BC_cardiotox: A cardiotoxicity dataset for breast cancer patients

Beatriz Piñeiro-Lamas¹ *, Ana López-Cheda¹, Ricardo Cao¹, Cayetana Barbeito-Caamaño², and Alberto Bouzas-Mosquera ²

¹Grupo MODES, Departamento de Matemáticas, CITIC, Universidade da Coruña, A Coruña, 15071, Spain
²Unidad de Imagen y Función Cardíaca, Servicio de Cardiología, Complexo Hospitalario Universitario de A Coruña, Instituto de Investigación Biomédica de A Coruña (INIBIC), A Coruña, 15006, Spain

*corresponding author: Beatriz Piñeiro-Lamas (b.pineiro.lamas@udc.es)

ABSTRACT

This database is a result of the collaboration between the University of A Coruña and the University Hospital of A Coruña. It contains information about 531 women diagnosed with HER2+ breast cancer, treated with potentially cardiotoxic oncologic therapies. These treatments can cause cardiovascular adverse events, including cardiac systolic dysfunction, the development of which has important clinical and prognostic implications. The availability of good predictors may enable early detection of these cardiac problems. Variables such as age, weight and height are available for each patient, as well as some measures obtained from echocardiography, a technique used prior and during the treatment to check the structure and function of the heart. Among them, there is a functional variable that measures the myocardial velocity during the cardiac cycle. For patients that experienced cancer therapy-related cardiac dysfunction during the treatment period, time until its appearance is known. This database aims to enable the scientific community in conducting new research on this cardiovascular side effect.

Background & Summary

Breast cancer is the most common cancer in women. According to the World Health Organization, 2.3 million of women were diagnosed with breast cancer around the world in 2020, and 685000 died due to this cause¹. Around 20% of all breast cancers have higher levels of the protein human epidermal growth factor receptor 2 (HER2), that helps breast cancer cells grow quickly. They are known as HER2-positive (HER2+). These cancers tend to grow and spread faster than breast cancers that are HER2-negative (HER2-), but are much more likely to respond to treatment with drugs that target the HER2 protein. Therapies that specifically target HER2 include trastuzumab. In fact, this monoclonal antibody is now considered the standard of care for the treatment of this disease in both the metastatic and adjuvant setting. It works by attaching to HER2, so it stops the cancer cells from growing and dividing. Previous studies have confirmed that this treatment has a strong antitumor effect, improving the overall survival and the progression-free survival². However, trastuzumab may cause cardiotoxicity³. In addition, it can be combined with other antineoplastic drugs that may also have cardiotoxic effects. The development of cancer therapy-related cardiac dysfunction (CTRCD) has important prognostic implications, so its early detection is an ongoing challenge. Previous studies found that clinical factors such as age, hypertension, previous anthracycline treatment or low baseline left ventricular ejection fraction (LVEF), among others, are risk factors⁴, but they are not sufficient to predict which patients will develop CTRCD during treatment. Therefore, in clinical practice it is recommended to monitor the appearance of CTRCD through blood laboratory tests and cardiac imaging tests. The echocardiogram is the fundamental test, since it allows to assess various parameters of the structure and function of the heart, such as LVEF. This parameter gives the percentage of blood leaving the heart each time it contracts. CTRCD is detected when the LVEF falls below 50 and by at least 10 points from the baseline LVEF⁵. In addition, the Tissue Doppler Imaging (TDI), which shows the velocity of contraction and relaxation of the cardiac muscle, can be obtained by using these techniques. This information, measured before treatment, may help in the early identification of CTRCD. Since the TDI shows a velocity as a function of time, it can be preprocessed to obtain a functional datum. However, the analysis of image and functional covariates as risk factors has not been broadly examined in the existing literature. This is one of the reasons that motivated building this database. Furthermore, in the study of cardiotoxicity risk factors, time until its appearance is often ignored. The availability of the aforementioned time in this database opens the opportunity to use survival analysis techniques in this context.

Figure 1 shows a flowchart of the treatment process. Baseline clinical and image variables are obtained during the checkup.
Methods

Participants

The authors declare that this study was conducted in accordance with the principles of the World Medical Association Declaration of Helsinki “Ethical Principles for Medical Research Involving Human Subjects”. The ethical approval for this research was obtained from the Drug Research Ethical Committee of Galicia (CEIm-G), code 2022/492.

Our data consists of information about 531 women diagnosed with HER2+ breast cancer between 2007 and 2021 and treated with potentially cardiotoxic oncologic drugs at the University Hospital of A Coruña. Among those patients, 54 (10.17%) experienced CTRCD during the follow-up period. The age group with the highest prevalence (17.29%) is 60-69. Table 1 shows the list of available variables, along with the acronym description (if needed), their type (numerical, binary or image) and their measure unit. All variables were measured prior to treatment. A detailed description of the participants’ numerical characteristics is presented in Table 2. The variable time contains, for patients who experienced CTRCD, the time (in days) until its appearance. For patients who did not experience this side effect, time contains the length of the follow-up period.

