HiCN Households in Conflict Network
hicn.org

How to conduct impact evaluations in humanitarian and conflict settings

Aysegül Kayaoglu[1,2,3], Ghassan Baliki[1], Melodie Al Daccache[4], Dorothee Weiffen[1], Tilman Brück[1,5,6]

HiCN Working Paper 387

March 2023

Keywords: Impact evaluation, research design, machine learning, conflict setting, humanitarian emergencies

JEL classification: C18, C30, C80, D04, D74, Q34

Abstract
Methodological, ethical and practical challenges make it difficult to use experimental and rigorous quasi-experimental approaches to conduct impact evaluations in humanitarian emergencies and conflict settings (HECS). This paper discusses recent developments in the design, measurement, data and analysis of impact evaluations that can overcome these challenges and provide concrete examples from our recent research where we analyse the impact of agricultural emergency interventions in post-war Syria. More specifically, the paper offers solutions: First, discuss the challenges in designing rapid and rigorous impact evaluations in HECS. By doing so, we mainly show alternative ways to construct counterfactuals in the absence of meaningful control groups; Second, we review how researchers can use additional data sources to create a counterfactual or even data on treated units when it is difficult to collect data and in some cases provide ethical and methodological benefits in addition to providing cost-effectiveness. Third, we argue that finding and fine-tuning proxy measures for the 'unmeasurable' concepts and outcomes such as resilience and fragility are crucial. Fourth, we highlight how adaptive machine learning algorithms are helpful in rigorous impact evaluations in HECS to overcome the drawbacks related to data availability and heterogeneity analysis. We

[1] ISDC - International Security and Development Center, Germany
[2] Department of Economics, Istanbul Technical University, Türkiye
[3] IMIS, University of Osnabrück, Germany
[4] American University of Beirut, Lebanon
[5] Thaer-Institute, Humboldt-University of Berlin, Germany
[6] Leibniz Institute of Vegetable and Ornamental Crops (IGZ), Germany

provide an example from our recent work where we use honest causal forest estimation to test the heterogeneous impact of an agricultural intervention when sample sizes are small. Fifth, we discuss how standardisation across methods, data and measures ensures the external validity and transferability of the evidence to other complex settings where impact evaluation is challenging to conduct. Finally, the paper recommends how future research and policy can adapt these tools to ensure significant and effective learning in conflict-affected and humanitarian settings.

Funding
This work is supported by the Centre of Excellence for Development Impact and Learning (CEDIL), funded by UK Aid.

Conflict of interest
The authors declare no conflict of interest.

1. Introduction

In 2022, an estimated 302 million individuals need humanitarian assistance worldwide, of which 80% are from conflict-affected settings (UNOCHA, 2022; World Bank, 2022). These figures underscore the global magnitude of the threat to lives and livelihoods caused by conflict and other humanitarian emergencies. The international community estimates that it requires USD 46.06 billion for humanitarian assistance only for the targeted 202 million of those 302 million people in need, which is seldom met (UNOCHA, 2022). Hence, in a context where millions live in such fragile settings and resources to help them fall far shy, efficient and impactful humanitarian assistance becomes more important than ever.

Impact evaluation is "an assessment of how the intervention being evaluated affects outcomes, whether these effects are intended or unintended" (OECD, 2006:1). Beyond monitoring dashboards and ex-post assessments, which have been used extensively in the humanitarian sector in the past two decades, rigorous and theory-based impact evaluations are a simple yet powerful tool to measure if and how humanitarian assistance or aid can causally contribute to achieving targeted outcomes of a pre-specified theory of change (White, 2009). According to the impact evaluation repository of the International Initiative for Impact Evaluation (3ie), 10,374 impact evaluations were completed between 1990 and 2021. And only 7.4% of them were conducted in fragile and conflict-affected countries where 2 billion people live (United Nations, 2022). Besides, there are essential evidence gaps across sectors, as seen in ***Table 1***.

That said, impact evaluations have been well-established and used for development programming in many contexts, yet they remain limited in the Humanitarian Emergency and Conflict Settings (HECS) due to a myriad of methodological, ethical, and practical challenges such as selection bias, information bias, contamination bias, non-random attrition and response, need for rapid evaluations, attribution problem, and (un)intentionally harming vulnerable populations (Puri et al., 2017). Although methodological challenges can be overcome using randomisation in assigning treatment

and control groups, using Randomised Control Trials (RCTs) is particularly difficult in HECS due to security, ethical, financial and political reasons. Therefore, observational data and quasi-experimental designs are preferred to establish a better counterfactual without randomisation and, thus, estimate the causal impact. Recent years have presented enormous improvements in creating and improving these tools to identify a counterfactual, even without randomization. In addition, regression discontinuity designs, difference-in-differences method, synthetic controls, and machine learning techniques have strengthened their positions in the field of causal inference.

Table 1. *Sectoral Distribution of Impact Evaluations Conducted in Fragile and Conflict-affected Countries by the year of publication*

Sector	***1990-2000***	***2001-2010***	***2011-2021***	***Total***
Agriculture, fishing and forestry	0	5	86	91
Education	0	5	46	51
Energy and extractives	0	0	7	7
Financial Sector	0	1	23	24
Health	15	48	311	374
Industry, trade and services	0	1	15	16
Information and communication technologies	0	0	5	5
Social protection	0	8	100	108
Public administration	0	1	58	59
Transportation	0	0	2	2
Water, sanitation and waste management	0	5	26	31

Notes: *Authors' calculations using the impact evaluation repository of 3ie.*

Puri et al. (2017) asked the question of whether rigorous impact evaluations can improve humanitarian assistance or not. Revisiting this question five years later, this paper presents if there was indeed an improvement in the area of rigorous impact evaluations in

the HECS and, if not, what are the main challenges for that. Therefore, in this paper, we review, synthesise, and present the knowledge, insights and evidence generated from two distinct streams – the existing literature and our own experiences of conducting impact evaluations in HECS. As seen from the discussion below, we argue that it is possible and necessary to conduct rigorous impact evaluations in HECS, which are still at deficient levels compared to other settings.

That said, rapid and rigorous impact evaluations in HECS can be designed, for example, through utilising innovative methods to construct a counterfactual group, being flexible to use methods that are suitable in settings where multiple actors actively provide similar interventions in the same field and at the same time as in emergency situations; having an improved collaboration of different stakeholders during different phases of impact evaluations; employing mixed methods more and effectively; and, generating and improving the inclusivity of publication processes to motivate future researchers to conduct the effective and efficient impact evaluation methodology. Moreover, we also argue that big data sources such as remote-sensing, geo-spatial and administrative data should be utilised more, not as a substitute but as a complement to the 'traditional' data sources. In addition to design and data-related issues, it is also essential to have fine-tuning measurements so that practitioners and researchers can use simple but standardised measures to compare different settings, which is particularly important for HECS. Lastly, machine learning algorithms can be utilised to construct a well-defined counterfactual, which is arduous to manage in orthodox impact evaluation methods such as RCTs, especially due to the ethical challenges of conducting impact evaluations in HECS.

The structure of the paper is designed around four themes. Section 2 reviews the current pitfalls and developments in designing rapid and rigorous impact evaluations. Section 3 focuses on data and discusses the possible ways to generate and use novel data to conduct impact evaluations in HECS. Section 4 explains the possibility and necessity of fine-tuning measurements, while Section 5 provides examples of how researchers can benefit from machine learning tools in impact evaluations.

2. Designing Rapid and Rigorous Impact Evaluations

Expanding Buttenheim (2009)'s framework for the HECS, Puri et al. (2007) list three phases after the time of emergency where impact evaluations can be used. Those are the relief, recovery and resilience phases, where impact evaluations present specific practical, ethical and technical issues to consider, as project implementers might have different goals in each. For example, urgent recovery from a disaster is indisputably key in the relief phase, which necessitates a very rapid and leave-no-one-behind type of assistance, such as providing latrines to all of the camp residents in Bentiu, South Sudan, after a natural disaster. That is why randomised controlled trials (RCTs) – otherwise regarded as the gold standard among all impact evaluation methods and "now entirely dominate development economics"[1] -- cannot be applied in such a setting because the random allocation of humanitarian assistance and leaving some emergency-affected populations as a control group are both ethically and practically unfeasible. Therefore, flexibility and willingness to use innovative data and methodologies to overcome complex challenges to a rigorous impact evaluation in different phases in HECS are important.

