A multiple testing framework for diagnostic accuracy studies with co-primary endpoints

Max Westphal1 | Antonia Zapf2 | Werner Brannath1,3

1Institute for Statistics, University of Bremen, Bremen, Germany
2Institute of Medical Biometry and Epidemiology, UKE Hamburg, Hamburg, Germany
3Competence Center for Clinical Trials Bremen, University of Bremen, Bremen, Germany

Correspondence
Max Westphal, Institute for Statistics, University of Bremen, Bremen, Germany. Email: max.westphal@mevis.fraunhofer.de

Present address
Max Westphal, Fraunhofer Institute for Digital Medicine MEVIS, Max-Von-Laue-Straße 2 28359, Bremen, Germany

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 281474342/GRK2224/1

INTRODUCTION

Research projects concerned with the application of machine learning and deep learning techniques for disease diagnosis and prognosis have steadily grown in number over the last years. This is indicated, among others, by several review and overview publications.1-5 Despite all the recent success of these modern techniques, challenges regarding over-optimistic conclusions drawn from finite datasets remain. For instance, it is still sometimes neglected that the true performance of a predictive model is only empirically estimated. This makes a sound quantification of uncertainty important when
the data quantity is severely limited. Bias induced by selection (e.g., data source, data splitting modalities, performance measure, comparator, etc.) is also a recognized issue. The less of the defining properties of a study are specified in advance, the more ‘opportunities’ present themselves to the researchers to influence the results to match their expectations.6-9 This is especially important because trying out a wide variety of learning algorithms is common practice in applied machine learning. This approach is somewhat justified by the no free lunch theorem of statistical learning which states that no single algorithm gives universally best results for all prediction tasks.10 Modern algorithms additionally involve the tuning of several hyperparameters. For instance, the performance of neural networks depends on depth (number of layers), width (neurons per layer), activation function(s) and further hyperparameters.11 In effect, usually dozens, hundreds or even thousands of candidate models are trained and compared. A common recommendation to mitigate the resulting selection induced bias is to split data for development, selection and evaluation.12-17 In the machine learning community, the according data sets are commonly denoted as training, validation and test set.

Further sources of bias related to study design have been identified in the literature.18-22 Relevant sources for overoptimistic conclusions are selection of nonconsecutive patients, analysis of retrospective data and focus on severe cases and healthy controls.20 Another issue is nontransparent communication of the results.23-25 This may prevent other research teams to replicate the findings on similar problems (replicability) or even on the same data (reproducibility).26,27

The focus in this work are statistical methods for diagnostic accuracy studies. While the literature often targets traditional diagnostic or prognostic tests, the majority of concepts translate to the evaluation of (machine-learned) prediction models for diagnostic or prognostic purposes. In the terminology of Knottnerus and Buntinx (chapter 2)28 our work is mainly tailored toward cross-sectional phase III (comparative) diagnostic test accuracy studies which ask the question:

“Among patients in whom it is clinically sensible to suspect the target disorder, does the level of the test result distinguish those with and without the target disorder?”28

Translated to our scenario, the goal of such a study is to provide evidence that a novel diagnostic or prognostic model (test) outperforms either a given comparator or, if no such competing model (test) is available, a given performance threshold. Performance, that is, sensitivity and specificity, is measured with regard to the (defined) ground truth which is in the optimal case derived by a so-called gold standard. It may however be infeasible to implement such a gold standard in the study or there may not even exist one. Thus, typically a reference standard is defined as the best available approximation of the gold standard. Common reasons to utilize an new model (test) in clinical practice are either lower invasiveness or costs. This balances the fact that, by definition, the index model (test) cannot have better performance than the reference standard. We refer to the literature for a general introduction to diagnostic research and a throughout treatment of diagnostic accuracy studies28-31 and a discussion of issues related to study design and reporting.23-25

A recent example of a diagnostic accuracy study in the context of applied machine learning covers the IDx-DR device which allows diagnosis of diabetic retinopathy in diabetic patients via deep learning based on retinal color images.32 Diagnostic accuracy was assessed in an extensive observational clinical trial involving 900 patients, which lead to a marketing permit by the FDA*. An overview over the current regulatory framework for machine learning empowered medical devices in the EU and the United States from the viewpoint of radiology has recently been provided.33 Important developments are in particular the new Medical Devices Regulation (MDR) and the new In Vitro Diagnostic Medical Device Regulation (IVDR) in the EU.34,35 These new regulations are certainly relevant for the application of machine learning methods for medical testing purposes and apply from May 2020 and May 2022, respectively. In the IDx-DR study, only a single model was (successfully) prospectively evaluated on the independent test dataset.32 This default strategy is often advised in machine learning and diagnostic research to avoid any selection induced bias. It is reasonable if previously available data for model training and selection is (a) large in number and (b) representative of the intended target population. However, when potentially hundreds of modeling approaches are compared on (quantitatively and/or qualitatively) modest datasets, the model selection process can rarely be concluded with confidence. While the default approach enables an unbiased estimation and simple statistical inference, one is thus bound to this one-time model choice under uncertainty. In effect, this strategy is quite inflexible as it is impossible to retrospectively correct a flawed model selection without compromising the statistical inference in the evaluation study.

To address this issue, an established multiple testing approach36 was recently adapted to explicitly take into account that multiple models are assessed simultaneously on the same evaluation dataset.37 In effect, model selection can be improved with help of the test data.38 The employed simultaneous test procedure is based on the (approximate) multivariate normal distribution of performance estimates. An advantage of this procedure is that the multiplicity adjustment needs to be less strict when candidate models give (highly) similar predictions. This approach allows approximate control of the overall type I error rate and construction of simultaneous confidence regions as well as corrected (median-conservative)
point estimates. Moreover, it was found that selecting multiple promising models can increase statistical power for the evaluation study compared to the default approach where only the best (cross-)validation model is evaluated.

Multiple testing methodology is not particularly popular, neither for medical test evaluation nor in machine learning. As indicated above, the recommendation to evaluate a single final model or test on independent data has prevailed in both domains. In this case, no adjustments for multiplicity are necessary as the final model is selected independently of (i.e., prior to) the evaluation study. Our approach has some overlap to so-called benchmark experiments in predictive modeling. Several authors showcase different approaches for the comparison of learning algorithms over multiple (real or artificial) datasets and inference regarding their expected performances. In contrast, we aim to evaluate prediction models (conditional on the learning data), which is arguably more important at the end of the model building process—right before implementation of a specific model in (clinical) practice. In medical testing applications, it is natural to consider only a single index (candidate) test. This is particularly the case, when different tests are based on separate biological samples. It would then often not be considered ethically justifiable to assess more than a single index test and the reference test on the same patients. In contrast, when all index test are derived from the same data sample, for example, retinal images, the number of index tests (diagnostic models) should be primarily guided by statistical considerations as long as the type I error rate can be controlled.

This work covers two distinct contributions. Firstly, we extend the existing multiple testing approach to diagnostic accuracy studies with co-primary endpoints sensitivity and specificity. Secondly, we describe a novel Bayesian approach to determine the optimal number of models to include in the final evaluation study based on preliminary data. The remainder of this work is structured as follows. In Section 2, we motivate our work by a real-world application, namely the prediction of abnormal fetal health status with machine learning methods. In Section 3, we introduce core assumptions and notation and establish our inference framework. In Section 4, we describe a novel Bayesian approach to determine the optimal number of models to include in the final evaluation study based on preliminary data. In Section 5, we show several numerical results. We apply our methods to the real data introduced in Section 2 (Section 5.1). We show results from a large simulation study in the machine learning context with the aim to compare different approaches for model selection prior to the evaluation study (Section 5.2). Lastly, we assess our multiple testing approach with regards to finite sample type I error rates under least-favorable parameter configurations (Section 5.3). Finally, in Section 6, we summarize our results, point out limitations of our framework, and give an outlook on possible extensions.

2 | MOTIVATING EXAMPLE: SCREENING FOR ABNORMAL FETAL STATE

Cardiotocography (CTG) measurements capture the fetal heart rate and uterine contractions during pregnancy and delivery. The use of CTG can contribute to improve the management of high-risk pregnancies. In the following, we utilize the publicly available Cardiotocography Data Set to illustrate the approach introduced in this work as well as its use case in clinical research.

The dataset consists of 2126 observations and is labeled with a measurement date which allows us to emulate a prospective evaluation study. For that matter, the chronologically first $n_c = 1594$ observations are utilized for model development. The prediction target is an abnormal (suspect or pathologic) fetal state which has a prevalence of 24% in the learning data. The reference standard is a consensus vote of medical experts. Five different classes of machine learning models are utilized, namely penalized logistic regression (or elastic net) models, decision trees, random forests, support vector machines, and gradient boosted trees.

