Submitted to the Annals of Statistics

ANALYSIS OF "LEARN-AS-YOU-GO" (LAGO) STUDIES*

By Daniel Nevo[†,¶], Judith J. Lok[‡,¶] and Donna Spiegelman[§,¶]

[†] *Tel Aviv University,* [‡] *Boston University,* [§] *Yale School of Public Health and* [¶] *Harvard T. H. Chan School of Public Health*

In Learn-As-you-GO (LAGO) adaptive studies, the intervention is a complex multi-component package, and is adapted in stages during the study based on past outcome data. This design formalizes standard practice in public health intervention studies. An effective intervention package is sought, while minimizing intervention package cost. In LAGO study data, the interventions in later stages depend upon the outcomes in the previous stages, violating standard statistical theory. We develop an estimator for the intervention effects, and prove consistency and asymptotic normality using a novel coupling argument, ensuring the validity of the test for the hypothesis of no overall intervention effect. We develop a confidence set for the optimal intervention package and confidence bands for the success probabilities under alternative package compositions. We illustrate our methods in the BetterBirth Study, which aimed to improve maternal and neonatal outcomes among 157,689 births in Uttar Pradesh, India through a multi-component intervention package.

1. Introduction. Adaptive designs have been developed and have been available for use in clinical trials for decades. The U.S. Food and Drug Administration defines an adaptive design as "...a clinical study design that allows for prospectively planned modifications based on accumulating study data without undermining the study's integrity and validity" (FDA, 2016).

The existing literature on adaptive designs has thus far considered several types of prospectively planned design modifications, including blinded sample size reassessment, group sequential testing, interim analysis for benefit or futility, successive re-randomization, changing subgroup proportions or eligibility criteria of the trial (Rosenblum and van der Laan, 2011) and dropping treatment arms. Prominent among the techniques developed to

*This work was supported by the Director's Office and the National Institute of Environmental Health Sciences, National Institutes of Health (DP1ES025459) and by the National Institute of Allergy and Infectious Diseases, National Institutes of Health (R01AI112339). The BetterBirth Study was supported with funding from the Bill & Melinda Gates Foundation. The article contents are the sole responsibility of the authors and may not necessarily represent the official views of the Bill & Melinda Gates Foundation.

Keywords and phrases: Adaptive Designs, Dependent Sample, Coupling, Public Health

preserve the validity of statistical inference when design adaption has occurred is the conditional error function (Proschan and Hunsberger, 1995; Müller and Schäfer, 2001, 2004), and combination functions have been used to aggregate p-values from multiple stages (Bauer and Kohne, 1994; Brannath, Posch and Bauer, 2002). See Kairalla et al. (2012); Bauer et al. (2016) for recent comprehensive reviews of adaptive designs in clinical trials. In addition to valid testing, methods have been developed for estimation in an adaptive group sequential design (e.g. Gao, Liu and Mehta, 2013).

The present work is motivated by large-scale public health intervention studies of complex multi-component intervention packages. In the newly proposed "Learn-As-you-GO" (LAGO) design, the intervention, which can e.g. be a treatment, a device, a new way to organize care, or, more likely, a combination thereof, is composed of several components. While subject matter experts have some knowledge with regard to the preferred intervention package, in LAGO, optimal development of the intervention package is an inherent part of the study goals. A LAGO study is conducted in stages. After each stage, the data collected so far are analyzed, the intervention package is reassessed, and a revised intervention package is rolled out in the next stage. Unlike previous adaptive designs, in the LAGO design, the composition of the intervention package in later stages depends on the outcomes from previous stages. The lack of suitable framework, estimation and associated theory motivating the research in this paper, with focus on new estimators and asymptotic theory utilizing a novel coupling argument.

Response-adaptive designs (Rosenberger, Flournoy and Durham, 1997; Hu and Rosenberger, 2003) focus on binary or discrete treatments and, according to accumulated data, change treatment allocation probabilities, not (as in LAGO) treatment options. Thus, response-adaptive designs do not concern a multivariate intervention package, the composition of which changes with trial stage in LAGO studies.

The Sequential Multiple Assignment Randomized Trial (SMART) design (Murphy, 2005; Murphy et al., 2007) randomizes study participants at more than one time point to pre-specified randomization options with probabilities that depend on participant's past characteristics and outcomes. The aim of a SMART trial is to estimate the optimal sequence of treatments for each patient given the patient's covariate and response histories up to the present. It is a *non-adaptive* design method which optimizes a personalized and dynamic intervention, in part by restricting randomization options at each step. In contrast, LAGO identifies a complex *static*, possibly 'cluster-personalized', intervention package where, unlike in SMART, the options are unknown at the start of the trial and are estimated anew as a result of

trial data up to the current stage. In addition, LAGO studies will add new centers, with new participants, entering at each stage, while in SMART the same individuals are repeatedly re-randomized.

The multiphase optimization strategy (MOST, Collins, Murphy and Strecher, 2007; Collins, Nahum-Shani and Almirall, 2014) consists of three phases: preparation, optimization and evaluation. The optimal intervention package is developed during the optimization phase, followed by its formal statistical evaluation in a randomized controlled trial. The aim of MOST is similar to LAGO: to develop an optimal intervention package and estimate its impact. However, in MOST, the outcomes of the past are used at most in one stage, to determine the optimal package in the optimization phase. The resulting package is then independently studied through a controlled trial in the evaluation phase, using no prior data.

At face value, phase I dose-finding studies have perhaps the greatest similarity to the LAGO design paradigm. In dose-finding studies, the goal is to find the maximum tolerated dose, that is, the highest dose of a drug such that adverse effects of the drug are below a pre-determined threshold. Dose values are assigned to patients in a sequential manner, and in each step a decision is made to stop and declare that the maximum tolerated dose has been found, or to continue, and if so, with which dose. The more traditionally used methods include the "3 + 3" and "accelerated titration" designs (Simon et al., 1997; Wong, Capasso and Eckhardt, 2016). Another popular method is the continual reassessment method (O'Quigley, Pepe and Fisher, 1990; O'Quigley and Shen, 1996), which assigns each patient the current estimated maximum tolerated dose. Methods were also developed for the optimal dose of two drugs simultaneously (Thall et al., 2003; Wang and Ivanova, 2005). Rosenberger and Haines (2002) provide a review of the continual reassessment method and additional statistical methods for dose finding studies. Dose-finding studies are generally too small for the application of asymptotic statistical methods, and typically Bayesian approaches have been used. In contrast, in public health intervention studies, the magnitude of the per-stage sample size is typically much larger than the sample size in dose-finding studies, while the maximum number of stages will be limited. Additionally, unlike dose-finding studies, where methods are considered for a single or at most dual treatments, the complex public health interventions motivating the development of the LAGO design feature multiple components, some of which are continuous, while others are binary.

An ad hoc example of a precursor to a formal LAGO study is the "BetterBirth Study" (Hirschhorn et al., 2015; Semrau et al., 2017) of Ariadne Labs, a joint center of the Brigham and Women's Hospital and the Harvard

T.H. Chan School of Public Health, led by Atul Gawande (Gawande, 2014). The BetterBirth Study assessed the use of the World Health Organization's (WHO) Safe ChildBirth checklist, a 31-item checklist of best labor and delivery practices believed to be feasible in resource-limited settings, to reduce maternal and neonatal mortality. The intervention was adapted and tested in a three phase process in Uttar Pradesh, India, where neonatal mortality is 32 per 1000 live births and maternal mortality is 258 per 100,000 births (Semrau et al., 2017). During the first two phases, the intervention was adapted, and a final version was tested in a cluster randomized trial, that included 157,689 mothers and newborns.

The first goal of a LAGO study is to identify the optimal intervention package such that the cost of the intervention is minimized and the probability of a desired binary outcome is above a given threshold. For example, in the BetterBirth Study, the outcome could be the use of the WHO Safe ChildBirth checklist, with the aim being, for example, that the checklist is used during at least 90% of the births. In the illustrative example included in this paper, we investigate a process outcome, oxytocin administration after delivery, with the aim being that 85% of mothers will receive oxytocin after delivery. Oxytocin is recommended by the WHO, as a proven intervention for preventing postpartum hemorrhage. We determine whether the use of a multiple component intervention package that includes on-site coaching visits and an intervention launch of a particular duration, increases the administration of oxytocin, compared to standard of care.

The second goal of a LAGO study is to assess the overall impact of the intervention strategy, as well as that of its individual components. We present methodology to achieve both goals.

In a LAGO study, the data are not an independent sample. Beginning with the second stage, the recommended intervention package is itself a random variable that depends on previous outcomes. In the final analysis, a LAGO study uses the data from all stages. When considering the asymptotic behavior of the estimators, we assume that the sample size in each stage increases at a similar rate. In addition, we assume that the intervention in each stage converges in probability to a constant as the number of observations in the previous stages goes to infinity. This would happen, for example and under the usual regularity conditions, if the intervention in each stage is based on a maximum likelihood estimator obtained from the data collected in previous stages.

LAGO studies can be further characterized by a key design feature which determines the strength of the causal inferences that can be made. In an *uncontrolled* LAGO study, there are neither baseline data available to permit a

quasi-experimented before-after comparison nor randomized or nonrandomized planned variation in the implementation of the intervention package. Thus, unplanned variation, which is widespread in large-scale public health interventions, serves as the basis for estimating causal contrasts. Under unplanned variation, causal inference methods will be needed to adjust for possible confounding bias (Hernan and Robins, 2019; Spiegelman and Zhou, 2018). In a *controlled* LAGO study, baseline outcome data are collected before the intervention is implemented, or in additional centers in which no intervention was implemented. These additional centers may be randomized or not, to be included in the study as controls. When baseline data serves as the control, the quasi-experimental before-after design provides the data for causal contrasts. The before-after design relies on the untestable assumption that there are no time trends in the data, so changes in mean outcomes can be solely attributed to intervention effects (Cox, 1958). If, instead or in addition to baseline data, there are concurrent control centers, stronger causal inference is permitted by design, with the strongest design in this context being being a randomized controlled LAGO trial.

We propose estimators for a LAGO study allowing for several stages, multiple centers or sites, multiple component complex interventions, and center-specific baseline covariates that affect the outcome rate, or random center-specific deviations from the recommended intervention, or both. We show that even in this setup, the optimal intervention can be learned from the combined data from all stages. Even when the optimal intervention in the last stage does not achieve the pre-specified study goal, the optimal intervention is estimated. We prove consistency and asymptotic normality of the new estimators utilizing a novel coupling argument. We further establish the validity of tests for an overall intervention effect. In addition, we develop a confidence set for the optimal intervention package and confidence bands for the target outcome probability under various observed or hypothesized intervention packages.

The rest of the paper is as follows. In Section 2, we describe the LAGO design and our key assumptions (Section 2.1), propose a relevant estimator and study its asymptotic properties (Section 2.2), which we then use for construction of hypothesis tests (Section 2.3) and confidence intervals (Section 2.4). In Section 3, we report the results of a simulation study and in Section 4 we present an illustrative analysis of the BetterBirth Study. In Section 5 we discuss our results and future research. Proofs of our two main theorems are given in the appendix. Additional proofs and simulation study results are given in the supplementary materials.

2. LAGO design - theoretical development.

2.1. *Description of the learn-as-you-go design.* The methods we develop in this paper cover an arbitrary number of stages, K. At each stage k, a version of the intervention package is implemented in each of J_k centers. Let n_{jk} denote the sample size (e.g. the number of births) in the j-th center at stage k. We assume that each center is included in one stage only. In a randomized controlled trial, centers may be randomized to either intervention or control. Alternatively, data might be collected pre and post the implementation of the intervention package and then a center contributes data to both the intervention and the control.

Asymptotic theory is developed for the setting where the number of patients per center goes to infinity at the same rate in all stages, leading to reliable approximations when the number patients in each center is relatively large. Let $n_k = \sum_{j=1}^{J_k} n_{jk}$ be the number of participants in stage k and $n = \sum_{k=1}^{K} n_k$ be the total number of participants. Our asymptotic inference assumes that the ratio between the number of patients in each center and the total sample size n converges to a constant, and we write $\alpha_{jk} = \lim_{n\to\infty} n_{jk}/n$; then, $\sum_{k=1}^{K}\sum_{j=1}^{J_k} \alpha_{jk} = 1$. Define also $\bar{n}_k = (n_1, ..., n_k)$. For ease of presentation, we first develop methodology for a LAGO study consisting of $K = 2$ stages. Section 3 of the supplementary materials covers studies with $K > 2$.

The multivariate intervention package consists of p components. Let $\mathcal{X}$ be the support of the intervention, that is, all possible intervention values. For example, if all p intervention components are continuous and each is constrained to be within a given interval $[\mathcal{L}_r, \mathcal{U}_r], r = 1, ..., p$, then $\mathcal{X} = [\mathcal{L}_1, \mathcal{U}_1] \times [\mathcal{L}_2, \mathcal{U}_2] \times \cdots \times [\mathcal{L}_p, \mathcal{U}_p]$. Throughout this paper, as would ordinarily be the case in practice, we assume that $\mathcal{X}$ is bounded.

For stage 1, an initial $\boldsymbol{x}^{(1)}$ (or $\boldsymbol{x}_j^{(1)}$ for each center j) is chosen by the investigators, based on their best judgment. We distinguish between the *recommended* intervention and the *actual* intervention. In large scale public health settings, the actual intervention, denoted by $\boldsymbol{A}_j$, may differ from the recommended intervention, due to local constraints or preferences. We denote $\boldsymbol{z}_j$ for center-specific characteristics reflecting baseline heterogeneity between centers with respect to the outcome of interest and we consider the $\boldsymbol{z}_j$ fixed, i.e., they are not part of the intervention package. For each center, $\boldsymbol{z}_j$ could be, for example, the district of the health center or its monthly birth volume.

