PROCTO-CLUSTER v.1.2 (20/01/2026)

Title:
Data-driven clustering in hemorrhoid surgery: retrospective monocentric study for the identification of clinical phenotypes

Principal Investigator:

Dr. Angelo Stuto

Co-Investigators:

Dr. Carlotta La Raja
Dr. Francesca Lecchi
Dr. Giorgio La Greca
Dr. Paola Cellerino

Sponsor:

IRCCS Policlinico San Donato

Index

1.0 Introduction

Hemorrhoidal disease (HD) is one of the most common benign anorectal conditions worldwide, with lifetime prevalence estimates reaching up to one-third of the adult population.[1,2] It represents a substantial proportion of proctologic consultations and elective anorectal procedures, spanning from outpatient management of symptomatic internal hemorrhoids to complex surgery in patients with advanced prolapse or recurrent disease. Despite its benign nature, HD imposes a significant burden on healthcare systems, generating high volumes of outpatient visits, diagnostic procedures, day-surgery activity, and postoperative follow-up.[2,3]

Clinically, HD is a heterogeneous condition rather than a single, uniform entity.[1] Patients may present with very different symptom constellations, including bleeding, prolapse, pain, itching, soiling, and discomfort, often in association with constipation, pelvic floor dysfunction, or other anorectal disorders.[4] Anatomically, variations in the distribution and size of hemorrhoidal cushions, degree and circumferential extent of prolapse, associated skin tags, and coexisting anal pathology (e.g., fissures, fistulas) further contribute to this heterogeneity. Additionally, local inflammatory patterns, vascular congestion, and tissue fragility may vary significantly between patients, potentially influencing both the natural course of the disease and the response to surgical treatment.[5]

The impact of HD on quality of life is frequently underestimated.[6] Persistent bleeding and prolapse can lead to embarrassment, activity limitation, and social withdrawal, while pain and discomfort may interfere with work productivity, mobility, and sleep. In severe or recurrent cases, patients may experience chronic symptoms for years, with cumulative effects on physical, psychological, and sexual health.[6] Standard clinical scores and grading systems only partially capture this multidimensional burden, and patient-reported outcomes are not routinely integrated into decision-making in everyday practice.

From a health-service perspective, HD generates substantial resource utilization.[7] The same “degree” of HD can be treated with different techniques, such as conventional hemorrhoidectomy, stapled procedures, Doppler-guided ligation, or combined approaches, mainly based on surgeon preference and local expertise rather than on standardized risk stratification.[4–7] This variability in indications and techniques can translate into heterogeneous postoperative trajectories, including differences in pain, complications, and recurrence, with direct implications for costs, length of stay, unplanned readmissions, and time away from work. The widespread reliance on traditional degree-based classifications implicitly assumes that these systems adequately reflect the true clinical complexity of HD and could guide optimal therapeutic choices.[3] However, growing clinical experience suggests that patients

with the same nominal “grade” may differ substantially in symptom burden, comorbidity profile, anatomical features, and surgical risk, indicating that current grading systems are too coarse to support genuinely personalized care.

The most widely used framework for categorizing hemorrhoidal disease remains the Goligher classification, a system introduced more than half a century ago and based exclusively on the extent of prolapse during defecation.[8,9] Although pragmatic and easy to apply, the Goligher scale is inherently reductionist: it captures only a single anatomical dimension of a multifactorial condition and offers no insight into symptom severity, comorbidity burden, inflammatory status, or patient-specific risk factors. As a consequence, it provides a narrow and often misleading representation of the clinical complexity encountered in contemporary proctologic practice.[8,9] A fundamental limitation of the Goligher system is its inability to incorporate the symptomatic heterogeneity that characterizes hemorrhoidal disease. Patients with the same nominal degree may experience vastly different levels of bleeding, pain, prolapse-related discomfort, pruritus, or soiling. These symptoms, in turn, do not correlate consistently with the degree of prolapse, making the scale poorly reflective of patient-reported burden and insufficient for guiding therapeutic decisions. Similarly, the classification does not account for comorbidities, such as obesity, chronic constipation, anticoagulation, pelvic floor dysfunction, or inflammatory bowel conditions, that may modify both the natural history of the disease and the risk profile associated with surgical interventions.

