Contemporary Clinical Trials Communications 3 (2016) 48–54

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Genomic and clinical predictors for improving estimator precision in randomized trials of breast cancer treatments

Prasad Patil, Elizabeth Colantuoni, Jeffrey T. Leek*, Michael Rosenblum*

Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA

ARTICLE INFO

Article history:
Received 28 September 2015
Received in revised form
7 December 2015
Accepted 21 March 2016
Available online 31 March 2016

Keywords:
Adjustment
Genomics
Precision
Translation

ABSTRACT

Background: The hope that genomic biomarkers would dramatically and immediately improve care for common, complex diseases has been tempered by slow progress in their translation beyond bioinformatics. We propose a novel use of genomic information where the goal is to improve estimator precision in a randomized trial. We analyze the potential precision gains from the popular MammaPrint genomic signature and clinical variables in simulations of randomized trials analyzed using covariate adjustment.
Methods: We apply an estimator for the average treatment effect in the trial that adjusts for prognostic baseline variables to improve precision [1]. This precision gain can be translated directly into sample size reduction and corresponding cost savings. We conduct simulation studies based on resampling genomic and clinical data of breast cancer patients from four publicly available observational studies.
Results: Separate simulation studies were conducted based on each of the four data sets, with sample sizes ranging from 115 to 307. Adjusting only for clinical variables provided gains of −1%, 5%, 6%, 17%, compared to the unadjusted estimator. Adjusting only for the MammaPrint genomic signature provided gains of 2%, 4%, 4%, 5%. Simultaneously adjusting for clinical variables and the genomic signature provided gains of 2%, 6%, 7%, 16%. The differences between precision gains from adjusting for genomic plus clinical variables, versus only clinical variables, were −1%, 0%, 1%, 3%.
Conclusions: Adjusting only for clinical variables led to substantial precision gains (at least 5%) in three of the four data sets, with a 1% precision loss in the remaining data set. These gains were unchanged or increased when sample sizes were doubled in our simulations. The precision gains due to incorporating genomic information, beyond the gains from adjusting for clinical variables, were not substantial.
© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The announcement of the Precision Medicine Initiative [2] stated that “Precision medicine's more individualized, molecular approach to cancer will enrich and modify, but not replace, the successful staples of oncology – prevention, diagnostics, some screening methods, and effective treatments – while providing a strong framework for accelerating the adoption of precision medicine in other spheres.” In the realm of genomic biomarker development, this mandate puts an explicit focus on “enrichment”, i.e. how much *additional* information a new marker can provide to supplement the standard course of care. The uncertain value of genomic measurements for improving clinical practice has been a critical roadblock in the translation of genomic markers to the clinic [3], in addition to problems with reproducibility [4], interpretability [5], and cost [6]. A small number of laboratory tests based on genomic signatures have been approved for clinical use. Tests such as MammaPrint [7], Oncotype DX [8], and Prosigna [9] rely on measurement of expression for a set of genes that are associated with differential survival and severity of breast cancer cases.

It is difficult to evaluate the clinical value that these genomic signatures add beyond standard clinical factors measured for all breast cancer patients, such as age, estrogen receptor status, tumor size, and tumor grade. It is also known that tests based on genomic signatures are not part of the standard of care in many cases [10]; [3]. Ongoing clinical trials are being performed to ascertain the value of some of these signatures to make adaptive treatment decisions [11]. We propose to evaluate the use of genomic signatures in a different setting by considering the prognostic value added from adjusting for a genomic signature in a randomized clinical trial of a new treatment versus control.

* Corresponding authors. 615 N. Wolfe St., Baltimore, MD 21205, USA.
E-mail addresses: jtleek@gmail.com (J.T. Leek), mrosen@jhu.edu (M. Rosenblum).

http://dx.doi.org/10.1016/j.conctc.2016.03.001
2451-8654/© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

In a randomized trial the primary analysis typically involves estimating the average treatment effect. Adjusting for baseline variables that are prognostic for the outcome can lead to improved precision in estimating the average treatment effect at large sample sizes (i.e., asymptotically as sample size grows). Yang and Tsiatis [12] showed that for continuous outcomes and a linear model with main terms, the analysis of covariance (ANCOVA) estimator is guaranteed to be consistent and as or more precise than the standard unadjusted estimator, even if the linear model is not correctly specified, i.e., the true distribution of the outcome given baseline covariates may be much more complex than the linear model used, and still the guarantee holds.

More recently, estimators with the same desirable property as the ANCOVA procedure have been extended to binary and count outcomes; see Cao et al. [13]; Tan [14]; Rotnitzky et al. [15] and Gruber and van der Laan [16]. Colantuoni and Rosenblum [1] provide a review of these recent estimators, which are designed to estimate an average treatment effect in the general setting of an observational study, where the probability of being assigned to treatment is not randomized and must be learned from the data. These estimators may also be applied to randomized trials, where their guarantees on improved precision require fewer assumptions than in an observational study since in a randomized trial the assignment probability is known (and set by design).

