Finite-Sample Two-Group Composite Hypothesis Testing via Machine Learning

Tianyu Zhan[a], **Jian Kang**[b]

[a]Data and Statistical Sciences, AbbVie Inc., North Chicago, IL;

[b]Department of Biostatistics, University of Michigan, Ann Arbor, MI

Abstract

In the problem of composite hypothesis testing, identifying the potential uniformly most powerful (UMP) unbiased test is of great interest. Beyond typical hypothesis settings with exponential family, it is usually challenging to prove the existence and further construct such UMP unbiased tests with finite sample size. For example in the COVID-19 pandemic with limited previous assumptions on the treatment for investigation and the standard of care, adaptive clinical trials are appealing due to ethical considerations, and the ability to accommodate uncertainty while conducting the trial. Although several methods have been proposed to control Type I error rates, how to find a more powerful hypothesis testing strategy is still an open question. Motivated by this problem, we propose an automatic framework of constructing test statistics and corresponding critical values via machine learning methods to enhance power in a finite sample. In this article, we particularly illustrate the performance using Deep Neural Networks (DNN) and discuss its advantages. Simulations and two case studies of adaptive designs demonstrate that our method is automatic, general and prespecified to construct statistics with satisfactory power in finite-sample. Supplemental materials are available online including R code and an R shiny app.

Keywords

Confirmatory adaptive clinical trials; Deep neural networks; Efficient inference methods; Neyman-Pearson Lemma; Research assistant tools

---

✉**CONTACT** Tianyu Zhan tianyu.zhan.stats@gmail.com Data and Statistical Sciences, AbbVie Inc., 1 Waukegan Road, North Chicago, IL 60064.

1. Introduction

In simple hypothesis testing, the uniformly most powerful (UMP) level test can be constructed based on Neyman-Pearson Fundamental Lemma (Neyman and Pearson 1933). For a large class of composite hypothesis testing problems for which a UMP test does not exist, there sometimes exists a UMP test among all unbiased tests, under which no alternative has the probability of rejection be less than the size of the test (Lehmann and Romano 2006). However, even in exponential family, it can be challenging to prove the existence and further construct such UMP unbiased tests in finite-sample. For example in the *Behrens-Fisher* problem of testing equality of the means from two normal distributions with unknown variances and unknown variance ratio, the *Welch approximate t-solution* is one of approximate solutions for practical purposes (Welch 1951).

Another important application of composite hypothesis testing is to understand the effect of a treatment relative to the standard of care in randomized clinical trials (RCTs) (EMA 2007; Chen et al. 2014; FDA 2019). In the COVID-19 pandemic with limited knowledge of the treatment profiles, adaptive designs, for example the Adaptive COVID-19 Treatment Trial (ACTT) (National Institutes of Health 2020a), can be more efficient than traditional RCTs by allowing for prospectively planned modifications to design aspects based on accumulated unblinded data (Bretz et al. 2009; Chen et al. 2010). Despite several proposed statistical methods to control the Type I error rate (Bauer and Kohne 1994; Cui, Hung, and Wang 1999; Bretz et al. 2009), how to find a more powerful hypothesis testing strategy remains open. The improved hypothesis testing strategy is especially attractive to patients, because a safe and efficacious drug can be delivered more efficiently and more ethically to meet unfulfilled medical needs.

Motivated by this problem, we propose a general two-stage method via machine learning approaches to conduct two-group composite hypothesis testing when the theoretically most powerful test does not exist or is hard to characterize. We first construct test statistics and then estimate critical values by simulated null samples. Our computational method leverages machine intelligence to establish an automatic framework of performing hypothesis testing with only some basic knowledge of a problem at hand. The proposed method is also general in the sense that it can incorporate existing statistics to seek power improvement if possible. Additionally, our method constructs prespecified decision function for hypothesis testing before observing current data. This is especially appealing to ensure integrity of adaptive clinical trials as considered in Section 5.

In this article, we particularly apply deep neural networks (DNN) to the proposed framework due to its strong functional representation and scalability to large datasets (Goodfellow et al. 2016; Chollet and Allaire 2018). Recently, DNN has been applied to hypothesis testing based on the maximum mean discrepancy (MMD) and its variants (Cheng and Cloninger 2019; Kirchler et al. 2020; Kübler et al. 2020; Liu et al. 2020) and on testing nonlinear effects (Liu and Coull 2017) in large samples, but our aim is to enhance power in a finite sample. More motivations for using DNN are illustrated in Section 3.3, and its advantages are demonstrated by simulations in Appendix D, supplementary material.

The remainder of this article is organized as follows. In Section 2, we describe the problem setup and the motivation of our method. Then we introduce the proposed two-stage DNN-guided hypothesis testing framework in Section 3. Simulations are performed to demonstrate advantages of the proposed method in Sections 4.1 and 4.2. We further apply our method to the ACTT on COVID-19 in Section 5.1, and another adaptive trial Multiple Sclerosis and Extract of Cannabis (MUSEC) in Section 5.2. Concluding remarks are provided in Section 6.

2. Problem Setup

2.1. Composite Hypothesis Testing with Two Groups of Data

Based on the motivating example of the ACTT on COVID-19, we consider a composite hypothesis testing problem using two groups of independent data $x_j, j = 1, 2$, of equal size n where samples in group j are independent and identically distributed (i.i.d.) with a probability mass function (pmf) or a probability density function (pdf) denoted as $f(x; \theta_j, \eta_j)$. The parameter of interest $\theta_j \in \mathbb{R}$ is considered as a scaler quantity, while the nuisance parameters $\eta_j \in \mathbb{R}^w$ is a vector of dimension w in group j. For example in the ACTT, group 1 is the standard of care, group 2 is the remdesivir under evaluation, and θ_j can be the probability of achieving hospital discharge at Day 14 (National Institutes of Health 2020c; Gilead Inc. 2020) in group j, $j = 1, 2$. Suppose that the following composite null hypothesis H_0 is to be tested against a composite alternative hypothesis H_1 with a one-sided Type I error rate controlled at a,

$$H_0: \theta_1 = \theta_2, \text{ versus } H_1: \theta_1 < \theta_2, \tag{1}$$

where a hypothesis is said to be composite if it contains a class of distributions (Lehmann and Romano 2006). Under H_0 in (1), the class of distribution is denoted as $\mathscr{F}_0 = \{ f(x; \theta_{12}, \eta_1), f(x; \theta_{12}, \eta_2); \theta_{12} \in \mathbb{R}, \eta_1 \in \mathbb{R}^w, \eta_2 \in \mathbb{R}^w \}$; while $\mathscr{F}_1 = \{ f(x; \theta_1, \eta_1), f(x; \theta_2, \eta_2); \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}, \theta_1 < \theta_2, \eta_1 \in \mathbb{R}^w, \eta_2 \in \mathbb{R}^w \}$ under H_1.

With a finite sample, one cannot simultaneously control the probability of making a Type I error of rejecting H_0 when it is true and the probability of making a Type II error of accepting H_0 when it is false (Lehmann and Romano 2006). It is customary to minimize the Type II error rate, which is one minus *power*, subject to an upper bound a on the Type I error rate (Lehmann and Romano 2006). However, in the composite hypothesis testing, the UMP level a test or the UMP unbiased level a test may not exist when there is no single test that is the most powerful one among all tests or all unbiased tests under every distribution in $\mathscr{F}_1$ under H_1. For example, on testing the location parameter with a single observation from Cauchy distribution, the functional form of the most powerful test depends on the underlying location parameter and hence, no UMP test exists (Lehmann and Romano 2006). Even within the exponential distribution family, it can be challenging to prove the existence and further characterize such optimal test in a finite sample (Lehmann and Romano 2006). For example in the famous *Behrens-Fisher problem* of testing the equality of the means from two normal distributions with unknown variances and unknown ratio of variances, the *Welch approximate t*-test is a popular approximate solution in practice (Welch 1951).

In this article, we propose a novel two-stage method to perform hypothesis testing with the aim of increasing power in finite-sample problems when the theoretically optimal test does not exist or is hard to characterize.

