Datenqualität bei Primärdatenerhebungen – Konzept und Implementation

Workshop Datenqualität TMF, 03.05.2018

Carsten Oliver Schmidt

Universitätsmedizin Greifswald
ICM-SHIP-KEF
Funktionsbereich Qualität in der Gesundheitsforschung

Inhalte

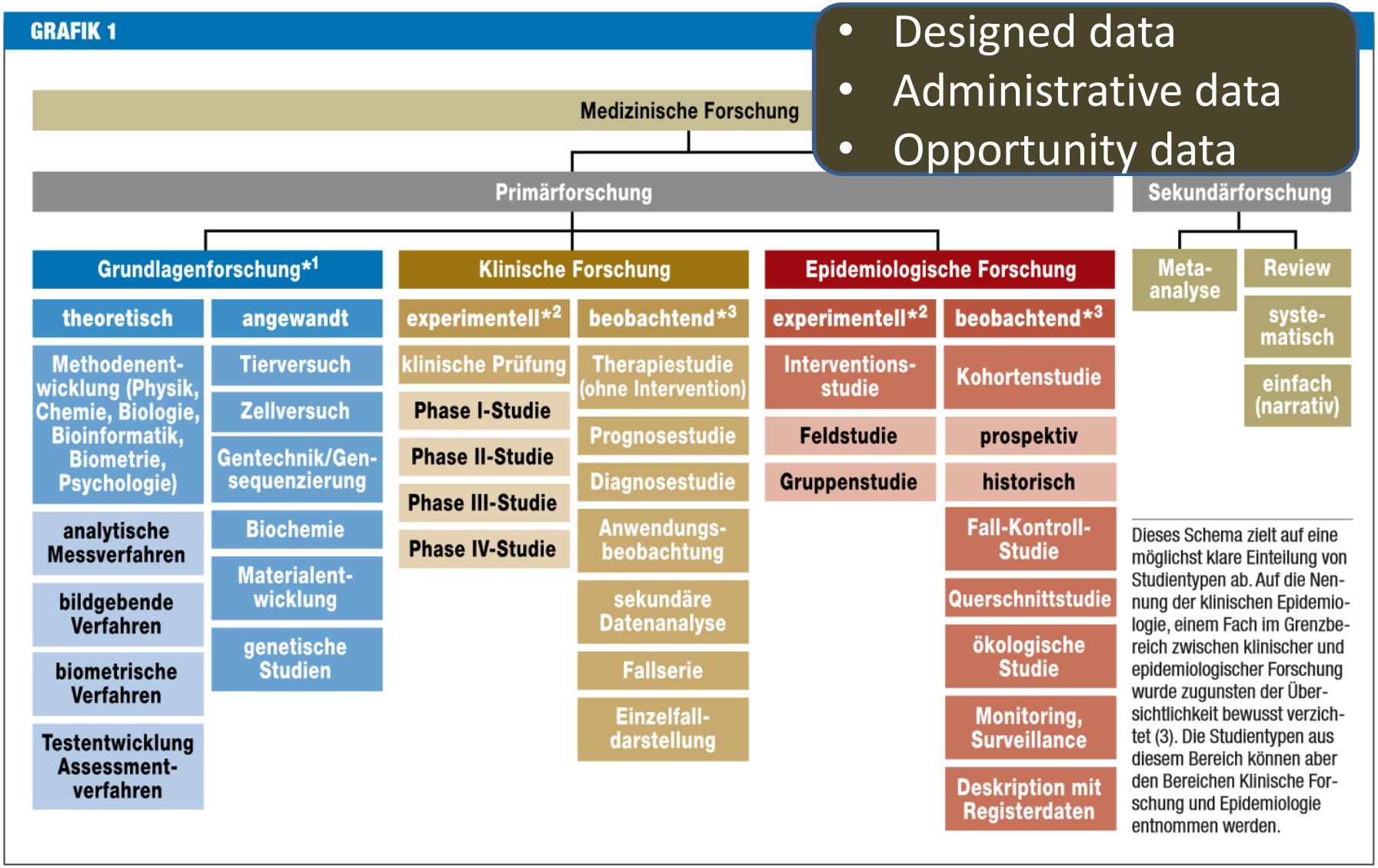
1. Anwendungsszenario und Datenqualität
2. Anwendungsbeispiel Study of Health in Pomerania
3. Konzept zur Bewertung von Datenqualität
4. Vom Konzept zur Statistik
5. Implementation in einer Kohortenstudie

Inhalte des Vortrags

1. **Anwendungsszenario und Datenqualität**
2. Anwendungsbeispiel Study of Health in Pomerania
3. Konzept zur Bewertung von Datenqualität
4. Vom Konzept zur Statistik
5. Implementation in einer Kohortenstudie

Anwendungsszenarien für Datenqualitätsbewertungen

- Designed data
- Administrative data
- Opportunity data



Einteilung verschiedener Studientypen

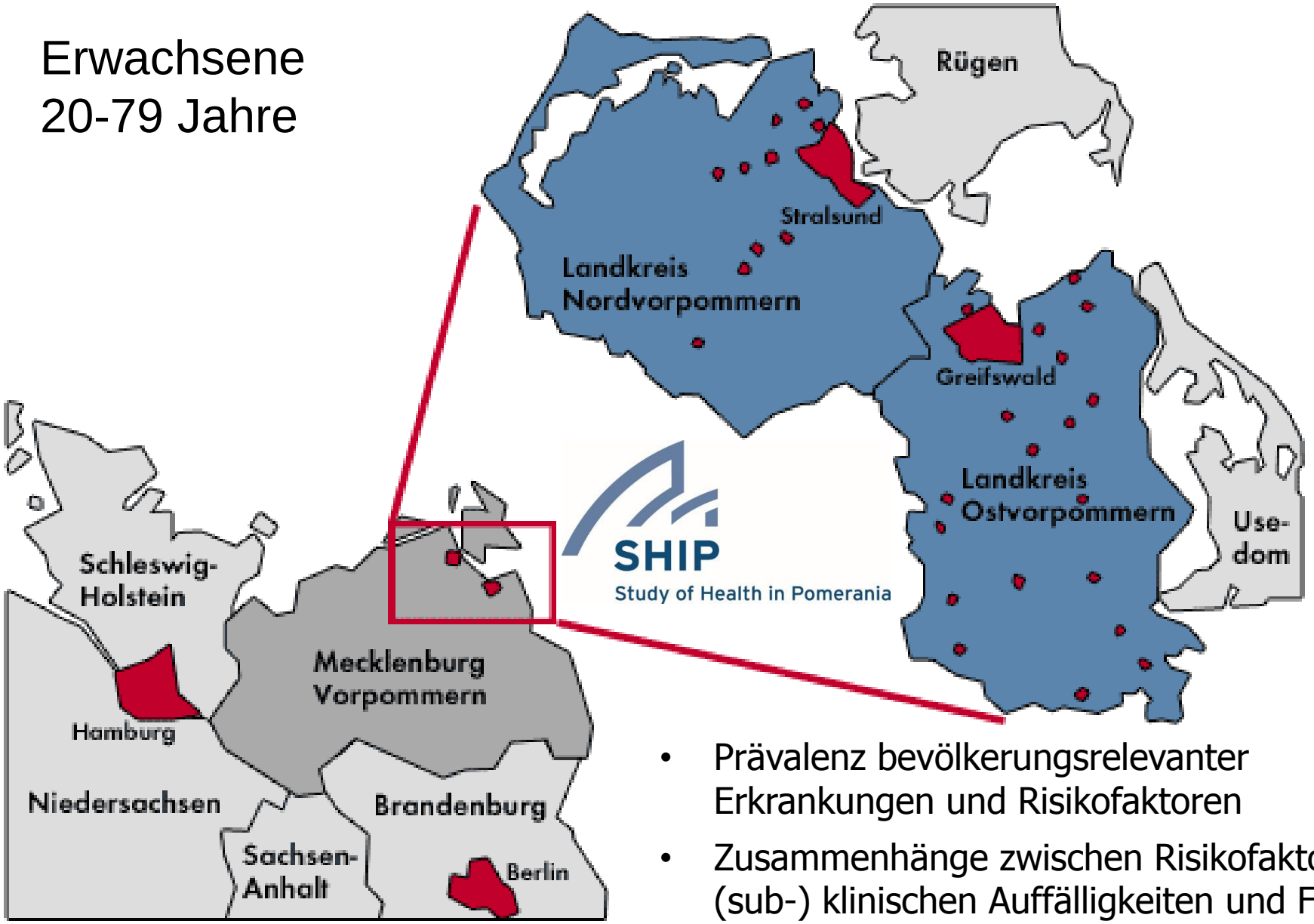
*1 häufig synonym verwendet: Experimentelle Forschung; *2 analoger Begriff: interventionell; *3 analoger Begriff: nicht interventionell/nicht experimentell

Röhrig et al. 2009 Dt. Ärzteblatt

Inhalte

1. Anwendungsszenario und Datenqualität
2. **Anwendungsbeispiel Study of Health in Pomerania**
3. Konzept zur Bewertung von Datenqualität
4. Vom Konzept zur Statistik
5. Implementation in einer Kohortenstudie

Erwachsene
20-79 Jahre



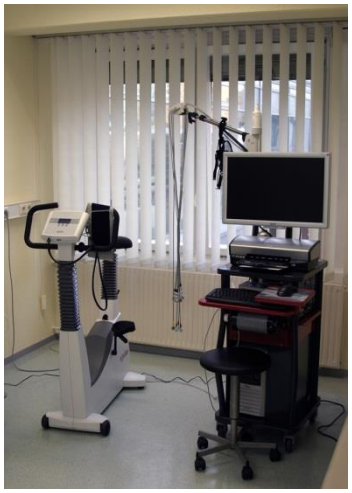
- Prävalenz bevölkerungsrelevanter Erkrankungen und Risikofaktoren
- Zusammenhänge zwischen Risikofaktoren, (sub-) klinischen Auffälligkeiten und Folgen analysieren

