Division of Comalcalco, Universidad Juárez Autónoma de Tabasco, Rancheria Sur 4th Section, Comalcalco 86658, Tabasco, Mexico; erick.delacruz@ujat.mx

* Correspondence: juana.canul@ujat.mx

Abstract

Background: This study developed a Causal Graphical Model (CGM) to analyze Bacterial Vaginosis (BV), a condition caused by an imbalance in the vaginal microbiota, whose bacterial composition varies among women. While previous studies used variable selection, clustering, and association rules to identify BV-associated bacteria, these approaches lack visual tools to explore causal relationships and determine which are the most relevant. In contrast, the CGM generated in this study allows visualization of associated bacteria and their causal links, thereby identifying those most influential. **Methods:** Path Analysis (PA), a statistical structural equation modeling method, was used to construct the CGM, with emphasis on observable variables and to assess direct and indirect effects through correlations and covariances. PA was applied to an already-collected third-party dataset related to BV diagnosis, consisting of data from 132 pregnant women between 4 and 24 weeks of gestation. **Results:** The CGM, built using a theoretical model based on the Spearman correlation matrix, was validated through statistical metrics and by a clinical-biological expert. The resultant model highlights bacteria influencing BV diagnosis, specifically *Mycoplasma hominis* (Mh), *Atopobium vaginae* (Av), *Gardnerella vaginalis* (Gv), *Megasphaera Type 1* (MT1), and *Bacteria Associated with Bacterial Vaginosis Type 2* (BVAB2). Among them, MT1 and BVAB2 showed the strongest association with BV. **Conclusions:** The CGM effectively identifies causal associations among bacteria related to BV.

Keywords: path analysis; correlation matrix; variability; statistical metrics



Academic Editor: Fernando Monroy

Received: 25 September 2025

Revised: 7 November 2025

Accepted: 10 November 2025

Published: 15 November 2025

Citation: García-Avalos, M.;

Canul-Reich, J.; Rodríguez-Henríquez, L.M.X.; De la Cruz-Hernández, E.N.

Causal Graphical Model of Bacterial Vaginosis in Pregnant Women. *Diseases* **2025**, *13*, 375. <https://doi.org/10.3390/diseases13110375>

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license

(<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Bacterial Vaginosis (BV) is a common microbiological disorder in women of reproductive age, affecting over 30% of the population, with symptoms occurring in approximately half of the cases [1,2]. When symptomatic, BV manifests as a homogeneous, foul-smelling vaginal discharge, with a distinctive fishy odor that intensifies after sexual intercourse [3,4].

In addition to the impact on women's quality of life, BV is associated with an imbalance in the vaginal microbiota, which can lead to gynecological and obstetric complications [3]. BV is not caused by a single pathogen but rather by a consortium of anaerobic bacteria,

including *Gardnerella vaginalis*, *Atopobium vaginae*, *Bacteria associated with bacterial vaginosis Type 2*, *Megasphaera Types 1 and 2*, whose roles vary across populations and clinical contexts [4]. Moreover, recent studies highlight that BV may follow multiple etiological pathways depending on host–microbe interactions, and that microbial profiles differ significantly across geographic and ethnic populations [2,5]. This study aims to identify coexisting bacterial species influencing BV diagnosis in pregnant women.

Machine learning techniques, including attribute selection, clustering, and association rules, have been used to identify bacteria associated with BV [6–8]. These approaches help analyze large datasets, improving diagnostic efficiency and treatment strategies [9]. However, while these methods identify relevant associations, they lack visual tools to represent causal relationships and determine which bacteria are most influential. Incorporating such representations would support healthcare professionals in making more accurate diagnoses and defining appropriate treatments for pregnant women with BV.

Recently, a study applied Bayesian Networks (BNs) to a BV dataset, generating a graphical model that, like Path Analysis (PA), represents the relationships between variables. However, while BNs show probabilistic dependencies between bacterial species and clinical factors [10], PA quantifies the strength and direction of direct and indirect effects through path coefficients, providing the total causal effect between bacteria and the diagnosis. Therefore, Path Analysis provides a more informative and suitable framework for causal inference in this study.

Path Analysis (PA) is a multivariate statistical method within structural equation modeling that examines direct and indirect effects among observable variables, using correlations and covariances [11–13]. In this study, PA was applied using R software [14] on BV diagnosis data from pregnant women in Tabasco, Mexico. The resulting Causal Graphical Model (CGM) identified five bacteria influencing BV diagnosis: *Mycoplasma hominis* (Mh), *Atopobium vaginae* (Av), *Gardnerella vaginalis* (Gv), *Megasphaera Type 1* (MT1), and *Bacteria Associated with Bacterial Vaginosis Type 2* (BVAB2). The first four bacteria align with previous findings from machine learning models [6–8].

Unlike previous techniques, our CGM provides a deeper causal analysis, revealing that MT1 and BVAB2 are the most influential bacteria in BV diagnosis with total effect values of 0.4900 and 0.5750, respectively. In contrast, the values of Mh (−0.1937), Av (0.2627), and Gv (0.0482) reflect differentiated contributions: Mh presents an inverse association, Av a moderate influence, and Gv a low participation.

This study is described in the rest of the paper as follows: Section 2 presents the materials and methods used. Section 3 describes the experimental study, data preprocessing, and the application of Path Analysis to construct the Causal Graphical Model. Section 4 reports the results with focus on the most influential bacteria in the diagnosis of Bacterial Vaginosis. Section 5 discusses the findings with respect to previous studies. Finally, Section 6 outlines the study's conclusions.

2. Materials and Methods

2.1. Spearman Correlation Matrix

Correlation determines the relationship between two quantitative variables, assessing both direction and intensity. Covariance measures how two random variables vary together, and through standardization, it gives out the correlation coefficients [15].

These coefficients range from −1 to 1, where 0 indicates no correlation. A positive value represents a direct relationship, while a negative value denotes an inverse correlation [15].

The Spearman correlation coefficient [16,17], unlike Pearson's, does not require normality or a strictly linear relationship. It is computed using Equation (1).

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (1)$$

where

- ρ_s : Spearman correlation coefficient.
- d_i : Difference between ranks of observation pairs.
- n : Total number of observations.

2.2. Path Analysis

Structural Equation Models (SEM) are multivariate statistical tools, also known as covariance structure analysis, that examine direct (causal) and indirect relationships between variables. SEM incorporates both observable and latent variables, with Path Analysis (PA) being a method specifically focused on observable variables, identifying direct and indirect effects among them [18,19].

The graphical conventions of the PA model [13,19] are illustrated in Figure 1, which visually represents causal relationships among variables through observable associations.

