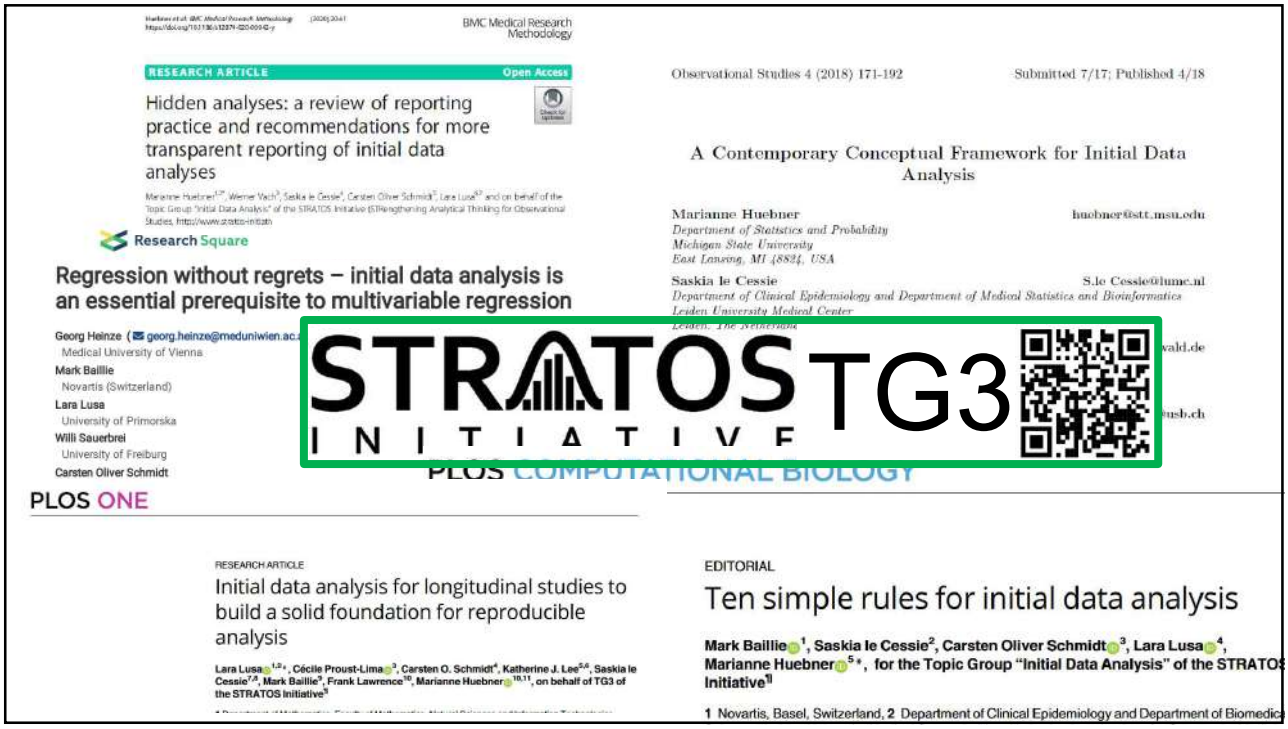

nfdi 4 health
Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
STRATOS INITIATIVE
TG3
Universitätsmedizin GREIFSWALD
Accessible structured Initial Data Analysis (IDA) and Data Quality Assessments (DQA)
Carsten Oliver Schmidt, Lara Lusa, Marianne Huebner on behalf of STRATOS TG3
DAGSTat 2025


