



Quantifying Disclosure Risks

Jerry Reiter
Department of Statistical Sciences
Duke University, Durham NC, USA
jerry@stat.duke.edu



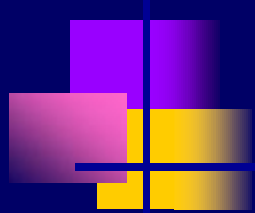
Setting for problem

- n Data producer collects data on individuals, including biomedical/genetic variables.
- n Data producer seeks to share collected data after removing obvious identifiers.
- n Data producer concerned about risk of deductive disclosures.



Types of disclosure risks

- n *Identification disclosure*
Match record in released data with target.
- n *Perceived identification disclosure*
Match record in released data with incorrect target.
- n *Attribute disclosure*
Learn value of sensitive variable for target.
- n *Inferential disclosure*
Closely estimate value of sensitive variable for target.



Identification disclosure

n For particular target, snooper knows and matches on values of key variables that are in shared data, such as:

geography,

race, sex, marital status, age,

occupation, housing, income, family,

medical/genetic profiles



Measuring identification disclosure risk

- n Uniqueness: estimate which records are unique in population on available keys.

(Duplicates also at risk.)

- n Re-identification experiments: attempt to match released records to other databases.

(Match to existing data set if snooper knows target is in released data.)



Uniqueness

- n Select (discrete) variables deemed to be key identifiers.
- n Fit models to sample counts (e.g., Poisson regression models).
- n Predict population counts in cells of interest.

Elamir E., Skinner C. (2006). Record level measures of disclosure risk for survey microdata, *Journal of Official Statistics*.



Re-identification experiments

- n Find database with overlapping key identifiers.
- n Use exact record linkage, or probabilistic record linkage, to match sampled records to the database (compute probabilities of match for each record).
- n Count of percentage of correct matches.
- n Can be adapted for masked data.

Reiter J. (2005). Estimating risks of identification disclosure for microdata. *Journal of the American Statistical Association*.



Perceived identification disclosure

- n Almost never done.
- n Lambert (1993) suggests taking maximum of match probabilities as perceived disclosure risk measure.

Lambert D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*.



Attribute and inferential disclosure risks

- n Measures are specific to attribute.
- n Given identification, for attribute values considered at risk: compute distance between released/estimated attribute values and true attribute values for those considered at risk.
- n Set minimum threshold.



Special concerns for biomedical/genetic data

- n Geographic information may be known even when not released (e.g., Framingham study).
- n Knowledge of who is in the data may be widespread.
- n Genetic information is identifying when known by others.
- n Linked data: safe in one data set may not be safe after linkage.
- n Potential harm from disclosures of biomedical/genetic information.