



Automatisierung des
Datenreinigungsprozesses in einer
internationalen multizentrischen
Follow-up-Studie (I.Family-Studie)

Willempje Hummel-Bartenschlager

Leibniz-Institut für Präventionsforschung und Epidemiologie - BIPS



This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 266044



Überblick

- Eckpunkte der I.Family-Studie
- Aufgaben des BIPS Datenmanagements innerhalb der Studie
- Datenbereinigung in der Vorgänger-Studie IDEFICS
- Umstrukturierung der Eingabe-Datenbanken in I.Family
- Datenbereinigungsprozess in I.Family
- Ergebnisse



I.Family

- Multizentrische Follow-Up-Studie in 8 europäischen Ländern
- Erforschung von Einflussfaktoren auf das Ernährungsverhalten europäischer Kinder, Jugendlicher und ihrer Eltern
- BIPS als Koordinator
- 2013 Start der Erhebung
 - über 2000 Variablen bei mehr als 17000 Probanden
- 10 Eingabe-Datenbanken
 - auf zentralem Server
 - Erst- und Zweiteingabe
 - diverse Interviews und Tests auf lokalen Computern



Aufgaben des Datenmanagements im BIPS in I.Family

- Programmierung und Bereitstellung aller Datenbankapplikationen
- Aufbereitung der in Erst- und Zweiteingabe-DB eingegebenen Daten
 - Bereinigung von Dopplern
 - Löschung leerer Datensätze/falsch eingegebener Daten
 - erste Datenkorrekturen
 - Erstellung von Dateien als Grundlage für die Weiterverarbeitung
- Datenbereinigung
 - Vergleich Erst- und Zweiteingabe
 - Prüfung auf unplausible Werte
 - Bereitstellung von Korrektur-Datenbanken
- Bereitstellung Analyse-Datenbank

Datenbereinigung in der IDEFICS-Studie

- Vorgängerstudie von I.Family
 - Erforschung und Prävention von Übergewicht und Lebensstilbedingten Erkrankungen bei Kindern in acht europäischen Ländern
 - Erhebung von 2007 bis 2010
- T0 (Studienphase 1)
 - SAS-Programme
 - Prüfung auf Unterschiede zwischen Erst- und Zweiteingabe
 - Nach Beendigung der gesamten Datenerhebung
 - Prüfung auf unplausible Werte
- Übermittlung der Korrekturlisten
- Manuelle Eingabe der Korrekturen durch die Zentren in die Ersteingabe-Datenbank



Datenbereinigung in der IDEFICS-Studie

- T1 (Studienphase 2)
 - Erster Schritt zur Automatisierung
 - Einsatz von Access-Datenbanken für die Korrekturen
 - Automatischer Export der Korrekturen in die Ersteingabe-Datenbank

Country	Sweden									
ID Number	Counter/ Current Number	Questionnaire	Question Number	Description	First Entry	Second Entry	Choice <small>(red = enter a choice!)</small>	New Value <small>(orange = obligatory input!)</small>	Reason for Choice/Reason for New Value/Reason for No Value	
Press key 'F4' to re-drop list down, if necessary										
		Parental Questionnaire	04.	Mother's height [cm]	169	109	1 First Entry 2 Second Entry 3 New Value 4 No Value 5 In Process			Delete inputs for this record
		Parental Questionnaire	04.	Father's height [cm]	169	109				Delete inputs for this record
		Parental Questionnaire	07.	Number of persons in household	5	4				Delete inputs for this record



Vorteile durch Einsatz von Korrektur-Datenbanken

- Weniger fehlerträchtig
- Erhebliche Zeitersparnis
- Elektronische Dokumentation der Korrekturen gemäß GEP

Umstrukturierung der Eingabe-Datenbanken in I.Family

- Access-Backend-Datenbanken wurden durch MySQL-Datenbanken ersetzt
 - Erhöhte Datensicherheit
 - Unbegrenzte Dateigröße
 - Bessere Performance durch schnelleren Zugriff und Datentransfer
 - Änderungen an den Daten-Tabellen zentral
 - Einsatz von Trigger
 - Nachverfolgung von Änderungen an den Daten