BMC Medical Research Methodology
RESEARCH ARTICLE
Open Access
Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses
Research Square
Regression without regrets – initial data analysis is an essential prerequisite to multivariable regression
Georg Heinze
Medical University of Vienna
Mark Baillie
Novartis (Switzerland)
Lara Lusa
University of Primorska
Willi Sauerbrei
University of Freiburg
Carsten Oliver Schmidt
PLOS ONE
Observational Studies 4 (2018) 171-192
Submitted 7/17; Published 4/18
A Contemporary Conceptual Framework for Initial Data Analysis
Marianne Huebner
huebner@stt.msu.edu
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA
Saskia le Cessie
S.le Cessie@lumc.nl
Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
STRATOS INITIATIVE TG3
PLOS COMPUTATIONAL BIOLOGY
RESEARCH ARTICLE
Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis
EDITORIAL
Ten simple rules for initial data analysis
Mark Baillie[1], Saskia le Cessie[2], Carsten Oliver Schmidt[3], Lara Lusa[4], Marianne Huebner[5]*, for the Topic Group "Initial Data Analysis" of the STRATOS Initiative[1]
1 Novartis, Basel, Switzerland, 2 Department of Clinical Epidemiology and Department of Biomedica

Aim of IDA

The aim of IDA is to provide a data set and reliable findings on this data set which allows researchers to work with this data set in a responsible manner.

Huebner et al. 2018

Common (implicit) assumptions underlying analyses

1. Informative data
2. Sufficient data quantity
3. Limited missing data
4. Little systematic noise
5. Correct data labels
6. Class balance
7. Non-adversarial data
8. Independent and Identically Distributed (i.i.d.)
9. Linearity
10. Homoscedasticity
11. Non-multicollinearity
12. Consistent scaling of features
13. Sample representativeness

…….

Background information
Study design
Variable descriptions

Range violations
Contradictions
Inadmissible values

Volume
Unit / item missingness
Missing patterns
Missing mechanisms

Univariate descriptions
Multivariate descriptions
Associations

…….

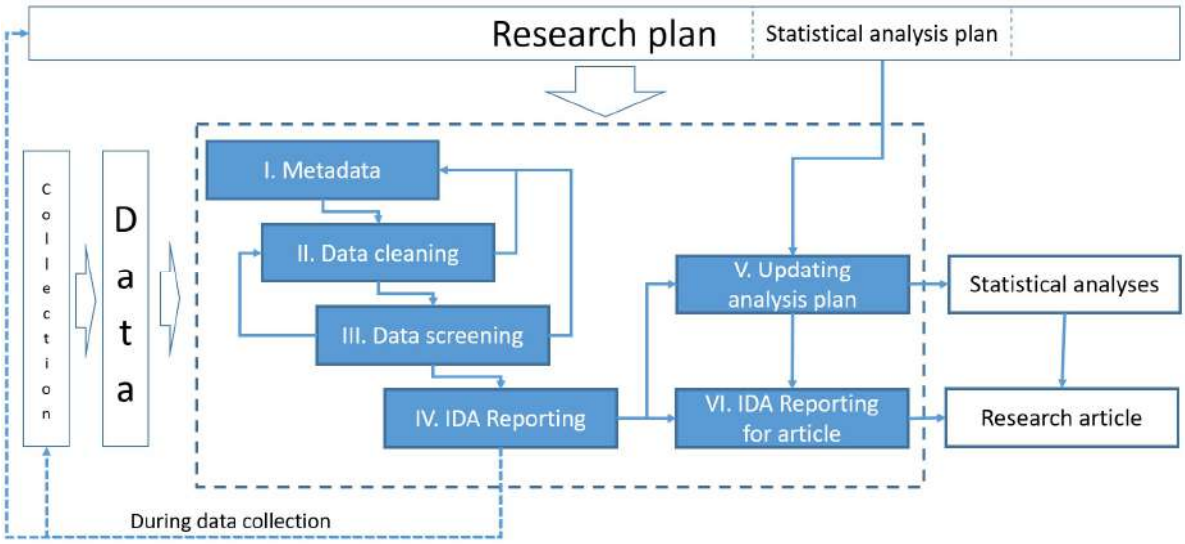
Data Quality Framework for EU medicines regulation 2023

