Title
A three-groups model for high-throughput survival screens

Journal
Biometrics, 72(3)

ISSN
0006-341X

Authors
Shaby, Benjamin A
Skibinski, Gaia
Ando, Michael
et al.

Publication Date
2016-09-01

DOI
10.1111/biom.12479

Peer reviewed

HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2016 September 07.

Published in final edited form as:

Biometrics. 2016 September ; 72(3): 936–944. doi:10.1111/biom.12479.

A Three-groups Model for High Throughput Survival Screens

Benjamin A. Shaby[1,*], **Gaia Skibinski**[2], **Michael Ando**[2], **Eva S. LaDow**[3], and **Steven Finkbeiner**[2]

[1]Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A

[2]Gladstone Institute of Neurological Disease, J. David Gladstone Institutes, San Francisco, California 94158, U.S.A.

[3]Department of Neuroscience, Univeristy of Texas at Dallas, Richardson, Texas 75080, U.S.A.

Summary

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative condition characterized by the progressive deterioration of motor neurons in the cortex and spinal cord. Using an automated robotic microscope platform that enables the longitudinal tracking of thousands of single neurons, we examine the effects a large library of compounds on modulating the survival of primary neurons expressing a mutation known to cause ALS. The goal of our analysis is to identify the few potentially beneficial compounds among the many assayed, the vast majority of which do not extend neuronal survival. This resembles the large-scale simultaneous inference scenario familiar from microarray analysis, but transferred to the survival analysis setting due to the novel experimental setup. We apply a three component mixture model to censored survival times of thousands of individual neurons subjected to hundreds of different compounds. The shrinkage induced by our model significantly improves performance in simulations relative to performing treatment-wise survival analysis and subsequent multiple testing adjustment. Our analysis identified compounds that provide insight into potential novel therapeutic strategies for ALS.

Keywords

Bayesian; High throughput data; Mixture model; Multiple testing; Shrinkage; Survival analysis

1. Introduction

The neurodegenerative condition amyotrophic lateral sclerosis (ALS) is characterized by the progressive deterioration of motor neurons in the cortex and spinal cord that leads to muscle atrophy and respiratory failure (Shook and Pioro, 2009). The need for an effective treatment is urgent, as very few treatments are currently available. One medication, Riluzole, a glutamate antagonist, slows deterioration by approximately two months (Miller et al., 2007),

* bshaby@psu.edu.

6. Supplementary Materials

The R and JAGS code used to run all simulations is available and may be accessed at the *Biometrics* website on Wiley Online Library. The JAGS code used for the data analysis is identical to that of the simulations.

however no treatment stops the course of this disease that leaves patients with a median survival of approximately 3–4 years from onset of symptoms (Bäumer et al., 2014). In most cases of ALS, the RNA-binding protein TDP43 accumulates within the cytoplasm of neurons and glia (Arai et al., 2006; Neumann et al., 2006). Mutations in the gene that codes for TDP43 cause familial ALS and lead to changes in TDP43 localization and a reduction in neuronal health (Barmada et al., 2010).

To identify novel compounds that could mitigate neuron death caused by mutant TDP43, a library of FDA-approved compounds was screened in a neuronal model of mutant TDP43 that recapitulates ALS pathology. Primary rodent neurons were cultured and induced to degenerate by the introduction and ectopic expression of TDP43 carrying the ALS-associated mutation M337V. Using a robotic microscope platform invented by the Finkbeiner laboratory (Arrasate et al., 2004) thousands of individual neurons expressing mutant TDP43 were imaged at regular intervals, enabling them to be tracked over their lifetimes in a high-throughput and fully automated manner. In-house image processing software analyzed the images taken by the robotic microscope to determine, among other things, in which interval death occurred for each individual neuron. In comparison to conventional single snapshot approaches, the longitudinal tracking of individual neurons is substantially more sensitive at identifying the effects of disease associated phenotypes or small molecule therapeutics.

We wish to use the variation in survival times between groups of cells exposed to different compounds to quantify the role of each compound on modulating the survival of primary neurons expressing mutant TDP43. The key feature of the problem is that we wish to do simultaneous inference on hundreds or thousands of different treatments. The structure of the data is similar to the canonical genomics setup, where one tries to detect meaningful differences between experimental groups in some quantity associated with individual genes, with the assumption that only a few from among thousands of candidate genes manifest real changes.

The naive way to proceed would be to test each treatment against a control using, say, Cox proportional hazards regression, and declare treatments significantly different from the control using a standard p-value cutoff. A widely recognized problem with this approach is that it results in an enormous multiple comparison issue. The strategy for dealing with large numbers of comparisons that has become fairly standard, thanks to highly active microarray and other “omics” research, is to compute many test statistics, and adjust the resultant p-values by controlling something like false discovery rate (Benjamini and Hochberg, 1995).

Post hoc adjustment of p-values has proven highly successful. However, recent studies have demonstrated that performing shrinkage when computing test statistics, i.e. before calculating p-values, can improve performance substantially. Various shrinkage strategies have been proposed in the statistical genomics context. Many employ James-Stein type forms or random effects to pool information in the mean component of the test statistic, the standard error component of the test statistic, or both (see Bar et al., 2010, for a taxonomy of such approaches).

Like Bar et al. (2010), we take a completely model-based approach that uses random effects to induce shrinkage, expressing the distribution of the response as a mixture over treatment-specific indicators. Inference is then based on the posterior probability of the indicators, thereby bypassing traditional hypothesis testing completely. Our setup is slightly different from the classical genomics setting considered in Bar et al. (2010) in that for us, the treatments take on the role traditionally played by genes. That is, each inference consists of determining whether one treatment shows a significant effect relative to a control, which is analogous to asking whether one gene shows a significant effect between treatments. In addition, Bar et al. (2010) consider normal responses that are measured without censoring, while our data, as is common in survival analysis, is non-normal and features both interval and right censoring. Hence, while our modeling approach inherits the key features found in Bar et al. (2010), the key differences described above in the structure of the data necessitate extensions to the Bar et al. (2010) model.

2. The Core Model

Conceptually, our model assumes that the treatments come from three distinct populations: those with no effect on survival (the null group), those that have a positive effect on survival (the beneficial group), and those that have a negative effect on survival (the deleterious group). Although the sharp divisions of the three-group structure may seem contrived, they need not literally hold for the shrinkage induced by the modeling conceit to prove effective (Efron, 2008).

