# Replicability of research results in anonymized data of the GCKD study

## Anonymization framework

* Controlled-access model: data are shared among researches with certain conditions (e.g. data-sharing agreement, within European Union)1

## Metadata

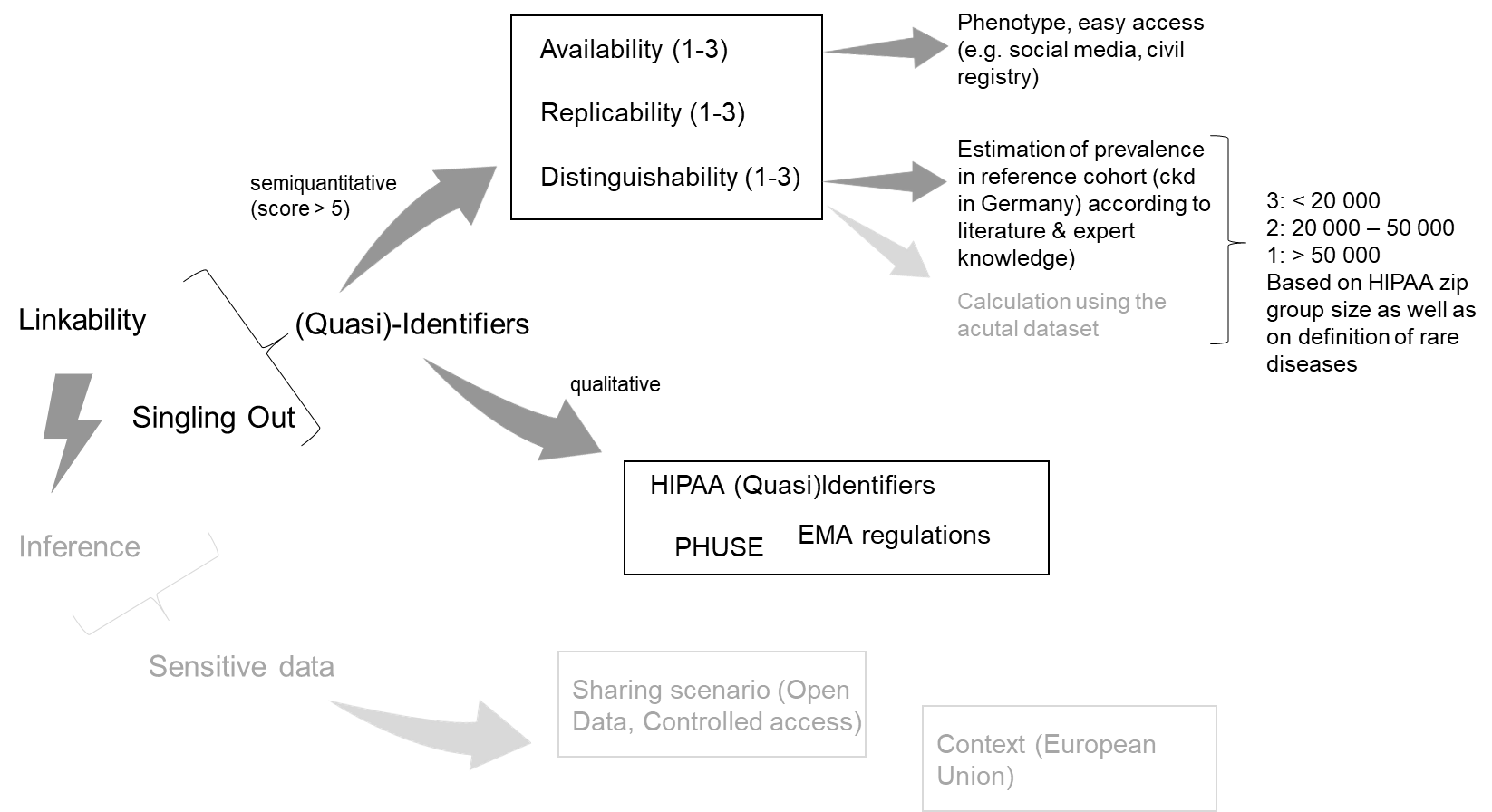
* Overall cohort: n=5217
* Total dataset: x variables
* Transferred dataset
  + includes data collected at baseline
  + n=5217
  + variables 361
* Use-case specific dataset (df1) = original dataset
  + includes data relevant for the specific use-case2 after
    - data cleansing: data-type, declaration of NA
    - creation of new variables: calculation, aggregation
    - variable selection
    - script on GitHub:
  + n=5217
  + variables 76
* Anonymized dataset
  + includes data after anonymization procedure
  + n=5217
  + variables 76

## Utility assessment

* Use-case specific utility assessment2
  + Each variable is scored 1 (low) – 3 (high) according to its relevance in the research scenario

