Publications and Research

CUNY Graduate School of Public Health & Health Policy

2015

BAYESIAN NONPARAMETRIC CROSS-STUDY VALIDATION OF PREDICTION METHODS

Lorenzo Trippa
Harvard University

Levi Waldron
CUNY School of Public Health

Curtis Huttenhower
Harvard University

Giovanni Parmigiani
Harvard University

The Annals of Applied Statistics
2015, Vol. 9, No. 1, 402–428
DOI: 10.1214/14-AOAS798

BAYESIAN NONPARAMETRIC CROSS-STUDY VALIDATION OF PREDICTION METHODS

BY LORENZO TRIPPA[*,†], LEVI WALDRON[‡], CURTIS HUTTENHOWER[*] AND GIOVANNI PARMIGIANI[*,†]

Harvard School of Public Health[*], *Dana-Farber Cancer Institute*[†] *and School of Urban Public Health at Hunter College, City University of New York*[‡]

We consider comparisons of statistical learning algorithms using multiple data sets, via leave-one-in cross-study validation: each of the algorithms is trained on one data set; the resulting model is then validated on each remaining data set. This poses two statistical challenges that need to be addressed simultaneously. The first is the assessment of study heterogeneity, with the aim of identifying a subset of studies within which algorithm comparisons can be reliably carried out. The second is the comparison of algorithms using the ensemble of data sets. We address both problems by integrating clustering and model comparison. We formulate a Bayesian model for the array of cross-study validation statistics, which defines clusters of studies with similar properties and provides the basis for meaningful algorithm comparison in the presence of study heterogeneity. We illustrate our approach through simulations involving studies with varying severity of systematic errors, and in the context of medical prognosis for patients diagnosed with cancer, using high-throughput measurements of the transcriptional activity of the tumor's genes.

1. Introduction.

Predictive models, in most cases, need to be validated using data from independent studies. In many disciplines it is common for research communities to generate multiple data sets that address similar prediction problems. The availability of multiple data sets makes it possible to systematically compare the performance of alternative statistical learning algorithms, and to characterize their strengths and limitations in the context of a specific area of application.

Here, the term learning algorithm is used for any procedure, say, linear regression or nearest neighbor classification, that produces prediction rules. We consider the task of assessing learning algorithms, via what we call leave-one-in cross-study validation: the algorithm is trained on one data set; the resulting prediction model is then validated on each remaining data set, and a validation performance statistic (such as the classification error rate or the mean squared error of prediction) is recorded. By repeating this over all possible training data sets one generates a square array Z of validation statistics. Computation of leave-one-in matrices Z is,

Received December 2013; revised September 2014.

Key words and phrases. Reproducibility, validation analysis, meta-analysis, random partitions, Bayesian nonparametrics, cancer signatures.

in most cases, straightforward. Our goal is to develop a statistical framework for the analysis of leave-one-in matrices.

Our motivation comes from earlier experience in clinical genomics [Garrett-Mayer et al. (2008)] where the goal is to predict individual outcomes based on high-dimensional features of the genome. Leave-one-in cross-study validation is well suited to this context for two reasons. First, while different studies address the same prediction question, they may do so using different sampling designs or technological platforms, generating heterogeneity that makes it difficult to directly combine all data. Second, it is not uncommon for studies to be affected by unknown artifactual variation, such as the so-called batch effects, making it important to use methodologies that allow identification and separate handling of studies that show poor concordance with the majority of the rest [Baggerly, Coombes and Neeley (2008)].

Our perspective is therefore that cross-study validation should simultaneously be concerned about two questions: the identification of heterogeneity and outliers among studies, and the comparison of alternative algorithms, done in a way that accounts for heterogeneity across studies. We achieve this by modeling directly each of the algorithm-specific Z matrices. Variability in the validation measures contained in a Z matrix may arise from several sources, including differences in study design, study populations and measurement technologies, as well as accidental causes that may have affected data quality in individual studies. To illustrate, imagine the outcome of interest is determined by a different set of predictors in different geographical areas. A collection of studies may include two major clusters of studies, each confined to a given area. Performance evaluations are best handled by considering cross-study validation within each of these clusters, as a good algorithm should not be required to generate models that predict well across geographical areas when trained on data from a single area. Similar considerations apply to clusters defined by technological platforms.

We propose a two-stage procedure. The first stage addresses sampling variation in the Z array via Bootstrap. The second stage infers a latent partition of the studies defined by a Dirichlet process. Studies will be assigned to the same subset when the corresponding vectors of validation statistics are similar. Conversely, if the Z array provides evidence of heterogeneity between two studies, then these will tend to be assigned to separate clusters. Our model achieves two goals: (i) to cluster studies using Z, generating hypotheses on the sources of heterogeneity; and (ii) to provide cluster-based summaries of algorithm performance, allowing for comparisons that account for heterogeneity and possible systematic artifacts in the study pool.

Clustering based on the Z matrix is perhaps most attractive in the context of prediction problems with a large number of predictors. High dimensionality makes it difficult to spot the important differences between studies and to understand the factors hindering cross-study replicability. In this scenario, it is important to provide a solid evaluation of prediction strategies using distinct training and validation data sets. This evaluation should be rooted in the context of a specific application. The Z matrix helps in this: its strengths and limitations arise from reducing the problem to a single figure of merit for prediction performance. It is simple to interpret and easy to visualize. Also, it is not affected by subtle issues such as overfitting, batch effects and selection of favorable training/testing combinations. The goal of our Bayesian procedure is to retain these advantages of the Z matrix, to provide an accurate uncertainty analysis and to suggest clusters for further inquiry.

While the motivation and examples for our methodology come from clinical genomics, the only requirement for its application is the availability of independent studies using similar approaches to measure predictors.

2. Bayesian cross-study validation analysis.

2.1. *The leave-one-in validation performance matrix Z.*

We consider a set of S studies, indexed by s and including n_s samples, indexed by i. For study s, we have measurements on outcomes $Y_{s,i}$ and predictors $X_{s,i}$. Our focus is the two-dimensional array of validation statistics $Z = (Z_{s,v}; s, v = 1, \dots, S, s \neq v)$. We use the term algorithm to refer to a training methodology (such as CART or ridge regression) and the term model to refer to a specific prediction rule, resulting from using the algorithm on a training data set. For a given algorithm, the statistic $Z_{s,v}$ measures the predictive performance of the model trained on data set s, when validated on a different data set v. Typical definitions of $Z_{s,v}$ with binary outcomes include the classification error rate and, if the model generates risk scores for binary outcomes, the area under the operating characteristic curve (AUC). Validation statistics for time-to-event outcomes include versions of the concordance index [Uno et al. (2011) and references therein]. Our approach is based on the Z matrix and does not include direct modeling of the data at the individual level. This choice is motivated by the goal of obtaining easily interpretable results with modest computational effort.

In addition to $Z_{s,v}$, with $s \neq v$, one can also consider the variables $Z_{s,s}$, obtained by standard cross-validation, iteratively splitting the data set into training and validation components. Here we do not use the variables $Z_{s,s}$ to avoid summary statistics that might be inflated by systematic errors or batch effects.

2.2. *Relation to Bayesian meta-analysis.*

There are important points of contact, as well as differences, between our approach and existing ideas in Bayesian meta-analysis.

Bayesian modeling allows one to easily account for study heterogeneity. Several approaches are based on hierarchical models [Berry (1990)]. For example, Warn, Thompson and Spiegelhalter (2002) consider $S = 31$ randomized trials for assessing the analgesic Ibuprofen. The data for each study consist of sample size, number of individuals randomized to placebo and number of events (pain relief) for each arm. Treatment assignments $X_{s,i}$ and outcomes $Y_{s,i}$ are binary. They specify a hierarchical model with latent parameters θ_s describing success rates in each study and an unknown distribution F describing variability in the study specific parameters, that is, $\theta_s|F \overset{\text{i.i.d.}}{\sim} F$. The assumption that, conditionally on these parameters, individual observations within each study are independent completes the model.

Heterogeneity of study-specific parameters is often better understood via clustering, as we will propose here. Berry and Christensen (1979) introduced the idea of using a Dirichlet prior for F. A practical advantage of the Dirichlet process in this context is the resulting discreteness of F. This implies that when $(\theta_1, \dots, \theta_S)$ are sampled either from the prior or from the posterior, one observes clusters of studies: for every pair (s, v) the event $\theta_s = \theta_v$ has positive probability. Thus, one obtains *a posteriori* the distribution of a latent random partition of the studies $\{1, \dots, S\}$ dictated by ties in the values of the parameters $(\theta_1, \dots, \theta_S)$. While evidence synthesis may average over the distribution of this partition, cluster analysis can be performed by selecting a single representative partition. Model-based clustering and the use of a latent partition are effective for dealing with questions and hypotheses such as (i) the response probabilities are the same across studies, (ii) there exists a large group of studies sharing identical response probabilities and (iii) there are studies that should be considered outliers.

2.3. *Two-stage analysis.* Our validation analysis uses a summary of the data, consisting of (i) the Z array and (ii) a parametric estimate $\hat{d}$ of the unknown joint distribution d of the zero mean random variables $Z_{s,v} - \zeta_{s,v}$, where $s, v = 1, \dots, S$, $s \neq v$, and $\zeta_{s,v}$ is the expected value of $Z_{s,v}$. The expected values $\zeta_{s,v} = \mathbb{E}_{P_s,P_v}(Z_{s,v})$ refer to the true unknown distributions of the data P_s and P_v within studies s and v. These are joint distributions including both predictors and outcomes, and might vary across studies.

Our approach is in two stages. The first stage estimates the dispersion of the $Z_{s,v}$ random variables. The second stage is based on a Bayesian model, specified using a Dirichlet prior and the dispersion of the Z's estimated in the first stage.

We propose a simple hierarchical model for Z that balances (i) the need, as in any validation study, of easily interpretable summary statistics that are free of questionable assumptions and (ii) the goal of detecting clusters of studies and possible outliers. We chose a prior model for Z with a minimal level of complexity in order to avoid difficulties in the interpretation of the resulting estimates. Similar to Bayesian meta-analysis, we use latent parameters for the unknown means of our Z random variables. The posterior distribution of these parameters, as discussed in Section 3, allows clustering of the studies. The goal of the model is to cluster studies with similar data quality, as well as studies sharing similarities in their designs and implementations. We will first provide a description of our model and clustering approach in Section 3, assuming identical sample sizes $n_1 = \cdots = n_S$ across studies, and then remove this constraint.

