Datenabgleich mit dem Deutschen Kinderkrebsregister

Erfahrungsbericht aus einer epidemiologischen Kohortenstudie

Dr. Gaël Hammer

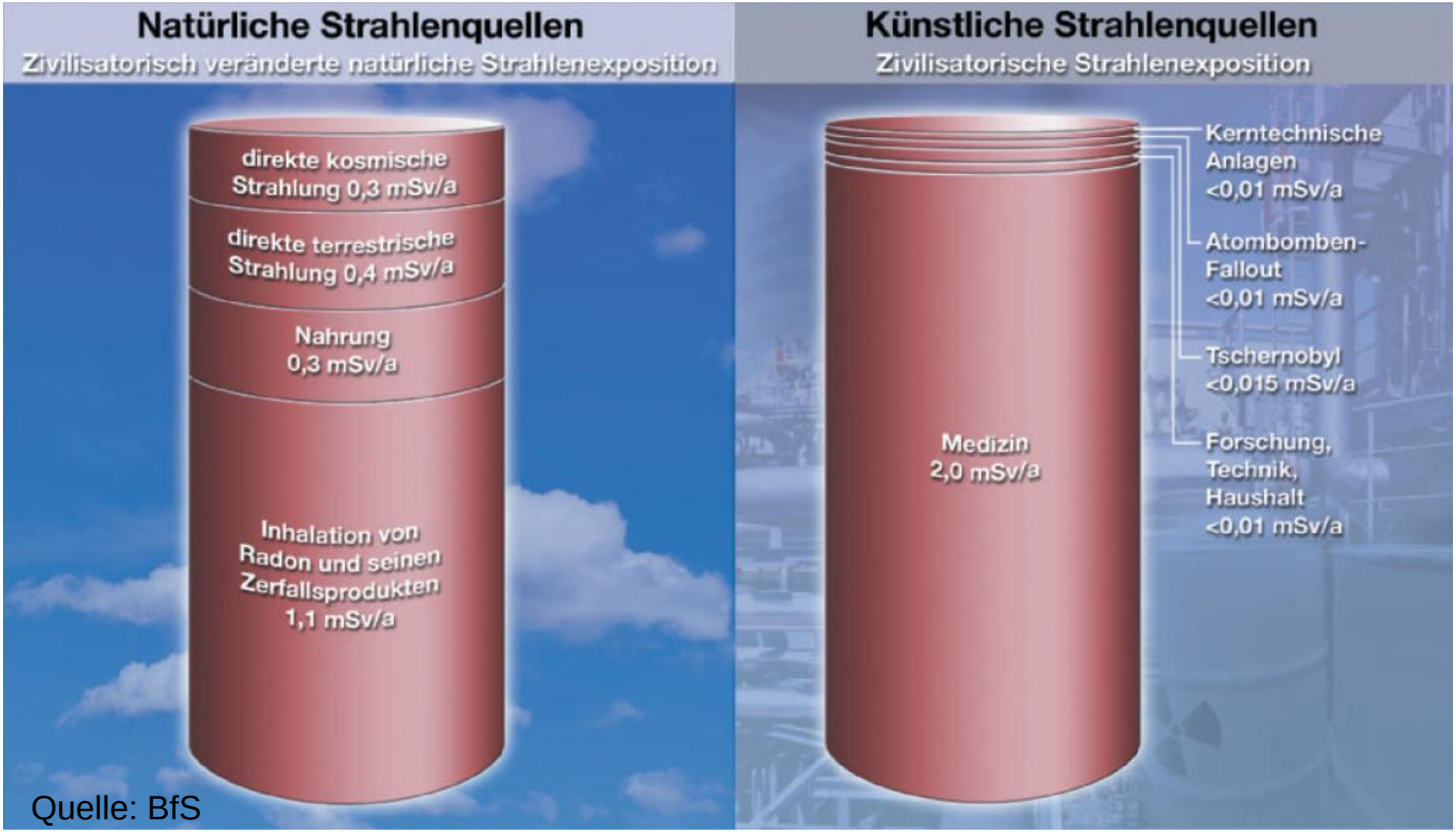
Gliederung

- Die RICC-Studie
- Epidemiologische Krebsregister in Deutschland
- Datenschutz in der RICC-Studie
- Garbage in …
- Stochastisches Record-Linkage
- … und was hinten 'rauskommt

Die RICC-Studie

RICC-Studie – Hintergrund

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik



RICC-Studie - Hintergrund

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

Quelle: Wikipedia: Kingofears

RICC-Studie - Material

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

- Röntgenverordnung 1973
- Datensammlung am Dr. von Haunerschen Kinderspital der LMU München
- Protokoll aller Röntgenaufnahmen seit 1976 in Datenbanken (MINDIUS I-III, ab 1998 RIS-System)
- Auswertungen
  1. Strahlendosen
  2. Strahlenrisiko

RICC-Studie – Kohorte

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

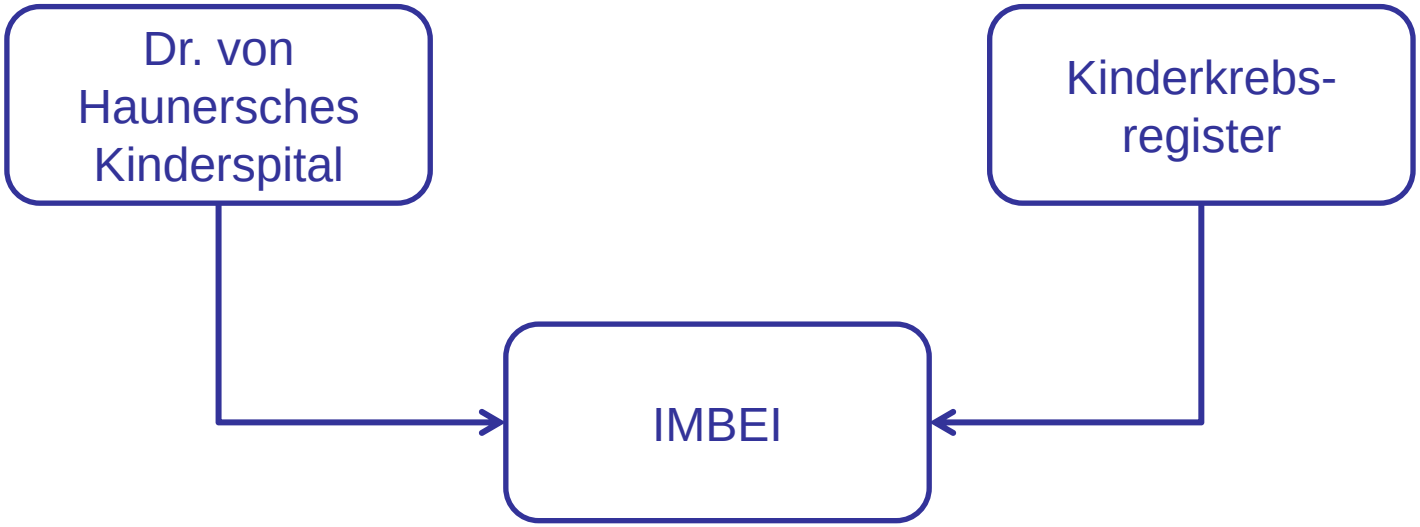
Einschlusskriterien

- Rekrutierungszeitraum 1976-2003
- krebsfrei zu Beginn (inkl. erste 6 Monate)
- Hauptwohnsitz in Deutschland
- Alter bei Einschluss ≤ 14,5 Jahre
- Am 1.1.1980 noch ≤ 15 Jahre alt

