False Positive Cardiotoxicity Events in Cancer-Related Clinical Trials: Risks Related to Imperfect Noninvasive Parameters

Michael S. Ewer^a Jay Herson^b

^aDepartment of Cardiology, Division of Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA;

^bDepartment of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Keywords

Adverse drug reactions · Cardiac monitoring in clinical trials · Cardiac ultrasound · Cardiotoxicity · False positive tests

Abstract

Purpose: Cardiac ultrasound provides important structural and functional information that makes identification of cardiac abnormalities possible. Left ventricular ejection fraction (LVEF) provides the most commonly used parameter for recognition of treatment-related cardiac dysfunction. Random reading variance and physiologic factors influence LVEF and make the reported value imperfect. We attempt to quantify the likelihood of false positive events by computer simulation. **Methods:** We simulated four visits on hypothetical trials. We assumed a baseline LVEF of 55% and normal distribution with regard to reading error and physiologic variation. 1,000 trials of sample size 1,500 were simulated. In a separate simulation, 1,000 patients entered with LVEFs of 45, 43, and 41% to estimate true positive incidence. **Results:** At each examination, less than 1.0% of false positives were noted. The cumulative false positive rate over four visits was 3.60%. True cardiotoxicity identification is satisfactory only when LVEF declines substantially. **Conclusion:** A 3.60% false

positive rate in trials where the expected level of toxicity is low suggests that false positives are troubling and may exceed true positive results. Strategies to reduce the number of false positive results include making confirmatory studies mandatory. Evaluating increases along with decreases obtains some estimation of variance.

© 2019 S. Karger AG, Basel

Introduction

Cardiac evaluation of patients undergoing cancer treatment with agents that are potentially cardiotoxic presents many interpretative challenges. While cardiac injury from agents such as anthracyclines is well understood and is now known to occur early after exposure, the clinical presentation may be delayed for months or years due to extensive cardiac reserves [1]. Myocyte death may remain clinically undetectable as we use the generally available noninvasive tests, as a considerable degree of cell loss initially may not alter systolic or diastolic function. While in the past, cardiac biopsy allowed for early estimations of true damage long before the injury became clinically manifest, biopsies have become a

rarity as cumulative dosages have declined, and as non-invasive imaging techniques have improved. Nevertheless, attempts to estimate cardiac damage noninvasively continue to play a highly important role as we assess cancer patients; we make important decisions based on these data – the consequences of a wrong decision can be huge and go beyond the treatment of individual patients in that they may preclude or delay allowing a potentially curative drug to receive approval. Notwithstanding genuine improvements in ultrasound technology, the techniques remain suboptimal.

Cardiac ultrasound, presently the most frequently applied noninvasive imaging modality, is plagued with two important confounding detractors when used to estimate the extent of true cardiac dysfunction specifically related to a potentially cardiotoxic agent. The first is inter- and intra-observer variability, and that problem has yet to be overcome even in the age of three-dimensional imaging, contrast enhancement, and strain techniques. The second confounding factor is that many physiologic variables affect the ejection fraction, and we do not have the ability to separate these variables from the effect of the potentially cardiotoxic agent under investigation. We recognize that the heart compensates for early damage, and subtle compensatory alterations are often lost in the noise of suboptimal specificity and sensitivity related to these confounding influences.

Both interpretive imprecision and physiologic variation may augment or diminish the reported ejection fraction as an indicator of cardiac contractility. Variations related to physiologic or metabolic alteration may be the consequence of hormonal effects, medications, high-output states related to shunting of blood through the tumors, vasodilatory effects of cancer drugs, or anemia, as well as increases in systemic resistance to vasoconstrictive drugs; they represent true alteration in systolic function, but these alterations should not imply a direct toxic effect of an agent under consideration. These changes cannot be separated easily from the effects of the specific agent we seek to study or evaluate. It is in this setting that we attempt to investigate the extent that these factors, unrelated to the cardiotoxic agent itself that we are studying, play in our assessment, and how they may lead to conclusions that are incorrect with regard to the extent of cardiac injury or toxicity that results from a specific agent. Misinterpretation of the etiology of changes in cardiac function derived non-invasively leads to decisions that may be detrimental to individual patients or compromise the data generated as part of clinical trials.