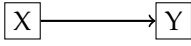
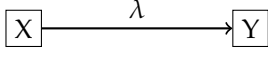
Representation	Description
	Observable variables are represented using squares or rectangles.
	Unidirectional arrows indicate a direct causal effect of one variable on another.
	Curves with double arrows indicate an association between the two variables (covariance).
	The model parameters are represented on the corresponding arrow.

Figure 1. Graphical representation of the PA model, illustrating the causal relationships among variables through observable effects and associations, adapted from [13,19].

In PA, variables are categorized into two types [19,20]:

- Endogenous: Affected by other variables within the model.
- Exogenous: Not influenced by other variables in the same context.

Types of effects in a path diagram, classified according to their influence on variable interactions [13,19]:

- Direct effect: Influence of one variable on another without intermediaries.
- Indirect effect: A mediating variable connects independent and dependent variables.
- Total effect: Sum of direct and indirect effects.

2.3. Statistical Metrics in Path Analysis

Statistical metrics assess model performance relative to observed data by quantifying its fit and predictive capacity. Table 1 presents metrics used in PA [18,21] to evaluate whether the theoretical model adequately represents the underlying data structure.

RMSEA and SRMR are absolute fit indices that evaluate how well the covariance matrix aligns with the implicit model, where lower values indicate better fit quality. GFI, NFI, and CFI are incremental fit indices that measure improvements over a baseline model [21,22].

Table 1. Statistical metrics to assess model fit in path analysis.

Metric	Description	Expected Value
RMSEA	Root Mean Square Error of Approximation	<0.05
SRMR	Standardized Root Mean Square Residual	<0.05
GFI	Goodness of Fit Index	≥0.95
NFI	Normalized Fit Index	≥0.95
CFI	Comparative Fit Index	≥0.95

2.4. Kolmogorov–Smirnov Test

In statistics, parametric and non-parametric tests are widely used for data analysis. One nonparametric test is the Kolmogorov–Smirnov test, which assesses whether a sample follows a normal distribution.

This test considers two hypotheses at a significance level of $p \geq 0.05$:

- H_0 : The data follow a normal distribution.
- H_1 : The data do not follow a normal distribution.

The Kolmogorov–Smirnov statistic, presented in Equation (2), measures the maximum vertical distance between empirical cumulative distribution functions or between an empirical and a theoretical cumulative distribution function [23,24]. The test determines whether the null hypothesis should be rejected based on the significance level.

$$D_n = \sup_x |F_n(x) - F(x)| \quad (2)$$

where

- D_n : Kolmogorov–Smirnov statistic, which measures the maximum distance.
- $\sup_x$: Supremum (the maximum value) of the absolute differences between the functions.
- $F_n(x)$: Empirical cumulative distribution function based on the sample data.
- $F(x)$: Theoretical cumulative distribution function under the null hypothesis.

2.5. Q-Q Plots

Normal quantile plots, or Q-Q plots, compare a dataset's distribution to a theoretical distribution, such as the normal distribution [23]. In these plots, observed data quantiles are on the horizontal axis, while expected quantiles from the theoretical distribution are on the vertical axis. When the data follow the theoretical distribution, the points align diagonally. Deviations indicate potential mismatches between the dataset and the theoretical model.

2.6. Data Transformation

Transformations are applied when data do not follow a normal distribution, in order to improve the stability and interpretability of the models. The transformations applied in this study are as follows [25,26]:

- Logarithmic transformation: Used for right-skewed positive data; a constant can be added if zero or negative values are present. It is defined as Equation (3).

$$y = \log(x + a). \quad (3)$$

- Square root transformation: Suitable for non-negative count data with moderate right skewness, especially when the variance increases with the mean, and represented as Equation (4).

$$y = \sqrt{x}. \quad (4)$$

- Reciprocal transformation: Helps reduce the impact of large values but cannot be applied to zero, as the reciprocal is undefined. It is given by Equation (5).

$$y = \frac{1}{x}. \quad (5)$$

- Box-Cox transformation: A generalized approach requiring positive data, expressed in Equation (6).

$$y = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{si } \lambda \neq 0. \\ \log(x), & \text{si } \lambda = 0. \end{cases} \quad (6)$$

where x denotes the original value, y the transformed value, a shift constant and λ a parameter estimated from the data to enhance normality.

3. Experimental Study

The study was conducted in two phases: the first involved preprocessing the dataset, and the second consisted of implementing path analysis to build the causal graphical model. All analyses were performed in R version 4.3.2 (Windows 11), using the following functions and packages:

- Statistical functions.
- Lavaan for SEM/PA.
- SemPlot for plotting.

The dataset and R scripts are available upon request.

3.1. Data Preprocessing

Data preprocessing is essential for any statistical analysis or machine learning model, as it includes steps such as imputation of missing values, handling of outliers, removal of irrelevant variables, among others. These procedures enhance dataset quality, reduce bias, and ensure more reliable and reproducible results [27,28]. It is important to note that the dataset does not exclusively contain variables related to bacterial vaginosis, as it was designed to support multiple types of studies.

As shown in Figure 2, data preprocessing involved five steps: (a) understanding the source of the dataset, (b) analyzing its structure, (c) removing irrelevant variables, (d) addressing outliers, and (e) imputing missing values to ensure consistency and analytical readiness.

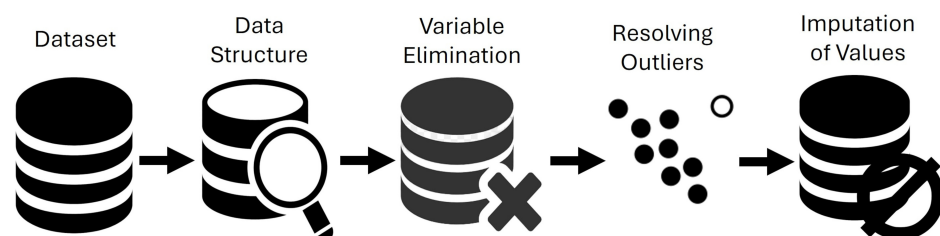


Figure 2. Stages applied during the development of the analysis.

(a) Dataset: Data were collected by a researcher from the Multidisciplinary Academic Division of Comalcalco (DAMC), Universidad Juárez Autónoma de Tabasco (UJAT). This cross-sectional dataset included one sample per participant, collected at varying gestational weeks (4–24), without longitudinal follow-up, and was not generated by the authors of the present study. Sample collection followed the standardized clinical protocol previously described by [4], conducted at the Laboratory of Research in Metabolic and Infectious

Diseases (UJAT). Vaginal and cervical samples were obtained by a trained gynecologist using sterile swabs from the ectocervix and posterior vaginal fornix, and stored at 4 °C until genomic DNA extraction. Bacterial detection and quantification were performed using standardized molecular methods applied consistently across all samples. The dataset contains information from pregnant women participating in healthy pregnancy campaigns (August 2018–January 2020) across rural and urban communities in Tabasco, Mexico. It includes sociodemographic data, bacterial presence, Human Papillomavirus, and BV diagnosis.

