Full Matching Approach to Instrumental Variables Estimation with Application to the Effect of Malaria on Stunting

Hyunseung Kang

University of Pennsylvania, Philadelphia, USA.

Benno Kreuels

Research Group Infectious Disease Epidemiology, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany

Division of Tropical Medicine, I. Department of Internal Medicine, University Medical Centre, Hamburg Eppendorf, Germany

Jürgen May

Research Group Infectious Disease Epidemiology, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany

and Dylan S. Small

University of Pennsylvania, Philadelphia, USA.

Summary. Most previous studies of the causal relationship between malaria and stunting have been limited to studies where potential confounders are controlled via regression-based methods, but may have been biased by unobserved

†Address for correspondence: Hyunseung Kang, Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA. E-mail: khyuns@wharton.upenn.edu.
confounders. Instrumental variables (IV) regression offers a way to control for unmeasured confounders. With malaria and stunting, we consider the use of the sickle cell trait as an instrumental variable. For the sickle cell trait to be a valid instrument, it may still be important to account for measured confounders. The most commonly used instrumental variable regression method, two-stage least squares, relies on parametric assumptions on the measured confounders to account for them. Additionally, two-stage least squares does not make transparent whether there is covariate balance or how different subjects are weighed in the analysis and does not blind the researcher to the outcome data. To address these drawbacks, we propose an alternative method for IV estimation based on full matching. We also discuss efficiency of various full matching schemes when they are used in IV estimation. We evaluate our procedure on simulated data and apply our method on the causal effect of malaria on stunting among children. We estimate that the risk of stunting increases by 0.22 per malaria episode (p-value: 0.011, 95% CI: 0.044, 1).

**Keywords:** Full matching; Instrumental variables; Malaria; Stunting; Two-stage least squares

1. **Introduction**

1.1. *Does malaria cause stunting?*

Malaria is a parasitic disease that is common in Africa and other tropical regions; there were an estimated 174 million cases of malaria in sub-Saharan Africa in 2010, with most cases occurring in children under the age of 5 years (World Health Organization, 2012). Consequently, there is great public health interest in understanding the causal effect of malaria on childhood development. One such indicator of childhood development is stunting, defined as a child’s height being two standard deviations below the mean for his/her age (Mercedes, 2006).
The focus of this paper is to examine the causal effect of malaria on stunting using a new statistical method, instrumental variables with full matching, on a recently collected data set from Ghana, Africa.

Previous studies analyzed the causal relationship between malaria and stunting by collecting characteristics of children that might affect malaria and stunting and “controlling” for them (McGregor et al., 1956; Bradley-Moore et al., 1985; Snow et al., 1991; Genton et al., 1998; Deen et al., 2002; ter Kuile et al., 2003; Nyakeriga et al., 2004; Ehrhardt et al., 2006; Fillol et al., 2009; Deribew et al., 2010; Crookston et al., 2010; Arinaitwe et al., 2012). For example, Crookston et al. (2010) controlled for sex, age, and other physical measurements. Deribew et al. (2010) not only controlled for age and sex, but also included birth order and family income as covariates. Fillol et al. (2009) controlled for sex, age, and place of residence. Nyakeriga et al. (2004) controlled for age, sickle cell group, season, and ethnic group. ter Kuile et al. (2003) controlled for educational status, rainfall, sickle cell trait, and age. Deen et al. (2002) controlled for age, sex and ethnicity. Genton et al. (1998) controlled age and bed nets.

However, there is always the possibility that some important confounders were not controlled for and that unmeasured confounders are present. For example, Fillol et al. (2009) and Deribew et al. (2010) stated that a limitation in their studies was not controlling for diet, specifically a child’s intake of micronutrients such as vitamins, zinc, or iron as these micronutrients could impact a child’s growth as well as his immune system’s ability to fight off a malaria episode. In addition, Ehrhardt et al. (2006) and Crookston et al. (2010) suggested controlling for socioeconomic factors in future studies of malaria and malnutrition because affluent families are more likely to provide mosquito nets and nutritious food to their children compared to impoverished families. In summary, unmeasured confounders are likely present in all the aforementioned
studies since it’s impractical to measure all possible confounders.

1.2. Instrumental variables with two stage least squares

Instrumental variables (IVs) is an alternative method to estimate the causal effect of an exposure on the outcome when there are measured and unmeasured confounders (Angrist et al., 1996; Brookhart and Schneeweiss, 2007; Cheng et al., 2009; Swanson and Hernán, 2013; Baiocchi et al., 2014). In a nutshell, an instrument is a variable that (A1) is associated with the exposure, (A2) has no direct pathways to the outcome, and (A3) is not associated with any unmeasured confounders after controlling for the unmeasured confounders (See Figure 1 and Section 2.3).

Recently, genotypic variation has been successfully used as an instrument in an approach called Mendelian randomization (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008). Specifically, the idea is to find genotypic variations in the form of single nucleotide polymorphisms (SNPs) that are robustly associated with the exposure, usually from genome wide association studies (GWAS) that have been replicated from an independent sample. For our study of analyzing the causal effect of malaria on stunting, the presence of the sickle cell trait (HbAS) versus the normal hemoglobin type (HbAA) is plausibly a valid instrument (Kang et al., 2013). In a nutshell, the sickle cell trait (HbAS) is a condition where a person inherits from one parent a mutated copy of the hemoglobin, beta (HBB) gene called the sickle cell gene mutation that bends red blood cells into a sickle (crescent) shape, but inherits a normal copy of the HBB gene from the other parent. Section 2.3 discusses the sickle cell trait in detail along with its plausibility as an IV.

With regards to plausibility, a potential instrumental variable may only be independent of the unmeasured confounders after conditioning on measured co-
variates (i.e. adhere to (A3)), in which case it is important to control for the measured covariates. Table 1 lists the measured covariates that were available for our analysis. The most popular and well-studied among methods that use an IV to estimate causal effects is two-stage least squares (2SLS). 2SLS first estimates, via least squares, the predicted exposure given the instrument and measured covariates and second, regresses the outcome on this predicted exposure and the measured covariates; the 2SLS estimate of the causal effect is the coefficient on the predicted exposure in the second regression. Standard results in econometrics show 2SLS estimators are consistent and efficient under linear single-variable structural equation models with a constant treatment effect (Davidson and MacKinnon, 1993; Wooldridge, 2010). Similarly, Angrist and Imbens (1995) showed that under assumptions based on potential outcomes and where the constant treatment effect assumption is replaced by assumptions on the relationship between the exposure and the instrument such as monotonicity, 2SLS converges to a weighted average of per-unit causal response to a change in treatment intensity.

Despite its attractive estimation properties, 2SLS has some drawbacks in (i) lack of transparency of the population to which the estimate applies, (ii) lack of blinding of the analyst/researcher and (iii) dependence on parametric assumptions. First, with regards to transparency, if a saturated model for the covariates is used (i.e., a dummy variable for each possible value of the vector of the covariates) and there is a constant treatment effect for each value of the covariates, then 2SLS converges to a weighted average of the covariate-specific treatment effects with the weights proportional to the average conditional variance of the expected value of the treatment given the covariates and the instrument (Angrist and Imbens, 1995, Theorem 3). Suppose that there are some values of the covariates for which the instrument is almost always low, some values for
which the instrument is almost always high and some values of the covariates for which the instrument takes on both low and high values. Then, the 2SLS estimate will put most of its weight on the causal effect for subjects with the values of the covariates for which the instrument takes on both low and high values, and little weight on subjects with the values of the covariates for which the instrument usually takes on low (or high) values. In our malaria study, this would mean that there might be some villages, a covariate, that are receiving little weight in the 2SLS estimate and the estimate might not be helpful for understanding the effect of malaria on stunting in these villages even though these villages might have contributed many subjects to the analysis. Although the weighting function in 2SLS can be studied, there is nothing in the 2SLS estimation procedure itself that warns us when some values of the covariates are receiving little weight and it is rare to see discussion of the weighting function for 2SLS in empirical papers.