In an effort to estimate the extent that such factors play in altering the predictive value of cardiac testing in the cancer patient, we created a group of statistical models to ascertain the extent that patients with normal hearts might be deemed to have experienced cardiotoxicity due to interpretative and biologic variables when a hypothetical agent without cardiac effect was administered. Our model is analogous with regard to multiple assessments of ejection fraction as have been found in clinical trials as well as when clinical guidelines are followed in the treatment of individual patients.

Methods

A group of 1,500 hypothetical subjects in each of 1,000 clinical trials was included in this computer simulation. The assumptions were that all had an equal baseline left ventricular ejection fraction (LVEF) of 55%, and that the distribution of interpretive error followed a normal distribution with mean zero and standard deviation 2.5. This reference distribution is consistent with a 95% confidence interval of $\pm 5\%$ and approximates a conservative estimation of the interpretive variance of clinical reading of cardiac ultrasound [2]. We made the same distributional assumption regarding biologic variance, but unlike the reading error, the biologic variation was applied only at the four post-baseline hypothetical visits. These characteristics were chosen so as to simulate what might be seen in the initial and subsequent evaluations of patients as part of a clinical trial of a new potentially cardiotoxic agent. The baseline LVEF adjusted by random reading error was considered the baseline LVEF for all subsequent calculations. A small number of patients with an initial reading of $<50\%$, a scenario that might ensue when interpretive error was unusually high, were dropped from further inclusion; in this hypothetical model, they were deemed to have met exclusion criteria as may be defined in a protocol for a clinical trial.

LVEF assessment was repeated four times to simulate what is often recommended in clinical trials or what might be incorporated in the cardiac follow-up during and after treatment in a clinical setting. The cardiac ultrasound is repeated every 3 months during 1 year of treatment, and a further measurement is undertaken during the post-treatment period. All patients would be evaluated a total of five times including the baseline assessment; those who met the threshold of a cardiac event (described below) would continue to be observed as per the predetermined hypothetical follow-ups to assess the degree of presumed reversibility of what had been deemed a cardiac event. Any of the patients who met toxicity criteria at any of the four determinations following the baseline assessment would be identified as having experienced a cardiac adverse event; all would represent false positive adjudications.

As in a number of clinical trials, the thresholds were a decrease of 10 percentage points to a value of less than 50%, or any absolute decrease of 15 percentage points. Recovery after an event was defined as an increase in LVEF to within 5 percentage points of the baseline value [3]. In our assumptions, no agent was given, simulating a hypothetical situation of no direct or indirect temporary or permanent cardiac effect. In that regard, all patients who met

Table 1. Percentiles of cumulative cardiotoxicity incidence (%) by visit – 1,000 replicates of clinical trials of 1,500 patients

Visit No.	Minimum	Percentiles					Maximum
		2.5%	25%	50% (median)	75%	97.5%	
1	0.27	0.47	0.80	0.93	1.13	1.47	1.93
2	0.80	1.20	1.60	1.87	2.07	2.60	3.00
3	1.60	1.93	2.47	2.73	3.00	3.60	4.07
4	2.13	2.67	3.27	3.60	3.93	4.60	5.00

At visit 4, mean incidence = 3.60%, SD = 0.48, 95% CI (2.66%, 4.54%).

the criteria for an event, even those who had experienced a true decline due to biologic factors, would constitute false positive results; it is implied that none of these hypothetical subjects had actual declines in ejection fraction related to the drug under consideration. We report the cumulative incidence of presumed adverse events over the four determinations. Cumulative incidence was computed by the Kaplan-Meier life table method [4]. As noted, all are false positives. The trial was replicated 1,000 times. Calculations regarding statistical power were made with reference to the normal distribution using common frequentist statistical methods as provided by PASS software (version 15, NCSS, 2017).

