Record Linkage von unterschiedlichen Datenarten:

Beispiele aus der epidemiologischen Forschung

Timm Intemann, Bianca Kollhorst, Stefan Rach, Wolfgang Ahrens und Iris Pigeot

68. GMDS-Jahrestagung

Workshop: Record Linkage von unterschiedlichen Datenarten

20. September 2023

Leibniz-Institut
für Präventionsforschung und
Epidemiologie – BIPS

nfdi 4 health

Motivation: Record Linkage bietet großartige Möglichkeiten

z. B. Krebsregister

ID	Var1	Var2
1		
2		
3		
4		

z. B. Versicherungs-daten

ID	Var3	Var4
1		
2		
3		
4		

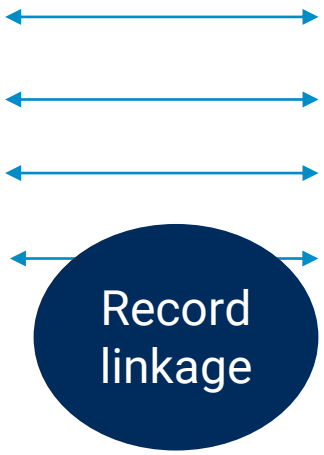
Mit einer Datenquelle:

- Verteilung von *Var1*
- „Wie viele Krebsfälle?“

Motivation: Record Linkage bietet großartige Möglichkeiten

z. B. Krebsregister

ID	Var1	Var2
1		
2		
3		
4		



z. B. Versicherungs-daten

ID	Var3	Var4
1		
2		
3		
4		

Mit einer Datenquelle:

- Verteilung von *Var1*
- „Wie viele Krebsfälle?“

Motivation: Record Linkage bietet großartige Möglichkeiten

z. B. Krebsregister

ID	Var1	Var2
1		
2		
3		
4		

z. B. Versicherungs-daten

ID	Var3	Var4
1		
2		
3		
4		

Record linkage

ID	Var1	Var2	Var3	Var4
1				
2				
3				
4				

Mit einer Datenquelle:

- Verteilung von *Var1*
- „Wie viele Krebsfälle?“

Mit den verknüpften Daten:

- Zusammenhang zwischen *Var1* und *Var3*
- „Gibt es eine Assoziation zwischen Medikament und Krebs?“

Motivation: Record Linkage bietet großartige Möglichkeiten

"Anonymous data and statistical information ***have some value*** *especially for descriptive epidemiological purposes, but the* ***real value*** of register-based information systems depends on the possibility for record linkages."

(Sund et al. 2014)

Record Linkage mit unterschiedlichen Datenarten: Beispiele aus der epidemiologischen Forschung

Ohlmeier et al. (2015) Verknüpfung von Routinedaten der Gesetzlichen Krankenversicherung mit Daten eines Krankenhausinformationssystems: Machbar, aber auch „nützlich“? doi:10.1055/s-0034-1395644
Dreger et al. (2020) Cohort study of occupational cosmic radiation dose and cancer mortality in German aircrew, 1960–2014. doi:10.1136/oemed-2019-106165
Riedel et al. (2023) Quality of life in bariatric patients up to twelve years after surgery - Results from a nationwide retrospective cohort study. doi:10.1016/j.orcp.2023.08.001
Langner et al. (2019) Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. doi:10.1136/bmjopen-2018-026834
Kollhorst et al. (2022) Record linkage of claims and cancer registries data — evaluation of a deterministic linkage approach based on indirect personal identifiers. doi:10.1002/pds.5545

Record Linkage mit unterschiedlichen Datenarten: Beispiele aus der epidemiologischen Forschung

- Krankenkassendaten ⇿ Daten aus Krankenhausinformationssystem (Ohlmeier et al. 2015)
  - Machbarkeitsanalyse eines indirekten Linkage-Ansatzes

Ohlmeier et al. (2015) Verknüpfung von Routinedaten der Gesetzlichen Krankenversicherung mit Daten eines Krankenhausinformationssystems: Machbar, aber auch „nützlich“? doi:10.1055/s-0034-1395644
Dreger et al. (2020) Cohort study of occupational cosmic radiation dose and cancer mortality in German aircrew, 1960–2014. doi:10.1136/oemed-2019-106165
Riedel et al. (2023) Quality of life in bariatric patients up to twelve years after surgery - Results from a nationwide retrospective cohort study. doi:10.1016/j.orcp.2023.08.001
Langner et al. (2019) Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. doi:10.1136/bmjopen-2018-026834
Kollhorst et al. (2022) Record linkage of claims and cancer registries data — evaluation of a deterministic linkage approach based on indirect personal identifiers. doi:10.1002/pds.5545

