Anschreiben Ganslandt

dqrep: Facilitating harmonized data-quality assessments with Stata

Carsten Oliver Schmidt
Stephan Struckmann, Birgit Schauer
Institute for Community Medicine
SHIP/KEF

Take #1
Population-based epidemiologic cohort studies

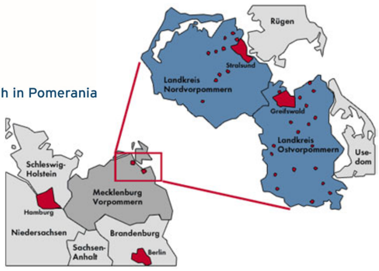
SHIP
Study of Health in Pomerania

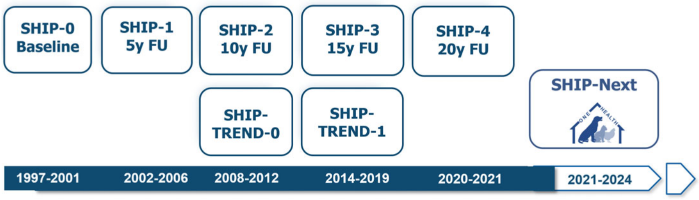
Rügen
Stralsund
Landkreis Nordvorpommern
Greifswald
Landkreis Ostvorpommern
Use-dom
Schleswig-Holstein
Hamburg
Mecklenburg Vorpommern
Niedersachsen
Brandenburg
Sachsen-Anhalt
Berlin

Photos: © Carsten Oliver Schmidt

Take #1
Population-based epidemiologic cohort studies

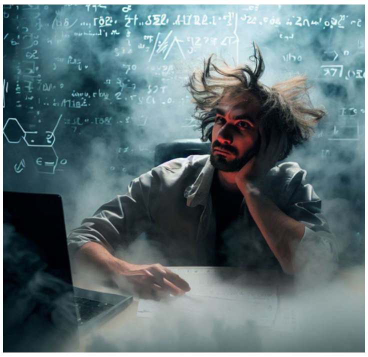
SHIP
Study of Health in Pomerania

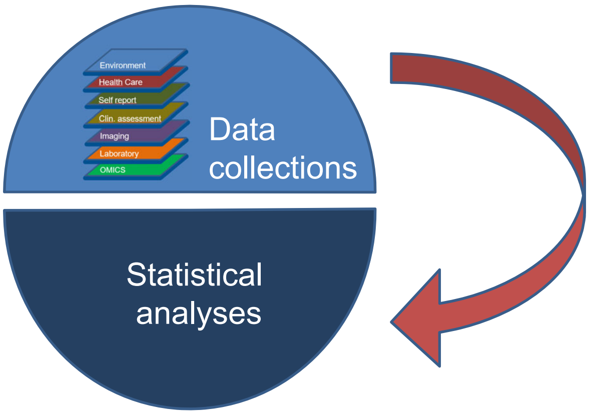



SHIP-0 Baseline
SHIP-1 5y FU
SHIP-2 10y FU
SHIP-3 15y FU
SHIP-4 20y FU
SHIP-TREND-0
SHIP-TREND-1
SHIP-Next