Regarding TDI, since it shows a velocity, it should be considered as a function, in spite of its image nature. A preprocessing was proposed and implemented in the statistical software R to obtain a functional covariate from the image. Although 531 images are available, only 270 were preprocessed. The reason is that the TDIs were obtained with different echocardiographic devices. For images provided by the most used device, a semi-automatic algorithm was designed to preprocess them and obtain the corresponding functions. For images provided by the rest of machines, automation was not achieved. Therefore, they were not preprocessed because they required excessive manual work. Figure 2 shows an overview of the number of available data based on the nature of the variable (clinical or functional) and the CTRCD status of the patient. A detailed description of the preprocessing procedure is presented in the following subsection.

Preprocessing of image data

The baseline TDI, which informs about the condition of the heart just before the treatment begins, may be a good predictor of CTRCD. Figure 3 shows an example of a TDI. Although it is stored as an image, the main interest relies on the yellow signal, which shows the velocity of contraction and relaxation of the cardiac muscle. Note that the cardiac cycle is represented by 3 waveforms, as illustrated in Figure 4: S (systolic), above zero velocity; and E’ and A’ (early and late diastolic), with negative velocity. In order to extract the underlying function, a preprocessing of the image was proposed by taking into account the colour of each pixel in HSV format (Figure 5). Note that HSV stands for Hue-Saturation-Value:

- Hue ranges from 0 to 360 degrees, and it is the attribute of a colour by virtue of which it is discernible as red, green, yellow, etc. It depends on the dominant wavelength, but not on the intensity or lightness.
- Saturation describes the amount of gray in a particular colour, ranging from 0 (gray) to 1 (primary colour).
- Value describes the brightness or intensity, from 0 to 1, where 0 is completely black, and 1 is the brightest.

The preprocessing procedure consists of the following steps:

**Step 1. Detection of the yellow pixels**

The signal of interest has different yellow tones. Unlike the well-known colour plane representation red, green and blue (RGB), HSV is closer to how humans perceive colours. In HSV, the pixels in yellow can be identified just by controlling the range of movement of the first parameter (Hue). However, with RGB it is not that easy, since very different combinations of the three components may result in similar final colours. Therefore, it is challenging to choose a proper range for the three primary colours so that their mixture is yellow.

**Step 2. Transformation into a function**

Once the yellow pixels have been detected, the result is still not a function, because for each value in the horizontal axis there is more than one yellow pixel in vertical. To overcome this drawback, the following procedure has been considered. For each pixel in the horizontal axis, let us define by \( a \) the lowest yellow pixel in vertical and by \( b \) the highest one (Figure 6). Then, there will be three situations:

- If \( 0 < a < b \), \( b \) is considered.
- If \( a < b < 0 \), \( a \) is considered.
If $a < b$, consider $a \frac{e^{\lambda(b-a)}}{e^{\lambda(b-a)+\lambda b(b-a)+\lambda a}} + b \frac{e^{\lambda(b-a)}}{e^{\lambda(b-a)+\lambda b(b-a)+\lambda a}}$, for some $\lambda > 0$.

Preliminary studies (supervised by cardiology experts) suggested that a good choice for $\lambda$ is 10. The result is a collection of pixel position pairs. To obtain their coordinates (time on the horizontal axis, velocity on the vertical one), we have taken into account the known coordinates of some pixels in the image. This way, a function for each patient can be obtained from the TDI image.

**Step 3. Selection of one cycle per patient**

The resulting function is a periodic combination of S, E’ and A’ waves, as illustrated in Figure 4. The end of the A’ wave determines the end of a cycle and the beginning of the subsequent one. To standardize the data, just one cycle for each patient is selected considering the following criterion:

- If a patient’s function only contains one full cycle, this one is selected.
- If it has two cycles, the one that describes better the S, E’ and A’ waveforms (supervised by experts) is chosen.
- If it has three or more, the selected one is the one verifying that its distance to the mean cycle is the lowest.

All the cycles have been standardized to have the same length (one cycle). Figure 7 shows the standardized cycles for the 270 patients. The difference between the cycles of both groups of patients (with and without CTRCD) is more remarkable in the central zone. In particular, for the 27 patients with CTRCD detected during the follow-up period, E’ and A’ waves seem smoother, as well as slightly shifted to the right.