30 October 2023
Data Analytics and Methods Task Force
EMA/326985/2023

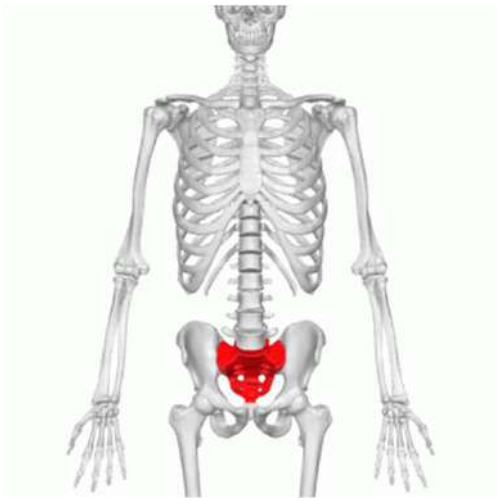
What does accessible IDA and DQA mean?

A transparent approach to planning, checking, reporting, data handling, and decision-making processes that govern the creation and use of analysis datasets as part of the Main Data Analysis (MDA).

Initial Data Analysis Workflow

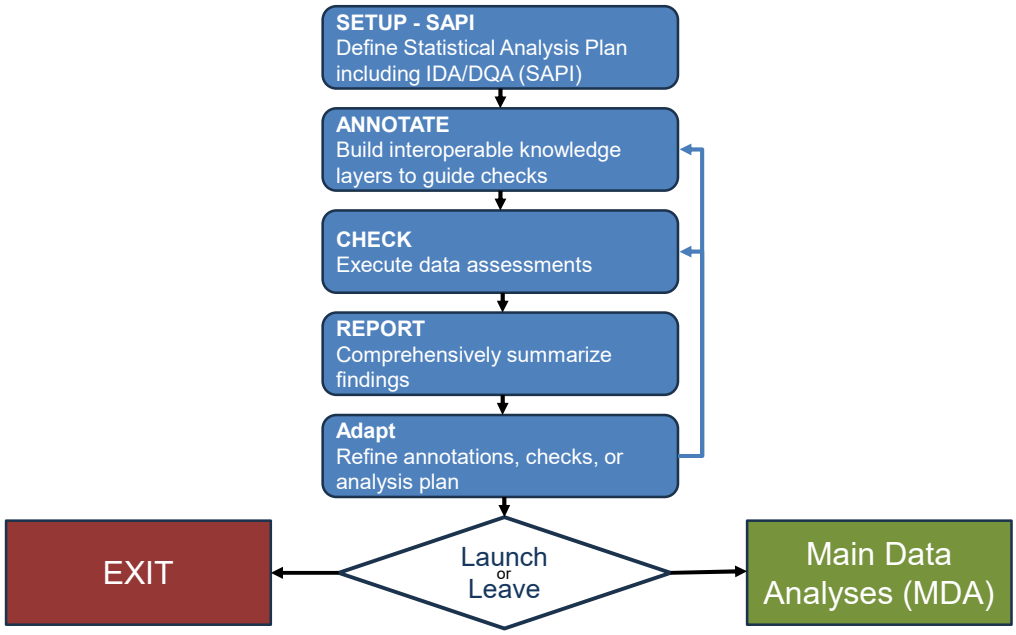


Huebner et al. 2018



https://en.wikipedia.org/wiki/Sacrum

The SACRAL-Cycle



Setup – Define Statistical Analysis Plan including IDA/DQA (SAPI)

ADMINISTRATIVE INFORMATION
1.1 Project title
1.2 Project documents
1.3 Ethics approval
1.4 Names and contact

PROJECT BACKGROUND
2.1 Research aims
2.2 Target population
2.3 Research objectives

DATA
3.1 Data sources
3.2 Data sets

OBSERVATION UNITS
4.1 Inclusion and exclusion criteria
4.2 Sampling

VARIABLES
5.1 Variables in the main data analysis
5.2 Other variables

MAIN DATA ANALYSIS (MDA)
6.1 Description of observation units
6.2 Main data analysis methods
6.3 Assumptions and diagnostics
6.4 Sample size
6.5 Software