1997-2001
2002-2006
2008-2012
2014-2019
2020-2021
2021-2024

Take 1:
Support data collections and statistical analyses more efficiently



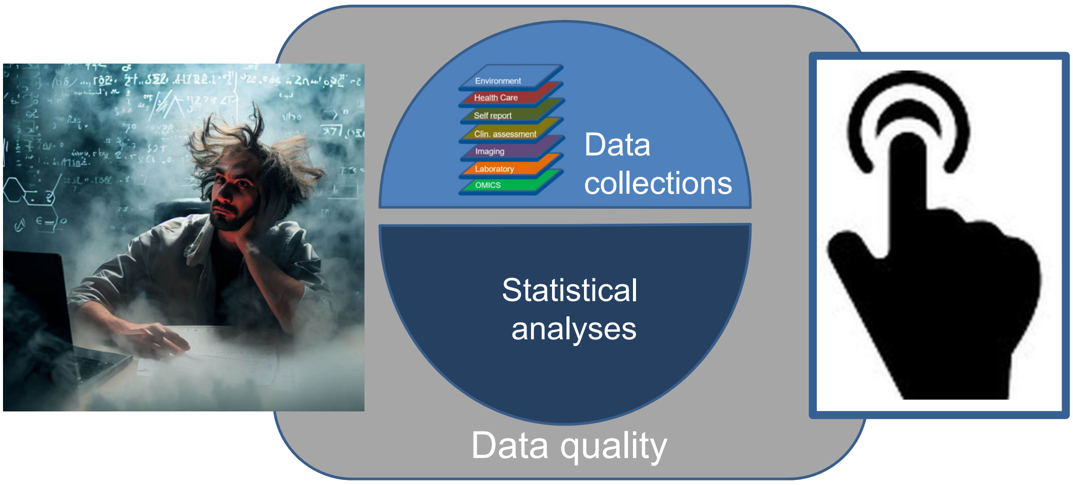


Data collections

Statistical analyses

Files?
Variables?
Missing data?
Violations?
Distributions?
Associations?

Take 1:
Support data collections and statistical analyses more efficiently



ANALYSIS PERSPECTIVE

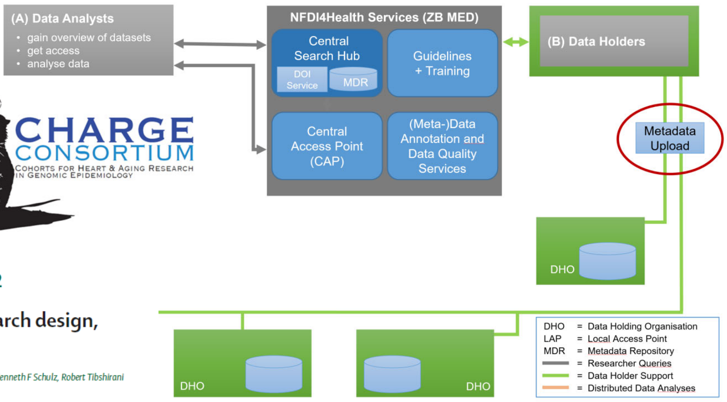
Take #2
Make science more transparent

nfdi4health

Kiel
Greifswald
Lübeck
Hamburg
Neubrandenburg
Bremen
Berlin
Hannover
Potsdam
Magdeburg
Münster
Essen
Witten
Göttingen
Halle
Düsseldorf
Leipzig
Köln
Dresden
Jena
Mittweida
Marburg
Aachen
Bonn
Mainz
Mannheim
Heidelberg
Erlangen

An EU-Canada joint infrastructure for next-generation multi-Study Heart research

CHARGE CONSORTIUM
COHORTS FOR HEART & AGING RESEARCH IN GENOMIC EPIDEMIOLOGY



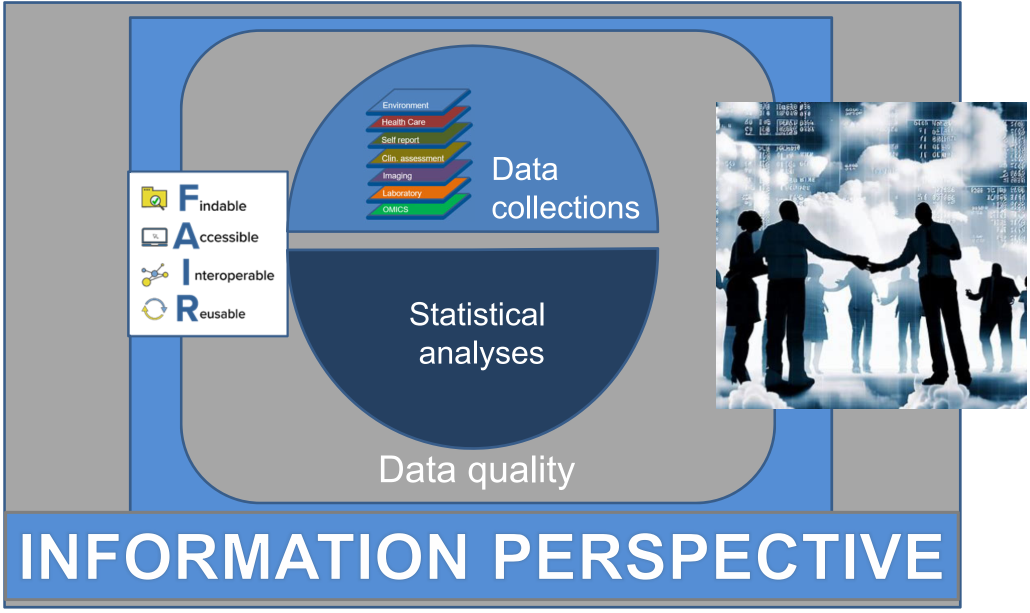
DHO = Data Holding Organisation
LAP = Local Access Point
MDR = Metadata Repository
▬ = Researcher Queries
▬ = Data Holder Support
▬ = Distributed Data Analyses