(b) Data Structure: To understand the structure of the dataset, an exploratory analysis was performed. Table 2 summarizes its main aspects, showing an imbalance between classes, with BV− as the majority class. However, this imbalance does not affect PA methodology, as PA relies on covariance or correlation matrices rather than class proportions [13,18].

The data set comprises three classes of BV diagnosis:

- BV+: Positive BV, indicating presence of the condition.
- Indeterminate: Cases without a clear diagnosis.
- BV−: Normal microbiota.

Table 2. Summary of the data structure.

Aspect	Description
Attributes	87 variables
Instances	132 observations
Attribute Types	41 categorical and 46 numerical
Observations by Class	BV+ (32), Indeterminate (5), and BV− (95)
Missing Data	609 values
Outliers	143 values

In addition to class distribution, missing values and outliers were examined as part of the exploratory analysis. Box-and-whisker diagrams were used to identify outliers by visualizing distribution through minimum, first quartile, median, third quartile, and maximum values [29], see Figure 3.

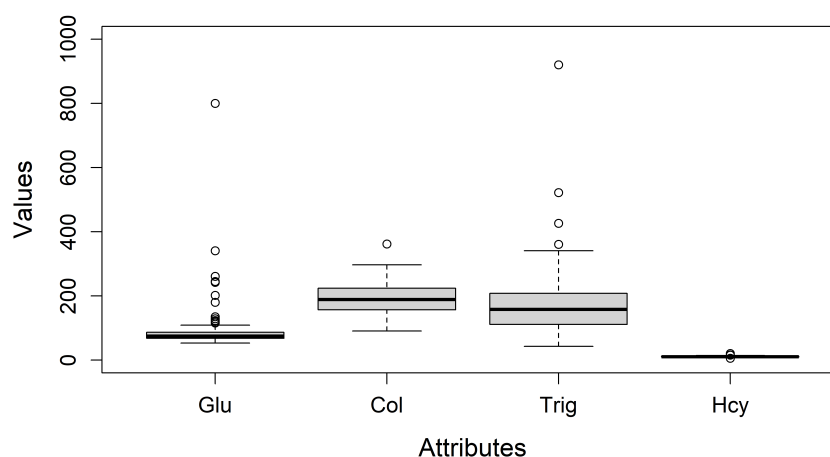


Figure 3. Box and whisker plots of Glu, Col, Trig, and Hcy, showing the distribution and variability of Glucose, Cholesterol, Triglycerides, and Homocysteine.

(c) Variable Elimination: Next, consultations were held with the clinical-biological team responsible for data creation to identify variables unrelated to BV. Variables associated with Human Papillomavirus (HPV) were removed, as previous studies found no

significant association with BV [4]. This refinement reduced the dataset to 72 attributes and 132 instances, retaining only variables relevant to the study objectives.

(d) Resolving Outliers: To ensure result quality and reliability, missing values and outliers were addressed [30]. A detailed analysis based on boxplot charts, identified outliers in Glucose (Glu), Cholesterol (Col), Triglycerides (Trig), and Homocysteine (Hcy). These attributes could be real values and may indicate additional patient conditions; thus, the clinical-biological expert recommends retaining them for a more comprehensive analysis. Figure 3 presents box plots illustrating their distribution and preserved out-of-range values.

(e) Imputation of Values: Subsequently, missing values were addressed through imputation, replacing null entries to prevent gaps from affecting the analysis. For numerical variables (e.g., age, salary), missing values were substituted with the mean of the attribute, while for categorical variables (e.g., gender, color), the most frequent value (mode) was used [31,32].

In this study, imputation was performed based on the BV diagnosis. For categorical variables, complete instances of each attribute were grouped by diagnosis (BV+, Indeterminate, or BV−), and the mode of each group was assigned to the missing values corresponding to that specific diagnosis. For numerical variables, the mean was applied following a similar process.

Tables 3 and 4 display the imputed data for the first ten observations of the variables number of pregnancies (N.EMBARAZO) and vaginal discharge (FlujoV). The column DxVBNomh (Diagnosis of Bacterial Vaginosis without *Mycoplasma hominis*) indicates the BV diagnosis: 1 for BV+ (positive), 2 for Indeterminate, and 3 for BV− (normal microbiota).

Table 3. Data imputed with the mean of the attribute N.EMBARAZO.

Instance	Original Value	Imputed Value	DxVBNomh
1	1	1.000000	3
2	NA	2.500000	1
3	NA	2.571429	3
4	NA	2.571429	3
5	2	2.000000	3
6	NA	2.571429	3
7	1	1.000000	3
8	1	1.000000	3
9	1	1.000000	3
10	3	3.000000	3

Table 4. Data imputed with the mode of the attribute FlujoV.

Instance	Original Value	Imputed Value	DxVBNomh
1	2	2	3
2	NA	2	1
3	NA	2	3
4	NA	2	3
5	1	1	3
6	NA	2	3
7	1	1	3
8	2	2	3
9	2	2	3
10	2	2	3

The original dataset included 87 variables and 132 instances. After removing irrelevant attributes, 72 variables remained, with the number of instances unchanged. Outliers were identified and retained based on recommendations from the clinical-biological team, who

deemed them relevant for interpretation and analysis. Missing values were imputed using the mode for categorical variables and the mean for numerical ones, ensuring completeness and coherence for subsequent analysis.

3.2. Application of Path Analysis

Figure 4 illustrates the steps of the second phase of the study, which focused on path analysis. This stage included (a) variability assessment, (b) distributional analysis, (c) exploration of the correlation structure, (d) theoretical modeling, (e) validation using statistical metrics, (f) construction of the causal graphical model and (g) review by the expert team.

