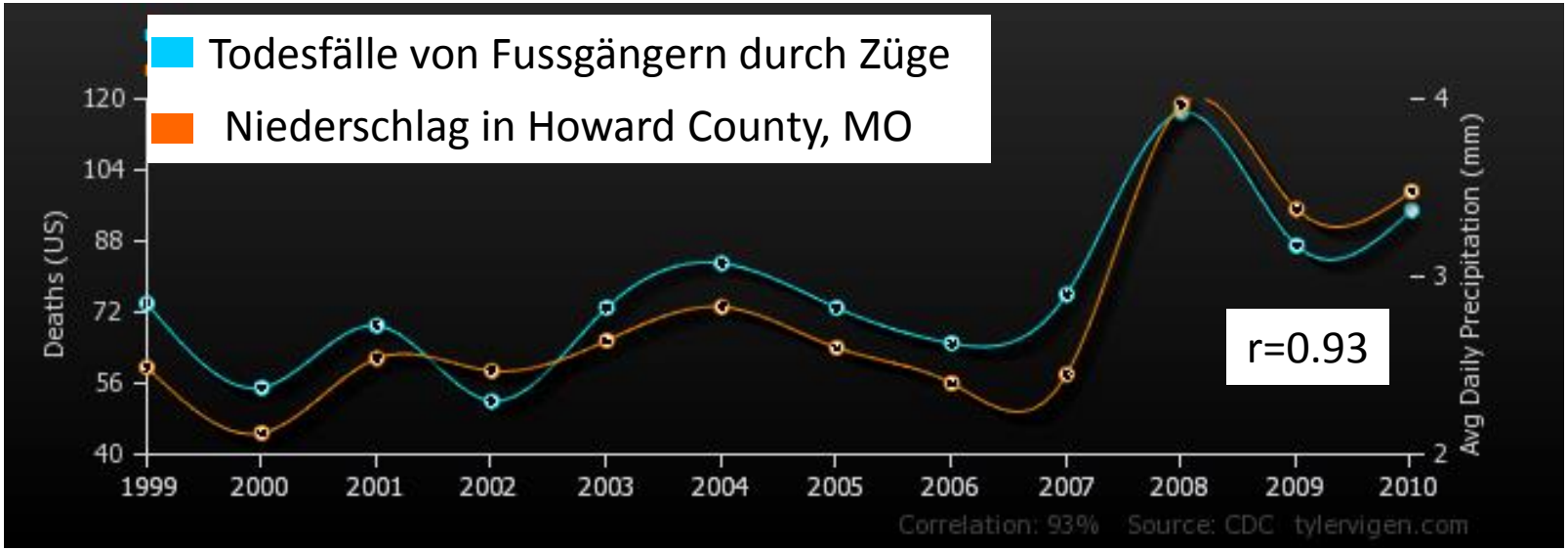
Sieg der Korrelation über Kausalität? Methodische Fragen im Kontext von Big Data



Dr. Amke Caliebe

Institut für Medizinische Informatik und Statistik
Christian-Albrechts-Universität zu Kiel

Big Data konkret - bvitg, smart data, tmf – Berlin, 13. Dezember 2016

Big data in der Medizin

"In short, **translational research** is the process of **turning appropriate biological discoveries into drugs and medical devices** that can be used in the treatment of patients."