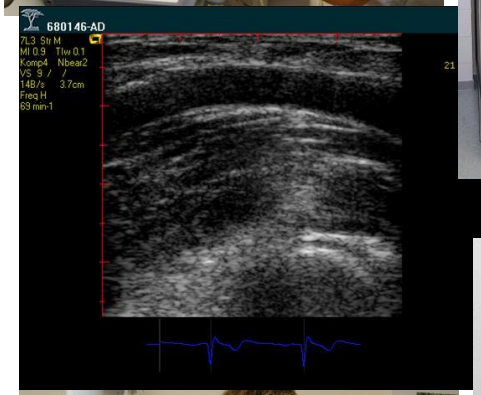
Wiederholte Messungen in SHIP

SHIP
Study of Health in Pomerania

SHIP-0 Baseline	SHIP-1 5y Follow-up	SHIP-2 10y Follow-up	SHIP-3 15y Follow-up
n= 4308	*n= 3300*	*n= 2333*	*n= 1718*
		SHIP-TREND-0 Baseline	SHIP-TREND-1 5y Follow-up
		n= 4422	
1997-2001	2002-2006	2008-2012	2014-2019

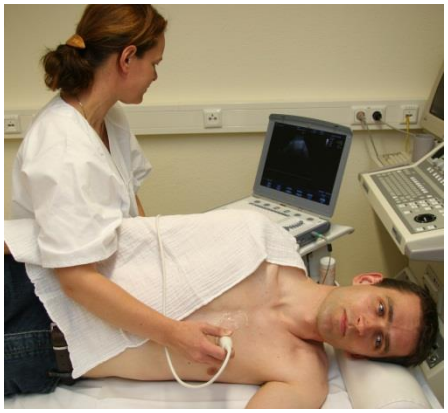


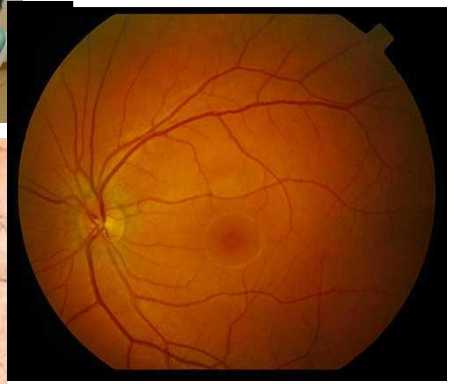
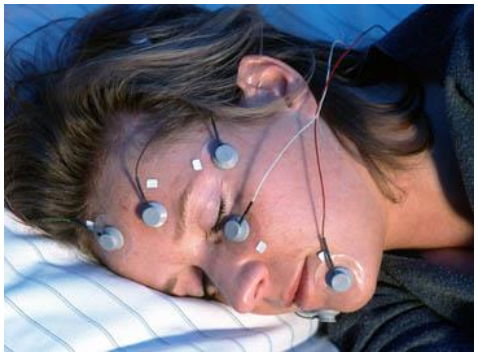









Völzke et al.; Int J Epidemiol 2011

Elemente des Qualitätsmanagements

Tabelle 1 Beispielhafte Elemente der Qualitätssicherung in Kohortenstudien

a. **Maßnahmen vor der Datenerhebung**
Studienhandbücher für Untersuchungen, Datenmanagement, QS
Checklisten zur Untersuchungsdurchführung
Training / (Re-) Zertifizierung von Untersuchern, Befundern und ggfs. Trainern bei multizentrischen Projekten
Gerätekalibrierung, Gerätevergleiche, Gerätewartung
Phantommessungen bei bildgebenden Verfahren
Gespräche/Qualitätszirkel zu qualitätsrelevanten Aspekten der Untersuchung
Pilotierung der Studie/ Prätest einzelner Untersuchungsmodule
b. **Maßnahmen während und nach der Datenerhebung**
Kontrolle der Teilnahmevoraussetzungen (Identität und Einverständnisse)
Standardisierte Dateneingaben (z.B. webbasiert mit automatischer Plausibilitätskontrolle)
Mehrfachbefundungen (v.a. bei bildgebenden Verfahren)
Mehrfacheingaben (Standard bei Eingabe von Papiervorlagen, z.B. Fragebögen)
On-Site Monitoring (Räumlichkeiten sowie Untersuchungsprozess)
Standardisierte und zentralisierte Erfassung von Auffälligkeiten (z.B. datenbankbasiert)
Qualitätsberichte
Monitoring der laufenden Datenerhebung (Fokus auf Rohdaten)
Datenmanagement, syntaxbasierte Datenbereinigungen
Externes Monitoring (Advisory Board)

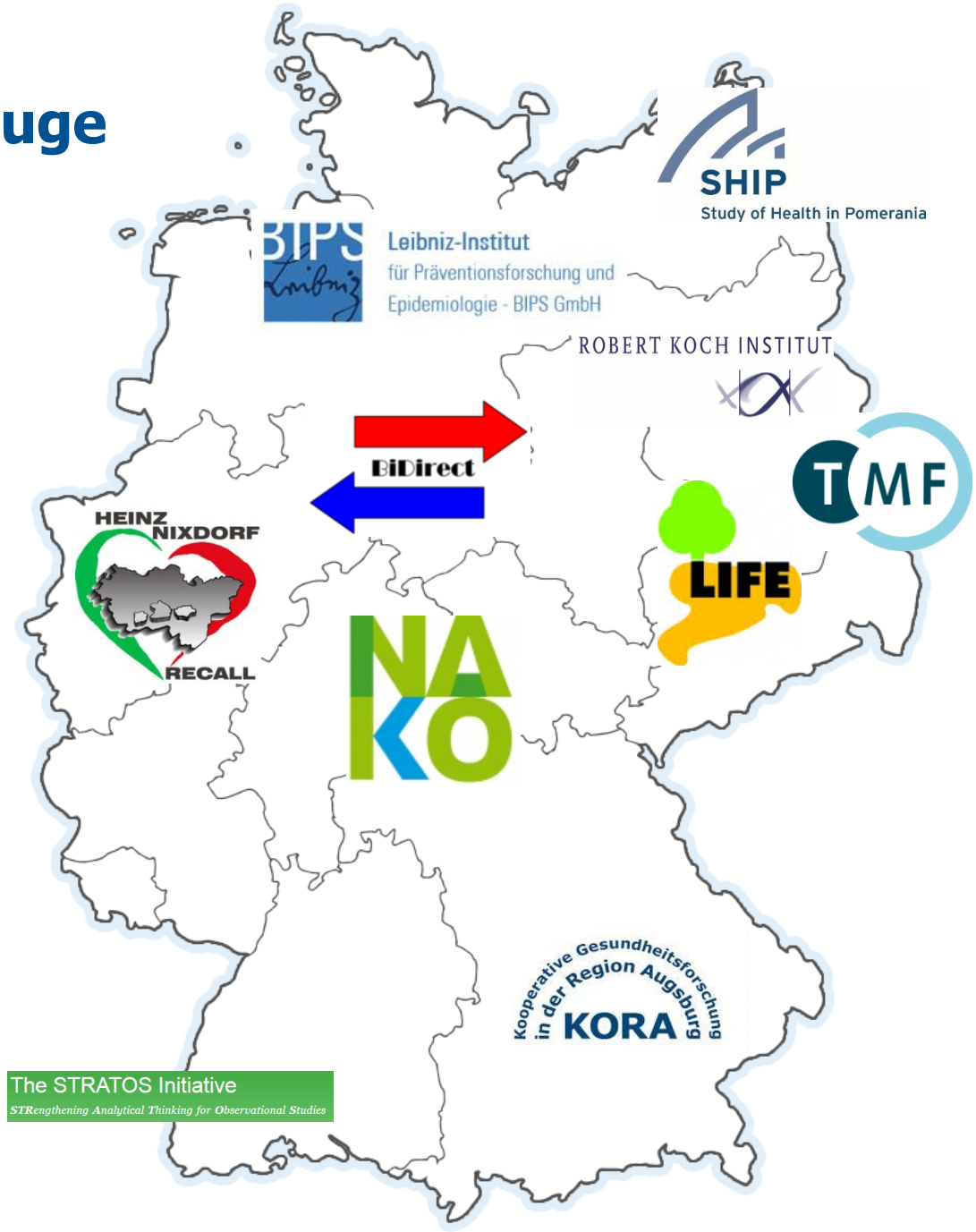
Quelle: Schmidt 2014 Qualitätssicherung in Kohortenstudien. Aus Leitlinie zum Adaptiven Datenmanagement in Kohortenstudien und Registern. TMF e.V.

Inhalte

1. Anwendungsszenario und Datenqualität
2. Anwendungsbeispiel Study of Health in Pomerania
3. **Konzept zur Bewertung von Datenqualität**
4. Vom Konzept zur Statistik
5. Implementation in einer Kohortenstudie

Standards und Werkzeuge zur Beurteilung der Datenqualität in komplexen epidemiologischen Studien