For each learning algorithm, 200 distinct hyperparameters are investigated which are sampled randomly as implemented in the R package caret. To assess model performance, a hold out validation set with $n_v = 318$ observations is used to estimate out-of-sample sensitivity and specificity of each of the 1000 candidate models. Empirical validation performance is quite high with 82 models exceeding 90% sensitivity and specificity of which 48 models are random forests (18 elastic net models, 2 support vector machines, 14 gradient boosted trees).

Assuming a sensitivity of 80% and a specificity of 70% have been declared as the absolute minimal acceptance criteria for implementation in clinical practice, a typical research question is now: "Which of the developed prediction models should be evaluated with regards to these co-primary endpoints in an independent, prospective accuracy study?" To counter the uncertainty in the preliminary validation estimates, in this work, we will rather select multiple promising models instead of a single model, as often recommended (Section 4). The selected models are then evaluated simultaneously while adjusting for multiple comparisons (Section 3). The evaluation study for our example utilizing the new methodology is illustrated in Section 5.1.
3 | STATISTICAL INFERENCE FRAMEWORK

3.1 | Prerequisites

We consider predicting a binary label $Y \in \{0, 1\}$ based on $P \in \mathbb{N}$ features $X \in \mathbb{R}^P$. Our canonical example in this work will be the distinction between diseased (having a given target condition, $Y = 1$) and healthy ($Y = 0$) subjects. In medical applications, the target $Y$ should be obtained by the (best available) reference standard and forms the ground truth for learning and evaluation. The goal of supervised machine learning is to provide a prediction model $\hat{f} : x \mapsto \hat{y}$ with high performance. In practice, this is achieved by a learning algorithm (e.g., stochastic gradient descent) which trains a prediction model $\hat{f}$ (e.g., a neural network) based on learning data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{n}$. We emphasize that we are concerned with the assessment of a specific model $\hat{f}$ trained on specific data $\mathcal{L}$ and not with the properties of the algorithm that was used to learn $\hat{f}$. Finally, we note that, from the viewpoint of predictive modeling, disease diagnosis and prognosis essentially only differ regarding the time lag between capturing the features $x$ and the label $y$—at least if this lag is approximately equal in the population of interest. Otherwise, a time to event analysis may be more appropriate. For the sake of brevity, we will focus on diagnosis tasks in the following. In medical diagnosis, sensitivity ($Se$) and specificity ($Sp$) are often both assessed simultaneously. We thus consider the tuple $\theta = (Se, Sp)$ as the performance measure of interest where

$$Se = \mathbb{P}(\hat{f}(X) = 1 | Y = 1) \quad \text{and} \quad Sp = \mathbb{P}(\hat{f}(X) = 0 | Y = 0).$$

This is advantageous compared to just considering the overall accuracy $Acc = \mathbb{P}(\hat{f}(X) = Y) = \rho Se + (1 - \rho)Sp$ in the sense that neither the accuracy in healthy or diseased alone dominates our perception of the predictive performance if the disease prevalence $\rho = \mathbb{P}(Y = 1)$ is small.

As we have outlined in Section 1, it is usually recommended to conduct a final performance assessment on independent evaluation or test data $\mathcal{E} = \{(x_i, y_i)\}_{i=1}^{n}$. Performance estimates on the learning data are generally a bad indicator of the true performance, in particular when a highly complex model might have overfitted the learning data. We assume that $\mathcal{E}$ is an i.i.d. sample from the unknown joint distribution $\mathcal{D}_{(X,Y)}$ of $X$ and $Y$. Moreover, we assume that $M \in \mathbb{N}$ models $\hat{f}_m$, $m \in M = \{1, \ldots, M\}$, have been initially trained in the learning phase via different learning algorithms. A subset $S \subset M$ of these initial candidate models is selected for evaluation. To simplify the notation we assume that the candidate models are ordered such that $S = \{1, \ldots, S\}$ with $1 \leq S \leq M$. Different approaches to identify $S$ are discussed in Section 4. This is implemented in practice by training models first only on the so-called training data $\mathcal{T} \subset \mathcal{L}$, a subset of the learning data. The remaining validation data $\mathcal{V} = \mathcal{L} \setminus \mathcal{V}$ is then used for a performance assessment of the resulting preliminary models $\hat{f}_m$ (only trained on $\mathcal{T}$). The model(s) selected for evaluation are than re-trained with all available data $\mathcal{L} = \mathcal{T} \cup \mathcal{V}$ before going into the evaluation study with the expectation that this (slightly) increases their performance. This procedure is known as (simple) hold-out validation. There exist several variations and alternative strategies such as cross-validation or bootstrapping. They can stabilize model selection, are however computationally more expensive as models have to be retrained several times which is not always feasible in practice. For simplicity, we will focus on simple hold-out validation in this work.

3.2 | Study goal

Our overall goal is to identify a prediction model with high sensitivity and specificity from the $M$ initially trained models. Additionally, we aim to show superiority of at least one model compared to prespecified thresholds $\theta_0 = (Se_0, Sp_0) \in (0, 1)^2$. Alternatively, $\theta_0 = \theta(\hat{f}_0)$ is the unknown performance of a comparator $\hat{f}_0$, that is, an established diagnostic testing procedure which is also estimated in the evaluation study. In the following, we will focus on the former scenario for brevity. Throughout this work, we assume that the evaluation study is declared as successful if and only if superiority in both endpoints for at least one candidate model $m \in S$ relative to the the comparator $\theta_0$ can be demonstrated. This so-called co-primary endpoint analysis is the standard approach in confirmatory diagnostic accuracy studies. Other approaches have also been discussed in the literature. The system of null hypotheses is thus given by

$$H_S = \{H_m : H_m^{Se} \cup H_m^{Sp}, \ m \in S\}$$

where $H_m^{Se} : Se_m \leq Se_0$ and $H_m^{Sp} : Sp_m \leq Sp_0$. (1)
This corresponds to an union-intersection test over models and an intersection-union test over parameters \((\text{Se}, \text{Sp})\). Equivalently, the hypothesis system can be expressed as

\[ H_S = \{ H_m : \min(\Delta_m, \Delta_m^\text{Sp}) \leq 0, \ m \in S \}. \]

whereby \(\Delta_m^\text{Se} = \text{Se}_m - \text{Se}_0\) and \(\Delta_m^\text{Sp} = \text{Sp}_m - \text{Sp}_0\). The goal of the evaluation study is to estimate the unknown sensitivities and specificities, provide confidence bounds for these estimates and obtain a multiple test decision for the hypothesis system (1).

A multiple test \(\varphi = (\varphi_1, \ldots, \varphi_S)\) depends on the evaluation data \(\mathcal{E}\) and results in a binary vector \(\varphi \in \{0, 1\}^S\) of test decisions whereby \(H_m\) is rejected if and only if \(\varphi_m = 1\). In a regulatory setting, it is usually required that \(\varphi\) shall control the family-wise error rate (FWER) strongly at significance level \(\alpha\), i.e. we require for all possible parameter configurations \(\theta\)

\[ \text{FWER}_\varphi(\varphi) = \mathbb{P}_\theta \left( \bigcup_{m \in S} \{ \varphi_m = 1 \} \right) \leq \alpha. \]  

(2)

Hereby, \(S_0 = S_0(\theta, \theta_0) \subset S\) is the index set of true null hypothesis, that is, we have either \(\text{Se}_m \leq \text{Se}_0\) or \(\text{Sp}_m \leq \text{Sp}_0\) for all \(m \in S_0\). Any parameter \(\theta\) for which \(\text{FWER}_\varphi(\varphi)\) becomes maximal is called a least favorable parameter configuration (LFC). In this work, we will restrict our attention to asymptotic control of the FWER, i.e. (2) shall only hold as \(n = n_\varepsilon \to \infty\). We will investigate the finite sample FWER in realistic and least-favorable settings in section 5. We extend \(\varphi\) to a multiple test for the actually relevant hypothesis system

\[ H = H_M = \{ H_m : \theta_m \leq 0, \ m \in M \}, \]  

(3)

concerning all initial candidate models \(\hat{f}_m, m \in M\), by setting \(\varphi_m = 0\) for all \(m \in M \setminus S\). That is to say, a model \(m\) cannot be positively evaluated \((\varphi_m = 1)\) when it is not selected for evaluation. This natural definition has two consequences. Firstly, the extended test retains (asymptotic) FWER control as only non-rejections are added. Secondly, we can compare different model selection strategies because the extended multiple test always operates on \(H = H_M\) and not only on \(H_S\). That is to say, a model \(m\) cannot be positively evaluated \((\varphi_m = 1)\) when it is not selected for evaluation. This natural definition has two consequences. Firstly, the extended test retains (asymptotic) FWER control as only non-rejections are added. Secondly, we can compare different model selection strategies because the extended multiple test always operates on \(H = H_M\) and not only on \(H_S\).