2.2. Motivated Problem: Simple Hypothesis Testing

We motivate our proposed method from the Neyman–Pearson Lemma in a simple hypothesis testing problem on θ with known nuisance parameters $\boldsymbol{\eta}_0$ using a single group of data $\boldsymbol{x}$ with pmf or pdf $f(\boldsymbol{x}; \theta, \boldsymbol{\eta}_0)$. The objective is to test the following simple H'_0 against simple H'_1,

$$H'_0: \theta = t_0, \text{ versus } H'_1: \theta = t_1, \tag{2}$$

where $t_0 < t_1$ are two constants. We define a test function $\phi(\boldsymbol{x}) = 1$ if H'_0 is rejected, and $\phi(\boldsymbol{x}) = 0$ otherwise. The rejection region is given by $R(\phi) = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = 1\}$. Based on the Neyman–Pearson Lemma, a test $\phi(\boldsymbol{x})$ that satisfies

$$\begin{aligned} \phi(\boldsymbol{x}) &= I\left\{ f(\boldsymbol{x}; t_1, \boldsymbol{\eta}_0) > c^\dagger f(\boldsymbol{x}; t_0, \boldsymbol{\eta}_0) \right\} \text{ and} \\ \alpha &= \Pr_{H'_0}\left\{ \boldsymbol{x} \in R(\phi) \right\}, \end{aligned} \tag{3}$$

for some $c^\dagger \geq 0$ and $I(\cdot)$ as an event indicator function, is a UMP level α test (Lehmann and Romano 2006). For example, when data $\boldsymbol{x}$ of size n are assumed to follow a normal distribution with unknown mean θ and known variance σ_0^2, the UMP level α test of (2) is the z-test. It rejects H'_0 if $\hat{\theta}(\boldsymbol{x}) > \sigma_0 z_{1-\alpha}/\sqrt{n} + t_0$, where $\hat{\theta}(\boldsymbol{x})$ is the sample mean as a sufficient statistic for θ, $z_u = \Phi^{-1}(u)$, and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution.

As an alternative, we formulate the hypothesis testing in the context of a binary classification problem to categorize whether $\boldsymbol{x}$ is sampled from H'_1 or from H'_0. We introduce a latent variable $y \in \{0, 1\}$ indicating where $\boldsymbol{x}$ is drawn. Given $y = k$, the pdf or pmf is equal to $f(\boldsymbol{x}; t_k, \boldsymbol{\eta}_0)$, for $k = 0, 1$. Therefore, the rejection region in (3) can be expressed as

$$\begin{aligned} f(\boldsymbol{x}; t_1, \boldsymbol{\eta}_0) > c^\dagger f(\boldsymbol{x}; t_0, \boldsymbol{\eta}_0) &\Leftrightarrow \frac{f(\boldsymbol{x}; t_1, \boldsymbol{\eta}_0)\Pr(y=1)}{f(\boldsymbol{x}; t_0, \boldsymbol{\eta}_0)\Pr(y=0)} > c' \\ &\Leftrightarrow \frac{\Pr(y=1 \mid \boldsymbol{x}, \boldsymbol{\eta}_0)}{1 - \Pr(y=1 \mid \boldsymbol{x}, \boldsymbol{\eta}_0)} > c' \\ &\Leftrightarrow d(\boldsymbol{x}; \boldsymbol{\eta}_0) > \log(c') \equiv c, \end{aligned} \tag{4}$$

where $\Leftrightarrow$" reads if and only if, $d(\boldsymbol{x}; \boldsymbol{\eta}_0) = \text{logit} \{\Pr(y = 1|\boldsymbol{x}, \boldsymbol{\eta}_0)\}$, and $\text{logit}(u) = \log [u/(1-u)]$ for some constants $c^\dagger$, c' and c. A larger value of $d(\boldsymbol{x}; \boldsymbol{\eta}_0)$ indicates that $\boldsymbol{x}$ is more likely to be drawn from H'_1 as compared to H'_0. The constant c in (4) is computed to control the Type I error rate at α,

$$\Pr_{H'_0}\{d(\boldsymbol{x}; \boldsymbol{\eta}_0) > c\} = \alpha. \tag{5}$$

Based on the sufficient conditions of the Neyman-Pearson Lemma, any test that satisfies (4) and (5) is most powerful for testing the simple null H'_0 at level α.

Before considering the composite hypothesis testing in (1), we first define a set of sufficient statistics $t^{(s)} = \left\{ \hat{\theta}(x_1), \hat{\theta}(x_2), \hat{\eta}(x_1), \hat{\eta}(x_2) \right\}$ of dimension $w^{(s)}$, where $\hat{\theta}(x_j)$ and $\hat{\eta}(x_j)$ are sufficient statistics of their true parameters θ_j and η_j based on data x_j from the distribution function $f(x; \theta_j, \eta_j)$, for group $j = 1, 2$. The superscript "(s)" indicates that $t^{(s)}$ is used in the formulation of statistics in Section 3.2. In some problems, for example the scale-uniform distribution considered in Section 4.1, $\hat{\theta}(x_j)$ can be a vector even if θ_j is a scalar.

A direct generalization from the simple hypothesis testing (2) to the composite hypothesis testing (1) is challenging when the likelihood ratio in (4) depends on unknown parameters θ_1, θ_2, η_1 and η_2, because statistics are functions of only data. Following the formulation of $d(x; \eta_0)$ in (4), we intend to identify a statistic $d\{t^{(s)}\}$ for composite hypothesis testing such that

$$d\left\{t^{(s)}\right\} = \text{logit}\left[\text{Pr}\left\{y = 1 \mid t^{(s)}\right\}\right], \tag{6}$$

and then obtain the critical value c to control Type I error rates at α under H_0,

$$\text{Pr}_{H_0}\left[d\left\{t^{(s)}\right\} > c\right] = \alpha. \tag{7}$$

However, the functional form of $d\{t^{(s)}\}$ in (6) may not be known explicitly or is even intractable. Another complication is to study the distribution of $d\{t^{(s)}\}$ under H_0 in a finite sample in order to calculate the critical value of c in (7). In the following section, we introduce our two-stage method to first characterize the statistic $d\{t^{(s)}\}$ and then estimates the corresponding critical value c. We particularly leverage DNN in the proposed method and discuss its advantages.

3.   DNN-Guided Hypothesis Testing Method

We first provide a short review on DNN in Section 3.1. Then we illustrate our DNN-guided hypothesis testing method by first approximating the test statistics in Section 3.2 and then estimating critical values in Section 3.3. The conduct of the proposed method on observed data is demonstrated in Section 3.4.

3.1.   Review on Deep Neural Networks (DNN)

DNN defines a mapping $y = q(t; \psi)$ and learns the value of the parameters ψ that result in the best function approximation of output label y based on input data t (Goodfellow et al. 2016). The *deep* in DNN stands for successive layers of representations. The last-layer activation function can be chosen as the sigmoid (expit) function sigmoid$(u) = 1/[1 + \exp(-u)]$ for binary classification and the *linear* function for a continuous variable approximation, while the inner-layer activation function is usually the *Rectified Linear Unit (ReLU)* function defined by ReLU$(u) = \max(0, u)$ (Chollet and Allaire 2018). In obtaining $\hat{\psi}$ to estimate ψ based on a nonconvex loss function, we use RMSProp (Hinton 2012) which has been shown to be an effective and practical optimization algorithm for deep neural networks (Goodfellow et al. 2016).

A proper DNN structure and other hyperparameters are usually chosen by cross-validation with 80% as the training data and the remaining 20% as the validation data (Goodfellow et al. 2016). We start with an architecture with a relatively large capability and a large number of training epochs in the optimization algorithm to reduce the training error. We then apply regulation approaches to increase the generalizability of the model, for example the dropout technique which randomly sets a number of features of the layer as zeros during training, and the mini-batch approach which stochastically selects a small batch of data in computing gradient in the algorithm. We further propose several structures around this suboptimal solution as the candidate pool and select the final skeleton with the smallest model fitting error. Sensitivity analysis in Section 4.1 shows that the performance of our method is robust to different DNN structures (supplementary Table 3).

3.2. Approximating the Test Statistics via DNN

In the first stage, we train a DNN to construct the test statistic $d\{t^{(s)}\}$ in (6) by using Monte Carlo samples. Note that $t^{(s)}$ does not require minimal sufficient statistics, but only sufficient statistics, which are usually straightforward to obtain based on parametric assumptions. One can also substitute them by order statistics as trivial sufficient statistics, or a vector of key summary statistics, such as mean, median, standard deviation, sample quantiles, etc. In practice, one may evaluate different choices of $t^{(s)}$ to determine the empirically optimal one. For a composite hypothesis testing problem in (1), we define $\Theta_1 \subseteq \mathbb{R}$ as a neighborhood of the true value of θ_1. The corresponding notations for θ_2, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are $\Theta_2 \subseteq \mathbb{R}$, $\boldsymbol{H}_1 \subseteq \mathbb{R}^w$ and $\boldsymbol{H}_2 \subseteq \mathbb{R}^w$, respectively. We assume that Θ_1, Θ_2, $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are all compact.

