The Causal Effects of U.N. Peacekeeping Operations on Economic Development

Scholars in the various literatures that comprise the study of conflict management are united in a common goal. Regardless of the type of conflict management strategy studied, researchers wish to know if these strategies *work*. Unlike other areas of conflict studies, such as conflict onset, where researchers are often interested in structural factors such as poverty, contiguity, etc., the independent variables of interest in the conflict management literatures are almost always policy-applicable strategies. Whether a scholar is studying the effects of mediation, military intervention, foreign aid, or combatant re-integration programs, the logic extending from the conclusion of such studies is near universal. “We found that conflict management strategy X is associated (or is not associated) with a favorable outcome. Therefore, we encourage policymakers to pursue (or consider alternatives to) conflict management strategy X.” However, associations and correlations are not causal effects, which are primarily what scholars, policymakers, and the public are interested in. Given that conflict management strategies are similar to experimental treatments (albeit, treatment, a conflict management strategy, is not randomized) in that they are explicit events with a before-and-after period where some units (countries, regions, individuals) are treated and others are not, a great opportunity exists for scholars to apply the logic of experimental design to evaluate the causal effects of conflict management techniques.

In the following document, I execute a quasi-experimental design, evaluating the long-term causal effects of PKO deployments and withdrawals on economic development as an example of causal conflict management research. Many scholars have documented the phenomena of “peacekeeping economies”, where higher levels of economic activity in a host country are directly attributed to the deployment of an international peacekeeping force (Jennings 2015, Rolandsen 2015, Jennings and Bøås 2015, Jennings 2018). Following the onset of a PKO deployment, there are numerous channels through which a PKO may contribute to economic growth and development. PKOs can assist in the peace agreement process and police and monitor armed and criminal actors in the post-conflict environment, leading to a safer environment in which economic activity can increase. Emboldened by the stabilizing environment partially produced by a foreign force, individuals can feel physically safer in public to purchase and sell goods and services. Further, a consistently stabler environment may also incentivize increased business activity as the rule of law and property protection is fostered. Additionally, the presence of a large foreign force artificially stimulates demand for a variety of goods and services such as housing, food, transportation, retail, housework, etc. PKOs can assist in the development of civil infrastructure projects such as schools and hospitals and can directly or indirectly (through INGOs) provide immediate relief through humanitarian aid. Local individuals may even receive direct vocational opportunities from a PKO itself.

Alternatively, it may also be the case that alleged economic benefits of peacekeeping economies are not sustainable and are PKO-dependent. For example, when a PKO withdraws, the artificial demand for a variety of goods and services from the PKO personnel subsequently go away. If the core issues of the prior conflict were not resolved and an enduring peace was not established, the stabilization that fostered economic activity provided by a PKO may likewise erode following a PKO withdrawal. Despite sound theoretical reasoning for a causal effect of UN PKO deployments and withdrawals on economic development, prior quasi-experimental evidence of such effects are mixed (Bove and Elia 2017, Caruso et al. 2017, Beber et al. 2019, Mvukiyehe and Samii 2020, Bove et al. 2021).

Prior to the analysis of the quasi-experimental design of this paper, a number of topics are reviewed to establish a coherent argument advocating the plausibility and necessity of executing causal research in the conflict management literatures. First, an attempt at estimating causal effects warrants discussion on the nature of causal inference and quasi-experimental design in observational settings. As such, the following section discusses statistical control, the standard practice for covariate adjustment in conflict management literatures and its limitations in causal research. Second, I provide a brief review on the prospects of causal research with non-experimental data. Third, I review existing studies in the PKO literature employing quasi-experimental methodology and discuss the virtues and shortcomings of existing strategies. Fourth, I discuss inverse probability weighting and the novel panel data matching methodology developed by Imai et al. (2021) and develop a research design that applies these tools to examine the economic effects of UN PKOs. Counter to much existing theoretical and empirical work, this study presents moderate evidence suggesting PKO deployments cause a decrease in GDP per capita in the immediate eight years post-intervention. Although the long-term direction is negative, the estimated effect of UN PKO withdrawals is not statistically significant. In the results and conclusion section, I discuss numerous concerns relating to the execution of causal research in both the PKO and broader conflict management literatures, acknowledging that the estimates of this study and similar studies should be interpreted with caution. With more awareness devoted towards the contemporary issues complicating causal research in the conflict management literatures, researchers may be able to collectively establish solutions and finally answer the million-dollar question, “does this conflict management program *work*?”

**Limits of Statistical Control**

The standard approach in the conflict management literature (and the social sciences generally) for addressing alternative explanations of an outcome of interest is statistical control. Researchers adjust for alternative explanations by including such control variables in a specified regression formula which includes the main explanatory variable of interest. Following the estimation of a model with this approach, researchers focus on the coefficient for the explanatory variable of interest after covariate adjustment. Researchers may then proclaim something to the effect of, “after controlling for various alternative explanations of the dependent variable, we find that the explanatory is associated with a X-unit change in the dependent variable”. Despite the prominence of this approach, it is lacking in many ways.

A common critique of statistical control is the dependence on the strict assumptions of linear regression. Researchers often insert control variables in a regression equation reflexively without assessing whether the relationship between the control variable(s) and dependent variable is linear when a non-linear specification may be more appropriate. While researchers can create polynomial transformations of variables and specify these transformations in the regression model, such specifications can only occur if researchers consciously choose to do so. A similar concern can be made for interactive effects where the validity of a model’s output is dependent on the researcher to specify such interactive effects. Additionally, compared to alternative mainstream methods, statistical control preforms poorly at balancing covariates, rendering the quality of causal inferences from such designs poor (if researchers attempt to make causal inferences at all). Alternative methods such as matching and weighting are often viewed as favorable for making causal inferences from observational data as they deliberately attempt to create a balanced data set where observations with similar values in the covariates are deliberately matched with other observations that differ only in treatment (main explanatory variable) status. Not only is the logic of such preferred methods more intuitive for the purposes of causal inference, matching and weighting are likewise not burdened by the assumption of linear relationships.

