2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)

Breast Cancer Digital Patient Model to Capture and Visualize Real World Data

Nekane Larburu[1,2], Mónica Arrúe[1,2], Iván Macía[1,2], Jon Kerexeta[1,2], Naiara Muro[1,2,3]
[1]eHealth and Biomedical Applications, Vicomtech, Donostia-San Sebastian, Spain
[2]Biodonostia, Donostia-San Sebastian, Spain
[3]Sorbonne Universités, UPMC Univ Paris 06, INSERM, Université Paris 13, Sorbonne, Paris Cité, UMR S 1142, LIMICS, Paris, France
nlarburu@vicomtech.org

Abstract— **Digital revolution in health enables clinicians to access huge amount of data that can be exploited for decision making. However, the lack of integration of the various data sources, the existence of data sources not directly exploitable (e.g. free text, image, signals, genomic sequences) and the lack of digital data models (i.e. digital representation of the data) make such exploitation difficult. The development of effective Decision Support Systems (DSS) in concrete clinical contexts involves the development of appropriate and integrated representations of them, together with new paradigms for the exploitation, modeling and visualization of data oriented to decision-making. The European project DESIREE aims to contribute to the development of a system with these characteristics that has application to decision making by the Breast Committee. In particular, the visual analytics tool can contribute to the exploitation of clinical data in Breast Cancer.**

Keywords-breast cancer, data model, visual analytics

I. INTRODUCTION

Due the complexity of breast cancer, a multi-disciplinary team is needed to make treatment decisions. A breast cancer case may last for years and accumulated information is extensive and heterogeneous, differing greatly from one patient to another depending on the evolution of the disease.

Clinical Decision Support Systems (CDSS) based on guidelines may support clinicians during the decision-making process, but these guidelines are sometimes too inflexible and do not provide relevant information about the case. It may also happen that there is simply not enough evidence for the current case, limiting their applicability. In order to address these problems, the DESIREE[1] project has been developed; a web-based software that aims to improve the coordination and management of information in the Breast Units (BU). The DESIREE software includes a novel CDSS based on an advanced digital patient model, which integrates all significant sources of information to support decisions on the case and relevant information from the patient's context [1]. In addition, the system is able to model the experience, improving and supporting the process of knowledge discovery [2].

[1] http://www.desiree-project.eu/

Besides a CDSS, it contains a visual analytics module that represents statistical information of the patients that can support clinicians and clinical managers. This system includes an advanced visual analytics tool to study clinical hypothesis based on the acquired real world data in breast cancer.

This paper is structured in the following way. In Section II, the state of the art in relation to Digital Breast Cancer Patient (DBCP) model and visual analytics tools in healthcare are presented. Section III discusses the DBCP model used to build the solution. Section IV presents the visual analytics tool for a specific use case and analyses the results obtained. Lastly, Section V closes the paper with benefits and limitations of the developed tool, and future work.

II. STATE OF THE ART

Throughout the years multiple techniques have been used to assist clinicians in the prediction of the prognosis of a disease and to create a subsequent correct visualization of the results.

Therefore, this section presents A) digital patient models in breast cancer and B) existing visual analytics tools relevant to this research.

A. Digital Breast Cancer Patient Model

Data formalization and structuring has been identified as a requirement for illness modelling and clinical support. Semantic Web Technologies, such as Ontologies, have shown to be a very powerful tool for not only modelling the raw data but also all the relationships among all concepts. The use of ontologies for an illness representation allows making complex queries and inferences among data and model complex clinical pathways as when modelling a concrete illness or clinical condition. Several applications have been developed over the years. The project called AlzPharm [3] presented a Semantic Web approach for integrating different neurodegeneration data sources using RDF. The work of Younesi et al. [4], modelled the environmental exposure focusing on the tobacco smoke exposure risk using an ontology. Zhang et al. [5] described an ontology-based approach to integrate datasets for cancer research. The scope of all these formalizations is to build up a digital formalization of illness or patient models for the

2372-9198/19/$31.00 ©2019 IEEE
DOI 10.1109/CBMS.2019.00132

best predictive and personalized healthcare. In this context the Digital Breast Cancer Patient (DBCP) model is a shared conceptualization of the domain of breast cancer diagnosis and treatment. It provides a set of inter-related concepts that can be used to describe the knowledge about the domain, which will be the basis of clinical decision making.

