Exploratory Data Analysis of the Ebola dataset

Ridha Chahed[1] and Mary-Anne Hartley[1]

[1] *intelligent Global Health, Machine Learning and Optimization Laboratory, EPFL.*

Abstract

This report summaries the analysis of the Ebola data assembled and curated by the Infectious Diseases Data Observatory (IDDO). We thank all members of the research, health and humanitarian communities who participated in the data collection.

Keywords: Data, Ebola, health, epidemic

1 Demographics

The data set consists of individual level data from **9472** individuals from **7** different studies at **14** sites across **3** countries. See Figure 1

Table 1. Summary of the cohort origin

Study Identifier	Contributor	Country	City	Number of patients
EJPDEJ	Médecins Sans Frontières (MSF)	Guinea	Conakry	2301
EOPNOJ	Alliance for International Medical Action (ALIMA)		Nzerekore	147
EORKWS	Oxford University	Sierra Leone	Port Loko	35
EQJJGF	Médecins Sans Frontières (MSF)	Liberia	Monrovia	1909
ERFCVU	International Medical Corps (IMC)	Liberia	Bong County	550
			Margibi County	292
		Sierra Leone	Makeni	1085
			Lunsar	549
			Kambia	273
ESYADD	Save the Children International (SCI)	Sierra Leone	Kerry Town	456
EUZJTB	Médecins Sans Frontières (MSF)	Sierra Leone	Kailahun	1189
			Bo	529
			Magburaka	157

Table 1 gives us details about the data processed by IDDO. By comparing it to the inventory we should have **13562** individuals, we have therefore **4090** missing individuals. We have two sources of errors, the first one is that some studies are missing from the curated data (EGOYQN, EBOPHA, ESBMRS, EIXUZQ, EPGLFV, EFFVXT) and the other one is that for some studies the count between the inventory and the curated data is different. The latter discrepancy is summarized in in Table 2

Of the 9420/9472 (n=52 missing) individuals in whom the sex is known, **53, 9%** are male (n=5082/9420). The median age was **29** years (range 0-102). An odd value of −1 has been observed. In Figure 2 we observe peaks at some rounded number like 25,30,40 which may be due to the fact that sometimes the medical personal needs to approximate the patient's age.

Regarding the mortality, **2538** died during the study period giving a death rate of **26, 8%**.

Table 2. Differences between curated data and inventory

Study Identifier	City	Curated data	Inventory	Difference
EGOYQN	Guéckédou	0	2500	-2500
EBOPHA	Monrovia	0	4	-4
ESBMRS	Donka	0	102	-102
EIXUZQ	Foya	0	870	-870
EPGLFV	Freetown	0	171	-171
EFFVXT	Donka	0	418	-418
EQJJGF	Monrovia	1909	1907	2
EUZJTB	Kailahun	1189	1219	-30
	Bo	529	524	5
	Magburaka	157	159	-2
Total				-4090

Chronologically the studies span from March 2014 to October 2015. The Reference Start Date variable describes the date of the start of the Subject Reference Period. The Subject Reference Period is defined by IDDO as starting with the subject's first study encounter and ending with the subject's final study encounter. RFSTDTC corresponds with the time and date of the subject's first study encounter (e.g., screening, enrollment, admission). This date will be used to calculate the relative days in other variables.