Our three-groups scheme is generalization of the two-groups concept of Bar et al. (2010), which assumes that the positive and negative non-null groups are symmetric in their deviation from the null group. However, we expect this symmetry not to hold in our survival screens. Essentially, it is much easier to kill a cell than it is to prolong its life. Compounds that are detrimental can easily shorten the life of the neurons drastically, some of them killing almost instantly, but drugs that are beneficial, which we expect to be far fewer, cannot be expected to extend life very much. Put another way, a perfect drug would extend life to that of healthy cells, and the lifetimes of healthy cells are closer to those of un-treated diseased cells than they are to those killed instantly. Furthermore, there is no reason to expect the variance of the lifetimes of the cells treated with toxic compounds to be the same as the variance of the lifetimes of the cells treated with beneficial compounds. It is therefore important that the beneficial and deleterious groups not be considered mirror images of each other, as they are in a two-groups model.

Since our response variable is a collection of survival times, we may choose from a wealth of existing survival analysis models in which to embed our three-groups structure. While it is not conceptually difficult to build upon flexible nonparametric Bayesian models (e.g. Kottas, 2006) or Bayesian flavors of the classic proportional hazards model (Kalbfleisch, 1978; Sinha et al., 1999), for simplicity and ease of implementation we work with parametric accelerated failure time (AFT) models. This class of models considers failure time distributions of the form $\log Y = \mu + \sigma W$, where Y denotes time of death and W has some parametric form (see Kalbfleisch and Prentice, 2002, Chapter 2, e.g.). Different choices of W lead to different survival time distributions, but regardless of the specific

distribution of W, the location parameter μ and the scale parameter σ are convenient platforms for hierarchical model building. We will denote the generic class of distributions of log Y induced by the AFT form as $f_{\text{AFT}}(\mu, \sigma)$.

The core component of the three-groups AFT model is the following. Let N be the total number of individual cells, and M be the number of treatments. The null, deleterious, and beneficial groups are labeled, respectively 1, 2, and 3. Let y_{ij} be the survival time of the ith individual cell, which was exposed to the jth treatment. Then

$$\begin{aligned} log \quad y_{ij} &\sim f_{AFT}\left(\mu_{ij}, \sigma_j\right) \\ \mu_{ij} &= \omega_{G_j} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \delta_{\mathbf{j}} \\ G_j &\sim \text{Categorical}\left(\mathbf{p}\right) \end{aligned} \quad (1)$$

where $\mathbf{p} = (p_1, p_2, p_3)^{\mathrm{T}}$. The categorical variable G_j, $j = 1, \ldots, M$, takes a value of either 1, 2, or 3, and indicates membership of treatment j in either the null, deleterious, or beneficial group, respectively. It is the posterior distribution of the category assignments $G_1, \ldots, G_M$ that will ultimately be of primary interest.

The location parameter μ_{ij} corresponding to cell i (which was exposed to treatment j) has three components: a group mean ω_{Gj}, which is constant across treatments with the same group assignment; a cell-specific covariate effect $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$; and a treatment-specific offset δ_j, which represents the systematic deviation of treatment j from the mean of the group to which it belongs. Each treatment j is assigned its own σ_j, rather than assuming a single shared σ, because assuming this form of homogeneity (i.e. that $\sigma_j \equiv \sigma$) is known to result in severe performance degradation when it does not hold (Bar et al., 2010). Our indexing scheme is unorthodox (each i is paired with only one j, so that the total number of calls is N rather than the more typical NM) to accommodate the lack of balance in the ALS data. The number of individual cells is not constant across treatments because the robotic microscope tracks as many cells as it can find in each experimental well, and that number varies from well to well.

As is typical of survival data, we need to account for right censoring that occurs when cells live beyond the monitoring period. In addition, because the robot revisits each cell at discrete time intervals rather than monitoring all cells continuously, each time of death is only known up to an interval that is fairly wide relative to the duration of the experiment. As a result, all data that is not right censored is interval censored. Both right and interval censoring are easily handled through straightforward modification of the data likelihood.

Continuing with the model, we parametrize the overall group means as

$$\begin{aligned} \omega_1 &= \mu_0 \\ \omega_2 &= \mu_0 + \psi_1 \\ \omega_3 &= \mu_0 + \psi_2, \end{aligned}$$

with prior distributions

$$\begin{aligned}\mu_0 &\sim \mathrm{N}\left(0, 10^6\right)\\ \psi_1 &\sim TN^-\left(\sigma^2_{\psi_1}\right)\\ \psi_2 &\sim TN^+\left(\sigma^2_{\psi_2}\right),\end{aligned}$$

where the notation $TN^+\left(\sigma^2_{\psi_1}\right)$ denotes a mean-zero normal distribution with variance $\sigma^2_{\psi_1}$, truncated to have only positive support, and analogously for $TN^-\left(\sigma^2_{\psi_2}\right)$. This parametrization is minimally informative about ω_1, ω_2, and ω_3, other than enforcing the ordering $\omega_2 < \omega_1 < \omega_3$ that characterizes the three groups as deleterious, null, and beneficial, respectively.

Sharing of information about the treatment-level scale parameters of the survival time distributions is accomplished by assigning them a common inverse-Gamma prior distribution whose parameters are assigned half Cauchy hyper-priors,

$$\begin{aligned}\sigma^2_j &\sim \text{Inverse- Gamma}\left(\alpha_{\sigma^2}, \beta_{\sigma^2}\right), \quad j=1,\ldots,M\\ \alpha_{\sigma^2}, \beta_{\sigma^2} &\sim \text{TCauchy}^+\left(1\right),\end{aligned}$$

where TCauchy$^+$(1) denotes the positive half of a standard Cauchy distribution. In this way, the scale parameters $\sigma_1, \ldots, \sigma_M$ are shrunken towards each other but remain inhomogeneous.