Anatomical and morphological variations, including circumferential involvement, presence of external components, associated fissures, and tissue fragility, are likewise omitted from the Goligher framework, despite their known implications for technical planning and postoperative outcomes.[9] The underlying inflammatory phenotype, characterized by degrees of vascular congestion, edema, or mucosal fragility, also remains unmeasured, even though it may partly explain differences in postoperative bleeding, pain trajectories, and recurrence risk. Another critical omission is the variability in postoperative recovery patterns: pain, time to return to normal activity, and complication rates show substantial interpatient variation that is not predicted by prolapse grade alone.[9] Increasing evidence underscores the weak correlation between the Goligher degree and clinically meaningful outcomes.[10] Studies have demonstrated that the degree of prolapse does not reliably predict the severity of symptoms, responsiveness to office-based procedures, recurrence after surgery, or the intensity of postoperative pain.[9,10] As a result, patients with “similar degrees” often behave differently across the entire continuum of care, from symptom presentation to surgical indication, operative complexity, and recovery course, highlighting a persistent mismatch between the current classification and real-world patient trajectories.

The limitations of degree-based classifications highlight a broader methodological gap in the conceptualization and management of HD.[8–10] In current practice, treatment decisions are frequently guided by a coarse anatomical descriptor (i.e., prolapse extent) rather than by a nuanced understanding of the underlying clinical phenotypes.[11] This stands in contrast with recent developments in other surgical and medical fields, such as colorectal surgery, oncology, bariatric and cardiovascular care, where phenotype-driven approaches have been increasingly embraced to improve risk stratification, tailor interventions, and personalize postoperative management.[12] Patients may cluster along distinct trajectories, for example: bleeding-predominant, prolapse-dominant, pain-dominant, mixed symptomatic, or comorbidity-driven presentations, each potentially associated with different responses to treatment and different postoperative courses. However, these latent clinical subgroups remain largely unexplored because current frameworks do not systematically capture the multidimensionality of the disease.[1,4–6,13]

A more precise characterization of HD is clinically relevant for several reasons. First, better phenotyping could improve surgical indication: identifying which patients benefit from conventional hemorrhoidectomy versus minimally invasive approaches would reduce overtreatment in some cases and undertreatment in others. Second, phenotype-aware planning could help anticipate perioperative complexity, for example, higher bleeding risk in inflammatory or vascular-congestive profiles, or higher postoperative pain in patients with external components, fibrotic changes, or associated anal fissures. Third, postoperative trajectories vary widely, and understanding phenotype-dependent recovery patterns (e.g., pain duration, risk of edema, risk of thrombosis, recurrence probability) has direct implications for patient counseling, work leave planning, and resource allocation. Critically, the assumption that anatomical degree alone determines clinical behavior has obscured the possibility that hemorrhoidal disease contains biologically and clinically meaningful subtypes. Patients classified as “Grade III” may follow completely different clinical paths depending on factors that remain invisible to the current classification, such as bowel habits, pelvic floor coordination, chronic straining, comorbidity burden, or local inflammatory state. The absence of a standardized, multidimensional phenotyping framework prevents proctologists from anticipating this heterogeneity and limits opportunities to move toward precision surgery. In this context, identifying reproducible clinical phenotypes is not a purely exploratory exercise but a necessary step toward modernizing proctologic care.

A phenotype-based approach has the potential to: (i) improve the granularity of preoperative assessment; (ii) guide technique selection on the basis of patient rather than surgeon factors; (iii) enhance prognostication about postoperative pain, bleeding, or recurrence; (iv) provide a foundation for developing predictive models for surgical outcomes; (v) and, in the long term, support the formulation of evidence-based, personalized care pathways. The current gap between the

complexity of real-world hemorrhoidal disease and the simplicity of existing classification systems, therefore, represents a significant unmet need. This provides the rationale for adopting data-driven, unsupervised methods as a means of capturing the multidimensional nature of HD and revealing clinically meaningful phenotypes that are currently undetected by traditional scoring systems.

1.1 Study Aim

The overarching aim of this study is to identify and characterize distinct clinical phenotypes among patients undergoing surgical treatment for hemorrhoidal disease by applying unsupervised machine learning techniques to routinely collected perioperative data. The study aims to identify latent patient subgroups that are not captured by traditional degree-based classifications and may be relevant for surgical planning and predicting postoperative outcomes.