The above estimators all have the aforementioned consistency and precision guarantee. One difference among them is that the estimators of Colantuoni and Rosenblum [12]; Tan [14]; and Colantuoni and Rosenblum [1] do not require solving a non-convex (and therefore computationally challenging) optimization problem; however, the benefit of solving such a problem, as done by the estimators of Cao et al. [13]; Rotnitzky et al. [15] and Gruber and van der Laan [16]; is that they have potential for further precision gains, so there is a computation versus precision tradeoff.

The precision gains provided by adjusting for baseline variables depend on how correlated the baseline variables are with the outcome and the degree of chance imbalance in the baseline variables across the treatment groups. To the best of our knowledge, the value of such adjustment has not yet been assessed using simulations based on resampling from breast cancer patient data sets, as we do here. We resample in a way that preserves correlations between baseline variables and the outcome in order to give a realistic assessment (as best as we can using simulations and our data sets) of the magnitude of precision gains likely to be observed in practice.

We aim to determine the prognostic value of clinical and/or genomic variables measured at baseline (pre-randomization). Of particular interest is the additional gain from adjusting for the genomic signature beyond that obtained by adjusting for standard clinical baseline variables. Our definition of precision gain in this setting equals the percent sample size reduction from using the adjusted estimator compared to the unadjusted estimator in order to attain the same power, asymptotically. Although perhaps not as groundbreaking of a result as once hoped, this approach represents a realistic attempt to assess the value of the information provided by a genomic signature.

2. Methods

2.1. Data

Microarray data used to validate the MammaPrint model [17] were gathered as described in the appendix of Marchionni et al. [18]. The MammaPrint validation data set consists of 307 breast cancer patients. Table 1 summarizes the key clinical factors recorded for these patients as well as their MammaPrint risk prediction, which is a binary classification based on the risk score calculated by the MammaPrint model [7]. We dropped 11 patients whose estrogen receptor (ER) status or tumor grade were unknown and conducted our analysis using the 296 remaining patients.

Table 1
MammaPrint validation data set. ER - estrogen receptor status, Grade - tumor severity grading (3 is most severe), Five-Year Recurrence - whether or not cancer has reappeared after five years, MammaPrint risk prediction - high or low risk for cancer recurrence. Age and Tumor Size are given as means with standard deviations in parentheses.

Characteristic	Summary
n	307
Age (years)	47.08 (7.27)
Five-Year Recurrence	
Yes	47
No	260
Tumor Size (mm)	21.48 (7.71)
Grade	
1	47
2	126
3	126
Unknown	8
ER	
+	212
−	90
Unknown	5
MammaPrint Risk Prediction	
High	194
Low	113

We also conduct simulations based on three external breast cancer data sets described in the Supplementary Material. These are called GSE19615, GSE11121, GSE7390, with sample sizes 115, 200, 198, respectively.

2.2. Statistical method to adjust for baseline covariates

We define the average treatment effect to be the difference between the population mean of the primary outcome under assignment to treatment and the population mean under assignment to control. The term “covariate adjustment” means that information from baseline variables is used to improve the precision in estimating the average treatment effect. This is done by adjusting for chance imbalances in baseline variables between treatment and control arms. Since our focus is improved precision for estimating the average treatment effect, we do not consider effects within subgroups; investigating the latter is an area for future research.

Increased precision for estimation of the average treatment effect can lead to a trial with fewer participants and shorter duration, compared to a trial with the same power that uses a less precise estimator. This is because the sample size for a trial is typically selected in order to achieve a desired power, e.g., 80% or 90%, at an alternative (e.g., the minimum, clinically meaningful effect size); using a more precise estimator leads to a smaller required sample size to achieve the power goal. More precise estimators can be used to reduce the sample size even when the average treatment effect is zero, which is the setting of our simulation study. This can be achieved by prespecifying the sample size as that which achieves a desired power at a given alternative, taking into account the percent variance reduction from using the adjusted estimator compared to the unadjusted estimator. A more flexible approach is to use information based monitoring, where the trial runs until a preplanned information level has accrued (see, e.g., Jennison and Turnbull [19]. Information with respect to a given estimator, defined as the reciprocal of its variance, accrues faster for estimators with greater precision, leading to smaller sample sizes.

We assume each participant in the trial contributes a data vector $D = (W,A,Y)$, where $W = (W_1, \ldots, W_j)$ is a vector of covariates measured at baseline, A is an indicator of study arm (0 = control, 1 = treatment), and Y is a binary outcome of interest which in our case is the indicator of cancer recurrence within 5 years from baseline. We assume the trial data consist of n independent, identically distributed participant data vectors $\{D_i\}_{i=1}^n$ drawn from unknown joint distribution P on (W, A, Y). We assume a nonparametric model except that W and A are independent by randomization (called the randomization assumption), and we assume the regularity conditions in Ref. [1]; Section 3.2).