RICC-Studie – Methoden

Institut für Biometrie, Epidemiologie
und Medizinische Informatik

- Abgleich Pseudonymisierter Daten mit dem Gesamtbestand des deutschen Kinderkrebsregisters (erweiterte Registerbevölkerung)
- Beobachtungszeitraum: 1980-2006

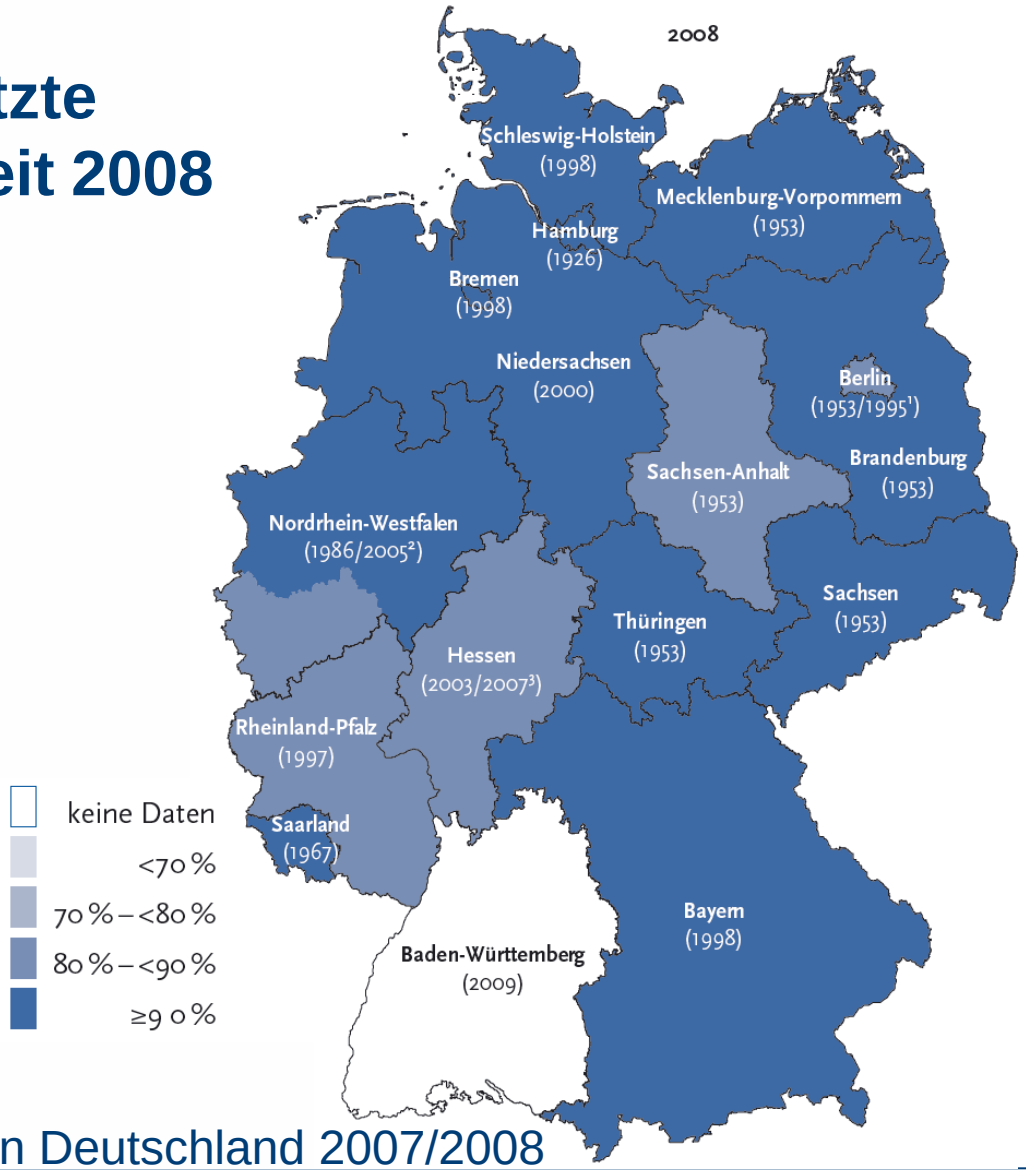


Epidemiologische Krebsregister in Deutschland

Epidemiol. Krebsregister in Deutschland

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik

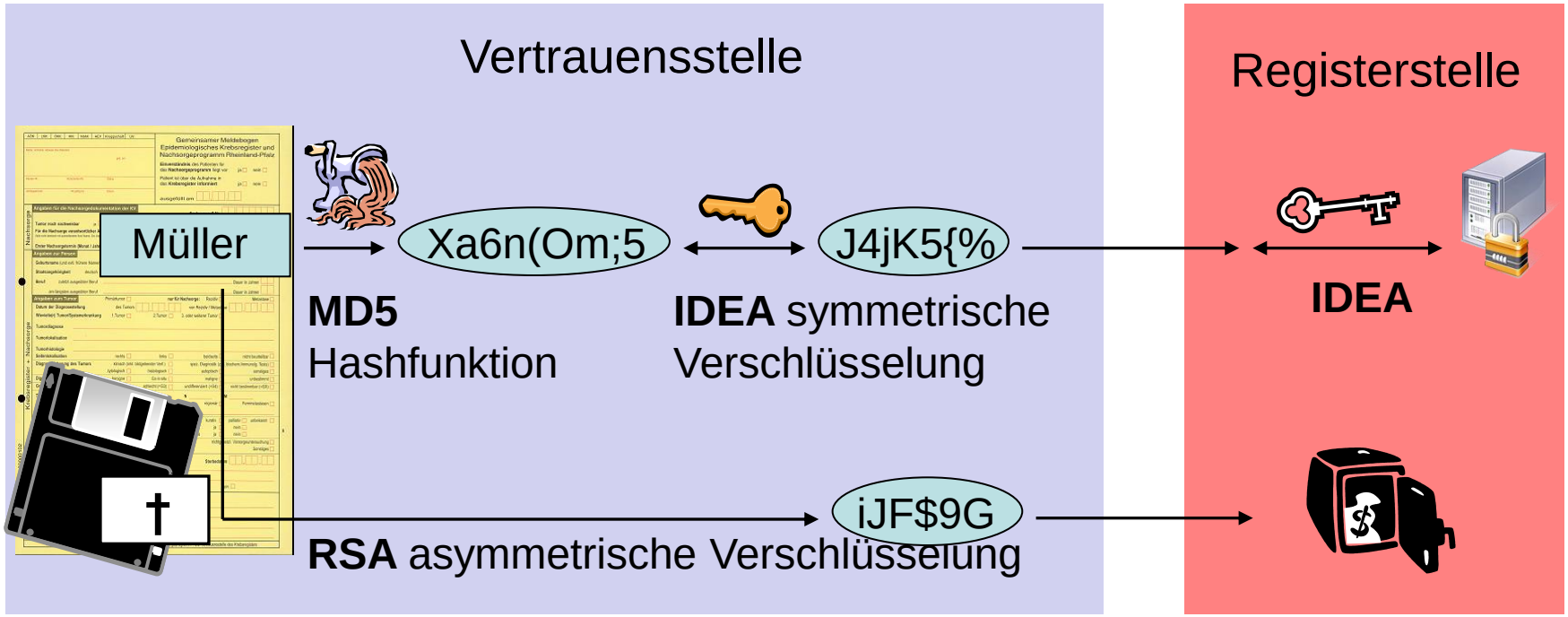
geschätzte Vollzähligkeit 2008



Quelle: RKI: Krebs in Deutschland 2007/2008

Epidemiol. Krebsregister Datenschutz

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik



Epidemiol. Krebsregister gespeicherte Daten

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik

Personenidentifizierende Merkmale

- Vorname(n)
