

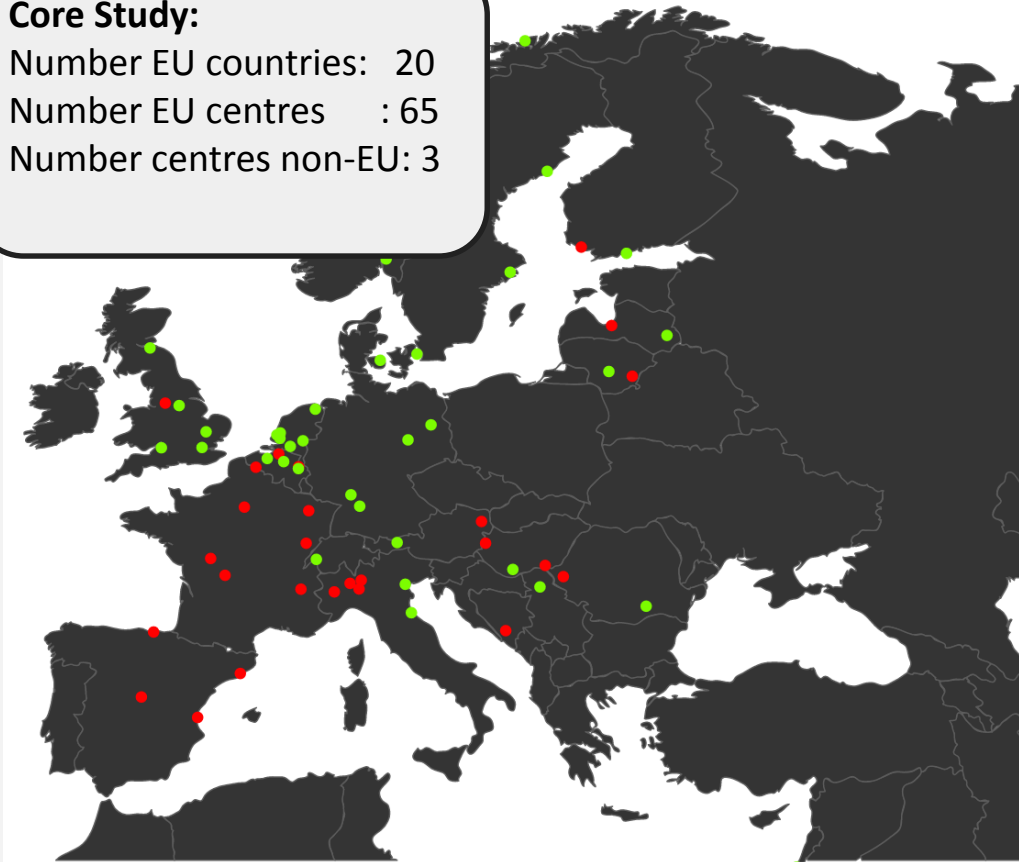
CHALLENGES AND SOLUTIONS IN DATA CURATION: THE CENTER EXPERIENCE

**Andrew I.R. Maas on behalf of the CENTER Collaborators
International Initiative for Traumatic Brain Injury Research – 2017
6th Annual Meeting; October 30 – 31, Bethesda, MD**

Evidence Generation

Core Study:

Number EU countries: 20
Number EU centres : 65
Number centres non-EU: 3



FPI: 19 December 2014

Current Recruitment Status

CORE Data Collection	5,041
-----------------------------	--------------

ER	1,056
----	-------

ADM	1,533
-----	-------

ICU	2,462
-----	-------

Registry

EU	21,724
----	--------

China	13,624
-------	--------

India	2,568
-------	-------

Recruitment end: December 1st 2017

ER: Discharge from ER

Adm: Primary admission to ward

ICU: Primary admission to ICU

Background

- CENTER-TBI dataset is really rather complicated
- Multiple centres, multiple countries cf Framingham
- >5000 patients (not counting registry)
 - Heterogeneous patients: ICU, ED, ward
 - Heterogeneous data: Clinical, outcomes, interventions, imaging, lab... cf '100,000 genomes'
- Individually 'quite big' but together unparalleled