To generate the training data, we first simulate A sets of features $\{\theta_{1,a}, \theta_{2,a}, \boldsymbol{\eta}_{1,a}, \boldsymbol{\eta}_{2,a}\}_{a=1}^{A}$ from uniform distributions in their corresponding parameter spaces. Then within each set a, for $a = 1, \ldots, A$, we simulate B_0 samples under H_0 with distribution $f(x; \theta_{1,a}, \boldsymbol{\eta}_{1,a})$ for group 1 and $f(x; \theta_{1,a}, \boldsymbol{\eta}_{2,a})$ for group 2, and simulate B_1 samples under H_1 with distribution $f(x; \theta_{1,a}, \boldsymbol{\eta}_{1,a})$ for group 1 and $f(x; \theta_{2,a}, \boldsymbol{\eta}_{2,a})$ for group 2. In each sample b, for $b = 1, \ldots, \{A \times (B_0 + B_1)\}$, we calculate the vector of sufficient statistics $t_b^{(s)} = \{\widehat{\theta}(\boldsymbol{x}_{1,b}), \widehat{\theta}(\boldsymbol{x}_{2,b}), \widehat{\boldsymbol{\eta}}(\boldsymbol{x}_{1,b}), \widehat{\boldsymbol{\eta}}(\boldsymbol{x}_{2,b})\}$ based on data $\boldsymbol{x}_{1,b}$ and $\boldsymbol{x}_{2,b}$ from two groups to establish the training data $\{t_b^{(s)}\}_{b=1}^{A \times (B_0 + B_1)}$. The support of $t^{(s)}$ is denoted as $\boldsymbol{T}^{(s)} \subseteq R^{w^{(s)}}$. For a sample with index b, we further define a classification label $y_b^{(s)}$ taking value either 0 or 1, where the event $\{y_b^{(s)} = k\}$ indicates a sample being drawn from the distribution under H_k, for $k = 0, 1$.

Next, we train the test statistic DNN (TS-DNN) with the *ReLU* as the inner-layer activation function and the *sigmoid* as the last-layer activation function. In the training process, DNN seeks a solution $\widehat{\boldsymbol{\psi}}^{(s)}$ which maximizes the log-likelihood function,

$$\widehat{\boldsymbol{\psi}}^{(s)} = \operatorname{argmax}_{\boldsymbol{\psi}^{(s)}} \sum_{b=1}^{A \times (B_0 + B_1)} \left(\left\{ 1 - y_b^{(s)} \right\} \log \left[1 - p \left\{ t_b^{(s)}; \boldsymbol{\psi}^{(s)} \right\} \right] \right.$$
$$\left. + y_b^{(s)} \log p \left\{ t_b^{(s)}; \boldsymbol{\psi}^{(s)} \right\} \right),$$

(8)

where $p\{t^{(s)}; \boldsymbol{\psi}^{(s)}\} = \text{sigmoid}\,[q\{t^{(s)}; \boldsymbol{\psi}^{(s)}\}]$, $q\{t^{(s)}; \boldsymbol{\psi}^{(s)}\}$ is the linear predictor of DNN, and $\boldsymbol{\psi}^{(s)}$ is a stack of the weight and the bias parameters from all layers in DNN. Essentially, DNN obtains $\widehat{\boldsymbol{\psi}}^{(s)}$ by (8) such that sigmoid $\left[q\left\{ t^{(s)}; \widehat{\boldsymbol{\psi}}^{(s)} \right\} \right]$ approximates the underlying classification probability $\Pr\{y = 1 | t^{(s)}\} = \text{sigmoid}\,[d\{t^{(s)}\}]$ in (6). We use *sigmoid* as the last-layer activation function for TS-DNN to follow the formulation based on Equation (4). Before DNN training, we normalize training data with mean zero and unit standard deviation to mitigate the potential gradient issue of *sigmoid*. In supplementary Table 7, we also evaluate DNN with *softmax* as the last-layer activation function.

The approximation error of using a DNN with linear predictors $q\{t^{(s)}; \boldsymbol{\psi}^{(s)}\}$ to approximate the objective function $d\{t^{(s)}\}$ in (6) is defined by the following uniform maximum error (Yarotsky 2017),

$$\left| d - q \right|_\infty = \sup_{t^{(s)} \in T^{(s)}} \left| d\left\{ t^{(s)} \right\} - q\left\{ t^{(s)}; \boldsymbol{\psi}^{(s)} \right\} \right|.$$

(9)

Many theoretical investigations have been done to show that an underlying DNN $q\{t^{(s)};$ $\boldsymbol{\psi}^{(s)}\}$ can approximate an objective function $d\{t^{(s)}\}$ in a certain function class to have $|d - q|_\infty = $ upper bounded by a given tolerance, for example Hölder functions (Chen et al. 2019; Shen, Yang, and Zhang 2019), functions in Sobolev spaces (Mhaskar 1996; Yarotsky 2017), and Lipschitz-continuous functions (Bach 2017).

We provide some discussion on choosing the training features $\{\theta_{1,a}, \theta_{2,a}, \boldsymbol{\eta}_{1,a}, \boldsymbol{\eta}_{2,a}\}_{a=1}^A$. If one sets $\theta_{1,a} = \theta_{2,a}$ and $\boldsymbol{\eta}_{1,a} = \boldsymbol{\eta}_{2,a}$, then the distribution family $\mathscr{F}_0$ under H_0 is exactly the same with $\mathscr{F}_1$ under H_1. The resulting solution will be a random classifier with no practical use. On the other hand, if $\theta_{2,a}$ is way larger than $\theta_{1,a}$, then the classification error goes to zero but the trained DNN loses generalizability when $\theta_{2,a}$ is moderately larger than $\theta_{1,a}$. For a set of given $\{\theta_{1,a}, \boldsymbol{\eta}_{1,a}, \boldsymbol{\eta}_{2,a}\}$, we suggest choosing a $\theta_{2,a}$ such that our DNN-based method reaches a moderate level of power. This can be approximated by some known tests, for example the Student's *t*-test. As further demonstrated in Sections 4 and 5, our method has a satisfactory performance in validations when θ_2 is different from this training magnitude.

3.3. Approximating the Critical Values via DNN

In the second stage, we train another DNN to estimate the critical value c in (7) based on simulated samples under H_0.

We construct the training data as $\left\{ t_a^{(c)} \right\}_{a=1}^A$ of dimension $(1+2 \times w)$ and size A, where $t_a^{(c)} = \left(\theta_{1,a}, \boldsymbol{\eta}_{1,a}, \boldsymbol{\eta}_{2,a} \right)$ are the design features from the previous section. The superscript "(c)"

in $t^{(c)}$ and other notations indicates that they pertain to the estimation of critical values. Within each feature $t_a^{(c)}$, we further simulate B' null data under H_0 with group 1 data from the distribution function $f(x;\ \theta_{1,a},\ \boldsymbol{\eta}_{1,a})$ and group 2 data from $f(x;\ \theta_{1,a},\ \boldsymbol{\eta}_{2,a})$. Their test statistics based on the TS-DNN in the first stage are computed at $\left\{q\left[t_{b'}^{(s)};\widehat{\boldsymbol{\psi}}^{(s)}\right]\right\}_{b'=1}^{B'}$. The output label $y_a^{(c)}$, for $a = 1,\ \ldots,\ A$, is set at the empirical upper α quantile in those test statistics to satisfy (7). This procedure is performed in a similar fashion as the parametric bootstrap (Efron and Tibshirani 1994) to construct the null distribution of the statistic $q\left[t^{(s)};\widehat{\boldsymbol{\psi}}^{(s)}\right]$, and then to obtain the corresponding critical values. Under a general composite null hypothesis H_0 in (1), the critical value $y_a^{(c)}$ may depend on the unknown θ_{12}, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$, where θ_{12} is the common value of θ_1 and θ_2 under H_0. Therefore, we train a critical value DNN (CV-DNN) $q\left\{t^{(c)};\widehat{\boldsymbol{\psi}}^{(c)}\right\}$ to estimate $y_a^{(c)}$ by *linear* function as the last-layer activation function, and the mean squared error (MSE) as the loss function. As illustrated in Appendix E, supplementary material, the proposed procedure saves computational time as compared with the parametric bootstrap method.