Beyond the model-based critiques of statistical control, a culture of “controlling” and adding a host of covariates into a regression model has also emerged with negative consequences for the estimation of causal and representative effects. In part, this culture has emerged due to the emphasis on omitted variable bias (OVB). While handling OVB is important, many scholars have discussed the inverse issue of including “too many” or “bad” controls (Achen 2005, Schrodt 2014, Rohrer 2018, Cinelli et al. 2022). As Cinelli et al. (2022) note, even when “bad” controls are discussed, the scope of discussion is often limited to critiques of researchers including controls that are essentially alternative operationalizations of the outcome itself. However, issues of casually including control variables are oftentimes more severe than this critique. Casually including a laundry list of controls can create wildly inaccurate results due to collinearity and incorrect specification of the relationship between the covariates and the outcome variable (Achen 2005, Schrodt 2014). With respect to estimating causal effects, researchers may blindly create bias in a variety of contexts, such as including controls that are correlated with the treatment, but not the outcome, including controls that are influenced by the treatment *and* outcome, or including controls that are caused by the treatment and affect the outcome (Rohrer 2018, Cinelli et al. 2022). Much work has been executed on the topics of statistical control, its limitations, and the bad practices unintentionally created as a result. Such topics are not the core issue of this paper, although, they are important to review as a justification for embracing alternative methods, encouraging a culture of thinking clearly about the logic of which covariates to include and how to adjust for them, and attempting to make causal inferences from observational data.

**Causal Inference with Observational Data**

The ability to establish causation is not one achieved with ease. Two variables may covary with one another but establishing the exact causal connection between two variables is much more challenging. Within a single unit, a treatment can be applied, and the outcome can subsequently alter. However, other factors beyond the treatment that vary across time may also be a part of the causal story. Across spatial dimensions in which a treatment is applied to numerous units, any change in the outcome could simply be a result of unit-specific factors. If the treatment and outcome of interest are both influenced by a common confounding factor, it becomes difficult to disentangle the effect of the treatment on the outcome from the effect the confounder on the outcome. In an ideal world, researchers would have access to a time machine in which they could apply a treatment to a unit and record the outcome. Following this, the researchers could go back in time, ensure that the treatment never occurred, and record the same outcome. The difference in the outcome with and without treatment would represent the causal effect of the treatment. Obviously, this is not an option, and this impossibility is known as the fundamental problem of causal inference.

In the absence of time machines, researchers have relied on randomized controlled trials (RCTs) as the “gold standard” for making causal inferences and resolving the fundamental problem of causal inference. In theory, RCTs resolve two serious problems that arise when attempting to discover a causal effect of a treatment on an outcome. First, because researchers can assign treatment in an RCT, issues of reverse causality in which the outcome effects the treatment are assuaged. Second, because access to treatment is randomized, concerns of confounding are largely ameliorated as any confounding effect cannot be correlated with treatment exposure due to the randomness in treatment assignment. With randomization, pure chance is the only factor correlated with treatment. With confounding effects resolved, the average difference in the outcome between the treated and control units represents the average treatment effect (ATE). While RCTs have their issues and shortcomings (Deaton and Cartwright 2018), their capacity to effectively eliminate the concern of unobserved confounding is impressive. Unfortunatley, for most conflict management research questions, the possibility of RCTs is either impossible or unethical. Researchers do not have the capacity to randomly assign peacekeeping operations, mediated peace agreements, or foreign aid packages. If they did, random assignment would hardly be an ethical method of applying potentially life-saving treatments. As a result, conflict management researchers must attempt to make causal inferences with treatments that are non-random. Consequentially, confounders of an impossible to know quantity - some obvious and some obscure - are introduced that complicate the exact causal relationship between a treatment and an outcome.

From these limitations, both historic and contemporary research in conflict management has tended to neglect causal research. True, while experimental designs in conflict management studies are incredibly difficult to execute, *quasi*-experimental methodology is abundant and has a rich and growing role in social science research. Further, following the emergence and popularization of *do*-calculus and directed acyclic graphs (DAGs), researchers have the capacity to isolate confounders, frame their research designs in experimental language, and isolate causal effects using observational data (Pearl 1995, Rohrer 2018). In the following paragraphs, I discuss two assumptions of causal inference, the ignorability assumption and the stable unit value treatment assumption (SUTVA).[[1]](#footnote-1) Dealing with the former assumption is made possible with DAGs and sensitivity analysis. However, especially for conflict and conflict management research, satisfying SUTVA remains a challenge.

Recall that the threat of unobserved confounding seriously jeopardizes the capacity of a researcher to make causal claims about a particular treatment with observational data. The ignorability assumption (also known as the exchangeability assumption or the assumption of no unmeasured confounding) is the prime culprit for why researchers interpret regression coefficients as associations instead of causal effects. Since it is impossible to control for *all* confounding effects, a researcher cannot be sure that the unmeasured *Nth* confounding effect would not have seriously influenced a causal estimate had it been specified in a model. It should be noted that the possibility of unobserved confounding is *always* present with observational data. Indeed, it is this observation, in part, that led many researchers to avoid causal inference with observational data entirely. DAGs are helpful because they execute the behind-the-scenes math for a user to isolate confounding effects. However, DAGs are only capable of doing so with variables that users have directly specified. Generating a DAG does not solve the issue of unobserved confounding. To resolve this, sensitivity analysis is crucial. While there are at least dozens of different iterations of sensitivity analysis, the common goal with such methods is to examine how sensitive causal estimates are to various hypothetical unobserved confounders. Researchers specify various thresholds in magnitude of unobserved confounders and assess whether the likelihood of a confounder of such a magnitude plausibly exists and, if so, to what degree would the inclusion of said confounder mathematically alter the causal estimate. Together, DAGs and sensitivity analysis do not completely satisfy the ignorability assumption. With observational data, this is not possible. However, the combined use of these tools allows researchers to specify their models correctly and assess the extent to which unobserved confounding is a legitimate threat to observed causal estimates.

On the other hand, SUTVA is perhaps a much more threatening assumption for the prospects of causal inference in conflict management research. SUTVA implies that the outcome of one unit is impacted by their treatment status alone and not the treatment status of other units. Such “spillover effects” are naturally problematic for the study of conflict management given the large amount of research dedicated to the role of geographic contiguity and distance in conflict and peace. It is not too difficult to imagine a certain conflict management strategy, such as a UN peacekeeping operation, being implemented in Country A and this specific PKO impacting levels of violence and stability in contiguous, non-treated countries B and C. Indeed, this very scenario has some empirical support (Beardsley 2011). When SUTVA is violated, we cannot assure that non-treated units are serving as control units at all, seriously complicating any causal effect of a particular treatment. Unfortunatley, attempts at satisfying SUTVA are not yet commonplace or standardized. While satisfying SUTVA can become a problem for RCTs as well, the inability of researchers to control and monitor treated units in a quasi-experimental setup renders current efforts at satisfying SUTVA difficult when working with observational data. As a result, I do not claim to satisfyingly address SUTVA in this research project. Nonetheless, considering SUTVA and strategies to satisfy SUTVA for conflict management research questions with observational data should be at the forefront of methodological concerns in this literature.