The goal is to estimate the average treatment effect defined as the difference between 5 year survival probabilities comparing treatment versus control, i.e.,

$$\psi = E[Y|A=1] - E[Y|A=0] = P(Y=1|A=1) - P(Y=1|A=0). \quad (1)$$

Another possible treatment effect, which we do not consider, is the hazard ratio under a proportional hazards model. This would have the advantage that the recurrence time (rather than only the indicator Y of recurrence by 5 years) is fully used; however, a disadvantage is that inferences depend on the proportional hazards assumption being correct, and these inferences would typically be biased (even at large sample sizes) if that assumption fails to hold.

The unadjusted estimator of ψ is defined as

$$\widehat{\psi}_{una} = \frac{\sum_{i=1}^n Y_i A_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n Y_i (1-A_i)}{\sum_{i=1}^n (1-A_i)}.$$

This estimator is consistent (i.e., converges in probability to the population average treatment effect ψ) but ignores the baseline variables W. If W is prognostic for Y then it is possible to improve precision by appropriately adjusting for W. Throughout, we do not assume that W contains information about treatment effect heterogeneity, i.e., who benefits more or less from treatment; we only use W as prognostic variables that may explain some of the variation in Y. This variation could be unrelated to treatment.

To leverage the information in W, we apply the enhanced efficiency, doubly-robust estimator of Colantuoni and Rosenblum ([1], Section 4.2), which is a special case of the class of estimators from Rotnitzky et al. [15] that is slightly modified for use in the randomized trial context. We denote this estimator by $\widehat{\psi}_{adj}$. Software to compute this estimator is given in R and SAS by Colantuoni and Rosenblum [1]. The R code we used is available at the link in Section 2.5.

The estimator $\widehat{\psi}_{adj}$ uses parametric working models for the mean of the outcome given baseline variables and study arm. We call these working models since we do not assume they are correctly specified. The true data generating distribution may differ arbitrarily from the functional form of the model.

Computation of $\widehat{\psi}_{adj}$ is accomplished via the following steps:

1. Let $\alpha = (\alpha_0, \ldots, \alpha_j)^T$. Fit the following propensity score working model for $P(A=1|W)$: $g(W,\alpha) = \text{logit}^{-1}(\alpha_0 + \alpha_1 W_1 + \ldots + \alpha_j W_j)$ via maximum likelihood estimation and denote the estimator of α by $\widehat{\alpha} = (\widehat{\alpha}_0, \ldots, \widehat{\alpha}_j)^T$.
2. For each arm $a \in \{0,1\}$, define the following working model for $E(Y|A=a,W)$:$Q^{(a)}(W,\beta^{(a)}) = \text{logit}^{-1}(\beta_0^{(a)} + \beta_1^{(a)} W_1 + \ldots \beta_j^{(a)} W_j)$. Fit the above model at $a=1$ using weighted logistic regression with weights $\frac{1}{g(W,\widehat{\alpha})}$ and using only participants with $A=1$ to obtain estimated coefficients $\widehat{\beta}^{(1)} = (\widehat{\beta}_0^{(1)}, \ldots, \widehat{\beta}_j^{(1)})$. Define the initial estimator for $E[Y|A=1]$ as $\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n Q^{(1)}(W_i, \widehat{\beta}^{(1)})$, where the sum is taken over all participants. The estimator $\widehat{\mu}_0$ for $E[Y|A=0]$ is obtained analogously by setting $a=0$, replacing $A=1$ with $A=0$, and replacing $\frac{1}{g(W,\widehat{\alpha})}$ by $\frac{1}{1-g(W,\widehat{\alpha})}$ above.
3. Define the new covariate $\mu_a(W) = Q^{(a)}(W, \widehat{\beta}^{(a)}) - \widehat{\mu}_a$ for each $a \in \{0,1\}$, which uses $\widehat{\mu}_a, \widehat{\beta}^{(a)}$ as estimated in step 2. Fit the following augmented propensity score model for $P(A=1|W)$: $g_{aug}(W,\alpha,\gamma) = \text{logit}^{-1}(\alpha_0 + \alpha_1 W_1 + \ldots + \alpha_j W_j + \gamma_0 \mu_0(W) + \gamma_1 \mu_1(W))$ using maximum likelihood estimation to obtain estimated coefficients $\tilde{\alpha}$ and $\tilde{\gamma} = (\tilde{\gamma}_0, \tilde{\gamma}_1)$.
4. Recompute step 2 using $g_{aug}(W, \tilde{\alpha}, \tilde{\gamma})$ in place of $g(W, \widehat{\alpha})$ in the weights to obtain new estimates $\tilde{\mu}_1, \tilde{\mu}_0$. Define the adjusted estimator of the average treatment effect as $\widehat{\psi}_{adj} = \tilde{\mu}_1 - \tilde{\mu}_0$.

Throughout, we assume there are no missing data and the vector (W_i,A_i,Y_i) is observed for each participant i. The models g and g_{aug} are correctly specified as long as each contains an intercept, due to the randomization assumption. By design, each participant is assigned to treatment or control with probability 0.5, independent of his/her baseline variables, so $P(A=1|W=w) = P(A=1) = 0.5$ for all values of w. Consider the model

$$g(W,\alpha) = P(A=1|W) = \text{logit}^{-1}(\alpha_0 + \alpha_1 W_1 + \ldots + \alpha_k W_k)$$

Setting $\alpha_1 \ldots \alpha_k = 0$ and $\alpha_0 = \text{logit}(1/2)$ yields correct specification of the model, i.e., the model at these parameter values equals the true distribution $P(A=1|W) = P(A=1) = 1/2$. The same holds for g_{aug}. Though the data generating distribution has A independent of W, in any given realization of the data there can be imbalances in W across arms due to chance variation.