We assume that the probability of success for a single unit i (e.g., participant or birth) in a center j with characteristics $\boldsymbol{z}_j$ under intervention

$\boldsymbol{A} = \boldsymbol{a}_j$, $p_{\boldsymbol{a}_j}(\boldsymbol{\beta}; \boldsymbol{z}) = pr(Y_{ij} = 1 \mid \boldsymbol{A}_j = \boldsymbol{a}_j, \boldsymbol{X}_j = \boldsymbol{x}_j, \boldsymbol{z}_j; \boldsymbol{\beta})$, does not depend on the recommended intervention $\boldsymbol{x}_j$, except through the actual intervention $\boldsymbol{a}_j$, and follows a logistic regression model

$$\text{logit}\, p_{\boldsymbol{a}_j}(\boldsymbol{\beta}; \boldsymbol{z}_j) = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{a}_j + \boldsymbol{\beta}_2^T \boldsymbol{z}_j, \tag{2.1}$$

where $\boldsymbol{\beta}^T = (\beta_0, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$ is a vector of unknown parameters, such that $\boldsymbol{\beta}_1$ describes the effects of the p intervention package components. For centers in the control arm or for pre-intervention data, if available, $\boldsymbol{a} = \boldsymbol{x} = \boldsymbol{0}$. We assume that in each stage, conditionally on all $\boldsymbol{a}_j$ and $\boldsymbol{z}_j$, outcomes are independent within and between centers. Learning the intervention, however, causes dependence between stages, which we consider below.

A main goal of the LAGO design is to identify the optimal intervention package. Let $\tilde{p}$ be a pre-specified outcome probability goal and $C(\boldsymbol{x})$ be a known cost function. For example, in the BetterBirth Study, one may want to find the minimal number of on-site coaching visits to ensure that oxytocin is administrated to the mother right after delivery in at least 85% of births ($\tilde{p} = 0.85$). If $\boldsymbol{\beta}$ were known, an optimal intervention for a center with covariates $\boldsymbol{z}_j$ could be the solution to the center-specific optimization problem

$$\min_{\boldsymbol{x}_j} C(\boldsymbol{x}_j) \quad \text{subject to} \quad p_{\boldsymbol{x}_j}(\boldsymbol{\beta}; \boldsymbol{z}_j) \geq \tilde{p} \quad \& \quad \boldsymbol{x}_j \in \mathcal{X}. \tag{2.2}$$

Computational issues regarding solving (2.2) will be discussed in Section 2.5. We assume that for the true parameter values, there is a unique solution to (2.2). For example, if the intervention has two components with unit costs c_1 and c_2 and a linear cost function, we assume that $\beta_{11}/c_1 \neq \beta_{12}/c_2$. Other optimization criteria can be considered. For example, the optimal intervention could require that the intervention results in an outcome probability $\tilde{p}$ when calculating a weighed average over a group of centers $\{j = 1, ..., J\}$, with sample sizes n_j. That is,

$$\min_{\boldsymbol{x}_1, ..., \boldsymbol{x}_J} \sum_{j=1}^{J} C(\boldsymbol{x}_j) \quad \text{subject to} \quad \frac{1}{N} \sum_{j=1}^{J} n_j p_{\boldsymbol{x}_j}(\boldsymbol{\beta}; \boldsymbol{z}_j) \geq \tilde{p} \quad \& \quad \boldsymbol{x}_j \in \mathcal{X} \quad \forall j$$

where $N = \sum_{j=1}^{J} n_j$. In this paper we focus on (2.2).

We continue our description of the data and model. Let $\bar{\boldsymbol{z}}^{(k)} = (\boldsymbol{z}_1^{(k)}, ..., \boldsymbol{z}_{J_k}^{(k)})$ be the observed center characteristics in each of the J_k stage k centers. We start with stage 1. Let $\boldsymbol{x}_j^{(1)}$ be the recommended (multivariate) intervention package for center j in stage 1, which in the absence of $\boldsymbol{z}$, may be the same

for all centers. We assume that the stage 1 recommended interventions $\boldsymbol{x}_j^{(1)}$, $j = 1, .., J_1$, are determined before the trial starts. The actual intervention in center j of stage 1 is, however, $\boldsymbol{a}_j^{(1)} = h_j^{(1)}(\boldsymbol{x}_j^{(1)})$, where $h_j^{(1)}$ is a deterministic center-specific continuous function from $\mathcal{X}$ to $\mathcal{X}$ that determines how center j implements the actual intervention based on the recommendation $\boldsymbol{x}_j^{(1)}$. We do not require that the $h_j^{(1)}$ are known, but only that the $\boldsymbol{a}_j^{(1)}$ are observed. Let $Y_{ij}^{(1)}$ be the binary outcome of interest for patient i in center j of stage 1, each following model (2.1), and let the outcome vector in center j of stage 1 be $\boldsymbol{Y}_j^{(1)} = (Y_{1j}^{(1)}, ..., Y_{n_{j1}j}^{(1)})$. Let $\bar{\boldsymbol{a}}^{(1)} = (\boldsymbol{a}_1^{(1)}, ..., \boldsymbol{a}_{J_1}^{(1)})$ and $\bar{\boldsymbol{Y}}^{(1)} = (\boldsymbol{Y}_1^{(1)}, ..., \boldsymbol{Y}_{J_1}^{(1)})$ be the stage 1 actual interventions and outcomes, respectively.

Following the stage 1 data collection, a stage 1 analysis is conducted to determine the recommended interventions for the new centers in stage 2, denoted by $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}, j = 1, ..., J_2$. If there are control centers, their recommended intervention and their actual intervention are zero. The value $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$ is chosen through a function, g, that takes as input the stage 1 data, the goal of the intervention, and the center-specific covariates and returns a recommended intervention, which is usually the estimated optimal intervention $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)} = g(\bar{\boldsymbol{a}}^{(1)}, \bar{\boldsymbol{Y}}^{(1)}, \bar{\boldsymbol{z}}^{(1)}, \boldsymbol{z}_j^{(2)})$. Then, $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$ can be obtained by solving the optimization problem given in (2.2) for each center, with $\boldsymbol{\beta}$ replaced by an estimator $\hat{\boldsymbol{\beta}}^{(1)}$ based on the stage 1 data alone. The superscript, n_1, in $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$ reminds us that $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$ is a random variable that is a function of the data from the n_1 participants in stage 1.

The actual intervention implemented in center j of stage 2 is $\boldsymbol{A}_j^{(2,n_1)} = h_j^{(2)}(\hat{\boldsymbol{x}}_j^{opt,(2,n_1)})$, where $h_j^{(2)}$ are the analogues of $h_j^{(1)}$, but now for the stage 2 centers. Let $\bar{\hat{\boldsymbol{x}}}^{opt,(2,n_1)} = (\hat{\boldsymbol{x}}_1^{opt,(2,n_1)}, ..., \hat{\boldsymbol{x}}_{J_2}^{opt,(2,n_1)})$ be the recommended interventions at the J_2 stage 2 centers. Once $\bar{\hat{\boldsymbol{x}}}^{opt,(2,n_1)}$ are determined, stage 2 outcomes are collected under the actual interventions $\bar{\boldsymbol{A}}^{(2,n_1)} = (\boldsymbol{A}_1^{(2,n_1)}, ..., \boldsymbol{A}_{J_2}^{(2,n_1)})$, which may be the same as $\bar{\hat{\boldsymbol{x}}}^{opt,(2,n_1)}$. Let $\boldsymbol{Y}_j^{(2,n_1)} = (Y_{1j}^{(2,n_1)}, ..., Y_{n_{j2}j}^{(2,n_1)})$ be the stage 2 outcomes in center j, each following model (2.1), and $\bar{\boldsymbol{Y}}^{(2,n_1)} = (\boldsymbol{Y}_1^{(2,n_1)}, ..., \boldsymbol{Y}_{J_2}^{(2,n_1)})$ be all the stage 2 outcomes. Our two main assumptions are

ASSUMPTION 2.1. *Conditionally on* $\bar{\hat{\boldsymbol{x}}}^{opt,(2,n_1)}$, $(\bar{\boldsymbol{A}}^{(2,n_1)}, \bar{\boldsymbol{Y}}^{(2,n_1)})$ *are independent of the stage 1 data* $(\bar{\boldsymbol{a}}^{(1)}, \bar{\boldsymbol{Y}}^{(1)})$.

ASSUMPTION 2.2. *For each* $j = 1, ..., J_2$, *the stage 2 recommended in-*

tervention $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$ *converges in probability to a center-specific limit* $\boldsymbol{x}_j^{(2)}$.

Assumption 2.1 assumes that learning takes place only through the determination of the recommended intervention. It ensures that the dependence between the stage 1 data and stage 2 outcomes is solely due to the dependence of the $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$ on the stage 1 data. It specifically means that, given $\bar{\hat{\boldsymbol{x}}}^{opt,(2,n_1)}$, the actual intervention in a stage 2 center is conditionally independent of $\bar{\boldsymbol{Y}}^{(1)}$. Under Assumption 2.1, and the aforementioned assumption that conditionally on the actual interventions, the outcomes do not depend on the recommended interventions, we can conclude that in stage 2, $pr(\bar{\boldsymbol{Y}}^{(2,n_1)} \mid \bar{\boldsymbol{A}}^{(2,n_1)}, \bar{\hat{\boldsymbol{x}}}^{opt,(2,n_1)}, \bar{\boldsymbol{z}}^{(2)}, \bar{\boldsymbol{Y}}^{(1)}) = pr(\bar{\boldsymbol{Y}}^{(2,n_1)} \mid \bar{\boldsymbol{A}}^{(2,n_1)}, \bar{\boldsymbol{z}}^{(2)})$, so the logistic regression model (2.1) holds for the stage 2 data. Assumption 2.2 implies that in the presence of more and more stage 1 data under $\boldsymbol{a}_j^{(1)}, j = 1, ..., J_1$, each of the estimated optimal intervention packages $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}, j = 1, ...J_2$, converges in probability to a fixed value $\boldsymbol{x}_j^{(2)}$. For example, Assumption 2.2 will hold if $\bar{\hat{\boldsymbol{x}}}^{opt,(2,n_1)}$ are continuous functions of the stage 1 maximum likelihood estimator, $\hat{\boldsymbol{\beta}}_1$, as is the case if $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$ solves (2.2) and $\beta_{11}/c_1 \neq \beta_{12}/c_2$. Under Assumption 2.2 and continuity of the h_j's, the Continuous Mapping Theorem implies that $\boldsymbol{A}_j^{(2,n_1)} = h_j^{(2)}(\hat{\boldsymbol{x}}_j^{opt,(2,n_1)})$ converges in probability to $\boldsymbol{a}_j^{(2)} = h_j^{(2)}(\boldsymbol{x}_j^{(2)})$. We additionally assume that there is no separation or quasi-separation of the data. This assumption ensures that the estimator is unique and alleviates identifiability concerns (Albert and Anderson, 1984; Wedderburn, 1976).

In fact, the results we prove in this paper regarding the estimators obtained at the end of the study hold not only for $g(\bar{\boldsymbol{a}}^{(1)}, \bar{\boldsymbol{Y}}^{(1)}, \bar{\boldsymbol{z}}^{(1)}, \boldsymbol{z}_j^{(2)}) = \hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$, but under any choice of function g for the recommended intervention, as long as Assumption 2.2 holds.

2.2. $\hat{\boldsymbol{\beta}}$ *and its asymptotic properties.* We estimate $\boldsymbol{\beta}$ after the K stages are concluded. As in previous sections, for ease of development, we consider here $K = 2$. Section 3 of the supplementary materials covers the case of $K > 2$.

We propose to estimate $\boldsymbol{\beta}$ by solving the estimating equations

$$
\begin{aligned}
0 = \boldsymbol{U}(\boldsymbol{\beta}) = \frac{1}{n} \Bigg\{ & \sum_{j=1}^{J_1} \sum_{i=1}^{n_{j1}} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \left(Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(1)}) \right) \\
& + \sum_{j=1}^{J_2} \sum_{i=1}^{n_{j2}} \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \left(Y_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)}) \right) \Bigg\}.
\end{aligned} \tag{2.3}
$$

In Section 2 of the supplementary materials, we show that the estimator $\hat{\boldsymbol{\beta}}$ that solves (2.3) is also a maximum partial likelihood estimator, although that is not needed for the proofs below. The estimating equations (2.3) also arise if the interventions A were determined a priori, so $\hat{\boldsymbol{\beta}}$ can be estimated using standard software.

Asymptotic theory for $\hat{\boldsymbol{\beta}}$ is complicated, however, by the fact that $\bar{\boldsymbol{Y}}^{(1)}$ and $(\bar{\boldsymbol{A}}^{(2,n_1)}, \bar{\boldsymbol{Y}}^{(2,n_1)})$ are not independent. Thus, the score function $U(\boldsymbol{\beta})$ is not a sum of independent random variables.

Let $\mathcal{B}$ be the parameter space for $\boldsymbol{\beta}$. A conditional expectations argument (Equation (A.6) in the appendix) shows that the score function has mean zero when evaluated at the true value, denoted by $\boldsymbol{\beta}^\star$. Furthermore, we show in the appendix (Equation (A.7)) that the two terms in (2.3), although dependent, are uncorrelated. These two properties are useful for proving that $\hat{\boldsymbol{\beta}}$ is consistent:

THEOREM 2.1. *(Consistency). Assume $\mathcal{B}$ is compact. Under Assumptions 2.1 and 2.2, $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^\star$.*

The proof is given in Section A.1 of the appendix.