Primary objective:
To assess the feasibility and clinical interpretability of a fully unsupervised clustering pipeline, based on t-distributed stochastic neighbor embedding (t-SNE) for dimensionality reduction followed by k-means clustering, to derive reproducible, data-driven phenotypes of hemorrhoidal disease from real-world perioperative datasets.

Secondary objectives:

- To describe the clinical, anatomical, and perioperative characteristics that differentiate the identified phenotypes, including symptom profiles, comorbidity burden, bowel habits, anatomical variations, and surgical technique selection.

- To evaluate whether the emergent clusters correspond to different levels of perioperative complexity, as reflected by operative duration, intraoperative events, and technical modifications performed during surgery.

- To compare postoperative outcomes across clusters, including pain intensity trajectories, early and late complications (classified according to Clavien–Dindo), time to return to routine activities, and recurrence patterns at short-term follow-up.

- To assess the contribution of demographic, clinical, and surgical variables to cluster formation, exploring their relative importance in defining latent phenotypes through post-hoc feature relevance analyses.

- To examine the internal validity and stability of the clustering solution, including silhouette coefficients, cluster separation metrics, and reproducibility across multiple seeds and perplexity settings.

- To generate hypothesis-driven insights for future predictive modeling, identifying candidate variables or phenotypic profiles that may inform individualized surgical planning, risk stratification, and postoperative counseling.

2.0 Materials and Methods

2.1 Study Design

This study is designed as a retrospective, observational, monocentric analysis conducted at the IRCCS Policlinico San Donato, a high-volume tertiary referral center for proctologic surgery. The study utilizes routinely collected perioperative data from all eligible patients who underwent surgical treatment for hemorrhoidal disease within the predefined study period. No intervention, modification of clinical practice, or additional patient contact was required.

Data extraction was conducted according to a predefined protocol to ensure consistency, completeness, and adherence to Good Clinical Practice and institutional data protection standards. The retrospective nature of the study allows for the analysis of real-world surgical variability and postoperative outcomes, providing an ideal framework for evaluating the feasibility and interpretability of unsupervised machine learning methodologies applied to heterogeneous clinical datasets.

2.2 Study Population

The study population consists of all consecutive adult patients who underwent surgical treatment for hemorrhoidal disease at the IRCCS Policlinico San Donato between December 2024 and June 2025. Consecutive-case sampling minimizes selection bias, enhances representativeness of routine clinical practice, and ensures

that the derived phenotypes reflect the full spectrum of hemorrhoidal disease severity and clinical presentations seen in surgical settings.

Inclusion criteria:
Patients were eligible for inclusion if they met all of the following criteria:

- Age ≥ 18 years
- Clinical and/or surgical diagnosis of hemorrhoidal disease, documented by the treating surgeon
- Complete perioperative records available, including demographic, clinical, surgical, and postoperative data necessary for machine learning analyses

Exclusion criteria:
Patients were excluded if any of the following conditions were present:

- Incomplete or missing clinical data preventing reliable integration into the analytic pipeline
- Neoplastic anorectal conditions (e.g., anal carcinoma, rectal cancer), given their distinct pathophysiology, management, and postoperative course
- History of anal surgery within the previous 6 months, due to potential confounding effects on anatomy, symptomatology, and recovery patterns.

These criteria were selected to ensure a homogeneous dataset suitable for unsupervised clustering and to avoid biases related to incomplete information or comorbid anorectal conditions with different disease trajectories.

2.3 Data sources and data collection

As this is a retrospective observational study, all data were obtained exclusively from existing institutional clinical records. Specifically, information was sourced from the electronic health record (EHR) system and the internal perioperative database used routinely at the IRCCS Policlinico San Donato for documentation of preoperative assessments, operative reports, postoperative monitoring, and scheduled follow-up visits. These data were originally collected for clinical care and administrative purposes, not for research. Prior to extraction for analysis, all records underwent a standardized anonymization process to remove direct identifiers (e.g., name, date of birth, contact information) and to re-code indirect identifiers in compliance with institutional GDPR procedures. No new data were generated, and no additional patient contact occurred. Only de-identified variables relevant to the study aims were included in the analytic dataset, ensuring full adherence to the methodological requirements of retrospective research and to the applicable data-protection framework.