Research: increasing value, reducing waste 2

Increasing value and reducing waste in research design, conduct, and analysis

John P A Ioannidis, Sander Greenland, Mark A Hlatky, Muin J Khoury, Malcolm R Macleod, David Moher, Kenneth F Schulz, Robert Tibshirani

Take #2
Improved information management



Findable
Accessible
Interoperable
Reusable

Data collections

Statistical analyses

Data quality

INFORMATION PERSPECTIVE

ANALYSIS PERSPECTIVE

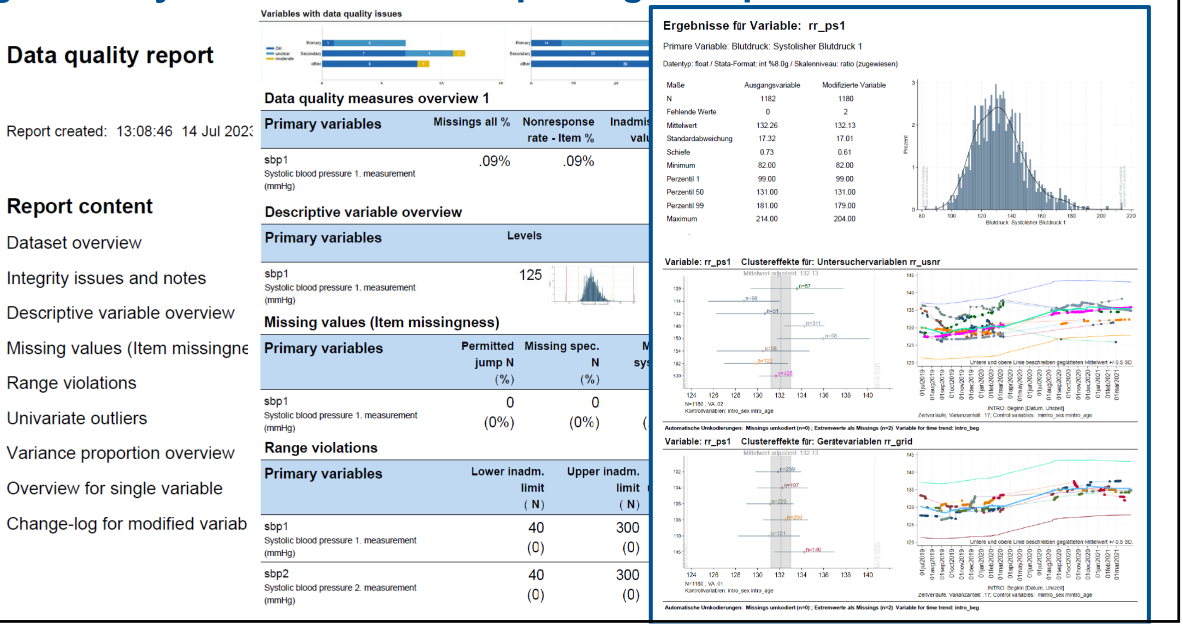
Analysis perspective
Single survey/examination DQ reporting

```
net from https://packages.qihs.uni-greifswald.de/repository/stata/dqrep
net install dqrep, replace
```

`dqrep,`

- >60 ados
- pdf, docx reports + result files (spreadsheet + graphs)

Analysis perspective
Single survey/examination DQ reporting - Output



Data quality report

Report created: 13:08:46 14 Jul 202

Report content

- Dataset overview
- Integrity issues and notes
- Descriptive variable overview
- Missing values (Item missingne
- Range violations
- Univariate outliers
- Variance proportion overview
- Overview for single variable
- Change-log for modified variab

Variables with data quality issues

Data quality measures overview 1

Primary variables	Missings all %	Nonresponse rate - Item %
sbp1 Systolic blood pressure 1. measurement (mmHg)	.09%	.09%

Descriptive variable overview

Primary variables	Levels
sbp1 Systolic blood pressure 1. measurement (mmHg)	125

Missing values (Item missingness)

Primary variables	Permitted jump N (%)	Missing spec. N (%)
sbp1 Systolic blood pressure 1. measurement (mmHg)	0 (0%)	0 (0%)

Range violations

Primary variables	Lower inadm. limit (N)	Upper inadm. limit (N)
sbp1 Systolic blood pressure 1. measurement (mmHg)	40 (0)	300 (0)
sbp2 Systolic blood pressure 2. measurement (mmHg)	40 (0)	300 (0)