**Attempts at Making Causal Inferences in the PKO Literature with Observational Data**

After reviewing the logic of causal inference and assumptions required for causal inference, it is worth evaluating popular quasi-experimental designs employed in the peacekeeping literature to estimate causal effects. As noted earlier, statistical control often falls short compared to alternative methods. Many researchers have acknowledged this in the peacekeeping literature (albeit, for reasons that are oftentimes not explicitly linked to the improvement of causal inferences). In the following section, I review these alternative strategies, briefly explaining the logic of such quasi-experimental design methods and their limitations as it relates to the peacekeeping literature.

Fixed effects (FE) are often employed in the peacekeeping literature (Joshi 2013, Hultman et al. 2014, Kocher 2014, Bove and Ruggeri 2016, 2018, Fjedle et al. 2018, Haas and Ansorg 2018, Beber et al. 2019, Blair 2019, Di Salvatore 2019, Phayal 2019, Phayal and Prins 2019, Bara 2020, Bove et al. 2021), oftentimes not explicitly for causal inference purposes (in many cases, FE is employed to account for omitted variable bias). The implementation of fixed effects can be helpful for making causal inferences due to its capacity to control for all unobserved *time-invariant* factors of a specified unit. By creating a dummy variable for each unit, researchers can remove confounding effects that are unit-specific. Confounding effects such as these are often hard, if not impossible, to identify individually, which lends credit to the implementation of fixed effects. However, two glaring issues with the implementation of fixed effects for making causal inferences in the PKO literature should be noted. First, for the study of post-conflict environments, the implementation of fixed effects for the study of PKOs *as an event* is impossible given that the presence of a PKO in a prior conflict spell is a *time-invariant variable*. In other words, it is *fixed*, meaning that a scholar studying PKOs would be unable to determine the effect of PKOs independent of the other unit-specific fixed factors. This problem can be avoided if one alters their measure of PKOs. If one chooses not to measure PKOs using a dummy, opting to include a fluid measure such as the number of personnel involved in the PKO instead, fixed effects can still be employed given that the PKO measure is no longer a time-invariant variable. Still, while fixed effects accounts for all time-invariant aspects of a unit, it does not eliminate the potential for a *time-variant* confounder to slip through the cracks and bias estimates. One may be tempted to use the two-way fixed effects (TWFE) to account for both unit-specific and time-specific confounders. However, as Imai and Kim (2021) have demonstrated, numerous issues arise when attempting to use TWFE for causal inference.

In the absence of experimental data, many scholars in the field have adopted an instrumental variables approach to making causal inferences concerning the effect of PKOs (Sambanis 2008, Vivalt 2015, Caruso et al. 2017, Ruggeri et al. 2017, Blair 2019, Bove et al. 2021). The instrumental variables approach seeks to remove aspects of the treatment that are endogenous (associated with confounders) and retain the exogenous aspects (not associated with confounders) of the treatment. The instrumental variables approach does so by identifying a variable (an instrument) that is correlated with the treatment, is not correlated with other confounding factors, and is correlated with the outcome *only through the* treatment. If these conditions are met, it can be assumed that the instrument reflects a portion of the exogenous aspects of the treatment and is untainted by confounding factors. Predicted values are generated by regressing the treatment on the instrument and these subsequent predicted values are used to estimate the causal effect of the treatment on the outcome. While this method is appealing when there are theoretical reasons to believe that confounding variables are present that current data either does not or cannot account for, this approach has not been implemented without controversy. Gilligan and Sergenti (2008) criticized the use of instrumental variables, referring to causal estimates from such an approach as invalid. These authors argued that the literature has a good grasp on the confounders that complicate the relationship between PKOs and peace, rendering the concern of unknown confounders relatively unimportant. Further, the authors were also skeptical that an instrument for this type of research *could* exist on the grounds that "Any factor that affects how long a war or its subsequent peace will last should also be taken into account by the UN Security Council when it is deciding whether or not to allocate a mission" (Gilligan and Sergenti 2007, p. 91). Essentially, the authors argued that there are no exogenous aspects of the treatment (UN PKO) given that the authorization of PKOs are heavily influenced by endogenous factors related to conflict and peace duration. Indeed, the discovery of valid instruments are particularly difficult given the challenge of satisfying the excludability assumption in which the instrument effects the outcome solely through the treatment. For example, weather is commonly used as an instrument in conflict studies employing an instrument variables approach. However, recent work has suggested that this once-reliable instrument heavily violates the excludability assumption (Mellon 2021). Such findings present a fundamental problem with the use of instrumental variables. Instruments are as valid as our ability to argue that the instrument effects the outcome solely through the treatment, rendering the validity of these instruments incredibly sensitive and subjective.

Instead of instrumental variables, Gilligan and Sergenti (2007) suggested the adoption of matching as an approach to improve causal estimates in the peacekeeping literature. The virtues of matching, as the authors claimed, can be attributed to the relative simplicity and transparency of the technique. Units are matched to each other according to their similarity with a specific number of confounding factors. They differ, however, with respect to their treatment status. Given the similarity between matched units, the difference in outcome between matched units *may* be indicative of a causal effect of the treatment. Matches can be made transparent along with the variables on which they are matched. Indeed, given the intuitive nature of this approach, matching is widely employed in the peacekeeping literature (Sambanis 2008, Kathman and Wood 2011, Hultman et al. 2013, Hultman et al. 2014, Ruggeri et al. 2017, Di Salvatore 2018, Fjelde et al. 2018, Haass and Ansorg 2018, Beber et al. 2019, Bara 2020, Mvukiyehe and Samii 2020), albeit, not always as a method to explicitly improve causal interpretation (oftentimes, matching is employed as a “robustness check” to assess the strength of results estimated with adjusted covariates). However, matching strategies have two notable drawbacks. First, because matching *matches* observations and weights these observations in a regression framework accordingly, many observations are naturally omitted from the analysis if a comparable match is not found within the data set. Especially in areas of conflict management research where the number of observations cannot afford to be significantly trimmed, this represents a major problem.

