Infantile Predictors of Functional Gastrointestinal Disorders: A Machine Learning Approach to Risk Assessment

Flavia Indrio[1], Elio Masciari[2], Flavia Marchese[3], Matteo Rinaldi[4], Gianfranco Maffei[4], Enea Vincenzo Napolitano[2,*], Isadora Beghetti[5], Luigi Corvaglia[5] and Arianna Aceti[5]

[1]Department of Experimental Medicine School of Medicine, University of Salento, Lecce, Italy

[2]Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, Italy

[3]Department of Medical and Surgical Science Pediatric Section, University of Foggia, Foggia, Italy

[4]Department of Neonatology and NICU, Ospedali Riuniti Foggia, Foggia, Italy

[5]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy

Abstract

This study examines the considerable impact of Functional Gastrointestinal Disorders (FGIDs) on children, their families and healthcare systems, and highlights the historic challenge of identifying children at risk due to unclear pathophysiology. The research aims to identify early-life risk factors for FGIDs, specifically infantile colic, regurgitation, and functional constipation, within the first year of life. Using a prospective observational cohort design, the study enrolled term and preterm infants from a tertiary care university hospital in Foggia, Italy, between 1 January 2020 and 31 December 2022, excluding infants with severe disease or major neonatal complications. By using conventional statistical methods and artificial intelligence, specifically a random forest classification model, this study identified birth weight, cord blood pH, and maternal age as significant predictors for FGIDs. A logistic regression predictive model also established an inverse relationship between these variables and the occurrence of FGIDs. Using these findings, the study created an AI-based predictive model and a practical, user-friendly web interface for risk assessment. This enables clinicians to identify infants at a higher risk for FGIDs. The approach is innovative and marks a pioneering step in FGID risk prediction.

Keywords

Neonatal Health, Early Diagnosis, Risk Factors, Health Informatics,

1. Introduction

Functional Gastrointestinal Disorders (FGIDs) are a significant challenge in pediatric healthcare due to their prevalence and impact on infants. FGIDs refer to a range of conditions, including infant colic, regurgitation, functional diarrhea, and functional constipation, that are defined

by the absence of identifiable biochemical or structural anomalies. These conditions affect almost 50% of infants in their first year of life [1, 2, 3, 4, 5, 6, 7]. Functional gastrointestinal disorders cause distress and discomfort in infants and place substantial burdens on families and healthcare systems worldwide. Despite their classification based on the Rome IV criteria, the underlying pathophysiology of FGIDs remains unclear. Potential contributing factors include genetic predispositions, psychosocial stressors, and early life events such as delivery type and feeding practices [8, 9, 10, 11, 12, 13, 14]. It is important to note that any evaluation of contributing factors should be objective and clearly marked as such.

In recent years, the use of artificial intelligence (AI) and machine learning (ML) has revolutionised biomedical research, providing innovative tools for analysing complex health issues [15]. This is particularly true in paediatric healthcare, where AI and ML have the potential to transform early detection, risk assessment, and intervention strategies for FGIDs. A Machine Learning Approach to Risk Assessment aims to utilise machine learning to identify early-life predictors of FGIDs. The research analyses a comprehensive area-based cohort to identify multifaceted risk factors present during the first year of life that predispose infants to FGIDs.

Using an AI-based predictive model, our aim is to gain a detailed understanding of the early-life factors that contribute to FGIDs. This will enable the development of a practical risk assessment tool to help clinicians identify infants who are at a higher risk of developing FGIDs. This, in turn, will facilitate early and targeted interventions. Proactive measures have the potential to improve the immediate symptoms of disorders and mitigate their long-term impact on children's health and well-being.

This study aims to contribute significantly to the field of pediatric gastroenterology by using machine learning to analyze early-life factors associated with the development of FGIDs. The statement represents progress towards a predictive and preventative approach to pediatric healthcare, moving beyond the symptom-based classification of FGIDs.

2. Methods

2.1. Study Design, Participants, and Recruitment

This study, a prospective observational cohort study, was conducted at the Obstetric and Neonatal Unit of the "Ospedale Riuniti" in Foggia, Italy, from 1 January 2020 to 31 December 2023. The study adhered strictly to institutional data protection requirements, and informed consent was obtained from the legal representatives of the infants before their participation. The study's protocol was approved by the Institutional Review Board to ensure compliance with the ethical standards outlined in the Declaration of Helsinki.

The study included both term and preterm infants who met the eligibility criteria. Newborns presenting with severe clinical conditions, major neonatal complications such as inherited metabolic disorders, congenital anomalies, or those who died during hospitalization were excluded. Newborns who met the inclusion and exclusion criteria were consecutively recruited within their first few days of life, as shown in the patient enrollment flow chart 1.