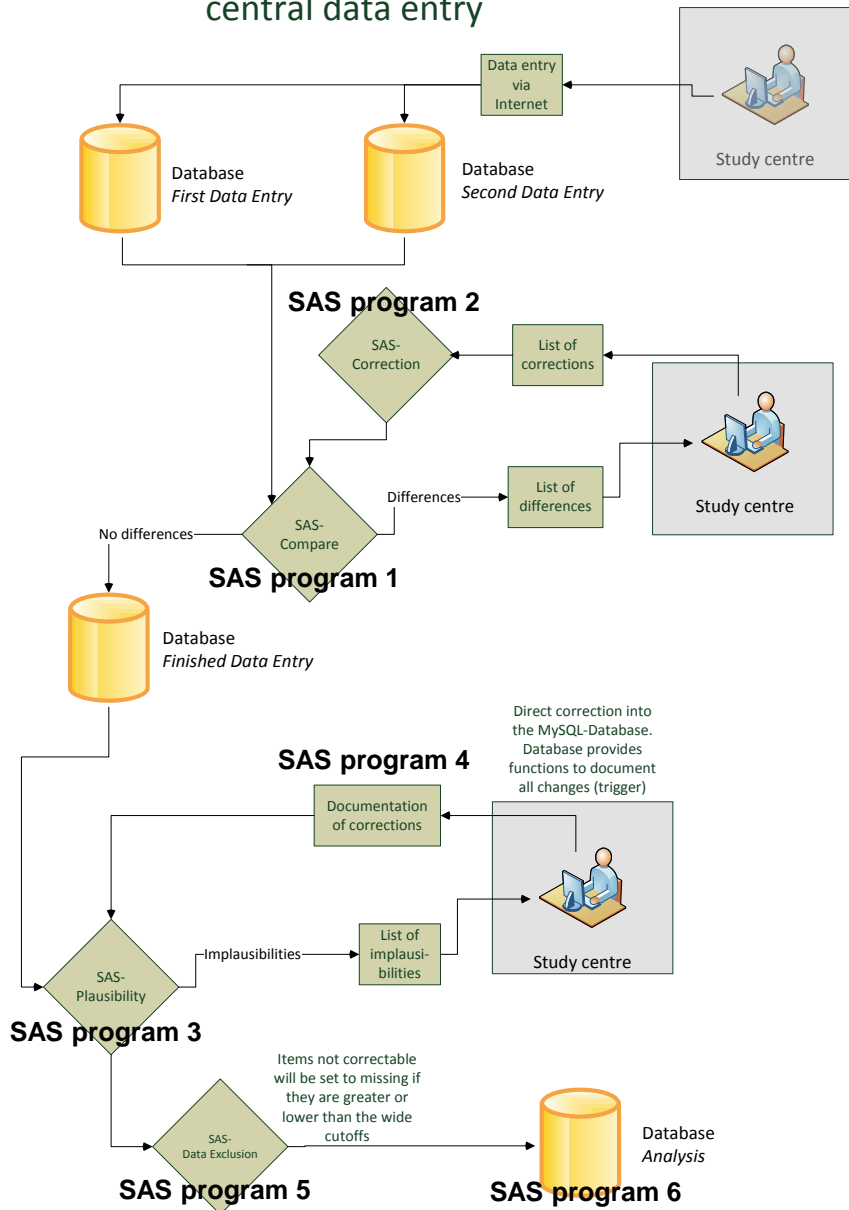
Einsatz eines zentralen Datenservers

- Bereitstellung der Eingabe- und Korrektur-Datenbanken auf zentralem Server
 - Frontend: Access-Datenbank
 - Programmierte erste Plausibilitätsprüfungen
 - Backend: MySQL-Datenbank
 - eine Datenbank für alle Zentren
 - Zusammenspielen der Daten entfällt

Vorteile eines zentralen Datenservers

- Bereitstellung der Anwendungsprogramme
- Problemlose und schnelle Bereitstellung von Versionsänderungen
- Unkomplizierter Datentransfer ohne Zwischenschritte
 - Verfügbarkeit der Daten für Datenmanagement direkt nach Eingabe
 - Vermeidung unterschiedlicher Versionen