Bakir M. (2011) Bosn J Basic Med Sci 11: 73

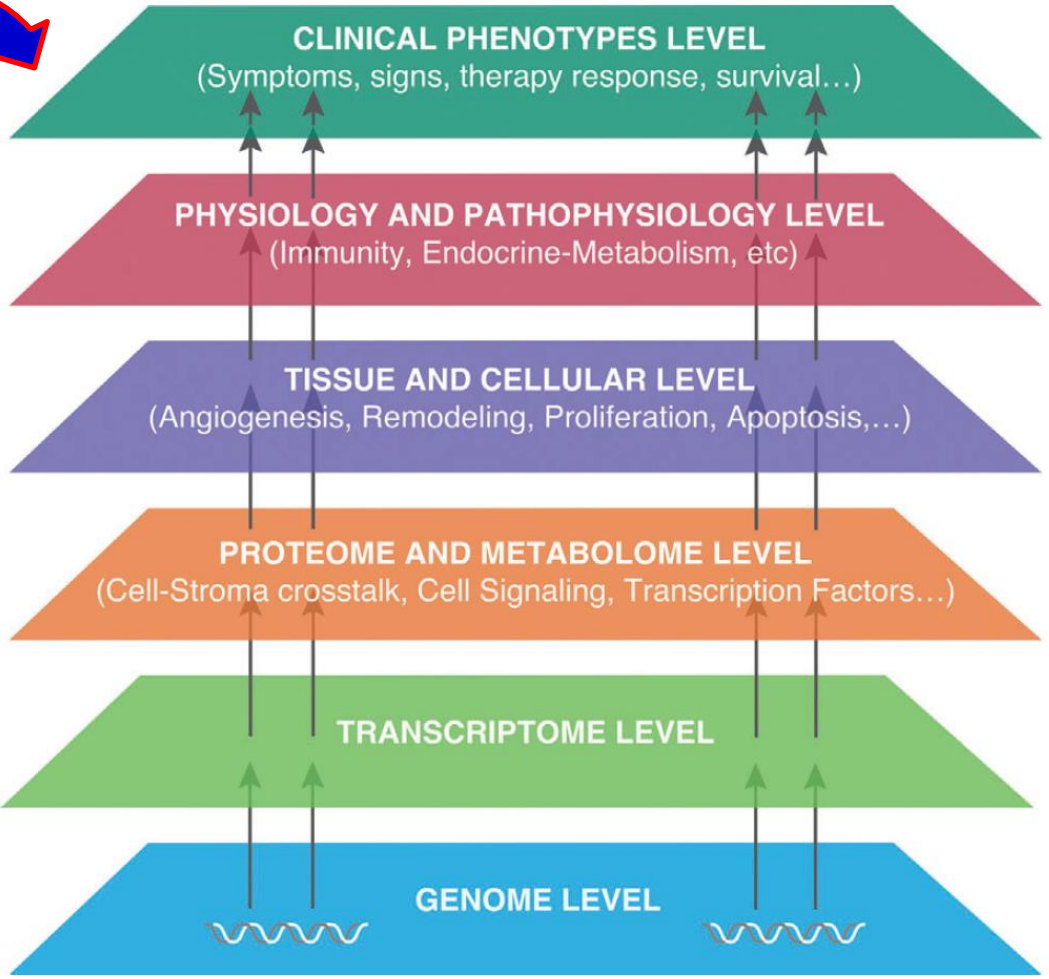
"**Precision medicine** is an emerging approach for disease treatment and prevention that takes into account **individual variability in genes, environment, and lifestyle** for each person."

www.nih.gov, 2016

Big data in der Medizin

ENVIRONMENT

smoking,
nutrition,
physical activity,
air pollution, …



Blanco-Gomez A et al. (2016) Bioessays 38: 664-673

US Präsidentenwahl 1936
Landon vs. Roosevelt

60% Landon

Literary Digest-Umfrage (**10,000,000** Wähler):
60% Landon

US Präsidentenwahl 1936
Landon vs. Roosevelt

60% Roosevelt

Ergebnis: 60% Roosevelt

US Präsidentenwahl 1936
Landon vs. Roosevelt

"The answer, very simply, was the Digest's reliance on **voluntary response**. Ten million sample ballots were mailed to prospective voters, but only 2.3 million were returned. As everyone ought to know, such samples are practically always **biased**."

