Rev Saúde Pública 2007;41(Supl. 1)

Ana Luiza Bierrenbach[I]

Antony Peter Stevens[I]

Adriana Bacelar Ferreira Gomes[I]

Elza Ferreira Noronha[II]

Ruth Glatt[I]

Carolina Novaes Carvalho[I]

João Gregório de Oliveira Junior[I]

Maria de Fátima Marinho de Souza[I]

Impact on tuberculosis incidence rates of removal of repeat notification records

ABSTRACT

OBJECTIVE: To evaluate the impact on tuberculosis (TB) incidence rates of removal of improper duplicate records from the notification system.

METHODS: Data from the Sistema de Informação de Agravos de Notificação (Brazilian Information System for Tuberculosis Notification) from 2000 to 2004 were analyzed. Repeat records were identified through probabilistic record linkage and classified into six mutually exclusive categories and then kept, combined or removed from database.

RESULTS: Of all TB records, 73.7% had no duplicate, 18.9% were duplicate, 4.7% were triplicate, and 2.7% were quadruplicate or more. Of all repeat records, 47.3% were classified as transfer in/out; 23.6% return after default, 16.4% true duplicates, 10% relapse, 2.5% inconclusive and 0.2% had missing data. These proportions were different in Brazilian states. Removal of improper duplicate records reduced TB incidence rate per 100.000 inhabitants by 6.1% in the year 2000 (from 44 to 41.3), 8.3% in 2001 (from 44.5 to 40.8), 9.4% in 2002 (from 45.8 to 41.5), 9.2% in 2003 (from 46.9 to 42.6) and 8.4% in 2004 (from 45.4 to 41.6).

CONCLUSIONS: The study results indicate that the observed tuberculosis incidence rates represent estimates that would be closer to the actual rates than those obtained from the raw database at state and country level. The use of record linkage approach should be promoted for better quality of notification system data.

KEY WORDS: Tuberculosis, epidemiology. Disease Notification. Diseases registries. Data sources. Information Systems. Brazil.

INTRODUCTION

The *Sistema de Informação de Agravos de Notificação* (SINAN – Brazilian Information System for Disease Notification) collects and processes data on compulsory disease notification nationwide.* Improper repeat records in health information systems jeopardize correct interpretation of epidemiological surveillance data.

Repeat notification of chronic diseases such as tuberculosis (TB) can be attributed to data entry or processing errors. Also, the same patient can be reported repeated times by different health units due to authorized or voluntary transfers between units during treatment or different treatments due to relapse after cure or return after default.** Although they concern to the same patient, relapses and

[I] Secretaria de Vigilância em Saúde. Ministério da Saúde. Brasília, DF, Brasil

[II] Faculdade de Medicina. Universidade de Brasília. Brasília, DF, Brasil

Correspondence:
Ana L Bierrenbach
Esplanada dos Ministérios, Bloco G
Edifício Sede, 1º andar, sala 150
70058-900 Brasília, DF, Brasil
Telefone: 061-33153496.
E-mail: ana.bierrenbach@saude.gov.br

Received: 1/30/2007
Approved: 3/27/2007

* Ministério da Saúde. Secretaria de Vigilância em Saúde. Sistema de Informações de Agravos de Notificação Normas e Rotinas. Brasília; 2004. (Série A: normas e manuais técnicos)
** Ministério da Saúde. Fundação Nacional de Saúde. Tuberculose - Guia de Vigilância Epidemiológica. Brasília; 2002

returns are considered valid entries in this database as they are new TB episodes. But all other repeat records must be removed.

The objetive of the present study was to assess the impact on TB incidence rates of removal of repeat improper records from the notification system.

METHODS

Nationwide TB notification records for the period 2000–2004 were studied. Data provided by health departments at state level were made available by SINAN-TB National Management on February 2006.

The following steps were taken to identify repeat records: 1) database pre-processing; 2) identification of matched records (matches) using record linkage Link-Plus software; 3) ascertainment whether matched records concerned the same patient (links); 4) post-processing with regrouping of records concerning the same patient. Linked records concerning the same patient were considered repeat records.

