

441P Real-world characteristics and outcomes in a NATALEE-eligible population at Akershus University Hospital: AI-based data extraction from clinical records

Agnar Martin Bjørnstad^{1,2}, S. Aas^{1,2}, Ø. Sundby³, B. Paulsson⁴, C. Hentschke⁵, S. B. Geisler⁶, J. Geisler^{1,6}, A. Ranjbar², J. Ravn²

¹Faculty of Medicine, University of Oslo, Oslo, Norway; ²Medical Technology & E-health, Akershus University Hospital HF, Lørenskog, Norway; ³Medical Affairs, Novartis Norway, Oslo, Norway; ⁴Medical Affairs, Novartis AB, Kista, Sweden; ⁵Evidence Generation, Novartis Pharma GmbH, Nürnberg, Germany; ⁶Department of Oncology, Akershus University Hospital HF, Lørenskog, Norway.

2026 **ESMO BREAST CANCER**
Annual Congress

BERLIN GERMANY
6-8 MAY 2026



BACKGROUND

Real-world observational data in early breast cancer are essential for monitoring patient outcomes and validating treatment effects observed in clinical trials. However, retrospective extraction of data from hospital systems is challenging due to large volumes of sensitive data, heterogeneous data structures, siloed storage, and changes in clinical practice, reporting standards, and IT infrastructure over time. Advances in artificial intelligence offer new opportunities to facilitate extraction and structuring of real-world clinical data.

In this study, we applied an AI-based model to extract and structure real-world clinical and pathology data from patients with HR+/HER2- early breast cancer fulfilling the eligibility criteria of the NATALEE trial. Using this approach, we characterized NATALEE-eligible patients in a Norwegian real-world cohort and descriptively evaluated recurrence patterns across stage and nodal subgroups.

METHODS

Patients retrospectively identified through the breast cancer diagnostic pathway at Akershus University Hospital (Jan 2015–Aug 2024) were included. Data were extracted from EHR and pathology reports, comprising structured tabular data, semi-structured text, and scanned documents. EHR data were obtained directly, while pathology data were extracted from tabular data, semi-structured text, and scanned reports, as detailed in Figure 2. Patients with invasive eBC and surgery ≤12 months from diagnosis were included. Patients with distant recurrence ≤3 months post-surgery were excluded to avoid occult metastatic disease. Eligibility was defined per NATALEE criteria (HR+/HER2-, stage II–III or high-risk N0); only patients fulfilling NATALEE criteria were included in the analyses [1]. Variables used for high-risk N0 classification (grade, Prosigna, KI-67) were conditionally required; missingness was defined only when these variables were necessary for eligibility assessment.

