Received: 10 May 2023 | Revised: 18 December 2023 | Accepted: 2 January 2024

DOI: 10.1002/sim.10012

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Statistical plasmode simulations–Potentials, challenges and recommendations

Nicholas Schreck[1] | **Alla Slynko**[2] | **Maral Saadati**[1] | **Axel Benner**[1]

[1]Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

Correspondence
Nicholas Schreck, Division of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.
Email: nicholas.schreck@dkfz-heidelberg.de

Statistical data simulation is essential in the development of statistical models and methods as well as in their performance evaluation. To capture complex data structures, in particular for high-dimensional data, a variety of simulation approaches have been introduced including parametric and the so-called plasmode simulations. While there are concerns about the realism of parametrically simulated data, it is widely claimed that plasmodes come very close to reality with some aspects of the "truth" known. However, there are no explicit guidelines or state-of-the-art on how to perform plasmode data simulations. In the present paper, we first review existing literature and introduce the concept of statistical plasmode simulation. We then discuss advantages and challenges of statistical plasmodes and provide a step-wise procedure for their generation, including key steps to their implementation and reporting. Finally, we illustrate the concept of statistical plasmodes as well as the proposed plasmode generation procedure by means of a public real RNA data set on breast carcinoma patients.

KEYWORDS

data-generating process, outcome-generating model, parametric simulations, resampling, statistical plasmodes

1 | INTRODUCTION

Data availability is a crucial issue that arises in the context of statistical model development and validation, inference derivation, introduction of statistical concepts and many others.[1-3] In some cases, especially for high-dimensional data (HDD) where the number of features is substantially larger than the number of observations, the number of data samples and data sets is not large enough to properly and reliably perform all required tasks. To overcome that deficiency, alternative data generation approaches such as the generation of artificial data are required. The generated data should match as close as possible the real-life data underlying the research question of interest, in particular with respect to its probabilistic structure as well as possible dependencies.

A common approach for data generation is statistical data simulation. In this paper, we interpret data simulation as a data generation procedure that follows a data-generating process (DGP), with marginal distributions and a dependence structure as its basic components. This paper focuses on explanatory and prediction models, including propensity-based models. For these, data generation procedures should necessarily include steps on the generation of outcome variables.

Nicholas Schreck and Alla Slynko are equally contributing authors.

To this end, probabilistic outcome-generating models (OGMs) are usually utilized to generate the outcomes using the available covariate information. Parameters of the OGM are to be estimated from real data, taken from literature or just set based on investigator's choice. Examples for outcome data generation can be found in Reeb and Steibel,[4] Franklin et al,[5] Schulz et al,[6] Atiquzzaman et al[7] and many others. Shmueli[8] depicts the application of OGMs both for explaining and prediction purposes.

For simulated data, the "truth" is assumed to be known a priori, at least to a some extent.[9] That "truth" can be represented by simulation parameters or prespecified effect sizes, and is used to reliably evaluate the obtained estimates or predictions.

The most detailed practical introduction to simulation studies that includes structural approaches for their planning and reporting as well as the discussion on the appropriate performance measures is presented in Morris et al.[2] In particular, the authors introduce a coherent terminology on simulation studies, data-generating mechanisms, and provide guidance on coding simulation studies. Most prominently, they introduce the aims, data-generating mechanisms, estimands, methods, and performance (ADEMP) criteria as a guidance for the planning and performing of simulation studies.

For those researchers who use the results of simulation studies without being familiar with the entire simulation process, the discussion presented in Boulesteix et al[10] can be of great assistance. The paper not only contains many useful examples and applications, but also describes basic principles of simulation studies, gives insights into sampling variability and data generating processes, and demonstrates the role of statistical simulations in health research.

A systematic comparison of the performance of statistical methods can also be made using a benchmarking data set.[11,12] The comparison is based on a realistic but unknown DGP where the underlying truth might be partially known or suspected. However, good benchmarking data sets are rarely available, results are specific and contextual, and comparisons are mainly limited to prediction performance. This results in the need to generate data, as we focus on in this paper.

Two common approaches for data generation are parametric and plasmode simulation, with the first approach being the most extensively studied and widely used one. Parametric simulations assume that the parametric stochastic model used to generate data is realistic and representative, with parameters of interest estimated from real data, derived from the literature or even set up by the user, in order to model specific scenarios.[1,2] Plasmode data generation usually begins by resampling covariate information from the original real data.[13,14] External (parametric) "truth" such as effect sizes or model parameter values in explanatory or prediction models can then be added to the covariable data sets to define the relationship between the covariables and the outcome. With the OGM being a part of the plasmode data generation procedure, the resulting plasmode simulation can then be viewed as a semi-parametric data generation procedure. Of note, parametrically simulated data may often be considered to be purely artificial, whereas plasmode data is claimed to reflect reality in the most close way.[9,13,15]

In the present paper, we provide an extensive literature overview of parametric and plasmode simulations. We introduce the concept of *statistical plasmodes* and discuss its differences to biological plasmodes. We address advantages and challenges of parametric and statistical plasmode simulation approaches in various contexts and provide step-by-step recommendations for the generation of statistical plasmodes.

At this point, we have no intention to demonstrate the superiority of a plasmode simulation over a parametric simulation or vice versa. We aim to analyze advantages and challenges of both data generation methods, and to illustrate the usefulness of plasmode simulations as a complement to and possible extension of parametric simulation studies.

The paper is organized as follows: Section 2 discusses parametric and plasmode simulations, compares their characteristics and provides an extensive literature review on both data generation methods. Section 3 analyzes statistical plasmodes in more detail by discussing their challenges. Section 4 provides recommendations for planning, performing and reporting of statistical plasmode simulations. In Section 5, we present a numerical example to illustrate the application of such a data generation approach. Section 6 concludes with a discussion.

2 | FROM PARAMETRIC TO PLASMODE SIMULATION STUDIES

This section provides a comparative introduction to parametric and plasmode simulations. In particular, we start with a description of the main properties of parametric simulations as one of the most well-established types of data generation. Then we move on to plasmode simulations which are often claimed to be a more close-to-reality approach for data generation.

2.1 | Parametric data simulation

In cases where an underlying DGP is defined in closed form and represented by a parametric stochastic model, we speak of parametric simulations.

The main asset of parametric simulations is their flexibility in terms of the chosen DGP. That is, one can easily generate a variety of independent data sets by varying the assumptions imposed on the DGP and its crucial parameters. As a result, the simulated data sets can cover many complex but relevant scenarios as well as extreme situations that do not reflect reality. This feature becomes particularly important whenever we aim to analyze the behaviour and performance of different statistical methods. In addition, the knowledge of the DGP makes the corresponding parametric simulation more transparent and plausible.

Parametric simulations depreciate the sample size issue as an unlimited amount of data can be generated by means of a particular DGP. For instance, when applied to simulation of continuous random variables, an "infinite" number of distinct data points can be generated without much effort.

The corresponding DGP represents a cornerstone of any parametric simulation. However, the existence of an appropriate model for the DGP that best fits some underlying real data cannot be taken for granted. On the other hand, even with a DGP available, the conclusions based on parametric simulations might be limited or even biased by the parameters of the chosen DGP model.

Obviously, the quality of a parametric simulation depends on the level of our comprehension of the underlying processes, distributions and possible dependencies. For instance, Vaughan et al[15] state that simulations might not fully reflect the complexity of the biological data that originates from nonrandom mating, recombination, hot spots, and other genetic mechanisms. Boulesteix et al[10] provides a similar statement and claims that many simulation studies are too simplified to describe the complexity of the real life data and thus may lead to inaccurate or even misleading findings.

When generating new data, we strive to preserve not only the marginal distributions but also the underlying dependence structure. In this context, the specification of appropriate dependence metrics, such as correlation, emerge. One of the options for such a specification is estimation of the dependence structure from the data at hand. Computationally, such an estimation can be very expensive and time-consuming, especially in case of large data sets. For parametric simulations, this computational issue is one of possible reasons for making independence assumptions on certain variables that then leads to a block diagonal structure for the corresponding correlation matrix.[13] Furthermore, a multivariate normal distribution provides the simplest model for multivariate covariate data with a pre-specified mean and correlation structure.[1] For generating non-normal correlated data, diverse copulas or the extended Fleischman power method can be utilized.[16] Obviously, certain concerns about the accuracy and realism of the underlying modelling assumptions emerge for all such approaches.

One of the assumptions imposed in the context of parametric simulations is that the "truth" must be known a priori. Such an assumption may not hold in both low- and high-dimensional situations. However, high-dimensionality may even exacerbate this issue making parametric simulations inapplicable. For instance, the "truth" about the set of biological markers truly associated with a given outcome may be unknown.[17]

Within the framework of a parametric simulation, the underlying dependence structure has to be completely specified in advance. The specification of such a dependence structure for high-dimensional data may become a challenge. Possible reasons can be not only computational efficiency issues, but also spurious correlations,[18] sparsity of the data, some nonlinear or even hidden dependencies and other issues related to large covariance matrices; for more discussion and examples see Fan and Li,[19] Johnstone and Titterington[20] and the references therein. Pitfalls in the specification of the dependence structure may then lead to false research discoveries and incorrect statistical inferences.

Dimensionality reduction and feature extraction play pivotal roles and are often fundamental in many high-dimensional settings.[19] However, it is not obvious how a parametric simulation may impact the findings of those procedures considering possible non-representativeness of parametrically generated covariate data sets.

Altogether, in many cases parametric simulations may turn out to be infeasible for high-dimensional data generation as it is not obvious how such simulations would cope with features of high-dimensionality.

A number of papers share our concerns on the applicability of parametric simulations for the generation of high-dimensional data sets. For instance, Gadbury et al[14] question the applicability of standard simulations performed in a high-dimensional experiment where hundreds of hypotheses are to be tested. Also Franklin et al[5] sees the application of ordinary simulation methods as an issue when comparing high-dimensional variable selection strategies. In particular, the authors point out that the performance of those strategies depend "[…] on the information richness and complexity

of the underlying empirical data source" and it is doubtful whether a parametric simulation is able to capture the richness and complexity of the information.

To overcome, at least partially, the limitations of parametric simulations outlined above, plasmode simulations have been introduced.[5,13,15]

2.2 | Plasmode data simulation

The term "plasmode" has been first introduced in Cattell and Jaspers,[21] with a plasmode data set defined as "[…] a set of numerical values fitting a mathematico-theoretical model". In their seminal work, the authors emphasized that the certainty for the produced plasmode data set to fit the model comes either because there is a real life experiment producing the data of that kind or because the simulated data is produced mathematically to fit the functions. Two different approaches for plasmode generation, performed either in a lab experiment or by resampling, were also mentioned in Mehta et al.[13] In the present paper, we refine the discussions provided by these authors, introduce the concept of statistical plasmodes and analyze their properties. From our perspective, the classification of plasmodes in biological and statistical depends on the procedure used for their generation.

Biological plasmodes are those generated "[…] by natural biological processes, under experimental conditions that allow some aspects of the truth to be known".[15] Such plasmodes may be created, for example, in a wet lab by manipulating biological samples as in case of a "spike in" experiment.[13,15] The latter paper provides a very illustrative introduction to biological plasmodes. In their detailed definition, the authors state that "[…] a plasmode can be defined as a collection of data that (i) is the result of a real biological process and not merely the result of a computer simulation; (ii) has been constructed so that at least some aspect of the "truth" of the DGP is known".