B. *Visual Analytics Tools*

Considering the limited time of clinicians to explore data, it is very important to provide them intuitive tools. Visual analytics is defined as the science of displaying data by interactive interfaces in order to detect patterns in the data and to obtain relevant information [6]. One of the most common criteria for classifying visualizations is to consider the dimensionality of the visualization, that is, the number of attributes that the visualization shows. On the one hand, the univariate visualization (i.e. only one dimension) is the simplest visualization and its main objective is to obtain information about the distribution and the central tendency of an attribute. On the other hand, the main purpose of the multivariate visualization (i.e. two or more dimensions) is to provide insight into the relationship between attributes [7], [8]. This type of analysis, besides allowing to examine the distribution of the attributes, it also allows to analyze the relationship, patterns and correlations between these variables. One of the most popular graphs for displaying multivariate data is the parallel coordinates plot [9], [10].

In the parallel coordinates plot, each of the attributes is depicted by a vertical axis. These axes are placed in parallel in such way that they are at the same distance from each other. The samples of the dataset are depicted by horizontal lines in a way that they cross the axes of the attributes taking the corresponding value in each case. The lines are typically colored following a criterion defined by the research question to which the answer is to be given. The parallel coordinates plot is ideal for analyzing many variables at one time and finding patterns in the data [11].

As an example of parallel coordinates visualization is Fastbreak; a tool developed to analyze and visualize large amounts of genomic data [12]. This tool allows to analyze structural variations of next-generation sequencing data. One of the visualizations that this tool provides is a parallel coordinates plot (Figure 1). In this graph, the vertical axes represent the expression level that a gene may have in each particular cancer [13]. On the other hand, the horizontal lines depict different genes and they are colored according to their genetic expression: most expressed genes are colored red and least expressed ones are colored blue, which facilitates the exploration of gene expression similarities between different types of cancers.

With the aim of having a more global vision of the data, visualizations can be placed on a dashboard, which is defined as a comprehensive visual representation, usually composed by more than one visualization, that displays the most relevant information to reach specific goals [14]. There are several visual design guidelines, but one most widespread guidelines is the Shneiderman's Visual Information Seeking Mantra: "overview first, zoom and filter, then details on demand" [15]. These tasks can be summarized as follows:

- Overview: get an overview of all data
- Zoom: zoom in on interesting items
- Filter: filter out uninteresting elements
- Details-on-demand: select an item or group and get details when needed

This is the mantra on which the DESIREE team has been inspired to design the visual analytics tool.

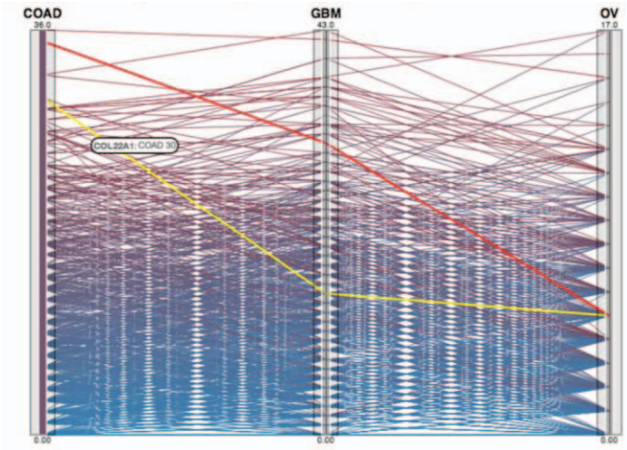

Figure 1. Fastbreak parallel coordinates plot showing differences in genetic expression between colon adenocarcinoma (COAD), glioblastoma multiforme (GBM) and ovarian cancer (OV) with two genes colored red and yellow.

III. DIGITAL BREAST CANCER PATIENT MODEL

In DESIREE we defined the Digital Breast Cancer Patient (DBCP) model as the necessary information about the patient for the decision making in BUs. The DBCP represents the clinical, diagnostic and therapeutic information in a structured way over time, and it is fundamentally oriented to the decision making in diverse relevant moments. To take into account all the information related to the course of the disease, the information is structured in different scenarios that are defined below.

A. *Scenarios*

- *Scenario A: After diagnosis - Without treatment*

This scenario collects information regarding the first therapeutic decision in the patient's sequence. It is assumed that the diagnosis has been made, and all the information necessary for the therapeutic decision is available considering that the patient has not been given any prior breast cancer related treatment. Hence, the decision may be surgery or neo-adjuvant therapy.

- *Scenario B: After neoadjuvant therapy - before surgery*

In this case, the patient has already received neo-adjuvant therapy but has not yet received any surgery.