## Risk assessment

* Declaration of Quasi-Identifiers (QI)
  + Defined by qualitative risk assessment and/or
  + Score > 5 in semi-quantitative risk assessment
* Qualitative risk assessment: based on HIPAA3, EMA regulations4 and further international recommendations 5,6
* Semi-quantitative risk assessment7:
  + Availability: based on phenotype and access (e.g. social media, civil registries)
    - 1=protected against access (e.g. creatinine value)
    - 2=possible observable (e.g. smoking)
    - 3=visible (e.g. gender)
  + Replicability: based on variability at different measuring points
    - 1=very variable (e.g. pulse)
    - 2=possible (e.g. drugs)
    - 3=mostly immutable (e.g. gender)
  + Distinguishability: based on estimated prevalence in German CKD reference cohort according to literature and expert knowledge; cut-off set in line with HIPAA zip group size and definitions of rare diseases
    - 1= > 50 000 (e.g. antihypertensive)
    - 2= 20 000-50 000 (e.g. primary glomerulopathy)
    - 3= < 20 000 (e.g. gfr)
* Currently not included
  + Distinguishability based on calculations using the actual dataset
  + Inference: Framework specific assessment of sensitive items



## Privacy models

* Requirement
  + Data must comply to k-anonymity with respect to the variables (=QI) age, gender, height, weight, bmi and biopsy performed
  + with k=2 / k=5 / k=11.
* Choice of k according to literature8,9
* Currently not included
  + L-diversity, t-closeness
  + Differential privacy

## Transformation models

* Models chosen: Generalization, suppression
  + Hierarchies on GitHub:
* Fixed settings
  + Top- & buttom-coding: de-activated
  + Utility measure: loss
  + Attribute weights: 0.5
  + Global transformation scheme
* Variable settings for fine tuning: Generic vs. use-case specific scenario
  + Minimum and maximum applied hierarchies: according to utility assessment
  + Suppression limit: according to generally accepted missing rates for statistical analyses10,11
* Currently not included
  + (Clustering &) Micro-Aggregation
  + Noise addition

## ARX configuration: Generic Scenario

* Sample extraction: 100%
* k-anonymity: k=2 / k=5 / k=11
* Generalization hierarchies
  + Minimum and maximum values
    - BL\_age: Minimum value 0 – Maximum value 150
    - BL\_ku\_weight: Minimum value 0 – Maximum value 300
  + No specification of minimum and maximum applied hierarchies
* Suppression limit: 100%
* Utility measure: loss
* Aggregate function: geometric mean
* Attribute weights: 0.5
* Population size
  + 2 000 000 (CKD in Germany)12
* Search strategy: Optimal
* Global transformation

## ARX configuration: Use-Case Specific Scenario

* Sample extraction: 100%
* k-anonymity: k=2 / k=5 / k=11
* Generalization hierarchies
  + Minimum and maximum values
    - BL\_age: Minimum value 18 – Maximum value 80
  + Specification of minimum and maximum applied hierarchies
    - BL\_age:
    - dem\_sex:
    - BL\_ku\_weight:
    - BL\_ku\_height:
    - BL\_ku\_bmi:
    - biopsy:
* Suppression limit: 5%
* Utility measure: loss
* Aggregate function: geometric mean
* Attribute weights: 0.5
* Population size
  + 2 000 000 (CKD in Germany)12
* Search strategy: Optimal
* Global transformation

## References

1 Keerie, C. *et al.* Data sharing in clinical trials - practical guidance on anonymising trial datasets. *Trials* **19**, 25, doi:10.1186/s13063-017-2382-9 (2018).

2 Titze, S. *et al.* Disease burden and risk profile in referred patients with moderate chronic kidney disease: composition of the German Chronic Kidney Disease (GCKD) cohort. *Nephrol Dial Transplant* **30**, 441-451, doi:10.1093/ndt/gfu294 (2015).

3 Garfinkel, S. L. De- Identification of Personally Identifiable Information. (National Institute of Standards and Technology (NIST), 2015).

4 European Medicines Agency (EMA). External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. (2018).

5 Institute of Medicine of the National Academies. in *Appendix B: Concepts and Methods for De-identifying Clinical Trial Data.* (2015).

6 PHUSE Data Transparency Working Group. Data Anonymisation and Risk Assessment Automation. (2020).

7 Malin, B., Loukides, G., Benitez, K. & Clayton, E. W. Identifiability in biobanks: models, measures, and mitigation strategies. *Hum Genet* **130**, 383-392, doi:10.1007/s00439-011-1042-5 (2011).

8 El Emam, K. & Dankar, F. K. Protecting privacy using k-anonymity. *J Am Med Inform Assoc* **15**, 627-637, doi:10.1197/jamia.M2716 (2008).

9 Prasser, F., Kohlmayer, F. & Kuhn, K. A. The Importance of Context: Risk-based De-identification of Biomedical Data. *Methods Inf Med* **55**, 347-355, doi:10.3414/ME16-01-0012 (2016).

10 Schafer, J. L. Multiple imputation: a primer. *Stat Methods Med Res* **8**, 3-15, doi:10.1177/096228029900800102 (1999).

11 Bennett, D. A. How can I deal with missing data in my study? *Aust N Z J Public Health* **25**, 464-469 (2001).

12 Girndt, M., Trocchi, P., Scheidt-Nave, C., Markau, S. & Stang, A. The Prevalence of Renal Failure. Results from the German Health Interview and Examination Survey for Adults, 2008-2011 (DEGS1). *Dtsch Arztebl Int* **113**, 85-91, doi:10.3238/arztebl.2016.0085 (2016).