Researchers generally have two options to resolve this issue. First, researchers can relax matching criterion to allow the top *k* best matches to serve as control units rather than relying on the first best match. In addition, researchers can opt to match with replacement, meaning that a control unit can serve as a control unit for more than one treated unit if it is an optimal match for other treated units. While these measures offer a larger *N* to work with, the quality of matching breaks down when implemented. Units that are increasingly different from one another will be matched (creating further issues with resolving confounding) and repeated control groups are overrepresented in the sample. Another approach is the use of weighting, particularly, inverse probability weighting. Like matching, IPW generates weights to balance the data set along specified confounders. Unlike matching, IPW does this in a manner that does not omit any observations.[[2]](#footnote-2) IPW does so by following a two-step process. First, a numeric value for the propensity of receiving treatment predicted by specified covariates is generated for each observation. Given that treatments are often binary, logistic regression is commonly employed for this purpose, although, other advanced machine learning classification algorithms can be employed for more precise propensity scores. Next, these observations are weighted. Observations are weighted more heavily when their propensity to receive treatment differs largely from their *actual* exposure to treatment. For example, according to the results of a logistic regression model, if a country was very unlikely to receive a UN PKO and *still received one*, this observation would be weighted heavily. Likewise, if the results of the model predicted a certain country had a very high chance of receiving a PKO and *did not*, then this observation is also weighted heavily. Observations that experience the treatment in accordance with the predicted propensity to receive treatment are less heavily weighted. Rather than relying on distance to another observation to be weighted, IPW does not require any data points to be dropped as weights are generated agnostic of other observations. While one may be tempted to match on the generated propensity scores themselves, using these scores themselves as weighting criteria creates a number of issues for making causal inferences (King and Nielsen 2019).

In theory, with a sufficiently large *N*, we could simply run models with matched and weighted data sets and leap right into results interpretation. However, as Imai et al. (2021) note, contemporary mainstream matching methods are designed to operate within a *cross-sectional* setting in which unit observations are not repeated over time. Given that most conflict management research utilizes panel (time-series cross-sectional) data, this is problematic. In the following section, I discuss the issues that arise when utilizing matching methods designed for cross-sectional research using panel data. I also discuss the novel approach developed by Imai et al. (2021) that allows researchers to follow a matching strategy with panel data. Lastly, I develop a research design to implement this novel method for the study of UN peacekeeping operations.

**Matching and Weighting with Panel Data**

In the canonical setting, matching methods are applied to evaluate a situation with a single treatment for various cross-sectional units at a given time. However, applying conventional matching techniques for settings when this is not the case introduces several complications. Within panel data, cross-sectional units can receive multiple treatments over time, units can reverse their treatment status over time, and the treatment itself can be applied at different time intervals for different cross-sectional units. To allow for matching given these possibilities, Imai et al. (2021) introduced a novel matching method that allows for matching analysis to be executed using panel data. Instead of a unit-time observation being matched with another similar unit-time observation, control *observations* are selected based on an identical pre-treatment history for a specified timespan (a matched set). For example, if a researcher specified a four-period lag, units would only be matched if their pre-treatment history is identical four-periods prior to treatment. Following this, a refinement method is applied that balances the covariates. For robustness purposes, I use nearest-neighbor matching with Mahalanobis distance and inverse probability weighting (IPW). I conduct two separate analyses for the nearest-neighbor matching, allowing up to 5 matches in one set of analysis and up to 10 matches in the other. Finally, a difference-in-differences (DID) estimator is applied to account for possible time trends in the estimation of the treatment effect.[[3]](#footnote-3) The causal estimate of interest is a contemporaneous and user-specified lead effect (eight post-treatment years in this analysis) as developed by the model. This is particularly helpful given that there are reasons to suspect that UN PKO deployments and withdrawals create economic effects beyond the immediate time of their deployment/withdrawal.

The range of cases avaliable in this study are all civil war and post-civil war cases as identified by the Uppsala Conflict Data Program (UCDP)/International Peace Research Institute in Oslo (PRIO) Armed Conflict Dataset, ranging from 1994 to 2007 (Gleditsch et al. 2002, Davies et al. 2022). To qualify as a conflict-state, a country must experience at least 25 battle-related deaths per year. A post-conflict state is subsequently any state that has experienced conflict prior and whose battle-related death count has fallen below the threshold. At any point, a post-conflict state may experience conflict recurrence, in which the 25-deaths threshold is crossed. These observations remain the data set, as they are both conflict and post-conflict cases. To avoid artificial cases of civil war caused by bloody military coups that result in at least 25 deaths, I recode instances of military coups as “non-civil war”. Criteria for coups is acquired from Powell and Thyne (2011). Finally, temporary lulls in conflict may result in the battle deaths threshold dropping below 25 despite conflict remaining ongoing. A consequence of leaving this unaddressed is an incorrectly inflated number of short post-conflict “peace” spells that actually represent short-term lulls in conflict intensity. To correct for this, I recode three-year (or less) spells of “peace” as conflict-level cases.

Because this study seeks to estimate the causal impact of UN PKOs on economic development, I follow standard convention and measure economic development as GDP per capita (log-transformed). Data on GDP per capita is acquired from Fariss et al. (2022) who employ latent variable modeling to account for missing data, a particularly useful feature of their data collection strategy for the purpose of this paper given the already comparatively small *N*.

Information on the location of UN peacekeeping operations is acquired from the Geocoded Peacekeeping Operations (Geo-PKO) Dataset v. 2.0 (Cil et al. 2020) which covers UN PKOs from 1994-2020. The Geo-PKO Dataset is aggregated at the sub-national level. However, I aggregate this data at the country-year level to account for confounding effects where data is aggregated at the country-year level. UN PKOs are operationalized as a dummy where a value of “1” denotes a country-year where a UN PKO is present and a value of “0” denotes a country-year where a UN PKO is not present.

