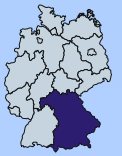
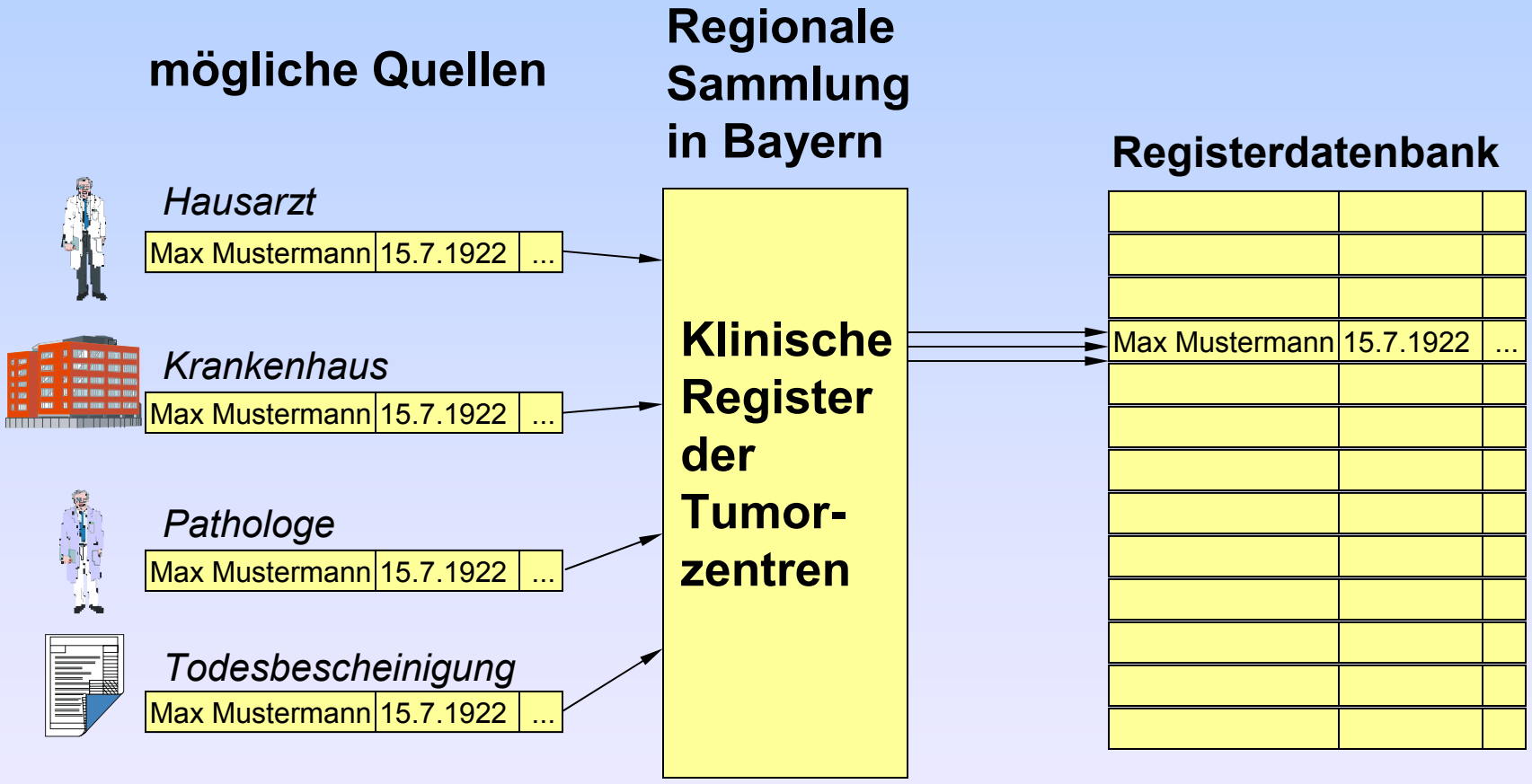


