Early Ovarian Cancer Detection

Alberto Esteban-Medina[1 2] , Laura Alejos[1 2], Víctor de la Oliva[1 2], Román Villegas Portero [3], Dolores Muñoyerro Muñiz [3] , Carlos Loucera[1 2], Joaquín Dopazo[1 2].

[1]Andalusian Platform for Computational Medicine, Fundación Progreso y Salud. Sevilla, España.

[2]IBiS - Instituto de Biomedicina de Sevilla, Spain.

[3]Subdirección Técnica Asesora de Gestión de la Información. Servicio Andaluz de Salud. 41071, Sevilla, España.

Introducction

Ovarian cancer is the **fourth** most **common gynecological cancer** and the **most lethal one**. Typically, this disease is **diagnosed** at **late stages**, which drastically **reduces** the **survival rate**, particularly in older age groups.

This **project** consists in developing a **Machine Learning tool** for **early ovarian cancer detection** based on Electronic Health Record (**EHR**) data from **Andalusian** patients extracted from the Population Health Base (**BPS**). To ensure **human-understandable** explanations at the patient-level resolution, we used an Explainable Boosting Machine (**EBM**) at the core of the system.

Electronic medical records from BPS

4 Informative datasets used:

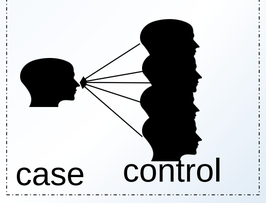
I. **BPS pathologies.**
II. **Diagnoses.**
III. **Specialists.**
IV. **Analytics**

