

When prediction is not enough: multimodal causal discovery from ICU vasospasm data

Maico Bruyneel

Student number: 02104123

Supervisor: Prof. dr. ir. Thomas Demeester

Counsellors: Adrick Tench, Paloma Rabaey, Henri Arno

Master's dissertation submitted in order to obtain the degree of Master of Science in Computer Science Engineering

Academic year 2025-2026

Preface

This master's dissertation was written as part of the Master of Science in Computer Science Engineering at Ghent University, Faculty of Engineering and Architecture, during the academic year 2025-2026.

I would like to thank prof. dr. ir. Thomas Demeester for supervising this thesis and for the guidance throughout the project. I am also grateful to Adrick Tench, Paloma Rabaey, and Henri Arno for their feedback, suggestions, and discussions during the development of the work. The author also thanks Kirsten Colpaert for the clinical discussion of the cohort construction, variable choices, and interpretation of the causal discovery findings.

This thesis brought together clinical data processing, information extraction from medical text, and causal discovery. The project required many methodological choices, and the feedback from the supervisory committee helped to keep the work focused and clinically interpretable.

Ghent, May 2026
Maico Bruyneel

Permission for Use

The author gives permission to make this master dissertation available for consultation and to copy parts of this master dissertation for personal use.

In all cases of other use, the copyright terms have to be respected, in particular with regard to the obligation to state explicitly the source when quoting results from this master dissertation.

Ghent, May 2026
Maico Bruyneel

Statement Concerning the Master's Dissertation and the Oral Defence

This master's dissertation is part of an exam. Any comments formulated by the assessment committee during the oral presentation of the master's dissertation are not included in this text.

Abstract

Cerebral vasospasm after aneurysmal subarachnoid hemorrhage (aSAH) is difficult to study in routine intensive-care data because it is not recorded as one consistent variable. Evidence may appear in diagnosis codes, discharge summaries, radiology reports, treatment records, or monitoring patterns, and these sources do not always describe the same clinical picture. This thesis examines whether such heterogeneous information can be combined into an admission-level representation for exploratory causal discovery around vasospasm.

Using MIMIC-IV and MIMIC-IV-Note, a cohort of likely aSAH admissions was constructed through structured candidate retrieval followed by large language model extraction from discharge summaries and radiology reports. The final cohort contained 470 admissions. Structured EHR variables were combined with note-derived information on vasospasm evidence, hemorrhage severity, hydrocephalus, aneurysm treatment, external ventricular drain presence, and physiological states. Cross-modal comparison showed that vasospasm was coded in 13.8% of admissions, while discharge summaries and radiology reports suggested vasospasm evidence in 48.3% and 52.6% of scored admissions, respectively. This suggests that LLM-based extraction from clinical text captured vasospasm evidence more completely than ICD coding alone, while still preserving the source-specific differences between discharge summaries and radiology reports.

Causal discovery was then applied using PC-stable, FGES, and FCIT across hyperparameter sweeps and stability analyses. The most recurrent structure linked clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, and vasospasm. Notably, clot burden was not recovered as a stable direct neighbour of vasospasm, despite being one of the strongest known clinical risk factors. Instead, it appeared mainly as an upstream severity signal connected to hydrocephalus, severe-care pathways, and later inflammatory-metabolic disturbance. Hyperglycemia was the most consistent systemic neighbour of vasospasm across FGES and FCIT, while leukocytosis was more often connected through hyperglycemia than as a stable direct neighbour. These findings suggest a biologically plausible hypothesis in which leukocytosis and hyperglycemia may be more directly related to vasospasm, potentially reflecting inflammatory and metabolic stress that could increase vascular vulnerability.

Overall, the results suggest that multimodal causal discovery can make administrative, textual, physiological, and care-process signals around vasospasm more explicit, and may provide useful hypotheses for targeted follow-up studies.

Keywords: aneurysmal subarachnoid hemorrhage, vasospasm, MIMIC-IV, multimodal clinical data, clinical text extraction, large language models, causal discovery, electronic health records

When prediction is not enough: multimodal causal discovery from ICU vasospasm data

Maico Bruyneel

Supervisor(s): Prof. dr. ir. Thomas Demeester

Abstract Cerebral vasospasm after aneurysmal subarachnoid hemorrhage (aSAH) is difficult to study in routine intensive-care data, because evidence for it is spread across several parts of the record. It may appear across codes, notes, imaging reports, treatments, and monitoring patterns, which do not always describe the same clinical picture. This work examines whether these heterogeneous sources can be combined into an admission-level representation that is useful for exploratory causal discovery. Using MIMIC-IV and MIMIC-IV-Note, a cohort of likely aSAH admissions was constructed through structured candidate retrieval and large language model-based adjudication of discharge summaries and radiology reports. Structured EHR variables were combined with note-derived vasospasm, severity, treatment, hydrocephalus, EVD, and physiological variables. Cross-modal comparisons suggested that ICD coding captured a relatively narrow part of the vasospasm evidence: vasospasm was coded in 13.8% of admissions, while discharge summaries and radiology reports suggested vasospasm evidence in 48.3% and 52.6% of scored admissions, respectively. Causal discovery was then applied using PC-stable, FGES, and FCIT across hyperparameter sweeps and stability analyses. The most recurrent structure linked clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, and vasospasm. Hyperglycemia was the most consistent systemic neighbour of vasospasm across FGES and FCIT, while leukocytosis was more often connected through hyperglycemia than recovered as a stable direct neighbour. These findings are best read as hypothesis-generating. They suggest that multimodal causal discovery may help make biological, administrative, textual, physiological, and care-process signals around vasospasm more explicit.

Keywords aneurysmal subarachnoid hemorrhage, vasospasm, MIMIC-IV, multimodal clinical data integration, clinical text extraction, large language models, causal discovery, electronic health records

I. INTRODUCTION

Aneurysmal subarachnoid hemorrhage, abbreviated as aSAH, is a severe form of bleeding around the brain caused by rupture of an intracranial aneurysm. Even after the aneurysm has been secured, patients may still develop delayed complications during the following days. One of the most clinically relevant complications is cerebral vasospasm, where cerebral arteries narrow and may reduce blood flow to brain tissue [1], [2]. Vasospasm can contribute to neurological deterioration, although the exact processes leading to it remain difficult to separate in observational intensive-care data.

Several factors have been associated with vasospasm after aSAH. These include the amount of subarachnoid blood, hydrocephalus, inflammation, metabolic stress, electrolyte disturbances, treatment decisions, and monitoring intensity [2], [3]. However, these relations are not straightforward to interpret. A variable may reflect a possible contributor, a

consequence, a marker of initial severity, or part of the care process. This makes routine EHR data promising for exploratory work, while also requiring caution in interpretation.

A major difficulty is how vasospasm is represented in the clinical record. In MIMIC-IV, vasospasm is not available as one clean variable. It may appear as an ICD diagnosis code, in a discharge summary, as vessel narrowing in a radiology report, or as part of a treatment and monitoring trajectory. These sources describe related but non-identical views of the same admission: diagnosis codes are structured but administratively shaped, discharge summaries provide retrospective clinical narratives, radiology reports capture imaging-specific observations, and medication or procedure records reflect parts of the care pathway.

A predictive model could be trained to classify admissions with documented vasospasm. Such a model may be useful for recognition, although it would leave the causal interpretation largely unresolved. In ICU data, both candidate predictors and outcomes are shaped by the care process. Patients with more imaging, more interventions, longer ICU stays, or more severe physiology may have more opportunities for vasospasm to be documented. This can improve prediction while still leaving uncertainty about what the learned relations mean clinically.

The goal of this work is to construct a multimodal admission-level representation of likely aSAH admissions and to use causal discovery to generate clinically interpretable hypotheses about possible relations around vasospasm. The analysis focuses on relations that recur across algorithms, hyperparameter settings, and stability analyses.

II. DATA AND METHODS

The study used MIMIC-IV version 3.1 together with MIMIC-IV-Note [4], [5]. The unit of analysis was the hospital admission, identified by the pair (subject_id, hadm_id). Candidate admissions were retrieved through two complementary pathways. The first pathway selected admissions with SAH-related ICD coding. The second pathway selected admissions without SAH ICD codes but with treatment signals compatible with aneurysm-related care, such as clipping, coiling, or nimodipine use. An overview of the full workflow, from candidate retrieval to causal discovery, is shown in Figure 1.

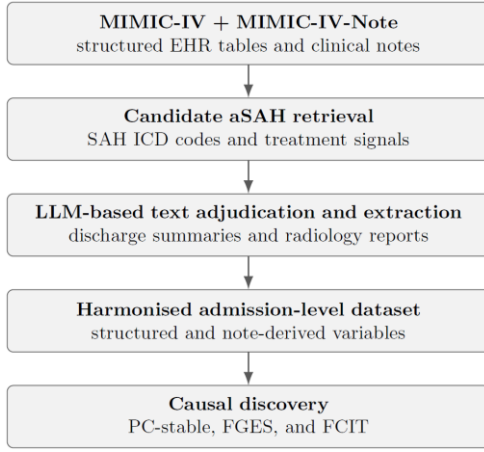


Figure 1 Overview of the multimodal workflow. Candidate admissions were retrieved from structured MIMIC-IV data, adjudicated using discharge summaries and radiology reports, and transformed into a harmonized admission-level dataset before causal discovery with PC-stable, FGES, and FCIT.

Structured data alone could not reliably distinguish aneurysmal from non-aneurysmal SAH. Therefore, a large language model was used to adjudicate candidate admissions based on discharge summaries and radiology reports. This adjudication produced an ordinal likelihood score and a categorical label. Admissions labelled as likely aSAH were retained for the downstream analysis. The final cohort contained 470 likely aSAH admissions, of which 426 came from the ICD-supported pathway and 44 from the treatment-driven pathway without explicit SAH ICD coding.

After cohort construction, structured variables were derived from MIMIC-IV hospital and ICU tables, including demographics, history, diagnosis and procedure codes, medication records, ICU stay structure, laboratory values, bedside measurements, GCS, EVD placement, aneurysm-directed treatment, and mortality. Repeated measurements were summarised into admission-level physiological state variables, including dysnatremia, hypotension, hypoxia, leukocytosis, hyperglycemia, and hypomagnesemia, while retaining source provenance where possible.

Clinical text was used both for aSAH adjudication and for feature extraction within the final cohort. Extracted variables described vasospasm evidence, hemorrhage severity, clot burden, aneurysm location and treatment type, hydrocephalus, EVD evidence, and vasospasm severity. The extraction pipeline used structured JSON outputs, predefined schemas, uncertainty categories, ordinal evidence scores, and supporting text fragments, making the variables inspectable and suitable for graph learning.

The structured and note-derived variables were then harmonised into a single admission-level dataset. Overlapping sources were preserved where useful and consolidated where a single analysis variable was needed. Disagreement between sources was treated as part of the analysis, especially for vasospasm, where ICD codes, discharge summaries, and radiology reports differed substantially in coverage.

Causal discovery was applied using three algorithms: PC-stable [6], FGES [7], and FCIT [8]. PC-stable was used as a constraint-based method, FGES as a score-based method, and FCIT as a PAG-based method that can represent possible

latent confounding. All variables were encoded as discrete categories. Background knowledge constraints were used to forbid clinically implausible directions, such as vasospasm causing baseline age or EVD placement causing the hydrocephalus that usually motivates it. Expected clinical edges were not forced into the graph.

The algorithms were run across hyperparameter sweeps. For PC-stable and FCIT, the main sweep dimensions included conditional independence thresholds, conditioning depth, and sparse-cell handling. For FGES, the sweep varied graph-complexity and BDeu prior settings. Selected configurations were also evaluated using bootstrap stability analysis. The interpretation focused on recurrence across algorithms, hyperparameter regimes, and resampling.

III. RESULTS

The final dataset had broad multimodal coverage. Discharge summaries were available for 456 of 470 admissions, and radiology report bundles were available for 468 admissions. Both note types were present in 454 admissions. Most admissions also had ICU exposure, bedside measurements, and routine laboratory data. At least one ICU stay was recorded in 445 admissions, and white blood cell count, glucose, sodium, and magnesium measurements were available in more than 96% of admissions.

The first main result was that the apparent prevalence of vasospasm depended strongly on the source used. ICD coding identified vasospasm in 65 admissions, corresponding to 13.8% of the cohort. Discharge summaries suggested vasospasm evidence in 48.3% of scored admissions, while radiology reports suggested evidence in 52.6%. This suggests that ICD coding represented a smaller and more administratively defined part of the vasospasm signal present in the full record (Figure 2).

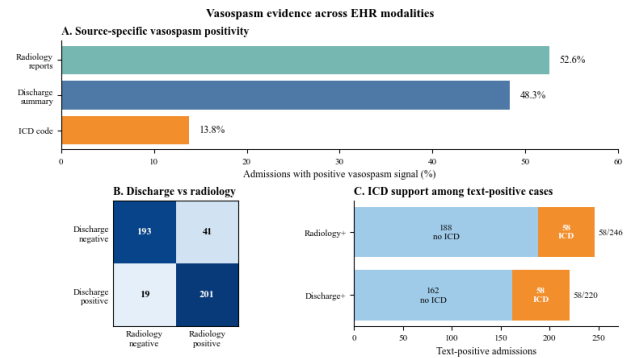


Figure 2 Vasospasm evidence across EHR modalities. ICD coding identified a smaller subset of admissions with vasospasm evidence than the note-derived sources. Discharge summaries and radiology reports showed substantial binary agreement, while many text-positive admissions had no corresponding ICD vasospasm code.

Discharge and radiology signals were broadly consistent. At the binary level, they agreed in 394 pair-complete cases, corresponding to 86.8% agreement, with strongest agreement among clear positive and clear negative cases. Disagreements were concentrated in a smaller subset and often reflected timing or source scope: radiology reports captured vessel-level narrowing at specific imaging time points, while

discharge summaries summarized the overall clinical course and management.

ICD comparison showed a more asymmetric pattern: ICD-positive cases were usually note-supported, but many note-positive cases had no vasospasm ICD code. ICD coding therefore appeared to capture a narrower administrative subset of the broader text-documented vasospasm signal.

Other extracted variables showed plausible cross-source consistency. Aneurysm location and treatment type generally aligned after coarse grouping, while EVD detection was more source-dependent because radiology reports could directly describe catheter position even when structured procedure records did not capture it.

The causal discovery results were interpreted at the level of recurrent structures. The most visible pattern was a broader trajectory linking hemorrhage burden, ventricular complications, severe-care markers, systemic disturbance, and vasospasm evidence (Figure 3):

Clot burden → Hydrocephalus → EVD → Leukocytosis → Hyperglycemia → Vasospasm

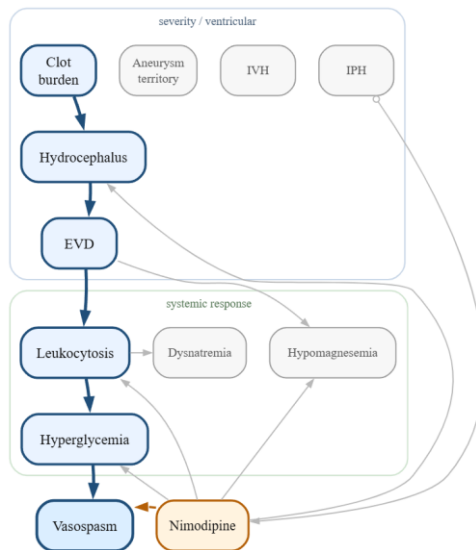


Figure 3 Representative FCIT neighbourhood around vasospasm. The highlighted path shows the recurrent structure linking clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, and vasospasm; lighter edges show nearby FCIT relations.

Clot burden repeatedly connected to hydrocephalus, and hydrocephalus strongly connected to EVD. This relation was not forced into the graph, although the reverse direction was excluded as clinically implausible, supporting the face validity of the learned structure.

The middle part linked EVD, leukocytosis, and hyperglycemia. The EVD-leukocytosis relation was recurrent but difficult to interpret, because EVD may mark a more severe-care subgroup with invasive monitoring, frequent laboratory testing, infection workups, or greater baseline severity. In contrast, leukocytosis → hyperglycemia was one of the more stable systemic relations and was usually oriented in that direction when orientation was recovered, consistent

with inflammation and physiological stress contributing to stress hyperglycemia.

The downstream vasospasm-adjacent finding was centered on the inflammatory-metabolic part of the graph. Hyperglycemia → vasospasm appeared in all FCIT runs and in most FGES runs, while it was less frequent in PC-stable. This makes hyperglycemia the most consistent systemic neighbour of vasospasm in the score-based and PAG-based analyses. One interpretation is that hyperglycemia captures systemic stress or overall disease severity after aSAH. However, this explanation is not necessarily sufficient on its own, because the variable set already included several other markers of hemorrhage burden, neurological injury, physiological instability, treatment, and monitoring intensity. A more clinically interesting hypothesis is therefore that hyperglycemia may not only reflect illness severity, but may also be part of a metabolic pathway that affects the blood vessels. High glucose levels can make vessels less able to relax, more prone to narrowing, and may strengthen inflammation or small clotting processes. Through these mechanisms, hyperglycemia could potentially contribute to vasospasm or make it more severe [9], [10].

Leukocytosis showed a related but less direct pattern. The direct leukocytosis → vasospasm edge was less stable than leukocytosis → hyperglycemia and hyperglycemia → vasospasm, appearing mainly in part of the PC-stable and FGES analyses. Still, leukocytosis remained repeatedly involved in the vasospasm-related structure. This supports the hypothesis that inflammatory activation may contribute to the same downstream process, either indirectly through stress hyperglycemia or, in some settings, more directly through vascular inflammation. Activated white blood cells may irritate or damage the inner lining of blood vessels. This can make the vessels less able to relax, more likely to narrow, and more prone to interactions with platelets or tiny clots [9]. In this reading, leukocytosis and hyperglycemia form a plausible inflammatory-metabolic bridge between severe aSAH physiology and vasospasm evidence, with hyperglycemia recovered as the more stable local neighbour of vasospasm. The analysis cannot prove that either variable directly causes vasospasm, but it does identify this pathway as a clinically plausible hypothesis for follow-up study.

Nimodipine → vasospasm was also highly recurrent, especially in FGES and FCIT. Clinically, this edge is best interpreted as a care-process and documentation-related signal. Nimodipine is given after aSAH as part of standard management and is linked to recognition, monitoring, and concern for delayed ischemic complications [1]. Its stable relation with vasospasm likely reflects how treatment and documentation are embedded in the EHR.

Several clinically relevant variables were not recovered as stable direct parents of vasospasm. Clot burden, for example, is an often mentioned potential risk factor in clinical literature [2], [3], but in these graphs it was more consistently placed upstream of hydrocephalus and EVD than directly adjacent to vasospasm. Dysnatremia, hypomagnesemia, hypoxia, and hypotension appeared in some vasospasm-adjacent structures, especially under more permissive PC-stable settings, but these relations were more sensitive to hyperparameter choices and were therefore treated as more tentative.

IV. DISCUSSION

The results suggest that multimodal representation is useful for studying vasospasm in routine EHR data. ICD coding alone gave a much narrower view than discharge summaries and radiology reports, while the note-derived signals themselves still required caution. They were therefore evaluated through cross-source agreement, disagreement review, and consistency checks with structured treatment and procedure records. This source-aware approach reflects that EHR variables combine measurement, coding, documentation, and clinical action.

The causal discovery results suggest that vasospasm-adjacent structure in this dataset is shaped by a mixture of biological severity, systemic response, and care-process effects. The most coherent recurrent structure placed clot burden upstream of hydrocephalus and EVD, followed by inflammatory and metabolic disturbance. This is compatible with a severe aSAH trajectory in which hemorrhage burden, ventricular complications, invasive management, inflammation, and stress hyperglycemia cluster around vasospasm evidence. The graph thus provides a useful structured hypothesis that can be examined in follow-up studies.

The inflammatory-metabolic part of this trajectory is the most clinically interesting. Hyperglycemia and leukocytosis may partly mark patients with more severe disease. However, they are also biologically plausible vascular stress signals, which makes the recovered structure more specific than a generic severity pattern. EVD should be read differently, mainly as a marker of hydrocephalus severity and intensive neurocritical care, not as a direct biological cause.

This interpretation was also supported by a discussion with an ICU vasospasm expert, who considered the recurrent clot burden-hydrocephalus-EVD structure clinically plausible and the leukocytosis-hyperglycemia findings interesting for further analysis.

Several limitations should be kept in mind. The analysis is admission-level, so repeated measurements and clinical events are compressed into one row per admission, reducing temporal precision. Some variables may occur before, during, or after vasospasm. The study is limited to a single EHR database and lacks external validation. Unmeasured confounding remains likely, especially for clinical severity, treatment selection, infection, sedation, vasopressor use, ventilation strategy, and clinician decision-making. Finally, LLM-derived variables depend on source-note quality and the extraction framework; cross-modal checks reduce some risk, but do not replace expert adjudication for every admission.

Future work could move toward more temporally detailed modelling. Vasospasm is a delayed event, so time-windowed representations may help distinguish early severity, physiology, and later outcome. External validation in another hospital cohort would also be useful. The recurrent relations identified here, could also be tested using targeted causal inference designs with explicit exposure definitions, timing, and confounder adjustment.

V. CONCLUSION

This work constructed a disease-specific multimodal dataset for likely aSAH admissions in MIMIC-IV and used it for exploratory causal discovery around cerebral vasospasm. The findings suggest that vasospasm is represented much more broadly in clinical text than in ICD coding alone, and that discharge summaries and radiology reports provide complementary evidence. Causal discovery across PC-stable, FGES, and FCIT recovered a recurrent structure linking clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, and vasospasm. These results help organise the heterogeneous EHR signals around vasospasm into clinically interpretable hypotheses. The approach may be useful as a source-aware workflow for moving from multimodal routine-care data toward more focused causal questions.

ACKNOWLEDGEMENTS

The author would like to thank Prof. dr. ir. Thomas Demeester for supervising this thesis, and Adrick Tench, Paloma Rabaey, and Henri Arno for their feedback, suggestions, and discussions during the development of the work.

REFERENCES

- [1] J. Claassen and S. Park, *Spontaneous subarachnoid haemorrhage*, The Lancet, vol. 400, no. 10355, pp. 846–862, 2022.
- [2] J. M. Findlay, J. Nisar, and T. E. Darsaut, *Cerebral vasospasm: A review*, Canadian Journal of Neurological Sciences, 2016.
- [3] T. Inagawa, K. Yahara, and N. Ohbayashi, *Risk factors associated with cerebral vasospasm following aneurysmal subarachnoid hemorrhage*, Neurologia Medico-Chirurgica, 2014.
- [4] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark, *MIMIC-IV*, version 3.1, PhysioNet, 2024. doi: 10.13026/kpb9-mt58.
- [5] A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark, *MIMIC-IV-Note: Deidentified free-text clinical notes*, version 2.2, PhysioNet, 2023. doi: 10.13026/1n74-ne17.
- [6] D. Colombo and M. H. Maathuis, *Order-independent constraint-based causal structure learning*, Journal of Machine Learning Research, 2014.
- [7] J. Ramsey, M. Glymour, R. Sánchez-Romero, and C. Glymour, *A million variables and more: The Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images*, Journal of Data Science, 2017.
- [8] J. Ramsey, B. Andrews, and P. Spirtes, *Efficient latent variable causal discovery: Combining score search and targeted testing*, arXiv preprint arXiv:2510.04263, 2025.
- [9] I. Buce-Satoba, D. Rozkalne, B. Mamaja, G. Krumina, and A. Ozolina, *Leukocytosis and C-reactive protein may predict development of secondary cerebral vasospasm in patients with aneurysmal subarachnoid hemorrhage*, Medicina, vol. 58, no. 2, article 323, 2022, doi: 10.3390/medicina58020323.
- [10] N. D. Kruyt, G. J. Biessels, J. H. DeVries, M. J. A. Luitse, M. Vermeulen, G. J. E. Rinkel, W. P. Vandertop, and Y. B. W. E. M. Roos, *Hyperglycemia in aneurysmal subarachnoid hemorrhage: a potentially modifiable risk factor for poor outcome*, Journal of Cerebral Blood Flow & Metabolism, vol. 30, no. 9, pp. 1577–1587, 2010, doi: 10.1038/jcbfm.2010.102.
- [11] L. E. Fernández-Garza, V. A. Fernández-Garza, D. Mares-Custodio, V. Gutiérrez-Ruano, A. Navarrete-Rodríguez, and J. J. Arias-Alzate, *Glucose/potassium ratio, a novel biomarker for the prognosis of patients with subarachnoid hemorrhage: a review*, Journal of Vascular Diseases, vol. 4, no. 4, article 48, 2025, doi: 10.3390/jvd4040048.

Contents

Preface	i
Permission for Use	ii
Statement Concerning the Master’s Dissertation and the Oral Defence	iii
Abstract	iv
List of Figures	ix
List of Tables	x
List of Abbreviations and Symbols	xi
1 Introduction	1
1.1 Motivation	1
1.2 Analytical framing	2
1.3 Research questions	2
1.4 Contributions	2
1.5 Scope and delimitation	3
1.6 Thesis structure	3
2 Literature Review	4
2.1 Clinical Background: aSAH and Vasospasm	4
2.1.1 aSAH Pathophysiology	4
2.1.2 Vasospasm	5
2.1.3 Known Risk Factors	6
2.1.4 Additional Clinical Context and Systemic Factors	7
2.1.5 Limitations of Clinical Knowledge	8
2.2 Clinical Data Representation in EHR-Based Research	8
2.2.1 EHR data as a research substrate	8
2.2.2 Modalities of clinical information	8
2.2.3 Structural limitations of EHR data	9
2.2.4 Cross-modal representation and disagreement	9
2.3 Information Extraction from Clinical Text	9
2.3.1 Traditional NLP approaches	10
2.3.2 Machine learning / deep learning NLP	10
2.3.3 Large Language Models in clinical NLP	10
2.3.4 Limitations of LLM-based extraction	11
2.4 Causal Inference in Observational Clinical Data	11
2.4.1 Correlation vs. Causation	12
2.4.2 Classical causal inference with a specified question	12

2.4.3	Why this motivates causal discovery	12
2.5	Causal Discovery Methods	13
2.5.1	Constraint-Based Methods	13
2.5.2	Score-Based Methods	14
2.5.3	PAG-Based Methods and Latent Confounding	14
2.5.4	Stability and Robustness	15
2.6	Causal Discovery in Clinical Applications	16
2.7	Synthesis and Research Gap	17
2.7.1	Clinical gap	17
2.7.2	Data gap	17
2.7.3	Methodological gap	18
2.7.4	Contribution	18
3	Data and Cohort Construction	19
3.1	Data source and environment	19
3.2	Raw data sources and modalities	19
3.3	Cohort construction	20
3.4	Structured data scope and coverage	20
3.5	Final cohort summary	20
4	Methodology	21
4.1	Structured variable retrieval pipeline	21
4.1.1	Retrieval logic and staged feature construction	22
4.1.2	Core structured variables	22
4.1.3	Compound physiologic feature blocks	24
4.1.4	Threshold definitions for physiologic feature blocks	26
4.1.5	Aggregation logic, temporal constraints, and missingness handling	26
4.1.6	Validation, quality control, and provenance	28
4.1.7	Role of the structured retrieval pipeline in the overall methodology	28
4.2	LLM-based information extraction from clinical text	28
4.2.1	Cohort-defining task: aSAH adjudication	29
4.2.2	Feature-generating tasks	30
4.2.3	Structured output design	31
4.2.4	Uncertainty representation	31
4.2.5	Few-shot example construction	32
4.2.6	Implementation safeguards	34
4.2.7	Construction of the combined dataset	34
4.3	Data Harmonization and Quality Control	34
4.3.1	Making extracted values consistent	35
4.3.2	Manual review of disagreement cases	35
4.3.3	Distributional and consistency checks	36
4.4	Causal discovery implementation	36
4.4.1	Role of causal discovery in the study	36
4.4.2	Variable set construction	37
4.4.3	Constraint design	38
4.4.4	Algorithms used	39
4.4.5	Hyperparameter sweeps and sensitivity analysis	40
4.4.6	Bootstrap and stability workflow	40
4.4.7	Output aggregation and summarization	41

5	Results	42
5.1	Cohort construction and label validation	42
5.1.1	Candidate sets and note availability	42
5.1.2	LLM adjudication of SAH-coded admissions	42
5.1.3	Recovery of admissions without SAH ICD coding	43
5.1.4	Manual inspection of adjudication quality	43
5.1.5	Why the final cohort was restricted to <i>likely aSAH</i>	45
5.2	Coverage and data availability across modalities	45
5.3	Structured variable distributions	46
5.3.1	Demographics and baseline history	47
5.3.2	Treatment and care trajectory	47
5.3.3	Neurological severity	48
5.3.4	Complications and physiologic states	48
5.4	LLM extraction results	48
5.4.1	Vasospasm signal across modalities	50
5.4.2	Consistency and cross-modal validation of additional LLM-extracted variables	56
5.5	Causal discovery results	62
5.5.1	Overall graph characteristics	62
5.5.2	Recurrent structures across algorithms	62
5.5.3	Robustness of the main structures	67
5.5.4	Limitations of the causal discovery results	69
6	Discussion and Conclusions	70
6.1	Discussion	70
6.1.1	Interpretation of cross-modal extraction results	70
6.1.2	Interpretation of causal discovery results	71
6.1.3	Clinical expert feedback	71
6.2	Limitations	73
6.3	Future work	74
6.4	Conclusion	74
7	Societal Reflection	76
7.1	Ethical considerations and equity	76
7.2	Environmental sustainability and green AI	77
7.3	Global context and international disparities	77
7.4	Alignment with the Sustainable Development Goals	78
7.5	Reflections and future directions	78
	References	79
A	ICD pattern groups used in structured retrieval	88
B	Core structured variables derived in the first retrieval stage	90
C	Derived physiologic state definitions used	94
D	Detailed validation checks for structured retrieval	97
E	Variable provenance mapping	99
F	LLM prompt templates used for clinical text extraction	101

G	LLM execution framework and runtime control	118
H	Background knowledge constraints used for causal discovery	121
I	Causal discovery variable sets	122
J	Causal discovery hyperparameter grids	124
K	Causal discovery output aggregation	126
L	Use of generative artificial intelligence	129

List of Figures

2.1	Schematic illustration of a cerebral aneurysm	4
2.2	Simplified schematic of vessel narrowing	5
2.3	Temporal course after aneurysmal subarachnoid hemorrhage	6
2.4	Examples of DAG and PAG edge notation	15
4.1	Overview of the study pipeline. Candidate admissions were retrieved from structured MIMIC-IV data, adjudicated using clinical text, and converted into a harmonised multimodal admission-level dataset before causal discovery was applied.	21
4.2	Illustrative admission-level representation from the final structured dataset.	27
5.1	LLM label distribution among ICD-coded SAH admissions, including admissions without available clinical notes.	42
5.2	Distribution of LLM-assigned aSAH likelihood scores among scored SAH-coded admissions.	43
5.3	Distribution of the ICD-based vasospasm indicator.	50
5.4	Distribution of discharge summary vasospasm score buckets.	51
5.5	Distribution of radiology vasospasm score buckets.	51
5.6	Binary vasospasm classification derived from discharge summaries. Scores ≥ 1 were treated as positive and scores ≤ 0 as negative.	52
5.7	Binary vasospasm classification derived from radiology reports. Scores ≥ 1 were treated as positive and scores ≤ 0 as negative.	52
5.8	Confusion matrix comparing discharge- and radiology-derived vasospasm using binary classification. Scores ≥ 1 were treated as positive and scores ≤ 0 as negative.	53
5.9	Alignment of discharge and radiology vasospasm scores at the bucket level.	53
5.10	Comparison of ICD-based vasospasm coding with discharge-derived and radiology-derived binary signals.	55
5.11	Agreement between discharge- and radiology-derived aneurysm location (anterior vs posterior).	57
5.12	Cross-modal validation of aneurysm treatment type across discharge summaries, structured records, and radiology.	58
5.13	Agreement between discharge- and radiology-derived hydrocephalus status.	59
5.14	Agreement between radiology-derived EVD presence and structured procedure-based EVD detection.	60
5.15	Relationship between radiology-derived vasospasm severity categories and ordinal scores.	61
5.16	Representative FCIT PAG with highlighted main path	64

List of Tables

3.1	Overview of MIMIC-IV modules used in this thesis	19
4.1	Derived physiological state variables and encoded categories.	25
4.2	Thresholds used to derive physiologic feature states.	26
5.1	Coverage of main modalities and selected derived variables in the final cohort. Percentages are calculated over 470 admissions.	46
5.2	Selected complications and physiologic states in the final cohort. Percentages are calculated over 470 admissions.	49
5.3	Summary of graph characteristics across causal discovery algorithms. All algorithms used the same 32-variable node set.	62
5.4	Summary of recurrent causal discovery structures and their robustness across methods and sensitivity analyses.	67
A.1	ICD diagnosis and procedure pattern groups used in the structured retrieval pipeline.	89
B.1	Core structured variables derived in the first retrieval stage.	90
C.1	Definition and distribution of leukocytosis categories.	94
C.2	Definition and distribution of hyperglycemia categories.	94
C.3	Definition and distribution of dysnatremia categories.	95
C.4	Definition and distribution of hypotension categories.	95
C.5	Definition and distribution of hypoxia categories.	95
C.6	Definition and distribution of hypomagnesemia categories.	95
C.7	Definition and distribution of hypercapnia categories.	96
E.1	Provenance mapping for core admission-level variables in the structured retrieval pipeline.	99
E.2	Provenance mapping for physiologic and complication-related feature blocks in the structured retrieval pipeline.	100
F.1	Overview of LLM prompt templates used in the information extraction pipeline.	101
H.1	Additional background-knowledge restrictions used in causal discovery, beyond the tier-based temporal ordering.	121
I.1	Candidate variables considered for causal discovery.	122
J.1	Full PC-stable hyperparameter sweep.	124
J.2	Full FGES hyperparameter sweep.	124
J.3	Full FCIT hyperparameter sweep.	125

List of Abbreviations and Symbols

Abbreviations

Abbreviation	Meaning
aSAH	Aneurysmal subarachnoid hemorrhage
SAH	Subarachnoid hemorrhage
DCI	Delayed cerebral ischemia
EBI	Early brain injury
IVH	Intraventricular hemorrhage
IPH	Intraparenchymal hemorrhage
SDH	Subdural hemorrhage
EVD	External ventricular drain
GCS	Glasgow Coma Scale
WFNS	World Federation of Neurosurgical Societies scale
WBC	White blood cell count
NO	Nitric oxide
EHR	Electronic health record
ICU	Intensive care unit
MIMIC-IV	Medical Information Mart for Intensive Care IV
MIMIC-IV-Note	Free-text clinical note component of MIMIC-IV
BIDMC	Beth Israel Deaconess Medical Center
ICD	International Classification of Diseases
CPT	Current Procedural Terminology
NLP	Natural language processing
LLM	Large language model
CoT	Chain-of-Thought
JSON	JavaScript Object Notation
JSONL	JSON Lines
RCT	Randomized controlled trial
DAG	Directed acyclic graph
PAG	Partial ancestral graph
CPDAG	Completed partially directed acyclic graph
CI	Conditional independence
PC	Peter–Clark algorithm
PC-stable	Order-independent variant of the PC algorithm
FGES	Fast Greedy Equivalence Search

Abbreviation	Meaning
FCI	Fast Causal Inference
FCIT	Targeted variant of Fast Causal Inference
BIC	Bayesian Information Criterion
BDeu	Bayesian Dirichlet equivalent uniform score
G^2	Likelihood-ratio test statistic used for discrete conditional independence testing
CT	Computed tomography
CTA	Computed tomography angiography
CTP	Computed tomography perfusion
MRI	Magnetic resonance imaging
MRA	Magnetic resonance angiography
MRV	Magnetic resonance venography
TCD	Transcranial Doppler ultrasonography
MAP	Mean arterial pressure
SBP	Systolic blood pressure
DBP	Diastolic blood pressure
PaO_2	Arterial partial pressure of oxygen
$PaCO_2$	Arterial partial pressure of carbon dioxide
SpO_2	Peripheral oxygen saturation
SaO_2	Arterial oxygen saturation
FiO_2	Fraction of inspired oxygen
$PbtO_2$	Brain tissue partial pressure of oxygen
Mg	Magnesium
Na^+	Sodium ion
EMAR	Electronic medication administration record
OMR	Online medical record table in MIMIC-IV
SQL	Structured Query Language

Symbols and notation

Symbol / notation	Meaning
$\rightarrow$	Directed edge in a graph. For example, $X \rightarrow Y$ indicates that X is represented as upstream of Y in the learned graph.
$\leftrightarrow$	Bidirected relation, often used to represent possible latent confounding or unresolved ancestral structure in PAG-based outputs.
$\circ$	Circle endpoint in a PAG, indicating that the endpoint orientation is not determined from the available conditional independence information.
α	Significance threshold used in conditional independence testing.
n	Number of observations, admissions, or samples.
ρ	Spearman rank correlation coefficient.

Symbol / notation	Meaning
mg/dL	Milligrams per decilitre; used for glucose and some laboratory measurements.
mmHg	Millimetres of mercury; used for blood pressure and blood gas measurements.
mmol/L	Millimoles per litre; used for electrolyte concentrations.
$\times 10^9/\text{L}$	Unit used for white blood cell count.

Chapter 1

Introduction

1.1 Motivation

Aneurysmal subarachnoid hemorrhage (aSAH) is a severe type of bleeding around the brain, caused by the rupture of an intracranial aneurysm, a weakened bulging blood vessel inside the skull. Even after the aneurysm has been secured, patients can still develop delayed complications during the following days. One of these complications is cerebral vasospasm. This means that blood vessels in the brain become narrowed. This narrowing can reduce blood flow to brain tissue and may contribute to neurological deterioration.

Clinicians have been trying to understand why vasospasm develops. If the processes leading to vasospasm were better understood, prevention and early treatment could be studied in a more targeted way. Some risk factors are already well described, such as the amount of blood visible after the hemorrhage. However, the causal mechanisms are still difficult to separate. Vasospasm does not seem to follow one simple pathway. It may involve hemorrhage severity, inflammation, metabolic stress, and treatment decisions. Many of these factors are associated with vasospasm, but an association alone does not show whether a variable is a possible cause, a consequence, a marker of severity, or part of the care process.

This thesis studies vasospasm using electronic health record (EHR) data from MIMIC-IV. These data are attractive for this question because they contain many views of the same admission: diagnosis codes, procedure records, medication data, laboratory values, bedside monitoring, radiology reports, and discharge summaries. At the same time, this is also what makes the problem difficult. Vasospasm is not recorded as one clean variable. It may appear as an ICD diagnosis code, as a statement in a discharge summary, as vessel narrowing in a radiology report, as a treatment decision, or as a reason for repeated monitoring. These sources may agree, but they may also capture different parts of the same clinical course.

A predictive model could be trained to recognise admissions with documented vasospasm. Such a model may be useful for classification, but it would not answer the main question in this thesis. In ICU data, both the outcome and many possible predictors are shaped by the care process. A model may learn that patients with more imaging, more interventions, more severe physiology, or longer ICU stays are more likely to have vasospasm documented. That can improve prediction, while still leaving the causal interpretation unclear.

For that reason, this thesis does not aim to build a deployable prediction model for vasospasm. Its aim is to construct a multimodal admission-level representation of aSAH admissions and to use causal discovery to generate clinically interpretable hypotheses about possible causal relations around vasospasm. The focus is on relations that appear repeatedly across algorithms and hyperparameter settings, and that remain plausible when interpreted against the clinical meaning of the variables. This leads to the central goal of the thesis:

To use multimodal EHR data and causal discovery to generate clinically interpretable hypotheses about possible causal relations around vasospasm after aSAH.

1.2 Analytical framing

The main analysis uses causal discovery. In this thesis, causal discovery is used as a hypothesis-generating method. The goal is to identify possible causal relations between clinically relevant variables, not only to describe which variables are statistically associated.

In the graph representation, variables are shown as nodes. The algorithms then search for graph structures that fit the observed data and the clinical background knowledge supplied to the model. The statistical evidence comes from patterns of dependence, conditional independence, and score-based graph fit. These patterns are useful for generating causal hypotheses.

Three causal discovery algorithms are used: PC-stable, FGES, and FCIT. These methods follow different discovery strategies and may behave differently with a finite sample, categorical variables, and possible unmeasured confounding. The interpretation therefore does not rely on one graph from one run. Instead, the analysis focuses on recurring patterns across algorithms, hyperparameter settings, bootstrap stability checks, and sensitivity analyses. Clinical background knowledge is used to exclude directions that would be difficult to defend, such as vasospasm causing baseline age or EVD placement causing the hydrocephalus that usually motivates it.

1.3 Research questions

The main research question is:

Which hypotheses about causal relations around cerebral vasospasm after aneurysmal subarachnoid hemorrhage can be generated from multimodal ICU electronic health record data?

This question is addressed through four subquestions:

1. How can an aSAH cohort and clinically interpretable admission-level variables be constructed from structured MIMIC-IV data and clinical text?
2. How do vasospasm and related clinical signals differ across ICD codes, discharge summaries, radiology reports, and structured treatment or procedure records?
3. Which possible causal relations around vasospasm recur across causal discovery algorithms, hyperparameter settings, and stability analyses?
4. Which recovered relations are compatible with existing clinical expectations, and which are more likely to reflect documentation, monitoring, treatment, or care-process effects?

1.4 Contributions

This thesis makes three main contributions.

First, it constructs a disease-specific multimodal dataset for aSAH in MIMIC-IV. Candidate admissions are retrieved through diagnosis-based and treatment-based signals, then refined using LLM-based extraction from discharge summaries and radiology reports. The final dataset combines structured EHR variables with note-derived variables, while preserving information about where each variable came from.

Second, it evaluates how clinically important variables are represented across different EHR modalities. The analysis compares ICD-based, discharge-derived, radiology-derived, and structured procedure or medication signals for vasospasm and related constructs. In this way, disagreement between sources becomes part of the analysis instead of being treated only as noise.

Third, it uses causal discovery to generate and evaluate hypotheses about possible causal relations around vasospasm. PC-stable, FGES, and FCIT are applied to the admission-level dataset, with hyperparameter sweeps, bootstrap stability checks, and sensitivity analyses. The interpretation focuses on recurrent relations involving clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, nimodipine, and vasospasm, while distinguishing relations that seem biologically plausible from relations that may mainly reflect monitoring, treatment, or documentation.

1.5 Scope and delimitation

This thesis is limited to aSAH admissions in MIMIC-IV. The hospital admission is the unit of analysis. Individual measurements, imaging studies, notes, and ICU days are summarised into admission-level variables.

This design makes it possible to combine structured data and clinical text, but it also reduces temporal precision. The causal discovery input does not fully encode whether every abnormality occurred before, during, or after vasospasm. The learned graphs should therefore be read as admission-level causal hypotheses.

The analysis is exploratory. It does not estimate treatment effects, compare interventions, or produce a deployable clinical prediction model. Questions such as clipping versus coiling, nimodipine exposure, EVD placement, or glucose-control strategies would require dedicated causal designs with explicit timing and confounder selection.

The study is also limited by the variables available in the pipeline. Factors such as aneurysm morphology, infection, fever, steroid exposure, vasopressor use, sedation, nutrition, detailed ventilation settings, and exact timing of vasospasm-directed interventions are not fully represented.

Finally, the empirical analysis is limited to MIMIC-IV. Although UZ Ghent data were part of the broader motivation, no direct external replication or cross-hospital comparison is performed in this thesis.

1.6 Thesis structure

Chapter 2 reviews the clinical and methodological background. Chapter 3 then describes MIMIC-IV, the data modalities used in the study, and the aSAH cohort construction process. Chapter 4 presents the structured variable retrieval pipeline, the LLM-based extraction framework, the harmonisation steps, and the causal discovery implementation. Chapter 5 reports the cohort construction results, cross-modal validation results, variable distributions, and causal discovery findings. Finally, chapter 6 discusses the interpretation of the findings, limitations of the approach, and possible directions for future work.

Chapter 2

Literature Review

2.1 Clinical Background: aSAH and Vasospasm

2.1.1 aSAH Pathophysiology

Aneurysmal subarachnoid hemorrhage occurs when a weakened blood vessel in the brain ruptures. More specifically, it is usually caused by the rupture of an intracranial aneurysm: a weak, balloon-like bulge in a blood vessel inside the skull. Most of these aneurysms are saccular, often called "berry" aneurysms because of their rounded shape. The bleeding enters the subarachnoid space, which is the fluid-filled space around the brain [16]. Because the rupture occurs in an artery, the bleeding can happen under high pressure. This can rapidly increase the pressure inside the skull and temporarily reduce the amount of blood reaching the brain [16]. The injury after aSAH therefore does not end at the moment of rupture. It develops over time, with an early phase caused by the initial bleeding and later phases involving secondary complications.