Record Linkage mit unterschiedlichen Datenarten: Beispiele aus der epidemiologischen Forschung

- Krankenkassendaten ⇿ Daten aus Krankenhausinformationssystem (Ohlmeier et al. 2015)
  - Machbarkeitsanalyse eines indirekten Linkage-Ansatzes
- Strahlenschutzregister ⇿ Flugpersonaldatenbanken (Dreger et al. 2020)
  - Untersuchung der Krebssterblichkeit beim Flugpersonal

Ohlmeier et al. (2015) Verknüpfung von Routinedaten der Gesetzlichen Krankenversicherung mit Daten eines Krankenhausinformationssystems: Machbar, aber auch „nützlich“? doi:10.1055/s-0034-1395644
Dreger et al. (2020) Cohort study of occupational cosmic radiation dose and cancer mortality in German aircrew, 1960–2014. doi:10.1136/oemed-2019-106165
Riedel et al. (2023) Quality of life in bariatric patients up to twelve years after surgery - Results from a nationwide retrospective cohort study. doi:10.1016/j.orcp.2023.08.001
Langner et al. (2019) Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. doi:10.1136/bmjopen-2018-026834
Kollhorst et al. (2022) Record linkage of claims and cancer registries data — evaluation of a deterministic linkage approach based on indirect personal identifiers. doi:10.1002/pds.5545

Record Linkage mit unterschiedlichen Datenarten: Beispiele aus der epidemiologischen Forschung

- Krankenkassendaten ⟷ Daten aus Krankenhausinformationssystem (Ohlmeier et al. 2015)
  - Machbarkeitsanalyse eines indirekten Linkage-Ansatzes
- Strahlenschutzregister ⟷ Flugpersonaldatenbanken (Dreger et al. 2020)
  - Untersuchung der Krebssterblichkeit beim Flugpersonal
- Krankenkassendaten ⟷ Epidemiologische Primärdaten
  - z. B. Untersuchung der Lebensqualität von Patient:innen nach bariatrischen Eingriffen (Riedel et al. 2023)
  - CoVerlauf und NAKO

Ohlmeier et al. (2015) Verknüpfung von Routinedaten der Gesetzlichen Krankenversicherung mit Daten eines Krankenhausinformationssystems: Machbar, aber auch „nützlich“? doi:10.1055/s-0034-1395644
Dreger et al. (2020) Cohort study of occupational cosmic radiation dose and cancer mortality in German aircrew, 1960–2014. doi:10.1136/oemed-2019-106165
Riedel et al. (2023) Quality of life in bariatric patients up to twelve years after surgery - Results from a nationwide retrospective cohort study. doi:10.1016/j.orcp.2023.08.001
Langner et al. (2019) Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. doi:10.1136/bmjopen-2018-026834
Kollhorst et al. (2022) Record linkage of claims and cancer registries data — evaluation of a deterministic linkage approach based on indirect personal identifiers. doi:10.1002/pds.5545

Record Linkage mit unterschiedlichen Datenarten: Beispiele aus der epidemiologischen Forschung

- Krankenkassendaten ⇿ Daten aus Krankenhausinformationssystem (Ohlmeier et al. 2015)
  - Machbarkeitsanalyse eines indirekten Linkage-Ansatzes
- Strahlenschutzregister ⇿ Flugpersonaldatenbanken (Dreger et al. 2020)
  - Untersuchung der Krebssterblichkeit beim Flugpersonal
- Krankenkassendaten ⇿ Epidemiologische Primärdaten
  - z. B. Untersuchung der Lebensqualität von Patient:innen nach bariatrischen Eingriffen (Riedel et al. 2023)
  - CoVerlauf und NAKO
- Krankenkassendaten ⇿ Krebsregisterdaten
  - Validierung der Angaben zur Todesursache in Krankenkassendaten (Langner et al. 2019)
  - DFG-Linkage (Kollhorst et al. 2022)