In database pre-processing content of variables "patient's name" and "patient's mother name" were corrected aiming to increase the likelihood of finding matched records. These procedures included: 1) correction of obvious typing errors; 2) elimination or replacement of special characters (%, /); 3) capitalization of names; 4) removal of any individual letters or prepositions from names, 5) removal of terms indicating lack of information on patient's name and patient's mother name (e.g. don't know, unknown).

Matched records were identified using the record linkage Link-Plus software (CDC, Atlanta, Georgia, USA)* through probabilistic search for repeat records. The probabilistic record linkage (PRL), developed by Fellegi & Sunter,[2] allowed to estimate the likelihood of agreement and disagreement of variables selected for record linkage (linkage variables).

The software was set up to search for repeat records. Variables such as "patient's name," "mother's name" and "date of birth" were included as matching variables. The variable "gender" was selected as blockage variable, i.e., a variable used for separating the file into smaller blocks to speed up linkage process.

Probabilities in the linkage process were obtained through an indirect approach, i.e., probability estimates were determined by the selection of records in SINAN-TB database undergoing linkage. Default probabilities or probabilities preset by the investigators were not used.

Link-Plus software estimates scores for each pair of matched records. The higher the score, the more likely a matched pair concerns the same patient. Scores above the set cutoff value are considered repeat records and score below the cutoff value are considered single records. A cutoff value of six was set. When linkage process is complete, reports with listings of pairs of matched records and single records are issued.

Three successive manual removals were conducted to ascertain whether pairs of matched records concerned the same patient, in which case they were called linked or repeat records. Those pairs with records that did not concern the same patient were broken down based on a set of information and criteria. For example, common misreporting of date of birth evidenced by inconsistencies between date of birth and patient's age. Records with inconsistent dates of birth have low negative predictive value in the ascertainment of a pair of linked records concerning the same patient while consistent dates have high positive predictive value. Investigators' knowledge on the composition of Brazilian proper names was also applied. For example, family customs of giving similar names to their children allowed, using Link-Plus program, to identify records of likely siblings as concerning to the same patient, and break them down during manual removal treatment. For uncertain cases, the investigators chose to take a conservative approach and not consider matched records as repeat.

The first two rounds of removal were based only on linkage variables and program scores. The third round of removal was carried out after regrouping of repeat records and other linkage variables were compared, such as municipality and notifying health unit or municipality and home address. In all steps, program scores helped to determine records requiring careful consideration during removal.

Link-Plus software yields results as paired records but some records are transitively paired. According to the transitive logic, if record A is associated to records B and C, then records B and C are also necessarily associated. Thus, records A, B and C were regrouped as a record triplet concerning the same patient even if records A and C had not been matched by the record linkage program.

As a result of this process, groups of three, four or more records were considered as concerning to one patient. The largest group of repeat records concerning the same patient included 15 records.

In the last step, records were classified as single (one notification and no repeat), duplicate (one notification

* Centers for Disease Control and Prevention. Link Plus fact sheet. Atlanta: 2004 [access on: Sept 02, 2005]. Available from: http://ftp.cdc.gov/pub/Software/RegistryPlus/Link_Plus/Link%20Plus.htm

and one repeat), triplicate (one notification and two repeats), and so on.

For the classification of repeat records, the following variables values were compared: notification number, date of notification, date of diagnosis, date of current notification, date of current treatment start, date of notification completion, code of notifying municipality, code of notifying health unit, code of health unit proving patient follow-up, type of system entry, TB clinical form and status at completion.

Repeat records were classified in six mutually exclusive categories as follows:

- Missing data: repeat records with missing information for variables "date of notification" and/or "type of system entry" and/or "code of notifying health unit".

- True duplication: repeat records with the same (but no missing) information for the variable "code of notifying municipality" and the same date of notification or time interval up to 60 days between notifications and were from the same notifying health unit. As there could have been concurrent use of two different charts for coding health units, records were considered from the same health unit if they had the same code or a corresponding code for both charts. All states were asked to provide their plan of health unit code change but only some of them provided it timely to be included in the study.