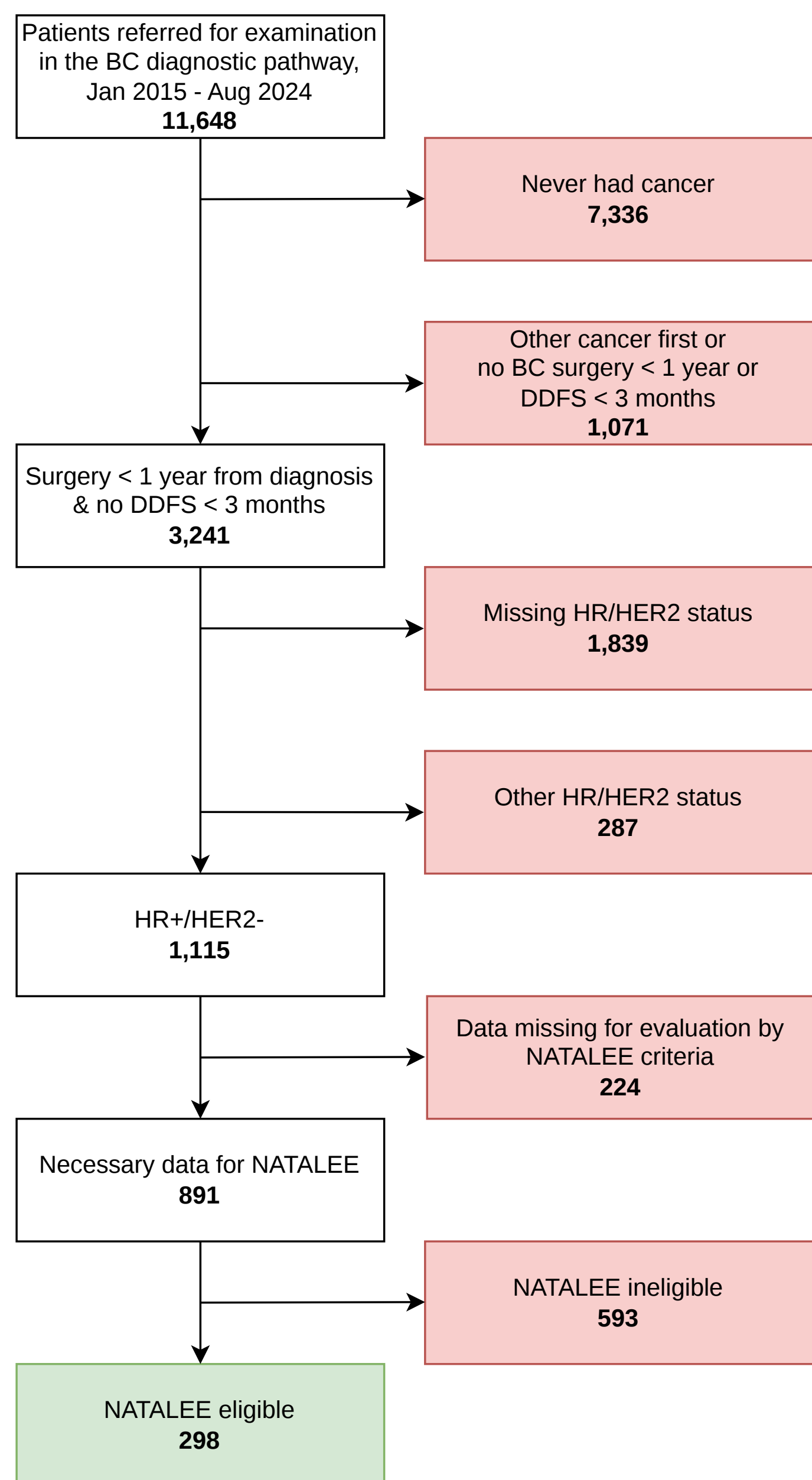
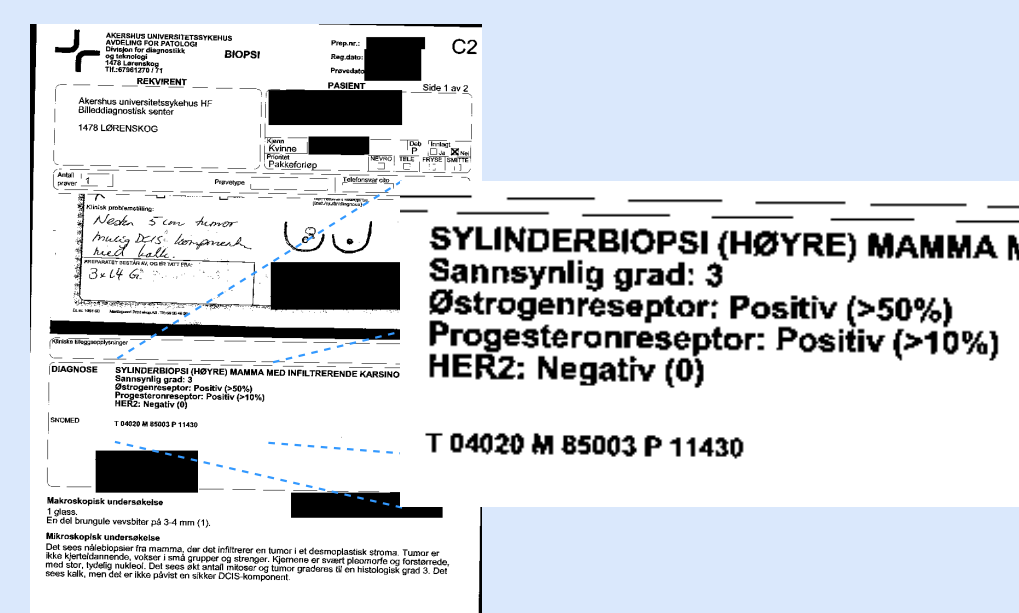