The therapeutic decision is based on criteria that describe the tumor and possible contraindications to surgery related to the patient's condition can be considered. Hence, usually the decisions are usually surgery.

- *Scenario C: After surgery - After neo-adjuvant therapy*

In this case, the patient has already been given neo-adjuvant therapy and surgery. The decision refers to adjuvant treatment including chemotherapy, radiotherapy and / or hormone therapy.

- *Scenario D: After surgery - Without neo-adjuvant therapy*

In this case, the patient has already had the surgery but not the neo-adjuvant therapy. Decisions refer to adjuvant treatment, including chemotherapy, radiotherapy and / or hormone therapy.

- *Scenario E: After surgery - Incomplete adjuvant therapy*

In this case, adjuvant therapy has not been well tolerated or any unforeseen event has occurred, so a new decision must be made, such as: change / stop the initially decided treatment.

B. Attributes to be studied.

572 real patients of Primary Breast Cancer from four different European hospitals were analyzed in this study. A multidisciplinary team of clinicians selected 194 attributes to use in the dataset, which are reviewed periodically. The majority of these attributes are already in the Electronic Health Records (EHR) but it had been expanded to include additional information of interest to clinicians. The types of attributes found in this dataset are:

- Numerical variables: integer and float (e.g. age)
- Categorical variables (e.g. gender)
- Logical variables (e.g. if the patient smokes)

Patients were treated with various therapies grouped in (i) surgical procedures, (ii) radiotherapy protocols, (iii) endocrine therapies, (iv) chemotherapy protocols and (v) clinical trials. This depends on the scenario and the patients' status.

C. Outcomes to be studied

Additionally, the system collects different type of outcomes, which is crucial for knowledge acquisition, assessment and new hypothesis discoveries:

a) Guideline Compliance: as DESIREE develops a CDSS based on guidelines, when BUs do not meet the recommendations of these guidelines, this information is stored in the system and is used to determine if there is a tendency not to follow the guideline for patients who have certain characteritics in common.

b) Adverse Events (AE) or Toxicities: is defined as "any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medical treatment or procedure that may or may not be considered related to the medical treatment or procedure" [16]. The US National Cancer Institute (NCI) develop one of the most relevant systems to determine the toxicities which is known as the NCI Common Terminology Criteria for Adverse Events or CTCAE2, which is used in this system.

c) Treatment Response: neoadjuvant therapy only can be evaluated "objectively", and then one of the aims is to explore whether certain patients may have a better or worse response to neoadjuvant treatments. The values that the treatment response can take are (from worst to best): disease progression, stable disease, partial response and complete response.

d) Clinical Outcomes: two different clinical outcomes are measured: the relapse and exitus (both caused or not by the breast cancer) and the survival (using the diagnosed date and currend date, the survival rate can be calculated).

The multidisciplinary team of clinicians has pointed out these outcomes as most significant ones. Nevertheless, other factors like Patient Reported Outcomes (PRO), should to be considered too [17] since they include a perception of the effectiveness of the treatment from the patient's point of view. But in this study this information is not available.

D. Requirements for the DBCP model

Achieving a consensus on what data is needed has been shown as an arduous task, with the advantage of having different clinical teams in the consortium, in order to compare the variability of clinical practice and the availability of information sources.

This definition of DBCP enables the storage of the information of the different cases, including aspects related to the decision making and follow-up of the clinical guide, as well as the results after the decision. The advantage is evident in the review and exploitation of cases, search of similar cases, monitoring of results, elaboration of working hypotheses and discovery, in a much more systematic way. To reduce complexity in the pilot phases, we started from what we consider a minimum data set, which goes beyond the parameters mentioned in the clinical guidelines. However, DBCP is a living structure, and therefore it is defined in an ontology, where novel sources of data, such as prognostic image biomarkers, genomic data or other aspects related to the patient's life and condition could be easily included, and therefore, it is the format used to integrate the concepts in the whole DESIREE system [1].

IV. VISUAL ANALYTICS

A. Visual Analytics for Statistical Results

The statistical dashboard is composed of 5 different charts. The data that these charts display can be filtered by two criteria: the unit (i.e. hospital) and the scenario, offering also the option to visualize the totality of the data. The Figure 2 shows the statistical dashboard.

Figure 2. Visual analytics statistical dashboard

This dashboard focuses on the representation of several features:

- The number of treatments given to patients in different scenarios, in *x, y* axis to represent the difference based on different patients' attributes (see Figure 2 up, left) and the percentage of treatments given (see Figure 2 up, right).
- When a type of treatment is selected in the bubble chart, the percentage of specific type of procedures given within that treatment is presented (see Figure 3 right). The remaining graphs of the dashboard display the information of that specific treatment as well when a treatment is selected in the bubble chart.