The model is completed with the vague conjugate prior distributions

$$\begin{aligned}\mathbf{p} &\sim \text{Dirichlet}\left(\kappa\boldsymbol{\alpha}\right),\\ \delta_j &\sim \mathrm{N}\left(0, \sigma^2_\delta\right), \quad j=1,\ldots,M\\ \sigma^2_{\psi_1}, \sigma^2_{\psi_1} &\sim \text{Inverse- Gamma}\left(0.01, 0.01\right),\end{aligned}$$

for some choice of hyperparameter $\boldsymbol{\alpha}$ with $\sum_{k=1}^{3}\alpha_k=1$, so that $\boldsymbol{\alpha}$ reflects prior beliefs about the proportion of null, deleterious, and beneficial treatments in the screen, and the concentration parameter $\boldsymbol{\kappa}$ reflects confidence in that belief.

2.1 Details of the ALS Experiment

Additional model components are necessary to accommodate details of the survival screen that are particular to the ALS study, which we now describe in depth. To model ALS in cells, rat cortical neurons were obtained from rat embryos at 20–21 days gestation. The cells were grown in plates with 96 independent wells at a density of approximately 100,000 cells per well. After 4 days neurons were transfected to enable them to ectopically express either an inert GFP protein (via a pGW1-GFP plasmid) or the ALS causing protein, GFP-TPD43-M337V (via a pGW1-GFP-TDP43 M337V plasmid). In order to visualize the cells by immunofluorescence, a red fluorescent protein was also ectopically expressed in the same cells (via a pGW1-mApple plasmid).

Each 96-well plate contained a set of positive and negative controls. Negative controls were composed of 8 wells of neurons expressing the inert GFP protein, and positive controls were composed of 8 wells of neurons expressing the ALS causing GFP-TDP43-M337V protein. The negative controls were not analyzed, but the positive controls were extremely valuable and informative because they consist of a large number of replications for which the group classification (i.e. 1, the null group) is known a priori.

The experiment considered 8 distinct plate design configurations, each replicated 4 times, for a total of 32 plates. Each of the 8 designs contains 96 wells, 16 of which are, again, controls, and the remaining 80 of which are treated with one compound each. Thus the number of treatments under consideration is 640. The average number of cells per treatment is around 265, making the total number of neuron survival times analyzed just short of 17,000.

A potential covariate that we considered including in the model is the initial concentration of TDP43, measured at the outset of the experiment, in each neuron. TDP43 is known to be toxic to neurons and clearly influences longevity (Barmada et al., 2010). However, the expression of TDP43 is highly confounded with any potential treatment effects because one way a drug could potentially improve survival is by lowering a cell's propensity to express TDP43, or conversely, a compound could shorten lifetimes by increasing the prevalence of TDP43. We therefore decided to omit measured concentration of TDP43 as a covariate, and in the end used no covariates at all. In future studies, TDP43 levels could be used to investigate how the drugs reduce TPD43 toxicity.

2.2 Additional Model Components

To accommodate between-plate variation described in Section 2.1, a plate-level random effect is defined as

$$\gamma_k \sim \mathrm{N}\left(0, \sigma_\gamma^2\right), \quad \text{subject to} \sum_{k=1}^{32} \gamma_k = 0. \tag{2}$$

where the sum-to-zero constraint is imposed to improve mixing. Furthermore, because the 8 treatment designs were prepared on different days using potentially different biological material, an additional random effect

$$\gamma_\ell^{\mathrm{D}} \sim \mathrm{N}\left(0, \sigma_{\gamma^{\mathrm{D}}}^2\right), \quad \text{for} \quad \ell = 1, \ldots 8, \tag{3}$$

is needed to model between-treatment design variability. The random effects defined by (2) and (3) are integrated into the model (1) for the f_{AFT} location parameter as

$$\mu_{ijk\ell} = \omega_{G_j} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \delta_{\mathbf{j}} + \gamma_{\mathbf{k}} + \gamma_\ell^{\mathrm{D}}.$$

The model was fit using MCMC, implemented with the JAGS package through its R interface (Plummer, 2014). To improve mixing, a hierarchically-centered equivalent parametrization (Gelfand et al., 1995) was used. A new parameter vector $\boldsymbol{\eta}$ was defined as $\eta_j = \omega_{G_j} + \delta_j \sim \mathrm{N}\left(\omega_{G_j}, \sigma^2_\delta\right)$, for $j = 1, \ldots, M$, so that sampling was performed on $\boldsymbol{\eta}$, which is "centered" at $\boldsymbol{\omega}$, instead of the vector $\boldsymbol{\delta}$.

3. Simulation

To test the performance of the three-groups random effects model relative to treatment-specific analysis followed by multiple testing adjustment, we conducted a simulation study. Our simulation design closely mirrors that of Bar et al. (2010). For each simulation, we drew 200 datasets of size $N = 5,000$, $M = 400$ from the three-groups model with a single covariate drawn from $N(0, 1)$, with f_{AFT} chosen so that the survival times were Weibull. Random effects matching those described in Section 2.2 were added to the log survival times. The data was then right censored at the 0.85 empirical quantile, with the remaining data divided into seven equally-spaced bins and interval censored accordingly. This censoring setup is fairly extreme and closely mimics that of the data generated by the automatic microscope. Just as in Bar et al. (2010), we varied model parameters to study classification performance under different scenarios. We considered a 2 × 2 design where we simulated a high and low proportion of null treatments (90% and 75%) and a high and low degree of inhomogeneity of the variance parameters $\sigma^2_1, \ldots, \sigma^2_M$ (drawn from Inverse-Gamma(6, 4) and Inverse-Gamma(11, 8)). The group mean vector for the null, deleterious, and beneficial groups was $\boldsymbol{\omega} = (0.0, -1.0, 1.0)^{\mathrm{T}}$. The covariate effect β was set at −0.5.

For each simulated dataset, we fit the three-groups model with a Weibull response and computed the posterior distributions of the group identification variables $G_1, \ldots, G_M$, as well as the fraction of null, deleterious, and beneficial treatments **p**. Posterior quantities were computed from 25,000 MCMC iterations (after discarding 10,000 as burn-in), a sample size that was determined from examination of trace plots to be well beyond what was necessary to achieve good convergence. To test sensitivity to the choice of f_{AFT}, we also fit the three-groups model with a lognormal response to the Weibull data. In addition, for each treatment in each dataset, we fit a suite of survival models to each treatment, with a covariate to indicate treatment versus control. The estimated coefficient from each treatment indicator was divided by its associated estimated standard error, and the set of resultant z-scores was run through the locfdr function (Efron et al., 2011). Local fdr can be interpreted as a posterior probability of being null, given a z-score, so a comparison with our posterior probabilities is natural. Since local fdr is based on a two-groups model, where positive and negative non-null treatments are lumped together, we compared local fdr scores to our posterior probabilities of $G_j = 1$ for each $j = 1, \ldots, M$.