This implies that researchers must use other quasi-experimental methods with **a well-defined counterfactual** necessary for a rigorous impact evaluation. The main idea is to construct or find a control group that will not receive the intervention but needs to be structurally similar to the treatment group both in trends and preferably in levels during the pre-treatment period and, by assumption, also in the post-intervention in the absence of a treatment. However, various ethical, practical and methodological challenges exist to construct a well-defined counterfactual in HECS. A crucial ethical challenge is, for example, the difficulty of implementing an intervention considering the do-no-harm principle since being pure counterfactual implies that no assistance will be received. Moreover, the high mobility rates of conflict-affected persons and a lack of registration or

[1] Press release: The Prize in Economic Sciences 2019. Available at: https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/ (published 14 October 2019).

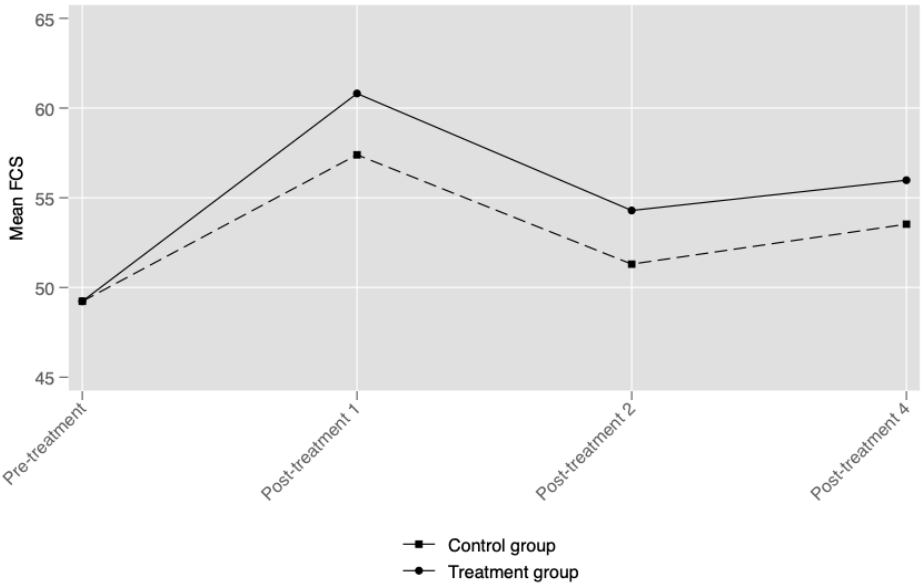
administrative data in HECS make longitudinal or representative data collection almost impossible. This complicates the application of several of the key quasi-experimental procedures in these circumstances, such as difference-in-differences with a never-treated group. Furthermore, the presence of various stakeholders in such settings and the need for quick and widespread interventions in HECS makes it hard, for example, to evaluate the causal effect of programs using RCT.

However, there are four potential remedies to the challenge of having a counterfactual in HECS. All these suggestions provide alternative ways to construct a well-defined control group instead of a traditional one. Firstly, **factorial designs** can be particularly beneficial in cases where ethical concerns are more pivotal, such as those implemented during the emergency phase. In such designs, the question of *which* intervention has more impact rather than *what* the effect of an intervention is more relevant. Answering this question is particularly crucial in the early phases of an intervention, where many vulnerable individuals wait for humanitarian assistance. Thus, such an impact evaluation would help implementers to use the most effective and efficient tools in the field once the research outcomes are rapidly obtained. Secondly, **phased-in designs** can be implemented **after the emergency relief phase**. In such circumstances, households are included in the intervention in a staggered roll-out base, which implies that late-comers act as a control group for the early-treated units. Thus, this is 'more' ethical than a purely randomised experiment once the staggered groups of treatment receivers are determined randomly and, although, in different phases, everyone eventually receives the treatment. This way of constructing counterfactuals is also preferable, considering the usual budget constraints of implementers. Moreover, **block randomization** can be incorporated in phased-in designs because it is a valuable tool to fairly allocate participants to receiving treatment in a given study and thus reduces local tensions and prevents spillover effects (Peter and Soetevent, 2019). The third way of creating a valid counterfactual is using **synthetic control groups**. It is a recent method but has received considerable attention since its first use in 2003. It is even branded as "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey and Imbens, 2017:3). This

method uses pre-intervention information to produce an optimally estimated control group (called a “synthetic” control group) by assigning weights to statistically chosen units (i.e. households) based on their similarity to treatment units. However, this approach works best if the observations are collected over a long period, both pre- and post-intervention, and requires a large number of control observations, which might not be feasible or available in many HECS settings. Finally, unpacking the potential power of **spatially disaggregated data or geographic information databases (GIS)** can help researchers to construct valid counterfactuals in projects where physiographic characteristics are crucial for treatment. They also have a tremendous potential to create instrumental variables (Puri et al. 2017) and enable using spatial RDD designs or geographic difference-in-discontinuities.

Moreover, even in rare cases where a never-treated control group is available in the data, targeting in HECS might result in statistically significant imbalances between treated and control units. The standard matching techniques to balance the distribution across these two groups might still cause a biased estimation in those cases. Recently developed methods such as **entropy balancing** (Hainmueller, 2012) or **hierarchically regularised entropy balancing** (Xu & Young, 2021) can be used as they enforce the balance across moments of distributions. Figure 1 below is from Kayaoglu, Baliki and Brück (2023), which compares the trends between control and treatment units where the entropy-weighted pre-treatment outcome in both groups, on average, is the same. This graph clearly shows how the impact of an agricultural intervention in post-war Syria affects households' food consumption scores (FCS) in the long term once their initial food security score, among other baseline covariates, are the same. In their analysis, they used entropy balancing with one period of pre-treatment data to estimate the weights that balance the distribution of both main socio-demographic characteristics of households and their food consumption scores. Then these entropy weights are used to estimate the average treatment effect. These methods can even be used to balance the observable treatment-invariant characteristics across groups in cases where pre-treatment data is unavailable.

Figure 1. Trends between treated and untreated groups using the Entropy Balancing



Source: Kayaoglu, Baliki and Brück (2023)

Moreover, in cases where researchers have pre-treatment data, then other available methods can be used to create valid counterfactuals. One of these important methods is the **kernel propensity score matching DiD,** where we can first generate kernel weights by estimating the propensity scores following Heckman et al. (1997, 1998), which are then used in the DiD estimations. Kayaoglu, Baliki and Brück (2023), for example, also use the kernel propensity score matching DiD method to analyse the long-term impact of an agricultural intervention in Syria and show that correcting the imbalances between treatment and control units through kernel weights eliminates the bias. In cases where more than one period of pre-treatment data is available, researchers can also use the recently developed **synthetic DiD** (Arkhangelsky et al., 2021) to estimate the average treatment impact of interventions.

In addition to the issues related to not having traditional counterfactuals, when designing impact evaluation in HECS, **power analyses and oversampling** are key to preventing

underpowered studies resulting from unprecedented decreases in sample size through attrition, lack of consent, or inaccessibility of the field. Underpowered studies are either unable to detect the effect or present a biased estimation even in cases where it reveals one (Rosch et al., 2021). Moreover, model specifications such as interaction terms and investigation of impact channels through heterogeneity analysis should be considered during the design phase of an evaluation to prevent unforeseen drops in statistical power.

Particularly in complex settings, designs often need to be adjusted in later stages for several reasons, such as to remain ethically acceptable and statistically efficient. The use of **adaptive strategies** allows researchers to use preliminary findings to change the allocation of participants across groups without compromising statistical power. This strategy is also ethically desirable, given that researchers can discontinue ineffective treatments while extending the coverage of the effective ones. Besides, it helps researchers understand how the interventions work along the causal chains or for particular groups, allowing for better resource allocation and withholding the use of unnecessary procedures. Despite the importance of this method, adaptive trials are rarely used due to the need for more expertise and familiarity among researchers and funders (Masset et al., 2021).