- Nachname(n)
- frühere(r) Name(n)
- Geschlecht
- Straße
- PLZ
- Ort
- Geburtsdatum
- Todesdatum
- Diagnosedatum

Epidemiologische Daten

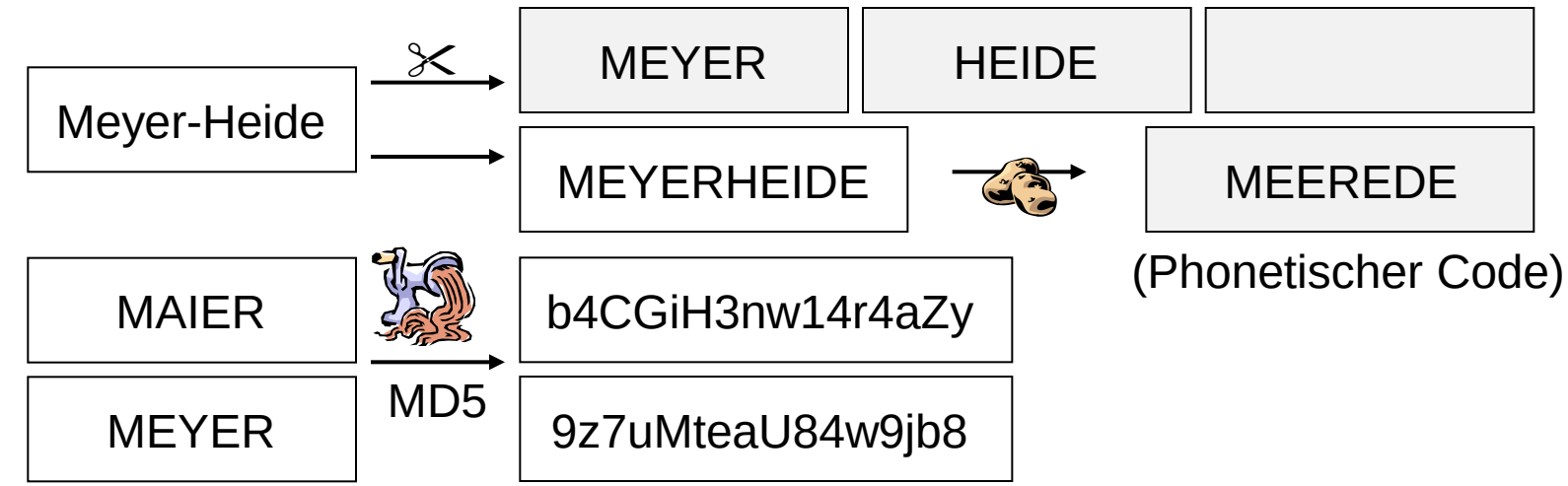
- Geschlecht
- Geburtsmonat und -jahr
- Gemeindekennziffer
- Nationalität
- Letzte und längste Beschäftigung(en)
- Tumor-Klassifikation (ICD-10)
- Topologie und Morphologie (ICD-O-3)
- Tumorlokalisation, Lateralität
- Diagnosemonat und -jahr
- Stadium (TNM)
- Basis der Diagnose
- Methode der Ersterkennung
- Erstbehandlung
- Todesmonat und -jahr
- Todesursachen
- Ersttumor ja/nein

Epidemiol. Krebsregister
Bildung von Kontrollnummern

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

1. Verarbeitung
   - Zerlegung in 3 Einzel-Bestandteile ✂
   - Phonetischer Code
2. Kodierung (MD5, IDEA)
   - Ergibt „KontrollNummern“