The suite of survival models we fit consisted of a parametric Weibull survival model, using the R function survreg (Therneau, 2013), which included the random effect terms that were used to generate the data; a Cox proportional hazards model with interval censoring, using the R package intcox (Henschel and Mansmann, 2013), which does not support random effects; and a Cox proportional hazards model with random effects using the R package

coxme (Therneau, 2015), which does not support interval censoring and for this reason was run using survival times as the midpoint of the censoring intervals.

The treatment-wise parametric Weibull model with random effects is a perfect model setup in the sense that for each treatment, the model that is fit is the same as the data-generating model. The two Cox models have the ability to fit data-generating model because the Weibull is a special case of the proportional hazards model, although both versions of the proportional hazards model considered here are handicapped because neither can simultaneously accommodate interval censoring and random effects. In contrast, the three-groups lognormal model cannot fit the data-generating model, so its success would reflect robustness to the choice f_{AFT}.

The classification performance for the three-groups model and local fdr were evaluated using several criteria. First, for each simulated dataset, we computed the relative classification accuracy between the three groups Weibull model and each of the competing models. Classification requires specification of a threshold below which a treatment is declared null versus non-null, and classification accuracy depends on choice of this threshold. Relative classification accuracy is defined for dataset k and posterior probability threshold p_t as $(\text{TP}_{\text{3-group}} + \text{TN}_{\text{3-group}})/(\text{TP}_{\text{competitor}} + \text{TP}_{\text{competitor}})$, where TP_C is the number of "true positives" (i.e., the number of correctly classified non-null treatments) under classifier C, and TN is the number of "true negatives" (i.e., the number of correctly classified null treatments) under classifier C.

Results for the high and low degree of inhomogeneity simulations were similar, so to avoid clutter we show only those from the high-inhomogeneity simulations. Figures 1(a) and 1(b) show accuracies, relative to the data-generating three-groups model, of a three-groups model with a mis-specified data likelihood f_{AFT}, as well as several different treatment-wise survival models followed by applying local fdr. Mis-specifying f_{AFT} seems to have little effect on accuracy. Compared to treatment-wise methods, the three-groups model resulted in gains in accuracy that were larger in the simulation with the smaller proportion of null treatments, although the improvements seen in the higher proportion simulation were still significant at smaller values of the threshold p_t. The exception was the Cox mixed effect model in the simulation with the greater proportion of null treatments, which had similar accuracy to the three-groups model.

We also computed the false discovery rate (FDR) for each simulated dataset as function of threshold p_t. FDR is defined as $\text{FP}_C/(\text{TP}_C + \text{FP}_C)$, where FP_C denotes the number of "false positives" (i.e., the number of null treatments that were declared non-null) for classifier C. Figures 2(a) and 2(b) show false discovery rates of competing models (dashed lines, hatched confidence regions), each plotted on the same axes as the FDR from the correctly specified three-groups model (solid lines, gray confidence regions). The three groups model resulted in FDRs that were broadly similar to those of the competing models. The Cox mixed effects and the parametric Weibull models both yielded lower FDR in the simulation with the lower proportion of null treatments, but this effect disappeared in the simulation with the higher proportion of null treatments. For both simulations FDR for the three-groups model was

acceptably low for low to moderate probability thresholds, which are commonly used in practice, where the gains in accuracy (Figures 1(a) and 1(b)) were appreciable.

A more systematic way to evaluate the performance of probabilistic classifications is to use proper scoring rules (Gneiting and Raftery, 2007). The two proper scoring rules we report are the Brier score, defined as

$$-\frac{1}{M}\sum_{i=1}^{M}\left[1_{\{G_i^0=1\}} - P\left(G_i{=}1|\mathbf{y}\right)\right]^2,$$

and the logarithmic (or log) score, defined as

$$\frac{1}{M}\sum_{i=1}^{M} log\left[P\left(G_i{=}G_i^0|\mathbf{y}\right)\right],$$

where G_i^0 is the true group classification of treatment i. Neither scoring rule depends on a chosen threshold p_t, but rather both more heavily reward correct classifications the more confident they are more severely penalize incorrect classifications the more confident they are. In both cases, higher scores are better.

Figure 3 compares Brier scores and log scores of the three-groups model and local fdr applied to treatment-wise z-scores. Each boxplot depicts scores for 200 simulated datasets. In panels 3(a) and 3(b), the left-hand set of boxplots show results for the simulation with the higher proportion of non-null treatments, and the right-hand set shows results for the simulation with the lower proportion. The pattern is the same for the Brier and log scores, as well as across simulations. In all cases, the three-groups model noticeably outperformed the procedure of applying local fdr after fitting M individual survival models, and was similar to the three-groups model with the mis-specified f_{AFT}. The discrepancy is more pronounced for the log than the Brier score, and for the simulation with the smaller than the larger proportion of null treatments. The shrinkage induced by the random effects seems to be effective at improving discrimination between null and non-null treatments. Again, the best of the treatment-wise survival models was the Cox mixed effects model. It is rather puzzling that the parametric Weibull mixed effect model seemed to be the weakest model, whether evaluated using relative accuracy or proper scoring rules. Since the Weibull model is the data-generating model, we expected it to perform better. We speculate that the poor performance is due to a problematic fitting routine used in the implementation in survival package.

4. Results

Exploratory analysis suggested that a Weibull AFT model was appropriate for the ALS dataset, so f_{AFT} was specified to yield a Weibull response for the survival times. The hyper-parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\alpha}$ were specified as 2.0 and $(0.5, 0.25, 0.25)^{\text{T}}$. We also ran the model with $\boldsymbol{\kappa} = 1$ and $\boldsymbol{\kappa} = 4$, which had no noticeable impact on the results. After discarding 10,000 MCMC iterations as burn-in, an additional 25,000 samples were retained for inference.