3.3 Parameter estimation

We assume that a subset of promising models \(S = \{1, \ldots, S\} \subset M\) has been selected prior to the evaluation study. Suitable strategies for that matter are presented in Section 4. The observed feature-label data \(\mathcal{E} = \{ (\mathbf{x}_i, y_i) \}_{i=1}^n\) from the evaluation study is transformed to the actual relevant binary similarity matrices \(Q^{\text{Se}} \in \{0, 1\} \times \times \) and \(Q^{\text{Sp}} \in \{0, 1\} \times \times \) for the diseased \((y = 1)\) and healthy \((y = 0)\) subpopulation, respectively. Hereby, \(n_1\) and \(n_0\) are the number of diseased and healthy subjects of the \(n = n_\varepsilon = n_1 + n_0\) evaluation study observations. The entry of \(Q\) in row \(i\) and column \(m\) is equal to one if the prediction of the \(i\)th observation by the \(m\)th model is correct and zero if it is wrong. The sample averages

\[ \hat{\text{Se}} = \frac{1}{n_1} \sum_{i=1}^n \mathbb{1}(\hat{f}(\mathbf{x}_i) = y_i = 1) \quad \text{and} \quad \hat{\text{Sp}} = \frac{1}{n_0} \sum_{i=1}^n \mathbb{1}(\hat{f}(\mathbf{x}_i) = y_i = 0), \]

(4)

can thus be calculated as the column means of these similarity matrices. Moreover, we can estimate the covariances \(\Sigma^{\text{Se}} = \text{cov}(\hat{\text{Se}})\) and \(\Sigma^{\text{Sp}} = \text{cov}(\hat{\text{Sp}})\) as the sample covariance matrices of the similarity matrices divided by factors of \(n_1\) and \(n_0\), respectively.

A problem arises when a sample proportion of one (or zero) is observed. In this case, say if \(\hat{\text{Se}}_m = 1\), the plug-in variance estimate \(\text{Se}_m(1 - \text{Se}_m)/n_1\) collapses to zero. A secondary effect are confidence intervals of zero width which are difficult to interpret. In this work, we tackle this problem by utilizing a shrinkage estimator which is based on the posterior distribution of a multivariate Beta-binomial model. A detailed description is provided in Appendix S1. Marginally, this approach has the same effect as established approaches which are based on adding pseudo observations (1 correct, 1 false prediction in our case). For finite sample sizes, the influence of this adaptation is conservative as mean estimates are slightly shrunk toward 0.5 and variance estimates are slightly inflated. In addition, our approach induces shrinkage of the empirical correlation matrix toward the identity matrix. Asymptotically, the adapted estimators are equivalent to (4) and the asymptotic properties described in the next section thus remain unchanged.
3.4 | Multiple comparison procedure

Our goal is to apply the so-called maxT-approach to hypothesis system (1). This simultaneous test procedure has previously been utilized for the assessment of the overall classification accuracy of several binary classifiers. This multiple test enables us to reduce the need for multiplicity adjustment in case several similar models are evaluated. Similarity is measured by the correlation of (empirical) performances which depends on the probability of a common correct prediction in our context. The maxT-approach relies on determining a common critical value \( c_o \), which depends on the empirical correlation structure, such that each hypothesis \( H_m \) is rejected if and only if

\[
T_m = \min(T_m^{Se}, T_m^{Sp}) > c_o.
\]

The test statistics are defined in a standard manner via

\[
T_m^{Se} = \frac{\hat{S}e_m - S_0}{se(\hat{S}e_m)},
\]

whereby \( \hat{S}e(\hat{S}e_m) = \sqrt{S_e_m(1 - \hat{S}e_m)/n_1} \) is the estimated SE of \( \hat{S}e_m \). The test statistics \( T_m^{Sp} \) are defined accordingly. When only a single model is evaluated with regard to co-primary endpoints the tests regarding sensitivity and specificity each need to be conducted at local level \( \alpha \). This is because the LFC is of the form

\[
Se = S_0 \land Sp = 1 \quad \text{or} \quad Se = 1 \land Sp = S_0,
\]

that is, one parameter (Se or Sp) is equal to one and the other parameter lies on the boundary of the null hypothesis. Similarly, for our case of \( S \) candidate models, the two cases described in (6) are possible for each dimension \( m \in S = \{1, \ldots, S\} \), resulting in \( 2^S \) potential LFCs. We could attempt to identify the single least favorable configuration in terms of multiple testing adjustment which depends on the true correlation matrix. However, the computational burden of this extensive approach increases exponentially in \( S \) as \( 2^S \) critical values would have to be computed.

We will instead focus on a more direct approach. Assume we knew if \( \Delta_m^{Se} = S_e_m - S_0 < S_p_m - S_0 = \Delta_m^{Sp} \) or if \( \Delta_m^{Se} > \Delta_m^{Sp} \) is true for each \( m \in S \). For simplicity, we assume that \( \Delta_m^{Se} \) and \( \Delta_m^{Sp} \) are not exactly equal in the following. Define the indicator variables \( b_m = 1(\Delta_m^{Se} < \Delta_m^{Sp}), b = (b_1, \ldots, b_S) \) and the corresponding diagonal matrix \( B = \text{diag}(b) \). Moreover, we define

\[
B_2 = \begin{pmatrix} B \\ I_S - B \end{pmatrix} \in \{0, 1\}^{2S \times S} \quad \text{and} \quad T_2 = \begin{pmatrix} T_m^{Se} \\ T_m^{Sp} \end{pmatrix} \in \mathbb{R}^{2S \times 1},
\]

whereby \( I_S \) is the \( S \)-dimensional identity matrix. Finally, we can define the \( S \)-dimensional vector of test statistics

\[
T^b = (B_2)^T T_2 \in \mathbb{R}^S.
\]

Assuming \( n_1/n \to 0 \) as \( n = n_1 + n_0 \to \infty \), under the LFC \( \theta^b = (S_e^b, S_p^b) \), it can be shown that

\[
T_{(n)}^b \overset{\mathcal{D}}{\to} \mathcal{N}(0, R^b), \quad n \to \infty,
\]

whereby \( R^b = BR^{Se}_b + (I_S - B)R^{Sp}_b (I_S - B) \). This result is the foundation for the application of the maxT-approach and proven in the supplementary material. Together with the consistent estimator

\[
\hat{R}^b = \hat{B}R^{Se}_b + (I_S - \hat{B})R^{Sp}_b (I_S - B),
\]

of the correlation matrix \( R^b \), we can calculate a common critical value \( c_o \) such that

\[
\mathbb{P}(\max_{m \in S} T_m^b \leq c_o) \approx \Phi_S(c_o, \hat{R}^b) = \int_{-\infty}^{c_o} \ldots \int_{-\infty}^{c_o} \phi_S(x, \hat{R}^b) dx = 1 - \alpha,
\]

(7)
under the 'estimated LFC' $\theta^b$. Hereby, $\hat{b}$ and $\hat{B}$ are simply defined by plugging in $\hat{Se}$ for $Se$ and $\hat{Sp}$ for $Sp$. $\phi_S$ and $\Phi_S$ denote the density and distribution function of the $S$-dimensional standard normal distribution with mean $\theta$ and covariance matrix $\hat{R}_b$. In practice, $c_a$ can be found by numerical integration, for example, with help of the $R$ package mvtnorm.50

Finally, we can define our actual test statistic $T$ with entries $T_m = \min(T_{m}^{Se}, T_{m}^{Sp})$ for which $T \leq T^b$ holds (deterministically). This allows us to define a multiple test $\varphi$ via $\varphi_m = 1 \Leftrightarrow T_m > c_a$ which asymptotically controls the FWER (2) at level $\alpha$. A rigorous formulation of this result is provided in Appendix S1.