Second, 2SLS lacks blinding with respect to the outcome data when adjusting for covariates. Cochran (1965), Rubin (2007) and Rosenbaum (2010) argue that the best observational studies resemble simple randomized experiments. An important feature of the design of randomized experiments is that when designing the study and the planing the analysis, the researcher is blinded to the outcome data. However, in regression based procedures for adjusting for covariates like 2SLS, there is often judgment that needs to be exercised in choosing covariate adjustment models, which require one to look at the outcome data and estimates of causal effects to exercise such judgment. It is difficult even for the most honest researcher to be completely objective in comparing models when the researcher has an a priori hypothesis or expectation about the direction of the causal effect from weighing various covariate adjustment models (Rubin and Waterman, 2006).
Third, 2SLS relies on proper specification of how the measured covariates affect the outcomes. Often, parametric modeling assumptions are made for how the measured confounders affect the outcome. In particular, 2SLS, as usually implemented, relies on the measured confounders having a linear effect on the expected outcome. Section 3 contains simulation evidence of 2SLS to demonstrate its reliance on linear, parametric assumptions.

1.3. Instrumental variables with full matching

Matching is an alternative method to adjust for measured covariates. A matching algorithm groups individuals in the data with different values of the instrument but similar values of the observed covariates, so that within each group, the only difference between the individuals is their values of the instrument (Haviland et al., 2007; Rosenbaum, 2010; Stuart, 2010). For example, in the malaria data, a matching algorithm seeks to produce matched sets so that in a matched set, individuals are born in the same village and similar on other measured covariates. The only difference between individuals in a matched set is their instrument values. We can then compare stunting between individuals with high and low values of the instrument within a matched set to assess the effect of malaria on stunting (Baiocchi et al., 2010).

Matching addresses the drawbacks of 2SLS discussed in the previous section as follows. First, if there are values of covariates for which almost all subjects have a high (or low) value of the IV, then the matching algorithm and associated diagnostics will tell us that matched sets cannot be formed when subjects in the matched sets have certain values of the covariates but different levels of the IV; thus, it will be transparent that for these values of the covariates, the causal effect cannot be estimated without extrapolation. Relatedly, matching allows us to control the weighting of subjects with different values of the covariates to
make the weighting transparent, such as weighting the covariates in proportion to their population frequency. Second, matching is blind to the outcome data; a matching algorithm only requires the measured covariates and the instrument values for each individual in the data. Diagnostics can be done and the matching can be adjusted until it is adequate, all without looking at the outcome data. Finally, when estimating the causal effect, matching does not use any assumptions about the outcome-covariate model, especially linearity and parametric assumptions on the model.

Previous work using matching in studying causality is abundant in non-IV settings; see Stuart (2010) for a complete overview. In contrast, work on using matching methods on IV estimation is limited to pair matching (Baiocchi et al., 2010) and fixed control matching, i.e. each unit with level 1 of the IV is matched to a fixed number of units with level 0 of the IV (Kang et al., 2013). A drawback to these methods is that these matching algorithms do not use the full data (Keele and Morgan, 2013; Zubizarreta et al., 2013). In particular, Kang et al. (2013) studied the same causal effect of interest, malaria on stunting, but with a smaller amount of data, because the statistical methodology was limited to matching with fixed controls. That is, out of the total of 884 individuals available, the matching algorithm dropped 25% of the individuals and the final statistical inference was based only on 660 individuals. In addition, the authors did not analyze statistical efficiency of the resulting causal estimate.

In this paper, we extend these two methods and propose an IV estimation based on full matching that uses the full data. Full matching is the most general, flexible, and optimal type of matching (Rosenbaum, 1991; Hansen, 2004; Rosenbaum, 2010). Specifically, full matching is the generalization of any type of matching, such as pair matching, matching with fixed controls, or matching with variable controls. Full matching is also flexible in that it can incorporate
constraints on matched set structures, such as limiting the number of individuals in each matched set, to improve statistical efficiency. Finally, full matching is optimal in the sense that it produces matched sets where within each set, measured covariates between individuals with different instrument values are most similar (Rosenbaum, 1991).

Under IV estimation with full matching, we derive a randomization-based testing procedure and sensitivity analysis based on the proposed test statistic. We also discuss efficiency of matching-based method that is used for IV estimation. In addition, we conduct simulation studies to study the performance of 2SLS versus full matching IV estimation, specifically analyzing the robustness of both methods to non-linearity (Section 3). Finally, we apply full matching IV estimation to analyze the causal effect of malaria on stunting and demonstrate the full matching method’s transparency in adjusting for covariates.

2. Methods

2.1. Notation

To introduce the idea of matching in IV estimation, we introduce the following notation. Let $i = 1, \ldots, I$ index the $I$ total matched sets that individuals are matched into. Each matched set $i$ contains $n_i \geq 2$ subjects who are indexed by $j = 1, \ldots, n_i$ and there are a total of $N = \sum_{i=1}^{I} n_i$ individuals in the data. Let $Z_{ij}$ denote a binary instrument for subject $j$ in matched set $i$. In each matched set $i$, there are $m_i$ subjects with $Z_{ij} = 1$ and $n_i - m_i$ subjects with $Z_{ij} = 0$. For instance, in the malaria data, for each $i$th matched set, there are $m_i$ children who inherited the sickle cell trait, HbAS (i.e. $Z_{ij} = 1$), and $n_i - m_i$ children who inherited HbAA (i.e. $Z_{ij} = 0$). Let $Z$ be a random variable that consists of the collection of $Z_{ij}$’s, $Z = (Z_{11}, Z_{12}, \ldots, Z_{I,n_i})$. Define $\Omega$ be the set that contains all possible values $z$ of $Z$, so $z \in \Omega$ if $z_{ij}$ is binary and $\sum_{j=1}^{n_i} z_{ij} = m_i$ for all $I$.
matched sets. Thus, the cardinality of $\Omega$, denoted as $|\Omega|$, is $|\Omega| = \prod_{i=1}^{I} \binom{n_i}{m_i}$.

Denote $\mathcal{Z}$ to be the event that $\mathcal{Z} \in \Omega$.

For individual $j$ in matched set $i$, define $r_{1ij}$ and $r_{0ij}$ to be the potential outcomes if the individual had instrument value $Z_{ij} = 1$ and $Z_{ij} = 0$, respectively. In the malaria data, $r_{1ij}$ is a binary variable that represents whether the $j$th child in the $i$th matched set would be stunted (i.e. 1) or not (i.e. 0) if the child carried the sickle cell trait (i.e. if $Z_{ij} = 1$). Similarly, $r_{0ij}$ is a binary variable that represents whether the child would be stunted or not if the child carried no sickle cell trait (i.e. if $Z_{ij} = 0$). Also, define $d_{1ij}$ and $d_{0ij}$ to be the potential exposure values under $Z_{ij} = 1$ or $Z_{ij} = 0$, respectively. Again, with the malaria data, $d_{1ij}$ and $d_{0ij}$ represent the number of malaria episodes the child would have if she had the sickle cell trait, $Z_{ij} = 1$, and no sickle cell trait, $Z_{ij} = 0$, respectively.

For individual $j$ in matched set $i$, let $R_{ij}$ be the binary observed outcome and $D_{ij}$ be the observed exposure. The potential outcomes $r_{1ij}, r_{0ij}, d_{1ij},$ and $d_{0ij}$ and the observed values $R_{ij}, D_{ij},$ and $Z_{ij}$ are related as follows:

$$R_{ij} = r_{1ij}Z_{ij} + r_{0ij}(1 - Z_{ij}) \quad D_{ij} = d_{1ij}Z_{ij} + d_{0ij}(1 - Z_{ij})$$

Equation (1) quantifies the fact that we only get to observe one value of the potential outcome, depending on the value of $Z_{ij}$. In the case of the malaria data, if child $j$ in matched set $i$ carries the sickle cell trait (HbAS), we observe $Z_{ij} = 1$ and we only get to see $r_{1ij}$ and $d_{1ij}$ where by equation (1), $R_{ij} = r_{1ij}$ and $D_{ij} = d_{1ij}$. On the other hand, if the child doesn’t carry the sickle cell trait, we observe $Z_{ij} = 0, r_{0ij},$ and $d_{0ij}$ where $R_{ij} = r_{0ij}$ and $D_{ij} = d_{0ij}$.

For individual $j$ in matched set $i$, let $X_{ij}$ be a vector of observed covariates and $u_{ij}$ be the unobserved covariates. For example, in the malaria data, $X_{ij}$ represents each child’s covariates listed in Table 1 while $u_{ij}$ is an unmeasured
confounder, like diet, which was mentioned in Section 1.1. We define the set
\[ \mathcal{F} = \{(r_{1ij}, r_{0ij}, d_{1ij}, d_{0ij}, X_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, \ldots, n_i \} \] to be the collection of potential outcomes and all covariates/confounders, observed and unobserved.

