

Clean Data – Behind the Scenes

Philipp Wittmann, ABCSG Studienzentrale – Clinical Data Management

„SCIENCE
FROM THE HEART
OF EUROPE“



Definition von ‚Clean Data‘

- Pharma-/Biotech-/Lebensmittelindustrie: **ALCOA(+)** Prinzipien

A	Attributable (zuordenbar)	Es muss klar ersichtlich sein, wer die Daten dokumentiert oder geändert hat (Audit Trail).
L	Legible (lesbar)	Die Daten müssen (dauerhaft) lesbar und nachvollziehbar sein (zB. keine unleserlichen handschriftlichen Korrekturen; klare, standardisierte Codierungen).
C	Contemporaneous (zeitgleich)	Die Daten müssen zeitgleich mit der Beobachtung/Durchführung aufgezeichnet werden (keine nachträgliche Eintragung von früher erhobenen Werten).
O	Original	Die Daten müssen das Original oder eine exakte, zertifizierte Kopie des Originals sein.
A	Accurate (genau)	Die Daten müssen korrekt und plausibel sein
+/C	Complete (vollständig)	Keine fehlenden oder unvollständigen Informationen.
+/C	Consistent (beständig)	Die Daten sind über die gesamte Studie hinweg einheitlich und widerspruchsfrei .
+/E	Enduring (dauerhaft)	Die Daten müssen über den gesamten vorgeschriebenen Zeitraum aufbewahrt werden.
+/A	Available (verfügbar)	Die Daten müssen bei Bedarf für Audits oder Überprüfungen zugänglich sein.

vollständiges &
korrektes
Datenset

Notwendigkeit und Relevanz von ‚Clean Data‘

- ‚Clean Data‘ ist relevant, sobald mit **größeren Datenmengen** gearbeitet wird und diese zur **Auswertung** kommen
- **Publizierte Daten** müssen korrekt und vollständig sein
- **Abhängigkeit derzeitiger und künftiger Therapieentscheidungen** von Ergebnissen ausgewerteter Studien (Überleben von PatientInnen, Lebensqualität)
- Gesetzliche/Regulatorische Vorgaben

No data, no science.

Know data, know science!

Gesetzliche/Regulatorische Vorgaben

- ICH E6 (R3) – Good Clinical Practice (GCP)
- EMA Clinical Trials Regulation (CTR)
- EMA Guidelines on Computerized Systems and Electronic Data in Clinical Trials
- EU GMP Annex 11 – Computerized Systems
- FDA 21 CFR Part 11 (Electronic Records, Electronic Signatures)
- Österreichisches Arzneimittelgesetz
- ...

Wie kann ‚Clean Data‘ erreicht werden?

- Prüfung der Daten **im gesamten Studienverlauf** auf Plausibilität, Vollständigkeit und Korrektheit
- Überlegungen und Prozesse zu ‘Clean Data’
 - Phase 1 – **eCRF Setup**
 - Phase 2 – **Data Cleaning**
(während des gesamten Studienverlaufs)
 - Phase 3 – **Database Lock/Closure**

- Nutzung eines **geeigneten Electronic Data Capture (EDC) Systems** erforderlich (FDA 21 CFR Part 11, EMA Guidelines)
 - Validated system
 - Security features, user role/access management, continuity, disaster recovery, archiving
 - Audit Trail
 - Electronic signatures
- Erstellung eines **klaren, gut strukturierten, vollständigen, protokollkonformen und voll funktionsfähigen eCRF**
 - eCRF Design in engem Austausch mit relevanten Abteilungen (Statistik, Medical & Safety, Monitoring, Projektmanagement)
 - Nutzung von Programmfeatures zu Fehlerminimierung ('Only collect-if' conditions, Real-time checks)
 - Verwendung standardisierter Metadaten i.e. Variablen-Codes, Datentypen/-struktur, Ausprägungen etc.
 - Validierung mittels User Acceptance Testing

Phase 1 – Setup & Dokumenterstellung

- Nutzung von Programmfeatures zu Fehlerminimierung
 - ∅ ~700 programmierte Real-time Cecks (PoEC)
 - ∅ ~1200 programmierte 'Only collect-if' Bedingungen (OCI)
 - ∅ ~150 programmierte Post-entry checks (PEC)
 - ∅ ~25 zusätzlich programmierte Data Line Listings (DLL)

Warnings

The following warnings have been generated:

Message

 T-Stage is not in line with documented tumor size. Please check and correct.