A number of confounding factors complicate the causal relationship between PKOs and development. To isolate a causal effect, it is necessary to account for these confounding factors. First, many authors have argued that the UN has a general aversion to entering armed combat with military forces of the state (Gilligan and Stedman 2003, Fjelde et al. 2018). Likewise, states with stronger military capacity may be able to enforce a greater degree of order over the territories they control, impacting the prospects of order and stability. Such order and stability may be more conducive to economic growth and development. To account for this confounding effect, government capacity is measured as the log-transformed number of military personnel per capita acquired from the Correlates of War National Military Capabilities data set v.5.0 (Singer et al. 1972, Singer 1987). Second, the duration of conflict also effects both the presence of a PKO and prospects for economic development. As the duration of a war increases, the authorization of a PKO may likewise increase as the need for international intervention may be viewed as necessary to terminate conflict. Additionally, the longer war continues, the farther away the process of post-conflict economic recovery is pushed. I measure war duration using the count of years a conflict (as defined by UCDP) has been ongoing. If a conflict is not ongoing, war duration is simply coded as “0”. Third, the size of a potential target country’s population may also influence the likelihood of development and receiving a UN PKO. Many scholars have argued that a connection exists between population growth and economic growth (Headey and Hodge 2009, Prettner and Prskawetz 2010, Peterson 2017). In addition, if the UN is concerned with its post-intervention success, smaller countries may represent a more tempting target as they might be easier to manage and govern. A log-transformed population variable is therefore introduced to account for this confounding effect. Fourth, levels of conflict intensity may impact prospects for development and the chances of a UN PKO being authorized. To measure conflict intensity, I use a log-transformed value of the count of all battle-related and one-sided violence-related deaths. Data on conflict intensity is acquired from UCDP’s Georeferenced Events Dataset (Davies et al. 2022, Sundberg and Melander 2013). Lastly, I use the Variety of Democracy’s (V-Dem) “electoral democracy index” to account for the effect that democracy may have on economic development (Teorell et al. 2019, Coppedge et al. 2022). While the theoretical logic linking democracy to development has not been firmly settled, the empirical linkage between the two *is* well-established (Barro 1996, Przeworski et al. 2000, Acemoglu et al. 2019). Given that there is not much extant research suggesting that UN PKOs should be more likely to intervene in democratic cases, I do not consider democracy to be a confounding factor. Nonetheless, including variables that affect the outcome *but not the treatment* (“neutral” controls) may decrease the variation in the outcome and improve the precision of estimates.

Initially, this short list of confounders may seem inefficient to capture all confounding effects complicating the relationship between PKOs and development. Fortunately, we do not have to blindly wonder if an estimated causal effect from observational data is legitimate or if it is biased by an unspecified confounder. Sensitivity analysis can be employed to quantitatively assess the extent to which unobserved confounding may bias observed results. While the variety of sensitivity analyses is large, I follow the approach of Cinelli and Hazlett (2020). In this approach, a researcher specifies a benchmark covariate. This covariate should be one that is observed in the model and one the researcher believes should have the largest substantive effect among the group of observed covariates. Hypothetical unobserved confounders are then constructed with varying levels of strength relative to the specified benchmark covariate (one times as large as the benchmark covariate, twice as large, three times as large, etc.). Following this, researchers are able to both numerically and graphically assess the extent to which an unobserved confounder of varying strengths has the capacity to reduce or flip a causal estimate. Unfortunately, this method is not directly applicable for matching with panel data. Given the novelty of the panel data matching approach, this is not surprising. Instead of estimating the sensitivity of the estimates generated from the panel data matching analysis, I utilize sensitivity analysis to demonstrate the sensitivity of an estimate using a standard linear regression model with statistical control to account for covariates. While this approach is not applicable for the results of the panel data matching analysis, it is insightful for quantitatively assessing the sensitivity of coefficients using the predominant approach in the literature to estimate effect sizes and directions.

**Results**

Because matching methods rely on a comparable control unit for every treated unit, the number of observations naturally shrink when compared to standard regression-based designs. If treatment is a fairly rare event (such as PKO deployments/withdrawals), then *N* will be, at minimum, twice the size of the number of treated units. If the number of treated units is particularly low, this may be problematic as the subsequent *N* in such a design would likewise be concerningly low. In addition, if treatment is clustered heavily within a certain timeframe or in a certain area, it may be difficult to generalize results beyond the temporal-spatial clusters. To visually assess such concerns, Figure 1 presents a treatment variation plot (Imai et al. 2021) that visualizes the distribution of treatment across time and space (blank spaces represent country-year observations where the country did not exist).

**Figure 1. Treatment Variation Plot**

[See Figure 1.]

A number of observations can be made from a glance at Figure 1. First, treatment (PKO) is clustered in time and in space. It is well-known that UN PKOs as they are conceived of today (large deployments of armed personnel to assist in the implementation of peace agreements and to police post-conflict environments) were practically non-existent prior to the end of the Cold War. Further, UN PKOs are heavily clustered in African countries. Given that the debate surrounding PKOs concerns their *contemporary* effectiveness (rather than a retrospective Cold War-era effectiveness), the temporal clustering of PKOs is a minor issue. However, the clustering of PKOs in African countries is problematic for generalizability given the logic of matching. Recall that in the panel data matching setup, units are matched based on a pre-specified number of years where treatment status is the same and values in covariates are similar. At time *t*,matched sets are allowed to differ based on their treatment status. Given that most treatments occur in Africa, matched sets will primarily consist of *African* countries as African countries share many baseline similarities to one another. While this is not an issue if one wishes to evaluate the effectiveness of PKOs *in Africa*, it is an issue for researchers wishing to make universal claims concerning the effectiveness of PKOs in *conflict-stricken areas*. Because of this, I proceed with the analysis under the assumption that such results generated from this analysis may not be generalizable beyond the post-Cold War African context.

The virtue of matching techniques over statistical control is derived from the former’s ability to more effectively balance covariates. The goal of refining techniques such as nearest-neighbor matching or inverse probability weighting is to construct a new data set where observations are weighted so that units *as similar as possible* as it relates to the pre-specified covariates (but different as it relates to treatment status) are compared to one another. However, refinement is not a guarantee that balance has been achieved. Figures 2 and 3 plot the effectiveness of select refinement methods such as nearest-neighbor matching using Mahalanobis distance (up to 5 and 10 matches, respectively) and inverse probability weighting. In particular, Figure 2 assesses covariate balance when PKO deployments are considered as treatment and Figure 3 assesses covariate balance when PKO withdrawals are considered as treatment. The x-axis reflects the standardized mean difference between covariates for the treated and control units prior to the execution of the specified refinement method. The y-axis reflects the standardized mean difference between covariates for the treated and control units following the execution of the specified refinement method. Each dot in the plot reflects a specific covariate. Dots falling below the 45-degree dashed line suggest an improvement in covariate balance post-refinement. Dots falling above the dashed line imply that the select refinement method *decreased* covariate balance. To reach an ideal level of covariate balance, dots should cluster at the lower end of the y-axis. Balance is assessed at various lag thresholds (one-to-four-year lags) to examine how sensitive balancing is to the criteria established for the generation of matched sets.

**Figure 2. Covariate Balance Post-Refinement for PKO Deployments as Treatment**

[See Figure 2.]

**Figure 3. Covariate Balance Post-Refinement for PKO Withdrawal as Treatment**

[See Figure 3.]