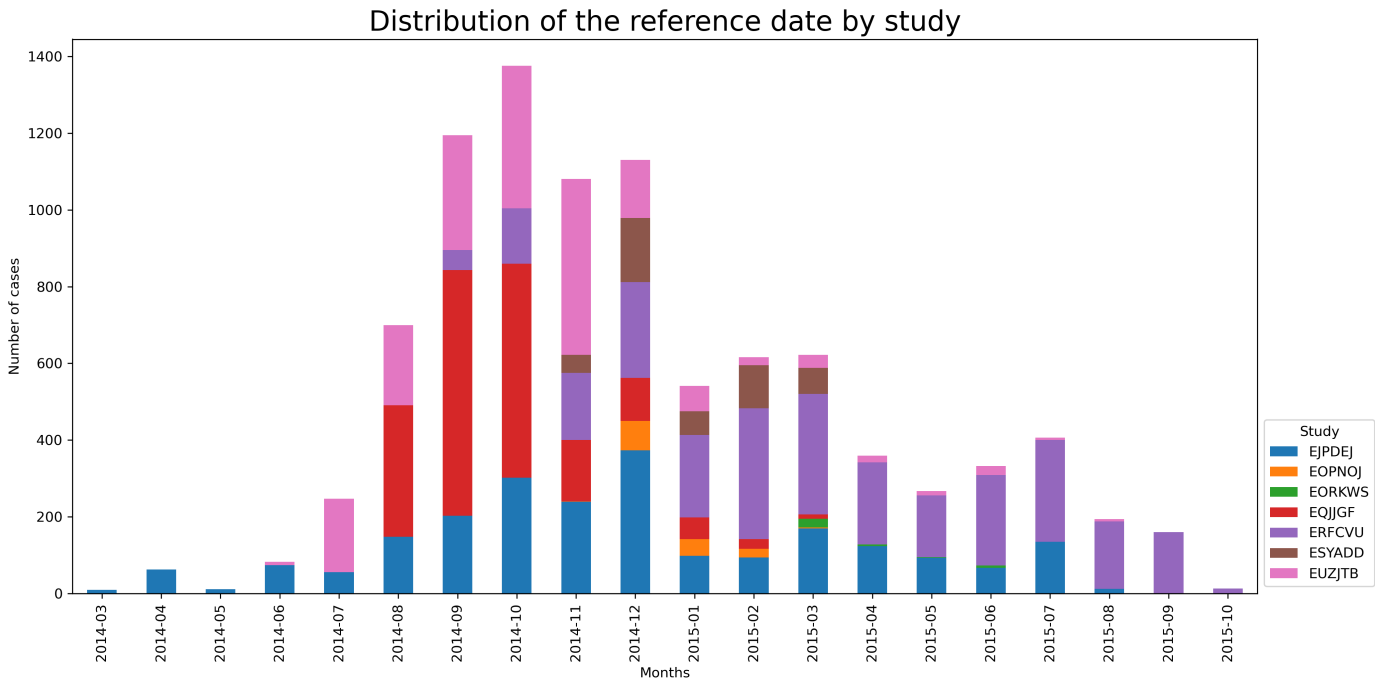

Figure 1. Distribution of the reference date per study

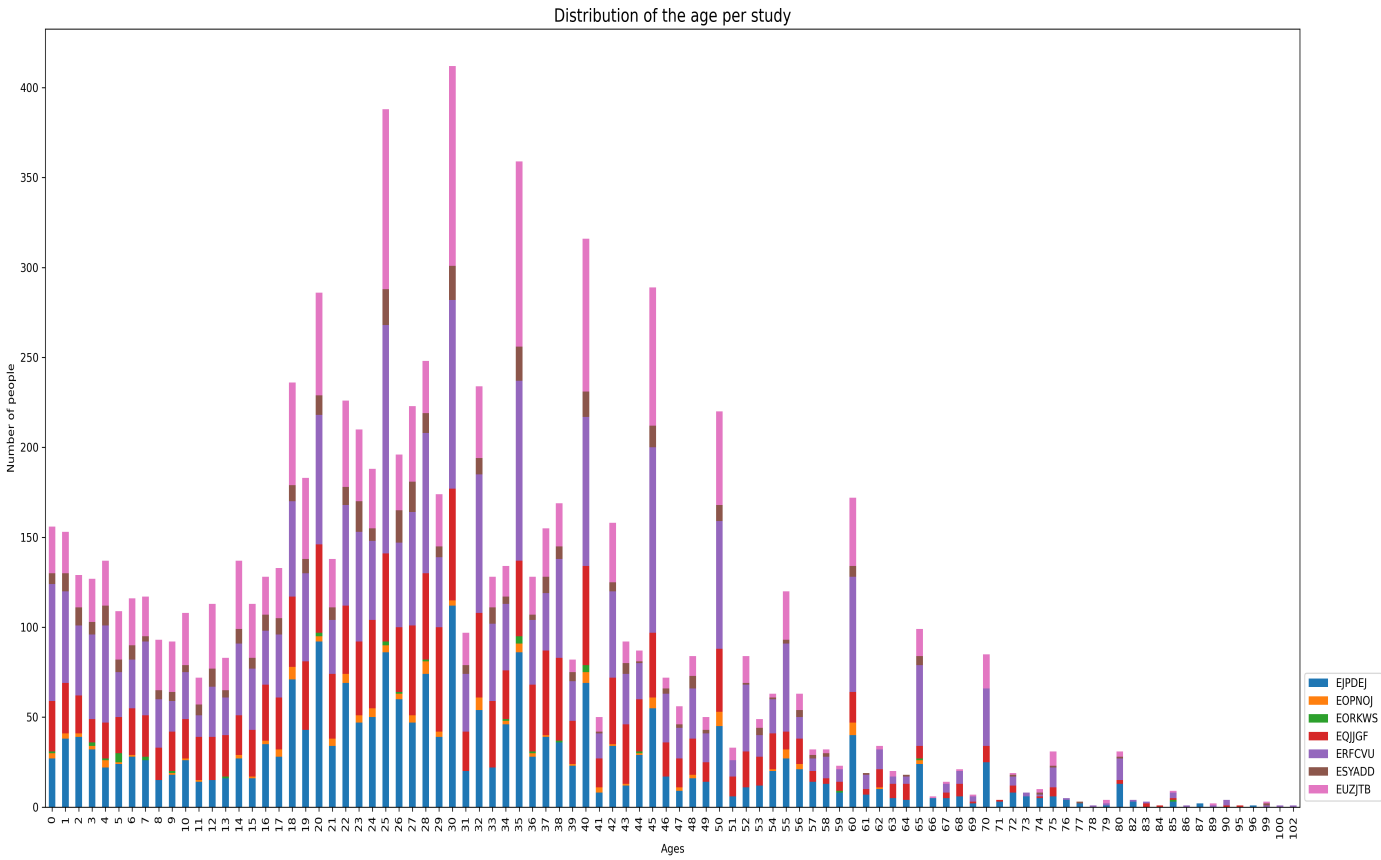

Figure 2. Distribution of the age per study