METHODS: INITIAL DATA ANALYSIS (IDA)
7.1 Data preparation
7.2 Unit missingness
7.3 Unit profile
7.4 Item missingness
7.5 Univariable descriptions
7.6 Multivariable descriptions

EVALUATION AND UPDATES
8.1 Evaluating the SAPI
8.2 Updating the SAPI

SUPPLEMENT
9.1 Key references
9.2 Sustainable handling of outputs

Huebner et al. Work in Progress

Setup – Define Statistical Analysis Plan including IDA/DQA (SAPI)

Table 1. Initial data analysis checklist for data screening in longitudinal studies.

Topic	Item	Features
IDA screening domain: Participation profile		
Time frame	P1	Provide number of time points and intervals at which measurements are taken, using the time metric that best reflects the time from inclusion in the study, or calendar time in studies that involve long enrollment times. Highlight the differences between the time of first measurements and follow-up times.
Time metric	P2	Describe the time metric and corresponding time points specified in the analysis strategy, if different from the time metric described in P1.
Participants	P3	Provide the number of participants who attended the assessment by time metric(s).
Optional extensions: Participation Profile		
Other time metrics	PE1	Use different time metric(s) to describe the time frame of the study, if applicable and appropriate, e.g. calendar time or data collection visits.
IDA screening domain: Missing data (outcome variable and explanatory variables)		
Non-enrollment	M1	Describe the non-enrolled, i.e., the participants that were selected but did not enter the study (and the reasons, if available), if applicable.
Drop-out	M2	Describe the participants who dropped out from the study during the follow-up (loss to follow-up and other possible reasons: death, withdrawal, missing by design, if applicable).
Intermittent visit missingness	M3	Describe the participants that have missing data for some of the measurements (intermittent, occasional omission, but do not drop out out of the study).

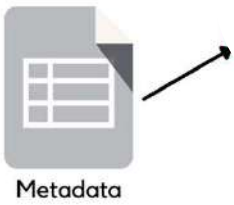
Lusa et al. 2024

Setup – Define Statistical Analysis Plan including IDA/DQA (SAPI)

Data sets	3.2	**Describe how the data is provided for this analysis project, for example format and content of data sets** • Describe how the data was processed from their raw state after data collection until transfer for use in this analysis project. For example, any preliminary data editing, deletion of cases and variables, or rules underlying the computation of new variables.

Huebner et al. Work in Progress

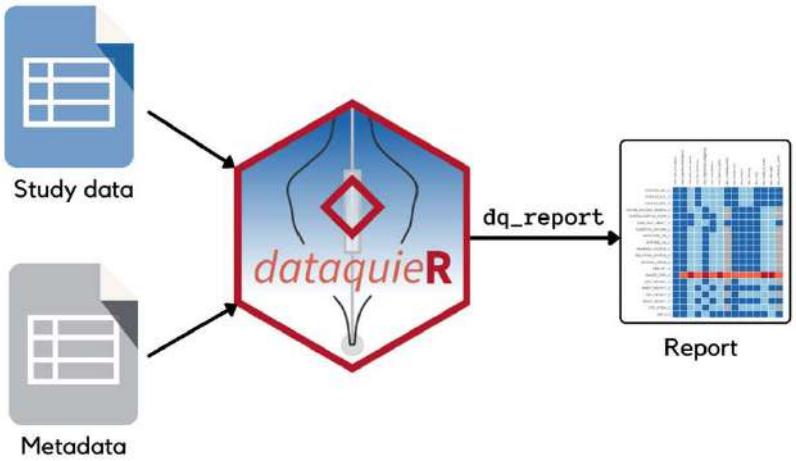
Annotate – Build interoperable knowledge layers to guide checks