Asymptotic normality also poses a challenge due to the dependence between the two summands in $\boldsymbol{U}(\boldsymbol{\beta})$. It can be shown that $\partial \boldsymbol{U}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ converges in probability to $-I(\boldsymbol{\beta})$, for all $\boldsymbol{\beta} \in \mathcal{B}$, with $I(\boldsymbol{\beta})$ given in Equation (A.13) of the appendix. The following theorem establishes asymptotic normality of $\hat{\boldsymbol{\beta}}$:

THEOREM 2.2. *(Asymptotic normality). Under Assumptions 2.1 and 2.2,*

$$
n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{\mathcal{D}} N\left(0, I^{-1}(\boldsymbol{\beta}^\star)\right). \tag{2.4}
$$

The full proof of Theorem 2 is given in Section A.2 of the appendix. Here we outline the main parts of the proof, which rests upon a novel coupling

argument. First, by the mean value theorem and further arguments, it can be shown that the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)$ is the same as the asymptotic distribution of

$$[I(\boldsymbol{\beta}^\star)]^{-1} n^{-1/2} \left[\sum_{j=1}^{J_1} \sum_{i=1}^{n_{j1}} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \left(Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) \right) \right. \tag{2.5}$$
$$\left. + \sum_{j=1}^{J_2} \sum_{i=1}^{n_{j2}} \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \left(Y_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right) \right].$$

We next show that the asymptotic distribution of the part of (2.5) that does not involve $I(\boldsymbol{\beta}^\star)$ is multivariate normal. The following coupling argument deals with the fact that the two summands in (2.5) are not independent. For each $j = 1, ..., J_2$, let $Y_{ij}^{(2)}$, $i = 1, ..., n_{j2}$, be independent Bernoulli random variables, independent of all stage 1 data, with success probability $p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})$, where, as defined before, $\boldsymbol{a}_j^{(2)} = h_j^{(2)}(\boldsymbol{x}_j^{(2)})$. We construct variables $\tilde{Y}_{ij}^{(2,n_1)}$ which, given the stage 1 data and the $\boldsymbol{A}_j^{(2,n_1)}$, have the same distribution as the original $Y_{ij}^{(2,n_1)}$, but coupled (see e.g. Lindvall (2002)) with the $Y_{ij}^{(2)}$ in the following way. Let W_{ij} be independent uniform $(0, 1)$ random variables, independent of all other variables introduced so far. For the case $p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) > p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})$, let

$$\tilde{Y}_{ij}^{(2,n_1)} = \begin{cases} 0 & \text{if } Y_{ij}^{(2)} = 0 \\ 0 & \text{if } Y_{ij}^{(2)} = 1 \text{ and } W_{ij} < \dfrac{p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})}{p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})} \\ 1 & \text{if } Y_{ij}^{(2)} = 1 \text{ and } W_{ij} \geq \dfrac{p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})}{p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})}. \end{cases} \tag{2.6}$$

A similar expression is given in equation (A.14) in the appendix for the case $p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \leq p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})$. The key property of the coupling argument is that given $\boldsymbol{A}_j^{(2,n_1)}$ and the stage 1 data, the distribution of the coupled $\tilde{Y}_{ij}^{(2,n_1)}$ is identical to the distribution of the original $Y_{ij}^{(2,n_1)}$. Therefore, when we replace $Y_{ij}^{(2,n_1)}$ with $\tilde{Y}_{ij}^{(2,n_1)}$ in (2.5), the distribution of (2.5) is unaffected. The coupled outcomes are used in Section A.2 to show that the part of (2.5) that does not involve $I(\boldsymbol{\beta}^\star)$ has the same asymptotic

distribution as

(2.7)
$$\frac{1}{\sqrt{n}}\left\{\sum_{j=1}^{J_1}\sum_{i=1}^{n_{j1}}\begin{pmatrix}1\\ \boldsymbol{a}_j^{(1)}\\ \boldsymbol{z}_j^{(1)}\end{pmatrix}\left(Y_{ij}^{(1)}-p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)+\sum_{j=1}^{J_2}\sum_{i=1}^{n_{j2}}\begin{pmatrix}1\\ \boldsymbol{a}_j^{(2)}\\ \boldsymbol{z}_j^{(2)}\end{pmatrix}\left(Y_{ij}^{(2)}-p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)\right\}.$$

The outcomes $\bar{\boldsymbol{Y}}^{(1)}$ and $\bar{\boldsymbol{Y}}^{(2)}=(\bar{\boldsymbol{Y}}_1^{(2)},...,\bar{\boldsymbol{Y}}_{J_2}^{(2)})$ are independent, because the $\boldsymbol{Y}_{ij}^{(2)}$ are the outcomes under the constant intervention $\boldsymbol{a}_j^{(2)}$. Therefore, by standard logistic regression theory, the expression in (2.7) converges in distribution to a normal random variable with mean zero and variance $I(\beta^\star)$. Combining the asymptotic normality of (2.7) with (2.5) implies that Theorem 2 holds.

The asymptotic variance can be consistently estimated from the data by replacing $\boldsymbol{a}_j^{(2)}$, $\boldsymbol{\beta}^\star$, α_{j1} and α_{j2} with $\boldsymbol{A}_j^{(2,n_1)}$, $\hat{\boldsymbol{\beta}}$, n_{j1}/n and n_{j2}/n, respectively, in $I(\boldsymbol{\beta}^\star)$. The asymptotic variance and its approximation are the same as if the interventions were fixed in advance and $\bar{\boldsymbol{Y}}^{(1)}$ and $\bar{\boldsymbol{Y}}^{(2,n_1)}$ were independent.

2.3. *Hypothesis testing.* A major goal of a LAGO study is to test the null hypothesis of no overall intervention effect. One way to test this is to carry out a test for the subvector of $\boldsymbol{\beta}$ characterizing the effect of the intervention. That is, to test $H_0:\boldsymbol{\beta}_1=0$ in model (2.1) using the asymptotic normality result of Section 2.2. Because of this asymptotic normality result, the Wald or likelihood ratio tests for $H_0:\boldsymbol{\beta}_1=\boldsymbol{\beta}_1^0$ are asymptotically valid for any constant $\boldsymbol{\beta}_1^0$.

Alternatively, in a controlled LAGO design, let Q be a group indicator that equals one for the intervention group and zero for the control, and let p_0 and p_1 be the success probabilities under $Q=0$ and $Q=1$, respectively. Then, an alternative test for an overall intervention effect, $H_0:\boldsymbol{\beta}_1=0$, can be carried out by testing $H_0:p_0=p_1$. The latter test is valid despite the adaption of the intervention package, under the assumption that the arm allocation ratio (i.e. the assignment to control versus intervention arms) does not depend on the prior data, but only the intervention package composition depends on data from previous stages. By Assumption 2.1, the dependence between the stage 2 and stage 1 data is solely due to the stage 1 data determining the stage 2 recommended intervention, which, in turn, affects the actual stage 2 intervention, and thus the stage 2 outcomes. However, under the null, there is no effect of the actual intervention on the stage 2 outcomes. Therefore, under the null, regardless of the way the intervention was adapted, the stage 1 and stage 2 outcomes are independent. Thus, a

standard test for equal probabilities in the control and the intervention arms is valid. While not needed due to our asymptotic results, the same arguments could have been used for the standard tests of $H_0 : \boldsymbol{\beta}_1 = 0$.

In a controlled LAGO design, an alternative, possibly more powerful, test for the overall effect of the intervention in the presence of center characteristics is to consider $H_0 : \gamma = 0$ in the model $\text{logit}\, \tilde{p}_Q(\boldsymbol{\beta}, \gamma; \boldsymbol{z}) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_2^T \boldsymbol{z} + \gamma Q$. As before, in light of the between-stages independence under the null, $\boldsymbol{\beta}_1 = 0$ in model (2.1) implies $\gamma = 0$.

2.4. *Confidence sets and confidence bands.* After the conclusion of the study, the optimal intervention is estimated as the solution to (2.2) with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}$. To obtain an asymptotic 95% confidence set for the optimal intervention $\boldsymbol{x}^{opt}$, we first obtain a confidence interval for $p_{\boldsymbol{x}}(\boldsymbol{\beta}^\star; \tilde{\boldsymbol{z}})$, for a given $\boldsymbol{z} = \tilde{\boldsymbol{z}}$ and for each $\boldsymbol{x} \in \mathcal{X}$. To do this, we calculate a 95% confidence interval for $\text{logit}(p_{\boldsymbol{x}}(\boldsymbol{\beta}^\star; \tilde{\boldsymbol{z}}))$, i.e., for $(1\ \boldsymbol{x}^T\ \tilde{\boldsymbol{z}}^T)\boldsymbol{\beta}^\star$:

$$CI_{\boldsymbol{x}} = (1\ \boldsymbol{x}^T\ \tilde{\boldsymbol{z}}^T)\hat{\boldsymbol{\beta}} \pm 1.96\sigma(\hat{\boldsymbol{\beta}}; \boldsymbol{x}, \tilde{\boldsymbol{z}}),$$

where $\sigma^2(\hat{\boldsymbol{\beta}}; \boldsymbol{x}, \tilde{\boldsymbol{z}}) = (1\ \boldsymbol{x}^T\ \tilde{\boldsymbol{z}}^T) n^{-1} \hat{I}^{-1}(\hat{\boldsymbol{\beta}})(1\ \boldsymbol{x}^T\ \tilde{\boldsymbol{z}}^T)^T$ is the estimated variance of $(1\ \boldsymbol{x}^T\ \tilde{\boldsymbol{z}}^T)\hat{\boldsymbol{\beta}}$, and $n^{-1}\hat{I}^{-1}(\hat{\boldsymbol{\beta}})$ is the estimated variance of $\hat{\boldsymbol{\beta}}$. The 95% confidence interval for $p_{\boldsymbol{x}}(\boldsymbol{\beta}^\star; \tilde{\boldsymbol{z}})$ is $CI_{p_{\boldsymbol{x}}} = \text{expit}(CI_{\boldsymbol{x}})$. Then, we obtain the confidence set for the optimal intervention as $CS(\boldsymbol{x}^{opt}) = \{\boldsymbol{x} : CI_{p_{\boldsymbol{x}}} \ni \tilde{p}\}$. That is, $CS(\boldsymbol{x}^{opt})$ includes intervention packages for which $\tilde{p}$ is inside the confidence interval for the success probability under those interventions.

We now show that the confidence set $CS(\boldsymbol{x}^{opt})$ contains $\boldsymbol{x}^{opt}$ with the specified probability of 0.95. Recall that under the assumption that $\tilde{p}$ can be achieved, $p_{\boldsymbol{x}^{opt}}(\boldsymbol{\beta}^\star; \tilde{\boldsymbol{z}}) = \text{expit}[(1\ \boldsymbol{x}^{opt^T}\ \tilde{\boldsymbol{z}}^T)\boldsymbol{\beta}^\star] = \tilde{p}$. Therefore,

$$pr(CS(\boldsymbol{x}^{opt}) \ni \boldsymbol{x}^{opt}) = Pr(CI_{p_{\boldsymbol{x}^{opt}}} \ni \tilde{p}) = Pr(CI_{p_{\boldsymbol{x}^{opt}}} \ni p_{\boldsymbol{x}^{opt}}(\boldsymbol{\beta}^\star; \tilde{\boldsymbol{z}})) = 0.95.$$

Implementing this procedure is simple and its calculation is fast. Because calculating $CS(\boldsymbol{x}^{opt})$ does not depend upon estimating $\boldsymbol{x}^{opt}$, it does not involve the optimization algorithm.

At the end of the study, researchers might be interested in a variety of potential intervention packages in $\mathcal{X}$ that were not necessarily identified as of interest a priori. We propose a method to develop confidence bands for the outcome probabilities $p_{\boldsymbol{x}}(\boldsymbol{\beta}; \tilde{\boldsymbol{z}})$ for a range of $\boldsymbol{x} \in \mathcal{X}$ of interest, simultaneously. These confidence bands allow researchers to study the entire intervention space when comparing potential choices of the intervention package. We propose a procedure that is based on the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and on Scheffé's method (Scheffé, 1959). First, for all $\boldsymbol{x} \in \mathcal{X}$, construct

$CB_{\boldsymbol{x}}$ to obtain 95% confidence bands for $\{(1\ \boldsymbol{x}^T\ \tilde{\boldsymbol{z}}^T)\boldsymbol{\beta}^\star : \boldsymbol{x} \in \mathcal{X}\}$,

$$CB_{\boldsymbol{x}} = (1\ \boldsymbol{x}^T\ \tilde{\boldsymbol{z}}^T)\hat{\boldsymbol{\beta}} \pm \sqrt{\chi^2_{0.95,p+q+1}}\sigma(\hat{\boldsymbol{\beta}}; \boldsymbol{x}, \tilde{\boldsymbol{z}}),$$

with $\sigma(\hat{\boldsymbol{\beta}}; \boldsymbol{x}, \tilde{\boldsymbol{z}})$ defined as before and $\chi^2_{0.95,p+q+1}$ the 95% quantile of a χ^2_{p+q+1} distribution. As before, we transform $CB_{\boldsymbol{x}}$ into confidence bands for $p_{\boldsymbol{x}}(\boldsymbol{\beta}; \tilde{\boldsymbol{z}})$ by setting $CB_{p_{\boldsymbol{x}}} = \text{expit}(CB_{\boldsymbol{x}})$. These confidence bands guarantee asymptotic simultaneous 95% coverage for all possible intervention package compositions; the proof is given in Section 4 of the supplementary materials.