The aim of the data collection was to identify potential perinatal risk factors for FGIDs. The risk factors included gestational age (GA), birth weight (BW), sex, mode of delivery, Apgar score, venous cord blood pH, maternal demographic characteristics, Neonatal Intensive Care
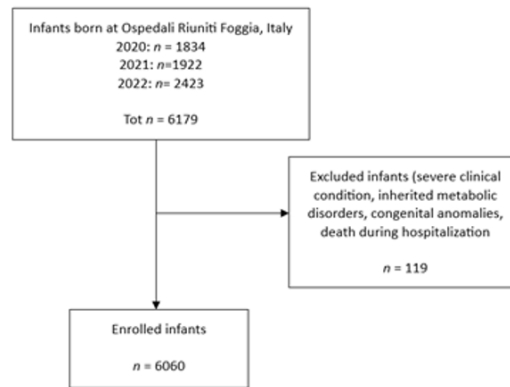
Figure 1: Patient enrollment flow chart

Unit (NICU) admission, antibiotic administration, and feeding practices at discharge. The data collection was performed using a randomised controlled trial design. Discharge feeding was categorised into exclusive breastfeeding and non-exclusive breastfeeding, which included both exclusive formula feeding and mixed feeding.

During follow-up visits in the first year of life, a dedicated pediatrician diagnosed and classified FGIDs according to the Rome IV criteria. Parents provided information on feeding practices at discharge from the nursery or NICU, and at subsequent milestones of 3, 6, and 12 months. This information included details on family history of allergic diseases and parental smoking habits.

The aim of this study was to investigate the relationship between the development of FGIDs (infantile colic, regurgitation, functional constipation) within the first year of life and various perinatal/neonatal characteristics.

2.2. Data Analyses

Potential associations between the development of each FGIDs (infantile colic, regurgitation, functional constipation) and perinatal characteristics were investigated through both conventional statistics and AI. An AI-based predictive model and a practical risk assessment tool for each FGIDs were then developed.

Differences between infants developing or non-developing each FGID were evaluated using the independent sample t test for continuous variables, and the chi-squared test for categorical data. A p value <.05 was considered as statistically significant. Statistical analyses were performed using IBM SPSS Statistics 28.0 (IBM Corp., Armonk, NY, USA).

A machine learning (ML) process was implemented for the analysis of the dataset. An accurate data preprocessing 2.2.1 was first performed to obtain a dataset suitable for ML analysis. After a cleaned dataset had been produced, it was partitioned into a training set composed of 4.242 instances and a test set composed of 1.818 instances. A Feature Selection step was performed to identify the most important variables for the output prediction. Finally, a Classification model based on Random Forest was produced.

2.2.1. Data Preprocessing

During the dataset preparation phase of our study, we prioritised the integrity and usability of the data. This required a comprehensive data cleaning process, which began with the exclusion of variables that had missing values exceeding 30% following the collection phase. We set this threshold to ensure the quality and reliability of the dataset for subsequent analysis.

Our approach to treating missing values varied based on the nature of the variables involved. Continuous variables with missing entries were imputed using the mean value of the respective variable to ensure a balanced representation of the data without introducing significant bias. Discrete variables with missing values were imputed with zero. Classification variables underwent a stricter process, where instances missing any values were entirely removed from the dataset. This strategy ensured that only complete and accurate data were included in the analysis.

To standardize the dataset and facilitate analysis, continuous variables were normalized using a standard scaler. This allowed for quantification on a uniform scale. Categorical variables were transformed using one-hot coding, which converts categories into a binary representation, simplifying their inclusion in statistical models.

After completing thorough data preparation and preliminary analysis, we selected specific variables to include in the predictive model. The variables selected for the study were chosen based on their relevance to the study objectives. These variables included birth weight (BW), term/preterm status, mode of delivery (vaginal vs. cesarean), Neonatal Intensive Care Unit (NICU) admission, sex, occurrence of twin births, maternal age, parity, 5-minute Apgar score, and feeding at discharge. The feeding at discharge variable was categorized into exclusive versus non-exclusive breastfeeding. In addition, we considered the smoking status of both parents due to its potential influence on the risk of FGIDs in infants. This selected set of variables will serve as the foundation for the development of an AI-based predictive model, which aims to shed light on the complex factors contributing to FGIDs during the first year of life.