A number of research papers such as Mehta et al,[9] Vaughan et al[15] deal with "spike in" experiments in microarray expression analysis as an example for a biological plasmode data set. As part of that experiment, real cases from one population are randomly assigned to two groups. Then, a known amount of transcript is added to serve as a positive control. As a result, distributions and correlations in the generated data are viewed as most realistic since being taken directly from real data. Besides others, Mehta et al[13] discusses application of plasmode data sets in high- dimensional biology. Vaughan et al[15] use plasmodes for the estimation of admixture, or the proportion of an individual's genome that originated from different founding populations and thus illustrates the utility of plasmodes in the evaluation of statistical genetics methodologies. Several authors such as Sokal et al[22] and Mehta et al[13] provide helpful insights into generation and application of biological plasmodes as those are expected to incorporate valuable information on biological variation and capture biological reality. Biological plasmode data sets have also been utilized to evaluate the performance of statistical methods[23] and their validity.[9] Plasmode data sets were also used to investigate the validity of multiple factor analysis in a known biological model.[22]

Despite their ability to create new and more advanced biological set-ups, for example, by crossing mice,[15] sometimes biological plasmodes may become not only very time-consuming but also require high experimental costs. Researchers might eventually do not have a lab available to construct biological plasmodes. In some cases, ethical reasons may also speak against the construction of biological plasmodes. In all such situations, *statistical plasmodes* offer an advantageous alternative.

Statistical plasmodes, being in focus of the present paper, begin with generation of covariate information performed by applying resampling-based methods to a real data set, see, for example, Tibshirani,[24] Reeb and Steibel,[4] Franklin et al;[5] note that no biologically new samples are created in the context of such resampling procedure. Further, an appropriate OGM has to be applied to generate outcomes based on the resampled covariates.[5,7,25-34]

Statistical plasmodes have often been utilized in causal (propensity-based) methods using weighted regression models. The application of these weighted regression models is often related to the presence of a complex covariance structure.[31] To calculate the weighted population, exposure (or treatment) modeling is also a part of the study,[5,25,26,29,31] and some known "truth" such as exposure effects can be added manually.[5,7,28,30,31] For instance, in Franklin et al[5] the authors create statistical plasmode data sets by "[…] resampling from the observed covariate and exposure data without modification to preserve the empirical associations among the variables." In that paper, the "true" treatment effect and the baseline hazard function are estimated from the empirical data. Furthermore, the baseline hazard has been modified to guarantee a desired rate of events in a certain time interval. In addition, the associations between event times and the covariates as well as between censoring times and the covariates have been defined by means of two Cox proportional hazard models. Such a modeling approach corresponds to the application of an OGM in our terminology.

According to our interpretation, statistical plasmode simulation utilizes aspects of resampling (when generating the covariate information) as well as parametric modeling (eg, application of OGM, modeling of exposure etc.) and thus can be interpreted as a semi-parametric method.

There are numerous applications of plasmodes generated by certain methods of data modifications in the literature. Some of those applications, including explanatory, causal and prediction studies, are based on statistical approaches in the sense of our definition. For instance, Tibshirani[24] utilize plasmodes to assess sample size requirements in microarray experiments when estimating the false discovery rate and false negative rate for a list of genes. Gadbury et al[14] illustrates use of plasmodes by comparing the performance of 15 statistical methods for estimating the false discovery rate in data from an high-dimensional experiment. Elobeid et al[35] employs plasmode data sets to analyze the performance of several statistical methods used to handle missing data in obesity randomized controlled trials. Reeb and Steibel[4] suggest an interesting application of plasmode data sets to complement the evaluation of statistical models for RNA-seq data. In their subsequent paper, Reeb et al[36] then use plasmode data sets to assess dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data. In Franklin et al,[5] plasmode-based studies are used for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Resampling in combination with outcome generation by a logistic model to compare the HDD propensity score method with ridge regression and lasso is used by Franklin et al.[25] Franklin et al[26] use plasmode-based studies to compare the performance of propensity score methods in the context of rare outcomes. In Desai et al,[28] the authors utilize plasmode data sets to analyze the uncertainty in using bootstrap methods for propensity score estimation whereas Liu et al[30] conduct a plasmode-based study to compare the validity and precision of marginal structural models estimates using complete case analysis, multiple imputation, and inverse probability weighting in the presence of missing data on time-independent and time-varying confounders. The issue of data imputation has also been addressed in Atiquzzaman et al[7] where the authors used plasmodes to compare two imputation techniques when imputing body mass index variable in osteoarthritis-cardiovascular disease relationship. In Ejima et al,[37] the authors use statistical plasmodes to assess type I and type II error rates of analyses commonly used in murine genetic models of obesity. Similarly, Alfaras et al[38] resample from the empirical distributions to create plasmode data sets for murine aging data. Those plasmodes are then utilized to compute type I error rates and power for commonly used statistical tests without assuming a normal distribution of residuals. In their most recent study, Hafermann et al[33] design a plasmode simulation study to investigate how random forest and machine learning methods may benefit from external information provided by prior variable selection studies. Rodriguez et al[34] evaluate plasmodes as being useful for preserving the underlying dependencies among hundreds of variables in real-world data used to evaluate the potential utility of novel risk prediction models in clinical practice; the authors generate plasmodes when studying lung transplant referral decisions in cystic fibrosis.

To our understanding, two central steps in the *statistical plasmode* generation procedure can be derived, namely:

(i) **Generation of the covariate data** by
  (i.1) Resampling from an original data set
  (i.2) Artificial covariates (treatment, exposure, etc.) by a parametric model
(ii) **Outcome generation** that includes
  (ii.1) Choice of an appropriate outcome generating model (OGM)
  (ii.2) Choice of covariate effects either by individual specification or by estimation based on the original data
  (ii.3) Generation of new outcomes by drawing from the OGM chosen in (ii.1), with the effects specified in (ii.2), applied to the covariate data generated in (i)

The discussion performed in this section is summarized in Table 1 that provides a comparative summary of parametric and plasmode simulation studies. That discussion as well as the supporting literature imply that plasmodes provide an attractive supplement to parametric simulations in data-based research. In particular, it is expected that plasmode data sets resemble the reality most closely, especially regarding the covariate dependence structure. In the following, we will analyze plasmodes to examine their strengths and weaknesses in more detail.

3 | CHALLENGES OF STATISTICAL PLASMODE SIMULATIONS

Simulations studies are, at least in the scope of the present work, designed to enable the practical analysis of statistical methods. To this end, data generation should satisfy several criteria, such as, amongst others, to provide the basis for

TABLE 1 Parametric simulations versus statistical plasmodes: Similarities and differences.

Feature	Parametric simulations	Statistical plasmodes
Data-generating process (DGP)	DGP is to be specified in advance	No DGP specification is required
Outcome-generating model (OGM)	Parameters of a chosen OGM to be estimated from data or derived from literature or set manually	Parameters of a chosen OGM to be estimated from the original data or derived from literature or set manually
Range of possible scenarios	Arbitrary scenarios, in particular, extreme and rare scenarios can be generated	Only reality bounded to the sample at hand can be generated
Knowledge of "truth"	"Truth" must be completely known in advance	At least some "truth" such as effect sizes should be known a priori
Data availability and representativeness	Irrelevant for simulations based on literature results or previous knowledge	Crucial, as the simulated data is always limited to the sample at hand
Reality reflection	Parametric simulations may not be able to capture the complexity of real life data	Plasmodes are expected to resemble the reality in the most accurate way
High-dimensional data simulations	Usually time- and cost-consuming. Latent dependencies may also become an issue	Mostly straightforward, as no estimation of distributions and/or dependencies is required
Small sample sizes	Essentially uncomplicated, but may become an issue in cases when simulation parameters are to be estimated from the real data at hand	Difficult due to resampling
Dependence structure	Becomes a challenge with complex dependencies	No modeling/estimation of dependence structure is required

subsequent undistorted model comparisons or to enable a specific covariate dependence structure. Constructing, reporting and comprehending a parametric simulation study is mostly straightforward and transparent, as the resulting data is artificially generated in a target-oriented way. Critical steps in the construction of parametric simulations include, for instance, the investigator's choice of the outcome-covariable association. While this ambiguity is shared by statistical plasmode simulations, many of the properties of statistical plasmodes are typically less obvious and verifiable because statistical plasmodes are designed with the complex task to mimic reality in the closest way while simultaneously specifying some aspects of the truth. The main advantage of statistical plasmodes lies in their ability to generate data with specific distributions and dependence structures without the need for explicit assumptions. The assumption that statistical plasmodes can faithfully generate data that closely resemble reality has rarely been questioned. For instance, the lack of statistical analyses or simulations to verify the preservation of dependence structures can undermine the reliability of the generated data. Consequently, the advantages attributed to statistical plasmodes can also transform into challenges. Further potentially critical steps in the construction of plasmode data include the representativeness of the underlying data and the choice of the resampling scheme. Below, we theoretically discuss these potential pitfalls in more detail while also providing corresponding examples from literature.

3.1 | Resampling of covariate information

In our concept of statistical plasmodes, the simulation is based on the generation of covariable information by resampling from a real data set. This has the intention to preserve the characteristics of the original underlying data set such as, amongst others, the number and type of covariables and the corresponding dependence structure, see for example, Franklin et al,[25] Atiquzzaman et al,[7] Conover et al.[31] This preservation is primarily achieved through the use of appropriate resampling techniques. Consequently, the applied resampling scheme, which consists of specifying the number of generated data sets (N) and the resampling technique, has central importance for the generated plasmode (covariable) data sets.

Of note, while utilizing resampling, statistical plasmodes are arguably even more complicated to analyze because of the additional artificial outcome generation. Consequently, not all established theoretical results concerning resampling might be transferable to the full plasmode data set but only to the plasmode covariable data sets. We use the terminology resampling and bootstrap interchangeably and indicate the concrete resampling/bootstrapping technique if necessary. The analysis of resampling methods is almost exclusively formulated in terms of the asymptotic performance of the bootstrap distribution L^* of an estimator T (eg, variance, confidence interval) applied to the empirical distribution of the resampled data.[39] For statistical plasmode covariable data sets, the estimator T could be, for example, some function of the covariance matrix of the covariables (preservation of correlation structure). When considering the statistical plasmode procedure as a whole, the estimator of interest T typically utilizes the artificial outcomes, for example, T could be the linear predictor in ridge regression when investigating its performance compared to other models. The resampling is said to "have worked", if L^* convergences weakly to L (the theoretical distribution of T) for increasing sample size n of the underlying data.[40] Otherwise, one speaks of "bootstrap failure" which implies that the bootstrap estimator should not be trusted because it does not provide the correct value even asymptotically. This could imply that some characteristic of the population which we want to preserve, for instance some complex association structure, might in fact not be preserved in the bootstrap samples. In the following, we discuss the influence of the chosen resampling scheme on the generated plasmodes in more detail, focusing mainly on the preserveness of the covariable information.