M. Meyer, M. Radespiel-Tröger, C. Vogel

Probabilistisches Record-Linkage mit anonymisierten Krebsregistermeldungen

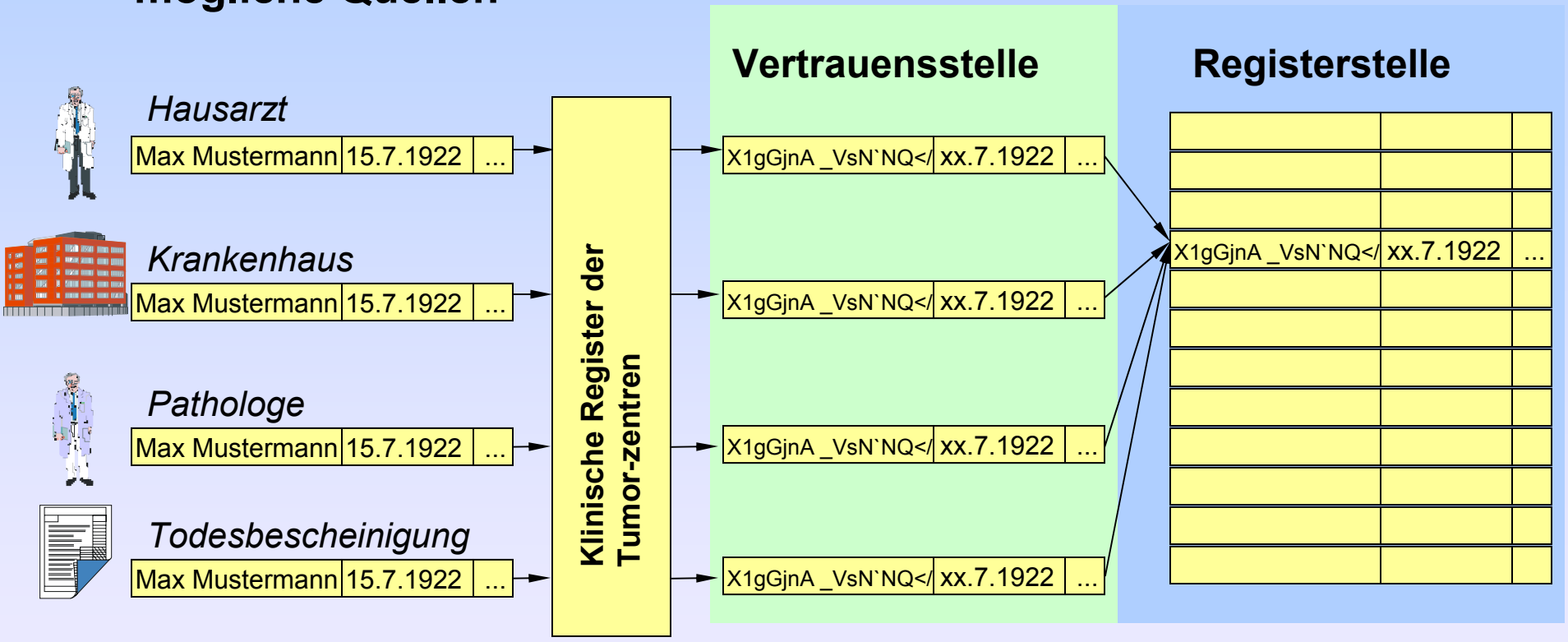
Eingehende Meldungen im Krebsregister (nicht anonymisiert)



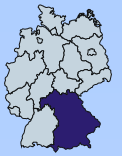
Meldungen treffen im Allgemeinen zu verschiedenen Zeiten ohne Berücksichtigung einer logischen Reihenfolge (auch bei Zweittumoren) ein.

Eingehende Meldungen im Krebsregister (anonymisiert)

mögliche Quellen



Durch die Anonymisierung kann keine Person mehr namentlich identifiziert werden, aber es bleibt für die wissenschaftliche Auswertung möglich zu unterscheiden, ob sich zwei Meldungen auf dieselbe oder auf verschiedene Personen beziehen.



