Global Repository for Injury Data Discrepancies (GRID) Across Low- and Middle-Income Countries and Harmonization Strategies

by

Armstrong Mbi Obale

Duke Global Health Institute
Duke University

Date:_____
Approved:

_____
Joao Vissoci, Supervisor

_____
Catherine Staton

_____
Joseph Egger

_____
Blandina Mmbaga

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Duke Global Health Institute
in the Graduate School of Duke University

2023

ABSTRACT

Global Repository for Injury Data Discrepancies Across Low- and Middle-Income Countries and Harmonization Strategies

by

Armstrong Mbi Obale

Duke Global Health Institute
Duke University

Date:_____

Approved:

_____
Joao Vissoci, Supervisor

_____
Catherine Staton

_____
Joseph Egger

_____
Blandina Mmbaga

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Science in the Duke Global Health Institute
in the Graduate School of Duke University

2023

Abstract

Introduction: There is a high burden of injuries in low- and middle-income countries (LMICs). Data management and storage systems are suboptimal, making it hard to share data. Also, small data sizes hinder the application of modern data science methods to draw insight. **Methods:** We performed a document analysis of injury data dictionaries from 4 institutions located in 4 different LMICs for discrepancies among data elements. We also compared each of the dictionaries with the World Health Organization (WHO) data set for injury (DSI) and then explored harmonization strategies of injury data, given the discrepancies. **Results:** Of the 949 data elements across the dictionaries, there were 6 (0.72%) shared data elements when considered by presence and 4 (0.45%) when considered by equality across the 4 dictionaries. The number of shared data elements varied when the dictionaries were compared in pairs and triads. We identified four methods of ensuring harmonization of injury data; 1) using the WHO DSI common data elements, 2) using the National Institute of Neurologic Disorders and Stroke (NINDS) common data elements, 3) using functions written in software like R and python, and 4) adopting a prospective and somewhat promising development of a framework that includes common data elements specific to injuries. We successfully harmonized three injury data sets from the Kilimanjaro Christian Medical Centre (KCMC). **Conclusion:** There are huge discrepancies across injury data dictionaries in LMICs. Harmonization of

injury data from LMICs is however achievable and can result in a larger dataset. More research is needed in this area for the development of tools that facilitate injury data harmonization.

Dedication

This work is dedicated to inhabitants of LMICs suffering from preventable injuries.

Contents

List of Tables

List of Figures

Acknowledgements

Dr. João Vissoci: Thank you for passionately guiding me through to a successful completion of this work. You have indeed been a mentor in every sense of the word. I will continue to sit on your shoulders so I will see even farther. A brilliant person, you truly are.

Dr. Catherine Staton: Words cannot express my gratitude to you. Without saying many words, you have taught me several lessons. Your commitment and persistence are worth emulating. The energy that flowed from you saw me through to the completion of this thesis.

Dr. João Vitor Perez de Souza: Thank you for being a reliable collaborator in this work.

Dr. Blandina Mmbaga, Dr. Anna Tupetz, Ashley Philips, Paige O'Leary, Dr. Elizabeth Keating, and all other members of the GEMINI lab: Thank you all for making it easy at every step of this project. I appreciate your mentorship, edits, and regular checking on my progress.

Laura Bey, Kathryn Sullivan, Erin Gauldin, and Emilee Kerr, Sarah Martin, and Lysa Mackeen: I am grateful to you all for standing by me as advisors through the program. The happy conversations kept me going.

The Duke Global Health Institute (DGHI) and Duke University: I am thankful for funding every step of this thesis and creating a supportive learning environment.

The Kilimanjaro Christian Medical Center (KCMC): Thank you for the warmth created during my field work.

Sandy Hatoum: Thank you for translating the Arabic section of the Egyptian documents to English.

Family and friends: I am thankful to family and friends for emotional support throughout the course of this work.

Finally, I am thankful to God Almighty for the gift of life and for enabling me to meet all the amazing persons and institutions that contributed to the success of this work.

1. Introduction

1.1 Burden of injuries

Each year, 4.4 million people die from injuries globally (World Health Organization, 2021). Moreover, according to the Global Burden of Disease 2017 report, injuries were responsible for 252 million disability adjusted life years (DALYs) (Roth et al., 2018). The extent to which these alarming statistics reflect the true burden of injuries is questionable, given other sequelae and the financial burden associated with their management (Rosenkrantz et al., 2019). The burden of injuries is unevenly distributed across countries (Roth et al., 2018), with low- and middle-income countries (LMICs) experiencing a higher burden compared to high-income countries (Gosselin, 2009; James et al., 2020). Despite a higher burden, limited attention is provided to injuries in LMICs (Gosselin, 2009; James et al., 2020).

For decades, high-income countries have utilized high-quality data for injury surveillance and clinical policy making (Rosenkrantz et al., 2020). These countries have well-established trauma care systems that generate quality data which is used for research (Reynolds et al., 2017). Limitations in data collection, management, and storage methods delay and obscure efforts aimed at improving injury surveillance, clinical policy making, and impact the overall population health and wellbeing. Researchers in LMICs have acknowledged that an available and functional trauma registry can result in

better injury surveillance and patient outcome (Bommakanti et al., 2018). However, injury surveillance systems are generally poorly developed in LMICs (Mock et al., 2004). This may be related to limited funding (St-Louis et al., 2018).

1.2 Trauma registries in LMICs

In LMICs, trauma registry models fall within 4 basic categories; "emergency department (ED) only, ED with pre-hospital data, ED with in-hospital data, and ED with pre-hospital, in-hospital, and post discharge indicators" (Pyda et al., 2018). Successful implementation of the registry is more likely among facilities that utilize the ED-only model (Pyda et al., 2018). Most LMICs generally do not have the required resources such as finances to set up robust trauma registries comparable to those in high-income countries (St-Louis et al., 2018). Despite the resource constraint, there has been an increase in the number of trauma registries in LMICs within the last decade (Rosenkrantz et al., 2019; St-Louis et al., 2018). There are, however, marked differences in the registries (Pyda et al., 2018; Rosenkrantz et al., 2020).

However, there are limitations in leveraging trauma registries. Inconsistencies in data elements have been found across trauma data bases (Simko et al., 2019). The challenges faced by LMICs to set up and sustain trauma registries are unique within specific countries (Bommakanti et al., 2018). To set up and sustain data harmonization

pipelines, there is need for the existence of properly functional trauma registries within countries.

Generally, data are collected and stored differently across countries and institutions within countries. This seems not to only be limited to trauma. In cancer data harmonization, there existed a possibility for a variable name to mean different things in different places (Rolland et al., 2015). Data harmonization requires that there be a consensus on data variables and units across countries (Burkhauser & Lillard, 2005). The benefits of data harmonization are countless (Doiron et al., 2013) and include attainment of large sample sizes (Hamilton et al., 2011; Khoury, 2004; Smith-Warner et al., 2006; Thompson, 2009), greater generalizability (Hamilton et al., 2011; Khoury, 2004; Noale et al., 2005), and improved validity of comparative analysis (Bath et al., 2010; Serra-Majem et al., 2003).

1.4 Purpose of this study

This work is among the first of a series of efforts aimed at establishing a Global Repository for Injury Data (GRID) for LMICs. The GRID is intended to be developed and implemented using the FAIR data principles: Findability, Accessibility, Interoperability, and Reusability. Given that injuries remain a global health crisis (Roth et al., 2018) and a leading cause of disability-adjusted life years (DALYs) in LMICs (Gosselin, 2009; James et al., 2020), establishment of the GRID is a major step towards

understanding injuries in LMICs. With high quality and sufficient data, interventions that target prevention, control, and policy establishment in LMICs will be better informed.