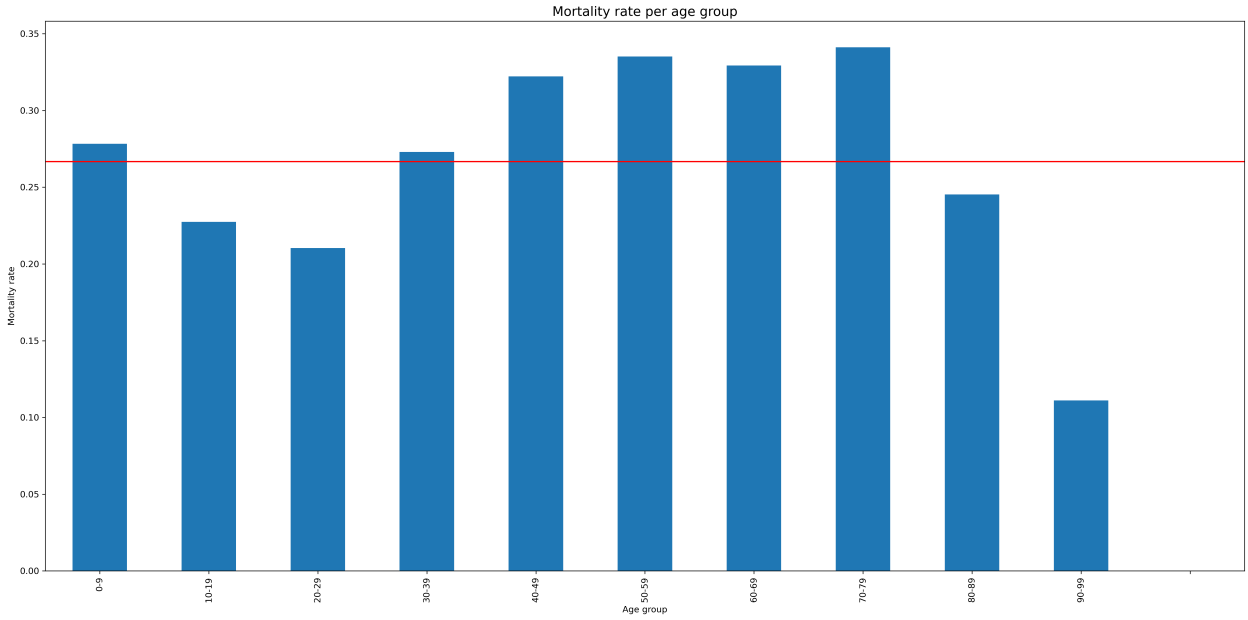

Figure 3. Mortality rate per group

2 Death Details

We have **1028** records of death details for **661** individuals from **6** studies. See Table 3.

