**Supplementary Material**

**TITLE: Overview of National Health Laboratory Service Corporate Data Warehouse Patient Linking Algorithm**

The aim of the patient linking process is to identify specimens that were taken from the same individual and link them together by assigning a unique patient id to all these episodes. The linking process uses as its basic principle the following assumption: From the demographic attributes available from the laboratory database, an individual is distinguished from another based on the following attributes: (1) surname, (2) first name, and (3) birthdate.

The linking steps are as follows; the data cleaning and preparation stage, the exact linking step and the fuzzy linking step.

*Data cleaning and preparation stage*

The record received from the laboratory information system is cleaned and loaded to a raw demographics folder on the corporate data warehouse (CDW) system. The cleaning process includes trimming of blank spaces, removal of titles in name fields, removal of special characters, splitting of multiple given names in one variable into multiple variables, joining of surnames with prefixes into one name, converting all string variables to lower case and standardization of date formats.

*Exact linkage stage*

If the raw records have valid values for first name, surname and date of birth these attributes are matched against the values for existing unique patients in an archived folder with existing patients and if matched the unique patient identifier (id) number is assigned to the specimen. After this stage all raw records that have not been assigned an unique patient id number and have valid values for first name, surname and date of birth are assigned a new unique patient id number and records with the same first name, surname and date of birth are assigned the same unique patient id number. Lastly, all remaining raw records will be records for which insufficient data exists for any type of meaningful linking to be possible. Each of these is assigned its own unique patient id as it is impossible for the process to determine if this relates to another specimen or not.

*Fuzzy linkage stage*

The fuzzy linking process uses as input the records generated by the last run on the exact linking process. The purpose of the fuzzy linking is to attempt to get a better level of linking by accommodating different spelling of names and surnames due to factors such as typing errors or complex spelling of surnames etc. In order to achieve this, rows of data are compared with each other and using the fuzzy algorithm to determine the probability of the same values from the 2 rows being the same, these probabilities are then weighted and the match is scored. Records that score above a certain cut-off level are then assumed to be the same and linked together. Since name and surname is considered in exact matching, exact matches on these attributes are excluded in fuzzy matching. The probability score is calculated for first name, surname and date\_ of\_ birth. The probability score is weighted as follows: First Name -= 40%; Surname =40% and Date of birth -= 20%. If probability multiplied by weights is greater than 90% (cut-off level) then the record linked. All records that score a 90% or above match are extracted as matches to be resolved into a single hierarchy of linked demographics. This scoring and extraction of matched values results in a grid of unique patients ids that are matched to each other. This is resolved into a single hierarchy resolving all values to a single unique patient id so that this id can be assigned to the other episodes to effectively link them together.

The detailed linking algorithm is available by request from the National Health Laboratory Service, South Africa. The patient linking process in its current release represents the best that has been achieved programmatically to date given the constraints involved such as volume of data, and especially the quality of data. It has been designed to favour linking in cases where concise attributes that reasonably should identify a patient such as the national id number is utilized. A conservative approach was used to rather underlink than overlink. However due to factors such as the ability and limitations of the probability algorithm both overlinking and underlinking is possible.

References

Jaco Grobler, NHLS Corporate Data Warehouse Patient Linking Technical Document, 2014.