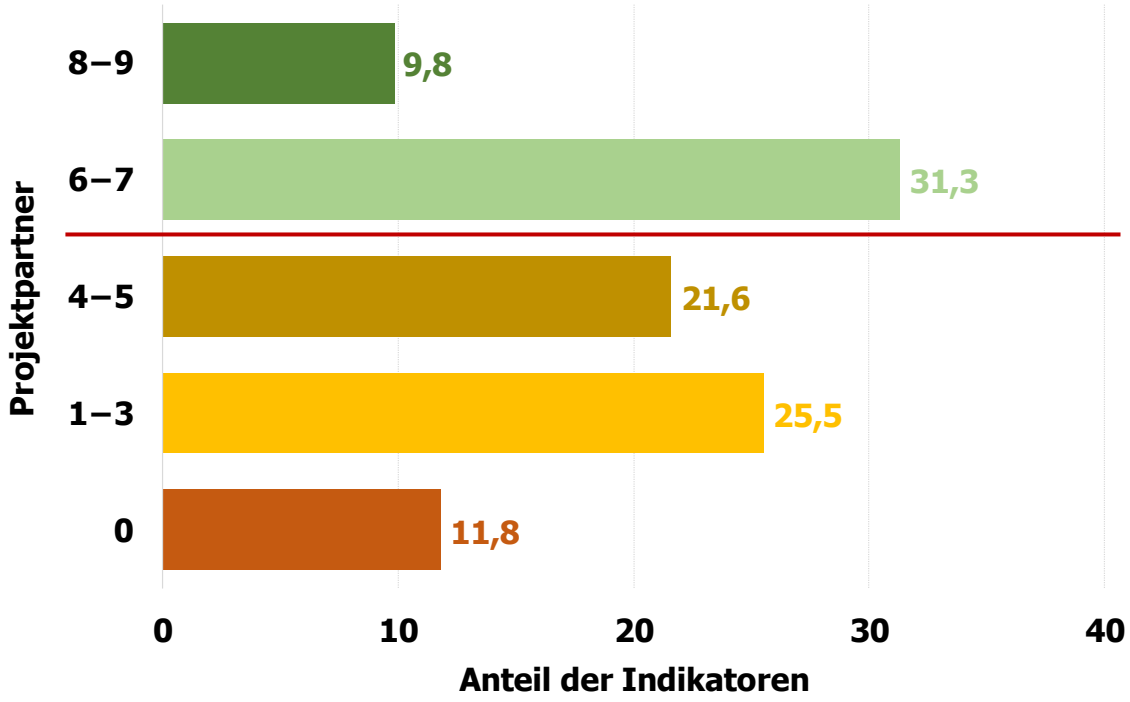
Schmidt, Carsten Oliver; Prof. Dr. Dr.
Bamberg, Fabian; Prof. Dr.
Berger, Klaus; Prof. Dr.
Hoffmann, Wolfgang; Prof. Dr.
Jöckel, Karl-Heinz; Prof. Dr.
Kurth, Bärbel-Maria; Prof. Dr.
Löffler, Markus; Prof. Dr.
Meisinger, Christa; Prof. Dr.
Pigeot, Iris; Prof. Dr.
Stausberg, Jürgen; Prof. Dr.



Kennziffern zur Qualität – TMF Leitlinie, Auszug

Qualitätsindikator: Integrität	ID neu
Werteverteilung	
Bevorzugung bestimmter Endziffern	
Werteverteilung der durch Untersucher erfassten Parameter	
Werteverteilung der durch Geräte erfassten Parameter	
Werteverteilung von Befunden	
Werteverteilung von Parametern zwischen Zentren	
Fehlende Einträge	
Fehlende Module	TMF-1012
Fehlende Werte bei Datenelementen	TMF-1013
Fehlende Werte bei mandatorischen Datenelementen	TMF-1014
Fehlende Werte bei optionalen Datenelementen	TMF-1015
Datenelemente mit Wert unbekannt o. ä.	TMF-1016
Ausreißer bei stetigen Datenelementen	TMF-1018
Werte, die die Messbarkeitsgrenzen von Verfahren unter- oder überschreiten	TMF-1019
Unerlaubte Werte	
Unerlaubte Werte bei qualitativen Datenelementen[1]	TMF-1021
Unerlaubte Werte bei qualitativen Datenelementen zur Kodierung von Missings	TMF-1022
Unerlaubte Werte bei quantitativen Datenelementen	TMF-1024

Berücksichtigung inhaltliche Bereiche aus TMF Leitlinie



Sichere Widersprüche
Werteverteilungen nach Untersucher/Gerät
Ausreißer

Übereinstimmung mit Vorwert
Fehlende Werte bei Datenelementen
Nachweis bekannter Korrelationen
Rekrutierungsrate

Widerspruchsfreiheit
Fehlende Werte bei mandatorischen Datenelementen
Übereinstimmung mit Verfahrensregeln
Repräsentativität

Untersuchungen am Wochenende
Umfang der Metadaten bei Untersuchungen
Aktualität der Daten
Genauigkeit der Registerdaten

Beobachtungseinheiten mit unbekanntem Primärtumor
Solitäre Meldungen von Pathologen
Zurückgewiesene Meldungen

Bewertung TMF Leitlinie 2.0 – Ausgewählte Kommentare

- Viele **relevante inhaltliche Bereiche** werden getroffen
- **Abstraktionsniveau** der Indikatoren sehr heterogen
- **Definitionen** aus Ausführungen teilweise unklar
- **Registerbezug** sehr eng
- Eignung vorgeschlagener **Berechnungen** („Raten") teilweise fraglich
- Interpretation der **Ebenen** Integrität, Organisation und Richtigkeit bei Primärdatenerhebung unklar
- **Relevanz** der Indikatoren nicht benannt

Revision der Indikatoren im Kontext Primärdatenerhebungen

Anforderungen

- Indikatoren mit vergleichbarer Komplexität
- Möglichst wenig Überlappung zwischen Indikatoren
- Hierarchische Strukturierung
  - Um verschiedene Komplexitäten abzubilden
- Bestehende TMF Indikatoren sollen gemappt werden
- Definition zur vereinfachung der Anwendbarkeit

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Heterogenes Abstraktionsniveau bestehender Indikatoren

Konzeptüberlegung End Digit Preference

Application examples

Laboratory parameters may follow a normal distribution. A particular device may generate a large number of measurement values equalling "3".

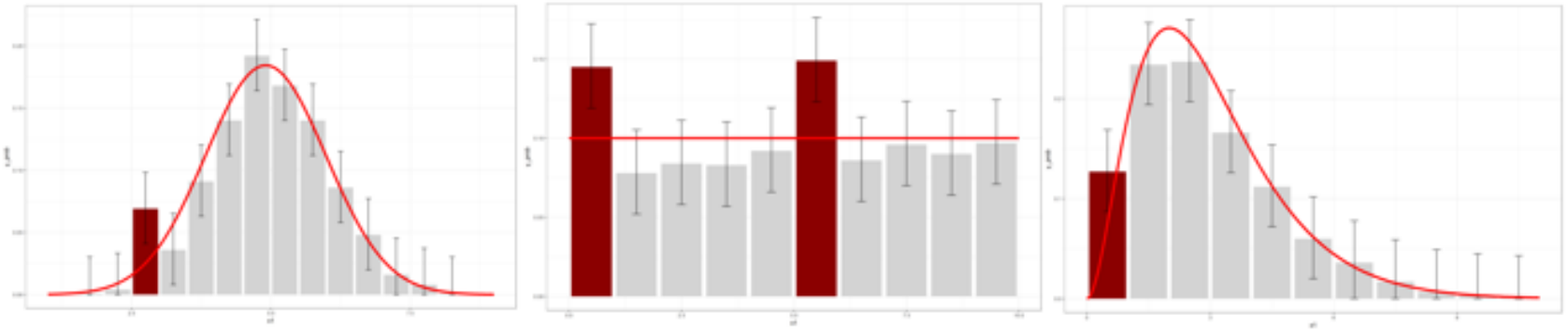


Figure 1: Normal distribution; Figure 2: Uniform distribution; Figure 3: Gamma distribution.

End digit preference: Body weight measurements may be rounded to "0" or "5" instead of using last digits provided by digital balance (Figure 2).

Heterogenes Abstraktionsniveau bestehender Indikatoren

Hierarchical Concept

Group level

Indicator domain

Indicator	Unexpected probability distribution
Subtype	End-digit-preference

Hierarchical Concept

Group level	
Indicator domain	Unexpected value distributions
Indicator	Unexpected probability distribution
Subtype	End-digit-preference

Begriffe zur Beschreibung von Datenqualität

Completeness	Accessibility	Trueness
Data completeness	Timeliness	Correctness
Model completeness	Currency	Accuracy
Data volume	Volatility	Validity
Uniqueness	Availability	Reliability
Appropr. Amount of data	Granularity	Reputation
Comprehensiveness	Resolution	Objectivity
Naturalness	Consistency	Plausibility
Rate of enrollment	Concordance	Precision
Utility	Conformance	Agreement
Contextualization	Integrity	Verifiability
Usefulness	Conciseness	Comparability
Relevance	Spatial stability	Standardization
Appropriateness	Predictive value	Generalizability
Informativeness	Coherence	Redundancy
Maintainability	Traceability	Believability
Responsiveness	Interpretability	Credibility
Usability	Complexity	Flexibility
Security	Cohesiveness	Portability

Hierarchical Concept

Group level	…provides an ontological framework.
Indicator domain	…provides a descriptive classification of methods to approach data issues.
Indicator	…is the level at which data quality indicators are defined.
Subtype	…classifies different application scenarios of an indicator which do not merit the definition of own indicators.