Preprocessing

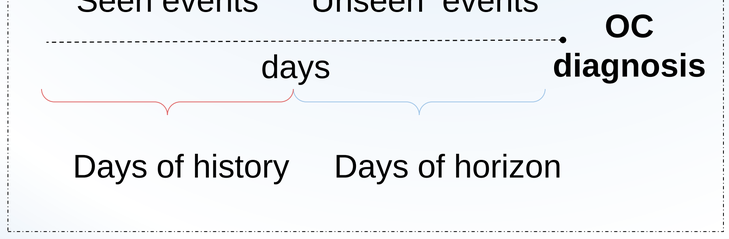
Train

Train
Patients with analytics 2017-2020

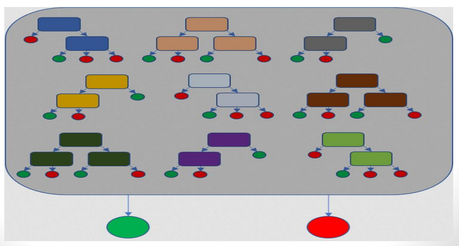
Matching



Time induced model

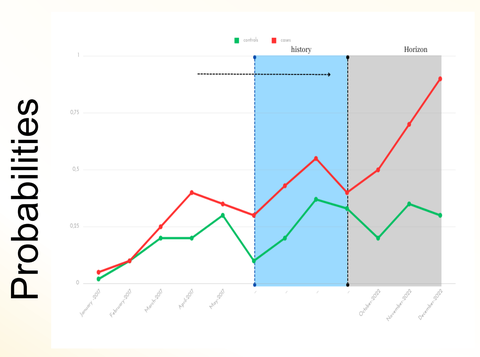


Fit EBM-time-induced pipeline

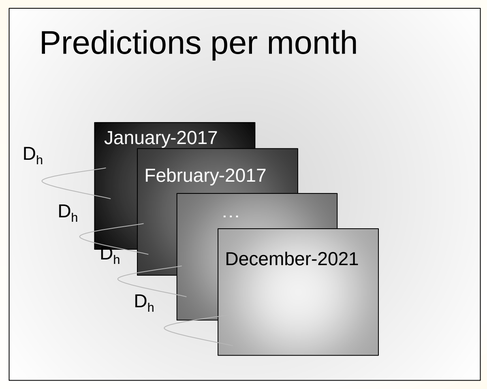


Validation

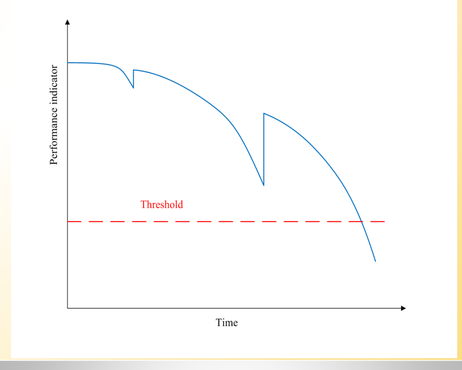
Validation
Patients with analytics 2021



Probabilities time series



Obtaining Threshold



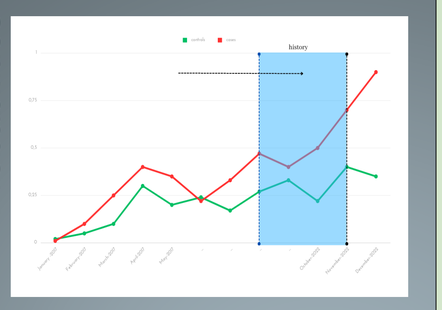
Maximum of the last 6 predictions

Test

Test
- Monitoring controls
- New data 2022

Time induced model

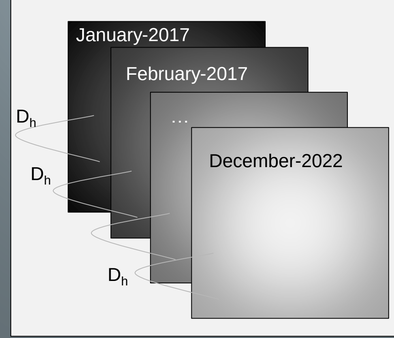
Probabilities time series



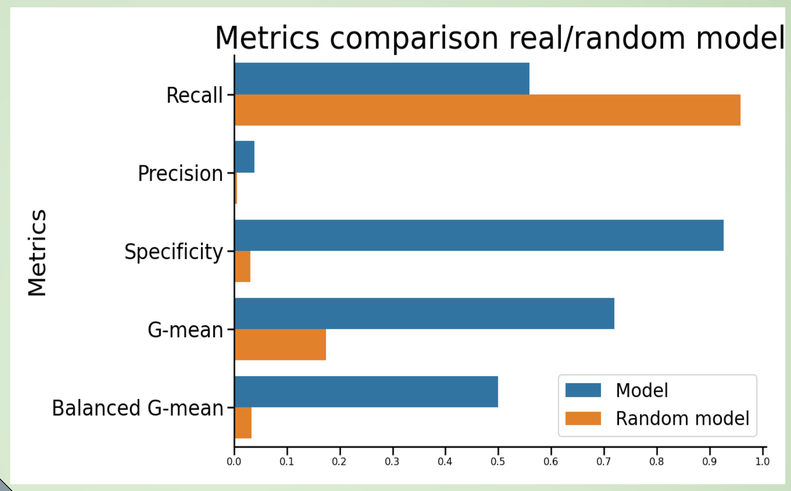
Months

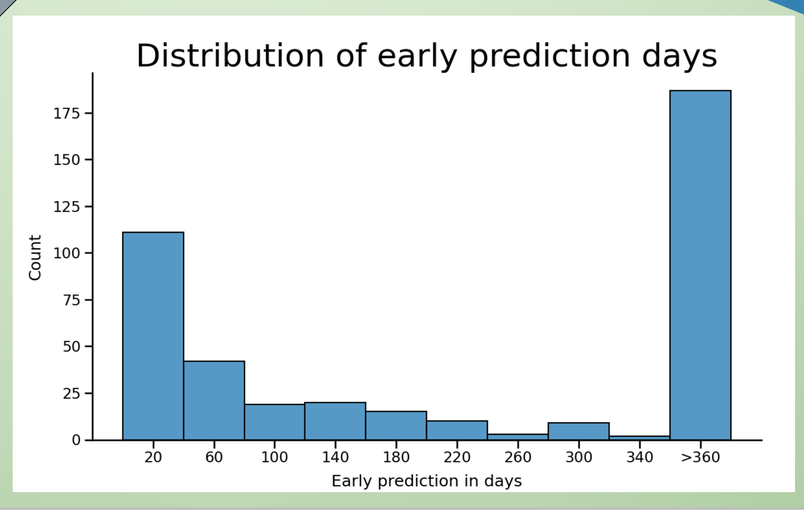
Applying Threshold

Predictions per month



Results





Conclusions

- The knowledge acquired/obtained by the algorithm from the data aligns with the current existing literature.
- The results obtained by the algorithm have a significant value considering the low cost of false positives and speed of prediction.
- There is a problem with the temporal resolution of the data, which makes it challenging to treat it as a time series problem.
- Out of 747 new cases in 2022, 418 were early diagnosed (208.5 median days)

BPS
IBIS INSTITUTODEBIOMEDICINADESEVILLA
JABI JORNADAS ANDALUZAS DE BIOINFORMÁTICA SEVILLA, 3-5 DE JULIO 2023
@clinicalbioinfo
Junta de Andalucía
Consejería de Salud y Consumo
Fundación Progreso y Salud