Appendix

Derivation of formulae for projection to larger database

Let *h* be the probability to link a random pair of records erroneously. In a database with N records there are *N*(*N*-1)/2 possible pairs of records. Therefore an upper bound for the number of erroneously linked pairs of records is given by h*N(N-1)/2 and hence an upper bound for the homonym errors rate by h*(N-1)/2. From this we can estimate *h* given the observed number of homonyms N_H: $H=N_H/N_G \approx h*N(N-1)/2$ or $h \approx 2\, N_H/(\, N_G(N_G-1))$ where N_G is the number of persons in the gold standard.

Let *s* be the probability that an error occurs in the data that precludes linkage of two records referring to the same person. Assuming that such errors are independent and the probability that the same error occurs twice can be neglected the expected number of synonyms is given by $\mathrm{E}\left(N_S\right)=sN-\sum_{i\geq 1} n_{Gi}s^i$ (8) where n_{Gi} is the number of cases with exactly I notifications in the gold standard.

With q_{Gi}, the proportion of cases with exactly *i* notifications we get $\mathrm{E}\left(N_S\right)=sN_G\sum_{i\geq 1} q_i\left(i-s^{i-1}\right)$. The synonym rates S_1 and S_2 can be expressed as

$$S_1=\frac{N_S}{N_G}=\frac{sN-\sum_{i\geq 1} n_i s^i}{N_G}=\frac{sN_G\sum_{i\geq 1} q_i\left(i-s^{i-1}\right)}{N_G}=s\sum_{i\geq 1} q_i\left(i-s^{i-1}\right)=s\sum_{i\geq 2} q_i\left(i-s^{i-1}\right),$$

$$S_2=\frac{N_S}{\sum_{i\geq 2} n_i}=\frac{sN_G\sum_{i\geq 1} q_i\left(i-s^{i-1}\right)}{N_G\sum_{i\geq 2} q_i}=\frac{s\sum_{i\geq 2} q_i\left(i-s^{i-1}\right)}{\sum_{i\geq 2} q_i}=\frac{s\sum_{i\geq 2} q_i\left(i-s^{i-1}\right)}{1-q_1}.$$

From this we can iteratively estimate s given the observed number of synonyms and the q_i's. Once *h* and *s* have been determined it is possible to project error rates to larger databases and notification schemes with different proportions of multiple notifications.

Attachment 1 to: Schmidtmann I, Sariyar M, Borg A, Gerold-Ay A, Heidinger O, Hense HW, Krieg V, Hammer GP. Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data. GMS Med Inform Biom Epidemiol. 2016;12(1):Doc02. DOI: 10.3205/mibe000164. Available from: http://www.egms.de/en/journals/mibe/2016-12/mibe000163.shtml