A diagram is provided to streamline our two-stage method of approximating the test statistics and estimating the critical values training two different DNNs (Figure 1).

At this point, we provide some remarks on using DNN in this framework. First of all, as compared with some simple models like the generalized linear model (GLM) or the linear model (LM), DNN has a stronger functional representation to approximate Lipschitz-continuous functions (Bach 2017) or functions in Sobolev spaces (Yarotsky 2017). As demonstrated in Section 4.1, if one substitutes the TS-DNN with GLM and substitutes the CV-DNN with LM in our proposed method, then the Type I error rate in validation is not controlled at α (Appendix D, supplementary material). Secondly, compared to other nonparametric and machine learning methods, DNN provides a scalable way of training on large datasets (Goodfellow et al. 2016). The number of simulated training features (A, B_0 and B_1 in Figure 1) can be set sufficiently large to give DNN satisfactory performance (Goodfellow et al. 2016). We compare its performance with support vector machine (SVM; Boser, Guyon, and Vapnik 1992) and random forest (RF; Ho 1995) with varying hyperparameters and with the same computational cost in Section 4.1 (Appendix D, supplementary material). DNN has a more accurate Type I error rate control even with reduced numbers of training features. Moreover, DNN is able to automatically identify important and relevant features from data to characterize the objective function without manual feature engineering (Chollet and Allaire 2018).

3.4. Hypothesis Testing based on Observed Data $\tilde{x}_1$ and $\tilde{x}_2$

Now we are ready to conduct hypothesis testing on (1) with observed data $\tilde{x}_1$ from group 1 and $\tilde{x}_2$ from group 2. We first calculate the input data for the TS-DNN at $\tilde{t}^{(s)} = \left\{\widehat{\theta}(\tilde{x}_1), \widehat{\theta}(\tilde{x}_2), \widehat{\boldsymbol{\eta}}(\tilde{x}_1), \widehat{\boldsymbol{\eta}}(\tilde{x}_2)\right\}$ and then compute its test statistic $q\left\{\tilde{t}^{(s)};\widehat{\boldsymbol{\psi}}^{(s)}\right\}$. Let $\tilde{\theta}(x)$ and $\tilde{\boldsymbol{\eta}}(x)$ be unbiased or consistent estimators for θ and $\boldsymbol{\eta}$, respectively. The critical value is

computed at $q\left\{\tilde{\boldsymbol{t}}^{(c)}; \widehat{\boldsymbol{\psi}}^{(c)}\right\}$, where $\tilde{\boldsymbol{t}}^{(c)} = \left\{\tilde{\theta}(\tilde{\boldsymbol{x}}_{12}), \tilde{\boldsymbol{\eta}}(\tilde{\boldsymbol{x}}_1), \tilde{\boldsymbol{\eta}}(\tilde{\boldsymbol{x}}_2)\right\}$ and $\tilde{\boldsymbol{x}}_{12} = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2)$. Note that $\tilde{\theta}(\tilde{\boldsymbol{x}}_{12})$ is an unbiased or consistent estimator of $\theta_1 = \theta_2$ under H_0. Finally, H_0 in (1) is rejected if $q\left\{\tilde{\boldsymbol{t}}^{(s)}; \widehat{\boldsymbol{\psi}}^{(s)}\right\} > q\left\{\tilde{\boldsymbol{t}}^{(c)}; \widehat{\boldsymbol{\psi}}^{(c)}\right\}$, but not rejected otherwise.

4. Experiments

4.1. Scale-Uniform Distribution

In this section, we consider the scale-uniform distribution unif($[1 − k]\theta, [1 + k]\theta$) with the positive parameter of interest θ and a known design parameter $k \in (0, 1)$, where unif denotes the uniform distribution (Galili and Meilijson 2016). This distribution has wide applications, for example the product inventory management in economics (Wanke 2008) and the inverse transform sampling (Vogel 2002). This distribution is also an example where the distribution family does not satisfy the usual differentiability assumptions leading to the Cramér-Rao bound and efficiency of MLEs (Lehmann and Romano 2006; Galili and Meilijson 2016). We demonstrate how to use our proposed DNN-based method to assist statistical research.

With two groups of data $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of equal size $n = 20$, we are interested in testing H_0 against H_1 in (1) with a one-sided Type I error rate $\alpha = 0.05$. Although the likelihood ratio test (LRT) based on the asymptotic Chi-square distribution is not valid (Lehmann and Romano 2006), one can construct a LRT type statistic $T_1 = \max(\boldsymbol{x}_2)/\max(\boldsymbol{x}_1)$ because the likelihood ratio is a monotonically increasing function of $\max(\boldsymbol{x}_2)/\max(\boldsymbol{x}_1)$. The critical value of T_1 given known k can be computed by simulations, because T_1 is a pivotal quantity in the sense that its distribution is independent from the unknown quantity θ (Lehmann and Romano 2006). Next, we apply our proposed method, denoted as "DNN," as a benchmark to evaluate the power performance of T_1.

The neighborhoods of the true values of θ_1 and θ_2 are considered at $\Theta = (0.5, 10) \subseteq \mathbb{R}$, and $K = (0, 1) \subseteq \mathbb{R}$ for k. Following the procedures as described in Section 3, we simulate $A = 500$ features with θ_1 from Θ and k from K, and further set θ_2 under H_1 at a value for our approach to reach approximately 90% power. We choose $B_0 = B_1 = 10^4$. In this case, the training data size for TS-DNN is $A \times (B_0 + B_1) = 10^7$, while the training data size for CV-DNN is $A = 500$. The input vector for the TS-DNN is $\boldsymbol{t}^{(s)} = \left\{\widehat{\theta}(\boldsymbol{x}_1), \widehat{\theta}(\boldsymbol{x}_2), k\right\}$, where $\widehat{\theta}(\boldsymbol{x}_j) = \left\{\min(\boldsymbol{x}_j), \max(\boldsymbol{x}_j)\right\}$ are sufficient statistics for θ_j in group $j = 1, 2$ (Galili and Meilijson 2016). By cross-validation, the final DNN structure is selected as the one with the smallest validation error from 6 candidate structures, which are all combinations of the number of layers at 2 and 3, and the number of nodes per layer at 50, 100 and 150. The number of epochs is 10, the batch size is 10^4, and the dropout rate is set at 0.1. As illustrated in Section 3.1, the above candidate structure pool is formulated by evaluating a wider and deeper DNN structure with a certain dropout rate and a small batch size introduced to reduce overfitting (Chollet and Allaire 2018). Sensitivity analyses in Appendix C, supplementary material show that the performance of our method is robust under different properly chosen hyperparameters of DNN, including DNN structure, batch size, dropout rate, last-layer activation function of TS-DNN. Specifically, the power performance is consistent with varying batch size of TS-DNN at 100 and 1000. For other problems in general, one can

implement cross-validation to choose the empirically optimal batch size, which may be less than 100. The whole training process is implemented by the R package `keras` (Allaire and Chollet 2020).

In the CV-DNN, the training data is of size A with $t^{(c)} = (\theta_1, k)$, and the output is computed on $B' = 10^6$ samples under H_0. A number of epochs at 10^3, a batch size of 10, and a dropout rate of 0.1 are used in this DNN training. When estimating the critical value, we use the sample mean as an unbiased estimator of $\theta_1 = \theta_2$ under H_0. The number of validation iterations is 10^6. Unless otherwise specified, the above set-up parameters are used throughout this article. Note that the size of training data, for example A, can be sufficiently large to give DNN a satisfactory performance, and the number of iterations, for example B', can be increased to improve the precision of numeric calculations.