2.5. *Computation of the optimal intervention.* The algorithm used to solve (2.2) after stage k, using $\hat{\boldsymbol{\beta}}^{(k)}$, depends on the form of $C(\boldsymbol{x})$. Under a linear cost function with unit costs c_r for the r–th intervention component, the solution is achieved by 1. setting all components to their minimal value $\mathcal{L}_r$, 2. ordering the components by their estimated cost-efficiency $\hat{\beta}_{1r}/c_r$, and 3. increasing the most cost-efficient component until either $\tilde{p}$ is achieved or until this component reaches its maximal value, then moving to the next most cost-efficient component among the remaining components, and so on. For non-linear cost functions, standard non-linear optimization algorithms can be used.

3. Simulations. We conducted simulation studies to investigate the finite sample properties of our methods. We simulated 2000 data sets per simulation scenario. We considered three main scenarios.

1. In **Scenario 1**, we considered a two-stage controlled LAGO design with equal number of centers per stage J, with half the centers in the intervention arm and half in the control arm. The total sample size available at the end of the study is $J(n_{1j} + n_{2j})$. We considered the values $J = 6, 10, 20$, $n_{1j} = 50, 100, 200$, and $n_{2j} = 100, 200, 500, 1000$. The intervention had two components, $\boldsymbol{x} = (x_1, x_2)$, with unit costs $c_1 = 1$ and $c_2 = 8$. The minimum and maximum values of X_1 and X_2 were $[\mathcal{L}_1, \mathcal{U}_1] = [0, 2]$ and $[\mathcal{L}_2, \mathcal{U}_2] = [0, 5]$. We considered the following values for $\exp(\boldsymbol{\beta}_1^\star) = (\exp(\beta_{11}^\star), \exp(\beta_{12}^\star))$: $(1, 1)$ (the null), $(1, 1.2)$, $(1, 1.5)$, $(1.2, 1.5)$, and $(1.2, 2)$. A single center covariate z was normally distributed with mean 0 and variance 1 and its coefficient was taken to be $\beta_2^\star = \log(0.75)$. For simplicity, we did not include an intercept in model (2.1), although each center had its own baseline success probability due to z. For $z = 0$, the probability of success in the control arm was 0.5. The stage 2 recommended intervention was based on solving

the optimization problem (2.2) using the stage 1 estimates of $\boldsymbol{\beta}$. Section 5.1 of the supplementary materials provides the details on what was done when no solution existed for which $\tilde{p}$ was reached.

2. **Scenario 2** is similar to Scenario 1 with respect to true parameter values, cost functions and the center covariate. However, in Scenario 2, the per-center sample size is lower in stage 1 than in stage 2, and the number of centers is also lower in stage 1 than in stage 2. Thus, Scenario 2 reflects the potential desire in practice to learn the optimal intervention faster. This scenario is divided to Scenario 2a with $J_1 = 6$ and $J_2 = 12$ centers in stages 1 and 2, respectively, and Scenario 2b ($J_1 = 10$ and $J_2 = 20$). The per-center sample sizes are $n_{1j} = 50$ and $n_{2j} = 200$.
3. **Scenario 3** is carefully modeled after the illustrative example, the BetterBirth Study, described in Section 4. While the BetterBirth Study did not use a LAGO design, in this simulation study we investigated how LAGO would have performed had LAGO been used. All non-adaptive design parameters were determined by this study, including the stage 1 center-specific interventions, number of centers, per-center sample size, intervention arm allocation in each of the three stages of the trial, and the distribution of the center-specific covariate z, monthly birth volume, by taking them to be exactly as in the BetterBirth data. The true parameter values in Scenario 3 were the final estimators from the data (last column in Table 4). In each simulation iteration, stage 1 outcome data was first simulated, and then analyzed to determine stage 2 intervention. Then, stage 2 outcome data was simulated, and data from both stages were analyzed to derive stage 3 interventions. Stage 3 outcomes were simulated, and the entire data were analyzed to obtained the final estimators. In all stages, variation in the uptake of the intervention (specifically in the number of coaching visits) was simulated according to the actual variation in the data, at that specific stage.

Selected results for Scenarios 1 and 2 are presented in Tables 1 and 2. Table 1 presents results on the performance of $\hat{\beta}$, and shows that for $J > 6$, the finite sample bias was minimal, the mean estimated standard error was very close to the empirical standard deviation, and the empirical coverage rate of the confidence intervals for the effects of the individual package components was very close to 95%. With 2000 replicates per simulation scenario, the empirical coverage of 95% confidence intervals should lie between 94% and 96% (in 95% of the scenarios). This was indeed the case (Table 1). Moreover, in Section 5.2 of the supplementary materials, we found that the

type I error rate of the tests discussed in Section 2.3 was close to the nominal value of 0.05. However, in several scenarios explored, the finite sample bias was beyond that which could have been expected due to random simulation sampling error for 2000 replicates per simulation scenario, that is, in absolute value beyond $1.96SD(\hat{\beta})/\sqrt{2000}$, where $SD(\hat{\beta})$ is the empirical standard deviation of $\hat{\beta}$. This occurred more frequently for β_1 than for β_2 and for lower sample sizes and per-stage number of centers. When we further increased the sample size, this bias disappeared.

Table 2 presents bias and root mean square errors for the second-stage recommended intervention and the final estimated optimal intervention, calculated for a typical center with $z = 0$; additional results for Scenario 1 with $J = 6, 10$ are presented in Section 5.2 of the supplementary materials. The finite sample bias and the root mean squared errors of the final $\hat{\boldsymbol{x}}^{opt}$ were generally small and decreased as the number of centers per stage and the sample size increased. The bias of the second-stage recommended intervention was often much more substantial. Table 3 presents information about success probabilities under the second-stage recommended intervention and the final estimated optimal intervention. The empirical 2.5% and 97.5% quantiles of the true success rate show that the desired 90% was generally achieved with the final estimated optimal intervention, but less so with the second-stage recommended intervention. The nominal coverage rate of the confidence set for $\boldsymbol{x}^{opt}$ was approximately 95%, with the set typically including between 3 to 15 percent of $\mathcal{X}$, as a measure of precision in the scenarios studied. We also compared the cost of the estimated optimal intervention to the cost of the true optimal intervention and found it to be almost the same for the scenarios presented in Table 2; see Section 5.2 of the supplementary materials. Table 2 also shows that the empirical coverage rate of the confidence bands for $p_{\boldsymbol{x}}(\boldsymbol{\beta}^{\star}; z = 0)$ was very close to 95%.

The results from Scenario 3 are summarized in Section 5.2 of the supplementary. The results generally agreed with the results of Scenarios 1 and 2. Minimal bias was observed for the final estimated intervention component effects and estimated optimal intervention. However, the estimated optimal intervention in the earlier stages were generally biased, especially when stage 1 sample size was small. It should be noted that the intermediate recommended interventions or intervention effect estimates are not the goal of LAGO. Rather, the final estimated optimal intervention and final intervention effect estimates are the main output of a LAGO study.

4. Illustrative example. The BetterBirth Study consisted of three stages. The first two stages were pilot stages used to develop the intervention

TABLE 1

Simulation study: results for individual package component effects. Unit costs were $c_1 = 1$ and $c_2 = 8$.

$e^{\boldsymbol{\beta}^\star}$	n_{1j}	n_{2j}	J	$\hat{\beta}_{11}$ %RelBias	$\hat{\beta}_{11}$ $\frac{SE}{EMP.SD}$ ($\times 100$)	$\hat{\beta}_{11}$ CP95	$\hat{\beta}_{12}$ %RelBias	$\hat{\beta}_{12}$ $\frac{SE}{EMP.SD}$ ($\times 100$)	$\hat{\beta}_{12}$ CP95
Scenario 1 ($J_1 = J_2 = J$)									
$(1.2, 1.5)$	50	100	6	-2.3	96.5	95.1	-1.9	84.1	94.0
			10	-2.7	98.8	94.9	-1.2	92.2	95.2
			20	-1.4	101.3	95.2	-0.3	102.7	95.6
		200	6	-1.8	95.0	94.9	-2.6	81.0	95.4
			10	-4.4	92.7	94.2	-1.0	91.9	95.2
			20	-2.1	102.2	95.5	-0.2	99.7	95.2
	100	100	6	-1.7	92.9	94.7	-1.5	86.2	95.5
			10	2.8	101.9	95.7	-1.4	100.9	95.4
			20	2.1	101.1	95.5	-0.5	101.6	95.0
		200	6	-3.2	91.4	94.6	-0.8	83.6	95.5
			10	-1.6	99.5	95.4	-0.6	94.9	95.3
			20	-0.4	98.4	95.0	-0.3	97.5	94.5
$(1.2, 2)$	50	100	6	-16.0	91.6	95.4	0.7	86.0	96.0
			10	-7.4	101.4	95.8	0.2	102.2	96.0
			20	-3.6	99.6	95.2	-0.1	101.4	94.8
		200	6	-11.8	89.9	95.1	0.7	89.7	95.1
			10	-9.2	94.9	95.5	0.1	97.6	96.0
			20	-2.7	100.0	95.0	-0.2	101.4	96.2
	100	100	6	-7.6	94.5	95.8	-0.1	94.1	95.2
			10	-2.1	98.2	94.8	-0.0	102.7	95.2
			20	-3.7	100.3	95.2	0.2	102.7	95.5
		200	6	-7.1	84.6	95.2	0.3	95.8	95.9
			10	-4.6	96.4	94.7	0.0	99.6	95.5
			20	-3.5	98.0	94.6	0.1	104.8	95.9
Scenario 2a ($J_1 = 6$, $J_2 = 12$)									
$(1.2, 1.5)$	50	200		-3.8	96.4	95.5	-0.5	91.0	94.8
$(1.2, 2)$	50	200		-7.4	95.6	95.9	0.7	94.7	95.5
Scenario 2b ($J_1 = 10$, $J_2 = 20$)									
$(1.2, 1.5)$	50	200		-3.1	96.9	94.6	-0.7	95.5	95.5
$(1.2, 2)$	50	200		-6.2	93.4	94.7	0.2	100.1	95.2

%RelBias, percent relative bias $100(\hat{\beta} - \beta^\star)/\beta^\star$; SE, mean estimated standard error; EMP.SD, empirical standard deviation; CP95, empirical coverage rate of 95% confidence intervals.

TABLE 2

Simulation study: results for estimated optimal intervention package in stages 1 and 2. Unit costs were $c_1 = 1$ and $c_2 = 8$.

$e^{\boldsymbol{\beta}^\star}$	$\boldsymbol{x}^{opt}$	n_{1j}	n_{2j}	Stage 1			Stage 2		
				Bias_1 (×100)	Bias_2 (×100)	RMSE (×100)	Bias_1 (×100)	Bias_2 (×100)	RMSE (×100)
Scenario 1 ($J_1 = J_2 = 20$)									
(1, 2)	(0, 3.2)	50	100	52.8	-10.0	110.6	34.5	-4.7	85.0
			500	52.6	-11.5	110.5	16.5	-2.1	58.5
		100	100	35.0	-5.8	89.0	24.0	-2.5	71.0
			500	38.9	-7.5	93.0	10.6	-0.9	47.0
(1.2, 1.5)	(2, 4.5)	50	100	-30.0	-9.9	94.5	-9.5	2.7	51.6
			500	-30.7	-9.8	94.8	-2.7	2.1	27.8
		100	100	-14.9	-3.1	68.6	-3.6	1.2	35.9
			500	-16.6	-2.5	70.9	-0.7	1.7	18.1
(1.2, 2)	(2, 2.6)	50	100	-50.2	-0.5	106.3	-33.1	4.5	84.0
			500	-51.4	0.5	107.1	-14.9	3.3	56.6
		100	100	-35.8	1.7	88.2	-23.2	3.3	70.3
			500	-35.0	1.7	87.5	-8.8	2.3	43.6
Scenario 2a ($J_1 = 6$, $J_2 = 12$)									
(1, 2)	(0, 3.2)	50	200	76.0	-43.0	168.6	42.7	-8.1	96.6
(1.2, 1.5)	(2, 4.5)	50	200	-65.4	-92.2	210.8	-18.6	1.2	71.3
(1.2, 2)	(2, 2.6)	50	200	-81.0	-29.3	163.9	-44.4	3.0	98.4
Scenario 2b ($J_1 = 10$, $J_2 = 20$)									
(1, 2)	(0, 3.2)	50	200	66.4	-20.1	134.4	32.1	-4.8	82.2
(1.2, 1.5)	(2, 4.5)	50	200	-49.3	-33.1	141.3	-10.4	4.6	52.4
(1.2, 2)	(2, 2.6)	50	200	-68.6	-8.3	133.4	-32.6	4.2	83.3

Bias_1, bias of $\hat{x}_1^{opt}$; Bias_2, bias of $\hat{x}_2^{opt}$; RMSE, root of mean squared errors $\{\text{mean}(||\hat{\boldsymbol{x}}^{opt} - \boldsymbol{x}^{opt}||^2)\}^{1/2}$, mean taken over simulation iterations;

TABLE 3

Simulation study: results for estimated optimal intervention package in stages 1 and 2 and coverage of 95% confidence bands for success probabilities. .Unit costs were $c_1 = 1$ *and* $c_2 = 8$.