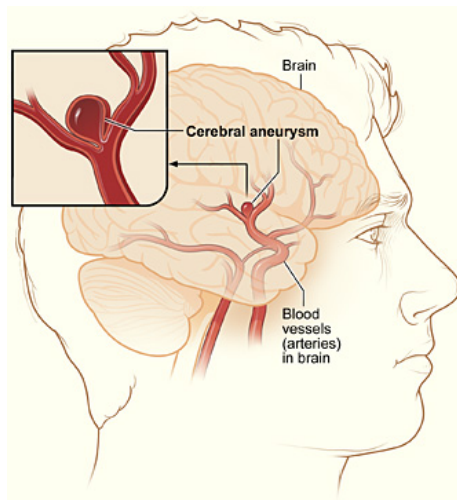


Figure 2.1: Schematic illustration of a cerebral aneurysm. The figure is used to show the type of vascular abnormality that can rupture and lead to aneurysmal subarachnoid hemorrhage. Image source: National Institutes of Health, via Wikimedia Commons, public domain [70].

Early Brain Injury The first 24 to 72 hours after the event are commonly referred to as early brain injury (EBI) [22,29]. During this period, the sudden rise in pressure inside the skull can temporarily limit blood flow to the brain, a situation known as global ischemia [29]. This can damage the inner lining of blood vessels and disrupt the blood-brain barrier, which normally regulates what enters and leaves brain tissue [29]. Signs of cell damage, including cell death, can already be observed early on.

A surge in stress hormones (catecholamines) is often seen during this phase. This response is thought to be related to effects on brain regions that regulate stress and vital functions [29]. Other findings reported during EBI include brain swelling (cerebral edema), reduced ability of blood vessels to regulate flow, formation of small clots, and activation of inflammatory responses [22, 29].

Delayed Processes and DCI In the days following the acute phase, additional changes begin to affect the brain. Delayed cerebral ischemia (DCI), which refers to a reduction in blood supply leading to new neurological problems, is typically observed between 4 and 12 days after the hemorrhage [16, 29]. While vasospasm remains the most significant factor, DCI can also occur in patients without clear narrowing of these vessels [29].

Blood that has entered the subarachnoid space gradually breaks down over time. As red blood cells degrade, they release substances such as oxyhemoglobin, with levels often peaking around day 7 [29]. Oxyhemoglobin has been linked to constriction of blood vessels, partly by reducing the availability of nitric oxide (NO), a molecule that normally helps vessels relax [29, 34].

Other mechanisms have also been proposed, including problems in the smallest blood vessels (microcirculation), reduced clearance of waste products from the brain, waves of abnormal electrical activity (cortical spreading depolarizations), and inflammatory processes involving clotting [22, 29]. These processes can occur together, and their relative importance may differ between patients. Overall, DCI is generally described as a multifactorial condition, where the brain's energy needs are not fully met by the available blood supply.

2.1.2 Vasospasm

Definition and Clinical Distinction The terms vasospasm and DCI were often used interchangeably in earlier work, but they are now usually treated as distinct [26, 29, 109]. A useful distinction is between angiographic vasospasm and clinical (or symptomatic) vasospasm.

Angiographic vasospasm refers to narrowing of the larger brain arteries that can be seen on imaging, for example with CT or MRI angiography [1, 34, 109]. Clinical vasospasm refers to neurological deterioration, such as new deficits or reduced consciousness, that is thought to result from reduced blood flow, when no other clear cause is identified [34]. The term DCI is now commonly used for clinical vasospasm, or for evidence of brain infarction on imaging [109].

Figure 2.2 gives a simplified visual illustration of the core idea behind vasospasm: narrowing of an artery, which can reduce downstream blood flow.

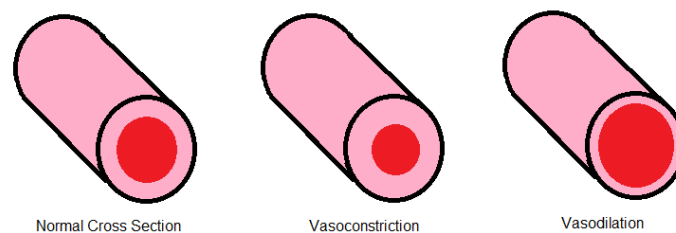


Figure 2.2: Simplified schematic of vessel narrowing and widening. In cerebral vasospasm, the relevant idea is abnormal narrowing of a cerebral artery, which can reduce downstream blood flow. Image source: Elizabeth2424, via Wikimedia Commons, licensed under CC BY-SA 3.0 [31].

Timing Vasospasm tends to follow a recognizable time pattern. It is usually observed a few days after the hemorrhage, most often between days 3 and 14, with a peak around the end of the first week [1, 20, 105]. It is less commonly seen in the first 48 hours, likely because it depends on the gradual breakdown of blood in the subarachnoid space [105]. In many cases, the narrowing

resolves over the following weeks [105]. Figure 2.3 summarizes the approximate timing of the main clinical phases discussed above.

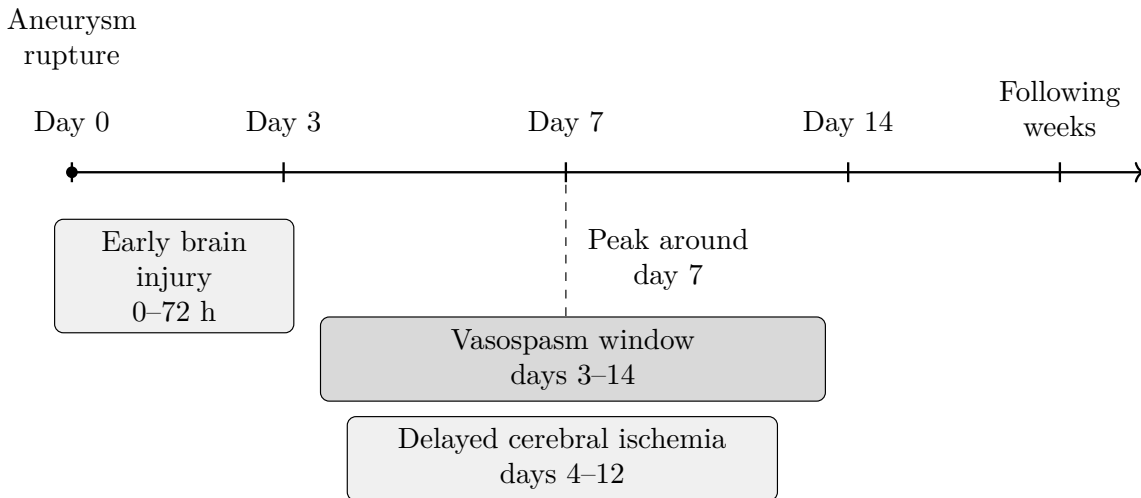


Figure 2.3: Simplified temporal course after aneurysmal subarachnoid hemorrhage. Early brain injury mainly occurs during the first 24–72 hours after rupture, while vasospasm and delayed cerebral ischemia generally occur later, with peak vasospasm risk around the end of the first week. The figure is schematic and shows approximate clinical time windows only.

Mechanisms The processes involved in vasospasm are not fully understood and likely involve several interacting factors [34, 105].

Inflammatory responses are often described. After the hemorrhage, immune cells become active and release signaling molecules that may influence both blood vessels and surrounding brain tissue [34, 82].

There are also changes in how blood vessels regulate their diameter. Substances released as blood breaks down, including oxyhemoglobin, are thought to reduce the availability of NO, which normally helps vessels relax. At the same time, signals that promote vessel constriction may become more prominent [20, 34].

In addition, smaller blood vessels may be affected. Reduced flow in these vessels, sometimes linked to small clots or local disturbances, could contribute to reduced oxygen delivery in brain tissue. This may help explain why some patients develop DCI even when large arteries do not appear severely narrowed [29, 102].

2.1.3 Known Risk Factors

The course of aSAH can vary considerably between patients. A number of clinical and physiological factors have been linked to the development of later complications [34]. These factors are commonly used to get a sense of patient risk and are also reflected in the variables used in this study.

Blood Burden (Clot and IVH) The amount of blood visible on the initial brain scan appears to matter [34, 50]. Larger clots in the subarachnoid space have been associated with a higher risk of DCI [21]. Blood in the brain’s fluid-filled spaces (intraventricular hemorrhage, IVH) may further increase this risk [21]. In practice, scales such as the Modified Fisher Scale combine these findings to give an overall impression of severity [34].

Hydrocephalus Hydrocephalus refers to an accumulation of fluid in the brain, often because blood interferes with normal drainage [102]. It is relatively common after aSAH and is generally

seen in more severe cases [22]. Management often involves placing a temporary drain (external ventricular drain, EVD) [16, 102].

Inflammation Signs of inflammation are frequently observed after aSAH [82]. For example, higher white blood cell counts (leukocytosis) have been linked to an increased likelihood of vasospasm [14]. Elevated C-reactive protein (CRP) levels have also been associated with DCI and poorer recovery, although these markers are not specific [14].

Metabolic Stress (Hyperglycemia) Elevated blood sugar levels are common, even in patients without a history of diabetes [57]. This is usually interpreted as part of the body's stress response. Higher glucose levels have been associated with worse outcomes, though it is not always clear whether they play a direct role or reflect overall severity [57].

Electrolyte and Fluid Balance Disturbances in sodium levels, especially low sodium (hyponatremia), occur frequently after aSAH [16, 34]. These changes are often accompanied by shifts in fluid balance, which may affect blood flow to the brain and could contribute to the risk of DCI [7, 34].

Treatment Factors Treatment may also influence how patients progress. Nimodipine is widely used and has been linked to a lower risk of DCI [16, 29]. Aneurysm repair can be done through endovascular coiling or surgical clipping. There is some evidence suggesting differences in outcomes between these approaches, although this can depend on patient and aneurysm characteristics [34].

2.1.4 Additional Clinical Context and Systemic Factors

Neurological severity The level of consciousness at presentation is commonly used to get a sense of how severely a patient is affected after aSAH [16, 83]. It reflects, at least in part, the extent of the initial brain injury and the broader physiological response. Reduced consciousness is often seen in patients with larger volumes of subarachnoid blood or with signs of cerebral edema, although this relationship is not always straightforward [16, 22]. In clinical practice, the Glasgow Coma Scale (GCS) is typically used as a practical proxy for this severity, mainly because it is standardised and relatively consistent between observers [16, 26]. Lower GCS scores are generally associated with poorer outcomes, but in this setting the measure is used more as a baseline description of neurological status than as a specific indicator of delayed complications [83, 105].

Physiological instability Following the initial hemorrhage, many patients show signs of broader physiological instability. Blood pressure can become harder to keep stable, and periods of hypotension (low blood pressure) can reduce blood flow to the brain (cerebral perfusion), especially when the brain can no longer regulate its own blood flow effectively [10, 18]. Respiratory factors such as hypoxia (low oxygen levels) or hypercapnia (elevated carbon dioxide) can also affect how blood vessels behave and how well oxygen is delivered to the brain [10, 112]. These features are not necessarily direct drivers of vasospasm, but they do give an indication of how stable or unstable the overall clinical picture is [33].

Broader clinical and baseline context The way patients progress after aSAH also seems to depend on their underlying health status. Factors such as smoking and hypertension (chronically high blood pressure) are linked to aneurysm formation and rupture, and may influence vascular behaviour more broadly [22, 50]. Other conditions, including diabetes, thromboembolic history, or cardiac rhythm disturbances, form part of the clinical background at the time of admission

[16, 22, 34]. Additional elements, such as nutritional status or the occurrence of seizures, can be seen as markers of the patient’s overall condition [16].

2.1.5 Limitations of Clinical Knowledge

Incomplete understanding Despite decades of research, the mechanisms behind cerebral vasospasm after aSAH are still not fully understood [18, 34]. Blood in the subarachnoid space is generally considered an important trigger, but the steps from this initial bleeding to sustained arterial narrowing, reduced cerebral perfusion, and later brain injury remain difficult to separate [34]. One reason this is challenging is that vasospasm is not an isolated process. It occurs together with early brain injury, inflammation, hydrocephalus, disturbed vascular regulation, metabolic stress, and intensive monitoring or treatment.

Association and causation Many reported risk factors for vasospasm and poor outcome after aSAH come from observational studies. These associations are clinically useful, but they do not automatically show whether a variable contributes causally to vasospasm, reflects more severe initial injury, or is part of the subsequent care pathway. For example, hyperglycemia has been associated with poor recovery after aSAH, but this does not by itself prove that high glucose is a direct cause of worse outcome [57]. Similarly, associations between fluid management strategies and clinical outcomes do not necessarily imply that the management strategy itself caused the observed outcome [22].

This distinction is important for the present thesis. Variables may be related in the EHR, but the meaning of these relations is not obvious. Some may reflect biological processes, while others may reflect severity, treatment decisions, monitoring intensity, or documentation patterns. This uncertainty motivates the use of causal discovery as a hypothesis-generating approach.

2.2 Clinical Data Representation in EHR-Based Research

2.2.1 EHR data as a research substrate

Electronic Health Records (EHRs) are longitudinal digital collections of patient health information originally developed to support clinical and administrative workflows, including registration, scheduling, and billing [52]. Unlike clinical trial data, EHR data are not collected according to a predefined research protocol. They are generated when patients interact with the healthcare system, and the timing and content of the record depend on clinical need, documentation habits, and administrative processes [4]. This is especially relevant for causal research. In a randomized trial, treatment assignment and follow-up are designed to make causal effects easier to estimate. In routine EHR data, by contrast, treatments, measurements, and diagnoses are all shaped by the patient’s condition and by the care process itself. This makes it harder to separate cause, consequence, confounding, and documentation effects.

EHR data therefore provide a rich but difficult research substrate. They contain heterogeneous data types, ranging from numerical laboratory values to free-text narratives, recorded at irregular intervals depending on clinical context [87, 92]. Because these data arise from routine care, they do not directly form a coherent research dataset and usually require substantial cleaning, normalization, and transformation before analysis [52, 87].

2.2.2 Modalities of clinical information

Structured EHR data consist of organized fields that can be queried and analyzed more directly, such as diagnosis codes, procedure codes, medication records, laboratory results, vital-sign measurements, and ICU flowsheet entries [53, 87].

Diagnoses and Procedures Clinical states are commonly encoded using standardized terminologies such as the International Classification of Diseases (ICD) for diagnoses and Current Procedural Terminology (CPT) for procedures. These coding systems are primarily used for administrative and billing purposes [43, 52, 87, 92].

Laboratory Measurements Laboratory data include results from blood and fluid testing (like electrolyte levels, cell counts). These measurements are often mapped to standards such as Logical Observation Identifiers Names and Codes (LOINC) to support consistency across institutions [43].

ICU Monitoring and Charted Variables In intensive care settings, high-frequency data are recorded, including vital signs (temperature, pulse rate, blood pressure), sensor measurements, and flowsheets capturing fluid balance or ventilator settings [87].

Unstructured data A large part of clinically relevant information is stored in unstructured, free-text narratives that provide context not captured in coded fields [53, 104]. Discharge Summaries describe the course of a patient’s hospital stay, including presenting symptoms, treatments, and conclusions at discharge or transfer [53, 87]. Radiology reports, on the other hand are written interpretations of imaging studies such as CT, MRI, or X-ray. They describe observed findings and often note the presence or absence of specific conditions [17, 52].

2.2.3 Structural limitations of EHR data

The use of EHR data for research is influenced by the way in which the data are recorded. Since documentation is shaped by clinical workflows and administrative needs, it represents an indirect view of a patient’s underlying health state [43].

Temporal ambiguity and irregular sampling are common challenges. Data are not recorded at fixed intervals but follow the timing of clinical encounters, meaning that the recorded time of a diagnosis may not reflect when the condition actually began [43, 87]. Coding practices can also vary, as ICD codes are assigned differently across settings and may be influenced by administrative considerations or by financial incentives [69, 71].

EHR data also exhibit non-random missingness, often referred to as informative missingness [61, 119]. The absence of a measurement may reflect clinical judgment, access to care, or other factors [61]. In addition, patient information can be fragmented across different healthcare systems that do not share data, which can result in incomplete longitudinal records [61].

2.2.4 Cross-modal representation and disagreement

Because information is distributed across different data types, the same clinical state may be represented in different ways. This can make it difficult to define a single ground truth [52]. For example, a condition may be described in a radiology report or clinical note, while the corresponding ICD code is absent or does not fully match [87].

Signals across modalities may also be inconsistent. A diagnostic code might be recorded to justify a test, even if the test results do not support the diagnosis, yet the code remains part of the record [43]. As a result, structured data and narrative text often provide partial and sometimes conflicting views, without a standard way to reconcile them.

2.3 Information Extraction from Clinical Text

Clinical text contains information that is difficult to recover from structured EHR tables alone. In this thesis, this is especially relevant for variables such as vasospasm evidence, aneurysm

characteristics, and treatment context, which are often described in discharge summaries or radiology reports and not stored as clean database fields. This section briefly reviews how information extraction methods have been used to turn such clinical narratives into structured variables, moving from rule-based systems toward more recent LLM-based approaches.

2.3.1 Traditional NLP approaches

Early work in clinical natural language processing mainly relied on rule-based extraction and concept mapping [59]. Systems such as cTAKES and MetaMap use handcrafted rules, pattern matching, and domain-specific lexicons to identify medical entities in text [81]. After extracting the information, it is translated into a common medical dictionary (like UMLS or SNOMED CT) so that different terms all mean the same thing and can be compared consistently [81,116].

A limitation of these rule-based approaches is that they are sensitive to the variability of clinical text. Clinical notes often contain abbreviations, informal phrasing, spelling variation, and institution-specific wording [81,116]. Because rule-based systems depend on predefined patterns, they can achieve high precision when the wording matches the rule, but they may miss relevant information when the same clinical concept is expressed differently [62].

They are also difficult to scale to more interpretive extraction tasks. In this thesis, for example, some variables require more than recognising a keyword. Distinguishing aneurysmal from non-aneurysmal SAH, separating routine preventive treatment from treatment given because vasospasm is already suspected, or interpreting temporal evidence across reports requires context that is hard to capture with fixed rules. Rule-based systems therefore remain useful for narrow and well-defined extraction tasks, but adapting them to broader clinical reasoning tasks can require substantial manual work [116].

2.3.2 Machine learning / deep learning NLP

Later approaches introduced supervised learning and deep learning models, including sequence-based architectures such as Recurrent Neural Networks and Long Short-Term Memory networks [13, 58, 80]. With the introduction of transformer models, it became possible to learn more context-sensitive representations of text [8,80]. Models such as BERT, along with domain-specific variants such as BioBERT (trained on biomedical literature) and ClinicalBERT (trained on EHR data such as MIMIC-III), have been widely used for tasks such as named entity recognition and relation extraction [46,58].

A limitation of many supervised machine learning and deep learning approaches is that they require annotated training data. For clinical text, such annotations are often expensive to create because they need domain expertise and careful review [44,45]. This becomes a practical problem when the target variables are specific to one clinical setting, such as vasospasm evidence, aneurysm rupture context, or hydrocephalus interpretation.

Another limitation is input length. Many transformer-based clinical NLP models operate within a restricted context window, often around the length of a single short text segment rather than a full admission record. This makes it harder to process long discharge summaries or combine information across multiple radiology reports from the same admission. Although some architectures aim to extend this limit, standard clinical BERT-style models often remain constrained in this respect [60].

2.3.3 Large Language Models in clinical NLP

LLMs such as the GPT series and Llama family have recently been applied to clinical information extraction through the use of zero-shot and few-shot learning [2,5]. These models can perform extraction tasks with little or no task-specific training, using prompts and a small number of examples [2,94]. This differs from earlier approaches that typically rely on large annotated datasets.

Prompt-based Extraction Performance depends to a large extent on how tasks are formulated. Prompting strategies, including Chain-of-Thought (CoT) approaches, can help guide the model when dealing with more complex clinical text [88, 95]. LLMs are often able to identify entities such as medications or laboratory values, even when no explicit training data for the task is available [5]. Some studies suggest that a small number of examples can already lead to reasonable performance, in some cases comparable to more specialized systems [2].

Structured Output LLMs can be instructed to return outputs in structured formats, such as JSON or predefined schemas [2, 9]. This makes it easier to integrate extracted information into downstream analyses. In many cases, this reduces the amount of post-processing that would otherwise be needed [5].

Reasoning Across Documents More recent models support longer input contexts, which makes it possible to process multiple documents together [56]. This can be useful in clinical settings, where relevant information is often spread across notes and reports. In practice, approaches such as chunking are still used to manage longer texts while preserving context [116].

Dealing with Uncertainty Reliability can be improved through uncertainty quantification, where models can explicitly communicate their confidence in an extraction or diagnosis [54]. In addition, models can be prompted to provide supporting text fragments, which makes it easier to trace outputs back to the original documentation [54, 116].

2.3.4 Limitations of LLM-based extraction

Applying LLMs to clinical text also introduces risks that need to be controlled. One concern is hallucination: a model may produce information that appears plausible but is not supported by the source text [2, 9, 54]. In clinical information extraction, this is especially problematic because an unsupported output can look like a real clinical fact once it has been converted into a structured variable.

A second limitation is inconsistency. Outputs can vary between runs, even when the input is the same, and they can be sensitive to prompt wording, formatting, and the order in which information is presented [86]. When extracted variables are used downstream, small changes in extraction behaviour can propagate into cohort construction, validation, or causal discovery.

Evaluation is also difficult. For many clinical extraction tasks, there is no single unquestionable ground truth. Clinical notes may be incomplete, different sources may disagree, and expert annotators may interpret ambiguous cases differently [54]. In this thesis, these risks are addressed by using structured output schemas, low-temperature inference, evidence quotes, uncertainty categories, cross-modal comparisons, and manual review of selected disagreement cases.

2.4 Causal Inference in Observational Clinical Data

The increasing availability of EHRs, billing data, and clinical registries has made observational research more prominent in clinical studies, especially in situations where randomized controlled trials (RCTs) are not feasible due to cost, ethical constraints, or practical limitations [3, 72, 96]. While RCTs are often considered the reference standard because randomization helps make treatment groups comparable, observational studies require explicit causal assumptions and adjustment strategies to address confounding [23].

A useful way to think about causal inference is as a missing data problem: for any individual, we can only observe the outcome under the treatment they actually received, not the outcome under alternative treatments [41, 49].

2.4.1 Correlation vs. Causation

A central point in causal inference is that association does not imply causation [42]. An association reflects differences observed between subsets of the population (for example, treated versus untreated), whereas causation involves a contrast of potential outcomes within the same population (counterfactual worlds) [41, 49]. Causal effects are quantified by comparing what would happen if the entire population were exposed versus what would happen if the same population remained unexposed.

Two main sources of bias often prevent observed associations from reflecting causal effects:

Confounding Confounding occurs when both the treatment and the outcome are influenced by a common cause [41, 49]. This can introduce associations that are not causal in nature. In clinical settings, examples include confounding by indication, where sicker patients are more likely to receive a treatment, and also more likely to experience adverse outcomes [96].

Selection Bias Selection bias can arise from how individuals enter a study, or from conditioning on variables that are affected by both treatment and outcome [3]. In particular, conditioning on a common effect (a collider) can introduce artificial associations that would not otherwise be present [3].

2.4.2 Classical causal inference with a specified question

Classical causal inference usually starts from a defined causal question. A treatment or exposure is chosen, an outcome is specified, and the analysis then tries to estimate what would have happened under different exposure conditions. In the potential-outcomes framework, this is often described as comparing the outcome observed under the treatment a patient actually received with the unobserved outcome that would have occurred under an alternative treatment [41, 49].

In observational clinical data, this comparison is difficult because treated and untreated patients are usually not comparable at baseline. Sicker patients may be more likely to receive an intervention, and they may also be more likely to experience poor outcomes. Classical approaches therefore try to make comparisons more meaningful by adjusting for measured confounders. Regression models do this by conditioning on covariates in the outcome model [49]. Propensity-score methods instead model the probability of receiving treatment given observed variables, and then use matching or weighting to balance treated and untreated groups [84, 91].

Directed acyclic graphs (DAGs) are often used alongside these methods to make the assumed causal structure explicit [28]. A DAG can help decide which variables should be adjusted for and which variables should not, for example by identifying backdoor paths or possible collider structures [15]. In that sense, DAGs are not an estimation method by themselves, but a way to reason about the assumptions behind an effect estimate.

These approaches are highly useful when the exposure, outcome, and causal assumptions are already well defined. For example, they could be used to study a specific question such as whether early aneurysm treatment affects mortality, or whether a glucose-control strategy affects outcome, provided that timing and confounding can be represented carefully.

2.4.3 Why this motivates causal discovery

The question in this thesis is different. The aim is not to estimate one predefined treatment effect. Instead, the aim is to explore which causal relations around vasospasm are plausible enough to become hypotheses for further study. The causal structure around vasospasm is not fully known in advance. Variables may be connected through several pathways, including biological mechanisms, disease severity, treatment decisions, monitoring intensity, and documentation practices.

Classical causal inference methods depend strongly on the structure supplied by the analyst. If the assumed graph is incomplete or wrong, the chosen adjustment set may also be wrong. This can lead to residual confounding, adjustment for mediators, or adjustment for colliders [3, 99]. In a setting where the relevant causal pathways are still uncertain, the difficulty is therefore not only how to estimate an effect, but also which relations should be considered in the first place.

Causal discovery addresses this earlier-stage problem. It uses patterns in the data, together with background knowledge, to generate candidate graph structures. These structures do not prove causal effects, and they still depend on modelling assumptions. However, they can help identify recurring relations that are worth interpreting clinically and testing more directly in later work. In this thesis, causal discovery is therefore used as a hypothesis-generating step: it helps move from a large multimodal EHR dataset toward a smaller set of candidate causal relations around vasospasm.

2.5 Causal Discovery Methods

Causal discovery studies how causal relations between variables can be inferred from observational data [47]. The output is typically a graph, in which nodes represent variables and edges represent relations between them. The same graphical notation can be used for purely probabilistic models, but in causal discovery the graph is interpreted as a candidate causal structure. Under this interpretation, a directed edge such as $A \rightarrow B$ represents a direct causal relation from A to B , relative to the variables included in the analysis. The absence of an edge indicates that the method did not retain a direct relation between the two variables in the learned structure.

This interpretation is not automatic proof of causality. It depends on the discovery method, the available variables, and the assumptions that connect statistical patterns in the data to causal structure [117]. Common assumptions include the Causal Markov condition, which states that each variable is independent of its non-effects once its direct causes are known, and the Faithfulness assumption, which states that the independencies observed in the data reflect the causal structure rather than accidental cancellations of effects. These assumptions are used in many causal discovery settings, including both constraint-based and score-based approaches [89, 98].

PC-style methods also typically assume causal sufficiency, whereas FCI-type methods relax this by allowing for latent confounding and selection effects [98, 118]. In clinical data, these assumptions should be treated cautiously because unmeasured severity, measurement error, treatment-selection effects, and temporal aggregation can all influence the recovered graph [68].

2.5.1 Constraint-Based Methods

Constraint-based methods, such as the PC algorithm and its variant PC-stable, infer graph structure through conditional independence (CI) testing [24]. These approaches typically start from a fully connected undirected graph and remove edges when conditional independence is detected for some conditioning set [117]. The remaining edges are then oriented using rules based on v-structures and additional logical constraints [117].

PC-stable modifies the original PC algorithm by storing adjacency sets at each level of testing. This reduces sensitivity to variable ordering and tends to produce more stable skeletons, especially in higher-dimensional settings [24].

Strengths A practical advantage of constraint-based methods is that the edge-removal process is closely tied to explicit CI tests, which makes the learned skeleton relatively transparent. This test-based structure also allows missing data to be handled at the level of individual CI tests: with test-wise deletion, each test uses only the records complete for the variables involved in that

specific test, instead of discarding all rows with any missing value across the full dataset [113]. This can be useful in clinical datasets, where missingness often differs across measurement blocks.

Weaknesses Their performance depends strongly on the reliability of the CI tests. Since many tests are performed across different conditioning sets, errors can accumulate and affect both the learned skeleton and later edge orientations. This makes the results sensitive to sample size, noise, variable discretisation, and the choice of significance threshold, especially in smaller or heterogeneous clinical datasets [55, 100].

2.5.2 Score-Based Methods

Score-based methods, such as Fast Greedy Equivalence Search (FGES), frame causal discovery as an optimization problem over the space of possible graphs. Candidate structures are assigned a goodness-of-fit score, such as the Bayesian Information Criterion (BIC), which rewards graphs that fit the observed data well while penalizing unnecessary complexity. The algorithm then iteratively adds, removes, or reorients edges in an attempt to improve this score. [77].

Strengths Score-based methods evaluate candidate structures using a graph-level objective, so they are not limited to a sequence of local CI-test decisions. During the search, earlier edge choices can be revised if another structure improves the score. Under suitable assumptions and scoring choices, score-based methods can also be statistically consistent in the large-sample limit, meaning that they can recover the correct equivalence class when the required conditions hold [89].

Weaknesses Their behaviour depends on the assumptions behind the chosen scoring function and on the way the graph search is carried out. Distributional assumptions, penalty terms, and search restrictions can all influence the resulting structure [89]. In practice, greedy search procedures may also converge to locally optimal graphs, so the learned structure can still depend on algorithm settings and initialization choices.

2.5.3 PAG-Based Methods and Latent Confounding

PC-style methods return a graph under the assumption that all relevant common causes are observed. This assumption is often difficult to defend in clinical data. Fast Causal Inference (FCI) and related PAG-based methods address this limitation by allowing for latent confounders and selection effects [98, 118].

Instead of producing a single directed acyclic graph (DAG), FCI-type methods return a Partial Ancestral Graph (PAG), which represents a class of causal structures compatible with the observed conditional independencies [118]. Different edge markings encode different levels of orientation certainty: arrowheads rule out ancestry in a given direction, tails indicate ancestry, and circle endpoints indicate that the direction cannot be determined from the data alone [98].

Figure 2.4 gives a compact visual summary of the graph notation used in this thesis and highlights how DAG and PAG outputs encode different levels of causal certainty.

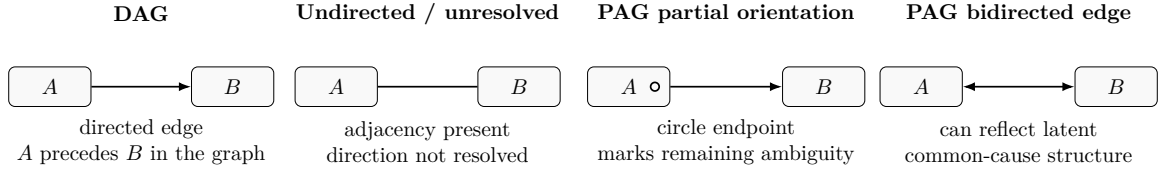


Figure 2.4: Examples of DAG and PAG edge notation. In these graphs, nodes represent variables and edges represent hypothesized direct relations. DAGs encode fully directed relations when the orientation is taken as resolved, whereas PAGs allow partial orientation and ambiguity when the data are compatible with multiple causal structures or possible latent confounding.

FCIT further modifies this approach by combining a score-based initialization with targeted CI testing. It starts from a score-based CPDAG, often obtained with BOSS, converts this structure into a preliminary PAG, and then refines it using targeted tests that are chosen from path-based information rather than from exhaustive adjacency-subset search [76]. This is intended to reduce the repeated-testing problem in standard FCI-style algorithms, where many CI tests can increase the chance of unstable or spurious independence decisions in finite samples [76].

Strengths The main advantage of FCIT is that it retains the PAG-based representation of FCI-type methods, and therefore does not require the same causal sufficiency assumption as PC-stable or FGES. This is useful in clinical data, where unmeasured severity, treatment-selection processes, or undocumented clinical decisions may act as latent common causes. Compared with standard FCI-style searches, FCIT uses targeted, score-guided testing to reduce the number of CI tests and includes checks that maintain a well-formed PAG during refinement [76]. This can improve finite-sample stability and make the output easier to interpret than a malformed or overly ambiguous PAG.

Weaknesses FCIT still depends on several modelling choices, including the score-based initialization, the CI test, and the assumptions behind both. If the initial score-based structure is poor, the later targeted refinement may also be affected. In addition, although FCIT is designed to be more efficient than exhaustive FCI-style testing, it remains more complex than simpler methods such as PC-stable or FGES. Finally, bidirected edges should still be interpreted cautiously: even in simulations designed for latent-variable discovery, bidirected latent-edge precision can remain low in larger graphs [76].

2.5.4 Stability and Robustness

A recurring challenge in causal discovery is that results can vary depending on the data sample, algorithm, or modelling choices. Small changes in the data, or in how the algorithm is configured, may lead to different structures [67]. This is especially important in clinical applications, where the true causal graph is usually not known and evaluation cannot rely on direct comparison with ground truth.

Bootstrapping and Resampling Resampling techniques, such as bootstrapping, are often used to assess stability. Edges that appear consistently across multiple resampled datasets are generally considered more robust than edges that appear only sporadically [11, 67]. However, stability should not be interpreted as proof of causality, since a stable edge can still reflect bias, unmeasured confounding, or consistent measurement practices.

Hyperparameter Sensitivity Most methods require the specification of hyperparameters, such as significance thresholds for CI tests or penalty terms in score-based methods. The choice

of these parameters can have a noticeable impact on the resulting graph, and in some cases may influence results as much as the choice of algorithm itself [64].

Evaluation without a known ground truth A further difficulty is that causal discovery results are hard to evaluate when the true causal graph is unknown [68]. In synthetic studies, recovered graphs can be compared with a predefined data-generating structure, but this is rarely possible in real clinical data [100]. For ICU and EHR data, the true graph is not observed, interventions may be ethically or practically impossible, and observational data may only identify a set of statistically equivalent graphs rather than one unique DAG [97,117]. Evaluation therefore becomes less a matter of proving that a discovered graph is correct, and more a matter of checking whether the result is plausible, stable, and not easily falsified [68,90]. Relevant checks include consistency with temporal ordering and clinical knowledge, robustness to hyperparameter choices and resampling, comparison across algorithmic families, sensitivity to the included variable set, and external validation when suitable data or prior evidence are available [68].

2.6 Causal Discovery in Clinical Applications

Applying causal discovery methods to clinical data, particularly from ICU and EHR, has received increasing attention in recent years. The goal is to move beyond purely predictive models and toward representations that may offer more insight into underlying clinical processes. In this context, identifying patterns that remain stable across different settings is often of interest, as interventions based solely on statistical associations may not translate into meaningful clinical effects [68,90].

Several studies have used large observational clinical datasets, including MIMIC-III and MIMIC-IV, to study causal questions in intensive care [30,36]. A recent sepsis study by Petrun-garo et al. used Bayesian network structure learning to investigate potential causes of sepsis, combining clinical expertise with score-based, constraint-based, and hybrid algorithms to obtain a consensus structure [74]. Such work illustrates the growing interest in graph-learning methods for clinical hypothesis generation, although much of the broader literature still focuses on predefined interventions or outcomes rather than exploratory recovery of candidate clinical structures. Across applied studies that use graph-learning approaches, several recurring methodological patterns can be observed. These patterns help position the present work, which places more emphasis on cross-method comparison, multimodal variable construction, and stability of recovered relations.

However, in applied clinical causal discovery, studies often rely on one algorithm or one family of algorithms, such as constraint-based methods, score-based methods, or Bayesian network structure learning. This can make the results difficult to interpret in isolation, because different discovery paradigms may recover different graph structures from the same data. Comparisons between methods such as PC-style algorithms, score-based search, and FCI/PAG-based approaches are therefore useful for assessing how dependent the recovered relations are on specific modelling choices [97].

Moreover, stability is an ongoing concern in this area. Inferred structures can change when the data are perturbed, for example through noise, sampling variation, or shifts in patient populations. Not all studies systematically evaluate this sensitivity, and procedures such as extensive hyperparameter exploration or resampling-based validation are not always included. As a result, findings may depend more strongly on specific modeling decisions than is immediately apparent [64].

Compared with the broader literature on multimodal prediction in healthcare, causal discovery from multimodal clinical data remains less established [63]. Multimodal prediction studies usually combine sources such as structured EHR variables, imaging, waveforms, and clinical text to improve outcome prediction. Causal discovery has a different aim: it tries to infer pos-

sible causal structure from observed data under explicit assumptions. This makes the analysis more sensitive to how variables are defined, how heterogeneous sources are combined, and which assumptions are imposed on the graph [63].

Most applied causal discovery studies in medicine still rely mainly on structured tabular variables. This is understandable, since standard algorithms such as PC or FCI usually require a fixed set of variables with well-defined values across observations [106]. Clinical text does not naturally have this form. Discharge summaries and radiology reports contain temporal descriptions, uncertainty, negation, clinical interpretation, and source-specific wording [104]. They can include information that is missing from coded data, but this information must first be converted into structured variables before it can be used in a graph-learning workflow.

2.7 Synthesis and Research Gap

The literature reviewed above points to a gap at the intersection of three areas: vasospasm after aSAH, multimodal EHR representation, and causal discovery. Vasospasm is clinically important because it can contribute to delayed neurological deterioration after the initial hemorrhage, yet the mechanisms behind it remain difficult to separate. Several factors, including hemorrhage severity, early secondary complications, treatment decisions, and monitoring intensity, are associated with vasospasm. However, these associations do not by themselves show which variables may be causes, consequences, severity markers, or parts of the care process.

At the same time, routine ICU data contain more information than structured codes alone. In aSAH, relevant evidence may appear in diagnosis codes, procedure records, medication tables, laboratory measurements, bedside monitoring, discharge summaries, and radiology reports. This is useful for studying vasospasm, but it also creates a representation problem. The same clinical state can be recorded differently across sources, and clinical text often contains timing, uncertainty, negation, and interpretation that are not available in coded fields.

Causal discovery is relevant in this setting because it is designed to generate hypotheses about possible causal structure when the relations between variables are not fully known in advance. However, most applied clinical causal discovery work still relies mainly on structured tabular variables. The use of multimodal clinical information, especially variables extracted from free-text notes and radiology reports, is less developed. This means that the uncertainty introduced by information extraction, source disagreement, and variable harmonisation needs to be considered as part of the causal discovery problem itself.

2.7.1 Clinical gap

For aSAH, the causal structure around vasospasm is still not fully clear. Existing clinical knowledge identifies several risk factors and associated processes, but it remains difficult to distinguish direct contributors from markers of severe hemorrhage, treatment responses, or documentation effects. This motivates an exploratory analysis focused on hypotheses about possible causal relations around vasospasm, rather than only on prediction or association.

2.7.2 Data gap

Most structured EHR variables provide only a partial view of vasospasm and related complications. ICD codes, procedure records, medication tables, laboratory values, discharge summaries, and radiology reports each capture different parts of the admission. In particular, clinical notes can contain information about vasospasm evidence, aneurysm context, hydrocephalus, EVD presence, and treatment interpretation that is missing or incomplete in structured tables. A major data challenge is therefore how to combine these modalities into an admission-level representation that remains clinically interpretable and traceable to its source.

2.7.3 Methodological gap

The methodological gap is not only that causal discovery has rarely been applied to this specific disease setting. It is also that multimodal causal discovery in clinical data raises extra difficulties. The graph depends on how variables are extracted from text, how overlapping sources are harmonised, and how disagreements between modalities are handled. In addition, different causal discovery algorithms can recover different structures from the same dataset, and results can be sensitive to hyperparameters or resampling. This makes cross-method comparison, source provenance, and stability analysis important parts of the methodology, not just optional checks.

2.7.4 Contribution

This thesis addresses these gaps by constructing a disease-specific multimodal causal discovery pipeline for aSAH admissions in MIMIC-IV. Structured EHR variables are combined with variables extracted from discharge summaries and radiology reports, with explicit attention to how signals from different sources align or differ.

The resulting admission-level dataset is then used for causal discovery with PC-stable, FGES, and FCIT across hyperparameter sweeps and stability analyses. The aim is not to produce a definitive causal model of vasospasm, but to identify recurrent and clinically interpretable candidate structures that can guide further clinical interpretation and follow-up analysis.

Chapter 3

Data and Cohort Construction

3.1 Data source and environment

The dataset combines hospital-wide administrative data, ICU-specific monitoring data, and free-text clinical notes. These sources differ not only in structure but also in how clinical information is recorded, introducing variation in timing, granularity, and completeness.

Data processing was implemented with DuckDB and Python. DuckDB was used for SQL-based extraction and joins across large MIMIC-IV tables, while Python was used for orchestration, feature construction, validation checks, and figure generation. Intermediate outputs were stored as parquet files so that each pipeline stage could be inspected, rerun, and reused without repeating the full extraction process.

3.2 Raw data sources and modalities

The analysis draws on multiple components of MIMIC-IV. Table 3.1 summarizes the modules used and their role in the pipeline.

Table 3.1: Overview of MIMIC-IV modules used in this thesis

Module	Role in pipeline
<code>admissions, patients</code>	Demographics, admission timing, mortality
<code>diagnoses_icd, procedures_icd</code>	Diagnoses and coded procedures
<code>labevents</code>	Laboratory measurements
<code>emar, prescriptions</code>	Medication administration and orders
<code>icustays</code>	ICU episode structure
<code>chartevents</code>	High-frequency physiological measurements
<code>procedureevents</code>	ICU procedures (e.g. EVD placement)
Discharge summaries	Narrative clinical summaries
Radiology reports	Imaging findings and interpretations

While all modules contribute to the final dataset, they differ in how directly they reflect clinical reality. Administrative codes capture standardized diagnoses and procedures but may miss nuance or timing. ICU and laboratory data provide detailed physiological signals but are often sparse or irregular outside critical care. Free-text notes contain the richest clinical descriptions, though they require interpretation.

Variables are constructed by combining evidence across modalities where appropriate to avoid privileging a single source. This is particularly relevant for conditions such as vasospasm, where no single structured signal is consistently reliable.

3.3 Cohort construction

The target cohort for this thesis consists of hospital admissions that most likely represent aSAH. Subarachnoid hemorrhage codes in MIMIC-IV can also include non-aneurysmal, traumatic, angiogram-negative, or otherwise mixed hemorrhage contexts. Since the later analysis focuses on vasospasm after aSAH, the cohort construction step aimed to retain admissions with evidence for aneurysmal origin, while excluding broader SAH admissions where this was not supported.

The cohort was defined at the admission level, using `(subject_id, hadm_id)` as the unit of analysis. All data sources were merged under the constraint that each admission appears at most once in the final dataset. This admission-level uniqueness constraint was explicitly checked after each merge step to avoid silent duplication or misalignment.

Candidate admissions were identified through two complementary strategies. The first relied on ICD-based identification of SAH, selecting admissions with relevant diagnosis codes. The second captured admissions without SAH coding but with evidence of aneurysm-related treatment or medication use, such as clipping, coiling, or nimodipine administration.

Because ICD coding alone does not reliably distinguish aneurysmal from non-aneurysmal SAH, a further adjudication step was applied using language models over discharge summaries and radiology reports. This step operates at the admission level and produces a structured assessment of the likelihood of aneurysmal SAH, together with supporting evidence extracted from the notes. The LLM-based extraction and adjudication procedure is described in more detail in Chapter 4.

3.4 Structured data scope and coverage

After cohort construction, all structured sources were linked at admission level using `(subject_id, hadm_id)` as the unit of analysis. At this stage, the purpose was not to define every analytical variable in detail, but to establish which source tables were available for each admission and how complete the different modalities were.

Coverage varied across sources. Administrative tables were available for the full cohort, while laboratory, ICU, and bedside charted variables depended on clinical measurement and ICU exposure. Free-text notes were also linked at admission level, with discharge summaries and radiology reports serving different roles: discharge summaries provided retrospective clinical summaries, while radiology reports captured imaging-specific evidence.

These coverage differences were retained for later interpretation. The detailed construction of structured variables, including temporal constraints, aggregation rules, physiologic thresholds, missingness handling, and validation checks, is described in Section 4.1.

3.5 Final cohort summary

After applying the ICD-based selection and the LLM adjudication step, the final cohort consists of 470 unique admissions. Most admissions (426) come from the ICD-based SAH group, with an additional 44 identified through the treatment-based pathway without explicit SAH coding.

This cohort is used in the following steps of the analysis.

Chapter 4

Methodology

This chapter describes the methodological workflow used to move from the final likely aSAH cohort to the causal discovery results. The work consisted of several connected steps: retrieving clinically relevant variables from structured MIMIC-IV tables, extracting additional information from discharge summaries and radiology reports, harmonising overlapping signals across modalities, preparing a discrete admission-level matrix, and applying causal discovery with stability and sensitivity analyses. Because these steps depend on each other, Figure 4.1 first gives a compact overview of the full workflow. The sections that follow then describe each component in more detail, starting with the structured variable retrieval pipeline and ending with the causal discovery implementation.