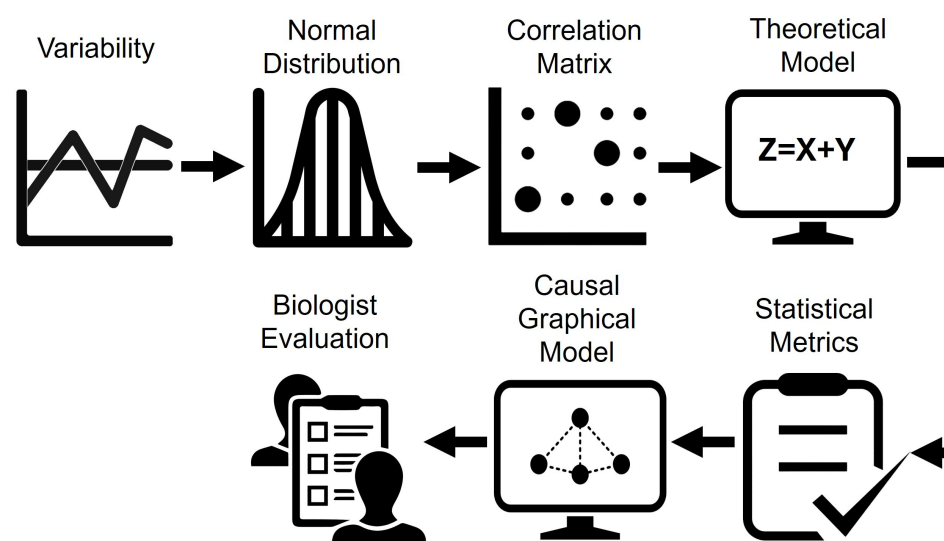


Figure 4. Applied methodology for trajectory analysis: from statistical exploration to evaluation of the causal graphical model.

(a) Variability: Ensuring data variability is crucial for PA validity, as parameter estimation in the path diagram depends on variance and covariances [18,22]. Thus, variance was calculated to confirm that the variables exhibited sufficient variability. Table 5 shows the variance values for some variables in the dataset.

Table 5. Variance of illustrative variables in the dataset.

Attribute	Variance
Age	74.06
Seeks of gestation (SemanasGesta)	55.83
Glucose (Glu)	5431.24
Cholesterol (Col)	2335.83
<i>Gardnerella vaginalis</i> (Gv)	116.56
<i>Atopobium vaginae</i> (Av)	111.25
<i>Bacteria Associated with Bacterial Vaginosis Type 2</i> (BVAB2)	92.58
<i>Lactobacillus crispatus</i> (Lcrispatus)	158.32
<i>Lactobacillus iners</i> (Liners)	115.07
<i>Lactobacillus jensenii</i> (Ljensenii)	143.58

(b) Normal Distribution: After verifying variability, the distribution of the variables was analyzed to determine whether Pearson's or Spearman's correlation matrix would be appropriate for assessing correlations. The Q-Q plot and Kolmogorov-Smirnov normality test evaluated adherence to a normal distribution and were applied to each attribute in the dataset, revealing that none exhibited normality.

For example, Figures 5 and 6 display the Q-Q plots for weeks of gestation (SemanaGesta) and *Atopobium vaginae* (Av), both of which indicate a non-normal distribution. The test results showed $p = 0.0001741$ with $D = 0.18819$ for SemanaGesta and $p < 2.2 \times 10^{-16}$ with $D = 0.40474$ for Av, both with p -values below 0.05, confirming a deviation from normality.

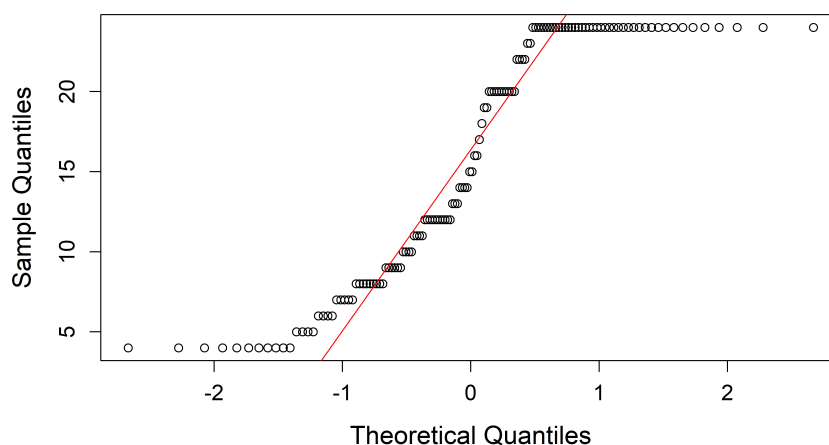


Figure 5. Q-Q plots of SemanasGesta, assessing the distribution of gestational weeks compared to a theoretical normal distribution.

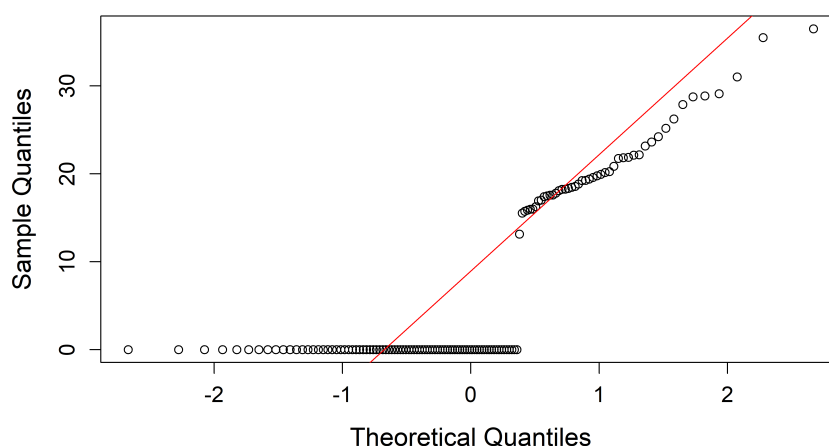


Figure 6. Q-Q plots of Av, assessing the distribution of *Atopobium vaginae* compared to a theoretical normal distribution.

Different transformation techniques, including logarithmic, square root, reciprocal, and Box-Cox transformations, were applied to improve distribution characteristics. However, due to the nature of the data, no one successfully achieved normality. Table 6 presents the p -values of the Kolmogorov-Smirnov test applied to each illustrative variable: weeks of gestation (SemanasGesta), *Mycoplasma hominis* (Mh), *Atopobium vaginae* (Av), *Bacteria Associated with Bacterial Vaginosis Type 2* (BVAB2) and *Lactobacillus crispatus* (Lcrispatus), before and after transformation. N/A values indicate that the transformation was not applicable due to the presence of zeros.

(c) Correlation Matrix: Since the dataset does not follow a normal distribution, the Spearman correlation matrix assessed relationships between variables and their association with BV diagnosis. Five key variables linked to the class attribute (DxVBNoMh) were identified: *Mycoplasma hominis* (Mh), *Atopobium vaginae* (Av), *Gardnerella vaginalis* (Gv), *Megasphaera Type 1* (MT1), and *Bacteria Associated with Bacterial Vaginosis Type 2* (BVAB2) [33].