The models $Q^{(0)}$,$Q^{(1)}$ will typically be misspecified if any of the baseline variables is continuous valued or has many discrete levels. An important feature of the estimator $\widehat{\psi}_{adj}$ is that it is consistent regardless of whether the parametric models $Q^{(0)}$,$Q^{(1)}$ are correctly specified; that is, consistency holds even when the true data generating distribution $E(Y|A=a,W)$ does not have the form $Q^{(a)}(W,\beta^{(a)})$ for any β. Furthermore, the estimator $\widehat{\psi}_{adj}$ is guaranteed to have asymptotic precision equal to or greater than that of the unadjusted estimator as proved by Refs. [15]; [1]. However, depending on the number of baseline covariates and the sample size, the precision may be worse for the adjusted estimator compared to the unadjusted estimator; this can happen if the baseline variables are only weakly (or not at all) prognostic, there are more than a few of them, and the sample size is relatively small.

It is also possible to use the output of step 2 to construct the simpler estimator $\widehat{\mu}_1 - \widehat{\mu}_0$ of the average treatment effect. This estimator is called the double-robust weighted least squares estimator (DR-WLS) and is attributed to Marshall Joffe by Robins et al. [20]. The value of adding steps 3 and 4 is that the resulting estimator has been proved to be asymptotically as or more precise than the unadjusted estimator [15]; [1].

2.3. Baseline covariates used for adjustment

The baseline variables W used in the estimators defined above must be pre-specified. They can be any functions of measurements made prior to randomization. We define four sets of covariates that we will adjust for using the procedure described in Section 2.2:

- W_{-ER}: {Age, Tumor Size, I(Tumor Grade = 2), I(Tumor Grade = 3)}
- W_C: {Age, Tumor Size, I(Tumor Grade = 2), I(Tumor Grade = 3), ER Status}
- W_G: {MammaPrint Risk Category}

- W_{CG}: {Age, Tumor Size, I(Tumor Grade = 2), I(Tumor Grade = 3), ER Status, MammaPrint Risk Category}

Here, I(Tumor Grade = 2) is an indicator of whether or not the patient's tumor is severity grade 2.

With these four sets of covariates, we are able to contrast gains in precision from different covariate sources. We compare adjusting for W_C versus W_{-ER} to determine how much adding the clinical covariate ER status to other clinical covariates improves precision. We also compare the prognostic value of the genomic predictor plus clinical covariates (W_{CG}) versus clinical covariates alone (W_C).

We consider the clinical covariates above because they reflect quantities that clinicians may commonly use to evaluate cancer-related risks and courses of therapy. The number of covariates we are adjusting for here exceeds the conservative approach recommended by Ref. [1]. They recommend 2-3 adjustment covariates at sample sizes such as ours. The potential downside to adjusting for greater numbers of covariates is that we risk non-negligible increases in estimator variance if our covariates turn out to be non-prognostic for the outcome, as shown in Section 3. We chose to include larger numbers of covariates here in order to compare the added value of MammaPrint above the prognostic value of the full set of relevant clinical covariates available in our data sets.

2.4. Simulations

We conducted a simulation study with the goal of comparing the variance of the unadjusted and adjusted estimators to determine how much precision we may gain from adjusting for clinical and genomic covariates. For each of the four data sets described in Section 2.1 and in the supplement, we construct a data generating distribution that mimics the observed correlation between baseline variables and outcomes.

To preserve the relationship between outcome and potentially prognostic covariates from the original data set, we resample participants with replacement and create a new sample of the size of our data set (296 for the MammaPrint validation data) for each simulated trial; we record (W,Y) for each resampled participant. This maintains the empirical joint distribution of (W,Y), preserving the correlation of these variables. In each simulated trial, the study arm assignment A of each participant is a random draw from the Bernoulli distribution with probability 1/2 of being 0 or 1, independent of (W,Y). The population average treatment effect defined in (1) corresponding to the above data generating distribution is therefore ψ=0.

The reason we do not simply resample patient data vectors (W,A,Y) with replacement from a given data set is that the resulting data generating distribution would not have treatment A independent of baseline variables W (as in a randomized trial). This is because our data sets are from observational studies, as opposed to randomized trials. Though it would be preferable to use data from randomized trials, we were not able to obtain data from any such trials that also recorded the MammaPrint predictor at baseline. Observational studies still can provide a rough approximation to the magnitude of potential precision gains from covariate adjustment, since these gains are directly related to the variance of Y explained by W [1].