Another important challenge in conducting rigorous impact evaluations in HECS is delineating the treatment while a multiplicity of actors is in the field. When Holland (1986) said “no causation without manipulation”, he implied that a well-defined (one without vagueness) intervention is needed for an impact to be **attributed** to a specific treatment. However, it is challenging to attribute an estimated impact to a specific intervention when **multiple actors** often implement their interventions simultaneously. In such a case, coordination among different actors, particularly donor coordination, is vital, which might lead to using factorial designs in impact evaluation. In addition, concerning the problem with statistical power noted above, such contamination needs large sample sizes so that researchers can distinguish the effects of a single intervention from those of several comparable interventions.

Furthermore, the transportability of the impact is of paramount concern in HECS because people's behaviour in HECS differs systematically from peaceful settings (Verwimp et al., 2019). Therefore, the rich evidence from rigorous impact evaluation from peaceful settings cannot be directly transferred to HECS. This implies that researchers should find the most appropriate method to conduct rigorous impact evaluations in HECS and help build up the evidence pool for these distinct settings where billions of people continue to live.

Moreover, continuous, delayed and concurrent interventions do not have a clear baseline or endline and mostly lack a clear control group. Researchers tend to assess the impact of a multi-component intervention as a whole intervention at the end of the program. This approach is helpful for accountability purposes but fails to identify the most effective component or combination of components. For this reason, the results cannot be extrapolated to other similar contexts. A potential solution is to run **multisite trials** where the intervention is implemented in different contexts and simultaneously, which allows summarising the results using meta-analyses. However, using such a design is uncommon because of the high costs, the existence of unique aspects of humanitarian emergency and conflict settings, and the difficulty in standardising the measures. A similar approach is the **Metaketa initiative** which aims to address the question of policy importance by coordinating a cluster of field experiments implemented independently (Dunning et al., 2019; Hartman and Kern, 2020).

A standardised, flexible, and open-source design that incorporates all the important elements for rigorous yet feasible impact evaluation is critical to estimate the causal effects of interventions (WHO, 2017). As resources are scarce and there is an immense need for speedy and impactful interventions in HECS, researchers and implementers must work closely in all phases of impact evaluations to make instant improvements to an ongoing assistance program through this embeddedness approach. This implies that even the results of short-term impact evaluations can be used to improve project implementations in the near future. Moreover, engaging program implementers in different stages of impact evaluation might also increase their awareness of the benefits

of conducting a rigorous impact evaluation in HECS and their willingness to include researchers in the project's risk management and security procedures.

Since a unique aspect of HECS is the possibility of encountering drastic changes in the circumstances of impact evaluation design, it is also helpful to include an emergency response preparedness (ERP) plan during the research design, which highlights practical aspects to help implementers and researchers systematically and collaboratively respond to unexpected shocks (OCHA, 2019). This off-the-shelf design template would allow reproducibility and comparability across similar programs from different settings, which is key to increasing the number and enhancing the quality of meta-analysis and systematic reviews currently available.

Moreover, researchers have limited access to information about control villages and households and, most importantly, about other interventions from other organisations (which are usually geographically concentrated in HECS). Therefore, a web portal of interventions that summarises each intervention's characteristics, such as the implementing organisation, inputs, outputs, time frame, and geographical coverage, can be significant not only to developing better research designs in HECS but also to improving knowledge transfer between implementers and to cultivate collaborations in such complex settings. This kind of a portal can also feature information if any intervention had an impact evaluation and transparently present its outcome to the wider audience.

Another important issue is the role and importance of mixed methods in conducting rigorous impact evaluations in HECS. Most researchers specialise in one methodological approach and often use only quantitative or qualitative methods to track changes in conflict settings, missing out on all the additional analytical strength **that could be obtained from a mixed research inference**. This is primarily the result of the inability to conduct quantitative and qualitative data collection simultaneously due to logistic reasons, high costs, security, legal restrictions and the overall sensitivity of the subjects to be discussed. In addition, HECS is different from (more) stable contexts as it has its

specific risks, vulnerabilities and unpredictabilities, making it difficult to benefit from the evidence generated in stable settings (Blanchet et al. 2017).

Yet, on the one hand, the mere use of quantitative methods fails to provide a complete understanding of the complex nature of the problems, potential solutions, and unintended consequences. On the other hand, small-N studies have various disadvantages when used alone, such as not being representative of the population studied, making it impossible to draw conclusions for the impact evaluation. However, they can be instrumental when complementing the quantitative methods. Using narratives to complement quantitative research enables researchers to pinpoint the conditions for success and build evidence about what does or does not work in an intervention (Wood, 2021). In other words, qualitative methods can potentially suggest mechanisms and help derive an (implicit) theory of change, while quantitative impact evaluation methods can identify the average impacts of the interventions. Employing advanced mixed methods approaches can, therefore, expand and triangulate research findings, particularly in complex, real-world interactions. This necessity is another reason why strengthening the capacity of local researchers and organisations is critical for an efficient and effective impact evaluation in HECS. However, additional ethical issues might arise if local researchers and organisations do not represent all sides of the population in conflict-affected regions. Thus, preserving this neutrality is essential in deciding with whom to collaborate in those settings.

Furthermore, as mentioned above, a counterfactual-based identification strategy is critical for impact evaluations. Researchers need to be creative to overcome many methodological, practical and ethical challenges during impact evaluations in HECS. However, peer-reviewed journals' review processes are biased towards RCTs, which is usually not preferable in HECS. Using the 3ie repository data, Ravallion (2018) shows that 60% of all impact evaluations after 2000 used RCTs. This domination of RCTs leads to evidence gaps and biases in topics decisive for the well-being of the most vulnerable people in the world. According to the impact evaluation repository of 3ie, around 65% of all published impact evaluations in fragile and conflict-affected countries employ RCTs. This apparent tendency to use RCTs implies that relief phase or even recovery phase

interventions are not preferably evaluated as they are. In many cases, it is not feasible to have RCTs in those stages of an emergency. This mismatch between real conditions in HECS and journal requirements creates a crucial evidence gap and makes it more difficult for implementers and policymakers to learn from impact evaluations and for researchers to innovate new methods and designs. Therefore, it is crucial for peer-reviewed journals to be more inclusive of sub-optimal designs other than RCTs, which are ethically more suitable given the context it analyses. Besides, these analyses can imply important contributions to the literature, knowing that there are crucial knowledge gaps in HECS, as summarised in Table 1. This inclusivity of the publication process will also motivate and train future researchers to innovatively adapt methods and designs more suitable to HECS. This would also result in meticulous designs over time.

Finally, many ethical challenges arise in conflict settings, specifically in the design, conduct of research and fieldwork. To overcome these concerns, the **'do no harm'** approach must be fundamental in emergency response preparedness plans (Anderson, 1999). Ethically responsible research includes complete transparency in all stages, the consideration of the trade-off between learning and ethics, informed consent to respect confidentiality, ensuring privacy, and collaborative partnerships with local stakeholders. However, researchers must also remember that respondents might be unable to give 'actual' informed consent if they fear jeopardising any assistance they already receive or hope to receive in the future. Moreover, legal responsibilities and getting approval from local authorities are also crucial for impact evaluations in fragile settings. Lastly, some ethical considerations arise, such as the lack of targeting criteria in identifying the population most in need and the exclusion of the control group from humanitarian assistance, at least for the duration of the study.

We argue that complete transparency is key in the research design and program implementation as a solution to these ethical concerns. And as mentioned above, if there is a need to use a phased-in or staggered roll-out design, then block randomisation can be used not to exacerbate the disagreements between different conflict-affected population

groups inhabiting the same region. Moreover, collaborations with the actors in the field will also be vital to reducing ethical concerns related to control groups.

3.Generating and Using Novel Data: Data Hubris or Data Supplement?

High-quality data is key for a precise rigorous impact evaluation. However, conventional face-to-face survey data collection in humanitarian emergencies and conflict settings is prone to enumerator, selection, contamination, attrition and information biases, and recall or response errors (Brück et al., 2016, Puri et al., 2017). In other instances, face-to-face data collection is even impossible due to safety risks emerging from heightened conflicts or public health crises. Alternative sources of data such as online or phone surveys (Stojetz et al., 2022), demographic data (Corsi, 2012), administrative data (Altındağ et al., 2021), crowdsourcing (Baliki, 2017), open access data such as ACLED (Raleigh et al., 2010), geospatial, satellite and remote-sensing data (Breunig et al., 2020) social media data (Anson et al. 2017), virtual communication/telecommunication tools (Van der Windt and Humphreys, 2016) and data footprints (Shiells et al., 2020; Aiken et al. 2022) are powerful tools where conventional household survey data collection is impeded or difficult to conduct. In the following, we discuss a selection of these data sources and their usefulness in analysing impacts in HECS.