Ergebnisse für Variable: rr_ps1

Primäre Variable: Blutdruck: Systolisher Blutdruck 1

Datentyp: float / Stata-Format: int %8.0g / Skalenniveau: ratio (zugewiesen)

Maße	Ausgangsvariable	Modifizierte Variable
N	1182	1180
Fehlende Werte	0	2
Mittelwert	132.26	132.13
Standardabweichung	17.32	17.01
Schiefe	0.73	0.61
Minimum	82.00	82.00
Perzentil 1	99.00	99.00
Perzentil 50	131.00	131.00
Perzentil 99	181.00	179.00
Maximum	214.00	204.00

Variable: rr_ps1 Clustereffekte für: Untersuchervariablen rr_usnr

Mittelwert adjustiert: 132.13
109
114
132
148
150
154
192
530
n=57
n=98
n=51
n=311
n=58
n=58
n=122
n=425
N=1180 ; VA .02
Kontrollvariablen: intro_sex intro_age
Untere und obere Linie beschreiben geglätteten Mittelwert +/-0.5 SD.
INTRO: Beginn [Datum, Uhrzeit]
Zeitverläufe, Varianzanteil: .17; Control variables: mintro_sex mintro_age

Automatische Umkodierungen: Missings umkodiert (n=0) ; Extremwerte als Missings (n=2) Variable for time trend: intro_beg

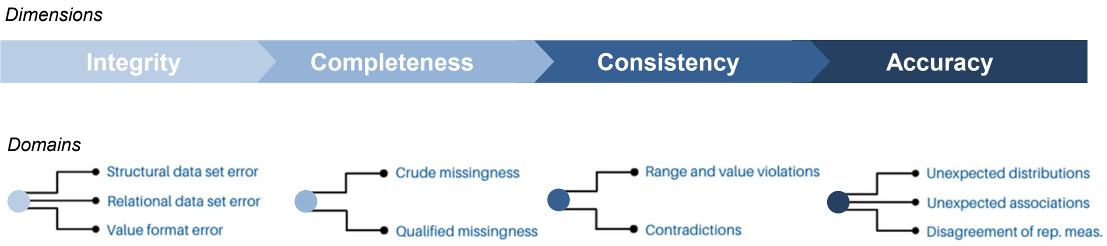
Variable: rr_ps1 Clustereffekte für: Gerätevariablen rr_grid

Mittelwert adjustiert: 132.13
102
104
106
108
110
145
n=239
n=197
n=223
n=250
n=131
n=140
N=1180 ; VA .01
Kontrollvariablen: intro_sex intro_age
Untere und obere Linie beschreiben geglätteten Mittelwert +/-0.5 SD.
INTRO: Beginn [Datum, Uhrzeit]
Zeitverläufe, Varianzanteil: .17; Control variables: mintro_sex mintro_age

Automatische Umkodierungen: Missings umkodiert (n=0) ; Extremwerte als Missings (n=2) Variable for time trend: intro_beg

Formal background: Data quality framework

Universitätsmedizin
GREIFSWALD



Dimensions

Integrity
Completeness
Consistency
Accuracy

Domains

Structural data set error
Relational data set error
Value format error

Crude missingness
Qualified missingness

Range and value violations
Contradictions

Unexpected distributions
Unexpected associations
Disagreement of rep. meas.

Schmidt *et al. BMC Medical Research Methodology* (2021) 21:63
https://doi.org/10.1186/s12874-021-01252-7

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R

Carsten Oliver Schmidt[1*], Stephan Struckmann[1], Cornelia Enzenbach[2], Achim Reineke[3], Jürgen Stausberg[4], Stefan Damerow[5], Marianne Huebner[6], Börge Schmidt[4], Willi Sauerbrei[7] and Adrian Richter[1]