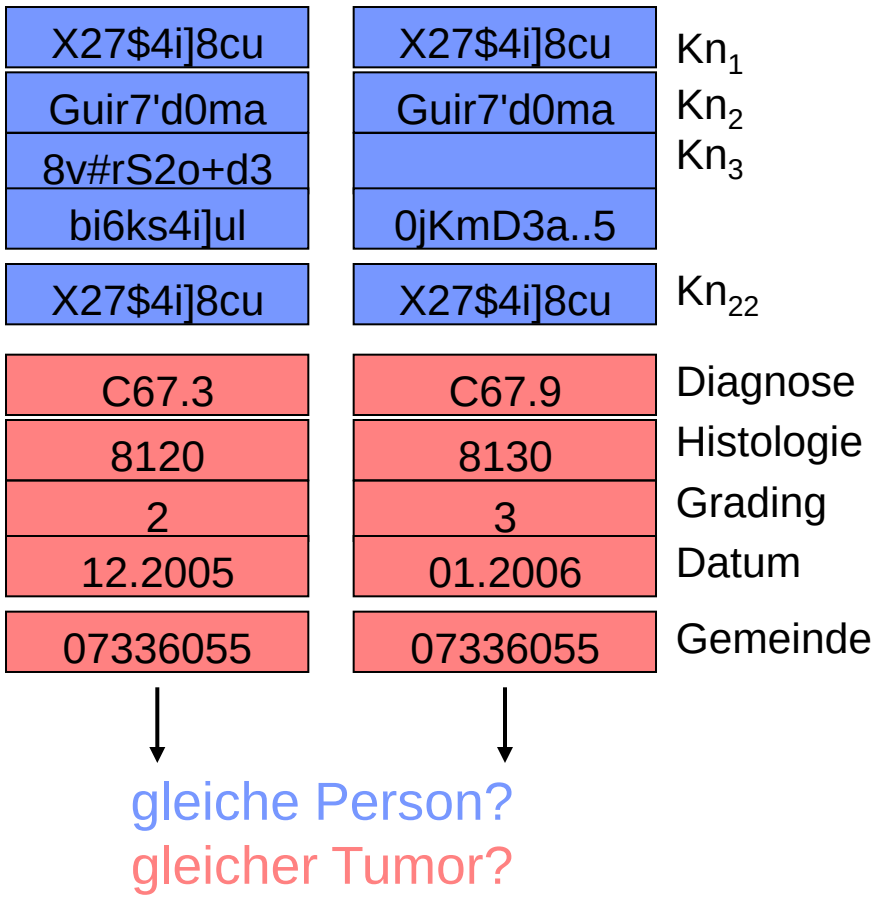
Kölner Phonetik	
a, ae, ai, au, ay, e, ei, eu, ey, i, ie, j, oe, oi, ue, y	→ e
b, p	→ b
c, ck, g, k, q	→ k
ch, cz, s, sc, sz, tz, x, z	→ s
d, dt, t	→ d
f, ph, v, w	→ f
m, n	→ m
o, ou, u, uo	→ u
l	→ l
r	→ r
h, Leerzeichen, Akzente, Spezialzeichen	→ weg



Epidemiol. Krebsregister
Record Linkage

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

- Fragen: Gibt es mehrere Meldungen
  - zur gleichen Person?
  - zum gleichen Tumor?
- Vergleich der neuen Meldungen mit allen Bestandsmeldungen
- Problem: Abweichungen, Daten passen *niemals* perfekt
- Verfahren: Stochastisches Record-Linkage