High-frequency phone and online surveys have gained attention in the past three years during the Covid-19 and have been used extensively since (Stojetz et al., 2022; Gouraly et al., 2021). There are advantages and disadvantages to using remote survey data collection methods (Brück & Regessa, 2022; Hensen, et al 2021). On the one hand, remote survey data is more affordable, is rapidly and easily set up compared to face-to-face survey data collection, and it enables researchers to access valuable information from households living in challenging and hard-to-reach areas during acute phases of emergencies. On the other hand, remote data collection might discriminate against vulnerable groups who do not have access to communication tools like mobiles or the internet have regular access to

information networks, which increases self-selection bias. The literature on survey methodology provides evidence of selection bias stemming from phone ownership in low-income countries. For example, comparing different data collection modes in Nigeria, Lau et al. (2019) show that mobile phone data had a significantly lower representation of women, older people, less educated and rural population. Another option is to combine multiple sources of conventional and modern survey data carefully. For example, researchers and practitioners can use administrative data to sample households in an impact evaluation study and collect baseline data using face-to-face interviews where sufficient information on the households (e.g., mobile number) is collected. Follow-up surveys can be then done more cost-effectively and frequently to revisit households at different times during the programme implementation or after specific events or shocks. A shorter, concise version of the questionnaire can be used to minimise non-response rates over the phone but designed such that it can be merged and compared to the baseline or the main administrative survey data.

Using geospatial data is another promising solution to access information on communities and households living in challenging settings in the absence of conventional survey data. Low-resolution satellite imagery is readily and publicly often accessible but not sufficient to provide accurate, precise, and helpful information for assessing impact. While high-resolution satellite data provides accurate remote-sensed information, it is still more costly, particularly if the area of interest is large and the researchers require multiple images of the exact location to conduct time-series analysis. To overcome the financial constraints, expensive high-resolution Earth observation data[2] can be used to train machine learning algorithms to use the free low-resolution data to generate larger datasets for the outcome of interest. Earth observation data are increasingly popular in development economics and social sciences, particularly once combined with machine learning algorithms. Available high-value data with gaps can be used to train and then predict target data (Paul et al., 2018) and single encoding of satellite images can overcome

[2] Earth observation is defined as "the gathering of information about planet Earth's physical, chemical and biological systems via remote sensing technologies, usually involving satellites carrying imaging devices" by the EU Scientific Commission.
Please see https://joint-research-centre.ec.europa.eu/scientific-activities-z/earth-observation_en for details.

the limitations of accessibility and use of satellite imagery with machine learning (Rolf et al., 2021). Furthermore, remote-sensed data is used to measure outcome variables or key independent variables such as agricultural mapping and monitoring (Behrer and Lobell, 2022), or even to predict compliance to the treatment (Jack et al., 2022). Thus, machine learning provides not only a methodological solution for conducting impact evaluations in HECS but also facilitates and increases the utilisation of satellite imaginary data for impact evaluation.

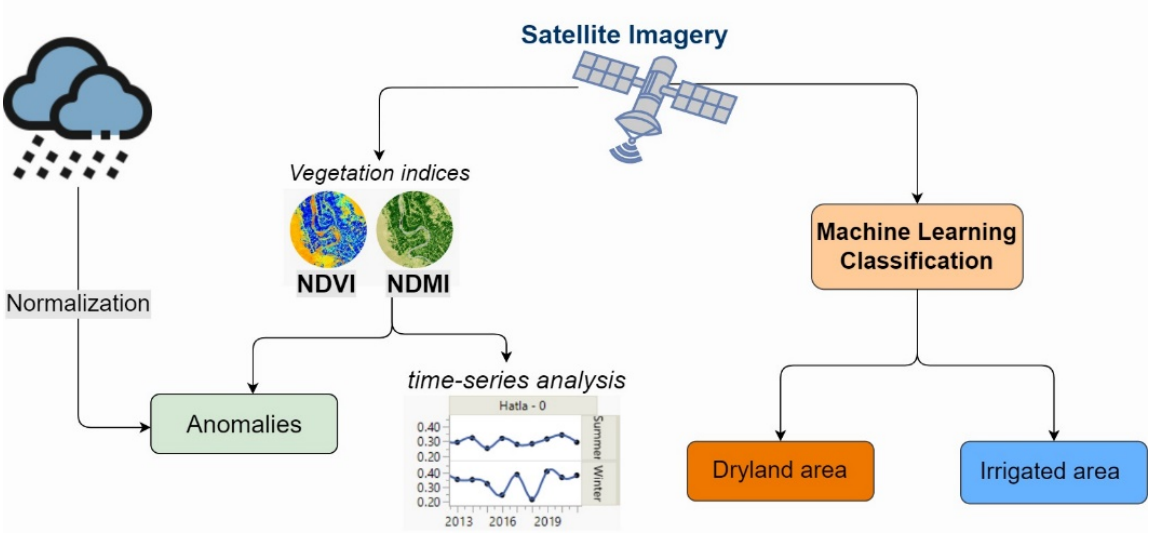
In the case where baseline data is not available, which is not uncommon in HECS, various existing data sources can be leveraged to evaluate large-scale interventions. For example, geo-spatial data can be used as a source of pre-intervention information to estimate needs, which can be matched and incorporated with existing survey data to provide deeper insights where key information and programme-level data are missing or not available (Bunte et al., 2017). Once geo-coded data of all interventions in a geographic unit are available for each year, then existing **geo-referenced observational survey data** such as the Demographic and Health Surveys (DHS), Afrobarometer or the Living Standard Measurement Survey (LSMS) can be used to assess pre- and post-intervention changes in outcome variables using quasi-experimental tools. Howell et al. (2020) combined the DHS with the geocoded Social Conflict Analysis Database (Salehyan et al., 2012) and found that living closer to a conflict zone increases the acute malnutrition of Nigerian children.

Furthermore, **assignment to treatment** and **compliance** are not always clear to the researchers either due to the lack of communication with the program implementers or the possibility of multiple interventions in the same region, particularly in HECS. Geospatial data help to estimate programme participation through environmental changes, such as land or water use of households, for impact evaluations in agriculture and environment sectors (Berger, 2017). Besides that, **remote-sensing** can complement conventional survey data to measure outcomes that are traditionally hard to measure in humanitarian and conflict-affected settings, such as agricultural productivity or water usage (Hoffman et al., 2011; Santika et al., 2019) or provide supplementary information that can be used to refine the statistical models used to assess impacts. Like this, in-field

work can be complemented or even replaced using geospatial data. Granular datasets such as IPUMS TERRA and NASA's Earth Observing System Data and Information System (NASA EOSDIS) can help researchers measure reliable outcome variables which are otherwise difficult or simply impossible to collect in HECS.

However, one must note that obtaining ground-truth data in HECS is very difficult. To overcome this drawback in HECS, unsupervised classification algorithms can be employed. Sujud et al. (2022) is a good example of using satellite imagery with unsupervised classification algorithms to predict agricultural productivity and to identify fields with access to irrigation. Through these methods, they conducted a rigorous impact assessment of an agricultural intervention in one of the conflict-affected regions of Syria. As shown in Figure 2, they normalised and standardised the satellite imagery data by rainfall to derive irrigated areas and agricultural activity across time. In other words, they could use only satellite imagery and deep learning image classification techniques to differentiate which areas were agriculturally productive due to rainfall and which areas were irrigated.