Data for this study were extracted from an anonymized institutional database prospectively maintained at the IRCCS Policlinico San Donato. The database captures all routinely collected perioperative and follow-up variables for patients undergoing proctologic surgery and is curated through standardized data-entry procedures implemented by the clinical team. For the purposes of this retrospective analysis, all data were accessed only after irreversible anonymization, with removal of direct and indirect identifiers in accordance with institutional data-protection policies.
The analytic dataset encompassed four major domains:

1. Demographic and anthropometric variables
These included age, sex, and body mass index (BMI). These data were systematically recorded during the preoperative assessment and served as essential baseline descriptors that potentially influenced clinical presentation, surgical indications, and postoperative trajectories.
2. Clinical history and symptom profile
A detailed set of preoperative clinical variables was collected, including:

- comorbidities (e.g., cardiovascular disease, metabolic disorders, gastrointestinal conditions),
- current medications (e.g., anticoagulants, steroids, analgesics),
- history of previous surgeries (abdominal, pelvic, or anorectal),
- presenting symptoms (bleeding, prolapse, pain, pruritus, soiling),
- duration and pattern of symptoms, including chronicity, episodic exacerbations, and bowel habit characteristics.

3. Surgical characteristics
All intraoperative data were derived from standardized operative reports and included:

- type of procedure performed (standard hemorrhoidectomy, advanced hemorrhoidectomy, prolapsectomy, Doppler-guided procedures, or combined techniques),
- presence of external components or associated anorectal pathology addressed during surgery,
- operative duration and intraoperative events where available.

4. Postoperative outcomes and follow-up
Postoperative variables included:

- complications classified according to the Clavien–Dindo grading system,
- symptom persistence or resolution at early (1 month) and intermediate (6 months) follow-up,

- indicators of pain, bleeding, edema, or recurrence when documented,
- need for re-evaluation, medical therapy, or additional interventions.
  These outcomes were extracted from follow-up visits performed as part of routine postoperative care.

All data were reviewed for completeness and internal consistency prior to analysis. De-identified records were stored in a secure research environment, ensuring full compliance with Good Clinical Practice (GCP), GDPR, and institutional policies. Only anonymized variables were imported into the machine learning workflow.

2.4 Statistical and Machine Learning Analysis

All statistical analyses will be conducted on an anonymized retrospective dataset extracted from institutional clinical records. The final cohort will consist of approximately 100 patients, selected based on the availability of comprehensive perioperative data. Prior to analysis, data will be inspected for completeness, distributional characteristics, and potential inconsistencies. Continuous variables will be summarized using means and standard deviations or medians and interquartile ranges, depending on their distribution, while categorical variables will be described using frequencies and percentages. Missing values will be quantified for each variable, and cases with insufficient perioperative information will be excluded based on predefined criteria. No imputation procedures will be applied, given the retrospective nature of the dataset and the primary aim of assessing feasibility and identifying potential clinical phenotypes through exploratory clustering. A formal sample size calculation will not be performed, as the study is exploratory in nature and not intended to test predefined hypotheses.

The machine learning workflow will follow a fully unsupervised analytic approach designed to identify latent clinical phenotypes. All eligible variables—demographic, clinical, anatomical, surgical, and postoperative—will be standardizedto ensure comparability across scales. A dimensionality reduction step will be performed using t-distributed stochastic neighbor embedding (t-SNE), initialized with principal component analysis (PCA) to enhance stability and reduce noise.14,15 t-SNE will be applied to project high-dimensional clinical data into a two-dimensional embedding while preserving the local structure essential for phenotype discovery. Hyperparameters (including perplexity) will be selectedthrough an iterative process of inspecting embedding stability and cluster separation. Following dimensionality reduction, k-means clustering will be applied to the t-SNE coordinates to identify homogeneous patient groups. Multiple candidate values for k (range 2–10) will be explored.

The optimal number of clusters will be determined through a combination of internal validity metrics, primarily silhouette coefficients, alongside consideration of clinical interpretability and cluster stability across multiple random seeds. The resulting clusters will be treated as emergent clinical phenotypes. After cluster assignment, comparative analyses will be conducted to characterize differences across phenotypes. Continuous variables will be compared using ANOVA or Kruskal–Wallis tests, and categorical variables will be analyzed using chi-square or Fisher's exact tests, as appropriate. Given the exploratory aim of the study, emphasis will be placed on effect sizes and observed patterns rather than on inferential statistical significance alone.

All analyses will be performed using R, Orange Data Mining, and SPSS v.28 in a controlled analytic environment that ensures reproducibility through fixed random seeds, version tracking, and complete documentation of all parameter settings. This workflow reflects the primary feasibility objective of applying unsupervised machine learning to real-world perioperative data to identify latent and clinically meaningful phenotypes in hemorrhoidal disease.