Primäre Gruppen von Qualitätsindikatoren

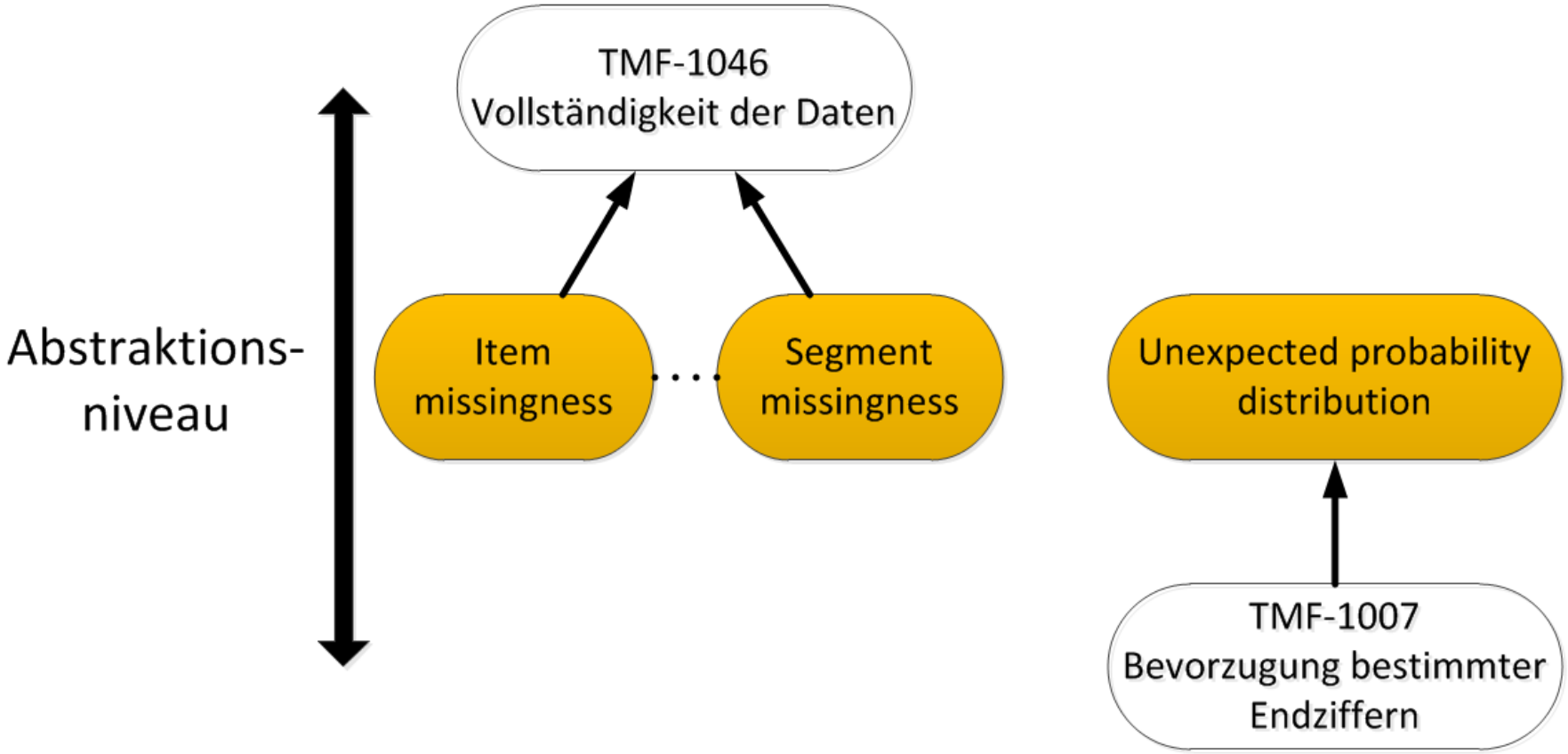
Completeness	The degree to which data values are present in a data collection.
Accuracy	The closeness of agreement between data values and the reference values.
Consistency	The degree to which data values are free of contradictions or convention breaks.

....

Indicators– Group „Completeness“

Unit Missingness	The degree to which measurements from an entire data collection are missing.
Segment missingness	The degree to which measurements from an entire segment (e.g. examinations or case report) of a data collection are missing.
Item missingness	The degree to which measurements are partially missing in segments of a data collection.

Heterogenes Abstraktionsniveau bestehender Indikatoren



Inhalte des Vortrags

1. Anwendungsszenario und Datenqualität
2. Anwendungsbeispiel Study of Health in Pomerania
3. Konzept zur Bewertung von Datenqualität
4. **Vom Konzept zur Statistik**
5. Implementation in einer Kohortenstudie

Übersetzung in Statistiken / Metriken

Completeness Boolean, absolute, relative Häufigkeiten

Consistency Boolean, absolute, relative Häufigkeiten

Accuracy Vielfältige Metriken

- ICC, Korrelationen
- (Nicht-) parametrische Regressionen
- Sensitivität, Spezifität, NPW, PPW
- Stat. Tests
- …

Übersetzung in Statistiken / Metriken

Completeness

Consistency

Accuracy

Prüffokus: Data values

Boolean, absolute, relative Häufigkeiten

Boolean, absolute, relative Häufigkeiten

Vielfältige Metriken

- ICC, Korrelationen
- (Nicht-) parametrische Regressionen
- Sensitivität, Spezifität, NPW, PPW
- Stat. Tests
- …

Prüffokus:
Verteilungen, Assoziationen

Metadata

Konzeptüberlegung End Digit Preference

Application examples

Laboratory parameters may follow a normal distribution. A particular device may generate a large number of measurement values equalling "3".

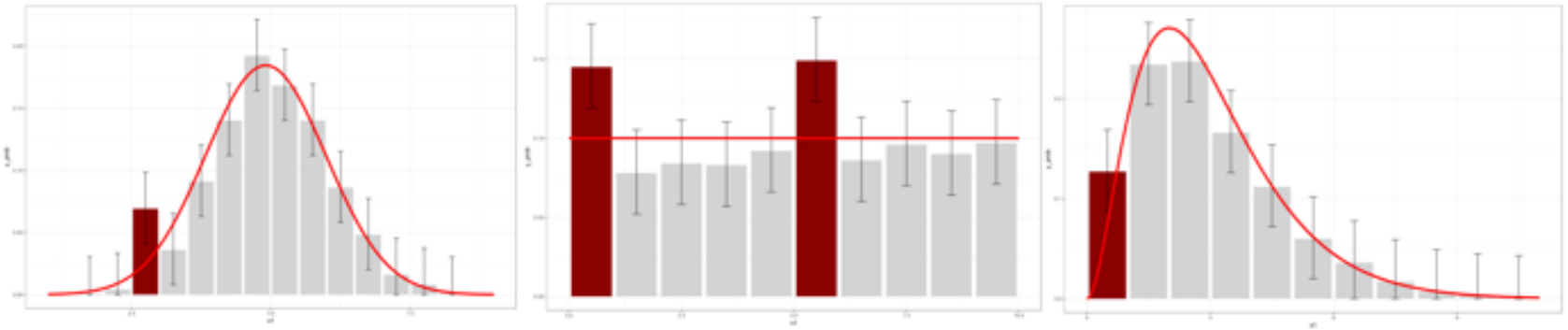


Figure 1: Normal distribution; Figure 2: Uniform distribution; Figure 3: Gamma distribution.

End digit preference: Body weight measurements may be rounded to "0" or "5" instead of using last digits provided by digital balance (Figure 2).

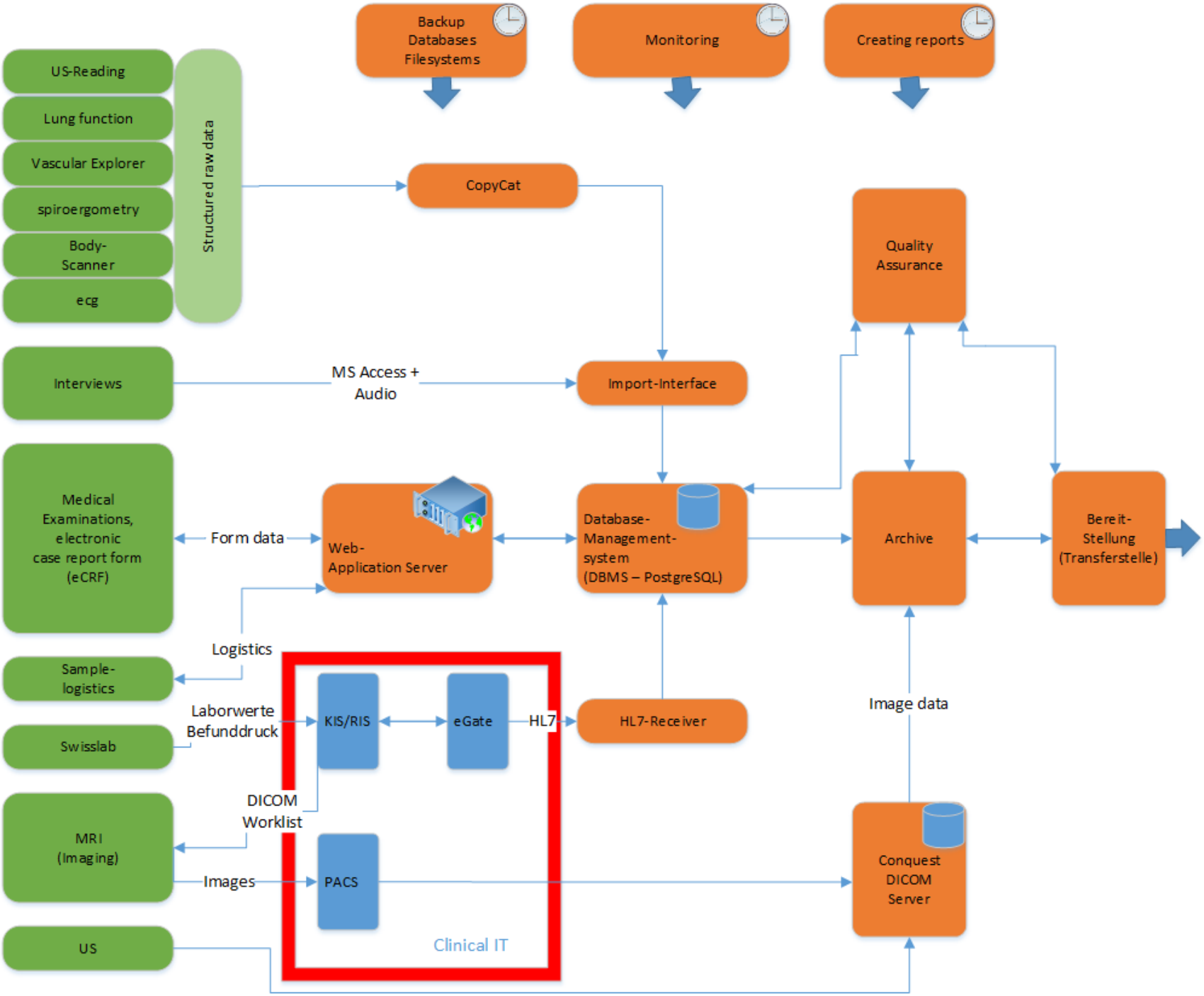
Anwendungen Metadaten in SHIP

	Description	Application examples
Auxiliary variables	related to study design and implementation	
Observer, Device	Identifier of the examiner, reader, etc.	Observer/Device-Effects
Auxiliary variables	related to the environmental measurement conditions	
Date-time stamps	Date and time stamps related to a visit or examination, pre-analytic processing time	Compliance with procedural rules Time trends
Environmental conditions	e.g. temperature, humidity, luminance	Compliance with procedural rules
Metadata attributes	related to completeness	
Missing codes	List of reasons for missing measurements	Item missingness
Jump codes	Conditionally missing measurements	To compute item missingness
Metadata attributes	related to data consistency / precision	
Value list	Variables with predefined categories	Inadmissible measurements
Validity limits	Upper and/or lower validity limits	Inadmissible measurements
Distribution	Expected prob. distribution	Unexpected prob. distribution
Metadata attributes	related to selection of quality statistics	
Data type	e.g. categorical, count, continuous, string	to select appropriate statistics
Metadata attributes	related to interpretation /standardized reporting	
Measurement unit	Continuous variables, e.g. mg/l	Implausible measurements