Table 6. Kolmogorov-Smirnov (*p*-values) test before and after transformations applied to each illustrative variable.

Attribute	Original	Logarithmic	Square Root	Reciprocal	Box-Cox
SemanasGesta	0.0001741	7.878×10^{-5}	0.0001983	6.869×10^{-6}	$<2.2 \times 10^{-16}$
Mh	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
Av	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	N/A	N/A
BVAB2	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	N/A	N/A
Lcrispatus	1.583×10^{-7}	3.553×10^{-12}	8.09×10^{-9}	N/A	N/A

The correlation matrix presents the coefficients between these variables and the BV diagnosis. Figure 7 includes three diagnostic categories: BV+ (presence of condition), Indeterminate (unclear diagnosis), and BV− (normal microbiota), while Figure 8 considers only BV+ and BV−. Comparing the matrices reveals that the correlation values are highly similar, suggesting that the Indeterminate class has minimal impact and can be omitted without affecting the results.

To investigate the indeterminate class further, future research will conduct an exploratory analysis to determine whether specific patterns justify its inclusion in refined diagnostic models. Thus, Figure 8 is this study's matrix of primary interest. The last column presents correlation coefficients between the variables and the BV diagnosis: DxVBNoMh with Av (−0.70), Gv (−0.41), MT1 (−0.89), BVAB2 (−0.83), and Mh (0.55). Although some values are low, their clinical relevance lies in their potential to provide essential information for diagnostic and therapeutic decisions [34,35].

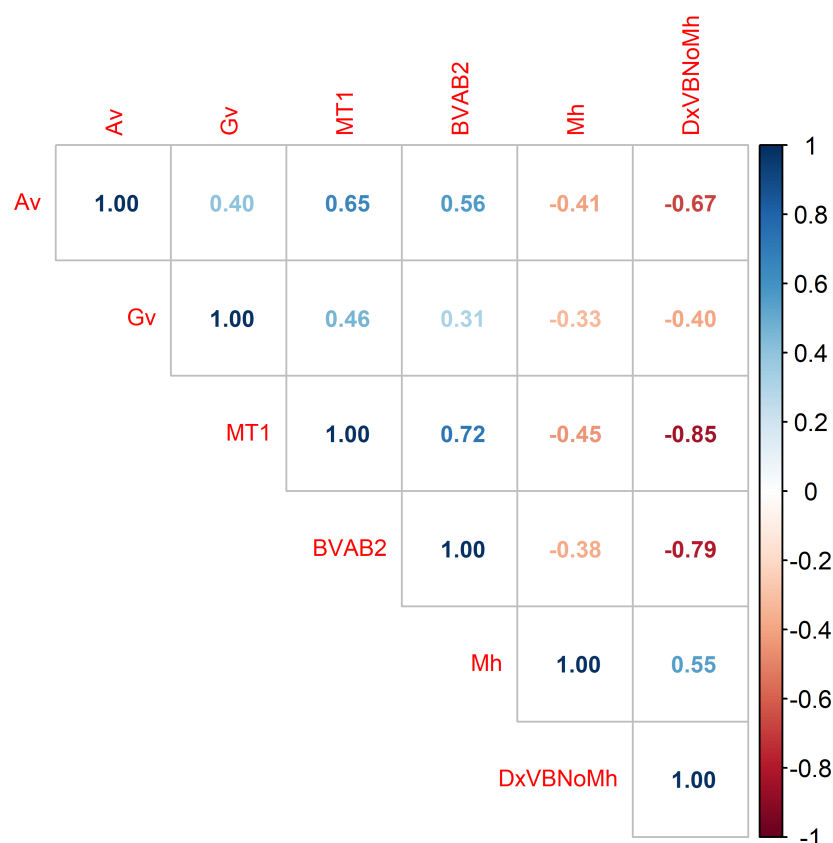
**Figure 7.** Correlation matrix of DxVBNoMh (BV diagnosis categories: BV+, Indeterminate, BV−) and bacterial presence, highlighting their associations.



Figure 8. Correlation matrix of DxVBNomh (BV diagnosis categories: BV+, BV−) and bacterial presence, highlighting their associations.

(d) Theoretical Model: Based on the correlation matrix, a theoretical model was built as a prerequisite for obtaining the CGM from the dataset. This model is represented by the Equation (7) which specifies the structure of the theoretical model used to construct the CGM. The symbol $\sim$ denotes a directional relationship, where the variable on the left is modeled as a function of those on the right. The first line indicates that the BV diagnosis without *Mycoplasma hominis* (DxVBNomh) is predicted using five bacterial variables: Mh, Av, Gv, MT1, and BVAB2. The subsequent lines model MT1 as a dependent variable influenced by each of the other bacteria individually, reflecting its mediating role in the causal pathway. This structure suggests that MT1 may act as a conduit through which other bacteria exert influence on BV diagnosis.

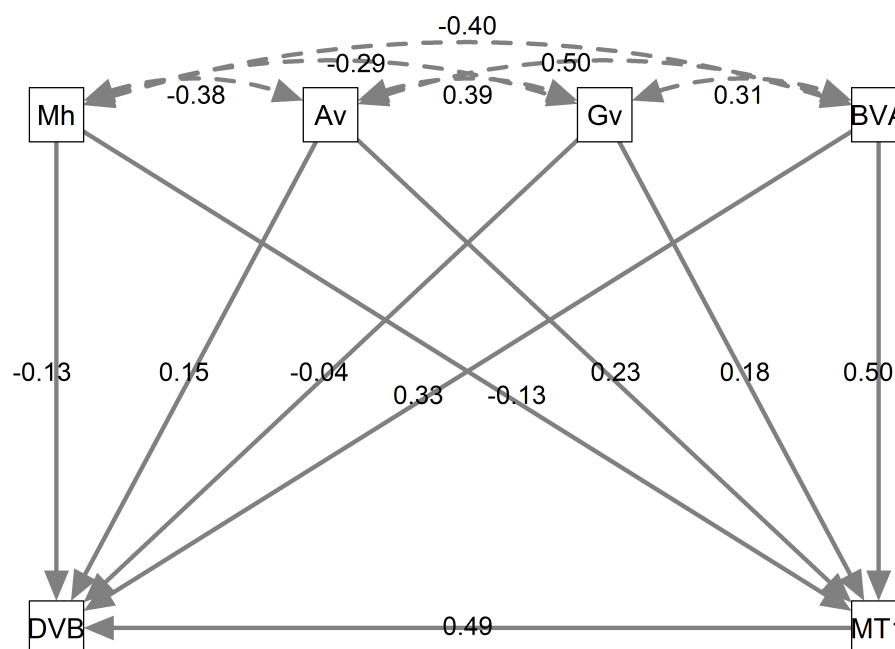
$$\begin{aligned}
 &\text{ModC} \leftarrow \\
 &\text{'DxVBNomh} \sim \text{Mh} + \text{Av} + \text{Gv} + \text{MT1} + \text{BVAB2} \\
 &\text{MT1} \sim \text{Mh} \\
 &\text{MT1} \sim \text{Av} \\
 &\text{MT1} \sim \text{Gv} \\
 &\text{MT1} \sim \text{BVAB2}'
 \end{aligned} \tag{7}$$

(e) Statistical Metrics: After obtaining the theoretical model, its fit must be evaluated. Table 7 presents the values derived from applying the statistical metrics in Table 1 to Equation (7).