For each data generating distribution described above, we construct J = 100,000 simulated trial data sets, each of sample size equal to the original data set (excluding patients with missing data). Using the j^{th} simulated data set, we compute the unadjusted estimator $\widehat{\psi}^j_{una}$ and the adjusted estimator $\widehat{\psi}^j_{adj}$ using each of the covariate sets W_{-ER},W_C,W_G,W_{CG}. We then approximate the bias and variance of each of these estimators based on its values over the 100,000 simulated trials. Since ψ=0, the bias B of an estimator $\widehat{\psi}$ is $E(\widehat{\psi}) - \psi = E(\widehat{\psi})$, which is approximated by the average of $\widehat{\psi}$ over the 100,000 simulated trials we conducted. We similarly approximate the variance of each estimator. For the unadjusted estimator, the approxmiate bias and variance based on our simulation study are denoted by $B_{una} = \frac{1}{J}\sum_{j=1}^{J}\widehat{\psi}^j_{una}$ and $\sigma^2_{una} = \frac{1}{J-1}\sum_{j=1}^{J}(\widehat{\psi}^j_{una} - B_{una})^2$, respectively. The bias and variance approximations for the adjusted estimator $\widehat{\psi}_{adj}$ are denoted similarly: $B_{adj} = \frac{1}{J}\sum_{j=1}^{J}\widehat{\psi}^j_{adj}$, $\sigma^2_{adj} = \frac{1}{J-1}\sum_{j=1}^{J}(\widehat{\psi}^j_{adj} - B_{adj})^2$. For conciseness, we refer to these approximations as the bias and variance of the corresponding estimator, rather than writing “approximate bias” and “approximate variance”.

We define the (percent) precision gain due to the adjusted estimator in comparison to the unadjusted estimator, as approximated by simulation, as $G_{adj} = \frac{\sigma^2_{una} - \sigma^2_{adj}}{\sigma^2_{una}} \times 100\%$. The precision gain equals, asymptotically (as sample size goes to infinity), the percent reduction in sample size to achieve a desired power at a local alternative comparing the adjusted versus unadjusted estimator. It equals $1 - 1/RE$, where RE is the asymptotic relative efficiency. Negative values of G_{adj} correspond to efficiency losses, which can occur if baseline variables are only weakly (or not at all) prognostic for the outcome. Asymptotically (as sample size goes to infinity), G_{adj} converges to a nonnegative value, which represents zero or positive precision gain, as proved by Rotnitzky et al. [15]; Colantuoni and Rosenblum [1].

Simulations were conducted via the BatchJobs R package [21], which allows for an interface between R and a cluster queuing system. We parallelized such that 1000 simulated data sets were constructed concurrently by each of 100 processors on a Sun Grid Engine (SGE) cluster, which sped up the computation of our approximations.

We also conducted simulation studies as above except where the sample size in each simulated trial is double that of the original data set. In all of our simulation studies, each simulated participant's data is an independent, identically distributed draw from a joint distribution P (which depends on the data set being resampled from) on (W,A,Y). Therefore, even though we are resampling (with replacement) double the sample size n from the original data set, the effective sample size is $2n$ (i.e., each estimator's variance is roughly cut in half compared to its variance at the original sample size.) To illustrate this point, consider the analogy of drawing n independent, identically distributed realizations $Y_1,\ldots,Y_n$ from a Bernoulli distribution with true probability 1/4 of being 1. Though this is equivalent to resampling n times with replacement from the four person data set {0,0,0,1} (with equal chance of each), each draw is independent and the effective sample size equals the number of draws n. The precision gains from adjustment are expected to be similar when the original sample sizes are used.

2.5. Reproducibility

Our analyses are reproducible. Code, data files, and supplementary results are available at https://github.com/leekgroup/genesigprecision.

3. Results

Table 2 presents variances for each estimator and the precision gain G_{adj}, using different sets of baseline covariates, for the MammaPrint validation data set and the data sets GSE19615, GSE11121, GSE7390. All precision gains G_{adj} are rounded to the nearest

Table 2
Precision gains due to adjustment for different sets of baseline covariates.

Covariate set	Original sample size			Double sample size		
	σ^2_{una}	σ^2_{adj}	G_{adj}	σ^2_{una}	σ^2_{adj}	G_{adj}
MammaPrint data set						
W_{-ER}	0.0018	0.0017	4%	0.00089	0.00084	6%
W_C	0.0018	0.0017	5%	0.00089	0.00083	6%
W_G	0.0018	0.0017	5%	0.00089	0.00084	5%
W_{CG}	0.0018	0.0017	6%	0.00089	0.00082	7%
GSE19615 data set						
W_{-ER}	0.0088	0.0078	11%	0.0044	0.0037	14%
W_C	0.0088	0.0073	17%	0.0044	0.0035	21%
W_G	0.0088	0.0084	4%	0.0044	0.0042	4%
W_{CG}	0.0088	0.0074	16%	0.0044	0.0035	21%
GSE11121 data set						
W_{-ER}	0.0036	0.0034	7%	0.0018	0.0016	9%
W_C	0.0036	0.0034	6%	0.0018	0.0017	9%
W_G	0.0036	0.0036	2%	0.0018	0.0018	2%
W_{CG}	0.0036	0.0034	7%	0.0018	0.0016	9%
GSE7390 data set						
W_{-ER}	0.0045	0.0045	−1%	0.0022	0.0022	1%
W_C	0.0045	0.0045	−1%	0.0022	0.0022	1%
W_G	0.0045	0.0043	4%	0.0022	0.0022	4%
W_{CG}	0.0045	0.0044	2%	0.0022	0.0021	5%

percent.