Considerations regarding an Agent That Exhibits True Toxicity

Additionally, we strove to answer a second question as to when, i.e. at what level of ejection fraction change, true toxicity could be detected with statistical accuracy. We assumed the same definitions of toxicity, but modeled our hypothetical population as having received an agent that did have, by definition, true toxicity. In that scenario, a patient enters visit 2 having simulated ejection fractions of 45, 43, and 41%; we assumed a hypothetical cohort of 1,000 patients. Again, we applied the reader error and physiologic adjustment as described above and recorded the positive and negative outcome. We then repeated the simulation analogous to an independent retest 1 week later, to quantify the importance of retesting. We report a summary, in clinical terms, of bivariate positive/negative outcomes for the tests.

Results

Table 1 presents the results of the computer simulation of 1,000 clinical trials each with 1,500 patients on a treatment with no direct toxic effects and who enter the trial with an LVEF of 55%. The percentiles of cumulative cardiotoxicity incidence are shown as a percent through visits 1 through 4. The median cumulative incidence increases linearly from visits 1 (0.93%) through 4 (3.60%). At visit 4, the median reported cardiotoxicity is 3.60% with the interquartile range spanning from 3.27 to 3.93%. The mean incidence of cardiotoxicity at visit 4

Table 2. Sample size necessary to reject random toxicity hypothesis at selected levels of observed cardiotoxicity

Power	Observed cardiotoxicity		
	5.5%	7.5%	9.5%
0.80	722	210	107
0.90	1,013	281	147

One-sided test, significance level = 0.05. Random toxicity hypothesis level = 3.6%.

was also 3.60% with standard deviation 0.48. This results in a 95% confidence interval of 2.66–4.54%. This range compares favorably with the central percentile span 2.5–97.5 which was 2.67–4.60%. Thus, it is reasonable to assume that the true cumulative false positive cardiotoxicity incidence is 3.60%.

Sample sizes for oncology clinical trials often are estimated to detect an effect size for efficacy variables such as progression-free survival or overall survival. In designing metastatic breast cancer clinical trials oncologists would want to know if these efficacy-based sample sizes in the experimental arm would be sufficient to deliver the power to reject a false positive cardiotoxicity hypothesis. Sample size guidelines are shown in Table 2 for one-sided hypothesis tests against a false positive cardiotoxicity rate of 3.60% at significance level 0.05. If we wanted to distinguish a 5.5% cardiotoxicity from an observed 3.6% with power 0.80, a sample size of 722 would be required; for a 90% power, n would be 1,013. The sample size requirement would be less to distinguish cardiotoxicities 7.5 and 9.5%. All sample sizes in Table 2 are smaller than the typical efficacy-based sample sizes so that inference with considerable power is supported in practice.

Table 3. Cardiotoxicity classification performance under test-retest for patients entering visit 2 with compromised LVEF (based on simulation of 1,000 patients)

LVEF on entry to visit 2	45%	43%	41%
Classified as cardiotoxicity on visit 2 and confirmed on retest (+, +)	63.5%	72.3%	85.5%
Incorrectly classified on visit 2 but corrected on retest (–, +)	40.8%	56.8%	77.1%
Incorrectly classified on visit 2 and not corrected on retest (–, –)	59.2%	43.2%	22.9%
Correctly classified as cardiotoxicity by test-retest sensitivity (–, +) and (+, +)	52.1%	69.2%	84.1%

+, correct classification of cardiotoxicity; –, incorrectly classifying as no cardiotoxicity.

In the second part of our simulations where we assume toxicity, it was important to examine the performance of the cardiotoxicity definition when patients enter visit 2 with a compromised LVEF that we presume to be treatment induced. We assumed the clinical situation where a patient classified with cardiotoxicity is retested 1 week later.