3.1.1 | Number of plasmode data sets N

The specification of the number of resampled plasmode data sets N is often performed ad-hoc and potentially leads to different answers to the same question, in particular if N is specified as too small.[41] In the framework of bootstrap tests, Davidson and MacKinnon[42] propose a pretest procedure for choosing the number of bootstrap samples to minimize the loss of power due to N being finite. A more general, data-dependent three-step procedure is proposed by Andrews and Buchinsky[41] who estimate N to achieve a desired accuracy of the approximation of the bootstrap to the ideal ($N \to \infty$) distribution of the estimator of interest. However, to the best of our knowledge, there is no general guideline to theoretically specify the number N of data sets to be generated in a data-independent way (ie, without already performing the resampling scheme) such that asymptotic resampling results hold with sufficient accuracy. Moreover, existing results might not be valid for statistical plasmodes due to the additional artificial outcome-generation procedure.

In the plasmode literature, $N = 500$[5,7,28] and $N = 1000$[29,30,33] seem to be popular ad-hoc choices. We have not seen any application where the choice of N was explicitly justified or the convergence or stability of the subsequent analyses applied to the plasmode data have been checked for increasing N. In summary, the number of data sets can be a critical aspect in the generation of *statistical plasmodes*, in particular if convergence of T is not reached. In Section 4 we provide some recommendations for determining N which we further illustrate in Section 5.

3.1.2 | Resampling technique

Resampling can be performed without replacement as in the n-over-m bootstrap (subsampling with $m < n$) and sample-splitting (cross-validation) bootstrap. Subsampling draws from the data-generating process of the original data[43] and has been shown to lead to consistent estimators under minimal conditions, see theorem 1 in Bickel et al,[39] as long as the subsampling size m and the size of the original data set n are appropriately specified. Alternatively, resampling with replacement such as the n-out-of-n bootstrap (also called nonparametric bootstrap) can be utilized. Resampling schemes based on drawing with replacement draw from the empirical probability distribution derived from the underlying data[43] and require additional assumptions for consistent estimation, but are more efficient if the bootstrap "works".[39] However, the nonparametric bootstrap can fail, for example when the limiting distribution of the estimator has discontinuities, when estimating extremes and when setting critical values for some test statistics.[39,44,45] In that case, the nonparametric bootstrap estimator should not be trusted to provide asymptotically correct results. As a remedy, the m-out-of-n bootstrap (sampling $m \leq n$ with replacement) has been introduced to prevent bootstrap failure while losing efficiency if the nonparametric bootstrap was consistent. Sampling fewer than n observations has since been treated as a "cure-all" method (being asymptotically valid under weak assumptions, not failing and hence showing less asymptotic problems) which has been critically discussed, for instance, in Andrews and Guggenberger.[45] A comprehensive overview of resampling techniques is provided, for example, in Bickel et al.[39]

For increasing sample size n, estimators based on subsampling and the m-out-of-n bootstrap become more similar as the probability of repeating observations decreases. Note that, contrary to subsampling, resampling with replacement allows for $m = n$. The additional requirements for the consistency of sampling with replacement compared to sampling without replacement mainly state, informally speaking, that the influence of tied observations on the bootstrap estimator should be small.[40]

In the majority of literature concerning plasmode generation, the m-out-of-n bootstrap has been used,[5,27,32] whereas the nonparametric bootstrap has been used by Rodriguez et al,[34] the sample-split bootstrap by Gerard[46] and subsampling by Hafermann et al.[33] In some publications we did not find indications whether resampling was performed with or without replacement, for example, in Ju et al,[47] Ripollone et al.[29] All in all, the type of resampling technique influences asymptotic properties of estimators including the covariables and hence the plasmode data sets, effects whether the resampling "has worked" (ie, whether estimators or predictions can be trusted) and consequently impacts subsequent analyses on the generated plasmodes data sets. However, to the best of our knowledge, we have not seen any application in the literature concerning plasmode data generation in which the choice of a particular resampling technique has been explicitly justified.

3.1.3 | Resampling size m

Using resampling with replacement of size m, with $m \to \infty$ and $m/n \to 0$, typically resolves failure of the n-out-of-n bootstrap, but requires the specific choice of m as a key issue.[40] An adaptive rule for the choice of m for subsampling and the m-out-of-n bootstrap in the case of independent observations has been proposed by Bickel and Sakov[40] and is further illustrated in our example in Section 5. Informally speaking, if m is in the right range of values, the bootstrap distributions of the estimator for similar m's are close to each other, indicating consistency of the estimator. The rule provides an adaptive estimator $m^*(n)$ and leads to optimal convergence rates of the estimator irrespective whether the nonparametric bootstrap would work in the example (then $m^*(n)/n \to 1$ as $n \to \infty$) or would fail (then $m^*(n)/n \to 0$).

To the best of our knowledge, only fixed resampling sizes m have been chosen in the plasmode literature, and we have not observed any explicit justification of the specific value of m. In other words, m appeared to be chosen arbitrarily. For instance, Hafermann et al[33] used a selection of $m's$ (250, 500, 1000, 2000, 4000) which are small compared to the number of observations $n = 198,895$ while Liu et al.[30] chose $m = 500$ for a data set with $n = 646$ and Atiquzzaman et al[7] sampled $m = 75,000$ out of $n = 84,452$. Interestingly, different authors used different values of m which have been chosen without justification for the same underlying data set, as exemplified for the NSAID data set with $n = 49,653$. While $m = 30,000$ has been picked by Franklin et al[5] and Ripollone et al[29] used a comparably large $m = 25,000$ as well, Ju et al[47] chose a much smaller value in $m = 1,000$ and Wyss et al[32] set $m = 10,000$. In summary, when applying subsampling or the m-out-of-n bootstrap, the value of m matters for the consistency of T, and should be properly justified and adapted to the underlying data and estimator(s) of interest.

3.1.4 | Covariable dependence structure and HDD

Resampling the covariable information has, amongst others, the aim of preserving the covariable dependence structure of the underlying data set; see for example, Franklin et al,[5] Karim et al,[27] Conover et al.[31] Under some assumptions such as i. i. d. observations and finite fourth moments of the covariables, Beran and Srivastava[48] have shown that the resampled covariance matrix converges to the original covariance matrix for the nonparametric bootstrap when n increases and the number of covariables p is fixed (ie, most HDD situations excluded, see also below). However, for other resampling schemes similar results have, to the best of our knowledge, not been shown. For the m-out-of-n bootstrap, the optimal m could be estimated with the estimator T specified to represent the covariable covariance matrix in order to investigate and ensure that the resampling scheme works (at least for that certain aspect of the data), see also our example in Section 5. However, other aspects of the covariable information, such as extreme values, might be more important in some applications.

A well-discussed issue, in particular in the context of HDD, is the occurrence of spurious correlations. Amongst others, Fan et al[18] have shown that sampling p independent normal n-vectors leads to empirical covariance structures strongly deviating from a diagonal matrix, in particular if $p \gg n$. However, the risk of spurious correlations is not limited to parametric simulations. An increasing number of covariates p increases the risk that the underlying data sample suffers from

spurious correlations, which may be propagated to the generated plasmode data sets by resampling. Further, spurious correlation is likely to distort the empirical covariance structure of the statistical plasmodes, leading to even stronger deviations from the population covariance matrix.

For a fixed number of covariables p, the bootstrap has been shown to work in linear models if p/n is small.[49] If the number of covariables growths with the number of observations, Mammen[50] has shown that the bootstrap works for effect estimates in high-dimensional linear models if $p(n) \to \infty$ and $p(n)/n \to 0$ as $n \to \infty$, and Karoui and Purdom[51] have shown that the confidence intervals of the pairs and residual bootstrap in linear models are too wide if $p(n)/n \to c, c \in \mathbb{R}$.

In summary, the goal of preserving the covariable dependence structure can be used to determine an optimal resampling scheme. In particular in high-dimensions, the covariance structure could, however, be distorted by spurious correlations and whether resampling and subsequently statistical plasmodes work in these scenarios might require additional research.

3.2 | Representativity of the underlying data sample

One of the main assets attributed to statistical plasmode simulations is that they are expected to preserve the complex real-world data structure by resampling the covariable information from a real data set. Naturally, an appropriate representative data set has to be available and constitutes the basis for the entire plasmode simulation study. Parametric simulations, on the other hand, can be artificially constructed without requiring representative data. The data sample is expected to represent the population of interest. This limits the generalizability of the results of the analyses that the plasmode simulation study was designed for, which has been acknowledged, amongst other, by Franklin et al,[5] Liu et al,[30] Atiquzzaman et al.[7]

The data sample should satisfy the assumptions of the applied resampling technique. As a result, the choice of the resampling technique depends strongly on the underlying data set at hand. Standard resampling techniques, such as discussed above, assume that the observations are independent.[39] This assumption is violated, for instance, if the observations show clusters, repeated measures, population structure or longitudinal measurements. In this context, more sophisticated resampling schemes including block-wise resampling have to be applied, for which most of the asymptotic results are not explicitly formulated.[40]

Depending on the underlying data and the resampling scheme, the characteristics of the original data set to be conserved might not be reflected by the generated data. This is acknowledged by Karim et al[27] who state that "[…] it is possible that important confounders in the empirical study might not remain important in the plasmode samples".

In summary, the generated statistical plasmode data sets depend strongly on the representativeness of the underlying real data and are limited to the population represented by the data sample. The resampling scheme should be adaptive to the characteristics of the real data such as population structure, which have to be identified and reported.

3.3 | Investigator's choice of the "truth"

The concept of plasmode simulations is mainly based on preserving the complex but realistic structure of the underlying data while inserting some "truth" by investigator's choice. This specification can be manifold in type, and potentially distort the real-word characteristics of the generated plasmode data.

3.3.1 | Artificial covariables

Additional to resampling covariable information, important covariables such as exposure or treatment variables have been artificially created using an exposure (or treatment) generating model. This is in particular useful when investigating causal models where the influence of (high-dimensional) confounders on the exposure and the outcome is to be modeled. For instance, Franklin et al[26] and Conover et al[31] model a binary exposure variable in relation to confounder variables via logistic regression (exposure-generating model), while Rodriguez et al[34] simulate covariables at a later stage of a longitudinal study. In particular, it is possible to control the exposure prevalence which could be important in

the area of causal inference, for example, for propensity-based methods. For instance, Franklin et al,[26] Conover et al[31] modified the intercept in the OGM for the binary exposure to generate simulation setups with a variety of exposure prevalences. Naturally, artificially generating covariables does not preserve the full real-world setting and should be performed with care.

3.3.2 | Artificial outcome generation

With the covariable information generated (by resampling or artificially), corresponding artificial outcomes are created. In our concept of statistical plasmodes this is mainly done according to some outcome-covariable association specified by the investigator. A straightforward way to create a transparent association between resampled covariates and the artificial outcomes is to utilize regression models specified by the combination of a link function (type of OGM) and the linear predictor (effect structure).

The OGM determines the type of the artificial outcome (eg, binary, survival) and strongly influences subsequent results of the analyses of the generated data. For instance, if the aim of the study is the performance assessment of several models, the model closest to the chosen type of OGM has an advantage induced by the investigator, leading to potentially distorted comparisons. Most commonly, logistic regression is used for binary outcomes[7,26,27] and the Cox model for survival endpoints,[5,25,28] whereas Rodriguez et al[34] apply an exponential survival model. Normal linear regression is used by Liu et al,[30] and Conover et al[31] use a Poisson model for the number of events.