Ohlmeier et al. (2015) Verknüpfung von Routinedaten der Gesetzlichen Krankenversicherung mit Daten eines Krankenhausinformationssystems: Machbar, aber auch „nützlich“? doi:10.1055/s-0034-1395644
Dreger et al. (2020) Cohort study of occupational cosmic radiation dose and cancer mortality in German aircrew, 1960–2014. doi:10.1136/oemed-2019-106165
Riedel et al. (2023) Quality of life in bariatric patients up to twelve years after surgery - Results from a nationwide retrospective cohort study. doi:10.1016/j.orcp.2023.08.001
Langner et al. (2019) Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. doi:10.1136/bmjopen-2018-026834
Kollhorst et al. (2022) Record linkage of claims and cancer registries data — evaluation of a deterministic linkage approach based on indirect personal identifiers. doi:10.1002/pds.5545

Krankenkassendaten

- 96 gesetzliche Krankenkassen (GKV) (Mai 2022)
- 73 Mio. GKV-versichert (87% der Bevölkerung)
- > 46 private Krankenkassen

- Individuelle Daten zu Abrechnungszwecken
  - Demografische Daten
  - Arzneimittelverordnungen, ambulante/stationäre Leistungen, Diagnosen

- Vorteile gegenüber Befragungsdaten:
  - Objektiv
  - Kein Aufwand für Studienteilnehmer:innen
  - Keine Erinnerungs- oder Wissenslücken

Fallbeispiel 1:
Krankenkassendaten ⟷ Epidemiologische Primärdaten

CoVerlauf: Hintergrund und Ziele

- Studie zum Erkrankungsverlauf bei Personen mit einer COVID-19-Erkrankung bzw. einem positiven Test auf SARS-CoV-2
- Häufigkeit schwerer Erkrankungsverläufe
- Determinanten schwerer Erkrankungsverläufe
- Demonstration: Krankenkassendaten zur Untersuchung von Erkrankungsverläufen nützlich?

CoVerlauf: epidemiologische Primärdaten

- Kontaktaufnahme über Gesundheitsamt Bremen
- Einladungsschreiben an 12.995 Personen
- 1908 Personen in Studie eingewilligt (Responsequote: 15%)
- Befragungsdaten u.a. zu
  - Soziodemografie
  - Vorerkrankungen und Medikamenten
  - Lebensstil

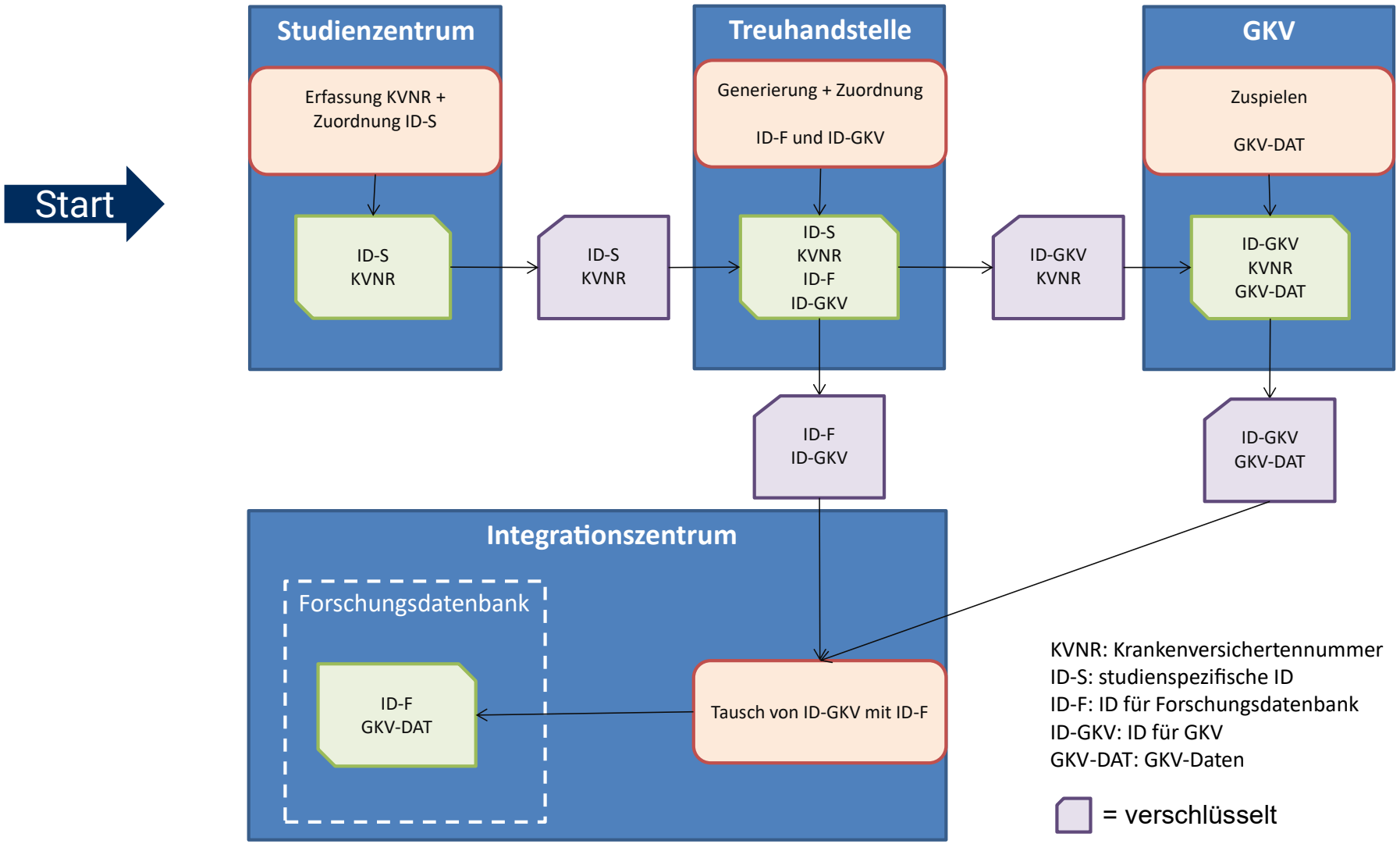
➢1617 (85%) Personen in Verknüpfung mit Krankenkassendaten eingewilligt und Krankenversichertennummer (KVNR) übermittelt