	A	B	C	D	E	F	G	
1	VAR_NAMES	LABEL	DATA_TYPE	SCALE_LEVEL	VALUE_LABELS	MISSING_LIST_TABLE	HARD_LIMITS	DETECTI
2	v00000	CENTER_0	integer	nominal	1 = Berlin | 2 = Hamburg | 3 = Leipzig | 4 = Cologne | 5 = Munich			
3	v00001	PSEUDO_ID	string	na				
4	v00002	SEX_0	integer	nominal	0 = females | 1 = males			
5	v00003	AGE_0	integer	ratio			[18;Inf)	
6	v00103	AGE_GROUP_0	string	ordinal				
7	v01003	AGE_1	integer	ratio			[18;Inf)	
8	v01002	SEX_1	integer	nominal	0 = females | 1 = males			
9	v10000	PART_STUDY	integer	nominal	0 = no | 1 = yes			
10	v00004	SBP_0	float	ratio		missing_table	[80;180]	[0;265]
11	v00005	DBP_0	float	ratio		missing_table	[50;Inf)	[0;265]
12	v00006	GLOBAL_HEALTH_V	float	ratio		missing_table	[0;10]	
13	v00007	ASTHMA_0	integer	nominal	0 = no | 1 = yes	missing_table	[0;1]	
14	v00008	VO2_CAPCAT_0	string	ordinal	A = excellent < B = good	missing_table		
15	v00009	ARM_CIRC_0	float	ratio		missing_table	[0;Inf)	
16	v00109	ARM_CIRC_DISC_0	integer	ordinal	1 = (-Inf,20] < 2 = (20,30	missing_table	[1;3]	
17	v00010	ARM_CUFF_0	integer	ordinal	1 = (-Inf,20] < 2 = (20,30	missing_table	[1;3]	
18	v00011	USR_VO2_0	string	nominal	USR_101 | USR_103 | U	missing_table		
19	v00012	USR_BP_0	string	nominal	USR_121 | USR_123 | U	missing_table		
20	v00013	EXAM_DT_0	datetime	interval			[2018-01-01 00:00:00 CET	
21	v20000	PART_PHYS_EXAM	integer	nominal	0 = no | 1 = yes			
22	v00014	CRP_0	float	ratio		missing_table	[0;Inf)	[0.16;Inf
23	v00015	BSG_0	float	ratio		missing_table	[0;100]	
24	v00016	DEV_NO_0	integer	nominal				
25	v00017	LAB_DT_0	datetime	interval			[2018-01-01 00:00:00 CET	

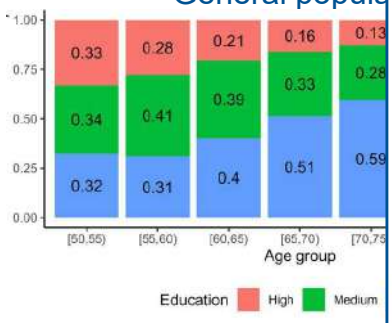
item_level | missing_table | cross-item_level | segment_level | dataframe_level | expected_id