We simulated 1,000 pairs of test-retest results using the same cardiotoxicity definition as in Table 1. Table 3 presents these simulation-derived performance characteristics for three levels of treatment-affected visit 2 LVEF: 45, 43, and 41%. When LVEF is 45%, the definition is not sufficiently specific in that we could correctly classify and confirm only 63.5% of patients, and 59.2% of patients who initially are incorrectly classified are not corrected on initial test. The resulting sensitivity (incorrect-correct plus correct-correct) is only 52.1%. When LVEF on visit 2 is 41%, the definition performs much better correctly classifying and confirming 85.5% of patients, correcting 77.1% of patients initially misclassified. Only 22.9% of patients incorrectly classified on entry are not corrected on retest. This results in a sensitivity estimate of 84.1%. Thus, the definition appears to provide satisfactory detection of cardiotoxicity only when the degree of cardiac impairment due to the treatment is substantial.

Discussion

The primary goals of collaborative efforts between oncologists and cardiologists are to maximize survival while maintaining the best possible quality of life. Minimizing cardiac dysfunction must always be balanced with the potential for impaired oncologic efficacy, and the calculus strives to define, both for groups and for individual patients, an optimal balance. Both an overemphasis on cardioprotection as well as a failure to appreciate true cardiac risk may impact optimal care recommendations. The

balance would be much easier to achieve if we had parameters to measure and categorize myocyte injury or dysfunction that were sufficiently sensitive and specific so that the true cardiac burden, to an extent that it impacts on longevity or quality of life, could be ascertained with a high degree of accuracy. Unfortunately, these methods do not exist; even using cardiac biopsy we are left with ambiguities regarding long-term implications of the observed changes. Noninvasive techniques have substantial uncertainties, yet the reported values continue to have considerable impact on both immediate care and long-term surveillance. This paper attempts to quantitate these uncertainties when using cardiac ultrasound, presently the most common test used both for the evaluation of populations enrolled in clinical trials and for individual patients. It is clearly evident, however, that techniques such as three-dimensional imaging and cardiac magnetic resonance imaging have improved predictive value; yet, traditional 2-dimensional cardiac ultrasound remains widely used both in the assessment of individual patients and in the cardiac adjudication of clinical trial data. The information provided herein should help both researchers and primary oncologists understand the shortcomings of cardiac ultrasound methodology.

A number of breast cancer trials have incorporated noninvasive testing; some of these are included in Table 4, depicting the variance in defining criteria for cardiac events and the extent of subjects that met event criteria. Some have event rates that may have been skewed because of false positive endpoints as suggested by our simulations.

Understanding of the extent of false positive results in clinical trials has important application. While secondary cardiac endpoints may not be the primary concern of a trial undertaken to study efficacy, knowing the likelihood that a secondary cardiac endpoint represents a true positive result is hugely important when describing the extent of reported cardiotoxicity. Clinicians and re-

Table 4. Cardiotoxicity criteria and results from breast cancer clinical trials (treatment arms combined)