$e^{\boldsymbol{\beta}^\star}$	$\boldsymbol{x}^{opt}$	n_{1j}	n_{2j}	PrOpt1 (Q2.5,Q97.5)	PrOpt2 (Q2.5,Q97.5)	SetCP95	SetPerc%	BandsCP95
Scenario 1 ($J_1 = J_2 = 20$)								
$(1, 2)$	$(0, 3.2)$	50	100	(83.6, 93.8)	(87.2, 91.8)	94.0	7.6	97.0
			500	(83.5, 93.7)	(88.2, 91.1)	95.0	4.0	97.2
		100	100	(85.2, 93.1)	(87.8, 91.6)	94.8	6.3	96.5
			500	(85.6, 92.8)	(88.8, 91.0)	95.3	3.7	97.4
$(1.2, 1.5)$	$(2, 4.5)$	50	100	(81.1, 91.6)	(87.3, 91.6)	94.8	13.3	96.0
			500	(81.9, 91.6)	(88.8, 91.3)	95.1	7.6	95.9
		100	100	(84.7, 91.6)	(87.9, 91.6)	94.8	12.3	95.4
			500	(84.0, 91.6)	(89.0, 91.1)	95.3	7.1	95.4
$(1.2, 2)$	$(2, 2.6)$	50	100	(83.3, 93.2)	(87.2, 91.7)	94.6	14.3	95.5
			500	(83.7, 93.3)	(88.5, 91.2)	94.4	8.1	95.3
		100	100	(85.6, 92.4)	(87.7, 91.5)	95.6	12.4	96.0
			500	(85.3, 92.5)	(88.7, 91.1)	95.1	7.5	95.8
Scenario 2a ($J_1 = 6$ $J_2 = 12$)								
$(1, 2)$	$(0, 3.2)$	50	200	(50.0, 97.0)	(85.6, 92.2)	94.7	9.8	97.5
$(1.2, 1.5)$	$(2, 4.5)$	50	200	(56.8, 91.6)	(85.8, 91.6)	95.1	17.3	95.7
$(1.2, 2)$	$(2, 2.6)$	50	200	(56.7, 97.3)	(85.5, 92.0)	95.8	17.1	97.2
Scenario 2b ($J_1 = 10$ $J_2 = 20$)								
$(1, 2)$	$(0, 3.2)$	50	200	(78.7,95.5)	(87.1, 91.6)	94.7	6.6	96.8
$(1.2, 1.5)$	$(2, 4.5)$	50	200	(70.0,91.6)	(87.5, 91.6)	95.6	11.8	95.4
$(1.2, 2)$	$(2, 2.6)$	50	200	(75.6,95.2)	(87.2, 91.4)	95.2	12.4	96.3

PrOpt1, success probability of the second-stage recommended intervention, calculated using true coefficient values; PrOpt2, success probability of the final estimated optimal intervention, calculated using true coefficient values; Q2.5 and Q97.5, 2.5% and 97.5% quantiles; SetCP95, empirical coverage percentage of confidence set for optimal intervention; SetPerc%, mean percent of $\mathcal{X}$ covered by the confidence set; BandsCP95, empirical coverage rate of 95% confidence bands for $\{p_{\boldsymbol{x}}(\boldsymbol{\beta}; \boldsymbol{z} = 0) : x \in \mathcal{X}\}$.

package. Stage 3 was a randomized controlled trial. The development of the recommended intervention package was conducted qualitatively, as described in Hirschhorn et al. (2015), and the intervention package was adjusted after each pilot stage. The results of stage 3, the randomized controlled trial, were presented and discussed in Semrau et al. (2017). The number of centers with data on oxytocin administration in the first, second, and third stages was 2, 4 and 30, respectively. In the first two stages, data in each center were collected before and after the intervention was implemented. In stage 3, there were 15 centers in the control arm and 15 centers in the intervention arm. In 5 intervention arm centers, outcome data were also collected before the intervention was implemented.

Here, we focus on the binary outcome of oxytocin administration immediately after delivery, as recommended by the WHO (WHO, 2012) to prevent postpartum hemorrhage, a major cause of maternal mortality. The intervention package components were the duration of the on-site intervention launch (in days), the number of coaching visits after the intervention was launched, leadership engagement (non-standardized initial engagement, standardized initial engagement, and standardized initial engagement with follow-up visits) and data feedback (none; ongoing, paper-based; ongoing, app-based). The four components were adapted in a way that resulted in near multicollinearity. Therefore, for illustration purposes, we considered the first two components only, launch duration and number of coaching visits. The launch duration was 3 days in stage 1 and 2 days in stages 2 and 3. Compared to stage 1, the intensity of coaching visits was increased in stage 2, and further increased in stage 3. For illustrative purposes, we truncated the data at 40 coaching visits or less. The baseline center characteristic we included was the approximate monthly birth volume, given that large facilities might be likely to follow WHO recommendations about oxytocin administration more closely, regardless of the intervention package implemented. Other available center characteristics, e.g. number of staff nurses, were highly correlated with the monthly birth volume.

Table 4 provides the estimated effects of the intervention package components after each of the stages, using all available data at that point. The sample size in stage 1 was relatively small, explaining the wide confidence intervals for the odds ratios. The final results imply that both package components had an effect. Tests for the overall effect of the package yielded a highly significant p-value, regardless of the test we used.

After consulting with the study investigators, we assigned unit costs of \$800 per launch day and \$170 per coaching visit. In practice, implementation costs may also depend on center size and, if so, $C(\boldsymbol{x})$ could be replaced with

TABLE 4
Package component effect estimates and confidence intervals, calculated after each stage.

	Stage 1 $n_1 = 73$ OR (CI-OR)	Stages 1-2 $(n_1 + n_2 = 1780)$ OR (CI-OR)	Stages 1-3 $(n_1 + n_2 + n_3 = 6124)$ OR (CI-OR)
Intercept	1.07 (0.00, 280.80)	0.10 (0.07,0.15)	0.10 (0.09,0.11)
Coaching Visits (per 3 visits)	7.95 (1.77,73.95)	1.11 (0.96,1.28)	1.08 (1.04,1.12)
Launch Duration (days)	1.41 (0.76,2.64)	2.65 (1.95,3.77)	2.79 (2.41,3.23)
Birth Volume (monthly, per 100)	0.37 (0.00,32.33)	2.11 (1.93,2.33)	1.94 (1.84,2.06)
	$\hat{\boldsymbol{x}}^{opt,(2,n_1)} = (1, 5)$	$\hat{\boldsymbol{x}}^{opt,(3,(n_1,n_2))} = (3, 1)$	$\hat{\boldsymbol{x}}^{opt} = (3, 1)$

OR, estimated odds ratio $\exp(\hat{\beta})$; CI-OR, 95% Confidence interval for the odds ratio. In the estimated optimal interventions, the first component is the launch duration (in days) and the second component is the number of coaching visits .

$C_{\boldsymbol{z}}(\boldsymbol{x})$.

The estimation of the optimal intervention package with linear cost $C(\boldsymbol{x}) = c_1x_1 + c_2x_2$ was conducted as in the simulation study. Assuming that at least 1 launch day and 1 coaching visit are needed, and that a launch duration of more than 5 days or having more than 40 coaching visits is impractical, we estimated the optimal intervention for a center with average birth volume ($z = 175$) to be a launch duration of 2.78 days and 1 coaching visit. We also carried out optimization over all possible combinations of discrete values within $\mathcal{X}$, which are $1, ..., 40$ for coaching visits and $1, 1.5, 2, 2.5, ..., 5$ for duration of intervention launch and obtained the optimal intervention as launch duration of three days with one coaching visit, $\hat{\boldsymbol{x}}^{opt} = (3, 1)$. The total cost of the estimated optimal intervention package, $\hat{\boldsymbol{x}}^{opt}$, was \$2570.

We calculated a 95% confidence set for the optimal intervention $CS(\boldsymbol{x}^{opt})$ over the grid of $\mathcal{X}$, taking all possible numbers of coaching visits, $1, ..., 40$, and $1, 1.5., 2, 2.5, ..., 5$ for intervention launch duration. Out of 360 potential intervention packages, 38 (10.5%) were included in the 95% confidence set. The set included the following combinations: 1.5 days launch duration and 40 coaching visits; 2 days launch durations and 27 or more coaching visits; 2.5 days launch duration and less than 20 coaching visits; and 3 days launch duration and less than 5 coaching visits. The first, second and third quartiles of the cost distribution within $CS(\boldsymbol{x}^{opt})$ were $Q1$=\$2462, $Q2$=\$4035, and $Q3$=\$6797. We also calculated 95% simultaneous confidence bands for the probability of success under all 360 intervention compositions; plots are shown in Section 6 of the supplementary materials. For the estimated optimal intervention $\hat{\boldsymbol{x}}^{opt} = (1, 3)$, the obtained confidence interval (within the

bands) for the probability of oxytocin administration was $(0.79, 0.93)$. The mean difference between the top and bottom of the confidence band over all 360 intervention compositions was 0.07.

5. Discussion. We developed the LAGO design for multiple component intervention studies with a binary outcome, where the intervention package composition is systematically adapted as part of the design. The goals of studies using the LAGO design are to find the optimal intervention package, to test its effect on the outcome of interest, and to estimate its effect as well as the effects of the individual components.

The methodology in this paper was developed for scenarios with a stage-wise analysis that does not include formal interim hypothesis testing. However, the LAGO design allows for futility stops, since stopping the trial for futility between stages preserves the type I error. The type I error can only decrease from the nominal level when futility stops are included, because when stopping for futility, the null hypothesis is not rejected (Snapinn et al., 2006).

For clear presentation of the design, methods, and theory, we focused on a general yet practical design. Our work opens the way for further research. For example, it would be interesting to develop methods for studies with further dependence because centers contribute data to more than one stage. The results in this paper could also be extended to continuous, count, or survival outcome data. Adapting the LAGO framework to paired data would also be useful. Additionally, many design problems arise, in terms of identifying the optimal K, J_k and n_{jk} for given settings. It should be noted that the performance of estimators obtained from a LAGO trial depends on the choice of the function g, which determines how the later stage interventions depend on the data from previous stages. Therefore, an important topic for future research is the choice of g.

Our asymptotic results use the assumption that the sample sizes in the different stages increase at a similar rate, in the sense that the ratio between the sample size in each of the stages and the overall sample size converges to a constant, which can be small. Even when the stage 1 sample size was relatively small, we showed in simulation Scenario 3 that the asymptotic properties were still good approximations of the finite sample behavior of the final estimators. On the other hand, even when the stage 1 sample size is large, further data collection in a second stage is often desirable to avoid excessive extrapolation of the outcome model to intervention packages that have not been implemented in stage 1, minimizing the potential for bias due to model misspecification. In practice, researchers will usually prefer to

observe the performance of the optimal intervention before reaching final conclusions.

In this paper, we assumed that center effects can be fully captured by observed covariates and that the intervention effects are fixed across centers. In the BetterBirth Study, for example, we assumed that the monthly birth volume captured center effects. This is a limitation of the presented work, because, in practice, center effects often cannot be captured solely by observed covariates. Therefore, future work will consider generalizing LAGO to allow for clustered data.

Van der Laan (2008) provides rigorous proofs for specific adaptive designs which do not include LAGO, while providing "templates and conditions" for more general settings. As in LAGO, van der Laan (2008) considers settings where the intervention of patient i depends on the information available of previous patients, and where the limiting design is a fixed design. However, van der Laan (2008) is not directly applicable to LAGO as developed in this paper. In the LAGO design, the number of stages, i.e., the number of times the intervention could be adapted, is finite and fixed, while van der Laan (2008) would require that the number of stages tends to infinity. However, following the same arguments as in van der Laan (2008) page 11, the LAGO estimating equations form a martingale and it might be possible to apply a triangular Martingale Central Limit Theorem instead of the Martingale Central Limit Theorem referenced in van der Laan (2008), to develop theory for LAGO both for settings with a large number of patients per stage and for settings with smaller numbers of patients per stage; it might also be useful for extending LAGO to continuous and time-to-event outcomes.

In this paper, we considered the model parameter values fixed, and not dependent on the sample size n. As a result, the limiting design, that is, the probability limit of the intervention package composition, is constant in all stages. An interesting direction for future research involves studying the asymptotic regime when the parameter values themselves change with n, and specifically sequences of distributions where the intervention component effects ($\boldsymbol{\beta}$) go to zero at rate $n^{-1/2}$ (known as local alternatives, see e.g. Chapter 14 in van der Vaart (1998)). In this setting of local alternatives, even in the limit for large n, the later stage interventions will not converge to a constant but may have a limiting distribution. The resulting asymptotic theory might lead to better approximations for finite sample situations where there is less certainty about the later stage interventions.

Many large effectiveness and implementation trials fail because current design methodology does not permit adaptation of the intervention in the face of implementation failure as in, for example, the BetterBirth (Semrau

et al., 2017) and the TasP (Iwuji et al., 2017) studies. The LAGO design rigorously formalizes practices in public health research that are presently conducted in an ad hoc manner, with unknown consequences for the validity of the subsequent standard analysis (Escoffery et al., 2018). We expect widespread use of the LAGO design as a result, with potential gain for many randomized clinical trials.

APPENDIX A: PROOFS OF THEOREMS 1 AND 2

As previously explained, we prove the results in the paper for a general recommended interventions $\boldsymbol{X}_j^{(2,n_1)} = g(\bar{\boldsymbol{A}}^{(1)}, \bar{\boldsymbol{Y}}^{(1)}, \bar{\boldsymbol{z}}^{(1)}, \boldsymbol{z}_j^{(2)})$. Usually $\boldsymbol{X}_j^{(2,n_1)}$ will be the estimated optimal intervention (previously denoted as $\hat{\boldsymbol{x}}_j^{opt,(2,n_1)}$). The proof works, however, for any function of the data such that $\boldsymbol{X}_j^{(2,n_1)}$ converges in probability to a center-specific limit $\boldsymbol{x}_j^{(2)}$, for all $j = 1, ..., J_2$. Let $\bar{\boldsymbol{X}}^{(2,n_1)} = (\boldsymbol{X}_j^{(2,n_1)}, ..., \boldsymbol{X}_{J_1}^{(2,n_1)})$ be all the stage 2 recommended interventions.

A.1. Proof of Theorem 1: consistency of $\hat{\boldsymbol{\beta}}$. The following Lemma will be useful for the proof of Theorem 1.