- Relapse: repeat records where categories in the variables related to "type of system entry" and/or "status at completion" indicated prior cure.

- Return: repeat records where categories in the variables related to "type of system entry" and/or "status at completion" indicated prior default.

- Transfer between health units: repeat records notified by different health units with information in the variables related to "type of system entry" and/or "status at completion" indicating case transfer. Repeat records that, although with same (or corresponding) codes for notifying health unit, showed different code for health unit providing patient follow-up were also classified as transfer between health units.

- Inconclusive: classification was not possible even though variables did not have any missing information.

Repeat records classified as "transfer between health units" were grouped as within municipalities, when the notifying health units belonged to the same municipality; between municipalities, when they were from different municipalities but within the same state; and inter-state when they were from different states.

Score comparison and classification were carried out using Stata 8.2 software.

After classification, repeat records were then either excluded or remained in the database following SINAN working guidelines. Hence, records classified as relapse, return, and inconclusive remained in the database. For "true duplication," the oldest record (or the most complete one, if both had same date of notification) was left in the database. For "transfer between health units," notification form information of the oldest record was joined to follow-up form information of the most recent record.* A database was defined as "complete" when it included all notified records and "lean" when it included non-excluded records only.

Following SINAN guidelines for epidemiological surveillance actions,** a new TB case was defined when: 1) any notification where the variable "system entry" reported "new case" or "don't know"; 2) the variable "status at completion" was left blank in the category "diagnosis change".

TB incidence rates were estimated as the number of new cases living in a given area diagnosed in a given year, divided by the population living in this area in the same year and multiplied by 100,000. Population-based data were provided by the *Instituto Brasileiro de Geografia e Estatística* (Brazilian Institute of Geography and Statistics – IBGE).***

RESULTS

TB notification database for the period 2000–2004 included 482,501 records comprising all types of system entries and all TB clinical forms. Of these, more than 70% were single records and no clear trend was seen in single, duplicate, triplicate, and quadruplicate or more records (Table 1). In all Brazilian regions, the proportion of single, duplicate, triplicate and quadruplicate or more records did not vary much over the years studied but it varied widely in some states.

Table 2 shows that, in 2003, states with the lowest and highest rates of single records were Goiás (21.1%) and Roraima (86.9%), respectively.

Table 3 displays the annual proportions in the six repeat record classifications. "Transfers between health units"

* Ministério da Saúde. Fundação Nacional de Saúde. Tuberculose - Guia de Vigilância Epidemiológica. Brasília; 2002
** Ministério da Saúde. Secretaria de Vigilância em Saúde. Sistema de Informações de Agravos de Notificação Normas e Rotinas. Brasília; 2004. (Série A: normas e manuais técnicos)
*** Departamento de Informática do Sistema Único de Saúde. Informações de saúde: demográficas e socioeconômicas. Brasília; 2005. [Access on Sept 2, 2005]. Available from: http://w3.datasus.gov.br/datasus/datasus.php?area=359A1B379C6D0E0F359G23HIJd6L26M0N&VInclude=../site/infsaude.php

Table 1. Number of records for each patient in *Sistema de Informação de Agravos de Notificação - Tuberculose,* by year of notification. Brazil, 2000–2004.

Year	Number of records for each patient								Total
	Single		Duplicate		Triplicate		Quadruplicate or more		
	N	%	N	%	N	%	N	%	
2000	70,151	77.9	14,911	16.6	3,189	3.5	1,795	2.0	90,046
2001	68,975	74.2	17,071	18.3	4,353	4.7	2,620	2.8	93,019
2002	70,491	71.8	19,377	19.7	5,160	5.3	3,116	3.2	98,144
2003	72,468	71.4	20,577	20.3	5,399	5.3	3,054	3.0	101,498
2004	73,259	73.4	19,422	19.5	4,625	4.6	2,488	2,5	99,794
Total	355,344	73.7	91,358	18.9	22,726	4.7	13,073	2.7	482,501

Source: Brazilian Information System/Health Surveillance Department/Brazilian Ministry of Health (SINAN/SVS/MS)

was the most prevalent category in the study period, accounting for 55.4% of all repeat records in the first year and then remaining around 47% in subsequent years. There were 12% of returns in 2000 and then they remained constant around 25%. Overall, true duplications decreased and relapses increased over the period studied.

