Exercise 2: Repeated records and misclassification of TB records
Impact on tuberculosis notification rates of removal of repeat notification records: a record linkage study in Brazil

Objectives:

• To understand the value of identifying repeated records concerning the same patient in the TB routine notification system

• To understand how it is possible to correct the misclassified type of entry status of TB records in the routine notification system, by comparing data from repeated consecutive records for the same patient

• To calculate the difference in the number of new notified cases of tuberculosis (TB) when repeated records are removed

• To identify possible reasons why duplicate or misclassified records may exist in the TB notification system of your country

• To discuss possible methods to assess the number of duplicate or misclassified TB records, in your country

Setting:

• Brazil has an electronic, case-based TB notification system since the beginning of the 90s.

• Records available at national level include several variables that allow clear/unique identification of individual patients (e.g. name, date of birth, ..), as well as information about where and when the patient was diagnosed and notified to the system, where the patient was treated, the clinical form(s) of TB and laboratory results, treatment outcome and when treatment ended.

• Names of states, cities and health care units are recorded using standardized codes.

• Duplicate or misclassified records may exist in the TB notification system if there are problems with data entry and data management, and when cases are not correctly categorized (e.g. when a re-treatment case is incorrectly classified as a new case).

• Although available in the system, mechanisms to avoid double-counting cases within and across health care units were known to be inadequate.

• In 2005, WHO estimated that Brazil had a smear positive TB case detection rate within DOTS areas of 64%.
Methods:

- Data from the Brazilian Information System for Tuberculosis Notification from 2000 to 2006.
- Repeated records were identified through electronic record linkage, using the variables name of patient, name of mother and date of birth as "matching variables", and city and address of residence and health care unit as "checking variables".
- Repeated records were put in chronological order and classified into six mutually exclusive categories, based on comparing the values of key variables belonging to two consecutive records. The exclusive categories were: true duplicate, return after default, relapse after cure, transfer between health care units, inconclusive, and impossible to classify due to missing data.
- Depending on the classification, repeated records were kept, combined or removed from the database. Figure 1 depicts what to do with possible duplications, according to the Brazilian recording and reporting system.

Figure 1. Flow diagram. What to do in case of possible repetitions?
Exercise:

1. Open the Excel file "Brazilian_repeated_records.xls" and read the contents of the sheet "Dictionary" to become familiar with the variables names and values.

2. In the sheet "Data" you will find the repeated records of 12 patients\(^1\) registered in the Brazilian Information System for Tuberculosis Notification from 2000 to 2006. Repeated records are sorted by notification date and diagnosis date.

3. Compare the values of the variables name of patient, name of mother, date of birth, city of residence, address of residence and health care unit to verify whether you agree that these repeated records really belong to the same people.

4. Complete the values of the variable "classification" by classifying the repeated records into six mutually exclusive categories (do not classify the first one, just from the second record onwards): true duplicate, return after default, relapse after cure, transfer between health care units, inconclusive, and impossible to classify due to missing data. In order to do this, you will need to compare the values of key variables belonging to two consecutive records (e.g. records 1 with record 2, record 2 with record 3, and so on). The most important variables to be considered are: notification date, diagnosis date, notification city, health care unit, type of entry of the actual record being classified and treatment outcome of the record immediately prior to the one being classified.

   - 'True duplicates' are repeated records with the same information for the variables notification city and health care unit code. Usually the notification number and the notification dates are the same, but occasionally there may be differences. In any case, notification dates should not be more than 60 days apart. After 60 days, it is more likely that the patient has defaulted from the initial treatment and has returned to the same unit to initiate a new treatment, and therefore he/she should be newly notified as such;

   - 'Relapses' are records where the treatment outcome of the record immediately prior to the one being classified indicates that the patient has been cured and/or the type of entry of the record being classified indicates that the patient has relapsed;

   - 'Returns after default' are repeated records where the variables treatment outcome of the record immediately prior to the one being classified indicates that the patient has defaulted and/or the type of entry of the actual record being classified indicates that the patient has returned after default;

   - 'Transfers' are repeated records notified by different health care units, with treatment outcome of the record immediately prior to the one being classified, and/or the type of entry of the actual record, indicating that the case was transferred;

---

\(^1\) To ensure confidentiality, the names of the patients and their addresses have been modified for the purposes of this exercise.
• 'Inconclusive' are repeated records that could not be classified, even when data for key variables was available;

• 'Impossible to classify due to missing data' are repeated records with missing information for key variables: notification date and/or type of entry and/or treatment outcome.