Besides the type of OGM, the determination of the effect structure of the corresponding linear predictor is vital. Elements of the effect vector have been specified by literature review,[31] by sampling from independent standard normal distributions,[47] by estimation on the original data set[30,32] or manually by investigator's choice.[5,25,28] Some authors specify the treatment or exposure effect by hand while estimating the confounder effects on the original data.[26,27] In propensity-based methods, effects of confounders on exposure or treatment can be set in the exposure generating model while the effect of the same confounders on the outcome can be set in the OGM. Specifying the value of the effects of covariables might represent a strong intervention in the generation process of a realistic data set. Potential problems include creating artificial outcome-covariate associations and invalidating or even nullifying existing "real" associations between the covariables and the novel, artificial outcomes, in particular if the effects are set manually by investigator's choice. Also, the effect size alone does not fully describe the relevance of a covariate for the exposure or the outcome. When considering the coefficient of determination, for instance, the variances of and the covariances between covariates strongly influence the explanatory potential in terms of variance. Hence, manually picking effects might lead to an unrealistic association structure in terms of explanatory potential. If the effects are estimated, they depend on the underlying data sample and estimation uncertainty is ignored. Additionally, the specification of the type of OGM influences effect estimates and subsequent analyses might become problematic. For example, an effect vector estimated by a sparse method will induce advantages of sparse methods in model comparisons on the generated data.

Outcomes on the scale of the observations might then be generated by drawing from the probabilistic family corresponding to the OGM. In the context of logistic regression, for instance, the event rate is typically drawn from a Bernoulli variable with probability determined by the link function applied to the calculated linear predictor (in the respective plasmode covariate dataset).[25,26]

By specifying the type of the OGM and the effect structure we might be able to control outcome rates, for example, in the case of a binary outcome such as response to treatment. For instance, Franklin et al[25] and Franklin et al[26] specify the probability of the outcome generating Bernoulli distribution via the estimated OGM. A variety of simulation scenarios were generated by setting the prevalence of the outcome to desired values by modifying the intercept in the linear predictor of the OGM. For time-to-event endpoints, the estimated baseline hazard function can be adjusted in order to approximately control the event rate in a specified time period.[5]

In summary, a crucial assumption for both parametric and plasmode simulations is that the chosen outcome generation reflects realistic, natural or biological associations between outcome and covariables. Whereas the choice of the OGM and the effect structure is a natural aspect of parametric simulation studies, plasmodes are often described as closely depicting reality. However, the outcome data in statistical plasmodes are also artificially created while inducing some investigator's choice "truth". These manipulations of the real-data are harder to assess and less transparent than in parametric simulations studies as the structure of the data is typically more complex. In the end, plasmode generation also leads, at least in part, to artificial data and constructed associations.

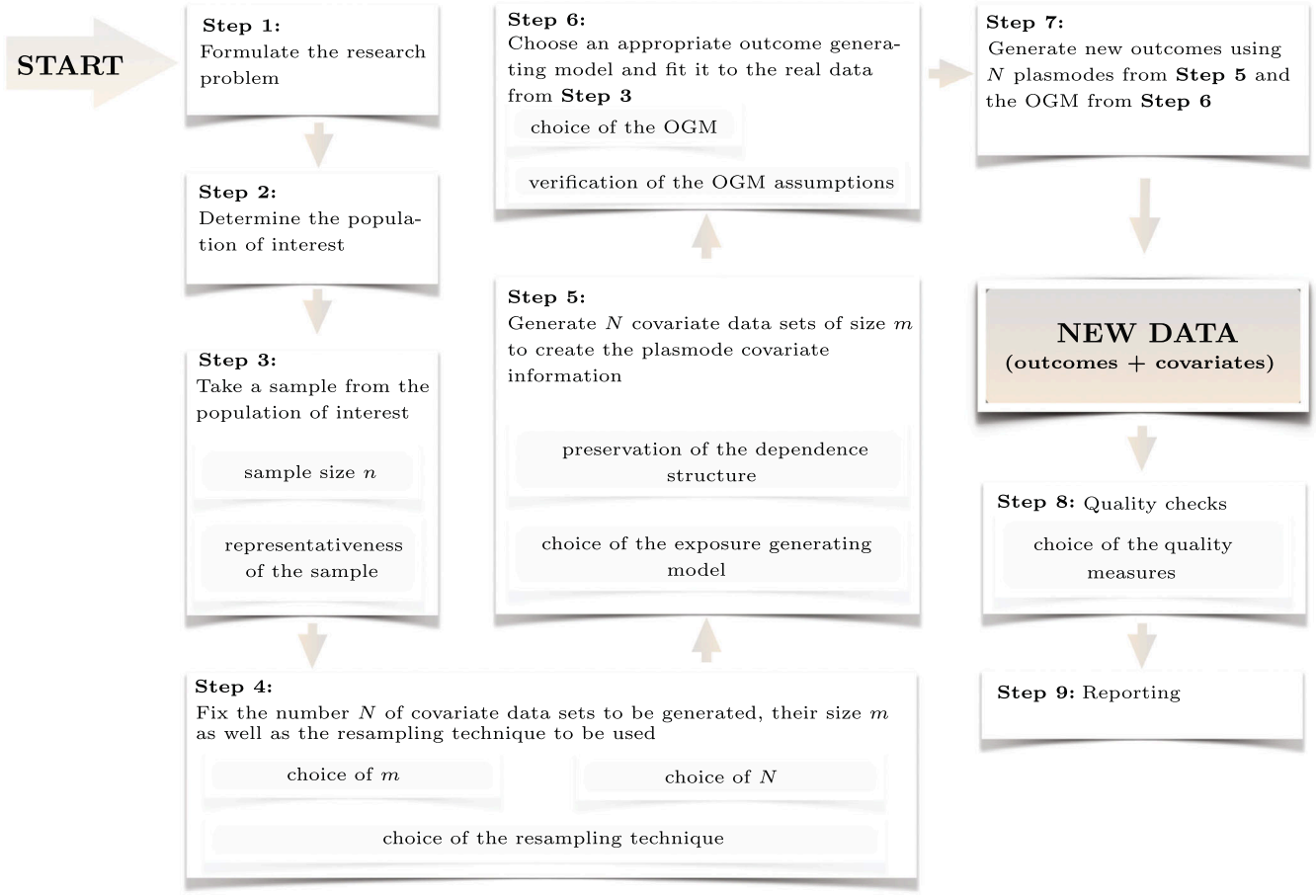

FIGURE 1 Statistical plasmode data generation procedure step-by-step.

4 | STATISTICAL PLASMODES: STEP-BY-STEP RECOMMENDATIONS

We provide a hands-on overview of our recommendations for the generation and reporting of our concept of statistical plasmodes in Figure 1. This summary extends the basic plasmode generation procedure described in Section 2 and addresses the critical steps discussed in the previous section. We theoretically discuss our step-by-step procedure below and illustrate its application in a real data example in Section 5.

4.1 | Step 1: Planning of the simulation study

We recommend to clearly formulate the research problem and to plan the simulation study using the ADEMP criteria.[2] In a first step, researchers should specify the general aims of the simulation study, for example, whether the authors are concerned with specific properties of estimators such as precision or efficiency or prediction performance of competing models. Choosing a data-generating mechanism, either explicitly using parametric models or by resampling, might be the most time-consuming part, depends on the specified aims and mainly influences the applicability and generalizability of the results of the study. The estimand might describe a population parameter to be estimated or targets concerned with method evaluation such as prediction performance or model selection. In the design of the study it is required to specify the methods to be investigated, each of which should be appropriate and aim for the same estimand. Naturally, all relevant methods for the determined estimand should be included in the study. The performance measures are used to evaluate the methods and should correspond to the identified estimand. Morris et al[2] recommend that the relevance of the performance measures should be justified, that the formulae should be stated, and that uncertainty estimates for the estimate of the performance measures should be provided. Additionally, the implementation of the performance measures might be helpful. Fixing the aims and the data-generating processes aids in the choice whether statistical plasmodes

are needed in the first place or whether a parametric simulation study might be more appropriate. Additionally, it guides the choice of the population of interest in Step 2 and potentially the choice of the resampling technique in Step 4. Importantly, the methods and performance measures indicate the subsequent choice of the OGM in Step 6. For instance, if we plan to assess the prediction performance of several models, we should make sure that the chosen OGM does not bias the subsequent model comparisons. Also, the choices of the OGM and the DGP determine the scenarios in which the properties of novel statistical methods can be empirically assessed in the context of the plasmode simulation study.

4.2 | Step 2: Population of interest

The population of interest can be of primary interest and consequently be strongly connected to the aim of the simulation study and the DGPs determined in the previous step, in particular if methods are developed to deal with populations with certain characteristics (eg, many missing values, complex covariance structure, high-dimensionality). However, we might also be mainly interested in the analysis of statistical methods such that the population of interest serves primarily as an illustration and is not necessarily connected to Step 1. In the latter case, particular effort should be taken clarify why the chosen population covers those situations in which the methods under consideration are claimed to work. In any case, the hypothetical population should be stated and described as clearly as possible, for example, the entity of interest, covariables and population structure. Since the data sets generated by the statistical plasmodes should be representative of the population of interest, this step influences several of the following steps, in particular Steps 3, 4 and 6.

4.3 | Step 3: Representative sample

It is a central aspect of statistical plasmodes that the underlying sample is representative of the population of interest clarified in the previous step, refer also to the discussion in Section 3.2. Consequently, it is vital to investigate and communicate why the utilized data sample represents the population of interest and which potential limitations arise. In particular, it should be stated how the data was sampled, which covariables are included and what the endpoint of interest is. Additionally, the sample size should be justified and potential population structures investigated, as this can influence the choice of the resampling technique, see Steps 4 and 5 and the discussion in Section 3.1. Note that even if the sample is representative, the generated plasmode data sets might not be, for example, as a result of a poor resampling plan or a outcome generation that distorts either the relationship between the covariables or the outcome-covariable association.

4.4 | Step 4: Resampling scheme

The resampling scheme consists of the number of bootstrap samples, the type of resampling technique used and, if applicable, the justification of the resampling size, see also Section 3.1. It determines, together with the data sample, the plasmode covariate data sets. Each of the aspects of the resampling scheme plays a crucial role for the asymptotic properties of the estimators applied to the data generated by resampling and to properties such as the preserveness of the covariable correlation structure, see the discussions in Section 3.1. Unfortunately, the resampling scheme has to be decided on for each application individually, while keeping those research aims and properties of the population of interest, that should be preserved with high priority, in mind. Ensuring the plasmode sets are drawn from the hypothetical population is only possible by applying subsampling,[43] whereas sampling with replacement draws from the empirical distribution of the data sample specified in Step 3. However, if the underlying data set is representative of the population of interest, drawing with replacement might become preferable due to its increased efficiency and second-order properties.[39] As described in Section 3.1, the nonparametric bootstrap potentially fails, although this is often impossible to know before the application. To avoid bootstrap failure, it is often recommended to utilize the m-out-of-n bootstrap although this might lead to efficiency losses. The optimal resampling size m can be determined by applying the algorithm introduced in Bickel and Sakov[40] while using those properties of priority as estimator in Step 2 of the optimization algorithm for m. It has to be noted that the estimation of m might require high additional computational cost. The algorithm requires to use an estimator of some characteristic that we want to be able to estimate in a consistent way based on our simulated datasets. In many applications it might be meaningful to opt for some function of the covariate covariance structure as an estimator, as it is often stressed that the empirical dependence structure of the original data set should be preserved. This

might, however, lead to the nonparametric bootstrap (optimal m equal to n) at least for low-dimensional situations as discussed in Section 3.1. The choice of an appropriate measure for the specification of the subsampling size m for statistical plasmode simulations is beyond the scope of this paper and remains part of future research.

