



DE-IDENTIFICATION OF PERSONAL DATA

1 June 2015

1. Introduction

In order for the Health Market Inquiry (the “Inquiry”) to carry out an analysis of competition in several markets falling within its scope, a large amount of transactional data from a number of service providers needs to be processed. It is the intention of the Inquiry to keep to the minimum the processing of personal information that may be contained within the relevant transactional data. Where, consistent with maintaining the integrity of the data and the effectiveness of its analysis, the Inquiry is able to receive any relevant transactional data in a format from which identifiable personal information has already been eliminated, that route will be preferred.

For the purpose of minimising the receipt and processing of identifiable personal information by the Inquiry, a procedure of “de-identification” of data has been developed.

This procedure will allow healthcare providers to submit patient data to the Inquiry in such a format that individuals’ personal identities (“Personal Individual Information”) and residential addresses (“Personal Address Location”) will be “de-identified”, while keeping each individual patient’s records distinct for analytical purposes.

There are two scenarios to be catered for. One is where the service provider wishes to de-identify their own data before submission to the Inquiry, with the necessary assistance and auditing of the process by the Inquiry. The other is where the service provider wishes the Inquiry to de-identify the data after submission.

For both scenarios, a software package (the “De-Identification Tool”) has been developed to achieve the task of removing personal identifiers in the data. The Inquiry will provide the De-Identification Tool to the relevant service providers by arrangement, or will implement it itself as the case may be.

The process is explained as follows:

2. De-Identification of Personal Address Location

The De-Identification Tool contains an address table with all known South African street addresses which are aggregated to a relevant census enumerator area code (“EA”). EA’s are areas compiled by Statistics South Africa for National Census and de-identification purposes which represent homogeneous areas on average of approximately 150 to 250 households per EA. *(In instances where an EA contain less than the required number of households, neighbouring EA’s are grouped together and referred to as Small Areas. For the majority of EA’s, the Small Areas are the same, and for purposes of this Inquiry, Small Area boundaries will actually be used. The concept of EAs is however more commonly understood – and this term will thus be used.)*

An illustrative example follows:

Data Reference Table:

No	Street Number	Street Name	Suburb	Town Boundary	Province	EA_CODE
1	15	Sonja	Moreleta Park	Tshwane	Gauteng	79910906
2	33	Edwin	Moreleta Park	Tshwane	Gauteng	79910906
3	64	Melissa	Moreleta Park	Tshwane	Gauteng	79910906
4	50	Vincent	Moreleta Park	Tshwane	Gauteng	79910906

Service Provider Patient Address Table:

RECORD	NUMBER	STREET	SUBURB	TOWN	PROVINCE
1		15 Sonja	Moreleta	Tshwane	Gauteng
2		33 Edwin	Moreleta	Tshwane	Gauteng

Search and Replace all Service Provider Address with EA_CODE:

No	Street Number	Street Name	Suburb	Town Boundary	Province	EA_CODE
1	15	Sonja	Moreleta Park	Tshwane	Gauteng	79910906
2	33	Edwin	Moreleta Park	Tshwane	Gauteng	79910906
3	64	Melissa	Moreleta Park	Tshwane	Gauteng	79910906
4	50	Vincent	Moreleta Park	Tshwane	Gauteng	79910906

Once the Personal Address Location has been de-identified in this way, it will not be possible to re-identify that location.

3. De-Identification of Personal Individual Information:

The second part of the de-identification procedure is to remove the individual's name and identity details from the relevant transactional data. The De-Identification Tool will do a #Hash key conversion (non-standard) on the patient's personal information to de-identify patient identifiers. This will de-identify patient identifiers in such a manner as to allow for pattern matching between different service providers whilst ensuring individual anonymity. E.g. unique codes will only be generated for unique clients. The same client in two different databases will have the same unique code. If necessary, the Medical Aid number, if available, can be included in the #Hash key conversion, being an alternative to ID or merely added as another field, so as to ensure complete de-identification and remove the possibility of later re-identification.

Run DE-ID #Hash key generator:

Name	Surname	ID	#HASHKEY
W	Ximiya	23123121233	#673721
D	Snyman	12312123112	#871112
E	Magobe	23532432423	#918771

Following this step the service provider can provide the Inquiry with de-Identified data:

#HASHKEY	EA_CODE	OTHER REQUESTED FIELDS
#673721	79910906	ASNDASNDADAD
#871112	79910906	ADADASDADAAS
#918771	79910906	DADSADADADASDA

For those service providers which do not have the technical capacity to run the De-Identification Tool themselves, this function can alternatively be performed by the Inquiry.

4. Data verification and quality

The Inquiry proposes a two-way communication process between the Inquiry and the service provider with regards to data verification and data quality. This means that data will either be received in a de-identified state or still-required-to-be de-identified state. From here the data will be processed according to its relevance to the Inquiry. Access by the service provider to its de-identified data will be allowed so as to enable the service provider to verify and validate the results of the de-identification process.

If any service provider wishes to engage with the Inquiry on the above process, please contact the Inquiry Director.

Graphic Presentation of Process:

