Published in final edited form as:

J Pharm Sci. 2016 June ; 105(6): 2005–2008. doi:10.1016/j.xphs.2016.03.013.

Development of a CART Model to Predict the Synthesis of Cardiotoxic Daunorubicinol in Heart Tissue Samples from Donors with and without Down Syndrome

Carrie C. Hoefer*, Rachael Hageman Blair*, and Javier G. Blanco

Department of Pharmaceutical Sciences, School of Pharmacy and Pharmaceutical Sciences, The State University of New York at Buffalo, Buffalo, NY 14260, USA (C.C.H, J.G.B). Department of Biostatistics, School of Public Health and Health Professions, The State University of New York at Buffalo, Buffalo, New York, 14260, USA (R.H.B)

Abstract

Daunorubicin (DAUN) and doxorubicin (DOX) are used to treat a variety of cancers. The use of DAUN and DOX is hampered by the development of cardiotoxicity. Clinical evidence suggests that patients with leukemia and Down syndrome (DS) are at increased risk for anthracycline-related cardiotoxicity. Carbonyl reductases (CBRs) and aldo-keto reductases (AKRs) catalyze the reduction of DAUN and DOX into cardiotoxic C-13 alcohol metabolites. Anthracyclines also exert cardiotoxicity by triggering mitochondrial dysfunction.

In recent studies, a collection of heart samples from donors with and without DS was used to investigate determinants for anthracycline-related cardiotoxicity including cardiac daunorubicin reductase activity (DA), CBRs/AKRs protein expression, mitochondrial DNA content (mtDNA), and *AKR7A2* DNA methylation status. In this study, the available demographic, biochemical, genetic, and epigenetic data were integrated through classification and regression trees analysis (CART) with the aim of pinpointing the most relevant variables for the synthesis of cardiotoxic daunorubicinol (i.e., DA). Seventeen variables were considered as potential predictors. Leave One Out Cross Validation was performed for model selection and to estimate the generalization error. The CART model and variable importance measures suggest that cardiac mtDNA content, mtDNA⁴⁹⁷⁷ deletion frequency, and *AKR7A2* protein content are the most important variables in determining DA.

Keywords

Classification & Regression Trees; anthracyclines; heart

Corresponding author: Javier G. Blanco, Clin. Biochem., PhD, Department of Pharmaceutical Sciences, The State University of New York at Buffalo; 471 Kapoor Hall, Buffalo, NY 14214, Telephone: 716-645-4820, Fax: 716-645-3693, ; Email: jgblanco@buffalo.edu, URL: <http://pharmacy.buffalo.edu>

*Denotes equal contribution

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

The use of anthracyclines (e.g., doxorubicin and daunorubicin) for the chemotherapy of a wide variety of solid and hematological cancers is limited by the development of cardiotoxicity in some patients. A meta-analysis of randomized controlled trials for breast cancer, ovarian cancer, lymphomas and osteosarcomas showed that anthracyclines increased the risk of clinical cardiotoxicity by 5.43 fold, subclinical cardiotoxicity by 6.25 fold, and the risk of cardiac death by 4.94 fold compared to non-anthracycline regimens ¹. Clinical reports have shown that the presence of DS (trisomy 21) increases the risk of anthracycline related cardiotoxicity in pediatric patients with leukemia ^{2; 3}.

The complex pathogenesis of anthracycline-related cardiotoxicity is mediated by a combination of oxidative stress and metabolic perturbations in myocardial tissue induced by anthracycline C-13 alcohol metabolites, whose formation is catalyzed by carbonyl reductases (CBRs) and aldo-keto reductases (AKRs) ^{4; 5}. Anthracyclines also exert cardiotoxicity by triggering mitochondrial dysfunction. For example, anthracyclines cause an irreversible decrease in mitochondrial Ca^{2+} loading and ATP content. There are two main types of cardiotoxicity typically associated with the clinical use of anthracyclines. Clinically important declines in cardiac function that occur within one year of treatment are considered early-onset cardiotoxicity, and declines occurring after one year of treatment are considered late-onset cardiotoxicity. Early and late-onset cardiotoxicity are characterized by progressive declines in left ventricular function, which in some cases results in heart failure ^{6; 7}.

In recent studies, we analyzed a collection of heart samples from donors with and without DS with the aim of investigating tissue-specific determinants for anthracycline-related cardiotoxicity. We profiled the expression of the main cardiac anthracycline reductases (CBR1, CBR3, AKR1A1, AKR1C3 and AKR7A2), quantitated daunorubicin reductase activity, and measured mitochondrial DNA (mtDNA) content ^{8–10}. AKR7A2 was found to be the most abundant anthracycline reductase accounting for approximately 36% of the total reductase content in hearts from donors with and without DS. Additional studies have shown that cardiac DNA methylation status at specific sites in the *AKR7A2* gene are covariates that account for a fraction of the observed variability in the synthesis of cardiotoxic daunorubicinol ¹¹.

Classification and Regression Trees (CART) are widely used in medical decision-making because of their interpretive nature and flexibility ¹². CART models are represented by a decision tree, and can jointly model a feature space that consists of a diverse set of data ¹³. The tree inference is based on a recursive partitioning algorithm, and exhibits natural features of model selection through the inclusion (and exclusion) of variables in the splitting process. A defining feature in the flexibility of CART models is their ability to handle missing data. Traditional regression approaches require elimination of observations with missing data in the predictor set of variables, or imputation. In the elimination scenario, a single predictor with missing data can effectively cut down the sample size, regardless of missing pieces in other predictors. In contrast, the recursive partitioning algorithm used to build CART models splits the feature space for a single predictor at a time, and is not influenced by the missing pieces of other predictors in the model. Moreover, predictions

using the CART model for new data that may be incomplete can often be carried out using surrogate variables and splits. Surrogate variables are widely recognized as a practical way of managing missing data and also offer valuable insights into the relative importance of each predictor.

To our knowledge, no CART-based models have been developed to predict the synthesis of cardiotoxic daunorubicinol in heart, a tissue relevant to the variable pharmacodynamics of anthracyclines. Thus, the objective of this study was to develop a decision tree model, using CART, for the prediction of cardiac synthesis of daunorubicinol in heart tissue. The CART model was fit using a broad spectrum of tissue-specific data, which included genetic, epigenetic, and phenotypic data for the CBRs and AKRs involved in the cardiac metabolism of anthracyclines. Data on cardiac mtDNA (i.e., mtDNA content and percent frequency of the “common” mtDNA⁴⁹⁹⁷ deletion) was also incorporated into the CART model because of the functional links between alterations in mtDNA and the expression of genes involved in the metabolism of drugs ¹⁴.

Methods

Heart samples

The Institutional Review Board of the State University of New York at Buffalo approved this research. Heart samples from donors with (n = 11) and without DS (n = 33) were procured from The National Disease Research Interchange (NDRI, funded by the National Center for Research Resources), The Cooperative Human Tissue Network (CHTN, funded by the National Cancer Institute), and The National Institute of Child Health and Human Development (NICHD) Brain and Tissue Bank. The postmortem recovery interval was 10 h. Samples (2 – 20 g, myocardium, left ventricle only) were frozen immediately after recovery and stored in liquid nitrogen until further processing. Down syndrome status (yes/no) was obtained from anonymous medical histories and confirmed by array comparative genomic hybridization (aCGH) as described ^{8–10}. Cardiac daunorubicin reductase activity, and the expressions of CBR1, CBR3, AKR1A1, AKR1C3 and AKR7A2 (mRNA and protein) have been reported recently ¹⁰. Cardiac DNA samples were genotyped for the functional *CBR1* (rs9024, G>A) and *CBR3* (rs1056892, G>A, V244M) polymorphisms as described previously ^{9; 15; 16}. Cardiac mtDNA content (i.e., the ratios between the mitochondrial gene *MT-ND1* and the nuclear gene *18S rRNA*), and the percent frequency of the “common” mtDNA⁴⁹⁷⁷ (Mdel) have been reported by us ⁸. Cardiac *AKR7A2* methylation analysis was performed with Sequenom’s MassARRAY EpiTyper technology (Sequenom, San Diego, CA) as previously described ¹¹. The demographics of heart donors and counts of the missing values are given in Supplemental Table 1. Descriptive statistics for the variables are summarized in Supplemental Table 2.

Statistical analysis

Cardiac daunorubicin reductase activity (DA) was discretized into three levels, based on quantiles: high daunorubicin reductase activity (≥ 3.02 nmol daunorubicinol/min.mg), intermediate daunorubicin reductase activity (1.6 nmol daunorubicinol /min.mg and < 3.01 nmol daunorubicinol /min.mg), and low daunorubicin reductase activity (<1.6 nmol

daunorubicinol /min.mg). In total 17 variables were considered as potential predictors of DA for decision tree modeling (Table 1). Each variable coded was either continuous or categorical and include genetic, epigenetic, and subject demographic data (Table 1). CBRs and AKRs protein levels were transformed into a log 10 scale.

A CART model is a decision tree, in which the model space is divided at each node in the tree¹² (Supplemental File). The objective is to identify sub-regions in the model space that represent the outcome well. The splits in the tree are determined in a recursive way through a minimization of a loss function, which is the gini-index for our calculations. That is, at each node in the tree, all of the variables in the model are scanned for partitioning across their range. The variable and the split point are selected as the pair that is most effective in reducing the loss function, and they are represented as an internal node on the tree. The process is then repeated to find subsequent internal nodes, but under constraint to operate within the already partitioned model space. The CART models were fit using the library rpart available in CRAN (<https://cran.r-project.org>) for the R programming language (V3.12, www.rproject.org). In the model fitting process, we utilized an additional constraint that the terminal nodes in the tree were required to have at least three observations in them. That is, if less than three observations exist in a branch, no further splits will be attempted in the tree, which further protects from over-fitting.

Leave One Out Cross Validation (LOOCV) was performed to estimate the generalization error that guides in model selection (i.e., where to cut the tree)¹⁷. This approach to model selection protects from over-fitting, or selection of an overly complex (deep tree) model that will not generalize well to a new population, and is appropriate for small sample sizes. In the recursive partitioning, it is important to note that variables that do not improve the objective function in the optimization are left out of the tree (Supplemental File). The accuracy (1-misclassification rate) was computed from LOOCV as well as the sensitivity and specificity. Each terminal node in the tree corresponds to a classification label of a low, medium, or high DA category. The majority class of the data in that node determined the label and the node purity was calculated as the proportion of majority class in the terminal node. Variable importance was calculated for each predictor in the model. Variable importance is a univariate measure of the contribution of each predictor as a splitting node in the decision tree, and as a surrogate variable. Briefly, surrogate variables are predictors at each node that may be used in place of the internal node as a splitter, for example, in the case of missing data. Surrogates are identified through partitioning in the non-recursive sense, see¹⁸ and the Supplemental File for more details. Importantly, since each variable is tallied at every split in the tree, as a splitting node or a potential surrogate, there will not necessarily be a 1:1 correspondence between the tree and the variable importance plot. In fact, the variable importance plot is often viewed as complementary to the tree, and insightful in terms of model interpretation. The model fitting, LOOCV method, and variable importance measure are detailed in the supplemental file.

The decision tree represents a predictive model. When a new data point is available, it can be inputted into the tree, and traced down through the branches accordingly. The terminal region determines the class through the majority vote for the training data that is available in that region.

Results & Discussion

CART was used to build a predictive model for cardiac DA in a collection of heart tissue samples from donors with and without DS. The resulting classification tree splits on seven different variables in the data set. The tree along with the confidence for each branch is shown in Figure 1A. The overall accuracy of the model (1-misclassification rate) was estimated from LOOCV to be 0.82. The accuracy for the high daunorubicin reductase activity ($3.02 \text{ nmol daunorubicinol/min.mg}$) was 0.88 (sensitivity = 0.90 and specificity = 0.85). The accuracy for the intermediate daunorubicin reductase activity ($1.6 \text{ nmol daunorubicinol /min.mg}$ and $3.01 \text{ nmol daunorubicinol /min.mg}$) was 0.85 (sensitivity = 0.77 and specificity = 0.92). Finally, the accuracy for low daunorubicin reductase activity ($<1.6 \text{ nmol daunorubicinol /min.mg}$) was 0.89 (sensitivity = 0.81 and specificity = 0.96).

The model suggests that heart samples with relatively low frequency of the mtDNA⁴⁹⁷⁷ deletion ($\text{Mdel} < 3.1 \times 10^{-4} \%$) would have low DA ($<1.6 \text{ nmol daunorubicinol /min.mg}$). Approximately 79% of the heart samples had Mdel frequencies $>3.1 \times 10^{-4} \%$ (Hefti et al., 2014). In this subset of samples, mtDNA content (*MT-ND1/18S rRNA*) becomes significant to discriminate samples with intermediate DA activity. Samples with mtDNA content <0.765 exhibited intermediate DA activity (Fig 1A). AKR7A2 protein content was the third most important variable for classification of cardiac DA activity. Only 7% of the heart samples ($n = 3$) exhibited AKR7A2 protein levels $<11.5 \text{ nmol/g protein}$ (i.e., log transformed AKR7A2 $<1.06 \text{ nmol/g}$), and were classified as having high DA. Analysis of cardiac protein expression profiles suggest that high DA in samples with low AKR7A2 protein levels may result from increases in the expression of other AKRs (Supplemental Figure 1). This subset included samples from relatively old donors with- (63 years old) and without-DS (75 and 80 years old). At this point in the regression, the CART model classified 15 donors into the high, medium, and low DA categories. Following the second split from the AKR7A2 protein node, the mitochondrial parameters mtDNA content and mtDNA⁴⁹⁷⁷ deletion frequency become significant again for classifying samples into the categories of cardiac DA (Figure 1A). The splits suggest that relatively high cardiac mtDNA content would result in low DA, and high mtDNA⁴⁹⁷⁷ deletion frequency would result in intermediate DA (Figure 1A). Up to this point, only 3 parameters (AKR7A protein content, mtDNA, and mtDNA⁴⁹⁷⁷ deletion frequency) have been identified by CART as significant variables for predicting cardiac DA for over 50% of the samples. The CART model suggests that gender and Down syndrome status are significant variables for predicting cardiac DA in the remaining subset of 19 samples (Figure 1A). Comprehensive details of fit at the various split point and surrogate variables are available in the Supplemental File.

An attractive feature of CART methodology is that variables that do not improve the prediction in a recursive sense do not necessarily enter into the final model, but may still be important in the generation of the final model. Variable importance is a commonly used measure that captures the relative impact of the predictors on the response variable. This calculation assesses the relative importance of each variable across the entire tree, as a split point and surrogate, which provides complementary (and univariate) insights into the relationship between predictors and response. The ranking of the predictors in terms of variable importance suggests that variables such as age, cardiac AKR1A1 protein content,

and methylation status in specific CpG sites in the promoter region of *AKR7A2* (i.e., % methylation at CpG site 1) are important, although not used for splitting in the model itself (Figure 1B). In previous work, we have shown that these high-ranking variables (e.g., *AKR7A2* protein content and *AKR7A2* methylation status) do indeed impact DA in heart tissue^{10; 11}. There are several reasons why these variables may not arise in the final decision tree (Supplemental File).

This pilot study is limited by the small sample size, and missing information for some variables. For example, the CART model was developed by incorporating data on CBR3 protein content available for only 25% of the samples; however, leaving CBR3 protein data out did not impact the overall performance of the final model. Non-parametric CART regression is well-suited to handle the “small n, large p” problem that is common among parametric approaches¹⁹. CART-based models are suitable for carrying out partitioning with missing information leading to relatively accurate predictions for the dependent variable¹².

Findings from the CART model suggest that *AKR7A2* protein content, mtDNA content, and mtDNA⁴⁹⁷⁷ deletion frequency are the most significant factors in determining the synthesis of cardiotoxic daunorubicinol in cardiac tissue. Further studies including a larger sample size, and robust estimates for the most relevant variables are needed to confirm this pilot model. This study represents a first step towards the development of more comprehensive models to characterize the system’s response at the organ level during pharmacotherapy with anthracyclines²⁰.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research in this report was supported by the National Institute of General Medical Sciences and the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under awards R01GM073646 and R03HD076055. RHB was also supported through NSF DMS 1557594 and NSF DMS 1312250. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

References

1. Smith LA, Cornelius VR, Plummer CJ, Levitt G, Verrill M, Canney P, Jones A. Cardiotoxicity of anthracycline agents for the treatment of cancer: systematic review and meta-analysis of randomised controlled trials. *BMC Cancer*. 2010; 10:337. [PubMed: 20587042]
2. Krischer JP, Epstein S, Cuthbertson DD, Goorin AM, Epstein ML, Lipshultz SE. Clinical cardiotoxicity following anthracycline treatment for childhood cancer: the Pediatric Oncology Group experience. *J Clin Oncol*. 1997; 15:1544–1552. [PubMed: 9193351]
3. O’Brien MM, Taub JW, Chang MN, Massey GV, Stine KC, Raimondi SC, Becton D, Ravindranath Y, Dahl GV. Cardiomyopathy in children with Down syndrome treated for acute myeloid leukemia: a report from the Children’s Oncology Group Study POG 9421. *J Clin Oncol*. 2008; 26:414–420. [PubMed: 18202418]
4. Minotti G, Parlani M, Salvatorelli E, Menna P, Cipollone A, Animati F, Maggi CA, Manzini S. Impairment of myocardial contractility by anticancer anthracyclines: role of secondary alcohol metabolites and evidence of reduced toxicity by a novel disaccharide analogue. *Br J Pharmacol*. 2001; 134:1271–1278. [PubMed: 11704647]

5. Olson RD, Mushlin PS, Brenner DE, Fleischer S, Cusack BJ, Chang BK, Boucek RJ Jr. Doxorubicin cardiotoxicity may be caused by its metabolite, doxorubicinol. *Proc Natl Acad Sci U S A*. 1988; 85:3585–3589. [PubMed: 2897122]
6. Barry E, Alvarez JA, Scully RE, Miller TL, Lipshultz SE. Anthracycline-induced cardiotoxicity: course, pathophysiology, prevention and management. *Expert opinion on pharmacotherapy*. 2007; 8:1039–1058. [PubMed: 17516870]
7. Lipshultz SE, Lipsitz SR, Sallan SE, Dalton VM, Mone SM, Gelber RD, Colan SD. Chronic progressive cardiac dysfunction years after doxorubicin therapy for childhood acute lymphoblastic leukemia. *J Clin Oncol*. 2005; 23:2629–2636. [PubMed: 15837978]
8. Hefti E, Quinones-Lombrana A, Redzematovic A, Hui J, Blanco JG. Analysis of mtDNA, miR-155 and BACH1 expression in hearts from donors with and without Down syndrome. *Mitochondrial DNA*. 2014:1–8.
9. Kalabus JL, Sanborn CC, Jamil RG, Cheng Q, Blanco JG. Expression of the anthracycline-metabolizing enzyme carbonyl reductase 1 in hearts from donors with Down syndrome. *Drug Metab Dispos*. 2010; 38:2096–2099. [PubMed: 20729274]
10. Quinones-Lombrana A, Ferguson D, Hageman Blair R, Kalabus JL, Redzematovic A, Blanco JG. Interindividual variability in the cardiac expression of anthracycline reductases in donors with and without down syndrome. *Pharm Res*. 2014; 31:1644–1655. [PubMed: 24562808]
11. Hoefer CC, Quinones-Lombrana A, Blair RH, Blanco JG. Role of DNA Methylation on the Expression of the Anthracycline Metabolizing Enzyme AKR7A2 in Human Heart. *Cardiovasc Toxicol*. 2015
12. Breiman, L.; Friedman, J.; Stone, CJ.; Olshen, RA. *Classification and Regression Trees*. Taylor & Francis; 1984.
13. Auffray C, Adcock IM, Chung KF, Djukanovic R, Pison C, Sterk PJ. An integrative systems biology approach to understanding pulmonary diseases. *Chest*. 2010; 137:1410–1416. [PubMed: 20525651]
14. Hsu CW, Yin PH, Lee HC, Chi CW, Tseng LM. Mitochondrial DNA content as a potential marker to predict response to anthracycline in breast cancer patients. *The breast journal*. 2010; 16:264–270. [PubMed: 20408822]
15. Gonzalez-Covarrubias V, Zhang J, Kalabus JL, Relling MV, Blanco JG. Pharmacogenetics of human carbonyl reductase 1 (CBR1) in livers from black and white donors. *Drug Metab Dispos*. 2009; 37:400–407. [PubMed: 19022938]
16. Lakhman SS, Ghosh D, Blanco JG. Functional significance of a natural allelic variant of human carbonyl reductase 3 (*CBR3*). *Drug Metab Dispos*. 2005; 33:254–257. [PubMed: 15537833]
17. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*. Springer; 2009.
18. Therneau, TM.; Atkinson, EJ. *An introduction to recursive partitioning using the RPART routines*. Mayo Foundation; 1997. p Technical Report
19. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*. 2009; 14:323–348. [PubMed: 19968396]
20. Weiss M. Functional characterization of drug uptake and metabolism in the heart. *Expert Opin Drug Metab Toxicol*. 2011; 7:1295–1306. [PubMed: 21870896]

Figure 1.
(A) CART model for cardiac Daunorubicin Activity (DA). Circles represent predicting variants and rectangles represent terminal nodes designating cardiac donors into high, medium or low DA. (B) Variable importance in the CART model for cardiac DA.

Table 1

Variables considered as potential predictors in CART modeling.

Item	Definition	Type of Data	Coding
<i>DA</i>	Daunorubicin reductase activity (nmol DAUNol/mg*min)	Continuous	
<i>AKK7A2</i>	Protein levels of AKR7A2 (nmol/g protein)	Continuous	
<i>AKR1C3</i>	Protein levels of AKR1C3 (nmol/g protein)	Continuous	
<i>AKR1A1</i>	Protein levels of AKR1A1 (nmol/g protein)	Continuous	
<i>CBR3</i>	Protein levels of CBR3 (nmol/g protein)	Continuous	
<i>CBR1</i>	Protein levels of CBR1 (nmol/g protein)	Continuous	
<i>CONDITION</i>	Non-Down syndrome and Down syndrome	Categorical	0= Non-DS, 1=DS
<i>AGE</i>	Age of donors	Continuous	
<i>GENDER</i>	Male or Female	Categorical	0=Male, 1=Female
<i>RACE</i>	White and Non-whites	Categorical	1= White, 2=Non-White
<i>mtDNA</i>	Mitochondrial DNA ratio = MT-ND1/18S rRNA	Continuous	
<i>CpG1</i>	CpG Site-877 in AKR7A2 (0–100% methylated)	Continuous	
<i>CpG2</i>	CpG Site-865 in AKR7A2 (0–100% methylated)	Continuous	
<i>CpG3</i>	CpG Site-852 in AKR7A2 (0–100% methylated)	Continuous	
<i>CBR1 rs9024</i>	<i>CBR1</i> genotype rs9024	Categorical	0=GG(G), 1=G(G)A, 2=AA(A), where (X) is for the extra copy of chromosome 21 in DS
<i>mtDNA deletion</i>	$\% \text{ frequency } 4977 \text{ bp mtDNA deletion} = \frac{\text{Copy number } 4977 \text{ bp mtDNA deletion}}{\text{Copy number total mtDNA}} \times 100$		
<i>CBR3 V244M</i>	<i>CBR3</i> genotype V244M (rs1056892)	Categorical	0=GG(G), 1= G(G)A or G(A)A, 2=AA(A), where (X) is for the extra copy of chromosome 21 in DS