Bryson MC (1976) Am Stat 30: 184-185

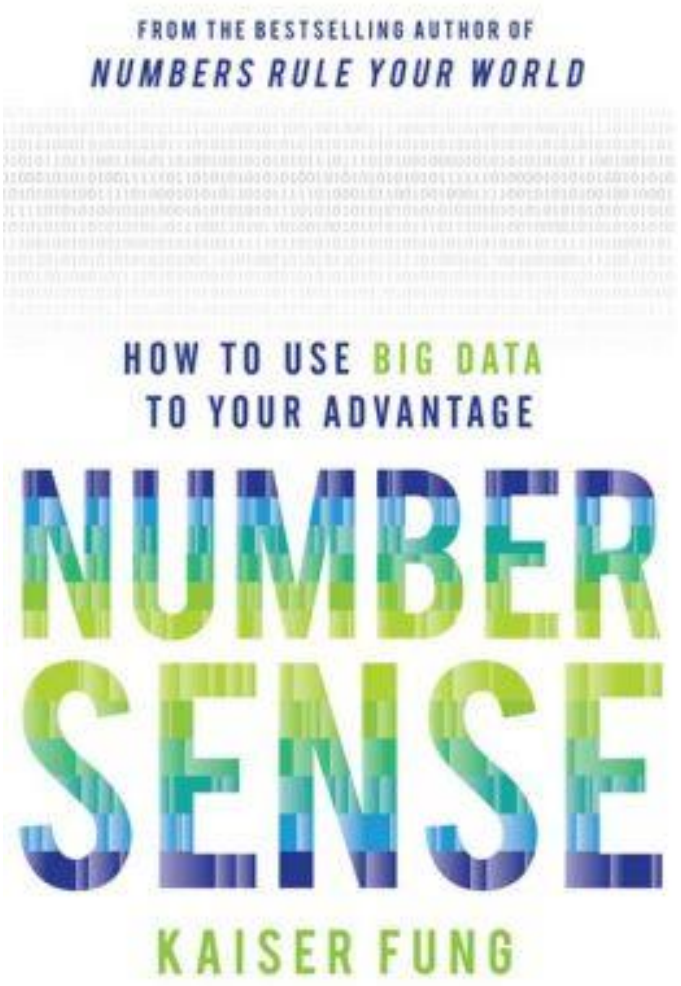
garbage in – garbage out

→ big garbage in – big garbage out

Datenqualität

Big data ist typischerweise

- observational
- ohne Kontroll-Daten
- scheinbar vollständig
- aus dritter Hand
- zusammengemischt



Datenqualität

- Wenn die Datenerfassung nicht verstanden ist, werden unzutreffende Schlüsse gezogen und die Evidenz der Ergebnisse ist zweifelhaft

- "Rebuttal to Simply Aggregating Data"
  Brian S. Yandell, Head of Department of Statistics, University of Wisconsin-Madison, 2013

→ Good Data ist wichtiger als Big Data

Datenqualität

- **Veracity**
  Sind die Daten für das Problem angemessen?

- **Validity**
  Sind die Daten valide und von hoher Qualität?

- **Volatility**
  Wie lange sind die Daten aktuell?

Inderpal Bhandar, Chief Data Officer at Express Scripts
Big Data Innovation Summit in Boston, 2013

Fallzahlen

- Overfitting
- Google Flu Trends (GFT):

50 Mio Suchbegriffe für **1152 Datenpunkte**

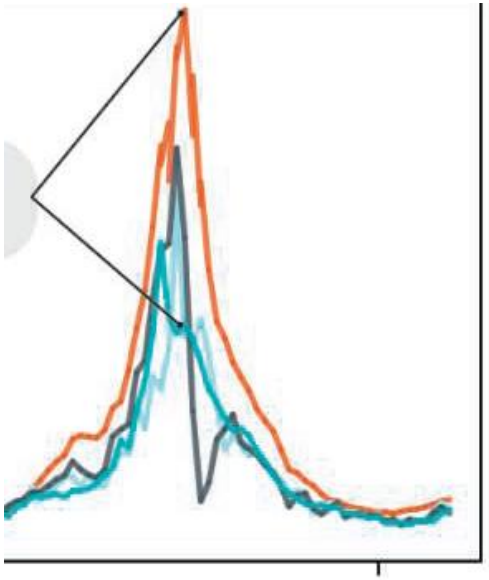
„What's exciting about Flu Trends is that it lets anybody -- epidemiologists, health officials, moms with sick children -- learn about the current flu activity level in their own state based on data that's coming in this week."

