

# DATA MANAGEMENT

**Workshop on implementation of operational research conditions and  
an all-oral shorter regimen**

10<sup>th</sup> July 2019 – Dubai

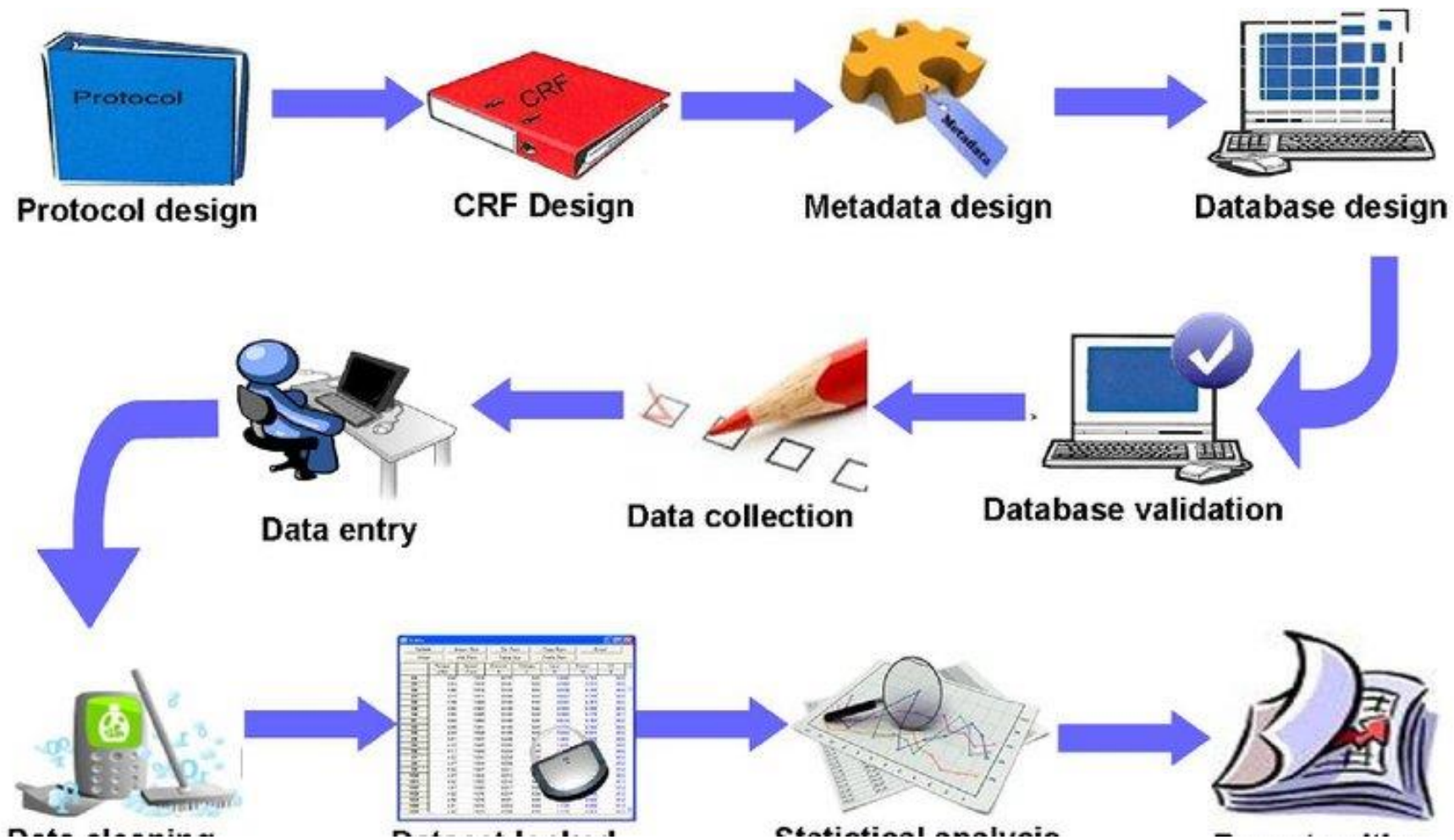
Mathieu Bastard

Epicentre – Médecins sans Frontières



CENTER FOR GLOBAL  
HEALTH DELIVERY-DUBAI  
HARVARD MEDICAL SCHOOL











Thousands of  
data collected

# Data flow

- Important to clearly identify who will be responsible for what
  - Filling the forms
  - Organization of the transportation of the forms (froms the clinic to data entry team)
  - Data entry
  - User rights and access to data collection too or EMR
- Process for data sharing
  - Ownership of the data
  - To who and how data will be shared or sent
    - Data sharing agreement

# Security

- Need to ensure confidentiality and protect privacy of patients

## DATA TO PROTECT

Whose data?