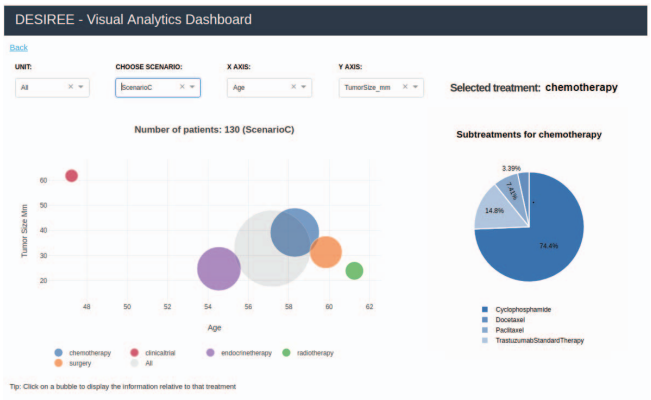

Figure 3. Bubble chart and chemotherapy procedures pie chart

- Outcomes results are also analyzed, focusing on results as relapses and toxicities, where we provide not only the toxicity rates but also the kind of toxicity observed (e.g. toxicities can be described following the CTCAE terminology, where five different grades describe the severity of the reported toxicities, see Figure 4).

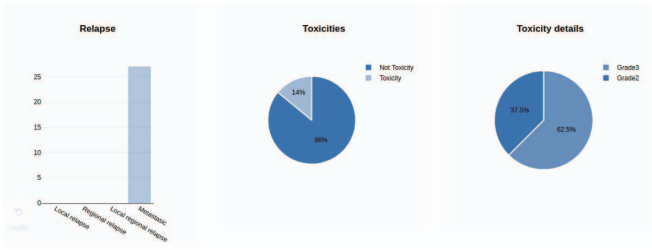

Figure 4. Toxicity charts

B. *Visual Analytics for Pattern recognition*

As presented in [18], besides visual representation of statistical results, DESIREE also implements a system that represents patterns associated with different outcomes intuitively (e.g. treatment response and treatment related adverse events).

For that, first data mining techniques have been implemented. This includes a preprocessing stage to detect the category of each attribute, exclusion of non-representative attributes, exclusion of null values and grouping of same clinical meaning attributes.

After preprocessing the data, it is necessary to obtain the relationship between all attributes and the outcome to be studied (i.e. color of the lines in the parallel coordinates plot). All these relationships between attributes and outcomes are calculated by means of a *correlation matrix*. Additionally, by detecting the most significant attributes for a certain outcome, the relationships between these relevant attributes are also obtained for later visualization.

Once the correlation matrix is calculated, to show the most relevant attributes, the user has selected the scenario and treatment(s) to be analyzed. Consequently, the attributes most correlated ("n") with the outcome selected by the user are given. Lastly, the "n" attributes (i.e. the axes) most correlated with the selected outcome are sorted according to their level of correlation between them. That is, these "n" attributes are placed in the parallel coordinates plot in such way that the most correlated attributes are side by side. In this way, the pattern obtained is illustrated in a more intuitive way to the user.

C. *Use Case for Pattern recognition*

This section presents the obtained results for scenario D - *After surgery, without neo-adjuvant therapy* – and as the criterion to color the parallel coordinates lines the death of the patient (Figure 5). As illustrated in Figure 5, this criterion is of a categorical type and its possible values are: true ("red") or false ("blue").

The plot in Figure 5 shows the attributes with the highest correlation to the selected criterion (i.e. death) obtained in the data mining process. Each of the axes in the parallel coordinates plot depicts one of the most correlated attributes and each of the lines shown horizontally depicts a patient.

These horizontal lines are colored in red if the patient has died and in blue if the patient has not died.

It is noteworthy that only lines (i.e. patients) that have data in the corresponding most correlated attributes (i.e. axes) and the selected criterion (i.e. lines color) are shown. In the case that a patient is missing one of the most correlated attribute data, this patient's line will not be plotted. In this case study, for the scenario D and the selected outcome (i.e. death), only 180 patients (i.e. samples lines) are shown.

Notice that this tool also gives the clinicians the option of filtering the data by unit, scenario and treatment. It also allows the clinician to select the number of axes (i.e. attributes) to be displayed and adding or removing axes to the plot. Thus, even if the most correlated attributes to the selected outcome are shown by default, clinicians can also explore the data freely.