LEMMA A.1. *Let $f(\boldsymbol{x};\boldsymbol{\beta}) : \mathcal{X} \to \boldsymbol{R}^q$ be a differentiable function of $\boldsymbol{x}$ with continuous and bounded first partial derivatives for all $\boldsymbol{x} \in \mathcal{X}$ ($\boldsymbol{\beta} \in \mathcal{B}$), uniformly bounded over $\mathcal{X} \times \mathcal{B}$, where $\mathcal{X}$ and $\mathcal{B}$ are compact sets in $\boldsymbol{R}^p$. Let $\boldsymbol{X}_n$ be a sequence of random vectors with support in $\boldsymbol{R}^d$. If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\sup_{\boldsymbol{\beta}} ||f(\boldsymbol{X}_n;\boldsymbol{\beta}) - f(\boldsymbol{X};\boldsymbol{\beta})|| \xrightarrow{P} 0$.*

PROOF. First, observe that

$$\begin{aligned} \sup_{\boldsymbol{\beta}} ||f(\boldsymbol{X}_n;\boldsymbol{\beta}) - f(\boldsymbol{X};\boldsymbol{\beta})|| &= \sup_{\boldsymbol{\beta}} \sqrt{\sum_{r=1}^{q} [f_r(\boldsymbol{X}_n;\boldsymbol{\beta}) - f_r(\boldsymbol{X};\boldsymbol{\beta})]^2} \\ &= \sqrt{\sup_{\boldsymbol{\beta}} \sum_{r=1}^{q} [f_r(\boldsymbol{X}_n;\boldsymbol{\beta}) - f_r(\boldsymbol{X};\boldsymbol{\beta})]^2}. \end{aligned} \tag{A.1}$$

We will show that $[\sup_{\boldsymbol{\beta}} ||f(\boldsymbol{X}_n;\boldsymbol{\beta}) - f(\boldsymbol{X};\boldsymbol{\beta})||]^2 \xrightarrow{P} 0$ and hence $\sup_{\boldsymbol{\beta}} ||f(\boldsymbol{X}_n;\boldsymbol{\beta}) - f(\boldsymbol{X};\boldsymbol{\beta})|| \xrightarrow{P} 0$. We have

$$\sup_{\boldsymbol{\beta}} \sum_{r=1}^{q} [f_r(\boldsymbol{X}_n;\boldsymbol{\beta}) - f_r(\boldsymbol{X};\boldsymbol{\beta})]^2 \leq \sum_{r=1}^{q} \sup_{\boldsymbol{\beta}} [f_r(\boldsymbol{X}_n;\boldsymbol{\beta}) - f_r(\boldsymbol{X};\boldsymbol{\beta})]^2. \tag{A.2}$$

For each $r = 1, ..., q$, because of the mean value theorem for f_r, there exists $\tilde{\boldsymbol{X}}_r(\boldsymbol{\beta})$ between $\boldsymbol{X}_n$ and $\boldsymbol{X}$ such that

$$f_r(\boldsymbol{X}_n;\boldsymbol{\beta}) - f_r(\boldsymbol{X};\boldsymbol{\beta}) = \left[\frac{\partial}{\partial \boldsymbol{x}} f_r(\tilde{\boldsymbol{X}}_r(\boldsymbol{\beta}),\boldsymbol{\beta})\right]^T (\boldsymbol{X}_n - \boldsymbol{X}). \tag{A.3}$$

Combining (A.1), (A.2) and (A.3), we have

$$\begin{aligned}
[\sup_{\boldsymbol{\beta}} ||f(\boldsymbol{X}_n;\boldsymbol{\beta}) - f(\boldsymbol{X};\boldsymbol{\beta})||]^2 &\leq \sum_{r=1}^{q} \sup_{\boldsymbol{\beta}} \left\{ \left[\left(\frac{\partial}{\partial \boldsymbol{x}} f_r(\tilde{\boldsymbol{X}}_r(\boldsymbol{\beta});\boldsymbol{\beta}) \right)^T (\boldsymbol{X}_n - \boldsymbol{X}) \right]^2 \right\} \\
&\leq \sum_{r=1}^{q} \sup_{\boldsymbol{\beta}} \left[\left|\left| \frac{\partial}{\partial \boldsymbol{x}} f_r(\tilde{\boldsymbol{X}}_r(\boldsymbol{\beta}),\boldsymbol{\beta}) \right|\right|^2 ||\boldsymbol{X}_n - \boldsymbol{X}||^2 \right] \\
&= ||\boldsymbol{X}_n - \boldsymbol{X}||^2 \sum_{r=1}^{q} \sup_{\boldsymbol{\beta}} \left|\left| \frac{\partial}{\partial \boldsymbol{x}} f_r(\tilde{\boldsymbol{X}}_r(\boldsymbol{\beta}),\boldsymbol{\beta}) \right|\right|^2,
\end{aligned}$$

where the second line follows by the Cauchy–Schwarz inequality. Lemma A.1 follows, because $||\boldsymbol{X}_n - \boldsymbol{X}||^2 \xrightarrow{P} 0$ and because the components of $\frac{\partial}{\partial \boldsymbol{x}} f_r(\boldsymbol{x};\boldsymbol{\beta})$ are bounded uniformly in $\boldsymbol{x}$ and $\boldsymbol{\beta}$ since $\boldsymbol{X}$ and $\boldsymbol{\beta}$ take values in a compact space. □

We are now ready to prove Theorem 1 (consistency of $\hat{\boldsymbol{\beta}}$).

PROOF. To prove consistency of $\hat{\boldsymbol{\beta}}$, we invoke Theorem 5.9 of van der Vaart (1998). Let

$$\begin{aligned}
\boldsymbol{u}(\boldsymbol{\beta}) = &\sum_{j=1}^{J_1} \alpha_{j1} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \left(p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(1)}) \right) \\
&+ \sum_{j=1}^{J_2} \alpha_{j2} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(2)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \left(p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)}) \right).
\end{aligned} \tag{A.4}$$

We show that the two conditions needed for Theorem 5.9 of van der Vaart (1998) hold. First, we prove uniform convergence over $\mathcal{B}$ of $\boldsymbol{U}(\boldsymbol{\beta})$ to $\boldsymbol{u}(\boldsymbol{\beta})$:

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} ||\boldsymbol{U}(\boldsymbol{\beta}) - \boldsymbol{u}(\boldsymbol{\beta})|| \xrightarrow{P} 0. \tag{A.5}$$

Recall Equation (2.3) and rewrite $\boldsymbol{U}(\boldsymbol{\beta})$ as

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{\beta}^\star) + \sum_{j=1}^{J_1} \frac{n_{j1}}{n} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \left(p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(1)}) \right) + \sum_{j=1}^{J_2} \frac{n_{j2}}{n} \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \left(p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)}) \right).$$

Therefore,

$$\boldsymbol{U}(\boldsymbol{\beta}) - \boldsymbol{u}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{\beta}^\star) + \boldsymbol{G}_1 + \boldsymbol{G}_2 + \boldsymbol{G}_3 + \boldsymbol{G}_4 + \boldsymbol{G}_5$$

where

$$\boldsymbol{G}_1 = \sum_{j=1}^{J_1} \left(\frac{n_{j1}}{n} - \alpha_{j1} \right) \left[\begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \left(p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(1)}) \right) \right]$$

$$\boldsymbol{G}_2 = \sum_{j=1}^{J_2} \alpha_{j2} \left[\begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(2)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right]$$

$$\boldsymbol{G}_3 = \sum_{j=1}^{J_2} \alpha_{j2} \left[\begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(2)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)}) - \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)}) \right]$$

$$\boldsymbol{G}_4 = \sum_{j=1}^{J_2} \left(\frac{n_{j2}}{n} - \alpha_{j2} \right) \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})$$

$$\boldsymbol{G}_5 = \sum_{j=1}^{J_2} \left(\alpha_{j2} - \frac{n_{j2}}{n} \right) \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)}).$$

By the triangular inequality for the supremum norm, we can analyze each of the terms $\boldsymbol{U}(\boldsymbol{\beta}^\star), \boldsymbol{G}_1, ..., \boldsymbol{G}_5$, separately.

Regarding $\boldsymbol{U}(\boldsymbol{\beta}^\star)$, we show that its expectation is zero and the variance of each of the $1 + p + q$ components of $\boldsymbol{U}(\boldsymbol{\beta}^\star)$ converges to zero, and thus, by applying Chebychev's inequality, $\boldsymbol{U}(\boldsymbol{\beta}^\star) \xrightarrow{P} 0$.

By the law of iterated expectations, we have

$$
\begin{aligned}
E(\boldsymbol{U}(\boldsymbol{\beta}^\star)) = \frac{1}{n}\Bigg\{ & \sum_{j=1}^{J_1}\sum_{i=1}^{n_{j1}} \left[\begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} E\left(Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) \right) \right] \\
& + \sum_{j=1}^{J_2}\sum_{i=1}^{n_{j2}} E\left[\begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} E\left[Y_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \middle| \boldsymbol{A}_j^{(2,n_1)} \right] \right] \Bigg\} = 0.
\end{aligned} \tag{A.6}
$$

We now turn to the variance. The random vector $\boldsymbol{U}(\boldsymbol{\beta}^\star)$ is a sum of two vectors, one for each stage. We first show that these two vectors are uncorrelated. Let $Q_{j,j'} = \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{A}_{j'}^{(2,n_1)} \\ \boldsymbol{z}_{j'}^{(1)} \end{pmatrix}^T$. For any i, i', j and j', we have

$$
\begin{aligned}
& E\left[\begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \left(Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) \right) \begin{pmatrix} 1 \\ \boldsymbol{A}_{j'}^{(2,n_1)} \\ \boldsymbol{z}_{j'}^{(1)} \end{pmatrix}^T \left(Y_{i'j'}^{(2,n_1)} - p_{\boldsymbol{A}_{j'}^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_{j'}^{(2)}) \right) \right] \\
& = E\left\{ Q_{j,j'} E\left[\left(Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) \right) \left(Y_{i'j'}^{(2,n_1)} - p_{\boldsymbol{A}_{j'}^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_{j'}^{(2)}) \right) \middle| \boldsymbol{X}_{j'}^{(2,n_1)} \right] \right\} \\
& = E\left\{ Q_{j,j'} E\left[Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) \middle| \boldsymbol{X}_{j'}^{(2,n_1)} \right] E\left[Y_{i'j'}^{(2,n_1)} - p_{\boldsymbol{A}_{j'}^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_{j'}^{(2)}) \middle| \boldsymbol{X}_{j'}^{(2,n_1)} \right] \right\} \\
& = E\left\{ Q_{j,j'} E\left[Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) \middle| \boldsymbol{X}_{j'}^{(2,n_1)} \right] E\left[Y_{i'j'}^{(2,n_1)} - p_{\boldsymbol{A}_{j'}^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_{j'}^{(2)}) \middle| \boldsymbol{X}_{j'}^{(2,n_1)}, \boldsymbol{A}_{j'}^{(2,n_1)} \right] \right\} \\
& = E\left\{ Q_{j,j'} E\left[Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) \middle| \boldsymbol{X}_{j'}^{(2,n_1)} \right] \cdot 0 \right\} = 0,
\end{aligned} \tag{A.7}
$$

where the second equality is justified since the two factors are conditionally independent given $\boldsymbol{X}_j^{(2,n_1)}$ by Assumption 1. Then, by the linearity of the covariance, we get that the two vectors in $\boldsymbol{U}(\boldsymbol{\beta}^\star)$ are uncorrelated.

Denote $DiagVar(\boldsymbol{V})$ for the diagonal of the covariance matrix of a random vector $\boldsymbol{V}$. Define $\tau^2(\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{\beta})$ as

$$
\tau^2(\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{\beta}) = p_{\boldsymbol{a}}(\boldsymbol{\beta}; \boldsymbol{z})(1 - p_{\boldsymbol{a}}(\boldsymbol{\beta}; \boldsymbol{z})), \tag{A.8}
$$

and observe that for each $j = 1, ..., J_2$, by the law of total variance, we have

$$
\begin{aligned}
&DiagVar\left[\frac{1}{\sqrt{n}}\begin{pmatrix}1\\ \boldsymbol{A}_j^{(2,n_1)}\\ \boldsymbol{z}_j^{(2)}\end{pmatrix}\sum_{i=1}^{n_{j2}}\left(Y_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})\right)\right] \\
&= \frac{n_{j2}}{n}E\left\{DiagVar\left[\frac{1}{\sqrt{n_{j2}}}\begin{pmatrix}1\\ \boldsymbol{A}_j^{(2,n_1)}\\ \boldsymbol{z}_j^{(2)}\end{pmatrix}\sum_{i=1}^{n_{j2}}\left(Y_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})\right)\middle|\boldsymbol{A}_j^{(2,n_1)}\right]\right\} \\
&+ DiagVar\left\{\frac{1}{\sqrt{n}}E\left[\begin{pmatrix}1\\ \boldsymbol{A}_j^{(2,n_1)}\\ \boldsymbol{z}_j^{(2)}\end{pmatrix}\sum_{i=1}^{n_{j2}}\left(Y_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})\right)\middle|\boldsymbol{A}_j^{(2,n_1)}\right]\right\} \\
&= \frac{n_{j2}}{n}E\left[\begin{pmatrix}1\\ \boldsymbol{A}_j^{(2,n_1)}\\ \boldsymbol{z}_j^{(2)}\end{pmatrix}\circ\begin{pmatrix}1\\ \boldsymbol{A}_j^{(2,n_1)}\\ \boldsymbol{z}_j^{(2)}\end{pmatrix}\tau^2(\boldsymbol{A}_j^{(2,n_1)}, \boldsymbol{z}_j^{(2)}, \boldsymbol{\beta}^\star)\right] + DiagVar\left(\frac{1}{\sqrt{n}}0\right)
\end{aligned}
$$

$$
\to \alpha_{j2}\begin{pmatrix}1\\ \boldsymbol{a}_j^{(2)}\\ \boldsymbol{z}_j^{(2)}\end{pmatrix}\circ\begin{pmatrix}1\\ \boldsymbol{a}_j^{(2)}\\ \boldsymbol{z}_j^{(2)}\end{pmatrix}\tau^2(\boldsymbol{a}_j^{(2)}, \boldsymbol{z}_j^{(2)}, \boldsymbol{\beta}^\star), \tag{A.9}
$$

with $\circ$ being the element-wise Schur product, $(\boldsymbol{u}\circ\boldsymbol{v})_i = u_i v_i$, for any two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, and where the last line is justified by Lebesgue's Dominated Convergence Theorem, because the $\boldsymbol{A}_j^{(2,n_1)}$'s take values in a compact space, the $\boldsymbol{z}_j^{(2)}$'s are finite, and $\boldsymbol{A}_j^{(2,n_1)} \xrightarrow{P} a_j^{(2)}$. It is easy to see that similar reasoning can be applied to the variance of the first term, leading to

$$
\begin{aligned}
&DiagVar\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{J_1}\begin{pmatrix}1\\ \boldsymbol{a}_j^{(1)}\\ \boldsymbol{z}_j^{(1)}\end{pmatrix}\sum_{i=1}^{n_{j1}}\left(Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)})\right)\right] \\
&\to \sum_{j=1}^{J_1}\alpha_{j1}\begin{pmatrix}1\\ (\boldsymbol{a}_j^{(1)})\\ (\boldsymbol{z}_j^{(1)})\end{pmatrix}\circ\begin{pmatrix}1\\ (\boldsymbol{a}_j^{(1)})\\ (\boldsymbol{z}_j^{(1)})\end{pmatrix}\tau^2(\boldsymbol{a}_j^{(1)}, \boldsymbol{z}_j^{(1)}, \boldsymbol{\beta}^\star).
\end{aligned} \tag{A.10}
$$