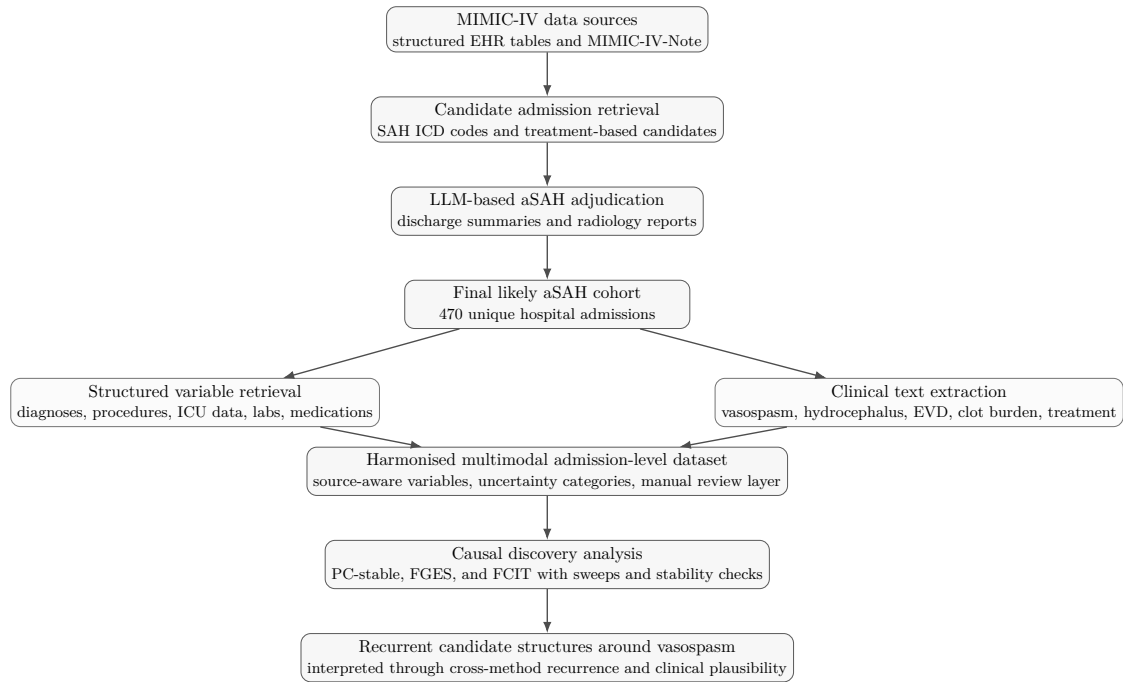


Figure 4.1: Overview of the study pipeline. Candidate admissions were retrieved from structured MIMIC-IV data, adjudicated using clinical text, and converted into a harmonised multimodal admission-level dataset before causal discovery was applied.

4.1 Structured variable retrieval pipeline

Chapter 3 described how the likely aSAH cohort was assembled and how the available MIMIC-IV sources were linked at admission level. The next step was to turn these linked records into

clinically interpretable variables.

The purpose of this structured retrieval step was deliberately broad. At this stage, the goal was not yet to select the final causal discovery node set, but to recover as much relevant structured information as possible about baseline status, hemorrhage severity, treatment, ICU course, physiological instability, complications, and outcomes. Variables could then be inspected, validated, combined with note-derived signals, and later filtered for the causal discovery analysis.

The pipeline was organised around feature groups. Each group started from one or more MIMIC-IV source tables and produced admission-level variables with explicit provenance. This made the dataset easier to audit and helped preserve the link between a final encoded value and the measurements, codes, procedures, or medication records from which it was derived.

4.1.1 Retrieval logic and staged feature construction

The structured retrieval proceeded in two linked stages. The first stage attached variables that could be derived directly from diagnosis tables, procedure tables, medication records, ICU stay information, bedside charting, note linkage, and admission metadata. These variables covered cohort provenance, administrative vasospasm coding, demographics, baseline history, aneurysm treatment, imaging-related coded findings, ICU exposure, neurological severity, treatment timing, EVD placement, antifibrinolytic therapy, ECG summaries, linked note identifiers, and death-related outcomes.

The second stage constructed compound physiological feature blocks. These blocks were used for clinical states that could not be represented well by a single table entry. For example, hyperglycemia, leukocytosis, dysnatremia, oxygenation, hypotension, hypercapnia, hypovolemia, and hypomagnesemia could appear through repeated laboratory measurements, ICU charted values, diagnosis codes, or treatment signals. The goal was to summarise these repeated and source-specific signals into interpretable admission-level states, while retaining the supporting measurements and indicators where useful.

The level of detail retained depended on the variable. For variables with a relatively direct meaning, such as sex, documented death, or the presence of an ICD-coded diagnosis, a binary or categorical indicator was usually sufficient. For variables based on repeated physiology, a single binary value would hide too much information. In those cases, the pipeline retained measurement counts, extrema, threshold crossings, coded evidence, and treatment-supported evidence alongside the final categorical state. This made it possible to distinguish, for example, one isolated abnormal glucose measurement from repeated or more severe hyperglycemia.

The output of this stage was therefore not a minimal causal discovery table, but a source-aware structured feature set. Later methodology sections describe how these variables were harmonised with LLM-derived variables and reduced to the final discovery-ready representation.

4.1.2 Core structured variables

The first block of structured retrieval attached variables that were either directly available at admission level or could be reduced to admission-level indicators through grouped queries and constrained joins. These variables covered cohort provenance, administrative vasospasm coding, demographics, baseline history, aneurysm treatment, imaging-related coded findings, ICU exposure, neurological severity, treatment timing, EVD placement, antifibrinolytic therapy, ECG summaries, linked note identifiers, and death-related outcomes. The ICD codes used for this extraction is provided in Appendix A, whereas the full variable list is provided in Appendix B.

Cohort provenance. The variable `cohort_source` recorded how the indexed admission entered the final unified aSAH cohort: through the ICD-restricted LLM-confirmed branch, through

the LLM-only branch without qualifying SAH ICD coding, or through overlap between both sources.

Administrative vasospasm indicator. An initial target-side vasospasm signal was derived from `diagnoses_icd`. This variable does not represent the final vasospasm definition used throughout the thesis, but a conservative administrative trace of documented vasospasm. The pipeline attached both a binary flag, `vasospasm_flag`, and the first matching code, `vasospasm_icd_code`. Keeping this ICD-based signal separate made it possible to compare administrative coding with later note-derived vasospasm evidence from discharge summaries and radiology reports.

Demographics. Sex was obtained directly from `patients`. Age at admission was reconstructed from the MIMIC-IV deidentification anchors by combining `anchor_age`, `anchor_year`, and the year of `admittime`. This yielded an age variable consistent with the shifted but internally coherent patient timeline.

Baseline history. Baseline history variables were retrieved from `diagnoses_icd` and linked back to `admissions` in order to recover the timing of the coded encounter. Four flags were created: `smoking_flag`, `hypertension_flag`, `diabetes_flag`, and `tee_flag` for thromboembolic history. These variables were intended to capture background conditions before or at the indexed admission, so diagnoses were only counted when the associated diagnosis admission time was on or before the indexed admission time.

Treatment variables. Treatment-related retrieval combined hospital procedure coding and medication administration. Nimodipine exposure was identified from `emar` through medication-name matching. Aneurysm-directed procedure variables were derived from `procedures_icd` and encoded as `clipping_flag`, `coiling_flag`, and `open_ligation_flag`. These variables were kept separate because they describe different treatment pathways.

Imaging-related coded findings. Three coded imaging-related complication variables were retrieved from `diagnoses_icd`: `ivh_flag`, `iph_flag`, and `hydrocephalus_flag`. These were used as structured proxies for bleed phenotype and secondary injury burden. As with baseline history, the coded diagnoses were interpreted in relation to admission timing.

Additional coded disturbances. A broader set of diagnosis-derived complication flags was then attached from `diagnoses_icd`. These included leukocytosis, erythrocytosis, cerebral salt wasting proxy, polyuria, hyponatremia, hypernatremia, epilepsy, malnutrition, hypovolemia, hypoxia, hypotension, hypercapnia, hypomagnesemia, and hyperglycemia. In addition to these individual variables, the pipeline computed `additional_diagnosis_burden` as the simple count of positive coded disturbance indicators per admission.

Neurological severity. Glasgow Coma Scale variables were derived from `chartevents`. The pipeline first extracted eye, verbal, and motor component item IDs, then pivoted repeated charted values into time-aligned GCS triples. Two representations were retained. The first was worst pre-treatment GCS. The second was admission GCS, based on the earliest available component values for the admission. The pre-treatment restriction was used to limit contamination by sedation, intubation, surgery, or later deterioration after aneurysm treatment.

ICU exposure. ICU structure was derived from `icustays`. Instead of immediately reducing ICU exposure to one scalar, the pipeline retained `icu_los_vector`, the full per-admission vector of ICU stay durations in hours. This allowed multiple ICU episodes, transfers, and returns to intensive care to remain visible.

Treatment timing. Treatment timing was derived by linking `admissions` and `procedures_icd`. The earliest aneurysm-securing procedure date was stored as `first_treatment_date`, then compared with `admittime` to calculate `treatment_delay_days`. A categorical version, `treatment_delay_cat`, was also constructed to distinguish same-day treatment, short delay, moderate delay, longer delay, and admissions without qualifying recorded treatment. Negative or unusually long delays were retained for review because these may reflect date granularity, transfer effects, or documentation artefacts.

Antifibrinolytic therapy. Antifibrinolytic exposure was derived from both prescribing and administration pathways. The pipeline created `afd_ordered_flag`, `afd_admin_flag`, and `afd_any_flag`. Detection combined medication-name matching and product-code matching so that synonymous naming conventions would be less likely to be missed. Ordered and administered therapy were kept separate to preserve the distinction between treatment intent and documented delivery.

External ventricular drainage. The variable `evd_flag` combined two structured sources: EVD-related procedure codes from `procedures_icd` and ICU bedside procedure evidence from `procedureevents` matched through item labels referring to ventricular drains. Using both sources made the variable less dependent on a single documentation pathway.

ECG-derived variables. A limited cardiac monitoring block was derived from `chartevents`. Heart rate summaries were represented through the mean, maximum and minimum heart rate. Rhythm information was derived from rhythm text values and mapped into a coarse categorical variable, `rhythm_major_class`, together with `any_arrhythmia_flag` and `any_major_arrhythmia_flag`.

Notes linkage. Although the current section focuses on structured retrieval, the core pipeline also attached `discharge_note_id`, `discharge_text`, `radiology_note_ids`, and `radiology_text`. This step provided the linkage needed for later multimodal extraction. It also made it possible to compare structured variables against narrative evidence without rebuilding the linkage.

Death-related outcomes. Mortality variables were derived by combining `admissions`, `patients`, and `icustays`. The pipeline produced flags for any documented death, in-hospital death, ICU-window death, death within the hospitalisation window, death after discharge, and early death before the typical vasospasm window, alongside provenance variables for death source and time resolution. These distinctions were retained because different mortality timings correspond to different clinical and analytical interpretations.

4.1.3 Compound physiologic feature blocks

The second stage of the structured retrieval workflow focused on variables that are harder to capture through a single source or a single observation. In ICU data, many clinically relevant processes show up across repeated laboratory values, bedside monitoring, coded diagnoses, and sometimes treatment records.

These variables were constructed using feature blocks. Each block followed a similar idea: relevant raw signals were extracted, repeated values were summarised, and these summaries were then used to derive a state variable. In most cases, this included both the magnitude of abnormal values and how often they occurred, so that isolated measurements and more persistent patterns were not treated the same.

Overview of derived state variables. Table 4.1 lists the final state variables together with their encoded values.

Table 4.1: Derived physiological state variables and encoded categories.

Feature block	Derived variable	Encoded categories
Dysnatremia	dysnatremia_state	normal, hyponatremia, hypernatremia, both
Hypotension	hypotension	none, possible, likely, severe
Oxygenation	hypoxia_state	none, possible, likely, severe
Leukocytosis	leukocytosis	none, possible, likely, severe
Hyperglycemia	hyperglycemia_state	none, moderate, severe
Hypovolemia	hypovolemia	none, possible, present
Hypomagnesemia	hypomagnesemia	none, borderline_or_treated, mild, severe

Table 4.1 gives the encoded state variables used in the downstream dataset. The exact rule-based category definitions, including the priority order between severity levels and the resulting category counts in the final 470-admission matrix, are provided in Appendix C.

Dysnatremia. Dysnatremia combined sodium laboratory measurements, sodium-related charted measurements, and ICD-derived hyponatremia and hypernatremia flags. The laboratory component looks at the number of measurements and the observed range, while threshold indicators captured low, very low, and high sodium values. This allows hypo- and hypernatremic patterns to be represented separately, including cases where both occur within the same admission.

Hypotension. Hypotension combined repeated ICU blood pressure measurements, outpatient or baseline blood pressure records from `omr`, and the ICD-derived hypotension flag. The block summarised mean blood pressure (MAP), peak blood pressure (SBP), and rest blood pressure (DBP), together with counts of values below clinically used thresholds. This provides a simple way to reflect both overall blood pressure levels and the presence of low-pressure episodes.

Oxygenation. Oxygenation combined PaO_2 laboratory values with ICU charted SpO_2 , SaO_2 , FiO_2 and PbtO_2 measurements. For each source, summary statistics and threshold counts were included. Bringing these measurements together gives a more complete picture of oxygenation than any single source alone.

Leukocytosis. Leukocytosis combined repeated white blood cell measurements with the ICD-derived leukocytosis flag. The block used measurement counts, summary statistics, and threshold indicators for elevated WBC values to represent inflammatory activity over the admission.

Hyperglycemia. Hyperglycemia combined laboratory glucose values, ICU charted glucose values, and the ICD-derived hyperglycemia flag. Measurement counts, summary statistics, and

threshold indicators for elevated glucose were included across both sources, allowing the final state to reflect both presence and degree.

Hypovolemia. Hypovolemia was treated as a lower-resolution block, since direct measurement of intravascular volume status is not available in the structured tables. The implementation therefore relied on ICD-coded hypovolemia and polyuria, keeping supporting codes and titles. Direct hypovolemia codes were given more weight, while polyuria (excessive urine output that may indicate fluid loss) was used as a weaker proxy.

Hypomagnesemia. Hypomagnesemia combined serum magnesium measurements, the ICD-derived hypomagnesemia flag, and magnesium sulfate administration from ICU `inputevent` s. The laboratory component contains measurement counts, central tendency measures, and threshold indicators for lower magnesium levels. Magnesium sulfate administration was included as additional context, so that both observed deficiencies and treatment responses are reflected.

Across the feature blocks, the final representations were based on abnormal measurements, their severity, how often they occurred, and, where available, supporting diagnoses or treatments. The aim was to summarise repeated observations into a manageable set of state variables, while keeping the underlying signals available for inspection.

To make the resulting representation more concrete, Figure 4.2 shows one example admission from the final structured dataset.

4.1.4 Threshold definitions for physiologic feature blocks

Continuous measurements were converted into clinically interpretable features using predefined thresholds. These thresholds defined abnormal values and supported repeated-abnormality counts and graded state construction.

Variable	Thresholds	Sources
Hypercapnia	pCO ₂ : > 45, > 50 mmHg	[78]
Dysnatremia	Na ⁺ : < 135, < 130; > 150 mmol/L	[16]
Hypotension	MAP < 65, < 60; SBP < 90; DBP < 60	[73, 108, 111]
Oxygenation	PaO ₂ : < 80, < 60, < 55 SpO ₂ : < 94, < 92, < 90, < 88% SaO ₂ : < 90% FiO ₂ : ≥ 40, ≥ 60% PbtO ₂ : < 15, < 10	[34, 73, 85]
Leukocytosis	WBC: ≥ 10, ≥ 12, ≥ 15	[50, 65, 93]
Hyperglycemia	Glucose: > 140, > 180 mg/dL	[16, 32]
Hypomagnesemia	Mg: < 1.7, < 1.4, < 1.2 mg/dL	[38]

Table 4.2: Thresholds used to derive physiologic feature states.

4.1.5 Aggregation logic, temporal constraints, and missingness handling

Across feature blocks, the retrieval process followed a small number of recurring patterns. Binary indicators were usually constructed through grouped aggregations, while repeated measurements were summarised using counts, extrema, medians, observation times, and threshold-crossing counts. This made it possible to represent not only whether an abnormality occurred, but also how often it was measured and how severe it appeared.

likely_aSAH · score 5.0

subject_id	18903459
hadm_id	21077260
aSAH likelihood score	5.0
final label	likely_aSAH
support source	both
cohort source	icd_asah_1lm
SAH ICD code	I6031

additional diagnosis burden	2
leukocytosis	0
hyponatremia	1
hypernatremia	0
hypoxia	0
hypotension	0
hypercapnia	0
hypomagnesemia	0
hyperglycemia	0

sex	F
age	43
smoking	0
hypertension	0
diabetes	0
thromboembolic history	0

hypotension	severe
hypoxia	possible
dysnatremia	hyponatremia
leukocytosis	severe
hyperglycemia	moderate
hypovolemia	none
hypomagnesemia	mild

vasospasm flag	1
vasospasm ICD code	167848
nimodipine	1
clipping	1
coiling	0
open ligation	0
external ventricular drain	1

heart rate mean	100.1
heart rate min	62.0
heart rate max	129.0
major rhythm class	sinus
any arrhythmia	0
any major arrhythmia	0
antifibrinolytic administered	0
antifibrinolytic ordered	0
any antifibrinolytic exposure	0

intraventricular haemorrhage	1
intraparenchymal haemorrhage	0
hydrocephalus	1

worst pre-treatment GCS eye	3.0
worst pre-treatment GCS verbal	1.0
worst pre-treatment GCS motor	6.0
worst pre-treatment GCS total	10.0
admission GCS total	10.0
ICU length of stay	587 h
first treatment date	2182-06-11
treatment delay	20 days
treatment delay category	6+ days

any documented death	0
in-hospital death	0
ICU-window death	0
death after discharge	0
early death before vasospasm window	0

Figure 4.2: Illustrative admission-level representation from the final structured dataset.

Temporal restrictions were applied when the interpretation of a variable depended on timing. Baseline-history variables were restricted to diagnoses recorded before or at the indexed admission. Worst pre-treatment GCS was limited to values recorded before the first aneurysm-securing procedure, reducing the chance that sedation, surgery, or later deterioration would be encoded as baseline neurological severity.

Missingness was handled according to the meaning of the variable. For event-based flags, such as diagnosis, procedure, or medication indicators, the absence of a matching source record was treated as no documented event. For continuous measurements, missing values were generally left missing, because an unmeasured laboratory or charted value is not the same as a normal result. Derived categorical states were assigned explicit default categories only after the relevant source signals had been combined.

4.1.6 Validation, quality control, and provenance

Validation was integrated into the retrieval workflow and applied after each major step. Checks focused on maintaining admission-level consistency, inspecting distributions of derived variables, and comparing related fields within the same feature block.

These steps were used to flag merge issues, implausible values, and inconsistencies before variables were carried forward. The detailed checks, including feature-specific examples, are described in Appendix D.

The structured retrieval pipeline retained source-level provenance for all main variables and physiologic feature blocks. To keep the methodology chapter focused on the construction logic, the full provenance mapping is provided in Appendix E.

4.1.7 Role of the structured retrieval pipeline in the overall methodology

By the end of this stage, the primary and secondary structured pipelines had turned the final aSAH cohort into a dataset covering provenance, diagnoses, procedures, medications, neurological severity, ICU course, physiological measurements, derived state variables, linked notes, and death-related outcomes.

The pipeline keeps track of where variables came from and how they were constructed. Each feature remained tied to its source, with its own aggregation logic and intermediate summaries, so that it could still be inspected or traced back if needed.

This became important in the later stages. Comparing structured data with notes only makes sense if the underlying signals are still interpretable. Validation also depends on being able to look beyond a single encoded value and see the supporting measurements or codes. For the causal analysis, it helped that the variables reflected something clinically meaningful and followed a consistent temporal structure.

4.2 LLM-based information extraction from clinical text

The causal discovery methods used in this thesis require a tabular representation: each admission must be described by a fixed set of interpretable variables that can serve as graph nodes. The structured variables described in the previous section provide part of this representation, but they are mostly based on coded diagnoses, procedures, medication records, laboratory values, and ICU measurements. These sources are useful, but they do not capture all clinically relevant information needed to study vasospasm after aSAH.

Several important concepts are recorded mainly in narrative clinical documentation. Vasospasm interpretation, aneurysm characteristics, clot burden, hydrocephalus, EVD evidence, and treatment context often appear in discharge summaries or radiology reports as part of clinical interpretation. In MIMIC-IV, these concepts are usually not available as separate, consistently coded database fields.

LLMs were therefore used in two distinct parts of the pipeline. First, they were used for cohort adjudication. Candidate admissions retrieved from structured data were assessed to determine whether the admission was likely to represent aSAH. This step helped define the final study cohort and is separate from the later variables used to describe or explain vasospasm. Second, LLMs were used for feature extraction within the final cohort. In this stage, discharge summaries and radiology reports were converted into structured admission-level variables describing vasospasm evidence, hemorrhage severity, aneurysm characteristics, hydrocephalus, EVD presence, and treatment-related information.

These two uses have different methodological roles. The aSAH adjudication task controls which admissions enter the study, while the later extraction tasks create clinical variables for description, cross-modal validation, and causal discovery. Keeping these roles separate reduces the risk of confusing cohort definition with downstream explanatory features.

The extraction was implemented as prompt-based structured output instead of text embeddings. Embeddings may be useful for similarity search or prediction, but they are harder to interpret as causal discovery nodes and harder to trace back to specific clinical statements. For this thesis, the downstream analyses required explicit admission-level variables with defined categories, uncertainty states, and supporting evidence from the source text.

The extraction pipeline was divided into separate tasks. Each task focused on a specific type of clinical inference and produced a predefined group of output variables. This separation was used because different types of information require different levels of interpretation. Some variables depend on judgement across several sources, while others depend on careful extraction of explicitly stated facts. Keeping the tasks separate made the outputs easier to inspect and allowed uncertainty to remain visible. The full prompt templates are provided in Appendix F.

4.2.1 Cohort-defining task: aSAH adjudication

The first LLM task addressed the cohort-definition problem: deciding whether a candidate admission was likely to represent aneurysmal subarachnoid hemorrhage. Candidate admissions were first retrieved from structured data, as described in Section 3.3. This candidate-generation step was intentionally broad, so that admissions plausibly related to SAH or aneurysm care were available for review before the final likely-aSAH cohort was fixed. However, structured data alone could not reliably distinguish aneurysmal from non-aneurysmal SAH.

The adjudication task used both discharge summaries and radiology reports. The model combined evidence across these sources and produced an ordinal `asah_likelihood_score` in the range $[-5, +5]$, together with a categorical label: `likely_aSAH`, `possible_aSAH`, `indeterminate`, or `likely_not_aSAH`. Additional structured sub-signals were also extracted, including aneurysm presence, rupture evidence, treatment indicators, and explicit negative evidence such as traumatic or non-aneurysmal SAH.

This task was used to define the final likely-aSAH cohort and was therefore treated differently from the later feature-extraction tasks. The aSAH label determined whether an admission entered the study, while the later LLM-derived variables described clinical properties within that cohort.

The task was designed to favour recall, but with explicit limits. Positive classification could be supported by direct statements such as "ruptured aneurysm", treatment trajectories such as clipping or coiling, and consistent radiological evidence. At the same time, weak proxies such as vasospasm, hydrocephalus, or prophylactic treatment were not allowed to determine aneurysmal origin by themselves. When sources disagreed, the output reflected the combined evidence, with supporting quotes retained for later review.

4.2.2 Feature-generating tasks

After the final likely-aSAH cohort had been defined, the remaining LLM tasks were used to extract clinical variables from the available notes. These variables were not used to decide whether an admission belonged to the cohort. Instead, they described aspects of the admission that were relevant for cohort description, multimodal validation, and causal discovery.

Discharge-based structured extraction: narrative synthesis

The second task extracts structured variables from discharge summaries. These summaries are treated as narrative accounts of the admission, combining diagnostic reasoning, treatment decisions, complications, and clinical interpretation over time.

The extracted variables focus on constructs that can benefit from this broader narrative view. These include:

- **severity indicators:** WFNS, Hunt-Hess, and Fisher grades, and clot burden proxies;
- **aneurysm characteristics:** location and treatment type;
- **vasospasm evaluation and management strategies:** imaging modalities and therapeutic interventions;
- **baseline history variables:** smoking, hypertension, and diabetes;
- **hydrocephalus and EVD context;**
- **neuroimaging intensity** as described in narrative form.

Variables are linked to coarse time windows, such as admission, early phase, or vasospasm window. Where available, textual time anchors are also preserved.

The extraction is cautious. Values are assigned only when supported by the note, and ambiguous or conflicting statements result in outputs such as **unknown** or **unclear**. Short evidence quotes and brief boundary justifications are extracted as well so that the structured variables can be traced back to the source text.

Radiology-based structured extraction: imaging evidence

The third task focuses on radiology reports, which are treated as a separate source type from discharge summaries. Discharge summaries provide clinical synthesis, while radiology reports provide time-specific imaging observations.

Extraction from this source therefore focuses on explicit findings and aggregation across time. The variable groups include:

- **neuroimaging intensity**, defined as counts and categories of imaging modalities such as CT, CTA, CTP, MRA/MRV, and angiography;
- **vasospasm evidence and severity**, based on radiological language;
- **bleed phenotype**, including SAH, IVH, IPH, SDH, and clot burden proxies;
- **hydrocephalus and ventricular status;**
- **aneurysm presence** and its location;
- **supportive signals** such as treatment mentions and mechanical ventilation indicators.

This task also has to handle multiple reports within one admission. Radiology data are longitudinal, and admissions often contain several imaging studies performed at different moments. To deal with this, concatenated reports are first segmented, and then aggregated across the admission. When findings change over time, later reports usually receive more weight, while earlier contradictory evidence is retained as context.

Guardrails are used to reduce over-interpretation. For example, chronic vascular stenosis is not treated as vasospasm unless the report explicitly describes it as such. This keeps the extracted variables closer to observed findings than inferred diagnoses.

Vasospasm scoring: targeted adjudication

The fourth task provides a focused assessment of vasospasm based on discharge summaries. Vasospasm-related information is also captured in the broader extraction tasks, but this separate scoring step produces one ordinal variable summarising the overall evidence for vasospasm.

The output is a `vasospasm_score` in the range $[-5, +5]$, following the same ordinal logic as the aSAH adjudication task. Positive scores reflect increasing evidence for vasospasm, while negative scores reflect evidence against it. The scoring uses explicit mentions, imaging findings, monitoring data, and vasospasm-directed treatments. Prophylactic therapy, such as nimodipine alone, is not treated as sufficient evidence. As in the aSAH task, each score is accompanied by evidence quotes and boundary justifications.

4.2.3 Structured output design

All extraction tasks use predefined JSON schemas. These schemas specify the required fields, data types, allowed values, and relationships between fields. This makes the outputs machine-readable and more consistent across admissions.

The schemas contain several types of variables:

- **Discrete labels**, such as binary indicators for hydrocephalus or aneurysm presence;
- **Ordinal scores**, such as evidence scores ranging from -5 to $+5$;
- **Categorical fields**, such as aneurysm location groups or treatment types;
- **Structured sub-objects**, which capture intermediate signals and boundary justifications;
- **Evidence quotes**, consisting of verbatim text spans that support the extracted values.

4.2.4 Uncertainty representation

Clinical documentation often contains incomplete, ambiguous, or conflicting information. The extraction framework therefore keeps uncertainty visible instead of forcing all cases into definitive categories.

For some clinical phenomena, uncertainty is encoded through the ordinal evidence scales. For example, vasospasm evidence is represented on a scale from -5 to $+5$, which allows weaker, stronger, positive, and negative evidence to be distinguished. Categorical variables also include explicit uncertainty states such as `possible`, `indeterminate`, or `unknown`. This helps avoid treating missing information as evidence against a condition. Uncertainty is therefore preserved during extraction, so later analyses can decide how to handle it depending on the modelling question.

4.2.5 Few-shot example construction

The prompt-based extraction tasks described above rely on structured output constraints and carefully written instructions. However, clinically complex inference tasks often involve subtle distinctions between evidence types, for example the difference between a score of +3 and a score of +2. Prompt instructions alone may therefore not be enough to produce consistent model behaviour. To address this, each extraction task was supplemented with a curated set of few-shot examples that demonstrate how raw clinical text should be mapped to structured output. The examples were selected and refined by the author, based on the clinical definitions, literature review, and repeated inspection of the source notes.

Few-shot examples serve two related purposes. First, they anchor the model’s behaviour in concrete cases, reducing dependence on abstract instructions alone. Second, few-shot examples provide guidance on edge cases, including conflicting signals, incomplete documentation, and borderline interpretations. Encoding these cases in the prompt context helps produce more calibrated outputs and reduces recurring errors, such as over-reliance on single keywords or overly confident classification in ambiguous cases.

The construction of these few-shot examples followed the following process: candidate mining from the full note corpus, heuristic scoring and filtering, diversity-oriented selection, author review and refinement, and transformation into reusable structured formats. This pipeline was designed so that examples reflected real clinical variation while still matching the intended extraction logic.

Candidate note generation

Few-shot construction begins with large-scale candidate generation from the full corpus of discharge summaries and radiology reports. Notes are not sampled randomly. Instead, candidates are mined using heuristic pattern matching aimed at finding clinically informative documents.

For discharge summaries, candidate selection uses a weighted scoring system that detects clinically relevant concepts such as severity grades, aneurysm location, treatment mentions, vasospasm-related terminology, hydrocephalus, and EVD placement. Each note is assigned a composite score based on the presence of these features, combined with length-based and domain-coverage adjustments.

Radiology candidate selection follows a similar idea but is adapted to imaging-specific signals. Notes are categorized based on patterns such as explicit vasospasm findings, absence of vasospasm, ambiguous language, bleed phenotype descriptions, and EVD-related findings. Additional heuristics are used to filter out non-neuro imaging and to prioritise reports with rich modality coverage, such as CTA, angiography, and CT perfusion.

For the vasospasm scoring task, candidate notes are generated using explicit SQL-based pattern queries that map textual patterns to predefined score buckets. This produces a stratified candidate pool spanning the full ordinal range from strong negative to strong positive evidence.

Across all tasks, the aim of candidate generation is recall, not precision. The goal is to surface a broad and varied set of potentially informative notes that can later be refined through manual selection.

Heuristic scoring and diversity-oriented selection

After candidate generation, heuristic scoring was used to rank notes by informational richness. Uniform random sampling was not used as the main selection strategy because many randomly selected notes would contain little or no relevant evidence for the specific extraction targets. This would make the few-shot context less useful, especially for rare or borderline cases. At the same time, selecting only the highest-scoring notes would also be problematic, because it could produce a few-shot set dominated by prototypical cases.

To balance these issues, a diversity-oriented selection strategy was used. Candidate notes were grouped using the same clinical themes that the extraction tasks were designed to capture, so that the few-shot set reflected the range of information the model was expected to recognise. The buckets therefore mirrored the extraction targets, including severity grading, aneurysm treatment, vasospasm evaluation and management, hydrocephalus and EVD context, and aneurysm location.

Selection was then guided by coverage across these dimensions, not only by heuristic score. This meant that the final few-shot set included both more straightforward cases and notes where the signal was less clear or partially conflicting. The aim was to keep the examples aligned with the extraction schema while still capturing variation in how these variables appear in clinical text, without repeatedly selecting near-identical cases.

Manual selection and refinement

The final selection of few-shot examples was performed manually from the candidate pool. This step was needed because automated heuristics could not fully capture the subtleties required for useful few-shot examples.

During manual selection, each candidate note was evaluated along several criteria:

- clarity and completeness of the clinical signal;
- presence of interpretable evidence supporting the target variables;
- representativeness of a specific clinical scenario;
- absence of unrelated procedures or excessive noise.

After selection, the examples were converted into the same structured format expected from the extraction model. The few-shot examples did not only show the model which text fragments were relevant; they also demonstrated how those fragments should be translated into the output schema. For each example, the note evidence was mapped to explicit variable values, uncertainty labels, and short supporting justifications. Ambiguous cases were not forced into overly certain categories, but were annotated in a way that matched the intended handling of uncertainty in the final pipeline.

The annotations were created and refined by the author based on the extraction definitions, the literature review, and repeated inspection of the notes. They should therefore be understood as schema-aligned examples for improving consistency, not as an independently validated expert gold standard.

Transformation into reusable few-shot corpora

Once selected and annotated, examples are transformed into reusable structured formats. Each example is stored as a JSON object containing:

- identifiers linking back to the original admission and note;
- the raw note text;
- the structured target output, stored as `gold_json`.

These examples are then aggregated into JSONL files, which are used directly during prompt construction for model inference. The conversion process is automated through dedicated scripts that retrieve the corresponding note text, attach the annotated outputs, and serialize the result into a consistent format.

4.2.6 Implementation safeguards

The extraction tasks were implemented through a shared execution framework that handled prompt assembly, model calls, JSON parsing, schema validation, output flattening, checkpointing, and logging. These implementation details were kept separate from the clinical task definitions because they mainly describe runtime control rather than the conceptual extraction design.

Several safeguards were used to make the extraction outputs easier to audit. Model responses had to pass task-specific JSON validation before they were accepted. Invalid or malformed outputs were either repaired through a second-pass prompt or marked with an explicit failure status. Each processed admission also received a processing status such as `ok`, `fail`, `no_text`, or `skip_too_long`. The raw JSON outputs, flattened variables, runtime metadata, and error logs were retained so that extraction behaviour could be inspected after processing.

The full execution framework, including runtime settings, retry behaviour, repair prompting, token-budget handling, and resumable checkpointing, is described in Appendix G.

4.2.7 Construction of the combined dataset

After the structured retrieval pipeline and the note-based extraction pipeline had been completed separately, their outputs were integrated into a single analysis table.

The integration started from the structured base cohort produced by the secondary structured-feature pipeline. This base table was then merged sequentially with three note-derived sources: the discharge-extraction output, the discharge-based vasospasm-scoring output, and the radiology-extraction output. Because the discharge-related outputs were linked to specific discharge notes, these merges were performed on `(subject_id, hadm_id, discharge_note_id)`. The radiology-derived output had already been aggregated to admission level and was therefore merged on `(subject_id, hadm_id)`.

A further processing step was needed because the LLM outputs were stored partly as nested JSON objects, not only as flat columns. These JSON-like fields were parsed and systematically unpacked into explicit tabular columns. Nested keys were flattened into stable variable names, after which redundant aliases and exact duplicate columns were identified and removed.

After merging and flattening, a final naming and cleaning stage standardised the integrated schema. Variable names were rewritten into a consistent prefix-based format, technical runtime artefacts from the extraction process were removed, repeated list-like objects were converted into stable string representations where necessary, and selected columns were cast into more appropriate types such as categorical, nullable Boolean, or integer formats.

Time-window and time-anchor helper fields generated during extraction were removed at this stage. Due to the de-identification procedures applied in MIMIC, temporal information is shifted and inconsistently recorded across data sources. Many clinically relevant signals lack precise timestamps or contain only coarse references such as "on admission" or "during hospitalization". As a result, these helper fields did not provide reliable or consistently interpretable temporal structure beyond what was already encoded in the admission-level variables, and were therefore excluded from the final analysis dataset.

4.3 Data Harmonization and Quality Control

The dataset combines several types of clinical evidence. These sources do not capture information in the same way. They differ in level of detail, timing, and clinical perspective, and they do not always agree with each other. This leads to gaps and occasional contradictions in how the same clinical process is described.

Vasospasm illustrates this clearly. Evidence for it may appear in ICD codes, discharge summaries, and radiology reports, each reflecting a different part of the clinical workflow. ICD

codes are assigned after the fact and mainly follow billing conventions. Discharge summaries provide a narrative overview written at the end of the admission. Radiology reports are tied to imaging studies and focus on interpretation at specific time points. These differences mean that the same admission can be described in multiple, sometimes inconsistent ways.

Because of this, an additional harmonization step was introduced to bring these representations into a common format, make disagreements explicit, and prepare the data in a way that remains interpretable for later analysis.

4.3.1 Making extracted values consistent

After extraction, additional steps were needed to ensure that outputs could be compared across admissions and sources.

Some variables required normalization at the level of meaning. Treatment mentions were mapped to consistent labels so that different spellings or naming conventions refer to the same concept. Blood pressure targets described in text were translated into structured categories or numeric ranges, making them easier to compare across admissions.

Fields containing multiple values were also standardized. Extracted outputs sometimes appeared as JSON-like strings, delimiter-separated lists, or single entries. These were parsed into a consistent format, duplicates were removed, and a uniform representation was enforced.

4.3.2 Manual review of disagreement cases

A binary version of the vasospasm variables was created using a threshold, where positive values indicate evidence for vasospasm. In addition, a coarser categorical version grouped the scores into ordered levels: strong negative (-5 to -3), weak negative (-2 or -1), uncertain (0), weak positive (+1 or +2), and strong positive (+3 to +5). These representations allow comparisons at different levels of detail, depending on the analysis.

The same mapping was applied across sources. No adjustments were made to favour one source over another, so that differences between them remain visible in the data.

Since the dataset brings together multiple sources, disagreement between them was treated as something to examine directly. It was defined using the normalized ordinal scores.

In addition to disagreement between discharge summaries and radiology reports, ICD-based vasospasm coding was also compared with the free-text-derived vasospasm signals.

Three types of disagreement were considered. Strong disagreements refer to cases where one source indicates clear presence and another indicates clear absence. Weak-opposite disagreements describe situations where weaker evidence points in opposite directions. A third group includes cases where both sources provide only limited evidence, which may reflect uncertainty or lack of information.

For a subset of admissions with opposing signals, manual review was carried out to better understand how different sources arrived at their respective interpretations.

To support this, the underlying text was processed with a different objective than in the extraction stage. Instead of preparing text for structured variables, the focus here was on readability during inspection. Free-text fields in discharge summaries and radiology reports include a mix of clinical content, formatting elements, repeated sections, and administrative information, which can make direct review difficult.

Radiology reports were first split into individual studies using report delimiters. Each study was then classified based on its content, allowing non-neurological imaging, such as chest scans, to be excluded. Within the retained studies, sections without clinical relevance were removed, including dose information, acquisition details, and administrative headers. The remaining text was cleaned to remove formatting artefacts and standardize spacing, while keeping sections such as findings and impression intact.

A similar approach was applied to discharge summaries. Sections not directly related to the clinical course, such as follow-up instructions or administrative notes, were filtered out. The focus was on parts where clinical interpretation and reasoning are documented.

The outcome of the review was stored as separate annotations linked to the corresponding admissions. These annotations were kept alongside the original representations and were not used to overwrite them.

4.3.3 Distributional and consistency checks

After normalization and disagreement labeling, the dataset was examined through a series of distributional and consistency checks. This step was important because the extracted variables do not have a single external ground truth. For many clinical concepts in this thesis, such as vasospasm evidence, hydrocephalus, EVD presence, clot burden, or treatment context, there is no perfectly reliable reference variable in MIMIC-IV. ICD codes, discharge summaries, radiology reports, procedure records, and medication records each capture only part of the clinical picture.

For that reason, the goal of these checks was not simply to describe the dataset, but to increase confidence that the variables used later in the analysis were internally coherent, traceable to their source, and clinically plausible. Without this step, extraction errors or source-alignment problems could be carried into the causal discovery stage and appear as meaningful graph structure.

Missing values were first summarised for each variable, making it easier to compare how complete different sources were. Agreement between sources was then explored using cross-tabulations of binary and categorical representations. These summaries showed how often sources aligned, without assuming that one source was automatically correct.

Additional checks focused on variables that were expected to be related. For example, severity indicators were compared with their underlying ordinal scores, and procedure variables were checked against evidence found in text. These comparisons helped identify inconsistencies that could point to extraction errors, linkage problems, or clinically implausible combinations.

The outcome of these checks was therefore a stronger basis for using the extracted variables in later analyses. Variables were not treated as proven clinical truth, but as source-aware and internally checked representations that were sufficiently consistent to support descriptive comparisons and exploratory causal discovery.

4.4 Causal discovery implementation

The literature review made the three pillars of causal discovery within this thesis clear. First of all, aSAH and vasospasm involve several interacting clinical processes that are not fully understood. At the same time, EHR data only give a partial view of these processes. The same clinical state may appear differently in coded data, structured measurements, radiology reports, and discharge summaries. Thirdly, causal discovery methods depend strongly on how variables are represented, which assumptions are made by the algorithm, and how the algorithm is configured.

For that reason, causal discovery was only applied after the structured and note-derived signals had been converted into a reconciled admission-level dataset. At that point, source disagreements had been reviewed, overlapping representations had been reduced where possible, and the multimodal table had been normalized into variables that could be used for graph learning.

4.4.1 Role of causal discovery in the study

The causal discovery component was used to study conditional dependence structure among clinically relevant variables. The goal was not to estimate one isolated treatment effect, nor to use

the variables only for prediction. The analysis instead explored how baseline factors, hemorrhage severity, early secondary complications, treatment-related variables, and note-derived vasospasm signals could be causally connected.

Vasospasm after aSAH is unlikely to be represented well by one isolated pathway. Variables in the dataset may be related through initial injury severity, monitoring intensity, treatment decisions, physiological instability, and documentation practices. A graph-based approach provided a way to organize these relations and to examine which structures appeared repeatedly under different discovery settings.

The resulting graphs were interpreted as candidate structures. Their value depended on recurrence across algorithms, robustness across settings, and whether the relations could be interpreted in light of the clinical meaning of the variables.

4.4.2 Variable set construction

Candidate variables and dataset policies

The final merged dataset contained more variables than could reasonably be used in causal discovery. The discovery workflow therefore started from a curated candidate pool of clinically plausible variables.

The candidate pool was defined through a feature registry. This registry recorded each variable’s source scope, representation type, clinical tier, and construction rule. This made the selection process easier to inspect and reproduce, and avoided selecting columns manually from the final table. The full candidate pool is listed in Appendix I.

Several dataset policies were used to examine sensitivity to variable selection. The full admission cohort was used in the main policy variants, which differed for example in whether neuroimaging intensity was included. This variable was considered because it may capture part of the documentation pathway behind the vasospasm score: patients who undergo more neuroimaging may have more opportunities for vasospasm to be described, suspected, or ruled out in the record. In that sense, neuroimaging intensity could help explain variation in the observed vasospasm signal as recorded in the EHR. However, it is less suitable as a variable in the final causal representation of the clinical process itself. The broader goal of the thesis is to move as close as possible to the underlying clinical state, and in that target interpretation neuroimaging intensity should not be treated as a causal ancestor of vasospasm or its upstream biological mechanisms. It was therefore useful as a variable during development, but was removed from the leaner discovery policy used for the main clinical interpretation.

Exclusion and consolidation of variables

Several groups of variables were excluded before discovery. Technical and audit-oriented fields were removed first. This included raw note text, evidence quotes, comments, source row indices, processing metadata, and extraction helper fields. These fields were useful for review and traceability, but they did not represent clinical states that were suitable as graph nodes.

Highly source-specific helper variables were also excluded when their interpretation was too closely tied to the extraction process.

Variables with excessive missingness or very limited variation were removed during dataset construction. Variables with too few observed levels were also dropped, and a minimum number of retained features was required before a dataset was accepted. This screening was included because a variable could be clinically meaningful in principle, but still unsuitable for discovery if it was very sparse, almost constant, or mostly shaped by documentation availability.

The multimodal pipeline often produced several candidate representations for the same clinical construct. The feature registry therefore reduced these overlapping fields to a smaller set of discovery variables, using predefined builder functions instead of selecting columns directly from the merged table. For example, hydrocephalus combined the structured hydrocephalus

flag with radiology and discharge-derived hydrocephalus evidence. Similarly, nimodipine combined the structured medication flag with the discharge-derived nimodipine indicator. In these cases, positive evidence from any source was allowed to define presence, while absence was only assigned when the available sources gave sufficient explicit negative evidence.

Other variables were constructed by choosing between related source-specific representations or by mapping repeated measurements into clinical states. For example, clot burden was derived from combined the radiology and discharge clot-burden proxies. Laboratory and physiological variables like hypoxia, and leukocytosis were carried forward as categorical state variables derived earlier in the structured feature pipeline. The vasospasm outcome was also reconciled explicitly: vasospasm used manual review where available, then structured vasospasm evidence, and then the maximum of the radiology and discharge vasospasm scores since stronger evidence was considered more reliably than weaker evidence after manual inspection.

Encoding for discovery

Missing values were handled using source-scope-aware rules. Admissions without available notes were handled separately from admissions where the source was present but the specific variable could not be extracted. This distinction was important because part of the note-level missingness was caused by the coverage of MIMIC-IV-Note, which does not extend beyond 2020. In that setting, a missing discharge or radiology note was treated as unrelated to the clinical state itself and thus missing completely at random. When the relevant note source was available but a variable remained missing, the missingness was more likely to reflect documentation, extraction uncertainty, or workflow-related absence of information. Those cases were therefore assigned an explicit `MISSING` category, so that potentially informative missingness was retained in the discovery matrix instead of being treated as clinical absence.

The output of this stage was an analysis matrix prepared for the discovery algorithms, with variables selected, reconciled, type-constrained, and encoded.

4.4.3 Constraint design

The discovery algorithms were run with background knowledge constraints. These constraints were used to prevent directions that would conflict with the meaning of the variables, while still allowing the algorithms to learn structure from the data. The constraints were not used to force expected clinical edges into the graph. They only restricted directions that would be difficult to defend given the temporal or definitional meaning of the variables.

Tier-based temporal constraints

The first constraint layer used the tier structure from the feature registry. Variables were grouped into ordered tiers: **baseline**, **severity**, **workflow**, and **outcome**. In the final implementation, early secondary complications and admission-course physiological variables were grouped within the **workflow** tier, because the admission-level dataset could not always separate their precise timing from treatment, monitoring, and documentation processes. The variable-to-tier mapping used for this ordering is listed in Appendix I, Table I.1.

Edges from later tiers to earlier tiers were forbidden across all tier pairs. For example, vasospasm should not point to age, sex, or baseline comorbidities. Treatment and admission-course variables should not determine initial clot burden or aneurysm characteristics. EVD placement should not be placed upstream of the ventricular pathology that usually motivates it.