We can also construct (e.g.) one-sided simultaneous comparison regions for $Se$ and $Sp$ via

$$
\text{CR}_{1-a,m}^{Se} = \left( \hat{Se}_m - c_a \cdot \hat{Se}(\hat{Se}_m), 1 \right) \quad \text{and} \quad \text{CR}_{1-a,m}^{Sp} = \left( \hat{Sp}_m - c_a \cdot \hat{Sp}(\hat{Sp}_m), 1 \right).$$

(8)

Note, that $\text{CR}_{1-a,m}^{Se} = \prod_{m \in S} \text{CR}_{1-a,m}^{Se}$ and $\text{CR}_{1-a,m}^{Sp} = \prod_{m \in S} \text{CR}_{1-a,m}^{Sp}$ together do not define a $1 - \alpha$ confidence region in the full $(2S)$-dimensional parameter space as the coverage probability $\mathbb{P}_{\varphi}(Se \in \text{CR}_{1-a,m}^{Se} \land Sp \in \text{CR}_{1-a,m}^{Sp})$ may be smaller than $1 - \alpha$. Instead, we have asymptotically that

$$
\mathbb{P}_{\varphi}\left( \bigcup_{m \in S} \left\{ Se_m \notin \text{CR}_{1-a,m}^{Se} \land Sp_m \notin \text{CR}_{1-a,m}^{Sp} \right\} \right) \leq \alpha.
$$

This is why $\text{CR} = (\text{CR}_{1-a,m}^{Se}, \text{CR}_{1-a,m}^{Sp})$ should be reported as a comparison region and not as a confidence region or interval.51 A regular, wider confidence region may however also be obtained via the maxT-approach and reported in addition. For this, one simply needs to replace $\hat{R}_b$ in (7) with the block correlation matrix composed of $\hat{R}_b^{Se}$ and $\hat{R}_b^{Sp}$. When setting $\alpha = 0.5$, the lower bounds of (8) can be used as corrected point estimators $\hat{Se}$, $\hat{Sp}$ with the property that the overall probability that both $\hat{Se}_m$ and $\hat{Sp}_m$ overestimate their target for any $m \in S$ is asymptotically bounded by 50%.37

3.5 | Study conclusion

Our framework provides a multiple test decision for the hypotheses system (1). After the evaluation study, one might therefore partition the set of candidate models $M = P \cup \mathcal{N}$ into positively and negatively evaluated prediction models $P = \{ m \in S \subset M : \varphi_m = 1 \}$ and $\mathcal{N} = S \setminus P$. In the case $P = \emptyset$, no performance claim can be made and the evaluation study has failed in a strict confirmatory sense. When $|P| = 1$, only a single model meets the requirements and the situation is clear. However, if $|P| > 1$, the investigator has the choice which of the models $m \in P$ should be implemented in practice.

When applying the maxT-approach, $m^* = \arg\max_{m \in S} T_m$ is an obvious final model choice. In the co-primary endpoint setting, this corresponds to $m^* = \arg\max_{m \in S} \min(T_{m}^{Se}, T_{m}^{Sp})$, that is, we select the model for which the evidence that both endpoints is significant is maximal. It might also be reasonable to select the final model according to further criteria not directly tied to the discriminatory performance. For instance, the calibration of predicted probabilities is a distinct important attribute of prediction models. Several criteria for the assessment of model calibration have been proposed in the literature.52-55 Moreover, the interpretability of a prediction model is also recognized as an important requirement for implementation in clinical practice.56,57

4 | MODEL SELECTION PRIOR TO THE EVALUATION STUDY

A prominent recommendation in machine learning and in medical testing is to strictly separate model development (training and selection) and model evaluation (performance assessment), compare Section 1. The most straightforward implementation of this strategy is to select a single final model prior to the evaluation study. Often times this is done by choosing the model with the highest empirical performance on a (hold-out) validation dataset out of the $M$ initially trained models, as described in Section 3.1. When evaluating binary classifiers regarding their overall accuracy, previous work has shown that it is beneficial to evaluate multiple promising models simultaneously, for example, all models within one standard error of the best validation model.37,38 This so-called within 1 SE selection rule can lead to evaluation studies with higher power and improved (final) model selection. Upward biased accuracy estimation can be corrected in a conservative fashion. To obtain multiplicity adjusted test decisions and corrected point estimates, the so-called maxT-approach36-38 may be applied which we adopted to the co-primary endpoint setting in Section 3.4.
To arrive at a suitable selection of prediction models for the evaluation study, more work is necessary when sensitivity and specificity are assessed simultaneously. The within 1 SE rule\(^{37}\) can be adopted to the co-primary endpoint scenario by basing it on the weighted accuracy \(w_{\text{Acc}} = w_{\text{Se}} + (1 - w)_{\text{Sp}}\), \(w \in (0, 1)\) instead of the overall classification accuracy. We will investigate the performance of this approach with the weight \(w = (1 - w_{\text{Sp}}) + 1\)\(^{-1}\) in Section 5. This specific choice for \(w\) can be motivated by requiring that the two parameter configurations \((\text{Se}, \text{Sp}) = (\text{Se}_0, 1)\) and \((\text{Se}, \text{Sp}) = (1, \text{Sp}_0)\) are assigned an equal \(w_{\text{Acc}}\) value. If sensitivity and specificity are deemed equally important \((\text{Se}_0 = \text{Sp}_0)\), this implies a weight of \(w = 0.5\). While the heuristic within 1 SE approach is intuitive and has proven to work empirically, it lacks a throughout theoretical justification. In particular, the question which multiplier \(k \geq 0\) results in the best within \(k\) SE rule remains open.

In the following, we introduce a more elaborate algorithm which aims to derive the models to be evaluated in an optimal fashion. We are dealing with a subset selection problem, that is to say that our goal is to find the best subset \(S \subset M\), whereby "best" is yet to be defined. This topic has been extensively studied in the literature, e.g. from the viewpoint of Bayesian decision theory [58, chapter 9]. From this perspective, the usual approach is to pick the decision (here: the final model) \(\hat{\Theta}\) which maximizes the posterior expected utility. Several loss functions for the subset selection problem have been proposed in the literature\(^{58}\). This will be problematic in practice, because there is (again) no clear-cut solution how to specify the "hyperparameter" \(c\).

To overcome this issue, we will use a utility function which has no hyperparameter in the above sense that needs to be specified. For this, it is important to realize that we are really dealing with a two-stage selection problem. First, our goal is to select a suitable subset \(S \subset M\) of models before the evaluation study. Secondly, from \(S\) we aim to select a final model \(m^* \subset S\) with help of the evaluation study for implementation in practice. As our ultimate goal is to obtain a model or medical test with high performance, that is, high sensitivity and specificity, we propose to optimize the expected final model performance

\[
\text{EFP}(S) = \text{EFP}_{\Theta}(S) = \mathbb{E}_{p(\Theta | \pi)} \theta_{m^*} = \sum_{m \in S} \theta_{m^*} \mathbb{P}_{p(\Theta | \pi)}(m^* = m),
\]

depending on the subset of models \(S \subset M\) selected for evaluation. The final model \(m^* = m^*(\hat{\Theta})\) is chosen based on the empirical performances in the evaluation study and thus a random variable, unlike the fixed parameter \(\Theta = (\text{Se}, \text{Sp})\). Our default final model choice will be \(m^* = \arg\max_{m \in S} \min(T_{m}\text{Se}, T_{m}\text{Sp})\), compare Section 3.5. The expectation in (10) is taken with respect to \(p(\Theta | \pi)\), the distribution of \(\hat{\Theta}\) in the evaluation study with \(n\) observations which depends on \(\Theta\). Assuming the prediction models are deterministic, \(p\) is a deterministic transformation of \(\mathcal{G}_n^{(X,Y)}\), the distribution of the feature-label data \((X, Y)\) as described in the beginning of Section 3.3. Of course, both distributions are unknown in practice, because \(\Theta\) is not known.

In formulation (10), EFP depends on \(S \subset M\) for which in principle there are \(2^M\) choices. In order to simplify the decision problem which models to evaluate and its numerical optimization, we rank our models before choosing \(S\) which leaves us with only \(M\) choices instead of \(2^M\), namely \(S = \{1\}, S = \{1, 2\}, \ldots\) or \(S = M\). Consequently, we may write EFP(S) instead of EFP(S) after ranking the models. How to rank models is not as obvious as in the single endpoint case concerned with the overall classification accuracy. Our default choice will be a ranking according to \(\min(T_{m}\text{Se}(V), T_{m}\text{Sp}(V))\), whereby the test statistics are defined as in (5) but based on the validation data. In other words, the models are ranked according to the evidence before the evaluation study that \(H_m\) is false.