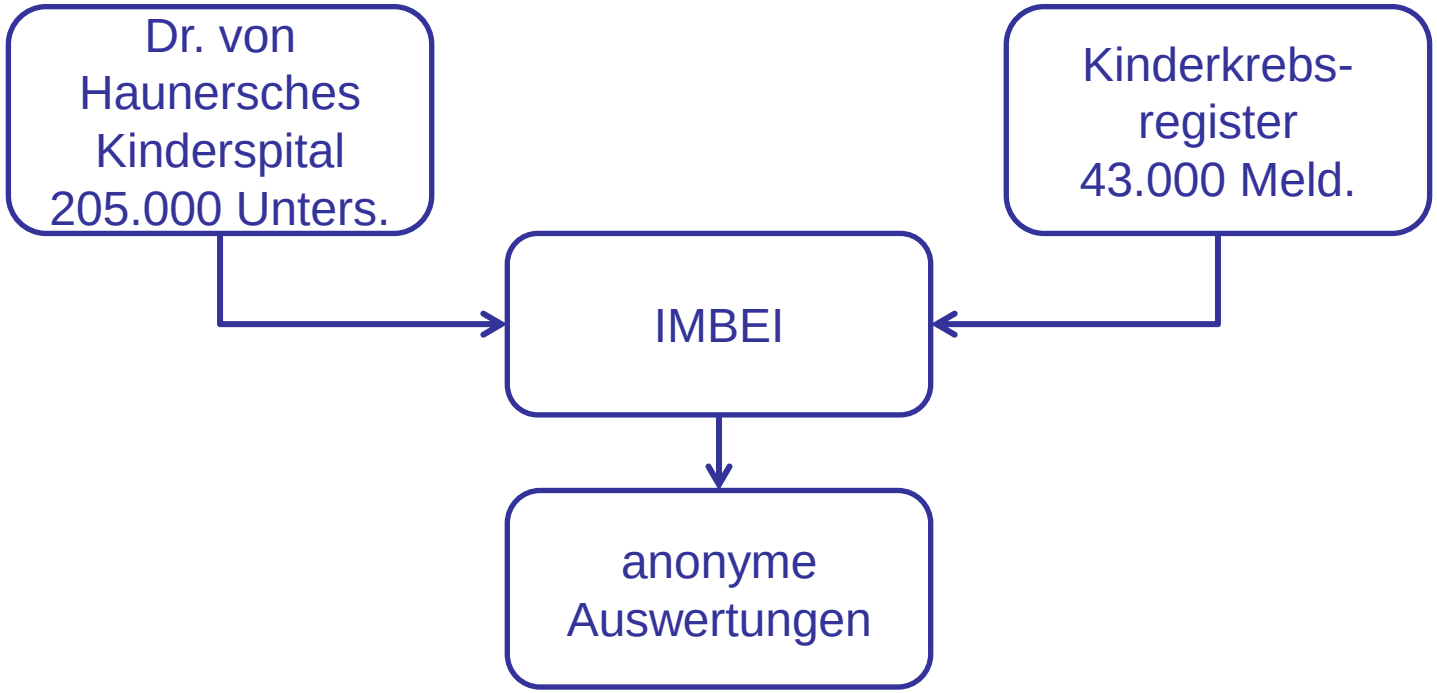


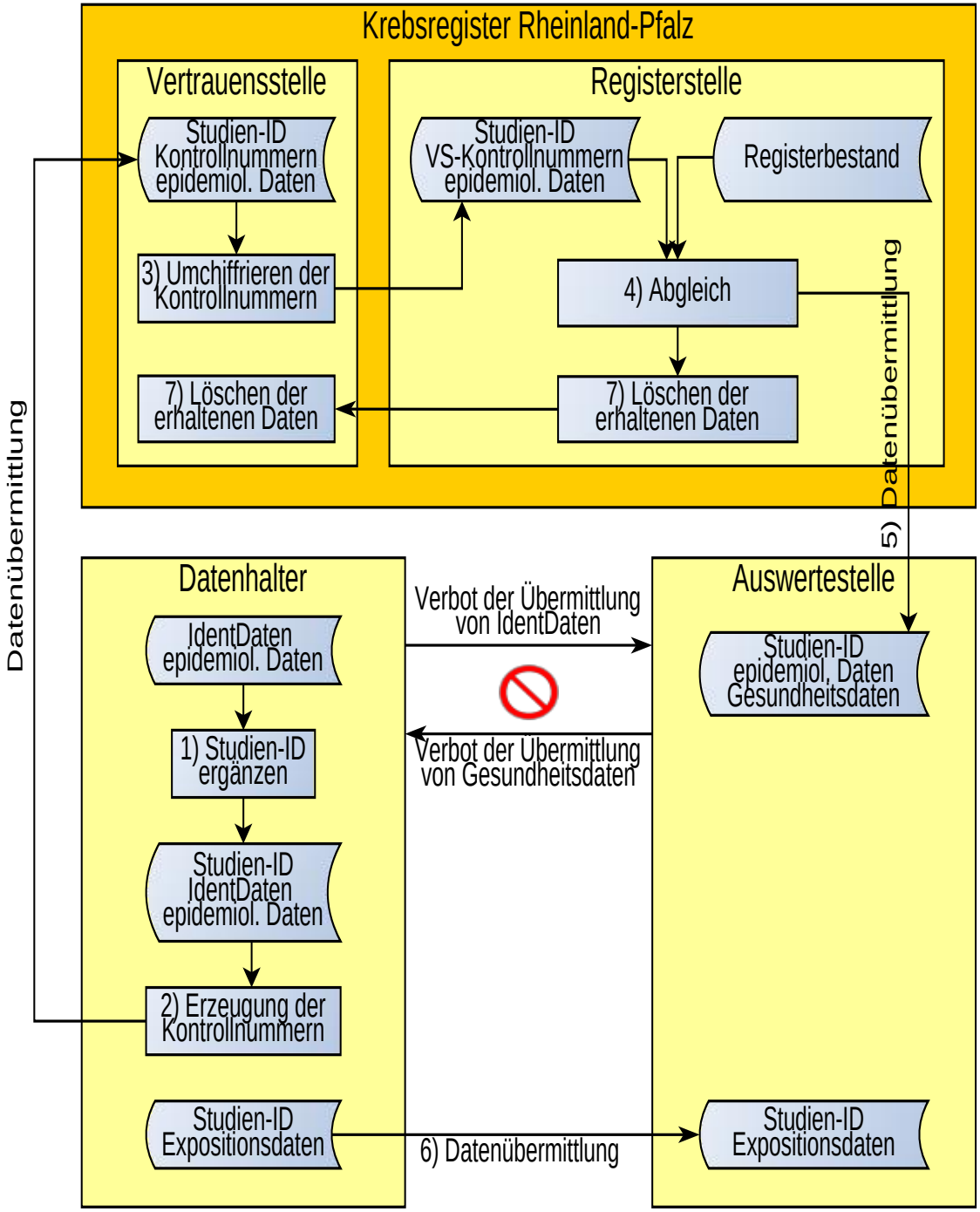
Datenschutz in der RICC-Studie

Datenschutz in der RICC-Studie

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

- IMBEI als Treuhänder und Auswerter
- Verwendung von Kontrollnummern mit studienspezifischem IDEA-Schlüssel
- Keine Rückübermittlung von Daten




Krebsregister Rheinland-Pfalz
Vertrauensstelle
Registerstelle
Studien-ID
Kontrollnummern
epidemiol. Daten
Studien-ID
VS-Kontrollnummern
epidemiol. Daten
Registerbestand
3) Umchiffrieren der Kontrollnummern
4) Abgleich
7) Löschen der erhaltenen Daten
7) Löschen der erhaltenen Daten
5) Datenübermittlung
Datenübermittlung
Datenhalter
Auswertestelle
Verbot der Übermittlung von IdentDaten
IdentDaten
epidemiol. Daten
Studien-ID
epidemiol. Daten
Gesundheitsdaten
1) Studien-ID ergänzen
Verbot der Übermittlung von Gesundheitsdaten
Studien-ID
IdentDaten
epidemiol. Daten
2) Erzeugung der Kontrollnummern
Studien-ID
Expositionsdaten
6) Datenübermittlung
Studien-ID
Expositionsdaten

Garbage in …

Datenaufbereitung

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

- 205.000 Untersuchungen in der Kohorte
- Deduplizieren: 103.000 Kinder
- Adressbereinigung

Vorname	Name	Geb.datum	PLZ	Ort	Straße
			8	M	Lindwurm 4
				Moos	3
				Oberschleißheim	
				Oberschlei heim	
			8	M	Konsulat

Stochastisches Record-Linkage in der RICC-Studie

Stochastisches Record-Linkage - Stufen

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik

- Verwendung von Kontrollnummern, binärer Vergleich
- Merge Toolbox
- Mehrstufige Strategie
  - In jeder Stufe werden festgelegt:
    1. Variablen deren Ausprägungen exakt übereinstimmen müssen (”Block-Variablen”) (sozusagen ein deterministischer Teil)
    2. Variablen zur Berechnung des Scores (”Match-Variablen”, "stochastischer Teil")
    3. Schwellenwerte für die Kategorisierung des Scores in “Match / Graubereich / Non-Match”

Klassisches Modell von Fellegi-Sunter

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik

- Definiere die bedingten Wahrscheinlichkeiten
  - m_i = P(Gleicher Wert in Variable *i* | Datensätze gehören zusammen)
  - u_i = P(Gleicher Wert in Variable *i* | Datensätze gehören nicht zusammen)
- Berechne Übereinstimmungs-Scores aus den Odds:

Werte von Variable *i*	Wahrscheinlichkeiten		Odds	$Score_i$
	Matches	Non-Matches		
gleich	m_i	u_i	$m_i : u_i$	$\ln(m_i : u_i)$
ungleich	$1- m_i$	$1- u_i$	$1-m_i : 1-u_i$	$\ln(1-m_i : 1-u_i)$