One of the advantages of modeling the Z array is the possibility of estimating, for any pair of studies (s, v), the distribution of $Z_{s,v}$ should both studies be performed a second time. Estimates can be derived using the hypothesis that data are newly generated under identical technical conditions and that the populations from which samples arise remain identical. When the estimates of $\zeta_{s,v}$ are combined with the inferred partition of the studies $\{1, \dots, S\}$, these contribute to interpretation of the observed values in our Z array.

Stage 1. The first stage estimates d, with the goal of obtaining an approximate Bayesian analysis for the observed Z array. The approximation consists of plugging an estimate of d into the Bayesian model (Stage 2) to bypass computationally intensive joint modeling of S data sets. A practical method is the Bootstrap, either in its frequentist [Efron (1979)] or Bayesian [Rubin (1981)] versions. The resulting distribution is representative of the sampling variability of the Z statistics. The observed variations across the Z's are due to both sampling variability and also to possible differences across the study-specific distributions $P_1, \dots, P_S$.

The only result from the first stage of our procedure that we use in the analysis of the leave-one-in array is the estimate $\hat{d}$. Alternative estimators of d could in principle be used. Here we use the bootstrap because of its broad applicability. It can be applied to Z matrices generated by a spectrum of training methods ranging from popular machine learning procedures to algorithms highly tailored to specific application areas. Also, the bootstrap can estimate the variability of a number of possible validation summaries, such as the misclassification error rate or the mean squared error, that can be used to define Z arrays. Finally, the bootstrap is applicable wether or not there exists a probability model consistent with the training algorithm.

The Bootstrap [Efron (1979)] for estimating d includes (i) the computation of the empirical distributions $\hat{P}_1, \dots, \hat{P}_S$, which (ii) are then iteratively used for obtaining S independent Bootstrap samples, one for each study, $(X^*_{1,i}, Y^*_{1,i}; i \leq n_1), \dots, (X^*_{S,i}, Y^*_{S,i}; i \leq n_S)$, with $(X^*_{s,i}, Y^*_{s,i}) \sim \hat{P}_s$. Here we avoid the use of an additional index enumerating Bootstrap iterations. At each iteration the validation statistics are computed on the basis of $(X^*_{1,i}, Y^*_{1,i}; i \leq n_1), \dots, (X^*_{S,i}, Y^*_{S,i}; i \leq n_S)$, that is, the Z array is resampled. At each cycle we compute a prediction model using $(X^*_{s,i}, Y^*_{s,i}; i \leq n_s)$ and then validate it on $(X^*_{v,i}, Y^*_{v,i}; i \leq n_v)$, $s \neq v$, to obtain $Z^*_{s,v}$. Finally, (iii) we estimate d by centering the empirical distribution of the iteratively resampled arrays. The Bootstrap estimate of d, as the number of iterations increases, converges to the nonparametric maximum likelihood estimate of d. In other words, by resampling we approximate the mapping of $\hat{P}_1, \dots, \hat{P}_S$ to the distribution of $(Z_{s,v} - \zeta_{s,v}; s, v = 1, \dots, S, s \neq v)$ under the assumption that $P_s = \hat{P}_s$ for every $s \leq S$.

When d is estimated by the Bayesian Bootstrap, the flow of the procedure remains identical, with the exception that the initial components $\hat{P}_1, \dots, \hat{P}_S$ are replaced by random distributions $P^*_1, \dots, P^*_S$. The random distributions $P^*_1, \dots, P^*_S$

are defined by $P_s^* \propto \sum_{i \le n_s} W_{s,i} I_{(X_{s,i},Y_{s,i})}$, where $W_{s,i}$, $i \le n_s$, are independent exponential variables with a fixed scale parameter. The Bayesian Bootstrap averages over iteratively generated random distributions $P_1^*, \dots, P_S^*$. In this case, the resampling scheme allows us to obtain the Bayesian estimate of the Z's dispersion under Dirichlet process priors with infinitesimal concentration parameters for $P_1, \dots, P_s$.

Stage 2. We specify a Bayesian model for the validation statistics Z. To simplify posterior computations, we plug in a zero mean multivariate normal distribution $\hat{d}$ into our model by matching the covariance matrix estimate from the Bootstrap algorithm in the previous paragraph. This choice, in several cases, is justified by convergence of the actual joint distribution of the validation statistics Z, for large sample sizes, to a Normal density. We will provide examples of such convergence.

We introduce an exchangeable random partition $\Pi = \{C_1, \dots, C_m\}$ of $\{1, \dots, S\}$, where C_j, $j = 1, \dots, m$ are groups of studies. The number of clusters m is a random variable. The random partition Π of $\{1, \dots, S\}$ is specified by S exchangeable variables sampled from a discrete random distribution; the Dirichlet process is an example. We refer to Lee et al. (2013) for an overview on exchangeable partitions. We use $C(s)$ for indicating the subset of the partition Π that includes study s. Also, we use p_Π to denote the law of the random partition. We state the probability model for Z; it includes a latent partition and a set of random variables $(\mu_{i,j}; i, j = 1, \dots)$ which play a role similar to the atom locations in a Dirichlet process mixture:

$$\begin{aligned}
&\mu = (\mu_{i,j}; i, j = 1, \dots) \overset{\text{i.i.d.}}{\sim} p_\mu, \\
&\Pi \sim p_\Pi, \\
&\varepsilon = (\varepsilon_{s,v}; s, v = 1, \dots, S, s \ne v) \sim \hat{d} \quad \text{and} \\
&Z_{s,v} = \mu_{C(s),C(v)} + \varepsilon_{s,v}, \qquad s, v = 1, \dots, S, s \ne v,
\end{aligned} \tag{2.1}$$

where the components μ, Π and ε are a priori independent and p_μ is a distribution on the real line.

The probability that the conditional expected values of a pair (s, v) of Z columns (or rows) are identical is strictly positive:

$$p\biggl(\bigcap_{r \le S} \{\mu_{C(s),C(r)} = \mu_{C(v),C(r)}, \mu_{C(r),C(s)} = \mu_{C(r),C(v)}\}\biggr) > 0.$$

Also, the distribution of the array $(\mu_{C(s),C(v)}; s, v = 1, \dots, S, s \ne v)$ is invariant with respect to any permutation $\sigma = (\sigma_1, \dots, \sigma_S)$ of $\{1, \dots, S\}$,

$$(\mu_{C(s),C(v)};\ s, v = 1, \dots, S) \overset{d}{=} (\mu_{C(\sigma_s),C(\sigma_v)}; s, v = 1, \dots, S).$$

The model can handle an arbitrary number of additional studies ($S+1$, $S+2,\dots$). Therefore, one can perform predictive inference by considering a future ($S+1$)th study and obtain, conditionally on the observed Z statistics, the distribution of $(\mu_{C(S+1),C(s)}, \mu_{C(s),C(S+1)}; s=1,\dots,S)$.

Arrays with exchangeable rows and columns have been studied in a series of papers beginning with the contributions of Aldous (1981) and Hoover (1982). These authors proved de Finetti-type representations for these processes. Random arrays invariant in distribution to any simultaneous permutation σ of rows and columns, such as $(\mu_{C(s),C(v)}; s, v \geq 1)$, are called jointly exchangeable. This type of arrays arises, for instance, when relationships between individuals are represented using two-way tables [Roy and Teh (2009)]. In our study, these representation theorems provide a formal justification to use latent cluster membership variables for modeling exchangeable arrays.

2.4. *Asymptotic normality of validation arrays.* The proposed model for Z is closely connected with Dirichlet process mixtures. Consider, for example, S studies designed for estimating $\theta_s = \mathbb{E}(Y_{s,i})$. A possible approach for exploring the hypothesis of multiple clusters defined by studies with identical means θ_s consists in combining approximate likelihood functions $N(\bar{Y}_s = \sum_i Y_{s,i}/n_s; \theta_s, \hat{\sigma}_s^2/\sqrt{n}_s)$ with a random distribution F for the means, that is, $\theta_s | F \overset{\text{i.i.d.}}{\sim} F$. See Burr and Doss (2005) for a detailed study of this approach, and Dersimonian and Laird (1986) for a frequentist perspective. The approximation, from a Bayesian standpoint, consists in using Normal kernels with scale parameters $\sqrt{\sum_i (Y_{s,i} - \bar{Y}_s)^2/n_s}$, and is supported by asymptotic arguments. Similarly, we combine an exchangeable random partition with a multivariate Normal kernel $\hat{d}$ justified, in several cases, by asymptotic arguments.

A smooth estimate of d is computationally convenient and circumvents artifacts that arise with a discrete one, including the possibility of posterior distributions assigning exactly null probability to most of the Π configurations. One can identify several cases in which the leave-one-in array is asymptotically Normal. Below we briefly discuss one case where Z converges to a multivariate Normal distribution on a linear subspace of $\mathbb{R}^{S\times(S-1)}$. We discuss results for logistic regression, Poisson regression, proportional hazards models and support vector machine procedures in the supplementary material [Trippa et al. (2015)].