2.2. Full matching

A matching algorithm controls the bias resulting from different observed covariates by creating \( I \) matched sets indexed by \( i, i = 1, \ldots, I \) such that individuals within each matched set have similar covariate values \( x_{ij} \). Each matched set has \( n_i \) individuals, indexed by \( j = 1, \ldots, n_i \), of which \( m_i \) individuals have \( Z_{ij} = 1 \) and \( n_i - m_i \) individuals have \( Z_{ij} = 0 \). Thus, the only difference between individuals in each matched set is their instrument values, \( Z_{ij} \), since covariate values are similar for each individual within each set.

Full matching is a type of matching algorithm where each matched set either contains \( m_i = 1 \) individual with \( Z_{ij} = 1 \) and \( n_i - 1 \) individuals with \( Z_{ij} = 0 \) or \( m_i = n_i - 1 \) individuals with \( Z_{ij} = 1 \) and 1 individual with \( Z_{ij} = 0 \). As stated in Section 1.3, full matching is general, flexible, and optimal. In particular, full matching is general in the sense that all other types of matching, such as pair matching and matching with variable controls, is a special case of full matching. Second, full matching is flexible in the sense that we can vary the size of the matched set to improve statistical efficiency; this is discussed in great detail in Section 2.7. Finally, full matching is optimal compared to other matching schemes in the sense that covariate differences between \( Z_{ij} = 1 \) and \( Z_{ij} = 0 \) within each matched set are the smallest when the sets are produced by a full matching algorithm. More precisely, for any matched set index \( i = 1, \ldots, I \), let \( \delta_i \) be the average distance between individuals with \( Z_{ij} = 1 \) and \( Z_{ij} = 0 \). Let \( \Delta \) be the weighted average of the distances across all the matched sets, weighed by the size of the matched sets, \( \Delta = \sum_{i=1}^{I} \delta_i / n_i \). The optimal full matching
(i.e. the full matching that minimizes $\Delta$) minimizes $\Delta$ among all stratifications of the data (Rosenbaum, 1991).

Rosenbaum (2002), Hansen (2004), Rosenbaum (2010), and Stuart (2010) provide an overview of matching and a discussion on various distance metrics and tools to measure similarity for observed and missing covariates. For the malaria data, Section 4.2 describes how we used propensity score caliper matching with rank-based Mahalanobis distance to measure covariate similarity. Once we obtained the distance matrix, we use an R package available on CRAN called optmatch developed by Hansen and Klopfer (2006) to find the optimal full matching.

2.3. Definition of instrument

We formalize the definition of an instrumental variable as follows (Didelez and Sheehan, 2007; Glymour et al., 2012) (see Figure 1).

(A1) The instrument must be associated with the exposure, or $E(d_{1ij} - d_{0ij}) \neq 0$ where the expectation is over the collected sample.

(A2) All directed pathways from the instrument to the outcome passes through the exposure, or if $d_{1ij} = d_{0ij}$, then $r_{1ij} = r_{0ij}$.

(A3) The instrument must be unassociated with any unmeasured variable that is associated with both outcome and exposure, conditional on the measured covariates, or $P(Z_{ij} = 1|F, Z) = m_i/n_i$ for each $i$.

Among the three assumptions, only assumption (A1) can be verified by data via running a regression between $D_{ij}$ and $Z_{ij}$. A strong association is preferred between $D_{ij}$ and $Z_{ij}$ because it provides lower-variance estimates of the causal effect and makes estimates less sensitive to bias (Wooldridge, 2010; Small and Rosenbaum, 2008). In contrast, assumptions (A2) and (A3) require knowledge
about \( \mathcal{F} \), specifically all the potential outcomes and the confounders. Assessment of (A2) and (A3) requires contextual knowledge.

For our analysis of malaria and stunting, we use the sickle cell trait as the instrument for the following reasons. For assumption (A1), the trait is mostly asymptomatic and does provide protection against malaria as compared to people with two normal copies of the HBB gene (HbAA) (Aidoo et al., 2002; Williams et al., 2005; May et al., 2007; Cholera et al., 2008; Kreuels et al., 2010). For assumption (A2), this could be violated if the sickle cell trait had effects on stunting other than through causing malaria, for instance, if the sickle cell trait was pleiotropic (Davey Smith and Ebrahim, 2003). We can partially test for this assumption by examining individuals who carry the sickle cell trait, but who grew up in a region where malaria is not present. That is, if assumption (A2) were violated, heights between individuals with HbAS and HbAA in such a region would be different since there would be a direct arrow between the sickle cell trait and height. Studies among African American children and children from the Dominican Republic and Jamaica for whom the sickle cell trait is common, but there is no malaria in the area, found no evidence that the sickle cell trait affected a child’s physical development (Ashcroft et al., 1976; Kramer et al., 1978; Ashcroft et al., 1978; Rehan, 1981). This supports the validity of assumption (A2). Note, however, that although the results of this test support the validity of (A2), (A2) could still be violated. For example, the sickle cell trait could have a direct effect that interacts with the environment in such a way that the direct effect is only present in Africa, but not in the United States, the Dominican Republic, or Jamaica.

For assumption (A3), this assumption would be questionable in our data if we did not control for any population stratification covariates. Population stratification is a condition where there are subpopulations, some of which are
more likely to have the sickle cell trait, and some of which are more likely to be stunted through mechanisms other than malaria (Davey Smith and Ebrahim, 2003). For example, in Table 1 which provides the baseline characteristics for our data, we observed that the village Tano-Odumasi had more children with HbAA than HbAS. It is possible that the there are other measured confounders besides HbAA that differ between the village Tano-Odumasi and other villages and affect stunting.

Hence, assumption (A3) is more plausible if we control for these measured confounders, like village of birth. Specifically, within the framework of full matching, for each matched set \( i \), if the observed confounders \( x_{ij} \) are similar among all \( n_i \) individuals, it may be more plausible that the unobserved confounders \( u_{ij} \) play no role in the distribution of \( Z_{ij} \) among the \( n_i \) children. If (A3) exactly holds and subjects are exactly matched for \( X_{ij} \), then within each matched set \( i \), \( Z_{ij} \) is simply a result of random assignment where \( Z_{ij} = 1 \) with probability \( m_i/n_i \) and \( Z_{ij} = 0 \) with probability \( (n_i - m_i)/n_i \). However, even after matching for observed confounders, it is entirely possible that unobserved confounders \( u_{ij} \) may still influence the assignment of \( Z_{ij} \) in each matched set \( i \) and assumption (A3) could be violated. We delay this discussion to Section 2.6.

There are also other assumptions associated with instrumental variables, mainly the Stable Unit Treatment Value Assumption (SUTVA) and monotonicity assumption in Angrist et al. (1996). SUTVA, within the framework of MR, states that one’s individual potential outcomes are not affected by the genotype assignment of another individual. This is fairly reasonable in MR and, more specifically with our instrument, the sickle cell genotype, since the instrument was determined at the conception of the child and a child’s genotype only affects his exposure and outcome, and not the exposures and outcomes of other children. Monotonicity, within the framework of MR, states that there are no
individuals who would have an adverse effect on the exposure from inheriting the genotype which is purported to bring positive effect on the exposure. In MR where the chosen genetic instruments usually bring about a positive effect on the exposure, monotonicity is reasonable. For example, with our instrument, the sickle cell genotype, it is widely believe that inheriting it provides individuals protection from malarial infection compared to not inheriting it and hence, the monotonicity assumption is reasonable for our instrument.

2.4. Effect ratio

We define the parameter of interest, called the generalized effect ratio, as follows

\[
\lambda = \frac{\sum_{i=1}^{I} w_i \sum_{j=1}^{n_i} r_{1ij} - r_{0ij}}{\sum_{i=1}^{I} w_i \sum_{j=1}^{n_i} d_{1ij} - d_{0ij}}
\] (2)

where variables \(w_1, \ldots, w_I\) are fixed non-negative weights and it is implicitly assumed that \(\sum_{i=1}^{I} w_i \sum_{j=1}^{n_i} d_{1ij} - d_{0ij} \neq 0\) (this will generally hold when assumption (A1) from Section 2.3 holds). The parameter \(\lambda\) is a generalization of the parameter considered in previous work in IV estimation based on matching (Baiocchi et al., 2010; Kang et al., 2013). In particular, if \(w_i\) are constant for all \(i\), (2) would be reduced to the regular effect ratio

\[
\lambda = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} r_{1ij} - r_{0ij}}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} d_{1ij} - d_{0ij}}
\] (3)

Furthermore, if \(n_i = 2\), the regular effect ratio is identical to the one in Baiocchi et al. (2010). If \(n_i = k\), the regular effect ratio is identical to the one in Kang et al. (2013). The effect ratio here, both the generalized and regular effect ratio, is a generalization to previous effect ratios in order to handle full matching. In particular, our technical results in the paper encompass the prior works on the effect ratio. Also, the results developed here allows non-binary outcomes.
and exposures, even though our malaria data have binary outcomes and whole-number exposures.