2.5 Study Limitations related to external validity

This study has several inherent limitations that should be taken into account when interpreting the findings. First, its retrospective and monocentric design may limit generalizability. All data were derived from a single high-volume tertiary center, and patterns observed in this population may not fully reflect those encountered in different clinical settings, healthcare systems, or patient demographics. The retrospective nature of the analysis also restricts control over data completeness and variable standardization, despite the use of predefined extraction procedures.

Second, the study does not include external validation or prospective follow-up, which constrains the ability to confirm the reproducibility and predictive value of the identified phenotypes. While unsupervised clustering can reveal latent structures within perioperative data, validation in independent cohorts and within prospective designs is necessary to determine whether these phenotypes are stable, clinically actionable, and predictive of long-term outcomes, such as recurrence, chronic symptoms, or the need for reintervention.

Finally, postoperative outcomes were limited to routinely collected variables at short- and intermediate-term follow-up, and finer-grained patient-reported outcomes were not available. These constraints reflect the real-world nature of the dataset but may limit the depth of phenotype characterization. Despite these limitations, the study

provides an important proof-of-concept for applying unsupervised machine learning to characterize clinical heterogeneity in hemorrhoidal disease.

3.0 Ethical considerations

This study was conducted in accordance with the principles outlined in the Declaration of Helsinki (2013 revision), Good Clinical Practice (ICH-GCP E6[R3]), and relevant European and national regulations governing retrospective research. Because the analysis relies exclusively on previously collected clinical data and involves no direct patient contact or modification of care, no additional procedures or interventions were performed.

All data were extracted only after irreversible anonymization, which involved removal of all direct identifiers (e.g., name, date of birth, contact details) and recoding of indirect identifiers in compliance with the General Data Protection Regulation (GDPR, EU Regulation 2016/679) and Italian privacy legislation (Legislative Decree 196/2003, as amended by Legislative Decree 101/2018). The key linking codes to patient identity are maintained solely within the clinical system and are not accessible to the research team.

Given the retrospective design and full anonymization of data, informed consent was not required, in accordance with institutional and national guidelines for secondary use of health data for research purposes. The study was reviewed and acknowledged by the institutional governance structures responsible for data protection and research oversight at IRCCS Policlinico San Donato.

No identifiable information will appear in any presentation, publication, or external dissemination of the results. All analyses were performed on de-identified datasets stored on secure institutional servers with restricted access, encrypted storage, and routine audit trails.

These safeguards ensure that the project meets all ethical and regulatory requirements for retrospective observational research involving healthcare data.

4.0 Data Protection

All data processing activities associated with this study will comply with the General Data Protection Regulation (GDPR, EU Regulation 2016/679), the Italian Privacy Code (Legislative Decree 196/2003, as amended by Legislative Decree 101/2018), and institutional policies governing the secondary use of clinical data for research. Because this is a retrospective observational study, all data will be accessed exclusively after irreversible anonymization, ensuring that neither the research team nor any external entity will be able to directly or indirectly re-identify individual patients.

Personal Data will be processed in accordance with art. 6.1.e and art. 9.2.j of the GDPR in conjunction with art. 110bis, c.4.
This rule allows IRCCS, as the Promoter, to carry out scientific research activities by further processing personal data previously collected for clinical activity.
The information relating to the processing of personal data for the retrospective observational study will be made available to participants through the publication of the Information form on the Promoter's website.
A Data Protection Impact Assessment (DPIA) will be prepared. The DPIA will be published on the Promoter's website.

Anonymization procedures will include the removal of all direct identifiers (e.g., names, birth dates, contact information) and the transformation of indirect identifiers through de-identification protocols validated by the institution’s Data Protection Officer (DPO). A unique study identification code will be assigned to each record in the analytic dataset, with the correspondence to original patient identifiers retained solely within the clinical information system and kept inaccessible to the research team.

The anonymized dataset will be stored on secure institutional servers at IRCCS Policlinico San Donato, protected by encrypted access, multi-factor authentication, and routine automated backups. Access rights will be restricted to authorized study personnel only, applying the principle of least privilege. All data handling and analytic operations will generate audit trails documenting access, modification, and export events.