Clinical trial	Treatments	[Ref.]	Criteria	Results
NSABP B-31	doxorubicin, cyclophosphamide	[3]	decrease in LVEF from baseline by ≥ 10 pct. pts. to below 55% or 5% below LLN	7.5% (148/1,972) for patients treated with anthracyclines prior to randomization
NCCTG N9831	doxorubicin, cyclophosphamide	[5]	decrease in LVEF from baseline by ≥ 15 pct. pts. or $\leq 15\%$ pts to a value below LLN	5.0% (151/2,992) for patients treated with anthracyclines prior to randomization
NOAH	trastuzumab, doxorubicin, paclitaxel, methotrexate, 5-FU	[6]	LVEF decrease by ≥ 10 pct. pts. from baseline to lower than 50%	0.9% (2/228)
TECHNO	epirubicin, cyclophosphamide, paclitaxel, trastuzumab	[7]	LVEF decrease by ≥ 10 pct. pts. from baseline	2.8% (6/217)
CLEOPATRA	trastuzumab, docetaxel, pertuzumab	[8]	LVEF decrease by ≥ 10 pct. pts. from baseline or to $< 50\%$	5.1% (41/804)
NeoSphere	trastuzumab, docetaxel, pertuzumab	[9]	LVEF decrease ≥ 10 pct. pts. from baseline to below 50%	1.0% (4/417) during neoadjuvant period
TRYPHAENA	5-FU, epirubicin, cyclophosphamide, docetaxel, trastuzumab, pertuzumab	[10]	LVEF decrease ≥ 10 pct. pts. to $< 50\%$	neoadjuvant 4.9% (11/223) adjuvant 7.5% (15/200)
MARIANNE	trastuzumab emtansine, trastuzumab, pertuzumab, docetaxel, paclitaxel	[11]	LVEF decrease ≥ 10 pct. pts. from baseline to $< 50\%$	2.6% (28/1,080)
APHINITY	taxane of choice, anthracycline, 5-FU, pertuzumab	[12]	LVEF decrease ≥ 10 pct. pts. from baseline to below 50%	5.5% (131/2,364)
BERENICE	pertuzumab, trastuzumab, taxane of choice, anthracycline	[13]	LVEF decrease ≥ 10 pct. pts. from baseline to below 50%	4.3% (17/400)
HERA (BIG 1-01)	anthracycline, taxane of choice, trastuzumab	[14]	LVEF decrease ≥ 10 pct. pts. from baseline to below 50%	4.4% (74/1,682)

pct. pts., percentage points; LLN, lower limit of normal.

searchers should consider ejection fraction results more in the light of a screening test rather than as diagnostic test; as with many such tests, the predictive value must be considered as a criterion for validation. Our simulation demonstrates that false positive results in clinical trials may be a bigger problem than was heretofore considered. While this simulation does not provide absolute numbers, the reported range should prove helpful in estimating the extent of cardiotoxicity in future trials. While intuitive, the confirmation of a reported test abnormality suggesting a cardiac event by a repeat measurement dra-

matically reduces the likelihood of false positive results. Confirmation should be incorporated more broadly in clinical trials.

Beyond clinical trials, positive tests for individual patients undergoing standard treatment should be confirmed prior to altering effective regimens, and prior to initiating cardiac medication in instances of asymptomatic patients. Finally, it should be recognized that in keeping with the imperfect nature of cardiac ultrasound, the extent of false positive results over time is directly proportional to the number of tests performed. Over-testing

is therefore potentially dangerous, and in instances where the likelihood of true toxicity is small, should be minimized.

There are limitations to our approach in estimating false positive rates. The results are based on the assumption of the normal distribution with standard deviation 2.5 for both echocardiographic reading error and physiologic variation. These assumptions seem reasonable and, perhaps, conservative but cannot be objectively verified. We also assume that patient LVEF readings are independent across visits. It is possible that the readings are correlated so that, reading and physiologic variation notwithstanding, a patient's LVEF reading on each visit is correlated with the reading on the visit before. However, this assumption of independence is analogous to the assumption of exponential events which is inherent in analysis of survival time in oncology trials.

