

Development and validation of EHR-based
machine learning models for predicting
survival and recurrence after radical surgery for
colorectal cancer

Sergey S. Gordeyev

N.N.Blokhin Russian Cancer Research Center

Machine Learning (ML) based prognostic models work for CRC

- Recurrence rates for CRC: ~7 % (stage I) to ~29 % (stage III) after curative surgery [1]
- Prior ML models show strong performance for OS:
 - CatBoost — AUC **0.92** for recurrence prediction [2]
 - XGBoost with longitudinal CEA — AUC **0.94** [3]
 - 5-year survival models — AUC **0.85–0.87** [4]
 - Largest cohort (31 916 patients): XGBoost AUC **0.86** [5]

[1] Nors J, et al. Incidence of recurrence and time to recurrence in stage I to III colorectal cancer: a nationwide Danish cohort study. JAMA Oncol. 2023;9(12):1735-1744. [2] Kayikcioglu E, et al. Machine learning for predicting colon cancer recurrence. Surg Oncol. 2024;54:102079

[3] Rodriguez PJ, et al. Using machine learning to leverage biomarker change and predict colorectal cancer recurrence. JCO Clin Cancer Inform. 2023;7:e2300066

[4] Susič D, et al. Artificial intelligence based personalized predictive survival among colorectal cancer patients. Comput Methods Programs Biomed. 2023;231:107435

[5] Buk Cardoso L, et al. Machine learning for predicting survival of colorectal cancer patients. Sci Rep. 2023;13(1):8874

ML-based models are complex, unique and difficult to validate



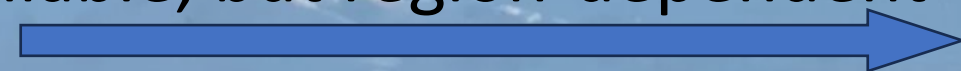
- *like snowflakes*
- Over-reliance on familiar data
- Overfitting problems
- Lack of models with external validation

The way ahead



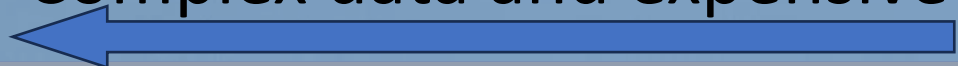
Electronic health records

- Simple and available, but region-dependent



Radiomics, pathomics, genomics

- Complex data and expensive calculations



The Data

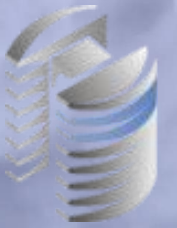
- 2 centers: 1 tertiary & 1 regional
- Accrual period: 01.01.2017–31.12.2024
- 42 clinicopathological and laboratory variables
- Split-sample validation 80:20 (140 recurrence events)
- Missing data:
 - Categorical variables: mode imputation (most frequent category)
 - Continuous variables: multiple imputation by chained equations (MICE) with LightGBM-based models (miceforest), 10 iterations

Inclusion:

- stage I–III CRC
- curative-intent resection
- histologically confirmed adenocarcinoma

Exclusion:

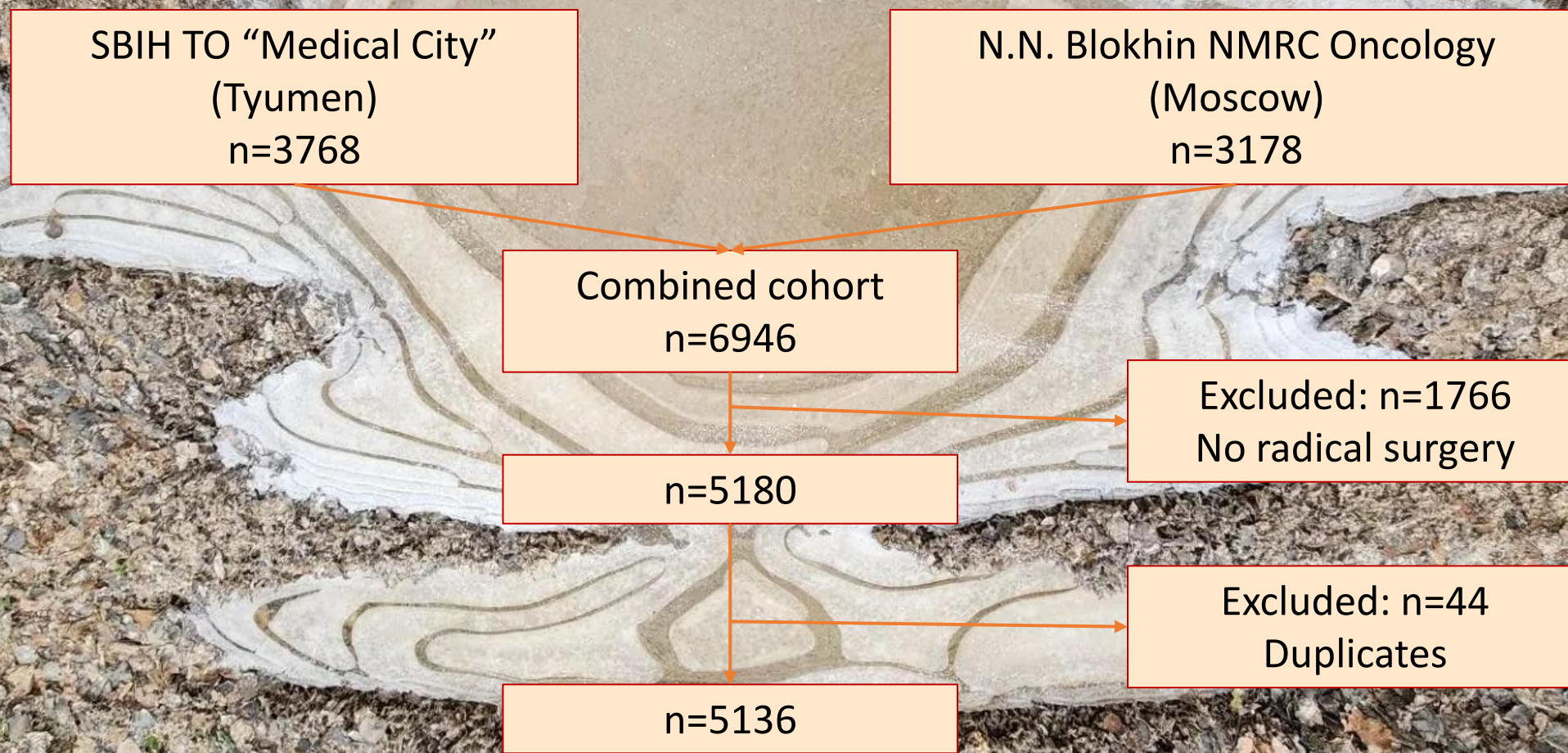
- multiple primary cancers
- palliative (R2) resections