Consider the left half of Table 2, which corresponds to simulated trials having the same sample size as the corresponding data set. Adjusting only for clinical variables (W_C) provided precision gains G_{adj} of −1%, 5%, 6%, 17% (from smallest to largest), compared to the unadjusted estimator, across the four data sets. Adjusting only for the MammaPrint genomic signature (W_G) provided gains of 2%, 4%, 4%, 5%. Simultaneously adjusting for clinical variables and the genomic signature (W_{CG}) provided gains of 2%, 6%, 7%, 16%.

Each of the above precision gains G_{adj} was unchanged or slightly increased when each simulated trial has double the sample size as the corresponding data set (right half of Table 2). This is to be expected, as described above. For each estimator, covariate set, and data set, the variance at double the sample size was approximately half of the corresponding variance at the original sample size, as expected.

The additional gain due to the genomic predictor is defined as the difference between the precision gain from W_{CG} versus W_C. First, consider the left half of Table 2, where each simulated trial has the same sample size as the corresponding data set. In simulations based on the MammaPrint validation data, the genomic predictor provided an additional gain of 1% above using all clinical factors. In two of the other data sets, the additional gains due to the MammaPrint predictor were 0% (GSE11121) and 3% (GSE7390). Using a third such data set, GSE19615, adjusting for the MammaPrint prediction in addition to the clinical covariates decreased precision by 1% compared to adjustment for clinical covariates alone. Such losses in precision can occur when adjusting for a variable that is only weakly prognostic (or not prognostic) for the outcome. The additional gains due to the genomic predictor were 0%, 0%, 1%, 4% when sample sizes in the simulations were doubled (right half of Table 2).

We also examined the additional gains due to ER status, defined as the difference between the precision gains from W_C versus W_{-ER}. These values were −1%, 0%, 1%, 6%, for the four data sets, based on simulations at the original sample size. Qualitatively, these were similar to the magnitudes of additional gains due to the genomic predictor.

We conducted additional simulations where we generated baseline covariates independent of the outcome, in order to determine the magnitude of precision losses due to adjusting for pure noise. This quantifies the loss that would occur if one were to prespecify an analysis that adjusts for variables conjectured to be prognostic, but these variables turn out to be non-prognostic. We generated 100,000 simulated trial data sets as above, except where the data generating distribution has baseline variables W independent of Y. This was done by resampling W with replacement from its marginal distribution in the MammaPrint data set, and similarly resampling Y from its marginal distribution. The results are shown in Table 3. As expected, all combinations of covariates produce zero or negative precision gains, with greater losses when adjusting for larger covariate sets (due to more degrees of freedom in the working models). The maximum loss in precision is 3% when using the original sample sizes (left half of Table 3). This is due to the inclusion of greater than the recommended number of adjustment covariates, as described in Section 2.3. The potential losses are smaller if the sample size is larger, as shown in the right half of Table 3 where the maximum loss is 1%. Larger sample sizes tend to decrease the magnitude of precision losses since asymptotically (as sample size goes to infinity), G_{adj} converges to a nonnegative value, which represents zero or positive precision gain, as proved by Refs. [15]; [1]. We present additional simulation results with W generated independent of Y in the Supplementary Material where we reduce the number of clinical covariates adjusted for, resulting in smaller precision losses.

The bias approximations B_{una} and B_{adj} were both quite small, with magnitudes of at most 0.0003 over the four simulation studies. We examined the distribution of the differences between $\widehat{\psi}^j_{una}$ and $\widehat{\psi}^j_{adj}$ over the $j = 1,\ldots,100{,}000$ iterations in the simulation using the Mammaprint validation data set; the histogram of $\widehat{\psi}^j_{una} - \widehat{\psi}^j_{adj}$ appears in Fig. 1, and analogous histograms for the other datasets along with a table comparing the distributions of differences across the four simulation studies are available in the supplement. For the simulation with the MammaPrint dataset, we saw an average difference of 0.00005 (standard deviation = 0.0145). The 2.5% and 97.5% quantiles of $\widehat{\psi}^j_{una} - \widehat{\psi}^j_{adj}$ were [-0.029, 0.029]; this implies that 95% of the differences between the unadjusted and adjusted estimators had magnitudes smaller than 3%. The correlation of the two estimators was 0.94.

In general, we expect the difference between the unadjusted and adjusted treatment effect estimators to be small unless there is substantial chance imbalance between treatment and control arms that is accounted for by the adjusted estimator. In that case, we would expect the adjusted estimator to be closer to the true effect. In our setting, the adjusted estimator was closer to the true effect of zero 53% of the time, suggesting a slight improvement over the unadjusted estimator.