No data will be transferred outside the institution, and no cloud-based or third-party storage services will be used. Analytical workflows in R, Orange Data Mining, and SPSS v.28 will be executed within a controlled computing environment operating on institutional infrastructure. Only aggregate results, summary statistics, and anonymized tables will be disseminated.

Given the anonymized nature of the dataset, the study will pose minimal risk to patient privacy. All procedures will be reviewed by the institutional DPO to ensure

compliance with the regulatory framework for retrospective research involving health data.

5.0 Data Management and Record Keeping

All study-related data will be managed in accordance with institutional data governance policies and international standards for observational research. Data extraction, cleaning, and preparation will be overseen by the Principal Investigator and an authorized data manager with expertise in handling clinical databases. The data manager will be responsible for ensuring the integrity, accuracy, and completeness of the dataset before analysis, including verifying variable ranges, conducting internal consistency checks, and cross-referencing with source documents as needed.

All documentation related to data management, including the data dictionary, coding schemes, preprocessing logs, clustering parameter configurations, and statistical analysis scripts, will be maintained in a secure institutional repository. Version control procedures will be applied to all scripts and derived datasets, ensuring full traceability of analytic decisions. Audit trails will be automatically generated to document access and modifications to the anonymized dataset.

The anonymized working dataset, together with all associated metadata (analysis logs, codebooks, and script files), will be stored on encrypted servers at IRCCS Policlinico San Donato with restricted, role-based access. No data will be stored on personal devices or external cloud platforms. Backup procedures will be executed automatically daily, and all files will be archived under controlled conditions.

All study records will be retained for a minimum of 10 years following study completion, in alignment with institutional requirements and international standards for data retention in clinical research. After the retention period, data and supporting files will be securely destroyed or permanently anonymized in accordance with institutional policies and GDPR-compliant procedures.

6.0 Bias Assessment related to internal validity

Given the retrospective nature of the study, several methodological risks and potential sources of bias are anticipated. First, the reliance on routinely collected clinical data introduces the possibility of misclassification bias, particularly in variables derived from operative notes or symptom descriptions. Although

standardized documentation procedures are routinely used in the institution, variability in clinical reporting cannot be entirely excluded.

Second, missing data represent an inherent limitation of retrospective studies. Missingness may occur in demographic variables, symptom profiles, or postoperative assessments, potentially leading to reduced cluster stability or biased comparisons between phenotypes. Because no imputation will be performed, cases with insufficient perioperative data will be excluded, which may introduce selection bias. The extent and pattern of missing data will be evaluated and reported, and sensitivity analyses will be performed when appropriate.

Third, the study design does not allow for control of confounding variables, such as bowel habits, lifestyle factors, or unmeasured clinical characteristics that may influence both surgical decisions and postoperative outcomes. Although clustering methods can help reveal latent structure, they cannot fully account for the confounding inherent in routine clinical practice.

Fourth, limitations intrinsic to EHR extraction must be acknowledged. Data availability is restricted to what was documented during routine care, and important patient-reported outcomes or detailed anatomical descriptors may be underrepresented or absent. Furthermore, variation in follow-up intensity may lead to underdetection of some postoperative events.

Finally, the monocentric nature of the study may limit external generalizability. Patterns identified in this cohort may not fully reflect those in different populations, institutions, or healthcare systems.

Despite these aspects, the study is expected to generate clinically meaningful insights and provide a foundational feasibility assessment for the application of unsupervised machine learning to hemorrhoidal disease. Transparent reporting of biases and methodological constraints will enhance interpretability and guide future multicenter or prospective validation efforts.

7.0 Publication Policy

The findings from this study will be disseminated through submission to peer-reviewed scientific journals and presentation at national and international conferences relevant to proctology, colorectal surgery, and data-driven clinical research. All dissemination activities will adhere to the International Committee of Medical Journal Editors (ICMJE) criteria for authorship and responsible reporting. Only fully anonymized and aggregated data will be reported. No identifiable patient

information, images, or data elements that could facilitate re-identification will be included in any publication or presentation. Results from the clustering analysis, postoperative outcomes, and perioperative characteristics will be presented exclusively at the group level. The Principal Investigator will coordinate manuscript preparation, with authorship allocated based on scientific contribution and in accordance with institutional policies. Any analytic code, supplementary documentation, or methodological appendices submitted with manuscripts will include only synthetic or fully anonymized content. The final dataset and all analysis scripts will be archived securely on institutional servers to ensure reproducibility and potential follow-up analyses. Data sharing with external researchers will not occur unless explicitly authorized by the institutional DPO and the Ethics Committee, and only after ensuring complete anonymity and full regulatory compliance.
All data and results will be property of the IRCSS Policlinico San Donato.