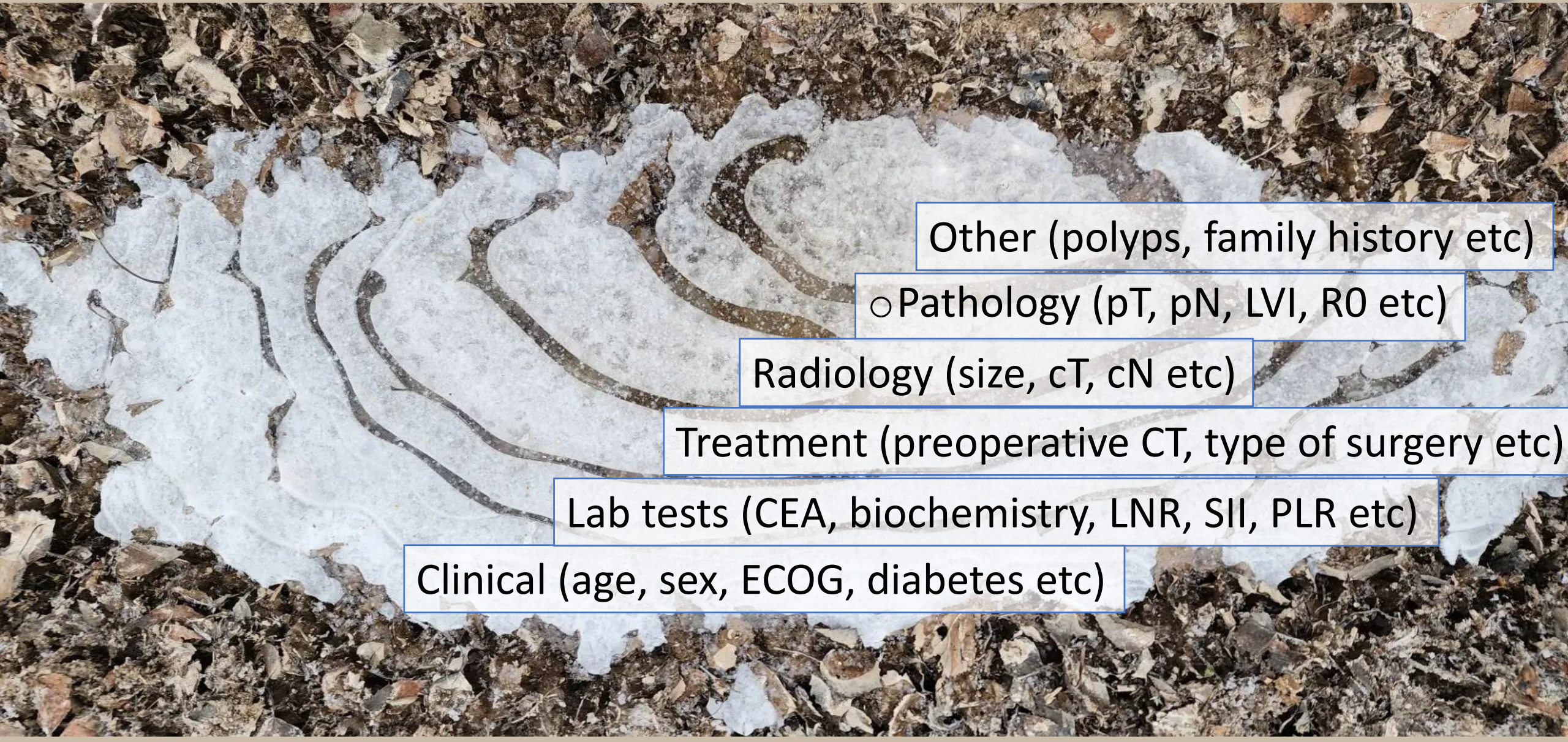
validation (n = 1,028)

training (n = 4,108)

Cohort selection flowchart



Data features (N=42)



Other (polyps, family history etc)

○ Pathology (pT, pN, LVI, R0 etc)

Radiology (size, cT, cN etc)

Treatment (preoperative CT, type of surgery etc)

Lab tests (CEA, biochemistry, LNR, SII, PLR etc)

Clinical (age, sex, ECOG, diabetes etc)

ML algorithms, target outcomes and comparators



TNM (via Cox-regression)

VS

Gradient Boost for OS

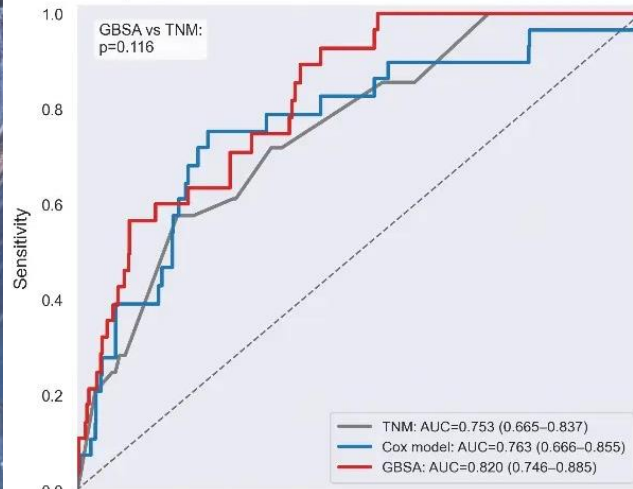
- *CatBoost* for recurrence risk & mortality
- At 1, 3, 5 years post-treatment