Compute time was approximately 24 hours on a 20-core 2.5GHz Xeon node. As expected, the vast majority of the compounds under consideration were classified as null with probability indistinguishable from one. Also as expected, far more compounds were classified as deleterious with high probability than beneficial; it is much easier to shorten the life of a cell than it is to prolong it.

To assess model fit, we use a modification of the posterior predictive checks of Gelman et al. (1996). The Gelman et al. (1996) procedure is to draw many synthetic datasets from the posterior predictive distribution. Then, if summary statistics calculated from the observed dataset fall within a reasonable range of the same summary statistics calculated from the synthetic datasets, the model is deemed to fit the data. For us, because the censored likelihood necessarily leads to posterior predictive draws that fall within the observed censoring intervals, any statistics calculated from the observed data will exactly match those of the (censored) synthetic data, making any comparisons trivial. Therefore, we instead draw synthetic datasets of y_{ij}, $j = 1, \ldots, M$, from Weibull distributions that are independent conditionally on posterior draws of μ_{ij} and σ_j. The result is a stylized version of a set of draws from the posterior predictive distribution, but ignoring the censoring. We simulate 1,000 such datasets, censor them, and compare them to the observed data according to three summary statistics of the left endpoints of the resulting intervals. The three statistics we consider are the overall mean survival time, the between plate sum of squares, and the between biological sample sum of squares. The first is meant to capture the marginal fit, and the sums of squares are meant to capture the fit of the random effects.

Table 1 displays the results of the model assessment. The two sum of squared statistics from the observed data both fall within the 0.025 and 0.0975 quantiles of the same statistics computed from the simulated datasets, indicating good fit. The overall mean statistic from the observed data falls slightly outside the 95% interval of simulated overall means, possibly indicating some lack of fit.

For the handful of compounds classified by the model as being in the beneficial group with non-negligible probability, we went back and examined the raw images taken by the robotic microscope. Several of them displayed visual artifacts that caused the image analysis algorithm to spuriously indicate that neurons were alive when in fact they had died and left brightly-colored clumps of debris. After removing the spurious hits, we were left with the compounds shown in Table 2.

For comparison, we also fit Weibull and Cox mixed effects models with frailties of the form (2) and (3) for each compound (where the entire suite of negative controls was included for identification of the coefficients in the frailties). We then ran the resulting z-scores through the local fdr procedure. The local fdr scores for the compounds detected by the three-groups model are also shown in Table 2. Of the compounds that the three-groups model estimated to have non-negligible probability of being beneficial, only Dextromethorphan would likely have been flagged using the Weibull model with local fdr, and only Leflunomide would have been flagged using the Cox mixed effects model with local fdr.

To further explore the differences between the results from the three-groups model and those from applying local fdr to treatment-wise tests, we plot posterior probabilities of being in the null group against local fdr scores. Figure 4 shows the three-groups model posterior probability of being null on the x-axis and the local fdr score on the y-axis, for the Weibull (Panel (a)) and Cox mixed effects (Panel (b)) treatment-wise models. The dotted lines represent a hypothetical, arbitrary, cutoff at a value of 0.5. In each panel, the "x" symbols denote treatments that both models declare as non-null (there are many more than those listed in Table 2 because the plot includes deleterious, as well as beneficial, non-null treatments). The squares are those treatments that the three-groups model declares non-null but local fdr declares null, and the triangles are those that local fdr declares non-null but that the three-groups model declares null. Figure 4(a) shows some degree of of agreement between the two sets of inference, but also a noticeable degree of disagreement. Figure 4(b) shows little agreement.

The posterior probabilities shown in Table 2 are not impressively large. However, given that in simulations the model is most accurate when using a moderate probability cutoff, and given the very small number of positive hits, in this context we are willing to trade higher power in exchange for accepting a higher false discovery rate. Thus, all four compounds shown in Table 2 are considered candidates for followup study.

The compound with the highest posterior probability of being in the beneficial group, dextromethorphan, is particularly interesting. Dextromethorphan is the major metabolite of the cough suppressant drug dextrorphan. More importantly in our context, it is known to act as an *N*-Methyl-D-aspartate (NMDA) receptor antagonist. The effect of NMDA antagonists is to reduce glutamate accumulation, a known mechanism for neuron toxicity in neurodegeneration (Lipton and Rosenberg, 1994).

The list of hits also includes the breast cancer drug Formestane, a steroidal aromatase inhibitor. It blocks the synthesis of estrogen, starving the growth of estrogen receptor-positive cancer cells (Winer et al., 2005). The other two hits are prednisone, the common synthetic corticosteroid, and leflunomide, which is a pyrimidine synthesis inhibitor that acts an immunosuppresive. The latter two compounds have an interesting connection in that they play a role in attenuating immune responses. One plausible hypothesis about their action in the ALS screen is that they act through a non cell autonomous role on the glia in the culture, rather than directly on the neurons. Dextromethorphan, Prednisone, and leflunomide are being followed up on in an independent dose-response experiment. If they show a dose-dependent reduction in TDP43-M337V induced neuron cell death, they could be considered candidates for potential for therapeutic utility.

5. Discussion

The three-groups modeling conceit induces sharing of information that is beneficial in large-scale simultaneous inference settings. In the high throughput survival analysis setting, our extension of the idea of Bar et al. (2010) to explicitly model the null treatments as coming from a distinct population seems to result in improved performance relative to the more standard practice of computing treatment-wise *z*-scores and subsequently applying a

multiple testing procedure. We have focused on the AFT family of models, but the three-groups structure is conceptually simple to incorporate into alternative survival models if the AFT family should prove unsuitable. The drawback of our fully Bayesian approach is that MCMC computations are expensive to run. However, putting the compute time in context, a screen like the ALS experiment that we describe takes many months to plan and execute, and requires tremendous financial and personnel resources to run. Given the scale of the experiment, investing a few days to perform analysis that yields improved results is a worthwhile effort. Even so, an alternative to MCMC might be to explore a Laplace approximation scheme, which could potentially yield substantial improvements in computational speed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank the helpful comments of the anonymous referees and the Associate Editor. This work was supported by NIH grants U01 MH105035, U54 NS091046, and R01 NS083390.