Code	Item type	Optional	Mandatory	Hidden	RFC	Only collect data if	Show Icons
SUBJIDDER	Question				x		x
REGDATDER	Question				x		x
CHEMPERF	Question				x		x
CHEMDAT	Question				x	CHEMPERF = Y	x
CHEMQG	Group					CHEMPERF = Y	
CHEMTEST	Group Question				x		x
CHEMSIGRES	Group Question	x			x	(question:cycle not_oneof [4, 5])or(question:cycle = 5 and not isknown(CHEMRES))	x
CHEMRES	Group Question				x	(question:cycle not_oneof [4, 5])or(question:cycle = 5 and not isknown(CHEMRES))	x
CHEMU	Group Question				x	isknown(CHEMRES(this))	x
CHEMUO	Group Question				x	CHEMU(this) = OTHER	x
CHEMSIGNRLO	Group Question	x			x	isknown(CHEMRES(this)) and (not isknown(CHEMSIGNRHI(this)))	x
CHEMNRLO	Group Question				x	isknown(CHEMRES(this)) and (not isknown(CHEMSIGNRHI(this)))	x
CHEMSIGNRHI	Group Question	x			x	isknown(CHEMRES(this)) and (not isknown(CHEMSIGNRLO(this)))	x
CHEMNRHI	Group Question				x	isknown(CHEMRES(this)) and (not isknown(CHEMSIGNRLO(this)))	x