Horizon 2020 Programme

DFG Deutsche Forschungsgemeinschaft

Analysis perspective
Methodological approach

Integrity
Completeness
Consistency

Focus: Data values

Boolean, abs., rel. Frequencies

Boolean, abs., rel. Frequencies

Accuracy

Focus: Variables

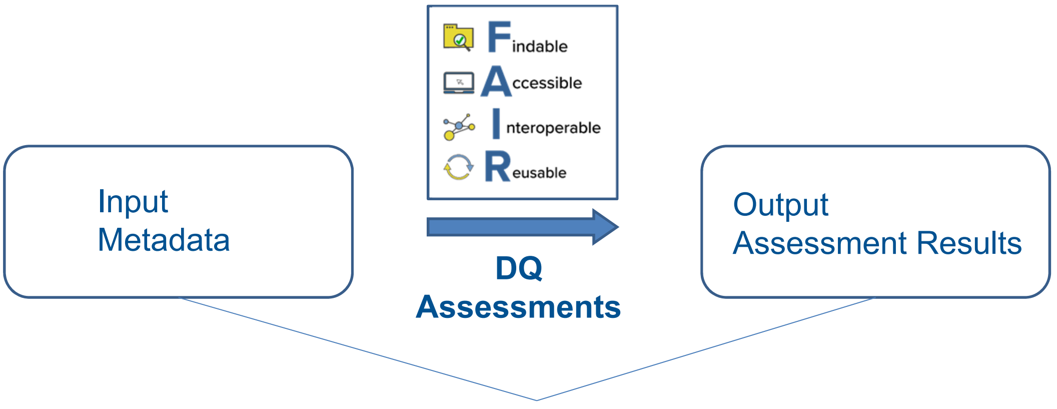
Diverse methods, metrics, e.g.:

- Mean, Median, SD, Min , Max
- Intra Class Correlations, Mixed Models
- (Non)-Parametric Regressions
- Outlier Assessments (e.g. Grubbs, Medcouple..)
- Confidence intervals

INFORMATION PERSPECTIVE

Take #2
Improved information management

Universitätsmedizin GREIFSWALD



Findable
Accessible
Interoperable
Reusable

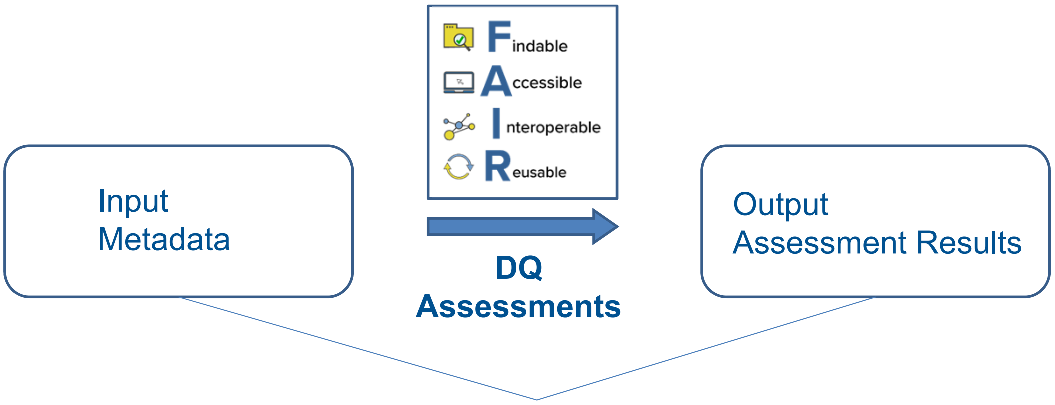
Input
Metadata

DQ Assessments

Output
Assessment Results

Machine Readibility

Take #2
Improved information management



Input
Metadata

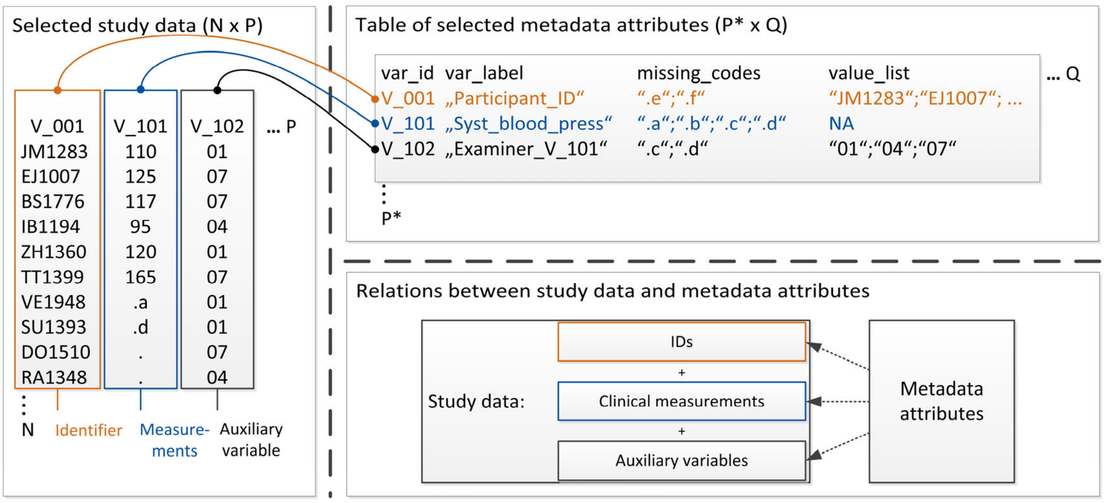
DQ
Assessments

Output
Assessment Results

Machine Readibility

e.g. as spreadsheet, ideally some CDM

Improved information management
INPUT



Selected study data (N x P)