4.5 | Step 5: Plasmode covariate information

After determining the details of the resampling scheme in the previous step we can perform resampling of the covariate information. Reproducibility should be enhanced by setting and reporting seeds and by making the resampling scheme publicly available. Artificial covariate information might be created by parametric simulations or by using an exposure-generating model as discussed in Section 3.3. We refer to the set of resampled and artificial covariate information as the plasmode covariate information.

4.6 | Step 6: Outcome generating model

The choice of the OGM includes the type of probabilistic model to determine the association between the plasmode covariables information and the novel artificial outcomes, as well as corresponding OGM components such as, for example, effect sizes. The OGM determines the artificial outcomes in type and value, and is a crucial component for many research questions formulated in Step 1. Also, it gives the investigator the opportunity to fix some aspects of the "truth", see also the discussion in Section 3.3. Special care should be taken that the OGM does not bias the subsequent analyses that the statistical plasmode simulations are generated for. To do so, it might be helpful to investigate the models or methods to be compared in detail and contrast them with the OGM. For instance, a sparse OGM will most likely support sparse models in subsequent model comparisons. If the effect structure is chosen in a sparse way, a sparse model might be more likely to correctly estimate the effect sizes or perform valid predictions. Additionally, if important relationships between variables have been detected, the effect structure should be chosen accordingly to preserve these. For instance, in linear predictor models, the observed outcome variation depends on the (co-)variances of the covariables weighted by their corresponding effects, stressing their influence on the artifical outcomes of the plasmode data sets.

4.7 | Step 7: Outcome generation

Each of the N plasmode covariate data sets sampled in Steps 4 and 5 is combined with the OGM determined in Step 6 to create N corresponding artificial plasmode outcome vectors. Depending on the type of model it might be necessary to sample from the probabilistic model corresponding to the OGM. For instance, if we consider logistic regression as the OGM then the natural link function (logit-link) applied to the calculated linear predictor results in a probability vector. The final outcomes might then be generated by drawing from Bernoulli distributions based on the derived probabilities. Note that this introduces a parametric source of variability, additionally to the variability introduced by resampling the covariates.

4.8 | Step 8: Quality checks

The quality of the covariables can be assured by appropriate resampling as described in Steps 4 and 5. It is, however, often not feasible to compare the original covariable covariance structure with those of the N statistical plasmode data sets. More research might be necessary to judge the distance of the original and the generated data. The original outcome values of the real data set are, if at all, only explicitly used to determine the effect structure in Step 6. The quality of the simulated data could be checked by comparing the generated outcomes of (some of) the statistical plasmode data sets with the original outcome. The type of potentially meaningful checks depends on the type of outcome. For continuous observations, the distributions of the two outcomes could be compared by the empirical densities or histograms as is done for example, in Franklin et al.[5] Additionally, the range of the data should be checked as well as potential outliers. For categorical (including binary) outcomes, the prevalence of the classes can be compared, see for example Franklin et al.[5]

4.9 | Step 9: Reporting

Within each step of the statistical plasmode generation every investigator's decision should be justified and reported to enhance reproducibility and transparency of the proposed data generation procedure. Additionally, the programming code including the seeds used to generated the statistical plasmodes should be made readily available to further increase transparent research. Whenever appropriate, we recommend that plasmode generation follows the scheme presented in Figure 1 and the corresponding descriptions provided in the present section. Finally, the results for the research question determined in Step 1 should be addressed.

5 | STATISTICAL PLASMODES: A NUMERICAL EXAMPLE

The following example has been constructed to illustrate the step-by-step procedure introduced in the previous section. Reproducible code is available as supplementary material.

Step 1. Assume that we are interested in the aim (A) of investigating the application of ridge regression[52] and the linear mixed model[53] in the context of high-dimensional RNA-expression data with sparse effects on a normal outcome (data-generating process, D). The estimands (E) are specified as the parameter vector and the predictions of the respective model implying that we are both concerned with the models explanatory (inferential) potential as well as their prediction accuracy. We split the sample once into training and test data (2:1), which we deem sufficient for our illustration purposes. The plasmode data sets are generated using the training data. Each plasmode data set of size m and number of covariates p is analysed (methods, M) using ridge regression of the form

$$y = \mu 1_m + X\beta + \varepsilon, \quad \|\beta\|_{L_2}^2 \leq \lambda, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_{m\times m}) \tag{1}$$

via penalized maximum likelihood with cross-validation for λ as implemented in the R-package `glmnet`,[54] as well as the linear mixed model in the variance components form

$$y = \mu 1_m + X\beta + \varepsilon, \quad \beta \sim \mathcal{N}(0, \sigma_\beta^2 I_{p\times p}), \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_{m\times m}) \tag{2}$$

with restricted maximum likelihood estimation as implemented in the R-package `sommer`.[55] Here, 1_m denotes the m-column vector of ones while $I_{p\times p}$ denotes the identity matrix of dimension p. As performance measure (P) for the explanatory potential and the accuracy of the methods we utilize the mean absolute bias

$$\text{MAB} = \frac{1}{p+1}\|(\hat{\mu}, \hat{\beta}) - (\mu, \beta)\|_{L_1} \tag{3}$$

where μ and β are known as part of the "truth". Naturally, a variety of performance measures, such as mean-square error for the precision of the estimators, might be investigated as well. Also, cross-validated or bootstrapped MSEP instead of a sample-split approach might be applied. As a performance measure (P) for the validity of the methods we use the sample-split mean squared error of prediction

$$\text{MSEP} = \frac{1}{m}\|\hat{y} - y_{\text{test}}\|_{L_2}^2, \quad \hat{y} = \hat{\mu} 1_m + X_{\text{test}}\hat{\beta} \tag{4}$$

where y corresponds to the artificial outcome in the test split. We estimate both measures using the mean of the estimates (indexed by superscript b) in the generated N statistical plasmode data sets

$$\widehat{\text{MAB}} = \frac{1}{N}\sum_{b=1}^{N}\frac{1}{p+1}\|(\hat{\mu}, \hat{\beta})^{(b)} - (\mu, \beta)\|_{L_1}, \quad \widehat{\text{MSEP}} = \frac{1}{N}\sum_{b=1}^{N}\frac{1}{m}\|\hat{y}^{(b)} - y_{\text{test}}\|_{L_2}^2. \tag{5}$$

and visualize the N individual measures via boxplots, see Step 9 and Figure 5.

Step 2. In the scope of this example, we are interested in the model choice for high-dimensional RNA-expression data with normal outcomes for female breast cancer patients which constitutes the population of interest.

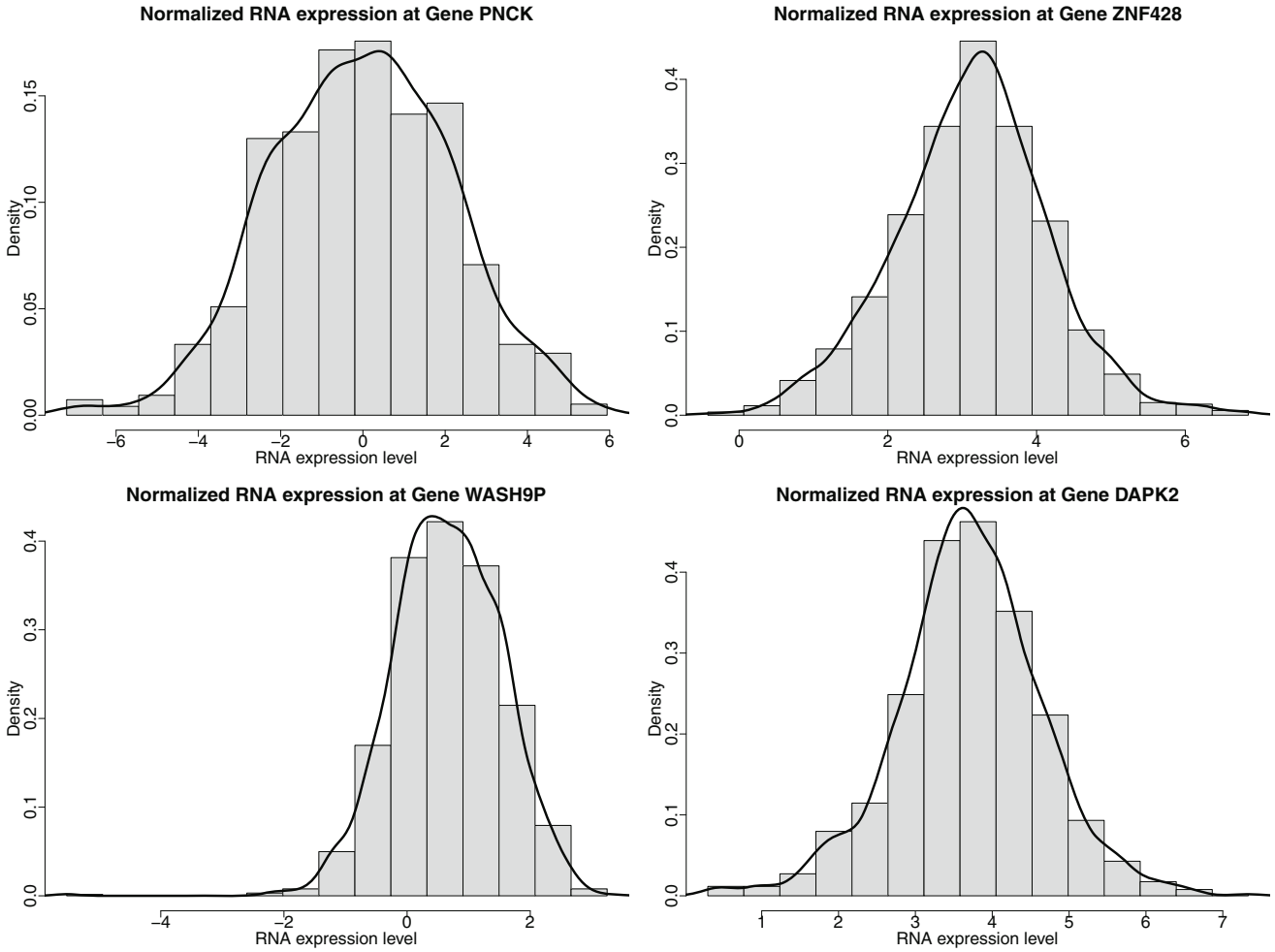

FIGURE 2 Empirical distributions illustrated by histograms (15 breaks each) and smoothed densities for four genes selected at random to illustrate the differences in location and shape.

Step 3. The data sample underlying the statistical plasmode simulation was generated by The Cancer Genome Atlas (TCGA) Research Network (https://www.cancer.gov/tcga). The breast carcinoma (BRCA) cohort which provides a basis for the following numerical example was last updated on May 31, 2016.

We restrict the publicly available data to $n = 1,098$ female patients with breast cancer with cancer tissue, excluding normal tissue and male patients. RNAseqV2 gene expression data and clinical data for BRCA were obtained from the TCGA Data Portal[56] via the R/Bioconductor package `TCGAbiolinks`.[57-59] For computational reasons, we choose $p = 5,000$ out of the 25,828 available genes at random for this illustration. Naturally, this introduces additional selection variability and a sensitivity analysis should be conducted if we move beyond illustration purposes. The R/Bioconductor package `limma`[60] has been utilized to normalize the RNA gene expression data. The expression levels can be assumed to be measured continuously and they show different shapes and ranges. This is illustrated in Figure 2 using their empirical distributions at four randomly chosen genes.

The outcome of interest is age at diagnosis date which can be considered to be approximately normally distributed, see Figure 3A. While the data set can be considered to be representative of a female breast cancer population from the United States of America, we acknowledge that RNA expression data from other populations (eg, different countries) might lead to different results for our research question.

Step 4. Before the analysis, we set the number of plasmode data sets to be generated to $N = 500$. In the final Step 9, we investigate the convergence of the estimators of the performance measures, see Equation (5), in the statistical plasmode data sets.

We choose the m-out-of-n bootstrap in order to prevent potential bootstrap failure but with the potential drawback of losing estimation efficiency. Performance analysis of resampling method and the estimation of the optimal m requires the

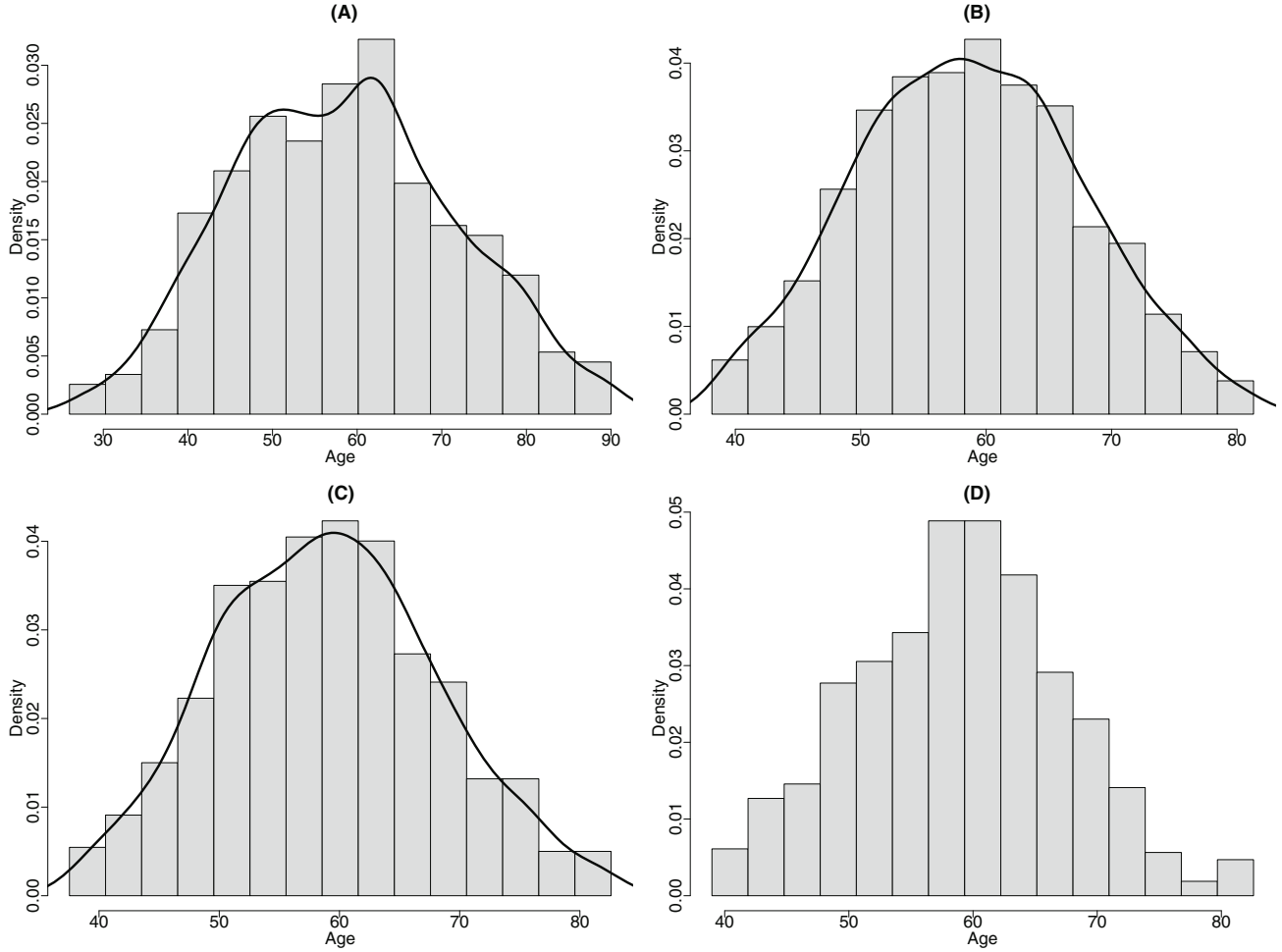

FIGURE 3 Empirical distributions illustrated by histograms (15 breaks each) and smoothed densities for (A) the original outcome (age at diagnosis) and (B–D) versus artificial outcomes of three plasmode data sets selected at random. (A) Original outcomes (age at diagnosis). (B) Artificial outcome plasmode dataset 1. (C) Artificial outcome plasmode dataset 2. (D) Artificial outcome plasmode dataset 3.

specification of an estimator which is applied to the generated data. Since resampling in statistical plasmodes is primarily concerned with the covariate information, already using the performance measures defined in Equation (5) as estimators is not feasible as they require the subsequent artificial outcome generation. Naturally, there are several reasonable estimators that could be considered. In this example, we opt for the covariate dependence structure as the measure of interest because the majority of publications which applied plasmodes referred to the advantage of the preserveness of the original covariable dependence structure.

We determine the resampling size m via the algorithm described in Bickel and Sakov.[40] In particular, to adopt that algorithm to our problem formulation, we specify the sequence of potential m's by setting $q = 0.97$, choose the L_2-norm of the covariance matrix (of the resampled covariate data) as a metric, and calculate the resulting empirical distribution functions. We estimate the covariance matrix using the Ledoit-Wolf linear shrinkage estimator[61] to obtain a more precise estimate which is necessary because the covariate data are high-dimensional. The optimal resampling size m^* is the one which minimizes the distance between the distributions of subsequent m's, where the distance is exemplarily measured by the Wasserstein metric. The optimal resampling size based on the Wasserstein metric using 100 iterations resulted in $m^* = 732$, hence $m^* = n_{\text{train}}$. Hence, we effectively apply the nonparametric bootstrap which is a special case of the m-out-of-n bootstrap.

We acknowledge that there is variety of optimal resampling sizes m^* if any of the parameters of the algorithms would be changed (such as, amongst others, estimator, distance metric for empirical distributions and sequence of potential m's).

Step 5. We apply resampling with replacement of size $m^* = 732$ to the matrix of covariable information to obtain $N = 500$ statistical plasmode covariable data sets. As we have determined the resampling size with optimality criterion as

the L_2-norm of the covariance matrix of the covariables, the empirical covariance structure of the original data set should be sufficiently preserved.

Step 6. We choose the LASSO[62] as an appropriate OGM to represent the sparse effect structure associated with the high-dimensional data as required in Step 1. Additionally, the LASSO most likely does not distort the comparison between ridge regression and the linear mixed model as both of these methods are shrinkage methods used to model polygenic effects. The "true" effect structure for the LASSO is chosen as the vector of estimated effect sizes obtained after a LASSO had been fit to the original data. The proportion of non-zero estimated effects was 93.56%. This implies that 322 covariables are selected in the investigator's choice "truth" while 4,678 genes are given a null effect. The effect sizes of the selected covariables has a median of −0.0196 (range [−1.0320, 1.6389]).

Step 7. We generate one artificial outcome vector of size $m^* = 732$ for each of the $N = 500$ plasmode covariate data sets by calculating the linear predictor based on the combination of the resampled covariable information (Step 5) and the "true" effects (Step 6). Thus, we obtain statistical plasmode simulations based on real covariate information with an investigor's choice "truth". Note that the normal linear model is the only representative in the class of linear predictor models which explicitly includes a random variable representing an error term. Hence, drawing from the probabilistic linear model to generate outcomes we might also add normally distributed random variables with expectation zero and variance equal to, for example, the estimated residual variance in the full dataset or some investigator's choice. Additionally to the variability in the artificial outcome induced by resampling the covariates this would represent additional