CoVerlauf: Einwilligungsquoten für Record Linkage zwischen epidemiologischen Primärdaten und Krankenkassendaten

Studie	Erhebungszeitraum	Stichprobenumfang	Einwilligungs-/ Einverständnisquote
KORA	1997-1999	796	64%
	2005	313	78%
Heinz-Nixdorf-Recall-Studie	Bis 2010	4.814	90%
AGil	Bis 2011	361	100%*
lida	2011	6.265	55%
	2014	4.244	63%
SHIP	Ab 2008	>5000	95%
NAKO	Ab 2014	~200.000	94%
LIFE (Follow-Up)	Ab 2017	5.665	88%
CoVerlauf	2021	1.908	**85%**

Intemann et al. (2023) White Paper: Verbesserung des Record Linkage für die Gesundheitsforschung in Deutschland. doi:10.4126/FRL01-006461895

CoVerlauf: Datenfluss analog zur NAKO Gesundheitsstudie



Intemann et al. (2023) White Paper: Verbesserung des Record Linkage für die Gesundheitsforschung in Deutschland. doi:10.4126/FRL01-006461895

CoVerlauf: Herausforderungen und Fazit

- Einholung der Einwilligungserklärungen über das Gesundheitsamt → Responseverluste
- Einwilligungsbereitschaft zum RL in Bevölkerung sehr hoch
- Organisatorischer Aufwand für die Kooperation mit 101 Krankenkassen
  - Einschluss: Kassen mit mindestens 20 Proband:innen:
    - ➢ Reduktion des Stichprobenumfangs auf 1352 (71%) und 14 Kassen
- Zweigliedriges Krankenversicherungssystem in Deutschland
  - GKV- und PKV-Daten fallen unter verschiedene Datenschutzregeln
  - Ausschluss der PKV
    - ➢ Reduktion des Stichprobenumfangs auf 1325 (69%)

➢ Unsicherere und möglicherweise verzerrte Ergebnisse aufgrund verringerter Fallzahlen und Selektionseffekten

Fallbeispiel 2:
Krankenkassendaten ⟷ Krebsregisterdaten

DFG-Linkage

DFG-Linkage: Hintergrund und Ziele

- Ziel: Evaluierung eines Linkage-Ansatz basierend auf **indirekten** Identifikatoren im Vergleich zu einem Ansatz mit **direkten** Identifikatoren
- Anhand **Beispielstudie** zum Risiko einer Krebsneuerkrankung und der Krebsmortalität bei Patient:innen mit Typ-2-Diabetes unter Behandlung mit verschiedenen Antidiabetika
- Antidiabetika: Sulfonylharnstoffe und Metformin
- Für zwei Krebsentitäten: Schilddrüsen- und Darmkrebs