Figure 1: Flowchart of inclusion and exclusion criteria for patients enrolled in the breast cancer (BC) diagnostic pathway at Akershus University Hospital from Jan 2015 to Aug 2024. Patients were retrospectively identified from existing hospital records; no prospective screening was performed.

Scanned pathology reports



Tabular data

Patient ID	Sample ID	Test Name	Test Result
2	#32	ER	40%
2	#32	PGR	50%
3	#37	HER2	1+
...	...	...	...

Semi-structured text

Ablation (mastectomy) of the right breast with infiltrating carcinoma, NST (no special type).
Tumor diameter: approx. 5.2 mm
Tumor grade: G1 (score: 1+1+1 = 3 points)
Mitoses (per mm²): 0
Resection margins: clear, with good distance

Biomarkers:
Estrogen receptor: positive (approx. 70%)
Progesterone receptor: positive (approx. 70%)
HER2: negative (IHC: 1+)
Ki-67 index (hot spot): < 5% (see assessment)

One sentinel lymph node without detected metastasis (0/1 SN, r. axilla)

pT1bN0(sn)Mx

Axillary resection (left) with nine examined lymph nodes, of which one contains metastasis from breast carcinoma (1/9).
Metastasis largest diameter: 8 mm
Perinodal growth: Early perinodal growth detected, 1.1 mm along the capsule, and 0.6 mm extending beyond the capsule.

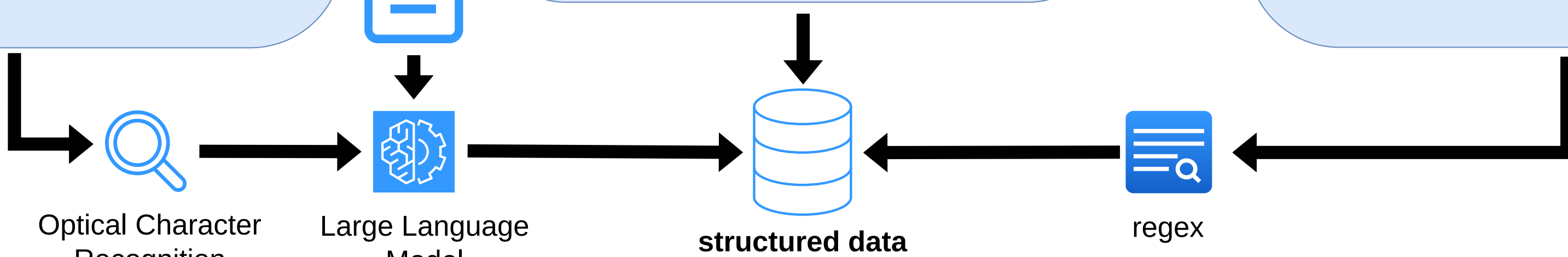


Figure 2: Pipeline for extracting pathology test results, with examples from each data modality. Structured fields were used directly, while semi-structured text and scanned pathology reports required automated extraction. Text from scanned pathology reports was obtained using PaddleOCR v3.3 [2], followed by structured extraction using Qwen3-4B-Thinking [3].