Table 3. Death Details

Study Identifier	Number of records	Number of individuals
EJPDEJ	1	1
EOPNOJ	15	12
EORKWS	26	13
EQJJGF	48	48
ERFCVU	808	457
ESYADD	130	130

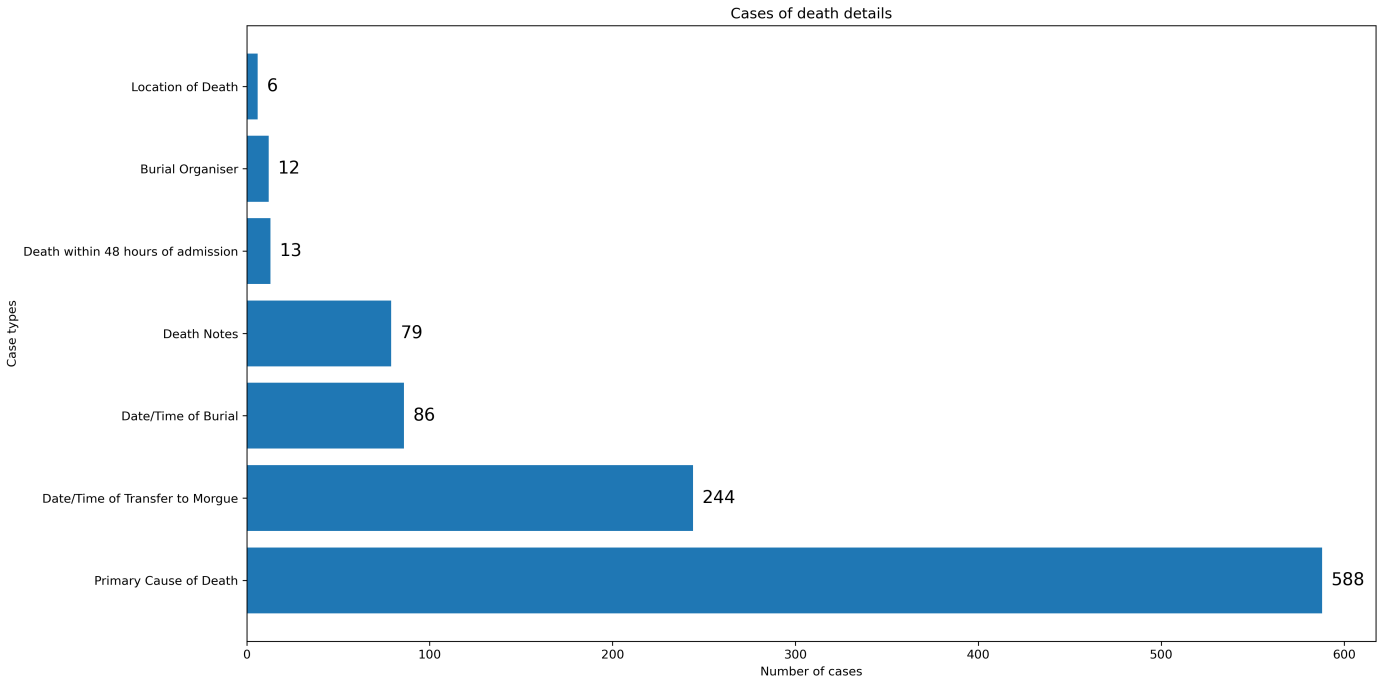

Figure 4. Short name of the death details

We can see that Primary Cause of Death is the most present death detail. If we look at the results of these primary cause of death we find that **68%** of them are related to Ebola.

3 Disposition

We have **10317** records of disposition for **8845** individuals from **6** studies. See Table 4.