8.0 References

1. Ashburn JH. Hemorrhoidal Disease: A Review. *JAMA*. 2025;334(17):1541. doi:10.1001/jama.2025.13083

2. Al-Masoudi RO, Shosho R, Alquhra D, Alzahrani M, Hemdi M, Alshareef L. Prevalence of Hemorrhoids and the Associated Risk Factors Among the General Adult Population in Makkah, Saudi Arabia. *Cureus*. 2024;16(1):e51612. doi:10.7759/cureus.51612

3. Wang L, Ni J, Hou C, et al. Time to change? Present and prospects of hemorrhoidal classification. *Front Med*. 2023;10:1252468. doi:10.3389/fmed.2023.1252468

4. Roberts K. What Are Hemorrhoids? *JAMA*. Published online November 6, 2025. doi:10.1001/jama.2025.17253

5. Pata F, Gallo G, Pellino G, et al. Evolution of Surgical Management of Hemorrhoidal Disease: An Historical Overview. *Front Surg*. 2021;8:727059. doi:10.3389/fsurg.2021.727059

6. Rørvik HD, Davidsen M, Gierløff MC, Brandstrup B, Olaison G. Quality of life in patients with hemorrhoidal disease. *Surgery Open Science*. 2023;12:22-28. doi:10.1016/j.sopen.2023.02.004

7. Božović B, Radoičić M, Janković S, Anđelković J, Kostić M. Pharmacoeconomic Aspects of Treating Hemorrhoidal Disease-Cost of Illness Study Based on Data from Balkan Country with Recent History of Social and Economic Transition. *Iran J Public Health*. 2021;50(6):1288-1290. doi:10.18502/ijph.v50i6.6433

8. Dekker L, Han-Geurts IJM, Grossi U, Gallo G, Veldkamp R. Is the Goligher classification a valid tool in clinical practice and research for hemorrhoidal

disease? *Tech Coloproctol*. 2022;26(5):387-392. doi:10.1007/s10151-022-02591-3

9. Van Oostendorp JY, Grossi U, Hoxhaj I, et al. Limitations of the Goligher classification in randomized trials for hemorrhoidal disease: a qualitative systematic review of selection criteria. *Tech Coloproctol*. 2025;29(1):133. doi:10.1007/s10151-025-03170-y

10. De Gregorio MA, Guirola JA, Serrano-Casorran C, et al. Catheter-directed hemorrhoidal embolization for rectal bleeding due to hemorrhoids (Goligher grade I–III): prospective outcomes from a Spanish emborrhoid registry. *Eur Radiol*. 2023;33(12):8754-8763. doi:10.1007/s00330-023-09923-3

11. Brillantino A, Renzi A, Talento P, et al. The Italian Unitary Society of Colon-Proctology (Società Italiana Unitaria di Colonproctologia) guidelines for the management of acute and chronic hemorrhoidal disease. *Ann Coloproctol*. 2024;40(4):287-320. doi:10.3393/ac.2023.00871.0124

12. Robinson PN, Mungall CJ, Haendel M. Capturing phenotypes for precision medicine. *Cold Spring Harb Mol Case Stud*. 2015;1(1):a000372. doi:10.1101/mcs.a000372

13. De Marco S, Tiso D. Lifestyle and Risk Factors in Hemorrhoidal Disease. *Front Surg*. 2021;8:729166. doi:10.3389/fsurg.2021.729166

14. Xiao C, Hong S, Huang W. Optimizing graph layout by t-SNE perplexity estimation. *Int J Data Sci Anal*. 2023;15(2):159-171. doi:10.1007/s41060-022-00348-7

15. Bernabé-Díaz JA, Franco M, Vivo JM, Fernández-Breis JT. Optimizing clustering-based analytical methods with trimmed and sparse clustering. *Computers in Biology and Medicine*. 2025;194:110436. doi:10.1016/j.compbiomed.2025.110436

9.0 Contacts

Principal Investigator:
Dr. Angelo Stuto – angelo.stuto@grupposandonato.it
Co-Investigators:
Dr. Carlotta La Raja – carlotta.laraja@grupposandonato.it