As demonstrated by Figure 2, when PKO deployment is considered as treatment, all three refinement methods generally improve balance across a variety of lag thresholds. However, among these, IPW is much more effective at reducing standardized mean differences between treated and control covariates. However, as Figure 3 establishes, the ability of all three refinement methods to improve balance generally breaks down when PKO withdrawal is considered as treatment. Overall, these results suggest that the refinement methods improve balance when PKO deployments are considered as treatment. However, when PKO withdrawals are considered as treatment, IPW preforms the best at both improving balance *and* reducing standardized mean differences (SMD) of covariates to zero. However, in Figures 2 and 3, covariate balance and increased SMD become a greater consequence as the number of lagged periods to match on increases. In particular, refinement methods appear no better at balancing covariates than non-refinement when PKO withdrawals are considered as treatment for nearest-neighbor refinement. However, in this circumstance, IPW generally does well at improving balance and keeping the SMD closer to zero up until four-lagged time periods as criterion for matched sets. Given this, observations will be matched if they share the same pre-treatment period, up to three time periods.

To account for time-invariant factors creating a baseline difference in outcome between the treated and control units (in addition to allowing this method to project the long-term impact of treatment), this method employs the difference-in-differences (DID) estimator. While a powerful estimator, DID crucially relies on the satisfaction of the parallel trends assumption. That is, prior to treatment, the trend in outcome between eventually-treated and control units was the same. Importantly, this does not require the *values* of the outcome to be the same for treated and control units. Rather, the parallel trends assumption merely requires that the trend in outcome for treated and control units be stable and increasing, decreasing, or remaining stagnant at the same rate. Figure 4 visualizes the trend in the standardized mean difference for log-transformed GDP per capita between treated and control units over three periods pre-treatment. Regardless of the size of the standardized mean difference between treated and control units, a stable line across the three time periods may suggest that the parallel trends assumption holds. Across different treatment specifications and refinement methods, the line remains fairly constant, lending support to the satisfaction of the parallel trends assumption.

**Figure 4. Trend in Log(GDP per capita) Pre-Treatment**

[See Figure 4.]

With improved balance and evidence supporting the satisfaction of the parallel trends assumption, we can now turn to the results of this analysis. However, prior to a discussion of the estimates themselves, it is worth clearly identifying *what* causal effect is being estimated in the first place. It is not possible to make any broad claims concerning the causal effect of UN PKOs if they were applied to any country within the population of all countries globally. That is, it is not possible to answer the question, “what would the average effect of UN PKOs on economic development be if applied to a cases that never received UN PKOs?”. This question is one that can be answered when the average treatment effect (ATE) is estimated. However, the ATE is not estimated because there are a range of cases in which UN PKOs would never intervene in for a variety of observed and unobserved reasons (we cannot extrapolate these results to a country such as Sweden or Japan in 2023) and the entire population of country-year observations is not included in the analysis due to the restrictions of matching. Barring randomization of treatment and the execution of a randomized controlled trial, the estimation of the ATE is not a possibility in this scenario given the issue of unobserved confounding. Instead, the average treatment effect on the treated (ATT) is generated, which can answer the question, “what was the average effect of UN PKOs on economic development *where UN PKOs were deployed*?”

Turning to the results as presented in Figure 5, no statistically significant effect is reported for PKO withdrawals in the eight years following PKO withdrawals, although as expected, the estimate is increasingly negative. This is true for all refinement methods, although, the substantive impact is the largest and the confidence intervals overlap with zero the least when IPW is employed for refinement. Given that, for PKO withdrawals, IPW improved balance the best, these results should be interpreted with the greatest degree of confidence. However, a perhaps more interesting finding are the estimates for the effects of PKO deployments. Across all three refinement methods, covariate balance tended to overwhelmingly improve when PKO deployments were considered as treatment. Although, by far, IPW preformed the best at reducing the SMD closer to zero. Consequentially, results obtained using IPW should be interpreted with more confidence than results obtained using nearest-neighbor matching. For nearest-neighbor matching-derived estimates, the effect of UN PKO deployments on economic development eight years following deployment is statistically significant. Confidence intervals for estimates derived from IPW marginally overlap with zero. Substantively speaking, the average effect of UN PKO deployments for units that received UN PKO deployments after eight years is a 34% (nearest-neighbor matching, up to 5 units), 28% (nearest-neighbor matching, up to 10 units), and a 27% (IPW) decrease in GDP per capita. These findings stand in stark contrast to prior research suggesting a positive relationship between PKOs and economic conditions (Caruso et al. 2017, Beber et al. 2019, Bove et al. 2021). In contrast, these results are well in line with other findings employing quasi-experimental designs where the economic effect of UN PKOs is largely negligible or even negative (Bove and Elia 2017, Mvukiyehe and Samii 2020). In the following paragraphs, I discuss what might be driving these contrasting findings.

**Figure 5. Estimated Average Effects of PKO Deployments and Withdrawals on Economic Development**

[See Figure 5.]

One clear explanation for the wide variation in findings is the wide variety of samples, methods, operationalizations of outcome, and levels of aggregation used in quasi-experimental designs seeking to isolate the economic effects of PKOs. Samples range from country-year level analyses (Bove and Elia 2017, Beber et al. 2019) to sub-national coverage including South Sudanese counties (Caruso et al. 2017) and household-level survey data in South Sudan (Bove et al. 2021) and Liberia (Beber et al. 2019, Mvukiyehe and Samii 2020). Quasi-experimental methods include matching (Beber et al. 2019, Mvukiyehe and Samii 2020), instrumental variables (Caruso et al. 2017, Bove et al. 2021), fixed effects (Beber et al. 2019, Mvukiyehe and Samii 2020, Bove et al. 2021), the synthetic control method (Bove and Elia 2017), and mediation analysis (Bove et al. 2021). In these studies, economic development has been operationalized as GDP per capita (Bove and Elia 2017), GDP per capita growth (Beber et al. 2019), cereal production (Caruso et al. 2017), and various household-level survey indicators of development and economic well-being (Beber et al. 2019, Mvukiyehe and Samii 2020, Bove et al. 2021). Like any emerging literature, more work simply needs to be done. The number of quasi-experimental designs studying the economic effects of PKOs is simply too small and too diverse in methods and data to create any declarative statements. Future research should experiment with a wide range of methods, data, and operationalizations of development to determine if any bias exists dependent on the selection of certain methods, data, etc.

