External and internal quality of Big Data

Beat Hulliger, FHNW School of Business

BigSurv18, 27.10.2018, Barcelona

Introduction

- Experience with three large datasets in relation with surveys
- External quality: timeliness, accessibility, relevance, coverage, comparability, accuracy
- Internal quality: reliability, coherence, completeness

1. Evaluation of pharmacy service

- PMC: **P**oly**m**edication **C**heck
- Check on compliance and issues with medication
- Offered by pharmacies
- Paid by health insurance
- On a provisional basis

(Hulliger et al., 2017)

Survey

- Survey among all Swiss pharmacies ($N = 1720$)
- Questionnaire on acceptance of PMC: Acceptance and issues of PMC by pharmacists
- Questionnaire on PMC records (if PMC done): View of pharmacists about effect of PMC and about satisfaction of patients
- Sample: $n = 585$ pharmacies ($r = 0.34$), $n = 345$ PMC records

Health Insurance Data

- Secondary data from three large health insurance companies, covering 3.5 million patients in 2013 (coverage ≈ 0.44).
- Socio-demographic data and medical history data over three years.
- Longitudinal analysis of PMC-patients
- Quasi-experimental analysis with matched non-PMC patients
- Treatment: Persons with PMC
- Control: Samples of matched persons not taking the service using socio-demographic and medical history

Health Insurance Data ctd.

- 1’707 PMC-Patients compared with 14’015 Non-PMC Patients
- Medical history data (e.g. every drug with ATC, quantity, price, date etc., and PMC)
- Response: Hospitalisations, Emergencies, Doctor visits, Expenditures for drugs

Lessons learned

- Survey of pharmacists
  - Low response rate of survey and missing values: Possible non-response bias.
  - Analysis and interpretation of survey straightforward but only viewpoint of pharmacists (proxy for patients).
  - Pharmacists mixed acceptance of PMC.
- Secondary analysis
  - Lack of harmonisation between companies: joint analysis impossible.
  - Coverage of about 44%: possible differences compared with smaller companies.
  - Analysis involved.
  - Significant effect over short period for cost of drugs. Otherwise no clear signal!

2. Imputation of turnover in business census

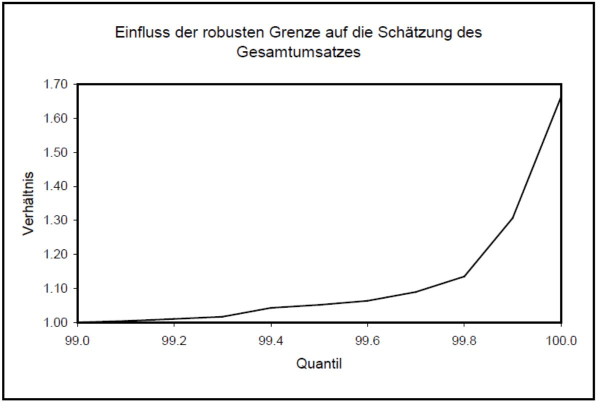
- Swiss business census 1995 (Hüsler and Müller, 2001)
- 277’331 enterprises
- 21% have missing turnover
- Various methods for the imputation (homogeneous groups, regression, robust variants)

Estimates of total GDP in BCHF:

	Lowest version	Highest version
Non-robust methods	550	1450
Robust methods	620	1040

- Standard error with multiple imputation: 10 BCHF
- Different models in different branches needed.
- Macro comparison with national accounts: All above 1000 BCHF excluded.
- For large data sets the problem of outliers and very skew distributions remains.

Figure: Total turnover vs. tuning constant for robustification



Source: Hüsler and Müller (2001)

3. TV-audience measurement

- Top-set box to register TV-audience.
- About 2000 panel members (households).
- Recruitment, instruction, installation, maintenance
- Highly sophisticated and detailed calibration to population.
- TV-audience measurement every 30 seconds (channel, persons).
- Analysis spells: day, week, month, trimester, semester.
- Problem: Are small TV-stations well covered by the audience measurement?

(Kuonen and Hulliger, 2013)

TV-audience measurement ctd.

- TV-audience measurement:
    - Big data in time dimension
    - Small survey in household dimension.
- Missing spells of measurements: Big data may help.
- Small area (households) estimation: Big data useless...?
- Small TV-channels: Rare event - Aggregation over time helps

Quality Overview

	PMC		BZ95	Mediapulse	
Dimension	Survey	Data	Imp.	Panel	Data
Relevance	2	2	3	3	3
Coverage	1	1	3	2	3
Comparability	2	1	3	2	1
Accuracy	2	1	1	1	3
Timeliness	3	2	2	1	3
Punctuality	3	1	2	3	3
Accessibility/Clarity	3	1	2	2	2
Reliability	1	3	3	2	3
Coherence	2	1	3	2	3
Completeness	2	1	2	1	2
Cost	2	3	3	1	1

(1=low, 2=middle, 3=high)

Conclusions

- Large datasets have the same problems of bias as any survey: coverage, non-response, robustness
- The bigger the data the larger the problems of comparability (definitions).
- Rare events may be captured by big data.
- Triangularisation may shed light on complex phenomena.
- Not the size of data makes the quality but how targeted the data is collected.

Bibliography

Hulliger, B., V. Butterweck, R. Niederer, M. Sterchi, and N. von Arx (2017). Studie zum Nachweis der Wirksamkeit, Zweckmässigkeit und Wirtschaftlichkeit des Polymedikationschecks (PMC).

Hüsler, J. and S. Müller (2001). Mehrfach imputierte Umsatzzahlen. Technical Report 338-0002, Swiss Federal Statistical Office. Schlussbericht Betriebszählung 1995.

Kuonen, D. and B. Hulliger (2013). Evaluation of the new mediapulse television panel with respect to its suitability for local television channels.