The main idea behind our novel optimal EFP selection rule is to maximize an estimate \(\hat{\text{EFP}}(S)\) of EFP(S) before the evaluation study. Taking the viewpoint of Bayesian decision theory, we aim to maximize the posterior expected utility

\[
\hat{\text{EFP}}(S, \pi(\Theta | V)) = \mathbb{E}_{\pi(\Theta | V)} \text{EFP}_{\Theta}(S) = \mathbb{E}_{\pi(\Theta | V)} \mathbb{E}_{p(\Theta | \pi)} \theta_{m^*}.
\]
at the time point of decision-making, right before the evaluation study. The expectation is taken with regard to the posterior distribution $\pi = \pi(\theta \mid V)$ of $\theta$ given the validation data $V$. Different Bayesian models could be applied to obtain said posterior distribution. In this work, we assume that $\pi = (\pi_{\text{Se}}, \pi_{\text{Sp}})$ is composed of two multivariate Beta (mBeta) distributions, $\text{Se} \sim \pi_{\text{Se}}$ and $\text{Sp} \sim \pi_{\text{Sp}}$, which allows to model arbitrary dependency structures between different Beta variables.

As an initial prior distribution $\pi(\theta)$ of $\theta$, we use a vague (uniform) prior by default. This approach allows us to sample $\theta = (\text{Se}, \text{Sp})$ from $\pi$ and subsequently to sample estimates $\hat{\theta} = (\hat{\text{Se}}, \hat{\text{Sp}})$ from the multivariate Binomial likelihood $p = p(\theta \mid \theta)$.

Utilizing this Bayesian model allows us to forecast both, true parameter values and their respective estimates including the dependency structure of both. This enables a simulation of the selection process in the evaluation study and the inherent selection-induced bias problem. In our numerical implementation, we iteratively repeat this process until our computational budget has been exhausted and finally average the results to approximate the (double) expectation in (11) for all $S$. Hereby, we specify a maximal number of models $S_{\text{max}}$ for evaluation in advance to reduce the computational burden of the optimization. A solution to the subset selection problem in dependence of the validation data $V$ is then simply $S = \{1, \ldots, S^*\}$ with

$$S^* = S^*(V) = \arg\max_{1 \leq S \leq S_{\text{max}}} \text{EFP}(S, \pi(\theta \mid V)).$$

Given the assumed prior distribution, the initial model ranking, the statistical modelling assumptions (concerning $\pi$ and $p$), the procedure results in an approximate Bayes action as we approximately maximize the posterior expected utility EFP($S$). The numerical implementation of the selection rule is described in more detail in Appendix S1. In Section 5, the optimal EFP approach is applied to the real data example from Section 2 and to synthetic data from an extensive simulation study.

5 | NUMERICAL EXPERIMENTS

5.1 | Model evaluation study for motivating example

We continue the real-world data example from section 2. We can now apply the optimal EFP selection rule from Section 4 to answer the previously stated question which prediction models to evaluate in the prospective accuracy study.

Based on empirical hold-out validation performance, $S = 16$ models are selected for evaluation. All of them are support vector machines as this model class was apparently best suited to train models with higher sensitivity than specificity which aligns with our study goals ($\text{Se}_0 = 0.80, \text{Sp}_0 = 0.70$). This can be explained by the fact that the particular caret implementation utilizes a hyperparameter to adjust for different misclassification costs. From the 16 selected models, no model empirically performs worse than $\hat{\text{Se}}(V) = 0.96$ and $\hat{\text{Sp}}(V) = 0.80$ in terms of hold-out validation performance.

In the subsequent evaluation study with $n_e = 532$ observations, the best model has an empirical performance of $\hat{\text{Se}}(\mathcal{E}) = 0.929$ and $\hat{\text{Sp}}(\mathcal{E}) = 0.771$. Adjusting for multiple comparisons using the maxT-approach from section 3.4, we obtain lower multiplicity adjusted comparison bounds of 0.863 and 0.722, respectively. As both bounds are greater than the respective thresholds $\text{Se}_0 = 0.80$ and $\text{Sp}_0 = 0.70$ the null hypothesis (1) can be rejected. This result can also be deducted from the corresponding multiplicity adjusted $P$-value which is $0.001 = \max(0.001, 4 \cdot 10^{-6})$ for the best model and as such smaller than the (one-sided) significance level of $\alpha = 0.025$.

Interestingly, all other models fail to replicate their good performance on the evaluation data, at least there is no further significant result. In particular, the seemingly best model before the evaluation study ($\hat{\text{Se}}(V) = 0.973, \hat{\text{Sp}}(V) = 0.881$) now fails to accurately classify healthy subjects as such ($\hat{\text{Se}}(\mathcal{E}) = 0.936, \hat{\text{Sp}}(\mathcal{E}) = 0.614$). In summary, in this illustrative example, the selection and evaluation of multiple promising models, saved the evaluation study from not being able to demonstrate a sufficient screening accuracy for at least one prediction model.

5.2 | Simulation study: machine learning and evaluation

5.2.1 | Setup

The goal of this first simulation study is to assess the properties of different model selection rules and our statistical inference approach under realistic parameter settings in the machine learning context. To this end, we investigate different
methods on a large simulation database which was generated for related methodological comparisons. While the overall binary classification accuracy was our only performance measure of interest in earlier work, we turn to the investigation of sensitivity and specificity as co-primary endpoints in the following. The simulation database consists of 144,000 instances of the complete machine learning pipeline. As shown in Figure 1, a single instance consists of training ($T$), validation ($V$), and evaluation ($E$) datasets which are all comprised of feature-label observations $(x, y)$.

Different binary classification models were trained on the so-called learning data $L = T \cup V$, compare Section 3.1. For that matter, we employed four popular learning algorithms (elastic net, classification trees, support vector machines, extreme gradient boosting) with help of the $R$ package `caret`. More concretely, we trained $M = 200$ initial candidate models, 50 models obtained from randomly sampled hyperparameters per algorithm, on the training data $T \subset L$. The distinct validation data $V = L \setminus T$ is then used for model selection. The models $S \subset M = \{1, \ldots, M\}$ selected for the evaluation study are then retrained on the complete learning data $L = T \cup V$ as this can be expected to slightly increase their predictive performance. The selected models $S \subset M$ then undergo a final assessment on the independent evaluation data $E$ which mimics a diagnostic accuracy study. The goal of this final evaluation study is to estimate sensitivities and specificities of the selected models and to obtain a test decision regarding hypotheses system (3). Ultimately, a final model $m^* \in S$ is selected to be implemented in practice, if positively evaluated, compare Section 3.5.

For each individual simulation instance, the distinct data sets $L$ and $E$ are sampled from the same joint probability distributions $\mathbb{D}_X(Y)$. To generate the whole database, different distributions $\mathbb{D}_X(Y)$ with varying characteristics have been specified. The database consists of linear and nonlinear tasks (ratio 1:2), independent and dependent feature distributions (ratio 1:1) and different disease prevalences $\rho \in \{0.15, 0.3, 0.5\}$ (ratio 1:1:1). The learning data $L$ is of size $n_L \in \{400, 800\}$ (ratio 1:1). The model selection was always conducted on a hold-out validation dataset $V \subset L$ of size $n_V = n_L / 4$. Each learning dataset (including the trained models) was used twice, once in connection with each of the considered evaluation data set sizes $n_E \in \{400, 800\}$ (ratio 1:1).

The true model performances $\theta = (Se, Sp)$ of all models are approximated with high precision on a large population data set ($n_P = 100,000$). This dataset was not used for any other purposes. The truly best model is defined in line with the study goal outlined in Section 3.2 as

$$m^B = \arg\max_{m \in M} \theta_m = \arg\max_{m \in M} \min(Se_m, Sp_m) + \Delta_0.$$  \hspace{1cm} (12)

The parameter $\Delta_0$ is set to 0 in this simulation study which expresses that sensitivity and specificity are deemed to be equally important. The corresponding maximal performance is denoted as $\theta_{m^B} = \theta_{m^*}$. The final model $m^* \in S$ is chosen based on the evaluation data as $m^* = \arg\max_{m \in S} T_m$, compare Section 3.5. In case of a tie, $m^*$ is chosen randomly from the set $\arg\max_{m \in S} T_m$. $R$ packages and scripts related to the simulation study are publicly accessible. 

We initially divided the simulation database into two parts via a stratified random split with regard to the above mentioned factors. The smaller subset with 24,000 instances was used in advance to perform a few methodological comparisons with the goal to fine-tune certain aspects of the optimal EFP selection rule described in Section 4. For instance, we altered the way models are ranked initially and several details associated to numerical complexity of the algorithm (number of iterations, convergence criterion) as outlined in the supplementary material. The final
algorithm was then fixed and investigated on the main part which consists of 120,000 instances, as described in the following.