2,641

Unique data variables

But then:

- Multiple timepoints
- Derived variables
- Standardization of all date/timevariables
- DICOM images and structured reports
- Biomarkers/genomics

X 5,400

Background

- Two problems:
 - How do we make this data usable?
 - Not surprising given onerous (human) data collection task that some data quality errors have crept in
- INCF / data curation / SWAT team have been working hard
 - To improve the data quality
 - To make tools to help researchers access the data in a (reasonably) manageable way

Data Curation

Three main streams:

- **Data Completeness**
- **Data Quality**
- **Extraction Tool for user-friendly access**

Data Completeness

October 2015:

- Concerns about substantial number of missings
- Actions to review missingness implemented

Missing Hypoxia

Date	15/10/2015	30/05/2016	18/04/2017	06/09/2017
Patients	N=1148	N=2473	N=4070	N=4447
Total (N/%)	265(23,1%)	224(9,1%)	155(3,8%)	341 (7,7%)

Missing Hypotension

Date	15/10/2015	30/05/2016	18/04/2017	06/09/2017
Patients	N=1148	N=2473	N=4070	N=4447
Total (N/%)	270(23,5%)	229(9,3%)	153(3,8%)	302 (6,8%)

Improving on Missingness

Missing Vomiting

Date	28/09/2015	30/05/2016	18/04/2017	06/09/2017
Patients	N=984	N=2473	N=4070	N=4447
Total (N/%)	218(22,2%)	262(10,6%)	241(5,9%)	505 (11,4%)

Missing cause of injury

Date	28/09/2015	30/05/2016	18/04/2017	06/09/2017
Patients	N=984	N=2473	N=4070	N=4447
Total (N/%)	167(17,0%)	184(7,4%)	117(2,9%)	198 (4,5%)

Follow-up rates in survivors

3mo Follow Up rates – Combined GOSE

Date	05/09/2016	06/01/2017	02/05/2017	03/09/2017
Patients	2228	2925	3471	3953
Total (N/%)	1331 (67%)	1712 (66%)	2083 (67%)	2429 (69%)

6mo Follow Up rates – Combined GOSE

Date	05/09/2016	06/01/2017	02/05/2017	03/09/2017
Patients	1662	2447	3040	3593
Total (N/%)	971 (66%)	1358 (62%)	1710 (63%)	2021 (63%)

So much more than 'missing'

- Database related
- Units
- Ambiguity in questions
- Language problems
- Site specific issues
- Systematic entry error
- Data hard to get
- Random entry error – but critical
- Impossible times
- Inclusion violations
- Free text
- Things we just did not think about

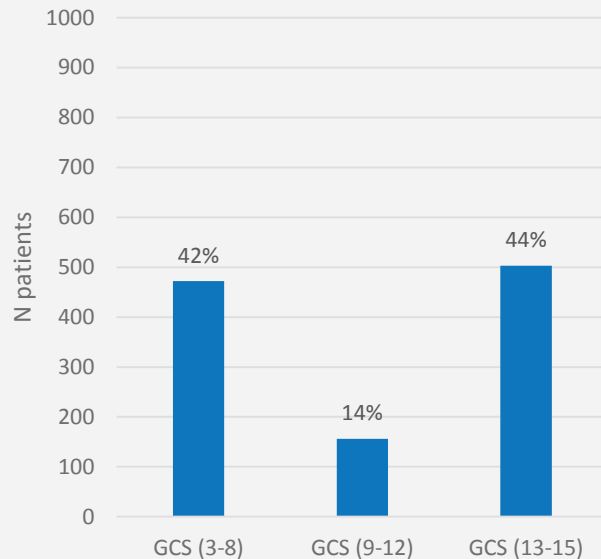
E-CRF ≠ DATABASE

Examples of curation issues:

- Inappropriate numeric rounding
- Problems with units
- Impossible dates/times
- Matching CT/MR dates/time
 - between DICOM records and e-CRF
- Intervariable datachecks

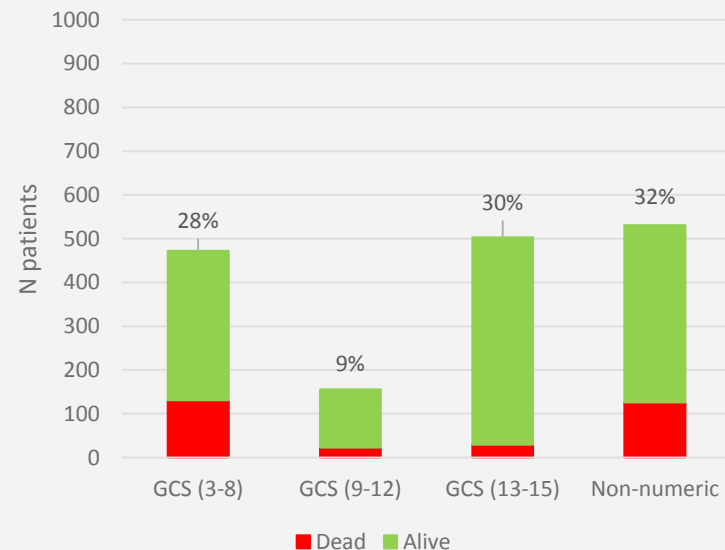
Distribution GCS SUM score in ICU stratum

GCS Sum score numeric
(ICU stratum)



	N	%
GCS (3-8)	472	42%
GCS (9-12)	156	14%
GCS (13-15)	503	44%
Total	1131	100%

GCS Sum score numeric + non-numeric
(ICU stratum)

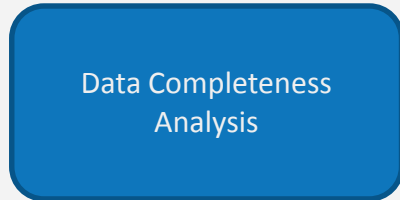


	ICU stratum (N/%)	Alive patients (N/%)	Dead patients (N/%)
GCS (3-8)	472 (28%)	341 (72%)	131 (28%)
GCS (9-12)	156 (9%)	133 (85%)	23 (15%)
GCS (13-15)	503 (30%)	474 (94%)	29 (6%)
Non-numeric	531 (32%)	405 (76%)	126 (24%)
Total	1662 (100%)	1353 (81%)	309 (19%)

DCTF

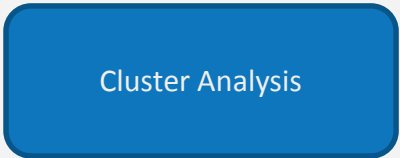
QUALITY IMPROVEMENT AND DATA CURATION COMMUNICATION FLOW

UZA

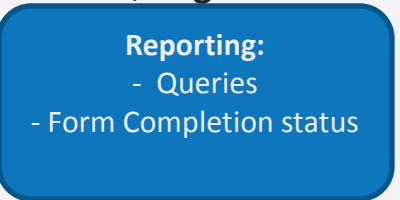


EMC/LUMC ↑

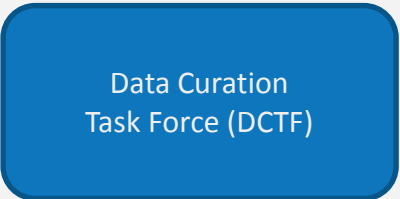
Inform frequency of analysis and reporting



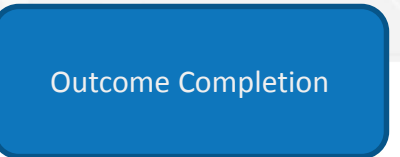
Quesgen



KI-INCF/ UCAM/ UZA



UOS/ UMG



SWAT



Supervision + Coordination



Work Flow