Of all 32,341 repeat records classified as "transfers between health units," 40.4% were within municipality; 47.8% between municipalities; and 11.8% between states.

Table 4 shows the classification of repeat records notified in 2003 by regions and states. Different proportions in each classification were found between states of the same region. Although some states had a small number of repeat records, Roraima, Amazonas and Amapá had the highest rates of transfers between health units, while Acre had the lowest rate. In Goiás, true duplication accounted for 74% of repeat records, more than twice the proportion found in Paraíba, ranked second in this category.

Table 5 shows a comparison of annual TB incidence rates between complete and lean databases, i.e., before and after removal of duplicate records and joining of transferred cases. With rare exceptions, different annual TB incidence rates were found in all states over the period studied. Differences were greater than 10% in at least one year in the states of Amapá, Goiás, Paraíba, Piauí, Rio Grande do Norte, São Paulo and Tocantins. Goiás showed a difference higher than 34% in all years studied. Nationwide, the observed incidence rates varied in the different databases, from 6.1% in 2000 to 9.4% in 2002 with no clear trend. Table 5 also shows rate differences between regions and states over the years that cannot be attributed to repeat records in database but this analysis is beyond the scope of this study.

DISCUSSION

SINAN was created in the beginning of 1990s and has undergone several updates to eliminate errors and make it more suitable to meet new demands in epidemiological surveillance. Although all Brazilian municipalities pass on their information to SINAN, around 70% carry out direct entry of electronic data. Database update at higher hierarchical levels is routinely conducted through vertical data transfers. Working guidelines and task description at local, state and country levels are regulated in official documents available to all users.*

In accordance with epidemiological surveillance guidelines, SINAN has implemented specific routine procedures for managing repeat TB patient records and has its own tools to help identification of potential duplicates as well as correction procedures. However, given the number of repeat records found in SINAN-TB database, these routine procedures are possibly not implemented as necessary and/or not adequately followed by system users, especially at local level. Implementation of routine procedures is a priority action that should be taken by TB surveillance officials at administrative level working together with information system managers.[3,]**

The study results showed quality issues in SINAN-TB databases in all Brazilian states. Reduction in annual TB incidence rates resulting from record linkage, classification and removal of improper repeat records from SINAN-TB database may have actually be even greater since there were unclassified repeat records and plans of health unit code changes were not available for all states. It is also likely that repeat records were left undetected in the linkage process as there is no gold standard to ascertain the sensitivity of Link-Plus software. Preliminary studies in SINAN database (unpublished data) showed its sensitivity was comparable

* Ministério da Saúde. Secretaria de Vigilância em Saúde. Sistema de Informação de Agravos de Notificação. Normas e rotinas. Brasília; 2004. (Série A: Normas e Manuais Técnicos)

** Glatt R. Análise da qualidade da base de dados de Aids do Sistema de Informação de Agravos de Notificação (Sinan) [master's dissertation]. Rio de Janeiro: Escola Nacional de Saúde Pública da FIOCRUZ; 2004.

Table 2. Number of records for each patient in *Sistema de Informação de Agravos de Notificação - Tuberculose,* by regions and states. Brazil, 2003.