Jeremy Ginsberg, Google, CNN, 2008

Fallzahlen

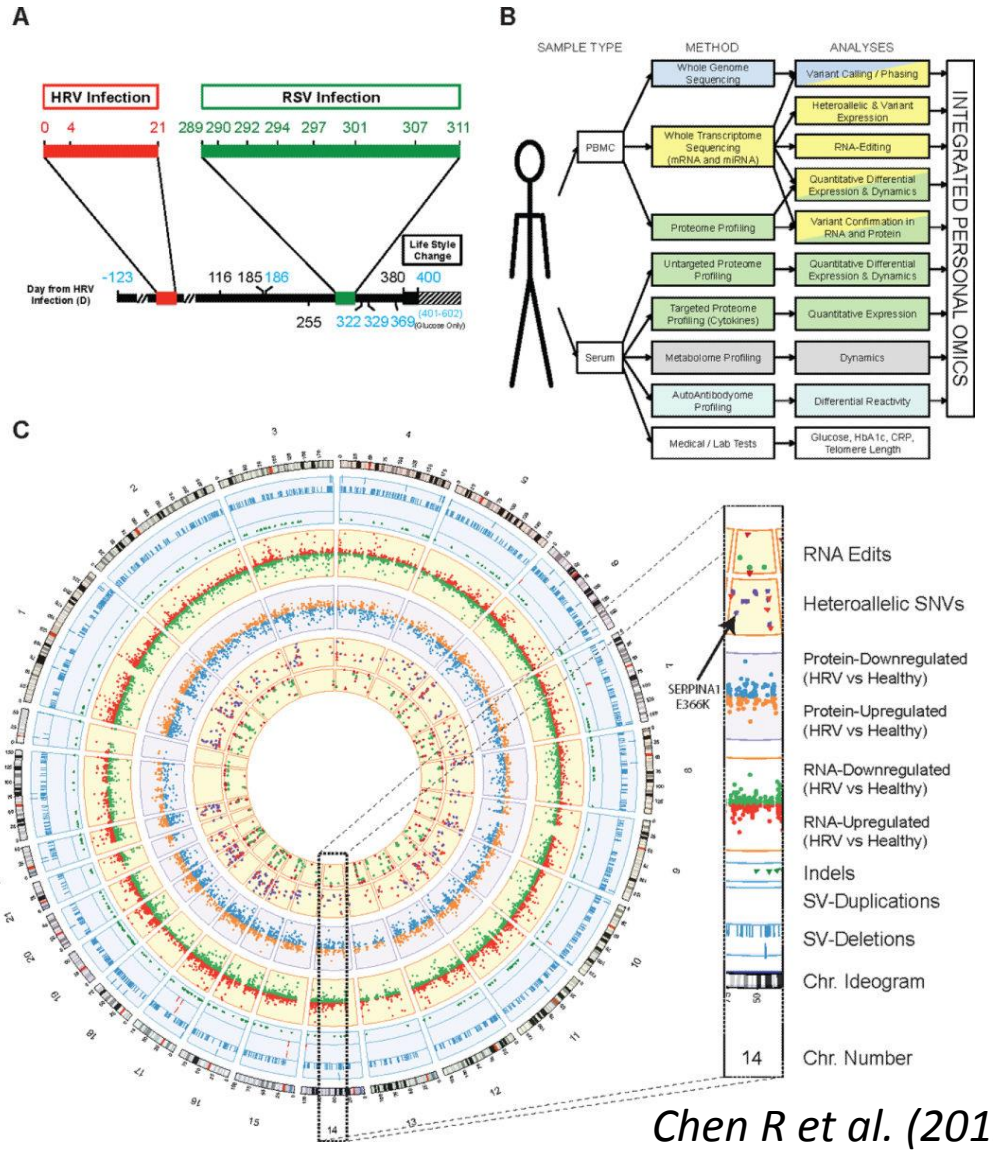
GFT überschätzte die Prävalenz für Grippe für 100 von 108 Wochen in der Saison 2011/2012.
Ein einfaches Vorhersagemodell – wie ein Modell das die Temperatur aus den Temperaturen der letzten Woche vorhersagt – erzielte deutlich bessere Resultate.