Editing: Validation condition for question T-Stage

Edit Check Options

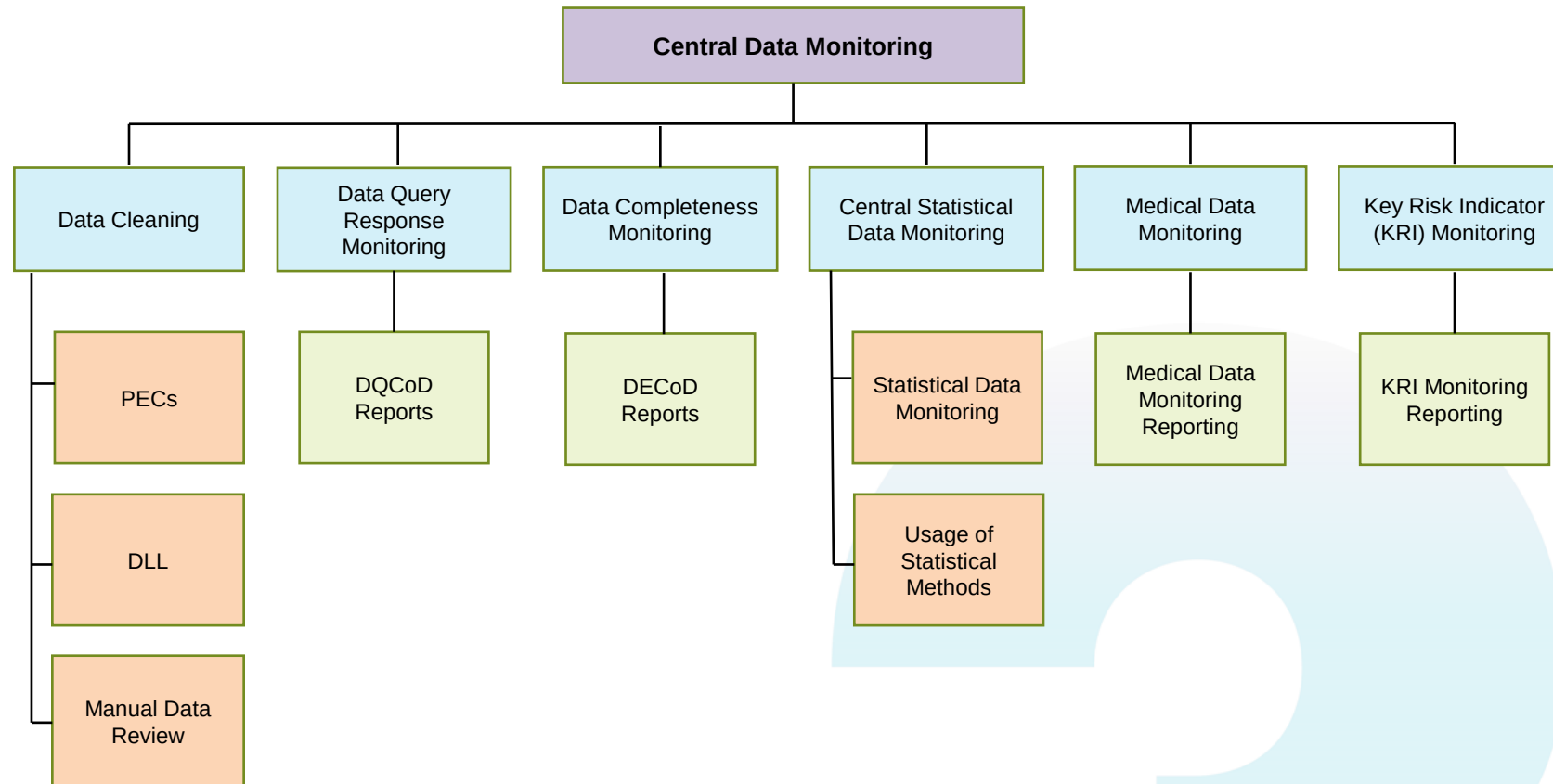
```

(me:value oneof ['0','60'] and isknown(TUMSIZE(this))) OR
(me:value = 10 and TUMSIZE(this) > 20) OR
(me:value = 11 and (TUMSIZE(this) <= 1 or TUMSIZE(this) > 5)) OR
(me:value = 12 and (TUMSIZE(this) <= 5 or TUMSIZE(this) > 10)) OR
(me:value = 13 and (TUMSIZE(this) <= 10 or TUMSIZE(this) > 20)) OR
(me:value = 14 and TUMSIZE(this) > 1) OR
(me:value = 20 and (TUMSIZE(this) <= 20 or TUMSIZE(this) > 50)) OR
(me:value = 30 and TUMSIZE(this) <= 50)
isknown( ANSTAGE(this) ) and
ANSTAGE(this) = 0 and me:value <> 60 or
(ANSTAGE(this) = 11 and me:value not_oneof ['10','11','12','13','14'] ) or
(ANSTAGE(this) = 12 and me:value not_oneof ['0','10','11','12','13','14'] ) or
(ANSTAGE(this) = 21 and me:value not_oneof ['0','10','11','12','13','14','20'] ) or
(ANSTAGE(this) = 22 and me:value not_oneof ['20','30'] ) or
(ANSTAGE(this) = 31 and me:value not_oneof ['0','10','11','12','13','14','20','30'] ) or
(ANSTAGE(this) = 32 and me:value not_oneof ['40','41','42','43','44'] )
                    
```

- Erstellung von **relevanten Studiendokumenten** zur einheitlichen Handhabung der Daten:
 - Data Completion Guideline (**DCG**) → Basisdokument zur Datenerfassung an den Zentren, dem Daten Review und dem Data Cleaning
 - Data Management Plan (**DMP**) → Beschreibung der verwendeten Datenbank sowie sämtlicher datenrelevanter Prozesse
 - Data Validation Plan (**DVP**) → Definition sämtlicher Checks, Listings und manueller Reviewprozesse während des Studienverlaufes
 - Spezifikationen, Risikobewertungen, Validierungsdokumente → Detailspezifikationen, Risikobewertungen und Validierungen der im DVP definierten Checks und Listings
- **Programmierung** sämtlicher Checks, Listings und Reports