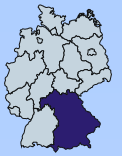
Aufgabe des Record-Linkage

- **Finden und Zusammenführen aller zu einer Person gehörenden Datensätze**

- **Fehlertoleranz:**

  Zusammenführung auch bei nicht völlig identischen Identifikationsmerkmalen

  z.B.
  - **Schreibfehler**
  - **Wohnortwechsel**
  - **Namenswechsel durch Heirat**
  - **Vertauschung von Geburtsmonat und -tag**



Mögliche Fehler beim Record-Linkage

Homonymfehler

Datensätze werden zusammengeführt, obwohl sie zu verschiedenen Personen gehören

Ursache: Zufällige Merkmalsidentität oder zu geringe Trennschärfe des Linkageverfahrens
z.B. häufiger Name und Großstadt als Wohnort

Synonymfehler

Datensätze werden nicht zusammengeführt, obwohl sie zu einer Person gehören

Ursache: Schreibfehler oder Änderungen in zu vielen Identifikationsmerkmalen

Linkage - Algorithmus (1)

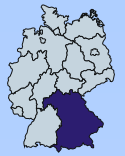
- **Quelle**

  Ivan P. Fellegi, Alan B. Sunter:

  A theory for record linkage.

  American Statistical Association Journal, December 1969; 40:1183-220

- **Grundlage für die meisten probabilistischen Record-Linkage-Anwendungen (auch kommerzieller Programme)**



Linkage - Algorithmus (2)

Ausgangssituation

Mengen A und B von Meldungen

Alle möglichen Paare:

$$A \times B = \{ (a,b);\ a \in A,\ b \in B\}$$

Gesucht werden Paare von Meldungen, die zur selben Person gehören („matched“):

$$M = \{ (a,b);\ a = b,\ a \in A,\ b \in B\}$$

Alle anderen Paare gehören zur Menge der nicht zusammengehörenden („unmatched“):

$$U = \{ (a,b);\ a \neq b,\ a \in A,\ b \in B\}$$

Anwendung beim Online-Record-Linkage im Krebsregister

A = {neu eingetroffene Meldung}
B = {alle schon vorhandenen Meldungen}

Linkage - Algorithmus (3)

Merkmalsvektor der Matchvariablen

$a = (a_1, ..., a_n) , \quad b = (b_1, ..., b_n) ,$

z.B. (Nachname, Vorname, Geburtsname, Geburtsdatum, Wohnort)

Linkage - Algorithmus (4)

Übereinstimmungswahrscheinlichkeiten m

$m_{ik} = P(a_i = b_i \ \wedge a_i = x_{ik} \mid (a,b) \in M)$

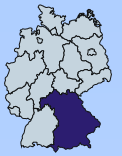
Wahrscheinlichkeit, dass zwei Meldungen a und b von der gleichen Person im i-ten Merkmal übereinstimmen und dieses Merkmal die Ausprägung x_{ik} besitzt

m_{ik} werden aus geprüftem Datenbestand ermittelt oder geschätzt, falls geschätzt, wird $m_{ik} = m_i$ für alle Merkmalsausprägungen angenommen

z.B. $m_{Geburtsdatum} = 0,99$

$m_{Vorname} = 0,975$

(1 - Wahrscheinlichkeit eines Schreibfehlers, Namensänderung, Umzugs, usw.)



Linkage - Algorithmus (5)

Übereinstimmungswahrscheinlichkeiten u

$u_{ik} = P(a_i = b_i \wedge a_i = x_{ik} \mid (a,b) \in U)$

Wahrscheinlichkeit, dass zwei Meldungen a und b von verschiedenen Personen im i-ten Merkmal übereinstimmen und Ausprägung x_{ik} haben

Auf Grund der wesentlich kleineren Mächtigkeit der Menge M im Vergleich zu U vereinfacht sich die Berechnung der u_{ik} : Diese können der Häufigkeitsverteilung für Merkmal i aus der Population entnommen werden (gegegebenfalls nach Anonymisierung der einzelnen Ausprägungen)

z.B.