Consider the linear model $Y_s | X_s \sim N(X_s\beta_s, I\sigma_s^2)$, with $(Y_s, X_s) = (Y_{s,i}, X_{s,i}; i \leq n_s)$, least squares estimates $\hat{\beta}_s$ and mean squared errors (MSE) of prediction

$$Z_{s,v} = \frac{\|Y_v - X_v\hat{\beta}_s\|^2}{n_v}.$$

Here, and in all the examples in the Supplementary Material, we let all sample sizes grow at the same rate, $n_s \approx c_s n_1$, $s = 2,\dots,S$, and fix $c_2,\dots,c_S$. Independence of $\|Y_v - X_v\hat{\beta}_v\|^2$ and $(X_v, \hat{\beta}_v)$ implies, under mild assumptions on the $X_{s,i}$ distributions, asymptotic normality. First, $n_v^{-1/2}(\|Y_v - X_v\hat{\beta}_v\|^2 - n_v\sigma_v^2) \to N(0, 2\sigma_v^2)$. Next, to obtain $Z_{s,v}$, we need to add to the in-sample mean squared error $n_v^{-1}\|Y_v - X_v\hat{\beta}_v\|^2$ a second term, $n_v^{-1}[\|X_v(\beta_v - \beta_s)\|^2 + 2(\delta_v - \delta_s)X_vX_v(\beta_v - \beta_s) + \|X_v(\delta_v - \delta_s)\|^2]$, with $\delta_v = (\hat{\beta}_v - \beta_v)$. It can be shown that $n_v^{-1/2}(\delta_v - \delta_s)[X_vX_v(\beta_v - \beta_s) - \mathbb{E}(X_vX_v(\beta_v - \beta_s))] \to 0$ and $n_v^{-1/2}(\delta_v - \delta_s)X_vX_v(\delta_v - \delta_s) \to 0$. Finally, both $n_v^{-1/2}(\delta_v - \delta_s)\mathbb{E}(X_vX_v(\beta_v - \beta_s))$ and $n_v^{-1/2}(\beta_v - \beta_s)(X_vX_v - \mathbb{E}(X_vX_v))(\beta_v - \beta_s)$ converge to normal densities. Asymptotic joint normality for Z follows from the asymptotic independence of δ_v and $n_v^{-1/2}(X_vX_v)$.

3. Cluster-based validation statistics. The procedure we propose generates a posterior distribution $p(\Pi|Z)$ for the unknown partition Π of our S studies. The tuning of the distribution p_Π and approaches for selecting the prior model are discussed in the supplementary material [Trippa et al. (2015)]. A representative partition summarizes the posterior distribution. We select an estimate $\hat{\Pi}$ that minimizes the expectation of a loss function $l(\hat{\Pi}, \Pi)$, that is, $\hat{\Pi} = \arg\min \mathbb{E}(l(\cdot, \Pi)|Z)$. The partition $\hat{\Pi}$ is a posterior point estimate. Quintana and Iglesias (2003) give a discussion on the decision theoretic paradigm applied to random partitions. Several loss functions $l(\hat{\Pi}, \Pi)$ have been proposed; see, for example, Denœud and Guénoche (2006).

We use the easily interpretable *maximum transfer metric*; see Charon et al. (2006) for a recent contribution. This metric $l(\Pi_1, \Pi_2)$ is defined as the minimum number of elementary corrections necessary to match the partitions Π_1 and Π_2; an elementary correction consists of moving a unit to a different (possibly empty) subset. If we consider, for example, $\Pi_1 = (\{1, 2\}, \{3, 4\})$ and $\Pi_2 = (\{1, 4\}, \{2, 3\})$, then $l(\Pi_1, \Pi_2) = 2$, and a possible chain of corrections is $(\{1, 2\}, \{3, 4\}) \to (\{1, 2, 3\}, \{4\}) \to (\{1, 3\}, \{2, 4\})$.

Our procedure tends to assign studies to separate clusters when they differ on aspects that affect the validation statistics Z. The dissimilarity captured by the clustering method might be due to different measurement techniques, different predictors distributions or other factors varying across studies. Interpretation of the inferred partition requires subsequent analyses to identify the primary causes of heterogeneity, such as data quality or experimental designs. The results can then inform the construction of models trained on multiple data sets. If, for instance, heterogeneity is driven by different distributions of relevant predictors, but the covariates effects on the outcome are consistent across studies, then it might be appropriate to combine the available data sets. In contrast, if heterogeneity is driven by measurement errors or batch effects, additional efforts may focus on data normalization steps.

We can now introduce the concept of clustering-based validation performance measure, by which we mean summary statistics aimed at assessing cross-study prediction taking into account study heterogeneity and within-cluster similarities.

Recall that model 2.1 formalizes the identity between the conditional expected values of $Z_{s,v}$ and $Z_{s',v'}$ when study s clusters together with s' and v clusters with v'. For example, we may be interested in the performance measure obtained when one trains on any of the studies in the cluster of study s and validates in any of the studies in the cluster of study v, that is, $\mu_{C(s),C(v)}$. The latent variable $C(s)$ indicates the cluster that includes s and $\mu_{C(s),C(v)}$ can be interpreted as the expectation of $Z_{s,v}$ assuming that studies s and v are repeated de novo. A point estimate $\mathbb{E}(\mu_{C(s),C(v)}|Z)$ can be obtained by averaging $\mathbb{E}(\mu_{C(s),C(v)}|\Pi, Z)$ with respect to the posterior distribution of the partition Π. Similarly, one may derive interval estimates.

We can also estimate the validation performance that one would obtain from training in a study from the set $C(s)$, and validating in a future $(S+1)$th study, by using $\mu_{C(s),C(S+1)}$. In particular, the joint posterior distribution of $\mu_{C(s),C(S+1)}$ and $\mu_{C(v),C(S+1)}$, with $s, v \leq S$, can be used for comparing studies s and v.

Let B be a subset of studies in $\{1, \dots, S\}$. We extend the definition of the validation statistic $Z_{s,v}$ to handle the case where a model is trained on the combination of the data from all the studies in B, and then validated on study v. We denote the resulting validation statistic by $Z_{B,v}$. If B includes v, then v is not used to train the model, and $Z_{B,v}$ is redefined to be the same as $Z_{B\setminus v,v}$, where $B \setminus v = \{s \leq S : s \in B \text{ and } s \neq v\}$. We also use $B(s) \subset \{1, \dots, S\}$ to denote the studies within the same Π latent cluster of s, that is, $B(s) = \{v \leq S : C(s) = C(v)\}$.

Clustering has the goal of identifying homogeneous groups of studies with similar sampling distributions. When this works, it is natural to train models by combining the studies in a cluster. However, the figure of merit used for the Z summary, not unlike a loss function, implies adopting a specific one-dimensional perspective in looking at the data. It is possible, for example, that two studies with different covariate distributions might be clustered together, or two studies which only differ in design, but not in the populations, may be allocated to separate clusters.

Clustering can be used to estimate the performance obtained when validating in study s after training on studies in $B(s)$, that is, using only data sets similar to s. This task reduces to estimating $Z_{B(s),s}$. The function $B \to Z_{B,s}$, over the collection of $\{1, \dots, S\}$ subsets, can be directly computed using our S data sets and is not related with the Bayesian model, but $Z_{B(s),s}$, the value of this function at $B(s)$, is estimated because $B(s)$ is an unknown latent component of the model. This approach is only useful when there is no strong evidence that s belongs to a singleton cluster. We thus estimate $Z_{B(s),s}$ by using the posterior distribution of the partition Π and conditioning on the event $B(s) \neq \{s\}$. We report both the estimate of $Z_{B(s),s}$ obtained by averaging over Π configurations with $B(s) \neq \{s\}$ and the posterior probability of the conditioning event $B(s) \neq \{s\}$. Alternatively, we can generate a plug-in estimate $Z_{\hat{B}(s),s}$ by focusing on $\hat{B}(s)$, the cluster in $\widehat{\Pi}$ that includes the sth study.

When we estimate $Z_{B(s),s}$ the goal is to evaluate a model trained by a homogeneous set of studies $B(s)$. Our clustering procedure uses validation statistics to detect study heterogeneity, and therefore the resulting partition is representative of differences between studies captured by the Z validation summaries. Studies included in the same cluster could still differ in important ways. We consider this point further in the discussion.

For comparing studies, we also need to be concerned about the potential for variations of clustering-based summaries, such as $Z_{B(s),s}$, driven by different total sample sizes within each cluster. Under the assumption of identical sample sizes $n_1 = \cdots = n_S$, which will be later removed, one can expect that the value $Z_{B(s),s}$ improves with the number of studies in $B(s)$. We thus define the *sample size adjusted* validation statistics $Z^j_{B,s}$. The definition of these statistics is analogous to that of $Z_{B,s}$. We randomly select j distinct samples from the ensemble of studies B. We train a model on these j samples and validate it on data set s to generate a performance measure, say, an AUC. We iterate this procedure, keeping fixed both B and s; $Z^j_{B,s}$ is the average of the accuracy measures obtained during these iterations. In this case, if B includes s, then the units in s are not selected for training the model. The index j can vary from a minimal size of interest up to the overall number of samples in $B \setminus s$.

Our interest is in the map $j \to Z^j_{B(s),s}$; recall that $B(s)$ is unknown but can be estimated using the posterior distribution of Π. The statistics $Z^j_{B(s),s}$ have an interpretation similar to $Z_{B(s),s}$; moreover, one can contrast the estimates of $Z^j_{B(s),s}$ and $Z^j_{B(v),v}$ to compare the sth study to the vth study. We can estimate $Z^j_{B(s),s}$ plugging in the point estimate $\hat{\Pi}$ or directly using the posterior distribution of Π. If we follow the first approach, the estimator is $Z^j_{\hat{B}(s),s}$, while the second approach averages with respect to the posterior distribution of $B(s)$. In both cases we estimate, assuming $\sum_{B(s)} n_v \geq j + n_s$, the mean value of the validation statistic when the algorithm is trained by j data points from the unknown subset $B(s) \setminus s$ and then validated on s. In the second case, we report the posterior probability of $\sum_{B(s)} n_v \geq j + n_s$, and compute our estimate conditionally on this event because $Z^j_{B(s),s}$ is well defined only when $B(s)$ includes at least $j + n_s$ units.

4. Simulation study.

4.1. *Scenario* 1. The goal of this simulation study is to illustrate the extent to which our model-based approach contributes to the interpretation of cross-study validation statistics, beyond what can be learned from direct visualization of Z. As this relies on estimating the unknown partition Π and the latent $\mu_{C(s),C(v)}$ variables, we also discuss our model's ability to reconstruct these.

The scenario is defined by 9 studies grouped into three clusters, $C_1 = \{1, 2, 3\}$, $C_2 = \{4, 5, 6\}$ and $C_3 = \{7, 8, 9\}$, which differ in the amount of measurement error in the predictors. All studies have a sample size of 300. For subject i from study s we have a binary outcome $Y_{s,i}$ and 50 candidate predictor variables $X_{s,i}$.

