# Requirements and Process for Integrating Data Analyses into MPI – Shaun Grannis

Below is a more detailed description of the requirements and process for the data analyses performed as part of a process to integrate into a master person Index:

1. A delimited dataset (ideally a pipe-delimited flat file) containing either a] (preferably) the entire set of identities in a cohort be matched, or b] a valid representative random sample of the cohort to be matched. Records in the data set should contain all fields that will be available to the MPI matching algorithm.

2. To aid in analyzing the data set, a description of each field (field definitions), and where possible, a description of the expectations (validation rules) should be included for each field. Hypothetical examples (for illustration purposes only) might include:

   National ID - definition: A nationally unique identifier assigned by the government at age 16. The National ID field should contain only digits 1-9, no alphabetic characters. Null values are permitted and are represented by '9999999999'
   Surname - definition: A name shared in common to identify the members of a family, as distinguished from each member's given name. The surname field should contain only alphabetic characters; hyphens and singe-quotes are permitted. Null values are not permitted.
   Gender - definition: Sexual identity indicating either female, male, or other. Valid values for the Gender field are limited to 'M' (male), 'F' (female), 'O' (other) and '' (Null).

3. Prior to performing a statistical analysis, we apply preprocessing and validation rules to the data set. We then compare the preprocessed data set to the original (typically using Unix utility "diff") to discover differences between the two datasets highlighting patterns that may indicate invalid assumptions about the data, or previously undiscovered vagaries within the data. Sample preprocessing rules are illustrated in the "preprocess.pl" routine attached.

4. The preprocessed data is then analyzed using the "field_metrics.R" package (attached), written for "R" - an open-source statistical analysis tool.
   The field metrics package calculates important informational characteristics of fields to be used for matching. In composite, these metrics help identify data fields that a) guide effective and efficient candidate pair selection, and b) inform matching algorithm configuration. Metrics include the following:

   - information entropy
   - percent of maximum entropy
   - token collision rates
   - average token frequencies
   - null rates
   - blocking efficiency

5. Using information gained from steps 3 and 4, we can a] more effectively identify and address any undiscovered data issues that may hinder accurate matching, and b] more efficiently configure the matching algorithm by identifying optimal fields for matching and candidate pair selection.

My understanding is that we will pre-populate the master person index with a database at the ministry of health, so we should analyze that. Additionally, since data captured in the OpenMRS installations will be used to identify patients, extracts from the OpenMRS pilot sites should be analyzed as well.