References

Arai T, Hasegawa M, Akiyama H, Ikeda K, Nonaka T, Mori H, Mann D, Tsuchiya K, Yoshida M, Hashizume Y, Oda T. Tdp-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Biochemical and biophysical research communications. 2006; 351:602–611. [PubMed: 17084815]

Arrasate M, Mitra S, Schweitzer ES, Segal MR, Finkbeiner S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. Nature. 2004; 431:805–810. [PubMed: 15483602]

Bar H, Booth J, Schifano E, Wells MT. Laplace approximated EM microarray analysis: an empirical Bayes approach for comparative microarray experiments. Statist. Sci. 2010; 25:388–407.

Barmada SJ, Skibinski G, Korb E, Rao EJ, Wu JY, Finkbeiner S. Cytoplasmic mislocalization of tdp-43 is toxic to neurons and enhanced by a mutation associated with familial amyotrophic lateral sclerosis. The Journal of Neuroscience. 2010; 30:639–649. [PubMed: 20071528]

Bäumer D, Talbot K, Turner MR. Advances in motor neurone disease. Journal of the Royal Society of Medicine. 2014; 107:14–21. [PubMed: 24399773]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B. 1995; 57:289–300.

Efron B. Microarrays, empirical Bayes and the two-groups model. Statist. Sci. 2008; 23:1–22.

Efron B, Turnbull BB, Narasimhan B. locfdr: Computes local false discovery rates. R package version. 2011; 1:1–7.

Gelfand AE, Sahu SK, Carlin BP. Efficient parameterisations for normal linear mixed models. Biometrika. 1995; 82:479–488.

Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Statist. Sinica. 1996; 6:733–807. With comments and a rejoinder by the authors.

Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 2007; 102:359–378.

Henschel V, Mansmann U. intcox: Iterated Convex Minorant Algorithm for interval censored event data. R package version 0.9.3. 2013

Kalbfleisch JD. Non-parametric Bayesian analysis of survival time data. J. Roy. Statist. Soc. Ser. B. 1978; 40:214–221.

Kalbfleisch, JD.; Prentice, RL. Wiley Series in Probability and Statistics. second edition. Wiley-Interscience [John Wiley & Sons]; Hoboken, NJ: 2002. The statistical analysis of failure time data..

Kottas A. Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. J. Statist. Plann. Inference. 2006; 136:578–596.

Lipton SA, Rosenberg PA. Excitatory amino acids as a final common pathway for neurologic disorders. New England Journal of Medicine. 1994; 330:613–622. [PubMed: 7905600]

Miller RG, Mitchell JD, Lyon M, Moore DH. Riluzole for amyotrophic lateral sclerosis (als)/motor neuron disease (mnd). Cochrane Database Syst Rev 1. 2007

Neumann M, Sampathu DM, Kwong LK, Truax AC, Micsenyi MC, Chou TT, Bruce J, Schuck T, Grossman M, Clark CM, McCluskey LF, Miller BL, Masliah E, Mackenzie IR, Feldman H, Feiden W, Kretzschmar HA, Trojanowski JQ, Lee VM-Y. Ubiquitinated tdp-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science. 2006; 314:130–133. [PubMed: 17023659]

Plummer M. rjags: Bayesian graphical models using MCMC. R package version 3-13. 2014

Shook SJ, Pioro EP. Racing against the clock: recognizing, differentiating, diagnosing, and referring the amyotrophic lateral sclerosis patient. Annals of neurology. 2009; 65:S10–S16. [PubMed: 19191305]

Sinha D, Chen M-H, Ghosh SK. Bayesian analysis and model selection for interval-censored survival data. Biometrics. 1999; 55:585–590. [PubMed: 11318218]

Therneau TM. A Package for Survival Analysis in S. R package version 2.37-4. 2013

Therneau TM. coxme: Mixed Effects Cox Models. R package version 2. 2015:2–4.

Winer EP, Hudis C, Burstein HJ, Wolff AC, Pritchard KI, Ingle JN, Chlebowski RT, Gelber R, Edge SB, Gralow J, Cobleigh MA, Mamounas EP, Goldstein LJ, Whelan TJ, Powles TJ, Bryant J, Perkins C, Perotti J, Braun S, Langer AS, Browman GP, Somerfield MR. American society of clinical oncology technology assessment on the use of aromatase inhibitors as adjuvant therapy for postmenopausal women with hormone receptor–positive breast cancer: status report 2004. Journal of Clinical Oncology. 2005; 23:619–629. [PubMed: 15545664]