5.2.2 Goals

Our overall goal is an assessment of important operating characteristics of the employed evaluation strategies. An evaluation strategy is comprised of a selection rule (before evaluation) and a statistical testing procedure (for evaluation). The latter will always be conducted via the maxT-approach as introduced in Section 3.4. The primary objective of this simulation study is therefore the comparison of different selection rules as summarized in Table 1. We are mainly interested in the performance of our novel Bayesian approach designed to optimize the EFP, compare Section 4. This approach is referred to as optimal EFP selection rule. Its main competitor is the within 1 SE rule which S is defined as all models \( m \in \mathcal{M} \) with preliminary balanced accuracy \( \text{bAcc}_m(\mathcal{V}) = (\hat{\text{Se}}(\mathcal{V}) + \hat{\text{Sp}}(\mathcal{V}))/2 \) (estimated on \( \mathcal{V} \)) not more than one standard error below the maximal \( \text{bAcc}_m(\mathcal{V}) \). The default rule selects the single best model in terms of \( \text{bAcc}(\mathcal{V}) \) for evaluation and will serve as a secondary comparator. Additionally, we consider the oracle selection rule which cannot be implemented in practice but gives an insight on the theoretically achievable performance (of selection rules). The oracle rule always selects the truly best model \( S = \{ m^\oplus \} \) as defined in (12). To reduce the computational complexity of the simulation study we hard-thresholded the number of evaluated models to \( S_{\text{max}} = \sqrt{n_\mathcal{L}} \).

Our main research hypothesis going into the simulation is that the optimal EFP improves the expected final performance (EFP) and statistical power compared to the within 1 SE rule while having comparable estimation bias. Moreover, we require that the FWER is still controlled at the desired level which is set to \( \alpha = 2.5\% \) (one-sided) in all scenarios.

5.2.3 Results

Firstly, we investigate how well the feature-label relationships can be learned by the considered algorithms. Figure 2 (left) shows the distribution of the truly best model performance \( \theta_{m^\oplus} \) over all unique 60,000 learning instances. The results are stratified for disease prevalence \( \rho \) and learning sample size \( n_\mathcal{L} \). We observe that \( \theta_{m^\oplus} \) is increasing in both these factors, as expected. Overall, the median of \( \theta_{m^\oplus} \) lies between 79% and 86% which we consider to be a very realistic range. Moreover, Figure 2 (right) shows the distribution of \( \Delta_{m^\oplus} = \hat{\text{Se}}_{m^\oplus} - \hat{\text{Sp}}_{m^\oplus} \) which indicates how balanced the accuracy of the best model is across the diseased and healthy subpopulations. Note that for the oracle rule, the model selection is \( S = \{ m^\oplus \} \) and thus \( m^* = m^\oplus \) is also the final model.

Figure 3 illustrates the true performance of the final selected model \( m^* \in S \) whereby \( S \) depends on the employed selection rule. Depending on the threshold value \( \theta_0 = \hat{\text{Se}}_0 = \hat{\text{Sp}}_0 \), Figure 3 shows the (simulated) probability \( \mathbb{P}(\theta_{m^*} \geq \theta_0 | \delta) \) that the final chosen model has at least a performance of \( \theta_0 = \theta_{m^\oplus} - \delta, \delta > 0 \). The general picture is that the optimal EFP and within 1 SE rules both clearly outperform the default rule. Concerning our main comparison, we observe that the optimal EFP approach uniformly outperforms the within 1 SE rule. The difference is however rather small. For instance, using the optimal EFP rule leads to a probability of 86.1% to select a final model with sensitivity and specificity higher than 0.75 compared to 80.9% with the within 1 SE rule (difference 5.2%; 99% CI: 5.0%-5.5%; average over all scenarios). Further detailed results are provided in Appendix S1.
FIGURE 2  Distribution of $\theta_\oplus$ and $\Delta_\oplus$ stratified for disease prevalence $\rho$ and learning sample size $n_L$. Each boxplot is based on 10,000 distinct points which amounts to 60,000 unique simulation instances in total.

FIGURE 3  Probability of the event that the final model performance $\theta_m = \min(Se_m, Sp_m + \Delta_0)$ exceeds a given threshold $\theta_0 = \theta_\oplus - \delta$. $\theta_\oplus$ is the true performance of the best candidate model. Each scenario of learning and evaluation sample sizes $(n_L, n_E)$ is based on 30,000 repetitions of the complete machine learning pipeline.

Figure 4 shows the rejection rate $rr(\delta) = P(\theta_m \geq \theta_0 | \delta)$ as a function of $\delta = \theta_\oplus - \theta_0$. The case $\delta = 0$ or $\theta_0 = \theta_\oplus$ corresponds to the smallest threshold value $\delta_0$ such that the global null hypothesis (none of the initially trained models has a high enough performance) is true, compare (3). Thus, for $\delta \leq 0$ the global null is true and the rejection rate $rr(\delta)$ should be bounded by $\alpha = 2.5\%$ for all selection rules. Note that this parametrization via $\delta$ is necessary as $\theta_\oplus$ is not fixed over all simulation instances as illustrated in Figure 2. The results clearly show that FWER control is given for all rules in all scenarios. For $\delta \leq 0$ the rejection rate $rr(\delta)$ is actually not zero but positive as reported in table 1 in Appendix S1. In the same table, we also show the FWER when the (pre-)selection process is not incorporated in the test decision. More precisely, we are then looking at the type I error rate of the final statistical test for hypothesis system (1) which is equal to
Rejection rate for the global null hypothesis (all prediction models have performance \( \theta_m \leq \theta_0 \)) in the evaluation study after models have been selected for evaluation by specified selection rule. \( \delta = \theta_0 - \theta_m \) specifies if the global null is true (\( \delta \leq 0 \)) or false (\( \delta > 0 \)).

Each scenario of learning and evaluation sample sizes \( (n_L, n_E) \) is based on 30,000 repetitions of the complete machine learning pipeline. The significance level is \( \alpha = 2.5\% \) (one-sided, dashed horizontal line).

\[ \mathbb{P}(\phi_m^* = 1 | \theta_0 = \theta_m^*). \]

We refer to this quantity as conditional FWER which is necessarily larger than \( rr(0) \) but still controlled, that is, smaller than \( \alpha \) in all cases. Strictly speaking, it is only an upper bound of the proper conditional FWER \( \mathbb{P}(\phi_m^* = 1 | \theta_0 = \max_{m \in s} \theta_m) \) which was not "recorded" in our simulation database.

In the situation \( \delta > 0 \), the threshold value \( \theta_0 \) is small enough such that the global null is no longer true, that is, there is at least one model \( m \in \mathcal{M} \) with true performance \( \theta_m > \theta_0 \) which we seek to detect. Larger rejection rates are thus positive as they correspond to an increased expected statistical power. The expectation hereby is taken over the generative distribution defined by the simulation design. Concerning expected power, the ranking of selection rules is the same as in the EFP comparison (Figure 3). The multiple testing approaches \textit{optimal EFP} and \textit{within 1 SE} outperform the \textit{default} rule clearly. The expected power increase is up to 30\%, depending on the scenario and \( \delta \). The \textit{optimal EFP} rule outperforms the \textit{within 1 SE} rule for all investigated values of \( \delta \). The difference in expected power is more pronounced in the case \( n_E = 800 \) compared to \( n_E = 400 \).

Additional results are reported in Appendix S1. For instance, the properties of the corrected point estimators \( \hat{\theta}_m^* = (\hat{\text{Se}}_m^*, \hat{\text{Sp}}_m^*) \) are assessed, compare Section 3.4. When only a single model is evaluated, \( \hat{\theta} \) corresponds to \( \hat{\theta} \), the raw (slightly regularized) point estimate, compare Section 3.3. For example, the combined mean absolute deviation

\[ \text{MAE}_2 = \mathbb{E}(|\hat{\text{Se}}_m^* - \text{Se}_m^*| + |\hat{\text{Sp}}_m^* - \text{Sp}_m^*|)/2. \]

is decreased when multiple models are evaluated simultaneously (\textit{optimal EFP}: 0.040; \textit{within 1 SE}: 0.048, averaged over both cases \( n_E \in \{400, 800\} \)) compared to the \textit{default rule} rule (0.077). This might seem counter-intuitive at first, but is in line with previous findings.\(^{38}\) This observation can be explained with the on average increased final model performance \( \theta_m^* = \min(\text{Se}_m^*, \text{Sp}_m^*) \) which results in a decreased variance of the binomially distributed estimators. In contrast, the estimation bias is more pronounced when multiple models are evaluated. Note that the bias \( \mathbb{E}(\hat{\theta}_m^* - \theta_m^*) \) is negative (by construction) as we use the corrected estimator \( \hat{\theta}_m^* = \min(\text{Se}_m^*, \text{Sp}_m^*) \), compare Section 3.4.