Pigeot et al (2021) Verknüpfung von Abrechnungsdaten gesetzlicher Krankenkassen mit Daten epidemiologischer Krebsregister: länderspezifische Möglichkeiten und Limitationen. doi:10.1007/s00103-021-03475-x
Pigeot et al (2021) Nutzung von Sekundärdaten für die pharmakoepidemiologische Forschung - machen wir das Beste draus! doi:10.1055/a-1633-3827

DFG-Linkage: Datenquellen

- Vier epidemiologische Krebsregister
  - Tumorstadium
  - Tumordiagnose
- Pharmakoepidemiologische Forschungsdatenbank (GePaRD): Abrechnungsdaten von ~25 Mio. Versicherten (TK, DAK, hkk, AOK Bremen)
  - Diabetesmedikation

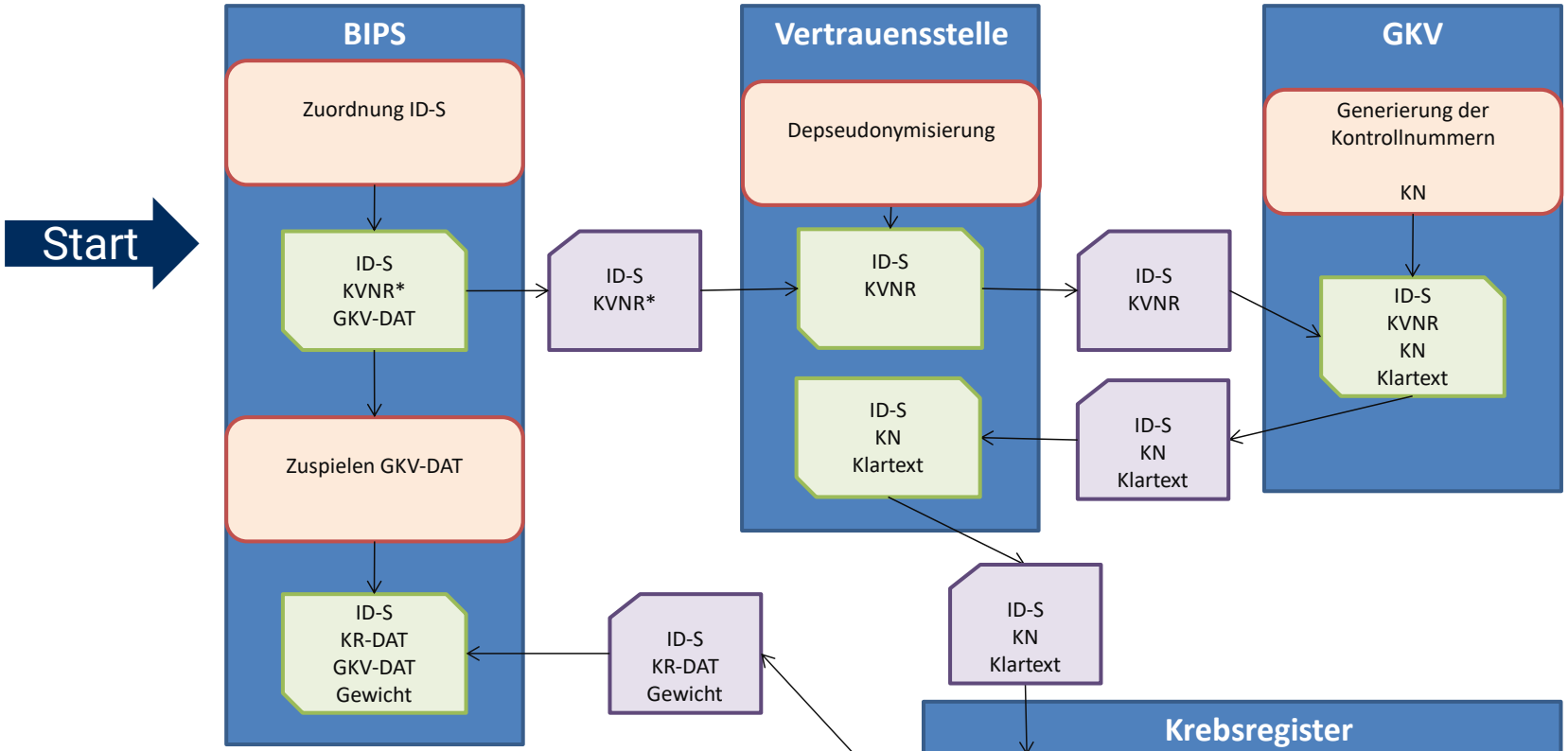
AOK Bremen/Bremerhaven

DAK-Gesundheit

hkk Krankenkasse

Techniker Krankenkasse (TK)