The objective of this study is to explore the discrepancies across current injury data collection registries in LMICs and explore strategies of harmonization of injury data given the observed discrepancies (Figure 1). In the context of this study, harmonization is the merging of two or more different data sets into one, such that the resulting data set has more data points, unique variable names, correct binding of corresponding variables, and preservation of non-corresponding variables. This aim is achieved through the following specific aims 1) To identify similarities and differences across injury data dictionaries from LMICs, 2) To explore harmonization strategies for Tanzania injury data, and 3) To harmonize injury data from Tanzania Kilimanjaro Christian Medical Centre (KCMC).

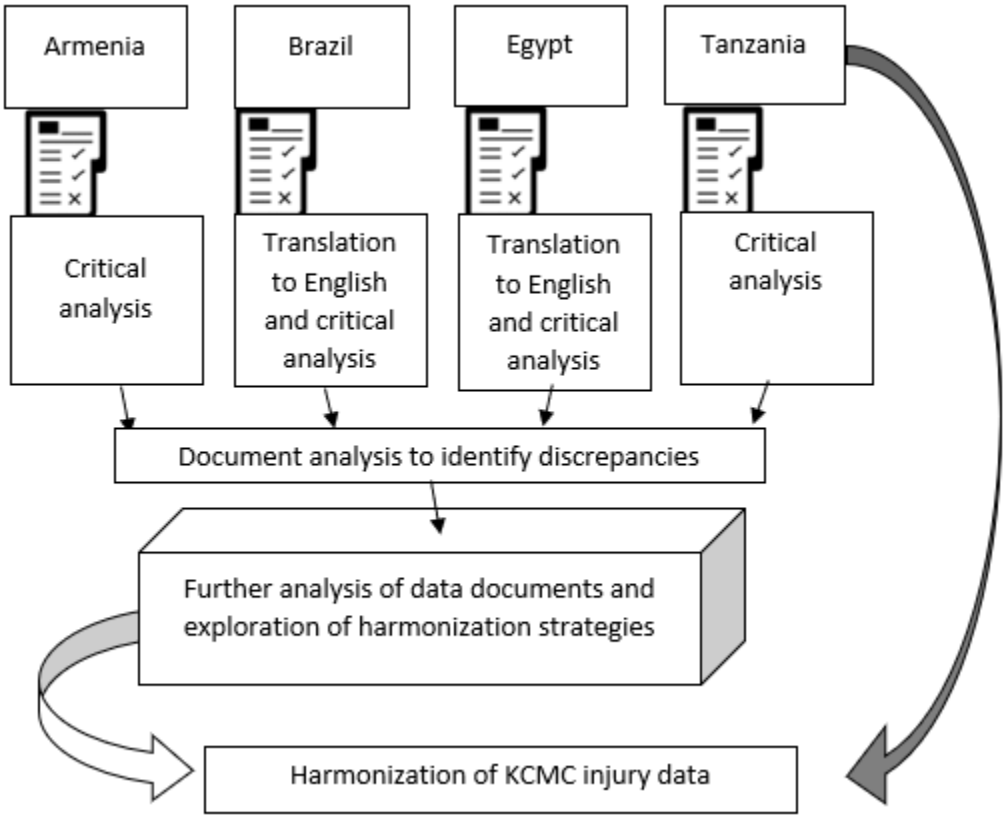


Figure 1: Study flowchart

2. Methods

2.1 Setting

The setting of this study consists of four hospitals located in four LMICs: Brazil, Egypt, Armenia, and Tanzania. Each of the selected institutions has an existing tool for the collection of injury data. The State University of Maringa (SUM) is in Maringa, Brazil, a city with a total population of 423,666, and a hospital with 1,577 beds. Each year, SUM receives about 1,000 trauma patients. Surb Astvatsamayr Medical Center is located in Yerevan, Armenia. It receives about 15,000 inpatients annually and serves a catchment area of 42,000 people (*Sourb Astvatsamayr Medical Center*, n.d.). Ain Shams University is in Cairo, Egypt. The metropolitan region of Greater Cairo has a population of 20.5 million people (*Cairo Population 2023*, n.d.). Every year, Ain Shams University Hospital treats more than 2000 trauma patients and has approximately 3200 hospital beds. Currently serving a catchment area of 15 million people, KCMC is a referral hospital located in the Kilimanjaro region of Tanzania. and is the third-largest hospital in the country (Zimmerman et al., 2023) and sees about 2,000 injury patients a year. KCMC currently has 2 trauma registries in active use for trauma data collection.

The WHO Data Set for Injury (DSI) (World Health Organization, 2020a) is a standard against which we compared the 4 data dictionaries. The WHO DSI consists of 76 data elements recommended by the WHO to be included in injury data collection registries. The WHO data set for injury data elements were developed and published by

the WHO to serve as a standardized case-based data collection tool in emergency departments (World Health Organization, 2020b). The overall purpose is to ensure interoperability and reusability of trauma data. The data elements are classified as core or extended.

2.2 Discrepancies across injury registries

A total of 4 documents (represented in this thesis by the names of the countries where the institutions are located) from the 4 institutions were analyzed for discrepancies. In addition to the discrepancies across the documents, each one was compared with the WHO DSI (World Health Organization, 2020). Data elements (variables) from each of the documents, including the WHO data set for injury were spread on a spreadsheet (Figure 2). The data elements were spread such that columns represented the documents (for example, "Tanzania", with elements highlighted blue, includes all data elements in the document from KCMC) and rows represented the data elements (for example, moving across the row highlighted in light yellow, we find shared data elements across the five documents; WHO to Tanzania).

Data elements from the respective dictionaries were the data points in our analysis. Each cell consisted of a data element from a specific document, or a combination of data elements, for matching purposes. The organization was such that

data elements that were shared in two or more documents could be found across the same row.

Data elements were matched at two levels. The first level consisted of matching according to **presence or absence** without regard to characteristics like data element type and number of levels. This first level was referred to as matching by presence. In the second matching level, referred to as matching by **equality**, data elements were matched if they had; 1) unambiguous meaning (neither included other variables nor meant something different, even if it could be inferred. This did not apply to date-time variables) 2), levels that agreed (similar parameterization), and 3) specified units, for data elements that contained units.

A	B	C	D	E
WHO Data set for injury (https://www.who.int/publication	**Brazil**	**Egypt**	**Armenia**	**Tanzania**
			TBI registry prospective data	
		Barcode		
Registry Case ID	(1) Paciente Patient numerical identification (text)	Patient number	(1) rcd_id Record ID (text)	(1)Participant ID (text)
				(2) Form version (int)
Patient residence	(6) Origem de entrada do paciente; patient's origin (text)	Address (text)	(9) Country of permanent residence (nominal - 4 levels)	(3) Where does the patient live (nominal - 3 levels), (4) Other (text)
Facility ID			(2) Health facility (nominal - 6 levels)	
			(3) Medical record number (text)	
date of birth		Date of birth		

Figure 2: Spreadsheet with data elements from the respective documents (represented by countries of origin)

Table 1 summarizes the documents analyzed. Of the documents, 3 were dictionaries while 1 was a set of data collection forms used for hospital data collection that included patient follow-up data. Two of the documents were completely in English. Two of the documents were divided into subsections. Moreover, one of the documents was adopted for use by multiple sites and 3 of the 4 were administered by an interviewer.