Gesamtscore = $\Sigma(Score_i)$

Stochastisches Record-Linkage - Beispiel

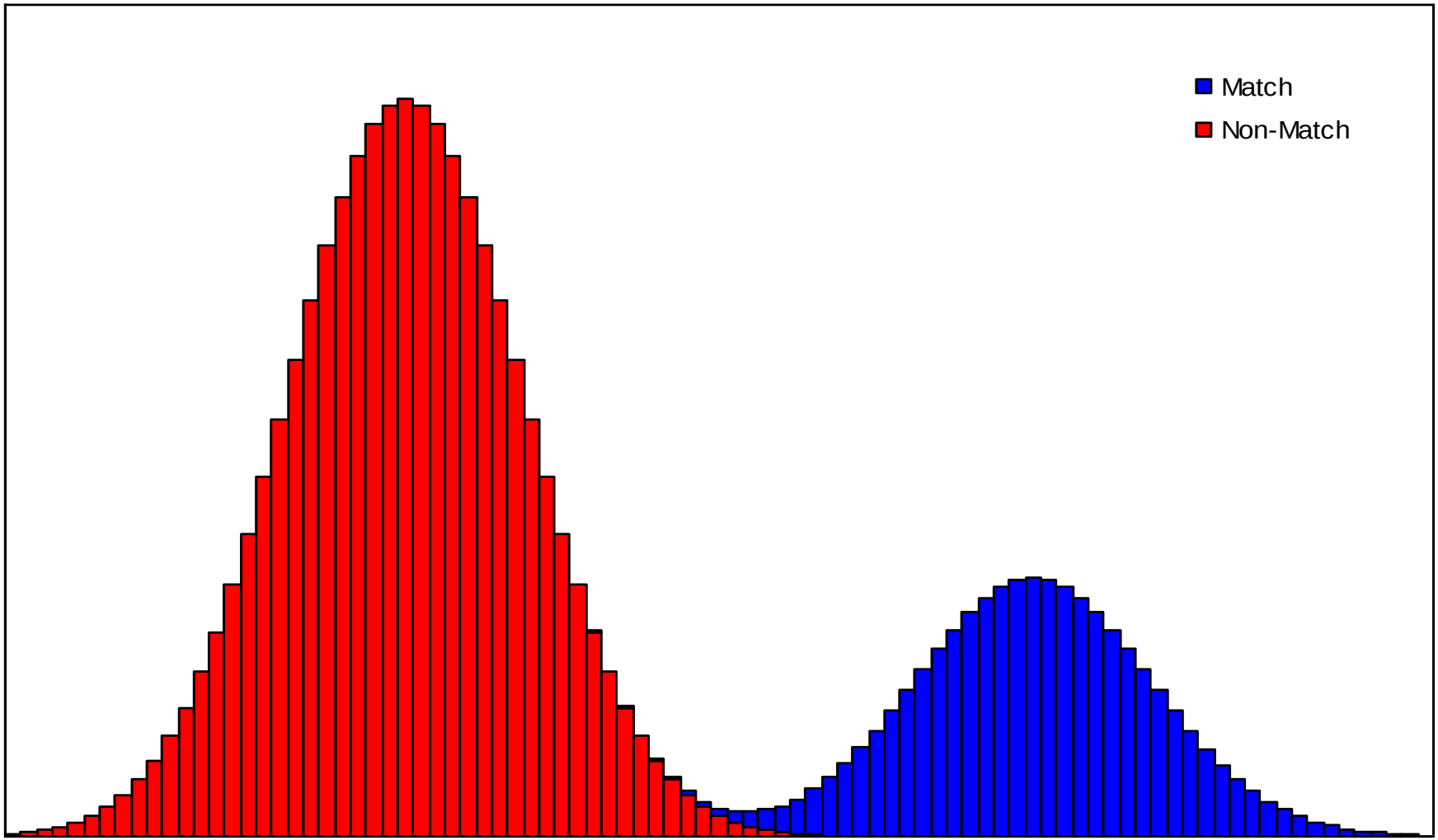
JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik

				Score	
				Gleichheit	Ungleichheit
Variable	Ausprägung	*m*	*u*	ln(*m*:*u*)	ln(1-*m*:1-*u*)
Vorname	Thomas	0,95	0,01	+4,55	-2,99
	Gaël	0,95	0,000001	+13,76	-3,00
	(andere)	0,95	0,01	+4,55	-2,99
Geschlecht		0,99	0,5	+0,7	-1,6

- Das Geschlecht trägt viel weniger Information bei als der Vorname, ob übereinstimmend oder nicht.
- Die Wahrscheinlichkeit *u* kann aus den Daten geschätzt werden als 1/Häufigkeit der jeweiligen Ausprägung

Beispiel Scores

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik



Stochastisches Record-Linkage - Stufen

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

- erste Stufe: perfekte Übereinstimmung
  - Alle Variablen sind Block-Variablen
- nächste Stufen: Relaxierungen
  - unterschiedliche Block-Variablen
- unveränderte Match-Variablen

Stochastisches Record-Linkage – Beispiel Stufen

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik

	Stufe							
	1	**2**	**3**	**4**	**5**	**6**	***m***	***u***
Block-Variablen	Vornam1	VornamΦ	Vornam1	VornamΦ	Vornam1	Geschlecht		
	Nachnam1	NachnamΦ	Nachnam1	NachnamΦ	Nachnam1	GebTag		
	GebTag	GebTag	GebJahr	GebJahr	PLZ5	GebMonat		
	GebMonat	GebMonat	Ort	Ort		GebJahr		
	GebJahr	GebJahr						
Match-Variablen	Vorname						0,95	0,01
	Nachname						0,90	0,01
	Geschlec						0,80	0,50
	GebTag						0,90	0,03
	GebMonat						0,90	0,08
	GebJahr						0,98	0,01
	PLZ4						0,80	0,02
	PLZ5						0,85	0,01
	Ort						0,80	0,01
	Diagnose-Monat+Jahr						0,90	0,01

Stochastisches Record-Linkage – Ergebnis (n:m)

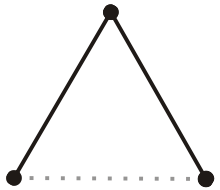
JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

Datei A	Datei B	
ID	ID	Score
1	198	51,28
1	217	45,38
1	1024	26,11
4	1024	49,70
6	37	33,45

Clerical Review

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

- RL-Software liefert: Liste von Paaren verglichener Meldungen + Übereinstimmungs-Score
- Aufbereitung (SAS): Gruppen potentiell zusammengehöriger Meldungen
- Handarbeit: Meldungen gehören zusammen/nicht



Grp	G	Vornam	Nach	Geb	Str	Ort	Tod	Diagn	Topo	Morph
9876	M	FRANZ	PUSICHA	13.02.1939	HERFORDER	BIELEFELD		11.2004	C349	80423
9876	W	*LUDWIG FRANZ*	PUSICHA	13.02.1939	HERFORDER	BIELEFELD		11.2004	C349	80423
9876	U	FRANZ	*PURSICHA*	*13.02.1938*	HERFORDER	BIELEFELD	18.11.2008			