Figure 2. *Standardised and normalised remotely sensed indices coupled with machine learning classifiers*



Source: *Sujud et al. (2022)*

Administrative data is logistically simpler to use in impact evaluations; however, access to it might be the main issue for researchers in HECS where governments might become more restrictive in sharing data due to security reasons. In addition, other technical, governance-related and logistical barriers cause the underutilisation of administrative data even in developed countries (McGrath-Lone et al., 2022) despite the important advantages of using them. For example, as it is regularly collected for administrative purposes and provides information with almost universal coverage, it enables researchers to estimate long-term impacts with high external validity (Harron et al. 2017). And through the objective assessment, we can exclude the selection, enumerator and recall biases (Isaksson, 2021). Therefore, administrative data is instrumental in measuring data with high levels of reporting biases, such as income and tax payments. However, we must keep in mind that it might only be useful for some research questions as it only provides registered data which is limited in nature and might not be useful to measure detailed outcomes or analyse complex analytical models. Thus, it cannot be beneficial if the

research is about informal markets or irregular migration because these activities are not declared to the authorities. In addition, administrative data could be outdated in countries that need more capacity to gather high-quality data. That said, it is often not available in countries affected by the conflict, which puts an additional barrier for researchers. However, collaborations with local and international institutions in those countries might ease the process of obtaining the administrative data. One recent example of using administrative data in a humanitarian crisis case is Peitz et al. (2022) which analyses the impact of formal labour market integration of Syrian refugees in Jordan. Another example is Kayaoglu (2022), who combines the administrative data on court cases and the refugee population at the province level and analyses the impact of refugees on crime rates in Türkiye.

Despite all these alternative advances in data sources that can be used in conducting impact evaluation in HECS, we identify several interrelated aspects that should be developed and addressed more thoroughly in data curation to improve the learning potential in crisis and conflict-affected settings.

First, impact evaluations of humanitarian emergency and development aid in HECS often lack **long-term effects** since fieldwork is usually limited in time, and post-intervention data is collected shortly after the aid distribution. The lack of long-term data in humanitarian impact evaluation impedes our understanding of the sustainability of targeted outcomes and whether humanitarian assistance leads to any expected or unexpected changes in the long term. Moreover, this drawback is one of the reasons that we do not have enough information about the impact of humanitarian assistance on development. The above-mentioned alternative data sources can extend impact insights, but this requires effort from all stakeholders to allocate additional interest and funding for collecting data not only when interventions ended or a few years after the treatment but also to develop data systems and advanced and safe tracking tools to follow up with households for longer periods.

Second, and related to the first point, one of the main issues in designing rigorous impact evaluation in HECS is clearly identifying the treated group. This is particularly challenging in settings where multiple organisations coordinate their efforts to provide emergency aid to crises-affected or displaced households. Broad selection criteria and lack of data on how beneficiaries were identified and selected in practice generate additional barriers. This is particularly prevalent in HECS, where multiple interventions co-occur at the same locations between different actor, and where the rapid response does not allow for rigorous selection, which further complicates developing good research design. Therefore, it is imperative to develop and maintain **standardised data sources about the interventions**, including location, target groups, and type of support, which is uniform across all involved humanitarian stakeholders. researchers and practitioners can co-develop strong yet implementable guidelines and co-create monitoring tools and platforms to collect intervention-level data that can be consistent across settings and actors.

Third, access to new data sources provides us with a novel richness of information. They also make it easier to conduct retrospective impact evaluations. However, there is a need to develop **data value chains (build data infrastructure)**; again, standardisation is key to getting the full benefit of this big data ecosystem. ***Measurement pluralism*** prevents researchers from comparing impact evaluations across different geographical settings, even within the same population. Thus, standardisation would help researchers and program implementers to benefit from other impact evaluations, which could help program effectiveness and research design.

Finally, improving the technical capacities in data collection, curation, and maintenance in fragile and conflict-affected settings and raising awareness on the importance of rigorous impact evaluation and evidence-based learning in informing policymakers and making decisions is important. Therefore, researchers of the Global North should increase their collaborations with local researchers in the Global South. Initiatives such as the Geo-Enabling Initiative for Monitoring and Supervision (GEMS) of the World Bank are

essential and should be thoroughly and methodically expanded in HECS to build local capacities in digital data collection (World Bank, 2021).

4. Fine-tuning measurements

In the past few years, measurement tools for impact evaluations in HECS have flourished. Measures in these settings need to be simple to be applicable for practitioners, comparable across different settings and informative to capture reality accurately. Dimensions of particular interest are demographics, experiences, behavioural parameters, economic well-being and social stability.

Gender-sensitive measurements are key to understanding the how and why impact of a program or policy changes across gender. We can measure gender-based welfare on the micro level, for example, through information on sex, marital status, household headship, household composition, female inclusion, empowerment, and access to assets and capital.

Experiences can be measured in HECS through exposure to adverse events. **Exposure to conflict or fragility** is of particular relevance (Baliki, et al 2022). On the macro level, critical information can be derived from geo-referenced conflict event data like ACLED[3] (Raleigh et al., 2010) or UCDP GED[4] (Sundberg et al., 2012) or Social Conflict Analysis Database (Salehyan et a., 2012), from big data such as phone data (Van der Windt and Humphreys, 2014) as well as from socioeconomic databases[5]. However, not all households living in the same sub-district, village or even community are equally exposed to conflict or are impacted by adverse events similarly. Thus different measurement instruments are needed for the micro level. Measuring exposure to conflict on the micro-level is feasible by directly assessing conflict measures through household surveys (Brück et al., 2016). In addition, behavioural measures such as risk preferences, coping and resilience strategies are important to measure with a minimum bias with concise and well-designed survey

[3] Eck (2012) advises researchers that the Armed Conflict Location Events Dataset (ACLED) may provide biased results if the subnational version of ACLED is used due to its uneven quality control.
[4] The Uppsala Conflict Data Program Georeferenced Events Dataset.
[5] Martin-Shields and Stojetz (2019) present a detailed review main conflict event data sources.

tools. Moreover, measures of economic well-being, poverty and social stability are crucial for fragile and conflict-affected countries but challenging to be captured. Yet, absolute poverty shifted from being a moral concept to a measurable indicator through a person's daily budget.

Furthermore, drawing macro implications from micro research is essential to reach a sufficient level of external validity to transfer findings to other similar contexts where in-field research might be challenging. **Micro, meso and macro** levels of measurement often require different indicators.

Another critical challenge of measurement in HECS is measuring the unmeasurable. There are decisive dimensions for research in humanitarian emergencies and conflict settings, such as resilience, fragility or adaptation, that are **not yet measurable** through available indicators. The FAO developed the Resilience Index Measurement and Analysis to measure resilience (FAO, 2020). However, this measure is criticised by interviewers for its complexity in fragile settings. At the same time, the reduced version is also criticised, for not being informative enough for rigorous research. That said, we have examples of using remote sensing data to measure damage (Bevington et al., 2010), community vulnerability (Flax et al., 2012) and community resilience (Burton 2012) in cases of natural disasters, and researchers in HECS can benefit from these tools in their impact assessments. Thus, there is still a long way to go to find suitable fine-tuning measures for HECS.

Moreover, many key figures in the context of humanitarian emergencies and conflicts cannot be categorised and could only be representative on a **continuous scale.** For example, merely binary classification of population groups into 'vulnerable' and 'non-vulnerable' is misleading. Assigning a vulnerability score as per their distinct challenges would be more informative. Furthermore, sometimes it is crucial to distinguish between **'within' and 'between' measures.** For example, the implications differ substantially between a person being poor and a country being poor.

There is a trade-off between comparability and innovation in measures. For the establishment of **new measures**, the correlation to the traditional measures often tends to

serve as a benchmark of goodness. This fosters comparability but conveys potential biases in measurements. Thus, it is vital to thoughtfully derive the objective and potential pitfalls when developing new measures. Where required values are not measurable, **proxies** might be used to approach an indicator. For example, food prices serve as a proxy for food security. Proxies are also essential to use to measure the unmeasurable outcomes in HECS, where survey questions have a risk of not being allowed or other government restrictions, such as data collection by drones, are forbidden. Moreover, in evaluations where the time frame is short, and longer-term data collection has budgetary and operational costs, the "surrogates" -which are short-term proxies for long-term outcomes- and "surrogate indices" can be used to estimate the long-term impacts (Athey et al. 2019).