In Table 1, we evaluate the performance of our method DNN versus the likelihood ratio based statistic T_1, the Student's t-test, Wilcoxon rank-sum test and the maximum mean discrepancy (MMD) considered in Cheng and Cloninger (2019), Kirchler et al. (2020), Kübler et al. (2020), and Liu et al. (2020) on testing means under four scenarios with varying k and varying θ_1. Since MMD is computationally intensive, we use 10^4 simulation iterations in its validation. In each scenario, the first row evaluates the Type I error rate under H_0, while the other three rows are for power under H_1. The value of θ_2 in the third row is the same as that in the training data. The second row captures a lower magnitude and the fourth row considers a higher one. Across all scenarios, all methods have controlled Type I error rates at $\alpha = 0.05$. Under H_1, DNN is generally more powerful than those alternatives. For example when $k = 0.2$, $\theta_1 = 5$ and $\theta_2 = 5.222$, DNN has a power of 80.2% in testing H_0, as compared with 67.7% for T_1, 31.0% for the Student's t-test, 29.2% for the Wilcoxon rank-sum test, and 10.8% for MMD. Sensitivity analyses in the supplemental materials show that our DNN-based method has a consistent power gain under varying scenarios (Appendix A, supplementary material). In Appendix B, supplementary material, we also evaluate different choices of $t^{(s)}$ as input data for TS-DNN, for example key summary statistics or order statistics, and observe that $t^{(s)}$ with sufficient statistics has the best power performance in this example. Advantages of DNN over other machine learning methods, for example, support vector machine (SVM; Boser, Guyon, and Vapnik 1992) and random forest (RF; Ho 1995) are also discussed (Appendix D, supplementary material). Additionally, we find that the LRT based statistic T_1 has a similar or slightly higher power than DNN when $k = 0.8$ and $\theta_1 = 1$, but is less powerful in other scenarios.

The superior benchmark performance of DNN suggests that maybe a better statistic can be constructed from T_1. Since the input data $t^{(s)}$ of TS-DNN contains both $\min(x_j)$ and $\max(x_j)$ as sufficient statistics for θ_j, we can modify T_1 with $\min(x_j)$ incorporated to obtain the following statistic T_2 as another pivotal quantity given known k,

$$T_2 = \frac{w\min(x_2)/(1-k) + (1-w)\max(x_2)/(1+k)}{w\min(x_1)/(1-k) + (1-w)\max(x_1)/(1+k)}, \tag{10}$$

where $\min(x_j)/(1-k)$ and $\max(x_j)/(1+k)$ are consistent estimators of θ_j (Galili and Meilijson 2016), and $w = [(1-k)^2]/[(1-k)^2 + (1+k)^2]$ as an inverse variance weight.

To investigate if we can further improve power based on T_2, we include it as another input data in $t^{(s)}$ when training TS-DNN, and denote this method as "DNN-T2." As can be seen from Table 1, DNN-T2 is more powerful than T_2 in most scenarios with a power gain as large as 2.7% when $\theta_1 = 5$ and $k = 0.2$, while T_2 is slightly more powerful than DNN-T2 with difference no larger than 0.2% when $\theta_1 = 1$ and $k = 0.2$. On the one hand, one may argue that there is essentially limited power improvement by constructing DNN-T2 as compared with T_2. This evidence supports that T_2 is a feasible statistic with satisfactory performance in practice. On the other hand, our well-trained DNN-T2 itself can also be applied in practice with its available functional form. Moreover, if not satisfying with the power performance of T_2 and DNN-T2, one may conduct further research to find a better statistic, for example replacing the consistent estimator of θ by its unbiased estimator (Galili and Meilijson 2016), accommodating correlation between $\min(x_j)$ and $\max(x_j)$ when calculating w in (10), etc. The new statistics can also be incorporated to the input data of TS-DNN training to find a better one if possible.

In this example, our proposed automatic method is used as a benchmark to evaluate other statistics which are carefully designed by human intelligence, and is also able to leverage existing statistics to identify a statistic with potentially higher power.

4.2. Student's t-Distribution

In this section, we consider a problem of testing the numbers of degrees of freedom in the Student's t-distribution with two groups of data. In robust estimation and modeling, the t-distribution provides a useful extension from normality assumption to mitigate the impact of outliers (Lange, Little, and Taylor 1989; Pinheiro, Liu, and Wu 2001). Since the Student's t-distribution is not an exponential family, we apply our method to find a better testing strategy as compared with common alternatives.

Two groups of data with equal size $n = 200$ are used to test H_0 in (1) with a one-sided Type I error rate 0.05 where θ_1 and θ_2 denote the number of degrees of freedom from two groups, respectively. In this example, we consider a constrained hypothesis testing problem where both θ_1 and θ_2 are within 3 and 10. Therefore, we simulate underlying θ_1 and θ_2 from $\Theta = (3, 10) \subseteq \mathbb{R}$ when generating training data for TS-DNN and CV-DNN. Since the sufficient statistics for θ are not common ones, we consider the training input data for TS-DNN as $t^{(s)} = \left\{ \widehat{\theta}(x_1), \widehat{\theta}(x_2) \right\}$, where $\widehat{\theta}(x_j)$ is a vector of key summary statistics: mean, median, standard deviation, minimum, maximum, first quartile, and third quartile for data x_j from group $j = 1, 2$. The training data for CV-DNN is $t^{(c)} = (\theta_1)$ as the common θ under H_0, and we set $B' = 10^5$ in this example. In the testing stage, the maximum likelihood estimator of θ within the constraint $(3, 10)$ is plugged into $t^{(c)}$ to estimate critical values. In Table 2, we compare our DNN method against the one-sided Fligner-Killeen test (Fligner and Killeen 1976) and the one-sided Levene's test (Levene 1967) on testing variance, and the one-sided likelihood ratio test (LRT) based on the asymptotic chi-square distribution with degree of freedom of one (Lehmann and Romano 2006). Note that for the Fligner-Killeen test and Levene's test, the one-sided alternative hypothesis is transformed to if the variance of group 1 is larger than that from group 2, because the variance of t-distribution $\theta/(\theta - 2)$ is a decreasing function of θ. Some tests that are sensitive to normality are not considered

due to potential Type I error rate inflation, for example the Bartlett's test. We also evaluate another method denoted as DNN-LRT, which incorporate LRT statistic into the training data $t^{(s)}$ of TS-DNN.

Within each of the two blocks in Table 2, the first row captures the Type I error rate, while the next three rows consider the power. The third row evaluates θ_2 under H_1 from the training stage. When $\theta_1 = 4$, all five methods have accurate Type I error rate controlled at 5%, and our DNN-LRT and DNN methods are more powerful than the other three comparators. For example when $\theta_1 = 4$ and $\theta_2 = 7$, DNN has more than 1% power gain than LRT, and over 10% gain than two other tests on variance. When $\theta_1 = 7$, we observe that LRT has a conservative Type I error rate at 3.4%, and leads to power loss under alternative hypothesis. The reason is that the asymptotic distribution of LRT may not be a single Chi-square distribution when θ is close to or on the boundary of its parameter space (Chen and Liang 2010). Without further derivation of the distribution of statistics in finite-sample or even asymptotically, our automatic framework leverages DNN to learn a feasible statistic with satisfactory power and controlled Type I error rate. When comparing DNN-LRT and DNN, we observe that DNN-LRT is generally more powerful than DNN with some numerical advantages. This study demonstrates that our framework is general and has the ability to integrate other existing statistics to identify a new test with potentially higher power.

4.3. Normal Distribution with Equal Variance Assumption

We consider a problem of testing means of two groups of data from the normal distribution $\mathcal{N}(\theta, \sigma^2)$ with equal variance assumption. The Student's t-test is the known UMP unbiased level α test for testing the composite hypothesis in (1) (Lehmann and Romano 2006). We implement our proposed method in this problem and compare its performance with this theoretically optimal test.

Since the normal distribution is in a location-scale family, then we set $t^{(s)} = \left\{\hat{\theta}(x_2) - \hat{\theta}(x_1), \hat{\sigma}(x_1), \hat{\sigma}(x_2)\right\}$ for the TS-DNN and $t^{(c)} = (\sigma)$ for the CV-DNN, where $\hat{\theta}(x)$ is the sample mean and $\hat{\sigma}(x)$ is the sample standard deviation. In the training stage, we fix θ_1 at 0, and consider a neighborhood of σ at $\sum = (0.2, 2) \subseteq \mathbb{R}$. In the validation stage, $\{\hat{\sigma}(x_1) + \hat{\sigma}(x_2)\}/2$ is plugged into $t^{(c)}$ to compute critical values.

Table 3 shows the Type I error rate and power of DNN and the theoretically optimal test Student's t-test with $n = 50$ per group, varying θ_1, θ_2 and σ. In addition to an accurate Type I error rate controlled at $\alpha = 5\%$, DNN achieves a similar power as compared with the Student's t-test with a deviance not exceeding 0.1% under all scenarios evaluated. Our proposed method well approximates the existing UMP unbiased level α test in this case.