Region / State	Number of records for each patient								
	Single		Duplicate		Triplicate		Quadruplicate or more		Total
	N	%	N	%	N	%	N	%	
Midwest	2,864	56.2	1,414	27.8	451	8.8	365	7.2	5,094
Federal District (DF)	511	75.8	126	18.7	27	4	10	1.5	674
Goiás (GO)	438	21.1	952	45.9	357	17.2	327	15.8	2,074
Mato Grosso do Sul (MS)	859	81.2	155	14.7	30	2.8	14	1.3	1,058
Northeast	22,090	73.7	5,812	19.4	1,370	4.6	701	2.3	29,973
Alagoas (AL)	1,118	78.2	243	17	48	3.4	20	1.4	1,429
Bahia (BA)	6,500	73.1	1,792	20.2	365	4.1	232	2.6	8,889
Ceará (CE)	4,670	76.2	1,026	16.7	283	4.6	153	2.5	6,132
Maranhão (MA)	2,405	76.1	566	17.9	142	4.5	48	1.5	3,161
Paraíba (PB)	1,030	66	425	27.2	73	4.7	32	2.1	1,560
Pernambuco (PE)	3,881	72.6	1,029	19.2	297	5.6	142	2.6	5,349
Piauí (PI)	966	66.8	371	25.6	82	5.7	27	1.9	1,446
Rio Grande do North (RN)	1,001	73.8	248	18.3	67	4.9	40	3	1,356
Sergipe (SE)	519	79.7	112	17.2	13	2	7	1.1	651
North	6,415	76.7	1,592	19	251	3	108	1.3	8,366
Acre (AC)	290	83.6	44	12.7	12	3.4	1	0.3	347
Amazonas (AM)	1,823	73.2	563	22.6	73	2.9	31	1.3	2,490
Amapá (AP)	203	71.2	73	25.6	6	2.1	3	1.1	285
Pará (PA)	3,199	78.4	716	17.5	117	2.9	50	1.2	4,082
Rondônia (RO)	537	79.6	105	15.5	21	3.1	12	1.8	675
Roraima (RR)	172	86.9	21	10.6	3	1.5	2	1	198
Tocantins (TO)	191	66.1	70	24.2	19	6.6	9	3.1	289
Southeast	32,629	70.0	9,704	20.8	2,751	5.9	1,543	3.3	46,627
Espírito Santo (ES)	1,276	84.8	176	11.7	33	2.2	20	1.3	1,505
Minas Gerais (MG)	5,173	80.3	1,001	15.5	206	3.2	65	1	6,445
Rio de Janeiro (RJ)	12,164	75.2	2,713	16.8	796	4.9	498	3.1	16,171
São Paulo (SP)	14,016	62.3	5,814	25.8	1,716	7.6	960	4.3	22,506
South	8,470	74.1	2,055	18	576	5	337	2.9	11,438
Paraná (PR)	2,617	76.3	585	17.1	135	3.9	91	2.7	3,428
Rio Grande do Sul (RS)	4,385	72.7	1,109	18.4	345	5.7	196	3.2	6,035
Santa Catarina (SC)	1,468	74.3	361	18.3	96	4.9	50	2.5	1,975
Brazil	72,468	71.4	20,577	20.3	5,399	5.3	3,054	3	101,498

Source: SINAN/SVS/MS

to that obtained using Levenshtein distance algorithm applied to patient's name, patient's mother name and date of birth.*

Alternatively, it is possible that the magnitude of reduction in TB annual incidence rates may have been overestimated if linked records of different patients were misclassified as repeat records. Misclassification of repeat records as true duplication or transfer between health units may have also contributed to overestimation. Though possible, these assumptions are unlikely given the study conservative approach.

In a probabilistic approach, accurate agreement between linkage variables is not required for record linkage. But improper classification of records as concerning

* Black PE. Levenshtein distance. In: Black PE, editor. Dictionary of Algorithms and Data Structures. Gaithersburg: National Institute of Standards and Technology; 2005. Available from: http://www.nist.gov/dads/HTML/Levenshtein.html [Accessed on Nov 3 2006]

Table 3. Classification of repeat records in *Sistema de Informação de Agravos de Notificação - Tuberculose,* by year of notification. Brazil, 2000–2004.

