Step1. Optimize pedigree information
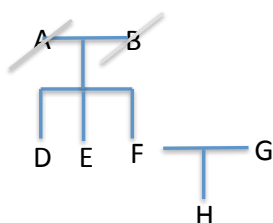
The Lifelines family database is based on information provided in the municipal registries (in Dutch Gemeenteeliike Basisadministratie (GBA)) and connects two relatives only if the relatives they have in common are also LifeLines volunteers. For example, in the figure below, samples D, E and F are siblings, but because their parents are not in Lifelines, they are listed in the database as being unrelated (i.e. having not other relatives in Lifelines).

Real scenario
one family, 5 members

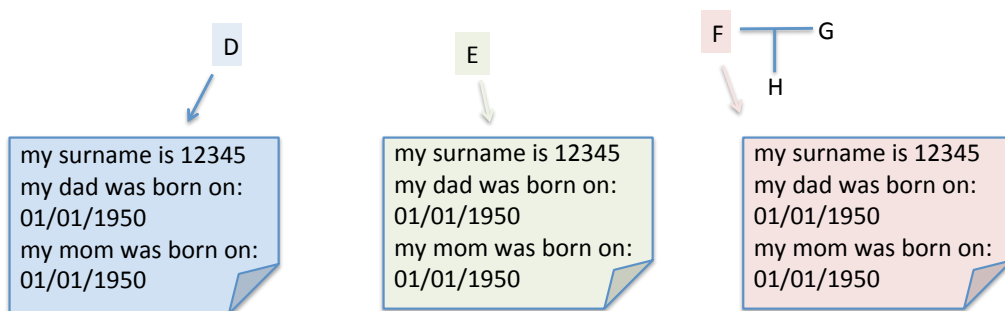*Current information from LifeLines*
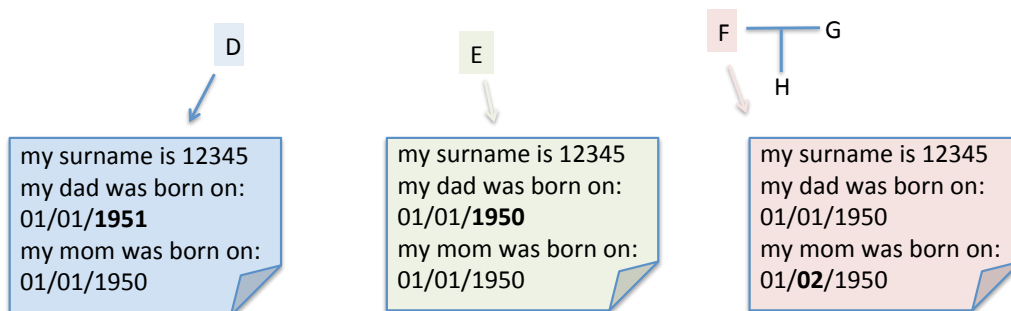two singletons and one family of 3 members

We used reported information from each volunteer to reconnect siblings. In fact, each adult volunteer (18 years or older) has reported the date of birth of his/her parents in the baseline questionnaire. In addition, LifeLines provided us the coded surname. We have thus searched for individuals with the same coded surname who reported the exact same date of birth for their parents (see figure).

Current information from LifeLines

two singletons and one family of 3 members

D
my surname is 12345
my dad was born on:
01/01/1950
my mom was born on:
01/01/1950

E
my surname is 12345
my dad was born on:
01/01/1950
my mom was born on:
01/01/1950

F   G
H
my surname is 12345
my dad was born on:
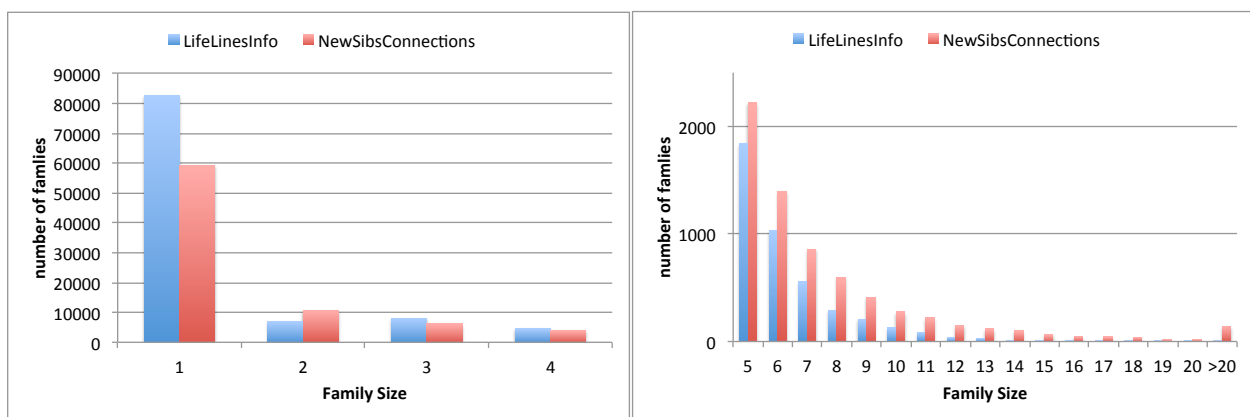01/01/1950
my mom was born on:
01/01/1950

Since we expect that some volunteers may not remember the exact date of birth of their parents, or that handwritten dates may be sometimes ambiguous (for example 1 and 7, 4 and 9), we also searched for pairs of individuals with one mismatch in reported date (day, month or year). See figure:

D
my surname is 12345
my dad was born on:
01/01/**1951**
my mom was born on:
01/01/1950

E
my surname is 12345
my dad was born on:
01/01/**1950**
my mom was born on:
01/01/1950

F   G
H
my surname is 12345
my dad was born on:
01/01/1950
my mom was born on:
01/**02**/1950

As additional condition to declare a sib-sib relation valid, we required the date of birth of two siblings to be at least 24 weeks apart or less than 2 days (twins can be born with one day difference).

With this procedure, we were able to re-allocate ~24,000 singletons (individuals for which there was not indication of other relatives in LifeLines) to other families, as well as merge small families to form bigger ones. Below is a bar plot that show the distribution of families by size, when families are reconstructed using only GBA familial relations as reported in the LifeLines database (blue) or when, additional sib-sib relations are detected using the reported parents' birth dates & surname. Of note, a family of size 2 can indicate 2 siblings or one parent and one child. A couple without children is not considered a family of size 2 but rather as 2 families of size 1. The largest family connects 71 LifeLines volunteers.

And here is a table that summarizes family composition when this step is completed:

Total Number of Singletons: 59358
Total Number of Families (size >=2): 27660 *(note: a couple without children, siblings or parents is not considered a family of 2). Summary:*

- **Family size (Number of LL participants):**

Average: 3.91 (2 to 71)
Distribution: 2 (38.6), 3 (22.3%) and 4 (14.7%)

- **Generations:**

Average: 2.37 (to 5)
Distribution: 2 (67.5%), 3 (28.4%) and 4 (4.1%)