**Projekt zur**

Entwicklung, Umsetzung und Evaluation von Leitlinien zum adaptiven Management von Datenqualität in Kohortenstudien und Registern

gefördert durch die Telematikplattform für Medizinische Forschungsnetze e. V.

# Fallzahlplanung Source Data Verification

**Markus Neuhäuser**

Institut für Medizinische Informatik, Biometrie und Epidemiologie

am Universitätsklinikum Essen

**Monitoring durch Source Data Verification (SDV)**

kann Qualitätsbewusstsein wecken und damit Datenqualität verbessern.

Ziel ist es nicht, Einträge zu korrigieren (da ohnehin nur SDV auf Stichprobenbasis).

Literaturrecherche:

Keine durch emprirische Untersuchungen begründete Empfehlungen

**Klinische Studien**

Monitoring ist grundsätzlich erforderlich (GCP),

aber es gibt eine Reihe von Faktoren, die bei der Entscheidung über Art und Umfang des Monitorings eine Rolle spielen

u.a.
- Phase der klinischen Prüfung
- Anzahl und geographische Lage der Prüfzentren
- zeitliche Dauer der Studie
- Datenerfassung elektronisch oder Papierform

**adaptives Monitoring**

(Ose et al.: Low budget GCP – am Beispiel Monitoring. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* 2004; 35: 54-62)

***häufigeres Monitoring bei niedriger Qualität,***

***selteneres Monitoring bei hoher Qualität.***

Pogash et al. (2001): bei mehr als 10 Abweichungen pro 10 000 Felder weitere 5% der CRFs

Zentren mit guter Datenqualität:

weniger große Stichprobe für die SDV

Datenqualität: anhand von Kenngrößen (Qualitätsscore)

und (wenn vorhanden) vorherigem SDV-Ergebnis

**Umfang der SDV je Zentrum**

Anzahl Patienten pro Zentrum nötig,

Patienten/Personen werden je Zentrum zufällig ausgewählt

Fallzahlplanung auf Basis des Anteils an Patienten/Personen mit mindestens einem fehlerhaften Eintrag (bzw. mind. x Fehleinträgen, oder mind. einem Fehleintrag in spezifizierten wichtigen Variablen)

→ Binomialverteilung kann angenommen werden,
wird durch Normalverteilung approximiert

(1-α)-Konfidenzintervall für den Anteil $p$:

$$\left(\hat{p} - \delta, \hat{p} + \delta\right)$$

erforderliche Fallzahl:

$$N \geq \frac{p(1-p)}{\delta^2} \cdot z_{\alpha/2}^2$$

$z_{\alpha/2}$ Quantil der Standard-Normalverteilung, z.B. 1.96 für $\alpha = 0.05$.

| Anteil (Annahme) | benötige Fallzahl für $\delta = 0.02$ |
|---|---|
| 0.01 | 96 |
| 0.02 | 189 |
| 0.03 | 280 |
| 0.04 | 369 |
| 0.05 | 457 |
| 0.06 | 542 |
| 0.08 | 707 |
| 0.10 | 865 |
| 0.15 | 1225 |
| 0.20 | 1537 |
| 0.30 | 2017 |
| 0.40 | 2305 |
| 0.50 und mehr | 2401 |

Fallzahl umso größer

- je größer der Anteil $p$,
- je kleiner $\delta$.

z.B. $p = 0.10$:

| $\delta$ | benötige Fallzahl |
|---|---|
| 0.01 | 3458 |
| 0.02 | 865 |
| 0.03 | 385 |
| 0.04 | 217 |
| 0.05 | 139 |
| 0.06 | 97 |

Zusätzlich zum Einfluss von $p$ könnte man bei schlechter Datenqualität ein kleineres $\delta$ fordern

→ noch stärkere Fallzahlunterschiede zwischen Zentren mit guter und schlechter Datenqualität.

Ab der 2. SDV:

Anteil $p$ aus der vorherigen SDV für die Fallzahlbestimmung bekannt

Statt Anteil an Patienten mit mindestens einem fehlerhaften Eintrag:

Anteil fehlerhafter Einträge insgesamt

→ Fallzahlschätzung analog, sofern Unabhängigkeit vorausgesetzt wird

evtl. 2 Fallzahlschätzungen

→ maximale Anzahl ergibt die Fallzahl

**Tiefe der SDV**

Wie viele und welche Variablen sollen überprüft werden?

- Alle (bzw. alle neuen) Einträge
- Auswahl nach Wichtigkeit
- Je Gruppe (z.B. Labor) mindestens ein Wert

aber auch formale Fallzahlplanung für einen zu schätzenden Anteil pro Patient möglich (unter Annahme der Unabhängigkeit, Binomialverteilung, ohne Approximation durch Normalverteilung)

z.B. $p = 0.10$, $\delta = 0.05$

100 Variablen insgesamt pro Patient

$\rightarrow$ 59 der 100 Variablen überpüfen für 95%-Konf.-intervall

| | | |
|---|---|---|
| 500 Variablen insgesamt | $\rightarrow$ | 109 der 500 Variablen |
| 1 000 Variablen insgesamt | $\rightarrow$ | 122 der 1 000 Variablen |
| 2 000 Variablen insgesamt | $\rightarrow$ | 130 der 2 000 Variablen |
| 5 000 Variablen insgesamt | $\rightarrow$ | 135 der 5 000 Variablen |
| 50 000 Variablen insgesamt | $\rightarrow$ | 138 der 50 000 Variablen |
| 100 000 Variablen insg. | $\rightarrow$ | 139 der 100 000 Variablen |

**Umfang der SDV je Zentrum**

falls kleine Zentren existieren evtl. Verzicht auf Approximation durch Normalverteilung

$\rightarrow$ Anzahl der Zentren berücksichtigen

z.B. 500 Patienten im Zentrum, $p = 0.10$, $\delta = 0.05$

**109 statt 139 Patienten reichen aus.**

**Frequenz der SDV**

Die aufgrund der Fallzahlplanung erforderliche SDV sollte gleichmäßig auf den zur Verfügung stehenden Zeitraum aufgeteilt werden.

Beispiel: 6 Monate Zeit für SDV mit Fallzahl 139

SDV bei 30 Patienten pro Besuch möglich

$\rightarrow$ 5 SDV-Besuche gleichmäßig auf 6 Monate aufteilen.