Lazer et al. (2014) Science 343: 1203-5



GFT mehr als doppelt so hoch

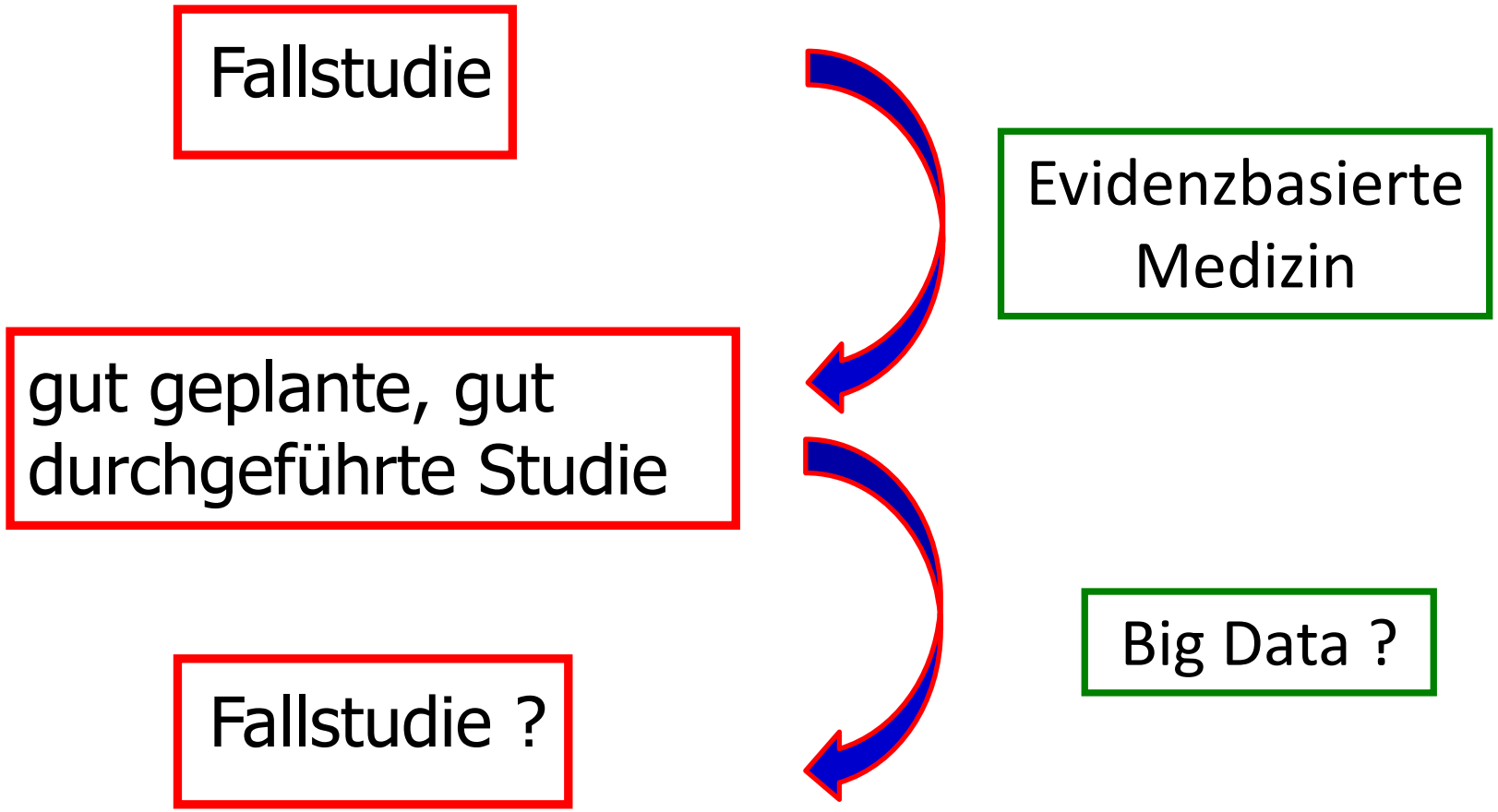
iPOP – Snyder



→ n=1

Chen R et al. (2012) Cell 148: 1293–1307

Fallzahlen


Fallstudie
Evidenzbasierte Medizin
gut geplante, gut durchgeführte Studie
Big Data ?
Fallstudie ?

Kausalität vs. Korrelation

© Wired Magazine

"All models are wrong, but some are useful."

George Box, Statistiker, University of Wisconsin, 1978

"All models are wrong, and increasingly you can succeed without them."

Peter Norvig, Director of Research, Google Inc., 2008

"Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."

Chris Anderson, Chief Editor, Wired Magazine, 2008

Kausalität vs. Korrelation

Vergleich der Medikamente $M_1,\ldots,M_n$
zur Behandlung einer Krankheit

Big-Data-Analyse:
M_5 hat die höchste Korrelation mit Behandlungserfolg

→ ~~M_5 ist das Medikament der Wahl~~

Confounders: Schweregrad der Krankheit, nationale Besonderheiten, sozialer Status, Preis der Medikamente, Subtyp der Krankheit, Vor/Begleiterkrankungen, Vorlieben der Ärzte,…

Kausalität vs. Korrelation

"Most studies in the health sciences aim to answer causal rather than associative questions."