- patients in MSF supported facilities
- communities served by MSF
- participants in research supported by MSF

Which data sources?



- files linked to individuals or containing individuals' personal data (patient records, images, video and audio)
- registers & tally sheets
- research data
- data used for advocacy
- GIS data

What level of data identification?



- identifiable
- indirectly identifiable (including de-identified data and data just with patient numbers)
- anonymous data that could cause harm to individuals, groups or MSF

## HIGHLY SENSITIVE DATA



Can cause harm to ...

- patients
- potential patients
- their family
- groups or communities



Can lead to ...

- harm
- stigmatisation
- discrimination
- violence

Example of MSF data protection policy (in dropbox)

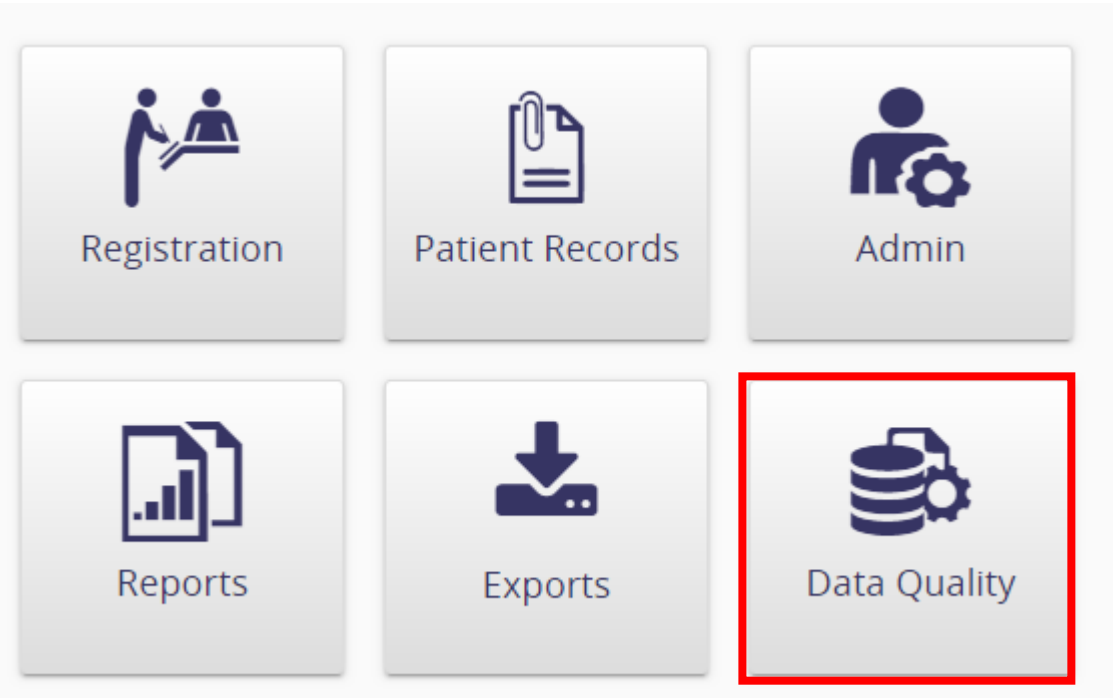
# Data cleaning, why?

- Data cleaning is important to ensure completeness and good quality of the data
- Data cleaning should be performed regularly
  - Retrieve missing data asap (age, start treatment date, culture results...)
  - Correct inconsistent data (dates, clinical assessment vs. laboratory values...)
  - Duplicate entries
- Important for monitoring (missing planned visits, tests...)

# Data cleaning

- Real time data quality: range values, inconsistent dates etc...
- TB-EMR has an **in-build data quality tool**

- Start treatment date missing
- DST missing
- Treatment outcome missing
- Etc...