2.3. Risk Prediction model development

Our study utilised a comprehensive approach to investigate the complex relationships between various factors and the occurrence of three specific target conditions: colic, constipation, and regurgitation. The classification models used in this study included logistic regression, support vector machine, decision tree classifier, extra tree classifier and random forest classifier. The Logistic Regression was found to be the most effective in terms of accurately representing the relationships between the variables and the target conditions. However, to address the challenge of dealing with an unbalanced dataset, both oversampling and undersampling techniques were applied to achieve a more balanced data representation. The selection of variables critical for prediction was conducted through a comparative analysis of various models under different configurations. Notably, all models consistently prioritised three variables: birth weight (BW), maternal age, and pH levels, due to their significant importance values. Depending on the specific model configuration, additional variables such as sex, mode of delivery, or the 5-minute Apgar score may have been included as either the fourth or fifth variable to improve the model's predictive accuracy. To construct the risk prediction model, we experimented with different sets

of variables. We maintained BW, maternal age, and pH as constant predictors while varying the inclusion of additional variables such as sex, mode of delivery, and the 5-minute Apgar score to create variable sets of sizes 4, 5, or 6. The model's predictive capacity was refined by identifying the most impactful combination of variables, thanks to this procedural flexibility.

The logistic regression encapsulates the foundation of our predictive models, particularly the relationship between the predictor variables (BW, maternal age, and pH) and the probability of a disorder's occurrence.

3. Results

3.1. Study Population and Clinical Outcomes

A total of 6060 infants participated in the study, with a slight male predominance (52.3%) and a twin birth rate of 6.0%. Among these infants, 488 (8.1%) were born preterm, indicating a significant proportion of the cohort experienced early birth. The delivery method statistics indicate a preference for vaginal birth, with 60.8% of infants delivered this way. The clinical evaluation conducted at birth showed a mean Apgar score of 8.9 (SD = 0.4, range 3-10) five minutes after delivery. Additionally, the mean venous cord blood pH was recorded at 7.32 (SD = 0.08, range 6.86-7.55), indicating favourable initial health outcomes for the newborns.

Within our cohort, 27.3% of infants experienced colic, 18.7% experienced regurgitation, and 10.2% experienced constipation. A deeper analysis revealed that preterm infants were significantly more likely to develop gastrointestinal conditions compared to their term counterparts. The incidences of colic, regurgitation, and constipation were higher in the preterm group (38.1% vs 25.8%, 35.8% vs 17.2%, and 21.6% vs 9.2%, respectively; p<0.001). These findings highlight the increased vulnerability of preterm infants to gastrointestinal disorders.

3.2. Model Insights

The analysis started by examining the correlation matrix to identify potential collinearity among variables in relation to the target conditions of colic, constipation, and regurgitation. This step was crucial to ensure the validity of the predictive model by excluding the possibility of collinear variables that could skew the results. The correlation matrix shows no collinearity affecting the target variables, indicating the reliability of the selected variables for further analysis.

Moderately strong positive correlations were found between certain variables, such as maternal age and parity (distinguished by nulliparous vs. multiparous mothers), as well as between the 5-minute Apgar score and birth weight (BW). The correlations suggest a relationship between the variables that could be significant in understanding infant health outcomes. However, some variables showed a moderately strong negative correlation, specifically between Neonatal Intensive Care Unit (NICU) admission and both birth weight (BW) and term birth. These inverse relationships highlight factors that may influence the likelihood of NICU admission and indicate the complex interplay of variables that affect infant health.

The results of our Random Forest Classifier provided additional insights. It identified BW, cord blood pH, and maternal age as the most influential variables in classifying the three target conditions. These variables are important in reflecting the health status of the infant and their

potential impact on the likelihood of developing colic, constipation, and regurgitation. The identification of key factors aids in understanding the investigated conditions and highlights critical variables for healthcare professionals to monitor closely. These findings offer valuable insights for clinical practice and future research into the development and diagnosis of functional gastrointestinal disorders in infants.

3.3. Risk Prediction Model

Our research took a novel approach to predicting three common infant conditions: colic, regurgitation, and constipation. We used a unified predictive modeling framework and carefully selected predictor variables. The selection process resulted in the identification of birth weight (BW), maternal age, and cord blood pH as the primary predictors, distinguished by their pronounced importance coefficients relative to other variables.

Three distinct yet interconnected predictive models were developed, each tailored to one of the conditions under study. Although each model is specific to a particular condition, a common set of predictor variables was incorporated across all models to comprehensively examine their interrelationships and impacts on different health outcomes. This approach ensured both efficiency and effectiveness of the models, while also providing insights into the nuanced roles played by the predictors in the context of infant health.