Table 7. Model metric values.

Indicator	Value	Ideal Value	Good Fit
RMSEA	0.00	<0.05	Yes
SRMR	0.00	<0.05	Yes
GFI	1.00	≥0.95	Yes
NFI	1.00	≥0.95	Yes
CFI	1.00	≥0.95	Yes

(f) Causal Graphical Model: Based on the theoretical model in Equation (7) and statistical metrics in Table 7, the CGM (Figure 9) was constructed to visually represent causal and indirect relationships among the investigated variables and their impact on DVB. Due to R programming simplifications, some variables were abbreviated: DxVBNoMh as DVB and BVAB2 as BVA. The CGM illustrates causal and indirect relationships (solid lines) between all variables and DVB, as well as covariations (dashed lines). Path coefficients indicate association strength, with values near ± 1 representing strong associations and values close to 0 indicating weak associations.

**Figure 9.** Causal graphical model illustrating the relationships between predictive bacterial attributes and BV diagnosis.

The strongest causal relationship observed was between MT1 and DVB, with a coefficient of 0.49, suggesting that an increase in MT1 level is associated with a higher probability of DVB diagnosis. The weakest causal relationship was between Gv and DVB, with a coefficient of -0.04 , indicating a weak negative association. Additionally, indirect relationships were identified, such as the connection between BVAB2 and DVB through MT1, where MT1 acts as an intermediate variable, with coefficients of 0.50 and 0.49, respectively.

Total effects were calculated by summing direct and indirect effects using standardized path coefficients [36,37]. The direct effect corresponds to the path coefficient linking one variable to another, representing their immediate relationship. Indirect effects were determined by multiplying coefficients along causal pathways. Table 8 summarizes the total effects of predictor variables on BV diagnosis.

Table 8. Total effects of bacteria with DxVBNoMh (DVB). Based on Figure 9, computed effect is calculated by algebraic sum of direct effect with indirect effect which in turn is obtained by multiplying the coefficients on the arrows.

Attributes	Computed Effect	Total Effects on DVB
Mh	$-0.13 + (-0.13 \times 0.49)$	-0.1937
Av	$0.15 + (0.23 \times 0.49)$	0.2627
Gv	$-0.04 + (0.18 \times 0.49)$	0.0482
MT1	0.49	0.4900
BVAB2	$0.33 + (0.50 \times 0.49)$	0.5750

An interpretation of Table 8 is provided below:

- Mh (−0.1937): Slightly reduces the probability of DVB, suggesting a modest protective effect.
- Av (0.2627): Positively associated with DVB, indicating a higher likelihood of diagnosis.
- Gv (0.0482): Weak positive correlation; a small increase in Gv slightly raises diagnosis probability.
- MT1 (0.4900): Strong positive impact, significantly increasing DVB likelihood.
- BVAB2 (0.5750): Highest positive correlation, indicating a strong association with DVB risk.

(g) Biologist Evaluation: Finally, the CGM was presented to the clinical-biological expert for interpretation, validating the identified causal relationships and assessing their relevance to BV diagnosis. The expert evaluated the causal links based on their biological plausibility, clinical coherence, and consistency with the specialized scientific literature.

Path analysis visualizes bacterial relationships through a diagram, illustrating their connection to BV. In this study, the selected bacteria (Av, Gv, MT1, BVAB2, and Mh) align with those most commonly associated with BV, as reported in specialized biologist literature. The diagram highlights direct and indirect interactions, allowing for calculating each bacterium's total effect on the diagnosis and identifying the most influential one.

4. Results

This study analyzed a dataset of pregnant women tested for BV, applying Path Analysis (PA) to identify coexisting bacteria influencing diagnosis. The original dataset contained 87 attributes and 132 instances, categorized into bacterial vaginosis, indeterminate, and normal microbiota. After exploratory analysis, it was refined to 72 attributes, considering only bacterial vaginosis and normal microbiota.

Given the non-normal distribution of the data, a Spearman correlation matrix identified five bacteria significantly associated with BV: *Mycoplasma hominis* (Mh), *Atopobium vaginae* (Av), *Gardnerella vaginalis* (Gv), *Megasphaera Type 1* (MT1), and *Bacteria Associated with Bacterial Vaginosis Type 2* (BVAB2).

Statistical evaluation confirmed the model's fit (RMSEA = 0.00, SRMR = 0.00, GFI = 1.00, NFI = 1.00, CFI = 1.00), supporting the construction of the Causal Graphical Model (CGM).

Among the identified bacteria, MT1, BVAB2 and Av exhibited the strongest effects on BV diagnosis, with total effect values of 0.4900, 0.5750 and 0.2627, respectively. The remaining bacteria contributed as follows: Mh (−0.1937) and Gv (0.0482), demonstrating unique interactions in the microbiological ecosystem.

Expert validation confirmed the CGM as an effective visual tool for analyzing bacterial interactions and their impact on BV diagnosis, reinforcing the relevance of causal modeling in microbiological research.

5. Discussion

The PA methodology identified five bacteria significantly associated with BV diagnosis, with MT1 and BVAB2 showing the most potent effects. While Av exhibited a positive relationship of lesser magnitude, Gv had a weak positive impact, and Mh showed a potential inverse association. Unlike machine learning models that identify bacterial correlations but do not quantify influence, PA enables the assessment of direct and indirect effects, providing deeper insights into BV diagnosis. The identified bacteria (Gv, Av, Mh, and MT1) align with previous findings [6–8], reinforcing the validity of the CGM.