DFG-Linkage: Datenfluss



KVNR: Krankenversichertennummer (*= pseudonymisiert)
ID-S: studienspezifische ID
GKV-DAT: GKV-Daten
KN: Kontrollnummern
Klartext: Gemeindekennziffer, Geschlecht, Geburtsjahr/-monat
KR-DAT: Krebsregisterdaten
Gewicht: Übereinstimmungsgewichte
= verschlüsselt

DFG-Linkage: Identifier und Record Linkage-Methoden

1. Probabilistisches Linkage basierend auf *direkten* Identifikatoren

- Etabliertes Kontrollnummernlinkage der Kresbregister
- Kontrollnummern abgeleitet aus u.a. *Vorname*, *Nachname*, *Tag der Geburt*…

DFG-Linkage: Identifier und Record Linkage-Methoden

1. Probabilistisches Linkage basierend auf *direkten* Identifikatoren
   - Etabliertes Kontrollnummernlinkage der Kresbregister
   - Kontrollnummern abgeleitet aus u.a. *Vorname*, *Nachname*, *Tag der Geburt*…
2. Deterministisches Linkage basierend auf *indirekten* Identifikatoren
   - Identifier:
     - *Geburtsjahr*,
     - *Geschlecht*,
     - *Gemeindekennziffer*,
     - *Krebsart und*
     - *Diagnosedatum*
   - Falls keine eindeutige Zuordnung: *Genauigkeit der ICD-Kodierung*, *stationäre Diagnose*, *Differenz des Diagnosedatums*

DFG-Linkage: Vergleich beider Linkage-Ansätze

- Linkage basierend auf indirekten Identifiern datenschutzrechtlich weniger aufwändig, da keine Vertrauensstelle notwendig

DFG-Linkage: Vergleich beider Linkage-Ansätze

- Linkage basierend auf indirekten Identifiern datenschutzrechtlich weniger aufwändig, da keine Vertrauensstelle notwendig
- Kontrollnummernlinkage als Goldstandard zur Evaluation des indirekten Ansatzes
- Sensitivität des indirekten Linkage:
  - Darmkrebs: 72%
  - Schilddrüsenkrebs: 67%

DFG-Linkage: Vergleich beider Linkage-Ansätze

- Linkage basierend auf indirekten Identifiern datenschutzrechtlich weniger aufwändig, da keine Vertrauensstelle notwendig
- Kontrollnummernlinkage als Goldstandard zur Evaluation des indirekten Ansatzes
- Sensitivität des indirekten Linkage:
  - Darmkrebs: 72%
  - Schilddrüsenkrebs: 67%
- Ereigniszeitanalyse für Darmkrebsrisiko mit beiden Ansätzen:

Record Linkage-Methode	Hazard Ratio (Sulfonylharnsäure vs. Metformin)
Prob. mit direkten Identifiern	0.71
Determ. mit indirekten Identifiern	0.81

- Unterschied zwischen Diabetesmedikation um 10%-Punkte unterschätzt

Kollhorst et al. (2022) Record linkage of claims and cancer registries data — evaluation of a deterministic linkage approach based on indirect personal identifiers. doi:10.1002/pds.5545

DFG-Linkage: Vergleich beider Linkage-Ansätze

- Linkage basierend auf indirekten Identifiern datenschutzrechtlich weniger aufwändig, da keine Vertrauensstelle notwendig
- Kontrollnummernlinkage als Goldstandard zur Evaluation des indirekten Ansatzes
- Sensitivität des indirekten Linkage:
  - Darmkrebs: 72%
  - Schilddrüsenkrebs: 67%
- Ereigniszeitanalyse für Darmkrebsrisiko mit beiden Ansätzen:

Record Linkage-Methode	Hazard Ratio (Sulfonylharnsäure vs. Metformin)
Prob. mit direkten Identifiern	0.71
Determ. mit indirekten Identifiern	0.81