Stochastisches Record-Linkage – Beispiel Clerical Review

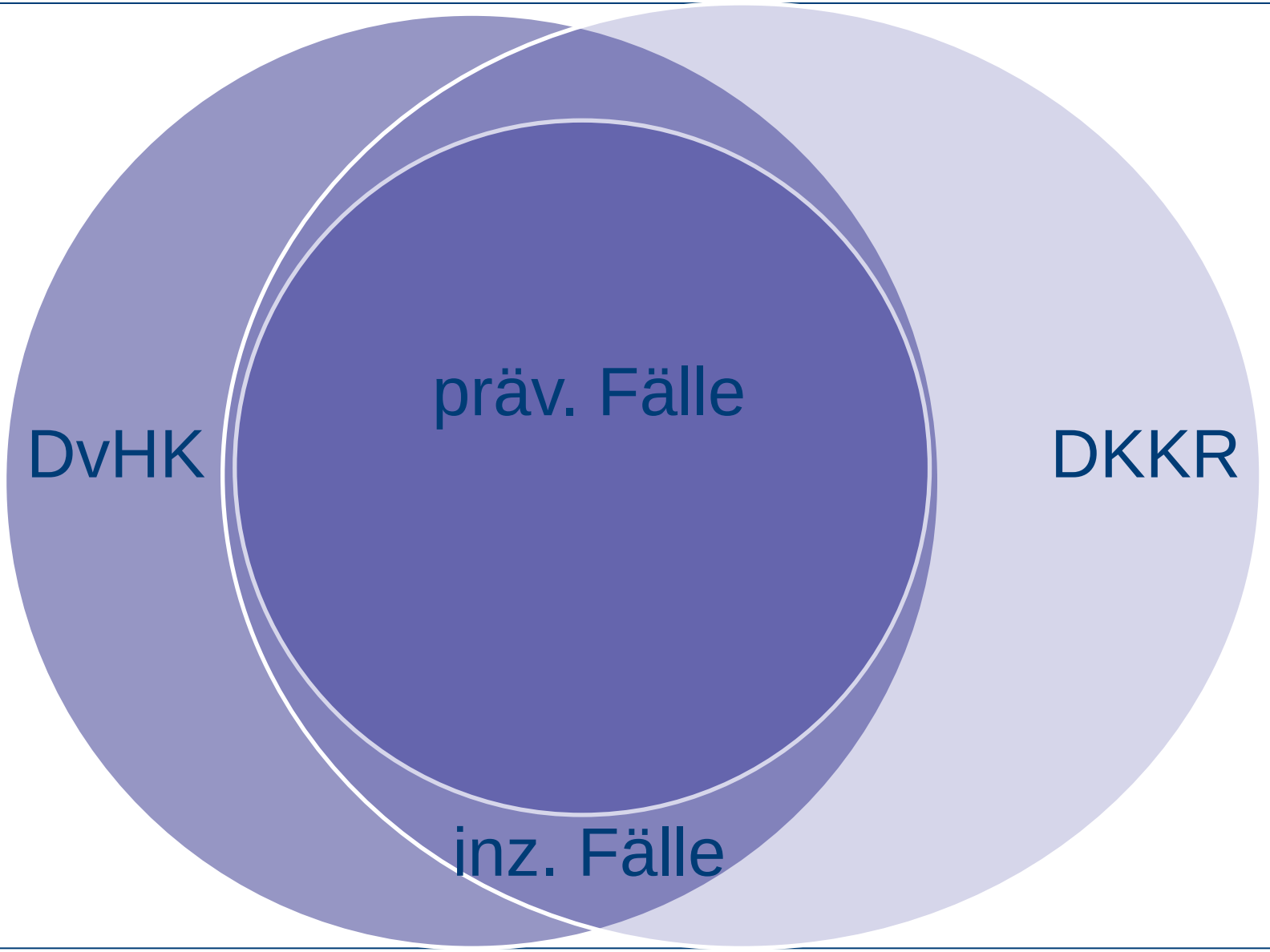
JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	Ma	Weight	Ident?	atient	Unters		Adr	G	Geburt	Vorname	Vers.	Nachname	VersName	PLZ	Ort	Unters/Dia	Alt	Indikation
2	ch	Weight			Malig	ZT		G	T MM JJJJ	1+b 2+b 3+b K+	Vornam	1+b 2+b 3+b	Frühname	4 5		TT MM JJJJ	er	ICD03 ICCC3
14699	MA	33.12		[illegible]	[illegible]		alt	M	F 01 1979	H J		G		I		B 02 1990	133	SCHMERZEN UN
14700	MB	33.12	1	[illegible]	[illegible]	1		M	F 01 1979	H J		G G	G G	I	D	01 03 1990		89203 91
14701	+A			[illegible]	[illegible]		alt	M	F 01 1979	H J		G		I		B 02 1990	133	SCHMERZEN UN
14702																		
14703	MA	32.92		[illegible]	[illegible]		neu	M	T 01 1975	I B L	E	D H R		S	F	C 08 1982	91	SCHMERZEN UN
14704	MB	32.92	2	[illegible]	[illegible]	1		M	T 01 1975	I L		Q Q	Q Q	P	F	12 08 1987		96913 22
14705																		
14706	MA	32.77		[illegible]	[illegible]		alt	M	L 04 1989	F H		G				A 02 1995	70	SEPTISCHER S
14707	MB	32.77	1	[illegible]	[illegible]	1		M	L 04 1989	F H		G G	G G	J	K	07 03 1994		96003 41
14708																		
14709	MA	32.53		[illegible]	[illegible]		alt	M	F 05 2002	G D		K				E 11 2003	18	
14710	MB	32.53	1	[illegible]	[illegible]	1		M	F 05 2002	G D		H H	H H	I	J	11 08 2003		89703 71
14711																		
14712	MA	32.41		[illegible]	[illegible]		alt	M	O 07 1992	H A		I				J 09 2001	110	
14713	MB	32.41	1	[illegible]	[illegible]	1		M	O 07 1992	H R		I I	I I	U	S	28 09 2001		96593 21
14714	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				G 10 2001	111	
14715	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				T 11 2001	112	
14716	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				B 02 2002	115	
14717	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				N 05 2002	118	
14718	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				V 08 2002	121	
14719	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				Q 11 2002	124	
14720	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				L 01 2003	126	
14721	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				K 05 2003	130	
14722	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				E 07 2003	132	
14723	+A			[illegible]	[illegible]		alt	M	O 07 1992	H A		I				P 10 2003	135	
14724																		
14725	MA	31.78		[illegible]	[illegible]		alt	W	G 04 2001	M A		D				K 06 2002	14	
14726	MB	31.78	1	[illegible]	[illegible]	1		W	G 04 2001	I L A		D E L	D E L	J	C	01 07 2005		89603 61
14727																		
14728	MA	31.48		[illegible]	[illegible]		alt	M	O 07 1992	N H		P		M	B	D 01 1993	6	HWI
14729	MB	31.48	1	[illegible]	[illegible]	1		M	O 07 1992	J H		P P	P P	I	B	03 08 2005		90643 101
14730	+A			[illegible]	[illegible]		alt	M	O 07 1992	N H	L	P	P	Q	B	G 12 1994	29	TRAUMA
14731																		
14732	MA	30.95		[illegible]	[illegible]		alt	W	J 08 1992	B D		M				K 07 2003	131	
14733	MB	30.95	1	[illegible]	[illegible]	1		W	J 08 1992	B I		M M	M M	N	C	20 10 1997		94703 33
14734	+A			[illegible]	[illegible]		alt	W	J 08 1992	B D		M				F 09 2001	109	
14735																		
14736	MA	30.57		[illegible]	[illegible]		alt	W	H 03 1997	A E		I				G 01 2001	46	
14737	MB	30.57	1	[illegible]	[illegible]	1		W	H 03 1997	B E		I I	I I	C	F	26 07 1999		98363 11