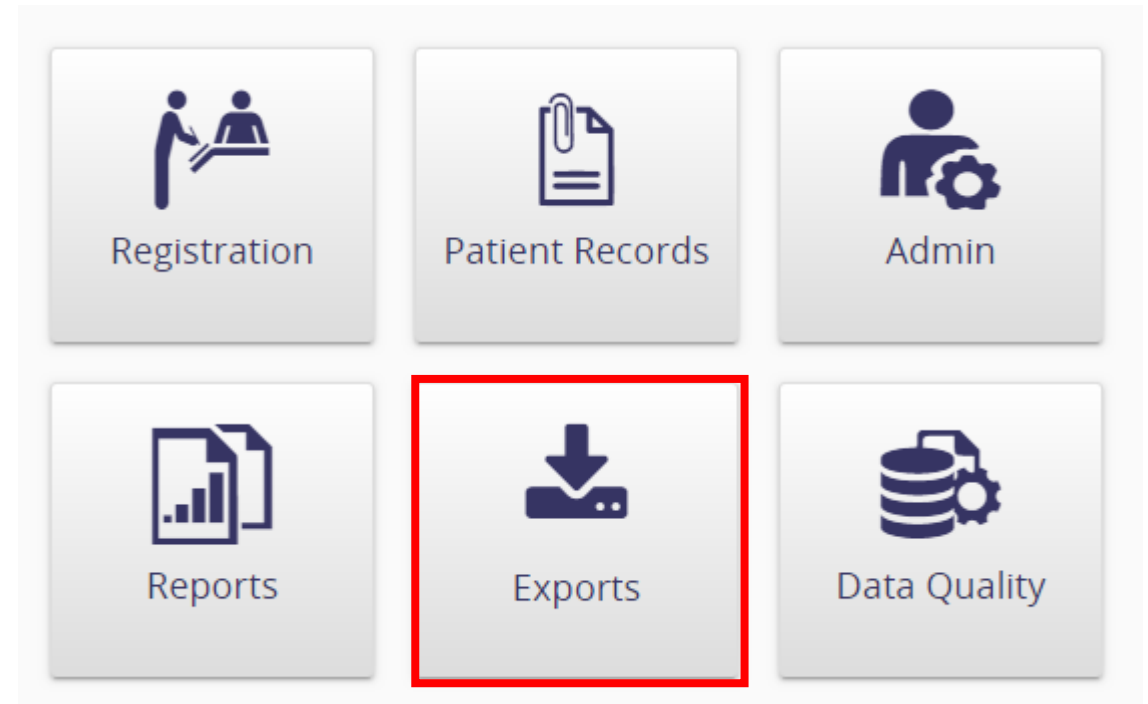


# Data quality tool

Rule Name	Treatment Reg No	Name	Treatment Facility	EMR ID	Notes	Action
Case Definition is missing	342434234	TestP Ptest		ARM7782	The Baseline form is not filled	<a href="#">Edit</a>
Baseline Firstline DST missing	56CF3EB5	56BAE630 56BAE637	NTCC DR UNIT 4, Abovyan, KOTAYK, Armenia	XYZ96	No results by the end of first month of treatment.	<a href="#">Edit</a>
Outcome Missing	56CF3F8F	56BAE551 56BAE558	NTCC DR UNIT 4, Abovyan, KOTAYK, Armenia	XYZ95	The Outcome - End of Treatment form is not filled	<a href="#">Edit</a>
Outcome Missing	56CF40F3	56BAE7FA 56BAE803	NOR-ARESH TBC, Yerevan-Nor-Aresh, YEREVAN, Armenia	XYZ98	The Outcome - End of Treatment form is not filled	<a href="#">Edit</a>
Baseline Firstline DST missing	56CF423D	56BAE92D 56BAE931	STEPANAVAN TBC, Stepanavan, LORI, Armenia	XYZ100	No results by the end of first month of treatment.	<a href="#">Edit</a>
Baseline Secondline DST missing	56CF423D	56BAE92D 56BAE931	STEPANAVAN TBC, Stepanavan, LORI, Armenia	XYZ100	No results by the end of first month of treatment.	<a href="#">Edit</a>
Culture Status At Start Missing	56CF423D	56BAE92D 56BAE931	STEPANAVAN TBC, Stepanavan, LORI, Armenia	XYZ100	There is no bacteriology culture for this patient in the first month	<a href="#">Edit</a>
Outcome Missing	56CF42D8	56BAE9AA 56BAE9AE	NTCC DR UNIT 4, Abovyan, KOTAYK, Armenia	XYZ101	The Outcome - End of Treatment form is not filled	<a href="#">Edit</a>
Baseline Firstline DST missing	56CF437F	56BAEA1F 56BAEA23	NTCC DR UNIT 4, Abovyan, KOTAYK, Armenia	XYZ102	No results by the end of first month of treatment.	<a href="#">Edit</a>
Baseline Firstline DST missing	56CF442C	56BAEAC2 56BAEAC6	VAGARSHAPAT POLICLINIC TBC, Echmiadzin, ARMAVIR, Armenia	XYZ103	No results by the end of first month of treatment.	<a href="#">Edit</a>