5. Once you have classified the records of all patients, complete the values of the variable "action" into three mutually exclusive categories: keep, combine and delete. You should keep all records classified as return after default, relapse after cure, inconclusive, and impossible to classify due to missing data; you should delete all records classified as true duplicate; and you should combine all records classified as transfer between health care units.

6. Calculate the difference in the number of new notified cases of tuberculosis (TB) between the original values of the variable "type of entry" and the revised values after removal of improper repeated records in a sample of the TB notification system. If this difference was representative of the entire notification system, how important do you think misclassification of patients was in the notification system in Brazil? Would this make a difference to the estimated case detection rate?

7. Suggest possible reasons why duplicate or misclassified records may exist in the TB notification system of your country.

8. Discuss how you could assess the number of duplicate or misclassified TB records in the TB routine notification system, of your own country.

Reference:
Variables in the electronic file (Dictionary):

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>Identity of repeated records</td>
</tr>
<tr>
<td>N_Id</td>
<td>Number of repeated records</td>
</tr>
<tr>
<td>Order_Id</td>
<td>Chronologic order of repeated records Sorted by dt_notif, dt_diag, nu_notif</td>
</tr>
<tr>
<td>name_patient</td>
<td>Name of the patient</td>
</tr>
<tr>
<td>name_mother</td>
<td>Name of the mother</td>
</tr>
<tr>
<td>dt_birth</td>
<td>Date of birth</td>
</tr>
<tr>
<td>State resid</td>
<td>State where patient lives</td>
</tr>
<tr>
<td>city resid</td>
<td>City where patient lives</td>
</tr>
<tr>
<td>address</td>
<td>Address where patient lives</td>
</tr>
<tr>
<td>age</td>
<td>Age</td>
</tr>
<tr>
<td>sex</td>
<td>Sex</td>
</tr>
<tr>
<td>nu_notif</td>
<td>Notification number</td>
</tr>
<tr>
<td>dt_notif</td>
<td>Date of notification</td>
</tr>
<tr>
<td>code HCF</td>
<td>Health Care facility code</td>
</tr>
<tr>
<td>state notif</td>
<td>State where patient was notified</td>
</tr>
<tr>
<td>city notif</td>
<td>City where patient was notified</td>
</tr>
<tr>
<td>dt_diag</td>
<td>Date of diagnosis</td>
</tr>
<tr>
<td>dt_start treatm</td>
<td>Date in which current treatment begun</td>
</tr>
<tr>
<td>dt_end treatm</td>
<td>Date in which current treatment ended</td>
</tr>
<tr>
<td>type entry</td>
<td>Type of entry</td>
</tr>
<tr>
<td>smear_1</td>
<td>Smear results at diagnosis</td>
</tr>
<tr>
<td>Clinical form</td>
<td>TB form at diagnosis</td>
</tr>
<tr>
<td>HIV</td>
<td>HIV results</td>
</tr>
<tr>
<td>outcome</td>
<td>Treatment outcome</td>
</tr>
</tbody>
</table>

- **type entry**
  - 1. New case
  - 2. Relapse
  - 3. Return after default
  - 4. Don’t know
  - 5. Transfer-in

- **smear_1**
  - 1. positive
  - 2. negative
  - 3. not done

- **Clinical form**
  - 1. pulmonar
  - 2. extrapulmonar
  - 3. pulmonar + extrapulmonar

- **HIV**
  - 1. positive
  - 2. negative
  - 3. pending results
  - 4. not done

- **outcome**
  - 1. Cure
  - 2. Default
  - 3. Death
  - 4. Transfer-out
  - 5. Change of diagnosis
  - 6. TB-MDR