In group C_1, the 50 covariates are simulated from a multivariate Normal distribution with null mean; all variances are equal to 17. The dependence between $X_{s,i}$ and $Y_{s,i}$, $s = 1, 2, 3$, is specified by a logistic regression function; 10 regression coefficients are equal to 0.1 and 40 are equal to 0. In group C_2 we add independent measurement errors with null mean and standard deviation equal to 14 to 50% of the covariates. In C_3 we add independent measurement errors with mean 0.33 and standard deviation 8 to all covariates.

For each study we obtain a prediction model by fitting a logistic function using ridge regression; we tune the penalization parameter with standard cross-validation. We then assess model performance using the mean absolute error (MAE) of prediction, that is, $Z_{s,v} = n_v^{-1} \sum_i \|Y_{v,i} - \text{logit}^{-1}(\hat{\beta}_s^o + \hat{\beta}_s X_{v,i})\|$, where $(\hat{\beta}_s^o, \hat{\beta}_s)$ denote the regression coefficients estimated using only data from study s.

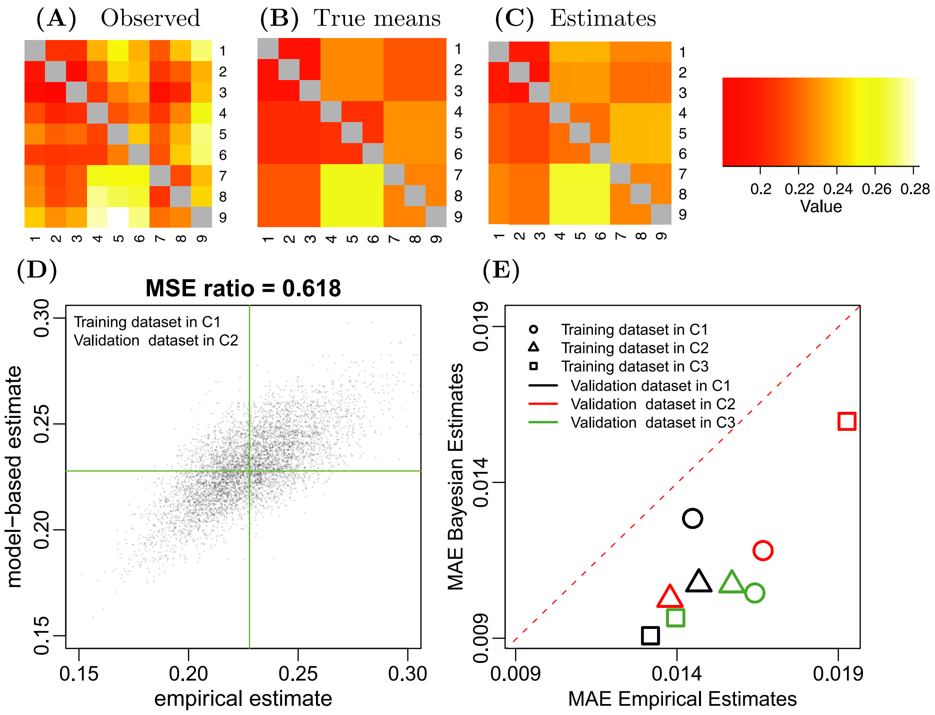

FIG. 1. *Leave-one-in array, with rows corresponding to training data sets and columns to validation data sets. Panel* (A) *shows the leave-one-in array Z for a single simulation. Panel* (B) *shows the true expected values $\zeta_{s,v}$ of $Z_{s,v}$. Panel* (C) *shows the Bayesian estimates $\mathbb{E}(\mu_{C(s),C(v)}|Z)$. The diagonals in panels* (A), (B) *and* (C) *are blank. Panel* (D) *considers* 500 *simulations and plots the empirical estimates $Z_{s,v}$ against the Bayesian estimates $\mathbb{E}(\mu_{C(s),C(v)}|Z)$. Panel* (D) *considers a training data set s in C_1 and a validation data set v in C_2. The green lines correspond to the true expected value $\zeta_{s,v}$. Panel* (D) *also reports the MSE ratio contrasting the Bayesian estimates with the empirical estimates. Panel* (E) *contrasts the Bayesian estimates of $\zeta_{s,v}$ with the empirical estimates by displaying the MAEs. Panel* (E) *considers all combinations with s and v in C_1, C_2 or C_3.*

Figure 1(A) shows the Z array for a single simulation, with rows corresponding to training data sets and columns corresponding to validation data sets. This array shows that sampling variability accounts for a relevant part of the observed differences across validation summaries, and the resulting panel is not easily interpretable by direct visual inspection. Figure 1(B) shows Monte Carlo approximations of the true expected values $\zeta_{s,v}$ of the $Z_{s,v}$ variables under the described sampling models. The expected value $\zeta_{s,v}$ is computed integrating with respect to the actual distributions (P_s, P_v) of $(X_{s,i}, Y_{s,i})$ and $(X_{v,i}, Y_{v,i})$. Figure 1(C) shows the cluster-based Bayesian estimates $\mathbb{E}(\mu_{C(s),C(v)}|Z)$ based on our two-stage procedure. In this simulation, our clustering procedure gives a point estimate $\hat{\Pi} = [\{1\}, \{2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}]$ of the latent partition. The distance $l(\hat{\Pi}, \Pi_{\text{TRUE}})$, measured with the maximum transfer metric, is equal to 1.

Comparison of panels (A) and (C) shows that the two-stage procedure correctly reconstructs the block structure of the true expected values $\zeta_{s,v}$ displayed in panel (B). Also, the procedure correctly identifies a group of studies, which are not affected by measurement errors, with estimated $\mu_{C(s),C(s)}$ value below 0.2.

We repeated the simulation 500 times. In each iteration, and for each pair (s, v), we estimated the unknown $\zeta_{s,v}$ means using our Bayesian estimator $\mathbb{E}(\mu_{C(s),C(v)}|Z)$ and the empirical estimator $Z_{s,v}$. The results are plotted in Figure 1(D) against each other for a single (s, v) combination, with s in C_1 and v in C_2. Then, for each (s, v) combination, we contrasted the MSEs and the MAEs of the Bayesian estimates with the empirical estimates. Across all (s, v) combinations the Bayesian estimator has lower MSE and MAE than the empirical estimates. These results are graphed in panel (E); each point corresponds to one (s, v) combination, and the MAEs of the Bayesian and empirical estimates are plotted against each other. In this comparison the Bayesian estimator achieves a substantially lower dispersion around the true expected value $\zeta_{s,v}$ compared to the empirical estimator.

For each simulation we computed $l(\hat{\Pi}, \Pi_{\text{TRUE}})$, the number of elementary set operations between the true and estimated latent partition. On average this distance is 1.63 and, in most iterations, $\hat{\Pi}$ has a distance of 2 set operations or less from Π.

4.2. *Scenario* 2. We consider a sampling model previously used in Waldron et al. (2011). We use it to investigate how the comparison of alternative algorithms is enhanced by Bayesian modeling of the Z arrays. Here we add measurement errors to the outcome variable in subsets of studies. We investigate how modeling of Z allows algorithm performance assessment for continuously varying training sample size. The main focus is on the maps $j \rightarrow Z^{j}_{B(s),s}$ to contrast methods. We also highlight how posterior inference on clustering based statistics, such as the estimates of $\mu_{C(s),C(v)}$, captures uncertainty on the algorithms' performances.

We simulated 540 zero-mean Normal predictors $X_{s,i}$ with a covariance matrix structured in blocks:

$$\sigma_{l,j} = \begin{cases} 1, & \text{if } l = j, \\ 0.2, & \text{if } l, j \leq 100 \text{ and } l \neq j, \\ 0.2, & \text{if } 100 < l, j \leq 200 \text{ and } l \neq j, \\ 0.2, & \text{if } 200 < l, j \leq 370 \text{ and } l \neq j, \\ 0, & \text{otherwise.} \end{cases}$$

Conditionally on these predictors, we then generated binary outcomes $Y_{s,i}$ with $\mathbb{E}(Y_{s,i}|X_{s,i}) = [1 + \exp(-\beta X_{s,i})]^{-1}$. Here $i \leq n_s = 100$ and $s = 1, \ldots, 9$. The regression coefficients $(\beta_1, \ldots, \beta_{540})$ are $\beta_j = 0.2$ for $j \leq 370$ and $\beta_j = 0$ for $j > 370$. Departing from the sampling model of Waldron et al. (2011), we added measurement errors to $Y_{s,i}$, by changing the value of $Y_{s,i}$ with probability 0.05 in $C_1 = \{1, 2, 3\}$, 0.25 in $C_2 = \{4, 5, 6\}$ and 0.5 in $C_3 = \{7, 8, 9\}$. These probabilities are independent of Y_s and X_s. Note that any classification approach applied to studies $s = 7, 8, 9$ has an average error rate of 0.5 because the binary outcomes $Y_{s,i}$, after measurement errors, become independent from the covariates and $\mathbb{E}(Y_{s,i}|X_{s,i}) = 0.5$.

We consider for illustration three classification methods: LASSO regression, ridge regression and a linear support vector machine; penalization parameters are tuned with cross-validation. We choose our validation statistics to be the classification error rates. For each study s, we computed the true clustering-based $Z^j_{B(s),s}$ statistics; in simulation studies the true latent partition, as well as $B(s)$, $s = 1, \ldots, S$, is known. If, for instance, $s = 1$, then $B(s) \setminus s = \{2, 3\}$, and $Z^j_{B(s),s}$ measures the average classification performance obtained when a model is trained by $j \leq 200$ records randomly sampled from $B(s) \setminus s$. The classification performance is obtained through empirical validation on data set s.

We then used the posterior distribution of $B(s) \setminus s$ and computed the estimates $\mathbb{E}(Z^j_{B(s),s}|Z, \sum_{v \in B(s)\setminus s} n_v \geq j)$. The first three panels in Figure 2 contrast $Z^j_{B(s),s}$ (dashed lines) with the Bayesian estimates (solid lines); each color corresponds to one of the three methods. Overall, the estimates correctly portray the differences that exist between the performances of the three algorithms; in this scenario, the support vector machine slightly outperforms ridge regression, which, in turn, has lower prediction errors than LASSO. These differences are shown in the third row of Figure 2 where we plot the maps $j \to \zeta^j_s$, with ζ^j_s equal to the expected value of $Z^j_{B(s),s}$. The second line of panels in Figure 2 shows the posterior probabilities $p(C(s) = C(v)|Z)$. In this example, the proposed model captures the underlying partition of the 9 studies and the differences across methods' performances.