5. Adaptive Clinical Trials

5.1. The Adaptive COVID-19 Treatment Trial (ACTT)

In this section, we apply our method to the Adaptive COVID-19 Treatment Trial (ACTT) to evaluate the efficacy of remdesivir from Gilead Inc. in hospitalized adults diagnosed

with COVID-19 (National Institutes of Health 2020a). As illustrated in Section 1, adaptive clinical trials are appealing under the pandemic with limited knowledge on COVID-19, because they are capable of accommodating uncertainty during study conduction. As an alternative to existing methods to control Type I error rates (Bauer and Kohne 1994; Cui, Hung, and Wang 1999), our DNN-based method builds prespecified function to seek power enhancement in order to make the adaptive clinical trials more efficient and more ethical.

In this case study based on ACTT, we consider the sample size reassessment adaptive design for illustrative purposes, which remains the adaptive design most frequently proposed to regulatory agencies for both Food and Drug Administration (FDA) (Lin et al. 2016) and European Medicines Agency (EMA) (Elsäßer et al. 2014). For demonstration, we consider a binary endpoint of achieving hospital discharge at Day 14 (National Institutes of Health 2020c; Gilead Inc. 2020). The goal is to test H_0 versus H_1 in (1) with a one-sided Type I error rate 0.05, where θ_1 is the response rate in the placebo, and θ_2 is from the treatment. The underlying true $\theta_1 = 0.47$ and $\theta_2 = 0.59$ are assumed based on approximations using exponential distributions with median recovery time from the preliminary interim results in National Institutes of Health (2020b). We consider a two-stage adaptive design with $n^{(1)} = 120$ as the sample size per group in the first stage. A Data and Safety Monitoring Board (DSMB) evaluates unblinded interim data of those 240 subjects and makes sample size adjustments based on the following rule,

$$n^{(2)} = \begin{cases} n_{\min}^{(2)}, & \text{if } \hat{\theta}\left\{x_2^{(1)}\right\} - \hat{\theta}\left\{x_1^{(1)}\right\} > \theta_{\min} \\ n_{\max}^{(2)}, & \text{otherwise} \end{cases} \tag{11}$$

where $\hat{\theta}\left\{x_j^{(h)}\right\}$ is the sample average, $x_j^{(h)}$ is a vector of observed binary data of size $n^{(h)}$ for group j, $j = 1, 2$ at stage h, $h = 1, 2$, and $n_{\min}^{(2)}$, $n_{\min}^{(2)}$ and $\theta_{\min}$ are prespecified design features. Basically, $n^{(2)}$ in the second stage will be decreased to $n_{\min}^{(2)}$ if a promising treatment effect larger than a clinically meaningful difference $\theta_{\min}$ is observed, but increased to $n_{\max}^{(2)}$ otherwise. Other adaptive measures can also be applied, for example the conditional power (Mehta and Pocock 2011).

We consider a design with $n_{\min}^{(2)} = 30$, $n_{\max}^{(2)} = 400$, $\theta_{\min} = 0.1$, and $\Theta = (0.15, 0.8)$ to cover the underlying $\theta_1 = 0.47$ and $\theta_2 = 0.59$. Our training vector is $t^{(s)} = \left[\hat{\theta}\left\{x_1^{(1)}\right\}, \hat{\theta}\left\{x_2^{(1)}\right\}, \hat{\theta}\left\{x_1^{(2)}\right\}, \hat{\theta}\left\{x_2^{(2)}\right\}, n^{(2)}\right]$ for the TS-DNN, and $t^c = (\theta_1)$ for the CV-DNN, where $\hat{\theta}\{x\}$ is the sample mean of x. With observed data in the first stage $\tilde{x}_{12}^{(1)} = \left\{\tilde{x}_1^{(1)}, \tilde{x}_2^{(1)}\right\}$, we use $\hat{\theta}\left\{\tilde{x}_{12}^{(1)}\right\}$ to estimate the critical value. We evaluate the performance of our DNN-based method versus two existing methods: the inverse normal combination test approach (INCTA; Bauer and Kohne 1994; Cui, Hung, and Wang 1999) and the empirical test (ET; Berry et al. 2010). The INCTA combines the p-values from two stages using prespecified weights, for example equal weights, such that the nominal level can still be applied (Bretz et al. 2009). The ET approach uses the traditional proportional test on the pooled data from two stages

and chooses the critical value in the p-value scale by a grid search method to control the Type I error rate in validation (Berry et al. 2010).

In Table 4a, we first study the Type I error rates under H_0 where the common θ in two groups takes the values 0.37, 0.47, 0.57, and 0.67 around the underlying $\theta_1 = 0.47$ (National Institutes of Health 2020b). All three methods have Type I error rates controlled at 0.05, where the critical value of ET in the p-value scale is 0.032. In terms of power evaluation, we fix θ_1 in group 1 at 0.47, and consider varying θ_2 at 0.58, 0.59 and 0.6. Under the true $\theta_2 = 0.59$, DNN consistently has a higher power than two methods, with 6.6% gain as compared with the INCTA and 4.3% gain versus the ET. This superior power performance of DNN is also available under varying designs as shown in Appendix F, supplementary material.

We provide some discussion on the superior power performance of the DNN-based approach. The existing methods INCTA and ET aim at Type I error protection, and hence, their power may not be optimal. Our proposed method, on the contrary, constructs a test statistic to category whether data come from H_1 or from H_0 to enhance power in Section 3.2, and further computes its corresponding critical value to control Type I error rate in Section 3.3. This formulation also leads to a natural interpretation of our DNN statistic: a measure to optimally category whether the observed data support H_1 (the study drug has a better efficacy profile than placebo) or support H_0 (there is no treatment effect).

We further calculate the average sample size (ASN) for each method to reach approximately 90% by varying $n_{\max}^{(2)}$ in (11). DNN requires the smallest ASN at 496, while 898 for INCTA and 604 for ET, demonstrating that our proposed method essentially leads to a more efficient and ethical adaptive clinical trial to evaluate treatment options for COVID-19. To ensure study integrity, regulatory agencies usually require that hypothesis testing strategy to be prespecified before the current Phase III trial conduct (Neuenschwander et al. 2010). The two well-trained DNNs from our method can be locked in files to satisfy this requirement. As demonstrated in the R shiny app (link provided in the Section of supplementary materials), one can instantly calculate the test statistic from TS-DNN and the critical value from CV-DNN to conduct hypothesis testing based on observed data.

5.2. The Multiple Sclerosis and Extract of Cannabis (MUSEC) Trial

We apply our method to the Multiple Sclerosis and Extract of Cannabis (MUSEC) trial, which implemented a sample size adaptive design (Zajicek et al. 2012). We consider a generic adaptive design with $n^{(1)} = 85$, $n_{\min}^{(2)} = 28$, $n_{\max}^{(2)} = 340$ and $\theta_{\min} = 0.1$ in (11). The underlying response rates for achieving relief from muscle stiffness after 12 weeks are assumed as $\theta_1 = 0.27$ for the placebo and $\theta_2 = 0.4$ for the treatment based on results in Zajicek et al. (2012) with the support $\Theta = (0.05, 0.6)$.

In Table 4b, we first evaluate Type I error rates under null response rates 0.17, 0.27, 0.37 and 0.47, and then consider power under the underlying placebo rate $\theta_1 = 0.27$. ET uses a critical value of 0.034 in the p-value scale to preserve Type I error rates at $\alpha = 0.05$ within the above range of θ in validation. Our method has consistently higher power than INCTA

and ET under different values of θ_2 in addition to well-controlled Type I error rates (Table 4b).

6. Concluding Remarks

In this article, we propose a novel two-stage hypothesis testing framework to enhance power in a finite sample. As an important application in the ACTT trial on COVID-19, our method can contribute to a study with a shortened timeline, saved resources, and most importantly, fewer patients involved for ethical consideration.

Motivated by the ACTT, this article focuses on a two-group comparison with equal sample size. Our method can be readily generalized to a two-group comparison with unequal sample size, paired samples, hypothesis with contrasts, and multiple hypotheses testings. In problems where the potential UMP level α test or the UMP unbiased level α test are hard to characterize, we acknowledge that one can construct a more powerful hypothesis testing strategy by studying the parametric assumption in a given problem; but the next complication is to understand the distribution of the test statistics in a finite sample to compute its critical value with a controlled Type I error rate. Our proposed method, on the other hand, provides an automatic learning framework of identifying and characterizing such test statistics and critical values based on DNNs to enhance power. It can be a reference measure to evaluate the performance of other proposed testing strategies based on either analytic derivations or numerical approximations.