In terms of the interpretation of the effect ratio, \( \lambda \) measures the ratio of the differences in the outcome to the differences in the exposure. For example, in the malaria data, if the instrument, the sickle cell trait, were to reduce malarial episodes by 0.5 per every child and this resulted in a reduction of stunting by 0.05, \( \lambda = 0.05/0.5 = 0.1 \). However, the generalized effect ratio allows each matched set’s differences to contribute differently to \( \lambda \) by the weights \( w_i \). For example, if \( w_i = m_i \), matched sets with more \( Z_{ij} = 1 \) than \( Z_{ij} = 0 \) would contribute more to the differences in the outcome and the exposure. This weighing scheme is useful when we would like to assess the effect of the treatment of people who were assigned the encouraging level of the IV \( (Z_{ij} = 1) \).

For simplicity, we only consider the regular effect ratio stated in (3) throughout the main paper. However, the theory holds for the generalized effect ratio and the proofs to our results in the Supplementary Materials are in terms of the generalized effect ratio.

2.5. Inference of effect ratio

We would like to conduct the following hypothesis test for the effect ratio \( \lambda \).

\[
H_0 : \lambda = \lambda_0, \quad H_a : \lambda \neq \lambda_0
\]

(4)

For example, if \( \lambda_0 = 0 \), (4) is a test of no effect between the exposure and the outcome. To test the hypothesis in (4), we propose the following test statistic

\[
T(\lambda_0) = \frac{1}{I} \sum_{i=1}^{I} V_i(\lambda_0)
\]

(5)

where

\[
V_i(\lambda_0) = \frac{n_i}{m_i} \sum_{j=1}^{n_i} Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - \frac{n_i}{n_i - m_i} \sum_{j=1}^{n_i} (1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij})
\]
and $S(\lambda_0)^2$, the estimator for the variance of the test statistic, $\text{Var}\{T(\lambda_0)|\mathcal{F}, Z\}$

$$S(\lambda_0)^2 = \frac{1}{I(I-1)} \sum_{i=1}^I \{V_i(\lambda_0) - T(\lambda_0)\}^2$$  \hspace{1cm} (6)

Each variable $V_i(\lambda_0)$ is the difference in adjusted responses, $R_{ij} - \lambda_0 D_{ij}$, of those individuals with $Z_{ij} = 1$ and those with $Z_{ij} = 0$. Under the null hypothesis in (4), these adjusted responses have the same expected value for $Z_{ij} = 1$ and $Z_{ij} = 0$ and thus, larger values of $T(\lambda_0)$ suggest $H_0$ is not true.

With equations (5) and (6), Proposition 1 in the Supplementary Materials states that under the regularity conditions, the asymptotic null distribution of $T(\lambda_0)/S(\lambda_0)$ is a standard Normal. This provides a point estimate as well as a confidence interval for the effect ratio. Specifically, setting $T(\lambda)/S(\lambda) = 0$ and solving for $\lambda$ gives an estimate for the effect ratio, $\hat{\lambda}$

$$\hat{\lambda} = \frac{\sum_{i=1}^I \frac{n_i^2}{m_i(n_i - m_i)} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)(R_{ij} - \bar{R}_i)}{\sum_{i=1}^I \frac{n_i^2}{m_i(n_i - m_i)} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)(D_{ij} - \bar{D}_i)}$$ \hspace{1cm} (7)

where $\bar{Z}_i$, $\bar{R}_i$, and $\bar{D}_i$ are averages of the instrument, response, and exposure, respectively, within each matched set. For confidence interval estimation, say 95% confidence interval, we can solve the equation $T(\lambda)/S(\lambda) = \pm 1.96$ for $\lambda$ to get the confidence interval for the effect ratio. In fact, from Corollary 1 in Supplementary Materials, for any value $q$, the solution to $T(\lambda)/S(\lambda) = q$ is a quadratic equation $A_2 \lambda^2 + A_1 \lambda + A_0 = 0$ where the coefficients of the polynomial are

$$A_2 = \bar{H}^2 - \frac{q^2}{I(I-1)} \sum_{i=1}^I (H_i - \bar{H})^2$$

$$A_1 = -2 \bar{G} \bar{H} + \frac{2q^2}{I(I-1)} \left\{ \sum_{i=1}^I (G_i - \bar{G})(H_i - \bar{H}) \right\}$$

$$A_0 = \bar{G}^2 - \frac{q^2}{I(I-1)} \sum_{i=1}^I (G_i - \bar{G})^2$$
where

\[
H_i = \frac{n_i^2}{m_i(n_i - m_i)} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)(D_{ij} - \bar{D}_i)
\]

\[
G_i = \frac{n_i^2}{m_i(n_i - m_i)} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)(R_{ij} - \bar{R}_i)
\]

\[
\bar{H} = \frac{1}{I} \sum_{i=1}^{I} H_i, \quad \bar{G} = \frac{1}{I} \sum_{i=1}^{I} G_i
\]

If \( q = 0 \), i.e. \( T(\lambda)/S(\lambda) = 0 \), we obtain the estimator for the effect ratio in equation (7).

We will use the result from Proposition 1 for our analysis of the malaria data. Specifically, since the responses are binary (i.e. stunted or not stunted) and the malaria episodes are bounded whole numbers, the regularity conditions, specifically the moment conditions in Proposition 1 are automatically met (i.e. \( V_i^4(\lambda) \) is uniformly bounded). Hence, Proposition 1 and Corollary 1 are used to compute the point estimate, the p-value, and the confidence intervals for the casual effect of malaria on stunting.

2.6. Sensitivity analysis

One benefit of using a matching framework for IV estimation is the ability to perform sensitivity analysis. Sensitivity analysis attempts to measure the influence of unobserved confounders on the inference on \( \lambda \). In the case of instrumental variables, a sensitivity analysis quantifies how a violation of assumption (A3) in Section 2.3 would impact the inference on \( \lambda \) (Rosenbaum, 2002). Specifically, under assumption (A3), the instrument is assumed to be free from unmeasured confounders or free after conditioning on observed confounders via matching. The latter implies that the instruments are assigned randomly, \( P(Z = z|\mathcal{F}, Z) = (|\Omega|)^{-1} \), i.e. that within each matched set \( i \),
\[ P(Z_{ij} = 1 | \mathcal{F}, Z) = \frac{m_i}{n_i}. \]

However, as discussed in Section 2.3, even after matching for observed confounders, unmeasured confounders may influence the viability of assumption (A3). For example, with the malaria study, within a matched set \( i \), two children, \( j \) and \( k \), may have the same birth weights, be from the same village, and have the same covariate values (\( x_{ij} = x_{ik} \)), but have different probabilities of carrying the HbAS genotype, \( P(Z_{ij} = 1 | \mathcal{F}) \neq P(Z_{ik} = 1 | \mathcal{F}) \) due to unmeasured confounders, denoted as \( u_{ij} \) and \( u_{ik} \) for the \( j \)th and \( k \)th unit, respectively. Despite our best efforts to minimize the observed differences in covariates and to adhere to assumption (A3) after conditioning on the matched sets, unmeasured confounders such as a child’s family’s ancestry could still be different between the \( j \)th and \( k \)th child, and this difference could lead to violations of assumption (A3). In particular, different family ancestry between two children may make the inheritance of the sickle cell trait depart from randomized assignment.