The direct relationship between MT1 and BV diagnosis, as identified by the CGM model, is consistent with the high prevalence of MT1 during pregnancy reported by Glascock et al. [38]. These findings suggest that, unlike other BV-associated bacteria, the physiological changes of pregnancy may favor colonization by MT1, increasing the risk of ascending infection of the upper genital tract. In addition, the indirect effect of MT1 on BVAB2 and Av highlights the role of the heterogeneity of microbial communities for BV development, supporting the proposal that MT1 is frequently associated with a vaginal high-diversity microbiome [38]. Gestational age is known to influence vaginal microbiota composition, potentially affecting the interpretation of infection dynamics throughout pregnancy. However, the cross-sectional design of the dataset restricts temporal follow-up. Future longitudinal applications of PA could capture these temporal shifts, providing a more complete understanding of influential bacterial trajectories.

Although the sample size is relatively small (132 real cases), PA relies on correct model specification given using correlation and covariance matrices rather than class balance, ensuring stable coefficient estimates. Future validation with larger datasets will strengthen its external validity and enhance clinical applicability, demonstrating the value of PA in microbiological research.

6. Conclusions

This study applied path analysis (PA) to a dataset of pregnant women to identify causal variables associated with the diagnosis of bacterial vaginosis. Before implementing the model, a structured data preprocessing stage was carried out to minimize bias and ensure optimal performance. Missing numerical values were imputed using the mean, and categorical values using the mode.

Outliers were retained based on the recommendation of a clinical-biological expert, as they reflected other conditions present in the patients. Additionally, variables unrelated to bacterial vaginosis such as those associated with Human Papillomavirus (HPV) were excluded based on prior evidence of non-association. These decisions allowed the dataset to be refined and the analysis to focus on relevant attributes.

PA effectively identified causal relationships between five bacterial variables and BV diagnosis. Among them, BVAB2 exhibited the strongest overall contribution (0.5750), followed by MT1 (0.4900), Av (0.2627), Gv (0.0482) and Mh (−0.1937). These findings align with the previous literature and were validated by a clinical-biological expert, confirming the methodology's reliability.

In addition to identifying significant associations, path analysis enabled the visualization of direct and indirect relationships between bacterial variables and the diagnosis, allowing the estimation of total effects that clarify each bacterium's contribution to BV. Beyond its analytical advantages, the CGM could offer clinically valuable insights, supporting gynecological assessments and prenatal care, particularly in low-resource settings.

In future work, we propose validating this CGM using larger and more diverse datasets, exploring its applicability in longitudinal and multicentric research to refine diagnostic precision and treatment strategies in clinical settings.

Author Contributions: Conceptualization, M.G.-A. and J.C.-R.; methodology, M.G.-A.; software, M.G.-A.; validation, J.C.-R., L.M.X.R.-H. and E.N.D.I.C.-H.; formal analysis, M.G.-A., J.C.-R. and L.M.X.R.-H.; investigation, M.G.-A.; resources, M.G.-A.; data curation, M.G.-A. and J.C.-R.; writing—original draft preparation, M.G.-A.; writing—review and editing, J.C.-R. and L.M.X.R.-H.; visualization, J.C.-R. and L.M.X.R.-H.; supervision, L.M.X.R.-H.; project administration, J.C.-R.; funding acquisition, none received. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Secretariat of Science, Humanities, Technology, and Innovation (SECIHTI) through scholarship number 842958.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the use of secondary data provided by DAMC-UJAT, with no direct involvement of human participants by the authors.

Informed Consent Statement: Informed consent for participation was not required as per institutional policy, since the authors did not collect data directly from human subjects.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to data privacy restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Salas, J.; Angulo, L.; Mendez, E. Vaginosis bacteriana: Actualización y novedad terapéutica. *UCIMED* **2022**, *5*, 85–93. [CrossRef]
- Cameron, C.; Dion, M.; Broce, J.; Ojo, D.; Drobin, E.; Louis, A.; Gilbert, J.; Fang, R.; Parker, I. Ethnic disparity in diagnosing asymptomatic bacterial vaginosis using machine learning. *npj Digit. Med.* **2023**, *6*, 211. [CrossRef]
- Morelli, I.; Gamboa, S. Vaginosis bacteriana en el embarazo: últimos avances hasta la fecha. *Rev. Méd. Sinerg.* **2022**, *7*, e838. [CrossRef]
- Sanchez, E.; Contreras, A.; Martinez, E.; Garcia, D.; Lizano, M.; Cruz, E. Molecular epidemiology of bacterial vaginosis and its association with genital micro-organisms in asymptomatic women. *J. Med. Microbiol.* **2019**, *68*, 1373–1382. [CrossRef]
- Sethi, N.; Narayanan, V.; Saaid, R.; Ahmad Adlan, A.S.; Ngoi, S.T.; Teh, C.S.J.; Hamidi, M.; WHOW Research Group. Prevalencia, factores de riesgo y resultados adversos de la vaginosis bacteriana entre mujeres embarazadas: Una revisión sistemática. *BMC Embarazo Parto* **2025**, *25*, 40. [CrossRef]
- Pérez, J.; Canul, J.; Cruz, E. Combination of Rankings as a Method for Biomarker Identification of Bacterial Vaginosis. *Res. Comput. Sci.* **2020**, *149*, 915–927. Available online: <https://dblp.org/rec/journals/rcs/Perez-GomezCD20> (accessed on 15 April 2025).
- Hernández, H.; Canul, J.; Cruz, E. An agglomerative hierarchical clustering approach to identify coexisting bacteria in groups of bacterial vaginosis patients. *Intell. Data Anal.* **2023**, *27*, 583–611. [CrossRef]
- Cruz, F.; Canul, J.; Rivera, R.; Cruz, E. Impact of data balancing a multiclass dataset before the creation of association rules to study bacterial vaginosis. *Intell. Med.* **2024**, *4*, 188–199. [CrossRef]
- Alvarez, M.; Quirós, L.; Cortés, M. Inteligencia artificial y aprendizaje automático en medicina. *Rev. Méd. Sinerg.* **2020**, *5*, e557. [CrossRef]
- Bautista Hernández, C.M.; Canul Reich, J.; López Ramírez, C.; De la Cruz Hernández, E. Un modelo de Red Bayesiana para datos cualitativos de Vaginosis Bacteriana en mujeres embarazadas. *Ideas Cienc. Ing.* **2024**, *2*, 62–77. [CrossRef]
- Sucar, L.E. *Probabilistic Graphical Models: Principles and Applications*, 2nd ed.; Springer: Cham, Switzerland, 2020; ISBN 978-3030619428.
- Sagaró, N.; Zamora, L. Métodos gráficos en la investigación biomédica de causalidad. *Rev. Electrón. Dr. Zoilo E. Mar. Vidaurreta* **2019**, *44*. Available online: <https://revzoilomarinello.sld.cu/index.php/zmv/article/view/1846> (accessed on 12 May 2025).
- Brand, Y. Modelos de Ecuaciones Estructurales: Conceptos y Aplicaciones. Ph.D. Thesis, Universidad Nacional de Colombia, Bogotá, Colombia, 2021. Available online: <https://repositorio.unal.edu.co/bitstream/handle/unal/80064/1059705148.2021.pdf> (accessed on 20 May 2025).
- Jahuey, F.; Herrera, J.; Paredes, F. El programa R: Una estrategia inicial para su entendimiento y aprendizaje. *Rev. Digit. Univ.* **2022**, *23*, 1–9. [CrossRef]
- Molina, M.; Ochoa, C.; Ortega, E. Correlación. Modelos de regresión. *Evid. Pediatr.* **2021**, *17*, 25. Available online: <https://evidenciasenpediatria.es/articulo/7827/correlacion-modelos-de-regresion> (accessed on 30 May 2025).
- Apaza, E.; Cazorla, S.; Condori, C.; Arpasi, F.; Tumi, I.; Yana, W.; Quispe, J. La Correlación de Pearson o de Spearman en caracteres físicos y textiles de la fibra de alpacas. *Rev. Inv. Vet. Perú* **2022**, *33*, e22908. [CrossRef]
- Mendivelso, F.; Rodríguez, M. Prueba no paramétrica de correlación de Spearman. *Rev. Méd. Sanitas* **2021**, *24*, 42–45. [CrossRef]

