Introduction
We have developed a machine learning approach to analyze heterogeneous clinical data in order to assist health professionals and researchers to visualize direct and indirect interrelations that can lead to a better understanding of complex diseases. With the help of physicians from Curie Hospital specialized in breast cancer, we have applied this network inference method to data on breast cancer which is the second leading cause of cancer death among women, and the most diagnosed cancer in women.

Materials and Methods
All data used in this work was provided by the Surveillance, Epidemiology, and End Results (SEER) Program supported by the National Cancer Institute. It was downloaded through SEER*Stat software and parsed and preprocessed with the vidente R package. The dataset contains raw data, engineered data and also data linked together from different datasets, with different notations. Therefore, a preprocessing step is crucial in order to analyze such data. Our preprocessing step led to the decrease in number of features from 544 to 52, and the final data (50,000 cases) was analyzed with MIIC (multivariate information-based inductive causation) algorithm (1). Differently from other methods, MIIC is able to infer both causal and non causal relationships through non-directed, directed and bidirected edges. It’s a novel approach in an attempt to take the best out of both search and score and constraint based methods.

Table 1: Information on initial and final number of samples and features.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of features</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEER18 Custom Data (Breast)</td>
<td>544</td>
<td>555,201</td>
</tr>
<tr>
<td>Carcinomas of Breast</td>
<td>544</td>
<td>440,715</td>
</tr>
<tr>
<td>Final dataset</td>
<td>52</td>
<td>50,000</td>
</tr>
</tbody>
</table>

Results

Figure 1: Network inferred using MIIC on a dataset with 52 features and 50,000 cases. Squared nodes represent discrete features, rounded nodes represent continuous features, red arrows indicate correlation, blue arrows indicate anticorrelation and gray arrows indicate that not possible to compare the signal of correlation. Nodes are colored according to the category they are involved with, according to the legend.

Figure 2: Subnetworks that were analyzed in more detail in this preliminary analysis. These subnetworks are of nodes directly associated with Marital Status and Age at Diagnosis (A) and Vital Status (B).

Discussion
Our results lead to the same direction of the literature, pointing to marital status as a prognostic factor (2). Physicians should take into consideration that the impact of surgical treatment on quality of life may vary according to patients’ sociodemographic and clinical characteristics. These results were also demonstrated for male breast cancer (3). Apart from that, progesterone receptor in breast cancer is understudied and controversial. Our results point to it as a putative protective/prognostic factor in breast cancer.

References