V_001	V_101	V_102	... P
JM1283	110	01	
EJ1007	125	07	
BS1776	117	07	
IB1194	95	04	
ZH1360	120	01	
TT1399	165	07	
VE1948	.a	01	
SU1393	.d	01	
DO1510	.	07	
RA1348	.	04	
⋮			
N			

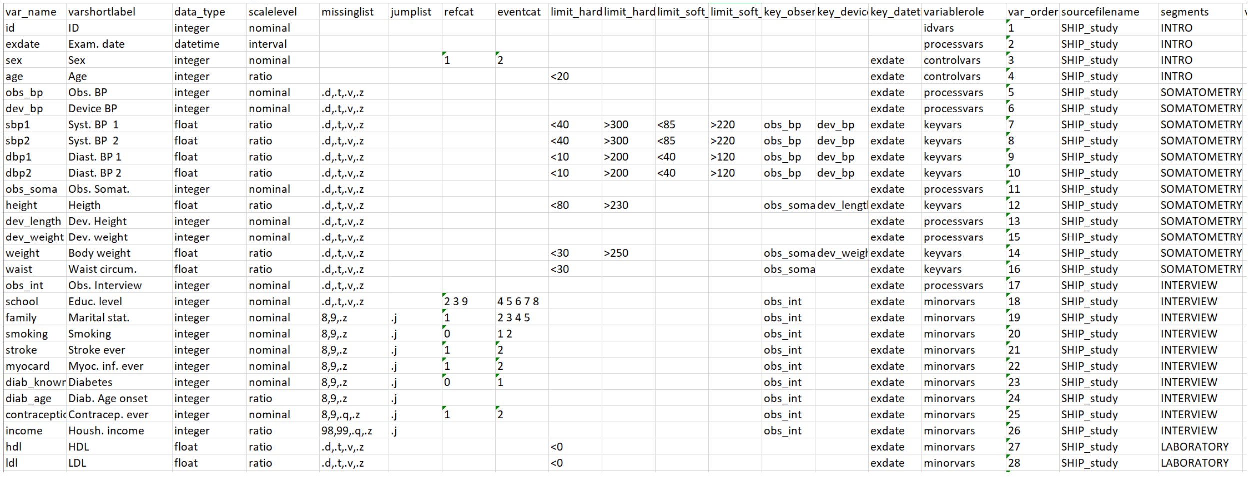
Table of selected metadata attributes (P* x Q)

var_id	var_label	missing_codes	value_list	... Q
V_001	„Participant_ID“	“.e“;“.f“	“JM1283“;“EJ1007“; ...	
V_101	„Syst_blood_press“	“.a“;“.b“;“.c“;“.d“	NA	
V_102	„Examiner_V_101“	“.c“;“.d“	“01“;“04“;“07“	
⋮				
P*				