To model this deviation from randomized assignment due to unmeasured confounders, let \( \pi_{ij} = P(Z_{ij} = 1 | \mathcal{F}) \) and \( \pi_{ik} = P(Z_{ik} = 1 | \mathcal{F}) \) for each unit \( j \) and \( k \) in the \( i \)th matched set. The odds that unit \( j \) will receive \( Z_{ij} = 1 \) instead of \( Z_{ij} = 0 \) is \( \pi_{ij}/(1 - \pi_{ij}) \). Similarly, the odds for unit \( k \) is \( \pi_{ik}/(1 - \pi_{ik}) \). Suppose the ratio of these odds is bounded by \( \Gamma \geq 1 \)

\[ \frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ik})}{\pi_{ik}(1 - \pi_{ij})} \leq \Gamma \]

(8)

If unmeasured confounders play no role in the assignment of \( Z_{ij} \), \( \Gamma = 1 \) and \( \pi_{ij} = \pi_{ik} \). That is, child \( j \) and \( k \) have the same probability of receiving \( Z_{ij} = 1 \) in matched set \( i \). If there are unmeasured confounders that affect the distribution of \( Z_{ij} \), then \( \pi_{ij} \neq \pi_{ik} \) and \( \Gamma > 1 \). By Rosenbaum (2002), equation (8) is equivalent to

\[ P(Z = z | \mathcal{F}, Z) = \frac{\exp(\gamma z^T u)}{\sum_{b \in \Omega} \exp(\gamma b^T u)} \]

(9)
where \( u = (u_{11}, u_{12}, \ldots, u_{I,n_i}) \). Unfortunately, the exact probability of (9) is unknown as it depends on the vector of unobserved confounders, \((u_{11}, \ldots, u_{I,n_i})\). However, for a fixed \( \Gamma > 1 \), we can obtain lower and upper bounds on (9). Furthermore, since the inference on the effect ratio \( \lambda \) is derived from the distribution of \( P(Z = z | F, Z) \), these bounds can be used to compute a range of possible p-values for the hypothesis in equation (4). The range of p-values indicates the effect of unmeasured confounders on the conclusions reached by the inference on \( \lambda \). If the range contains \( \alpha \), the significance value, then we cannot reject the null hypothesis at the \( \alpha \) level when there is an unmeasured confounder with an effect quantified by \( \Gamma \).

Specifically, consider Fisher’s sharp null hypothesis, \( H_0 : r_{1ij} = r_{0ij} \) for all \( i = 1, \ldots, n \) and \( j = 1, \ldots, n_i \). Note that this hypothesis implies the hypothesis \( H_0 : \lambda = 0 \). Furthermore, the test statistic in (5) simplifies to

\[
\tilde{T}(0) = \frac{1}{I} \sum_{i=1}^{I} \left\{ \frac{n_i}{m_i} \sum_{j=1}^{n_i} Z_{ij} R_{ij} - \frac{n_i}{n_i - m_i} \sum_{j=1}^{n_i} (1 - Z_{ij}) R_{ij} \right\}
\]

Regardless of the distribution of \( P(Z | F, Z) \), \( \frac{1}{I} \sum_{i=1}^{I} n_i / (n_i - m_i) \sum_{j=1}^{n_i} R_{ij} \) is a constant since \( r_{1ij} = r_{0ij} \) under Fisher’s sharp null hypothesis. Hence, we can use the simpler statistic, \( \tilde{T}(0) \),

\[
\tilde{T}(0) = \frac{1}{I} \sum_{i=1}^{I} \frac{n_i}{m_i (n_i - m_i)} \sum_{j=1}^{n_i} Z_{ij} R_{ij}
\]

(10)
to test the Fisher’s sharp null hypothesis. If the responses are binary, equation (10) is the sign-score test statistic for which exact bounds on p-values exist (Rosenbaum, 2002). If the responses are continuous, Gastwirth et al. (2000) and Small et al. (2009) provide an approximate bound on p-values.
Following Rosenbaum and Silber (2009), we can also reinterpret the sensitivity parameter $\Gamma$ by considering a binary unmeasured confounder with two values $\Delta$ and $\Lambda$ where $\Delta$ and $\Gamma$ have the following property

$$\Gamma = \frac{\Delta\Lambda + 1}{\Delta + \Lambda}, \quad \Delta > 0, \Lambda > 0 \quad (11)$$

The parameter $\Lambda$ refers to the odds of having one instrument value over another. The parameter $\Delta$ refers to the odds of having one outcome over another. For each $\Gamma$, we can use equation (11) and translate the interpretation of $\Gamma$ as the combined effect an unmeasured confounder must have on the instrument, $\Lambda$, and on the outcome, $\Delta$, to change the inference.

### 2.7. Efficiency

As discussed before, one of the advantages of full matching is its flexibility to accommodate various sizes of matched sets. All things being equal in terms of covariate balance, we would like an estimator of the effect ratio $\lambda$ that is as efficient as possible. This is particularly the case with full matching where an unconstrained full matching can create large matched sets which reduces efficiency (Hansen, 2004). However, we can constrain full matching to increase efficiency by restricting matched sets to have a maximum number of controls and/or treated units per matched set (Hansen, 2004). This section studies statistical efficiency of the estimator for $\lambda$ under different constraints on full matching.

In Section 4 of the Supplementary Materials, Proposition 2 allows us to estimate the impact of different matched sets of $n_i$ produced by different full matching schemes based on the asymptotic variance of the effect ratio estimator in (7). Specifically, under a linear model between $R_{ij}$, $D_{ij}$, and $Z_{ij}$ with
homoscedastic errors, the variance of the effect ratio estimator is roughly

$$Var(\hat{\lambda}) \approx K \frac{\sum_{i=1}^{I} \frac{n_i^3}{n_i^3 - 1}}{(\sum_{i=1}^{I} n_i)^2}$$

where $K$ is some constant that depends on the variance of $R_{ij}$ and the strength of the instrument. Section 4 of the Supplementary Materials also discusses the heteroscedastic error case. Under the linear linear model, this approximation is fairly accurate when the number of matched sets (i.e. large $I$) is large, or when the instrument is strong (see Section 4 of the Supplementary Materials).

Unfortunately, if the relationship between $R_{ij}$, $D_{ij}$, and $Z_{ij}$ is non-linear, the asymptotic variance of the effect ratio estimator is analytically difficult to derive, thereby making efficiency comparisons between matching schemes a challenge. For example, our data consists of binary responses $R_{ij}$ and whole-number malaria episodes $D_{ij}$, which cannot be reasonably generated by the linear link function between $R_{ij}$ and $D_{ij}$ or between $D_{ij}$ and $Z_{ij}$. In Section 4.2, we discuss an alternative, simulation-based strategy to compare the efficiency of different full matching schemes for our malaria data. We use this strategy to examine the tradeoffs between efficiency and covariate balance in the malaria study.

3. Simulation Study

One of the advantages of matching based IV estimation versus traditional IV estimation, such as conventional 2SLS without matching, is its robustness to non-linearity. Specifically, for conventional 2SLS that models the effect of covariates on the outcome linearly, in order for the estimate to be consistent, the covariates must have a linear effect on the expected outcome. In contrast, matching-based IV estimation puts no constraints on the structure of this relationship. In this section, we study this phenomena in detail through a simulation
Let the outcome $R_{ij}$, the exposure $D_{ij}$, the observed covariates $X_{ij}$, and the instrument $Z_{ij}$ be generated based on the following model known as structural equations model in econometrics (Wooldridge, 2010).

$$R_{ij} = \alpha + \beta D_{ij} + f(X_{ij}) + \epsilon_{ij}$$
$$D_{ij} = \kappa + \pi Z_{ij} + \rho^T X_{ij} + \xi_{ij}$$

where the parameters $\alpha, \beta, \kappa$ and $\rho$ are all fixed throughout the simulation. The parameters $\alpha$ and $\kappa$ are intercepts. The parameter $\beta$ is the quantity of interest, the effect of the exposure on the outcome, and is also equal to the effect ratio (see Section 1 of Supplementary Materials for details). The parameter $\pi$ quantifies the strength of the instrument. The function $f(\cdot)$ is a pre-defined function that takes in a vector of observed covariates $X_{ij}$ and produces a scalar value that affects the outcome, $R_{ij}$. In the simulation, $X_{ij}$, are five-dimensional vectors or $X_{ij} = (X_{ij1}, \ldots, X_{ij5})$. Also, we consider the following list of functions parametrized by $\gamma \in \mathbb{R}^5$

(a) Linear function: $f(X_{ij}) = \sum_{k=1}^{5} \gamma_k X_{ijk}$
(b) Quadratic function: $f(X_{ij}) = \sum_{k=1}^{5} \gamma_k X_{ijk}^2$
(c) Cubic function: $f(X_{ij}) = \sum_{k=1}^{5} \gamma_k X_{ijk}^3$
(d) Exponential function: $f(X_{ij}) = \sum_{k=1}^{5} \gamma_k \exp(X_{ijk})$
(e) Log function: $f(X_{ij}) = \sum_{k=1}^{5} \gamma_k \log(|X_{ijk}|)$
(f) Logistic function: $f(X_{ij}) = \frac{1}{1 + \exp(-\sum_{k=1}^{5} X_{ijk} \gamma_k)}$
(g) Truncated function: $f(X_{ij}) = \sum_{k=1}^{5} \gamma_k I(X_{ijk} \geq 0)$ where $I(\cdot)$ is an indicator function.
(h) Inverse function: $f(X_{ij}) = \sum_{k=1}^{5} \frac{\gamma_k}{X_{ijk}}$
(i) Square root function: $f(X_{ij}) = \sum_{k=1}^{5} \gamma_k \sqrt{|X_{ijk}|}$