We can have up to four records per individual. The reported term of the event is missing for only 27 records. Nonetheless, this feature is not very useful as it's not harmonized, the same event being described with different words. To solve this issue, several terms are grouped under the same Modified Reported Term for the Event. This process is not done correctly as it has 1404 missing values. We decide to correct it.

Table 4. Disposition

Study Identifier	Number of records	Number of individuals
EJPDEJ	2234	2234
EOPNOJ	147	147
EORKWS	39	35
EQJJGF	1907	1907
ERFCVU	4081	2647
ESYADD	1909	1875

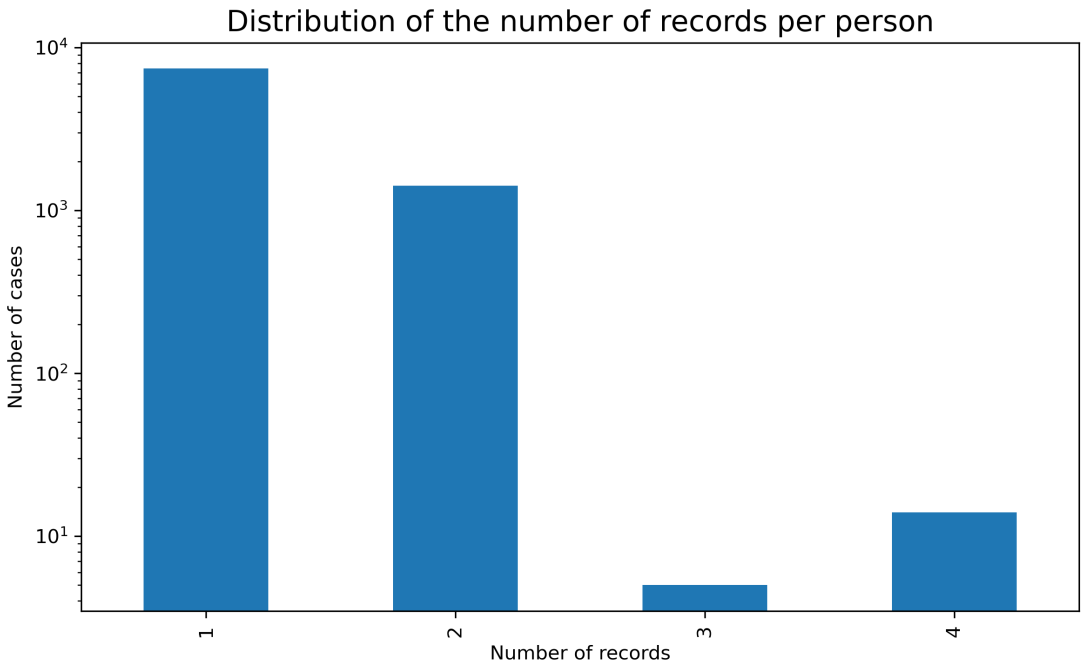

Figure 5. Distribution of the number of records per patient

Table 5. Imputation of Modified Reported Term for the Event

Reported term	Number of missing	New modify reported term
Death (positive, negative or unknown)	625	deceased
Not Recorded	559	unknown outcome
Deceased/Expired	85	deceased
Transferred	67	transferred
Defaulter/Escaped	32	escaped
Dead/Expired	19	deceased
Recovered/Cured	4	recovered
Defaulter	3	escaped
COMPLETED	3	discharged
DISCHARGED	3	discharged
SURVIVED	2	discharged recovered
LOST TO FOLLOW-UP	1	unknown outcome
Unknown	1	unknown outcome

We have a feature called Standardized Term that contains a dictionary-derived text description of the event. This is sometimes covered by CDISC Controlled Terminology and sometimes by IDDO Controlled Terminology. However, we have a lot of inconsistencies with this feature. First of all, it's composed of 6 categories but we suggest to remove two of them namely LOST TO FOLLOW-UP and COMPLETED that only have respectively 1 and 3 records. Also the same reported event can be found in two different standardized terms for the sake of example we have 978 cases of DISCHARGED in the standardized category OTHER but also 111 same cases are in RECOVERY. We decide to modify this categorization.