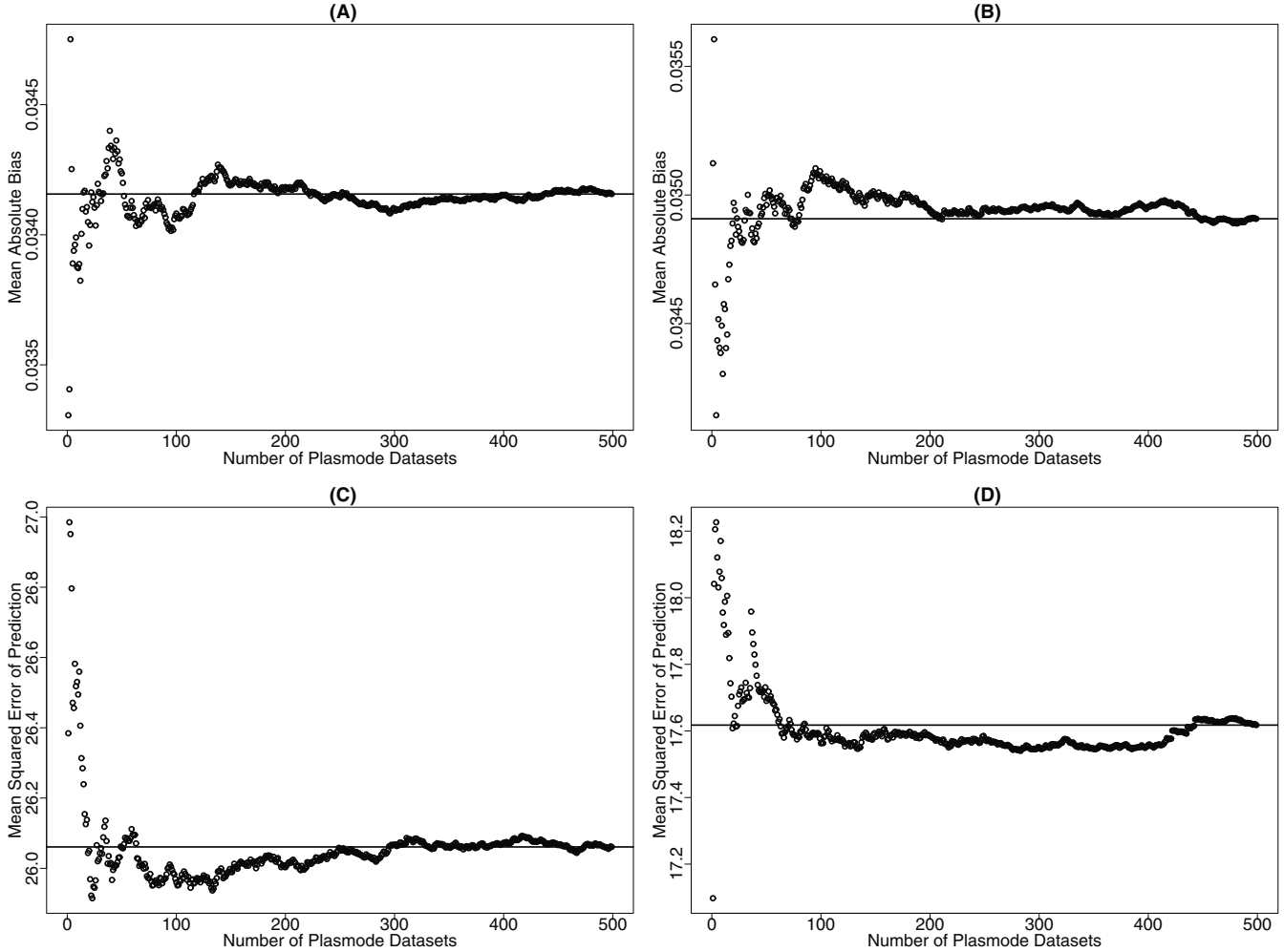

FIGURE 4 Convergence of the performance measures mean absolute bias (MAB) and mean squared error of prediction (MSEP) for increasing number N of statistical plasmode data sets. The horizontal lines illustrate the estimates of MAB and MSEP based on $N = 500$ Plasmode datasets. (A) Mean absolute bias of ridge regression. (B) Mean absolute bias of linear mixed model. (C) Mean squared error of prediction of ridge regression. (D) Mean squared error of prediction of linear mixed model.

parametric variability. For our example we have decided to abstain from modeling residuals, although we acknowledge that there might be valid reasons to include them.

Step 8. The artificial outcomes (of some of) the $N = 500$ plasmode data sets are compared with the original outcomes via histograms in Figure 3. The distribution of the original and artificial outcomes is very similar in shape and mean. The range of the original outcomes is larger than the range of the artificial outcome which can be explained by the outcome generation via resampled covariables and effects determined by LASSO (sparse and shrunken effects) which most likely will not lead to more extreme outcome values than contained in the underlying data set.

We conclude that the artificial outcome data come close to reality but might not properly reflect extreme values. The range of the artificial outcomes could be increased, for example, by manually altering some elements of the effect vector estimated by LASSO (as investigator's choice of the "truth"). By doing so, however, we would further alter the association between some of the covariables and the novel outcomes.

Step 9. We have described and justified the decisions for our plasmode simulations in each of the previous steps. The code for the generation of the plasmode simulation and their evaluation is available as supplemental material.

In Figure 4, we illustrate the convergence of the performance measures for increasing number of plasmode data sets. The estimators for MAB and MSEP for both ridge regression and the linear mixed model seem to have stabilized at about 200 generated simulations. Thus, we conclude that the generated number of statistical plasmode is sufficient to obtain stable estimates of the performance measures defined in Step 1.

Finally, we compare the performance of ridge regression and the linear mixed model in our statistical plasmode simulations. The MAB of the Ridge regression is estimated as 0.03415 while the estimate of the MAB of the linear mixed model

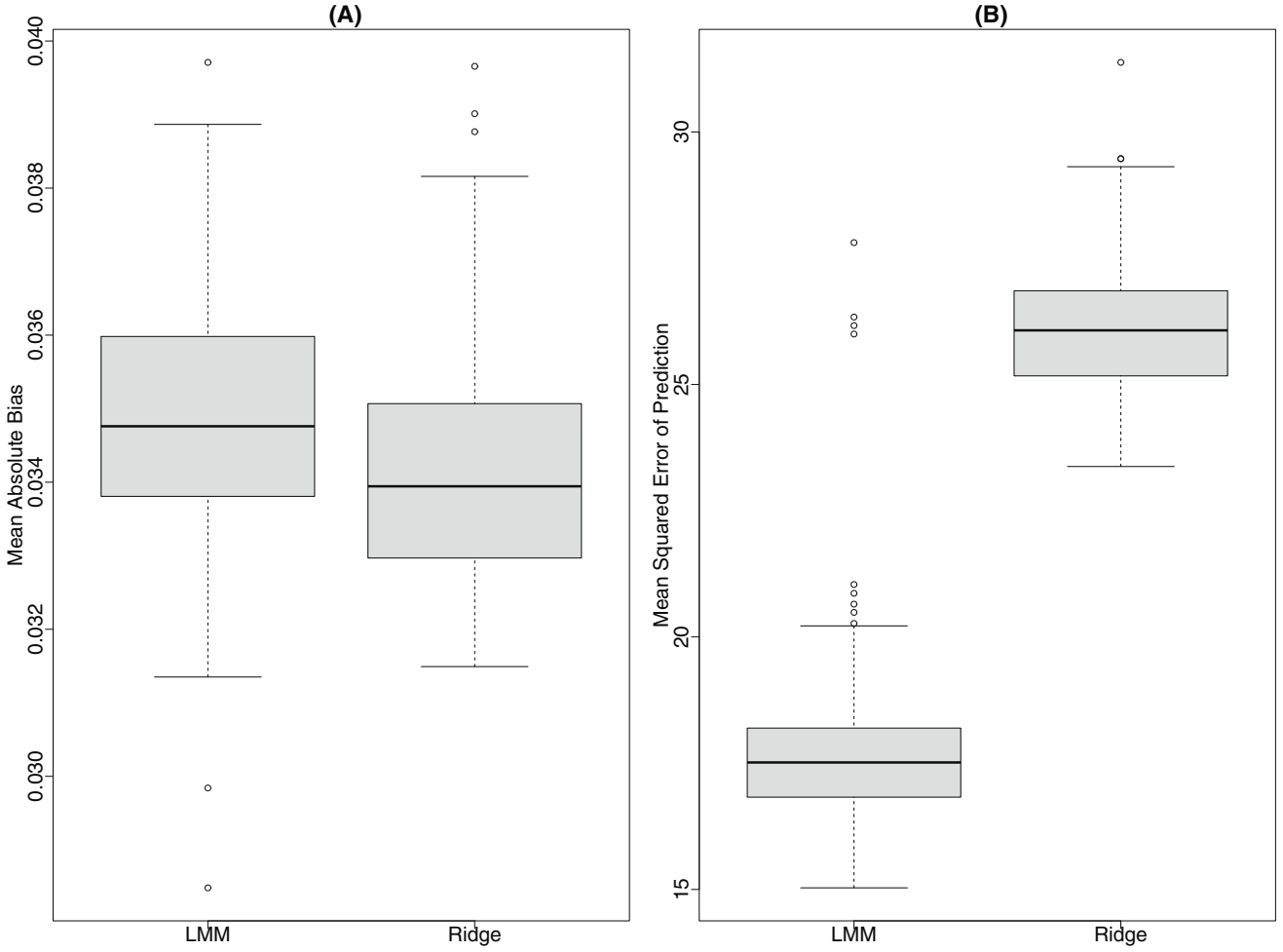

FIGURE 5 Boxplots of (A) the mean absolute bias for both the linear mixed model (LMM) and ridge regression (Ridge) in $N = 500$ statistical plasmode data sets; (B) the mean-squared error of prediction for both the linear mixed model and ridge regression in $N = 500$ statistical plasmode datasets. (A) Bias LMM versus ridge, $N = 500$. (B) Prediction error LMM versus ridge, $N = 500$.

is 0.03491, see also Equation (5). The sample-split MSEP of ridge regression is estimated as 26.01 while the sample-split MSEP of the linear mixed model is 17.62, see also Equation (5). In Figure 5, we depict the estimated values for each plasmode data set via boxplots. These results suggests that in our generated statistical plasmode simulations, which represent high-dimensional RNA expression data with sparse known effects and artificial normal outcomes, the linear mixed model performs superior to ridge regression with respect to prediction although the Ridge regression has a slightly smaller average bias.

6 | CONCLUSIONS AND OUTLOOK

Many simulation studies impose relatively strong assumptions regarding the nature of randomness in the data and its dependence structure. Mostly of theoretical kind, those assumptions primarily rely on the assumptions inherent in the statistical models applied to generate the data. Since not all assumptions can be justified in applied settings, the corresponding simulation studies may not be able to capture biologically meaningful relationships and thus result in misleading conclusions and research findings.

To avoid (at least some of) those issues, plasmode data sets are considered as an alternative data generation approach. While parametric simulations are known to provide only a partial representation of reality,[4] plasmodes have been declared to generate data that resemble reality in the closest way.[9] Highly appreciated for their ability to generate most realistic data, plasmode do not impose any specific model assumptions on their data generation process. Thus, no assumptions need to be justified to address the applicability of plasmodes. Nevertheless, a number of assumptions such as the representativeness of the underlying data sample have to be verified in order to guarantee the reliability of the generated plasmode data.

Plasmodes can accommodate unknown features such as dependence structure, distributions, and others, in particular, in the case of high-dimensional data. We recall that in case of parametric simulations most of those quantities are to be specified in advance. All in total, plasmode data sets may provide an attractive supplement to parametric simulations and can be applied in order to increase the reliability of the obtained research results.

In the present paper, we first discuss the concept of statistical plasmodes as those created by resampling of covariate information from empirical data at hand, optional parametric exposure generation and subsequent outcome generation using an appropriate outcome-generating model. This is what distinguishes them from biological plasmodes which are usually created by conducting lab experiments. We interpret statistical plasmodes as an intermediate step between the parametric and nonparametric simulations, with the parametric component mainly represented by the chosen outcome-generating model. After the introduction of statistical plasmodes, we discuss their main advantages and challenges and propose a step-by-step scheme for their generation and reporting. That scheme is then illustrated by means of a numerical example. All discussions in the present paper are presented in the context of prediction and explanatory models (including propensity-based methods such as propensity score weighted regression).

Plasmodes are bounded to the sample they are based on, and thus cannot produce the same variety of different scenarios as parametric simulations do. In this context, questions on the data availability and representativeness arise. In particular, even if plasmodes offer a flexible data generation procedure which creates realistic data, the representativeness of the generated data still substantially depends on the representativeness of the underlying real data set. To address this limitation, some authors such as Ejima et al[37] assume that the empirical data at hand represents the entire population of interest. Of course, such an assumption cannot be satisfied in each particular situation.

Spurious correlations are another issue closely related to the question of representativeness. Although plasmodes do not specify the underlying dependence structure explicitly, they do reproduce it to a certain extent while generating new data. Thus, if the sample at hand does not adequately represent the population of interest, the existing spurious correlations may be increased or even distorted for the generated plasmode data sets. As a result, the corresponding generated dependence structure will not represent the real one.

Statistical plasmodes as introduced in the present paper incorporate features from both parametric simulations and resampling approaches, and, as a result, inherit the strengths and weaknesses of each data generation method. On one hand, statistical plasmodes offer the advantage of creating more realistic data by generating covariate information through resampling techniques. On the other hand, they may also introduce certain challenges with respect to the subsequent model comparisons, as compared to purely parametric simulations.[32] Statistical plasmodes enable to control and manipulate certain aspects of the "truth" through the use of parametric OGMs, which can be advantageous over pure resampling methods. Nevertheless, asymptotic results established for resampling techniques may not be directly applicable to statistical plasmodes.