Relations between study data and metadata attributes

Richter et al. 2019

Improved information management
INPUT – dqrep standard

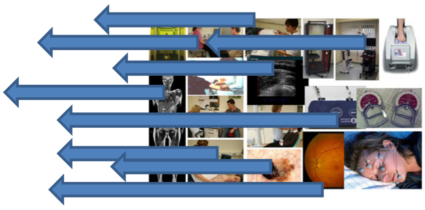


var_name	varshortlabel	data_type	scalelevel	missinglist	jumplist	refcat	eventcat	limit_hard	limit_hard	limit_soft_	limit_soft_	key_obser	key_devic	key_datet	variablerole	var_order	sourcefilename	segments
id	ID	integer	nominal												idvars	1	SHIP_study	INTRO
exdate	Exam. date	datetime	interval												processvars	2	SHIP_study	INTRO
sex	Sex	integer	nominal			1	2							exdate	controlvars	3	SHIP_study	INTRO
age	Age	integer	ratio					<20						exdate	controlvars	4	SHIP_study	INTRO
obs_bp	Obs. BP	integer	nominal	.d,.t,.v,.z										exdate	processvars	5	SHIP_study	SOMATOMETRY
dev_bp	Device BP	integer	nominal	.d,.t,.v,.z										exdate	processvars	6	SHIP_study	SOMATOMETRY
sbp1	Syst. BP 1	float	ratio	.d,.t,.v,.z				<40	>300	<85	>220	obs_bp	dev_bp	exdate	keyvars	7	SHIP_study	SOMATOMETRY
sbp2	Syst. BP 2	float	ratio	.d,.t,.v,.z				<40	>300	<85	>220	obs_bp	dev_bp	exdate	keyvars	8	SHIP_study	SOMATOMETRY
dbp1	Diast. BP 1	float	ratio	.d,.t,.v,.z				<10	>200	<40	>120	obs_bp	dev_bp	exdate	keyvars	9	SHIP_study	SOMATOMETRY
dbp2	Diast. BP 2	float	ratio	.d,.t,.v,.z				<10	>200	<40	>120	obs_bp	dev_bp	exdate	keyvars	10	SHIP_study	SOMATOMETRY
obs_soma	Obs. Somat.	integer	nominal	.d,.t,.v,.z										exdate	processvars	11	SHIP_study	SOMATOMETRY
height	Heigth	float	ratio	.d,.t,.v,.z				<80	>230			obs_soma	dev_lengt	exdate	keyvars	12	SHIP_study	SOMATOMETRY
dev_length	Dev. Height	integer	nominal	.d,.t,.v,.z										exdate	processvars	13	SHIP_study	SOMATOMETRY
dev_weight	Dev. weight	integer	nominal	.d,.t,.v,.z										exdate	processvars	15	SHIP_study	SOMATOMETRY
weight	Body weight	float	ratio	.d,.t,.v,.z				<30	>250			obs_soma	dev_weig	exdate	keyvars	14	SHIP_study	SOMATOMETRY
waist	Waist circum.	float	ratio	.d,.t,.v,.z				<30				obs_soma		exdate	keyvars	16	SHIP_study	SOMATOMETRY
obs_int	Obs. Interview	integer	nominal	.d,.t,.v,.z										exdate	processvars	17	SHIP_study	INTERVIEW
school	Educ. level	integer	nominal	.d,.t,.v,.z		2 3 9	4 5 6 7 8					obs_int		exdate	minorvars	18	SHIP_study	INTERVIEW
family	Marital stat.	integer	nominal	8,9,.z	.j	1	2 3 4 5					obs_int		exdate	minorvars	19	SHIP_study	INTERVIEW
smoking	Smoking	integer	nominal	8,9,.z	.j	0	1 2					obs_int		exdate	minorvars	20	SHIP_study	INTERVIEW
stroke	Stroke ever	integer	nominal	8,9,.z	.j	1	2					obs_int		exdate	minorvars	21	SHIP_study	INTERVIEW
myocard	Myoc. inf. ever	integer	nominal	8,9,.z	.j	1	2					obs_int		exdate	minorvars	22	SHIP_study	INTERVIEW
diab_know	Diabetes	integer	nominal	8,9,.z	.j	0	1					obs_int		exdate	minorvars	23	SHIP_study	INTERVIEW
diab_age	Diab. Age onset	integer	ratio	8,9,.z	.j							obs_int		exdate	minorvars	24	SHIP_study	INTERVIEW
contracepti	Contracep. ever	integer	nominal	8,9,.q,.z	.j	1	2					obs_int		exdate	minorvars	25	SHIP_study	INTERVIEW
income	Housh. income	integer	ratio	98,99,.q,.z	.j							obs_int		exdate	minorvars	26	SHIP_study	INTERVIEW
hdl	HDL	float	ratio	.d,.t,.v,.z				<0						exdate	minorvars	27	SHIP_study	LABORATORY
ldl	LDL	float	ratio	.d,.t,.v,.z				<0						exdate	minorvars	28	SHIP_study	LABORATORY

ANALYSIS PERSPECTIVE
Take II

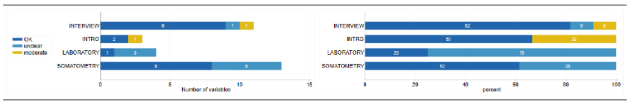
Analysis perspective
Multiple survey/examination DQ reporting

```
dqrep, rd(Example7) metadatafile("SHIP_metadata.xlsx") ///
        segmentname(segments) problemvarreport(4) benchmark(3)
```



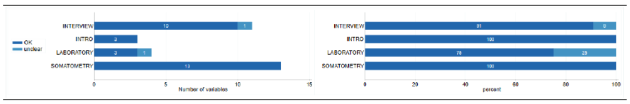
Analysis perspective
Multiple survey/examination DQ reporting – Output

Benchmark graphs on data quality gradings



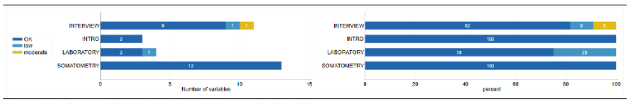
This graph shows the maximum problem category per variable.

