APPLYING BIG DATA ANALYTICS TO DEVELOP PREDICTIVE MODELS FOR ESTIMATING HER2+ BC INCIDENCE AND PREVALENCE IN BRAZIL

Rol R[1], França M[2], Bello E[2], [1]F. Hoffmann-La Roche Ltd, São Paulo, SP, Brazil, [2]F. Hoffmann-La Roche Ltd, São Paulo, Brazil

OBJECTIVES: Breast cancer (BC) is a high incident neoplasia worldwide with a high socio-economic impact and a significant pressure in the public healthcare system costs. The aim of this study is to apply Big Data Analytics for the development of an Oncology Structured Information System able to predict incidence, prevalence and costs of HER2+ BC in the Brazilian's public healthcare system at municipal level.

METHODS: A data lake was created from public databases, provided by Brazilian Public Health System (SUS), National Cancer Institute (INCA) and National Healthcare Agency (ANS), comprising healthcare and demographic information of all Brazilian municipalities. Data records were spliced into 8 different groups at anonymized patient level and based on treatment profile: Chemotherapy (C), Radiotherapy (R), Surgery (S), Chemo+Radio (C+R), Chemo+Surg (C+S), Radio+Surg (R+S), Chemo+Radio+Surg (C+R+S) and No Treatment (N) in order to cover 100% of the patients among the 03 different systems. An algorithm was developed in order to harmonize and prepare the data for statistical analysis. A restrospective period of 10 years (2017-2008) was used to build the linear regression methodology, using additive trend and seasonality, with a confidence interval of 95%.

1) Prevalence/Incidence core database selection

Due to the granularity of information, data volume and oncology driven characteristics of the chemotherapy dataset, it was selected as the core database for the study.

2) Prediction Tests

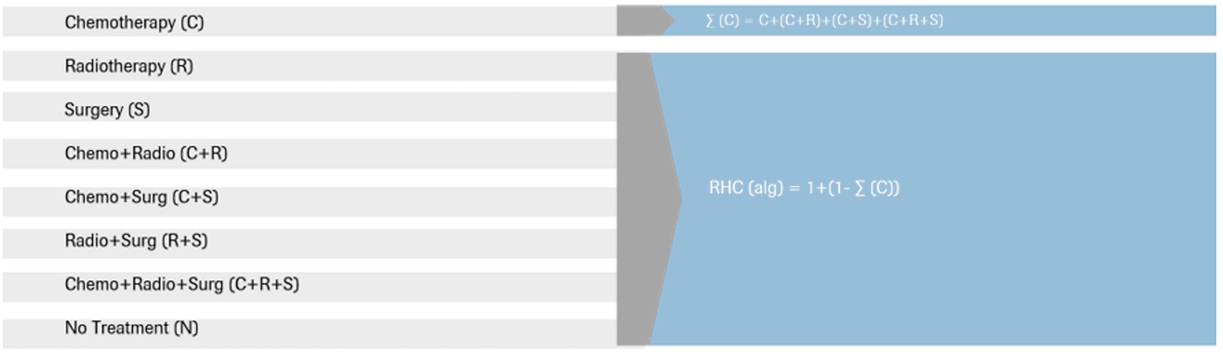
The main objective of this phase was to evaluate the accuracy of the prediction model applied over the DATASUS (Chemotherapy) database. We applied the predictive algorithm (based on SES (Simple Exponential Smoothing)) over a comprised dataset composed by data from 2008 to 2015. The prediction results (PR) was compared with the real results (RR) of 2016 and 2017. The accuracy was expressed by the ratio between PR and RR. The accuracy was measured in National, State and Capital spheres.

3) Harmonization Process

In despite of the Prev/Inc Tests have demonstrated a high accuracy, the harmonization of the database was necessary due to the aim of obtain precision at municipalities level. In this way, a range of algorithms were developed in order to harmonize the database in accord with the identified gap.

4) RHC Algorithm

Since the dataset build until here was developed based on the patients in the chemotherapy database, the model needs to extrapolate the numbers for those patients that are included in other types of treatment:



5) Private Algorithm

Since the dataset build until this point was based on public healthcare, the private perspective would be necessary in order to complete the final dataset. 4027 municipalities are mapped by ANS (Agencia Nacional de Saude) with beneficiaries' rate information. For each one of them was calculated a private index (PI). Since the oncology treatment depends on the availability of specialized services, an onco structure index (OSI) was also calculate in order to deflate the (PI) when necessary. Finally, since around 26% of oncological patients who have a private insurance search treatment in the public system, a private/public index (PP) was also created.

$$FDS = Pop(aRHC) * [(1+PI)*(1/OSI)*(1/PP)]$$

RESULTS: The model aplication resulted in an automated Oncology Structured Information System (OSIS) which comprises more than 3 billion records, including demographics, epidemics, cancer incidence, prevalence, healthcare procedures and infrastructure for all the 5,570 Brazilian municipalities. The OSIS is automatic updated in a monthly basis, presenting an accuracy of 97,8% for a 3 years prediction.

CONCLUSIONS: The use of big data analytics allow the development of models and tools for estimating BC incidence and prevalence of patients with HER2+ BC, at municipal level in Brazil. Such tool can support health managers in better planning and policies development in a more detailed and granular perspective.