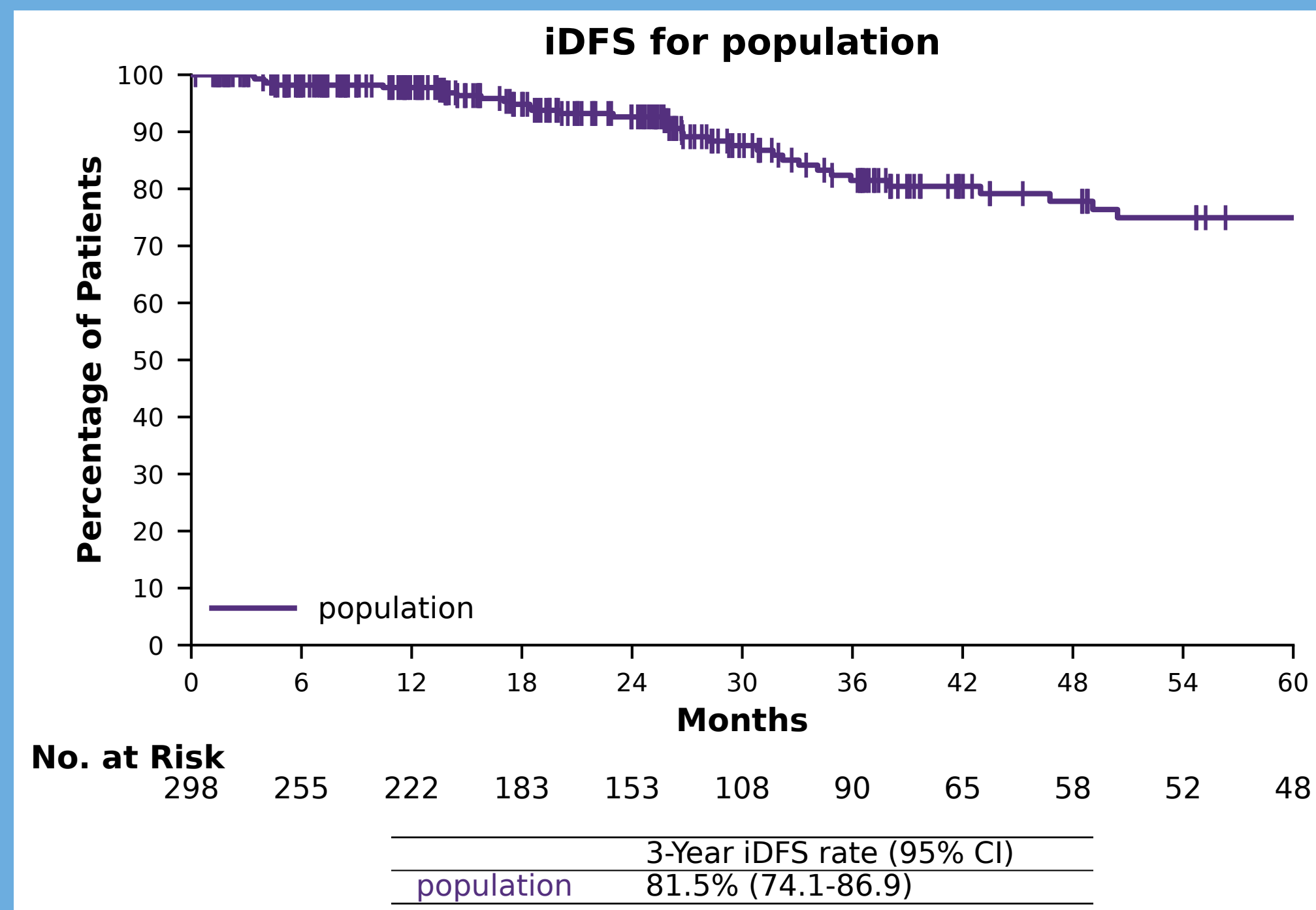


Figure 3: Exploratory invasive disease-free survival outcome in the overall study population.