Table 6. Modification of the Standardized Term

Reported term	Old standardized term	New standardized term	Count
COMPLETED	COMPLETED	RECOVERY	3
LOST TO FOLLOW-UP	LOST TO FOLLOW-UP	OTHER	1
SURVIVED	OTHER	RECOVERY	1749
DISCHARGED	OTHER	RECOVERY	978
Not A Case (discharged)	OTHER	RECOVERY	510
Recovered/Cured	OTHER	RECOVERY	502
Sorti_négatif	OTHER	RECOVERY	60
DECEASED	OTHER	DEATH	3
DIED ON ARRIVAL	OTHER	DEATH	5

We now obtain the following distribution.

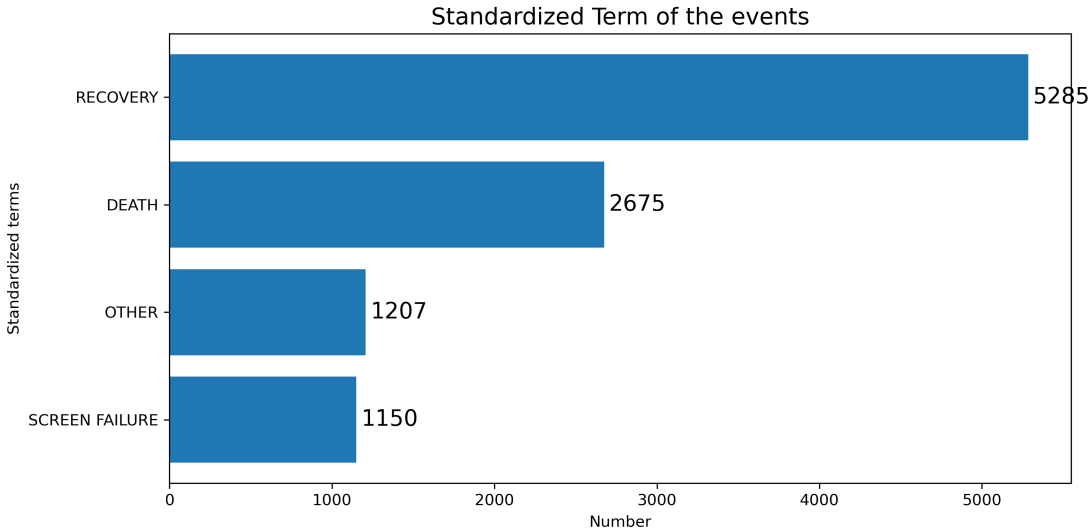

Figure 6. Distribution of the Modified Standardized Term

Chronologically we have two features Study Day of Observation/Collection (DSDY) and Study Day of Start of Observation (DSSTDY) which have a correlation of 1. DSDY have a lot of missing values (60%) whereas DSSTDY only have 5% of missing values. A major problem is that we observe negative values that we don't clearly understand for the DSSTDY feature.