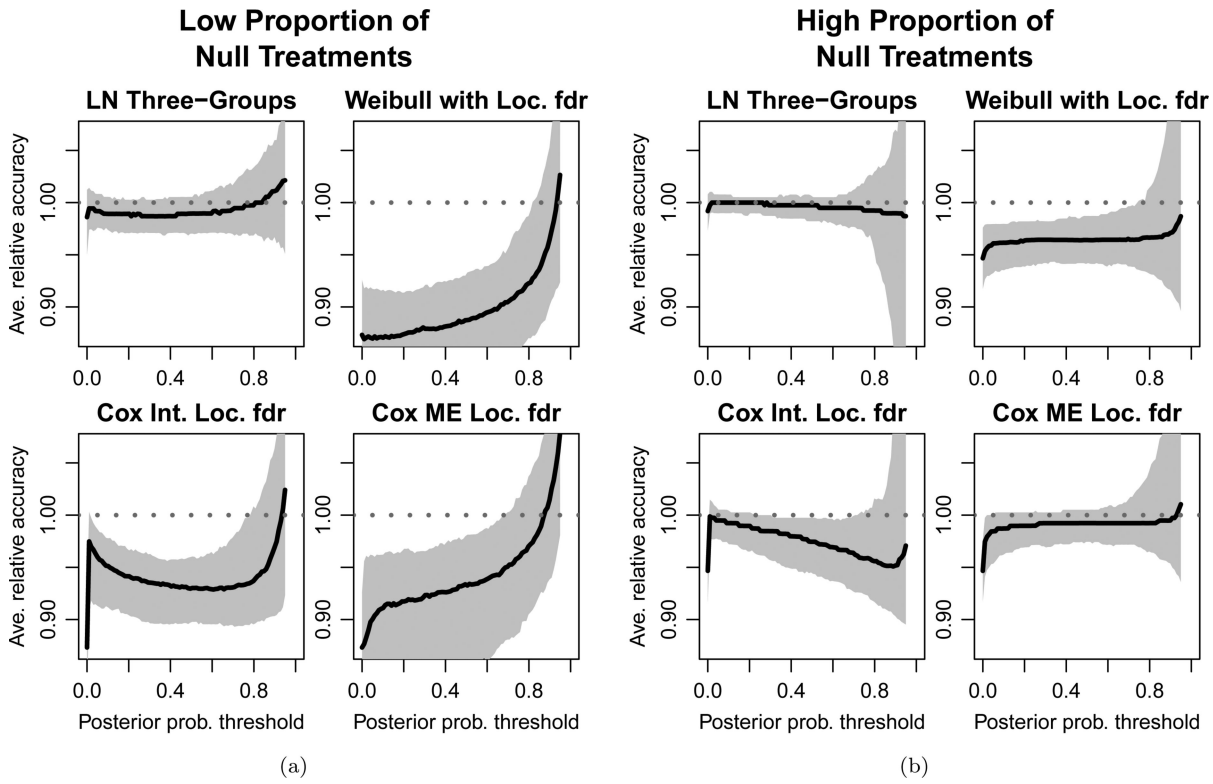


Figure 1.

Panels (a) and (b) plot accuracy of four competitors relative to the correctly-specified three-groups model, as a function of posterior probability threshold, for the simulations with low (Panel (a)) and high (Panel (b)) proportion of null treatments. The four competitors are, moving clockwise from the top left, the three-groups model with a mis-specified (lognormal) data likelihood, a parametric Weibull mixed effects model, a Cox mixed effects model that incorrectly assumes the data is right censored, and a Cox model with interval censoring and no random effects. In both scenarios, the incorrectly-specified three-groups model has similar accuracy to the correctly specified version. In the simulation with a higher proportion of null treatments, the Cox mixed effects also had a relative accuracy of close to one. In all other cases, the three-groups model showed greater accuracy, with differences more pronounced in the simulation with the lower proportion of null treatments. In all plots, the shaded areas represent pointwise 95% error bands. The analysis is based on 200 simulated datasets drawn for both the low and high proportion of null treatment scenarios.

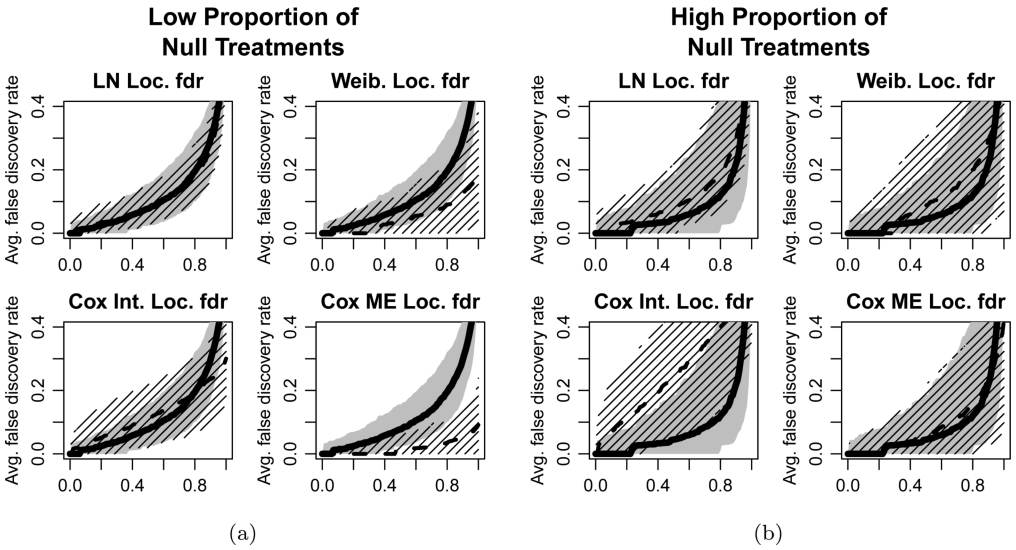
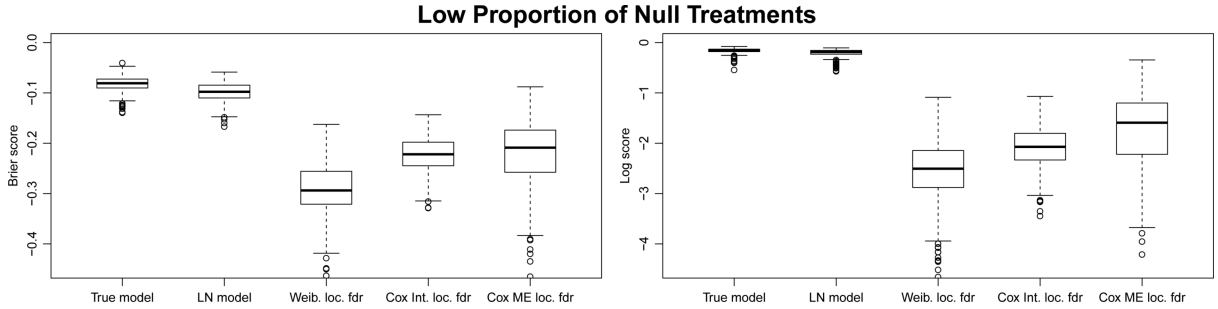
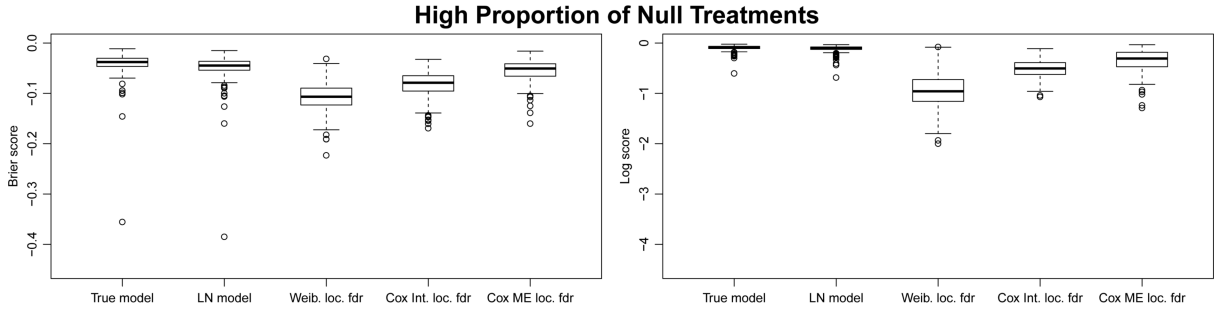


Figure 2.