Richter et al. submitted, Auszug

Inhalte

1. Anwendungsszenario und Datenqualität
2. Anwendungsbeispiel Study of Health in Pomerania
3. Konzept zur Bewertung von Datenqualität
4. Vom Konzept zur Statistik
5. **Implementation in einer Kohortenstudie**


Backup
Databases
Filesystems
Monitoring
Creating reports
US-Reading
Lung function
Vascular Explorer
spiroergometry
Body-
Scanner
ecg
Structured raw data
CopyCat
Quality
Assurance
Interviews
MS Access +
Audio
Import-Interface
Medical
Examinations,
electronic
case report form
(eCRF)
Form data
Web-
Application Server
Database-
Management-
system
(DBMS – PostgreSQL)
Archive
Bereit-
Stellung
(Transferstelle)
Logistics
Sample-
logistics
Laborwerte
Befunddruck
Swisslab
KIS/RIS
eGate
HL7
HL7-Receiver
Image data
DICOM
Worklist
MRI
(Imaging)
Images
PACS
Conquest
DICOM
Server
US
Clinical IT

Framework: Initiale Datenanalyse

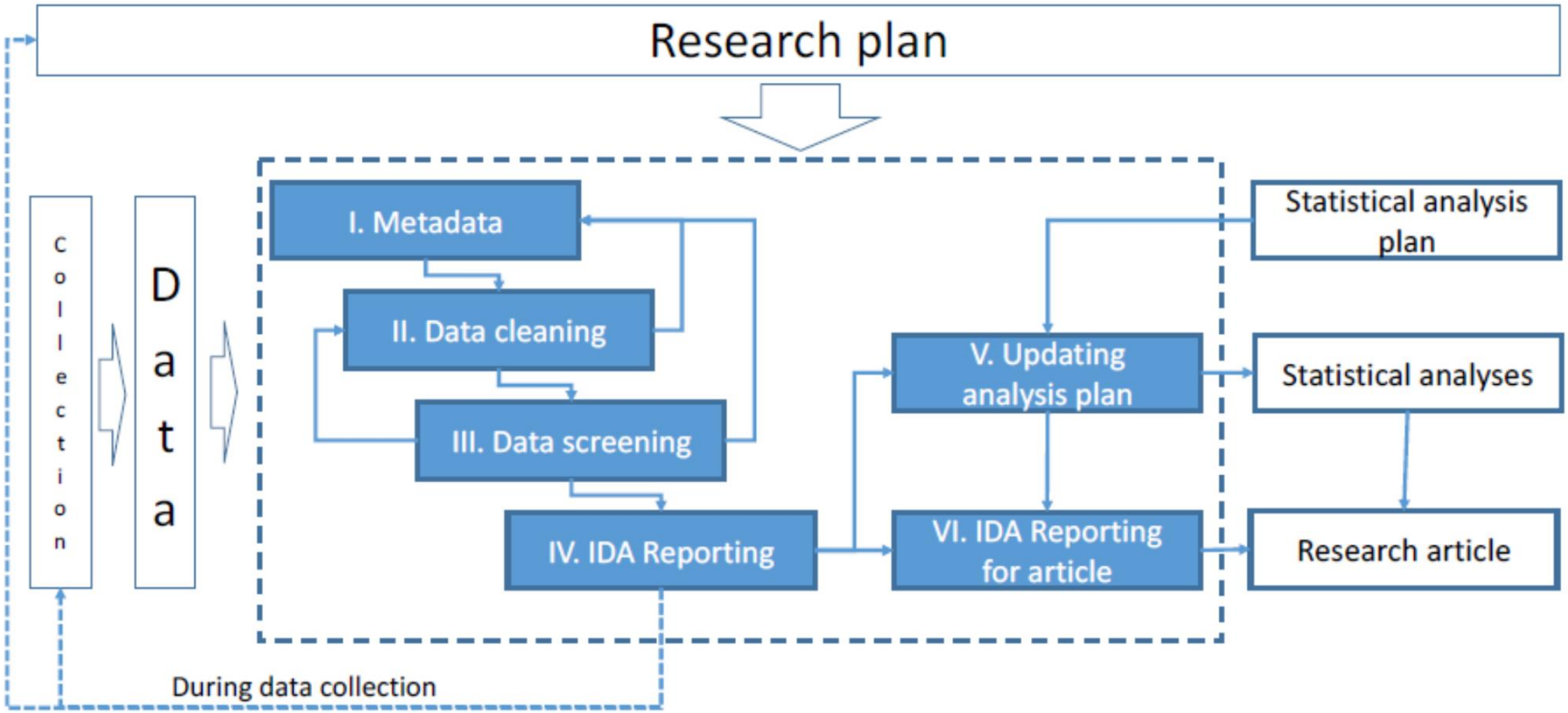
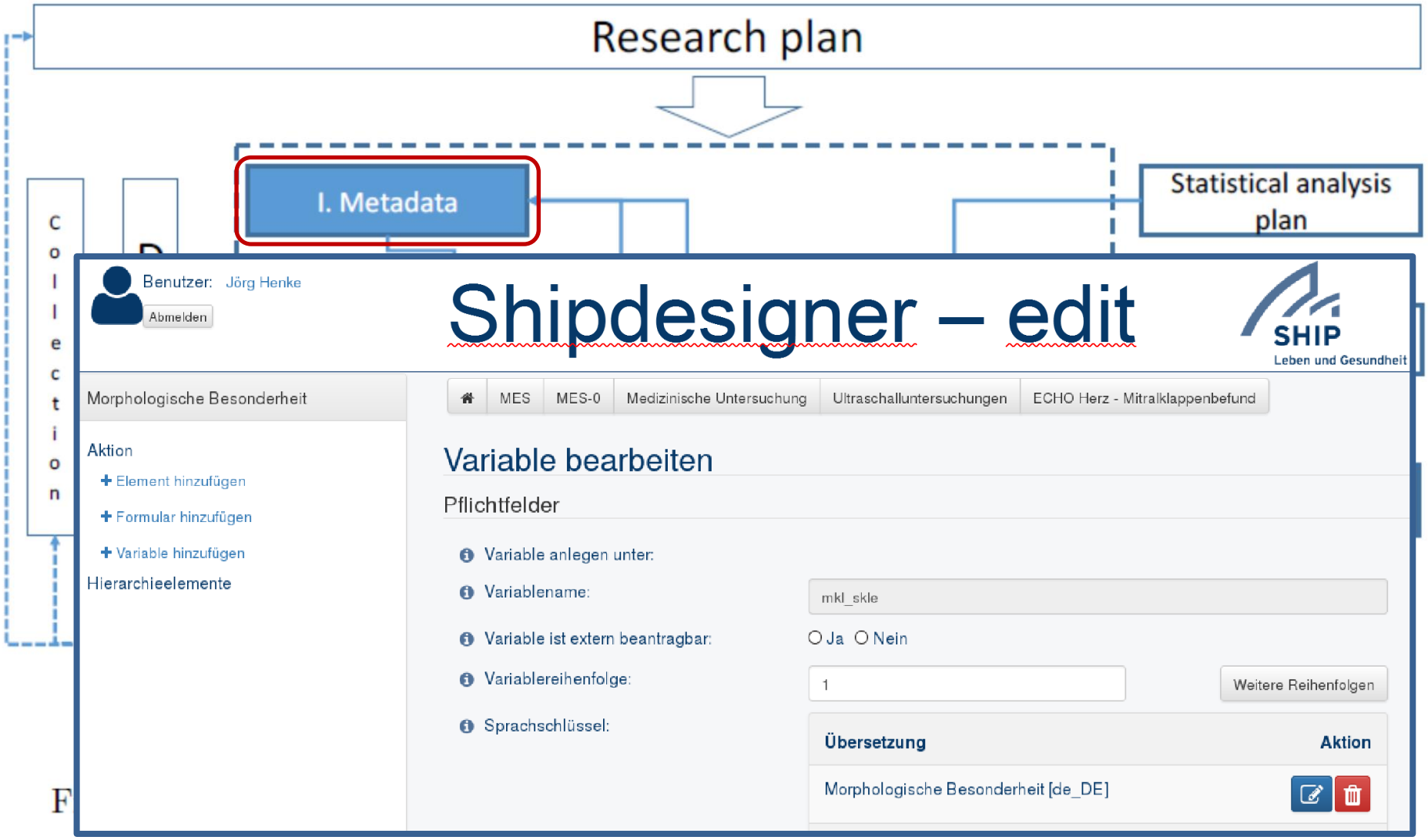