Check – Execute data assessments



Report – Comprehensively summarize findings

General popula



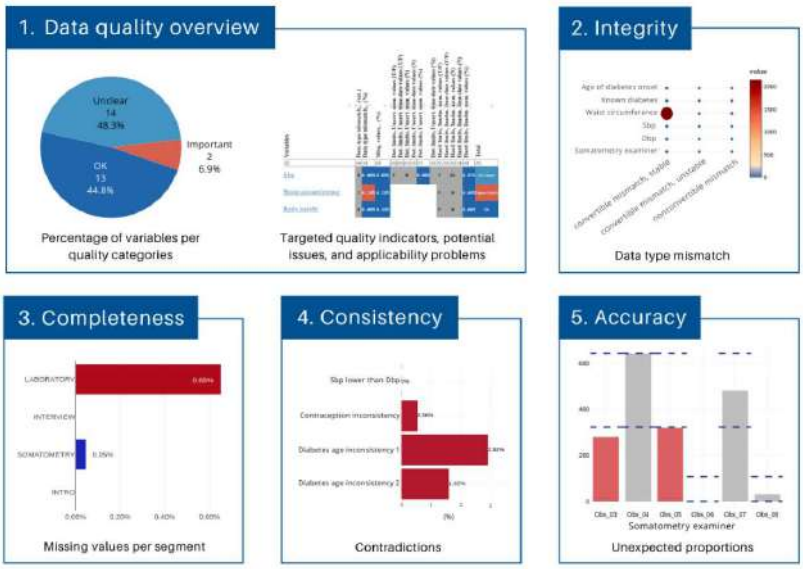
Lusa et al. 2024

Baseline characteristics by type of missingness.

	N	Complete N=2681	Death N=978	Intermittent missing N=476
gender : Female	5452	0.54	0.51	0.50
age_int	5452	52.00 58.00 66.00 60.28 ± 8.79	66.00 75.00 81.00 73.29 ± 10.44	52.00 58.00 64.00 59.55 ± 8.18
age_int_cat : 50-59	5452	0.54	0.12	0.59
60-69		0.29	0.21	0.27
70-80		0.14	0.41	0.13
80+		0.02	0.26	0.01
weight	5361	66.0 76.0 86.0 77.2 ± 15.2	62.5 71.0 81.0 72.7 ± 15.0	65.0 76.0 85.0 77.1 ± 15.6
height_imp	5418	165.00 172.00 178.00 171.82 ± 9.04	163.00 169.00 175.00 169.34 ± 8.80	165.00 172.00 178.00 171.66 ± 8.98
education_imp : Low	5428	0.17	0.38	0.19
Medium		0.38	0.39	0.41
High		0.45	0.23	0.40
pa_vig_freq	5423	0.67	0.35	0.66
pa_low_freq	5422	0.94	0.73	0.95
cusmoke_imp : Yes	5423	0.22	0.34	0.27
maxgrip	5272	29.0 36.0 48.0 38.5 ± 12.5	21.5 29.0 38.0 30.3 ± 11.9	28.0 37.5 49.0 38.5 ± 13.1

a b c represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous
is the number of non-missing values.

Report – Comprehensively summarize findings



https://dataquality.qihs.uni-greifswald.de/

Adapt – Refine annotations, checks, or analysis plan

ADMINISTRATIVE INFORMATION
1.1 Project title
1.2 Project documents
1.3 Ethics approval
1.4 Names and contact

PROJECT BACKGROUND
2.1 Research aims
2.2 Target population
2.3 Research objectives

DATA
3.1 Data sources
3.2 Data sets

OBSERVATION UNITS
4.1 Inclusion and exclusion criteria
4.2 Sampling

VARIABLES
5.1 Variables in the main data analysis
5.2 Other variables

MAIN DATA ANALYSIS (MDA)
6.1 Description of observation units
6.2 Main data analysis methods
6.3 Assumptions and diagnostics
6.4 Sample size
6.5 Software

METHODS: INITIAL DATA ANALYSIS (IDA)
7.1 Data preparation
7.2 Unit missingness
7.3 Unit profile
7.4 Item missingness
7.5 Univariable descriptions
7.6 Multivariable descriptions