- Unterschied zwischen Diabetesmedikation um 10%-Punkte unterschätzt

➢ Deswegen Linkage mit diesen indirekten Identifikatoren nicht zu empfehlen

DFG-Linkage: Herausforderungen

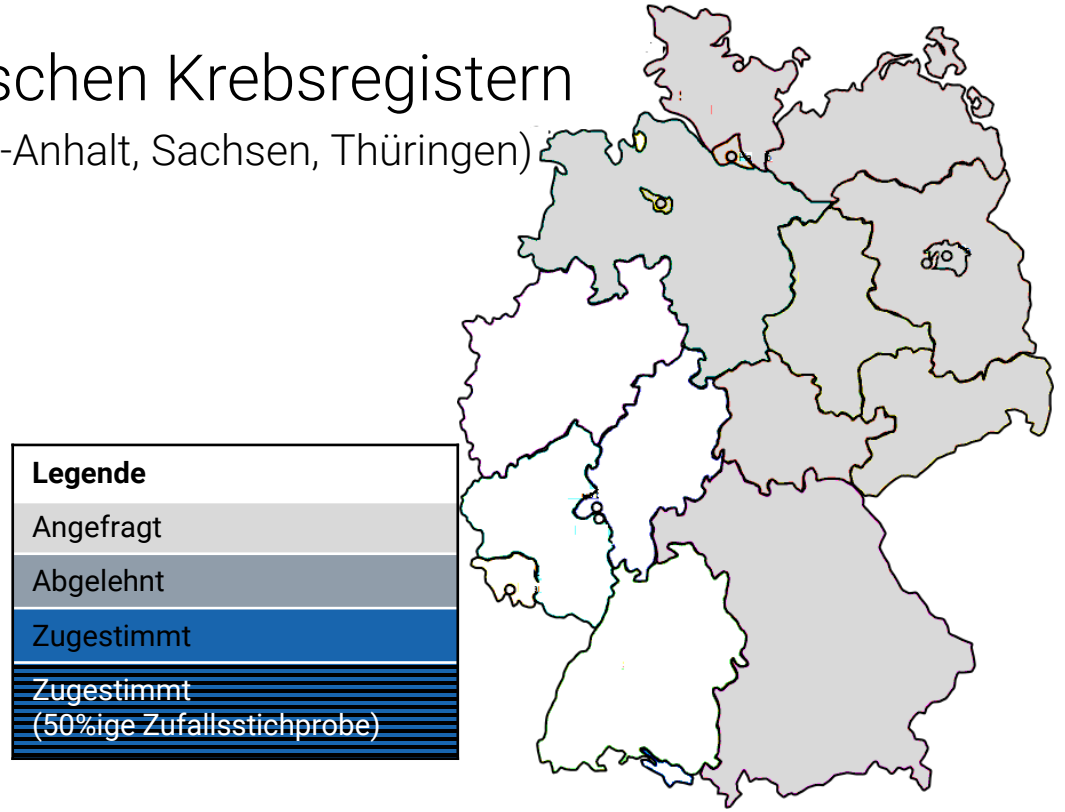
- Problem: keine Verlinkung über Zentrum für Krebsregisterdaten (ZfKD) aus rechtlichen Gründen möglich

DFG-Linkage: Herausforderungen

- **Problem:** keine Verlinkung über Zentrum für Krebsregisterdaten (ZfKD) aus rechtlichen Gründen möglich

➢ **Konsequenz:** Anfrage bei sechs epidemiologischen Krebsregistern

- ➢ GKR (Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen-Anhalt, Sachsen, Thüringen)
- ➢ Bayern
- ➢ Bremen
- ➢ Hamburg
- ➢ Niedersachsen
- ➢ Schleswig-Holstein