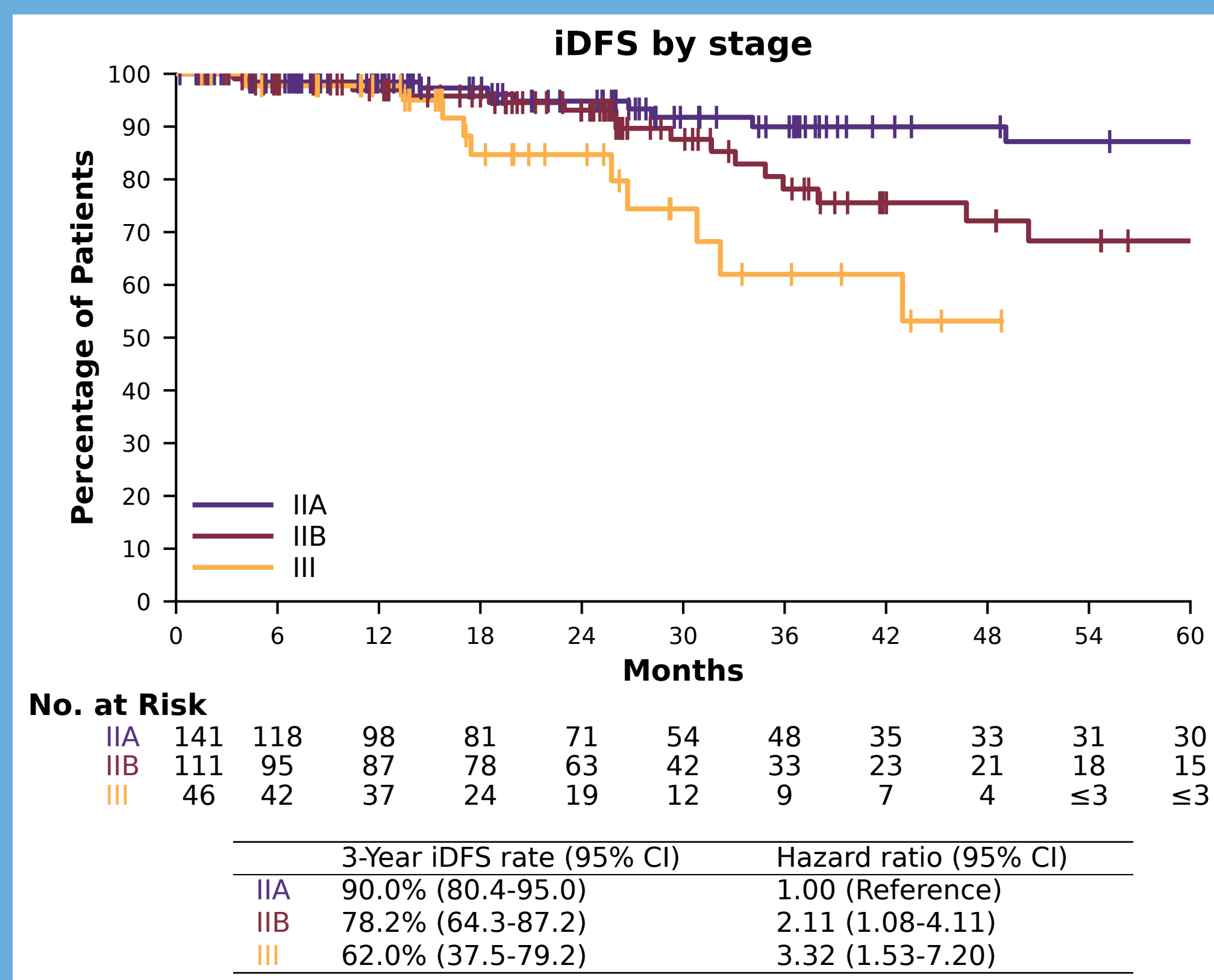


Figure 4: Exploratory invasive disease-free survival outcomes stratified by disease stage.

