



Public Research Module 2 Report

Authors

● Patricia Honorato Moreira - Estudante [Inteli]
● Flavia Santoro - Coorientadora [Inteli]
● Luciana Rodrigues Carvalho Barros - Orientadora [Fundação Faculdade de Medicina]
● Roger Chammas - Coorientador [Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP)]

1. Research Area

This research focuses on the development of **interpretable and clinically applicable machine learning models** to **predict recurrence and survival outcomes** in breast cancer patients.

The models are designed according to **TRIPOD** reporting guidelines and integrate clinical and hematological data to enhance prognostic accuracy.

In this module, the project transitioned to a new scope, expanding the analysis from a single **10-year recurrence horizon** to a **multi-horizon framework** that evaluates recurrence at **2, 5, and 10 years**.

This pivot required manuscript restructuring, reformatted figures, and new modeling preparation steps for the next module.

2. Overview of Sprints

Sprint 1

Objective: Establish target journal and produce the initial full manuscript draft.

Activities:

- Authored the **complete raw manuscript**(Introduction, Methods, Results, Discussion, Conclusions).
- Defined the **target journal (Cancer Research - AACR)** and publication window (**December 2025**).
- Outlined **format constraints** (≤ 250 -word abstract, $\leq 5,000$ -word text, ≤ 8 display items, ≤ 50 references).
- Re-scoped the module to prioritize **writing and formatting** before adding new analyses.

Outcome:

A comprehensive unformatted manuscript draft was completed and serves as the foundation for subsequent revisions.

Sprint 2

Objective: Advance manuscript structure and build reproducible model artifacts per tumor subtype.

Activities:

- Wrote the **Introduction, Abstract, and Statement of Significance** following **Cancer Research** requirements.
- Developed **three Jupyter notebooks per subtype (HR+, HER2+, TNBC)**:
 - 1. Data Preparation
 - 2. Model Development
 - 3. Model Evaluation

- Implemented metrics (AUC, AUPR, calibration) and explainability (SHAP).
- Version-controlled all notebooks in GitHub and mirrored to cloud storage.

Outcome:

A structured and transparent modeling framework was established for each subtype.

Sprint 3

Objective: Expand the manuscript with detailed Methods and Results sections.

Activities:

- Authored the full **Methods** section, describing:
 - Cohort selection, ethical compliance, and preprocessing workflow.
 - Feature engineering (e.g., NLR, PLR, MLR ratios, PCA/UMAP/GMM latent features).
 - Modeling strategies (Random Forest, XGBoost, SVM, and Stacked Ensemble).
 - Validation procedures and performance metrics.
- Wrote **Results** describing:
 - Subtype-specific model performance (AUC ~0.75–0.85).
 - Feature selection workflow and Kaplan–Meier survival stratification.

Outcome:

The manuscript achieved full technical completeness and methodological transparency.

Sprint 4

Objective: Refine manuscript text, prepare figures, and organize expanded dataset.

Activities:

- Rewrote and refined **Discussion** for clarity and coherence.
- Standardized **figure legends** for all panels:
- ****Figure 1:**** Workflow and modeling pipeline.

- ****Figure 2:**** Clinical and hematologic features table.
- ****Figures 3–6:**** ROC curves for general and subtype-specific models (HR+, HER2+, TNBC).
- Conducted **dataset expansion (+200 patients)** for the general model, including data validation.
- Reformatted and documented ****Jupyter notebooks**** for public release:
- Subtype-specific models and ****general recurrence model (≤ 10 years)****.
- Added a **dedicated notebook for early recurrence (≤ 2 years)**.
- Planned integration of updated results in the next sprint.

Outcome:

The manuscript and datasets were technically prepared for the upcoming pivot in modeling scope.

Sprint 5

Objective: Prepare the transition to multi-horizon modeling (2-, 5-, and 10-year recurrence) and reformat project notebooks for publication.

Activities:

- Defined a ****new modeling objective**** to handle recurrence at ****multiple time windows**** (2, 5, 10 years).
- Revised the ****manuscript structure and placeholders**** to reflect the new scope.
- Reformatted and standardized all ****notebooks**** for clarity, modularity, and publication readiness:
- General and subtype-specific models (HR+, HER2+, TNBC) for up to 10 years.
- Added an ****“early recurrence” notebook (≤ 2 years)****.
- Adjusted the ****data pipeline**** and labeling functions to support multi-horizon recurrence modeling.
- Verified data consistency across time horizons.

- Drafted internal notes on the reorganization of figures and the Discussion section for the next module.

Note:

- *The ****manuscript is still a working draft****, pending review by all collaborators.*
- *The ****figures and discussion sections are incomplete**** and will be rewritten in the next module (Module 4).*
- *****SMOTE analyses and final metrics have not yet been conducted****.*

Outcome:

The sprint established the methodological and organizational foundation for the multi-horizon models that will be developed in the next phase.

Project Status	Status	Notes
Dataset	✅ Ready	Includes recurrence labels for 2-, 5-, and 10-year horizons.
Notebooks (General + Subtypes + Early Recurrence)	✅ Reformatted	Modularized and documented for publication.
Manuscript	🟡 In progress	Text revised to reflect new modeling scope; not yet finalized.
Figures & Discussion	⚠️ Incomplete	Will be rewritten after new model analyses in Module 4.
SMOTE Analyses	⌚ Pending	To be conducted after retraining models for all time horizons.

4. Conclusions

Module 3 marked a significant **pivot in project direction**, shifting from single-horizon (10-year) modeling to a **multi-horizon recurrence framework (2, 5, and 10 years)**.

Key outcomes include:

- Reformatted and standardized notebooks for transparency and publication.
- Dataset preparation and labeling for time-window modeling.
- Revised manuscript structure aligned with the new analytical goals.
- Technical groundwork completed for final model development and manuscript completion in the next module.

Although the manuscript and figures remain incomplete, the project is now strategically positioned for the ****Finalization Phase (Module 4)****, where analyses and writing will converge toward submission readiness.

5. Next Steps (Module 4 – Finalization Phase)

- Train and evaluate models for ****2, 5, and 10-year recurrence horizons****.
- Conduct ****SMOTE and cross-validation analyses**** across general and subtype models.
- Regenerate ****figures and tables**** for all time horizons.
- Rewrite the ****Discussion**** to integrate new results and clinical interpretation.
- Finalize ****References**** and perform complete formatting check.
- Submit for ****team-wide review**** prior to journal submission.

6. Acknowledgments

This project is developed in collaboration between **Inteli – Instituto de Tecnologia e Liderança, Fundação Faculdade de Medicina**, and **Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP)**.

We thank the supervisors and data management teams for their continued support and technical contributions.

Report Version: Public Module 3 Report

Date: October 2025

Prepared by: Patricia Honorato Moreira

Supervised by: Luciana R. C. Barros, Flavia Santoro, Roger Chammas