# Additional data cleaning

- Additional data cleaning could be performed using export datasets
- All data entered into TB-EMR are exported into encoded csv files
- Use of external statistical package to program data cleaning





























# Example of additional data cleaning

Form	Original rule name	Type of query	Rule definition
Outcome - endTx	Missing date of end of treatment	Missing data	Flag if patient has a completed {outcome form}, but [Tx end date] is missing.
Outcome - endTx	Missing date of end of treatment decision	Missing data	Flag if patient has a completed {outcome form}, but [Tx outcome decision date] is missing.
Outcome - endTx	Missing date of death	Missing data	Flag if [tx outcome]==death, but [date of death] is missing.
Outcome - endTx	Missing cause of death	Missing data	Flag if [tx outcome]==death, but [cause of death] is missing.
Outcome - endTx	Missing reason for treatment failure	Missing data	Flag if [Tx outcome]==failed, but [reason for tx failure] is missing
Outcome - endTx	Missing reason for interruption	Missing data	Flag if [Tx outcome]==LTFU, but [reason for tx interruption] is missing
Outcome - endTx	Missing transfer out	Missing data	Flag if [Tx outcome]==Not evaluated, but no response to "did patient transfer out?"

# Exports from TB-EMR

- All data exported in csv files
- DRTB registration number to link all datasets
- All data coded according to the metadata
- Metadata is the source document to work with the export
- Prepare your datasets for analysis

 ae_form_other_causal_factor	Microsoft Excel Comma S...	1 KB	No
 ae_form_other_causal_factors_relat...	Microsoft Excel Comma S...	1 KB	No
 ae_form_related_test_result	Microsoft Excel Comma S...	14 KB	No
 ae_form_tb_drug_treatment	Microsoft Excel Comma S...	7 KB	No
 audiology_template	Microsoft Excel Comma S...	13 KB	No
 bacteriology_concept_set	Microsoft Excel Comma S...	20 KB	No
 bacteriology_culture_results_details	Microsoft Excel Comma S...	18 KB	No
 bacteriology_dst_result_details	Microsoft Excel Comma S...	4 KB	No
 bacteriology_dst_with_mic	Microsoft Excel Comma S...	1 KB	No
 bacteriology_hain_testpcr_results	Microsoft Excel Comma S...	4 KB	No
 bacteriology_other_drug_details	Microsoft Excel Comma S...	1 KB	No
 bacteriology_sequencing	Microsoft Excel Comma S...	1 KB	No
 bacteriology_smear_microscopy_te...	Microsoft Excel Comma S...	22 KB	No
 bacteriology_xpert_test_results	Microsoft Excel Comma S...	2 KB	No
 baseline_disease_site	Microsoft Excel Comma S...	2 KB	No
 baseline_drugs_used_in_arv_treatm	Microsoft Excel Comma S...	1 KB	No
 baseline_known_drug_allergies	Microsoft Excel Comma S...	2 KB	No
 baseline_list_of_drugs_taken_for_m...	Microsoft Excel Comma S...	6 KB	No
 baseline_method_of_mdr-tb_confir...	Microsoft Excel Comma S...	3 KB	No
 baseline_other_drug_taken_for_mor...	Microsoft Excel Comma S...	1 KB	No
 baseline_other_pre-existing_disease	Microsoft Excel Comma S...	3 KB	No
 baseline_past_tb_treatment_drug_r...	Microsoft Excel Comma S...	7 KB	No
 baseline_past_tb_treatment_table	Microsoft Excel Comma S...	5 KB	No
 baseline_template	Microsoft Excel Comma S...	9 KB	No
 depression_alcohol_score_template	Microsoft Excel Comma S...	1 KB	No
 documents template	Microsoft Excel Comma S...	1 KB	No