On the other side, a range of measures exists for other variables, which partially **overlap**. For example, while FAO measures food security through the Food Insecurity Experience Scale, WFP applies the Food Consumption Score. On the macro level, leading organisations agreed on the Integrated Food Security Phase Classification (IPC) as a measure for food security; however, this does not apply to the micro level. Using different indicators can lead to broader insights but also different conclusions. Therefore, the **comparability** and standardisation of measures are vital, especially in light of drawing meta-analyses and external validity. An essential dimension for comparability is also a standardised **time dimension.** In other words, there is a need for an agreement on the time frame used for each measure (Brück et al., 2016). For example, comparing the same variable measured in different periods might produce misleading conclusions. For example, there is a 'perverse incentive' problem since the impact for late arrivals (and 'low-quality administration') will exceed the impact for early arrival (and 'high-quality administration') (Puri et al., 2017). Besides, it is essential to interconnect different measures meaningfully and to elaborate on which measure serves the desired scope. Therefore, measure selection should always be built on the **theory of change.**

Measures can be **explanatory as well as explained variables** in HECS. To serve as an explanatory variable, independent indicators are needed. However, climate shocks or conflicts are mostly **anticipatable** to a certain extent**,** i.e., they are not entirely exogenous

(Puri et al. 2017). Likewise, humanitarian assistance is mainly targeted to the most vulnerable population, so full exogeneity is challenging to achieve. Measures should assess the **unintended and the intended as well as direct and indirect outcomes.** Further challenges in the development of measures are language and enumerator biases.

Lastly, gaps in the literature indicate the need for more adequate measures. **Evidence gap maps** proposed by IFAD are a powerful tool to unfold which outcome dimensions lack investigative addressing (IFAD, 2022).

5. Complementing Orthodox Evaluation Methods with Machine Learning

One of the toughest challenges to impact evaluation in HECS is finding the proper **evaluation method** which needs to fit the data's pitfalls. Orthodox impact evaluation methods such as RCTs, quasi-experimental designs and other observational studies can be easily implemented in peaceful settings but constructing a well-defined counterfactual is very tricky in HECS. Moreover, traditional linear regressions are prone to a lack of power in the attempt to include several interaction terms with a restricted sample size. Selection bias, information bias, contamination bias, non-random response, and high and systematic attrition are important methodological challenges in conflict settings and humanitarian emergencies (Puri et al., 2017). They can lead to inconsistent impact estimations and sample imbalances, which further challenge the credibility of counterfactuals. Systematic attrition and non-random response also make heterogeneity analysis impossible, which is crucial in learning the channels of impact.

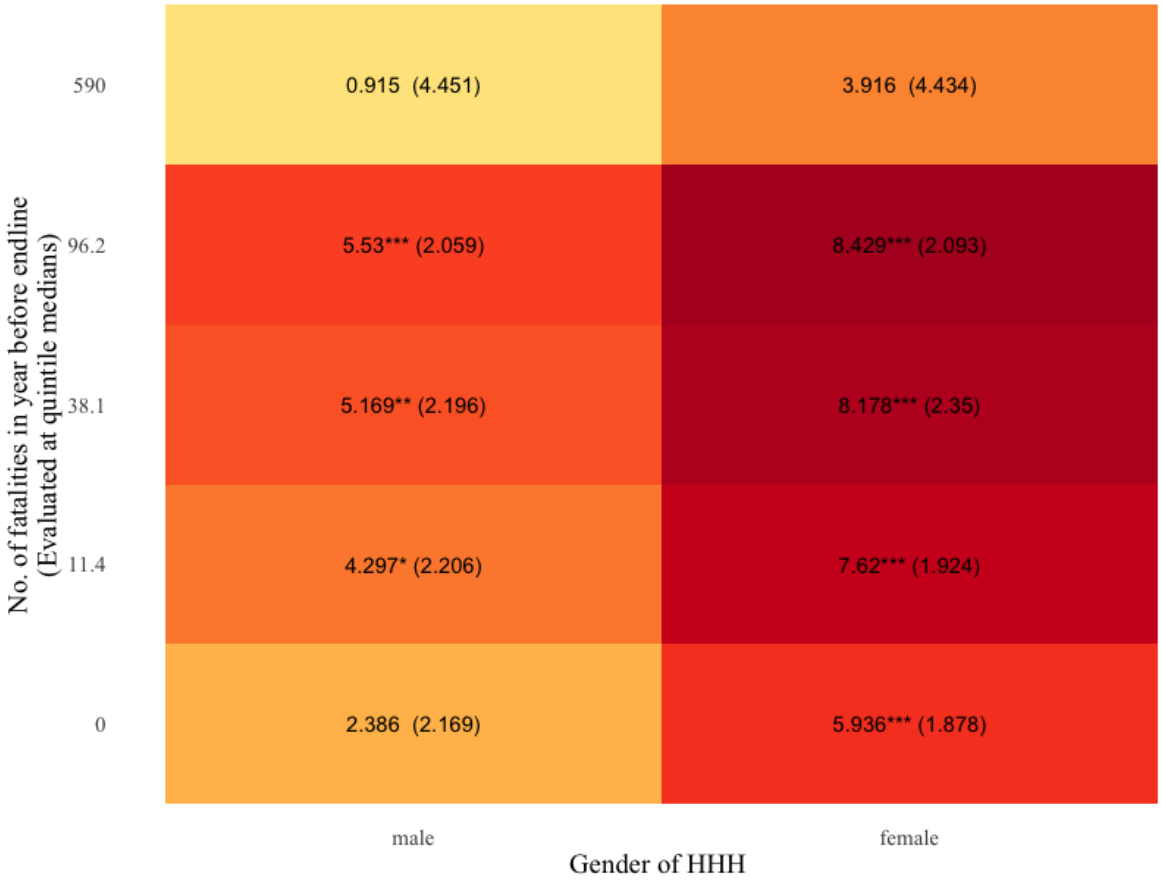
Innovative and adaptive **machine learning algorithms**, such as honest causal forest (Athey and Wager, 2018) and support vector machines (Imai and Ratkovic, 2013), can overcome these pitfalls, thanks to the availability of new (big) data sources. These algorithms enable us to unfold heterogeneous treatment effects, impute missing values, optimise targeting and increase external validity. Forecasting allows **anticipatory action** and efficient resource allocation. Wager and Athey (2019) present an empirical example of using

random forests to estimate the average treatment effect of a binary treatment. Moreover, machine learning can also be very useful for an objective model selection when researchers have many covariates (Samii et al. 2016). Another advantage of machine learning methods is that this data-driven approach can be used, for example, to calculate optimal weights that improve the balance between treatment and control groups in impact evaluation studies (Athey and Imbens, 2017).

Furthermore, in cases where the sample size is enough to estimate the average net impact but not across different demographic groups, then machine learning algorithms can be used to overcome this barrier for the heterogeneity analysis. Weiffen et al. (2022) use honest causal forest estimation to evaluate an agricultural intervention's short-term impact across gender and violent conflict levels in Syria. ***Figure 3*** below from Weiffen et al. (2022) is a nice example of visualising the conditional average treatment effects of the agricultural intervention in Syria across the gender of household heads and incidence of violence.

These novel methods still have **input and output limitations,** such as the lack of information on the underlying procedure of the output generation, the resulting uncertainty of the model quality, and the risk of overfitting, i.e., a lack of generalizability (Maleki et al., 2022). Therefore, it is essential to link machine learning to the corresponding theory of change, particularly in covariate selection, to construct a valid model. Moreover, it is key to cross-validate the model, ensure reliable data quality, and disclose the underlying data, assumptions and approximations. Chernozhukov et al. (2018) present a detailed review of machine learning theory and applications that can be used for impact evaluations.

Figure 3. *CATE by gender of the household head and incidence of violence.*



Source: Weiffen et al. (2022)

6. Concluding Remarks

RCTs are now dominating the development economics and impact evaluation field; however, it is often infeasible to use them in HECS. This domination results in very few impact evaluations done in such fragile contexts, which correspond to, according to the 3ie Impact Evaluation Repository, only 7.4% of all impact evaluations completed between 1990 and 2021. Still, the total share of RCTs in this very limited number of impact evaluations is high which clearly shows that this domination pushes researchers not to adopt sub-optimal methods which are ethically, practically, and methodologically more suitable to HECS, particularly during its relief and recovery phases. Moreover, the ethical

barriers to conducting an RCT in the emergency relief phase result in evaluations of interventions in later phases where randomisation is regarded as a lesser concern. This also creates critical evidence gaps in HECS for different phases of conflict and reconstruction.