In addition to the broad inconsistency in findings within existing literature, there are likewise a number of issues with this research design that, if improved, may generate alternative results. First, not all PKOs are the same. Mission goals, troop deployment, police deployment, and a host of various mission-specific factors will vary heavily across different missions. Treating all PKOs as a uniform binary treatment is convenient for the purposes of executing quasi-experimental designs, but it is not ideal for estimating an exact ATT. Future studies should incorporate the variation in treatment within research designs. While this does complicate the development of quasi-experimental designs, it is likely a necessary step in the right direction. Second, a brief return to the stable unit treatment value assumption (SUTVA) is warranted. Treatment spillover effects between countries remains a serious threat to causal inference. In particular, researchers extending quasi-experimental designs to the study of the potentially conflict-reducing effects of PKOs should proceed with great caution given the contagious nature of conflict and the geographically clustered nature of countries that receive UN PKOs.

As noted in the research design, extant sensitivity analyses are not currently supported for the panel data matching design as implemented by Imai et al. (2021). Rather than estimating the sensitivity of estimates from the matching analysis executed in this paper, I estimate the sensitivity of the PKO coefficient generated from a standard linear regression model. Importantly, the results of this sensitivity analysis *are not* generalizable to the analysis conducted in this paper. However, I choose to execute this analysis to demonstrate the sensitivity of coefficients in research designs that are much more common in the PKO literature than the analysis conducted for this paper. The execution of this sensitivity analysis is intended to demonstrate the issue of unobserved confounding in standard research designs using statistical control. For the following sensitivity analysis, I estimate a linear regression model where logged GDP per capita is regressed on a dummy variable indicating the presence of a PKO, logged population size, logged count of deaths (battle-related and one-sided violence), logged military personnel per capita, war duration, and democracy, measured using V-Dem’s electoral democracy index. The results of the sensitivity analysis are presented in Figure 6.

**Figure 6. Sensitivity Analysis for Unobserved Confounders Using Standard Statistical Control**

[See Figure 6.]

Because the size of a government’s military capability likely has a great degree of influence on the presence of a UN PKO (the UN may be hesitant to engage/stay in a conflict where the government is more capable) and the prospects for political stability and development, logged military personnel per capita (MPpc) is selected as a benchmark covariate. Simply put, this means that unobserved confounders are interpreted in reference to the size of the relationship between economic development and MPpc. The coefficient for PKO as determined by the linear regression model is -0.11 (the unadjusted black triangle in Figure 6). If an unobserved confounder with a relationship with GDP per capita 0.25x as strong as the relationship between MPpc and GDP per capita was specified in the linear regression model, the PKO coefficient would increase to -0.05. If an unobserved confounder was 0.5x as strong, the sign on the coefficient for PKO would *flip* (0.013), demonstrating a positive association between PKOs and economic development. To meaningfully interpret these results, two items must be considered. First, another researcher may easily be able to mention and include another confounder that is not specified in the regression model. An inclusion of such variables could easily alter the coefficient for PKO. Indeed, the results of the sensitivity analysis suggests the possibility of this is high, assuming an unspecified confounder at least 0.5x as strong as MPpc is plausible. This assumption leads to the second item to consider. Is an unspecified confounder 0.5x as strong as MPpc a realistic possibility? No statistic can answer this question. From here, expert domain knowledge, avaliable theoretical literature, and awareness of data availability must inform the interpretation of the plausibility of such a confounder.