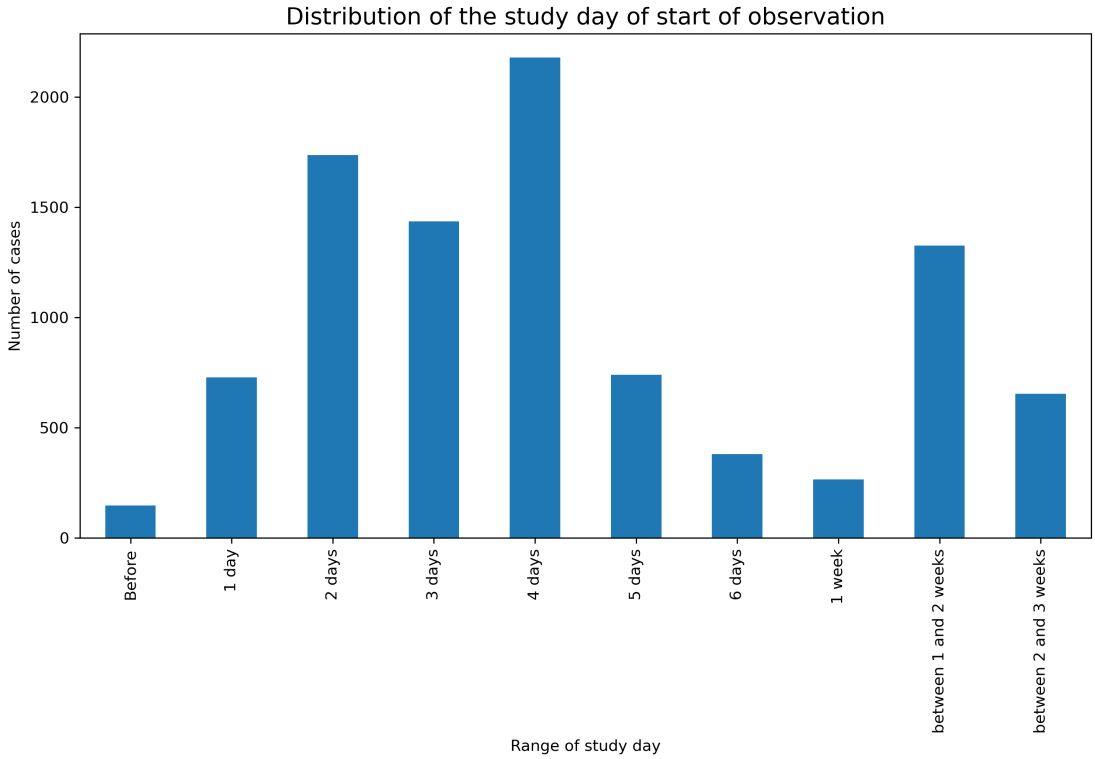

Figure 7. Distribution of the Study Day of Start of Observation

There is an error in the duration computation, and it can perhaps be resolved by releasing the date of the the Ebola test (or other dates).

4 Treatments and Interventions

We have **1195300** records of treatments and interventions for **5980** individuals from **6** studies. See Table 7.

Table 7. Treatments and Interventions

Study Identifier	Number of records	Number of individuals
EJPDEJ	223329	2301
EOPNOJ	6764	136
EORKWS	675	17
EQJJGF	185757	965
ERFCVU	770567	2105
ESYADD	8208	456

5 Disease Response and Clinical Classification

We have **57782** records of disease response and clinical classification for **4892** individuals from **5** studies. See Table 8.

Table 8. Disease Response and Clinical Classification

Study Identifier	Number of records	Number of individuals
EOPNOJ	301	147
EORKWS	419	17
ERFCVU	12470	2397
ESYADD	912	456
EUZJTB	43680	1875

In the test name (RSTEST) we have found a typo 'AVPU0101 - Responsiveness' should be 'AVPU01-Responsiveness', we correct it and get thus 3 categories of test. For the Result or Finding in Original Units (RSORRES) feature we have more than 28 different categories but it's not harmonized so by grouping equivalent ones we get down to 11.

Table 9. Distribution after modification of the Results and Findings

Test name	Result or finding	Count
AVPU01-Responsiveness	Alert	10410
	Verbal stimulus	349
	Physical stimulus	183
	Unresponsive	166
EVDNCL-Notification Classification	Suspected	559
	Highly Suspected	470
	Confirmed	433
	Dead on arrival	7
	Information and Prevention Visit	6
EVDFCL-Final Classification	Other	2002
	Ebola	553

6 Clinical and Adverse Events

We have **2161358** records of clinical and adverse events for **8966** individuals from **6** studies. See Table 10.

Table 10. Clinical and Adverse Events

Study Identifer	Number of records	Number of individuals
EJPDEJ	531832	2301
EOPNOJ	134526	147
EORKWS	1741	17
EQJJGF	245256	1909
ERFCVU	245256	2717
EUZJTB	819401	1875

Chronologically, we still have the problem of negative values for the Study Day of Observation/Collection (SADY) feature.

Reported term for the event (SATERM) describes the symptoms. We have 603 different categories that are not standardized and are really noisy with typos and discrepancies. Just by applying a lower case transformation we are down to 572 categories. An attempt to harmonize this feature is the Modified Reported Term for the Event (SAMODIFY) features but 69% of the data for this feature is missing, we need to correct that.

We face several issues :

- We have some SATERM that don't have SAMODIFY features due to the fact that they were written in UPPER case.
- We also have some SATERM that don't have SAMODIFY features even though that corresponds to an existing one.
- Finally we have some SAMODIFY terms that are redundant and need to be merged.