Feature importance in ML models

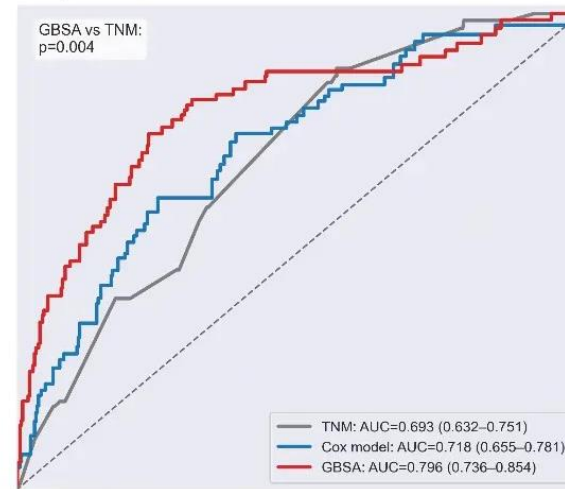
Feature	Importance for OS (%)	Importance for DFS (%)	Importance for Recurrence (%)
pN	9,56	16,13	3,21
Surgery type	8,22	6,77	6,72
LN positive	7,99	7,84	11,21
LN examined	7,97	6,87	21,60
pT	7,92	6,60	6,33
AST	6,51	1,53	2,02
Age	5,73	0,13	0,34
Albumin	3,80	2,52	0,14
Polyps (number)	3,70	2,43	0,91