… was hinten 'rauskommt

RICC-Studie – Ergebnis des Clerical Review

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik



RICC-Studie – Ergebnis des Clerical Review

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie und Medizinische Informatik

- 1210 Gruppen potentieller Treffer
  - Alle Gruppen wurden geprüft
  - 40 Gruppen telefonisch geklärt
- 25 Non-Matches
- 1185 Matches
  - 1098 Prävalente Krebsfälle = Ausschlüsse
  - 87 Inzidente Krebsfälle im Studienzeitraum
- 9 Weitere, potentielle Krebsfälle in der Kohorte
  - plausible Gründe, warum nicht an das Kinderkrebsregister gemeldet
- Kinderkrebsregister kennt weitere Fälle aus dem DvHK, die nicht in der Radiologie vorgestellt wurden

Ergebnisse der RICC-Studie

RICC-Studie – Ergebnisse

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

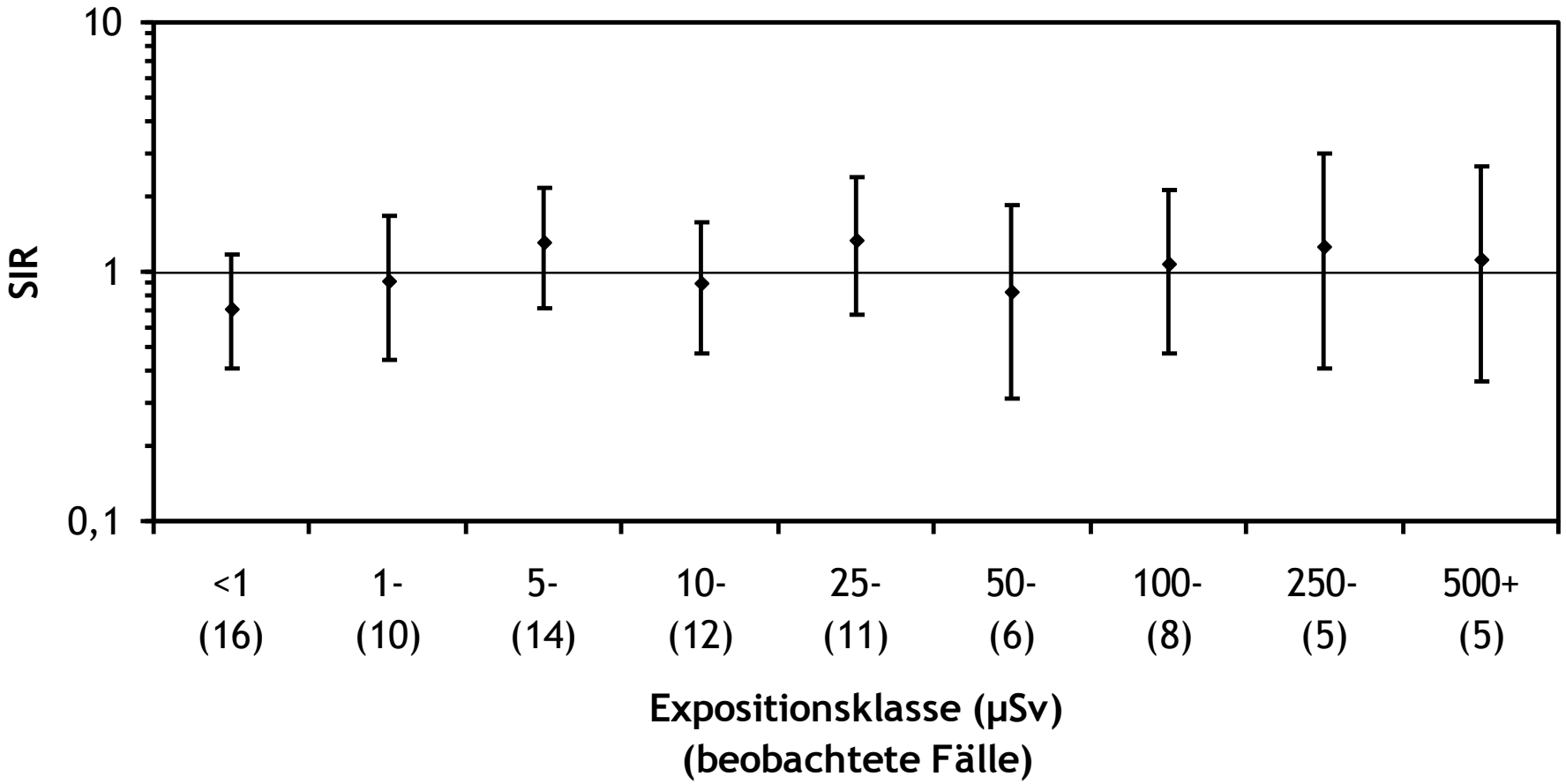
Standardisierte Inzidenz-Ratio (SIR)

	O	E	SIR	95%-KI
Geschlecht				
Jungen	52	52,8	0,99	0,74-1,29
Mädchen	35	35,2	1,00	0,69-1,38
Krebs insgesamt	87	88,0	0,99	0,79-1,22
Leukämie	33	30,5	1,08	0,74-1,52
lymphatische Leukämie	24	24,5	0,98	0,63-1,45
akute myeloische Leukämie	5	4,3	1,16	0,38-2,70
Lymphome	13	13,4	0,97	0,52-1,66
ZNS-Tumore	10	19,3	0,52	0,25-0,95
andere Tumore	31	24,8	1,25	0,85-1,77

RICC-Studie – Ergebnisse

UNIVERSITÄTSmedizin. MAINZ
Institut für Biometrie, Epidemiologie
und Medizinische Informatik

Dosis-Wirkungs-Beziehung



Fazit

Fazit

- Wichtig
  - Qualität, Qualität, Qualität der Eingangsdaten
  - Bildung von „Match-Gruppen“
  - Nutzung von Zusatzinformationen im Clerical Review

- Der Abgleich von Kohorten mit Krebsregistern
  - ist machbar
  - kann dem Register nützen
  - Kontrollnummern statt Klartext kein Hinderungsgrund

Der PID-Generator der TMF

TMF-Workshop „Tools zum ID-Management in der klinischen Forschung"

Berlin, 24. September 2010

Prof. Dr. Klaus Pommerening, Dr. Murat Sariyar
Universitätsmedizin Mainz, IMBEI
KN Pädiatrische Onkologie und Hämatologie
TMF-AG Datenschutz

Gefördert vom
Bundesministerium für Bildung und Forschung

http://www.tmf-ev.de/