Table 1: Summary of documents

Construct	**Document (Country of institution owning document used to represent the document)**			
	Brazil	**Egypt**	**Armenia**	**Tanzania**
Document type	Data dictionary	Data collection forms	Data Dictionary	Data Dictionary
Purpose	Collection of traumatic brain injury data	Collection of data on hospital follow up of admitted patients	Collection of traumatic brain injury data	Collection of traumatic injury data
Number of data elements	28	341	103	576
Original language of dictionary	Portuguese	Arabic and English	English	English
Number of Subsections	1	7	1	8
Hospital	Emergency	All departments,	Emergency	Emergency

departments included (Number)	department (1)	dependent on patients' diagnosis and treatment course	department (1)	department, Intensive care unit (ICU), SubICU, surgery (4)
Used by multiple sites (Number of sites)	No	No	Yes (6)	No
Used in at least one other country (Total number of countries)	No	No	Yes (3)	No
Data collections method	Interviewer administered	Data filled in health care provider(s) as part of care protocol	Interviewer administered	Interviewer administered

While spreading data elements on the Microsoft excel sheet, some adjustments were made to improve comparability. The adjustments resulted in a change in the number of rows of data elements used in making the Venn diagrams during analysis. In one instance, some data elements of documents were placed on more than one row, so they matched corresponding data elements of other documents. For example, "BP_/_" was duplicated to match "Systolic blood pressure on arrival" and "Diastolic blood pressure on arrival". Table 2 summarizes the adjustments made and final number of data elements included for each document.

Table 2: Adjustments made, and final number of data elements included for each document

	Brazil	Egypt	Armenia	Tanzania
Actual number of data elements on document	28	341	103	576
Number duplicated	2	1	0	0
Number merged	4	0	0	2
Number excluded because they appeared more than once	0	11	0	0
Excluded for other reasons	0	87 (signatures - 37, care plan -49, consent -1)	0	1 (instructions)
Number of entries for comparison in terms of number of equal data elements	28	244	103	574

2.3 Harmonization of data from Tanzania Kilimanjaro Christian medical Centre (KCMC)

Three data harmonization strategies were explored. First, the WHO Data Set for Injury (World Health Organization, 2020a) shares common data elements with injury data sets and could be used for data harmonization. Second, the National Institute of Neurologic Disorders and Stroke (NINDS) Common Data Elements (CDE) (National Institute of Neurologic Disorders and Stroke Common Data Elements, n.d.) is a collection of common data elements meant to ensure consistency in the collection and

recording of data across neuroscientific clinical studies. The NINDS also provides a means for injury data harmonization. Data elements from registries could be matched with the NINDS CDEs just as they would be matched with the WHO data set for injury. The conditions covered by the NINDS include traumatic brain injuries, epilepsy, and stroke.

Many software can be used for harmonization. To our knowledge, there are currently no known functions that can harmonize data with at least as many discrepancies as we have observed. To limit error associated with interpretation of variables, it may not be prudent to assign certain roles to machines. As such, the WHO data set for injury and the NINDS CDE harmonization approaches would work well when 5 or more data sets need to be harmonized while the use of functions would work well for the harmonization of less than 5 data sets.

2.3.1 Participants

The three registries harmonized in this study, pediatric injuries (702 rows, 912 columns), treatment of care (ToC) (1435 rows, 640 columns), and traumatic brain injuries (TBI) (3693 rows, 132 columns) are registries used in the collection of injury data at KCMC. The pediatric injuries registry is used to collect injury data from pediatric patients at KCMC. The TBI registry includes all ages, and the treatment of care is used to collect injury data from non-pediatric cases 18 and above presenting at the KCMC for

medical care. The harmonization was done with a function written in R version 4.2.1. (Appendix 1)

2.4 Analysis

The entries (data elements) in the spreadsheet were converted to numbers, such that data elements across rows were represented by unique numbers. The converted spreadsheet was imported into R version 4.2.1 for analysis. Levels of discrepancies were displayed on Venn diagrams. First, a Venn diagram that included all 4 data elements was used to display discrepancies in terms of available data elements across the four documents, triads of documents and pairs of documents. This displayed discrepancies by availability of data elements. Second, a similar analysis that included data elements by equality was done. Non-numeric discrepancies were analyzed using thematic analyses.

To compare each document with the WHO Data Set for Injury, we displayed the number of data elements of the WHO Data Set for Injury present in each document and expressed them as a percentage of the total number of data elements within the WHO Data Set for Injury. This information was displayed on a bar chart. We also displayed all data elements within the WHO Data Set for Injury that were not found in the documents on a table.

3. Results

3.1 Discrepancies across documents

The number of shared data elements were 6 (0.72%) based on presence (fig 3A) and 4 (0.45%) based on equality (Figure 3B). Across triads and pairs of dictionaries, the number of shared data elements were a minimum of 7 (0.84%) and a maximum of 13 (1.56%) and 57 (6.84%) respectively (Figure 3A). These numbers reduced when the data elements were analyzed by equality. The number of shared data elements varied by subsections (Table 4). There were no shared data elements under prehospital and prior facility care across any of the documents and a few data elements under clinical conditions.

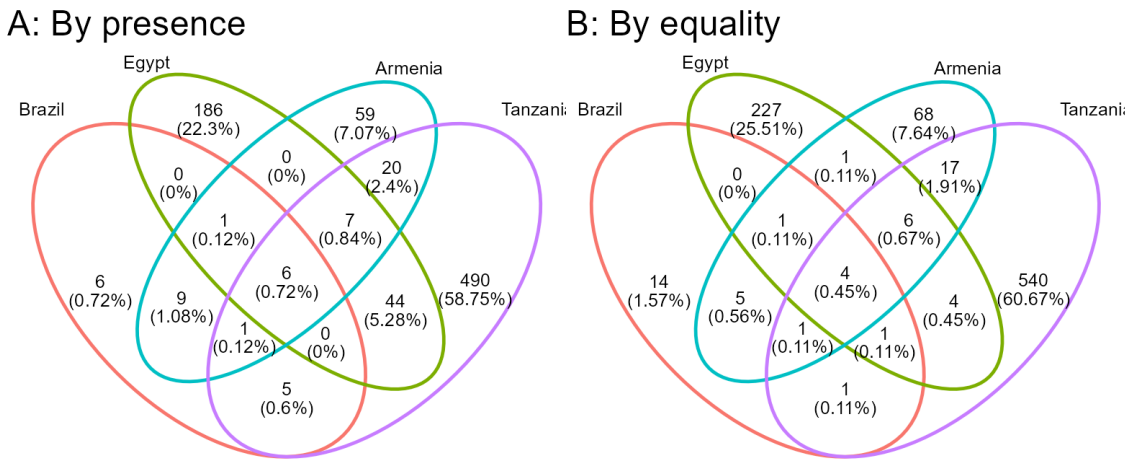


Figure 3: Comparison of dictionaries in terms of data elements. A) by presence (similarity), B) by equality

Table 3: Discrepancies across documents using the WHO minimum data for injury categories as themes

Theme	
Patient registration	Least number of discrepancies across the dictionaries as these data elements are essential to keep track of patients admitted. Noticeably, the terms "gender" and "sex" are used interchangeably.
Initial Clinical Condition	Commonly collected data elements such as Arrival Date, Initial Glasgow Coma Scale, Heart Rate, Oxygen Saturation, Initial Systolic Blood Pressure, and Initial Spontaneous Respiratory Rate were present in at least 3 datasets. Importantly, the WHO DSI data elements Signs of Life, First Vital Sign Assessment Time, and Time of First Provider Assessment were absent in all datasets. These could however be inferred from other data elements
Details of Injury	Data elements related to the Mechanism of Injury (MOI) and whether the injury was intentional, were present in 75 and 25% of dictionaries, respectively. Further description of MOI was mostly either absent or collected in different formats than recommended. Information regarding mass casualty events and the use of protective devices were completely absent. Brazil and Egypt reported 1 and 0 out of the 12 data elements, respectively.

Prehospital and Prior Facility Care	No agreement among the dictionaries. The Tanzanian dictionary contained 33.3% of the variables
Injury exam	Condition and Number of Provider-Defined Serious Injuries were completely absent from all dictionaries. Anatomic region injured was only present in one dataset.
Emergency unit details	Information regarding emergency unit Departure Date, Time, and Disposition was only disclosed in one of the data dictionaries
Inpatient details	Facility Discharge Date and Disposition were present in 3 and 2 dictionaries, respectively. Inpatient Interventions, Date, and Time of arrival to the Operating Theatre were completely absent. Brazil provided no information for the data elements in this section.