Combining (A.7)–(A.10), we obtain

$$DiagVar[\sqrt{n}\boldsymbol{U}(\boldsymbol{\beta}^\star)] \to \sum_{j=1}^{J_1} \alpha_{j1} \begin{pmatrix} 1 \\ (\boldsymbol{a}_j^{(1)}) \\ (\boldsymbol{z}_j^{(1)}) \end{pmatrix} \circ \begin{pmatrix} 1 \\ (\boldsymbol{a}_j^{(1)}) \\ (\boldsymbol{z}_j^{(1)}) \end{pmatrix} \tau^2(\boldsymbol{a}_j^{(1)}, \boldsymbol{z}_j^{(1)}, \boldsymbol{\beta}^\star) + \sum_{j=1}^{J_2} \alpha_{j2} \begin{pmatrix} 1 \\ (\boldsymbol{a}_j^{(2)}) \\ (\boldsymbol{z}_j^{(2)}) \end{pmatrix} \circ \begin{pmatrix} 1 \\ (\boldsymbol{a}_j^{(2)}) \\ (\boldsymbol{z}_j^{(2)}) \end{pmatrix} \tau^2(\boldsymbol{a}_j^{(2)}, \boldsymbol{z}_j^{(2)}, \boldsymbol{\beta}^\star),$$

which is finite, and we conclude that $DiagVar[\boldsymbol{U}(\boldsymbol{\beta}^\star)]$ is $o(1)$. Therefore, by applying Chebyshev's inequality to each component of $\boldsymbol{U}(\boldsymbol{\beta}^\star)$, we obtain $\boldsymbol{U}(\boldsymbol{\beta}^\star) \xrightarrow{P} 0$. Since $\boldsymbol{U}(\boldsymbol{\beta}^\star)$ is not a function of $\boldsymbol{\beta}$, its supremum over $\boldsymbol{\beta}$ is its value at $\boldsymbol{\beta}^\star$, which we just showed converges in probability to zero.

Regarding $\boldsymbol{G}_2$, like $\boldsymbol{U}(\boldsymbol{\beta}^\star)$, it does not involve $\boldsymbol{\beta}$. Recall that $\boldsymbol{A}_j^{(2,n_1)} \xrightarrow{P} \boldsymbol{a}_j^{(2)}$. Therefore, since $f_1(\boldsymbol{a}; \boldsymbol{\beta}, \boldsymbol{z}) = \boldsymbol{a} p_{\boldsymbol{a}}(\boldsymbol{\beta}; \boldsymbol{z})$ and $f_2(\boldsymbol{a}; \boldsymbol{\beta}, \boldsymbol{z}) = c p_{\boldsymbol{a}}(\boldsymbol{\beta}; \boldsymbol{z})$, for any constant c, are continuous in $\boldsymbol{a}$ for all $\boldsymbol{\beta} \in \mathcal{B}$, $\boldsymbol{G}_2 \xrightarrow{P} 0$ by the Continuous Mapping Theorem.

To show that the supremum over $\boldsymbol{\beta}$ of $\boldsymbol{G}_3$ converges to zero, we can use Lemma A.1 for each j, since the function $f(\boldsymbol{a}, \boldsymbol{\beta}; \alpha, \boldsymbol{z}) = \alpha \begin{pmatrix} 1 & \boldsymbol{a}^T & \boldsymbol{z}^T \end{pmatrix}^T p_{\boldsymbol{a}}(\boldsymbol{\beta}; \boldsymbol{z})$ is continuous with bounded derivatives with respect to $\boldsymbol{a}$ for all $\boldsymbol{\beta} \in \mathcal{B}$, and because $\mathcal{B}$ is compact, and because $\boldsymbol{A}_j^{(2,n_1)} \xrightarrow{P} \boldsymbol{a}_j^{(2)}$. Thus, $\sup_{\boldsymbol{\beta}} ||f(\boldsymbol{A}_j^{(2,n_1)}, \boldsymbol{\beta}; \alpha, \boldsymbol{z}) - f(\boldsymbol{a}_j^{(2)}, \boldsymbol{\beta}; \alpha, \boldsymbol{z})||$ converges in probability to zero for all j, and we assumed that J_2 is finite.

The convergence of n_{j2}/n to α_{j2}, and the boundedness of $f_1(\boldsymbol{a}; \boldsymbol{\beta})$ and $f_2(\boldsymbol{a}; \boldsymbol{\beta})$, uniformly in $\boldsymbol{\beta} \in \mathcal{B}$, implies that the supremums of $\boldsymbol{G}_1$, $\boldsymbol{G}_4$ and $\boldsymbol{G}_5$ each converges in probability to zero. Equation (A.5) follows.

The second condition in Theorem 5.9 of van der Vaart (1998) is

$$\inf_{\boldsymbol{\beta}: ||\boldsymbol{\beta} - \boldsymbol{\beta}^\star|| > 0} ||\boldsymbol{u}(\boldsymbol{\beta})|| > 0 = ||\boldsymbol{u}(\boldsymbol{\beta}^\star)||. \tag{A.11}$$

First, (A.4) implies $||\boldsymbol{u}(\boldsymbol{\beta}^\star)|| = 0$. Furthermore, $\boldsymbol{u}(\boldsymbol{\beta})$ is continuous, and its Jacobian matrix is negative definite, assuming no separation or quasi-separation of the data (Albert and Anderson, 1984; Wedderburn, 1976). Therefore, it has a unique zero (which is $\boldsymbol{\beta}^\star$), and condition (A.11) is fulfilled. Because of van der Vaart (1998), (A.5) and (A.11) imply that $\hat{\boldsymbol{\beta}}$ is consistent. □

A.2. Proof of Theorem 2: asymptotic normality of $\hat{\boldsymbol{\beta}}$.

PROOF. We start with a mean value theorem for each of the components of $\boldsymbol{U}(\boldsymbol{\beta})$:

$$0 = U_r(\hat{\boldsymbol{\beta}}) = U_r(\boldsymbol{\beta}^\star) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^T \frac{\partial}{\partial \boldsymbol{\beta}} U_r(\tilde{\boldsymbol{\beta}}_r) \tag{A.12}$$

for $r = 1, ..., p+q+1$, where each $\tilde{\boldsymbol{\beta}}_r$ is a point on the line between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^\star$. The square matrix of dimension $p+q+1$ $\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{U}(\boldsymbol{\beta})$ equals

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{U}(\boldsymbol{\beta}) = -\frac{1}{n} \Bigg[& \sum_{j=1}^{J_1} n_{j1} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix}^T [1 - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(1)})] p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(1)}) \\ & + \sum_{j=1}^{J_2} n_{j2} \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix}^T [1 - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)})] p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)}) \Bigg].\end{aligned}$$

Since under no separation or quasi-separation of the data (Albert and Anderson, 1984; Wedderburn, 1976), the logistic regression likelihood is strictly log-concave in $\boldsymbol{\beta}$, $\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{U}(\boldsymbol{\beta})$ is invertible. Furthermore, because of $\boldsymbol{A}_j^{(2,n_1)} \xrightarrow{P} \boldsymbol{a}_j^{(2)}$ and because the baseline covariates $z_j^{(1)}$ and $z_j^{(2)}$ are finite, we have that for all $\boldsymbol{\beta} \in \mathcal{B}$,

$$\begin{aligned} -\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{U}(\boldsymbol{\beta}) \xrightarrow{P} & \sum_{j=1}^{J_1} \alpha_{j1} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix}^T [1 - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(1)})] p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(1)}) \\ & + \sum_{j=1}^{J_2} \alpha_{j2} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(2)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(2)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix}^T [1 - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)})] p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}; \boldsymbol{z}_j^{(2)}) \\ & := I(\boldsymbol{\beta}), \end{aligned} \tag{A.13}$$

by Lebesgue's Dominated Convergence Theorem. Since $\tilde{\boldsymbol{\beta}}_r$ is between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^\star$ for all r, $\hat{\boldsymbol{\beta}}_r$ is consistent for each r. Since $I(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$ and uniformly bounded in $\boldsymbol{\beta} \in \mathcal{B}$, equations (A.12) and (A.13) imply that the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)$ is the same as the asymptotic distribution of f (2.5)

Regarding the part of (2.5) that does not involve $I(\boldsymbol{\beta}^\star)$, we will show that its asymptotic distribution is multivariate normal. We present a coupling argument (Lindvall, 2002) to deal with the fact the two summands are not independent. For each $j = 1, ..., J_2$, let $Y_{ij}^{(2)}$ be iid Bernoulli random variables

with success probability $p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})$. We construct variables $\tilde{Y}_{ij}^{(2,n_1)}$ which, given the stage 1 data and $\boldsymbol{A}_j^{(2,n_1)}$, have the same distribution as the original $Y_{ij}^{(2,n_1)}$, but coupled (see e.g. Lindvall (2002)) with the $Y_{ij}^{(2)}$ in the following way. Let W_{ij} be a uniform $(0,1)$ random variable independent of all other variables introduced so-far. For the case $p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) > p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})$, $\tilde{Y}_{ij}^{(2,n_1)}$ is defined by (2.6). For the case $p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \leq p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})$,

(A.14)
$$\tilde{Y}_{ij}^{(2,n_1)} = \begin{cases} 1 & \text{if } Y_{ij}^{(2)} = 1 \\ 1 & \text{if } Y_{ij}^{(2)} = 0 \text{ and } W_{ij} < \frac{p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})}{1 - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})} \\ 0 & \text{if } Y_{ij}^{(2)} = 0 \text{ and } W_{ij} \geq \frac{p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})}{1 - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})}. \end{cases}$$

The key ingredient of the coupling argument is that given $\boldsymbol{A}_j^{(2,n_1)}$ and all stage 1 data, the distribution of the $\tilde{Y}_{ij}^{(2,n_1)}$ is identical to the distribution of the $Y_{ij}^{(2,n_1)}$. Therefore, when replacing $Y_{ij}^{(2,n_1)}$ with $\tilde{Y}_{ij}^{(2,n_1)}$ in (2.5), the distribution of (2.5) is unaffected: the term of (2.5) that does not involve $I(\boldsymbol{\beta}^\star)$ has the same distribution as

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{J_1} \sum_{i=1}^{n_{j1}} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \left(Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(1)}) \right)$$
$$+ \frac{1}{\sqrt{n}} \sum_{j=1}^{J_2} \sum_{i=1}^{n_{j2}} \begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \left(\tilde{Y}_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right).$$

This equals

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{J_1} \sum_{i=1}^{n_{j1}} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(1)} \\ \boldsymbol{z}_j^{(1)} \end{pmatrix} \left(Y_{ij}^{(1)} - p_{\boldsymbol{a}_j^{(1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right)$$
$$+ \frac{1}{\sqrt{n}} \sum_{j=1}^{J_2} \sum_{i=1}^{n_{j2}} \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(2)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \left(Y_{ij}^{(2)} - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right) + \boldsymbol{D}_n,$$

where

$$\boldsymbol{D}_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{J_2} \sum_{i=1}^{n_{j2}} \left[\begin{pmatrix} 1 \\ \boldsymbol{A}_j^{(2,n_1)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \left(\tilde{Y}_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right) \right.$$
$$\left. - \begin{pmatrix} 1 \\ \boldsymbol{a}_j^{(2)} \\ \boldsymbol{z}_j^{(2)} \end{pmatrix} \left(Y_{ij}^{(2)} - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right) \right].$$

We will show that $\boldsymbol{D}_n \xrightarrow{P} 0$, using the fact that the $Y_{ij}^{(2)}$ and $\tilde{Y}_{ij}^{(2,n_1)}$ are coupled.