Additional domain-informed restrictions

A second constraint layer added a small number of domain-informed restrictions. These restrictions were kept sparse and were used only when the reverse direction would be difficult to

interpret clinically or when the direction would conflict with how a variable was defined.

For example, incoming edges were forbidden for age and sex. The direction `evd` $\rightarrow$ `hydrocephalus` was forbidden, since EVD placement is usually a response to ventricular enlargement or concern for fluid drainage. The direction `vasospasm_severity` $\rightarrow$ `vasospasm` was also excluded, since the severity variable refines the presence and strength of vasospasm evidence. Aneurysm-related variables were protected from incoming edges from non-aneurysm severity variables, reflecting their interpretation as upstream anatomical information. When present, `securing_technique` was not allowed to cause selected admission-course physiological or complication variables, since it was treated as part of the aneurysm-care pathway following earlier clinical assessment. This restriction is listed explicitly in Appendix H, because these variables were grouped within the same broad `workflow` tier in the final implementation.

The full set of additional non-tier restrictions is listed in Appendix H.

4.4.4 Algorithms used

The implementation used one representative from each of the causal discovery families discussed in the literature review: PC-stable as a constraint-based method, FGES as a score-based method, and FCIT as a hybrid/PAG-based method. This way, structures across methods with different assumptions and search behaviour could be compared.

All algorithms were run through Tetrad via `pytetrad`. The same discovery-ready datasets and background knowledge constraints were used across methods where applicable.

Data representation and implications

All variables were passed to the algorithms as discrete categories. This step was needed because the conditional independence testing used later relied on the likelihood-ratio (G^2) test, which was applied to discrete contingency tables. Continuous measurements were therefore not passed to the discovery algorithms in their raw form. The score-based parts of the workflow used the Bayesian Dirichlet equivalent uniform (BDeu) score, since it is designed for discrete Bayesian network structure learning and provides a natural fit for categorical variables. It also allowed the FGES and FCIT score components to evaluate candidate graph structures using the same type of data representation as the G^2 -based independence tests.

PC-stable

PC-stable was used as the constraint-based method. It was configured with the G^2 independence test for discrete variables, with sweeps over the significance threshold, conditioning depth, and sparse-cell handling.

The stable adjacency-search option was enabled because column order had no clinical meaning in the dataset. Bidirected edges were not allowed in the PC-stable output, which kept the representation closer to the intended partially directed structure for these runs.

FGES

FGES was used as the score-based method. It was configured with the BDeu score for discrete variables. The sweep varied settings related to graph complexity and scoring behaviour, including maximum degree, BDeu priors, the symmetric first step option, and the faithfulness setting.

FGES provided a complementary view to PC-stable because it searched through graph structures using a score, while PC-stable built structure through sequential conditional independence testing.

FCIT and PAG output

FCIT was used as the hybrid/PAG-based method. It combines an initial score-based CPDAG search with conditional independence testing and returns a PAG-style output, which is useful when latent confounding cannot be ruled out.

In this implementation, FCIT used the G^2 independence test together with the BDeu score. The initial CPDAG search was left at the Tetrad default, corresponding to BOSS in the Tetrad version used. The sweep varied the independence threshold, conditioning depth, sparse-cell handling, and path-related parameters used in the PAG-oriented search.

Cross-algorithm comparison

The three algorithms were used to examine which structures appeared under different discovery paradigms. Edges recurring across methods were treated as more stable patterns. Edges appearing only under one method were interpreted more cautiously, especially when they were also sensitive to hyperparameter settings or bootstrap resampling.

4.4.5 Hyperparameter sweeps and sensitivity analysis

The causal discovery analysis was run across explicit hyperparameter sweeps. These sweeps were not used in the same way as hyperparameter tuning in supervised prediction. There is no held-out ground-truth causal graph in this setting, and therefore no validation metric that can directly identify the "best" causal discovery configuration. Selecting one setting based on whether it produced an expected clinical edge would also risk circular interpretation.

The hyperparameter sweeps were therefore used as a sensitivity analysis. The goal was to examine how much the recovered structures depended on modelling choices such as conditional independence thresholds, conditioning depth, sparse-cell handling, graph-complexity limits, and score priors. Edges that appeared only under a narrow set of permissive or restrictive settings were interpreted more cautiously than edges that recurred across broader parts of the sweep.

For each algorithm, a broad sweep was used to inspect graph density, computational feasibility, and the general stability of recurring variable pairs. Settings that produced degenerate outputs, such as near-empty graphs, overly dense graphs, or unstable behaviour caused by sparse contingency tables, could not be treated as clinically informative..

For PC-stable and FCIT, the main sweep dimensions were the conditional independence threshold, conditioning depth, and sparse-cell handling for the G^2 test. For FGES, the sweep varied graph-complexity and scoring settings, including maximum degree, BDeu priors, the faithfulness setting, and the symmetric-first-step option. The full sweep grids are provided in Appendix J.

The results were interpreted at the level of recurrence across settings rather than as the output of one selected optimal model. This means that the causal discovery findings should be read as stability-informed hypotheses: relations gain credibility when they persist across algorithms, reasonable hyperparameter regimes, and bootstrap resampling, but they are not considered proven causal effects.

4.4.6 Bootstrap and stability workflow

The sweep analysis addressed sensitivity to modelling choices. To also examine sampling variability, selected configurations were supplemented with a nonparametric bootstrap procedure.

For each selected configuration, admissions were resampled with replacement. The bootstrap sample fraction was set to 1.0, so each resampled dataset had the same nominal size as the original dataset. A fixed random seed was used for reproducibility. The same graph search was then rerun on each bootstrap sample.

Because all variables were discrete, each bootstrap sample used the same category definitions as the original dataset. Bootstrap samples in which a variable collapsed to a single observed level were detected and rejected, since such samples could lead to invalid or unstable search behaviour.

For every bootstrap iteration, the discovered edges were stored in edge-list form. These edge lists were then aggregated across iterations to compute adjacency frequencies. For a given unordered pair of variables, the adjacency frequency was defined as the proportion of bootstrap runs in which the pair was connected by any edge.

Orientation was summarized separately from adjacency. For each adjacent pair, the workflow recorded the most frequent endpoint pattern across bootstrap graphs. It also computed the frequency of that pattern overall and conditional on the adjacency being present. This distinction separated pairs that rarely appeared from pairs that appeared frequently but had unstable direction.

Two thresholds were used to label bootstrap stability. An adjacency was labelled stable if it appeared in at least 0.70 of bootstrap runs. Conditional on stable adjacency, an orientation was labelled stable if the dominant endpoint pattern appeared in at least 0.80 of the runs in which the adjacency was present.

These thresholds were used as pragmatic descriptive cutoffs, not as formally optimized decision rules. In this setting there was no ground-truth causal graph or validation metric that could determine an optimal bootstrap threshold. The 0.70 adjacency threshold was chosen to mark edges that appeared in a clear majority of resampled datasets, while still allowing some sampling variability. The 0.80 orientation threshold was set more strictly because edge direction is usually less stable than adjacency in finite observational data.

Importantly, the thresholds were not used to discard all other edges from the analysis. The full bootstrap frequencies were retained and inspected as continuous stability measures. Edges below the stability threshold could still contribute to the interpretation when they recurred across algorithms or hyperparameter regimes, or when they formed part of a clinically coherent structure. The threshold labels were therefore used mainly to separate high-support patterns from more tentative ones in the summary tables and figures.

4.4.7 Output aggregation and summarization

The sweep and bootstrap workflow produced many run-level graphs. Before interpretation, these outputs were standardized into common edge, stability, and run-summary tables so that PC-stable, FGES, and FCIT results could be compared using the same format.

Edge recurrence was summarized separately for unordered variable pairs and oriented edge patterns. Bootstrap summaries were aggregated in the same way, preserving adjacency frequency, dominant orientation, and the assigned stability class. The aggregation retained both the binary stability class and the underlying bootstrap frequencies, so that interpretation did not depend only on whether an edge crossed a single cutoff.

Sensitivity to hyperparameters was examined by grouping runs into regimes, such as shared values of α , conditioning depth, maximum degree, or BDeu prior. Within each regime, edge and pair recurrence were summarized to identify relations that were broadly recurrent and relations that appeared mainly under specific settings.

Because vasospasm was the clinical endpoint of interest, additional summaries focused on edges and short paths involving `vasospasm` and `vasospasm_severity`. Directed and semidirected paths were used as structural summaries of recurring graph patterns. They were not interpreted as estimates of path-specific causal effects. Further details on the aggregation outputs are provided in Appendix K.

Chapter 5

Results

5.1 Cohort construction and label validation

5.1.1 Candidate sets and note availability

The candidate-generation procedure described in Chapter 3 produced two pools for aSAH adjudication. The diagnosis-driven branch contained 1,576 unique admissions with SAH-related ICD coding. The treatment-driven branch added 1,237 candidate admissions without SAH ICD codes, selected because they contained treatment signals that could be compatible with aneurysm-related care.

5.1.2 LLM adjudication of SAH-coded admissions

The resulting label distribution is shown in Figure 5.1. Among ICD-coded admissions, 426 admissions (27.0%) were labeled *likely aSAH*, while 719 admissions (45.6%) were labeled *likely not aSAH*. Only a small number of cases fell into intermediate categories: 12 admissions (0.8%) were labeled *possible aSAH* and 10 admissions (0.6%) were labeled *indeterminate*. 409 admissions (26.0%) could not be scored because no free-text notes were available.

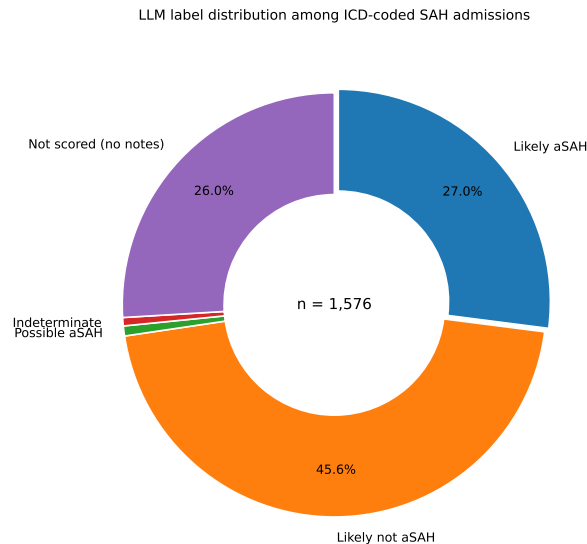


Figure 5.1: LLM label distribution among ICD-coded SAH admissions, including admissions without available clinical notes.

This distribution shows two important characteristics of the candidate pool. First, many SAH-coded admissions did not appear to represent aSAH after closer review of the available

clinical text. Second, a notable fraction of admissions could not be adjudicated because narrative documentation was unavailable. This missingness is consistent with the difference in temporal coverage between the structured MIMIC-IV tables and MIMIC-IV-Note.

The ordinal score distribution gives a more detailed view of how the available evidence was used. As shown in Figure 5.2, scores were concentrated near the extremes. In total, 399 admissions received the maximum score of +5, while 454 received the minimum score of −5. Only 35 admissions received scores between −3 and +3.

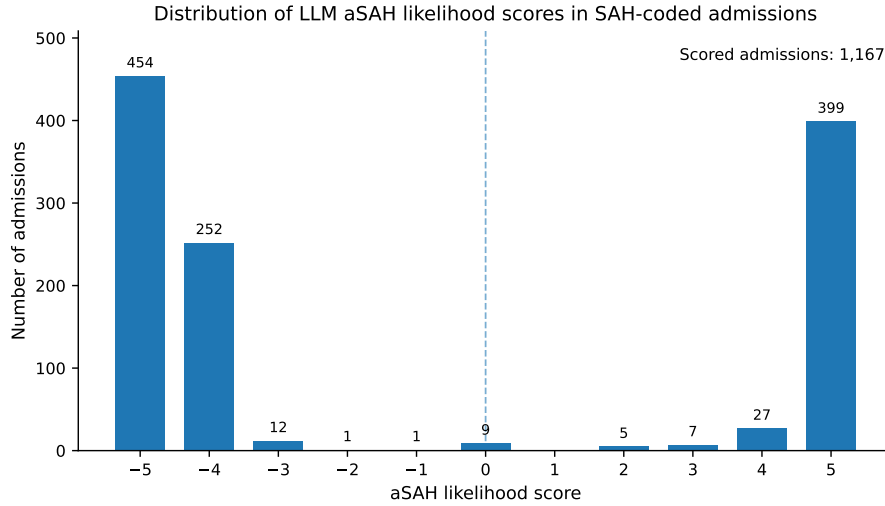


Figure 5.2: Distribution of LLM-assigned aSAH likelihood scores among scored SAH-coded admissions.

The 409 unscored admissions were not treated as clinically uncertain cases. Their missing scores reflected insufficient note coverage after 2020, meaning that the required information for LLM-based adjudication was not available. These admissions were therefore excluded from the adjudicated cohort instead of being assigned uncertainty labels.

After restricting the SAH-coded candidate pool to admissions labeled *likely aSAH*, 426 ICD-supported admissions remained. This formed the first component of the final cohort.

5.1.3 Recovery of admissions without SAH ICD coding

The treatment-driven, non-SAHCoded pathway identified 44 additional admissions classified as *likely aSAH*.

The final unified cohort therefore consisted of 470 unique admissions:

- 426 ICD-supported admissions labeled *likely aSAH*;
- 44 LLM-only admissions labeled *likely aSAH*.

All admissions retained in the final cohort were assigned the label *likely aSAH*.

5.1.4 Manual inspection of adjudication quality

To assess whether the score distribution was clinically plausible, a subset of cases was manually reviewed to check whether the model’s labels were broadly consistent with the evidence present in the records.

Strong positive cases. The strong positive group contained 426 admissions. Most of these cases were relatively straightforward on review. The records typically described a coherent clinical picture: subarachnoid bleeding was documented, an aneurysm was identified as the likely source, and the aneurysm was either treated or explicitly described as ruptured. Common supporting elements included rupture language in the discharge summary, radiology reports identifying a culprit aneurysm, and procedures such as clipping or coiling.

These cases did not appear to be driven by isolated keywords alone. In many admissions, the label seemed to depend on how several pieces of information fit together, including presentation, imaging findings, diagnostic interpretation, and treatment. For example, the presence of an aneurysm mention by itself was not necessarily enough. The relevant question was whether the aneurysm was linked to the hemorrhage in the clinical record.

Some inconsistencies were present, but these often had plausible explanations within the clinical course. An initial CTA could be negative, followed by a later angiogram identifying a small aneurysm. Similarly, an early description of the hemorrhage could be ambiguous, with later imaging or clinical documentation providing a more specific explanation. In most reviewed cases, the assigned label appeared to follow the more definitive later evidence.

Strong negative cases. The strong negative group contained 706 admissions that still had SAH ICD code. This was one of the more informative findings of the manual review.

Several recurring patterns were observed. In many cases, clinicians had investigated a possible aneurysm using CTA, MRA, or catheter angiography and did not identify one, sometimes after multiple studies. In other admissions, another explanation appeared more consistent with the record, such as trauma after a fall, an angiogram-negative SAH pattern, another vascular condition, bleeding after a procedure, hemorrhage after stroke treatment, or bleeding that began within the brain and extended into the subarachnoid space.

A similar issue appeared in the treatment-driven candidate group without SAH ICD coding. Although this pathway was intended to recover missed aSAH cases, many treatment-positive admissions did not represent current aneurysmal SAH. Clipping and coiling often reflected elective treatment of unruptured aneurysms, follow-up treatment of residual aneurysms after a previous rupture, or procedures unrelated to intracranial aneurysms, such as tumour embolization. Nimodipine appeared somewhat more informative, but it was still not specific; some cases involved traumatic or angiogram-negative SAH instead of confirmed aneurysmal rupture.

This helps explain the relatively low yield of the treatment-driven branch. Only a small fraction of clipping- and coiling-positive admissions were ultimately labeled *likely aSAH*, while nimodipine showed a higher, but still limited, proportion. This indicates that treatment signals alone are a weak proxy for acute aneurysmal SAH, especially when they didn't have a SAH-related ICD code.

SAH ICD codes thus seem to reflect the presence of subarachnoid blood in a broad sense, without consistently distinguishing the cause. Treatment signals, in turn, indicate aneurysm-related care, but do not necessarily imply a current rupture. In that context, many of the strong negative labels appear clinically plausible.

Mid-range cases. The mid-range group was small, with 35 admissions receiving scores between -3 and $+3$. These cases provide some insight into how the model handled uncertainty. On review, they generally did not look like obvious aSAH cases that had simply been missed. More often, the available evidence was incomplete, indirect, or partly conflicting.

Weak positive cases usually had some findings that made aneurysmal rupture possible, but the evidence was not strong enough to treat it as clearly established. For example, imaging might mention a possible aneurysm, but without clear confirmation that it had ruptured or without subsequent aneurysm-directed treatment. Indeterminate cases often involved major bleeding in or around the brain, including some blood in the subarachnoid space, but the underlying cause

remained unclear. In a few cases, the imaging was also difficult to interpret, for example because the patient had absent cerebral blood flow.

Weak negative cases usually had another explanation that seemed more likely than aneurysmal rupture. Examples included bleeding in patients using anticoagulant medication, bleeding related to high blood pressure that mainly started inside the brain and only partly extended into the subarachnoid space, imaging appearances after cardiac arrest that can mimic subarachnoid bleeding, or bleeding linked to another underlying condition. However, aneurysmal rupture was not always fully ruled out in these cases, which likely explains why they received weakly negative scores instead of strongly negative ones.

Only one case showed a clear conflict between strong positive and strong negative evidence. In that admission, the discharge summary suggested aneurysmal SAH, while multiple imaging studies did not identify an aneurysm. The intermediate score seemed to reflect this ambiguity.

5.1.5 Why the final cohort was restricted to *likely aSAH*

The final cohort was restricted to admissions labelled *likely aSAH* to keep the target phenotype as specific as possible. The distinction between *likely* and *possible* cases was based on the strength of evidence for aneurysmal causation in the available record, rather than on the presence of subarachnoid hemorrhage alone.

This restriction also helps explain why only 426 of the 1,576 SAH-coded admissions were retained. Although aneurysm rupture is often discussed as a common cause of spontaneous SAH, the ICD-based search used here was broader. It selected admissions where subarachnoid bleeding was coded or documented, but it did not by itself establish why the bleeding occurred. Inspection of the SAH-coded group showed several situations outside the intended phenotype, including bleeding after trauma, cases where no aneurysm was found despite investigation, bleeding patterns not typical of ruptured aneurysm, bleeding related to another intracranial hemorrhage, and bleeding after procedures.

Restricting the cohort to *likely aSAH* was therefore a conservative design choice. It reduced the chance that non-aneurysmal SAH cases entered the downstream analysis, at the cost of excluding some borderline admissions where aneurysmal origin could not be established with enough confidence from the available documentation. The *possible aSAH* group was small relative to the clearly retained and excluded groups, so this restriction mainly served to improve phenotype specificity before the multimodal variable construction and causal discovery steps.

5.2 Coverage and data availability across modalities

Since the study combines structured records, ICU measurements, laboratory values, and clinical text, the first step was to examine how often these different sources were available. In general, the cohort had good coverage across modalities, although availability differed by source and by variable type.

Discharge summaries were available for 456 admissions (97.0%), and bundled radiology reports were available for 468 admissions (99.6%). Both note types were present in 454 admissions (96.6%). No admission lacked both a discharge summary and radiology text.

ICU data were also available for most admissions. In total, 445 admissions (94.7%) had at least one ICU stay, while 25 admissions (5.3%) had no documented ICU stay. Most admissions had a single ICU episode, although 46 admissions had two ICU stays and one admission had three. The median total ICU length of stay was 243 hours, or approximately 10 days, with an interquartile range of 133-424 hours. This relatively long ICU exposure is compatible with aSAH care, where patients often need close neurological monitoring during the period in which hydrocephalus, rebleeding, vasospasm, and delayed cerebral ischemia may develop [34, 66].

Routine laboratory and bedside physiologic data were broadly available. White blood cell

count and glucose measurements were each available for 456 admissions (97.0%), sodium for 455 admissions (96.8%), and magnesium for 452 admissions (96.2%). Heart-rate data were available for 445 admissions (94.7%), which matched ICU-stay coverage. Oxygenation data were more uneven. A basic oxygen signal was available for 445 admissions, while more detailed oxygen-related blood test information was available for fewer admissions. However, these variables partly served as backups for the same broader question: whether there were signs that the patient had low oxygen levels. In most admissions, the routinely recorded bedside oxygen measurement was enough to screen for this.

Structured procedure and treatment data showed lower coverage than notes, ICU data, and laboratory measurements. Aneurysm-directed treatment was recorded in 332 admissions (70.6%), mostly through coiling codes. Nimodipine was recorded in 170 admissions (36.2%). Because nimodipine is standard treatment after aSAH [34], this relatively low structured rate should be interpreted cautiously. It probably reflects incomplete capture of medication exposure in the structured tables, especially for oral or ward-administered medication, more than true absence of treatment in most patients.

Table 5.1: Coverage of main modalities and selected derived variables in the final cohort. Percentages are calculated over 470 admissions.

Modality or variable	Admissions with data	Percentage
Discharge summary	456/470	97.0%
Radiology report bundle	468/470	99.6%
Both discharge and radiology text	454/470	96.6%
At least one ICU stay	445/470	94.7%
Heart-rate data	445/470	94.7%
White blood cell count	456/470	97.0%
Glucose measurement	456/470	97.0%
Sodium measurement	455/470	96.8%
Magnesium measurement	452/470	96.2%
Any SpO ₂ measurement	445/470	94.7%
Any PaO ₂ measurement	328/470	69.8%
Any FiO ₂ measurement	302/470	64.3%
Any aneurysm-directed procedure	332/470	70.6%
External ventricular drain	112/470	23.8%
Nimodipine exposure in structured data	170/470	36.2%

Overall, the final dataset is strongly multimodal. Almost all admissions had clinical text, and most had ICU exposure, bedside monitoring, and routine laboratory values. The most relevant gaps were more specific. First, ICU-derived variables were unavailable for admissions without an ICU stay. Second, the absence of FiO₂ limited the interpretation of oxygenation severity. Third, some treatment variables, especially nimodipine, appeared to be incompletely captured in structured data. These issues were taken into account when interpreting the feature distributions and the later modelling results.

5.3 Structured variable distributions

The structured variables were inspected to assess whether the final cohort showed a clinically plausible aSAH profile. This inspection focused on the overall distribution of patient characteristics, treatment patterns, complications, and physiologic states, while also considering where

the structured sources appeared incomplete.

5.3.1 Demographics and baseline history

The cohort had a mean age of 58.3 years, with a median of 57 years and an interquartile range of 49-69 years. The youngest patient was 19 and the oldest was 93. This age distribution fits reasonably well with aSAH, which most often affects adults in middle and later life [16, 107]. Women were more common than men: 317 admissions (67.4%) were female and 153 admissions (32.6%) were male. This is also in line with the female predominance often reported in aSAH cohorts [39, 79].

Baseline risk factors were common. Hypertension was documented in 252 admissions (53.6%), smoking in 160 admissions (34.0%), diabetes in 42 admissions (8.9%), and a history of thromboembolic events in 55 admissions (11.7%). The hypertension prevalence appears clinically plausible for an aSAH cohort [39, 93], and the same is true for diabetes [39, 93]. Smoking was also frequent, which is expected given its role as a modifiable risk factor for aneurysm formation and rupture [16].

The documented prevalence of thromboembolic history was 11.7%. This is higher than in at least one previously described cohort, a 10-year study at Ghent University Hospital where thromboembolic history was noted in 1.4% of patients [16]. Direct comparison is difficult, however, because pre-existing thromboembolic disease is not always reported in aSAH cohorts, and definitions may differ between studies.

The structured vasospasm flag was present in 65 admissions (13.8%). The relatively low vasospasm rate probably reflects under-ascertainment through ICD coding, at least in part, more than a truly low burden of vasospasm [103].

5.3.2 Treatment and care trajectory

Most admissions had evidence of aneurysm-directed treatment. Coiling was much more common than clipping: 280 admissions (59.6%) had coiling, 54 admissions (11.5%) had clipping, and one admission had open ligation. Three admissions had both clipping and coiling codes. On manual inspection, these appeared to represent combined or staged aneurysm management. In total, 332 admissions (70.6%) had at least one recorded aneurysm-directed procedure. This pattern is clinically plausible, as coiling is often favored when anatomy and patient condition allow it [16].

Treatment timing was generally early among admissions with recorded intervention. Same-day treatment occurred in 191 admissions (40.6%), and treatment within one to two days occurred in another 120 admissions (25.5%). Only 21 admissions had treatment delayed beyond two days. Among treated patients, the median treatment delay was 0 days, with an interquartile range of 0-1 day. This is consistent with aSAH care, where early aneurysm securement is used to reduce the risk of rebleeding [34]. A small number of negative or very long delays were retained as data-quality review cases, since they may reflect date granularity, transfer timing, or staged treatment.

Nimodipine was recorded in 170 admissions (36.2%). This is lower than expected clinically, because nimodipine is considered standard treatment after aSAH and is used to improve neurological outcome [16, 34, 39]. The most cautious interpretation is under-documentation in the structured medication source, rather than true absence of treatment in most patients. External ventricular drainage was recorded in 112 admissions (23.8%), which is low given the high hydrocephalus burden (49.1%) but plausible depending on protocol [16, 93]. Antifibrinolytic therapy was uncommon: only 8 admissions (1.7%) had any order or administration signal, and only one had a documented administration. This also fits with clinical expectations [16, 34].

5.3.3 Neurological severity

Neurological severity was represented mainly through GCS variables and ICU-care trajectory. The largest limitation was missingness in the pre-treatment GCS representation. Eye, verbal, motor, and total worst pre-treatment GCS were each missing in 234 admissions (49.8%).

5.3.4 Complications and physiologic states

Complications and physiologic derangements were frequent in the cohort, which fits with the high ICU exposure described above. To make these results easier to interpret, the main findings are grouped into imaging-related complications, electrolyte disturbances, blood pressure and oxygenation abnormalities, inflammatory or metabolic responses, and less frequent secondary states.

Table 5.2 reports the main threshold-level findings, while the results of the exact categorical definitions used are provided in Appendix C.

As shown in Table 5.2, complications and physiologic abnormalities were common in the final cohort. Hydrocephalus was the most frequent imaging-related complication, which is clinically plausible in aSAH [19]. IPH and IVH were less frequent, although the IVH rate may be underestimated [16, 93]. Overall, imaging-related complications were present in more than half of admissions, supporting the interpretation that this was a clinically severe, ICU-enriched cohort.

Sodium abnormalities were also frequent. Hyponatremia was more common than hypernatremia, and a smaller group of admissions showed both low and high sodium values during the admission. This pattern is plausible, since hyponatremia is a well-known complication after SAH, with reported prevalences commonly around 30–56% [16, 66]. Hypernatremia was less frequent, but still relevant because it may reflect severe illness, fluid shifts, diabetes insipidus, or active sodium management.

Blood pressure and oxygenation abnormalities were common when repeated bedside measurements were used. Low MAP values occurred frequently, including under stricter repeated-threshold definitions [32, 108]. Oxygenation abnormalities followed a similar pattern, with both routine bedside oxygen measurements and blood-test-based oxygen measurements contributing evidence [73]. These findings are plausible in an ICU-enriched aSAH cohort, where pulmonary complications, aspiration, neurogenic pulmonary edema, and mechanical ventilation may all occur [16].

Inflammatory and metabolic abnormalities were among the most frequent structured findings. Leukocytosis was very common, including at higher WBC thresholds [65]. This is clinically plausible after aSAH, although the high prevalence may also reflect frequent repeated laboratory testing in an ICU population [50, 93]. Hyperglycemia was also common and is expected after acute brain injury [16, 57]. The contrast between laboratory-defined hyperglycemia and ICD-coded hyperglycemia was large, suggesting that diagnosis codes captured only a small part of the measured physiologic burden.

Several less central states provided useful supporting context. Hypercapnia and hypomagnesemia were present in smaller subsets of admissions. Interpretation of magnesium values is more difficult because magnesium sulfate was frequently administered, meaning treatment may have prevented or corrected measured deficiency. Hypovolemia was rarely captured by codes, which likely reflects the difficulty of representing volume status in administrative data.

5.4 LLM extraction results

The main purpose of this section is not only to report the output of the LLM extraction pipeline, but to assess whether the extracted variables behave coherently enough to be used in the later multimodal analysis. Since there is no single ground-truth vasospasm label in MIMIC-IV, confidence in the extracted variables has to come from triangulation: comparison between discharge

Table 5.2: Selected complications and physiologic states in the final cohort. Percentages are calculated over 470 admissions.

Variable or state	Admissions	Percentage
<i>Imaging-related complications</i>		
Hydrocephalus	231/470	49.1%
IPH	64/470	13.6%
IVH	33/470	7.0%
Any hydrocephalus, IPH, or IVH	265/470	56.4%
<i>Electrolyte disturbances</i>		
Any hyponatremia	225/470	47.9%
Any hypernatremia	96/470	20.4%
No sodium abnormality	185/470	39.4%
Hyponatremia only	177/470	37.7%
Hypernatremia only	55/470	11.7%
Both hypo- and hypernatremia	53/470	11.3%
<i>Blood pressure and oxygenation</i>		
Any MAP < 65 mmHg	359/470	76.4%
At least three MAP values < 65 mmHg	259/470	55.1%
Repeated MAP < 60 mmHg	164/470	34.9%
SpO ₂ < 90%	228/470	48.5%
PaO ₂ < 80 mmHg	160/470	34.0%
PaO ₂ < 60 mmHg	69/470	14.7%
<i>Inflammatory and metabolic responses</i>		
WBC $\geq 10 \times 10^9/\text{L}$	402/470	85.5%
WBC $\geq 15 \times 10^9/\text{L}$	258/470	54.9%
Glucose > 140 mg/dL	362/470	77.0%
Glucose > 180 mg/dL	190/470	40.4%
ICD-coded hyperglycemia	14/470	3.0%
<i>Other secondary states</i>		
PaCO ₂ > 45 mmHg	99/470	21.1%
PaCO ₂ > 50 mmHg	72/470	15.3%
Magnesium < 1.7 mg/dL	147/470	31.3%
Severe hypomagnesemia	6/470	1.3%
Magnesium sulfate administration	307/470	65.3%
ICD-coded hypovolemia	11/470	2.3%

summaries, radiology reports, ICD codes, structured procedure records, and manual inspection of selected disagreement cases.

The main finding is that note-derived vasospasm evidence is much broader than ICD-based vasospasm coding. Discharge summaries and radiology reports identify many admissions with documented vasospasm evidence that are not captured by ICD codes. At the same time, the two text sources show substantial agreement overall, and most disagreements can be explained by differences in timing, source scope, or documentation style rather than by obvious extraction failure.

5.4.1 Vasospasm signal across modalities

Distribution of vasospasm scores

The ICD-based signal identified vasospasm in 65 of 470 admissions (13.8%). This was lower than the note-derived signals. Discharge summaries were positive for vasospasm in 48.3% of scored admissions, while radiology reports were positive in 52.6%. These proportions were much higher than the ICD-based estimate and are closer to reported rates of vasospasm after aSAH [34]. Figure 5.3 shows the distribution of the ICD-based indicator.

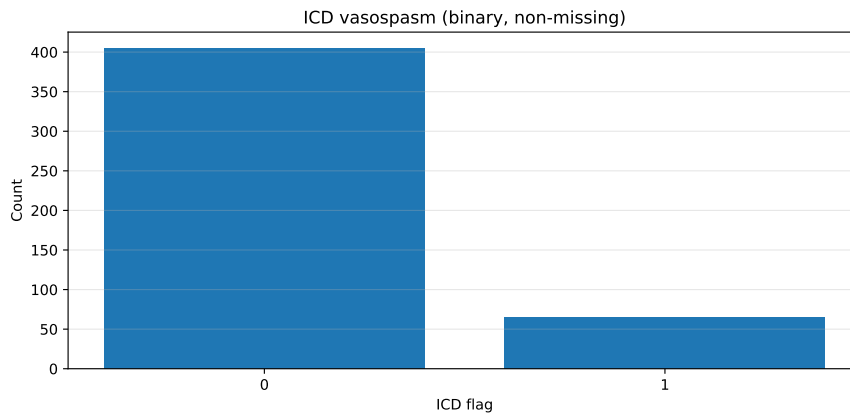


Figure 5.3: Distribution of the ICD-based vasospasm indicator.

The ordinal note-derived scores were grouped into five buckets: strong negative (≤ -3), weak negative (-2 to -1), neutral (0), weak positive (1 to 2), and strong positive (≥ 3). This grouping made it easier to compare the overall shape of the discharge- and radiology-derived signals without treating each score value as a separate category.

For discharge summaries, the distribution was concentrated mainly at the two extremes. Many admissions were assigned either a strong negative or a strong positive score, while the intermediate categories were less common. This pattern is consistent with the role of the discharge summary as a condensed account of the hospital course: when vasospasm was clearly present or clearly absent, it was often reflected in stronger language.

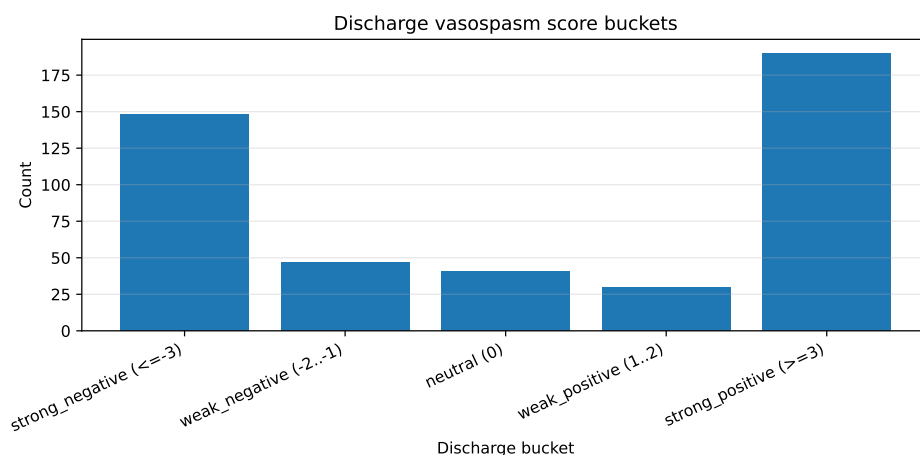


Figure 5.4: Distribution of discharge summary vasospasm score buckets.

Radiology reports showed a somewhat different distribution. The share of strong positive scores was similar to the discharge summaries, but there were fewer strong negative scores and more neutral scores. Weak negative radiology scores were uncommon. This may reflect the more time-specific nature of radiology reporting: reports often describe the findings of a particular scan or angiographic study, and not necessarily the full admission-level clinical course.

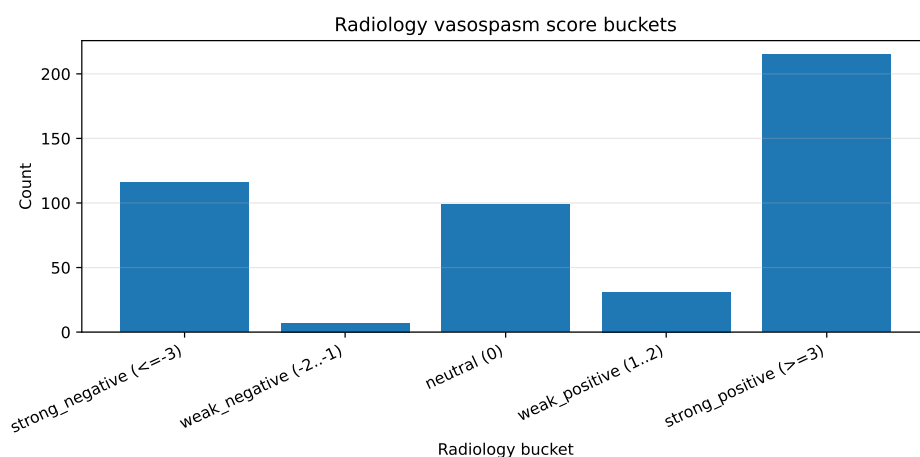


Figure 5.5: Distribution of radiology vasospasm score buckets.

For comparison with ICD coding and for the agreement analyses below, the note-derived ordinal scores were also converted into binary indicators. Scores ≥ 1 were treated as positive, while scores ≤ 0 were treated as negative. The resulting discharge-derived binary signal was more balanced than the ICD-based indicator, with almost half of scored admissions classified as positive.

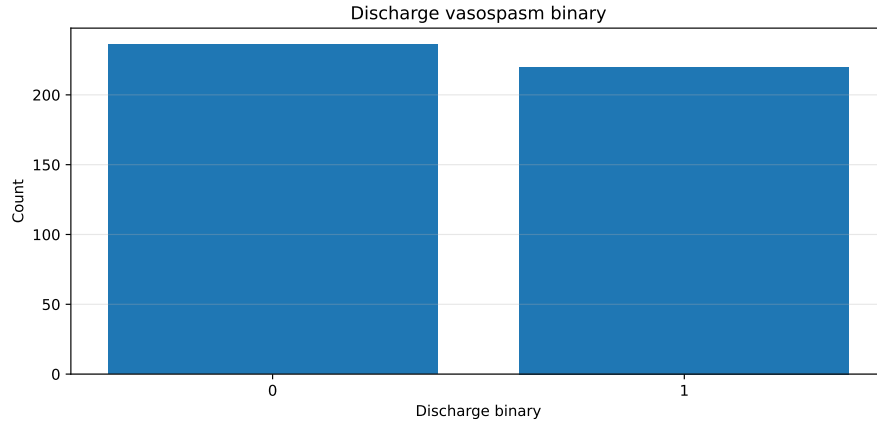


Figure 5.6: Binary vasospasm classification derived from discharge summaries. Scores ≥ 1 were treated as positive and scores ≤ 0 as negative.

The radiology-derived binary signal showed a similar pattern, with a slightly higher positive proportion than the discharge-derived signal. Taken together, the two note-derived modalities identified a broader set of vasospasm evidence than the ICD-based indicator.

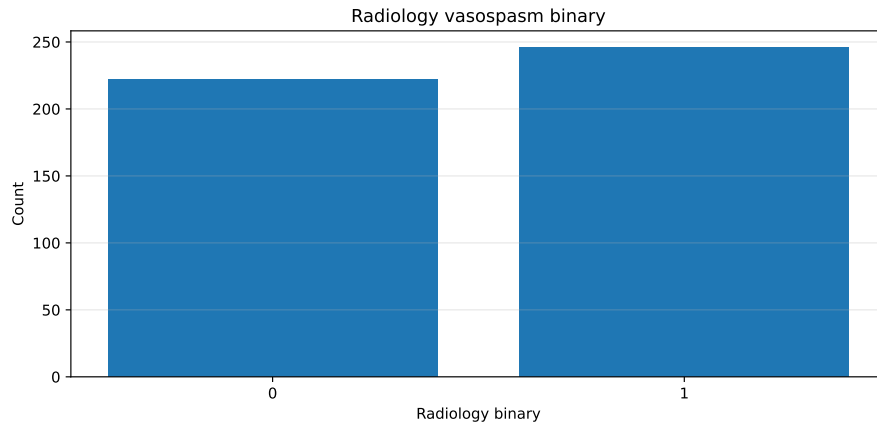


Figure 5.7: Binary vasospasm classification derived from radiology reports. Scores ≥ 1 were treated as positive and scores ≤ 0 as negative.

Agreement between discharge summaries and radiology reports

At the binary level, discharge summaries and radiology reports agreed in 394 cases (86.8%). Most disagreement cases were either jointly positive or jointly negative. Disagreement was less common and was somewhat asymmetric: radiology was positive while discharge was negative in 41 admissions, whereas the opposite pattern occurred in 19 admissions.

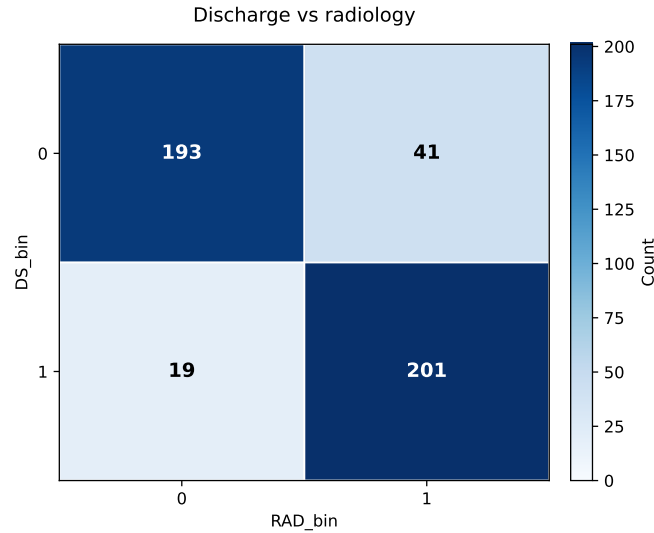


Figure 5.8: Confusion matrix comparing discharge- and radiology-derived vasospasm using binary classification. Scores ≥ 1 were treated as positive and scores ≤ 0 as negative.

The discharge and radiology scores were also compared at the bucket level, which gives more detail than the binary comparison.

Agreement was clearest when evidence was strong. Strong positive agreement was common, with 169 admissions placed in the strong positive category by both sources. Strong negative agreement was also seen, with 89 admissions placed in the strong negative category by both sources. Outside these extremes, the relationship was more spread out.

Discharge weak negative and neutral cases often mapped to neutral radiology categories. Some discharge strong positive cases were paired with weaker radiology signals. This may happen because discharge summaries can combine several observations across the hospital stay, while individual radiology reports are tied more closely to specific imaging moments.

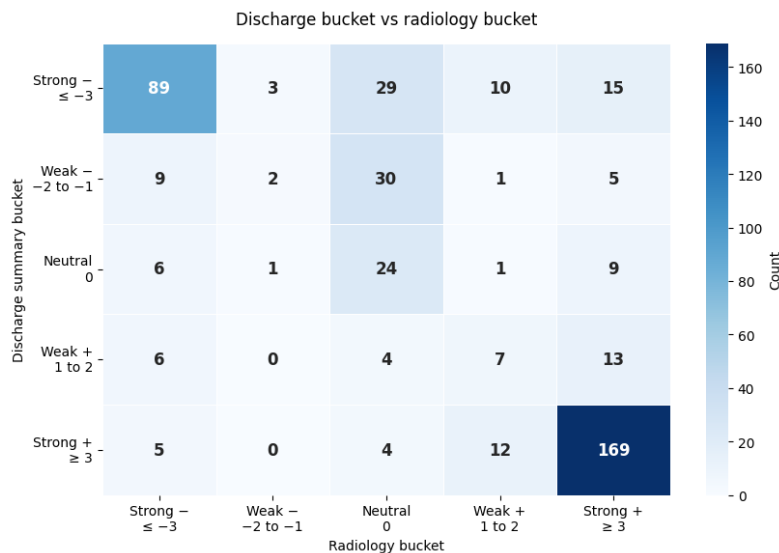


Figure 5.9: Alignment of discharge and radiology vasospasm scores at the bucket level.

Disagreement patterns

Disagreement was examined more directly using categories based on the original ordinal scores. Strong disagreement was defined as opposing strong signals across modalities: discharge score

≥ 3 with radiology score ≤ -3 , or discharge score ≤ -3 with radiology score ≥ 3 . A total of 20 admissions met this definition.

Another 11 admissions showed weak-opposite disagreement. In the implemented definition, this meant that the discharge score was weakly positive or weakly negative ($|score| = 1$ or 2), while the radiology score was strong in the opposite direction ($|score| \geq 3$). Finally, 10 admissions fell into a both-weak category, where both discharge and radiology scores were weak in magnitude ($|score| = 1$ or 2).

These categories show that disagreement was concentrated in a relatively small part of the cohort. Most admissions showed agreement, while clear opposition between the two sources was uncommon.

Manual review of strong opposition cases. Strongly discordant cases were reviewed to understand how opposing signals appeared in practice. The most common pattern involved time-dependent evidence. In several admissions, the discharge summary described CTA or TCD monitoring as negative for vasospasm, while later radiology reports documented mild, moderate, or severe vasospasm on CTA or conventional angiography. In these cases, radiology captured a positive finding at a specific time point that was absent from, or only briefly summarized in, the discharge narrative.

A second pattern involved differences in the level of evidence used by each source. Radiology reports sometimes contained direct vessel-level descriptions, such as narrowing, beading, or angiographic spasm. Discharge summaries more often focused on the overall clinical course and final neurological status. The reverse pattern also occurred: some discharge summaries described clinical vasospasm or vasospasm-directed treatment without a single radiology report using equally strong wording. These cases were therefore often related to documentation, timing, and summarisation, rather than clear extraction errors.

Manual review of weak-opposite cases. The weak-opposite cases showed a milder version of the same issue. These were usually admissions where one source contained a weak or qualified signal, while the other contained a stronger statement in the opposite direction. For example, a discharge summary could mention an early CTA showing diffuse vasospasm, while also describing later negative TCD or CTA findings and a stable clinical course. In that situation, the discharge score could remain weakly positive, while radiology could become strongly negative if the final neurovascular report explicitly stated that there was no vasospasm.