There were 9, 18, 24, and 31 shared data elements between Brazil, Egypt, Romania, and Tanzania respectively and the WHO data set for injury (Figure 4). Across the 7 sections of the WHO Data Set for Injury, there were several data elements that were absent across the documents (Table 4). Some of the data elements, such as "Age" could be inferred from other data elements, such as "Date of birth". They are different constructs anyway, given that "Date of birth" cannot be inferred from "Age".

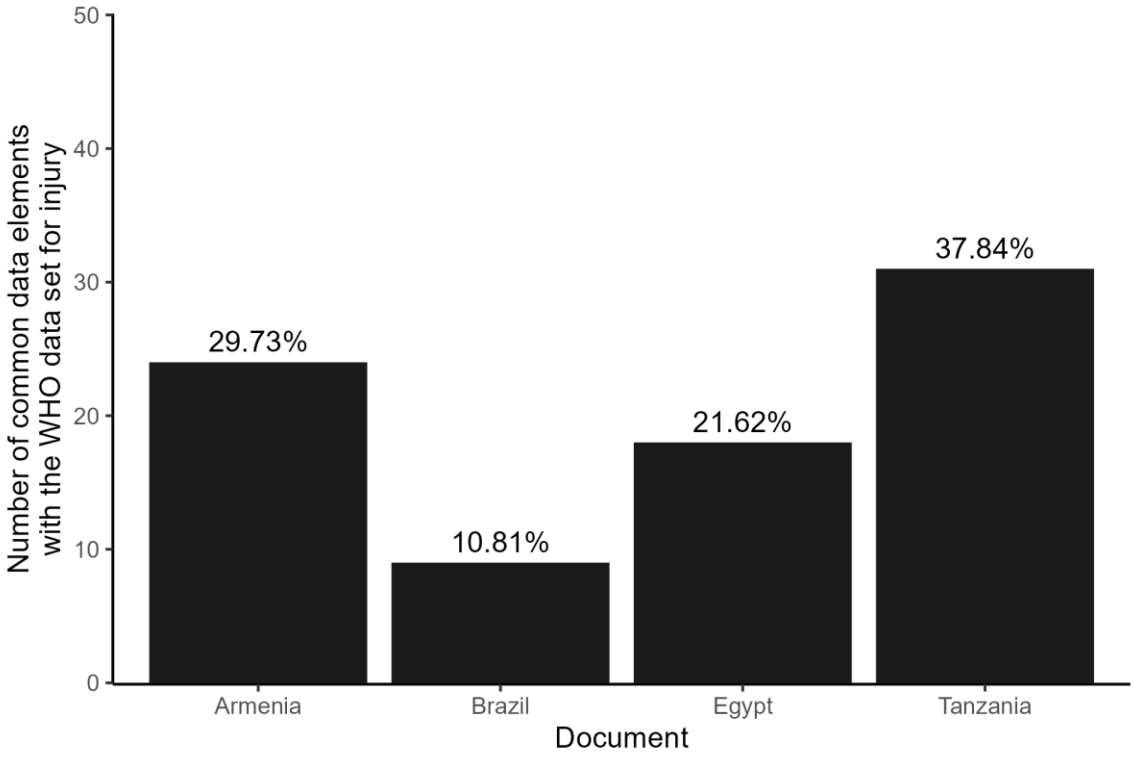


Figure 4: Shared data elements between documents and the WHO data set for injury

Table 4: WHO data set for injury data elements absent across documents

WHO minimum data set for injury category	**Excluded sections across data dictionaries**			
	Brazil	**Egypt**	**Armenia**	**Tanzania**
Patient registration	**Core:** If unknown, age category; Injury geographic location	**Core:** If unknown, age category; Injury geographic	**Core:** If unknown, age category	**Core:** If unknown, age category, Injury

	Extended: date of birth; patient occupation	location	**Extended:** date of birth	geographic location **Extended:** date of birth
Initial clinical condition	**Core:** Number of prior facilities; Arrival time; Time of first provider assessment; Emergency unit/facility arrival mode; Signs of life; First vital sign assessment time; Initial spontaneous respiratory rate; Initial heart rate; Initial systolic blood pressure; Initial oxygen saturation; initial AVPU; GCS qualifier; Supplemental oxygen administration **Extended:** Referring facility level; Number of major medical comorbidities; Is patient pregnant?; Number of pupils	**Core:** Number of prior facilities; Arrival time; Time of first provider assessment; Emergency unit/facility arrival mode; First vital sign assessment time; Initial oxygen saturation; initial AVPU; GCS qualifier **Extended:** Referring facility level; Number of major medical comorbidities; Is patient pregnant?; Number of pupils reactive to light; Initial gcs - eye; Initial gcs - verbal; Initial gcs - motor	**Core:** Number of prior facilities; Time of first provider assessment; Emergency unit/facility arrival mode; initial AVPU; GCS qualifier; Supplemental oxygen administration **Extended:** Referring facility level; Number of major medical comorbidities; Is patient pregnant?; Number of pupils reactive to light	**Core:** Number of prior facilities; Time of first provider assessment; Initial GCS; GCS qualifier; Supplemental oxygen administration **Extended:** Referring facility level; Number of major medical comorbidities; Is patient pregnant?; Number of pupils reactive to light

	reactive to light; Initial gcs - eye; Initial gcs - verbal; Initial gcs - motor			
Details of injury	**Core:** Setting; Activity; Injury event date; Injury event time; Mass casualty event; Intent; Protective devices; If drowning or submersion, intent of submersion; If poisoning or exposure, mode; Road user; If RTI, patient mode of transport **Extended:** If intentional assault, assault relationship: Counterpart to patient's crash; Drug use	**Core:** Setting; Activity; Injury event date; Injury event time; Mass casualty event; Intent; Protective devices; Mechanism; If drowning or submersion, intent of submersion; If poisoning or exposure, mode; Road user; If RTI, patient mode of transport; **Extended:** If intentional assault, assault relationship; Counterpart to patient's crash; Drug use	**Core:** Setting; Activity; Mass casualty event; Intent; Protective devices; Mechanism; If drowning or submersion, intent of submersion; If poisoning or exposure, mode; Road user **Extended:** If intentional assault, assault relationship; Counterpart to patient's crash; Drug use	**Core:** Setting; Activity; Mass casualty event; Protective devices; If drowning or submersion, intent of submersion; If poisoning or exposure, mode; Road user **Extended:** If intentional assault, assault relationship; Counterpart to patient's crash; Drug use
Prehospital and prior facility care	**Core:** Scene of first transport care **Extended:** Facility where first care sought - arrival date; Facility where first care	**Core:** Scene of first transport care **Extended:** Facility where first care sought - arrival date;	**Core:** Scene of first transport care **Extended:** Facility where first care sought - arrival date;	**Core:** Scene of first transport care **Extended:** Scene to first care transport mode; Priority

	sought - arrival time; Scene to first care transport mode; Priority facility interventions; Procedures performed at the scene or during first transport	Facility where first care sought - arrival time; Scene to first care transport mode; Priority facility interventions; Procedures performed at the scene or during first transport	Facility where first care sought - arrival time; Scene to first care transport mode; Priority facility interventions; Procedures performed at the scene or during first transport	facility interventions; Procedures performed at the scene or during first transport
Injury exam	**Core:** Conditions; Injury anatomic location; Number of provider-defined serious injuries	**Core:** Conditions; Injury anatomic location; Number of provider-defined serious injuries	**Core:** Conditions; Injury anatomic location; Number of provider-defined serious injuries	**Core:** Conditions; Injury anatomic location; Number of provider-defined serious injuries
Emergency unit details	**Core:** Emergency unit interventions; Emergency unit departure date; Emergency unit departure time; Emergency unit disposition **Extended:** If morgue or died official cause of death; Standardized	**Core:** Emergency unit interventions; Emergency unit departure date; Emergency unit departure time; Emergency unit disposition **Extended:** If morgue or died official cause of death;	**Core:** Emergency unit interventions; Emergency unit departure date; Emergency unit departure time; Emergency unit disposition **Extended:** If morgue or died official cause of death;	**Core:** Emergency unit interventions **Extended:** If morgue or died official cause of death; Standardized diagnosis