3.3.1. Model insights and findings

- The Colic Prediction Model revealed that higher values in birth weight, maternal age, and cord blood pH have a protective effect against colic onset. The model emphasises the significant role of pH levels in colic risk assessment, suggesting its potential as a key factor in preventive strategies.
- The Regurgitation Prediction Model showed a pronounced inverse relationship between the occurrence of regurgitation and the predictors, with birth weight and pH identified as the most substantial protective factors. It is important to closely monitor these variables in newborns to reduce the risk of regurgitation.
- The Constipation Prediction Model identified birth weight and cord blood pH as crucial predictors negatively associated with the risk of constipation, with maternal age playing a lesser role. The consistent significance of pH levels across all conditions emphasizes its relevance in the early identification and management of gastrointestinal issues in infants.
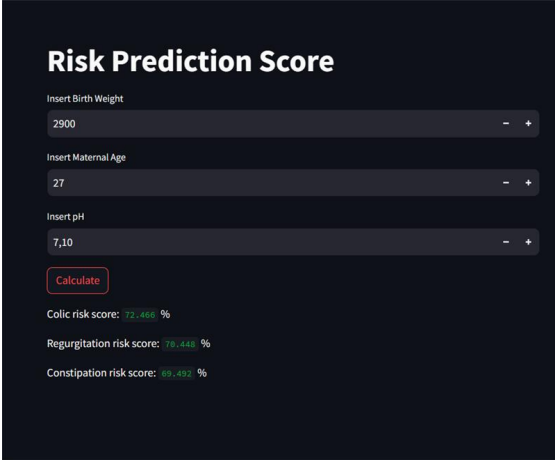
The use of a unified predictive modelling approach in our study is a significant advancement in paediatric healthcare, especially in the early detection and management of colic, regurgitation, and constipation in infants. Our models focus on a specific set of predictors, while also simplifying the interpretation and application of the findings. This methodology enables healthcare providers to gain a better understanding of the complex dynamics of these prevalent conditions and to implement more targeted and effective intervention strategies, ultimately resulting in improved infant health outcomes.

3.4. Risk prediction tool

The research team has developed a user-friendly web interface that translates insights from predictive models for infant colic, regurgitation, and constipation into a practical tool 2. This digital application streamlines the process of risk prediction by allowing users to input key variables such as birth weight (BW), maternal age, and cord blood pH, which have been identified as significant predictors of these conditions. The web application calculates a prediction score to reflect the probability of an infant developing any of the three disorders.

A risk stratification mechanism interprets the scores into three distinct risk categories. Scores under 33% are classified as low risk, indicating a minimal likelihood of the disorders. Scores between 33% and 66% are deemed medium risk, suggesting a moderate probability. Finally, scores above 66% are categorised as high risk, indicating a significant chance of occurrence. This categorisation helps to provide a clear and actionable evaluation of risk levels.

The development of this web-based tool simplifies complex statistical models for risk assessment and supports early identification and management of infant health issues. This approach aims to enhance preventative and diagnostic processes, ultimately contributing to better health outcomes for infants.

Figure 2: Risk prediction score web interface

4. Discussion and Conclusions

This study represents a significant step forward in the comprehension and treatment of functional gastrointestinal disorders in neonates and toddlers. We identified key risk factors for FGIDs using both conventional statistics and machine learning. The Functional Risk Index for Pediatric Subjects (FRIPS) is a novel, machine learning-based predictive model for early diagnosis of FGIDs in children. Healthcare practitioners can input patient data and receive risk coefficients for colic, regurgitation, and constipation, empowering them to make informed clinical decisions. This work should be validated on different populations.

The FRIPS predictive scores reflect the probability of developing any of the three conditions, providing a nuanced understanding of risk beyond simple binary outcomes. This approach employs an equation that considers the interplay between various factors rather than isolated variables, underscoring the complexity of FGIDs. The application of ML for feature selection has identified birth weight, cord blood pH, and maternal age as critical variables, indicating their significant cross influence on disease occurrence.

The incidence of FGIDs in our study population, particularly among preterm infants, is consistent with previous research. However, the reported incidence rates vary due to differences in diagnostic criteria and population stratification. Our findings indicate that preterm infants are especially susceptible to FGIDs, likely due to the crucial developmental processes that occur during the perinatal period, which affect the brain-gut-microbiota axis.

Furthermore, our research confirms that low venous cord blood pH is a risk factor for FGIDs. This suggests that increased surveillance for infants with low pH at birth could enable early detection and intervention, potentially mitigating the risk of FGIDs in the first year of life. This insight is particularly relevant given the established link between neonatal acidaemia and neurological issues, as well as the emerging evidence of its impact on gastrointestinal health.