ROC curves for OS & DFS

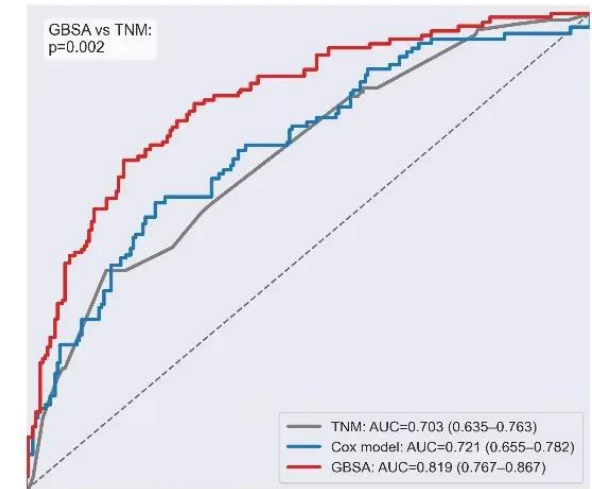
A. 1 year: OS



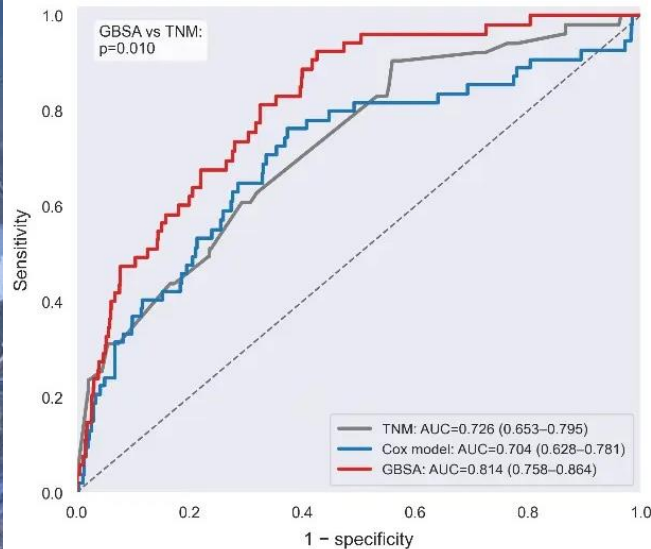
B. 3 years: OS



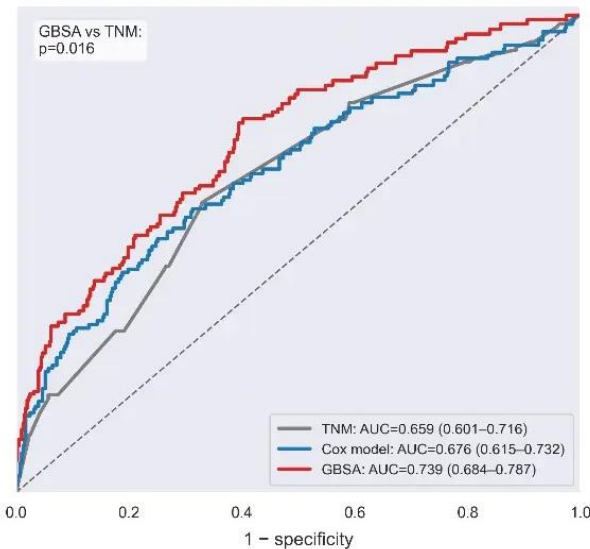
C. 5 years: OS



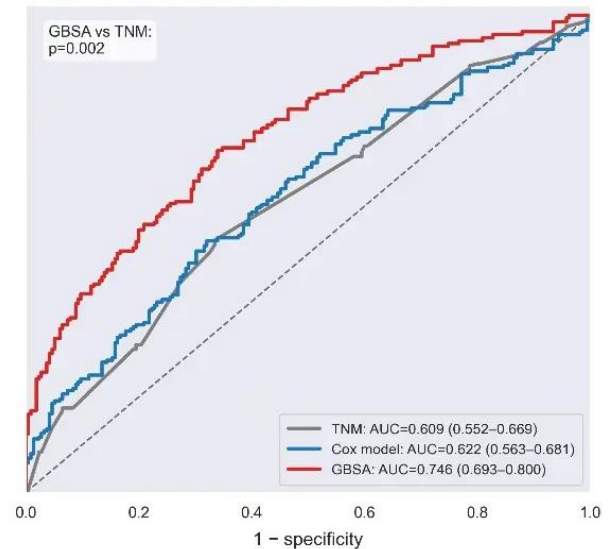
D. 1 year: DFS



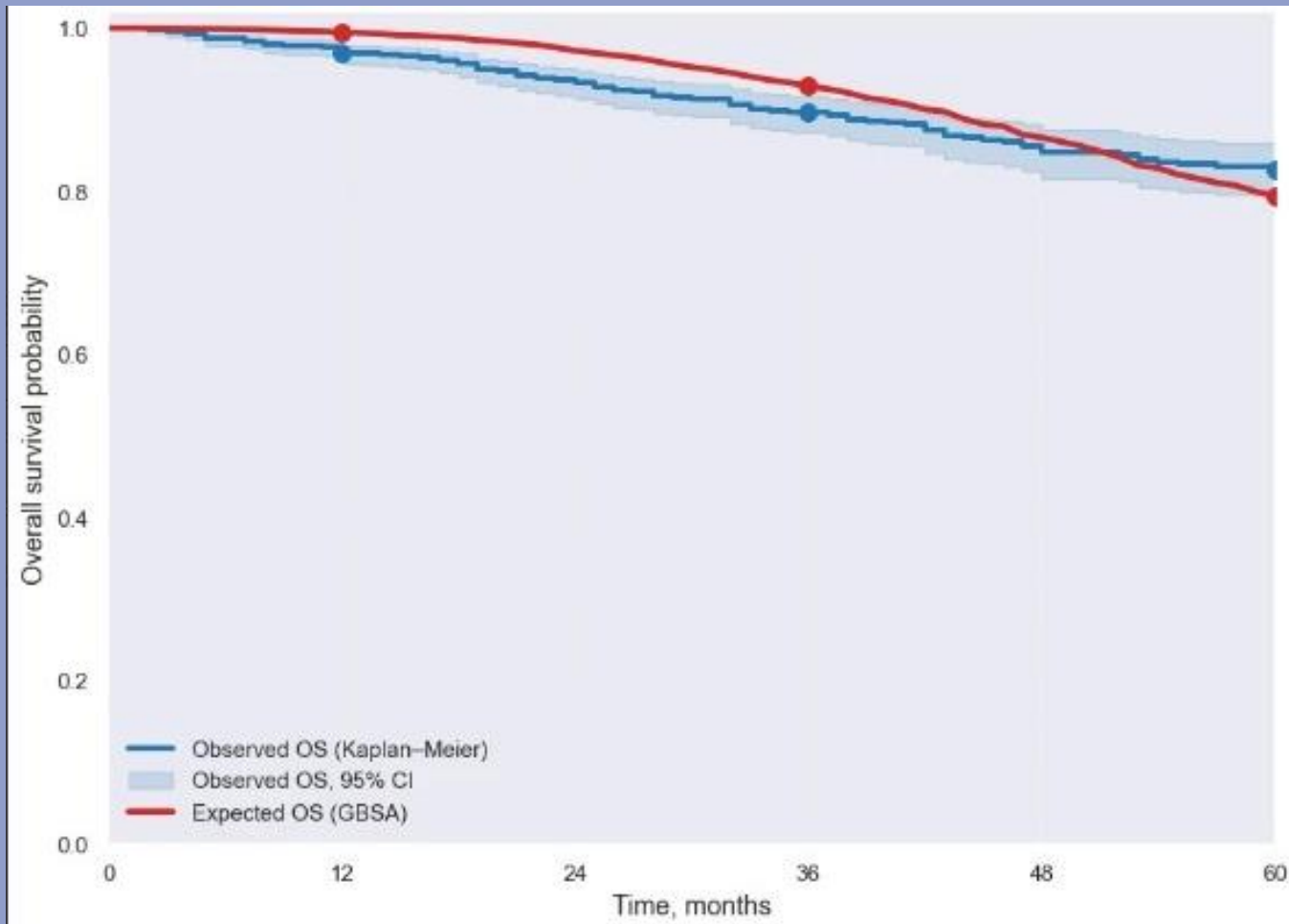