The reverse pattern also occurred. Some discharge summaries contained only indirect negative evidence, such as no vasospasm being mentioned in an otherwise detailed hospital course, while radiology reports documented vasospasm across serial CTA studies. These cases illustrate why weak scores were useful. They kept partial or uncertain evidence visible without forcing every admission into a firm positive or negative category.

Manual review of both-weak cases. The both-weak cases had a different character. They were not always true disagreements. Many reflected low-confidence evidence in both sources. Some admissions contained suggestive but non-definitive radiology language, such as vessel narrowing that "could reflect vasospasm" or findings where vasospasm was "not excluded". Others contained treatment hints, such as intra-arterial verapamil, without explicit use of the term vasospasm. In these cases, the evidence was stronger than routine surveillance alone, but weaker than direct documentation.

Several both-weak cases also had plausible alternative explanations for neurological decline, including mass effect, edema, herniation, infarction, infection, or poor baseline neurological status. This made attribution to vasospasm difficult. The both-weak category therefore mainly captured uncertainty and caution in interpretation.

Overall, manual review suggested that disagreement had a recognizable structure. Strong opposition often reflected conflicting time-point evidence or differences between imaging-level and summary-level documentation. Weak-opposite cases reflected partial conflict between weaker and more definitive evidence. Both-weak cases mostly represented ambiguous records.

Taken together, these checks suggest that the LLM-derived vasospasm variables should not be read as a replacement for a clinical gold standard. Instead, they provide a richer source-aware representation of documented vasospasm evidence than ICD coding alone. This is the representation carried forward into the multimodal comparison and causal discovery analyses.

Manual review of death cases. In addition to the discordance-based review, admissions in which the patient died during the typical vasospasm risk window were also selected for manual inspection. This added 49 cases. These admissions were clinically important because early death during this period can make vasospasm difficult to evaluate: deterioration may be caused by the initial hemorrhage before a vasospasm could actually develop, or exactly *because* of a vasospasm. After review by a medical professional, these cases were considered too difficult to manually assess reliably for vasospasm based on the available documentation. They were therefore not assigned a separate manual vasospasm score.

Comparison with ICD-based vasospasm signal

The ICD-based vasospasm indicator behaved differently from both note-derived signals.

When compared with discharge summaries ($n = 456$), ICD coding agreed well in negative cases, with 229 admissions jointly negative. Positive overlap was much lower. Of 220 discharge-positive admissions, only 58 were also ICD-positive.

A similar pattern was seen for radiology ($n = 468$). Of 246 radiology-positive admissions, only 58 had a corresponding ICD code. At the same time, ICD-positive cases were usually supported by note-based evidence: in the pair-complete comparisons, 58 of 65 ICD-positive admissions were also discharge-positive, and 58 of 65 were also radiology-positive.

This asymmetry suggests that ICD coding captured a narrower subset of vasospasm cases. Discharge summaries and radiology reports identified a broader set of documented vasospasm signals.

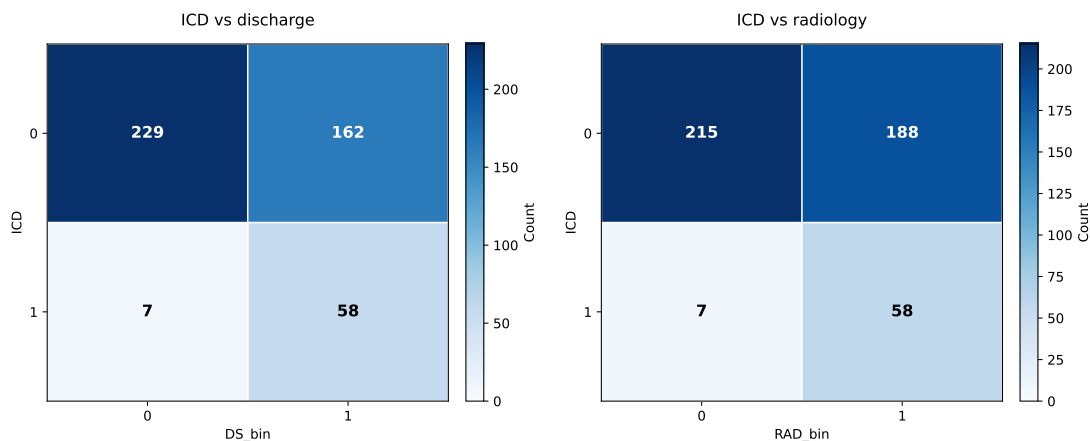


Figure 5.10: Comparison of ICD-based vasospasm coding with discharge-derived and radiology-derived binary signals.

Manual review of discordant ICD-positive cases

ICD-positive cases with negative discharge and radiology scores were manually reviewed to better understand the mismatch between administrative coding and note-derived evidence.

In the reviewed cases, both discharge summaries and radiology reports often contained repeated evidence against vasospasm. Serial CTA and TCD monitoring frequently documented absence of vasospasm. When transient concerns were mentioned, they were sometimes clarified by more definitive imaging such as conventional angiography.

These cases did not appear borderline based on the extracted text. Both modalities reached strong negative scores (-5), reflecting repeated negation across clinically relevant investigations. This suggests that the negative note-based classification was not simply due to missing text evidence. ICD coding may sometimes reflect suspected, monitored, or administratively recorded vasospasm, even when the available narrative record does not support a confirmed event. The exact coding rationale cannot be reconstructed from the available data. Still, these cases show why ICD-based vasospasm should be interpreted carefully when used as a proxy for clinically documented vasospasm.

This ICD comparison adds a second form of validation. The issue was not only whether discharge summaries and radiology reports agreed with each other, but also whether the note-derived signals captured information missed by administrative coding. The results suggest that ICD codes were relatively specific when present, but insensitive as a standalone vasospasm definition.

Manual review refinement

A subset of borderline and discordant admissions was manually re-evaluated and added to the dataset as an explicit annotation layer. This produced a `manual_review` variable for 14 admissions.

Some cases were shifted from weak negative to neutral or weak positive, reflecting ambiguous or evolving evidence. Others were upgraded to strong positive when supporting evidence was present but not fully captured by the initial extraction.

This annotation layer was used as a small correction and quality-control layer for cases where the original score did not fully match the reviewed evidence.

5.4.2 Consistency and cross-modal validation of additional LLM-extracted variables

The goal of this subsection is to evaluate whether the additional LLM-derived variables behave in a source-consistent and clinically plausible way. As with vasospasm, most of these variables do not have a single ground-truth field in MIMIC-IV. Confidence therefore has to come from agreement between discharge summaries, radiology reports, and structured records where overlap exists, and plausible disagreement patterns where sources differ.

Beyond vasospasm, the extraction pipeline produced variables related to aneurysm characteristics, hemorrhage severity, treatment, complications, baseline history, and medication exposure. These variables were evaluated in two ways. First, cross-modal agreement was examined where the same concept could be compared across discharge summaries, radiology reports, or structured data. Second, internal consistency was checked within extracted variable groups, for example whether severity scores aligned with severity categories or whether treatment labels matched procedure evidence. The purpose of these checks was not to prove that every extracted value was correct, but to assess whether the variables were coherent enough to support the later multimodal and causal discovery analyses.

Aneurysm location agreement

Aneurysm location, grouped as anterior versus posterior circulation, showed very high agreement between discharge summaries and radiology reports. Among 454 admissions with location information in both modalities, concordance exceeded 97% for both anterior and posterior classifications.

Strict anatomical disagreement was rare. Only one admission showed an anterior-versus-posterior mismatch. Most remaining discrepancies involved cases labeled as *unknown* in one modality. In many of these cases, the other modality provided a definitive classification.

This pattern suggests that aneurysm location was one of the more stable LLM-extracted features in this dataset. When location was explicitly documented, the two modalities usually gave the same broad anatomical classification.

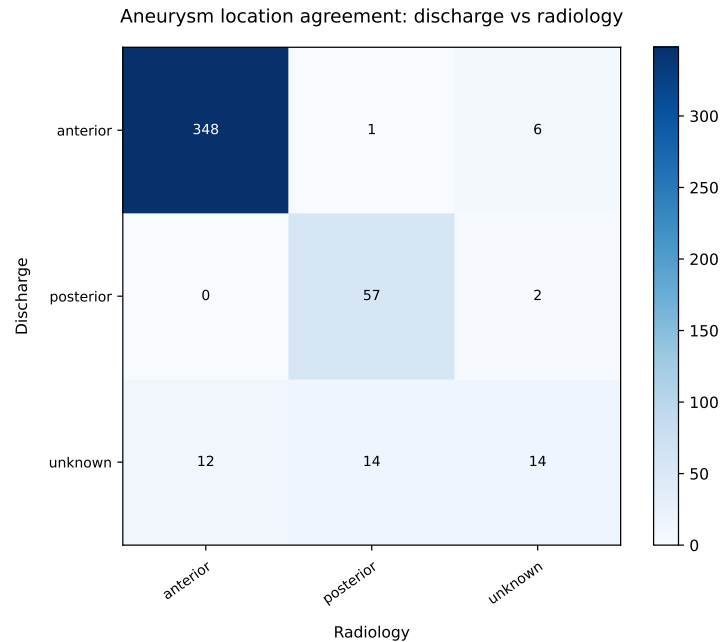


Figure 5.11: Agreement between discharge- and radiology-derived aneurysm location (anterior vs posterior).

Hemorrhage severity: clot burden and formal grading

Discharge summaries often contained enough information to assign a clot burden proxy, whereas formal grading systems were less consistently reported. A clot burden estimate was available in 424 of 456 admissions (93.0%). At least one formal grade, such as Fisher, Hunt-Hess, or WFNS, was present in 177 admissions (38.8%).

Despite this difference in availability, the two severity representations were broadly aligned. Higher clot burden categories tended to occur together with higher formal severity grades. This suggests that the discharge-derived clot burden variable captured a severity dimension that was related to established grading systems, even though it was not identical to them.

The clearest comparison was with Fisher grade. Higher Fisher grades were concentrated in the higher clot burden categories, with grades 3 and 4 occurring mainly in the highest clot burden group. This is in line with the clinical meaning of both variables, since they both reflect the extent of hemorrhage visible on imaging.

A similar pattern was seen for Hunt-Hess grade. Although Hunt-Hess is a clinical severity scale rather than a direct imaging-based measure, higher grades still tended to appear together with higher clot burden categories. The relation was therefore somewhat broader in meaning, but still clinically plausible.

The comparison with WFNS showed the same general tendency. Higher WFNS grades were more common in admissions with higher clot burden. Taken together, the three comparisons suggest that the clot burden proxy behaved in a coherent way across different severity frameworks.

This relationship was also reflected in moderate Spearman correlations (ρ between 0.51 and 0.59). These values suggest that the clot burden proxy captured a related severity dimension, while still remaining distinct from formal grading.

Comparison with radiology-derived clot burden showed that the two sources lined up well for the highest category. Most admissions fell in the highest category in both modalities, suggesting that extensive hemorrhage was usually recognised across sources. Differences were more common in the lower categories, where discharge summaries tended to be less precise and sometimes mapped to a higher category in radiology. The disagreement therefore seemed to concern fine-grained grading more than the presence of substantial clot burden.

The intermediate category was rarely used, which made it difficult to interpret as a meaningful separate level. For later analysis, a simpler binary representation therefore appeared more robust than the full ordinal scale.

Treatment extraction and cross-source validation

Treatment variables provided another useful cross-check between narrative extraction and structured procedure records.

Discharge-derived aneurysm treatment type aligned well with structured indicators for coiling and clipping. Among cases labeled as "coil" in discharge summaries, 85.1% had a corresponding structured coiling record. Among cases labeled as "clip", 77.6% had a corresponding structured clipping record. Radiology reports also supported this signal: in nearly all cases where a treatment type was specified in the discharge summary, radiology mentioned treatment as well.

A small number of mismatches remained between the core treatment labels and radiology-derived descriptions. On closer inspection, these often reflected more complex treatment trajectories. Some admissions involved multiple interventions over time, such as clipping followed by stenting or attempted coiling, while the structured core label reduced this to one dominant category. Radiology reports could mention a different procedure depending on the timing of the study or the focus of the report.

Other discrepancies were related to incomplete or uneven documentation. In a few cases, radiology did not explicitly mention a treatment that was present in the structured data, likely because the report focused on imaging findings and not procedural history. Overall, the remaining mismatches seemed to reflect differences in temporal scope, level of detail, and reporting emphasis.

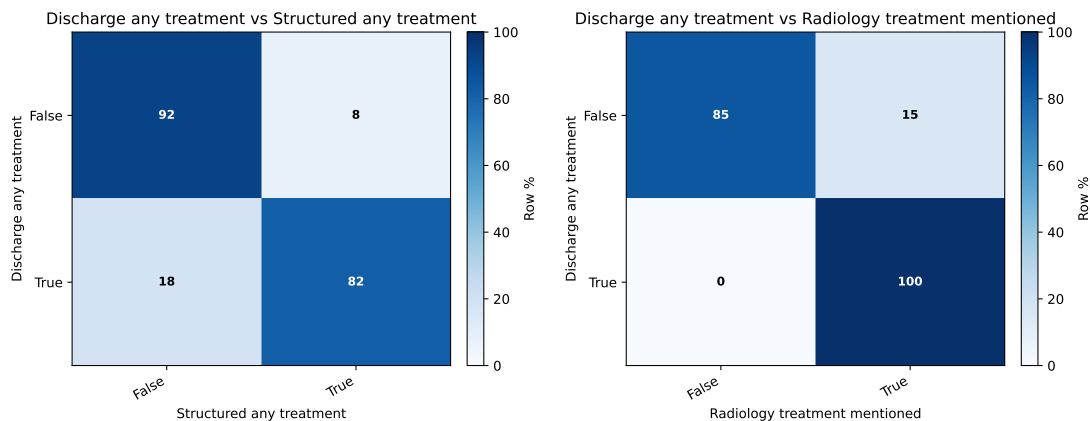


Figure 5.12: Cross-modal validation of aneurysm treatment type across discharge summaries, structured records, and radiology.

Hydrocephalus and EVD consistency

Hydrocephalus showed moderate agreement between discharge summaries and radiology reports. When discharge summaries explicitly stated hydrocephalus, radiology confirmed it in 71.0% of cases. When discharge summaries stated that hydrocephalus was absent, radiology rarely contradicted this. Most disagreement involved uncertainty categories. Cases labelled as *unclear* or *unknown* in discharge summaries often mapped to either present or absent in radiology, which is compatible with incomplete summarisation in the discharge narrative.

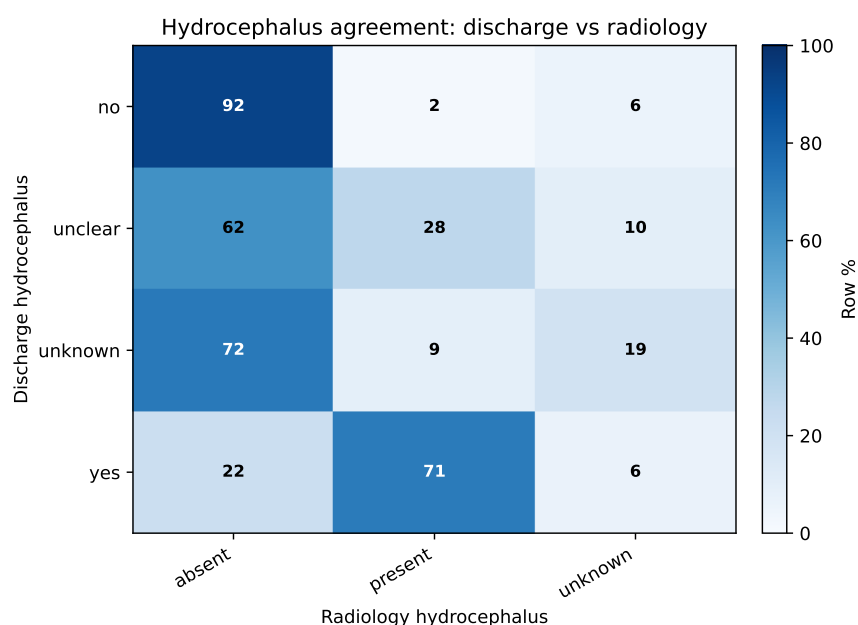


Figure 5.13: Agreement between discharge- and radiology-derived hydrocephalus status.

EVD detection showed a somewhat different pattern. Agreement was high when both sources indicated that an EVD was present, but radiology often identified EVD presence in cases where structured procedure records did not. This is plausible because radiology reports can identify devices directly, for example through catheter position or tip location, whereas structured data depends on procedure coding and on whether the relevant event was captured within the admission window.

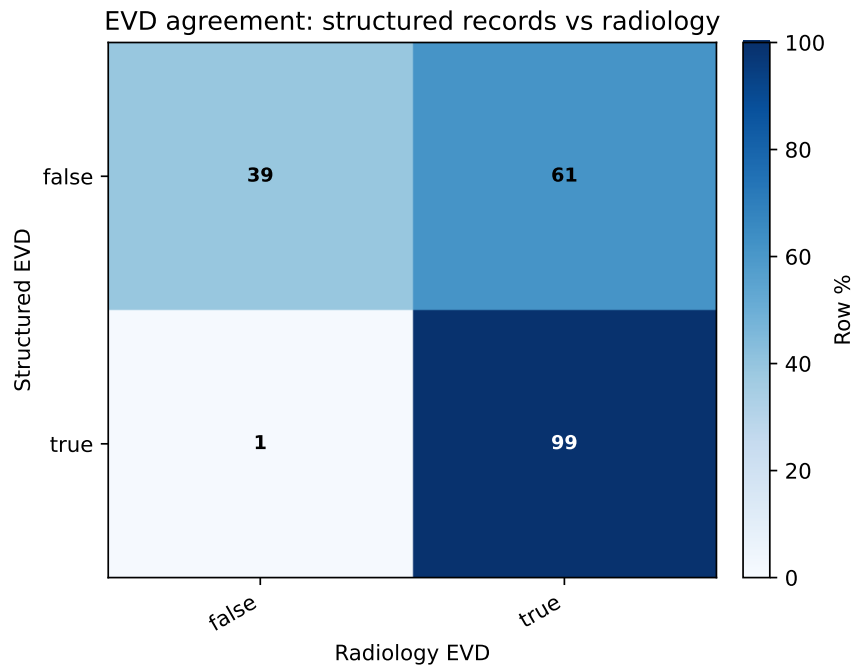


Figure 5.14: Agreement between radiology-derived EVD presence and structured procedure-based EVD detection.

Overall, these results suggest that hydrocephalus was captured reasonably consistently across narrative sources, with most disagreement concentrated in uncertain discharge labels. EVD showed a different type of mismatch, where radiology appeared more sensitive to device presence, while structured procedure fields provided a narrower administrative signal.

Baseline history extraction

Baseline comorbidities extracted from discharge summaries generally aligned with structured diagnosis-based history variables. For example, discharge-documented hypertension corresponded to structured hypertension in more than 90% of cases, and diabetes showed a similar pattern.

Discrepancies were often related to the *unknown* category in discharge summaries. In some of these cases, structured data still contained evidence of the comorbidity. This is clinically plausible, since absence of mention in a discharge summary does not necessarily mean that a condition was absent. The condition may simply not have been central to the admission narrative.

Medication consistency: nimodipine

Nimodipine administration showed substantial disagreement between discharge-derived and structured medication variables. Approximately half of discharge-positive cases did not have a corresponding structured medication record.

Manual inspection suggested that this discrepancy was probably related to documentation differences. Discharge summaries may state that nimodipine was given or prescribed as part of aSAH care, while structured medication tables may miss some short-term, ICU-administered, transferred, or irregularly documented exposures. The reverse also occurred: some structured-positive cases had no clear discharge mention, which is plausible because discharge summaries do not list every medication exposure with the same level of detail.

Internal consistency within radiology extraction

Internal validation within the radiology-derived variables showed mostly coherent patterns. The vasospasm severity indicator aligned well with the ordinal vasospasm score. Cases labeled as "severe" were almost always assigned high positive scores, while cases labeled as "none" tended to receive strongly negative scores.

Bleed phenotype variables also showed a clinically plausible structure. Extensive subarachnoid hemorrhage was usually associated with the highest clot burden category, and mixed hemorrhage phenotypes such as SAH with IVH were common.

One notable pattern was the presence of "unknown" severity labels despite high vasospasm scores. Manual inspection suggested that these cases often contained strong evidence for vasospasm in the report text, but no explicit severity term. The structured severity field therefore appeared more conservative than the ordinal score. The severity field required explicit severity wording, while the score could also reflect strong descriptive evidence.

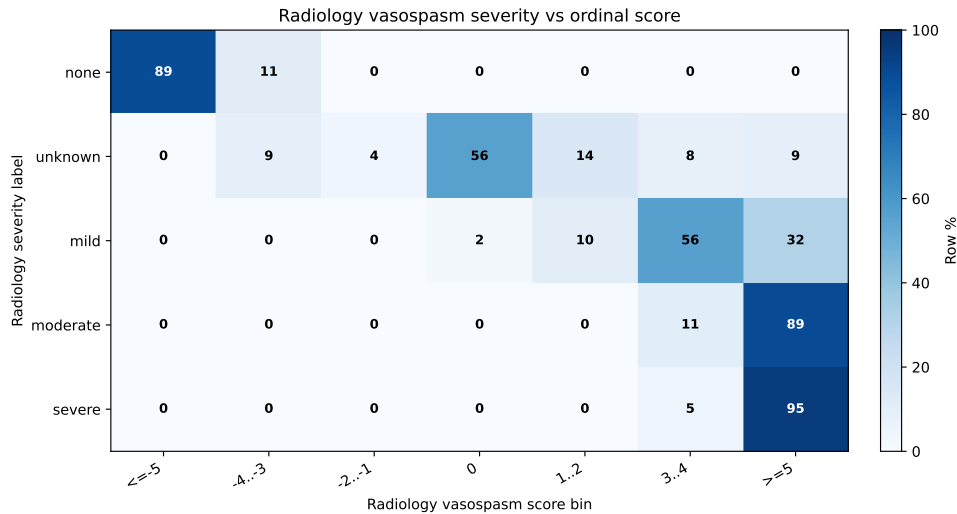


Figure 5.15: Relationship between radiology-derived vasospasm severity categories and ordinal scores.

Summary

Across the additional LLM-extracted variables, two broad patterns were observed.

First, agreement was highest for features that are usually well-defined and explicitly documented, such as aneurysm location and treatment type. These variables showed strong cross-modal consistency and little strict contradiction, suggesting that they were captured relatively reliably when the relevant information was present in the notes.

Second, disagreement was more common for variables that depend on timing, interpretation, or summarisation. This included vasospasm evidence, clot burden, hydrocephalus uncertainty, EVD capture, and medication administration. In these cases, disagreement did not usually indicate simple extraction failure. It often reflected the different roles of the sources: discharge summaries condense the clinical course, radiology reports describe specific imaging moments, and structured records depend on coding or medication/procedure capture.

Later causal discovery analysis treats the extracted variables as graph nodes. If the variables were internally inconsistent or contradicted their source modalities in unexplained ways, the resulting graph structures would be difficult to interpret. The validation results do not provide a definitive gold standard, but they suggest that the main extracted variables are sufficiently traceable, source-aware, and clinically coherent for exploratory downstream analysis.

5.5 Causal discovery results

5.5.1 Overall graph characteristics

Before interpreting individual relations, it is useful to compare the general type of graph produced by each algorithm. All algorithms used the same core set of 32 variables, spanning baseline factors, hemorrhage severity, early secondary complications, workflow and treatment variables, and vasospasm-related outcomes. Differences in graph size, density, and directedness therefore reflect the behaviour of the discovery procedures and their hyperparameter settings, not differences in the included variables.

Metric	PC-stable	FGES	FCIT
Edges per run	14-94; median approximately 28. Density varied with α , conditioning depth, and the G^2 minimum cell-count threshold.	16-58; mean approximately 31. Density varied mainly with maximum degree and BDeu prior settings.	39-42. Edge count varied comparatively little across the sweep.
Directedness	CPDAG-like output with directed and undirected edges; 149 directed and 24 undirected edge forms across the sweep.	Mostly directed in unrestricted runs; 72 directed and 5 undirected edge forms, with some equivalence-class ambiguity.	PAG output; 33 edges were oriented, bidirected, or partially oriented, while 11 remained undirected or circle-marked.

Table 5.3: Summary of graph characteristics across causal discovery algorithms. All algorithms used the same 32-variable node set.

PC-stable showed the widest variation in graph density. This is consistent with its dependence on conditional independence testing. More permissive α values and shallower conditioning depths allowed more adjacencies to remain, while stricter settings and higher minimum cell-count thresholds pruned the graph more aggressively. As a result, PC-stable produced both very sparse and relatively dense graphs across the sweep.

FGES produced graphs of intermediate density. Its sparsity was governed mainly by the maximum degree constraint and BDeu prior strength. Lower or moderate BDeu sample priors allowed more detailed local structures, whereas stronger priors favoured simpler graphs. FGES often returned clearer orientations than PC-stable, but these orientations remain conditional on the scoring assumptions, the priors, and the included variable set.

FCIT produced the most stable edge counts across its sweep, but also the most explicitly ambiguous edge marks. This ambiguity is useful for interpretation. Because FCIT allows for latent confounding, its PAG output is less likely to force a fully directed graph when multiple explanations remain compatible with the observed conditional independences.

Overall, the three algorithms behaved in expected and complementary ways. PC-stable was most sensitive to independence-test settings, FGES to score regularization and degree constraints, and FCIT returned comparatively stable but more ambiguous structures.

5.5.2 Recurrent structures across algorithms

The first interpretive layer concerns recurrent structures: adjacencies and directed edges that appeared repeatedly across algorithms and hyperparameter settings. The main recovered structure was not a single isolated edge, but a pathway linking hemorrhage burden to vasospasm through early complications, severe-care markers, and systemic physiological disturbance:

Clot burden $\rightarrow$ Hydrocephalus $\rightarrow$ EVD $\rightarrow$ Leukocytosis $\rightarrow$ Hyperglycemia $\rightarrow$ Vasospasm.

Figure 5.16 shows a representative constrained FCIT PAG output with this pathway highlighted. The full graph is included to show the broader structure returned by the PAG-based method, while the interpretation below focuses on recurrent structures rather than on every edge in this single graph.

This pathway was the main graph-level finding of the causal discovery analysis. It is also broadly in line with the clinical literature: greater hemorrhage burden is a known risk factor for later vasospasm and delayed ischemic complications, hydrocephalus and EVD placement commonly mark more severe aSAH courses, and inflammatory and metabolic disturbances such as leukocytosis and hyperglycemia have both been associated with poorer outcomes after aSAH [14, 22, 34, 50, 57].

The learned structure suggests that the relationship between clot burden and vasospasm was not mainly recovered as a direct clot burden $\rightarrow$ vasospasm edge. Instead, clot burden appeared at the start of a broader clinical trajectory involving hydrocephalus, EVD-marked severe care, inflammatory response, and metabolic stress. The separate edges in this pathway did not all have the same level of support, and the pathway should not be read as a fully established causal chain. However, the repeated recovery of its main components gives a coherent structure for interpreting the vasospasm-related results and a clinically plausible hypothesis for further study.

Severity and early management

The clearest severity-to-intervention structure was the relation between clot burden, hydrocephalus, and EVD. This forms the upstream part of the main recovered pathway. The edge clot burden $\rightarrow$ hydrocephalus appeared across discovery procedures, and hydrocephalus $\rightarrow$ EVD was among the most stable management-related edges. This is clinically plausible. Greater hemorrhage burden can contribute to obstruction of fluid pathways and ventricular enlargement, while EVD placement is a clinical response to hydrocephalus or related ventricular complications.

The direction EVD $\rightarrow$ hydrocephalus was explicitly forbidden, because it would contradict the intended clinical and temporal interpretation of these variables. However, the adjacency between hydrocephalus and EVD was not required. Its repeated recovery therefore acts as a useful sanity check: the algorithms identified a known severity-to-management relation without being forced to include it.

Hydrocephalus was also repeatedly linked with nimodipine, although the direction varied across algorithms. PC-stable often oriented the relation as hydrocephalus $\rightarrow$ nimodipine, whereas FGES and FCIT more often reversed it. This instability is understandable. Hydrocephalus, EVD, ICU monitoring, aneurysm treatment, and nimodipine documentation all belong to the early management trajectory of severe aSAH admissions. A direct edge between hydrocephalus and nimodipine may therefore reflect shared severity, care pathway, or documentation intensity. Because the orientation was not stable, this relation is better treated as part of the broader management structure than as an interpretable directed effect.

Inflammatory and metabolic disturbance

The next part of the main recovered pathway links EVD to leukocytosis, leukocytosis to hyperglycemia, and hyperglycemia to vasospasm. This section therefore covers the middle and downstream parts of the pathway:

EVD $\rightarrow$ Leukocytosis $\rightarrow$ Hyperglycemia $\rightarrow$ Vasospasm.

Representative FCIT PAG with highlighted main path

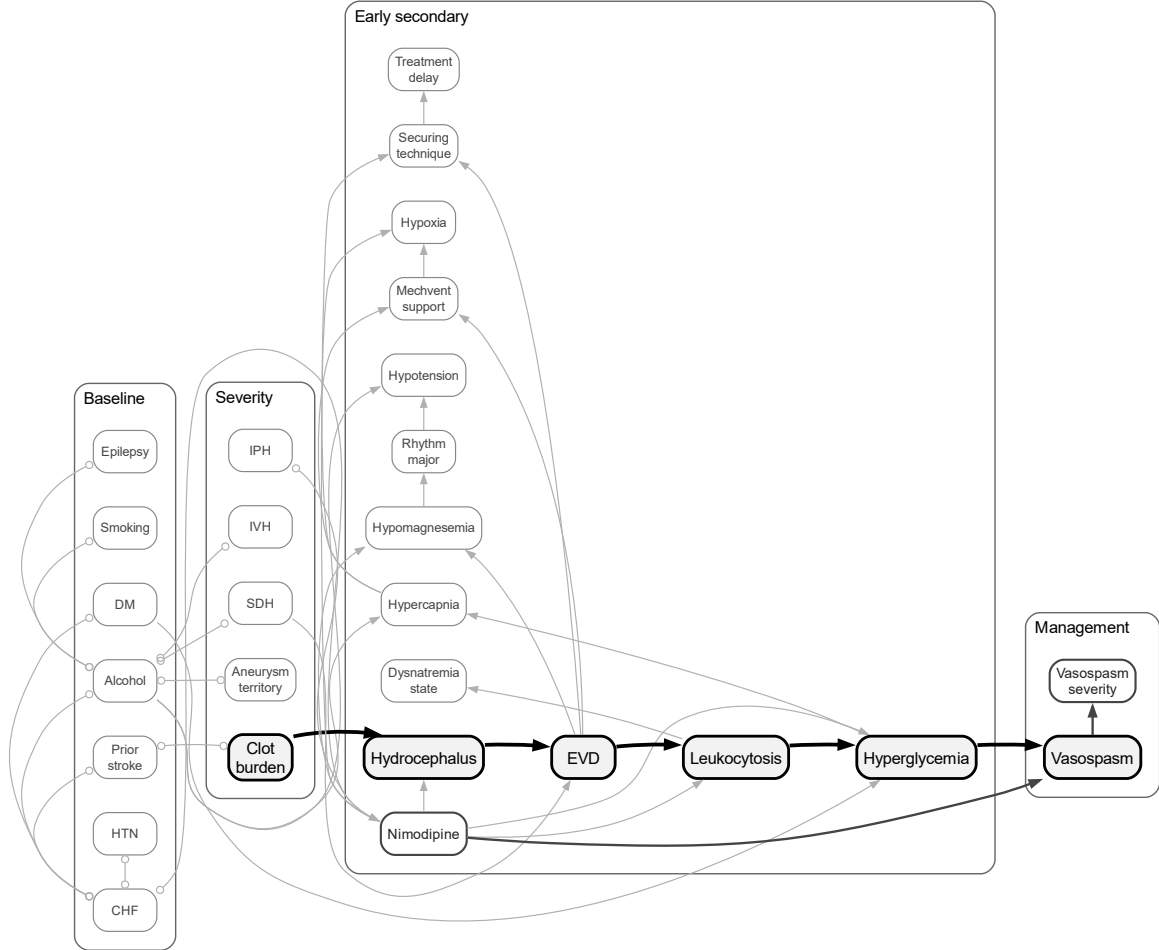


Figure 5.16: Representative constrained FCIT PAG output from the main sweep, using $\alpha = 0.001$, conditioning depth = 3, disc=3, block=3, starts=1, pag=1, G^2 minimum cell count = 1, G^2 table type = 2, BDeu sample prior = 10, and BDeu structure prior = 0. The bold path highlights the structure linking clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, and vasospasm. Lighter edges show additional relations from the same graph. This figure is used as a representative FCIT output; the interpretation in the text is based on recurrence across algorithms, hyperparameter settings, and stability analyses.

The edge $\text{EVD} \rightarrow \text{leukocytosis}$ appeared in approximately half of PC-stable runs, approximately two-thirds of FGES runs, and all FCIT runs. This makes it a recurrent structural association.

Several explanations are plausible. Patients with EVDs may undergo longer invasive monitoring, more infection workups, and more frequent laboratory testing. They may also be more severely ill at baseline, which increases both the probability of EVD placement and the probability of inflammatory abnormalities. The safe interpretation is therefore that EVD marks a severe-care subgroup with a stronger inflammatory or critical-care trajectory. The edge is robust as a graph feature, but it should not be read as a settled direct procedural effect.

The edge $\text{leukocytosis} \rightarrow \text{hyperglycemia}$ was one of the clearest systemic relations. It appeared in all FCIT runs, in approximately two-thirds of FGES runs, and in roughly half of PC-stable runs. When oriented, the direction was consistently $\text{leukocytosis} \rightarrow \text{hyperglycemia}$. This relation is clinically plausible. Systemic inflammation and physiological stress can contribute to stress hyperglycemia.

The edge $\text{hyperglycemia} \rightarrow \text{vasospasm}$ was one of the most important vasospasm-adjacent findings. It appeared in all FCIT runs and in most FGES runs, with an overall FGES run fraction around two-thirds across the full sweep. It was much less frequent in PC-stable. This difference between algorithms is informative. FGES and FCIT both supported hyperglycemia as a direct structural neighbour of vasospasm, while PC-stable often absorbed the relation into other severity, inflammatory, or workflow-related structures.

Clinically, this link is plausible, but it needs a careful interpretation. High blood sugar after a severe brain bleed does not necessarily mean that glucose itself is causing vasospasm. It may also reflect the body’s stress response, systemic inflammation, and the fact that some patients are more severely ill overall [57].

At the same time, the finding is not easily dismissed as only a general marker of severe illness. The causal discovery analysis already included several other variables that also capture severity, neurological injury, physiological instability, and treatment or monitoring intensity. Despite these competing explanations, hyperglycemia was one of the secondary physiological variables most consistently connected to vasospasm across the analyses. This suggests that the association may represent more than a broad distinction between mild and severe cases, at least within the measured variable set.

There are also possible biological reasons why high blood sugar could play a more active role. Hyperglycemia may affect the blood vessels by making them less able to relax, increasing signals that promote vessel narrowing, and supporting inflammation or small clotting processes. These changes could, in theory, make vasospasm more likely or more severe [33, 57].

The present analysis cannot separate a direct contribution from residual confounding or from unmeasured aspects of disease severity. A cautious interpretation is therefore that hyperglycemia emerged as a robust candidate in the clinical pattern around vasospasm.

The direct edge $\text{leukocytosis} \rightarrow \text{vasospasm}$ was less stable than the $\text{leukocytosis} \rightarrow \text{hyperglycemia}$ and $\text{hyperglycemia} \rightarrow \text{vasospasm}$ edges. It appeared in roughly half of PC-stable runs and in selected FGES regimes. In FGES, it was mainly recovered when the maximum degree was high and the BDeu sample prior was high. FCIT rarely placed leukocytosis directly adjacent to vasospasm.

This pattern suggests that leukocytosis was involved in the vasospasm-related structure, but that the analysis did not clearly determine whether this involvement was direct. A simple interpretation would be that leukocytosis marks patients who were more severely ill. However, as with hyperglycemia, this explanation is incomplete on its own because the variable set already included several other indicators of severity, hemorrhage burden, hydrocephalus, physiological instability, and treatment or monitoring intensity. Despite these competing severity-related variables, leukocytosis remained repeatedly connected to the vasospasm-related graph structure.

Leukocytosis also has a plausible link with vasospasm, but the idea should be explained

carefully. Blood vessels need to stay flexible. They must be able to widen and relax so that blood can keep flowing to the brain [35]. When many white blood cells are activated, they can disturb this balance in several ways. They may damage the inner lining of the blood vessels, reduce signals that normally help vessels relax, and increase signals that make vessels tighten [27, 75, 110]. They can also release reactive substances that irritate the vessel wall and interact with platelets, which may contribute to very small clotting processes [40, 101]. Together, these changes can create a vascular environment in which blood vessels are more likely to narrow too much.

The stronger and more stable leukocytosis $\rightarrow$ hyperglycemia edge also has a plausible interpretation. High white blood cell counts release alarm chemicals that act like a blockade, stopping your body's natural keys (insulin) from letting sugar into your cells for energy. These same chemicals trick your body into an emergency mode, which forces your liver to dump even more sugar into your blood. This creates a sugar traffic jam where your levels get stuck way too high because your body is producing extra sugar while being unable to use it [6, 115]. In this reading, leukocytosis may capture part of an inflammatory-stress response that feeds into hyperglycemia, while hyperglycemia forms the more consistently recovered local neighbour of vasospasm.

Other potential findings

The edge nimodipine $\rightarrow$ vasospasm was highly recurrent, appearing in approximately half of PC-stable runs and in all FGES and FCIT runs. This edge should not be interpreted as evidence that nimodipine causes vasospasm. Clinically, nimodipine is routinely given after aSAH to reduce the risk of poor outcome from delayed ischemic complications, so its presence is closely linked to recognition of aSAH, monitoring intensity, and concern for vasospasm. The learned direction is therefore more plausibly explained by the care pathway and documentation process around suspected vasospasm. In addition, because nimodipine is expected in standard aSAH care, apparent non-use may mark atypical or very severe admissions, early deterioration, limited treatment opportunity, transfer-related documentation gaps, or incomplete medication capture in the structured data. For this reason, the recurrent nimodipine-vasospasm relation was treated as a care-process signal rather than as a direct pharmacological causal effect.

Other physiologic instability variables, including hypotension, hypoxia, hypomagnesemia, and dysnatremia, appeared in some vasospasm-adjacent structures, especially in PC-stable under shallower or more permissive settings. These variables are clinically relevant in aSAH and ICU care, and in discussion with a clinical expert they might even have been expected to appear more often as causal parents of vasospasm than leukocytosis or hyperglycemia. For example, impaired oxygen delivery, low blood pressure, electrolyte disturbance, or low magnesium could plausibly affect cerebral perfusion, vascular tone, or tolerance to delayed ischemic injury. However, in the present discovery results these relations were less stable across algorithms, hyperparameter settings, and bootstrap analyses than the leukocytosis-hyperglycemia-vasospasm structure. This does not rule them out as clinically meaningful contributors. Their weaker recovery may reflect limited sample size, admission-level temporal aggregation, measurement heterogeneity, or the fact that these disturbances occur in more complex ICU contexts. Importantly, they were not absent from the graphs. They appeared in some vasospasm-adjacent structures and may become more consistently recoverable in a larger dataset with more precise timing and richer physiological detail.

Variables not recovered as stable direct parents of vasospasm

The recurrent structures are also informative for what they did not show. Several variables that are clinically important in aSAH were not recovered as stable direct parents of vasospasm. This is especially relevant for clot burden. Hemorrhage burden is one of the most established

risk factors associated with vasospasm in the clinical literature [34, 50]. However, in the learned graphs, clot burden was represented more consistently as an upstream severity variable than as a direct vasospasm parent.

Clot burden was repeatedly connected to hydrocephalus, hydrocephalus was strongly connected to EVD, and EVD was then linked to leukocytosis and hyperglycemia. In this representation, clot burden belongs to the severe-care trajectory around vasospasm, but the algorithms did not consistently preserve a separate direct clot burden $\rightarrow$ vasospasm edge after other related variables were included.

The same caution applies to baseline variables. Age and sex were allowed to point into later clinical variables because they precede the admission course, but they were not recovered as stable direct vasospasm parents. Age appeared in some PC-stable graphs under more permissive settings, but this edge disappeared under stricter configurations and was not treated as a core finding. This suggests that the learned vasospasm structure was not dominated by simple baseline demographic effects. Instead, the more reproducible vasospasm-adjacent variables were secondary physiological and care-process variables, especially hyperglycemia, leukocytosis, nimodipine, and monitoring-related pathways.

Structure	Recurrence across methods	Sensitivity
<i>Main recovered pathway</i>		
Clot burden $\rightarrow$ hydrocephalus $\rightarrow$ EVD $\rightarrow$ leukocytosis $\rightarrow$ hyperglycemia $\rightarrow$ vasospasm	High for the upstream and middle segments; strongest completion in FGES and FCIT	Low-moderate for the backbone; moderate near vasospasm
<i>Component structures</i>		
Clot burden $\rightarrow$ hydrocephalus $\rightarrow$ EVD	High	Low-moderate
EVD $\rightarrow$ leukocytosis	Moderate-high	Moderate
Leukocytosis $\rightarrow$ hyperglycemia	High	Low
Hyperglycemia $\rightarrow$ vasospasm	High in FGES and FCIT; lower in PC-stable	Moderate
Leukocytosis $\rightarrow$ vasospasm	Partial	High
Nimodipine $\rightarrow$ vasospasm	Very high	Low

Table 5.4: Summary of recurrent causal discovery structures and their robustness across methods and sensitivity analyses.

5.5.3 Robustness of the main structures

The recurrent structures were evaluated using the stability and sensitivity outputs described in Section 4.4. Edges were treated as more robust when they recurred across resampling, hyperparameter settings, or variable-set changes, especially when their orientation was also stable.

Bootstrap and run-fraction stability

The stability analysis largely supported the distinction made above between backbone structures and more fragile vasospasm-adjacent edges. In PC-stable, hydrocephalus $\rightarrow$ EVD showed near-complete adjacency and orientation stability, while leukocytosis $\rightarrow$ hyperglycemia had high adjacency stability, approximately 0.9, and very high directional consistency, approximately 0.98. In FCIT, the same inflammatory-metabolic structure remained stable, together with EVD $\rightarrow$ leukocytosis, hyperglycemia $\rightarrow$ vasospasm, and nimodipine $\rightarrow$ vasospasm.

FGES showed a similar pattern across its hyperparameter grid. Nimodipine $\rightarrow$ vasospasm appeared in every run, while hyperglycemia $\rightarrow$ vasospasm appeared mainly under low or mod-

erate BDeu sample priors and moderate maximum degree settings. The direct leukocytosis $\rightarrow$ vasospasm edge was more configuration-dependent, appearing mainly under high-degree or high-prior settings.

The vasospasm-adjacent findings were less uniform. Hyperglycemia was the most consistent systemic neighbour of vasospasm in FGES and FCIT, leukocytosis appeared either directly or through hyperglycemia, and nimodipine was highly stable but best interpreted as a treatment and documentation signal.

Hyperparameter sensitivity

The hyperparameter sweeps mainly affected the density of the graphs and the exact local neighbourhood of vasospasm. In PC-stable, more permissive α values and shallower conditioning depths produced additional vasospasm-adjacent edges, including age $\rightarrow$ vasospasm, dysnatremia $\rightarrow$ vasospasm, and hypomagnesemia $\rightarrow$ vasospasm. These edges disappeared under stricter settings and were therefore not treated as core findings. By contrast, leukocytosis $\rightarrow$ hyperglycemia was more persistent, including under stricter PC-stable configurations.

FGES sensitivity was driven mostly by maximum degree and the BDeu sample prior. Hyperglycemia $\rightarrow$ vasospasm was absent when the maximum degree was too restrictive, appeared most clearly at moderate degree settings, and disappeared under the highest sample prior. Under that stronger prior, FGES instead recovered leukocytosis $\rightarrow$ vasospasm, suggesting that the score-based search alternated between a more detailed inflammatory-metabolic path and a simpler direct inflammatory edge.

FCIT was comparatively insensitive to the tested settings. Hyperglycemia $\rightarrow$ vasospasm, nimodipine $\rightarrow$ vasospasm, leukocytosis $\rightarrow$ hyperglycemia, and EVD $\rightarrow$ leukocytosis appeared across essentially all FCIT configurations. Overall, the hyperparameter analysis supports the main backbone relations and the hyperglycemia-vasospasm link, while treating direct leukocytosis $\rightarrow$ vasospasm and the weaker physiologic neighbours as more configuration-dependent.

Early-window definitions of leukocytosis and hyperglycemia

As an additional robustness check, early-window versions of leukocytosis and hyperglycemia were also constructed using only measurements from the first 72 hours after admission. This was done because the main discovery variables were admission-level summaries, while vasospasm usually develops later in the hospital course. If the leukocytosis-hyperglycemia-vasospasm structure had been driven mainly by abnormalities occurring after vasospasm was already recognised, the early-window versions would be expected to behave differently.