We repeated the simulation 500 times, generating 9 independent data sets for each iteration. In the bottom three panels of Figure 2, we show medians and quartiles of the $Z^j_{B(s),s}$ posterior estimates, for $j = 100, 200$, obtained across these 500

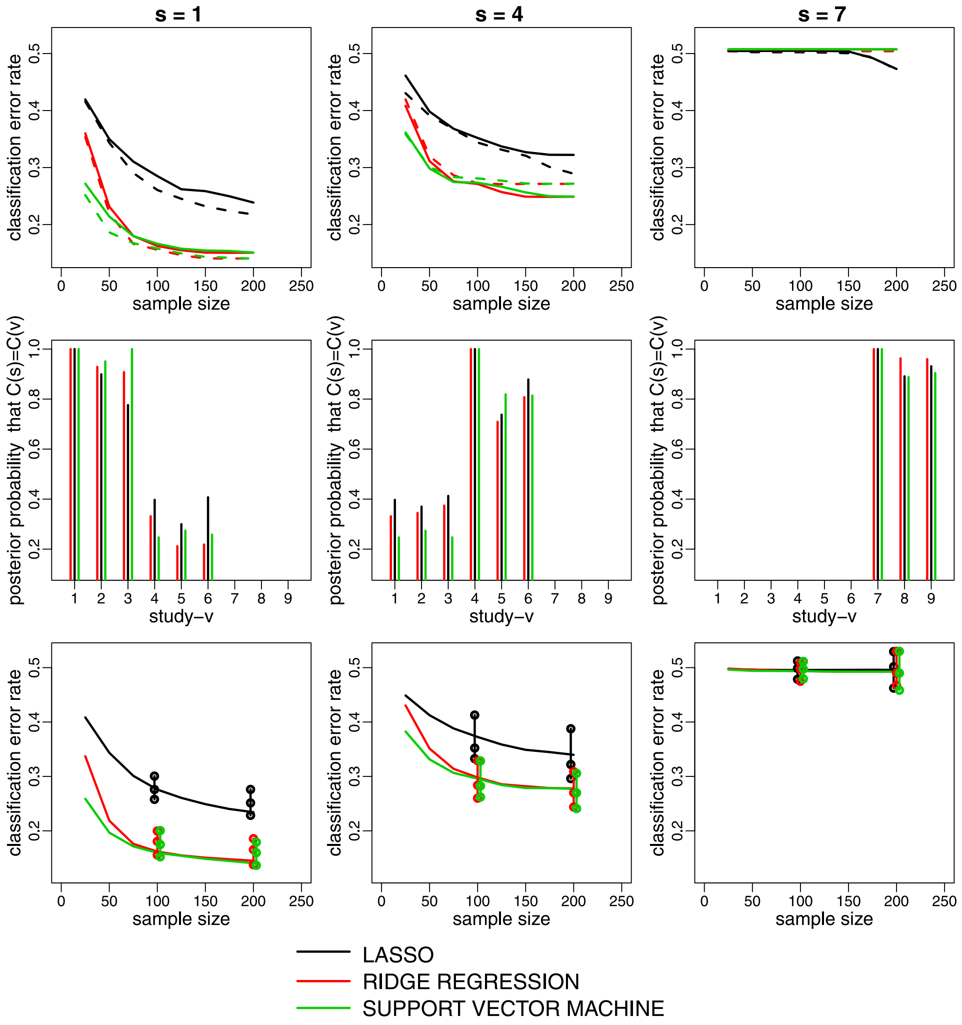

FIG. 2. *Clustering based validation statistics. The top row considers a single simulation and compares the true values of the validation statistics* $Z^j_{B(s),s}$ (*dashed lines*) *with our Bayesian estimates* (*solid lines*) *for varying size of training sets. Using the same data*, *the second row displays the posterior probabilities that two studies*, *s and v*, *are clustered together. The third row summarizes results from* 500 *simulations*; *solid lines display the true expected values* ζ^j_s *of* $Z^j_{B(s),s}$, *while dots marks medians and quartiles of the corresponding Bayesian estimates at* $j = 100$ *and* $j = 200$ *across simulations. Colors denote the three algorithms. In the* 1*st* (2*nd*, 3*rd*) *column* $s = 1$ (4, 7).

iterations. These are compared to approximations of the true maps $j \to \zeta^j_s$, obtained by averaging the true error rates $Z^j_{B(s),s}$ across simulations. These maps are displayed with solid lines in the third row of panels in Figure 2. These panels summarize the distribution across simulations of the estimated clustering validation measures $Z^j_{B(s),s}$ and confirm that the estimates are representative of the performances of the algorithms being compared.

5. Application to survival prediction in cancer.

We illustrate an application to the development of a prediction model for overall survival of ovarian cancer patients using microarray gene expression data. Ovarian cancer is the most lethal gynecological cancer, and numerous groups have undertaken microarray experiments to measure tumor gene expression for development of prognostic models of patient survival. It is widely accepted that gene signatures proposed for clinical application must be validated on independent data sets. In this area of research several strategies and methods have been proposed for prediction. Which one works best? How much uncertainty is involved in ranking methods? Posterior probabilities on the $\mu_{C(s),C(v)}$ random variables are suitable for answering these questions.

We identified nine previously curated studies utilizing five different microarray platforms, each providing patient overall survival for at least 40 late-stage, serous-type, ovarian tumors (Table 1). Microarray data were processed using standard normalization methods, after which probe identifiers were mapped to standard gene symbols, as provided by the *curatedOvarianData* library [Ganzfried et al. (2013)]. Only gene symbols represented on all platforms were considered for across-platform comparability. We noted that limiting consideration to those genes present across all platforms has a negligible effect on prediction performance. For example, we separately fitted Cox models with ridge penalty and estimated with cross-validation C-statistics, separately considering one study at a time; the average decrease of the C-statistics when only genes present in all platform were considered compared to using all available genes was less than 0.01.

5.1. *Accounting for different sample sizes.*

The sample size n_s varies across studies. One can therefore expect higher values of the validation statistics $Z_{s,v}$ for

TABLE 1
The nine ovarian cancer data sets considered in this study. We only considered late-stage serous tumors from these studies

s	**Study**	n_s	**Microarray platform**
1	Bentink et al. (2012)	117	Illumina Human v2
2	Crijns et al. (2009)	157	Operon Human v3
3	Yoshihara et al. (2010)	110	Affymetrix hgug4112a
4	Bonome et al. (2008)	185	Affymetrix hgu133a
5	Tothill et al. (2008)	139	Affymetrix hgu133plus2
6	The Cancer Genome Atlas Research Network (2011)	420	Affymetrix hthgu133a
7	Mok et al. (2009)	53	Affymetrix hgu133plus2
8	Konstantinopoulos et al. (2010)	42	Affymetrix hgu95av2
9	Dressman et al. (2007)	59	Affymetrix hgu133a

those models trained in the largest studies. To prevent this from creating artifactual clusters of studies, we apply an intuitive correction.

We selected a threshold of 110 samples and considered the 6 studies that have a sample size larger than the threshold. We then computed the empirical estimates $Z^{110} = (Z^{110}_{s,v}; s, v = 1, 2, 3, 4, 5, 6, s \neq v)$. The computation of $Z^{110}_{s,v}$ is straightforward. We iterate two steps: (i) we train a prediction model M^{110}_s with 110 data points sampled without replacement from the sth data set, then (ii) we validate the resulting model on the entire vth data set. We set $Z^{110}_{s,v}$ equal to the average value of the validation statistics across iterations. The statistic $Z^{110}_{s,v}$ estimates the performance of a model trained by 110 samples from P_s. We computed these estimates with 200 iterations.

The covariance matrix Σ^{110} of Z^{110} is then estimated by bootstrapping. The array Z^{110} and $\hat{\Sigma}^{110}$ are used for obtaining the posterior distribution of the random partition Π^{110} with the model proposed in Section 2; we only replace (Z, Π) with (Z^{110}, Π^{110}). The reported probability that two studies, say, $s = 1$ and $v = 6$, belong to the same cluster is provided by the posterior distribution $p(\Pi^{110}|Z^{110}, \hat{\Sigma}^{110})$.

Next, we need to extend this posterior, which refers to the 6 studies we selected, to the remaining 3 which have less than 110 samples. To achieve this goal, we compute $p(\Pi^{42}|Z^{42}, \hat{\Sigma}^{42})$ by reducing the threshold from 110 to 42, and report the following adjusted random partition:

$$\hat{p}(\Pi = \pi) = \frac{p(\Pi^{42} = \pi|Z^{42}, \hat{\Sigma}^{42}) \times p(\Pi^{110} = \Delta^{110}(\pi)|Z^{110}, \hat{\Sigma}^{110})}{\sum_{\pi'} \mathbf{1}(\Delta^{110}(\pi') = \Delta^{110}(\pi)) p(\Pi^{42} = \pi'|Z^{42}, \hat{\Sigma}^{42})}, \tag{5.1}$$

where the sum is over possible partitions of the 9 studies and the operator Δ^{110} projects them into partitions of the 6 studies $\{1, 2, 3, 4, 5, 6\}$ above the 110 samples threshold. Two of these 6 studies (s, v) are clustered together by π if and only if they are clustered together by $\Delta^{110}(\pi)$. Expression (5.1) implies $\hat{p}(\Delta^{110}(\Pi) = \cdot) = p(\Pi^{110} = \cdot|Z^{110}, \hat{\Sigma}^{110})$.

This correction for sample size effects preserves the interpretability of the clustering algorithm. It also avoids more complex constructions, such as replacing the latent random variables μ in (2.1) with latent functions for sample size-specific average validation statistics. The most computationally intensive stage of the procedure is the computation of Σ^{42} and Σ^{110}; the arrays Z^{42} and Z^{110} have been resampled 1000 times.