Figure 1: The main connections between the IDA steps and external components

Huebner, Cessie, Schmidt, Vach et al. 2018 Obs. Studies

Framework: Initiale Datenanalyse



Huebner, Cessie, Schmidt, Vach et al. 2018 Obs. Studies

Framework: Initiale Datenanalyse

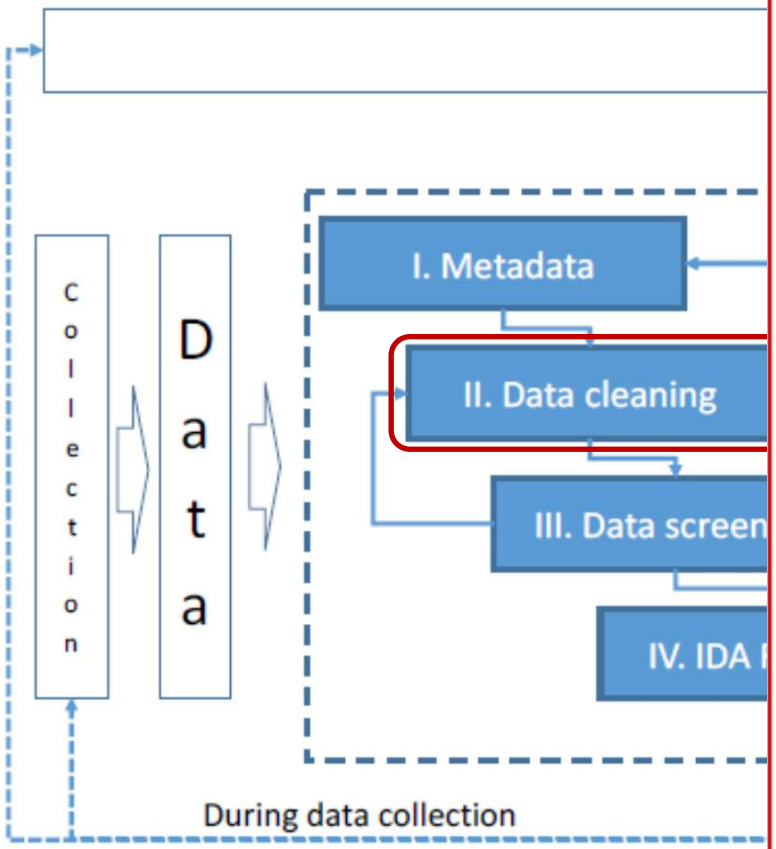
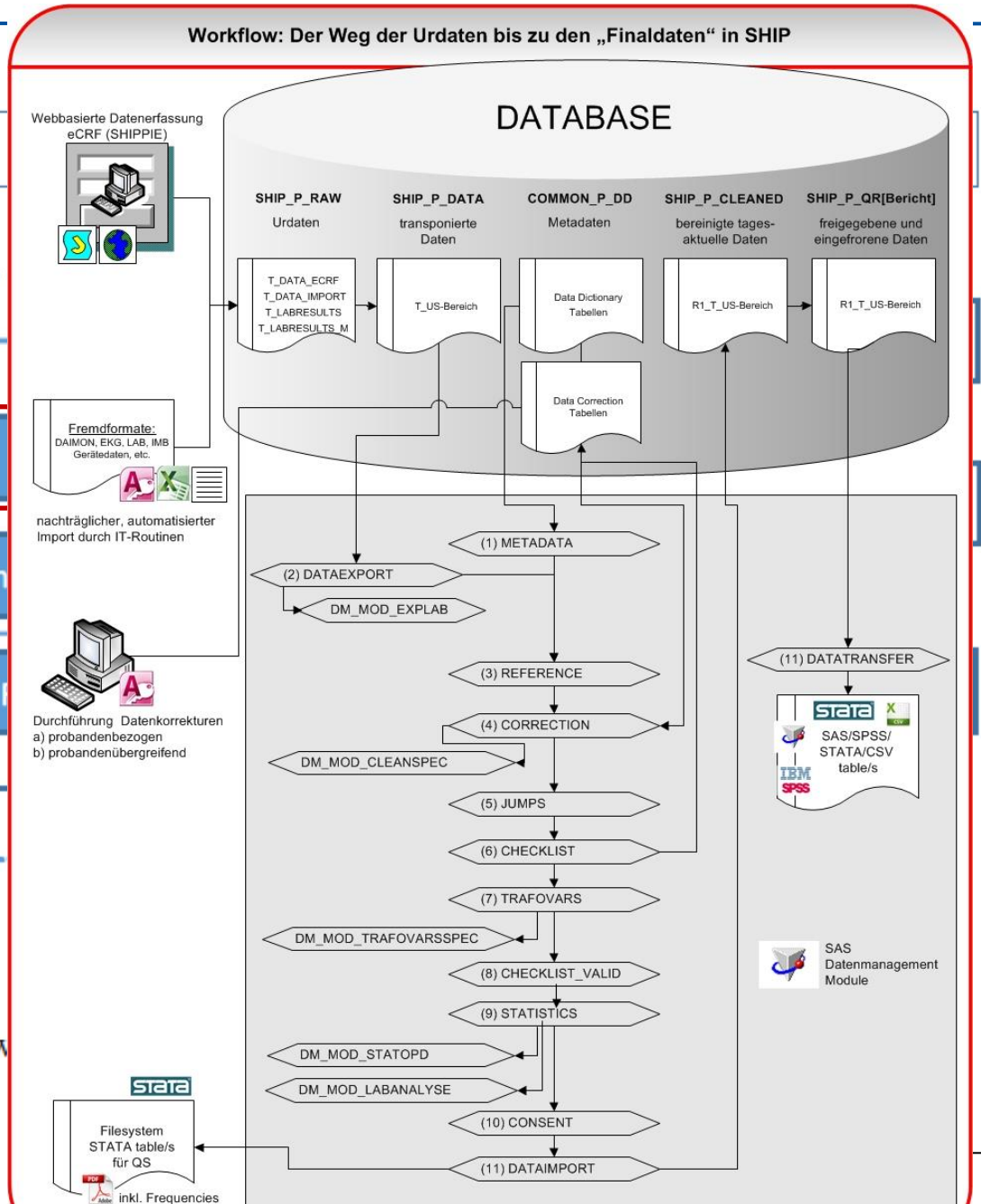