$u_{\text{Wohnort, „München“}} = 0{,}168$

$u_{\text{Wohnort, „Oberstdorf“}} = 0{,}001$

$u_{\text{Nachname, „Müller“}} = 0{,}01121$

$u_{\text{Nachname, „Hotzenplotz“}} = 0{,}00003$

Linkage - Algorithmus (6)

Übereinstimmungsgewicht

$$w = \sum w_i$$

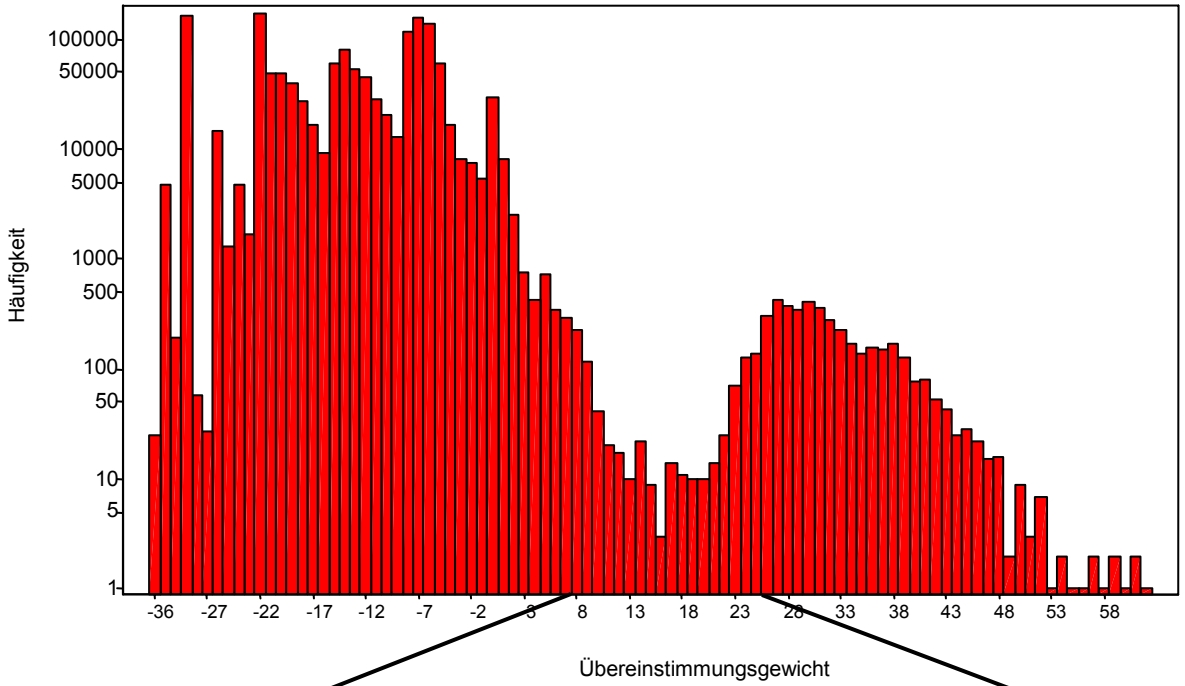
wobei

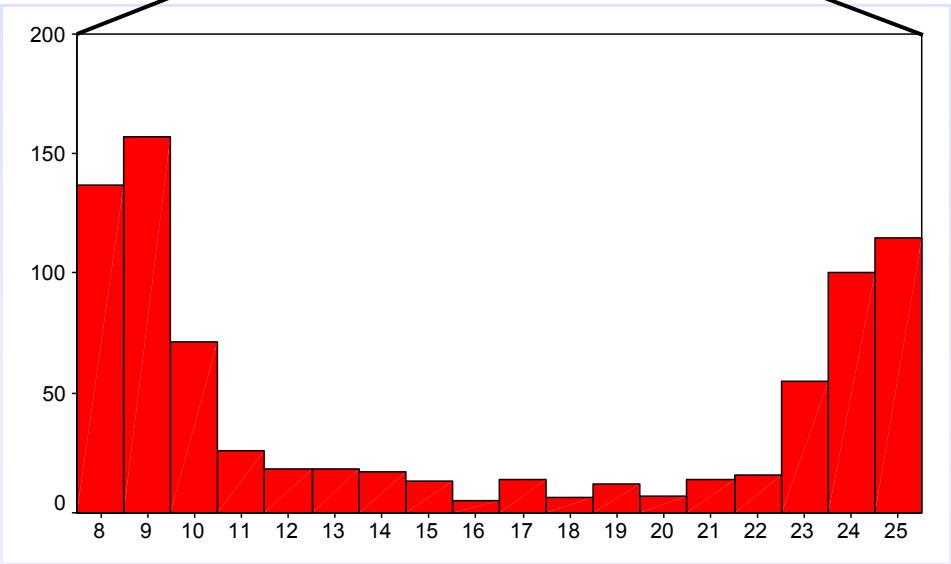
$$w_i = \log (m_i / u_{ik}) \quad , \text{falls } a_i = b_i \wedge a_i = x_{ik}$$