# Exports from TB-EMR

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id_baseline	regnum	d_baseform	civil	civil_oth	homeless	employ	employ_ot	refugee	prison	prison_wh	health_wc	alcohol_ba
2	1	CUE007	18/Apr/2014	3		0	3		0	0		0	0
3	304	CUE009	12.mars.16	3			4		0	1	1	0	0
4	1963	NTRM003	15/Jul/2015	1			4		0	0			0
5	2413	CUE006	29.janv.16	3					0	0		0	0
6	2874	CUE003	04/Apr/2014	3		0	3		0	0		0	0
7	6721	CUE001	05.nov.15	1		0			0			0	0
8	8513	CUE004	20.janv.16	4		0	4		0	0		0	1
9	10341	NTRM015	08/Apr/2016	1		0	4		0	0		0	1
10	11508	NTRY003	12/May/2015	5					0	0		0	0
11	11685	NTRM014	13/Apr/2016			0	2		0	0		0	0
12	13521	NTRM011	12/Feb/2016	1		0	4		0	0		0	0
13	15170	NTRM012	15/Feb/2016	1		0	4		0	0		0	0
14	15468	NTRY009	29.sept.15	1			4		0	0		0	1
15	16597	NTRM013	07/Apr/2016	1		0	1		0	0		0	0
16	19060	NTRY015	06/May/2016	1		0	4		1	0		0	1
17	20901	NTRY012	26/Feb/2016	1		0	5		0	0		0	0



# Metadata

heart			1.Baseline	This form is filled by ?				
Q No.	Core vs. Optional (C/O)	Organization(s) collecting this optional variable	Question Text	Short name	header	Data Type	Appearance	Options
Social history								
1	C		Date of baseline	Date baseline	d_baseform	date	choose date	
2	C		Marital status:	marital status	civil	numeric	check 1 option	1: married 2: living together 3: Single 4: divorced 5: widowed 6:separated 7: other
3	C		If other Marital status, specify:	other marital status	civil_oth	text	free entry	
4	C		Homeless within the past year?	homeless	homeless	numeric	check 1 option	0=No, 1=Yes, 99=unknown
5	C		Current employment status:	employed	employ	numeric	check 1 option	1: employed 2: unable to work 3: student 4: unemployed 5: housework 6: pensioner 7: other
6	C		If other employment status, specify	other employment	employ_oth	text	free entry	
7	C		Refugee, displaced person or migrant?	refugee, migrant,	refugee	numeric	check 1 option	0=No, 1=Yes, 99=unknown
8	C		Have you ever been in prison?	prison	prison	numeric	check 1 option	0=No, 1=Yes, 99=unknown
9	O	MSF	If yes,	If prison: past or present?	prison_when	numeric	check 1 option	1=Currently, 2=In the past
10	C		Have you ever been a health worker?	health worker	health_worker	numeric	check 1 option	0=Never, 1=In the past, 2=Currently, 99=unknown

# Metadata

s			4a. Outcome - End of Treatment			This form is filled by ?			
Q No.	Core vs. Optional (C/O)	Organization(s) collecting this optional variable	Question Text	Short name	header	Data Type	Appearance	Options	Length/Format
1	C		End of Treatment date	End of treatment	d_txend	date	choose date from calendar		dd/mm/yyyy
2	C	MSF, IRD	End of Treatment Outcome date	Treatment outcome	d_outdecision	date	choose date from calendar		dd/mm/yyyy
	C		Outcome:	Outcome	outcome	numeric	check 1 option	1=Cured, 2=completed, 3=Died, 4=Failed, 5=LTFU, 6=Not evaluated, 10=treatment adapted,	
3									
4	C		If Died: Date of death:	Date of death	d_death	date	choose date from calendar		dd/mm/yyyy
5	C		If Died: Suspected Primary cause of death:	Cause of death	deathcause	numeric	check 1 option	1=TB immediate cause of death, 2=Tb contributing to death 3=Surgery related death (type of surgery:___), 4=Cause other than TB (suspected cause:___), 5=Cause related to	
			If Died: surgery related:	Type of	surgerydeath	text	free entry		