Researchers conducting impact evaluations in HECS are expected, therefore, to be flexible and ready to use off-the-shelf designs and innovative methodologies to overcome those challenges in different phases of HECS – namely relief, recovery, and resilience. Alternative ways to construct a counterfactual such as factorial designs, phased-in designs, synthetic control groups, and granular spatial data, or using appropriate balancing methods can be used where possible instead of giving up once a 'traditional' control group is unavailable. Therefore, untraditional but still rigorous ways of conducting an impact evaluation should be preferred with the support of these unorthodox methods, standardisation of measurements, and measuring the otherwise 'unmeasurable' variables and outcomes thanks to the increased availability of innovative data sources with machine learning algorithms. Moreover, innovative and adaptive machine learning algorithms can help overcome causal challenges, particularly when used in tandem with the existing impact evaluation designs.

After discussing the challenges and summarising the new frontiers of impact evaluation in HECS, this paper discusses the need for standardisation across design, methods, data and measures for external validity and to transfer evidence to unavailable settings. Thus, developing a framework that enables the comparability of results across contexts and increases the possibility of external validity is crucial. That said, variable measurement in HECS needs to be simple, comparable across different settings, and informative to capture reality accurately. Finding and fine-tuning proxy measures for unmeasurable concepts and outcomes such as resilience and fragility are crucial. However, prioritising the standardisation of modules and measurements across various contexts is also important to check the external validity of findings. However, the appetite of donors for fuzzy concepts – which might imply different meanings to different people– can make it difficult to make these comparisons.

Given all the discussion in the paper, we provide recommendations for the main stakeholders in an impact evaluation framework. We believe researchers should engage more with local stakeholders, project designers and donors before research design and provide their feedback even during the implementation phase so that the impact of the applied intervention can be improved even in the very short run, which is key in HECS. This also implies that researchers should be flexible and anticipate all possible challenges. However, at the same time, collaboration is not a one-way issue. Program managers and implementing partners should also look for ways to collaborate with researchers. This is vital in all phases of program development and implementation so that the impact of projects can be estimated. And, it will help program developers ensure high-impact program development in an environment that demands rapid and cost-effective interventions. As rigorous impact evaluations require researchers to be rapid and flexible while working in challenging environments, funding agencies can also adjust their decision-making processes faster to support researchers. They can also increase the effectiveness of impact evaluations by investing in the capacity building of local researchers in HECS. That said, demand from aid agencies for rigorous impact evaluations of programs they fund is crucial to strengthen program managers' willingness and implementation partners to cooperate and collaborate with researchers to overcome the various challenges of establishing causality between inputs and outcomes in HECS. This, in turn, will help create and sustain high-impact humanitarian and emergency programs. Finally, we argue that research transparency is crucial in HECS as the low number of impact evaluations are highly important sources of information for everyone working in these fields. That is why researchers should try to be as transparent as possible in all stages of their impact evaluations and not avoid reporting study imperfections.

References

Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. Nature, 603(7903), 864-870.

Altındağ, O., O'Connell, S. D., Şaşmaz, A., Balcıoğlu, Z., Cadoni, P., Jerneck, M., & Foong, A. K. (2021). Targeting humanitarian aid using administrative data: Model design and validation. *Journal of Development Economics*, 148, 102564.

Anderson, M. B. (1999). *Do no harm: how aid can support peace--or war*. Lynne Rienner Publishers.

Anson, S., Watson, H., Wadhwa, K., & Metz, K. (2017). Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors. *International Journal of Disaster Risk Reduction*, 21, 131-139.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, *111*(12), 4088-4118.

Athey, S., and Imbens, G. W. (2017). "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685-725.

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *National Bureau of Economic Research*, No. w26463

Baliki, G. (2017). Empirical advances in the measurement and analysis of violent conflict. https://doi.org/10.18452/18363

Baliki, G., Brück, T., Ferguson, N. T., & Kebede, S. W. (2022). Fragility exposure index: Concepts, measurement, and application. Review of Development Economics, 26(2), 639-660.

Behrer, A. P., & Lobell, D. (2022). Higher levels of no-till agriculture associated with lower PM2. 5 in the Corn Belt. *Environmental Research Letters*, *17*(9), 094012.

Berger, D. (2017). *Water Effectiveness and Targeting Insights from a Geospatial Dataset on Uganda Water Projects*. University of California, Berkeley.

Bevington, J., Pyatt, S., Hill, A., Honey, M., Adams, B., Davidson, R., ... & Eguchi, R. (2010). *Uncovering community disruption using remote sensing: an assessment of early recovery in post-Earthquake Haiti*. Disaster Research Center.

Blanchet, K., Ramesh, A., Frison, S., Warren, E., Hossain, M., Smith, J., Knight, A., Post, N., Lewis, C., Woodward, A. & Dahab, M., (2017). Evidence on public health interventions in humanitarian crises. *The Lancet*, *390*(10109),2287-2296.

Breunig, M., Bradley, P. E., Jahn, M., Kuper, P., Mazroob, N., Rösch, N., ... & Jadidi, M. (2020). Geospatial data management research: Progress and future directions. ISPRS *International Journal of Geo-Information*, *9*(2), 95.

Brück, T., Justino, P., Verwimp, P., Avdeenko, A., & Tedesco, A. (2016). Measuring violent conflict in micro-level surveys: current practices and methodological challenges. The *World Bank Research Observer*, 31(1), 29-58.

Brück, T. and Stojetz, W. (2021). Data Options for Assessing Gender Dimensions of Forced Displacement: A Background Note. Washington, D.C.:World Bank

Brück, T., & Regassa, M. D. (2022). Usefulness and misrepresentation of phone surveys on COVID-19 and food security in Africa. Food Security, 1-31.

Bunte, Jonas B., Harsh Desai, Kanio Gbala, Brad Parks, Daniel Miller Runfola. 2017. Natural Resource Sector FDI and Growth in Post-Conflict Settings: Subnational Evidence from Liberia. AidData Working Paper #34. Williamsburg, VA: AidData.

Burton, C.G. (2012). The Development of Metrics for Community Resilience to Natural Disasters. Ph.D. Thesis, University of South Carolina, Columbia, SC, USA.

Buttenheim, A. (2010). Impact evaluation in the post-disaster setting: a case study of the 2005 Pakistan earthquake. *Journal of Development Effectiveness*, *2*(2), 197-227.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21: C1-CC68.

Corsi, D. J., Neuman, M., Finlay, J. E., & Subramanian, S. V. (2012). Demographic and health surveys: a profile. International journal of epidemiology, 41(6), 1602-1613.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., McIntosh, C., & Nellis, G. (Eds.). (2019). *Information, accountability, and cumulative learning: Lessons from Metaketa I*. Cambridge University Press.

Eck, K. (2012). In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Cooperation and Conflict*, *47*(1), 124-141.

Flax, L. K., Jackson, R. W., & Stein, D. N. (2002). Community vulnerability assessment tool methodology. *Natural Hazards Review*, *3*(4), 163-176.

Food and Agriculture Organization of the United Nations (2020). Resilience Index and Analysis (RIMA) - Short questionnaire. Available at https://www.fao.org/resilience/resources/resources-detail/en/c/1177512/

Gourlay, S., Kilic, T., Martuscelli, A., Wollburg, P., & Zezza, A. (2021). High-frequency phone surveys on COVID-19: good practices, open questions. Food Policy, 105, 102153.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis* 20 (1): 25–46.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, 4(2), 2053951717745678.

Hartman, A., & Kern, F. G. (2020). How to know what works in alleviating poverty: Learning from experimental approaches in qualitative research. *World Development*, *127*, 104804.

Hensen, B., Mackworth-Young, C. R. S., Simwinga, M., Abdelmagid, N., Banda, J., Mavodza, C., ... & Weiss, H. A. (2021). Remote data collection for public health research in a COVID-19 era: ethical implications, challenges and opportunities. Health Policy and Planning, 36(3), 360-368.