Data entry and plausibility checks in I.Family central data entry



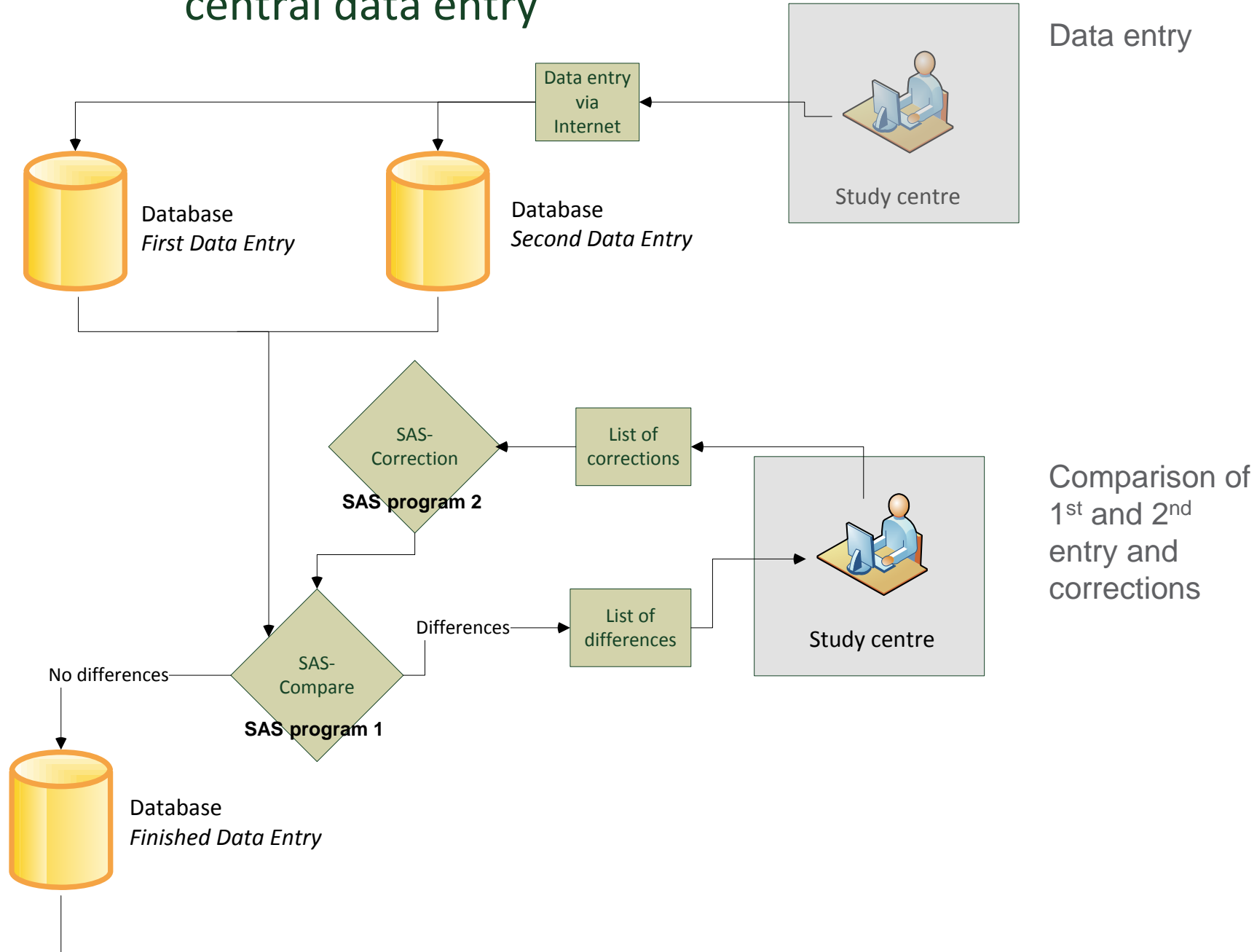
Data entry

Comparison of 1st and 2nd entry and corrections

Plausibility checks and corrections


Standardised corrections / exclusion of non-correctable values

Data entry and plausibility checks in I.Family central data entry





Korrektur-Datenbank für die Korrektur von Unterschieden zwischen Erst- und Zweiteingabe