	diagnosis	Standardized diagnosis	Standardized diagnosis	
Inpatient details	**Core:** Inpatient interventions; Date patient first arrived to operating theater; Time patient first arrived to operating theater; Facility disposition; Facility discharge date. **Extended:** inpatient intervention date; Total number of ventilator days; Number of operating theatre encounters; Complications; Functional status at disposition; if morgue or died official cause of death; Standardized diagnosis	**Core:** Date patient first arrived to operating theater; Time patient first arrived to operating theater; Facility disposition **Extended:** inpatient intervention date; Total number of ventilator days; Number of operating theatre encounters; Complications; Functional status at disposition; if morgue or died official cause of death; Standardized diagnosis	**Core:** Inpatient interventions; Date patient first arrived to operating theater; Time patient first arrived to operating theater **Extended:** inpatient intervention date; Total number of ventilator days; Number of operating theatre encounters; Complications; Functional status at disposition; if morgue or died official cause of death; Standardized diagnosis	**Core:** Inpatient interventions; Date patient first arrived to operating theater; Time patient first arrived to operating theater **Extended:** inpatient intervention date; Total number of ventilator days; Number of operating theatre encounters; Complications ; Functional status at disposition; if morgue or died official cause of death; Standardized diagnosis

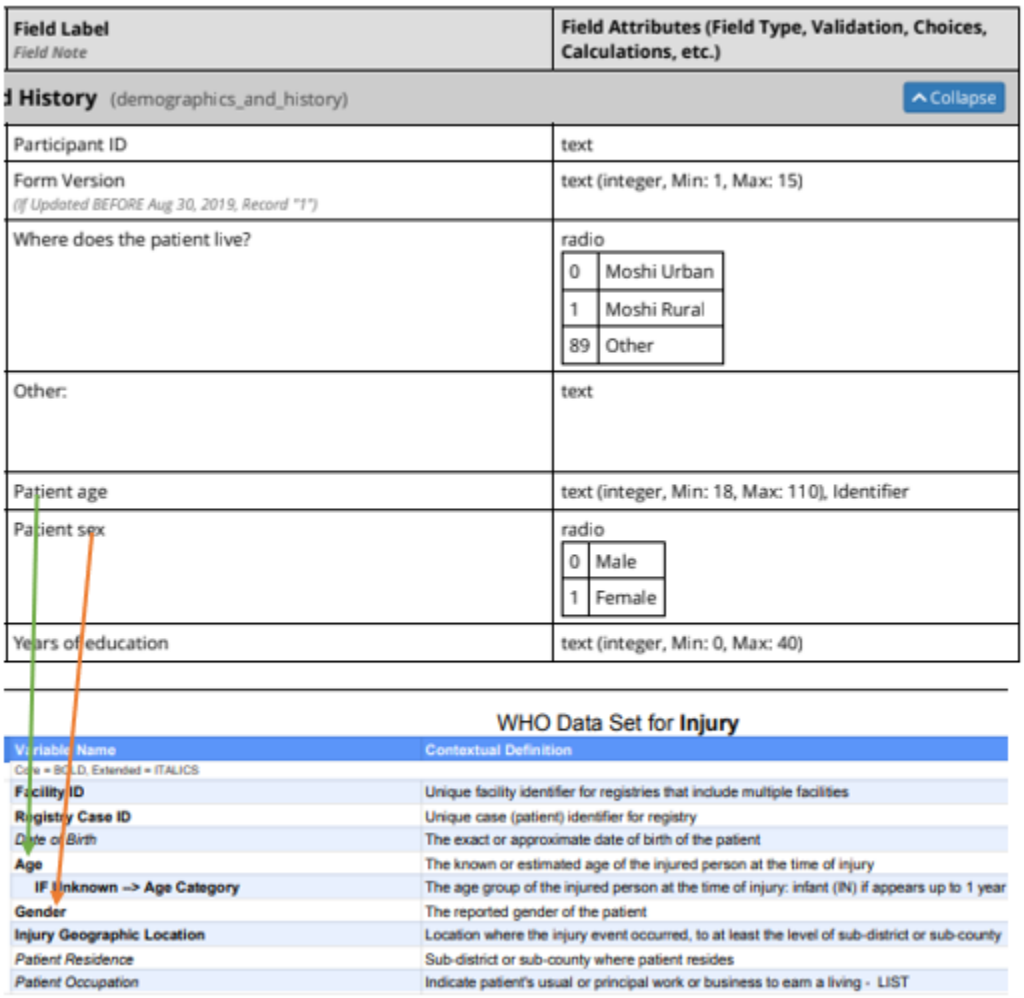
Core: Core data elements, **Extended:** Extended data elements

3.2 Data harmonization strategies

We found a good number of shared data elements between the WHO data set for injury and our data dictionaries. Therefore, mapping variables from different injury data sets to the WHO data set for injury common data elements (Figure 5, Figure 6) and then binding the corresponding data points together is a possible means of harmonizing injury data from LMICs. We did not perform any analysis on the NINDS CDE in this study. However, because just like the WHO data set for injury, it shares common data elements with the injury datasets, the same method described with the WHO Data set for injury could be applied to the NINDS CDE.

As an alternative means of harmonizing data, we wrote a function that harmonizes data sets in pairs. Our function takes as arguments two data sets to be harmonized (data 1 and data 2) and lists of shared data elements (list 1 and list 2) from the data sets such that the data elements are in the same order on both lists. We used the R function (appendix 1) which we developed for harmonization of injury data sets to harmonize three injury data sets from KCMC. This goal was completed through a series of steps. First, we simulated data that mimicked the discrepancies we would expect in injury data sets and wrote a function in R version 4.2.1 which could harmonize the data without any loss in data quality. The key deliverables of the function included 1) making sure that one name is adopted for each data element, 2) filling cells for data sets without particular data elements with NAs, 3) making sure that no data element is lost, and 4)

minimizing the risk of mistakes that may include combining variables that ought not to be combined and not combining variables that ought to be combined to the lowest possible level.



Field Label *Field Note*	**Field Attributes (Field Type, Validation, Choices, Calculations, etc.)**
History (demographics_and_history)	Collapse
Participant ID	text
Form Version *(If Updated BEFORE Aug 30, 2019, Record "1")*	text (integer, Min: 1, Max: 15)
Where does the patient live?	radio 0 Moshi Urban 1 Moshi Rural 89 Other
Other:	text
Patient age	text (integer, Min: 18, Max: 110), Identifier
Patient sex	radio 0 Male 1 Female
Years of education	text (integer, Min: 0, Max: 40)

WHO Data Set for **Injury**

Variable Name Core = BOLD, Extended = ITALICS	Contextual Definition
Facility ID	Unique facility identifier for registries that include multiple facilities
Registry Case ID	Unique case (patient) identifier for registry
Date of Birth	The exact or approximate date of birth of the patient
Age	The known or estimated age of the injured person at the time of injury
IF Unknown --> Age Category	The age group of the injured person at the time of injury: infant (IN) if appears up to 1 year
Gender	The reported gender of the patient
Injury Geographic Location	Location where the injury event occurred, to at least the level of sub-district or sub-county
Patient Residence	Sub-district or sub-county where patient resides
Patient Occupation	Indicate patient's usual or principal work or business to earn a living - LIST

Figure 5: Matching data elements with the WHO DSI common data elements. Top: Section of trauma registry, Bottom: Section of the WHO data set for injury

Our prospective solution is to create a collection of common data elements for injury data in a similar manner to what is within the NINDS CDE. We consider this as the most effective step towards ensuring future harmonization of injury data.