Finally, we note that the \textit{optimal EFP} rule on average selects fewer models for evaluation when \( n_E = 400 \) compared to the \textit{within 1 SE} rule. On the other hand, when \( n_E = 800 \) samples will be available in the evaluation study, the \textit{optimal EFP} rule selects slightly more models on average, compare table 1 in Appendix S1.
5.3 | Simulation study: LFCs

In the previous section, the operating characteristics of our simultaneous inference framework were assessed in a variety of realistic scenarios in the predictive modeling context. In contrast, we investigate the worst case type I error rate of the maxT-approach for different sample sizes in the following. For that matter, we simulate synthetic data under LFCs as described in Section 3.4 and also close to LFCs.

We assume that $S \in \{1, 10, 20\}$ models are assessed on the evaluation data and consider the case when $Se_0$ and $Sp_0$ are identical. For each simulation run, we randomly draw a binary vector $b$ such that $|b| = S/2$. When $S = 1$, we set $b = (1)$. The true parameters $Se$ and $Sp$ are defined as

$$Se_m^{b,c} = b_m(Se_0 - (m - 1)e) + (1 - b_m),$$
$$Sp_m^{b,c} = b_m + (1 - b_m)(Sp_0 - (S - m)e).$$

Hereby, the parameter $e \in \{0, 0.001, 0.002\}$ expresses how close we are to a true LFC which corresponds to $e = 0$. For instance, for $S = 10$, $Se_0 = Sp_0 = 0.8$, $e = 0.001$, a possible (nearly least favorable) parameter configuration is given by

$$b = (1, 1, 0, 1, 0, 0, 0, 1, 1, 0), \quad (13)$$

$$Se_m^{b,c} = (0.800, 0.799, 1, 0.797, 1, 1, 0.793, 0.792, 1, 1),$$
$$Sp_m^{b,c} = (1, 1, 0.793, 1, 0.795, 0.796, 0.797, 1, 1, 0.800).$$

If $e$ were 0 instead of 0.001 in (13), all elements in $Se^{b,c}$ and $Sp^{b,c}$ different from 1 would change to $Se_0 = Sp_0 = 0.8$ (assuming the same $b$). We mainly consider a (fixed) disease prevalence of $\rho = 0.2$. Hence, there are $n_1 = \rho n$ diseased and $n_0 = n - n_1$ healthy subjects, whereby $n \in \{200, 400, 800, 4000, 20000\}$ in our simulation. Additionally, the scenarios with $e = 0$ were also repeated with a balanced class distribution ($\rho = 0.5$). In all simulations, the true correlation between all estimators which belong to nontrivial variables (true mean not equal to one), for example, $\hat{Se}_1$ and $\hat{Se}_2$ in the example (13), is set to 0.5. Results from a sensitivity analysis regarding this specification are reported at the end of this section.

Any empirical sensitivity is of course independent from any empirical specificity. Data generation, that is, sampling of the similarity matrices $Q^Se$ and $Q^Sp$ (Section 3.3) based on the above specifications, was conducted with help of the bindata package in \texttt{R}.\textsuperscript{61,62} The number of simulation runs was set to $N_{sim} = 10,000$ which results in a (point-wise) upper bound for the SE of the simulated FWER of $0.5/\sqrt{10,000} = 0.005$.

Figure 5 shows the simulation results stratified by all relevant factors. In all scenarios, the results match the theoretical findings from Section 3.4 in that the asymptotic FWER under LFCs converges to $\alpha = 0.025$ as a function of the sample size $n$. When the disease prevalence is increased from $\rho = 0.2$ to 0.5, the FWER is (slightly) decreased. In comparison, the number of models $S$ and the parameter $e$ have a stronger influence on the FWER. In the most extreme case, the FWER is close to 14% for $S = 20$ and $Se_0 = Sp_0 = 0.9$ for $n = 200$ and reaches its target $\alpha = 0.025$ not before $n = 20,000$. Remember that there are only $n_1 = \rho n$ diseased subjects, so the “relevant” sample size is much smaller, particularly in the case $\rho = 0.2$. When we relax the LFCs ($e = 0$) to only slightly less favorable parameters ($e = 0.001$) such as in (13), the FWER declines much faster in $n$. For $Se_0 = Sp_0 = 0.8$ the target significance level is met already at around $n = 400 \approx \exp(6)$ total or $n_1 = 80$ diseased observations. For $Se_0 = Sp_0 = 0.9$, this number more than doubles to roughly $n = 1100 \approx \exp(7)$ for $S = 10$ and $n = 1800 \approx \exp(7.5)$ for $S = 20$.

In an ancillary simulation ($N_{sim} = 5000$), we also investigated results under different true correlation structures of the estimators $\hat{Se}, \hat{Sp}$ (independence, equicorrelation, autocorrelation) and strengths. None of these changes resulted in qualitatively different results than those shown in Figure 5. Additionally, we also investigated the cases $Se_0 = 0.8$ and $Sp_0 = 0.9$ and vice versa in the LFC scenario ($N_{sim} = 5000$). Hereby, it became apparent that the FWER development depends almost solely on the sensitivity value(s), at least in the low prevalence case ($\rho = 0.2$). This can be explained due to the lower number of diseased subjects, and hence lower “relevant” sample size. When $\rho = 0.5$, the problem becomes symmetric and the FWER curves for $\theta_0 = (0.8, 0.9)$ and $\theta_0 = (0.9, 0.8)$ are thus not distinguishable and lie in between those shown in Figure 5, that is, between the cases $\theta_0 = (0.8, 0.8)$ and $\theta_0 = (0.9, 0.9)$.

The worst-case projection of parameter values to one, for example, of $Sp_1, Sp_2, Se_3, \ldots$ in (13), is not really realistic. For instance, in the case $Se_m = 0.9, Sp_m = 1$ and $\rho = 0.2$, the overall accuracy is $\text{Acc}_m = 0.98$. In real-world problems, such a high accuracy is rarely attainable. We therefore also considered an alternative data-generating scenario where (13) is modified such that the overall accuracy for each model $m$ is bounded by 0.95. This change decreased the FWER by 0.5
to 2.0 percent points depending on the exact scenario. Otherwise, the results from this sensitivity analysis do not deviate noticeably from those shown in Figure 5.

6 | DISCUSSION

6.1 | Summary

In this work, we have investigated statistical and study design related methods to improve diagnostic accuracy studies with co-primary endpoints sensitivity and specificity. The main idea is to allow that multiple promising candidate models are assessed simultaneously on the evaluation data, thereby increasing the likelihood of identifying one that performs sufficiently well. In modern medical testing applications this is necessary because often hundreds or thousands of model architectures and preprocessing pipelines are compared on preliminary data, making a final decision before the evaluation study difficult and prone to error. Our approach allows to wait for the test data until a final decision is made and thereby possibly correcting an erroneous model ranking.

The main advantages of the multiple testing approach are the on average increased final model performance and the vastly improved statistical power. As multiple comparisons are now conducted on the final dataset, inferences need to be adjusted via a multiple testing procedure. For that matter, we employed a variation of the so-called maxT-approach which is based on a multivariate normal approximation.\(^3^6\) This framework also enables a corrected, conservative point, and interval estimation. On the downside, the unbiased estimation of the default single model evaluation strategy is lost. An important question of practical interest is how models should be selected for evaluation. Besides the heuristic within 1 SE rule, we also employed the novel optimal EFP rule. This Bayesian approach aims to optimize the EFP before the evaluation study. The underlying model for that matter is based on the hold-out validation data from the model development phase which is readily available before the evaluation study. The comparison regarding all important criteria (performance, power, estimation bias) turns out in favor for the more elaborate optimal EFP approach. However, the advantage relative to the within 1 SE rule was small to moderate in most scenarios. We conclude that the optimal EFP rule yields a vast improvement over the default approach but the conceptually and implementation wise simpler within 1 SE rule yields almost as good of results.
As was laid out in Section 4, an advantage of the optimal EFP selection rule is that it does not involve any tuning parameter, in contrast to other potential decision theoretic approaches and the within k SE rule \((k > 0)\). Another major advantage is that the optimal EFP rule takes into account the evaluation sample size, i.e. it selects more models when more data is (expected to be) available in the evaluation study. This is not the case for the simpler within 1 SE rule which only takes into account estimation uncertainty in the validation stage, before the evaluation study. A general advantage of evaluating multiple models simultaneously is that other decision criteria such as model calibration can be used to guide the final selection once superiority with regard to model accuracy has been established for multiple models. A disadvantage is the increased statistical (and computational) complexity.