frm_ComparedValues


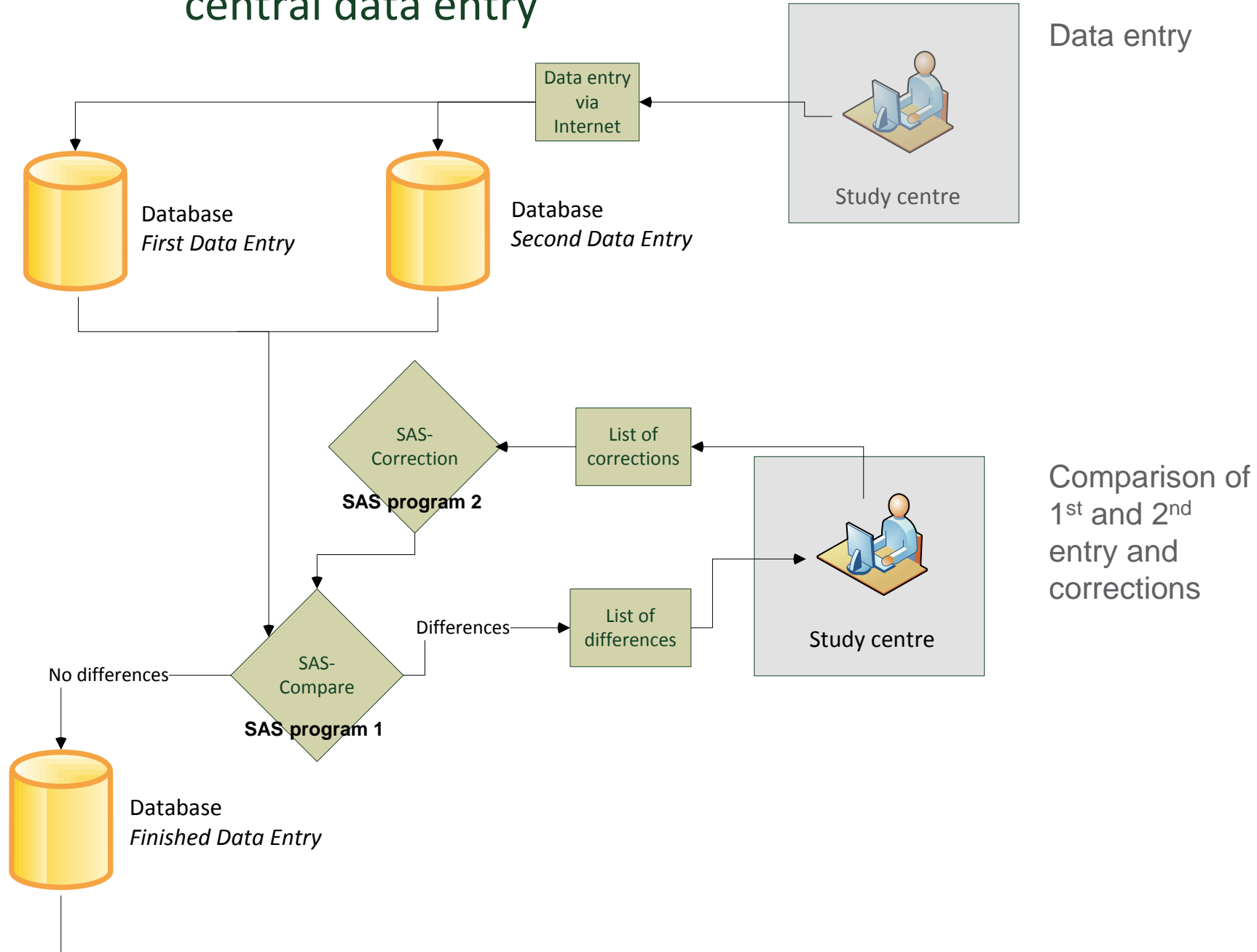
Country

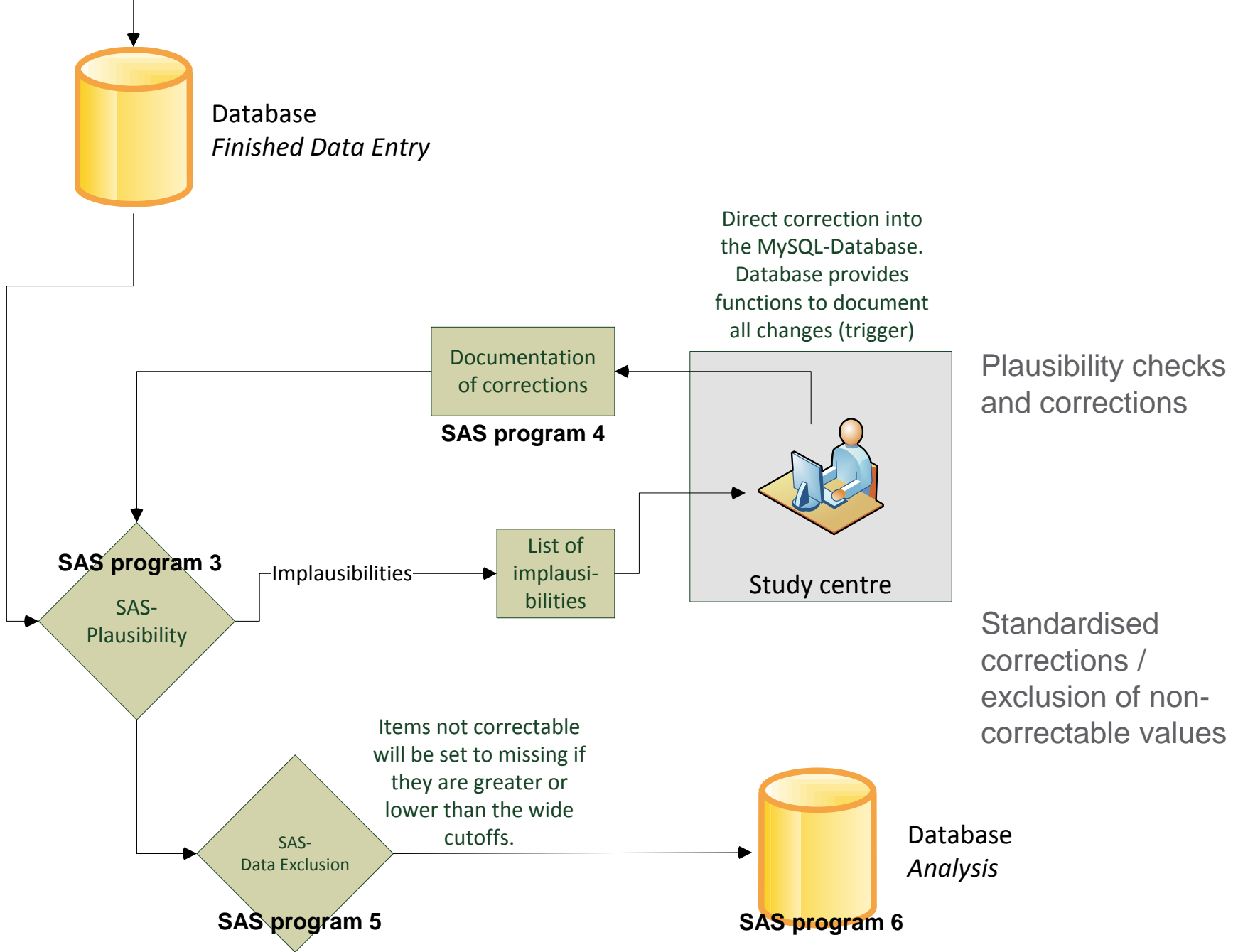
ID Number	Current Number/ Counter for day of wearing	Questionnaire	Question Number	Description	First Entry	Second Entry	Choice	New Value	Reason for Choice/Reason for New Value/Reason for No Value
<p>Incomplete Records</p> <p>Sort by ID Number Sort by Questionnaire</p> <p>Press key 'F4' to re-drop list down, if necessary</p>									
		Family Questionnaire	19.	Main occupational status of participant		2	▼		
									Delete inputs for this record
		Family Questionnaire	19.	Main occupational status of spouse/partner		3	▼		
									Delete inputs for this record