18. Lepera, A. Introducción a los Modelos de Ecuaciones Estructurales y su Implementación en R mediante un ejemplo. *Rev. Investig. Model. Matemáticos Apl. Gestión Econ.* **2021**, *1*, 15–37. Available online: <https://www.economicas.uba.ar/wp-content/uploads/2016/04/Lepera-Andrea-1.pdf> (accessed on 10 June 2025).
19. Civelek, M.E. *Essentials of Structural Equation Modeling*; Zea E-Books; University of Nebraska–Lincoln: Lincoln, NE, USA, 2018. Available online: <https://digitalcommons.unl.edu/zeabook/64> (accessed on 15 June 2025).
20. Barbeau, K.; Sarr, F.; Boileau, K.; Smith, K. Path analysis in Mplus: A tutorial using a conceptual model of psychological and behavioral antecedents of bulimic symptoms in young adults. *Quant. Methods Psychol.* **2019**, *15*, 38–53. [\[CrossRef\]](#)
21. Jordan, F. Valor de corte de los índices de ajuste en el análisis factorial confirmatorio. *Rev. Investig. Psicol. Soc.* **2021**, *7*, 66–71. Available online: <https://www.redalyc.org/articulo.oa?id=672371335005> (accessed on 26 June 2025).
22. Castillo Vargas, A.; Ruiz Sánchez, A. Entre lo observable y lo latente: Modelos de ecuaciones estructurales e investigación social. *Rev. Reflex.* **2025**, *104*, 186–213.
23. Luzuriaga, H.; Espinosa, C.; Haro, A.; Ortiz, H. Histograma y distribución normal: Shapiro-Wilk y Kolmogorov Smirnov aplicado en SPSS. *LATAM Rev. Lat. Cienc. Soc. Humanid.* **2023**, *4*, 596–607. [\[CrossRef\]](#)
24. Platas, V. Contrastes de Normalidad. Ph.D. Thesis, Universidad de Santiago de Compostela, Santiago de Compostela, Spain, 2021. Available online: <https://minerva.usc.gal/rest/api/core/bitstreams/c8efafac-12ac-446a-9443-bb3ca39fa5f0/content> (accessed on 29 June 2025).
25. Hamasha, M.; Ali, H.; Hamasha, S.; Ahmed, A. Ultra-fine transformation of data for normality. *Heliyon* **2022**, *8*, e09370. [\[CrossRef\]](#)
26. Ayala, J. Transformaciones. 2023. Available online: <https://rpubs.com/JairoAyala/TNM> (accessed on 4 July 2025).
27. Bhatia, P. *Data Mining and Data Warehousing: Principles and Practical Techniques*; Cambridge University Press: Cambridge, UK, 2019. [\[CrossRef\]](#)
28. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [\[CrossRef\]](#)
29. Madrid, A.; Valenzuela, S.; Batanero, C.; Carzón, J. Interpretación del diagrama de caja por estudiantes universitarios de ciencias de la actividad física y deporte. *Educ. Matemática* **2022**, *34*, 275–300. [\[CrossRef\]](#)
30. Bouza, C. *Análisis Exploratorios de Datos Univariados Para la Ciencia de los Datos*; ResearchGate: Berlin, Germany, 2023. [\[CrossRef\]](#)
31. Alwateer, M.; Atlam, E.; Abd, M.; Ghoneim, O.; Gad, I. Missing Data Imputation: A Comprehensive Review. *J. Comput. Commun.* **2024**, *12*, 53–75. [\[CrossRef\]](#)
32. Rashid, W.; Gupta, M.K. A Perspective of Missing Value Imputation Approaches. In *Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2019*; Springer: Singapore, 2021; Volume 1086, pp. 307–315. [\[CrossRef\]](#)
33. García, M.; Canul, J.; Rodríguez, L.; Cruz, E. Hacia un Modelo Gráfico Causal para el Análisis de la Vaginosis Bacteriana. *Ideas Cienc. Ing.* **2025**, *3*, 22–34.
34. Poaquiza, L.; Gavilanes, G. El estrés y su relación con la inteligencia emocional en docentes universitarios. *Cienc. Lat. Rev. Científica Multidiscip.* **2022**, *6*, 1357–1372. [\[CrossRef\]](#)
35. Cáceres, M.; Ponce, R. Dependencia emocional y su relación con el riesgo suicida en adultos jóvenes. *LATAM Rev. Lat. Cienc. Soc. Humanid.* **2023**, *4*, 329–339. [\[CrossRef\]](#)
36. Manzano, A. Introducción a los modelos de ecuaciones estructurales. *Investig. Educ. Méd.* **2017**, *7*, 67–72. [\[CrossRef\]](#)
37. Flora, D.; Crone, G.; Bell, S. Effect Size Interpretation in Structural Equation Models. *Struct. Equ. Model.* **2025**, *32*, 1069–1076. [\[CrossRef\]](#)
38. Glascock, A.L.; Jimenez, N.R.; Boundy, S.; Koparde, V.N.; Brooks, J.P.; Edwards, D.J.; Strauss, J.F., III; Jefferson, K.K.; Serrano, M.G.; Buck, G.A.; et al. Vaginal Microbiome Consortium. Unique roles of vaginal *Megasphaera* phylotypes in reproductive health. *Microb. Genom.* **2021**, *7*, 000526. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.