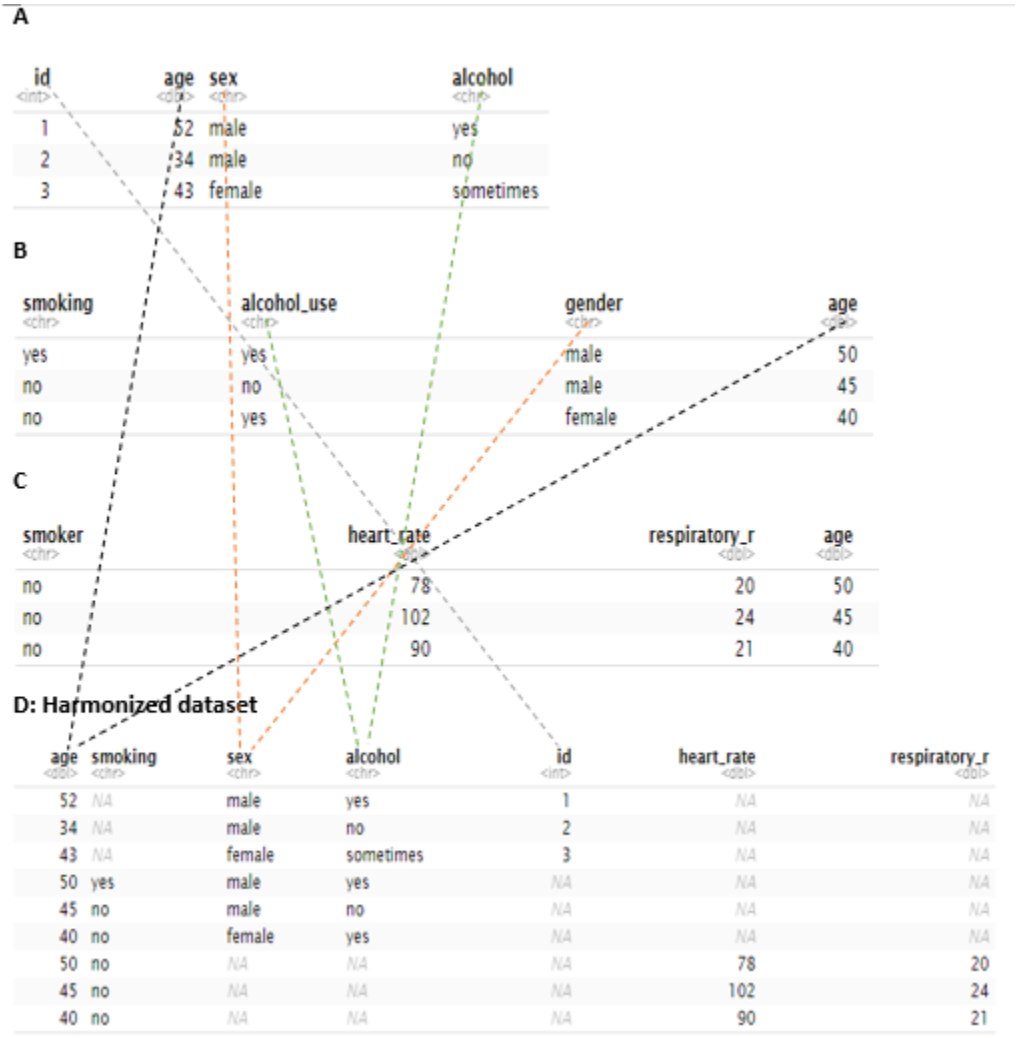


Figure 6: What data harmonization is expected to look like

3.3 Harmonization of data from KCMC

To harmonize the 3 data sets from KCMC, we followed the following steps: 1) data elements and associated labels from each of the registries were carefully extracted based and spread, based on similarities, on a spreadsheet such that shared data elements were found on the same rows (Figure 7). The row highlighted in yellow contains shared data elements across three dictionaries. Then, the spreadsheet was imported into R version 4.2.1. The respective matching data elements were pulled out and converted into lists and saved. Third, the data were imported into R and applied onto the function (Appendix 1). The result was a harmonized data set with 5830 rows and 1668 columns as well as preserved data quality.

A	B	C	D	E
SN	**ToC_label**	**Ped_I_label**	**PRACT_Label**	**TBI_label**
1	Participant ID	Pt 2020 Registry ID#	Pt 2020 Registry ID#	Study Idenfication
2				Time of Casualty MD Arrival
3				Time of Surgery MD Arrival
4		Research Assistant		
5			RN	
6	Form Version	Form Version (Bottom of page)	Form Version (Bottom of page)	
7		Patient Dead on Arrival?	Patient Dead on Arrival?	
8	Where does the patient live?	Where patient lives:	Where patient lives:	
9	Other:	Other:	Other:	
10	Patient age	Patient age ____ years	Patient age	Age
11		Patient age ____ months		
12		Patient age ____ days		

Figure 7: Data elements pulled out from the KCMC injury data dictionaries to be harmonized

4. Discussion

This paper is a part of a collection of early efforts that are aimed at establishing a global repository for injury data from LMICs. We discuss discrepancies across injury data dictionaries from 4 different LMICs, as well as with the WHO minimum data set for injury. We also discuss harmonization strategies for injury data registries and proceed to a description of the methods we adopted for the harmonization of injury data from KCMC.

Like previous studies (O'Reilly et al., 2016; Simko et al., 2019), we identified huge discrepancies across injury data dictionaries. The implications of this finding include a frustration of efforts to combine injury data from different LMIC facilities for injury research. This may limit the extent to which hidden patterns that occur across countries can be identified. As one would expect, the least area of discrepancy was in patient registration; it is obvious, given that data from every unique individual ought to be identifiable. "Gender" and "sex" were used to mean the same thing across dictionaries. Harmonization of the 'gender' and "sex" data elements would require adoption of either of them as the data element name in the harmonized data. However, had they meant "gender" and "sex" meant different things, they would be treated as different variables.

The WHO DSI has the potential of guiding the data harmonization process; data elements of different data sets could be matched with its minimum data elements,

enabling easy harmonization. However, as observed in our study, no document had the majority of its data elements found in the WHO DSI. This observation is not shocking, given that injury registries are traditionally designed within specific settings to collect data which are relevant in meeting specific needs. In most cases, culture and available diagnostic tools dictate, to some extent, the data that can be collected. Notwithstanding, the small number of shared data elements between the WHO DSI and the injury dictionaries may limit the extent to which we may depend on the WHO DSI for injury data harmonization.

Our observed limited agreement between each dictionary and the WHO DSI means that the injury dictionaries did not perform well at capturing the minimum data recommended by the WHO. This finding is in line with previous studies that have reported differences in the extent to which national trauma registries are in alignment with the WHO minimum dataset for injury in the collection of neurotrauma data (Barthélemy et al., 2021). This is somewhat expected , given that the WHO DSI was only recently created to ensure that defined characteristics are included in any injury data collection tool.

Currently, harmonization of injury data from LMICs will predominantly be done using the WHO DSI common data elements, the NINDS common data elements, or using statistical package functions like the one we developed and used in this study. The choice

will often depend on the available data. In the long run, however, the establishment of packages in R, Python, and other programming software that include functions that aid in performing the steps in the data harmonization process would be of great value. Injury data harmonization would be easier in the future, should a framework like the NINDS be established to facilitate the development of registries and subsequent data harmonization.