4. Conclusion

Appropriately adjusting for prognostic baseline covariates has potential to improve precision in estimating the average treatment effect in randomized trials. If baseline factors are collected for patients enrolled in a study, then adjusting for them can reduce the sample size necessary to obtain a desired precision in estimation of

Table 3
Precision gains under data generating distribution with W and Y independent, based on marginal distributions from MammaPrint validation data set.

Covar. Set	Original sample size			Double sample size		
	σ^2_{una}	σ^2_{adj}	G_{adj}	σ^2_{una}	σ^2_{adj}	G_{adj}
W_{-ER}	0.00177	0.00181	−2%	0.00090	0.00091	−1%
W_C	0.00177	0.00182	−2%	0.00090	0.00091	−1%
W_G	0.00177	0.00178	0%	0.00090	0.00090	0%
W_{CG}	0.00177	0.00183	−3%	0.00090	0.00091	−1%

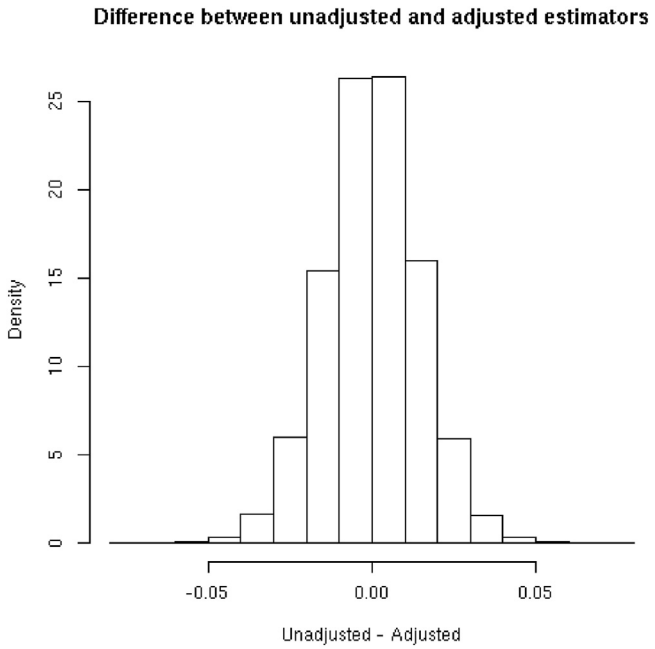

Fig. 1. Histogram of $\widehat{\psi}^j_{una} - \widehat{\psi}^j_{adj}$. The histogram of differences between the unadjusted and adjusted estimators is approximately normal and is centered close to the true effect of zero (mean = 0.00005, standard deviation = 0.0145). The adjusted estimator is closer than the unadjusted estimator to the true effect approximately 53% of the time. For this histogram, we considered the adjusted estimator using all available baseline covariates (clinical + genomic).

the average treatment effect and, therefore, the cost to run the trial.

The precision gains from adjusting for clinical variables were substantial (5%, 6%, 17%) in simulation studies based on three out of four data sets we considered; the last data set led to a loss in precision of 1%. These precision gains slightly increased when sample sizes were doubled, showing that covariate adjustment can be valuable both at moderate (115–307) and larger sample sizes, in the context of breast cancer treatment trials.

The additional gains from adjusting for the genomic predictor were quite small. We consider several possible explanations for this finding. First, our estimator may not have effectively extracted the additional prognostic information in the genomic marker; e.g., it may be that including interactions between the MammaPrint score and clinical variables, or using a less parametric model than logistic regression (e.g., splines), would lead to an adjusted estimator with better precision than we observed. This is difficult to evaluate, since using more flexible models could lead to overfit; this may be controllable via penalization or cross-validation, and is an area of future research. Another possible explanation is that the MammaPrint risk score is too coarse a summary measure of the 70 gene expression levels measured by the MammaPrint assay, for our purpose. The MammaPrint risk score was not designed for maximizing additional prognostic value beyond what is explained by clinical variables. It may be that a different function of the 70 gene expression levels would lead to greater precision gains, but this is beyond the scope of this paper. A third possible reason for the lackluster additional gains from the genomic predictor is that there may be little additional prognostic value in the genomic information for the outcome we considered. The MammaPrint score in the validation set examined here was 89% sensitive to high risk-of-recurrence patients, 42% specific to low risk-of-recurrence [18], but these measures (i.e., sensitivity and specificity) focus only on the MammaPrint score and do not separate out the variation that can be explained by clinical variables.

The additional gain due to the genomic predictor was roughly similar to the additional gain from including ER status over other clinical covariates. ER status may lack prognostic power if ER positive participants are treated with adjuvant tamoxifen [22]. Similarly, it is possible that the MammaPrint score influenced treatment decisions, which could lead to decreased prognostic value.

A limitation of our approach is that we used data from observational studies, rather than from randomized trials. If the prognostic value of baseline variables is similar in a randomized trial setting, then our results may shed light on the order of magnitude of precision gains that can be achieved from covariate adjustment. However, if the prognostic value of baseline variables for the outcome is systematically different in a randomized trial, then our results would not apply. Future work involves applying our simulation approach to randomized trial data sets. Another limitation of our approach is that we ignore censoring due to loss to follow up. It is possible to incorporate censoring into our estimator, under a missing at random assumption, but this is an area for future work.