Despite the contributions made, we acknowledge the limitations of our study, particularly the sample size, which may not fully represent the global population. Future research should aim to validate and refine FRIPS across diverse populations to enhance its applicability and accuracy in predicting FGIDs.

In conclusion, this study sheds light on the complex etiology and risk factors associated with functional gastrointestinal disorders (FGIDs) in neonates and toddlers. It also offers a practical tool for early diagnosis and management. The study integrates Machine Learning with traditional statistical methods, providing a robust framework for enhancing pediatric healthcare outcomes and improving the quality of life for children and their families worldwide.

Acknowledgments

References

[1] J. Hyams, C. Di Lorenzo, M. Saps, R. Shulman, A. Staiano, M. van Tilburg, Functional disorders: children and adolescents. gastroenterology, 2016.

[2] Y. Vandenplas, B. Hauser, S. Salvatore, Functional gastrointestinal disorders in infancy: impact on the health of the infant and family, Pediatric gastroenterology, hepatology & nutrition 22 (2019) 207–216.

[3] M. A. Benninga, S. Nurko, C. Faure, P. E. Hyman, I. S. J. Roberts, N. L. Schechter, Childhood functional gastrointestinal disorders: neonate/toddler, Gastroenterology 150 (2016) 1443–1455.

[4] Y. Vandenplas, A. Abkari, M. Bellaiche, M. Benninga, J. P. Chouraqui, F. Çokura, T. Harb, B. Hegar, C. Lifschitz, T. Ludwig, et al., Prevalence and health outcomes of functional gastrointestinal symptoms in infants from birth to 12 months of age, Journal of pediatric gastroenterology and nutrition 61 (2015) 531–537.

[5] L. A. Lestari, A. N. Rizal, W. Damayanti, Y. Wibowo, C. Ming, Y. Vandenplas, Prevalence and risk factors of functional gastrointestinal disorders in infants in indonesia, Pediatric Gastroenterology, Hepatology & Nutrition 26 (2023) 58.

[6] N. F. Steutel, J. Zeevenhooven, E. Scarpato, Y. Vandenplas, M. M. Tabbers, A. Staiano, M. A. Benninga, Prevalence of functional gastrointestinal disorders in european infants and toddlers, The Journal of pediatrics 221 (2020) 107–114.

[7] A. Chogle, C. A. Velasco-Benitez, I. J. Koppen, J. E. Moreno, C. R. R. Hernández, M. Saps, A population-based study on the epidemiology of functional gastrointestinal disorders in young children, The Journal of pediatrics 179 (2016) 139–143.

[8] M. A. van Tilburg, P. E. Hyman, L. Walker, A. Rouster, O. S. Palsson, S. M. Kim, W. E. Whitehead, Prevalence of functional gastrointestinal disorders in infants and toddlers, The Journal of pediatrics 166 (2015) 684–689.

[9] G. Holtmann, A. Shah, M. Morrison, Pathophysiology of functional gastrointestinal disorders: a holistic overview, Digestive Diseases 35 (2018) 5–13.

[10] I. Koppen, M. Benninga, M. Singendonk, Motility disorders in infants, Early human development 114 (2017) 1–6.

[11] R. Shamir, I. St James-Roberts, C. Di Lorenzo, A. J. Burns, N. Thapar, F. Indrio, G. Riezzo, F. Raimondi, A. Di Mauro, R. Francavilla, et al., Infant crying, colic, and gastrointestinal discomfort in early childhood: a review of the evidence and most plausible mechanisms, Journal of pediatric gastroenterology and nutrition 57 (2013) S1.

[12] S. Salvatore, M. E. Baldassarre, A. Di Mauro, N. Laforgia, S. Tafuri, F. P. Bianchi, E. Dattoli, L. Morando, L. Pensabene, F. Meneghin, et al., Neonatal antibiotics and prematurity are associated with an increased risk of functional gastrointestinal disorders in the first year of life, The Journal of pediatrics 212 (2019) 44–51.

[13] M. M. B. B. Gondim, A. L. Goulart, M. B. d. Morais, Prematurity and functional gastrointestinal disorders in infancy: a cross-sectional study, Sao Paulo Medical Journal 140 (2022) 540–546.

[14] D. Bi, H. Jiang, K. Yang, T. Guan, L. Hou, G. Shu, Neonatal risk factors for functional gastrointestinal disorders in preterm infants in the first year of life (2022).

[15] E. V. Napolitano, S. Fioretto, E. Masciari, A. Anniciello, How pandemic affected the adoption of e-health systems, in: Proceedings of the 27th International Database Engineered Applications Symposium, 2023, pp. 94–98.