To generate $X_{ij}$, we adopt the following scheme. For individuals with $Z_{ij} = 0,$
\( X_{ij} \) comes from a five-dimensional multivariate Normal distribution with mean \((0, \ldots, 0)\) and an identity covariance matrix. For individuals with \( Z_{ij} = 1 \), \( X_{ij} \) comes from a five-dimensional multivariate Normal with mean \((1, 0, \ldots, 0)\) and identity covariance matrix. The instruments, \( Z_{ij} \), are generated randomly with ratios between \( Z_{ij} = 1 \) and \( Z_{ij} = 0 \) fixed at 1 to 7, similar to the ratio observed in our malaria data. For each generated data set, we compute the estimate of \( \beta \) using 2SLS and our procedure. 2SLS is based on (i) regressing \( D_{ij} \) on \( Z_{ij} \) and \( X_{ij} \) to obtain the predicted value of \( D_{ij} \), say \( \hat{D}_{ij} \), and (ii) regressing \( R_{ij} \) on \( \hat{D}_{ij} \) and \( X_{ij} \). We simulate this process 5000 times and compute the estimates of \( \beta \) produced by the two procedures. We measure the performance of the two procedures by computing the median absolute deviation, the absolute bias of the median (i.e. the absolute value of the bias of the median estimate with respect to \( \beta \)), and the Type 1 error rate over 5000 simulations. For each simulation study, we vary the function \( f(\cdot) \) and \( \pi \).

Figures 2, 3, and 4 compare performances between 2SLS and our method when we fix the sample size, but vary the strength of the instrument (i.e. the strength of the effect of the instrument on the treatment) via \( \pi \). Specifically, we evaluate the strength of the instrument using a popular measure known as the concentration parameter (Bound et al., 1995). High values of the concentration parameter indicate a strong instrument while low values of it indicate a weak instrument. The concentration parameter is the population value of the first stage partial F statistic for the instruments when the treatment is regressed on the instruments and the measured covariates \( X_{ij} \); this first stage F statistic is often used to check instrument strength where an \( F \) below 10 suggests that the instruments are weak (Stock et al., 2002). The sample size is fixed at 800 where 100 individuals have \( Z_{ij} = 1 \) and the 700 individuals have \( Z_{ij} = 0 \), similar to the sample size presented in the malaria data. We also vary \( f(\cdot) \) based on the
functions listed in the previous paragraph.

Figure 2 measures the absolute bias of 2SLS and our method. When \( f(\cdot) \) is a linear function of the observed covariates \( x_{ij} \), 2SLS does better than our method, which is to be expected since 2SLS works best when the model is linear. However, if \( f(\cdot) \) is non-linear, our matching estimator does better than 2SLS and is never substantially worse for all instrument strength. For example, for quadratic, cubic, exponential, log, and square root functions, our method has lower bias than 2SLS for all strengths of the instrument. For logistic, truncated, and inverse functions, our method is similar in performance to 2SLS for all strengths of the instrument.

Figure 3 measures the median absolute deviation (MAD) of 2SLS and our method. Our method tends to have a slightly higher MAD than 2SLS. This higher variability of our method is to be expected since our method uses a nonparametric approach whereas 2SLS is a parametric approach. However, as the instrument gets stronger (i.e. high concentration parameter), the gap between the two MADs shrinks quickly.

Finally, Figure 4 measures the Type I error rate of 2SLS and our method. Regardless of the function type and the instrument strength, our method retains the nominal 0.05 rate. In fact, even for the linear case where 2SLS is designed to excel, our estimator has the correct Type I error rate for all instrument strengths while 2SLS has higher Type I error for weak instruments. For all the non-linear functions, the Type I error rate for 2SLS remains above the 0.05 line while our estimator maintains the nominal Type I error rate. This provides evidence that our estimator will have the correct 95% coverage for confidence intervals regardless of the non-linearity or instrument strength.

We have considered a range of \( f(\cdot) \) and for all the functions \( f(\cdot) \) in our list, the simulation study shows that our procedure is robust to non-linearity
between the outcome $R_{ij}$ and the covariates $X_{ij}$, especially with respect to Type I error rate and bias. For example, the inference from our estimator attains the nominal Type I error rate regardless of the strength of the instrument or the function $f(\cdot)$. In contrast, 2SLS suffers from higher than nominal Type I error rates, which implies that the coverage of its confidence interval will be lower than expected. Unfortunately, our estimator does pay a price in variability. In particular, our estimator tends to have a higher variability than 2SLS, but only slightly. In addition, the gap between the variability shrinks as instrument strength increases. In summary, the simulation study shows promise that our method is generally robust to assumptions about linearity between the outcome and the covariance and instrument strength at the small expense of variance.

4. Analysis of the Effect of Malaria on Stunting

4.1. Background information

The motivation for our development of the full matching IV method was based on the public health question of whether malaria causes stunted growth in children. If it is the case that malaria causes stunted growth, several intervention strategies can be implemented such as distribution of mosquito nets, control of the mosquito population during seasons of high malarial incidence, and surveillance of mosquito populations. Given the fact that randomized clinical trials are impossible and measuring all confounders is impractical in this study, we opted to use an instrumental variables approach, specifically Mendelian randomization, with the sickle cell trait used as an instrument. In this section, we utilize our new method to answer the original question, whether malaria causes stunting, and highlight some important advantages of our full matching IV method in contrast to more standard regression-based IV methods, such as 2SLS.

The data for analysis were collected from January 2003 to January 2004.
infants from Ghana were recruited to a clinical trial on Intermittent Preventative Treatment for malaria (IPT) (Kobbe et al., 2007). From the time of recruitment at 3 months of age until age two years, each child was monitored monthly for the presence of malaria parasites with 3-monthly measurements of length/height.

For this data analysis, only infants with the heterozygous strand HbAS, the sickle cell trait, or wildtype HbAA were considered, which reduced the sample size from 1070 to 884. Children with the homozygous strand (HbSS), which causes sickle cell disease rather than just the sickle cell trait, or a different mutation on the same gene leading to hemoglobin C (HbAC, HbCC, HbSC) were excluded following Kreuels et al. (2009). Among 884 children, 110 children carried HbAS and 774 children carried HbAA. Also, the difference in episodes of malaria between children with HbAS and HbAA is significant (Risk ratio: 0.82, p-value: 0.02, 95% CI: (0.70, 0.97)).

The instrument was a binary variable indicating either the HbAS or HbAA genotype. The exposure of interest was the malarial history, which was defined as the total number of malarial episodes during the study. A malaria episode was defined as having a parasite density of more than 500 parasites/µl and a body temperature greater than 38°C or the mother reported a fever within the last 48 hours. The outcome was whether the child was stunted at the last recorded visit, which occurred when the child was approximately two years old.

Table 1 summarizes all the measured confounders in this data. Specifically, they are: birth weight, sex, birth season, ethnic group, presence of alpha-thalassaemia, village of residence, mother’s occupation, mother’s education, family’s financial status, mosquito protection, and treatment arm of the original trial from which this data was collected (SP vs. placebo). We see that there are a few significant differences between the HbAS and HbAA groups, most notably in birth weight, village of birth, and mosquito protection status. Children with
the sickle cell trait (HbAS) tend to have high birth weights and lack any protection against mosquitoes compared to HbAA children. Also, children living in the village of Tano-Odumasi tend to inherit HbAA more frequently than HbAS. Any one of these differences can contribute to the violation of IV assumption (A3) in Section 2.3. For instance, it is possible that children with low birth weights were malnourished at birth, making them more prone to malarial episodes and stunted growth compared to children with high birth weights. We must control for these differences to eliminate this possibility, which we do in the next section.