5.2. *Comparative analysis of prediction methods.* Ovarian cancer studies for developing prognostic signatures are commonly based on two distinct groups of data sets, a training group, which in most cases only includes a single data set, and a group of publicly available validation data sets. A recent example that presents key questions related with our study is Kang, D'Andrea and Kozono (2012), and the subsequent comment Swisher, Taniguchi and Karlan (2012). The goal in Kang, D'Andrea and Kozono (2012) is to develop a molecular score based on expression of 151 genes that are involved in platinum-induced DNA damage repair to predict response to chemotherapy. This exemplifies using a biological hypothesis to preselect predictors for constructing prognostic models, thus avoiding some of the challenges arising in the "large p small n" setting. In Swisher, Taniguchi and Karlan (2012) authors point out that both independent validations and a suitable sample size of the validation data set are essential for assessing prediction models.

Prescreening the space of predictors has the advantages of parsimony and interpretability, but comes at the cost of some information loss. Our goal in this section is to quantify this trade-off using cross-study validation. We use the truncated C_τ statistic as proposed in Uno et al. (2011), truncated at $\tau = 3$ years, for measuring survival prediction accuracy. The C_τ statistic, given a prediction model M and independent (possibly censored) survival data with covariates (Y_i, X_i), $i = 1, \dots, n$, from an unknown distribution P, estimates the conditional probability $P(r_{n+1} \geq r_{n+2} | Y_{n+1} \leq Y_{n+2}, Y_{n+1} \leq \tau)$. The random variables (r_{n+1}, r_{n+2}) are risk scores computed from M based on individual covariates (X_{n+1}, X_{n+2}); if, for instance, M is a proportional hazards model with coefficients $\hat{\beta}$, then $r_{n+1} = \hat{\beta} X_{n+1}$ and $r_{n+2} = \hat{\beta} X_{n+2}$. The estimate converges, under the assumption of noninformative censoring, to the unknown conditional probability. It is only required that the censoring cumulative distribution function remains below 1 at τ.

We applied our method with prediction models constructed using several approaches. The first one is a direct application of survival ridge regression, using available gene expression data, under the assumption of proportional hazards with a linear link function. The second is similar to the approach followed in Kang, D'Andrea and Kozono (2012); we only use available gene expression data within the selective list proposed by the authors. Note that we do not attempt to reproduce their study; we follow a similar strategy. Also, in this case prediction models were derived using penalized maximum likelihood. Additionally, to these two approaches we also attempted the use of Kernel-based methods for estimating a smooth nonparametric link function [Li and Luan (2003)]. This produced results (i.e., values of the C_τ estimator) clearly inferior to the first two approaches.

Our goal is to show that the cross-study validation approach we present here facilitates methods comparison by estimating the easily interpretable μ latent variables. All the analyses were repeated separately under each method. Modeling of the Z array, in this example, produces an appreciable reduction of the uncertainty on the μ latent variables compared to the direct computation of the credible intervals by bootstrapping. All the model-based estimates of the μ latent variable are shrunk toward the average of the Z entries.

Figure 3 shows the observed validation statistics $Z^{110}_{s,v}$. As mentioned, these are empirical estimates of predictive performances adjusted for sample size variabil-

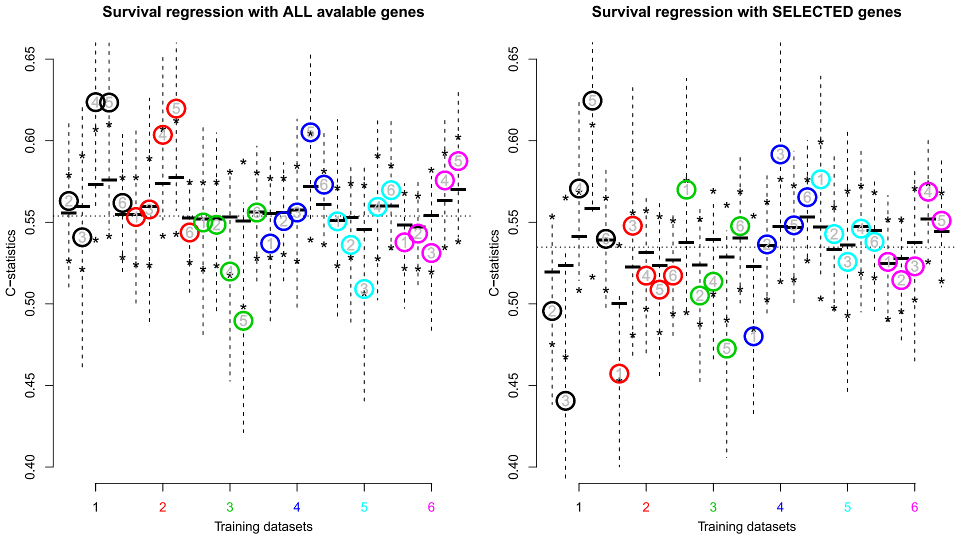

FIG. 3. *Validation analysis based on C_τ statistics. The left panel considers ridge regression based on all available gene expression data, while the right panel considers only a list of genes selected on the basis of the proposal in Kang, D'Andrea and Kozono* (2012)*. Each colored point illustrates a $Z^{110}_{s,v}$ validation statistic*; *colors indicate the training data set s while the integers in gray indicate the validation data set v. The "−" symbols indicate the corresponding Bayesian estimates $\mathbb{E}(\mu^{110}_{C(s),C(v)}|Z^{110})$. The dashed lines are* 80% *confidence intervals of the unknown means $\mathbb{E}(Z^{110}_{s,v})$ obtained by Bootstrapping. The "*" symbols indicate* 80% *credible intervals obtained from the posterior distribution of $\mu^{110}_{C(s),C(v)}$ given Z^{110}.*

ity. Each panel corresponds to one of the two approaches we compare, and colors indicate the training data sets, while the integers displayed in grey indicate the validation data sets. The plots show the 80% confidence intervals of the unknown means $\mathbb{E}(Z^{110}_{s,v})$ obtained by bootstrapping (dashed lines). They also display the model estimates (marked with the "−" symbol) of the $\mu^{110}_{C(s),C(v)}$ variables and the 80% credible intervals (marked with the "*" symbol). Under both approaches all $\mu^{110}_{C(s),C(v)}$ variables are estimated within the $(0.5, 0.6)$ interval. Our comparison suggests that models fitted after upfront selection of a subset of genes based on biological hypothesis perform worse than using all gene expression data. This indicates that genes other than those involved directly in DNA damage repair can contribute to explaining survival of ovarian cancer patients. All comparisons of the Bayesian estimates $\mathbb{E}(\mu^{110}_{C(s),C(v)}|Z^{110})$ under the two approaches are consistent with this evaluation. We also compared the posterior distributions of $\mu^{110}_{C(s),C(v)}$ under the two approaches; at each pair (s, v), when we sample from the posterior distributions we obtain inferior $\mu^{110}_{C(s),C(v)}$ values for the selective approach with probability above 0.67. If we use all genes, we obtain a probability of 0.78 that all $\mu^{110}_{C(s),C(v)}$ are larger than 0.5, meaning that all models perform better than assigning risk scores completely at random.

The posterior distribution of the latent partition of the 6 studies with sample sizes above 110 suggests the existence of two clusters, one including studies 4 and 5 and the other with all remaining studies. The estimate of the latent partition is identical under the two considered approaches for constructing predictive models but needs to be combined with relevant uncertainty. In particular, in both cases the partition constituted by a single degenerate subset with the six studies together accumulates posterior probabilities from both approaches above 0.15. When we added to the analysis the remaining three studies with sample sizes below 110, the resulting probabilities of the degenerate partition remained above 0.12. In summary, we found moderate evidence of a nondegenerate partition with heterogeneous subgroups.

In order to interpret the latent partition estimate of our leave-one-in analysis, we computed clustering-based validation statistics. Each solid line in Figure 4 is representative of a data set s and illustrates, for hypothetical sample sizes j from 100 to 600, estimates of how well outcomes in study s can be predicted by randomly selecting j data points within the cluster containing s. More formally, the y axis shows estimates of $Z^j_{B(s),s}$ with $j = 100, \dots, 600$. In this example the reported probabilities of the events $\sum_{v \in B(s) \setminus s} n_v \geq j$ are all above 0.6 for $j \leq 300$. These estimates are contrasted (dashed lines in Figure 4) with $Z^j_{\{1,\dots,S\},s}$. These are estimates of how well outcomes in s can be predicted by randomly selecting j data points from all available studies. We observe little difference between the solid and dashed lines. This similarity suggests that the clustering is not driven by heterogeneous data quality levels across studies (i.e., there is no evidence of clusters that

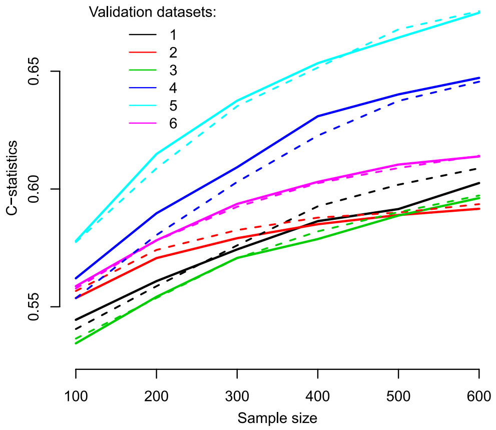

FIG. 4. *Sample size adjusted validation statistics. Solid lines display estimates of the clustering-based statistics $Z^j_{B(s),s}$ for values of j ranging between* 100 *and* 600. *Dashed lines display the validation statistics $Z^j_{\{1,\dots,S\},s}$, that is, the cluster $B(s)$ is replaced with the entire collection of* 9 *studies. Each color corresponds to a specific validation study s.*

produce prediction models with poor performance). Clustering is driven by studies 4 and 5, in which, due to covariates distributions, it appears relatively easier to achieve C-statistics above 0.6 compared to all other studies. Clustering-based statistics in Figure 4 suggest additional samples, above 600 and above the overall number of samples from the nine studies, might significantly contribute obtaining better prediction models.

For a comparison, we fitted the data sets with a hierarchical proportional hazards model, with studies clustered through a Dirichlet process, and Normal marginal priors for the regression coefficients. The prior assigns a vector of regression coefficients to each cluster of studies, while coefficients are independent across clusters. To facilitate the comparison, we tuned the Dirichlet process to match the estimate of the number of clusters in our leave-one-in analysis. We used the list of possible values for the latent partition and approximated the posterior using the approach discussed in Sinha, Ibrahim and Chen (2003). Our interest is in comparing the latent partitions obtained using the model just described to those from the validation analysis. If clustering in the leave-one-in analysis is driven by differences in the study-specific regression models, and not in the distributions of the predictors, then one expects the two approaches to infer similar partitions. Instead, the total variation distance between posterior distributions is 0.79, and we did not notice similarities. This is consistent with the interpretation of the partition inferred through the validation analysis that we discussed in the previous paragraph.