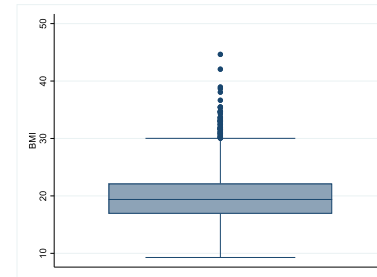
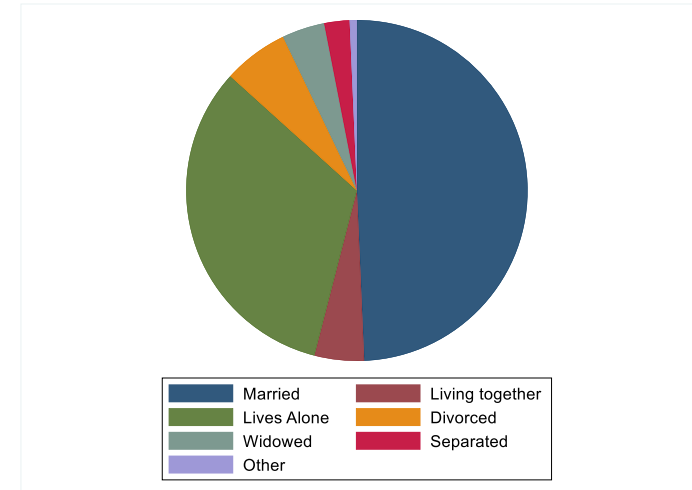
# Example of data imported into Stata

- Dates formatted
- Variable with categories coded and labeled
- Additional data management needed to create variable of interests

	regnum	d_baseform	civil	homeless	employ	refugee	prison
1	CUE001	05nov2015	Married	False	.	False	.
2	CUE002	29jan2016	Lives Alone	False	Unemployed	True	False
3	CUE003	04apr2014	Lives Alone	False	Student	False	False
4	CUE004	20jan2016	Divorced	False	Unemployed	False	False
5	CUE005	20jan2016	Lives Alone	False	Unemployed	False	False
6	CUE006	29jan2016	Lives Alone	.	.	False	False
7	CUE007	18apr2014	Lives Alone	False	Student	False	False
8	CUE008	29jan2016	Married	.	Unemployed	False	False
9	CUE009	12mar2016	Lives Alone	.	Unemployed	False	True