On a rather prospective note, just as we have observed with the NINDS CDE, a critical analysis could be done and clearly defined common data elements for injury data be established. This could mean improving the WHO data set for injury such that the data elements are more explicit in terms of at least variable class and number of levels. As such, future researchers would consult the common data elements when designing injury data collection tools. The established data elements would therefore be included in injury registries such that future injury data harmonization across LMIC would be facilitated.

4.1 Study strengths and limitations

With a focus on possibility for harmonization, this study reveals the discrepancies across injury dictionaries from 4 different countries. This is a major strength since the findings inform injury researchers on the discrepancies they should expect across injury data sets. This study also analyzes the documents by comparing each one of them with the WHO minimum data set for injury. This points to the extent to which current data collection tools in LMICs collect data that agree with both the

WHO data set for injury and with each other. Our study also identifies different harmonization strategies.

Our analysis focused more on the availability of data elements than on data element equality in terms of variable type and number of variable levels. As such, we would expect more discrepancies should we consider variable equality in terms of type and number of levels. In comparing the documents with the WHO minimum data set for injury, we focused on the presence of elements in a harmonizable manner. We would expect less discrepancies had the analysis been done considering variables that could be inferred. As an example, WHO DSI data elements "Signs of Life" and "First Vital Sign Assessment Time" were absent from all data dictionaries. They can however be inferred.

Moreover, while our study reveals harmonization strategies for injury data, we have not provided details on how this harmonization should be done. We would expect each of the harmonization strategies to have strengths and limitations. While the R function written and used for harmonization respects the principles of data harmonization, other unforeseen aspects such as the need to coerce the class of a variable to that of another for purposes of binding would have to be addressed.

4.2 Implications for policy and practice

Unlike prior studies, at least to our knowledge, this focused on quantifying the extent to which registries share data elements, as a means of exploring harmonization

strategies given the observed similarities and differences. We show possible means of harmonizing injury data. This may result in an increase in sample size, strengthening of research collaborations and facilitation of data storage and research.

To facilitate future harmonization of injury data, policy makers may need to incentivize the use of the WHO DSI common data elements. Funding institutions should make it their policy to require and mandate the inclusion of WHO DSI common data elements as a criterion for funding consideration.

4.3 Implications for further research

Our study findings have taken research on data structures a step further. An understanding of the current discrepancies that exist across data dictionaries is, in addition to findings from previous studies, arguably an effective means of moving towards injury data harmonization strategies. We acknowledge the need for research on the strengths and limitations of the data harmonization strategies which we have discussed. Furthermore, we can't undermine the need to explore and establish sustainable methods of implementing structures that can enable data harmonization and data sharing within ethically sound and socially acceptable norms.

5. Conclusion

Our study sought to identify discrepancies across injury data dictionaries and explore methods of harmonizing injury data, given the discrepancies. There are major discrepancies across injury data collection dictionaries that vary across pairs and triads of dictionaries. At a time when there is an increased need for data to inform policy and clinical decision making, data harmonization cannot be overlooked. Our study places data harmonization strategies in the spotlight. Despite the observed discrepancies, there exist methods of injury data harmonization. Injury data could be harmonized by binding datasets together using their 1) shared data elements, 2) common data elements of the WHO data set for injuries or the NINDS, and 3) applying functions written using programming languages. Prospectively, establishing a framework of common data elements for injury data would be a breakthrough in injury data harmonization. Such a framework will not only serve as a means of data harmonization but also a guide during development of injury data collection too

Appendix A

Appendix 1: The R function "harmonize_data()" used for data harmonization

```
harmonize_data <- function(data1, data2, var1_ , var2_common){

# data1 is the first data set. The function is designed such that the common variables in the new
data set will take the corresponding variable names of data1. data1 should therefore be chosen
wisely among the two data sets to be harmonized.

# data2 is the second data set.

# var1_common is the character string of variables in data1 which are present in data2 though
may not have the same name.

# var2_common is a character string of variables in data2 which are present in data1 though may
not have the same name (this means there must be manual inspection of the data sets prior to
harmonization such that the common variables will be identified).


 #In brief, this function harmonizes data by 1) matching two data sets by the data elements
(variables) provided. Data elements should the same variable called differently in different data
sets such as sex and gender, 2) aligning the common elements such that they fall on the same
index positions and share the same column names (the column names of the elements from data1),
3) binding the two data sets by rows such that the corresponding rows fall in place and the non-
corresponding rows are filled with 'NA's.
```

```
var1_common1 = c()


 for (i in var1_common){
 var1_common1[i] <- c(i)
} # iterating over the variables. This is done because the number may vary


var2_common1 = c()


 for (j in var2_common){
 var2_common1[j] <- c(j)
} # iterating over the variables. This is done because the number may vary


storage1 = c()
storage2 = c()


for (i in var1_common1){
    storage1[i] <- grep(i, colnames(data1))
} #this is for data1
```

```
for (j in var2_common1){
    storage2[j] <- grep(j, colnames(data2))
# We are pulling the index numbers of the variables such that we could rearrange them to match
the order of variables in data1
}

data1_new <- data1[, storage1]
data2_new <- data2[, storage2] #Here we are rearranging the common variables in data2 in a
way that they match those in data1

#other_variables <- setdiff(colnames(data2), var2_common)

colnames(data2_new) <- c(colnames(data1_new)) # The variables names of data1 are being
assigned to the corresponding variable names in data2

data2_other_vars = subset(data2, select = !colnames(data2) %in% var2_common) #
Variables in data2 which are not common to variables in data1 are separated from the data set and
saved in 'data2_other_vars'. This leaves only the variables that can be bound by rows arranged in
the same order as in data1. Remember that at this point the common variables already have the
same names as the variables in data1
```

```
data2_new <- bind_cols(data2_new, data2_other_vars) # The variables are bound to the new
data2 data set which has its variables arranged such that they match the variables of data1 and
have the same names as the data1 variables.



data1_other_vars = subset(data1, select = !colnames(data1) %in% var1_common) #
Variables in data1 which are not common to variables in data2 are separated from the data set and
saved in 'data1_other_vars'. This leaves only the variables that can be bound by rows arranged in
the same order as in data2.

data1_new <- bind_cols(data1_new, data1_other_vars) # The variables are being bound to
the new data1 data set which has its variables arranged such that they match the variables of
data2 and have the same names.

#At this point we can bind the two data sets by rows.

harmonized_data <- bind_rows(list(data1_new, data2_new)) #binding our datasets by row

return(harmonized_data) # This returns the harmonized variables.}
```

References

Bath, P. A., Deeg, D., & Poppelaars, J. (2010). The harmonisation of longitudinal data: A case study using data from cohort studies in The Netherlands and the United Kingdom. Ageing & Society, 30(8), 1419–1437. https://doi.org/10.1017/S0144686X1000070X

Bommakanti, K., Feldhaus, I., Motwani, G., Dicker, R. A., & Juillard, C. (2018). Trauma registry implementation in low- and middle-income countries: Challenges and opportunities. Journal of Surgical Research, 223, 72–86. https://doi.org/10.1016/j.jss.2017.09.039

Burkhauser, R. V., & Lillard, D. R. (2005). The contribution and potential of data harmonization for cross-national comparative research. Journal of Comparative Policy Analysis: Research and Practice, 7(4), 313–330. https://doi.org/10.1080/13876980500319436

Cairo Population 2023. (n.d.). Retrieved March 6, 2023, from https://worldpopulationreview.com/world-cities/cairo-population

Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H. R., Perola, M., Stolk, R. P., Foco, L., Minelli, C., Waldenberger, M., Holle, R., Kvaløy, K., Hillege, H. L., Tassé, A.-M., Ferretti, V., & Fortier, I. (2013). Data harmonization and federated analysis of population-based studies: The BioSHaRE project. Emerging Themes in Epidemiology, 10(1), 12. https://doi.org/10.1186/1742-7622-10-12