Panels (a) and (b) show average false discovery rate, as a function of posterior probability threshold, for the simulations with low and high proportion of null treatments, respectively. In each plot, the average FDR from the correctly-specified three-groups model is drawn as a solid line, with a 95% pointwise confidence region shaded in gray. Drawn on top of each three-groups average FDR curve is the average FDR (drawn as a dashed line) from, moving clockwise from the top left, the three-groups model with a mis-specified (lognormal) data likelihood, a parametric Weibull mixed effects model, a Cox mixed effects model that incorrectly assumes the data is right censored, and a Cox model with interval censoring and no random effects. In all cases, a 95% pointwise confidence region is crosshatched. Broadly, the three-group model has a similar average false discovery rate to the four competitors. The Cox mixed effects model has a lower average FDR in the simulation with the low proportion of null treatments, especially at high thresholds. The analysis is based on 200 simulated datasets drawn for both the low and high proportion of null treatment scenarios.



(a)



(b)

Figure 3.

Classification results. Panels (a) and (b) compare Brier and log scores for classifying null vs. non-null treatments for the three-groups model and local fdr applied to treatment-wise *z*-scores. Panel (a) shows results from the 200 datasets simulated with the higher percentage of non-null treatments, and panel (b) shows results from the 200 datasets simulated with the lower percentage of non-null treatments. In both panels, the performance of the three-groups model, as measured by the two proper scoring rules, appears significantly superior.

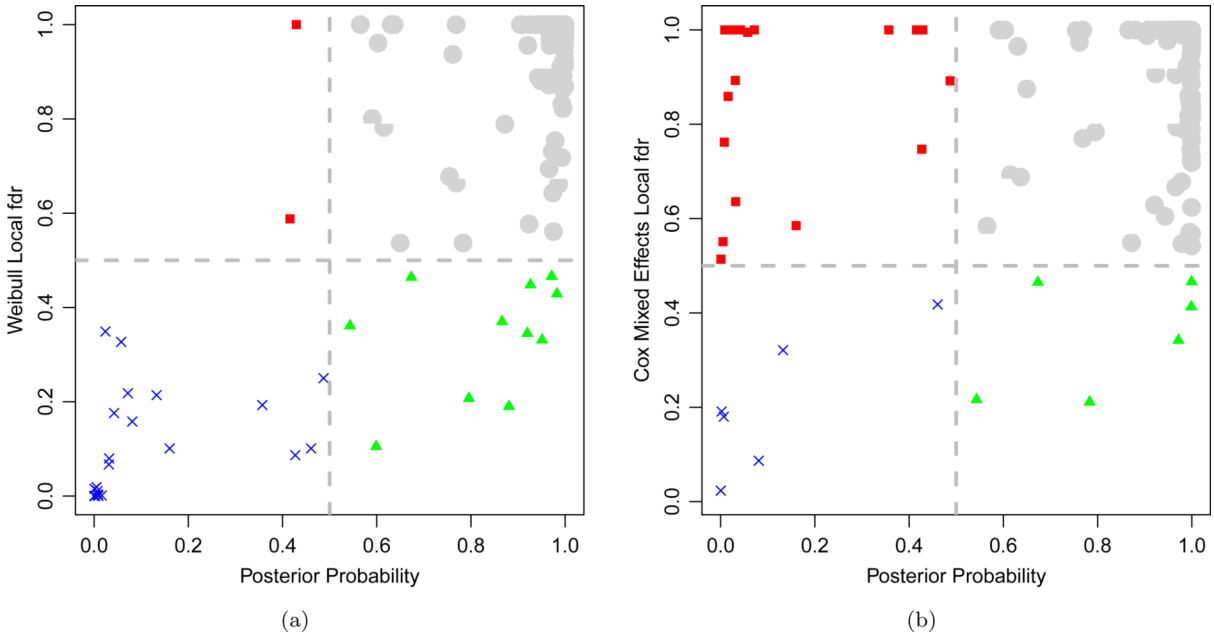


Figure 4.

Each point represents one compound in the ALS screen. The *x*-axis of each panel shows posterior probability of being null, as estimated by the three-groups model, and the *y*-axis shows the local fdr score from treatment-wise fitting of a Weibull frailty model (Panel (a)) and a Cox mixed effects model (Panel (b)). The dashed lines represent a hypothetical, arbitrary, cutoff value that might demark the boundary between declaring a compound as a hit vs. not a hit. Points in the lower left of each plot are compounds for which the three-groups model and local fdr agree are hits.

Table 1

Poster predictive assessment, similar to those advocated in Gelman et al. (1996). Observed statistics falling within the central 95% interval of statistics calculated from many synthetic draws from the model indicates good agreement with the data. The two variance statistics meet this criterion, while the mean survival time statistic falls slightly outside the interval.

Statistic	$q_{0.025}$	$q_{0.975}$	observed
Mean censored time	61.6	63.8	*60.6*
B/T plate SS (×l, 000)	20.8	33.0	29.3
B/T bio sample SS (×1, 000)	76.9	128.6	110.7

Table 2

Compounds that the three-groups model assigned non-negligible probability to membership in the beneficial group, after culling spurious hits due to visual artifacts. For each compound, the posterior probability of membership in the beneficial group is shown, as well as its associated local fdr score from the Weibull and Cox mixed effects models.

Compound name	Three groups model (Post. prob. beneficial)	Local fdr score Weibull model	Local fdr score Cox ME model
Dextromethorphan	0.513	0.250	0.892
Formestane	0.326	0.464	0.465
Prednisone	0.384	0.782	0.693
Leflunomide	0.217	0.537	0.211