There are some potential limitations of the proposed method. First of all, our approach numerically approximates test statistics and critical values with the aim of power enhancement. The power can be slightly lower than the available theoretical most powerful test, for example in Section 4.3, or other statistics well-designed by human intelligence for a given problem. Secondly, the interpretation of test statistics constructed by TS-DNN needs further investigation. Based on our current framework, it is interpretated as a measure to maximally category whether observed data come from alternative hypothesis as compared with null hypothesis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Allaire J, and Chollet F (2020), "keras: R Interface to 'Keras'," R package version 2.2.5.0. Available at https://keras.rstudio.com.

Bach F (2017). "Breaking the Curse of Dimensionality with Convex Neural Networks," The Journal of Machine Learning Research, 18, 629–681.

Bauer P, and Kohne K (1994), "Evaluation of Experiments with Adaptive Interim Analyses," Biometrics, 50, 1029–1041. [PubMed: 7786985]

Berry SM, Carlin BP, Lee JJ, and Muller P (2010), Bayesian Adaptive Methods for Clinical Trials, Boca Raton, FL: CRC Press.

Boser BE, Guyon IM, and Vapnik VN (1992), "A Training Algorithm for Optimal Margin Classifiers," in Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152.

Bretz F, Koenig F, Brannath W, Glimm E, and Posch M (2009), "Adaptive Designs for Confirmatory Clinical Trials," Statistics in Medicine, 28, 1181–1217. [PubMed: 19206095]

Chen MH, Dey DK, Müller P, Sun D, and Ye K (2010), "Bayesian Clinical Trials," in Frontiers of Statistical Decision Making and Bayesian Analysis, eds. Chen M-H, Müller P, Sun D, Ye K, and Dey DK, pp. 257–284, New York: Springer.

Chen MH, Ibrahim JG, Zeng D, Hu K, and Jia C (2014), "Bayesian Design of Superiority Clinical Trials for Recurrent Events Data with Applications to Bleeding and Transfusion Events in Myelodyplastic Syndrome," Biometrics, 70, 1003–1013. [PubMed: 25041037]

Chen M, Jiang H, Liao W, and Zhao T (2019), "Nonparametric Regression on Low-Dimensional Manifolds using Deep ReLU Networks," arXiv preprint, arXiv:1908.01842.

Chen Y, and Liang KY (2010), "On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic with Boundary Problems," Biometrika, 97, 603–620. [PubMed: 22822249]

Cheng X, and Cloninger A (2019), "Classification Logit Two-sample Testing by Neural Networks," arXiv preprint, arXiv:1909.11298.

Chollet F, and Allaire JJ (2018), Deep Learning with R, Shelter Island, NY: Manning Publications.

Cui L, Hung HJ, and Wang SJ (1999), "Modification of Sample Size in Group Sequential Clinical Trials," Biometrics, 55, 853–857. [PubMed: 11315017]

Efron B, and Tibshirani RJ (1994), An Introduction to the Bootstrap, Boca Raton, FL: CRC Press.

Elsäßer A, Regnstrom J, Vetter T, Koenig F, Hemmings RJ, Greco M, Papaluca-Amati M, and Posch M (2014), "Adaptive Clinical Trial Designs for European Marketing Authorization: A Survey of Scientific Advice Letters from the European Medicines Agency," Trials, 15, 383. [PubMed: 25278265]

EMA (2007), Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design, London: EMEA.

FDA (2019), "Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry." Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.

Fligner MA, and Killeen TJ (1976), Distribution-Free Two-Sample Tests for Scale," Journal of the American Statistical Association, 71, 210–213.

Galili T, and Meilijson I (2016), "An Example of an Improvable Rao–Blackwell Improvement, Inefficient Maximum Likelihood Estimator, and Unbiased Generalized Bayes Estimator," The American Statistician, 70, 108–113. [PubMed: 27499547]

Gilead Inc. (2020), "Gilead Announces Results From Phase 3 Trial of Investigational Antiviral Remdesivir in Patients With Severe COVID-19." Available at https://www.gilead.com/news-and-press/press-room/press-releases/2020/4/gilead-announces-results-from-phase-3-trial-of-investigational-antiviral-remdesivir-in-patients-with-severe-covid-19.

Goodfellow I, Bengio Y, and Courville A (2016), Deep Learning, Cambridge, MA: MIT Press.

Hinton G, Srivastava N, and Swersky K (2012), "Neural Networks for Machine Learning," Coursera, video lectures, 264, 1.

Ho TK (1995), "Random Decision Forests," in IEEE Proceedings of 3rd International Conference on Document Analysis and Recognition, 1, 278–282.

Pinheiro JC, Liu C, and Wu YN (2001), "Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using the Multivariate t Distribution," Journal of Computational and Graphical Statistics, 10, 249–276.

Kirchler M, Khorasani S, Kloft M, and Lippert C (2020), "Two-Sample Testing Using Deep Learning," in International Conference on Artificial Intelligence and Statistics, pp. 1387–1398.

Kübler JM, Jitkrittum W, Schölkopf B, and Muandet K (2020), "Learning Kernel Tests Without Data Splitting," arXiv preprint, arXiv:2006.02286.

Lange KL, Little RJ, and Taylor JM (1989), "Robust Statistical Modeling Using the t Distribution," Journal of the American Statistical Association, 84, 881–896.

Lehmann EL, and Romano JP (2006), Testing Statistical Hypotheses, New York: Springer.

Levene H (1961), "Robust Tests for Equality of Variances," Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, eds. Olkin I, Ghurye SG, Hoeffding W, Madow WG, and Mann HB, pp. 279–292, Palo Alto: Stanford University Press.

Lin M, Lee S, Zhen B, Scott J, Horne A, Solomon G, and RussekCohen E (2016), "CBER's Experience with Adaptive Design Clinical Trials," Therapeutic Innovation & Regulatory Science, 50, 195–203. [PubMed: 30227002]

Liu J, and Coull B (2017), "Robust Hypothesis Test for Nonlinear Effect with Gaussian Processes," in Advances in Neural Information Processing Systems, pp. 795–803.

Liu F, Xu W, Lu J, Zhang G, Gretton A, and Sutherland DJ (2020), "Learning Deep Kernels for Non-Parametric Two-Sample Tests," arXiv preprint, arXiv:2002.09116.

Mehta CR, and Pocock SJ (2011), "Adaptive Increase in Sample Size When Interim Results are Promising: A Practical Guide with Examples," Statistics in Medicine, 30, 3267–3284. [PubMed: 22105690]

Mhaskar HN (1996), "Neural Networks for Optimal Approximation of Smooth and Analytic Functions," Neural Computation, 8, 164–177.

National Institutes of Health. (2020a), "Adaptive COVID-19 Treatment Trial (ACTT)." Available at https://clinicaltrials.gov/ct2/show/NCT04280705.

———— (2020b), "NIH Clinical Trial Shows Remdesivir Accelerates Recovery from Advanced COVID-19." Available at https://www.niaid.nih.gov/news-events/nih-clinical-trial-shows-remdesivir-accelerates-recovery-advanced-covid-19.

———— (2020c), "Study to Evaluate the Safety and Antiviral Activity of Remdesivir (GS-5734™) in Participants With Severe Coronavirus Disease (COVID-19)." Available at https://clinicaltrials.gov/ct2/show/NCT04292899.

Neuenschwander B, Capkun-Niggli G, Branson M, and Spiegelhalter DJ (2010), "Summarizing Historical Information on Controls in Clinical Trials," Clinical Trials, 7, 5–18. [PubMed: 20156954]

Neyman J, and Pearson ES (1933), "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses," Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character, 231, 289–337.

Shen Z, Yang H, and Zhang S (2019), "Nonlinear Approximation via Compositions," Neural Networks, 119, 74–84. [PubMed: 31401528]

Vogel CR (2002), Computational Methods for Inverse Problems, Philadelphia: Society for Industrial and Applied Mathematics.

Wanke PF (2008), "The Uniform Distribution as a First Practical Approach to New Product Inventory Management," International Journal of Production Economics, 114, 811–819.

Welch BL (1951), "On the Comparison of Several Mean Values: An Alternative Approach," Biometrika, 38, 330–336.