We also considered *random survival forests* for constructing prediction models; we used methodology and software discussed in Ishwaran et al. (2008). This method directly provides mortality scores for each sample in a test data set. Under several choices of the tuning parameters involved in the application of random forests, including minimal final nodes sizes [see Ishwaran et al. (2008) for details], the resulting predictive models appeared inferior to ridge regression when compared using C_τ statistics. Under all considered choices of the tuning parameters at least 66% of the $\mu^{110}_{C(s),C(v)}$ estimates were inferior to ridge regression. Contrasting results with random survival forests based on all available gene expression data versus the use of selected genes as suggested in Kang, D'Andrea and Kozono (2012), we obtained again higher $\mu^{110}_{C(s),C(v)}$ estimates using all available gene expression data.

6. An example of heterogeneous studies. Next we discuss cross-study validation of four nonsmall cell lung cancer studies recently reviewed in Ferté et al. (2013), based on the data sets curated by the authors. The data structure is similar to the previous example and includes gene expression predictors and patient survival times. We refer to Ferté et al. (2013) for a detailed description. The four studies and corresponding samples sizes are as follows: the Director's challenge study Shedden et al. (2008) (299), Zhu et al. (2010) study (62), Hou et al. (2010) study (79) and the TCGA [Hammerman et al. (2012)] study (90). This example emphasizes the necessity of accounting for study heterogeneity and that averaging the $Z_{s,v}$ statistics does not provide a complete description of models' performances.

The Director's study [Shedden et al. (2008)] includes data from 4 institutions. Our first analysis investigates whether there are large differences in the data originating from these institutions. The posterior of the model assigns probability 0.83 to the event that these 4 data sets are clustered together. Then, we considered the hypothesis of clusters involving the remaining data sets [Hammerman et al. (2012), Hou et al. (2010), Zhu et al. (2010)]. The approach is identical to the description in the previous section: we trained models using gene expression data and validated using concordance C_τ statistics. The posterior of Π produced an estimate of two clusters, one including the Director's study and the Zhu et al. study, and the second including the remaining two studies. The posterior strongly supports the hypothesis of separate clusters. The posterior probability at the estimated configuration $\hat{\Pi}$ is 0.44, and 0.48 posterior probability accumulates on the neighborhood $\{\Pi : l(\Pi, \hat{\Pi}) = 1\}$.

We finally compared sample size adjusted statistics $Z_{s,v}^{j}$ to interpret the clustering configuration. Figure 5 summarizes the main discrepancies visualized with these comparisons. On average, models fitted with $50 \leq j \leq 290$ samples from the Director's study tend to achieve substantially higher validation results when validated on the Zhu et al. study (blue line) than when validated on the remaining two studies (black lines). In the latter case the validation statistics decrease with training data set sample size, and the fitted models fail to predict survival times. We tested this difference using the bootstrap covariance estimates. The evaluation of prediction models produced by the largest study [Shedden et al. (2008)] changes

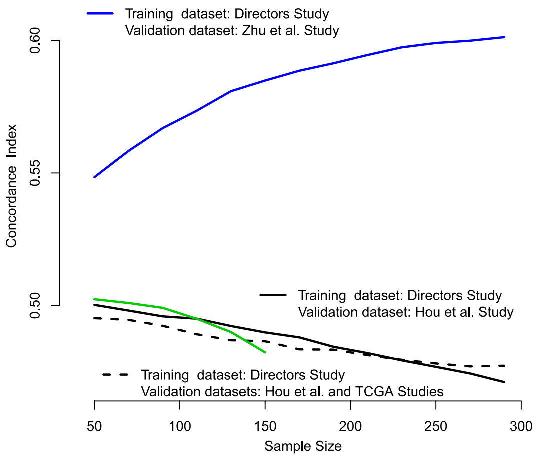

FIG. 5. *Sample size adjusted validation statistics for interpreting the clustering estimate $\hat{\Pi}$. The plot displays $Z_{s,v}^{j}$ validation statistics when the model is trained by the largest study [Shedden et al. (2008)] and validated on the remaining studies (black and blue lines). Additionally, it displays validation statistics when we train on the Hou et al. study and use the TCGA study for validation (green line). Prediction models have been trained using ridge regression.*

considerably if we only average the validation statistics across the three remaining studies, and it appears appropriate to report substantial discrepancies when we validate the Director's study results with the Zhu et al. study, versus validation on the Hou et al. and TCGA studies. While this is beyond the scope of our analysis, the next step is to investigate in depth the reason for these discrepancies.

7. Discussion.

Despite the availability of large collections of related data sets in many areas of application, articles that evaluate statistical learning algorithms based on a comprehensive analysis of available data sets remain a minority. Those using more than one data set are often based on cross-validation within each study due to heterogeneity between studies; see Demšar (2006), Japkowicz and Shah (2011) and Bernau et al. (2014) for discussions. Similar to meta-analyses for evidence synthesis, comprehensive model evaluations need to jointly consider study heterogeneity and algorithm performance. Here we propose a Bayesian approach to compare algorithms while incorporating relevant sources of uncertainty, including uncertainty on the comparability of independent studies.

The basis for our framework is the leave-one-in array Z of validation statistics. The concept is applicable to any validation statistic, such as concordance indices, classification errors and distances between predicted and observed responses. While it is certainly possible, and very useful, to simply use the leave-one-in array as a visualization tool without further modeling, our experience with evaluating genomic signatures in cancer suggests that modeling can substantially enhance interpretability of the leave-one-in analysis. Modeling addresses study heterogeneity, can prevent erroneous interpretations driven by sampling variability in the summary statistics, can help address multiplicity issues, and can formalize the process of identifying outlying studies requiring separate consideration. The analysis of the Z array helps interpreting the range of observed cross-study validation statistics, whether it is caused by differences in the study-specific distributions P_s or it reflects sampling variability.

Our two-stage procedure is based on a single figure of merit Z: this choice is motivated by the need for a simple strategy and by the consideration that this still accomplishes the main goal to control sources of overoptimism such as over-fitting, selection of favorable training/testing combinations and the use of internal cross-validations when the studies at hand are heterogeneous. Use of a one-dimensional figure of merit can, however, be a limitation. For example, if two studies generated data of poor quality, perhaps because of errors during sample processing and data management, our algorithm would likely cluster them together, because they both fail to produce accurate predictions and generate similarly poor Z scores when used for validating candidate models. These two studies might still be different in important ways; for example, they may consider two different populations. From this perspective, additional summaries of the data and potentially additional analyses may be advisable to identify differences between studies.

When multiple studies are available, a natural direction is to combine them. Bayesian hierarchical models, for instance, have emerged as a very useful paradigm to borrow information across studies [Lindley and Smith (1972) and Morris and Normand (1992)]. The leave-one-in analysis is not intended to replace combined analyses, but to address a different question: cross-study replicability of prediction. We consider the evaluation of prediction methods using leave-one-in matrices an important complementary goal and, in some cases, a prerequisite to the construction of predictive models based on multiple data sources. The analysis of leave-one-in matrices can be used not only to compare prediction methods, but also to select the most appropriate prediction methods for a subsequent combined analysis. In a related application to ovarian cancer prognosis using gene expression profiles, we illustrate a case where we first use cross-study validation to quantify the extent to which existing prognostic algorithms can produce results that hold up across studies [Waldron et al. (2014)], and then proceed to develop new prognostic algorithms based on a combined analysis [Riester et al. (2014)].

One advantage of the leave-one-in approach is that it can be used to evaluate any prediction approach, including heuristic procedures for which it might be challenging to construct hierarchical extensions. The modeling complexity that comes with constructing joint models for multiple studies varies across fields. In some cases the algorithms are based on probabilistic models and multi-study extensions are possible. In others they are not, and might be based on heuristics or be very specific to the field of application. The complexity and problem-specific competence necessary for developing joint models for heterogeneous data sets are greater compared to the analysis of Z matrices for off-the-shelf methods.

To address study heterogeneity, we cluster studies with similar validation profiles through a latent partition. The computation of the posterior distribution of the latent partition is straightforward and is a direct application of established computational strategies for fitting Dirichlet mixture models. We refer to the supplementary material [Trippa et al. (2015)] for more details. Clustering sharpens the interpretation of the cross-study validation results by allowing one to explore the maps $B \to Z_{B,s}$, focusing on either the estimates $\hat{B}(s)$ or on those partitions that a posteriori appear consistent with the dispersion estimate $\hat{d}$ and the observed array Z.

A simple alternative to formal Bayesian clustering of data sets is a reordering of rows and columns of Z, by maximizing objective functions, to obtain high values of the validation statistics close to the matrix diagonal. While this is perhaps simpler than what we propose, it can be dangerous to interpret the $Z_{s,v}$ validation summaries without consideration of the associated sampling variability, and it is easy to introduce an optimistic bias with clusters obtained by optimizing intracluster validation statistics.

In this article we only considered external validation statistics, where training and testing are performed on separate studies. Alternatively, one could integrate internal cross-validation into our framework by adding a diagonal to the Z array, with entries consisting of within study cross-validation statistics. A drawback of standard cross-validation techniques in this context is that they may result in overly optimistic assessments [Bernau et al. (2014)].

In this work we compared learning algorithms by separate analyses of the resulting Z arrays, but a natural extension is the joint analysis of multiple Z arrays corresponding to competing algorithms. A similar discussion applies to consideration of multiple validation statistics at the same time. A separate refinement could seek a data-driven approach for selecting the thresholds described in Section 5.1 to correct for sample size differences across studies.

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian nonparametric cross-study validation of prediction methods" (DOI: 10.1214/14-AOAS798SUPP; .pdf). We discuss results for logistic regression, Poisson regression, proportional hazards models and support vector machine procedures in the supplementary material.

REFERENCES

ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937

BAGGERLY, K. A., COOMBES, K. R. and NEELEY, E. S. (2008). Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. *J. Clin. Oncol.* **26** 1186–1187.