$$w_i = \log ((1-m_i) / (1-u_{ik})) \quad , \text{falls } a_i \neq b_i \wedge a_i = x_{ik}$$

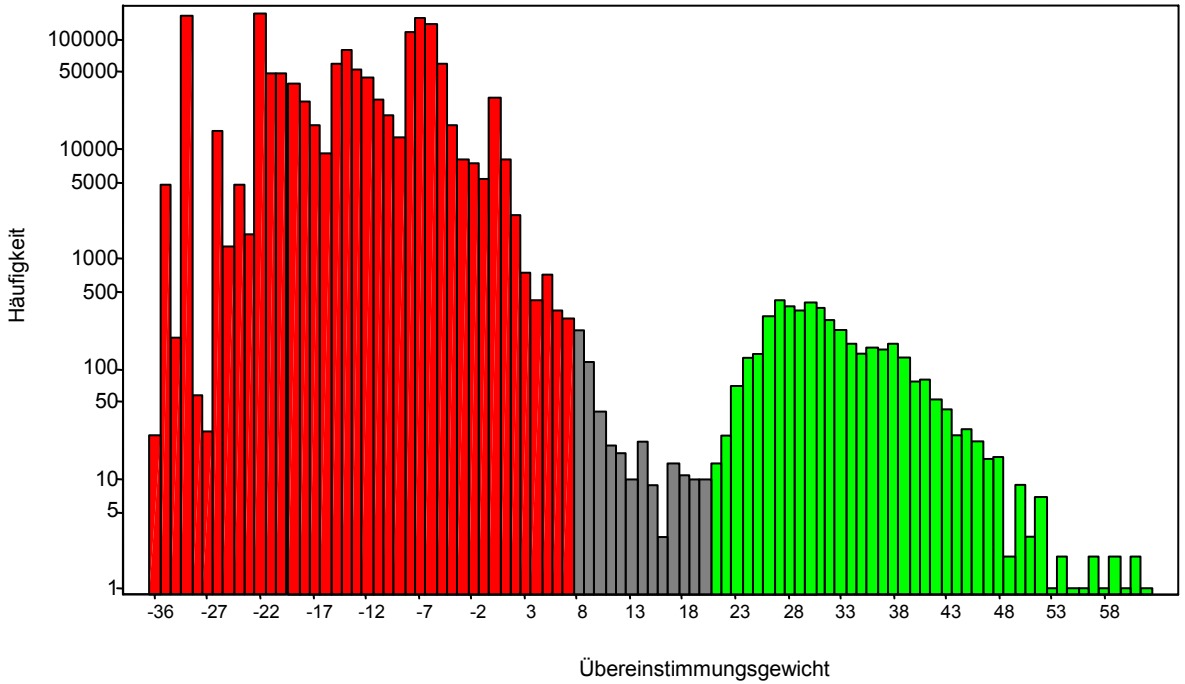
Die Gewichtsbeiträge w_i sind bei Übereinstimmung positiv, sonst negativ (Die Wahrscheinlichkeit, dass Übereinstimmung vorliegt, ist in der Menge M größer als die Wahrscheinlichkeit für zufällige Übereinstimmung).