Our focus was on the prognostic value of different variables, that is, the ability of these variables to explain variation in the outcome (5 year recurrence). In contrast, the more ambitious goal of personalized medicine is to find predictive variables, i.e., variables that discriminate between those who are likely to benefit from a specific treatment or not. Being prognostic is not a prerequisite for being predictive, e.g., as in the case of ER status. However, the MammaPrint score having little prognostic value beyond the variation explained by clinical covariates indicates that its utility for covariate adjustment is limited.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This research was supported by National Institutes of Health grant R01GM105705. This publication's contents are solely the responsibility of the authors and do not necessarily represent the official views of the above agency.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.conctc.2016.03.001.

References

[1] E. Colantuoni, M. Rosenblum, Leveraging prognostic baseline variables to gain precision in randomized trials, Stat. Med. 34 (18) (2015) 2602–2617.
[2] F.S. Collins, H. Varmus, A new initiative on precision medicine, N. Engl. J. Med. 372 (9) (2015) 793–795.
[3] W. Burke, D.M. Korngiebel, M. Cho, Closing the gap between knowledge and clinical application: challenges for genomic translation, PLoS Genet. 11 (2) (2015) e1004978–e1004978.
[4] K.A. Baggerly, K.R. Coombes, Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology, Ann. Appl. Stat. (2009) 1309–1334.
[5] B. Letham, C. Rudin, T.H. McCormick, D. Madigan, Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model, Ann. Appl. Stat. 9 (3) (2015) 1350–1371.
[6] R. Arnaout, T.P. Buck, P. Roulette, V.P. Sukhatme, Predicting the cost and pace of pharmacogenomic advances: an evidence-based study, Clin. Chem. 59 (4) (2013) 649–657.
[7] M.J. Van De Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, et al., A gene-expression signature as a predictor of survival in breast cancer, N. Engl. J. Med. 347 (25) (2002) 1999–2009.
[8] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F.L. Baehner, M.G. Walker, D. Watson, T. Park, et al., A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, N. Engl. J. Med. 351 (27)

(2004) 2817–2826.
[9] J.S. Parker, M. Mullins, M.C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J.F. Quackenbush, I.J. Stijleman, J. Palazzo, J.S. Marron, A.B. Nobel, E. Mardis, T.O. Nielsen, M.J. Ellis, C.M. Perou, P.S. Bernard, Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (8) (2009) 1160–1167.
[10] R.M. Connolly, Omics as useful tools in clinical practice: are we there yet? Oncology 27 (3) (2013).
[11] A. Barker, C. Sigman, G. Kelloff, N. Hylton, D. Berry, L. Esserman, I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy, Clin. Pharmacol. Ther. 86 (1) (2009) 97–100.
[12] L. Yang, A. Tsiatis, Efficiency study of estimators for a treatment effect in a pretest-posttest trial, Am. Stat.s 55 (2001) 314–321.
[13] W. Cao, A. Tsiatis, M. Davidian, Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data, Biometrika 96 (2009) 723–734.
[14] Z. Tan, Bounded, efficient and doubly robust estimating equations for marginal and nested structural models, Biometrika 97 (2010) 661–682.
[15] A. Rotnitzky, Q. Lei, M. Sued, J.M. Robins, Improved double-robust estimation in missing data and causal inference models, Biometrika 99 (2) (2012) 439–456.
[16] S. Gruber, M. van der Laan, Targeted minimum loss based estimator that outperforms a given estimator, Int. J. Biostat. 8 (1) (2012). Article 11.
[17] M. Buyse, S.M. Loi, L. Van't Veer, G. Viale, M. Delorenzi, A.M. Glas, M. Saghatchian d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. Harris, J. Bogaerts, P. Therasse, A. Floore, M. Amakrane, F. Piette, E. Rutgers, C. Sotiriou, F. Cardoso, M.J. Piccart, Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer, J. Natl. Cancer Inst. 98 (17) (2006) 1183–1192.
[18] L. Marchionni, B. Afsari, D. Geman, J.T. Leek, A simple and reproducible breast cancer prognostic test, BMC Genomics 14 (1) (2013) 336.
[19] C. Jennison, B. Turnbull, Group Sequential Methods with Applications to Clinical Trials, Chapman and Hall/CRC Press, Boca Raton, FL, 1999.
[20] J. Robins, M. Sued, Q. Lei-Gomez, A. Rotnizky, Comment: performance of double-robust estimators when inverse probability weights are highly variable, Stat. Sci. 22 (4) (2007) 544–559.
[21] B. Bischl, M. Lang, O. Mersmann, J. Rahnenfuehrer, C. Weihs, Computing on High Performance Clusters with R: Packages BatchJobs and BatchExperiments, Technical Report 1, TU Dortmund, 2011.
[22] M. Cianfrocca, L.J. Goldstein, Prognostic and predictive factors in early-stage breast cancer, Oncologist 9 (6) (2004) 606–616.