Pearl J (2010) Causal Inference 6: 1-59

Kausalität vs. Korrelation

- **Assoziation: aus den Daten allein zu berechnen**
  Korrelation, Regression, Odds Ratio, Adjustierung

- **Kausalität: Annahmen außerhalb der Daten**
  Randomisierung, Einfluss, Effekt, Confounding, Intervention

Pearl J (2010) Causal Inference 6: 1-59

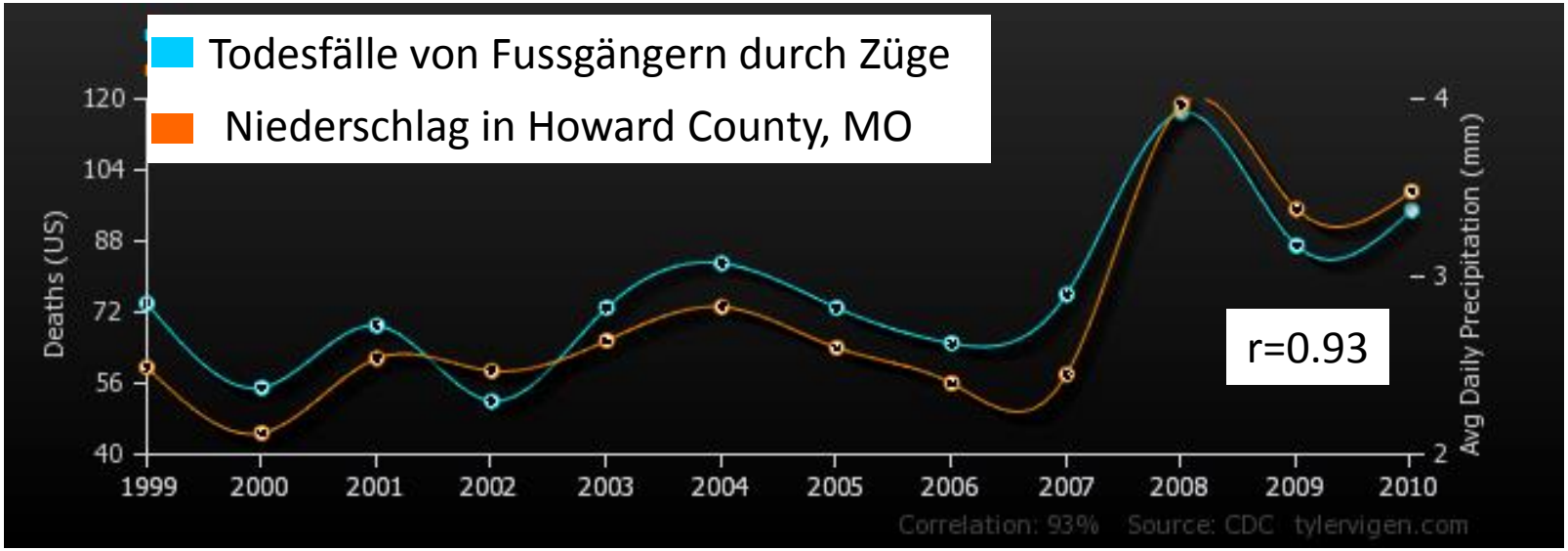
Von big data zu good data

- Ergebnisse von Big-Data-Analysen sind explorativ
- Bestätigung durch klinische Studien nötig
  - Randomisierte klinische Studien (Goldstandard)
  - Sinnvolle Fallzahl, sinnvolle Stichprobe
  - Definierte primäre/sekundäre Zielgrößen
- "To use big data in medicine, the results obtained from the analysis of big data should be validated by clinical studies."

Kanda E (2015) Use of big data in medicine. RenReplacTher

Fazit

- Big data in der Medizin benötigt qualitativ hochwertige, reichhaltige und umfangreiche Patientenkollektive
- Exzellenter Datenschutz ist ethisch und für die Akzeptanz unabdingbar
- Big-Data-Analysen können kontrollierte klinische Studien ergänzen, aber niemals ersetzen



Vielen Dank für Ihre Aufmerksamkeit!