Phase 2 – Central Data Monitoring

- Durchführung eines risikobasierten **Central Data Monitoring**



Phase 2 – Central Data Monitoring

- **Programm-gestütztes** Data Cleaning:
 - Post-Entry Checks (**PEC**)
 - Data Line Listings (**DLL**)
 - Statistical Data Monitoring
- **Manuelles** Data Cleaning
- **Medical** Monitoring
- Handling unplausibler Daten: **Data Queries**

ER/PR status assessed?			☒ Yes ✓	☐ No
	Target lesion?	Estrogen receptor status		Staining Intensity
✓	1. No	✓ Positive	✓	Moderate
✓	2. No	✓ Positive	✓	Strong

	Target lesion?	Estrogen receptor status		Staining Intensity
✓	1. No	✓ Positive	✓	Moderate
✓	2. Yes	✓ Positive	✓	Strong

	Target lesion?	Estrogen receptor status		Staining Intensity
✓	1. No	✓ Positive	✓	Moderate
✓	2. Yes	✓ Positive	✓	Strong

**CDM/Monitoring/
Medical&Safety:**
Rückfrage

Zentrum:
Response/Update

**CDM/Monitoring/
Medical&Safety:**
Closure

Text

The ER status of the target lesion is expected to be documented. Please check if any of the 2 already documented lines needs to be corrected or if an additional line with the target lesion needs to be added.

Second line is target lesion. Documentation corrected.

DCR closed

Phase 2 – Central Data Monitoring

- Programmierung & Erstellung von **Reports zur Datenqualität**:
 - Key Performance Indicators (**KPI**) Reports → Kennzahlen zur **Leistungsbewertung** (Performance)
 - Key Risk Indicators (**KRI**) Reports → Warnsignale zur **Früherkennung & Risiko-Minimierung**:
 - **Trends** (zB. sich wiederholende Stratifizierungsfehler)
 - **Patterns** (zB. Fehlende AE Dokumentation)
 - Statistische **Ausreißer** in der Dokumentation erkennen (zB. Laborwert in einer Visite viel höher als in allen anderen, aber noch unterhalb des erlaubten Maximums)
 - **Darstellung** der Ergebnisse → Studien-, Länder-, Zentren- und Patienten-**Ebene** (je nach Studie/ Fragestellung)

- Programmierung & Erstellung von **Reports zur vollständigen Datenerfassung:**

- **„Site Reminder“ Reports** —→ **Information** zum aktuellen und erwarteten Dokumentationsstatus
 - Schnelles Erkennen ausständiger Dokumentation, offener Data Queries, ...
 - Monatliche Aussendung

- **DECoD/DQCoD* Reports** —→ **Aufforderung zur Dokumentation**

*Data Entry Cut-off Date/
Data Query Cut-off Date

- Aufforderung zur Dokumentation aller ausständigen Daten **bis zu einem bestimmten Datum (CoD)**
- Aussendung zu definierten Zeitpunkten (großer allgemeiner Dokumentationsrückstand, vor IDMC Meetings, vor Analysen, etc.)

- ✓ **Regelmäßige Kommunikation und enge Zusammenarbeit** zwischen Data Management, Monitoring sowie Zentrenpersonal ist essentiell um ‘Clean Data’ sowohl während des Studienverlaufes als auch am Studienende zu gewährleisten.

Phase 3 – Database Lock/Closure

- **Verifikation ‘Clean Data’ Status** durch alle beteiligten Abteilungen (Data Management, Monitoring, Medical&Safety, Statistik)
- **Database Lock** (Daten nicht mehr veränderbar) → wichtig für die **Datenintegrität** (zum Zeitpunkt der Analyse/Publikation)
- **eCRF Signatures** (Principal Investigator oder Designee) → **Bestätigung der Korrektheit** der dokumentierten Daten (Verantwortung liegt immer beim Principal Investigator!)
- Erstellung von **Hashcodes** & **Download** aller Daten
- **Bereitstellung der Daten** für die Statistik (Verifikation der HashCodes für sämtliche Auswertungen erforderlich)

„SCIENCE
FROM THE HEART
OF EUROPE“