(red = enter a choice!) (orange = obligatory input!)
 1 = First Entry
 2 = Second Entry
 3 = New Value
 4 = No Value
 5 = In Process

If 'other', please specify

Data entry and plausibility checks in I.Family central data entry





Database
Finished Data Entry

Direct correction into the MySQL-Database.
Database provides functions to document all changes (trigger)

Plausibility checks and corrections

Documentation of corrections
SAS program 4

SAS program 3
SAS-Plausibility

List of implausibilities

Study centre

Standardised corrections / exclusion of non-correctable values

Items not correctable will be set to missing if they are greater or lower than the wide cutoffs.


SAS program 5

SAS program 6

Database
Analysis

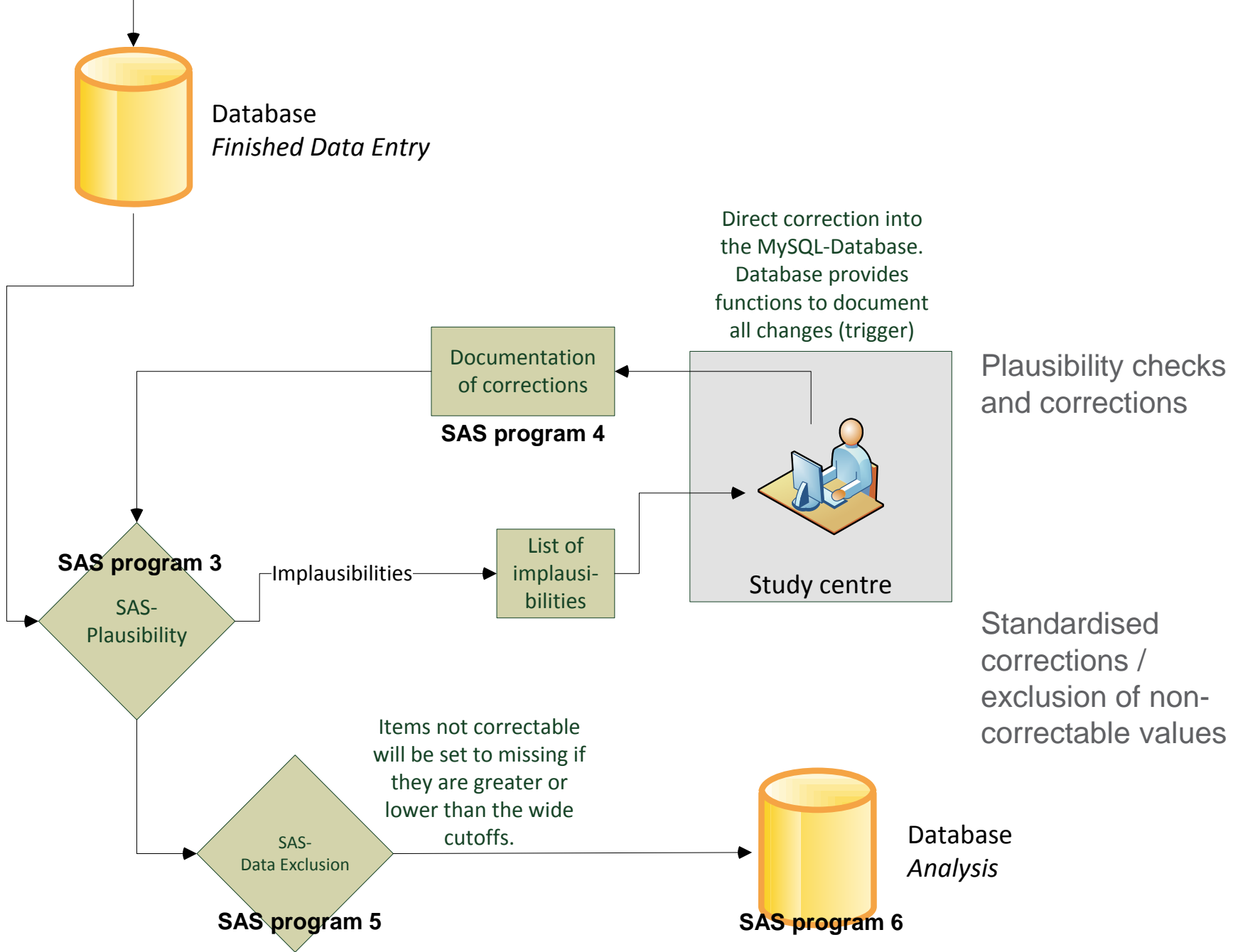


CorrectionDocuDB: Dokumentations-Datenbank für die Behandlung implausibler Werte

frm_Plausi


Country

ID Number	Current Number/Counter for day of wearing	Questionnaire	Question Number	Question Text	Error Description	Implausible Value	Decision	Reason for correction/no correction
<div style="display: flex; justify-content: space-between; align-items: center;"> <div style="border: 1px solid red; padding: 5px; color: red; font-weight: bold;">Incomplete Records</div> <div style="text-align: center;"> <p>Press key 'F4' to re-drop list down, if necessary</p> </div> <div style="text-align: right;"> <p>1 = Corrected 2 = Correction not possible 5 = In process</p> </div> </div>								
<div style="display: flex; justify-content: space-between;"> Sort by ID Number Sort by Questionnaire </div>								
[Redacted]	1	Activity diary for parents (Daily information)	Day 1	Time putting on accelerometer in the morning-hour	Using the bike (2nd) before wearing the accelerometer on	-	[Red]	[Dropdown]