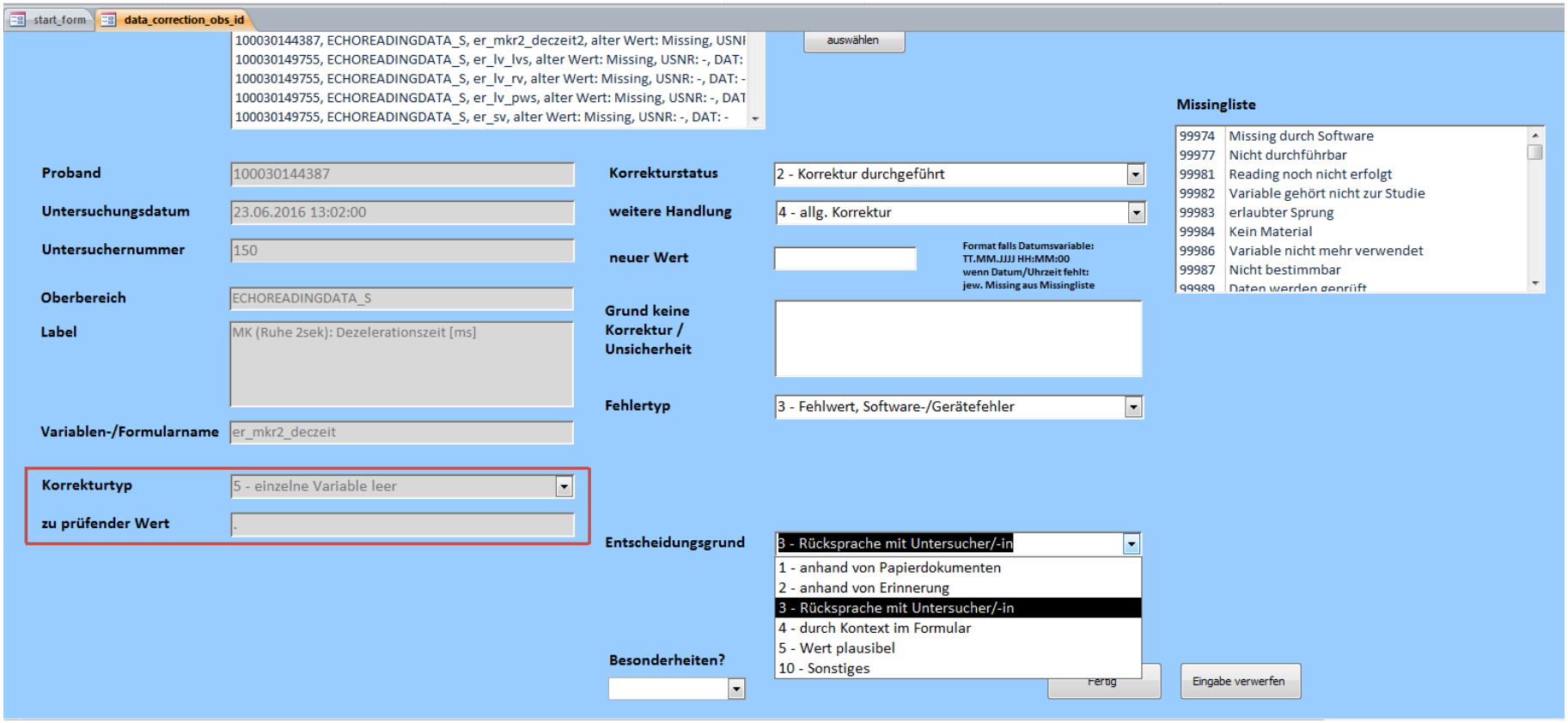
Figure 1: The main connections betw

Werner et al. 2016 KSFE



Modul – CHECKLIST / Prüffallkorrektur

- Schritt 2: Durchführung der Datenkorrektur



„Zero Tolerance“ für Auffälligkeiten, z.B. bei Consistency

Framework: Initiale Datenanalyse

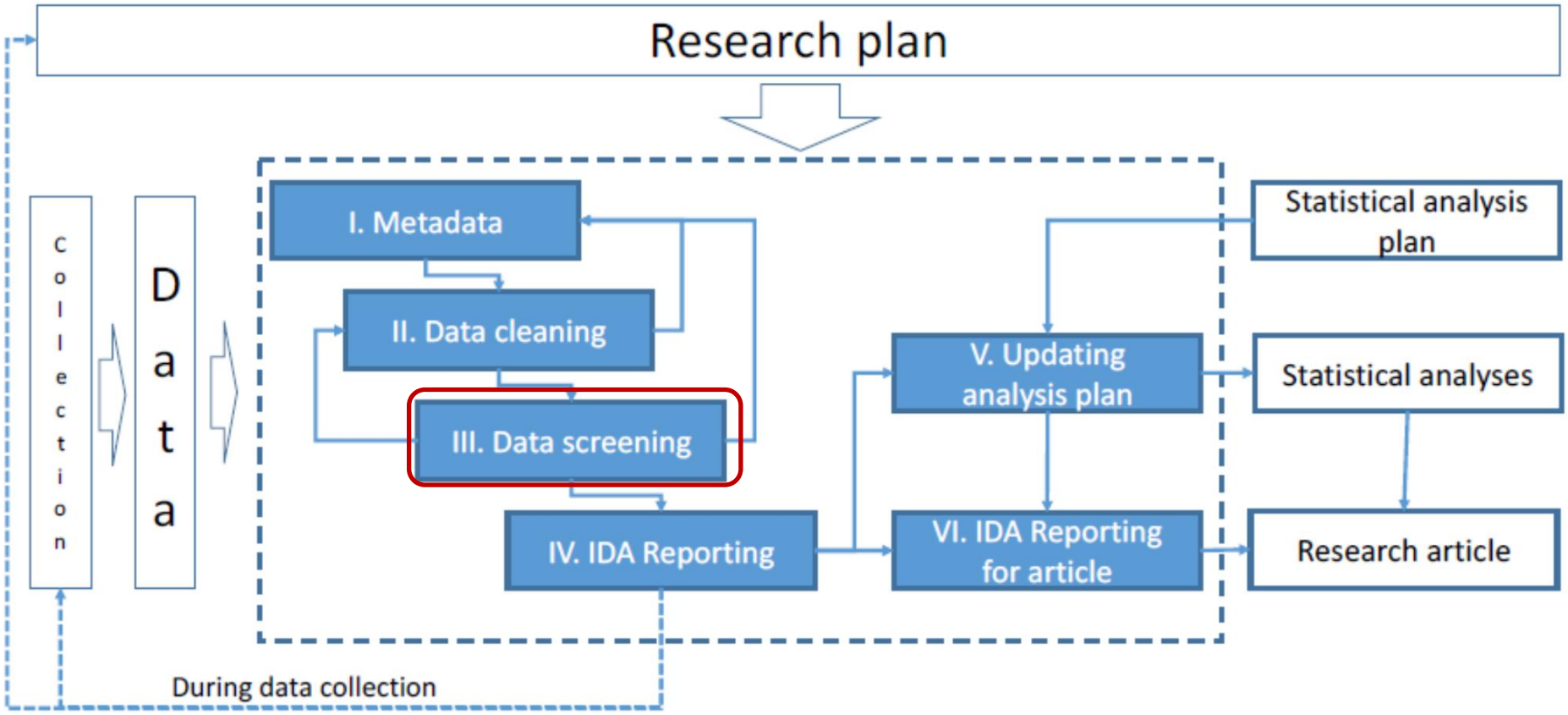
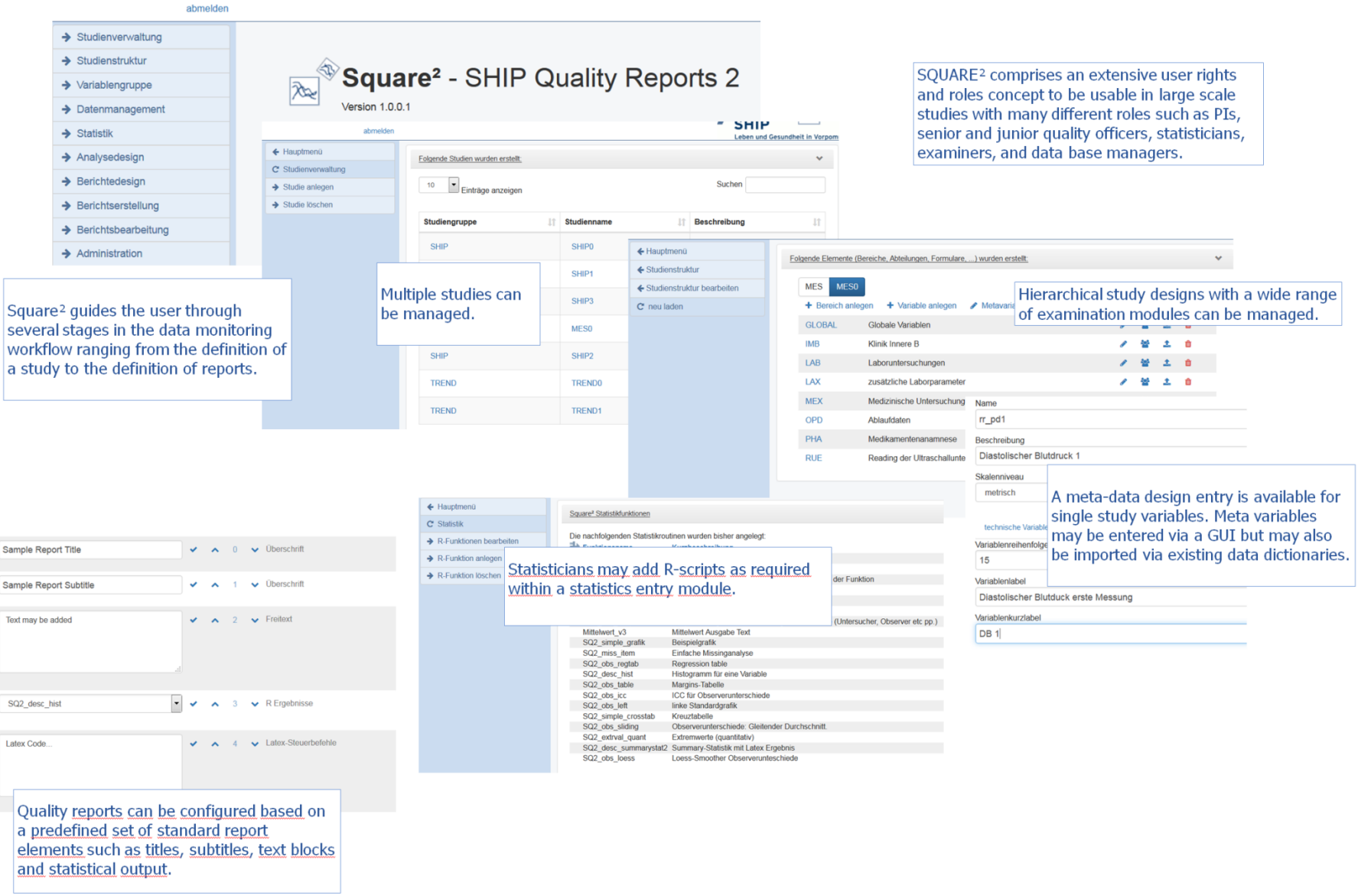


Figure 1: The main connections between the IDA steps and external components

Huebner, Cessie, Schmidt, Vach et al. 2018 Obs. Studies



SQUARE² comprises an extensive user rights and roles concept to be usable in large scale studies with many different roles such as PIs, senior and junior quality officers, statisticians, examiners, and data base managers.

Square² guides the user through several stages in the data monitoring workflow ranging from the definition of a study to the definition of reports.

Multiple studies can be managed.

Hierarchical study designs with a wide range of examination modules can be managed.

A meta-data design entry is available for single study variables. Meta variables may be entered via a GUI but may also be imported via existing data dictionaries.

Statisticians may add R-scripts as required within a statistics entry module.

Quality reports can be configured based on a predefined set of standard report elements such as titles, subtitles, text blocks and statistical output.

Schmidt et al. 2017 IOS press

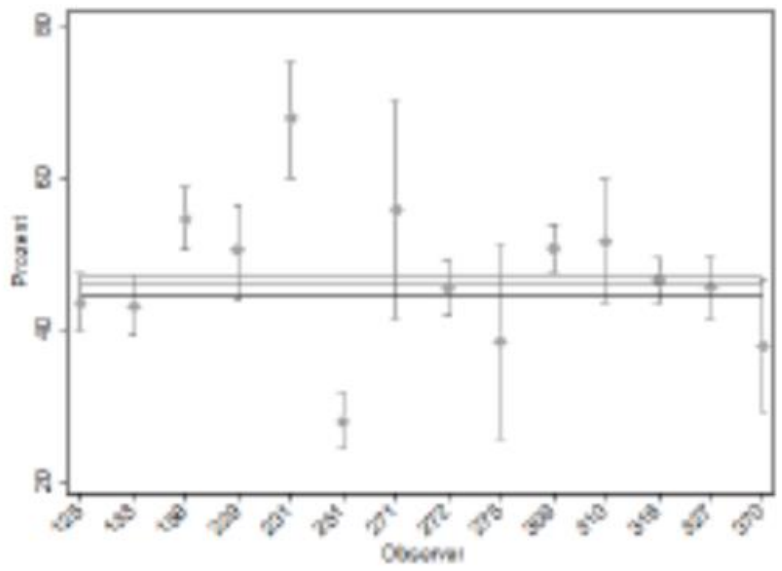
study administration | study structure | variable group | data management | statistics | templates | quality reports | administration

define report | analysis matrix | report design | generate report | show report

report 33

report

Statistikroutinenübersicht HGP S3



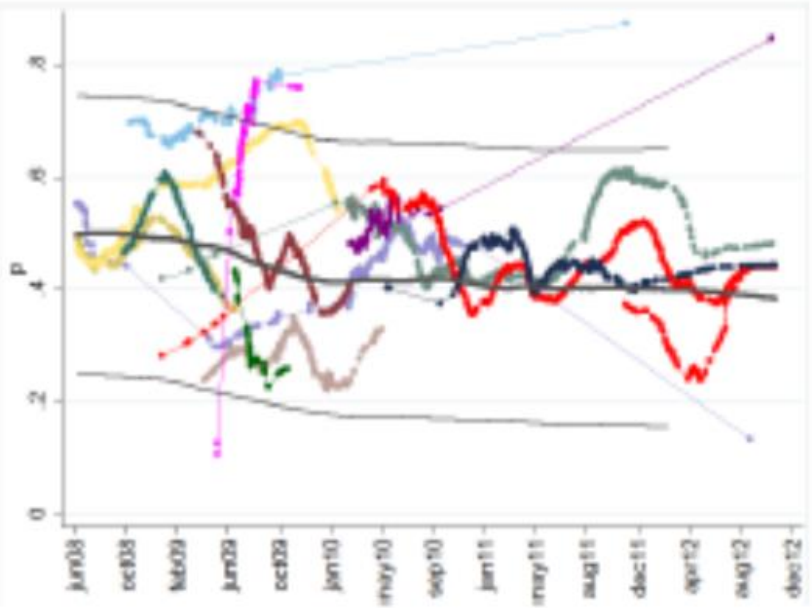


Abbildung 1.2: Mittlere Prävalenz für MRT_KOPF_DEMY adjustiert für Alter und Geschlecht nach logistischer Regression.