E. 3 years: DFS



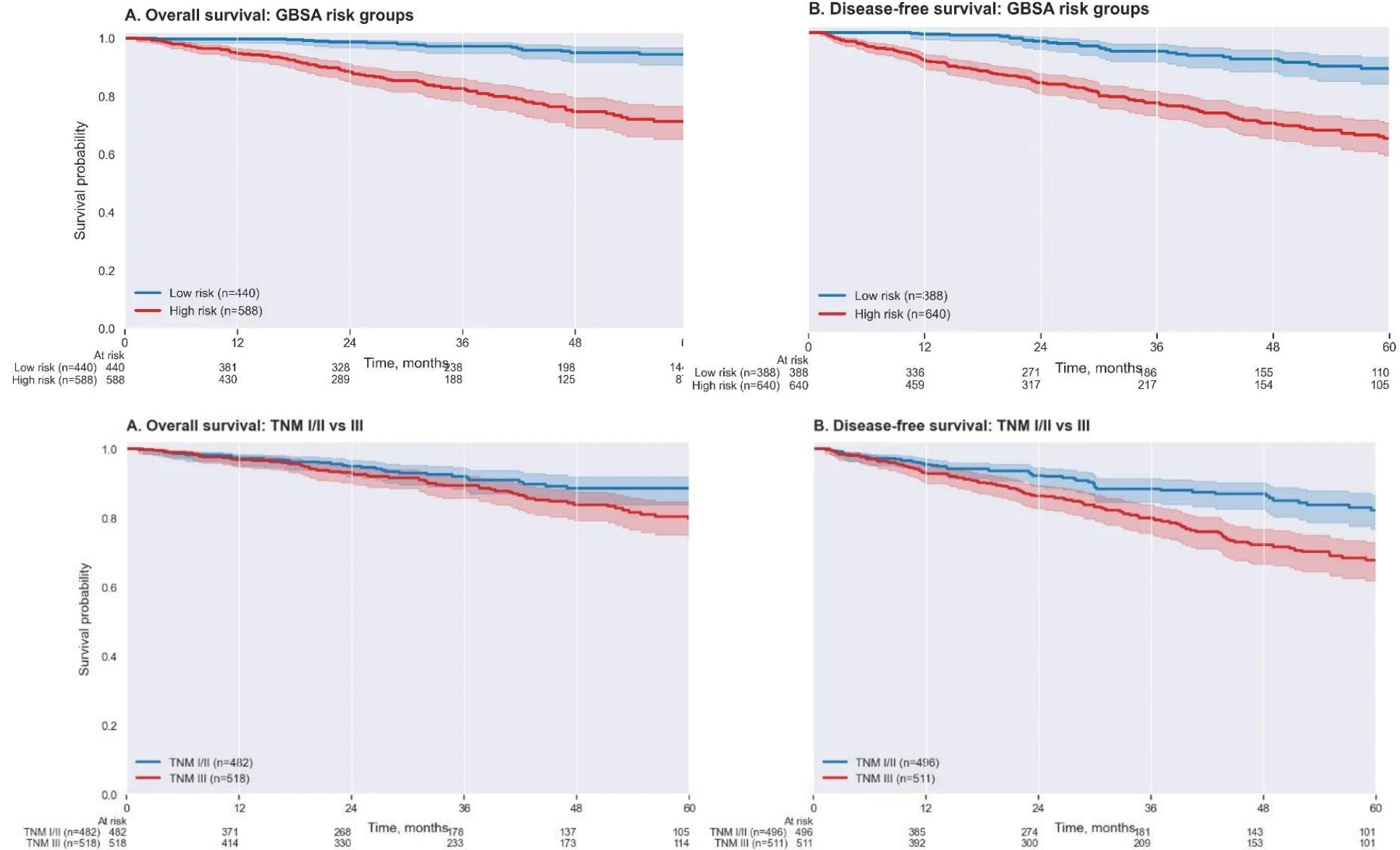
F. 5 years: DFS



Expected vs observed OS



Gradient boost vs TNM

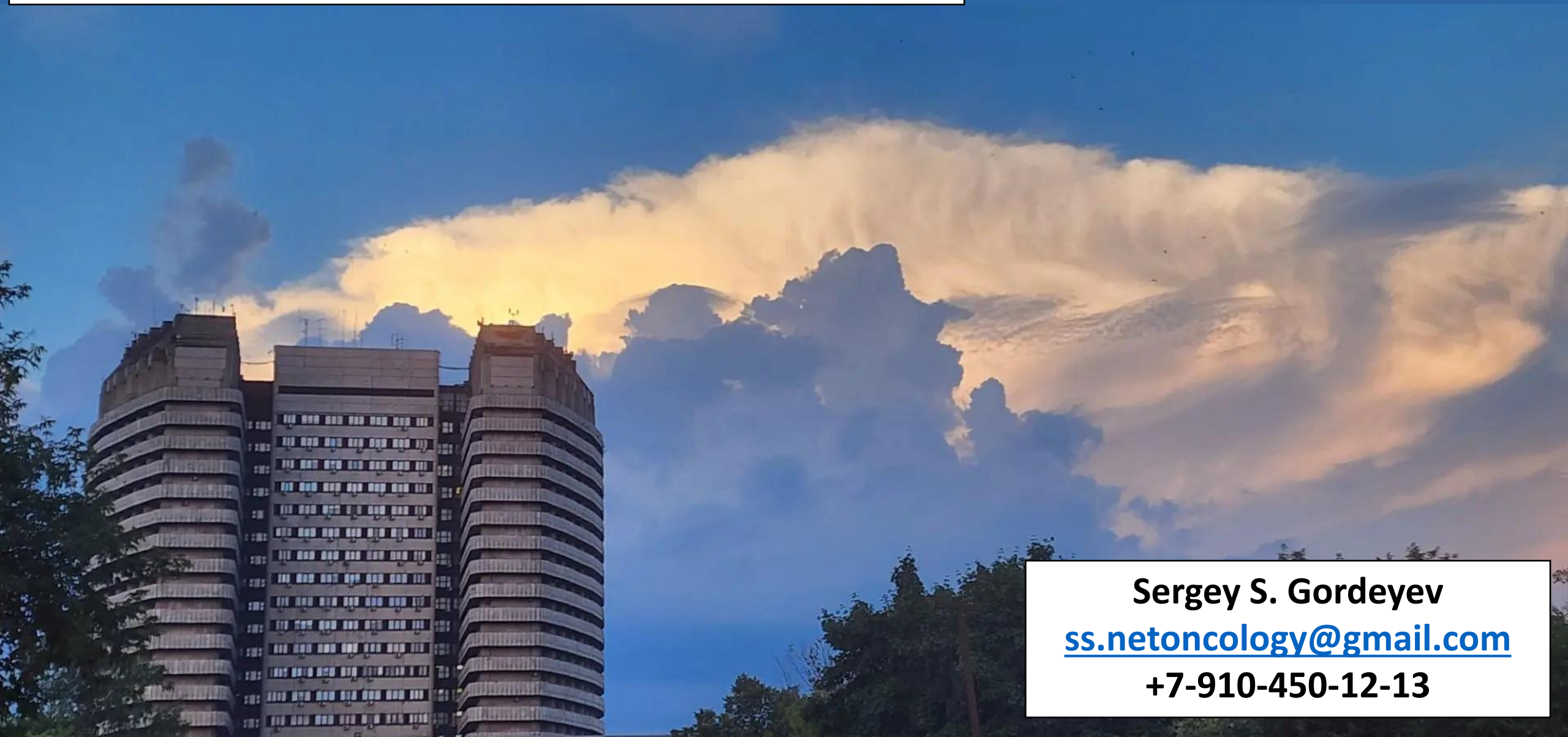


Take home message



- **EHR-based ML prognostic models can outperform TNM**
- **Clinical implementation is difficult and requires prospective trials (see cDNA)**
- **More data needed**
- **ML-models need constant re-education and validation in changing conditions
(new drugs, better pathology & diagnostics etc)**

Thank you for your attention!



Sergey S. Gordeyev
ss.netoncology@gmail.com
+7-910-450-12-13