6.2 Limitations

Our simultaneous test procedure guarantees asymptotic strong control of the FWER. As illustrated in Section 5.3, the finite sample performance under LFCs can be unsatisfactory, depending on the exact scenario. As expected, the FWER depends primarily on the threshold values \(S_{0}\) and \(S_{p0}\) because the employed normal approximation performs worse near the boundary of the unit interval. Control of the FWER is also worse when the number of models \(S\) is large or when the relevant smaller subsample size, usually the number of diseased, is small. In effect, our framework may not always be suitable in situations where only a very small number of diseased is available for evaluation purposes.

We could aim to improve this limitation. However, as indicated in the main simulation (Section 5.2), the average FWER under realistic conditions is far below the significance level. Note that in contrast to the simulation study presented in Section 5.3, the analysis in Section 5.2 is an average risk assessment. The average is taken with respect to different generative distributions over the relevant parameter values, that is, model performances and their dependency structure. These generative distributions are not known but rather implicit in the sense that the relevant parameters depend on the (feature-label) data generating distribution(s) \(\mathfrak{F}_{X,Y}\) (which are known) and properties of the learning algorithms. Their resulting characteristics are partially described in Figure 2 and deemed to be realistic for real-world prediction tasks. Note that the results from Section 5.2 can be generalized to other prediction tasks and learning algorithms which result in similar generative distribution over parameters \(\theta\) (and dependency structure thereof).

Together, the results from both simulation studies, realistic scenarios in Section 5.2 and the worst-case assessment in Section 5.3, indicate that parameter configurations in reality are rarely least favorable. An associated observation was made at the end of Section 5.3. When the LFC is made only slightly less unfavorable, the FWER declines faster in \(n\). In the machine learning context, it would be reasonable to assume that the larger the number of models \(S\), the less likely least favorable parameters become. After all, the LFC corresponds to the setting that from \(M\) initial candidate models, \(S\) have the exact same performance at the boundary of the null hypothesis and are all selected for evaluation. While this reasoning does apply to the practitioner, it does not apply from a purely frequentist viewpoint as probability statements over true parameters are irrelevant when (worst-case) FWER control in the strong sense is the goal.

The multiple testing approaches certainly increase the flexibility in the evaluation study. However, prediction models still need to be fixed entirely before the study. Verifying this could pose challenges in a regulatory context as merely specifying the learning algorithm and hyperparameters would not be enough in this regard. As the exact resulting models, that is, their weights, often depend also on the random initialization in the training process, only a full specification of model architecture and weights can be considered sufficient in that regard.

It can be seen as a limitation of this work that the learning algorithms implemented in the real data example and simulation study are not necessarily optimal for the considered classification tasks. On the other hand, this also applies to the real world as researchers rarely know which learning algorithms perform poorly, before applying them to the problem at hand. While we expect that the present numerical results generalize to more complex problems (e.g., imaging tasks) and algorithms (e.g., deep neural networks), this was not demonstrated in the present work and should thus be investigated in the future.

6.3 Extensions

Our framework can be extended to classification problems with more than two classes, for example, different disease types or severity grades. More generally, it can be employed in arbitrary prediction tasks for which not only the overall performance is important but also the performance in different subpopulations. So far, we have not considered such problems
in simulation studies. It can however be expected that the operating characteristics of our multiple testing framework will largely depend on the sample size of the smallest subpopulation. This could render the approach practically infeasible for problems with many or an unbalanced distribution of subclasses.

An adaptation which might be considered in the future is hypotheses weighting. The employed maxT-approach can rather easily be adopted for that matter. That is to say, rather then splitting up the significance level \( \alpha \) equally for each of the \( S \) hypotheses, one could spend more on the most promising models. A natural source of information to determine the \( \alpha \) ratio is of course the empirical data from the model development phase. Further research on an optimal weighting and its impact on, for example, statistical power is however necessary. A step further in connecting the evidence before and after the evaluation study would demand Bayesian methods.

Our optimal EFP subset selection algorithm is designed to maximize the expected final model performance. While this is a natural goal, there are many other potential utility functions that could reasonably be optimized before the evaluation study. In particular, utility functions based on statistical power, estimation bias or a weighted combination of different criteria appear reasonable. Our novel approach could also be used here to make the specification of the utility function independent of the subset size and thereby simpler for the practitioner. In its current form, our implementation uses several approximations and can still be improved regarding its numerical efficiency.

ACKNOWLEDGMENTS
The first and last author acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 281474342/GRK2224/1.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS
MW: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft preparation, Writing – review and editing, Visualization; AZ: Writing – review and editing; WB: Conceptualization, Methodology, Writing – review and editing, Supervision, Funding acquisition

DATA AVAILABILITY STATEMENT
The real data utilized for the motivating example is publicly available as described in Section 2. The utilized synthetic data is not shared directly, but R code to reproduce the generation of synthetic data is publicly available as described in Section 5.

ENDNOTES
\(^\ast\)https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf (last accessed 5 January 2022)
\(^\dagger\)https://archive.ics.uci.edu/ml/datasets/cardiotocography (last accessed 5 January 2022)
\(^\ddagger\)This example can be reproduced with code available at https://github.com/maxwestphal/SEPM.PUB (last accessed 5 January 2022).
\(^\S\)https://maxwestphal.github.io/SEPM.PUB/ (last accessed 5 January 2022)

ORCID
Max Westphal \(\odot\) https://orcid.org/0000-0002-8488-758X
Antonia Zapf \(\odot\) https://orcid.org/0000-0001-5339-2472
Werner Brannath \(\odot\) https://orcid.org/0000-0002-8622-3904

REFERENCES
1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.
2. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243.
3. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88.
4. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2017;19(6):1236-1246.
5. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141):20170387.
6. Boulesteix AL. Over-optimism in bioinformatics research. *Bioinformatics*. 2009;26(3):437-439.
7. Boulesteix AL, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Med Res Methodol*. 2009;9(1):85.
43. Committee for medicinal products for human use Guideline on clinical evaluation of diagnostic agents. Eur Med Agency. 2009;7. http://www.emea.europa.eu
44. Korevaar DA, Gopalakrishna G, Cohen JF, Bossuyt PM. Targeted test evaluation: a framework for designing diagnostic accuracy studies with clear study hypotheses. *Diagn Progn Res*. 2019;3(1):1-10.
45. Vach W, Gerke O, Høiland-Carlsen PF. Three principles to define the success of a diagnostic study could be identified. *J Clin Epidemiol*. 2012;65(3):293-300.
46. Westphal M. Simultaneous inference for multiple proportions: a multivariate beta-binomial model; 2019. arXiv preprint arXiv:1911.00098.
47. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. *Am Stat*. 1998;52(2):119-126.
48. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci*. 2001;16(2):101-133.
49. Bolstad WM, Curran JM. *Introduction to Bayesian Statistics*. Hoboken, NJ: John Wiley & Sons; 2016.
50. Genz A, Bretz F, Miwa T, et al. mvtnorm: multivariate normal and t distributions; 2019. R package version 1.0-10.
51. Eckert M, Vach W. On the use of comparison regions in visualizing stochastic uncertainty in some two-parameter estimation problems. *Biom J*. 2020;62(3):598-609.
52. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128.
53. Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *J Biomed Inform*. 2017;69:9-18.
54. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):1-7.
55. Vaicenavicius J, Widmann D, Andersson C, Lindsten F, Roll J, Schön T. Evaluating model calibration in classification; 2019:3459-3467; PMLR.
56. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. *Electronics*. 2019;8(8):832.
57. Molnar C. Interpretable machine learning; 2020. Lulu.com.
58. Lisee F, Miescke KJ. *Statistical Decision Theory: Estimation, Testing, and Selection*. New York, NY: Springer; 2008.
59. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102. doi:10.1002/sim.8086
60. Boulesteix AL, Groenwold RH, Abrahamowicz M, et al. Introduction to statistical simulations in health research. *BMJ Open*. 2020;10(12):e039921. doi:10.1136/bmjopen-2020-039921
61. Leisch F, Weingessel A, Hornik K. On the generation of correlated artificial binary data. Technical report, Working paper, Vienna University of Economics and Business Administration; 1998.
62. Leisch F, Weingessel A, Hornik K. Bindata: generation of artificial binary data; 2012. R package version 0.9-19.
63. Dickhaus T, Gierl J. Simultaneous test procedures in terms of P-value copulae. Technical report, Sonderforschungsbereich 649, Humboldt-Universität zu Berlin; 2012.

**SUPPORTING INFORMATION**
Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Westphal M, Zapf A, Brannath W. A multiple testing framework for diagnostic accuracy studies with co-primary endpoints. *Statistics in Medicine*. 2022;41(5):891-909. doi: 10.1002/sim.9308