4.2. Full Matching and Efficiency

We conduct full matching on all observed covariates. In particular, we group children with HbAS and without HbAS based on all the observed characteristics in Table 1 as well as match for patterns of missingness. To measure similarity of the observed and missing covariates, we use the rank-based Mahalanobis distance as the distance metric for covariate similarity (Rosenbaum, 2010). In addition, we compute propensity scores by logistic regression. Here, the propensity score is an instrumental propensity score, which is the probability of having the sickle cell trait given the measured confounders (Cheng, 2011). In addition, children with missing values in their covariates were matched to other children with similar patterns of missing data (Rosenbaum, 2010). Once covariate similarity was calculated, the matching algorithm optmatch in R (Hansen and Klopfer, 2006) matched children carrying HbAS with children carrying HbAA in a way that within each matched set, their covariates are similar.

Figure 5 shows covariate balance before and after full matching. Before matching, there are differences in birth weight, mosquito protection, and village of residence between children with HbAS and HbAA. After matching, these covariates are balanced. Specifically, the standardized differences for birth weight,
village of residence, and mosquito protection, are under 0.1 indicating balance (Normand et al., 2001). In fact, all the covariates are balanced after matching and the p-values used to test the differences between HbAS and HbAA in Table 1 are no longer significant after matching.

Because the response is binary, we cannot use the efficiency formulas in Section 2.7. Instead, we conduct a simulation study comparing efficiency and covariate balance for different full matching schemes that use all the data. For each matching scheme, we fix \( Z_{ij} \) and \( X_{ij} \), which, in turn, fixes the matched sets. For the other variables, \( D_{ij} \) and \( R_{ij} \), we assume a Poisson relationship between \( D_{ij} \) and \( Z_{ij} \) and a logistic relationship between \( D_{ij} \) and \( R_{ij} \). In particular, we use the following model

\[
P(R_{ij} = 1) = \frac{1}{1 + e^{-(\alpha_i + \beta D_{ij} + u_{ij})}}, \quad E(D_{ij}) = e^{\tau_i + \gamma Z_{ij}}
\]

We fix \( \beta \), the effect of malaria on stunting, to be 0.32 based on the estimates in Kang et al. (2013). We fix \( \gamma \), the strength of the instrument, to be -0.20, based on the risk ratio estimate in Section 4.1. We also randomly choose \( \alpha_i \) and \( \tau_i \), the intercepts, from Normal distributions with means \(-1.67\) and \(-0.19\), respectively, and variances \(0.12\) and \(0.027\), respectively. The mean and the variance for \( \alpha_i \) is from the intercept term and its corresponding standard error of the logistic regression between \( R_{ij} \) and \( D_{ij} \). Similarly, the mean and the variance for \( \tau_i \) is from the intercept term and its corresponding standard error of the Poisson regression between \( D_{ij} \) and \( Z_{ij} \). Once all the parameters are set, we sample 884 observations of \((R_{ij}, D_{ij})\) (i.e. the sample size of the malaria data set) and compute the effect ratio estimator in equation (7) based on the sample of 884. Note that the effect ratio estimator in equation (7) should be able to estimate \( \beta \) since it doesn’t rely on the functional form between stunting (i.e. outcome) and malaria episodes (i.e. exposure). We repeat the simulation 5000 times and
compute the median absolute deviation as a robust proxy for variance of the effect ratio estimator.

Table 2 shows the trade-off between efficiency and covariate balance for different full matching schemes that use all 884 samples of the malaria data. In particular, we restrict the matched set sizes to different values to see their impact on efficiency and standardized bias. We see that unrestricted full matching has the lowest bias among all other full matching schemes. However, full matching with restricted strata size of 9 has the lowest median absolute deviation, albeit by a little in comparison to other matching schemes. Given the large bias reduction by using unrestricted full matching with a small gain in median absolute deviation, we adopt to use unrestricted full matching.

4.3. Estimated causal effect of malaria on stunting

Table 3 shows the estimates of the causal effect of malaria on stunting from different methods, specifically our method, conventional two stage least squares (2SLS), and multiple regression. Our method computed the estimate by the procedure outlined in Section 2.5. 2SLS computed the estimate by regressing all the measured covariates and the instrument on the exposure and using the prediction from that regression and the measured covariates to obtain the estimated effect. Finally, the multiple regression estimate was derived by regressing the outcome on the exposure and the covariates. We compute the strength of the instrument for 2SLS by evaluating the first stage F-statistic along with its $R^2$ value. We also compute the strength of the instrument for our matching method by regressing the exposure (malaria episodes) onto the sickle cell trait and dummy variables that indicate which group a child belongs to and computing the F statistics from this regression.

We see that the full matching method estimates the risk of stunting i-
creases, on average, by 0.22 for every malarial episode. Furthermore, we reject the hypothesis $H_0 : \lambda = 0$, that malaria does not cause stunting, at the 0.05 significance level. The confidence interval $\lambda$ is (0.044, 1.0). Even the lower limit of this confidence interval means that malaria has a substantial effect on stunting and reduction of malaria would substantially reduce stunting; if $\lambda$ were equal to 0.044, the lower limit of the confidence interval, then an intervention which reduced malaria episodes over a child’s first two years of life by 1 per child would reduce the risk of stunting by 0.044, a substantial reduction in stunting.

In comparison to Kang et al. (2013) which used 1 to 5 matching and removed 25% of the original data (i.e. removed 224 subjects), the 95% confidence interval (0.09, 1.0) is similar to the 95% confidence interval using full matching (0.044, 1.0). Both confidence intervals also contain the point estimate in Kang et al. (2013), 0.32, and the point estimate using full matching. The estimate based on 2SLS is 0.21, similar to our method. However, our method achieves statistical significance in comparison to 2SLS. Also, multiple regression, which does not control for unmeasured confounders, estimates a much smaller effect of 0.018. For instrument strength for full matching, the $F$ statistic is 4.15 and its $R^2$ is 0.21. For instrument strength for TSLS, the $F$ statistic is 4.36 and its $R^2$ is 0.22.

Table 4 shows the sensitivity analysis due to unmeasured confounders. Specifically, we measure how sensitive our estimate and the p-value in Table 3 is to violation of assumption (A3) in Section 2.3, even after matching. We see that our results are somewhat sensitive to unmeasured confounders at the 0.05 significance level. If there is an unmeasured confounder that increases the odds of inheriting HbAS over HbAA by 10%, i.e. $\Gamma = 1.1$, then we would still have strong evidence that malaria causes stunting. But, if an unmeasured confounder increases the odds of inheriting HbAS over HbAA in a child by 20% (i.e. $\Gamma = 1.2$),
the range of possible p-values includes 0.05, the significance level, meaning that we would not reject the null hypothesis of $H_0 : \lambda = 0$, that malaria does not cause stunting.

Figure 6 shows another way to look at the sensitivity analysis by looking at the effect by unmeasured confounders on the odds of stunting and odds of inheriting HbAS over HbAA and on the inference. Specifically, the different values of $\Gamma$ in the sensitivity analysis provides us with range of possible p-values. By equation (11) in Section 2.6, each $\Gamma$ is associated with two other sensitivity parameters $\Delta$, odds of stunting, and $\Lambda$, odds of inheriting HbAS over HbAA, and can be presented as a two-dimensional plot with each axis representing $\Delta$ and $\Lambda$. For example, the point $(\Delta = 1.5, \Lambda = 1.5)$ on Figure 6 represents an unmeasured confounder that increases the odds of stunting and inheriting HbAS over HbAA by a factor of 1.5 and produces a p-value in between 0.025 and 0.05, which does not contain the significance level of 0.05. Hence, the null hypothesis would still be rejected despite having such an unmeasured confounder. In contrast, if the unmeasured confounder had an effect of $(2.0, 2.0)$ specified on the plot, the null hypothesis would be retained since the p-value contains the significance level of 0.05.

Overall, in contrast to regression-based procedures like 2SLS, our full matching method (i) provided a clear way to assess the balance of observed covariates and design the study without looking at the outcome data and (ii) provided a method to quantify the effect of unmeasured confounders on our inference of the causal effect. Our method made it explicitly clear how these covariates were adjusted by stratifying individuals based on similar covariate values. Finally, like in a randomized experiment, our analysis only looked at the outcome data once the balance was acceptable, i.e. once the differences in birth weight, village of residence, and mosquito protection between children with HbAS and HbAA
were controlled for. If the balance was unacceptable, then comparing the outcomes between the two groups would not provide reliable causal inference since any differences in the outcome can be attributed to the differences in the covariates. In contrast, conventional 2SLS can only analyze the causal relationship in the presence of outcome data, making the outcome data necessary throughout the entire analysis.