BENTINK, S., BENJAMIN, H., RISCH, T., FAN, J., HIRSCH, M., HOLTON, K., RUBIO, R., APRIL, C., CHEN, J., ELIZA, W., LIU, J., CULHANE, A., DRAPKIN, R., QUACKENBUSH, J. and MATULONIS, U. (2012). Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer. *PloS ONE* **7** e30269.

BERNAU, C., RIESTER, M., BOULESTEIX, A.-L., PARMIGIANI, G., HUTTENHOWER, C., WALDRON, L. and TRIPPA, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30** i105–i112.

BERRY, D. A. (1990). A Bayesian approach to multicenter trials and metaanalysis. ERIC, E0325480.

BERRY, D. A. and CHRISTENSEN, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7** 558–568. MR0527491

BONOME, T., LEVINE, D., SHIH, J., RANDONOVICH, M., CINDY, P., BOGOMOLNIY, F., OZBUN, L., BRADY, J., BARRETT, J., BOYD, J. and BIRRER, M. (2008). A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res.* **68** 5478–5486.

BURR, D. and DOSS, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *J. Amer. Statist. Assoc.* **100** 242–251. MR2156834

THE CANCER GENOME ATLAS RESEARCH NETWORK (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474** 609–615.

CHARON, I., DENŒUD, L., GUÉNOCHE, A. and HUDRY, O. (2006). Maximum transfer distance between partitions. *J. Classification* **23** 103–121. MR2280697

CRIJNS, A. P. G., FEHRMANN, R. S. N., DE JONG, S., GERBENS, F., MEERSMA, G. J., KLIP, H. G., HOLLEMA, H., HOFSTRA, R. M. W., TE MEERMAN, G. J., DE VRIES, E. G. E. and VAN DER ZEE, A. G. J. (2009). Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS Med.* **6** e24.

DEMŠAR, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7** 1–30. MR2274360

DENŒUD, L. and GUÉNOCHE, A. (2006). Comparison of distance indices between partitions. In *Data Science and Classification* 21–28. Springer, Berlin.

DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7** 177–188.

DRESSMAN, H., BERCHUCK, A., CHAN, G., ZHAI, J., BILD, A., SAYER, R., CRAGUN, J., CLARKE, J., WHITAKER, R., LI, L., GRAY, J., MARKS, J., GINSBURG, G., POTTI, A., WEST, M., NEVINS, J. and LANCASTER, J. (2007). An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J. Clin. Oncol.*: *Official Journal of the American Society of Clinical Oncology* **25** 517–525.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

FERTÉ, C., TRISTER, A. D., HUANG, E., BOT, B. M., GUINNEY, J., COMMO, F., SIEBERTS, S., ANDRÉ, F., BESSE, B., SORIA, J.-C. and FRIEND, S. H. (2013). Impact of bioinformatic procedures in the development and translation of high-throughput molecular classifiers in oncology. *Clin. Cancer Res.* **19** 4315–4325.

GANZFRIED, B. F., RIESTER, M., HAIBE-KAINS, B., RISCH, T., TYEKUCHEVA, S., JAZIC, I., WANG, X. V., AHMADIFAR, M., BIRRER, M. J., PARMIGIANI, G., HUTTENHOWER, C. and WALDRON, L. (2013). curatedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. *Database* **2013** bat013.

GARRETT-MAYER, E., PARMIGIANI, G., ZHONG, X., COPE, L. and GABRIELSON, E. (2008). Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics* (*Oxford*, *England*) **9** 333–354.

HAMMERMAN, P. S., LAWRENCE, M. S., VOET, D., JING, R., CIBULSKIS, K., SIVACHENKO, A., STOJANOV, P., MCKENNA, A., LANDER, E. S. GABRIEL, S. et al. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489** 519–525.

HOOVER, D. N. (1982). Row-column exchangeability and a generalized model for probability. In *Exchangeability in Probability and Statistics* (*Rome*, 1981) 281–291. North-Holland, Amsterdam. MR0675982

HOU, J., AERTS, J., DEN HAMER, B., VAN IJCKEN, W., DEN BAKKER, M., RIEGMAN, P., VAN DER LEEST, C., VAN DER SPEK, P., FOEKENS, J. A., HOOGSTEDEN, H. C., GROSVELD, F. and PHILIPSEN, S. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE* **5** e10312.

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796

JAPKOWICZ, N. and SHAH, M. (2011). *Evaluating Learning Algorithms*: *A Classification Perspective*. Cambridge Univ. Press, Cambridge.

KANG, J., D'ANDREA, A. D. and KOZONO, D. (2012). A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *J. Natl. Cancer Inst.* **104** 670–681.

KONSTANTINOPOULOS, P., SPENTZOS, D., KARLAN, B., TANIGUCHI, T., FOUNTZILAS, E., FRANCOEUR, N., LEVINE, D. and CANNISTRA, S. (2010). Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *J. Clin. Oncol.*: *Official Journal of the American Society of Clinical Oncology* **28** 3555–3561.

LEE, J., QUINTANA, F. A., MÜLLER, P. and TRIPPA, L. (2013). Defining predictive probability functions for species sampling models. *Statist. Sci.* **28** 209–222. MR3112406

LI, H. and LUAN, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pac. Symp. Biocomput.* 65–76.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B. Stat. Methodol.* **34** 1–41.

MOK, S., BONOME, T., VATHIPADIEKAL, V., BELL, A., JOHNSON, M., WONG, K.-K., PARK, D., HAO, K., YIP, D., DONNINGER, H., OZBUN, L., SAMIMI, G., BRADY, J., RANDONOVICH, M., CINDY, P., BARRETT, J., WONG, W., WELCH, W., BERKOWITZ, R. and BIRRER, M. (2009). A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: Microfibril-associated glycoprotein 2. *Cancer Cell* **16** 521–532.

MORRIS, C. N. and NORMAND, S. L. (1992). Hierarchical models for combining information and for meta-analyses. *Bayesian Stat.* **4** 321–344.

QUINTANA, F. A. and IGLESIAS, P. L. (2003). Bayesian clustering and product partition models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 557–574. MR1983764

RIESTER, M., WEI, W., WALDRON, L., CULHANE, A. C., TRIPPA, L., OLIVA, E., KIM, S.-H., MICHOR, F., HUTTENHOWER, C., PARMIGIANI, G. et al. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J. Natl. Cancer Inst.* dju048.

ROY, D. and TEH, Y. (2009). The mondrian process. *Adv. Neural Inf. Process. Syst.* **21** 27.

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134. MR0600538

SHEDDEN, K., TAYLOR, J. M. G., ENKEMANN, S. A., TSAO, M.-S., YEATMAN, T. J., GERALD, W. L., ESCHRICH, S., JURISICA, I., GIORDANO, T. J., MISEK, D. E. et al. (2008). Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat. Med.* **14** 822–827.

SINHA, D., IBRAHIM, J. G. and CHEN, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika* **90** 629–641. MR2006840

SWISHER, E., TANIGUCHI, T. and KARLAN, B. (2012). Molecular scores to predict ovarian cancer outcomes: A worthy goal, but not ready for prime time. *J. Natl. Cancer Inst.* **104** 642–645.

TOTHILL, R., TINKER, A., GEORGE, J., BROWN, R., FOX, S., LADE, S., JOHNSON, D., TRIVETT, M., ETEMADMOGHADAM, D., LOCANDRO, B., TRAFICANTE, N., FEREDAY, S., HUNG, J., CHIEW, Y., HAVIV, I., GROUP, A. OC. S., GERTIG, D., ANNA, D. and BOWTELL, D. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* **14** 5198–5208.

TRIPPA, L., WALDRON, L., HUTTENHOWER, C. and PARMIGIANI, G. (2015). Supplement to "Bayesian nonparametric cross-study validation of prediction methods." DOI:10.1214/14-AOAS798SUPP.

UNO, H., CAI, T., PENCINA, M. J., D'AGOSTINO, R. B. and WEI, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30** 1105–1117. MR2767845

WALDRON, L., PINTILIE, M., TSAO, M.-S., SHEPHERD, F. A., HUTTENHOWER, C. and JURISICA, I. (2011). Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics* **27** 3399–3406.

WALDRON, L., HAIBE-KAINS, B., CULHANE, A. C., RIESTER, M., DING, J., WANG, X. V., AHMADIFAR, M., TYEKUCHEVA, S., BERNAU, C., RISCH, T., GANZFRIED, B. F., HUTTENHOWER, C., BIRRER, M. and PARMIGIANI, G. (2014). Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J. Natl. Cancer Inst.* **106** dju049.

WARN, D. E., THOMPSON, S. G. and SPIEGELHALTER, D. J. (2002). Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Stat. Med.* **21** 1601–1623.

YOSHIHARA, K., TAJIMA, A., YAHATA, T., KODAMA, S., FUJIWARA, H., SUZUKI, M., ONISHI, Y., HATAE, M., SUEYOSHI, K., FUJIWARA, H., KUDO, Y., KOTERA, K., MASUZAKI, H., TASHIRO, H., KATABUCHI, H., INOUE, I. and TANAKA, K. (2010). Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. *PloS ONE* **5** e9615.

ZHU, C.-Q., DING, K., STRUMPF, D., WEIR, B. A., MEYERSON, M., PENNELL, N., THOMAS, R. K., NAOKI, K., LADD-ACOSTA, C., LIU, N., PINTILIE, M., DER, S., SEYMOUR, L., JURISICA, I., SHEPHERD, F. A. and TSAO, M.-S. (2010). Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J. Clin. Oncol.* **28** 4417–4424.

L. TRIPPA
G. PARMIGIANI
BIOSTATISTICS AND COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE
450 BROOKLINE AVE.
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: ltrippa@jimmy.harvard.edu
gp@jimmy.harvard.edu

L. WALDRON
CUNY SCHOOL OF PUBLIC HEALTH
AT HUNTER COLLEGE
2180 3RD AVE ROOM, 538
NEW YORK, NEW YORK 10035
USA
E-MAIL: levi.waldron@hunter.cuny.edu

C. HUTTENHOWER
BIOSTATISTICS DEPARTMENT
HARVARD SCHOOL OF PUBLIC HEALTH
BUILDING 1 #413
655 HUNTINGTON AVENUE
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: chuttenh@hsph.harvard.edu