Gosselin, R. (2009). Injuries: The neglected burden in developing countries. Bulletin of the World Health Organization, 87(4), 246–246. https://doi.org/10.2471/BLT.08.052290

Hamilton, C. M., Strader, L. C., Pratt, J. G., Maiese, D., Hendershot, T., Kwok, R. K., Hammond, J. A., Huggins, W., Jackman, D., Pan, H., Nettles, D. S., Beaty, T. H., Farrer, L. A., Kraft, P., Marazita, M. L., Ordovas, J. M., Pato, C. N., Spitz, M. R., Wagener, D., … Haines, J. (2011). The PhenX Toolkit: Get the Most From Your Measures. American Journal of Epidemiology, 174(3), 253–260. https://doi.org/10.1093/aje/kwr193

James, S. L., Castle, C. D., Dingels, Z. V., Fox, J. T., Hamilton, E. B., Liu, Z., Roberts, N. L. S., Sylte, D. O., Bertolacci, G. J., Cunningham, M., Henry, N. J., LeGrand, K. E., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abedi, A., Abegaz, K. H., Abosetugn, A. E., Abushouk, A. I., … Vos, T. (2020). Estimating global injuries morbidity and mortality: Methods and data used in the Global Burden of Disease 2017 study. Injury Prevention, 26(Suppl 2), i125–i153. https://doi.org/10.1136/injuryprev-2019-043531

Khoury, M. J. (2004). The case for a global human genome epidemiology initiative. Nature Genetics, 36(10), Article 10. https://doi.org/10.1038/ng1004-1027

Mock, C., Quansah, R., Krishnan, R., Arreola-Risa, C., & Rivara, F. (2004). Strengthening the prevention and care of injuries worldwide. The Lancet, 363(9427), 2172–2179. https://doi.org/10.1016/S0140-6736(04)16510-0

National Institute of Neurologic Disorders and Stroke Common Data Elements. (n.d.). Project Overview | NINDS Common Data Elements. Retrieved March 6, 2023, from https://www.commondataelements.ninds.nih.gov/ProjReview

Noale, M., Minicuci, N., Bardage, C., Gindin, J., Nikula, S., Pluijm, S., Rodríguez-Laso, A., Maggi, S., & for the CLESA working group. (2005). Predictors of mortality: An international comparison of socio-demographic and health characteristics from six longitudinal studies on aging: the CLESA project. Experimental Gerontology, 40(1–2), 89–99. https://doi.org/10.1016/j.exger.2004.09.003

O'Reilly, G. M., Gabbe, B., Moore, L., & Cameron, P. A. (2016). Classifying, measuring and improving the quality of data in trauma registries: A review of the literature. Injury, 47(3), 559–567. https://doi.org/10.1016/j.injury.2016.01.007

Pyda, J. S., Bhatt, M. N., Sonderman, K. A., Powell, B. L., Koch, R., Sanchez, S. E., Afshar, S., & Drake, F. T. (2018). Trauma Registries in Low- and Middle-Income Countries: Current Models. Journal of the American College of Surgeons, 227(4), e155. https://doi.org/10.1016/j.jamcollsurg.2018.08.422

Reynolds, T. A., Stewart, B., Drewett, I., Salerno, S., Sawe, H. R., Toroyan, T., & Mock, C. (2017). The Impact of Trauma Care Systems in Low- and Middle-Income Countries. Annual Review of Public Health, 38(1), 507–532. https://doi.org/10.1146/annurev-publhealth-032315-021412

Rolland, B., Reid, S., Stelling, D., Warnick, G., Thornquist, M., Feng, Z., & Potter, J. D. (2015). Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach. American Journal of Epidemiology, 182(12), 1033–1038. https://doi.org/10.1093/aje/kwv133

Rosenkrantz, L., Schuurman, N., Arenas, C., Nicol, A., & Hameed, M. S. (2020). Maximizing the potential of trauma registries in low-income and middle-income countries. Trauma Surgery & Acute Care Open, 5(1), e000469. https://doi.org/10.1136/tsaco-2020-000469

Rosenkrantz, L., Schuurman, N., & Hameed, M. (2019). Trauma registry implementation and operation in low and middle income countries: A scoping review. Global Public Health, 14(12), 1884–1897. https://doi.org/10.1080/17441692.2019.1622761

Roth, G. A., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, H. T., Abebe, M., Abebe, Z., Abejie, A. N., Abera, S. F., Abil, O. Z., Abraha, H. N., … Murray, C. J. L. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the Global Burden of Disease Study 2017. The Lancet, 392(10159), 1736–1788. https://doi.org/10.1016/S0140-6736(18)32203-7

Serra-Majem, L., MacLean, D., Ribas, L., Brulé, D., Sekula, W., Prattala, R., Garcia-Closas, R., Yngve, A., Lalonde, M., & Petrasovits, A. (2003). Comparative analysis of nutrition data from national, household, and individual levels: Results from a WHO-CINDI collaborative project in Canada, Finland, Poland, and Spain. Journal of Epidemiology & Community Health, 57(1), 74–80. https://doi.org/10.1136/jech.57.1.74

Simko, L. C., Chen, L., Amtmann, D., Gibran, N., Herndon, D., Kowalske, K., Miller, A. C., Bulger, E., Friedman, R., Wolfe, A., Chung, K. K., Mosier, M., Jeng, J., Giacino, J., Zafonte, R., Kazis, L. E., Schneider, J. C., & Ryan, C. M. (2019). Challenges to the Standardization of Trauma Data Collection in Burn, Traumatic Brain Injury, Spinal Cord Injury, and Other Trauma Populations: A Call for Common Data Elements for Acute and Longitudinal Trauma Databases. Archives of Physical Medicine and Rehabilitation, 100(5), 891–898. https://doi.org/10.1016/j.apmr.2018.10.004

Smith-Warner, S. A., Spiegelman, D., Ritz, J., Albanes, D., Beeson, W. L., Bernstein, L., Berrino, F., van den Brandt, P. A., Buring, J. E., Cho, E., Colditz, G. A., Folsom, A. R., Freudenheim, J. L., Giovannucci, E., Goldbohm, R. A., Graham, S., Harnack, L., Horn-Ross, P. L., Krogh, V., … Hunter, D. J. (2006). Methods for Pooling Results of Epidemiologic Studies. American Journal of Epidemiology, 163(11), 1053–1064. https://doi.org/10.1093/aje/kwj127

Sourb Astvatsamayr Medical Center. (n.d.). Retrieved March 25, 2023, from https://www.doctors.am/en/institution/sourb-astvatsamayr-medical-center

St-Louis, E., Paradis, T., Landry, T., & Poenaru, D. (2018). Factors contributing to successful trauma registry implementation in low- and middle-income countries: A systematic review. Injury, 49(12), 2100–2110. https://doi.org/10.1016/j.injury.2018.10.007

Thompson, A. (2009). Thinking big: Large-scale collaborative research in observational epidemiology. European Journal of Epidemiology, 24(12), 727. https://doi.org/10.1007/s10654-009-9412-1

World Health Organization. (2020a). WHO dataset for injury. https://www.who.int/publications/m/item/who-dataset-for-injury

World Health Organization. (2020b, October 12). WHO International registry for trauma and emergency care. https://www.who.int/publications/m/item/who-international-registry-for-trauma-and-emergency-care

World Health Organization. (2021). Injuries and violence. https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence

Zimmerman, A., Minnig, M. C., Meela, J., Tupetz, A., Bettger, J. P., Vissoci, J. R. N., & Staton, C. (2023). A systematic review and cross-sectional survey of rehabilitation resources for injury patients in the Kilimanjaro Region of Tanzania. Disability and Rehabilitation, 0(0), 1–8. https://doi.org/10.1080/09638288.2023.2179674