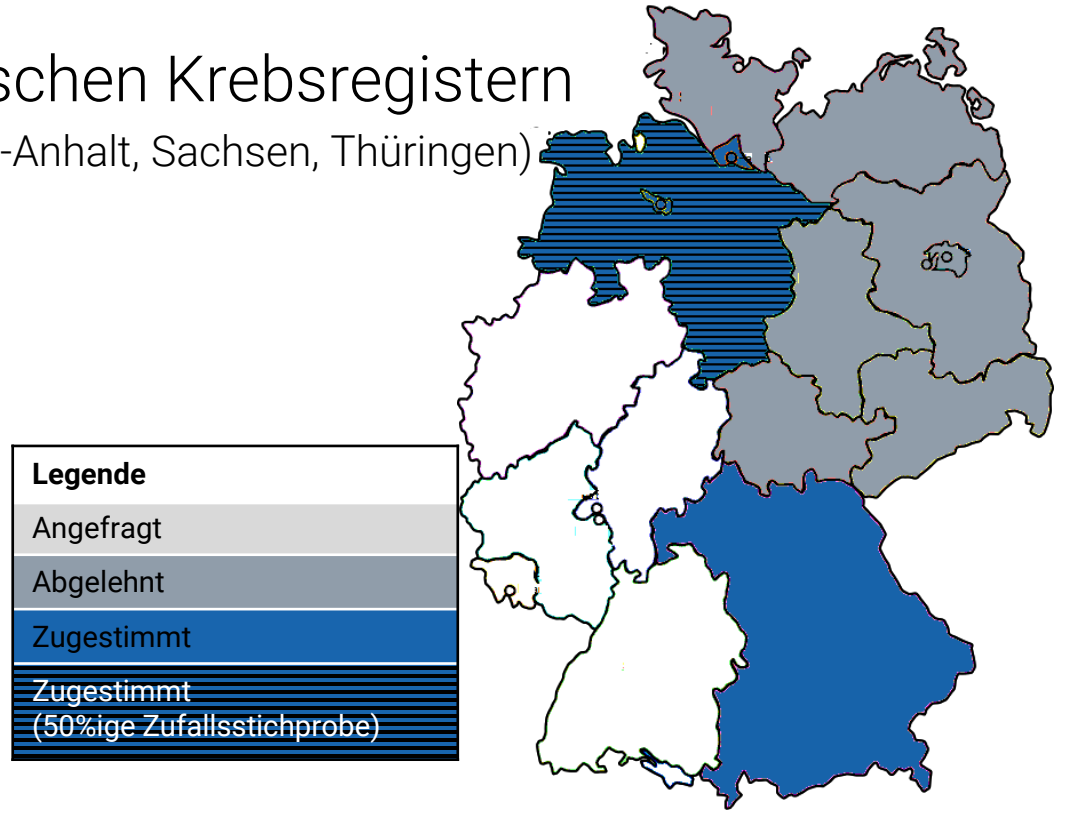
DFG-Linkage: Herausforderungen

- Problem: keine Verlinkung über Zentrum für Krebsregisterdaten (ZfKD) aus rechtlichen Gründen möglich

➢ Konsequenz: Anfrage bei sechs epidemiologischen Krebsregistern

  - ➢ GKR (Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen-Anhalt, Sachsen, Thüringen)
  - ➢ Bayern
  - ➢ Bremen
  - ➢ Hamburg
  - ➢ Niedersachsen
  - ➢ Schleswig-Holstein

- Unterschiede:
  - Art der Antragstellung
  - Antragsformulare
  - Zuständige Stelle
  - Geäußerte datenschutzrelevante Bedenken
  - Entscheidung nach Bundesland verschieden



DFG-Linkage: Herausforderungen und Fazit

- Kontrollnummernlinkage für Krebsregistrierung entwickelt, dafür nötige Merkmale nicht in gleicher Form in Versicherungsdaten enthalten
- Record Linkage-Projekte mit Krebsregister- und Krankenkassendaten zeit- und kostenintensiv: weitere Anträge und Genehmigungen
  - Zustimmung der Kassen
  - Zustimmung der Aufsichtsbehörden nach §75 SGB X

➢ Flächendeckende bundesweite Analysen mit verlinkten Daten aus Krebsregistern und von Krankenkassen erscheinen *momentan* unrealistisch

DFG-Linkage: Ausblick

- Mittlerweile **bundeseinheitliches Antragsformular** der Krebsregister verfügbar
  - ➢Antragstellung wird vereinfacht

- Zukünftig **KVNR als zusätzliche Kontrollnummer in Krebsregistern**
  - ➢Linkage-Qualität des Kontrollnummernlinkage erhöht

- **Gesundheitsdatennutzungsgesetz** löst Linkage-Probleme voraussichtlich nur beschränkt
  - ➢Linkage zwischen FDZ Gesundheit und Daten der klinischen Krebsregister vorgesehen

Link zum *White Paper: Verbesserung des Record Linkage für die Gesundheitsforschung in Deutschland*

https://doi.org/10.4126/FRL01-006461895

Vielen Dank für die Aufmerksamkeit!

Kontakt:

Timm Intemann

intemann@leibniz-bips.de

Leibniz-Institut für Präventionsforschung und Epidemiologie - BIPS

www.nfdi4health.de

Leibniz-Institut
für Präventionsforschung und
Epidemiologie – BIPS

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

nfdi 4 health

Link zum *White Paper: Verbesserung des Record Linkage für die Gesundheitsforschung in Deutschland*

https://doi.org/10.4126/FRL01-006461895