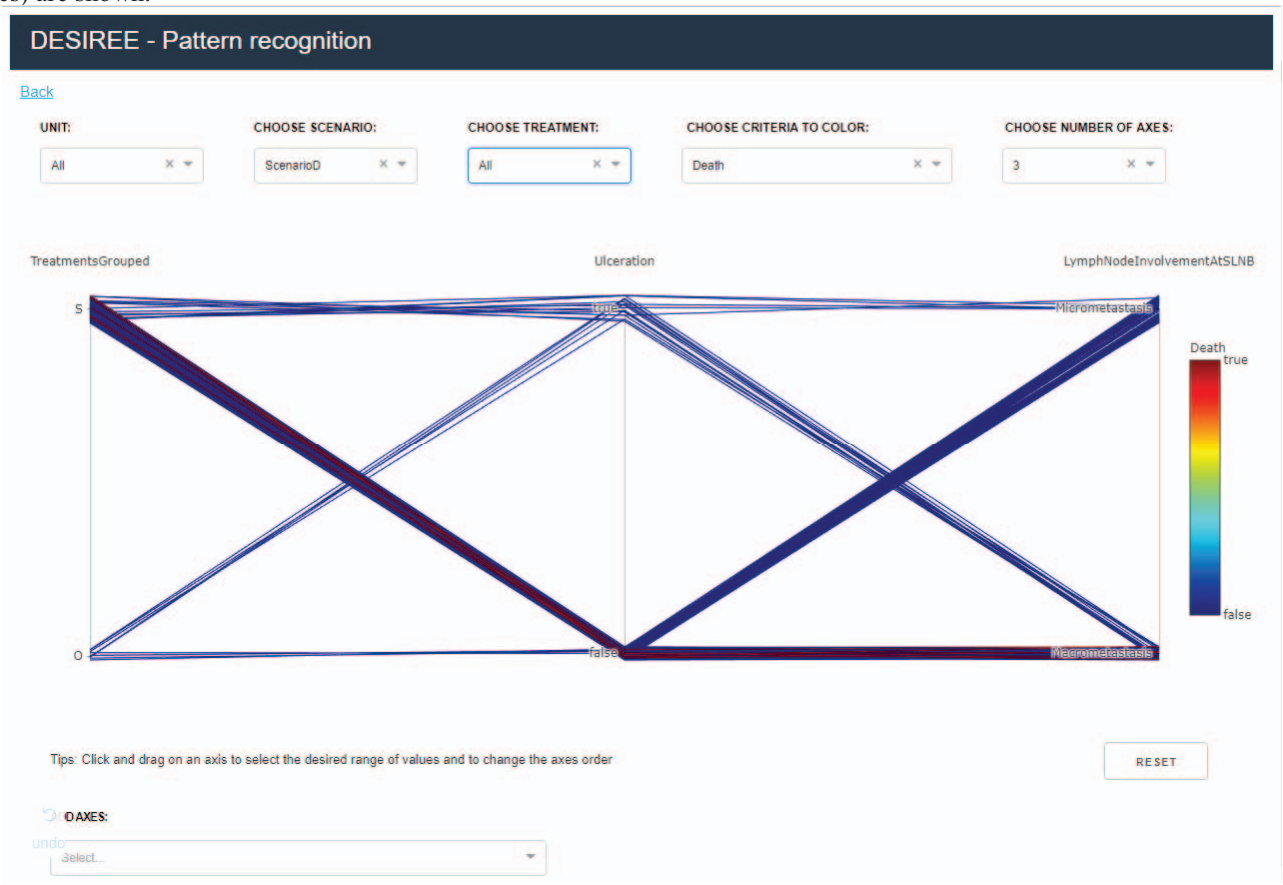

Figure 5. Parallel coordinates plot obtained for all the units, the scenario D and the death criterion

V. Discussion & Conclusion

The main goal of this study is to develop a digital breast cancer patient model to exploit the data and based on that to create intuitive and dynamic tools for clinicians that allows them to quickly explore data and patters within data related to outcomes, in such way that they can confirm their hypotheses.

This is a novel approach in the field of breast cancer. The data mining techniques used nowadays to develop models offer a very high accuracy. However, these models are too rigid and not accessible to clinicians. In addition, to the best of our knowledge, there is no tool that allow clinicians to explore the stored patient data quickly, dynamically and intuitively. For this reason, clinicians need a lot of time to carry out this type of studies.