In practice, the first-72-hour leukocytosis and hyperglycemia variables had very similar distributions to the full-admission versions. This suggests that, for most admissions, the relevant inflammatory and metabolic signal was already visible early in the admission and was not only a late consequence of the documented vasospasm course. When these early-window variables were used in the causal discovery workflow, the resulting graphs were also very similar to the main admission-level graphs.

This sensitivity analysis does not solve the temporal limitations of the admission-level design, since exact timing relative to vasospasm could still not be represented for every patient. However, it reduces the concern that the main inflammatory-metabolic findings were entirely produced by late measurements after vasospasm documentation. The similarity between the early-window and full-admission analyses supports treating leukocytosis and hyperglycemia as robust systemic signals in the recovered vasospasm-related structure.

Removing the constraints

As an additional sensitivity analysis, FCIT was also run without the temporal blacklist and without bootstrapping. This run was not intended to replace the constrained analysis, but to

check how much of the FCIT interpretation depended on the imposed clinical ordering.

Some structures remained recognizable. Clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, securing technique, nimodipine, and vasospasm still formed a connected severity-management-systemic disturbance region of the graph. EVD was again linked to leukocytosis and hyperglycemia, and vasospasm remained closely connected to management-related variables. In the unconstrained sweep, vasospasm had stable incoming edges from EVD and securing technique.

Importantly, removing the blacklist did not turn vasospasm into an upstream causal parents of any new nodes. This is relevant because the variable construction could not fully encode whether abnormalities such as leukocytosis or hyperglycemia occurred before or after vasospasm. If the unconstrained graph had consistently oriented vasospasm $\rightarrow$ leukocytosis or vasospasm $\rightarrow$ hyperglycemia, this would have raised a stronger concern that the constrained inflammatory-metabolic interpretation was mainly imposed by the temporal blacklist. That was not observed. In the unconstrained sweep, vasospasm did not become an ancestor of leukocytosis or hyperglycemia. The only stable outgoing vasospasm edges were vasospasm $\rightarrow$ vasospasm severity and vasospasm $\rightarrow$ hypomagnesemia. The first is expected, because vasospasm severity is derived from the presence and documented strength of vasospasm evidence. The second is clinically harder to interpret as a direct causal effect and is better treated as an implausible local orientation. Overall, vasospasm still behaved mainly as a downstream outcome-related node, rather than as a cause of the main systemic disturbance variables.

The unconstrained sweep did, however, produce several orientations that are difficult to defend clinically. For example, EVD $\rightarrow$ hydrocephalus appeared in all runs, even though EVD placement is normally a response to hydrocephalus or ventricular complications, not their cause. Other stable but implausible directions included nimodipine $\rightarrow$ clot burden, hyperglycemia $\rightarrow$ diabetes, dysnatremia state $\rightarrow$ sex, and EVD $\rightarrow$ smoking. These edges illustrate why repeated recovery alone is not sufficient for interpretation when temporal background knowledge is removed.

The unconstrained analysis therefore does not add a new substantive causal finding. Its main contribution is methodological. It shows that FCIT can recover a similarly sized graph without constraints, but that some stable orientations then violate the clinical ordering of the variables. For this reason, the constrained FCIT sweep is retained as the primary analysis, while the unconstrained sweep supports the use of the blacklist as necessary clinical background knowledge.

5.5.4 Limitations of the causal discovery results

Several limitations are important for interpreting these results.

First, the analysis is observational and admission-level. The graph cannot fully reconstruct within-admission temporal ordering. Variables such as leukocytosis and hyperglycemia were derived from repeated measurements across the admission, but the causal discovery input does not encode the exact time at which each abnormality occurred relative to vasospasm onset.

Second, several variables are partly proxies for clinical workflow. EVD, nimodipine, securing technique, neuroimaging intensity, and vasospasm documentation all reflect care processes as well as patient biology. Edges involving these nodes may therefore capture treatment decisions, monitoring intensity, or documentation patterns.

Third, unmeasured confounding is likely. Infection, steroid exposure, vasopressor use, ventilation settings, fever, sedation, nutrition, glucose management protocols, aneurysm morphology, and detailed timing of vasospasm treatment were not fully represented as graph nodes. FCIT partially addresses this by allowing latent confounding in the graph representation, but it cannot identify all unmeasured mechanisms.

Chapter 6

Discussion and Conclusions

6.1 Discussion

The goal of this thesis was to build a reproducible pipeline for extracting and analysing multimodal clinical data around aSAH and vasospasm. The work combined structured variables from electronic health record tables with LLM-derived variables from clinical text, resulting in an admission-level dataset of 470 likely aSAH admissions. This dataset was then used for cohort description, cross-modal validation, and causal discovery across three algorithmic approaches.

A recurring issue throughout the thesis is that vasospasm is not represented by one clean variable in routine ICU data. It can appear as an ICD code, a radiology finding, a discharge-summary statement, a treatment decision, or a monitoring concern. Predictive models therefore might be useful in other cases, but it does not solve the clinical interpretation problem on its own. A model can learn associations with documented vasospasm, yet those associations may reflect disease severity, care intensity, surveillance, treatment, or documentation. The causal discovery analysis was used to explore this structure more directly.

6.1.1 Interpretation of cross-modal extraction results

One important part of Chapter 5 was the comparison between administrative coding and clinical text. The ICD-based vasospasm indicator identified vasospasm in 65 of 470 admissions, or 13.8%. The note-derived signals were much broader: discharge summaries were positive in 48.3% of scored admissions, and radiology reports were positive in 52.6%. This difference suggests that ICD coding captured only a limited part of the vasospasm evidence present in the records.

The narrative sources gave a more detailed view, although they also had their own limitations. Discharge summaries and radiology reports agreed well at the binary level, with agreement in 394 cases, or 86.8%. Still, the disagreements were useful to inspect. Radiology reports sometimes captured vessel-level findings at a specific time point, while discharge summaries gave a broader account of the admission. In other cases, discharge summaries described clinical vasospasm or vasospasm-directed treatment without equally explicit wording in the radiology bundle. These differences seemed to reflect timing, summarisation, and source perspective more often than simple extraction failure.

Agreement also depended on the type of variable. Variables that are usually documented explicitly, such as aneurysm location and treatment type, showed high cross-modal consistency. More interpretive variables, including vasospasm evidence, clot burden, hydrocephalus, EVD presence, and nimodipine exposure, showed more disagreement. This was not necessarily a weakness of the dataset. In several places, the disagreement itself helped explain how clinical information moved through the record.

Overall, the LLM extraction pipeline seems to have produced variables that were usable for downstream analysis, especially when supported by cross-modal checks. The results also

show that LLM-derived variables should not be treated as automatically correct. Their value in this thesis comes partly from the fact that they were compared across sources, inspected in disagreement cases, and kept linked to their provenance.

6.1.2 Interpretation of causal discovery results

The causal discovery analyses examined how baseline characteristics, hemorrhage severity, early complications, treatment-related variables, systemic physiological states, and vasospasm outcomes were connected in the admission-level data. Across PC-stable, FGES, and FCIT, the most interpretable recurrent structure was:

Clot burden $\rightarrow$ Hydrocephalus $\rightarrow$ EVD $\rightarrow$ Leukocytosis $\rightarrow$ Hyperglycemia $\rightarrow$ Vasospasm.

This pathway should be read as a plausible candidate structure. Clot burden was not mainly recovered as a stable direct parent of vasospasm. It appeared more often as the start of a broader severity and care trajectory, linking hemorrhage burden to hydrocephalus, EVD placement, inflammatory disturbance, metabolic stress, and finally vasospasm. The final hyperglycemia-vasospasm link was strongest in FGES and FCIT, and less consistent in PC-stable.

The upstream part of the pathway fits clinical expectations. Greater clot burden can contribute to hydrocephalus, and EVD placement is commonly a response to hydrocephalus, raised intracranial pressure, or related ventricular complications. Since the hydrocephalus-EVD adjacency was not forced, its repeated recovery provides a useful check that the algorithms captured a clinically expected relation.

The middle and downstream parts need more caution. EVD should probably be read as part of a severe-care trajectory, not as evidence that the drain itself directly caused leukocytosis. The leukocytosis-hyperglycemia edge was one of the more stable systemic relations and fits a stress-response interpretation, where inflammation and acute illness contribute to insulin resistance and increased glucose production. hyperglycemia was then the most consistent systemic neighbour of vasospasm in FGES and FCIT. This does not prove that glucose directly contributes to vasospasm, but it makes hyperglycemia a reasonable candidate for further study, especially because it remained connected to vasospasm despite the presence of other severity and workflow variables.

The direct leukocytosis-vasospasm edge was less stable. It appeared in some PC-stable runs and selected FGES settings, while FCIT more often placed leukocytosis upstream of hyperglycemia. This suggests that leukocytosis belongs to the vasospasm-related structure, although the analysis does not clearly determine whether its role is direct, indirect, or mainly a marker of severity.

The sensitivity analyses helped separate stable patterns from more fragile ones. The backbone involving clot burden, hydrocephalus, EVD, leukocytosis, and hyperglycemia was more robust than several other vasospasm-adjacent edges. Hypotension, hypoxia, dysnatremia, and hypomagnesaemia appeared mainly under more permissive settings and were not treated as core findings. The unconstrained FCIT sweep also supported the use of clinical background knowledge: without the temporal blacklist, some difficult orientations appeared, including EVD $\rightarrow$ hydrocephalus. At the same time, vasospasm did not become a broad upstream driver of leukocytosis or hyperglycemia, which somewhat reduces the concern that the inflammatory-metabolic interpretation was only imposed by the constraints.

6.1.3 Clinical expert feedback

The cohort construction, variable choices, and main causal discovery findings were discussed with Prof. Dr. Kirsten Colpaert, intensivist at Ghent University Hospital, head of the Data Science Institute, and an expert in vasospasms. This discussion helped place the results in a

more clinical context and was especially useful for identifying which parts of the pipeline seemed reasonable, which parts should be interpreted cautiously, and which variables would be worth improving in future work.

Overall, the construction of the database seemed clinically acceptable within the limits of MIMIC-IV. The use of ICD codes alone was confirmed to be unreliable for vasospasm, which is consistent with the findings in this thesis. The broader use of discharge summaries and radiology reports therefore made sense as a way to capture vasospasm evidence that is not consistently coded administratively. At the same time, the discussion made clear that day-to-day hospital notes would probably be more useful than discharge summaries, especially for making the analysis more temporal. Discharge summaries often compress the entire admission into a retrospective narrative, while daily notes could better show when vasospasm was suspected, when physiological disturbances appeared, and how treatment decisions evolved. Since these daily reports are not available in MIMIC-IV-Note, the choices made in this thesis were considered reasonable, but clearly limited by the available source material.

The general characteristics of the aSAH cohort also seemed plausible. For example, the high proportion of female patients and the large share of coiled aneurysms were not surprising for this type of population. This gave some reassurance that the cohort construction did not produce an obviously unusual aSAH sample. SDH was mentioned as a node that would not usually be expected to play an important role in vasospasm-related mechanisms. For future versions of the analysis, it may therefore be reasonable to remove or deprioritise this variable. Similar questions could be asked for IVH and IPH, although their usefulness may depend more on the exact modelling goal and on whether better severity measures are available.

A recurring point in the discussion was that established clinical severity scores would be very valuable. Fisher grade, WFNS grade, and Hunt-Hess grade would likely describe the initial hemorrhage severity and clinical condition more directly than several of the proxy variables used here. In the current MIMIC-IV setting, these scores had too much missingness to be used reliably in the discovery matrix. In a future dataset, especially one from a neurocritical-care setting with richer structured documentation, these scores would probably be among the first variables to add. The expert also indicated that some of the physiological thresholds used in the thesis, for example for hypotension, might be defined somewhat differently from a clinical perspective. The current thresholds were useful for reproducible feature construction, but they should be seen as pragmatic definitions rather than definitive clinical cut-offs.

The recurrent part of the graph linking clot burden, hydrocephalus, and EVD was considered clinically plausible. This fits the interpretation that this part of the discovered structure mostly reflects hemorrhage severity and the subsequent care pathway. Greater clot burden can contribute to hydrocephalus, and EVD placement is a common response in that clinical context. The fact that this relation appeared repeatedly is therefore reassuring.

The more exploratory part of the graph involved leukocytosis, hyperglycemia, and vasospasm. This part of the result was considered clinically interesting. Their position close to vasospasm in several discovered structures therefore provides a useful starting point for further analysis, even though the exact interpretation remains open. The expert also noted that, from a clinical perspective, variables such as hypotension, hypoxia, dysnatremia, and hypomagnesaemia could also be expected to play an important role as potential contributors to vasospasm. The fact that these variables were less prominent in the stable results may therefore be informative in itself. Their role may depend strongly on timing, duration, and clinical thresholds, which are difficult to capture in the current admission-level setup. In addition, some of these states are actively prevented or corrected in neurocritical care. Hypotension, sodium disturbances, and low magnesium may therefore leave a weaker observational signal precisely because clinicians monitor and treat them closely. Their weaker presence in the graph should therefore not be read as evidence that they are clinically unimportant. In that sense, the leukocytosis and hyperglycemia findings were seen as interesting enough to motivate more targeted follow-up

analyses.

The discussion also pointed to several variables that were not included but could be clinically important. Polyuria and hypovolemia were mentioned in particular. These were not included in the current discovery matrix because ICD codes did not capture them reliably. Still, they may be recoverable in future work from combinations of urine output, fluid balance, sodium measurements, medication data, and other structured values. This could be especially relevant because polyuria, hypovolemia, and dysnatremia may be part of a clinically meaningful pathway after aSAH that is only indirectly represented in the current analysis.

The EVD-leukocytosis relation was another point where the expert feedback led to a more cautious interpretation. EVD should probably not be read too directly as a biological cause of leukocytosis. It may mainly indicate a more severe care pathway, where patients with hydrocephalus, closer monitoring, invasive devices, and longer or more complex ICU stays also have more opportunities to develop inflammatory signs. A reverse or protective interpretation is also possible in some cases, since CSF drainage could theoretically reduce intracranial pressure or inflammatory burden. More generally, the relation between clot burden, EVD, and leukocytosis is likely influenced by variables that were not included in the current graph.

Taken together, the expert discussion supported the broad direction of the thesis. The database construction seemed reasonable given the available MIMIC-IV sources, and the clot burden-hydrocephalus-EVD part of the graph aligned with expected clinical reasoning. The leukocytosis and hyperglycemia findings were considered interesting enough to motivate further study, but they should remain hypothesis-generating. The discussion also strengthened the argument that future work should move toward a more temporal dataset, richer severity measures, and additional variables such as polyuria and hypovolemia.

6.2 Limitations

The findings of this thesis should be interpreted as hypothesis-generating. The recurrent pathway linking clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, and vasospasm was clinically plausible and appeared repeatedly across the analysis, but it does not prove a fully timed biological mechanism.

The main limitation is the admission-level design. Measurements, procedures, clinical events, and note-derived findings were summarised into one row per admission. This made multimodal integration possible, but it compressed time. For variables such as leukocytosis, hyperglycemia, and vasospasm, the exact sequence of events matters, and this sequence could only be captured indirectly.

Several variables also combine clinical biology with care processes. EVD placement, nimodipine administration, neuroimaging intensity, and vasospasm documentation are shaped by patient severity, monitoring, treatment decisions, and documentation practices. Edges involving these variables may therefore reflect workflow or surveillance as well as biological relations.

Unmeasured confounding is likely. Factors such as infection, steroid exposure, vasopressor use, fever, sedation, ventilation settings, glucose management, aneurysm morphology, and detailed vasospasm treatment were not fully represented. This is especially relevant for the inflammatory and metabolic part of the recovered structure.

The analysis was limited to MIMIC-IV and was not externally replicated. Documentation practices, medication recording, imaging protocols, and patient mix may differ across hospitals. In addition, the LLM extraction layer introduces uncertainty, although structured schemas, evidence quotes, cross-modal checks, and manual review were used to reduce this risk.

Finally, the graphs should not be read as treatment-effect estimates. They suggest candidate relations for further study, but they do not show whether changing glucose, EVD management, or monitoring strategy would alter vasospasm risk or clinical outcome.

6.3 Future work

Several follow-up directions would strengthen this work.

- **Temporal modelling.** The most direct next step would be to move beyond admission-level summaries. Time-resolved glucose, WBC, sodium, oxygenation, blood pressure, imaging, EVD placement, and vasospasm documentation could help distinguish early risk factors from later consequences. Dynamic Bayesian networks, temporal causal discovery, or carefully defined landmark analyses could be useful here.
- **Moving from causal discovery to causal inference.** The graphs in this thesis were used to generate and organise candidate structures, not to estimate causal effects. A natural next step would be to select specific relations from the discovered structures and study them with a dedicated causal inference design. For example, the hyperglycemia-vasospasm relation could be translated into a more explicit question about whether early glucose control is associated with later vasospasm risk, under a clearly defined time window and adjustment set. This would require specifying exposures, outcomes, eligibility criteria, confounders, and possible mediators before estimation. Methods such as propensity-score weighting, outcome regression, doubly robust estimation, or target trial emulation could then be used to test focused hypotheses suggested by the discovery analysis.
- **External validation.** Applying the pipeline to another hospital dataset, such as UZ Ghent data, would show which parts of the extraction and graph structure are specific to MIMIC-IV and which parts are more stable across settings. This would also make it possible to compare documentation practices and treatment capture across institutions.
- **Richer clinical variables.** The current variable set could be expanded with aneurysm morphology, infection markers, steroid exposure, vasopressor use, ventilation parameters, fever, sedation, nutrition, glucose treatment, and more detailed vasospasm therapy. These additions would help address some of the likely unmeasured confounding around the inflammatory and metabolic pathway.
- **More systematic validation of LLM outputs.** The manual review in this thesis was useful but limited in scale. A larger clinician-reviewed reference set would allow more precise evaluation of LLM extraction errors, especially for borderline vasospasm, hydrocephalus, and treatment variables. It would also make it possible to compare prompts, models, and uncertainty thresholds more systematically.
- **Targeted causal studies.** The recurrent hyperglycemia and leukocytosis findings could motivate more focused analyses. For example, future work could define time windows before vasospasm onset and examine whether inflammatory or metabolic states remain associated with later vasospasm after stronger temporal adjustment. Interventional questions, such as glucose-control strategies, would need a separate design and should not be inferred directly from the discovery graph.

6.4 Conclusion

This thesis developed a multimodal pipeline for studying vasospasm after aSAH in MIMIC-IV by combining structured EHR variables with LLM-derived information from discharge summaries and radiology reports, while keeping source provenance visible. Taken together, these findings answer the research questions by showing how the cohort and variables can be constructed, how vasospasm differs across modalities, which graph structures recur, and how these structures should be interpreted clinically. The work showed that a likely aSAH cohort can be constructed

from MIMIC-IV by starting from broad structured candidate signals and then using clinical text to separate likely aneurysmal SAH from broader SAH or aneurysm-care contexts. It also showed that structured EHR tables and clinical notes can be converted into admission-level variables that remain clinically interpretable, provided that source provenance, uncertainty states, and cross-modal disagreement are kept visible.

The results suggest that ICD coding alone gives a narrow view of vasospasm in this cohort, since discharge summaries and radiology reports captured substantially more evidence, although their disagreements also showed how much timing and documentation context matter. This directly answers the cross-modal part of the thesis: vasospasm did not appear as one stable EHR signal, but as a construct distributed across administrative coding, narrative summaries, imaging interpretation, treatment context, and clinical monitoring. Disagreement between these sources was therefore not only a data-quality problem, but also an informative part of how vasospasm is represented in routine care.

The causal discovery analysis highlights structures that recurred across algorithms, hyperparameter settings, and stability checks. The most coherent pattern linked clot burden, hydrocephalus, EVD, leukocytosis, hyperglycemia, and vasospasm. The upstream part of this pathway aligned well with clinical expectations about hemorrhage severity and hydrocephalus management, while the downstream inflammatory and metabolic segment was more exploratory, with hyperglycemia emerging as the most stable systemic neighbour of vasospasm in FGES and FCIT. Other recurrent relations, especially those involving nimodipine or EVD, were clinically meaningful but required a more cautious interpretation because they may reflect treatment, monitoring, documentation, or severity-related care pathways as much as biological mechanisms.

Overall, the findings support the broader motivation of the thesis: prediction alone is not enough for this type of clinical question, because a predictive model may recognise documented vasospasm without showing whether the variables behind that prediction reflect biology, severity, treatment, surveillance, or documentation. Multimodal causal discovery offered a way to organise these signals and generate clinically meaningful hypotheses, provided the results are interpreted cautiously. A natural next step would be to bring more temporal detail into the analysis, since vasospasm is a delayed process and the distinction between cause, marker, and consequence depends heavily on when each signal appears.

Chapter 7

Societal Reflection

Earlier chapters focused mainly on the technical and clinical parts of this work. This chapter takes a broader view and reflects on the societal questions that come with this type of research. The thesis uses sensitive health data, automated text extraction, and graph-learning methods in a domain where incorrect interpretation could have clinical consequences. The reflection is organised around ethics, equity, sustainability, international context, and the Sustainable Development Goals.

7.1 Ethical considerations and equity

Research with health data starts from a position of trust. Clinical records contain information about diagnoses, treatments, complications, and outcomes that patients did not generate for research purposes. In this thesis, the data came from MIMIC-IV, which is de-identified and publicly available for credentialed researchers [51]. This reduces many privacy risks, but it does not remove the broader ethical responsibility. A similar pipeline used inside a hospital would need careful governance, secure storage, clear rules about access, and transparency about how patient data are being processed. This becomes especially relevant when narrative notes are used, because free text can contain details that are more personal and context-rich than structured codes. The World Health Organization also stresses that health AI should be governed in a way that keeps stakeholders accountable to healthcare workers, communities, and the people whose health may be affected by the technology [114].

The use of LLMs adds another layer of responsibility. These models can help extract information that would otherwise remain locked in narrative text, but their outputs are still interpretations of documentation. They can miss negations, overread uncertain statements, or give a structured answer that looks more definite than the source text allows. In this thesis, this risk was partly addressed by keeping source provenance, comparing discharge and radiology outputs, and manually reviewing disagreement cases. Even so, the extracted variables should be seen as research signals, not as automatically reliable clinical labels. In a clinical setting, this kind of system would need stronger validation, clinician oversight, and clear communication of uncertainty.

Bias is also a major concern. Models trained or evaluated on historical data may reflect existing inequalities in healthcare access, documentation, diagnosis, and treatment. If certain patient groups are less represented, less frequently investigated, or documented differently, automated methods may learn these patterns as if they were clinical facts. The CDC describes health equity in AI as the fair distribution of health technologies and their benefits across groups, regardless of socioeconomic status, race, sex or gender, ethnicity, disability status, or geographic location [25]. Bias can enter through data collection, labelling, preprocessing, model development, and validation. Suggested safeguards include more diverse data, explainable outputs, and continued monitoring for unequal performance. For this thesis, that means the MIMIC-IV re-

sults should be interpreted as findings from one hospital system and one documentation culture. They should not be assumed to transfer directly to other hospitals or patient populations.

7.2 Environmental sustainability and green AI

The environmental side of clinical AI is easy to overlook, especially in a thesis where the immediate focus is on data quality and clinical interpretation. Still, AI-based research uses computational resources, and this has a material footprint. Large models, repeated model calls, hyperparameter sweeps, and storage of intermediate outputs all require energy. Recent work on AI in healthcare points to energy consumption, carbon emissions, freshwater use, and electronic waste as relevant concerns, especially when data centres and large-scale computation are involved [12].

The footprint of this thesis was relatively limited compared with training a large model from scratch. The LLM component mainly used prompt-based extraction with an existing model, and the causal discovery experiments were run on a modest-sized admission-level dataset. Still, the project illustrates a broader pattern in medical AI research: it is tempting to increase model size, repeat more runs, store more intermediate files, and process more text whenever this is technically possible. That does not mean these choices are always justified. A more sustainable approach would weigh expected scientific value against computational cost. Practical steps include reusing intermediate outputs, avoiding unnecessary reruns, choosing smaller models when they are sufficient, reporting model and compute choices, and using infrastructure with lower environmental impact where possible. Energy monitoring, lifecycle assessment, low-carbon infrastructure, and circular use of hardware are all suggested as ways to reduce AI-related environmental harm in healthcare. More efficient algorithms and renewable-powered cloud infrastructure may also help reduce the footprint of AI workloads.

This connects directly to SDG 12 on responsible consumption and production and SDG 13 on climate action. In healthcare AI, sustainability should not be treated as separate from methodological design. A model that is only marginally better but much more resource-intensive may not be the best choice, especially when the clinical gain is uncertain.

7.3 Global context and international disparities

This thesis is based entirely on MIMIC-IV, a dataset from a single hospital system in the United States. MIMIC-IV is extremely valuable for reproducible research, but it reflects one healthcare system, one set of clinical workflows, one documentation style, and one patient population. The same pipeline might behave differently in Belgium, in other European hospitals, or in low-resource settings. Medication recording, ICU protocols, imaging availability, discharge-summary structure, and coding practice can all change how variables appear in the data.

This becomes more important when considering the global use of AI in healthcare. Many AI tools are developed using data from high-income countries, while the settings that might benefit from better access to expertise and decision support often face infrastructure, staffing, and data-quality constraints. A recent review of AI health applications in low-resource countries described economic and infrastructure barriers, data interoperability problems, and the need for regulation and surveillance of AI health applications [48]. It also noted that systems trained on high-income-country data may need recalibration before being used elsewhere. Barriers such as poor infrastructure, limited local expertise, and weak regulation can limit adoption, even when AI could support telemedicine or remote diagnostics.

For this thesis, the implication is modest but important: the pipeline should be seen as a research framework that needs local validation. Applying it to UZ Ghent or another hospital would not only test performance; it would also reveal how much of the structure is shaped by MIMIC-IV-specific documentation and care pathways. International use would require even

more care, with local clinicians involved in adapting definitions, validating extraction outputs, and deciding which questions are clinically meaningful.

7.4 Alignment with the Sustainable Development Goals

The SDGs offer a useful, although broad, framework for this reflection. The closest link is with SDG 3, which focuses on health and well-being. Better understanding of vasospasm after aSAH could eventually support improved monitoring, earlier recognition, or more focused research into modifiable risk factors. The contribution of this thesis is still far from direct clinical impact, since it is exploratory and based on retrospective data.

SDG 9, which concerns innovation and infrastructure, is also relevant. Reproducible pipelines, transparent variable construction, and open scientific methods can support more reliable digital health research. At the same time, technical innovation should not become an end in itself. In this thesis, causal discovery was useful because it helped organise a difficult clinical problem, not because the output graph should be treated as a final answer. This distinction matters socially, because overconfident use of AI methods in healthcare can reduce trust and may lead to poor decisions.

SDG 10, reducing inequalities, is relevant because AI systems can either reduce or reinforce existing disparities. If multimodal AI methods are validated only in well-resourced hospitals, their benefits may remain concentrated in those settings. If they are deployed without attention to bias, they may perform less well for groups who are already underserved. Finally, SDG 12 and SDG 13 are relevant because computational research has environmental costs. Responsible health AI should therefore consider who benefits, who carries the risks, and what resources are consumed.

7.5 Reflections and future directions

From a societal perspective, this thesis should be understood as an exploratory research contribution, not as a deployable clinical tool. The pipeline may help structure complex EHR data and generate hypotheses about vasospasm, but the results remain dependent on retrospective documentation, LLM extraction quality, modelling choices, and the characteristics of MIMIC-IV. Future work should therefore include external validation, clinician-adjudicated reference sets, and explicit checks for bias across patient groups. It would also be useful to report computational choices more transparently and to consider whether smaller or more efficient models can achieve similar extraction quality. If this line of research is extended toward clinical use, patients, clinicians, data stewards, and hospital governance bodies should be involved early. That would make the work less purely technical and more accountable to the people and systems it may eventually affect.

References

- [1] A. Abdulazim, M. Heilig, G. J. Rinkel, and N. Etminan. Diagnosis of delayed cerebral ischemia in patients with aneurysmal subarachnoid hemorrhage and triggers for intervention. *Neurocritical Care*, 2023.
- [2] H. Adam, J. Lin, J. Lin, H. Keenan, A. Wilson, and M. Ghassemi. Clinical information extraction with large language models: A case study on organ procurement. *AMIA Annual Symposium Proceedings*, 2024:115–123, May 2025.
- [3] A. J. Admon, A. M. Bohnert, C. R. Cooke, and S. P. Taylor. Beyond confounding: Identifying selection bias in observational pulmonary and critical care research. *Annals of the American Thoracic Society*, 2022.
- [4] D. Agniel, I. S. Kohane, and G. M. Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *British medical journal*, 2018.
- [5] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag. Large language models are few-shot clinical information extractors. *Conference on Empirical Methods in Natural Language Processing*, 2022.
- [6] I. Al-Hassani, N. A. Khan, E. Elmenyar, A. Al-Hassani, S. Rizoli, H. Al-Thani, and A. El-Menyar. The interaction and implication of stress-induced hyperglycemia and cytokine release following traumatic injury: A structured scoping review. *Diagnostics*, 2024.
- [7] M. Aleksandrowicz, E. Kozniewska, M. Aleksandrowicz, and E. Kozniewska. Hyponatremia as a risk factor for microvascular spasm following subarachnoid hemorrhage. *Experimental Neurology*, 2022.
- [8] E. Alsentzer, J. R. Murphy, W. Boag, W. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. Publicly available clinical bert embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019.
- [9] E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, and D. Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *medRxiv (Cold Spring Harbor Laboratory)*, 2024.
- [10] U. Athiraman, A. J. Norris, K. Jayaraman, A. Lele, R. Kentner, P. Singh, O. Higo, G. Zipfel, and R. Dhar. Intraoperative blood pressure and carbon dioxide values during aneurysmal repair and the outcomes after aneurysmal subarachnoid hemorrhage. *Journal of Clinical Medicine*, 2023.
- [11] R. Banerjee, B. Andrews, and E. Kummerfeld. An extensive simulation study evaluating the interaction of resampling techniques across multiple causal discovery contexts. *arXiv.org*, 2025.

- [12] E. Bignami, L. J. Darhour, and V. Bellini. Sustainable ai in medicine: Navigating innovation, challenges, and environmental impact. *Health Economics Review*, 15(1):110, Dec. 2025.
- [13] P. Bose, S. Srinivasan, W. Sleeman, J. Palta, R. Kapoor, and P. Ghosh. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 2021.
- [14] I. Buce-Satoba, D. Rozkalne, B. Mamaja, G. Krūmiņa, and A. Ozoliņa. Leukocytosis and c-reactive protein may predict development of secondary cerebral vasospasm in patients with aneurysmal subarachnoid hemorrhage. 2022.
- [15] S. Byeon and W. Lee. Directed acyclic graphs for clinical research: a tutorial. *Journal of Minimally Invasive Surgery*, 2023.
- [16] L. Capiiau, W. Pype, and H. Vanoverschelde. The epidemiology, treatment, complications and outcome of subarachnoid haemorrhages in uz gent: A retrospective cohort study. Master of medicine dissertation, Ghent University, Ghent, Belgium, 2019. Supervisor: Prof. dr. Kirsten Colpaert.
- [17] A. Casey, E. Davidson, M. T. C. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, H. Wu, and B. Alex. A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, 2021.
- [18] S. Chavali, J. R. Seshan, and G. P. Rath. Management of cerebral vasospasm after aneurysmal subarachnoid hemorrhage: An update. *Journal of Neuroanaesthesiology and Critical Care*, 2025.
- [19] L. Chen, Y. Meng, Q. Xue, Y. Zhao, X. Zhou, K. Hu, and H. He. Risk factors of shunt-dependent hydrocephalus after subarachnoid hemorrhage: a systematic review and meta-analysis based on observational cohort studies. *Neurosurgical review*, 2024.
- [20] A. V. Ciurea, C. Palade, D. C. Voinescu, and D. A. Nica. Subarachnoid hemorrhage and cerebral vasospasm – literature review. *Journal of Medicine and Life*, 2013.
- [21] J. Claassen, G. L. Bernardini, K. T. Kreiter, J. E. Bates, Y. Du, D. Copeland, E. S. Connolly, and S. A. Mayer. Effect of cisternal and ventricular blood on risk of delayed cerebral ischemia after subarachnoid hemorrhage:: The fisher scale revisited. *Stroke*, 2001.
- [22] J. Claassen and S. Park. Spontaneous subarachnoid haemorrhage. *The Lancet*, 400(10355):846–862, 2022.
- [23] B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv: Methodology*, 2020.
- [24] D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 2014.
- [25] I. Dankwa-Mullan. Health equity and ethical considerations in using artificial intelligence in public health and medicine. *Preventing Chronic Disease*, 21:E64, Aug. 2024.
- [26] J. B. C. de Andrade, N. de Lima Froio, P. S. Nascimento, S. O. Querobin, N. S. de Oliveira, and T. S. Carneiro. Cerebral vasospasm detection and delayed cerebral ischemia prediction after aneurysmal subarachnoid hemorrhage: A scoping review. *Neurocritical Care*, apr 2026. Epub ahead of print.

- [27] K. J. Diehl, B. R. Weil, J. J. Greiner, B. L. Stauffer, and C. A. DeSouza. White blood cell count and endothelin-1 vasoconstrictor tone in middle-aged and older adults. *Artery Research*, 6(2):65–70, June 2012.
- [28] J. C. Digitale, M. J. N. Martin, and M. Glymour. Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*, 2022.
- [29] W. Dodd, D. Laurent, A. S. Dumont, D. Hasan, P. Jabbour, R. M. Starke, K. Hosaka, A. Polifka, B. L. Hoh, and N. Chalouhi. Pathophysiology of delayed cerebral ischemia after subarachnoid hemorrhage: A review. *Journal of the American Heart Association : Cardiovascular and Cerebrovascular Disease*, 2021.
- [30] M. Doutreligne, T. Struja, J. Abecassis, C. Morgand, L. A. Celi, and G. Varoquaux. Step-by-step causal analysis of ehers to ground decision-making. *PLOS Digital Health*, 2025.
- [31] Elizabeth2424. Vasoconstriction and vasodilation.png. Wikimedia Commons, 2012. Licensed under CC BY-SA 3.0. Accessed May 2, 2026.
- [32] L. Evans, A. Rhodes, W. Alhazzani, M. Antonelli, C. Coopersmith, C. French, F. Machado, L. McIntyre, M. Ostermann, H. Prescott, C. Schorr, S. Simpson, W. Wiersinga, F. Alshamsi, D. Angus, Y. Arabi, L. Azevedo, R. Beale, G. Beilman, E. Belley-Cote, L. Burry, M. Cecconi, J. Centofanti, A. O. C. Yataco, J. D. D. Waele, R. Dellinger, K. Doi, B. Du, E. Estensoro, R. Ferrer, C. Gomersall, C. Hodgson, M. Møller, T. Iwashyna, S. T. Jacob, R. Kleinpell, M. Klompas, Y. Koh, A. Kumar, A. Kwizera, S. Lobo, H. Masur, S. Mcgloughlin, S. Mehta, Y. Mehta, M. Mer, M. Nunnally, S. Oczkowski, T. Osborn, E. Papathanassoglou, A. Perner, M. Puskarich, J. Roberts, W. Schweickert, M. A. Seckel, J. Sevransky, C. Sprung, T. Welte, J. Zimmerman, and M. Levy. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Medicine*, 2021.
- [33] L. E. Fernández-Garza, V. A. Fernández-Garza, D. Mares-Custodio, V. Gutiérrez-Ruano, A. Navarrete-Rodríguez, and J. J. Arias-Alzate. Glucose/potassium ratio, a novel biomarker for the prognosis of patients with subarachnoid hemorrhage: A review. *Journal of Vascular Diseases*, 4(4):48, 2025.
- [34] J. M. Findlay, J. Nisar, and T. E. Darsaut. Cerebral vasospasm: A review. *Canadian Journal of Neurological Sciences*, 2016.
- [35] B. Franczyk, J. Dybiec, W. Frąk, J. Krzemińska, J. Kućmierz, E. Młynarska, M. Szlagor, M. Wronka, and J. Rysz. Cellular mechanisms of coronary artery spasm. *Biomedicines*, 2022.
- [36] M. O. Gani, S. Kethireddy, R. Adib, U. Hasan, P. M. Griffin, and M. Adibuzzaman. Structural causal model with expert augmented knowledge to estimate the effect of oxygen therapy on mortality in the icu. *Artificial Intelligence in Medicine*, 2023.
- [37] Google Cloud. Vertex AI and zero data retention. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/vertex-ai-zero-data-retention>, 2026.
- [38] A. Gragossian, K. Bashir, B. S. Bhutta, et al. Hypomagnesemia. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2026. [Updated 2023 Nov 30].
- [39] Y. Han, B. Zhang, X. Qi, G. Yuan, X. Li, G. Hao, and G. Liang. Comparison of sex differences on outcomes after aneurysmal subarachnoid hemorrhage: a propensity score-matched analysis. *BMC Neurology*, 2024.

- [40] P. He, M. A. H. Talukder, and F. Gao. Oxidative stress and microvessel barrier dysfunction. *Frontiers in Physiology*, 11:472, May 2020.
- [41] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. Available online at <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>.
- [42] M. A. Hernán. The c-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, 2018.
- [43] J. Holmes, J. Beinlich, M. Boland, K. Bowles, Y. Chen, T. Cook, G. Demiris, M. Draugelis, L. Fluharty, P. Gabriel, R. Grundmeier, C. Hanson, D. Herman, B. Himes, R. Hubbard, C. E. Kahn, D. Kim, R. Koppel, Q. Long, N. Mirkovic, J. S. Morris, D. Mowery, M. Ritchie, R. Urbanowicz, and J. W. Moore. Why is the electronic health record so challenging for research and clinical care? *Methods of information in medicine*, 2021.
- [44] Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, K. Roberts, and H. Xu. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 2024.
- [45] J. Huang, D. M. Yang, R. Rong, K. Nezafati, C. Treager, Z. Chi, S. Wang, X. Cheng, Y. Guo, L. J. Klesse, G. Xiao, E. D. Peterson, X. Zhan, and Y. Xie. A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj digital medicine*, 2024.
- [46] K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019. Version 3, last revised November 29, 2020.
- [47] M. Huber. An introduction to causal discovery. *Zeitschrift für schweizerische Statistik und Volkswirtschaft/Schweizerische Zeitschrift für Volkswirtschaft und Statistik/Swiss journal of economics and statistics*, 2024.
- [48] S. A. Hussain, M. Bresnahan, and J. Zhuang. Can artificial intelligence revolutionize healthcare in the global south? a scoping review of opportunities and challenges. *Digital Health*, 11:20552076251348024, June 2025.
- [49] G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, 2015.
- [50] T. Inagawa, K. Yahara, and N. Ohbayashi. Risk factors associated with cerebral vasospasm following aneurysmal subarachnoid hemorrhage. *Neurologia Medico-chirurgica*, 2014.
- [51] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, and L. A. Celi. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 2023.
- [52] E. Kim, S. M. Rubinstein, K. T. Nead, A. P. Wojcieszynski, P. Gabriel, and J. L. Warner. The evolving use of electronic health records (EHR) for research. *Seminars in Radiation Oncology*, 2019.
- [53] M. Kim, C. Roupheal, J. McMichael, N. Welch, and S. Dasarathy. Challenges in and opportunities for electronic health record-based data analysis and interpretation. *Gut and Liver*, 2023.

- [54] Y. Kim, H. Jeong, S. Chen, S. S. Li, M. Lu, K. Alhamoud, J. Mun, C. Grau, M. Jung, R. R. Gameiro, L. Fan, E. Park, T. Lin, J.-H. Yoon, W. Yoon, M. Sap, Y. Tsvetkov, P. P. Liang, X. Xu, X. Liu, D. McDuff, H. Lee, S. R. Tulebaev, and C. Breazeal. Medical hallucinations in foundation models and their impact on healthcare. *arXiv.org*, 2025.
- [55] M. Kocaoglu. Characterization and learning of causal graphs with small conditioning sets. *Neural Information Processing Systems*, 2023.
- [56] M. Kruse, S. Hu, N. Derby, Y. Wu, S. Stonbraker, B. Yao, D. Wang, E. Goldberg, and Y. Gao. Large language models with temporal reasoning for longitudinal clinical summarization and prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20715–20735, Nov. 2025.
- [57] N. D. Kruyt, G. J. Biessels, J. H. DeVries, M. J. Luitse, M. Vermeulen, G. J. Rinkel, W. P. Vandertop, and Y. B. Roos. Hyperglycemia in aneurysmal subarachnoid hemorrhage: a potentially modifiable risk factor for poor outcome. *Journal of Cerebral Blood Flow and Metabolism*, 2010.
- [58] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 2020.
- [59] I. Lefkowitz, S. Walsh, L. J. Blank, N. Jetté, and B. Kummer. Direct clinical applications of natural language processing in common neurological disorders: Scoping review. *JMIR Neurotechnology*, 2024.
- [60] Y. Li, R. M. Wehbe, F. Ahmad, H. Wang, and Y. Luo. A comparative study of pretrained language models for long clinical text. *J. Am. Medical Informatics Assoc.*, 2022.
- [61] S. Lin, R. H. H. Groenwold, H. B. Mehta, J. S. Kim, and J. B. Segal. Addressing missingness in predictive models that use electronic health record data. *Annals of Internal Medicine*, 2025.
- [62] L. Liu, V. Blake, M. Barman, B. Gallego, T. Churches, G. Kennedy, S. Y. Ooi, G. P. Delaney, and L. Jorm. Using natural language processing to extract information from clinical text in electronic medical records for populating clinical registries: a systematic review. *Journal of the American Medical Informatics Association*, 33(2):484–499, Feb. 2026.
- [63] Z. Liu, T. Luo, H. Ma, J. Mo, X. Yang, Z. Huang, J. Li, and J. Kuang. Causal discovery in observational medical research: Scoping review. *JMIR Medical Informatics*, 14:e82499, 2026.
- [64] D. Machlanski, S. Samothrakis, and P. B. S. Clarke. Robustness of algorithms for causal structure learning to hyperparameter choice. *arXiv.org*, 2023.
- [65] V. Mank, W. Azhar, and K. Brown. Leukocytosis. <https://www.ncbi.nlm.nih.gov/books/NBK560882/>, 2024. [Updated 2024 Apr 21]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2026 Jan–.
- [66] N. I. Marupudi and S. Mittal. Diagnosis and management of hyponatremia in patients with aneurysmal subarachnoid hemorrhage. *Journal of clinical medicine*, 2015.
- [67] N. Meinshausen and P. Bühlmann. Stability selection. 2008.
- [68] M. Mesinovic, M. Buhlan, and T. Zhu. Causal graph neural networks for healthcare. *arXiv preprint arXiv:2511.02531*, 2026.

- [69] C. Milcent and C. Milcent. Bias due to re-used databases: Coding in hospital for extremely vulnerable patients. *HAL (Le Centre pour la Communication Scientifique Directe)*, 2023.
- [70] National Institutes of Health. Cerebral aneurysm nih.jpg. Wikimedia Commons, 2010. Public domain image; work of the U.S. federal government. Accessed May 2, 2026.
- [71] S. J. Nelson, Y. Yin, E. A. T. Rivera, Y. Shao, P. Ma, M. S. Tuttle, J. H. Garvin, and Q. Zeng-Treitler. Are icd codes reliable for observational studies? assessing coding consistency for data quality. *Digital Health*, 2024.
- [72] I. Olier, Y. Zhan, X. Liang, and V. Volovici. Causal inference and observational data. *BMC Medical Research Methodology*, 2023.
- [73] B. R. O’Driscoll, L. Howard, J. E. Earis, and V. Mak. Bts guideline for oxygen use in adults in healthcare and emergency settings. *Thorax*, 2017.
- [74] B. Petrungaro, N. K. Kitson, and A. C. Constantinou. Investigating potential causes of sepsis with bayesian network structure learning. *Applied Intelligence*, 55(496), 2025.
- [75] K. K. Ralota and J. Layland. Vasospastic angina: Pathophysiology, diagnosis, and emerging therapeutic approaches. *World Journal of Cardiology*, 17(10):108594, Oct. 2025.
- [76] J. Ramsey, B. Andrews, and P. Spirtes. Efficient latent variable causal discovery: Combining score search and targeted testing. *arXiv preprint arXiv:2510.04263*, 2025.
- [77] J. Ramsey, M. Glymour, R. Sánchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *Journal of data science*, 2017.
- [78] D. Rawat, P. Modi, and S. Sharma. Hypercapnia. <https://www.ncbi.nlm.nih.gov/books/NBK500012/>, 2023. [Updated 2023 Jul 24]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2026 Jan–.
- [79] S. Rehman, B. W. Sahle, R. V. Chandra, M. Dwyer, A. G. Thrift, M. L. Callisaya, M. Breslin, H. Phan, P. Otáhal, and S. Gall. Sex differences in risk factors for aneurysmal subarachnoid haemorrhage: Systematic review and meta-analysis. *Journal of the Neurological Sciences*, 2019.
- [80] D. Reichenpfader, H. Müller, and K. Denecke. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digital Medicine*, 2024.
- [81] R. Reátegui and S. Ratté. Comparison of metamap and ctakes for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making*, 2018.
- [82] M. Romoli, F. Giammello, M. G. Mosconi, A. D. Mase, G. D. Marco, A. Digiovanni, A. Ciacciarelli, R. Ornello, and B. Storti. Immunological profile of vasospasm after subarachnoid hemorrhage. *International Journal of Molecular Sciences*, 2023.
- [83] D. S. Rosen and R. Macdonald. Subarachnoid hemorrhage grading scales: a systematic review. *Neurocritical Care*, 2005.
- [84] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, apr 1983.
- [85] C. Roussos and A. Koutsoukou. Respiratory failure. *European Respiratory Journal*, 22(Suppl. 47):3s–14s, 2003.