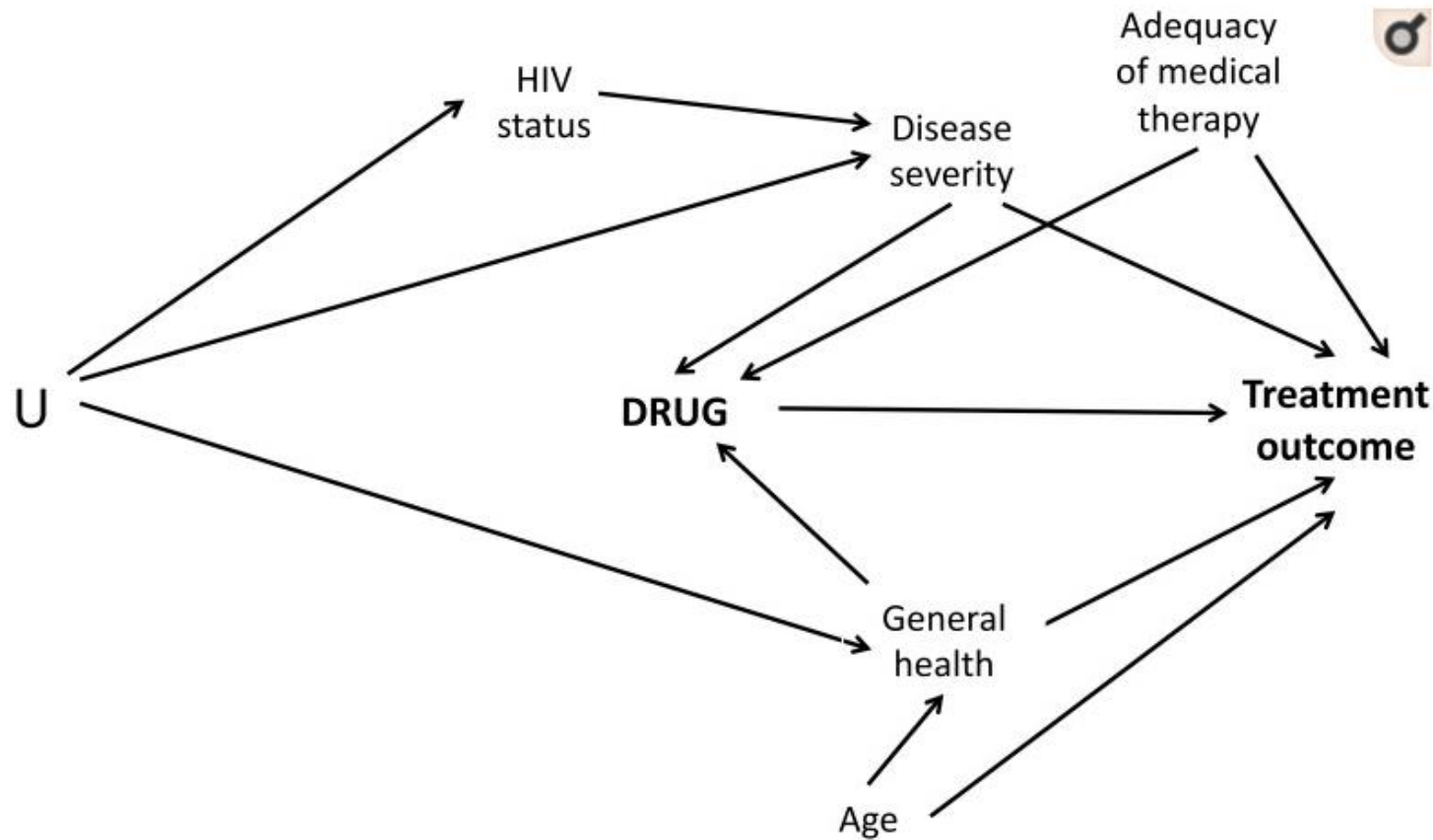
# Data analysis

- Analysis plan
  - Population
  - Definitions
  - Outcome of interest
- Adapted statistical methods to account for observational nature of the data
- Causality, regimen changes, missing data, unobserved confounders etc..



Marital status:	Freq.	Percent	Cum.
Married	1,302	49.32	49.32
Living together	124	4.70	54.02
Lives Alone	863	32.69	86.70
Divorced	163	6.17	92.88
Widowed	107	4.05	96.93
Separated	63	2.39	99.32
Other	18	0.68	100.00
Total	2,640	100.00	

# Confounding example



Directed Acyclic Graph describing the covariates affecting treatment outcomes for multi-drug resistant tuberculosis.

Legend: U = Unmeasured confounder; HIV = Human immunodeficiency virus.

**Citation:** Fox GJ, Benedetti A, Mitnick CD, Pai M, Menzies D, The Collaborative Group for Meta-Analysis of Individual Patient Data in MDR-TB (2016) Propensity Score-Based Approaches to Confounding by Indication in Individual Patient Data Meta-Analysis: Non-Standardized Treatment for Multidrug Resistant Tuberculosis. PLoS ONE 11(3): e0151724. doi:10.1371/journal.pone.0151724



## Reporting of observational studies



# ***STROBE Statement***

Strengthening the reporting of observational studies in epidemiology

---

## Observational Studies: Getting Clear about Transparency

The PLOS Medicine Editors 

Published: August 26, 2014 • <https://doi.org/10.1371/journal.pmed.1001711>

<https://www.strobe-statement.org>

## Reporting of observational studies



### *A Validated Checklist*

for Evaluating the Quality of Observational  
Cohort Studies for Decision-Making Support

#### GRACE: Good ReseArch for Comparative Effectiveness

The GRACE Checklist is designed for the assessment of observational studies of comparative effectiveness in terms of their quality and usefulness for decision-making. The checklist was developed from a review of the literature with guidance from recognized experts in this field. The content includes questions about data and methods. Validation activities have documented the usefulness of all 11 questions in this checklist. Approaches to scoring are under consideration.

<https://www.graceprinciples.org/>

# Challenges from the context

- Lack of **routine supervision** in some programs
- Insufficient **training** or no regular meetings
- No **data quality indicators** for routine monitoring
- Limited **skills for analysis (including knowledge of statistical package)** or data interpretation at the district
- High **rotation of staff** at district and health facility
- No **SOPs for data cleaning** at any level

# Challenges from the data

- Multiple variables from multiple **data sources**
- **Errors** with upload/**internet connectivity**
- No systematic mechanism for **reporting bugs**
- No **mandatory unique identifier** or incomplete in most programs
- No **linkage with laboratory** data (various formats)

**Thank you**