5. Discussion

Instrumental variables (IV) is a popular method to assess a potential causal relationship between an exposure and an outcome in the presence of measured and unmeasured confounding. The most popular IV estimation method, 2SLS, has some drawbacks, especially with regards to controlling for measured covariates and reliance on parametric modeling assumptions between the outcome and the observed covariates. In this paper, we propose an IV estimation method based on full matching that addresses some of the drawbacks. We also provide a formula to compute the efficiency of IV estimates based on full matching. In addition, we conducted simulation studies to study the performance of our estimator versus 2SLS when the relationship between the outcome and the covariates is non-linear. Through the simulation study, we find that our estimator is generally robust to modeling assumptions.

We used our full matching instrumental variables method to analyze the effect of malaria on stunting using the sickle cell trait as an instrumental variable. With our method, we were able to match individuals based on the similarity of their covariates and clearly demonstrate that covariates were balanced after matching, before looking at the outcome. We found that an increase in one malaria episode leads to, on average, an increase in the risk of stunting by 0.22, which was deemed significant at the 0.05 level and is also practically relevant.
In contrast, 2SLS provides a similar estimate of 0.21, but does not achieve statistical significance. Compared to previously developed matching methods for IV, the full matching IV method we developed in this paper leverages all the individuals in the data and estimates a treatment effect for all these individuals.

References


Table 1. Characteristics of study participants at recruitment. P-values were obtained by doing a Pearson’s chi-squared test for categorical covariates and two-sample t tests for numerical variables. *** corresponds to a p-value of less than 0.01, ** corresponds to a p-value between 0.01 and 0.05, and * corresponds to a p-value between 0.05 and 0.1.

<table>
<thead>
<tr>
<th></th>
<th>HbAS (n = 110)</th>
<th>HbAA (n = 774)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth weight (Mean,(SD))</td>
<td>3112.44 (381.9) (32 missing)</td>
<td>2978.7 (467.9) (239 missing) ***</td>
</tr>
<tr>
<td>Sex (Male/Female)</td>
<td>46.4% Male 51.0% Male</td>
<td>56.4% Dry 55.3% Dry</td>
</tr>
<tr>
<td>Birth season (Dry/Rainy)</td>
<td>56.4% Dry 55.3% Dry</td>
<td>56.4% Dry 55.3% Dry</td>
</tr>
<tr>
<td>Ethnic group (Akan/Northerner)</td>
<td>86.4% Akan 88.8% Akan (4 missing)</td>
<td></td>
</tr>
<tr>
<td>α-globin genotype (Norm/Hetero/Homo)</td>
<td>75.7% / 21.5% / 2.8% (3 missing)</td>
<td>74.4% / 23.1% / 2.6% (29 missing)</td>
</tr>
<tr>
<td>Village of residence:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afamanso</td>
<td>4.6%</td>
<td>4.8%</td>
</tr>
<tr>
<td>Agona</td>
<td>10.0%</td>
<td>13.6%</td>
</tr>
<tr>
<td>Asamang</td>
<td>13.6%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Bedomase</td>
<td>5.5%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Bipoa</td>
<td>14.5%</td>
<td>10.7%</td>
</tr>
<tr>
<td>Jamasi</td>
<td>15.5%</td>
<td>13.8%</td>
</tr>
<tr>
<td>Kona</td>
<td>16.4%</td>
<td>12.8%</td>
</tr>
<tr>
<td>Tano-Odumasi</td>
<td>4.5%</td>
<td>12.3%**</td>
</tr>
<tr>
<td>Wiamoase</td>
<td>15.5%</td>
<td>16.4%</td>
</tr>
<tr>
<td>Mother’s occupation (Nonfarmer/Farmer)</td>
<td>79.0% Nonfarmer 78.0% Nonfarmer (11 missing)</td>
<td></td>
</tr>
<tr>
<td>Mother’s education (Literate/Illiterate)</td>
<td>91.7% Literate (2 missing) 90.5% Literate (8 missing)</td>
<td></td>
</tr>
<tr>
<td>Family’s financial status (Good/Poor)</td>
<td>69.1% Good (13 missing) 70.1% Good (84 missing)</td>
<td></td>
</tr>
<tr>
<td>Mosquito protection (None/Net/Screen)</td>
<td>55.7% / 32.0% / 12.4% (13 missing)</td>
<td>45.4%* / 35.1% / 19.5% (76 missing)</td>
</tr>
<tr>
<td>Treatment of sulphadoxine pyrimethamine (Placebo/SP)</td>
<td>49.1% Placebo 50.1% Placebo</td>
<td></td>
</tr>
</tbody>
</table>

6. Tables and Figures
Table 2. Trade-off between efficiency and balance for different full matching schemes that use all the data based on simulation. Median absolute deviation measures the variability of the effect ratio estimates. The standardized bias is the standardized bias of the instrumental propensity score (Cheng, 2011) and was calculated by the same method as Figure 5 where the difference in propensity scores before and after matching is normalized by the within group standard deviation before matching (the square root of the average of the variances within the group).

<table>
<thead>
<tr>
<th>Matching</th>
<th>Median absolute deviation</th>
<th>Standardized bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full matching (max strata size is 9)</td>
<td>0.90</td>
<td>0.23</td>
</tr>
<tr>
<td>Full matching (max strata size is 10)</td>
<td>0.96</td>
<td>0.19</td>
</tr>
<tr>
<td>Full matching (max strata size is 15)</td>
<td>0.97</td>
<td>0.10</td>
</tr>
<tr>
<td>Full matching (unrestricted)</td>
<td>0.98</td>
<td>0.055</td>
</tr>
</tbody>
</table>

Table 3. Estimates of the causal effect using full matching compared to two-stage least squares and multiple regression. For our method, the estimate, the p-value, and the confidence interval were obtained based on Section 2.5. For two-stage least squares, the confidence interval and the p-value for two-stage least squares were computed based on asymptotic formulas. For multiple regression, the estimate, the p-value and the confidence interval were obtained by using standard t-tests.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Estimate</th>
<th>P-value</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>0.22</td>
<td>0.011</td>
<td>(0.044, 1)</td>
</tr>
<tr>
<td>Two stage least squares</td>
<td>0.21</td>
<td>0.14</td>
<td>(-0.065, 0.47)</td>
</tr>
<tr>
<td>Multiple regression</td>
<td>0.018</td>
<td>0.016</td>
<td>(0.0034, 0.033)</td>
</tr>
</tbody>
</table>
Table 4. Sensitivity analysis for instrumental variables with full matching. The range of significance is the range of p-values over the different possible distributions of the unmeasured confounder given a particular value of $\Gamma$, which represents the effect of unobserved confounders on the inference of $\lambda$.

<table>
<thead>
<tr>
<th>Gamma</th>
<th>Range of significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>(0.0082, 0.041)</td>
</tr>
<tr>
<td>1.2</td>
<td>(0.0034, 0.074)</td>
</tr>
<tr>
<td>1.3</td>
<td>(0.0015, 0.12)</td>
</tr>
</tbody>
</table>

Fig. 1. Causal diagram for the malaria study. Numbers (A1,A2,A3) represent MR assumptions.
Fig. 2. Absolute bias between our method and two stage least squares (2SLS) for different concentration parameters. The solid line indicates 2SLS and the dotted line indicates our method.
Fig. 3. Median absolute deviation between our method and two stage least squares (2SLS) for different concentration parameters. The solid line indicates 2SLS and the dotted line indicates our method. Note that the scaling of the y-axis for the inverse function plot is different than the rest of the plots.
Fig. 4. Type I error rate between our method and two stage least squares (2SLS) for different concentration parameters. The solid line indicates 2SLS and the dotted line indicates our method.
Fig. 5. Absolute standardized differences before and after full matching. Unfilled circles indicate differences before matching and filled circles indicate differences after matching. Absolute standardized differences before matching are computed by taking the difference of the means between children with HbAS and HbAA for each covariate, taking the absolute value of it, and normalizing it by the within group standard deviation before matching (the square root of the average of the variances within the groups). Absolute standardized differences after matching are computed by taking the differences of the means between children with HbAS and HbAA within each strata, averaging this difference across strata, taking the absolute value of it, and normalizing it by the same within group standard deviation before matching as before.
Fig. 6. Amplification of sensitivity analysis. Each point on the graph represents an effect by an unmeasured confounder on the instrument (HbAS) and on the outcome (stunting) to change the inference, specifically the p-value. Points within the two bold curves correspond to effects by unmeasured confounders that will give us p-values < 0.05 and points outside the two bold curves correspond to effects that will give us p-values > 0.05, thereby retaining our null hypothesis.