- [86] D. Roustan and F. Bastardot. The clinicians’ guide to large language models: A general perspective with a focus on hallucinations. *Interactive Journal of Medical Research*, 2025.
- [87] T. Sarwar, S. Seifollahi, J. Chan, X. Zhang, V. Aksakalli, I. Hudson, K. Verspoor, and L. Cavedon. The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys*, 2023.
- [88] T. Savage, A. Nayak, R. Gallo, E. S. Rangan, and J. Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine*, 2024.
- [89] M. Scutari, C. E. Graafland, and J. Gutiérrez. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 2018.
- [90] J. Segura-Egea, D. Cabanillas-Balsera, M. Jiménez-Sánchez, and J. Martín-González. Endodontics and diabetes: association versus causation. *International Endodontic Journal*, 2019.
- [91] K. Shiba and T. Kawahara. Using propensity scores for causal inference: Pitfalls and tips. *Journal of Epidemiology*, 2021.
- [92] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 2018.
- [93] R. Shrestha, S. Rayamajhi, S. Shrestha, A. Thakali, S. Bishokarma, R. Shrestha, S. Rayamajhi, S. Shrestha, A. Thakali, and S. Bishokarma. Peripheral leukocytosis and clinical outcomes after aneurysmal subarachnoid hemorrhage. *Cureus*, 2022.
- [94] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318, Apr. 2024.
- [95] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318, Apr. 2024.
- [96] M. W. Sjoding, K. Luo, M. A. Miller, and T. J. Iwashyna. When do confounding by indication and inadequate risk adjustment bias critical care studies? a simulation study. *Critical Care*, 2015.
- [97] J. M. Smit, J. H. Krijthe, W. M. R. Kant, J. Labrecque, M. Komorowski, D. Gommers, J. v. Bommel, M. J. T. Reinders, and M. E. v. Genderen. Causal inference using observational intensive care unit data: a scoping review and recommendations for future practice. *npj Digital Medicine*, 2023.
- [98] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2001.
- [99] P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 2016.
- [100] SpirtesPeter. Introduction to causal inference. *Journal of Machine Learning Research*, 2010.

- [101] r. Steyers, C. M. and J. Miller, F. J. Endothelial dysfunction in chronic inflammatory diseases. *International Journal of Molecular Sciences*, 15(7):11324–11349, June 2014.
- [102] H. Stragier, H. Vandersmissen, S. Ordies, S. Thiessen, D. Mesotten, D. Peuskens, and H. Ten Cate. Pathophysiological mechanisms underlying early brain injury and delayed cerebral ischemia in the aftermath of aneurysmal subarachnoid hemorrhage: a comprehensive analysis. *Frontiers in Neurology*, 16:1587091, 2025.
- [103] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 2020.
- [104] M. Tayefi, P. D. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Computational Statistics*, 2021.
- [105] S. Thilak, P. Brown, E. J. Lawrence, G. Nambi, R. Mullhi, C. Chacko, Z. Ahmed, R. Chelvarajah, J. S. Pooni, and T. Veenith. Delayed cerebral ischaemia after subarachnoid haemorrhage: pathophysiology, diagnosis and neurotherapeutic opportunities. *Deleted Journal*, 2025.
- [106] P. Upadhyaya, K. Zhang, C. Li, X. Jiang, and Y. Kim. Scalable causal structure learning: Scoping review of traditional and deep learning algorithms and new opportunities in biomedicine. *JMIR Medical Informatics*, 2021.
- [107] K. N. Vanent, T. M. Leslie-Mazwi, C. C. Matouk, K. N. Sheth, M. R. Levitt, and G. J. Falcone. Hypertension is associated with earlier onset of nontraumatic subarachnoid hemorrhage in women: A mendelian randomization study. *Stroke*, 2025.
- [108] W. V. d. Ven, J. Schuurmans, and J. Schenk. . . . Monitoring, management, and outcome of hypotension in intensive care unit patients, an international survey of the european society of intensive care *Journal of Critical . . .*, 2022.
- [109] M. D. Vergouwen, M. Vermeulen, J. v. Gijn, G. J. Rinkel, E. F. M. Wijicks, J. P. Muizelaar, A. D. Mendelow, S. Juvela, H. Yonas, K. G. terBrugge, R. L. Macdonald, M. N. Diringer, J. P. Broderick, J. P. Dreier, and Y. B. Roos. Definition of delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage as an outcome event in clinical trials and observational studies: Proposal of a multidisciplinary research group. *Stroke*, 2010.
- [110] A. E. Walker, S. M. Seibert, A. J. Donato, G. L. Pierce, and D. R. Seals. Vascular endothelial function is related to white blood cell count and myeloperoxidase among healthy middle-aged and older adults. *Hypertension*, 55(2):363–369, Feb. 2010. Epub 2010 Jan 4.
- [111] M. Walsh, P. Devereaux, A. X. Garg, A. Kurz, A. Turan, R. Rodseth, J. B. Cywiński, L. Thabane, and D. I. Sessler. Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: Toward an empirical definition of hypotension. *Anesthesiology*, 2013.
- [112] T. S. Wettervik, H. Engquist, A. Hånell, T. Howells, E. Rostami, E. Ronne-Engström, A. Lewén, and P. Enblad. Cerebral blood flow and oxygen delivery in aneurysmal subarachnoid hemorrhage: Relation to neurointensive care targets. *Neurocritical Care*, 2022.
- [113] J.-P. Witte, R. Foraita, and V. Didelez. Multiple imputation and test-wise deletion for causal discovery with incomplete cohort data. *Statistics in Medicine*, 2021.
- [114] World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. World Health Organization, Geneva, 2021.

- [115] F. Xiu, M. Stanojcic, L. Diao, and M. G. Jeschke. Stress hyperglycemia, insulin treatment, and innate immune cells. *International Journal of Endocrinology*, 2014:486403, 2014.
- [116] Z. Yang, H. Yuan, R. Sayeed, A. L. M. Tan, E. Cai, M. Moro, X. Li, H. Ying, N. Brown, G. Weber, S. Yu, I. Kohane, and T. Cai. Clines: Clinical llm-based information extraction and structuring agent. *medRxiv*, 2025.
- [117] A. Zanga, E. Özkırmılı, F. Stella, A. Zanga, E. Ozkirimli, and F. Stella. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 2022.
- [118] J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 2008.
- [119] Y. Zhou, J. Shi, R. Stein, X. Liu, R. Baldassano, C. B. Forrest, Y. Chen, and J. Huang. Missing data matter: an empirical evaluation of the impacts of missing ehr data in comparative effectiveness research. *JAMIA Journal of the American Medical Informatics Association*, 2023.

Appendix A

ICD pattern groups used in structured retrieval

All diagnosis- and procedure-based flags in the structured retrieval pipeline were implemented through explicit ICD pattern sets rather than loose descriptive matching.

These code groups were not used indiscriminately. For baseline-history variables and some coded complication variables, diagnosis timing was constrained relative to the indexed admission so that post-admission coding would not be misrepresented as pre-existing state. Procedure codes were used both for treatment pathway variables and for anchoring later temporal restrictions such as the pre-treatment GCS window.

Construct	ICD pattern set used in the pipeline
SAH cohort selection	I60%, 430%
Vasospasm	I6784%, I67841%, I67848%
Smoking history	3051%, V1582%, F17%, Z720%, Z87891%
Hypertension	401%, 402%, 403%, 404%, 405%, I1%
Diabetes	249%, 250%, E08%, E09%, E10%, E11%, E12%, E13%
Thromboembolic history	4151%, 444%, 451%, 453%, I26%, I74%, I80%, I82%
IVH	4329%, I615%
IPH	431%, I610%, I611%, I612%, I613%, I614%
Hydrocephalus	3313%, 3314%, 3315%, G91%
Leukocytosis	28860%, D72828%, D72829%
Erythrocytosis	2890%, D751%
Cerebral salt wasting proxy	E871%
Polyuria	78842%, R35%
Hyponatremia	2761%, E871%
Hypernatremia	2760%, E870%
Epilepsy	345%, G40%
Malnutrition	2699%, E639%
Hypovolemia	27652%, E861%
Hypoxia / respiratory failure	79902%, J9601%, J9611%, J9621%, J9691%, R0902%
Hypotension	458%, I95%
Hypercapnia	J9602%, J9612%, J9622%, J9692%
Hypomagnesemia	2752%, E8342%
Hyperglycemia	7902%, R739%
Aneurysm clipping	3951, 03VG0CZ
Aneurysm coiling	3972, 3975, 3976, 03VG3DZ, 03VG3BZ
Open aneurysm ligation	3952, 03VG0ZZ, 03VG3ZZ
External ventricular drain	02.21, 02.22, 00H00MZ, 00H03MZ, 00H033Z, 009630Z

Table A.1: ICD diagnosis and procedure pattern groups used in the structured retrieval pipeline.

Appendix B

Core structured variables derived in the first retrieval stage

Table B.1 gives the full overview of the main variables derived in the first structured retrieval stage. These variables are described in the methodology chapter, while this appendix keeps the complete variable-level overview.

Table B.1: Core structured variables derived in the first retrieval stage.

Variable	Source / derivation	Encoding / categories
cohort_source	Cohort construction output	icd_asah_llm, llm_only_no_sah_icd, both
sah_icd_code	First qualifying SAH ICD code from <code>diagnoses_icd</code>	ICD code; <code>no_sah_icd</code> for LLM-only admissions
vasospasm_flag	Vasospasm ICD codes from <code>diagnoses_icd</code>	Binary: 0/1
vasospasm_icd_code	First matching vasospasm ICD code	ICD code where present; missing otherwise
sex	<code>patients</code>	Recorded sex category
age	<code>anchor_age</code> , <code>anchor_year</code> , and admission year	Numeric age in years
smoking_flag	Smoking-related ICD codes before or at indexed admission	Binary: 0/1
hypertension_flag	Hypertension-related ICD codes before or at indexed admission	Binary: 0/1
diabetes_flag	Diabetes-related ICD codes before or at indexed admission	Binary: 0/1
tee_flag	Thromboembolic-history ICD codes before or at indexed admission	Binary: 0/1
nimodipine_flag	Medication-name matching in <code>emar</code>	Binary: 0/1

Variable	Source / derivation	Encoding / categories
clipping_flag	Aneurysm clipping ICD procedure codes	Binary: 0/1
coiling_flag	Aneurysm coiling ICD procedure codes	Binary: 0/1
open_ligation_flag	Open aneurysm procedure ICD codes	Binary: 0/1
ivh_flag	Intraventricular hemorrhage ICD diagnosis codes	Binary: 0/1
iph_flag	Intraparenchymal hemorrhage ICD diagnosis codes	Binary: 0/1
hydrocephalus_flag	Hydrocephalus ICD diagnosis codes	Binary: 0/1
leukocytosis_flag	Leukocytosis ICD diagnosis codes	Binary: 0/1
erythrocytosis_flag	Erythrocytosis ICD diagnosis codes	Binary: 0/1
csw_flag	Cerebral-salt-wasting proxy ICD codes	Binary: 0/1
polyuria_flag	Polyuria ICD diagnosis codes	Binary: 0/1
hyponatremia_flag	Hyponatremia ICD diagnosis codes	Binary: 0/1
hypernatremia_flag	Hypernatremia ICD diagnosis codes	Binary: 0/1
epilepsy_flag	Epilepsy or seizure-disorder ICD diagnosis codes	Binary: 0/1
bad_nutrition_flag	Malnutrition-related ICD diagnosis codes	Binary: 0/1
hypovolemia_flag	Hypovolemia ICD diagnosis codes	Binary: 0/1
hypoxia_flag	Hypoxia or respiratory-failure ICD diagnosis codes	Binary: 0/1
hypotension_flag	Hypotension ICD diagnosis codes	Binary: 0/1
hypercapnia_flag	Hypercapnia ICD diagnosis codes	Binary: 0/1
hypomagnesemia_flag	Hypomagnesemia ICD diagnosis codes	Binary: 0/1
hyperglycemia_flag	Hyperglycemia ICD diagnosis codes	Binary: 0/1
additional_diagnoses_burden	Sum of positive additional coded disturbance flags	Integer count

Variable	Source / derivation	Encoding / categories
<code>gcs_total_worst_pre_tx</code>	Worst complete GCS triple before first aneurysm-securing treatment	Numeric score where available
<code>gcs_total</code>	Earliest available complete GCS component set during admission	Numeric score where available
<code>icu_los_vector</code>	ICU stays from <code>icustays</code>	List of ICU stay durations in hours
<code>first_treatment_date</code>	Earliest aneurysm-securing procedure date	Date where available
<code>treatment_delay_days</code>	Difference between admission time and first treatment date	Numeric days; missing if no qualifying procedure
<code>treatment_delay_cat</code>	Categorised treatment delay	<code>same_day</code> , <code>1to2_days</code> , <code>3to5_days</code> , <code>6plus_days</code> , <code>no_procedure</code>
<code>afd_ordered_flag</code>	Antifibrinolytic prescriptions	Binary: 0/1
<code>afd_admin_flag</code>	Antifibrinolytic administrations	Binary: 0/1
<code>afd_any_flag</code>	Any ordered or administered antifibrinolytic exposure	Binary: 0/1
<code>evd_flag</code>	EVD ICD procedure codes and ICU procedure events	Binary: 0/1
<code>hr_mean</code>	Heart-rate values from <code>chartevents</code>	Numeric mean
<code>hr_min</code>	Heart-rate values from <code>chartevents</code>	Numeric minimum
<code>hr_max</code>	Heart-rate values from <code>chartevents</code>	Numeric maximum
<code>rhythm_major_class</code>	Rhythm text values from <code>chartevents</code>	<code>sinus</code> , <code>ectopy</code> , <code>atrial_arrhythmia</code> , <code>ventricular_arrhythmia</code> , <code>other</code> , <code>unknown</code>
<code>any_arrhythmia_flag</code>	Derived from rhythm category severity	Binary: 0/1
<code>any_major_arrhythmia_flag</code>	Derived from rhythm category severity	Binary: 0/1
<code>discharge_note_id</code>	Linked discharge note	Note identifier; missing if unavailable
<code>radiology_note_ids</code>	Chronologically aggregated linked radiology notes	List of note identifiers; missing if unavailable
<code>death_any_flag</code>	Hospital and patient death records	Binary: 0/1

Variable	Source / derivation	Encoding / categories
death_in_hosp_flag	Hospital deathtime	Binary: 0/1
death_in_icu_flag	Death time compared with ICU stay window	Binary: 0/1
death_within_hosp_window	Death time compared with hospital admission and discharge times	Binary: 0/1
death_after_discharge_flag	Death documented after hospital discharge window	Binary: 0/1
early_death_pre_vasospasm_flag	Death before the typical vasospasm window	Binary: 0/1
death_source	Source of death information	admissions.deathtime, patients.dod, none
death_time_resolution	Precision of death time information	timestamp, date, none

Appendix C

Derived physiologic state definitions used

All counts refer to the final encoded admission-level matrix with 470 admissions. Categories were assigned using priority-based rules: when an admission met the severe definition, it was assigned to the severe category even if it also met weaker criteria.

Table C.1: Definition and distribution of leukocytosis categories.

Category	Clinical definition used in the pipeline	n	%
Severe	WBC was ≥ 15 at least 3 times during the admission, or the median WBC was ≥ 15 .	167	35.5
Likely	Not severe, but WBC was ≥ 12 at least 3 times, or the median WBC was ≥ 12 .	109	23.2
Possible	Not severe or likely, but there was at least one WBC ≥ 10 , or an ICD-coded leukocytosis signal.	126	26.8
None	No WBC-based or ICD-based leukocytosis evidence.	68	14.5

Table C.2: Definition and distribution of hyperglycemia categories.

Category	Clinical definition used in the pipeline	n	%
Severe	At least one laboratory glucose value > 180 mg/dL during the admission.	190	40.4
Moderate	No laboratory glucose value > 180 mg/dL, but at least one laboratory glucose value > 140 mg/dL.	172	36.6
None	No laboratory glucose value > 140 mg/dL.	108	23.0

Table C.3: Definition and distribution of dysnatremia categories.

Category	Clinical definition used in the pipeline	n	%
Normal	No sodium-based or ICD-based evidence for hyponatremia or hypernatremia.	185	39.4
Hyponatremia	Sodium was < 135 mmol/L at least once, or there was an ICD-coded hyponatremia signal, without hypernatremia evidence.	177	37.7
Hypernatremia	Sodium was > 150 mmol/L at least once, or there was an ICD-coded hypernatremia signal, without hyponatremia evidence.	55	11.7
Both	Evidence for both hyponatremia and hypernatremia during the admission.	53	11.3

Severe hyponatremia was separately flagged when sodium was < 130 mmol/L, but it did not define a separate final dysnatremia category.

Table C.4: Definition and distribution of hypotension categories.

Category	Clinical definition used in the pipeline	n	%
Severe	MAP was < 60 mmHg at least 3 times, or the median MAP was < 65 mmHg.	165	35.1
Likely	Not severe, but MAP was < 65 mmHg at least 3 times.	95	20.2
Possible	Not severe or likely, but there was at least one MAP < 65 mmHg, at least one SBP < 90 mmHg, or an ICD-coded hypotension signal.	119	25.3
None	No blood-pressure or ICD-based hypotension evidence.	91	19.4

DBP < 60 mmHg was extracted as supporting information, but was not used directly to assign the final hypotension category.

Table C.5: Definition and distribution of hypoxia categories.

Category	Clinical definition used in the pipeline	n	%
Severe	At least one PaO ₂ < 60 mmHg, or SpO ₂ was $< 88\%$ at least 3 times, or SaO ₂ was $< 90\%$ at least 3 times.	107	22.8
Likely	Not severe, but at least one PaO ₂ < 80 mmHg, or SpO ₂ was $< 90\%$ at least 3 times.	94	20.0
Possible	Not severe or likely, but SpO ₂ was $< 92\%$ at least 3 times.	57	12.1
None	No PaO ₂ , SpO ₂ , or SaO ₂ evidence meeting these thresholds.	212	45.1

FiO₂ and PbtO₂ were extracted as additional oxygenation context, but were not used directly to assign the final hypoxia category.

Table C.6: Definition and distribution of hypomagnesemia categories.

Category	Clinical definition used in the pipeline	n	%
Severe	Magnesium was < 1.4 mg/dL at least once. A separate < 1.2 mg/dL flag was also kept, but both were included in the severe category.	25	5.3
Mild	Not severe, but magnesium was < 1.7 mg/dL at least once, or there was an ICD-coded hypomagnesemia signal.	123	26.2
Borderline or treated	No measured or ICD-coded hypomagnesemia, but magnesium sulfate was administered.	175	37.2
None	No magnesium-based, ICD-based, or treatment-based hypomagnesemia evidence.	147	31.3

Table C.7: Definition and distribution of hypercapnia categories.

Category	Clinical definition used in the pipeline	n	%
Present	Hypercapnia evidence was present. In the secondary feature extraction this was represented using $p\text{CO}_2$ threshold flags, especially $p\text{CO}_2 > 45$ and > 50 mmHg, together with ICD-derived evidence.	99	21.1
Absent	No hypercapnia evidence in the encoded variable used for the causal discovery matrix.	371	78.9

Unlike the other physiologic variables, hypercapnia was kept as a binary variable in the causal discovery matrix.

Appendix D

Detailed validation checks for structured retrieval

Validation was built into the structured retrieval pipeline to catch issues early on, such as extraction mistakes, unintended merge behaviour, or encoding inconsistencies. This appendix gives the more practical checks behind the shorter validation summary in Section 4.1.

Admission-level integrity checks

Each intermediate step was expected to keep the admission as the unit of analysis. Both the number of admissions and the uniqueness of the admission keys were verified after every merge. This mattered because many MIMIC-IV tables contain multiple rows per admission, for example repeated measurements, lab results, medication records, ICU stays, and procedures. These sources first had to be summarised before being joined back into a single admission-level dataset.

After joins, column names were also checked for unintended overlaps or suffixes. Without this, similarly named variables from different sources can end up coexisting in the same dataset and be confused later on.

Distribution and coverage checks

Binary variables were reviewed using simple prevalence summaries and cross-tabulations. This made it easier to spot values that looked unexpectedly rare, overly common, or inconsistent with related variables. For instance, treatment indicators were compared against evidence of aneurysm-securing procedures, and combined therapy flags such as `afd_any_flag` were checked against their underlying components.

For continuous variables, ranges, quantiles, and central values were inspected. These checks helped identify possible extraction issues, unit mismatches, or values that did not look plausible. This came up most often with physiologic measurements, where small differences in item selection or unit handling can carry through into the derived states.

Feature-specific consistency checks

For GCS, the focus was on whether the individual components (eye, verbal, motor) were present and aligned in time. Since the total score depends on all three, missing or mismatched components were examined before constructing admission-level summaries. The worst pre-treatment value was also checked against procedure timing, to limit the influence of sedation or interventions.

For ECG-related variables, both coverage and value ranges were reviewed to separate likely physiologic patterns from artefacts or inconsistent text entries. After text normalization, rhythm categories were inspected to see whether the grouping still made sense at a coarse level.

For treatment delay, negative values and very long delays were not removed automatically. Instead, they were listed for inspection, as they can reflect documentation timing, transfers, or differences in date granularity.

Compound feature-block checks

The secondary physiologic feature blocks combine multiple sources, so a few extra checks were needed to keep things consistent. Derived categorical variables were restricted to predefined categories. Measurement counts were checked against the presence of summary statistics, and threshold counts were compared to the total number of measurements. When thresholds were nested, stronger abnormalities were expected to align with weaker ones.

This was done for variables such as hypotension, oxygenation, leukocytosis, hyperglycemia, and hypomagnesemia, where the final state depends on both how abnormal the values are and how often they occur. Without these checks, the variables could still look fine at first glance while containing inconsistencies.

Terminal export check

Once all variable blocks had been merged, the dataset was loaded again and checked before export. This final pass confirmed that admissions were still unique, expected columns were present, and no unintended duplicates remained. The exported dataset therefore reflects a series of smaller checks across stages, not just a single check at the end.

Appendix E

Variable provenance mapping

Table E.1 summarises the provenance of the main admission-level variables outside the physiologic feature blocks. Table E.2 summarises the provenance of the physiologic and complication-related feature blocks.

Table E.1: Provenance mapping for core admission-level variables in the structured retrieval pipeline.

Variable(s)	Source table(s)	Notes
cohort_source	cohort labels / LLM outputs	ascertainment pathway retained
vasospasm_flag, vasospasm_icd_code	diagnoses_icd	administrative vasospasm indicator; first matching code retained
sex, age	patients, admissions	age reconstructed from deidentification anchors
smoking_flag, hypertension_flag, diabetes_flag, tee_flag nimodipine_flag, clipping_flag, coiling_flag, open_ligation_flag	diagnoses_icd, admissions emar, procedures_icd	temporally constrained baseline covariates medication and procedure pathways kept separate
ivh_flag, iph_flag, hydrocephalus_flag	diagnoses_icd, admissions	coded imaging-related proxies
additional coded disturbance flags	diagnoses_icd	broad coded complication block
additional_diagnosis_burden	derived	count of coded disturbance indicators
worst pre-treatment GCS variables	chartevents, procedures_icd	restricted to pre-treatment window
admission GCS variables	chartevents	earliest available components
icu_los_vector	icustays	full ICU stay-duration vector retained
first_treatment_date, treatment_delay_days, treatment_delay_cat	admissions, procedures_icd	delay measured from admission

Variable(s)	Source table(s)	Notes
afd_admin_flag, afd_ordered_flag, afd_any_flag	emar, emar_detail, prescriptions	order and administration separated
evd_flag	procedures_icd, procedureevents	coded and ICU bedside evidence combined
hr_mean, hr_min, hr_max, rhythm variables	chartevents, d_items	heart-rate summaries and rhythm classes
note linkage variables	discharge, radiology admissions,	multimodal linkage preserved
death-related variables	patients, icustays	multiple timing definitions retained

Table E.2: Provenance mapping for physiologic and complication-related feature blocks in the structured retrieval pipeline.

Feature block	Source table(s)	Notes
hypercapnia variables	labevents, chartevents, diagnoses_icd	pCO ₂ measurements, CO ₂ -related chart evidence, and coded hypercapnia
dysnatremia variables	labevents, chartevents, diagnoses_icd	sodium summaries and categorical dysnatremia state
hypotension variables	chartevents, omr, diagnoses_icd	ICU blood pressure, outpatient context, and coded hypotension
oxygenation variables	labevents, chartevents	laboratory and charted oxygenation summaries
leukocytosis variables	labevents, diagnoses_icd	repeated WBC thresholds and coded leukocytosis
hyperglycemia variables	labevents, chartevents, diagnoses_icd	laboratory and chart glucose evidence plus coded hyperglycemia
hypovolemia variables	diagnoses_icd	coded volume depletion and polyuria proxy
hypomagnesemia variables	labevents, inputevents, diagnoses_icd	magnesium measurements, replacement therapy, and coded hypomagnesemia

Appendix F

LLM prompt templates used for clinical text extraction

This appendix documents the prompt templates used in the clinical text extraction pipeline. The prompts were used in two different parts of the methodology. First, LLMs were used for cohort-defining adjudication, where candidate admissions were assessed for likely aneurysmal subarachnoid hemorrhage. Second, LLMs were used for feature-generating extraction within the final cohort, where discharge summaries and radiology reports were converted into structured admission-level variables.

All prompts used structured JSON output schemas. This made the extracted variables easier to validate, merge with structured EHR data, and inspect during downstream analysis. The prompts also required short supporting quotes from the source text where possible, so that extracted values could be traced back to the original clinical documentation.

Prompt	Methodological role	Main output
aSAH adjudication	Cohort definition	Admission-level likelihood score and label for likely aSAH, with supporting aneurysm, rupture, treatment, and negative-evidence signals.
Discharge extraction	Feature generation	Severity indicators, aneurysm characteristics, vasospasm evaluation mode, treatment context, baseline history, hydrocephalus, EVD context, and neuroimaging intensity.
Radiology extraction	Feature generation	Radiology-derived bleed phenotype, clot burden, hydrocephalus, EVD evidence, aneurysm information, neuroimaging intensity, and radiology-only vasospasm evidence.
Discharge vasospasm scoring	Feature generation / validation	Discharge-summary-based ordinal vasospasm evidence score with boundary justification and supporting quotes.

Table F.1: Overview of LLM prompt templates used in the information extraction pipeline.

Cohort-defining aSAH adjudication prompts

The aSAH adjudication prompts were used before the final cohort was fixed. They assessed whether a candidate admission likely represented aneurysmal subarachnoid hemorrhage. Two related prompt variants were used: one for admissions already selected through SAH-related structured evidence, and one for admissions selected through aneurysm treatment or nimodipine exposure despite lacking a qualifying SAH ICD code.

```

1  ASAH_PROMPT = ""
2  You are a clinical NLP assistant helping with retrospective research on aneurysmal subarachnoid
   hemorrhage (aSAH) in ICU patients.
3
4  You receive text from ONE hospital admission. The patient definitely had some form of subarachnoid
   hemorrhage (SAH), but the subtype is uncertain. Your task is to infer whether this admission was
   likely ANEURYSMAL SAH (aSAH), as opposed to traumatic SAH, perimesencephalic/nonaneurysmal SAH,
   or another non-aneurysmal SAH pattern.
5
6  The input may contain:
7  - a discharge summary
8  - radiology report text
9  - both
10 Some admissions may have only one of these.
11
12 MAIN GOAL
13 Be reasonably recall-oriented for aSAH:
14 - If the text gives credible evidence that the SAH was aneurysmal, ruptured from an aneurysm, or
   managed as an aneurysm case, lean positive.
15 - You do NOT need aneurysm location to call it likely aSAH.
16 - You should still avoid obvious false positives such as traumatic SAH or explicitly nonaneurysmal/
   perimesencephalic SAH.
17
18 WHAT COUNTS AS POSITIVE EVIDENCE FOR aSAH
19 Strong positive examples:
20 - explicit phrases such as:
21   - "aneurysmal subarachnoid hemorrhage"
22   - "aSAH"
23   - "ruptured aneurysm"
24   - "ruptured cerebral aneurysm"
25   - "ruptured saccular aneurysm"
26   - "subarachnoid hemorrhage secondary to aneurysm"
27   - "SAH from ruptured [artery] aneurysm"
28 - treatment trajectory strongly indicating aneurysm management:
29   - "status post coiling of aneurysm"
30   - "status post clipping of aneurysm"
31   - "underwent coil embolization of ruptured aneurysm"
32   - "aneurysm clipped/coiled/embolized"
33 - radiology explicitly identifying an intracranial aneurysm in the setting of SAH, especially if
   ruptured or treated
34
35 Moderate positive examples:
36 - explicit mention of an intracranial aneurysm in the same admission plus SAH, even if rupture is not
   spelled out
37 - discharge summary framing the hospitalization around aneurysm treatment after SAH
38 - radiology mention of an aneurysm with supportive treatment language or follow-up after coiling/
   clipping
39
40 WEAKER POSITIVE EXAMPLES
41 These can still support likely aSAH if nothing contradicts them:
42 - "aneurysm" clearly present in neurovascular context during this admission
43 - "s/p coiling" or "s/p clipping" where the target is clearly an intracranial aneurysm
44 - discharge text using aneurysm-centered language without full causal phrasing
45
46 WHAT COUNTS AS NEGATIVE EVIDENCE AGAINST aSAH
47 Strong negative examples:
48 - explicit traumatic SAH
49 - explicit perimesencephalic SAH / nonaneurysmal SAH / angiogram-negative SAH
50 - explicit statement that no aneurysm was found and the bleed was considered nonaneurysmal
51 - explicit alternative etiology incompatible with aneurysmal rupture
52
53 Moderate negative examples:
54 - repeated workup showing no aneurysm plus language favoring traumatic or nonaneurysmal pattern
55 - discharge summary or radiology clearly framing the case as traumatic SAH
56
57 IMPORTANT GUARDRAILS
58 - The patient definitely had SAH already. Do not spend effort deciding whether SAH itself existed.
59 - Do not require aneurysm location.
60 - Do not require the word "ruptured" if treatment or summary language clearly indicates aneurysm-
   related SAH.
61 - Do not call aSAH positive from unrelated aneurysm mentions outside clear intracranial/neuro context.
62 - Do not infer aSAH just from vasospasm, hydrocephalus, EVD, or nimodipine alone.
63 - If the note says "history of aneurysm clipping/coiling" but this appears unrelated to the current

```

```

        SAH admission, treat cautiously.
64 - If radiology says "no aneurysm identified" but discharge clearly says ruptured aneurysm treated this
        admission, weigh the full record and explain the conflict.
65 - If evidence conflicts, choose the overall best label for the admission and explain why.
66
67 OUTPUT SCALE
68 Return exactly one integer asah_likelihoood_score in [-5, 5].
69
70 Positive:
71 +5: explicit, unambiguous evidence the SAH was aneurysmal
72 +4: very strong evidence for aSAH, slightly less direct than +5
73 +3: moderate evidence for aSAH
74 +2: weak but meaningful evidence toward aSAH
75 +1: very weak leaning toward aSAH
76
77 Neutral:
78 0: insufficient, mixed, or truly ambiguous evidence
79
80 Negative:
81 -1: very weak leaning against aSAH
82 -2: weak but meaningful evidence against aSAH
83 -3: moderate evidence against aSAH
84 -4: strong evidence against aSAH
85 -5: explicit, unambiguous evidence this was not aneurysmal SAH
86
87 ALSO EXTRACT
88 - whether aneurysm is mentioned at all in a neuro context
89 - whether rupture/aneurysmal causation is explicit
90 - whether aneurysm treatment is mentioned
91 - whether there is explicit evidence against aSAH
92 - whether traumatic SAH is mentioned
93 - whether perimesencephalic or nonaneurysmal SAH is mentioned
94 - coarse source support:
95   - discharge_only
96   - radiology_only
97   - both
98   - neither
99
100 EVIDENCE QUOTES
101 Use short exact verbatim quotes copied from the provided text.
102 No paraphrasing.
103 No ellipses.
104 Each quote must be a contiguous span appearing exactly in the input.
105
106 STRICT JSON ONLY
107 Return exactly this schema:
108
109 {
110   "asah_likelihoood_score": integer from -5 to 5,
111   "final_label": "likely_aSAH" | "possible_aSAH" | "indeterminate" | "likely_not_aSAH",
112   "support_source": "discharge_only" | "radiology_only" | "both" | "neither",
113
114   "aneurysm_signal": {
115     "aneurysm_mentioned_neuro_context": true | false | null,
116     "rupture_or_aneurysmal_causation_explicit": true | false | null,
117     "aneurysm_treatment_mentioned": true | false | null,
118     "treatment_type": "coil" | "clip" | "other" | null,
119     "location_text": string | null
120   },
121
122   "negative_signal": {
123     "explicit_nonaneurysmal_sah": true | false | null,
124     "explicit_traumatic_sah": true | false | null,
125     "explicit_no_aneurysm_found": true | false | null
126   },
127
128   "evidence_quotes": [string, ...],
129   "why_not_lower": {
130     "summary": string,
131     "quotes": [string, ...]
132   },
133   "why_not_higher": {
134     "summary": string,

```

```

135     "quotes": [string, ...]
136 },
137 "comments": string
138 }
139
140 Rules:
141 - evidence_quotes: 1 to 6 short verbatim quotes when possible
142 - why_not_lower.quotes and why_not_higher.quotes: 0 to 2 short verbatim quotes
143 - If something is unknown, use null rather than guessing.
144 - Do not output any text outside the JSON object.
145 """.strip()
146
147
148 NON_SAH_TREATED_ASAH_PROMPT = """
149 You are a clinical NLP assistant helping with retrospective research on aneurysmal subarachnoid
150     hemorrhage (aSAH) in ICU patients.
151
152 You receive text from ONE hospital admission. This admission was selected because it had clipping,
153     coiling, and/or nimodipine exposure despite lacking an SAH ICD code. Therefore, the admission may
154     represent:
155     - a missed aneurysmal SAH
156     - a non-aneurysmal SAH
157     - traumatic SAH
158     - perimesencephalic SAH
159     - unruptured aneurysm treatment
160     - another neurovascular admission
161     - or a non-SAH case
162
163 Your task is to infer whether the admission supports CURRENT SAH during this admission, and if so,
164     whether it was likely ANEURYSMAL SAH (aSAH).
165
166 MAIN GOALS
167 1. Decide whether the text supports that SAH was present during the current admission.
168 2. Decide how likely it is that any current SAH was aneurysmal.
169 3. Remain reasonably recall-oriented for aSAH, but do not force an aSAH label if the admission appears
170     to be an unruptured aneurysm case without hemorrhage.
171
172 IMPORTANT
173 - Do not assume SAH is definitely present.
174 - Distinguish current admission findings from remote history when possible.
175 - Clipping/coiling alone does NOT prove aSAH; it may reflect treatment of an unruptured aneurysm.
176 - Nimodipine alone does NOT prove aSAH.
177 - If the note clearly describes aneurysm treatment in the setting of SAH, lean positive.
178 - If the note clearly describes aneurysm treatment without any evidence of current SAH, that is
179     evidence against current aSAH.
180 - If the text says traumatic SAH, perimesencephalic SAH, angiogram-negative SAH, or nonaneurysmal SAH,
181     weigh that as negative evidence against aSAH.
182 - If the text suggests prior history of clipping/coiling unrelated to the current admission, treat
183     cautiously.
184
185 WHAT COUNTS AS POSITIVE EVIDENCE FOR CURRENT SAH
186 Strong examples:
187 - explicit phrases such as:
188     - "subarachnoid hemorrhage"
189     - "SAH"
190     - "acute subarachnoid hemorrhage"
191     - "diffuse subarachnoid blood"
192     - "basal cistern subarachnoid hemorrhage"
193 - discharge summary framing the admission as SAH hospitalization
194 - radiology describing current intracranial subarachnoid blood
195
196 WHAT COUNTS AS POSITIVE EVIDENCE FOR aSAH
197 Strong positive examples:
198 - explicit phrases such as:
199     - "aneurysmal subarachnoid hemorrhage"
200     - "aSAH"
201     - "ruptured aneurysm"
202     - "ruptured cerebral aneurysm"
203     - "ruptured saccular aneurysm"
204     - "subarachnoid hemorrhage secondary to aneurysm"
205     - "SAH from ruptured [artery] aneurysm"
206 - treatment trajectory strongly indicating ruptured aneurysm management:
207     - "status post coiling of ruptured aneurysm"

```

```

200 - "status post clipping of ruptured aneurysm"
201 - "underwent coil embolization of ruptured aneurysm"
202 - radiology explicitly identifying an intracranial aneurysm in the setting of SAH
203
204 Moderate positive examples:
205 - explicit mention of an intracranial aneurysm in the same admission plus current SAH, even if rupture
    is not spelled out
206 - discharge summary framing the hospitalization around aneurysm treatment after SAH
207 - radiology mention of aneurysm with supportive treatment language in an SAH admission
208
209 WEAKER POSITIVE EXAMPLES
210 These can still support likely aSAH if nothing contradicts them:
211 - aneurysm clearly present in neurovascular context during this admission AND current SAH is also
    supported
212 - s/p coiling or s/p clipping where the target is clearly an intracranial aneurysm and the same
    admission also contains current SAH language
213
214 WHAT COUNTS AS NEGATIVE EVIDENCE AGAINST aSAH
215 Strong negative examples:
216 - no evidence of current SAH, only unruptured aneurysm treatment
217 - explicit traumatic SAH
218 - explicit perimesencephalic SAH / nonaneurysmal SAH / angiogram-negative SAH
219 - explicit statement that no aneurysm was found and bleed was considered nonaneurysmal
220 - explicit alternative etiology incompatible with aneurysmal rupture
221
222 Moderate negative examples:
223 - repeated workup showing no aneurysm plus language favoring traumatic or nonaneurysmal pattern
224 - discharge summary clearly framing the case as unruptured aneurysm repair without hemorrhage
225
226 OUTPUT SCALE
227 Return exactly one integer asah_likelihoood_score in [-5, 5].
228
229 Interpretation:
230 +5: explicit, unambiguous evidence of current aneurysmal SAH
231 +4: very strong evidence for current aSAH
232 +3: moderate evidence for current aSAH
233 +2: weak but meaningful evidence toward current aSAH
234 +1: very weak leaning toward current aSAH
235 0: insufficient, mixed, or ambiguous evidence
236 -1: very weak leaning against current aSAH
237 -2: weak but meaningful evidence against current aSAH
238 -3: moderate evidence against current aSAH
239 -4: strong evidence against current aSAH
240 -5: explicit, unambiguous evidence this admission was not current aneurysmal SAH
241
242 ALSO EXTRACT
243 - whether current SAH is supported at all in this admission
244 - whether SAH is explicit
245 - whether aneurysm is mentioned in neuro context
246 - whether rupture/aneurysmal causation is explicit
247 - whether aneurysm treatment is mentioned
248 - whether there is explicit evidence against aSAH
249 - whether traumatic SAH is mentioned
250 - whether perimesencephalic or nonaneurysmal SAH is mentioned
251 - coarse source support:
252   - discharge_only
253   - radiology_only
254   - both
255   - neither
256
257 EVIDENCE QUOTES
258 Use short exact verbatim quotes copied from the provided text.
259 No paraphrasing.
260 No ellipses.
261 Each quote must be a contiguous span appearing exactly in the input.
262
263 STRICT JSON ONLY
264 Return exactly this schema:
265
266 {
267   "asah_likelihoood_score": integer from -5 to 5,
268   "final_label": "likely_aSAH" | "possible_aSAH" | "indeterminate" | "likely_not_aSAH",
269   "support_source": "discharge_only" | "radiology_only" | "both" | "neither",

```

```

270
271 "sah_signal": {
272   "sah_present_current_admission": true | false | null,
273   "sah_explicit": true | false | null
274 },
275
276 "aneurysm_signal": {
277   "aneurysm_mentioned_neuro_context": true | false | null,
278   "rupture_or_aneurysmal_causation_explicit": true | false | null,
279   "aneurysm_treatment_mentioned": true | false | null,
280   "treatment_type": "coil" | "clip" | "other" | null,
281   "location_text": string | null
282 },
283
284 "negative_signal": {
285   "explicit_nonaneurysmal_sah": true | false | null,
286   "explicit_traumatic_sah": true | false | null,
287   "explicit_no_aneurysm_found": true | false | null
288 },
289
290 "evidence_quotes": [string, ...],
291 "why_not_lower": {
292   "summary": string,
293   "quotes": [string, ...]
294 },
295 "why_not_higher": {
296   "summary": string,
297   "quotes": [string, ...]
298 },
299 "comments": string
300 }
301
302 Rules:
303 - evidence_quotes: 1 to 6 short verbatim quotes when possible
304 - why_not_lower.quotes and why_not_higher.quotes: 0 to 2 short verbatim quotes
305 - If something is unknown, use null rather than guessing.
306 - Do not output any text outside the JSON object.""

```

Listing F.1: Prompt templates for aSAH cohort adjudication.

Discharge-summary structured extraction prompt

The discharge-summary extraction prompt was used after cohort definition. It extracted structured variables from narrative discharge summaries, including severity grades, clot burden proxies, aneurysm location, aneurysm treatment, vasospasm evaluation and management, baseline history, hydrocephalus, EVD context, and neuroimaging intensity.

```
1  You are a clinical NLP assistant helping with retrospective research on aneurysmal subarachnoid
   hemorrhage (SAH) in ICU patients.
2
3  You receive one English discharge summary per hospital admission, from the MIMIC-IV-Note dataset. The
   notes are de-identified and may contain three underscores (___) where protected health
   information was removed.
4
5  Your task is to extract a set of Tier-1 and Tier-2 variables from THIS discharge summary. You must be
   conservative: if a value is not explicitly supported by the text, mark it as unknown and explain
   why.
6
7  IMPORTANT PRINCIPLES
8  - This is NOT diagnosis. This is structured extraction from the discharge summary ONLY.
9  - Prefer explicit statements, standard scales (WFNS/Fisher/Hunt-Hess), and clear procedure phrasing.
10 - Handle negation and uncertainty carefully (e.g., no evidence of, cannot exclude, concern for, watch)
    .
11 - If multiple contradictory statements exist, keep the value as unknown/ambiguous and include both
    quotes.
12 - Time windows:
13   - Base: baseline / admission state
14   - D02: first 48h
15   - D314: typical vasospasm window
16   - Disch: discharge-time documentation
17 - If the note gives an explicit timing (on hospital day 7, on ___), capture it in a short free-text
    time_anchor field for that variable.
18
19 WHAT TO EXTRACT
20
21 1) SAH severity grades & bleed burden
22   - WFNS grade (if present)
23   - Hunt-Hess grade (if present)
24   - Fisher or modified Fisher grade (if present)
25   - clot_burden_proxy (map to 3-level ordinal):
26     1 = small/trace/thin/localized SAH
27     2 = moderate/diffuse SAH but not described as thick/packed
28     3 = thick/dense/diffuse/large clot burden OR Fisher 34 / thick descriptions
29     0 = unknown / not stated
30
31 2) aneurysm location
32   - aneurysm_location_tokens: list of raw location strings (e.g., Acomm, left ICA bifurcation, Pcomm,
     basilar tip)
33   - aneurysm_loc_group: one of ["anterior","posterior","unknown"]
34     *Anterior examples: Acomm, ACA, MCA, ICA, Pcomm
35     *Posterior examples: basilar, vertebral, PICA, PCA (non-fetal), SCA
36
37 3) aneurysm treatment summary
38   - aneurysm_tx_type: one of ["clip","coil","both","other_endovascular","diagnostic_angio_only","
     none_recorded","unknown"]
39   - tx_details_metadata: short structured flags where present:
40     - diagnostic_angio_done (true/false/unknown)
41     - aborted_attempt (true/false/unknown)
42     - staged_or_repeat_procedure (true/false/unknown)
43     - notes_free_text (short, only if needed)
44
45 4) Evidence mode for vasospasm evaluation (NOT re-scoring vasospasm)
46   - E_evidence_mode: categorize the main evidence channels mentioned:
47     one of ["TCD_only","CTA_only","Angio_DSA_only","TCD+CTA","CTA+Angio_DSA","TCD+Angio_DSA","TCD+CTA
     +Angio_DSA","text_only_watch","explicit_no_evidence","unknown"]
48   - This is about HOW vasospasm was evaluated/mentioned, not whether it happened.
49
50 5) Vasospasm-directed interventions (beyond nimodipine prophylaxis)
51   Extract whether the DS describes vasospasm-directed management:
52   - induced_hypertension_strategy: ["yes","no","unclear"]
53     (Examples: BP maintained 140/80 to promote perfusion, pressors for vasospasm)
54   - intra_arterial_therapy: ["yes","no","unclear"]
```

(Examples: intra-arterial verapamil/nicardipine/milrinone; slow infusion; angioplasty)

- angioplasty_for_spasm: ["yes", "no", "unclear"]
- bp_goal_ranges: list of any explicit BP targets/ranges mentioned (strings)
- spasm_treatment_agents: list of agents mentioned (e.g., verapamil, nicardipine, milrinone)

6) Baseline history flags (extract ONLY if clearly stated as patient history; not family history)

- smoking_hist_ds: ["yes", "no", "unknown"]
- alcohol_hist_ds: ["yes", "no", "unknown"]
- htn_hist_ds: ["yes", "no", "unknown"]
- dm_hist_ds: ["yes", "no", "unknown"]
- chf_hist_ds: ["yes", "no", "unknown"]
- stroke_hist_ds: ["yes", "no", "unknown"]

For each, capture brief supporting quote(s). If the DS is blanked (___), return unknown.