Yarotsky D (2017), "Error Bounds for Approximations with Deep ReLU Networks," Neural Networks, 94, 103–114. [PubMed: 28756334]

Zajicek JP, Hobart JC, Slade A, Barnes D, Mattison PG, and MUSEC Research Group. (2012), "Multiple Sclerosis and Extract of Cannabis: Results of the MUSEC Trial," Journal of Neurology, Neurosurgery & Psychiatryiometrics, 83, 1125–1132.
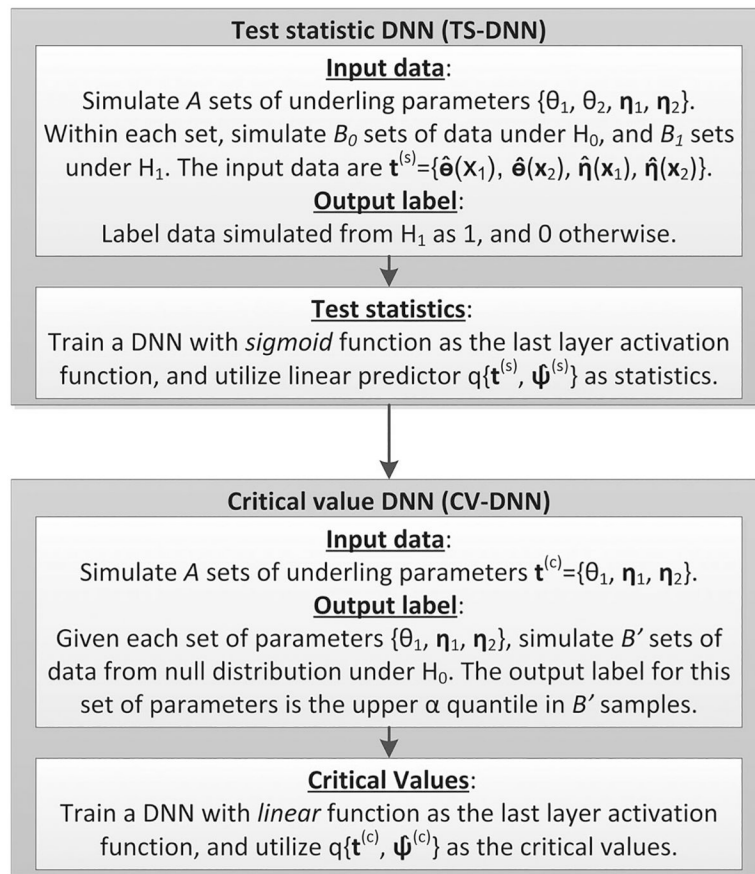
Figure 1.
Diagram of the proposed two-stage DNN-based method.

Table 1.

Type I error rate and power evaluation in the scale-uniform distribution with the method(s) with the highest power highlighted with bold font.

k	θ_1	θ_2	Type I error rate (*italicized*) / power						
			DNN-T2[1]	T_2[2]	DNN	T_1[3]	Student's t	Wilcoxon	MMD
0.2	1	1	*4.9%*	*5.0%*	*5.0%*	*5.0%*	*5.0%*	*4.8%*	*4.5%*
		1.044	79.2%	**79.4%**	76.2%	67.7%	31.0%	29.2%	11.3%
		1.055	91.2%	**91.3%**	89.9%	83.3%	41.5%	38.7%	15.0%
		1.061	94.4%	**94.5%**	93.7%	88.0%	47.1%	43.7%	17.8%
0.2	5	5	*5.0%*	*5.0%*	*4.8%*	*5.0%*	*5.0%*	*4.8%*	*5.0%*
		5.222	**82.1%**	79.4%	80.2%	67.7%	31.0%	29.2%	10.8%
		5.277	**92.5%**	91.3%	91.3%	83.3%	41.5%	38.6%	14.8%
		5.305	**95.2%**	94.5%	94.5%	88.0%	47.1%	43.7%	17.1%
0.8	1	1	*5.0%*	*5.0%*	*4.7%*	*5.0%*	*5.0%*	*4.8%*	*4.6%*
		1.178	**88.1%**	**88.1%**	87.5%	87.7%	28.4%	26.1%	12.4%
		1.222	**94.9%**	**94.9%**	94.7%	94.7%	37.1%	33.5%	17.5%
		1.244	**96.7%**	**96.7%**	96.6%	96.5%	41.6%	37.3%	20.9%
0.8	5	5	*5.0%*	*5.0%*	*4.9%*	*5.0%*	*5.0%*	*4.8%*	*4.8%*
		5.888	**88.4%**	88.1%	88.0%	87.7%	28.3%	26.0%	12.6%
		6.109	**95.3%**	94.9%	95.0%	94.7%	37.0%	33.5%	17.5%
		6.220	**96.9%**	96.7%	96.8%	96.5%	41.6%	37.3%	20.1%

[1]DNN-T2 uses T_2 in (10) as another input data of $t^{(s)}$ when training TS-DNN.

[2]T_2 has a functional form in (10) with critical value given k determined by simulations.

[3]$T_1 = \max(x_2)/\max(x_1)$ is a likelihood ratio test (LRT) statistic with critical value given k determined by simulations.

Table 2.

DNN-LRT and DNN achieve a higher power than the other three comparators with controlled Type I error rates in testing degrees of freedom in t-distribution under varying scenarios.

θ_1	θ_2	Type I error rate (*italicized*) / power				
		DNN-LRT[1]	DNN	LRT[2]	Fligner-Killeen	Levene
4	4	*5.1%*	*5.1%*	*5.0%*	*5.0%*	*5.0%*
	5	**17.1**%	16.9%	16.3%	10.8%	13.0%
	6	**31.5**%	31.0%	30.0%	16.8%	21.8%
	7	**44.6**%	43.9%	42.6%	22.2%	29.9%
7	7	*4.8%*	*4.8%*	*3.4%*	*4.9%*	*4.9%*
	8	**7.8**%	7.7%	5.2%	6.7%	7.1%
	9	**11.1**%	11.0%	7.0%	8.5%	9.3%
	10	**14.5**%	14.3%	8.7%	10.1%	11.3%

[1] DNN-LRT utilizes LRT statistic as another input data of $t^{(s)}$ when training TS-DNN.

[2] The one-sided likelihood ratio test (LRT) is based on asymptotic Chi-square distribution with one degree of freedom (Lehmann and Romano 2006).

Table 3.

DNN reaches the upper power limit from the Student's *t*-test as the UMP unbiased level α test.

θ_1	θ_2	σ	Type I error rate (*italicized*) / power	
			DNN	Student's *t*
−0.5	−0.5	1	*5.0%*	*5.0%*
	−0.1		63.0%	63.0%
	0		79.5%	79.5%
	0.1		90.6%	90.6%
−0.5	−0.5	1.5	*5.0%*	*5.0%*
	0.1		62.9%	63.0%
	0.25		79.4%	79.5%
	0.4		90.6%	90.7%
0	0	1	*5.0%*	*5.0%*
	0.4		63.0%	63.1%
	0.5		79.4%	79.5%
	0.6		90.6%	90.6%
0	0	1.5	*5.0%*	*5.0%*
	0.6		62.8%	62.9%
	0.75		79.5%	79.6%
	0.9		90.6%	90.6%

Table 4.

DNN consistently achieves a higher power than INCTA and ET in two adaptive designs with sample size reassessment: the ACTT on COVID-19 and the MUSEC on multiple sclerosis.

		(a) The ACTT					(b) The MUSEC		
θ_1	θ_2	Type I error rate			θ_1	θ_2	Type I error rate		
		DNN	INCTA[1]	ET[2]			DNN	INCTA[1]	ET[2]
0.37	0.37	4.9%	5.0%	4.8%	0.17	0.17	4.9%	5.0%	4.5%
0.47	0.47	4.9%	5.1%	4.8%	0.27	0.27	5.1%	5.0%	4.8%
0.57	0.57	5.0%	5.1%	4.9%	0.37	0.37	4.9%	5.0%	4.8%
0.67	0.67	5.0%	5.0%	4.7%	0.47	0.47	4.9%	5.0%	5.2%
θ_1	θ_2	Power			θ_1	θ_2	Power		
		DNN	INCTA[1]	ET[2]			DNN	INCTA[1]	ET[2]
0.47	0.58	**89.3%**	84.1%	86.2%	0.27	0.39	**87.4%**	82.9%	83.6%
0.47	0.59	**93.9%**	87.3%	89.6%	0.27	0.40	**91.3%**	85.9%	86.4%
0.47	0.60	**96.5%**	89.7%	92.0%	0.27	0.41	**93.8%**	88.2%	88.5%

[1] The inverse normal combination test approach (Bauer and Kohne 1994; Cui, Hung, and Wang 1999).

[2] The empirical test (Berry et al. 2010).