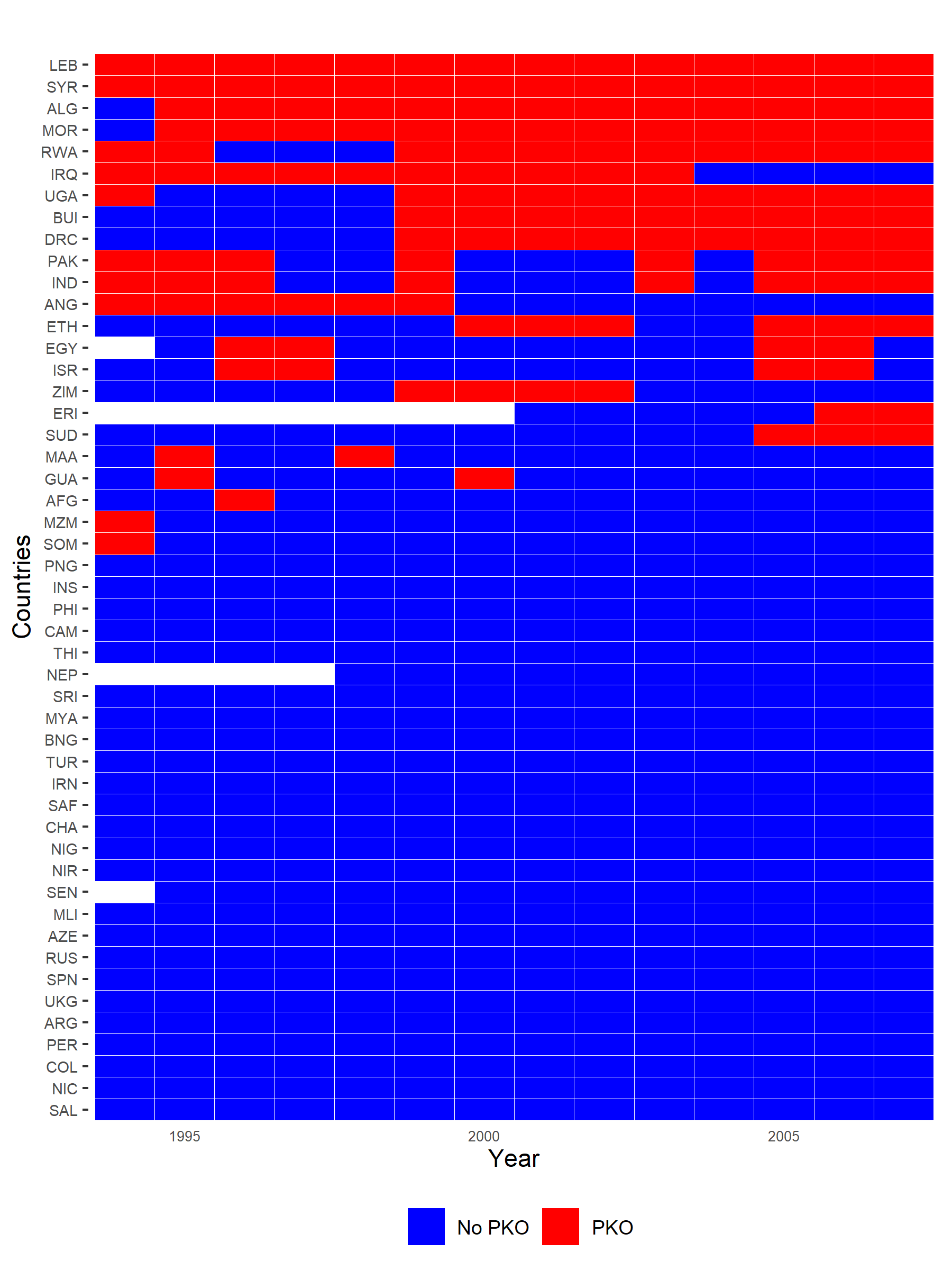
**Conclusion**

Traditionally, research on peacekeeping has been devoted to understanding its potentially pacifying effects. However, an emerging theoretical and empirical literature has likewise extended the study of peacekeeping to examine the potential economic effects of peacekeeping operations. Existing quantitative research is not clear, however, on whether UN PKOs have a substantive effect, negative or positive, on economic development. Employing a quasi-experimental design using conventional matching and weighting techniques extended to a panel data framework, the analysis in this paper compliments findings from other scholars employing quasi-experimental designs who generally fail to find a consistent, positive effect of UN PKOs on economic development. Further, in contrast to contemporary theory, PKO deployments were found to create a substantively negative impact in host country economies. Although, a number of methodological concerns are present that warrant placing asterisks next to these findings.

Beyond the empirical findings of this paper, I argue that a broader shift towards making causal inferences is necessary in the various literatures that comprise the study of conflict management. The study of conflict management lends itself well to the language of causality and quasi-experimental design given that scholars are often interested in some type of conflict management strategy which often manifests as a tangible event or program (PKOs, peace agreements, foreign aid packages, post-conflict justice processes, etc.) that can be considered a treatment. Further, given the high-stakes nature of conflict management research, scholars are generally interested in whether conflict management strategies *work*, rather than whether conflict management strategies are *correlated* or *associated* with success. As demonstrated in this paper, scholars studying conflict management can easily adapt causal inference tools such as DAGs, various quasi-experimental design methodologies, and sensitivity analysis.

Despite arguments laid out in this paper stressing a need to shift towards causal research, many practical concerns should be considered and addressed before scholars can place high levels of confidence in causal estimates generated from observational data. First, issues of data validity are equally as problematic for causal research as they are for correlational research. A causal *estimate* will not be a legitimate reflection of the *estimand* if the data it is estimated from does not reflect reality. Given that many outcomes of interest in the conflict management literature rely on data that is difficult to precisely gather (such as battle-related deaths), this remains an issue. Second, clusters in treatment potentially render generalizability of treatment effects low, especially when treatments such as UN PKOs are examined, which are highly temporally and spatially clustered. Third, spatial clustering of treatment presents a legitimate threat to causal inference in the conflict management disciplines as treatment oftentimes has the capacity to spill over into control units, violating SUTVA. If “control” units are affected by the treatment of treated units, then the credibility of these control units to serve as legitimate control units is diminished. Lastly, when working with observational data, scholars are never able to completely satisfy the ignorability assumption, regardless of how intentional a DAG was constructed. The possibility of an unknown confounder is always present and there is no way to assess whether *all* confounders have been identified and adjusted for. Instead, scholars using statistical control and quasi-experimental design alike should employ sensitivity analysis to quantitatively assess the potential impact of unobserved confounders. Of course, a push for sensitivity analysis is also a push for greater theorizing as the plausibility of an unknown confounder of an arbitrary magnitude will only be helpfully informed by greater theory and familiarity with the treatment and outcome being studied.

**Figure 1. Treatment Variation Plot (1994-2007)**



**Figure 2. Covariate Balance Post-Refinement for PKO Deployments as Treatment**

Chart, scatter chart

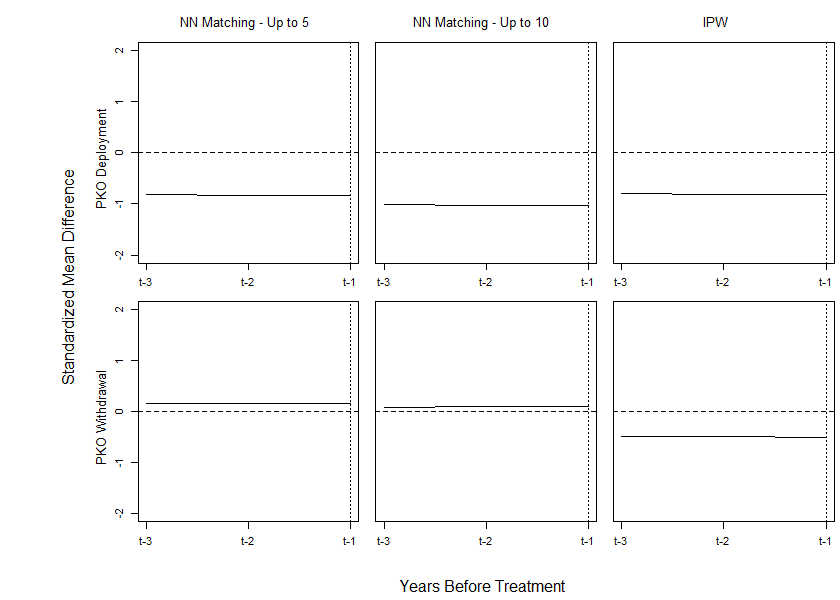
Description automatically generated

**Figure 3. Covariate Balance Post-Refinement for PKO Withdrawal as Treatment**

Chart, scatter chart

Description automatically generated

**Figure 4. Trend in Log(GDP per capita) Pre-Treatment**



**Figure 5. Estimated Average Effects of PKO Deployments and Withdrawals on Economic Development**

Chart, box and whisker chart

Description automatically generated

**Figure 6. Sensitivity Analysis for Unobserved Confounders Using Standard Statistical Control**

Chart

Description automatically generated

1. Making causal inferences requires other assumptions as well. However, these assumptions overlap with assumptions researchers familiar with regression and statistical inference should also be familiar with, such as the requirement of a sufficiently large sample size, the reduction of measurement error, and positivity (treatment varies within each possible combination of conditioned variables). As a result, I do not go into detail on these assumptions in this paper. [↑](#footnote-ref-1)
2. However, it is not uncommon to manually remove select observations if their generated weights are extreme. This occurs when a single non-treated observation, according to the propensity scores, was *very* likely to have been treated and vice versa (typically where the propensity score is .95 for a non-treated observation or .05 for a treated observation). [↑](#footnote-ref-2)
3. Assuming the parallel trends assumption holds (this will be discussed in the results section), the DID estimator naturally accounts for all unobserved time-invariant confounding. [↑](#footnote-ref-3)