7) Hydrocephalus and EVD mentions (backup)

- hydrocephalus_ds: ["yes", "no", "unclear", "unknown"]
(Yes if explicitly stated; No only if explicitly stated no hydrocephalus)
- evd_ds: ["yes", "no", "unknown"]
(Yes if EVD/ventriculostomy/external ventricular drain mentioned; include placement/removal if stated)
- evd_timing_notes: short free-text if explicit (placed on ___, removed on ___, clamped on ___)

8) Imaging intensity mention (backup; do NOT count exact studies)

- neuroimaging_intensity_ds: ["0", "1", "2+", "unknown"]
Use DS language only:
 - "0": explicitly indicates none or no imaging mentioned at all in the entire DS
 - "1": mentions a single neuroimaging episode/study type without repeats
 - "2+": mentions repeated/serial imaging (e.g., repeat CT, serial CT, daily CT, multiple CTA)
- imaging_repeat_language_quotes: quotes that justify "2+"

EVIDENCE QUOTES (VERBATIM ONLY)

- For each extracted variable, provide 13 SHORT exact quotes copied verbatim from the discharge summary.
- NO paraphrasing. NO rewriting. NO mixing fragments. NO ellipses (...).
- Each quote must be a contiguous span that appears exactly in the note text.
- If no quote exists for a variable, use an empty quotes list.

BOUNDARY JUSTIFICATION (MANDATORY, LIGHTWEIGHT)

For each variable that is NOT "unknown":

- briefly justify why you did NOT choose a more extreme/stronger interpretation (e.g., why unclear not yes, why anterior not posterior, why 2+ not 1)
- Provide 01 short verbatim quote if available; otherwise empty list.

OUTPUT FORMAT

Always answer in STRICT JSON with the following structure:

```
{
  "sah_severity": {
    "wfns_grade": {"value": string or null, "time_window": "Base" or "D0-2" or "unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
    "hunt_hess_grade": {"value": string or null, "time_window": "Base" or "D0-2" or "unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
    "fisher_grade": {"value": string or null, "time_window": "Base" or "D0-2" or "unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
    "clot_burden_proxy": {"value": 0 or 1 or 2 or 3, "time_window": "D0-2" or "Base" or "unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}}
  },
  "aneurysm_location": {
    "aneurysm_location_tokens": {"value": [list of strings], "time_window": "Base" or "unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
    "aneurysm_loc_group": {"value": "anterior" or "posterior" or "unknown", "time_window": "Base" or "unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}}
  },
  "aneurysm_treatment": {
    "aneurysm_tx_type": {"value": "clip" or "coil" or "both" or "other_endovascular" or "
```

```

    diagnostic_angio_only" or "none_recorded" or "unknown", "time_window": "D0-7" or "unknown", "
    time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "
    quotes": [0-1 quotes]}},
114 "tx_details_metadata": {
115     "diagnostic_angio_done": {"value": "true" or "false" or "unknown", "quotes": [0-2 quotes]},
116     "aborted_attempt": {"value": "true" or "false" or "unknown", "quotes": [0-2 quotes]},
117     "staged_or_repeat_procedure": {"value": "true" or "false" or "unknown", "quotes": [0-2 quotes]},
118     "notes_free_text": {"value": string or null, "quotes": [0-1 quotes]}
119 }
120 },
121
122 "vasospasm_evaluation_mode": {
123     "E_evidence_mode": {"value": one of the categories above, "time_window": "D3-14" or "D0-14" or "
    unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary":
    string, "quotes": [0-1 quotes]}}
124 },
125
126 "vasospasm_directed_interventions": {
127     "induced_hypertension_strategy": {"value": "yes" or "no" or "unclear", "time_window": "D3-14" or "
    unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary":
    string, "quotes": [0-1 quotes]}},
128     "intra_arterial_therapy": {"value": "yes" or "no" or "unclear", "time_window": "D3-14" or "unknown
    ", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string,
    "quotes": [0-1 quotes]}},
129     "angioplasty_for_spasm": {"value": "yes" or "no" or "unclear", "time_window": "D3-14" or "unknown
    ", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string,
    "quotes": [0-1 quotes]}},
130     "bp_goal_ranges": {"value": [list of strings], "time_window": "D3-14" or "unknown", "time_anchor":
    string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "quotes": [0-1
    quotes]}},
131     "spasm_treatment_agents": {"value": [list of strings], "time_window": "D3-14" or "unknown", "
    time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "
    quotes": [0-1 quotes]}}
132 },
133
134 "baseline_history_ds": {
135     "smoking_hist_ds": {"value": "yes" or "no" or "unknown", "time_window": "Base", "quotes": [0-2
    quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
136     "alcohol_hist_ds": {"value": "yes" or "no" or "unknown", "time_window": "Base", "quotes": [0-2
    quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
137     "htn_hist_ds": {"value": "yes" or "no" or "unknown", "time_window": "Base", "quotes": [0-2 quotes
    ], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
138     "dm_hist_ds": {"value": "yes" or "no" or "unknown", "time_window": "Base", "quotes": [0-2 quotes],
    "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
139     "chf_hist_ds": {"value": "yes" or "no" or "unknown", "time_window": "Base", "quotes": [0-2 quotes
    ], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}},
140     "stroke_hist_ds": {"value": "yes" or "no" or "unknown", "time_window": "Base", "quotes": [0-2
    quotes], "why_not_stronger": {"summary": string, "quotes": [0-1 quotes]}}
141 },
142 },
143
144 "hydrocephalus_evd_ds": {
145     "hydrocephalus_ds": {"value": "yes" or "no" or "unclear" or "unknown", "time_window": "D0-14" or "
    unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary":
    string, "quotes": [0-1 quotes]}},
146     "evd_ds": {"value": "yes" or "no" or "unknown", "time_window": "D0-14" or "unknown", "time_anchor
    ": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary": string, "quotes": [0-1
    quotes]}},
147     "evd_timing_notes": {"value": string or null, "quotes": [0-2 quotes]}
148 },
149
150 "neuroimaging_intensity_ds": {
151     "neuroimaging_intensity_ds": {"value": "0" or "1" or "2+" or "unknown", "time_window": "D0-14" or
    "unknown", "time_anchor": string or null, "quotes": [0-3 quotes], "why_not_stronger": {"summary":
    string, "quotes": [0-1 quotes]}},
152     "imaging_repeat_language_quotes": [0-3 short verbatim quotes]
153 },
154
155 "comments": "short explanation of overall reasoning and any major ambiguities"
156 }
157
158 - Do NOT include any text outside the JSON object.
159 - Do NOT add trailing commas.

```

```
160 - If a field is not mentioned in the discharge summary, set its value to null/unknown as specified and  
    leave quotes empty.
```

Listing F.2: Prompt template for structured extraction from discharge summaries.

Radiology structured extraction prompt

The radiology extraction prompt was used to convert bundled radiology reports into structured imaging-derived variables. It focused on neuroimaging intensity, radiology-only vasospasm evidence, vasospasm severity, bleed phenotype, clot burden, hydrocephalus and ventricular findings, EVD evidence, aneurysm location, aneurysm treatment mentions, and selected supportive signals.

```
1 You are a clinical NLP assistant helping with retrospective research on aneurysmal subarachnoid
  hemorrhage (SAH) in ICU patients.
2
3 You receive ONE radiology report (or a concatenated bundle of radiology reports) from the same
  hospital admission (MIMIC-IV-Note radiology).
4
5 The text is de-identified and may contain three underscores (___) where protected health information
  was removed.
6
7 GOAL
8 Extract a compact, structured summary of intracranial imaging evidence relevant to SAH phenotype,
  vasospasm evidence, aneurysm characterization, hydrocephalus/ventricles, and imaging intensity.
9 You must ONLY use the radiology text provided (no external knowledge, no guessing).
10
11 CORE RULES (MEASUREMENT BIAS + SAFETY AGAINST OVER-READING)
12 - Radiology exists because someone ordered imaging; it is NOT a complete clinical summary.
13 - If something is not explicitly stated, mark it as unknown (null) and explain in comments.
14 - Prefer explicit statements in the IMPRESSION section; use FINDINGS/TECHNIQUE only if necessary.
15 - If there are multiple reports bundled together, later-time reports override earlier ones when they
  contradict.
16   * When overriding, include BOTH quotes (earlier and later) and describe the conflict in comments.
17 - IMPORTANT: Many bundled radiology reports are non-neuro (CXR, extremity ultrasound, renal US, etc.).
18   * Ignore non-neuro reports for neuroimaging counts and neuro-specific fields.
19   * You may mention non-neuro imaging exists in comments.
20
21 CRITICAL GUARDRAILS FOR VASOSPASM (to avoid false positives)
22 - Do NOT interpret chronic vascular disease as vasospasm.
23   * If narrowing/stenosis/occlusion is described as atherosclerosis, plaque, chronic occlusion,
24     carotid disease, critical stenosis, or similar,
25     then it is NOT vasospasm evidence unless the text explicitly mentions vasospasm/spasm.
26 - Only treat vasospasm evidence as valid when the wording refers to intracranial arteries/branches in
  a neurovascular context
27   (CTA/CTP/MRA/MRV/conventional cerebral angiography), OR the text explicitly says vasospasm.
28 - Ambiguous phrases like may represent spasm vs congenital variation should be scored weakly (near 0),
  unless additional explicit vasospasm language exists.
29
30 INTERNAL PRE-PROCESSING (MANDATORY)
31 If the input contains multiple reports concatenated:
32 1) First, mentally segment into individual studies/exams using cues like:
33   - "EXAMINATION:", "REASON FOR EXAM:", "HISTORY:", "INDICATION:", timestamps, modality headers (CT/
34     CTA/MRA/ANGIO), "IMPRESSION:"
35 2) For each segmented report, identify:
36   - Is it neuro/head/brain/cerebral? If no, treat as non-neuro and do not use it for neuro counts or
37     neuro fields.
38 3) Then aggregate across neuro reports:
39   - Counts = sum across neuro reports.
40   - Bleed phenotype / hydrocephalus / aneurysm: if contradictory, later-time overrides earlier (
41     record both quotes).
42   - Vasospasm score: choose the score that best reflects the overall admission picture based on
43     neurovascular imaging,
44     prioritizing later-time neurovascular exams if they clarify earlier ambiguity.
45
46 WHAT TO EXTRACT
47
48 1) Neuroimaging intensity (proxy for diagnostic intensity)
49 Count NEURO/HEAD/BRAIN studies during the admission (ignore non-neuro):
50 - Non-contrast CT head/brain (includes CT head without contrast, noncontrast head CT, etc.)
51 - CTA head/brain (includes CT angiography head, CTA head, etc.)
52 - CT perfusion (CTP, perfusion CT)
53 - MRA and/or MRV
54 - Conventional cerebral angiography / arteriogram / diagnostic cerebral angiogram (intracranial)
55 Counting rule:
56 - Count a study if modality appears in a header-like region OR the report clearly describes that
  modality in TECHNIQUE.
```

52 - If one report describes multiple components (e.g., CT head + CTA head + CTP in one combined report),
 increment each applicable count.

53 Return counts + coarse bin for total neuroimaging: 0 / 1 / 2+.

54

55 Also return a counting confidence:

56 - "high" if the text has clear modality headers for neuro exams

57 - "medium" if modality is inferable but headers are inconsistent

58 - "low" if the bundle looks truncated or lacks clear exam boundaries

59

60 2) Vasospasm imaging evidence (radiology-only)

61 Determine whether the radiology text supports vasospasm, ONLY from radiology.

62 Positive examples:

63 - vasospasm, suggestive of vasospasm, persistent mild vasospasm, focal narrowing ... suggestive of
 vasospasm

64 Negative examples:

65 - no evidence of vasospasm, no occlusive vasospasm, no findings to suggest vasospasm

66 Ambiguous examples:

67 - may represent spasm vs congenital variation, cannot exclude, small caliber ...

68 Output an ordinal vasospasm evidence score in [-5..+5] ONLY based on radiology.

69 Also extract supporting modalities (CTA / Angiography / CTP / MRA/MRV / CT head only / Unclear).

70

71 2b) Vasospasm severity indicator (radiology-only; separates how bad from mere mention)

72 Extract E_vasospasm_severity {"none","mild","moderate","severe","unknown"} from radiology wording
 ONLY.

73

74 Rules:

75 - Mentions of vasospasm/spasm only in INDICATION/HISTORY (e.g., evaluate for vasospasm) do NOT count
 as documentation.

76 * Set E_vasospasm_severity="unknown" unless FINDINGS/IMPRESSION explicitly negate or affirm
 vasospasm.

77 - "none" ONLY if radiology explicitly negates vasospasm/spasm (e.g., no evidence of vasospasm).

78 - "mild"/"moderate"/"severe" ONLY if vasospasm/spasm is explicitly documented AND severity is
 explicitly stated

79 OR described with clear vasospasm-severity descriptors.

80 * Explicit severity words: mild, moderate, severe.

81 * Strong severity descriptors (map to "severe" unless stated otherwise): marked, diffuse, multifocal
 severe narrowing,
 critical vasospasm, pronounced vasospasm.

82 - If vasospasm/spasm is documented but WITHOUT severity adjectives E_vasospasm_severity="unknown".

83 - Do NOT interpret chronic vascular disease as vasospasm severity.

84 * If narrowing/stenosis/occlusion is described as atherosclerosis, plaque, chronic occlusion,
 carotid disease, etc.,

85 do NOT map to mild/moderate/severe; set "unknown" unless vasospasm/spasm is explicitly stated.

86 - NEVER infer severity from treatment intensity (angioplasty/IA vasodilators, etc.).

87 - If multiple neurovascular exams disagree across time, prefer later-time neurovascular imaging for
 the final severity,

88 and describe the conflict in comments with quotes.

89

90

91 3) Evidence mode (channel supporting evidence)

92 Classify which evidence channel(s) appear in the neuro radiology:

93 { "CTA", "CTP", "Angiography", "MRA/MRV", "CT head only", "Mixed/Multiple", "None/No neuroimaging" }.

94 Rules:

95 - "Mixed/Multiple" if >=2 of {CTA, CTP, Angiography, MRA/MRV} appear in neuro reports.

96 - "CT head only" if neuro imaging exists but none of CTA/CTP/MRA/MRV/Angiography is present.

97

98 4) Bleed phenotype & clot burden proxy (coarse)

99 Detect presence/extent of:

100 - SAH (subarachnoid hemorrhage OR equivalent descriptions like blood in basal cisterns/sylvian
 fissures)

101 - IVH (intraventricular hemorrhage)

102 - IPH (intraparenchymal hemorrhage)

103 - SDH (subdural hematoma)

104 SAH extent:

105 - "trace/small" vs "moderate" vs "extensive/diffuse" vs null (if not described)

106 Clot burden proxy:

107 - 0 = none/trace/small

108 - 1 = moderate

109 - 2 = extensive/diffuse/thick/residual extensive

110 - null if SAH is not described at all

111

112 5) Hydrocephalus / ventricles / CSF diversion

113 Determine:

114 - hydrocephalus: present/absent/unknown

```

115 - ventricle description: free text if described (e.g., mild ventriculomegaly, ventricles effaced,
    stable ventriculomegaly)
116 - EVD/ventriculostomy presence:
117   * detect terms: ventriculostomy catheter, ventricular drain, EVD, ventricular catheter, VP shunt,
    shunt catheter
118 - EVD tip location: free text if described
119
120 6) Aneurysm location group
121 Extract aneurysm location if present (e.g., PCOM/PCoM, ACoM, MCA, ACA, basilar, vertebrobasilar, PCA,
    ICA terminus).
122 Map to coarse group:
123 - anterior: ICA/ACA/ACoM/MCA/PCOM (circle of Willis anterior circulation)
124 - posterior: basilar/vertebral/PCA/PICA/AICA/SCA/vertebrobasilar
125 - unknown: not enough info
126 If the report says status post coiling of [location] aneurysm, treat aneurysm_present as true even if
    the word aneurysm is not repeated elsewhere.
127
128 7) Aneurysm treatment mention (support)
129 If radiology mentions coiling/clipping/embolization/stent or status post coiling, extract:
130 - aneurysm_treatment_mentioned true/false/null
131 - aneurysm_treatment_type: coil/clip/other/null
132
133 8) Mechanical ventilation evidence (support)
134 If radiology includes chest imaging with endotracheal tube, ET tube, intubated, or tracheostomy, set
    mechvent_support=true and quote it.
135 Otherwise null (do not force).
136
137 OUTPUT REQUIREMENTS
138
139 A) EVIDENCE QUOTES (VERBATIM ONLY)
140 - Provide SHORT exact quotes copied verbatim from the RADIOLOGY text (not discharge summaries).
141 - NO paraphrasing. NO rewriting. NO mixing fragments. NO ellipses (...).
142 - Each quote must be a contiguous span that appears exactly in the provided text.
143 - If a field is unknown, provide [] for its quotes.
144
145 B) STRICT JSON ONLY
146 Always answer in STRICT JSON, no extra text, no trailing commas.
147
148 JSON SCHEMA TO FOLLOW (EXACT KEYS)
149 {
150   "neuroimaging_intensity": {
151     "counts": {
152       "ct_head_noncontrast": int,
153       "cta_head": int,
154       "ct_perfusion": int,
155       "mra_mrv": int,
156       "conventional_angio": int,
157       "total_neuroimaging": int
158     },
159     "bin_total_neuroimaging": "0"|"1"|"2+",
160     "counting_confidence": "high"|"medium"|"low",
161     "quotes": [string, ...],
162     "comments": "Short note about uncertainties, contradictions across multiple reports, or non-neuro
    imaging present."
163   },
164
165   "evidence_mode": {
166     "mode": "CTA"|"CTP"|"Angiography"|"MRA/MRV"|"CT head only"|"Mixed/Multiple"|"None/No neuroimaging
    ",
167     "quotes": [string, ...],
168     "comments": "Short note about uncertainties, contradictions across multiple reports, or non-neuro
    imaging present."
169   },
170
171   "vasospasm_evidence": {
172     "vasospasm_score_rr": -5..5,
173     "supporting_modalities": ["CTA"|"Angiography"|"CTP"|"MRA/MRV"|"CT head only"|"Unclear", ...],
174     "summary": "1-3 sentence explanation based ONLY on radiology wording",
175     "evidence_quotes": [string, ...],
176     "why_not_lower": { "summary": string, "quotes": [string, ...] },
177     "why_not_higher": { "summary": string, "quotes": [string, ...] },
178     "comments": "Short note about uncertainties, contradictions across multiple reports, non-neuro
    imaging present, or chronic stenosis vs vasospasm guardrail application."

```

```

179 },
180
181 "vasospasm_severity_indicator": {
182   "E_vasospasm_severity": "none"|"mild"|"moderate"|"severe"|"unknown",
183   "evidence_quotes": [string, ...],
184   "comments": "Short note about why severity is unknown vs none, conflicts across time, and how
    chronic stenosis guardrails were applied."
185 },
186
187 "bleed_phenotype": {
188   "SAH": "present"|"absent"|"unknown",
189   "SAH_extent": "trace/small"|"moderate"|"extensive/diffuse"|null,
190   "IVH": "present"|"absent"|"unknown",
191   "IPH": "present"|"absent"|"unknown",
192   "SDH": "present"|"absent"|"unknown",
193   "clot_burden_proxy": 0|1|2|null,
194   "quotes": [string, ...],
195   "comments": "Short note about uncertainties, contradictions across multiple reports, or missing
    SAH extent language."
196 },
197
198 "hydrocephalus_ventricles": {
199   "hydrocephalus": "present"|"absent"|"unknown",
200   "ventricle_description": string|null,
201   "evd_or_ventriculostomy_present": true|false|null,
202   "evd_tip_location": string|null,
203   "quotes": [string, ...],
204   "comments": "Short note about uncertainties, contradictions across multiple reports, or CSF
    diversion terminology."
205 },
206
207 "aneurysm": {
208   "aneurysm_present": true|false|null,
209   "location_text": string|null,
210   "location_group": "anterior"|"posterior"|"unknown",
211   "quotes": [string, ...],
212   "comments": "Short note about uncertainties, contradictions across multiple reports, or post-
    treatment mentions."
213 },
214
215 "supportive_signals": {
216   "aneurysm_treatment_mentioned": true|false|null,
217   "aneurysm_treatment_type": "coil"|"clip"|"other"|null,
218   "mechvent_support": true|false|null,
219   "quotes": [string, ...],
220   "comments": "Short note about uncertainties or non-neuro context for mechvent signals."
221 }
222 }
223
224 SCORING RULES (-5 to +5) FOR vasospasm_score_rr
225 You must output exactly one integer score in [-5..5].
226
227 POSITIVE (vasospasm present)
228 +5: explicit, unambiguous radiology documentation of vasospasm (e.g., CTA/angio states vasospasm
    clearly).
229 +4: strong evidence of vasospasm but slightly indirect; still clearly supports vasospasm.
230 +3: moderate evidence; mention suggests vasospasm but wording leaves some uncertainty.
231 +2: weak evidence; concern/suspicion with some supportive context but not confirmed.
232 +1: very weak leaning toward vasospasm; minimal hints only.
233
234 UNCERTAIN
235 0: vasospasm not mentioned OR evidence too ambiguous/conflicting to lean yes or no.
236
237 NEGATIVE (vasospasm absent)
238 -1: very weak leaning toward no vasospasm; minimal hints only.
239 -2: evidence points toward no vasospasm but absence not explicitly stated.
240 -3: moderate evidence of no vasospasm (e.g., repeated no evidence style statements) but not fully
    definitive.
241 -4: strong evidence of no vasospasm with fairly explicit negation.
242 -5: explicit statement that vasospasm did NOT occur / was ruled out / no evidence of vasospasm.
243
244 SEVERITY MAPPING RULES for E_vasospasm_severity
245 Return exactly one of: "none","mild","moderate","severe","unknown".

```

```

246
247 - "none":
248   * Only if explicit negation of vasospasm/spasm exists (quote it).
249 - "mild"/"moderate"/"severe":
250   * Only if vasospasm/spasm is explicitly documented AND severity is explicitly stated OR described
      with clear severity descriptors.
251   * If only vasospasm/spasm is stated without any severity wording "unknown".
252   * Do NOT map chronic atherosclerotic stenosis/occlusion/plaque to vasospasm severity.
253 - "unknown":
254   * Default if vasospasm is not mentioned, only in indication, ambiguous (cannot exclude), or severity
      is not specified.
255
256 BOUNDARY JUSTIFICATION (MANDATORY)
257 In addition to your score, you MUST justify:
258 - why the score is NOT LOWER
259 - why the score is NOT HIGHER
260 For each:
261 - 13 sentences
262 - 02 SHORT verbatim quotes from the RADIOLOGY text (if available).
263 - If no quote exists, return an empty list and say so in the summary.
264
265 If there are both positive and negative statements across time, choose the score that best reflects
      the overall admission picture,
266 prioritizing later neurovascular imaging, and explain the conflict in comments.
267
268 Remember: STRICT JSON only.

```

Listing F.3: Prompt template for structured extraction from radiology reports.

Discharge-summary vasospasm scoring prompt

The vasospasm scoring prompt was used to assign an ordinal discharge-summary-based vasospasm evidence score. It distinguished explicit vasospasm documentation from prophylaxis, suspicion, absence, and ambiguous statements. The output included supporting quotes and boundary justifications explaining why the score was not lower or higher.

```
1  You are a clinical NLP assistant helping with retrospective research on aneurysmal subarachnoid
   hemorrhage (SAH) in ICU patients.
2
3  You receive one English discharge summary per hospital admission, from the MIMIC-IV-Note dataset. The
   notes are de-identified and may contain three underscores (___) where protected health
   information was removed.
4
5  Your task is to infer whether the patient had a cerebral vasospasm during THIS hospital admission.
6
7  DEFINITIONS
8  - "Cerebral vasospasm" means delayed narrowing of cerebral arteries after SAH, usually clinically
   relevant or radiologically confirmed.
9  - It may be described via:
10 - Terms: "vasospasm", "cerebral vasospasm", "angiographic vasospasm", "delayed cerebral ischemia", "
   DCI", "delayed ischemic neurologic deficit".
11 - Imaging: CTA/angiography showing vasospasm.
12 - Transcranial Doppler (TCD) findings consistent with vasospasm.
13 - Therapies: intra-arterial vasodilators, balloon angioplasty for vasospasm, "triple-H therapy"
   explicitly for vasospasm, etc.
14 - Prophylaxis alone (e.g. "nimodipine for vasospasm prophylaxis") DOES NOT count as vasospasm.
15
16 SCORING RULES (-5 to +5)
17 You must output exactly one integer score in [-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5]:
18
19 POSITIVE (vasospasm present)
20 +5: explicit, unambiguous documentation of vasospasm (imaging or treatment clearly confirms vasospasm)
   .
21 +4: strong evidence of vasospasm but somewhat indirect / less explicit phrasing; still clearly
   supports vasospasm.
22 +3: moderate evidence; mention suggests vasospasm and/or typical treatment but wording leaves some
   uncertainty.
23 +2: weak evidence toward vasospasm; concern/suspicion with some supportive context but not confirmed.
24 +1: very weak leaning toward vasospasm; minimal hints only.
25
26 UNCERTAIN
27 0: vasospasm not mentioned at all OR evidence is too ambiguous/conflicting to lean yes or no.
28
29 NEGATIVE (vasospasm absent)
30 -1: very weak leaning toward no vasospasm; minimal hints only.
31 -2: evidence points toward no vasospasm but absence is not explicitly stated.
32 -3: moderate evidence of no vasospasm (e.g. multiple normal studies/monitoring statements) but wording
   not fully definitive.
33 -4: strong evidence of no vasospasm with fairly explicit negation.
34 -5: explicit statement that vasospasm did NOT occur / was ruled out / no evidence of vasospasm.
35
36 EVIDENCE QUOTES (VERBATIM ONLY)
37 - Provide 14 SHORT exact quotes copied verbatim from the discharge summary.
38 - NO paraphrasing. NO rewriting. NO mixing fragments. NO ellipses (...).
39 - Each quote must be a contiguous span that appears exactly in the note text.
40
41 BOUNDARY JUSTIFICATION (MANDATORY)
42 In addition to your score, you MUST justify:
43 - why the score is NOT LOWER
44 - why the score is NOT HIGHER
45
46 For each boundary justification:
47 - Provide a short explanation (13 sentences)
48 - Provide 02 SHORT verbatim quotes from the discharge summary (if available).
49 - If no quote exists for that boundary, return an empty list for quotes and say so in the summary.
50
51 OUTPUT FORMAT
52 Always answer in STRICT JSON with the keys:
53 {
54   "vasospasm_score": integer from -5 to 5,
55   "evidence_quotes": [list of 1-4 short verbatim quotes copied from the note],
```

```
56
57   "why_not_lower": {
58     "summary": "short explanation",
59     "quotes": [list of 0-2 short verbatim quotes]
60   },
61
62   "why_not_higher": {
63     "summary": "short explanation",
64     "quotes": [list of 0-2 short verbatim quotes]
65   },
66
67   "comments": "short explanation of your reasoning"
68 }
69
70 - Do NOT include any text outside the JSON object.
71 - Do NOT add trailing commas.
```

Listing F.4: Prompt template for discharge-summary-based vasospasm scoring.

Appendix G

LLM execution framework and runtime control

Execution framework

The preceding sections described the clinical extraction tasks, the role of few-shot examples, and the design of the structured outputs. In the implemented pipeline, these elements are embedded within a broader execution framework that controls how note-level inference is performed, validated, logged, and reintegrated into the admission-level dataset.

At the highest level, each extraction task is defined as a reusable task object specifying the system prompt, the input text column, the expected output schema, the validation function, the flattening logic, and, where applicable, the few-shot formatter and repair behaviour. These task definitions are then executed through shared runner functions that apply the same processing loop to every admission: load the input parquet, construct the prompt for the current row, call the model, validate the response, flatten the result, and checkpoint the updated output parquet. Two closely related execution patterns were used. A general runner handled discharge extraction, radiology extraction, and vasospasm scoring, while a separate runner was used for the two aSAH adjudication tasks, where discharge and radiology evidence were jointly interpreted within a single prompt.

Model and runtime configuration

All extraction tasks were executed through the Vertex AI Gemini interface using a shared client construction function, rather than through a public web interface. This provided a controlled Google Cloud execution setting in which access to the model was managed through a project-level configuration, a fixed region, and explicit runtime parameters. This setup was also preferred from a data-governance perspective, since Vertex AI provides enterprise-oriented controls for generative AI workloads, does not use customer prompts or responses to train models without permission, and documents mechanisms for stricter zero-data-retention configurations when required [37]. Runtime settings were externalised through environment variables and converted into structured configuration objects before execution.

The default runtime configuration used location `europa-west4`, model `gemini-2.5-pro`, thinking budget 4096, temperature 0.0, checkpoint frequency 1, random seed 42, and response MIME type `application/json`. Throttling was controlled through a base sleep of 0.5 seconds, maximum sleep of 120.0 seconds, and relaxation factor of 0.70.

Retry behaviour was configured separately, with a maximum of 5 transport retries, an initial backoff of 5.0 seconds, a maximum backoff of 120.0 seconds, a maximum non-429 backoff of 60.0 seconds, escalation after 5 notes affected by 429 responses, and a 429 sleep multiplier of 2.0.

This configuration supported reproducibility and control. Keeping temperature at or near zero reduces unnecessary stochastic variation across otherwise identical prompts, while explicit JSON response mode aligns model behaviour with the schema-constrained extraction design. The thinking-budget parameter was also treated as part of the runtime specification because it affects how much reasoning capacity is available during extraction.

Inference-time prompt assembly

Although each task uses the same execution framework, prompt assembly remains task-specific. The aSAH adjudication tasks construct prompts by combining discharge summaries and radiology text within a single input, with the source boundaries explicitly marked in the prompt template. In the treatment-driven non-SAHA adjudication task, structured treatment indicators (clipping, coiling, and nimodipine) are injected into the prompt as supportive context, while the note text is explicitly designated as the evidence base. The SAHA-coded adjudication task uses discharge and radiology notes together but does not inject the additional treatment context block.

Discharge extraction and discharge-based vasospasm scoring operate only on discharge summaries, while radiology extraction only operates on radiology bundles. Few-shot examples are inserted before the new note text and are specific to the task.

Schema validation and malformed-output handling

A methodological safeguard in the implemented pipeline is that model outputs are not accepted merely because they appear JSON-like. After every model call, the raw text is first parsed into a JSON object and then passed through a task-specific validation function. These validators check more than top-level field presence. They also enforce score ranges, admissible label sets, type constraints, nested object structure, and length constraints on quote lists. For example, the aSAH adjudication validators require fields such as `asah_likelihood_score`, `final_label`, `support_source`, signal sub-objects, evidence quotes, and boundary-justification objects, and they reject outputs with invalid ranges or malformed nested fields. Similar validators are used for discharge extraction, radiology extraction, and vasospasm scoring.

Repair prompting

For tasks configured with repair enabled, a second-pass repair mechanism is used when the initial response fails JSON parsing or validation. In such cases, the framework constructs a repair prompt containing the original note text, the invalid output, and the same few-shot block used in the original call. The repair instruction requests a single strict JSON object matching the schema given in the system instructions, with no commentary and no extra keys. The repaired output is then parsed and validated again before acceptance.

Repair prompting is not meant to reinterpret the clinical note or change the target task, but to recover outputs that are substantively plausible yet formally invalid because of formatting errors or minor schema deviations. This is especially useful for the most complex structured tasks, discharge extraction, where nested outputs increase the chance of malformed JSON.

Retry, throttling, and rate-limit control

Large-scale note extraction can be affected not only by semantic model errors, but also by transport failures and rate limits. To address this, the execution framework includes an explicit retry layer with separate handling for rate-limit errors and non-rate-limit errors. Calls are retried with exponential backoff. When repeated 429 or `RESOURCE_EXHAUSTED` responses occur, the framework increases a persistent inter-note sleep interval through a throttle object. After successful calls, this sleep interval relaxes gradually toward a configured baseline.

Token-budget control and overlength-note handling

For tasks using few-shot prompting within the generic runner, prompt length is actively managed through an optional token-budget mechanism. Before inference, the framework estimates the combined size of the system prompt, few-shot block, and current note text. If the estimated prompt exceeds the configured input-token budget, the system first reduces the number of few-shot examples included in the prompt. If that is not enough and truncation is permitted by policy, the note text is shortened to a configured number of characters and the prompt is rebuilt. If the budget is still exceeded, the note is not silently discarded. It is assigned a dedicated `skip_too_long` status and logged separately.

Output flattening and integration into the dataset

Once a response passes validation, it is transformed in two parallel ways. First, the raw structured output is retained as a JSON string in the task output parquet. Second, a task-specific flattening function converts selected nested fields into admission-level columns that can be joined directly into the downstream analysis dataset.

Processing statuses, logging, and resumable execution

Each processed admission is assigned an explicit processing status. Depending on the task and outcome, rows may be marked as `ok`, `fail`, `no_text`, or `skip_too_long`. These statuses are written into the output parquet together with metadata such as runtime, model name, note length, whether repair was used, and comments or error messages. In parallel, the framework maintains task-specific debug logs and bad-output logs, and for overlength cases it can also write a dedicated skip log.

The extraction pipeline is also resumable. If an output parquet already exists, it is loaded at the beginning of a run, indexed by admission key, and rows whose latest status is already `ok` are skipped. New or updated results are upserted into the in-memory results list and checkpointed to parquet after a configurable number of newly processed rows. This avoids unnecessary recomputation and makes the extraction process more robust to interruption.

Appendix H

Background knowledge constraints used for causal discovery

The background knowledge used in causal discovery consisted of two layers. The first layer was the tier-based temporal ordering shown in Table I.1. The second layer consisted of additional forbidden directions that were added because the prohibited direction would conflict with the clinical timing or definitional meaning of the variables.

The tier ordering was:

baseline $\rightarrow$ severity $\rightarrow$ workflow $\rightarrow$ outcome.

For every pair of variables assigned to different tiers, directed edges from a later tier to an earlier tier were forbidden. The additional non-tier restrictions are listed in Table H.1.

Constraint group	Forbidden direction(s)
Protected demographic variables	$X \rightarrow \text{age}$ $X \rightarrow \text{sex}$ for all other variables X
Hydrocephalus and EVD ordering	$\text{evd} \rightarrow \text{hydrocephalus}$
Vasospasm severity definition	$\text{vasospasm_severity} \rightarrow \text{vasospasm}$
Aneurysm variables protected from downstream variables	$X \rightarrow \text{aneurysm_present}$ $X \rightarrow \text{aneurysm_territory}$ where X is a non-aneurysm severity, workflow, complication, or outcome variable
Securing technique restrictions	$\text{securing_technique} \rightarrow \text{hydrocephalus}$ $\text{securing_technique} \rightarrow \text{evd}$ $\text{securing_technique} \rightarrow \text{dysnatremia_state}$ $\text{securing_technique} \rightarrow \text{hypotension_state}$ $\text{securing_technique} \rightarrow \text{hypoxia_state}$ $\text{securing_technique} \rightarrow \text{hypercapnia_state}$ $\text{securing_technique} \rightarrow \text{leukocytosis_state}$ $\text{securing_technique} \rightarrow \text{hyperglycemia_state}$ $\text{securing_technique} \rightarrow \text{hypomagnesemia_state}$ $\text{securing_technique} \rightarrow \text{rhythm_class}$ $\text{securing_technique} \rightarrow \text{mechanical_ventilation}$

Table H.1: Additional background-knowledge restrictions used in causal discovery, beyond the tier-based temporal ordering.

Appendix I

Causal discovery variable sets

This appendix summarizes the candidate variables considered for causal discovery. The methodology chapter describes the selection logic; the tables below provide the corresponding variable-level overview.

Table I.1: Candidate variables considered for causal discovery.

Variable	Tier	Representation
age	Baseline	Continuous or binned
sex	Baseline	Binary / categorical
smoking	Baseline	Binary / categorical
hypertension	Baseline	Binary
diabetes	Baseline	Binary
alcohol_history	Baseline	Binary / categorical
heart_failure	Baseline	Binary
prior_stroke	Baseline	Binary
epilepsy	Baseline	Binary
clot_burden	Severity	Ordinal / categorical
aneurysm_territory	Severity	Categorical
ivh	Severity	Binary
iph	Severity	Binary
sdh	Severity	Binary
hydrocephalus	Severity	Binary
worst_pre_treatment_gcs	Severity	Ordinal / categorical
evd	Workflow	Binary
dysnatremia_state	Workflow	Categorical / ordinal
hypotension_state	Workflow	Categorical / ordinal
hypoxia_state	Workflow	Categorical / ordinal
hypercapnia_state	Workflow	Categorical / ordinal
leukocytosis_state	Workflow	Categorical / ordinal
hyperglycemia_state	Workflow	Categorical / ordinal
hypomagnesemia_state	Workflow	Categorical / ordinal

Continued on next page

Table I.1: Candidate variables considered for causal discovery. *Continued*

Variable	Tier	Representation
rhythm_class	Workflow	Categorical
mechanical_ventilation	Workflow	Binary / categorical
securing_technique	Workflow	Categorical
treatment_delay	Workflow	Categorical / ordinal
nimodipine	Workflow	Binary
neuroimaging_intensity	Workflow	Count / categorical
vasospasm	Outcome	Binary
vasospasm_severity	Outcome	Ordinal / categorical

The exact retained set differed by dataset policy. Variables were removed when they were not available in the relevant source scope, had excessive missingness, showed too little variation, or were excluded as part of a sensitivity policy. The main variable set included all variables except `neuroimaging_intensity` and `worst_pre_treatment_gcs`.

Appendix J

Causal discovery hyperparameter grids

The methodology chapter summarizes the hyperparameter sweep design. This appendix lists the exact full grids instantiated in the sweep registry.

Table J.1: Full PC-stable hyperparameter sweep.

Parameter	Values
Significance level α	{0.001, 0.0025, 0.005, 0.01, 0.02, 0.05, 0.1}
Conditioning depth	{1, 2, 3, 4, -1}
<code>stable_fas</code>	<code>True</code>
<code>allow_bidirected</code>	<code>False</code>
G^2 minimum cell count	{1, 2, 3, 4, 5}
G^2 cell table type	2

Table J.2: Full FGES hyperparameter sweep.

Parameter	Values
<code>symmetric_first_step</code>	{ <code>False</code> , <code>True</code> }
Maximum degree	{2, 3, 4, 6, -1}
<code>parallelized</code>	<code>True</code>
<code>faithfulness_assumed</code>	{ <code>True</code> , <code>False</code> }
BDeu sample prior	{10.0, 5.0, 20.0}
BDeu structure prior	{0.0, 1.0}

Table J.3: Full FCIT hyperparameter sweep.

Parameter	Values
Initial CPDAG search	Tetrad default, corresponding to BOSS in the Tetrad version used
Significance level α	{0.01, 0.005, 0.001}
Conditioning depth	{3, -1}
Maximum discriminating-path length	{3, -1}
Maximum blocking-path length	{3, -1}
<code>num_starts</code>	1
<code>guarantee_pag</code>	True
G^2 minimum cell count	{1, 4}
G^2 cell table type	2
BDeu sample prior	10.0
BDeu structure prior	0.0

Appendix K

Causal discovery output aggregation

The causal discovery workflow produced a large number of run-level graphs across algorithms, dataset policies, hyperparameter settings, and bootstrap samples. This appendix gives an overview of the aggregation outputs that were used to turn these run-level artefacts into comparative result tables. Most aggregation products were saved in both `.csv` and `.parquet` format.

Global and run-level summaries

Each completed run contributed one row to a run summary table. This table recorded the algorithm, dataset policy, hyperparameter configuration, completion status, and basic graph-level information, including the number of discovered edges. Where bootstrap results were available, the run summary also contained compact stability information, such as the number of bootstrap-derived stability rows and summary statistics of adjacency frequencies.

The aggregation workflow also produced global summary files. These files gave a compact overview of each aggregated output directory, including the number of runs, the available aggregation products, and broad graph-level characteristics. They were mainly used as sanity checks before working with the more detailed edge, stability, sensitivity, and path summaries.

Main-edge summaries

Edges from each run-level graph were converted into a standardized long-format edge table. Each edge record included the two endpoint variables, an unordered variable-pair identifier, the endpoint pattern, and an oriented-edge label when orientation was available.

Two recurrence summaries were computed from these edge tables:

1. pair-level recurrence, counting how often an unordered variable pair appeared across runs;
2. oriented-edge recurrence, counting how often the same endpoint pattern appeared across runs.

Bootstrap stability summaries

Bootstrap outputs were aggregated separately from the graph edges estimated on the original dataset. For each unordered variable pair, the workflow summarized:

- adjacency frequency across bootstrap samples;
- the dominant endpoint pattern;
- the frequency of the dominant endpoint pattern overall;
- the frequency of the dominant endpoint pattern among bootstrap samples where the adjacency was present;

- the assigned stability class.

An adjacency was treated as stable when it appeared in at least 0.70 of bootstrap runs. Among stable adjacencies, an orientation was treated as stable when the dominant endpoint pattern appeared in at least 0.80 of the runs in which the adjacency was present.

The resulting stability classes were:

1. `stable_adjacency_stable_direction`;
2. `stable_adjacency_unstable_direction`;
3. `unstable_adjacency`.

The workflow saved both long-format stability tables and summarized stability tables. The long-format files retained pair-level bootstrap information per run or configuration. The summary files aggregated these values to make comparison across algorithm settings more manageable.

Regime-based sensitivity summaries

Hyperparameter sensitivity was examined by grouping runs into regimes defined by shared parameter values, such as α , conditioning depth, maximum degree, BDeu sample prior, BDeu structure prior, faithfulness, symmetric-first-step settings, discretisation settings, or FCIT-specific options. Within each regime, the workflow summarized how often each unordered pair and each oriented edge appeared.

For each regime, the summaries included:

- the fraction of runs containing the unordered pair;
- the fraction of runs containing the oriented edge;
- the orientation fraction among runs where the pair was present.

The aggregation also produced edge-change tables for selected hyperparameters. These tables described how edge recurrence varied across parameter regimes, which helped identify relations that were mainly seen under particular settings.

In addition to the general sensitivity summaries, vasospasm-focused sensitivity files were produced. These applied the same type of regime-based comparison to edges or pairs involving the vasospasm-related target variables. This made it possible to inspect whether the clinical endpoint of interest was affected differently from the graph as a whole.

Path-based summaries

Because vasospasm was the clinical endpoint of interest, additional summaries focused on paths involving `vasospasm` and, where included in the search target set, `vasospasm_severity`. For each run, the edge list was converted into a graph representation and searched for simple paths leading to these target nodes up to a bounded length.

The directed path outputs summarized fully directed paths toward the target. These outputs were written as long-format path tables, run-level path summaries, path recurrence summaries, and prefix-path summaries. The prefix-path summaries were useful for finding recurring upstream segments of longer paths, even when the complete path did not recur in exactly the same form.

For FCIT, the workflow also produced semidirected path summaries. These were used for PAG-style outputs, where partially oriented edge marks can still allow traversal toward the target without requiring every edge to be fully directed. The semidirected path files were therefore

interpreted as structural reachability summaries under partial orientation, not as fully identified directed causal pathways.

The path summaries were used as descriptive summaries of recurring graph structure. They were not used to estimate path-specific causal effects.

Appendix L

Use of generative artificial intelligence

Generative artificial intelligence tools were used in a limited and supportive role during the preparation of this thesis.

Language models such as ChatGPT and GitHub Copilot were used to assist with tasks such as improving the clarity of written text, suggesting alternative formulations, autocompleting code, and helping to identify relevant programming functionalities. These tools were used during development, but all code and analysis were reviewed and adapted manually.

In addition, language models were used to help interpret unfamiliar medical terminology encountered in clinical notes and literature. These explanations were treated as guidance and were cross-checked against primary sources or clinical references before being incorporated into the work.

Research-oriented tools such as Research Rabbit were used to explore the literature, identify relevant publications, and navigate citation networks. All selected sources were evaluated and read by the author.

Generative artificial intelligence was also part of the research methodology itself. In particular, large language models, including Gemini, were used within the information extraction pipeline described in the methodology chapter. This use is distinct from the supportive role described above and is documented in detail in the thesis.

All core elements of the work, including the research design, data processing pipeline, variable construction, analysis, and interpretation of results, were developed independently by the author. Outputs generated by AI tools were critically reviewed and only incorporated when they aligned with the intended meaning and methodology. No content was included without verification.