Krebsregister: Übereinstimmungs-gewichte für alle in Frage kommenden Meldungspaare (1,4 Mio.)





Krebsregister: Übereinstimmungs-gewichte für alle in Frage kommenden Meldungspaare (1,4 Mio.)



positiver Nicht-Link | möglicher Link | positiver Link

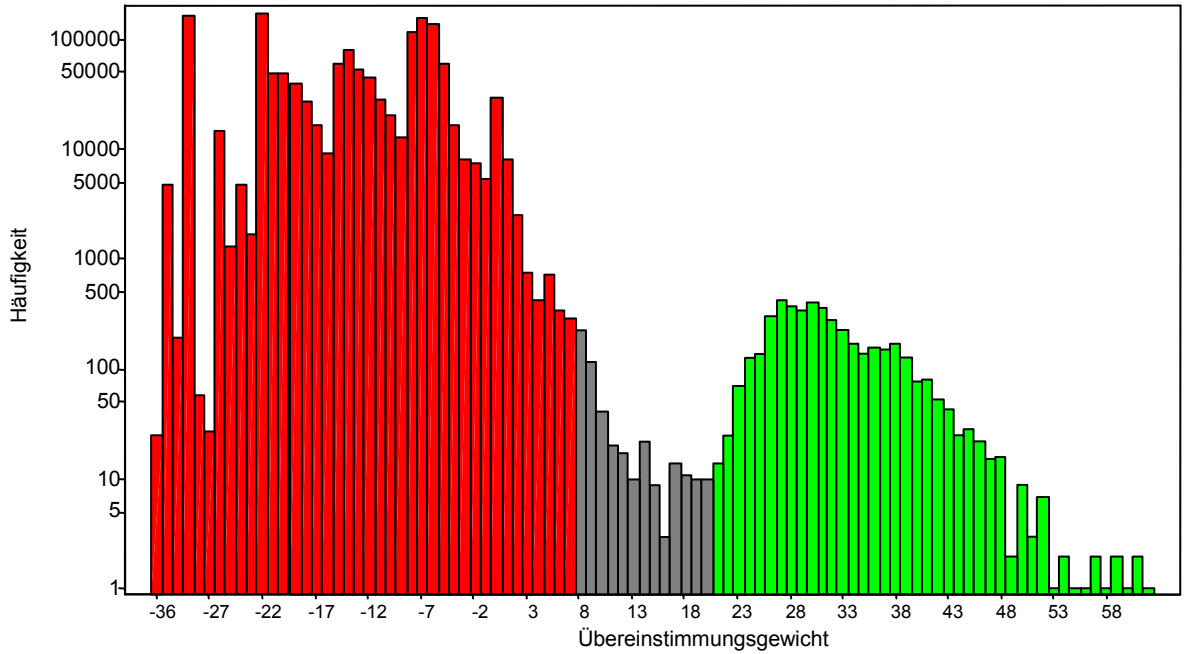
Unmatched | **manuelle Ent-schei-dung** | **Matched**