Hoffman, C., Melesse, A. M., & McClain, M. E. (2011). *Geospatial mapping and analysis of water availability, demand, and use within the Mara River Basin* (pp. 359-382). Springer Netherlands.

Holland, P.W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81:945-961.

Howell, E., Waidmann, T., Birdsall, N., Holla, N., & Jiang, K. (2020). The impact of civil conflict on infant and child malnutrition, Nigeria, 2013. *Maternal & child nutrition*, *16*(3), e12968.

IFAD (2022). How to do note: Knowledge Gap Mapping. Investing in rural people. Available at: https://www.ifad.org/en/web/knowledge/-/how-to-do-note-knowledge-gap-mapping

Imai, K., & Ratkovic, M. (2013). Estimating treatment effects heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1), 443-470.

Isaksson (2021). A Rapid and Rigorous Impact Evaluation: Advances in the Methods and Data Available for Timely and Cost-Efficient Evaluation. Center for Global Development Background paper. Accessed at https://cgdev.org/sites/default/files/Rapid-evaluation-background-paper.pdf

Jack, B. K., Jayachandran, S., Kala, N., & Pande, R. (2022). Money (Not) to Burn: Payments for Ecosystem Services to Reduce Crop Residue Burning. *National Bureau of Economic Research*, No. w30690.

Kayaoglu, A. (2022). Do refugees cause crime?. *World Development*, *154*, 105858.

Kayaoglu, A., Baliki, G., & Brück, T. (2023). Conducting (Long-term) Impact Evaluations in Humanitarian and Conflict Settings: Evidence from a complex agricultural intervention in Syria. Household in Conflict Network Paper Series, No. 386.

Lau, C. Q., Cronberg, A., Marks, L., & Amaya, A. (2019, December). In search of the optimal mode for mobile phone surveys in developing countries. A comparison of IVR, SMS, and CATI in Nigeria. In Survey Research Methods (Vol. 13, No. 3, pp. 305-318).

Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A., & Forghani, R. (2022). Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. arXiv preprint arXiv:2202.01337.

Martin-Shields, C. P., & Stojetz, W. (2019). Food security and conflict: Empirical challenges and future opportunities for research and policy making on food security and conflict. *World Development*, *119*, 150-164.

Masset, E., Shrestha, S. and Juden, M. (2021). 'Evaluating Complex Interventions in International Development'. *CEDIL Methods Working Paper 6*. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford [online]. Available from: https://doi.org/10.51744/CMWP6

Mc Grath-Lone, L., Jay, M. A., Blackburn, R., Gordon, E., Zylbersztejn, A., Wijlaars, L., & Gilbert, R. (2022). What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature. *International Journal of Population Data Science*, *7*(1).

Nobel Prize Press release: The Prize in Economic Sciences 2019. Available at: https://www.nobelprize.org/prizes/economic-sciences/2019/press-release (published 14 Oct 2019).

OECD. (2006). Outline of principles for impact evaluation. Paris: Organisation for Economic Cooperation and Development (OECD). Available at: www.oecd.org/dac/evaluation/dcdndep/37671602.pdf

OCHA. (2019). Humanitarian Response. Emergency Response Preparedness - Overview. Available at: https://www.humanitarianresponse.info/en/programme-cycle/space/page/preparedness

Paul, A., Jolley, C., & Anthony, A. (2018). Reflecting the past, shaping the future: making AI work for international development. Washington, DC: Center for Digital Development, USAID.

Petiz, L., Baliki, G., Ferguson, NTN., & Brück. T (2023). Do work permits work? The impacts of formal labour market integration of Syrian refugees in Jordan. *Unpublished manuscript*.

Peter, N., & Soetevent, A. R. (2019). Randomization in field experiments. In Handbook of Research Methods and Applications in Experimental Economics (pp. 121-140). Edward Elgar Publishing.

Puri, J., Aladysheva, A., Iversen, V., Ghorpade, Y., & Brück, T. (2017). Can rigorous impact evaluations improve humanitarian assistance? *Journal of Development Effectiveness*, 9(4), 519-542.

Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: an armed conflict location and event dataset: special data feature. Journal of peace research, 47(5), 651-660.

Ravallion, M. (2018). Should the Randomistas (Continue to) Rule? CDG Working Paper 492.

Robins, J. M., & Greenland, S. (2000). Causal inference without counterfactuals: comment. *Journal of the American Statistical Association*, *95*(450), 431-435.

Rosch, S., Raszap Skorbiansky, S., Weigel, C., Messer, K. D., & Hellerstein, D. (2021). Barriers to Using Economic Experiments in Evidence-Based Agricultural Policymaking. *Applied Economic Perspectives and Policy*, *43*(2), 531-555.

Samii, C., Paler, L., & Daly, S. Z. (2016). Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in Colombia. *Political Analysis*, 24(4), 434-456.

Salehyan, I., Hendrix, C. S., Hamner, J., Case, C., Linebarger, C., Stull, E., & Williams, J. (2012). Social conflict in Africa: A new database. *International Interactions*, *38*(4), 503-511.

Santika, T., Wilson, K. A., Budiharta, S., Law, E. A., Poh, T. M., Ancrenaz, M., ... & Meijaard, E. (2019). Does oil palm agriculture help alleviate poverty? A multidimensional counterfactual assessment of oil palm development in Indonesia. *World Development*, *120*, 105-117.

Shiells, K., Di Cara, N., Skatova, A., Davis, O. S., Haworth, C. M., Skinner, A. L., ... & Boyd, A. (2020). Participant acceptability of digital footprint data collection strategies: an exemplar approach to participant engagement and involvement in the ALSPAC birth cohort study. *International Journal of Population Data Science*, 5(3).

Stojetz, W., Ferguson, N. T., Baliki, G., Díaz, O., Elfes, J., Esenaliev, D., ... & Brück, T. (2022). The life with corona survey. *Social Science & Medicine*, 306, 115109.

Sujud,L., Jaafar, H., Baliki, G., and Brück, T. (2023). Evaluating the impact of humanitarian interventions on agriculture productivity in Syria using remote sensing and machine learning. *Unpublished manuscript*.

Sundberg, R., Eck, K., & Kreutz, J. (2012). Introducing the UCDP non-state conflict dataset. *Journal of Peace Research*, 49(2), 351-362.

The United Nations (2022). 'War's Greatest Cost Is Its Human Toll', Secretary-General Reminds Peacebuilding Commission, Warning of 'Perilous Impunity' Taking Hold. Available at: https://press.un.org/en/2022/sgsm21216.doc.htm

The World Bank (2021). Geo-Enabling initiative for Monitoring and Supervision (GEMS). Fragility, Conflict and Violence. available at: https://www.worldbank.org/en/topic/fragilityconflictviolence/brief/geo-enabling-initiative-for-monitoring-and-supervision-gems

The World Bank (2022). Overview. Fragility, Conflict and Violence. available at: https://www.worldbank.org/en/topic/fragilityconflictviolence/overview

UNOCHA. (2022). Global Humanitarian Overview 2022, Snapshot as of 31 May 2022. Accessed at https://reliefweb.int/report/world/global-humanitarian-overview-2022-may-update-snapshot-31-may-2022

Van der Windt, P., & Humphreys, M. (2016). Crowdseeding in Eastern Congo: Using cell phones to collect conflict events data in real time. *Journal of Conflict Resolution*, 60(4), 748-781.

Verwimp, P., Justino, P., & Brück, T. (2019). The microeconomics of violent conflict. *Journal of Development Economics*, 141, 102297.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228-1242.

Weiffen, D., Baliki, G., & Brück, T. (2022). Violent conflict moderates food security impacts of agricultural asset transfers in Syria: A heterogeneity analysis using machine learning (No. 381). Households in Conflict Network, Working Paper No. 381.

White, H. (2009). Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness*, 1 (3), 271–335.

Wood, N. (2021). Stop evidencing complex results with just numbers and embrace narrative. Chemonics. Available at: https://www.chemonics.com/blog/stop-trying-to-evidence-complex-development-results-with-just-numbers-and-embrace-narrative/

Xu, Y., & Yang, E., (2021). Hierarchically regularized entropy balancing. Available at SSRN: https://ssrn.com/abstract=3807620 or http://dx.doi.org/10.2139/ssrn.3807620