Variables with data quality issues



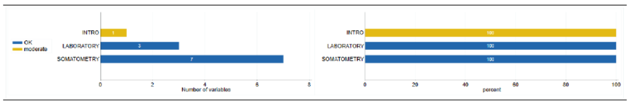
Categories for Item missing %: OK: <1; unclear: >=1;

Crude item missingness

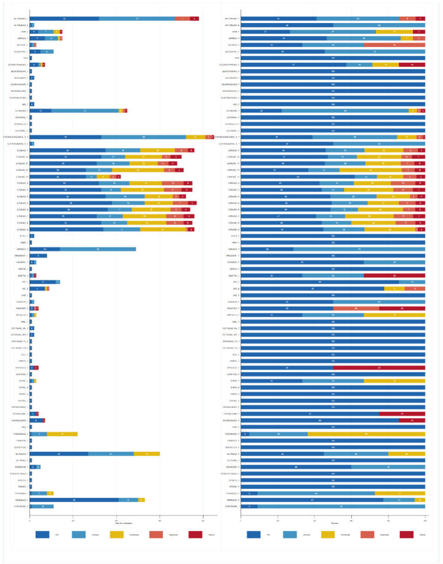


Categories for Item Response %: OK: <1; low: <5; moderate: <10; important: <25; critical: >=25;

Response rate - Item



Categories for Inadmissible values %: OK: ==0; moderate: <2; important: <5; critical: >=5;



Kategorien für Implausible values %: OK: ==0; unclear: <2; moderate: <5; important: <10; critical: >=10;

dqrep
Use scenarios
& Options

---

dqrep
Use scenarios

Active dataset

```
dqrep *bp*
```

Set of datasets
Metadata via command syntax
Variable invariant metadata

```
dqrep, rd(Example3) targetfiles("SHIP_study") ///
       itemmisslist(99900 99901 99902 99914) ///
       itemjumplist(99800 99801 99802) ///
       reportname("SHIP-Samplereport") ///
       reporttitle("SHIP-0 Data quality report") ///
       reportsubtitle("Report with anonymized SHIP-0 data") ///
       reportformat("docx") keyvars("sbp1 sbp2 dbp1 dbp2") ///
       minorvars(cholesterol stroke diab_known waist weight) ///
       observervars(obs_bp) devicevars(dev_bp) ///
       controlvars(age sex) idvars(id) timevars("exdate") store
```

Set of datasets
Metadata via spreadsheet
Variable variant metadata

```
dqrep, rd(Example4) metadatafile("SHIP_metadata.xlsx") store
```

dqrep
Options

#76

```
dqrep [varlist], [options]
```

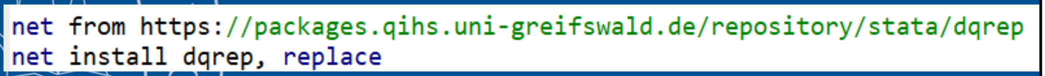
- Study data files and folders
- Result data files and folders
- Metadata files and folders
- Variable selections and variable roles
- Report formatting
- Analysis settings

+ adaptable
  - report structures, tables
  - languages
  - grading

Conclusion - dqrep

Strenghts
- A single command call suffices for complex DQ reporting
- Highly customizable → yet focus standard reports
- **dqrep relies on transparent information management**

Limitations
- No interactive assessments → formalized workflows
- Numerical variables (tries to convert strings)
- Not all data-quality related information automatically extracted
- Stata stability issues with single reports >100-150 variables
- Creating extensive metadata time-consuming



```
net from https://packages.qihs.uni-greifswald.de/repository/stata/dqrep
net install dqrep, replace
```

carsten.schmidt@uni-greifswald.de

Universitätsmedizin Greifswald . KöR

Carsten Oliver Schmidt
ICM SHIP/KEF
Fleischmannstraße 8 . 17475 Greifswald
www.medizin.uni-greifswald.de