**Data Records**

Data presented in this work consist of three files:

- **BC_cardiotox_clinical_variables.csv**: dataset with 531 rows (patients) and 11 columns (clinical variables in Table 1). The column headers are heart_rate, age, weight, height, CTRCD, time, LVEF, PWT, LAd, LVDD and LVSD.
- **BC_cardiotox_functional_variable.csv**: dataset with 270 rows (patients) and 1002 columns. The first column is CTRCD, and the remaining 1001 (named $t_1$, $t_2$, ..., $t_{1000}$, $t_{1001}$) contain the cycle extracted from the TDI discretized in 1001 equispaced points in the interval $[0,1]$.
- **BC_cardiotox_clinical_and_functional_variables.csv**: dataset with 531 rows (patients) and 1012 columns. The first 11 columns are the same as in **BC_cardiotox_clinical_variables**, and the remaining are $t_1$, $t_2$, ..., $t_{1000}$, $t_{1001}$. For the patients whose image has not been preprocessed, the last 1001 columns contain NAs.

These files are freely available online at Figshare\(^8\). A README file with a detailed description of each dataset is also available.

**Technical Validation**

Preliminary exploratory analyses have made it possible to detect (and correct) some outliers and fill in some apparently missing data. Regarding the TDI images, the corresponding functions obtained with the preprocessing procedure described before have been examined one by one. The algorithm worked automatically for 137 of the 270 images. However, for the remaining 133, the selection of the cycle had to be manually done. In general, the selection of the cycle can not be automated when the function crosses the horizontal axis several times in the gap between A’ and S waves. This gap, known as isovolumetric contraction time (IVCT), is the first phase of systole. When the function intersects the axis during the IVCT more than once, the beginning and end of the cycles have to be manually selected from the set of intersection points. Moreover, there are situations where the cycle does not begin at zero velocity, and therefore the intersection points with the axis are not useful. In these cases, the beginning and end of the cycles have to be manually chosen too. For the automatically preprocessed images, 77% of the outcomes were successful. The remaining images were preprocessed again manually, in order to make best use of the available data. Figure 8 shows an example of a TDI image along with its corresponding function (in red) and the beginning and end points of the cycles (with white crosses). Figure 9 shows the two resulting cycles (the first one in solid line, the second one in dotted line) in the same graph. Note that they have been standardized to have the same length. For this image, the selected cycle was the first one, although both of them are quite similar.

Despite the fact that low quality images were not considered in this study, the proposed algorithm still performs well when a portion of the image has a weak signal but contains one or more cycles with an acceptable quality.
Usage Notes

The preprocessing procedure was implemented in R (version 4.2.2). To reproduce the analyses, the installation of some R packages is required: png⁹, grDevices⁹, raster¹⁰, rasterVis¹¹, lattice¹² and pracma¹³.

The code can be easily adapted to other type of images. In our case, png images are read in R and the HSV colour of each pixel is considered. Then, the pixels of interest can be selected by taking into account their colours. For TDI images, since the goal are yellow pixels, only the H component is controlled. As the result contains some almost black isolated pixels, a denoising process was performed to get rid of them, by taking into account their own V component and the one of their neighbours. The threshold considered for V can be easily changed to better fit other images, if needed. Regarding the trasformation of pixel positions into velocities, it is necessary to know, at least, the velocity that corresponds to two pixels in the image. This way, one can find the conversion factor between pixels and velocities. All this information can be easily changed in the code.

Code availability

The commented R code for the preprocessing of image data is available at Figshare⁸. In addition, for reproducibility issues, two TDI images are provided. One of them requires a manual selection of the beginning and end points of the cycle.

References


Acknowledgements

B.P.L was sponsored by Axudas Predoutorais da Xunta de Galicia, with reference ED481A-2020/290. A.L.C. was sponsored by the BEATRIZ GALINDO JUNIOR Spanish Grant from MICINN (Ministerio de Ciencia e Innovación) with code BGP18/00154. B.P.L., A.L.C. and R.C. acknowledge partial support by the MICINN Grant PID2020-113578RB-I00 and partial support of Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14). B.P.L., A.L.C. and R.C. acknowledge the support received from the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union European Regional Development Fund (ERDF)-Galicia 2014-2020 Program, Grant ED431G 2019/01.
Author contributions statement

R.C., A.L.C. and A.B.M. identified the problem and its potential importance. A.B.M. selected the appropriate data, checked that all patients met the inclusion criteria and was responsible of the anonymization. Data validation, cleaning, and outlier detection were carried out by C.B.C. The image preprocessing procedure was proposed by R.C., A.L.C. and B.P.L., always taking into account A.B.M. and C.B.C.’s expert opinion. B.P.L. wrote the code for the preprocessing of the image data. Finally, B.P.L. drafted the manuscript with the help of all authors. Critical comments and revision of the manuscript have been made by the five authors.

Competing interests

The authors declare no competing interests.