<input type="button" value="Delete inputs for this record"/>								
[Redacted]	3	Activity diary for parents (Daily information)	Day 3	Starting time 1 (taking accelerometer off) to exercise-hour	Taking of (1st) before/after wearing accelerometer	-	[Red]	[Dropdown]
<input type="button" value="Delete inputs for this record"/>								

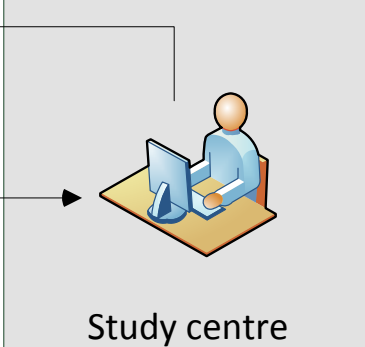


Database
Finished Data Entry

Direct correction into the MySQL-Database.
Database provides functions to document all changes (trigger)

Plausibility checks and corrections

Documentation of corrections
SAS program 4



Study centre

SAS program 3
SAS-Plausibility

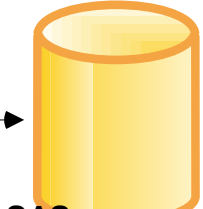
Implausibilities

List of implausibilities

Standardised corrections / exclusion of non-correctable values

Items not correctable will be set to missing if they are greater or lower than the wide cutoffs.

SAS program 5
SAS-Data Exclusion



SAS program 6

Database Analysis

Ablauf der SAS-Programme

- Alle SAS-Programme laufen nach Start automatisch nacheinander ab
- Speicherung von Zwischenergebnissen in temporären SAS-Dateien
 - Weiterverarbeitung vom nächsten SAS-Programm
- Import der Textdateien in Access-Korrektur-Datenbanken über programmierte Importfunktion



Ergebnisse der Automatisierung des Datenbereinigungsprozesses

- Effektive Nutzung von Ressourcen
- Zeitnahe Bereinigung der erhobenen Daten von Anfang an
- Frühzeitige Erkennung von Problemen bei der Datenerhebung bzw. der Dateneingabe
- Schnelle Bereitstellung des ersten bereinigten Datensatzes für die Auswertung

- Ziel: Automatischer Import der Outputs in die entsprechenden Datenbanken



Noch Fragen?

Noch Fragen?

Noch Fragen?

Noch

Fragen?



Danke für Ihre Aufmerksamkeit



Kontakt

Willempje Hummel-Bartenschlager
hummel@bips.uni-bremen.de