After correcting these three issues, we end up with only 2% of missing that remains because the corresponding SATERM is only mentioned once, so it can be imputed but it would take a lot of time. At the end we come up with 61 categories of SAMODIFY and as we can Figure 8 the same category can be found across different studies.

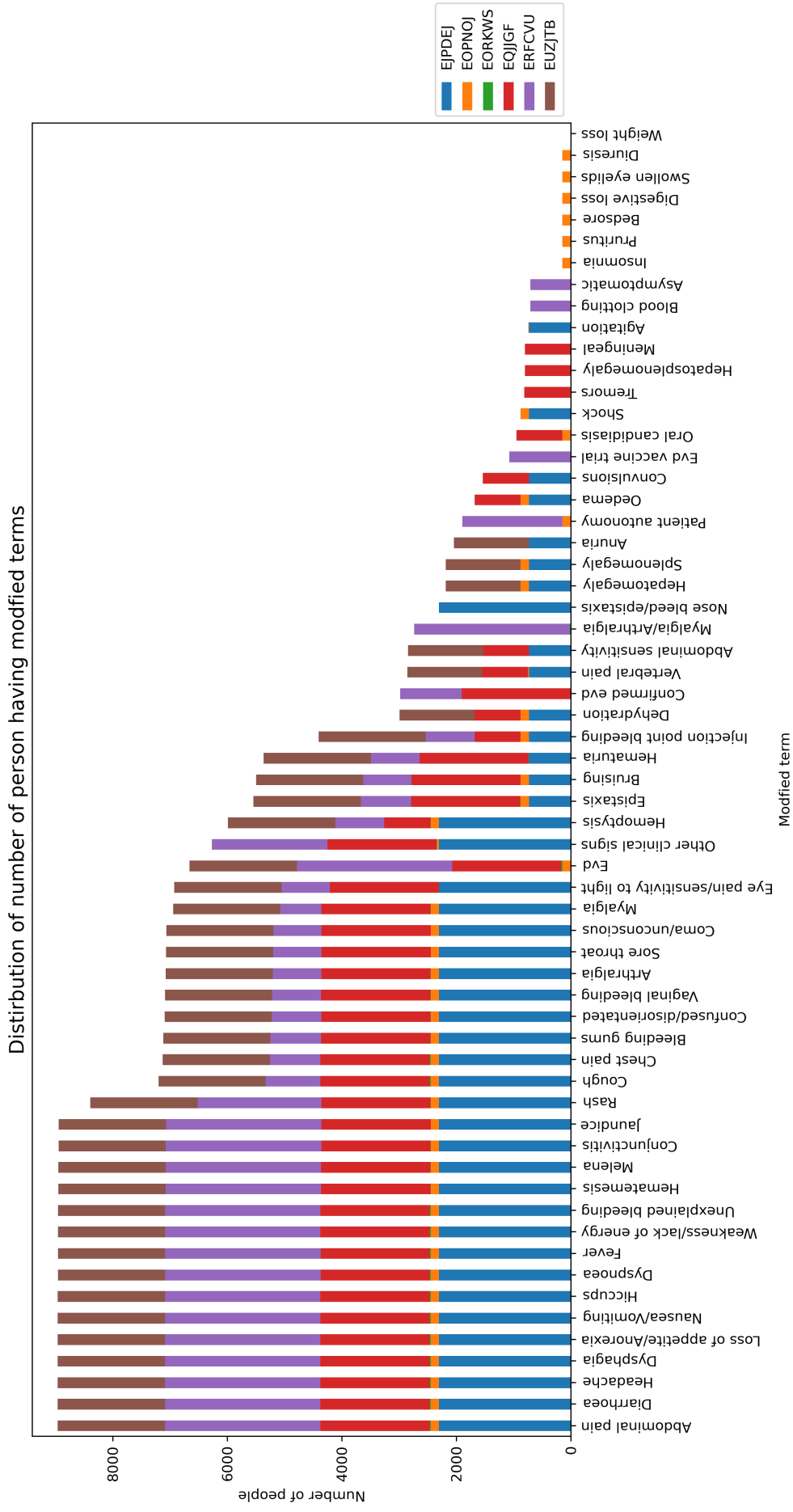

Figure 8. Distribution of the Modified Reported Term for the Event