The tool presented in this paper allows physicians to explore all important patient information in a simple manner and visually analyze the presence of patterns between data related to a particular criterion. It also allows them to contrast their hypotheses with the data. Nevertheless, the greatest limitation of this tool is when one of the attributes to be displayed (i.e. axes) contains a high number of null values. Therefore, in order to be able to visualize enough patients and gain insight on the relationship between attributes and patterns within data, it is essential to have enough samples "complete".

Another limitation is that this study requires the input of all values in DESIREE system by hand, as it is not integrated with the four clinical partners involved in this project. This is a widespread restriction, as most EHR are not interoperable and the data available there is not structured in a way that a computer can interpret it, so this

data it is not directly exploitable. However, the implementation of the digital breast cancer model in a semantically interoperable ontology, facilitates the integration of DESIREE system with EHRs.

ACKNOWLEDGMENT

The DESIREE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690238.

Additionally, the authors acknowledge support from the four clinical partners involved in the project: Onkologikoa, Eresa Grupo Médico, Hôpital Européen Georges-Pompidou and Hôpital Saint-Louis.

REFERENCES

[1] B. Séroussi *et al.*, "Reconciliation of multiple guidelines for decision support: a case study on the multidisciplinary management of breast cancer within the DESIREE project," *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2017, pp. 1527–1536, 2017.

[2] N. Muro *et al.*, "Augmenting Guideline Knowledge with Non-compliant Clinical Decisions: Experience-Based Decision Support," in *Innovation in Medicine and Healthcare 2017*, vol. 71, Y.-W. Chen, S. Tanaka, R. J. Howlett, and L. C. Jain, Eds. Cham: Springer International Publishing, 2018, pp. 217–226.

[3] H. Y. Lam *et al.*, "AlzPharm: integration of neurodegeneration data using RDF," *BMC Bioinformatics*, vol. 8, no. 3, p. S4, May 2007.

[4] E. Younesi *et al.*, "CSEO – the Cigarette Smoke Exposure Ontology," *J. Biomed. Semant.*, vol. 5, p. 31, Jul. 2014.

[5] H. Zhang *et al.*, "An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. Suppl 2, p. 41, Jul. 2018.

[6] R. May, P. Hanrahan, D. A. Keim, B. Shneiderman, and S. Card, "The state of visual analytics: Views on what visual analytics is and where it is going," in *2010 IEEE Symposium on Visual Analytics Science and Technology*, 2010, pp. 257–259.

[7] K. Nazemi, *Adaptive Semantics Visualization*. Springer, 2016.

[8] J. Brownlee, "Better Understand Your Data in R Using Visualization (10 recipes you can use today)," *Machine Learning Mastery*, 29-Jan-2016. .

[9] D. (DJ) Sarkar, "The Art of Effective Visualization of Multi-dimensional Data," *Towards Data Science*, 15-Jan-2018. [Online]. Available: https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57. [Accessed: 08-Jun-2018].

[10] A. Cuzzocrea and D. Zall, "Parallel Coordinates Technique in Visual Data Mining: Advantages, Disadvantages and Combinations," in *2013 17th International Conference on Information Visualisation*, 2013, pp. 278–284.

[11] "Parallel Coordinates Plot - Learn about this chart and tools." [Online]. Available: https://datavizcatalogue.com/methods/parallel_coordinates.html. [Accessed: 07-Jun-2018].

[12] "Fastbreak." [Online]. Available: http://fastbreak.systemsbiology.net/. [Accessed: 14-Jun-2018].

[13] R. Bressler *et al.*, "Fastbreak: a tool for analysis and visualization of structural variations in genomic data," *EURASIP J. Bioinforma. Syst. Biol.*, vol. 2012, p. 15, Oct. 2012.

[14] S. Few, "Dashboard Confusion Revisited," p. 6, 2007.

[15] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," p. 8.

[16] "Side Effects of Adjuvant Treatment of Breast Cancer | NEJM." [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJM200106283442607. [Accessed: 13-Feb-2019].

[17] N. Muro, N. Larburu, J. Bouaud, and B. Seroussi, "Weighting Experience-Based Decision Support on the Basis of Clinical Outcomes' Assessment," *Stud. Health Technol. Inform.*, vol. 244, pp. 33–37, 2017.

[18] N. Larburu, M. Arrue, N. Muro, R. Álvarez, and J. Kerexeta, "Exploring Breast Cancer Patterns for Different Outcomes using Artificial Intelligence," in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2018, pp. 1–6.