Figures & Tables

Figure 1. Schematic overview of the treatment process.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description (if needed)</th>
<th>Type</th>
<th>Unit of measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>heart rate</td>
<td>-</td>
<td>numerical</td>
<td>beats per minute (bpm)</td>
</tr>
<tr>
<td>age</td>
<td>-</td>
<td>numerical</td>
<td>years</td>
</tr>
<tr>
<td>weight</td>
<td>-</td>
<td>numerical</td>
<td>kg</td>
</tr>
<tr>
<td>height</td>
<td>-</td>
<td>numerical</td>
<td>cm</td>
</tr>
<tr>
<td>time</td>
<td>time from beginnig of treatment to CTRCD or end of follow-up</td>
<td>numerical</td>
<td>days</td>
</tr>
<tr>
<td>CTRCD</td>
<td>CTRCD experienced during the follow-up period</td>
<td>binary</td>
<td>0 (no) / 1 (yes)</td>
</tr>
<tr>
<td>LVEF</td>
<td>Left Ventricular Ejection Fraction</td>
<td>numerical</td>
<td>%</td>
</tr>
<tr>
<td>PWT</td>
<td>Posterior Wall Thickness</td>
<td>numerical</td>
<td>cm</td>
</tr>
<tr>
<td>LAd</td>
<td>Left Atrial diameter</td>
<td>numerical</td>
<td>cm</td>
</tr>
<tr>
<td>LVDd</td>
<td>Left Ventricular Diastolic diameter</td>
<td>numerical</td>
<td>cm</td>
</tr>
<tr>
<td>LVSd</td>
<td>Left Ventricular Systolic diameter</td>
<td>numerical</td>
<td>cm</td>
</tr>
<tr>
<td>TDI</td>
<td>Tissue Doppler Imaging</td>
<td>image</td>
<td>cm/s (once preprocessed)</td>
</tr>
</tbody>
</table>

Table 1. Acronym description and basic information of the variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>heart rate</td>
<td>46</td>
<td>65</td>
<td>73</td>
<td>74.61</td>
<td>82</td>
<td>138</td>
</tr>
<tr>
<td>age</td>
<td>26</td>
<td>45.5</td>
<td>55</td>
<td>55.05</td>
<td>65</td>
<td>86</td>
</tr>
<tr>
<td>weight</td>
<td>40</td>
<td>58</td>
<td>66</td>
<td>68.34</td>
<td>75</td>
<td>120</td>
</tr>
<tr>
<td>height</td>
<td>138</td>
<td>155</td>
<td>159.5</td>
<td>159.3</td>
<td>164</td>
<td>186</td>
</tr>
<tr>
<td>time</td>
<td>5</td>
<td>278</td>
<td>518</td>
<td>855</td>
<td>1146</td>
<td>4276</td>
</tr>
<tr>
<td>LVEF</td>
<td>36.81</td>
<td>61.02</td>
<td>65.37</td>
<td>65.28</td>
<td>70.31</td>
<td>78.87</td>
</tr>
<tr>
<td>PWT</td>
<td>0.4631</td>
<td>0.7638</td>
<td>0.8763</td>
<td>0.8757</td>
<td>0.9761</td>
<td>1.7100</td>
</tr>
<tr>
<td>LAd</td>
<td>1.662</td>
<td>3.000</td>
<td>3.317</td>
<td>3.325</td>
<td>3.636</td>
<td>5.000</td>
</tr>
<tr>
<td>LVDd</td>
<td>2.831</td>
<td>4.020</td>
<td>4.345</td>
<td>4.352</td>
<td>4.692</td>
<td>5.970</td>
</tr>
<tr>
<td>LVSd</td>
<td>1.258</td>
<td>2.459</td>
<td>2.746</td>
<td>2.765</td>
<td>3.000</td>
<td>4.700</td>
</tr>
</tbody>
</table>

Table 2. Baseline clinical characteristics of participants.
Figure 2. Schematic overview of the number of observations depending on the type of variable (clinical or functional) and the CTRCD status.

Figure 3. Example of a TDI.
Figure 4. Illustration of S, E’ and A’ waves in a TDI.

Figure 5. HSV colour model (cylinder type). Retrieved April 3, 2023, from [14].
Figure 6. Illustration of $a$ and $b$ points in a TDI.
Figure 7. In magenta, the cycles of the patients who experienced CTRCD during the follow-up period. In black, the cycles of the patients who did not experience this side effect.

Figure 8. TDI image with its corresponding function (in red) and the beginning and end points of the cycles (with white crosses).
Figure 9. Cycles extracted from the previous image (the first one in solid line, the second one in dotted line), after being standardized to have the same length.
Dataset

https://doi.org/10.6084/m9.figshare.22650748.v1