References

- 1 Ewer MS, Ewer SM. Cardiotoxicity of anti-cancer treatments. *Nat Rev Cardiol*. 2015 Sep; 12(9):547–58.
- 2 Nosir YF, Vletter WB, Boersma E, Frowijn R, Ten Cate FJ, Fioretti PM, et al. The apical long-axis rather than the two-chamber view should be used in combination with the four-chamber view for accurate assessment of left ventricular volumes and function. *Eur Heart J*. 1997 Jul;18(7):1175–85.
- 3 Tan-Chiu E, Yothers G, Romond E, Geyer CE Jr, Ewer M, Keefe D, et al. Assessment of cardiac dysfunction in a randomized trial comparing doxorubicin and cyclophosphamide followed by paclitaxel, with or without trastuzumab as adjuvant therapy in node-positive, human epidermal growth factor receptor 2-overexpressing breast cancer: NSABP B-31. *J Clin Oncol*. 2005 Nov;23(31):7811–9.
- 4 Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457–81.
- 5 Perez EA, Suman VJ, Davidson NE, Sledge GW, Kaufman PA, Hudis CA, et al. Cardiac safety analysis of doxorubicin and cyclophosphamide followed by paclitaxel with or without trastuzumab in the North Central Cancer Treatment Group NCCTG N9831 adjuvant breast cancer trial. *J Clin Oncol*. 2008;26:1231–8.
- 6 Gianni L, Eiermann W, Semiglazov V, Manikhas A, Lluch A, Tjulandin S, et al. Neoadjuvant chemotherapy with trastuzumab followed by adjuvant trastuzumab versus neoadjuvant chemotherapy alone, in patients with HER2-positive locally advanced breast cancer (the NOAH trial): a randomised controlled superiority trial with a parallel HER2-negative cohort. *Lancet*. 2010 Jan;375(9712):377–84.
- 7 Untch M, Fasching PA, Konecny GE, Haslmüller S, Lebeau A, Kreienberg R, et al. Pathologic complete response after neoadjuvant chemotherapy plus trastuzumab predicts favorable survival in human epidermal growth factor receptor 2-overexpressing breast cancer: results from the TECHNO trial of the AGO and GBG study groups. *J Clin Oncol*. 2011 Sep;29(25):3351–7.
- 8 Baselga J, Cortés J, Kim SB, Im SA, Hegg R, Im YH, et al.; CLEOPATRA Study Group. Pertuzumab plus trastuzumab plus docetaxel for metastatic breast cancer. *N Engl J Med*. 2012 Jan;366(2):109–19.
- 9 Gianni L, Pienkowski T, Im YH, Roman L, Tseng LM, Liu MC, et al. Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. *Lancet Oncol*. 2012 Jan;13(1):25–32.
- 10 Schneeweiss A, Chia S, Hickish T, Harvey V, Eniu A, Hegg R, et al. Pertuzumab plus trastuzumab in combination with standard neoadjuvant anthracycline-containing and anthracycline-free chemotherapy regimens in patients with HER2-positive early breast cancer: a randomized phase II cardiac safety study (TRYPHAENA). *Ann Oncol*. 2013 Sep;24(9):2278–84.
- 11 Perez EA, Barrios C, Eiermann W, Toi M, IM Y-H, Conte P, et al. Trastuzumab emtansine with or without pertuzumab versus trastuzumab plus taxane for human epidermal growth factor receptor 2-positive, advanced breast cancer: primary results from the phase III MARIANNE study. *J Clin Oncol*. 2017 Jan;35(2):142–8.
- 12 von Minckwitz G, Procter M, de Azambuja E, Zardavas D, Benyunes M, Viale G, et al.; APHINITY Steering Committee and Investigators. Adjuvant pertuzumab and trastuzumab in early HER2-positive breast cancer. *N Engl J Med*. 2017 Jul;377(2):122–31.
- 13 Swain SM, Ewer MS, Viale G, Delaloge S, Ferrero JM, Verrill M, et al. Pertuzumab, trastuzumab and standard anthracycline- and taxane-based chemotherapy for the neoadjuvant treatment of patients with HER2-positive localized breast cancer (BERENICE): A phase II, open-label, multicenter, multinational cardiac safety study. *Ann Oncol*. 2018 Mar;29(3):646–653.
- 14 Cameron D, Piccart-Gebhart MJ, Gelber RD, Procter M, Goldhirsch A, de Azambuja E, et al.; Herceptin Adjuvant (HERA) Trial Study Team. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. *Lancet*. 2017 Mar;389(10075):1195–205.

Statement of Ethics

No human subject data were accessed or analyzed in this study.

Disclosure Statement

The authors declare no conflicts of interest.

Funding Sources

There was no institutional or external funding used in the creation of this paper.