Fazit

1. Datenqualitätsanalyse muss Anwendungskontext berücksichtigen
2. Einheitliche Konzepte zur Datenqualitätsanalyse erforderlich
3. Umfassendes Metadatenmanagement zur Implementation automatisierter Datenqualitätschecks essentiell
4. Metadatenstandards für Datenqualitätschecks erforderlich
5. Umsetzung und Folgen aus Datenqualitätschecks auf den Ebenen Completeness, Consistency, Accuray sehr unterschiedlich
6. Umfassende infrastrukturelle Anforderungen zur Umsetzung automatisierter Datenqualitätschecks erforderlich
7. Standards für Routinen und Tools wichtig für Datenqualitätsanalyse

http://www.medizin.uni-greifswald.de/icm/

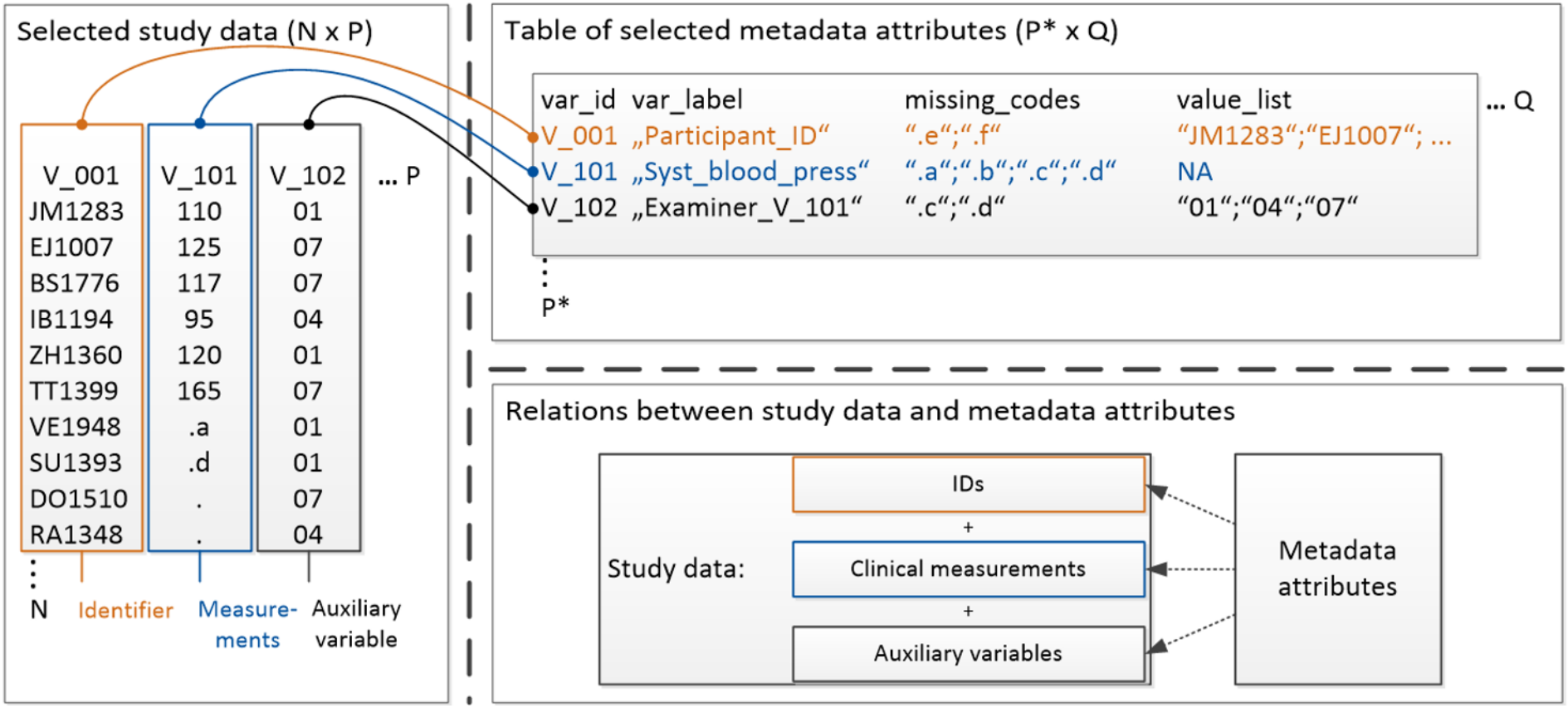
carsten.schmidt@uni-greisfwald.de

SHIP
Study of Health in Pomerania

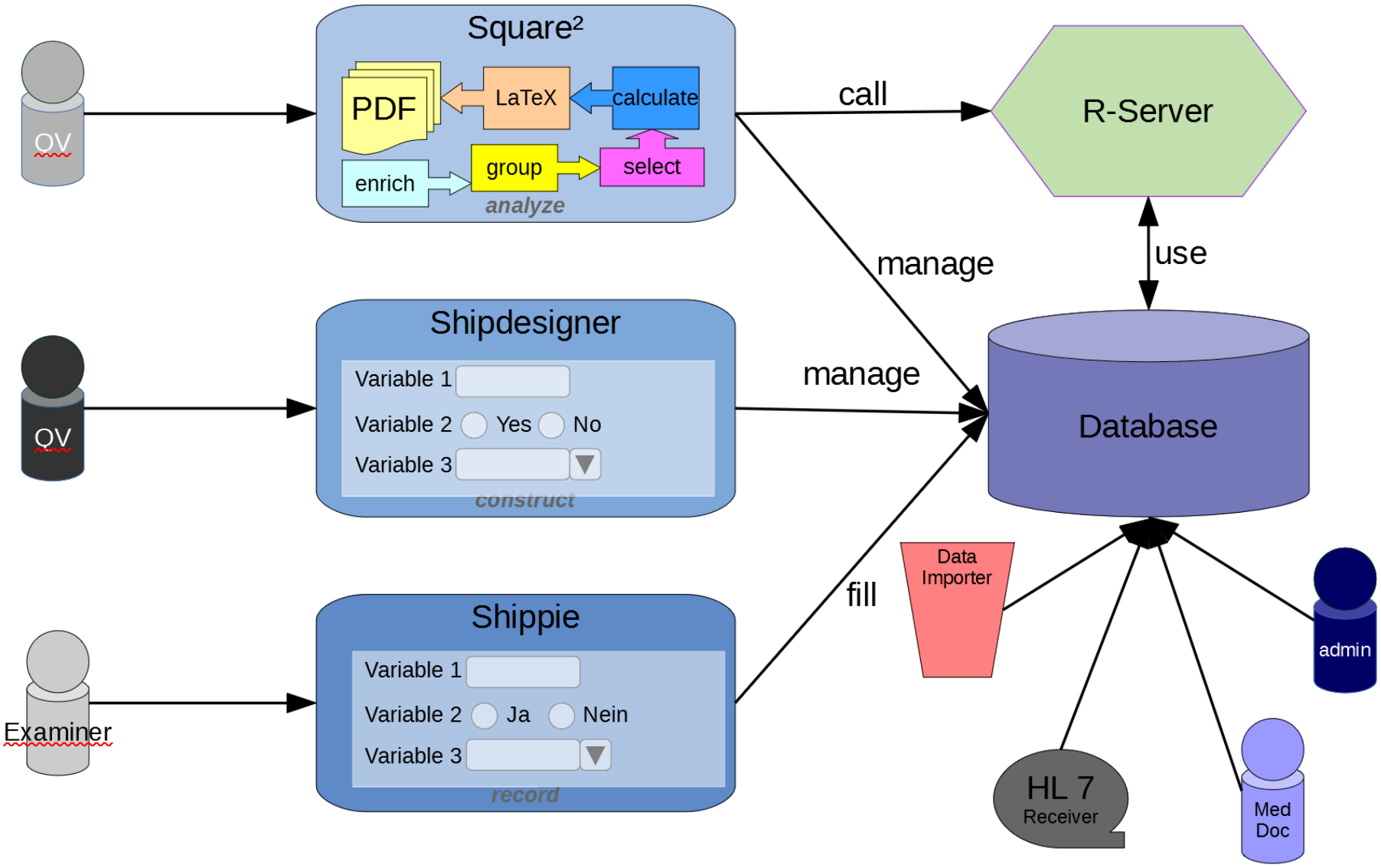
Community Medicine

Universitätsmedizin
GREIFSWALD

Anwendungen Metadaten



Richter et al. submitted, Auszug


Square²
PDF
LaTeX
calculate
enrich
group
select
analyze
OV
call
R-Server
manage
use
Shipdesigner
Variable 1
Variable 2
Yes
No
Variable 3
construct
OV
manage
Database
Data Importer
fill
Shippie
Variable 1
Variable 2
Ja
Nein
Variable 3
record
Examiner
admin
HL 7 Receiver
Med Doc