Our discussion points out several interesting options for future research. First, basic expectations placed on the plasmodes are related to their ability to preserve real data distributions, the underlying dependence structure and, as a result, the existing empirical associations. Those expectations are to be guaranteed by resampling from the observed covariate data at hand, without any additional data modification. However, it is not obvious how the choice of a particular resampling technique and specification of its parameters (such as the subsampling proportion in case of the subsampling technique) might impact the robustness of the obtained data generation results, for example, in the context of spurious correlations or sparse data. Additionally, the calculation of the optimal m might require high computational costs. A closer analysis of these impacts are possible topics for future research.

Second, a data generation method is considered to be realistic if it reflects the real data structure and the existing dependencies in the most accurate way. Thus, appropriate distance measures need to be specified in advance and also included into the reporting step of the data generation procedure. Such measures can then be used to measure the closeness of the generated plasmode data set to the underlying real data set. The choice of an appropriate distance measure, as well as the robustness of the plamode generation procedure with respect to that choice, can also be an interesting research topic.

Finally, the outcome-generating models present the major obstacle for plasmodes to become a purely non-parametric data generation approach. In the future we intend to analyze the impact of an OGM on the performance of the plasmode data generation procedure and to construct examples where the replacement of a parametric OGM with a non-parametric one improves the obtained data generation results. It is also of great interest to address possible "plasmode failure" for data sets generated through statistical plasmodes.

In total, our paper presents a comprehensive analysis of statistical plasmode simulations, discusses their potentials and central challenges and provides step-by-step recommendations for their generation. Our future research aims to address (at least some of) the pitfalls in the most close way to potentially provide more understanding and further novel insights into statistical plasmode generation.

ACKNOWLEDGEMENTS

The authors would like to thank Jörg Rahnenführer, Andrea Bommert and Marieke Stolte for excellent discussions. We are grateful to an anonymous referee and an associate editor for helpful comments and suggestions which helped us to enhance the quality of the manuscript. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The TCGA-BRCA data set is publicly available at the Genomic Data Commons. The R code to generate the data and to reproduce the example in Section 5 are available as supplementary material.

ORCID

Nicholas Schreck https://orcid.org/0000-0001-9277-4297
Alla Slynko https://orcid.org/0000-0001-6746-9039

REFERENCES

1. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25:4279-4292.
2. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38:2074-2102.
3. De Bin R, Boulesteix AL, Benner A, Becker N, Sauerbrei W. Combining clinical and molecular data in regression prediction models: insights from a simulation study. *Brief Bioinform*. 2020;21(6):1904-1919.
4. Reeb PD, Steibel JP. Evaluating statistical analysis models for RNA sequencing experiments. *Front Genet*. 2013;4:1-9.
5. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219-226.
6. Schulz A, Zöller D, Nickels S, et al. Simulation of complex data structures for planning of studies with focus on biomarker comparison. *BMC Med Res Methodol*. 2017;17:90.
7. Atiquzzaman M, Karim ME, Kopec J, Wong H, Vera MA, Anis AH. Using external data to incorporate unmeasured confounders: a plasmode simulation study comparing alternative approaches to impute body max index in a study of the relationship between osteoarthritis and cardiovascular disease. *J Stat Res*. 2020;54(2):131-145.
8. Shmueli G. To explain or to predict? *Stat Sci*. 2010;25:289-310.
9. Mehta T, Tanik M, Allison DB. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat Genet*. 2004;36:943-947.
10. Boulesteix AL, Groenwold RH, Abrahamowicz M, et al. Introduction to statistical simulations in health research. *BMJ Open*. 2020;10:e039921. doi:10.1136/bmjopen-2020-039921

11. Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene set analysis. *BMC Bioinform*. 2021;22(191):96-106.
12. Friedrich S, Friede T. On the role of benchmarking data sets and simulations in method comparison studies. *Biom J*. 2023;66:e2200212.
13. Mehta T, Zakharkin SO, Gadbury GL, Allison DB. Epistemological issues in omics and high-dimensional biology: give the people what they want. *Phys Genom*. 2006;28:24-32.
14. Gadbury GL, Xiang Q, Yang L, Barnes S, Page GP, Allison DB. Evaluating statistical methods using plasmode data sets in the age of massive public databses: an illustration using false discovery rates. *PLoS Genet*. 2008;6:1-8.
15. Vaughan LK, Divers J, Padilla MA, et al. The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. *Comput Stat Data Anal*. 2009;53:1755-1766.
16. Headrick TC, Sawilowsky SS. Simulating correlated multivariate nonnormal distributions: extending the Fleschman power method. *Psychometrika*. 1999;64(1):25-35.
17. Azuero A, Redden DT, Tiwari HK, Asmelash SG, Piyathilake CJ. A simple distribution-free algorithm for generating simulated high-dimensional correlated data with an autoregressive structure. *Commun Stat Simul Comput*. 2012;41(1):89-98.
18. Fan J, Han F, Liu H. Challenges of big data analysis. *Natl Sci Rev*. 2014;1(2):293-314.
19. Fan J, Li R. Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. Proceedings of the International Congress of Mathematicians, Madrid, Spain. 2006 595-622.
20. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. *Phil Trans R Soc A*. 2009;367:4237-4253.
21. Cattell R, Jaspers J. A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. 1967 63-67.
22. Sokal RR, Rohlf FJ, Zang E, Osness W. Reification in factor analysis: a plasmode based on human physiology-of-exercise variables. *Multivar Behav Res*. 1980;2:181-202.
23. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4:249-264.
24. Tibshirani R. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinform*. 2006;7:106.
25. Franklin JM, Eddings W, Glynn RJ, Schneeweiss S. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol*. 2015;182(7):651-659.
26. Franklin JM, Eddings W, Austin PC, Stuart EA, Schneeweiss S. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Stat Med*. 2017;36:1946-1963.
27. Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm. *Epidemiology*. 2018;29:191-198.
28. Desai RJ, Wyss R, Abdia Y, et al. Evaluating the use of bootstrapping in cohort studies conducted with 1:1 propensity score matching-a plasmode simulation study. *Pharmacoepidemiol Drug Saf*. 2019;28:879-886.
29. Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Franklin JM. Evaluating the utility of coarsened exact matching for pharmacoepidemiology using real and simulated claims data. *Pract Epidemiol*. 2019;189(6):613-622.
30. Liu SH, Chrysanthopoulou SA, Chang Q, Hunnicutt JN, Lapane KL. Missing data in marginal structural models: a plasmode simulation study comparing multiple imputation and inverse probability weighting. *Med Care*. 2019;57(3):237-243.
31. Conover MM, Rothman KJ, Stürmer T, Ellis AR, Poole C, Funk MJ. Propensity score trimming mitigates bias due to covariate measurement error in inverse probability of treatment weighted analyses: a plasmode simulation. *Stat Med*. 2021;40:2101-2112.
32. Wyss R, Schneeweiss S, van der Laan M, Lendle SD, Ju C, Franklin JM. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*. 2021;29:96-106.
33. Hafermann L, Klein N, Rauch G, Kammer M, Heinze G. Using background knowledge from preceding studies for building a random forest prediction model: a plasmode simulation study. *Entropy*. 2022;24:847.
34. Rodriguez PJ, Veenstra DL, Heagerty PJ, Goss CH, Ramos KJ, Bansal A. A framework for using real-world data and health outcomes modeling to evaluate machine learning based risk prediction models. *Value Health*. 2022;25(3):350-358.
35. Elobeid MA, Padilla MA, McVie T, et al. Missing data in randomized clinical trials for weight loss: scope of the problem, state of the field, and performance of statistical methods. *PloS One*. 2009;4(8):e6624.
36. Reeb PD, Bramardi SJ, Steibel JP. Assessing dissimilarity measures for sample based hierarchical clustering of RNA sequencing data using plasmode datasets. *PLoS Genet*. 2015;10(7):e0132310.
37. Ejima K, Brown AW, Smith DL, Beyaztas U, Allison DB. Murine genetic models of obesity: type I error rates and the power of commonly used analyses as assessed by plasmode-based simulation. *Int J Obes (Lond)*. 2020;44:1440-1449.
38. Alfaras I, Ejima K, Teixeira CVL, et al. Empirical versus theoretical power and type I error (false-positive) rates estimated from real murine aging research data. *Cell Rep*. 2021;36(7):109560.
39. Bickel BJ, Götze F, van Zwet WR. Resampling fewer than n observations: gains, losses, and remedies for losses. *Stat Sin*. 1997;7:1-31.
40. Bickel BJ, Sakov A. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Stat Sin*. 2008;18:967-985.
41. Andrews DWK, Buchinsky M. A three-step method for choosing the number of bootstrap repetitions. *Econometrica*. 2000;68:23-51.
42. Davidson R, MacKinnon JG. Bootstrap tests: how many bootstraps? *Econom Rev*. 2000;19:55-68.
43. Politis D, Romano J, Wolf M. *Subsampling*. New York: Springer; 1999.
44. Bickel BJ, Ren JJ. The bootstrap in hypothesis testing. Lecture Notes-Monograph Series. *State Art Probab Stat*. 2001;36:91-112.
45. Andrews DWK, Guggenberger P. Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Economet Theor*. 2010;26:426-468.
46. Gerard D. Data-based RNA-seq simulations by binomial thinning. *BMC Boinform*. 2020;21:206-220.

47. Ju C, Wyss R, Franklin JM, Schneeweiss S, Häggström J, van der Laan MJ. Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data. *Stat Methods Med Res*. 2019;28:1044-1063.
48. Beran R, Srivastava MS. Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann Stat*. 1985;13:95-115.
49. Bickel PJ, Freedman DA. Bootstrapping regression models with many parameters. Paper presented at: A Festschrift for Erich L. Lehmann in Honor of his Sixty-Fifth Birthday. Wadsworth; Belmont, CA. 1983 28-48.
50. Mammen E. Bootstrap and wild bootstrap for high dimensional linear models. *Ann Stat*. 1993;21:255-285.
51. Karoui NE, Purdom E. Can we trust the bootstrap in high-dimensions? The case of linear models. *J Mach Learn Res*. 2018;19:1-66.
52. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Dent Tech*. 1970;12:55-67.
53. Searle SR, Casella G, McCulloch CE. *Variance Components*. New Jersey: Wiley Interscience; 1992.
54. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1-22.
55. Covarrubias-Pazaran G. Genome assisted prediction of quantitative traits using the R package sommer. *PloS One*. 2017;11:1-15.
56. Weinstein J, Collisson E, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45:1113-1120.
57. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2015;44(8):e71.
58. Silva TC, Colaprico A, Olsen C, et al. TCGA workflow: analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research*. 2016;5:1542.
59. Mounir M, Lucchetta M, Silva TC, et al. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput Biol*. 2019;15(3):e1006701.
60. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
61. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *J Multivar Anal*. 2004;88(2):365-411.
62. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc B*. 1996;58:267-288.

How to cite this article: Schreck N, Slynko A, Saadati M, Benner A. Statistical plasmode simulations–Potentials, challenges and recommendations. *Statistics in Medicine*. 2024;43(9):1804-1825. doi: 10.1002/sim.10012