Automatische Weiterverarbeitung einzelner Meldungen

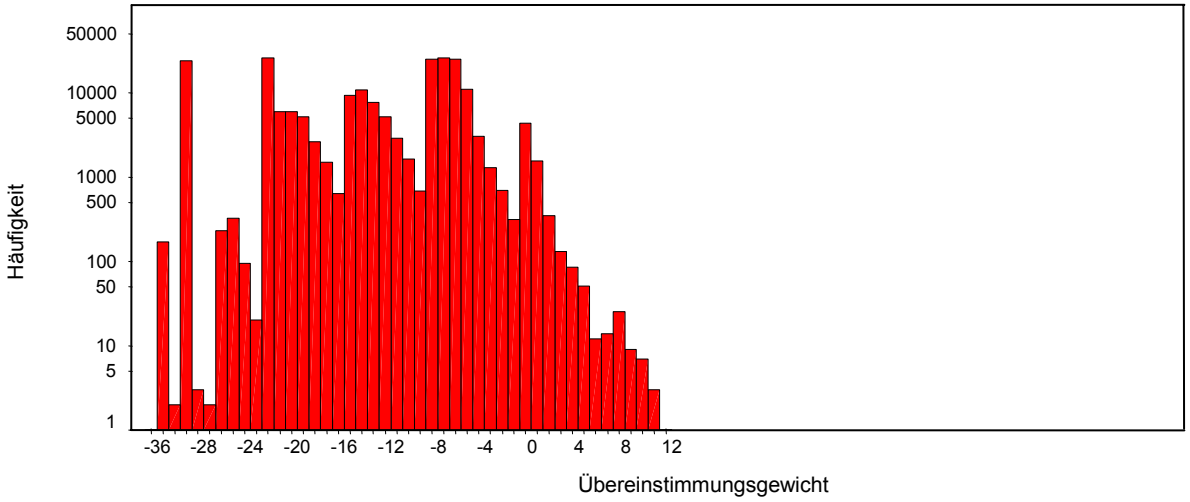
Interaktive Gewinnung der besten Information aus zusammengehörigen Meldungen
(automatisch generierte Vorschläge)