Table 1: Baseline characteristics of the included cohort. Observation time was defined as the period from date of surgery to last documented contact or death.

	population
Patients	298
Age (median, (IQR))	58 (49-68)
Observation time in years (median, (IQR))	2.0 (0.9–3.5)
Stage	
IIA	141 (47%)
IIB	111 (37%)
III	46 (15%)
Nodal status	
N0	69 (23%)
N1	195 (65%)
N2+	34 (11%)

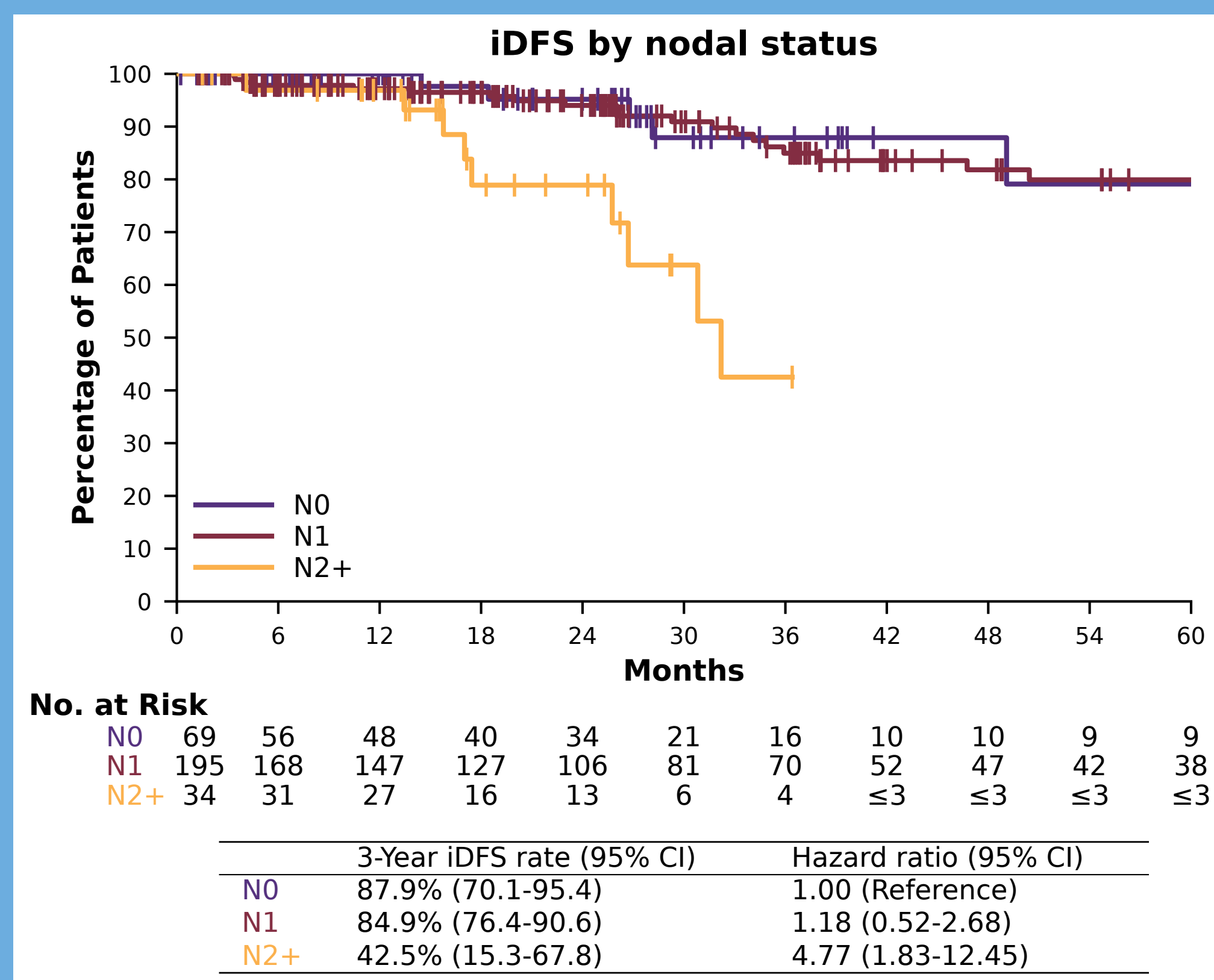


Figure 5: Exploratory invasive disease-free survival outcomes stratified by nodal status.