Conditionally on $\boldsymbol{A}_j^{(2,n_1)}$ for the respective terms the expectation of the first term of $\boldsymbol{D}_n$ is zero, and conditioning on $\boldsymbol{a}_j^{(2)}$ for the respective terms implies the expectation of the second term is also zero. Therefore, $E(\boldsymbol{D}_n) = 0$. We will show that the expectation of the square of each entry in the vector $\boldsymbol{D}_n$ converges to 0, so that Chebyshev's inequality implies that $\boldsymbol{D}_n \xrightarrow{P} 0$. We concentrate on the component of the vector that is led by $\boldsymbol{A}_j^{(2,n_1)}$, as the proof for the other terms is similar, yet simpler.

The expectation of the square of each of the m–th components of

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{J_2} \sum_{i=1}^{n_{j2}} \left[\boldsymbol{A}_j^{(2,n_1)} \left(\tilde{Y}_{ij}^{(2,n_1)} - p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right) - \boldsymbol{a}_j^{(2)} \left(Y_{ij}^{(2)} - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \right) \right]$$

equals to

$$\frac{1}{n}E\left[\sum_{j=1}^{J_2}\sum_{i=1}^{n_{j2}}A_{jm}^{(2,n_1)}\left(\tilde{Y}_{ij}^{(2,n_1)}-p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)-a_{jm}^{(2)}\left(Y_{ij}^{(2)}-p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)\right]^2$$

(A.15)

$$=\frac{1}{n}\sum_{j=1}^{J_2}\sum_{i=1}^{n_{j2}}E\left[A_{jm}^{(2,n_1)}\left(\tilde{Y}_{ij}^{(2,n_1)}-p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)-a_{jm}^{(2)}\left(Y_{ij}^{(2)}-p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)\right]^2$$

$$=\sum_{j=1}^{J_2}\frac{n_{j2}}{n}E\Big[\left(A_{jm}^{(2,n_1)}\tilde{Y}_{ij}^{(2,n_1)}-a_{jm}^{(2)}Y_{ij}^{(2)}\right)$$

$$-\left(A_{jm}^{(2,n_1)}p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})-a_{jm}^{(2)}p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)\Big]^2$$

$$=\sum_{j=1}^{J_2}\frac{n_{j2}}{n}E\Big[a_{jm}^{(2)}\left((\tilde{Y}_{ij}^{(2,n_1)}-Y_{ij}^{(2)})-(p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})-p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)$$

$$+\left((A_{jm}^{(2,n_1)}-a_{jm}^{(2)})(\tilde{Y}_{ij}^{(2,n_1)}-p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)}))\right)\Big]^2$$

$$=\sum_{j=1}^{J_2}\frac{n_{j2}}{n}\Bigg\{E\left[a_{jm}^{(2)}\left((\tilde{Y}_{ij}^{(2,n_1)}-Y_{ij}^{(2)})-(p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})-p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)\right]^2$$

(A.16)

$$+E\left[\left((A_{jm}^{(2,n_1)}-a_{jm}^{(2)})(\tilde{Y}_{ij}^{(2,n_1)}-p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)}))\right)\right]^2$$

(A.17)

$$+2E\Big[a_{jm}^{(2)}\left((\tilde{Y}_{ij}^{(2,n_1)}-Y_{ij}^{(2)})-(p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})-p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})\right)$$

$$\cdot\left((A_{jm}^{(2,n_1)}-a_{jm}^{(2)})(\tilde{Y}_{ij}^{(2,n_1)}-p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)}))\right)\Big]\Bigg\}$$

$$=\sum_{j=1}^{J_2}\frac{n_{j2}}{n}E\left[a_{jm}^{(2)}\left((\tilde{Y}_{ij}^{(2,n_1)}-Y_{ij}^{(2)})-(p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})-p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)}))\right)\right]^2+o(1)$$

$$=\sum_{j=1}^{J_2}\frac{n_{j2}}{n}(a_{jm}^{(2)})^2E\left\{E\left[\left((\tilde{Y}_{ij}^{(2,n_1)}-Y_{ij}^{(2)})-(p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)})-p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star;\boldsymbol{z}_j^{(2)}))\right)^2\Big|\boldsymbol{A}_j^{(2,n_1)}\right]\right\}$$

$$+o(1)$$

(A.18)

$$= \sum_{j=1}^{J_2} \frac{n_{j2}}{n} (a_{jm}^{(2)})^2 E\left\{ Var\left[(\tilde{Y}_{ij}^{(2,n_1)} - Y_{ij}^{(2)}) - (p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})) \middle| \boldsymbol{A}_j^{(2,n_1)} \right] \right\}$$
$$+ o(1)$$
(A.19)
$$= \sum_{j=1}^{J_2} \frac{n_{j2}}{n} (a_{jm}^{(2)})^2 E\left[\left| p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{a}_{j'}^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_{j'}^{(2)}) \right| \left(1 - \left| p_{\boldsymbol{A}_{j'}^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_{j'}^{(2)}) - p_{\boldsymbol{a}_{j'}^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_{j'}^{(2)}) \right| \right) \right] + o(1)$$
$$\to 0.$$

In (A.15), all terms with $j' \neq j$ and $i' \neq i$ vanish by conditioning on $(\boldsymbol{A}_j^{(2,n_1)}, \boldsymbol{A}_{j'}^{(2,n_1)})$, for all $j, j' = 1, ..., J_2$ ($\boldsymbol{a}_j^{(2)}$ are constants). Because $\boldsymbol{A}_j^{(2,n_1)} \xrightarrow{P} \boldsymbol{a}_j^{(2)}$ and $\boldsymbol{A}_j^{(2,n_1)}$ has bounded support, the expectations (A.16) and (A.17) are $o(1)$ by Lebesgue's Dominated Convergence Theorem. In both expressions, (A.16) and (A.17), all the components are bounded and $(A_{jm}^{(2,n_1)} - a_{jm}^{(2)}) \xrightarrow{P} 0$, both (A.16) and (A.17) are $o(1)$. In (A.19), we utilize the coupling: conditionally on $\boldsymbol{A}_j^{(2,n_1)}$ and $\boldsymbol{a}_j^{(2)}$, $(\tilde{Y}_{ij}^{(2,n_1)} - Y_{ij}^{(2)})$ is a plus or minus a Bernoulli random variable with corresponding probability $|p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) - p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})|$. By $\boldsymbol{A}_j^{(2,n_1)} \xrightarrow{P} \boldsymbol{a}_j^{(2)}, p_{\boldsymbol{A}_j^{(2,n_1)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)}) \xrightarrow{P} p_{\boldsymbol{a}_j^{(2)}}(\boldsymbol{\beta}^\star; \boldsymbol{z}_j^{(2)})$, so that Lebesgue's Dominated Convergence Theorem implies that the expectation converges to zero. Because $((a_{jm}^{(2)})^2$ is bounded and n_{j2}/n is bounded by 1, then $\boldsymbol{D}_n \xrightarrow{P} 0$.

We conclude that the asymptotic distribution of the term of (2.5) that does not involve $I(\boldsymbol{\beta}^\star)$ has the same asymptotic distribution as (2.7). The asymptotic normal distribution of (2.7) follows from standard theory about logistic regression because the $\boldsymbol{a}_j^{(1)}$ and $\boldsymbol{a}_{j'}^{(2)}$ are fixed for all j, j', so the outcomes are independent. Standard theory also implies that the asymptotic variance of (2.7) is equal to $I(\boldsymbol{\beta}^\star)$. Combining with (2.5), we conclude that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{\mathcal{D}} N\left(0, I^{-1}(\boldsymbol{\beta}^\star)\right).$$

The variance can be consistently estimated from the data by replacing $\boldsymbol{a}_j^{(2)}$, $\boldsymbol{\beta}^\star$, α_{j1} and α_{j2} with $\boldsymbol{A}_j^{(2,n_1)}$, $\hat{\boldsymbol{\beta}}$, n_{j1}/n and n_{j2}/n, respectively, in $I(\boldsymbol{\beta}^\star)$. This asymptotic variance is the same as the asymptotic variance that one would obtain if the interventions were fixed in advance (and thus $\boldsymbol{Y}_j^{(1)}$ and $\boldsymbol{Y}_{j'}^{(2,n_1)}$ were independent (for all j, j')). □

ACKNOWLEDGMENTS

The authors thank Dr. Katherine Semrau for her assistance with sharing and interpreting the results from the BetterBirth Study. The authors also thank the Editor, Associate Editor and two anonymous reviewers for helpful comments that improved the paper.

SUPPLEMENTARY MATERIAL

The supplementary material includes additional proofs, extension of the results to general number of stages, and results of further simulation studies. ().

REFERENCES

Albert, A. and Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10.

Bauer, P. and Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* 1029–1041.

Bauer, P., Bretz, F., Dragalin, V., König, F. and Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in medicine* **35** 325–347.

Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97** 236–244.

Collins, L. M., Murphy, S. A. and Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *American journal of preventive medicine* **32** S112–S118.

Collins, L. M., Nahum-Shani, I. and Almirall, D. (2014). Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (SMART). *Clinical Trials* **11** 426–434.

Cox, D. R. (1958). Planning of experiments.

Escoffery, C., Lebow-Skelley, E., Udelson, H., Böing, E. A., Wood, R., Fernandez, M. E. and Mullen, P. D. (2018). A scoping study of frameworks for adapting public health evidence-based interventions. *Translational behavioral medicine*.

FDA (2016). *Adaptive Designs for Medical Device Clinical Studies: Guidance for Industry and Foodand Drug Administration Staff*.

Gao, P., Liu, L. and Mehta, C. (2013). Exact inference for adaptive group sequential designs. *Statistics in medicine* **32** 3991–4005.

Gawande, A. (2014). *Being mortal: medicine and what matters in the end*. Metropolitan Books.

Hernan, M. A. and Robins, J. M. (2019). *Causal inference*. CRC Boca Raton, Chapman & Hall/CRC, forthcoming.

Hirschhorn, L. R., Semrau, K., Kodkany, B., Churchill, R., Kapoor, A., Spector, J., Ringer, S., Firestone, R., Kumar, V. and Gawande, A. (2015). Learning before leaping: integration of an adaptive study design process prior to initiation of BetterBirth, a large-scale randomized controlled trial in Uttar Pradesh, India. *Implementation Science* **10** 1.

Hu, F. and Rosenberger, W. F. (2003). Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association* **98** 671–678.

Iwuji, C. C., Orne-Gliemann, J., Larmarange, J., Balestre, E., Thiebaut, R., Tanser, F., Okesola, N., Makowa, T., Dreyer, J., Herbst, K. et al. (2017). Universal test and treat and the HIV epidemic in rural South Africa: a phase 4, open-label, community cluster randomised trial. *The Lancet HIV*.

Kairalla, J. A., Coffey, C. S., Thomann, M. A. and Muller, K. E. (2012). Adaptive trial designs: a review of barriers and opportunities. *Trials* **13** 145.

Lindvall, T. (2002). *Lectures on the coupling method*. Courier Corporation.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57** 886–891.

Müller, H.-H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in medicine* **23** 2497–2508.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine* **24** 1455–1481.

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R. and TenHave, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and alcohol dependence* **88** S24–S30.

O'Quigley, J., Pepe, M. and Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* 33–48.

O'Quigley, J. and Shen, L. Z. (1996). Continual reassessment method: a likelihood approach. *Biometrics* 673–684.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* 1315–1324.

Rosenberger, W. F., Flournoy, N. and Durham, S. D. (1997). Asymptotic normality of maximum likelihood estimators from multiparameter response-driven designs. *Journal of Statistical Planning and Inference* **60** 69–76.

Rosenberger, W. F. and Haines, L. M. (2002). Competing designs for phase I clinical trials: a review. *Statistics in medicine* **21** 2757–2770.

Rosenblum, M. and van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika* **98** 845–860.

Scheffé, H. (1959). *The analysis of variance* **72**. John Wiley & Sons.

Semrau, K. E., Hirschhorn, L. R., Marx Delaney, M., Singh, V. P., Saurastri, R., Sharma, N., Tuller, D. E., Firestone, R., Lipsitz, S., Dhingra-Kumar, N. et al. (2017). Outcomes of a Coaching-Based WHO Safe Childbirth Checklist Program in India. *New England Journal of Medicine* **377** 2313–2324.

Simon, R., Rubinstein, L., Arbuck, S. G., Christian, M. C., Freidlin, B. and Collins, J. (1997). Accelerated titration designs for phase I clinical trials in oncology. *Journal of the National Cancer Institute* **89** 1138–1147.

Snapinn, S., Chen, M.-G., Jiang, Q. and Koutsoukos, T. (2006). Assessment of futility in clinical trials. *Pharmaceutical Statistics* **5** 273–281.

Spiegelman, D. and Zhou, X. (2018). Evaluating Public Health Interventions: 8. Causal Inference for Time-Invariant Interventions. *American journal of public health* **108** 1187–1190.

Thall, P. F., Millikan, R. E., Mueller, P. and Lee, S.-J. (2003). Dose-Finding with Two Agents in Phase I Oncology Trials. *Biometrics* **59** 487–496.

van der Laan, M. J. (2008). The Construction and Analysis of Adaptive Group Sequential Designs Technical Report, Technical Report 232, Division of Biostatistics, UC

Berkeley.

VAN DER VAART, A. W. (1998). *Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press, Cambridge.

WANG, K. and IVANOVA, A. (2005). Two-Dimensional Dose Finding in Discrete Dose Space. *Biometrics* **61** 217–222.

WEDDERBURN, R. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* 27–32.

WHO (2012). *WHO recommendations for the prevention and treatment of postpartum haemorrhage*. World Health Organization.

WONG, K. M., CAPASSO, A. and ECKHARDT, S. G. (2016). The changing landscape of phase I trials in oncology. *Nature Reviews Clinical Oncology* **13** 106–117.