Validierung

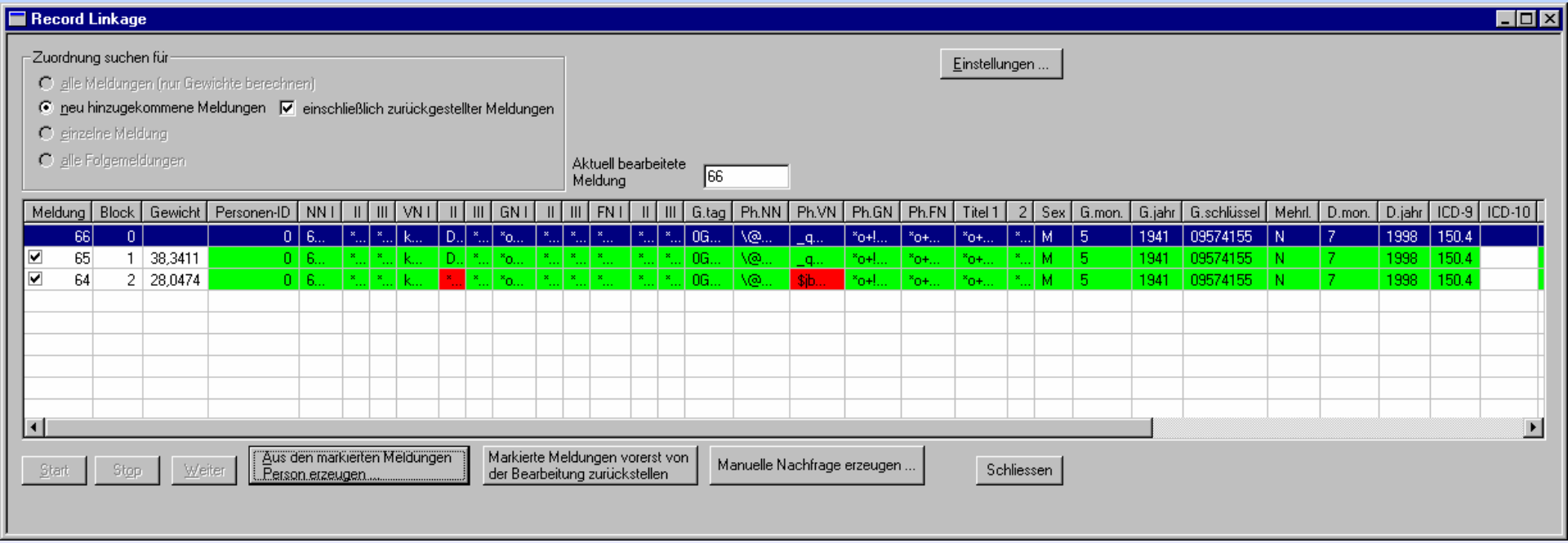
Alle Gewichte



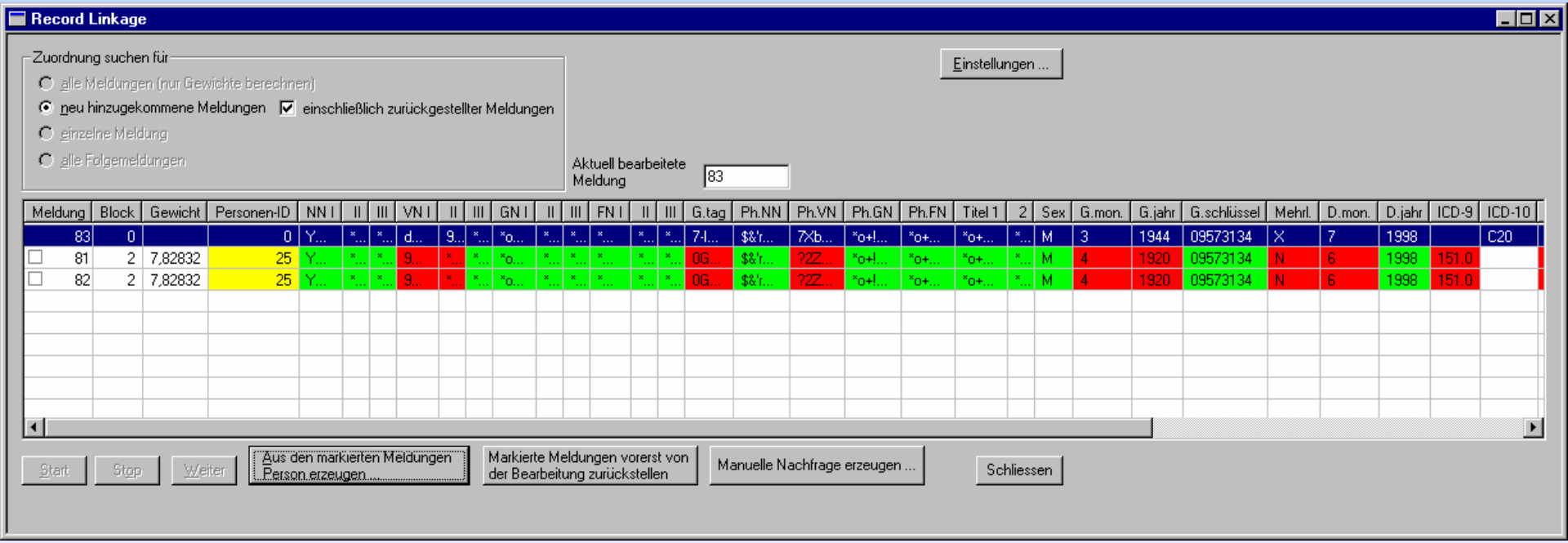
Gewichte für Meldungen aus einer einzigen Auswertung von Totenscheinen (hier kann es keine Mehrfachmeldungen geben)



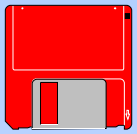
Interaktives Zusammenführen



Interaktives Zusammenführen (2)



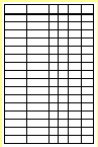
Population Based Cancer Registry Bavaria



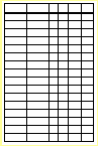
Incidence records from confidential office

External Dictionary

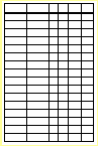
Data Dictionary



Validation rules



Linkage parameters



Import of data records

Validation

Record linkage (patients)

Record linkage (tumours)

Processes within the bavarian registration office

Automated	Interactive
Matching of imported to internal variables Formal validation (e.g. column format)	
Simple validation Cross-field validation Computer generated error reports	
Positive non-links Positive links	Possible links Cross-record validation
Computer aided suggestions Post-update validation	Final decision

M. Meyer, Registerstelle, 30.1.2003