RESULTS

Table 2: Extraction performance evaluated against manually annotated reference standards. Performance metrics are reported separately by extraction component and data modality, using OCR+LLM on 125 randomly selected scanned pathology reports and a regex-based pipeline on 500 randomly selected free-text pathology fields.

	OCR + LLM			regex		
	Cond. acc. ¹	Prec.	Rec.	Cond. acc. ¹	Prec.	Rec.
T	100	56	83	100	100	79
N	100	82	95	88	100	53
ER	100	98	93	100	100	84
PGR	98	98	93	100	100	79
HER2	95	100	91	100	100	80
KI67	100	100	93	100	100	85
GRADE	100	95	90	100	100	76
Prosigna	-	-	-	100	100	100

¹Calculated on correctly recalled cases.

Missing variables required for NATALEE classification were more common in earlier years, reducing the evaluable cohort to 298 patients, 76% of whom were diagnosed from 2020 onwards. Baseline characteristics are summarized in Table 1. In total, 4,995 scanned pathology reports, text from 5,351 unique sample IDs, and 19,380 test results stored in tables were used in this project. Extraction performance is presented in Table 2. The regex-based pipeline achieved high overall precision but demonstrated lower recall for variables frequently described in free-text, particularly T and N stage. The LLM-based method had a higher recall for these variables, although at the expense of reduced precision for most categories. Kaplan-Meier analysis showed a 3-year iDFS rate of 81.5% for the full cohort. iDFS varied across stage groups and by nodal status, with lower survival observed in patients with more advanced stage and nodal involvement.

DISCUSSION

The findings highlight a trade-off between rule-based and generative extraction approaches. The regex-based pipeline achieved high precision but lower recall for variables such as T and N status, likely due to free-text documentation, whereas the LLM-based method improved recall but reduced precision, particularly for T status. Stage-stratified analyses used combined outputs from both pipelines and tabular data, leveraging complementary strengths across modalities. Missingness was more common in earlier calendar years and may have influenced cohort composition. The results support the feasibility of AI-assisted real-world data extraction for cohort characterization, while underscoring challenges related to data completeness and heterogeneous clinical documentation. Findings should be interpreted in light of limited observation time and missing historical data inherent to retrospective analyses.

CONCLUSION

Overall, application of an AI-based model for automated extraction of real-world data enabled characterization of a NATALEE-eligible HR+/HER2- early breast cancer population. Exploratory iDFS patterns were directionally consistent with published outcomes from contemporary adjuvant endocrine therapy control arms, including NATALEE [4]. Despite limited observation time and retrospective data constraints, this automated approach proved useful for extracting unstructured clinical data. Recurrence risk remained substantial across stage and nodal subgroups, including patients with high-risk N0 disease.

REFERENCES

- Slamon, D. J. *et al.* Rationale and trial design of NATALEE: a Phase III trial of adjuvant ribociclib + endocrine therapy versus endocrine therapy alone in patients with HR+/HER2- early breast cancer. *Therapeutic Advances in Medical Oncology* **15**, 17588359231178125 (2023).
- Cui, C. *et al.* PaddleOCR 3.0 Technical Report. *arXiv preprint arXiv:2507.05595* (2025).
- Yang, A. *et al.* Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).
- Crown, J. *et al.* Adjuvant ribociclib plus nonsteroidal aromatase inhibitor therapy in patients with HR-positive/HER2-negative early breast cancer: 5-year follow-up of NATALEE efficacy outcomes and updated overall survival. *ESMO Open* **10**, 105858 (2025).