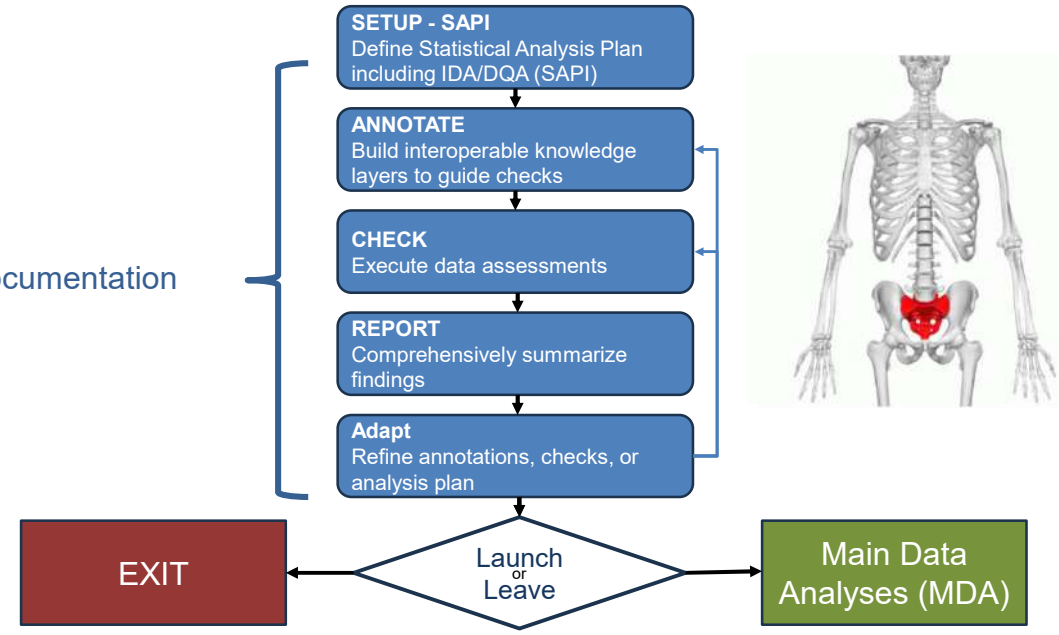
EVALUATION AND UPDATES
8.1 Evaluating the SAPI
8.2 Updating the SAPI

SUPPLEMENT
9.1 Key references
9.2 Sustainable handling of outputs

Launch or Leave – Conduct or discard inteded analyses

Section 63 Spatial and spatio-temporal Statistics I 02:00 pm - 03:20 pm **Topic :** 63 Spatial and Spatio-temporal Statistics **Form :** Oral	**Section 46 Advances in Item Response Theory-based measurement of abilities** 02:00 pm - 03:20 pm **Topic :** 46 Testing and Scaling	**Section 21 Tree-based AI** 02:00 pm - 03:20 pm **Topic :** 21 Artificial Intelligence and Machine Learning **Form :** Oral	**Section 11 Causal inference – mixed topics** 02:00 pm - 03:20 pm **Topic :** 11 Causal Inference **Form :** Oral **Chair(s):** Dr. Claudia	**Section 45 Robust Statistical Methods for Model Fitting and Data Analysis I** 02:00 pm - 03:20 pm **Topic :** 45 Robust and Nonparametric Statistics **Form :** Oral
Break	**Break**	**Break**	**Break**	**Break**
Section 63 Spatial and spatio-temporal Statistics II 03:50 pm - 05:10 pm **Topic :** 63 Spatial and Spatio-temporal Statistics **Form :** Oral	**Section 46 Linking, equating, and norming in the context of IRT** 03:50 pm - 05:10 pm **Topic :** 46 Testing and Scaling **Form :** Oral	**Section 21 xAI, Interpretability** 03:50 pm - 05:10 pm **Topic :** 21 Artificial Intelligence and Machine Learning **Form :** Oral	**Emil Gumbel: The prediction of extreme events (Documentary Movie)** 03:50 pm - 05:10 pm **Topic :** 98 Special Session **Form :** Oral	**Section 45 Robust Statistical Methods for Model Fitting and Data Analysis II** 03:50 pm - 05:10 pm **Topic :** 45 Robust and Nonparametric Statistics **Form :** Oral
Break	**Break**	**Break**	**Break**	**Break**

Accessible…. Transparent….. Reproducible



Thanks to the many involved

Carsten.schmidt@uni-greifswald.de
Universitätsmedizin Greifswald · KöR
ICM SHIP/KEF
Fleischmannstraße 8 · 17475 Greifswald
www.medizin.uni-greifswald.de