Year	Classification of repeat records												
	Transfer		Return		Duplicate		Relapse		Inconclusive		Missing data		Total
	N	%	N	%	N	%	N	%	N	%	N	%	N
2000	3,985	55.4	884	12.3	2,119	29.4	102	1.4	76	1.1	28	0.4	7,194
2001	5,654	47.5	2,917	24.5	2,159	18.2	831	7	270	2.3	59	0.5	11,890
2002	6,903	45	3,670	23.9	2,831	18.5	1,510	9.9	385	2.5	37	0.2	15,336
2003	7,925	46.4	4,141	24.3	2,487	14.6	1,991	11.7	493	2.9	18	0.1	17,055
2004	7,874	46.4	4,555	26.9	1,606	9.5	2,418	14.3	478	2.8	9	0.1	16,940
Total	32,341	47.3	16,167	23.6	11,202	16.4	6,852	10	1,702	2.5	151	0.2	68,415

Source: SINAN/SVS/MS

the same patient was prevented by the investigators' subsequent check of matched records. Thorough manual removal of matched records helped to improve specificity without affecting its sensitivity in finding repeat records in SINAN-TB database.

With respect to repeat records classification, only relapses, returns after default and transfers between health units in different states would be actually expected in the core national database. The other categories found reflect flawed operation and management of information system at the different levels engaged in TB surveillance and control.

Although their reporting to SINAN is mandatory, there was missing information for the variables "date of notification," "type of system entry," and "health unit code". This can be explained by faulty system operation where corrupted files are generated due to inadequate use of tools to access the original database (Sinanw.GDB) which eventually damages the system. Errors may also occur due to the fact that some states use parallel reporting systems and data are passed on to SINAN with missing mandatory fields producing incomplete databases.

Record true duplication can be generated at the time when a patient receives care from different providers in the same health unit after the visit that elicited the first notification, for example when the patient comes to the unit once again to sputum collection or medicine supply. These are the times when health providers can make a new reporting for assurance purposes and both records are eventually entered in the database. However, if main fields have any different information (notification number, date of notification, notifying municipality and unit), the system will not recognize the records as concerning the same patient and duplication will be generated.

Potential duplication in SINAN database can be ascertained using two different approaches. The first one is from listings of notifications including patient's name or their mothers' name in alphabetical order. The second approach is from listings of potential duplicates identified as having same information in a variable automatically created by the program. This automatic variable consists of a combination of patient's first and last name, gender and date of birth. Health providers engaged in TB surveillance are required to check these listings and investigate potential duplicates by contacting notifying health units so as to take the proper action. When such procedures are not routinely implemented, duplicates amass at all system levels.

The finding of records with codes of different health units but same information for the remaining variables was attributed to the introduction of new health unit codes and flawed standardization of new codes. Records with old codes were not replaced with records with new codes during vertical data transfer and thus duplicates were generated. After this programming failure was identified, SINAN national management provided the states an explanatory technical note and program correction application. The number of duplications generated due to this program failure yet to be removed from the database is now small. Therefore, the authors chose to classify this information together with other repeat records in the true duplication category. However, this program application was not widely used in the state of Goiás at the time of the study, producing 97.6% of true duplication and affecting the state's TB incidence rates.

In regard to repeat records related to transfers between health units, almost 90% were within municipalities or within the same state and these records should have been joined at local or state level, respectively. Routine procedures available in SINAN for identification and joining of transferred patient records are not automatically implemented and involvement of surveillance data management officials is necessary as these procedures require knowledge on specific TB surveillance notions. For adequate intervention the reasons why

joining routine procedures are not available should be investigated.

It is also likely that, among repeat records classified as inconclusive, there may be transfers between health units or returns after defaults which were not identified as such by the heath system and therefore not properly recorded in SINAN. To overcome this problem, better TB patient follow-up is needed as well as surveillance staff reporting to source health units of any case transfer or return after default.

The variations observed between states of data quality of SINAN-TB databases should be carefully assessed as all data management levels are equally responsible for generating repeat records. Moreover, the interpretation of data presented here is limited to the comparison of data quality related to repeat records.

Table 4. Classification of repeat records in *Sistema de Informação de Agravos de Notificação - Tuberculose,* after removal of duplicates and linkage of transferred case records by region and state. Brazil, 2003.

Region/ State	Classification of repeat records												
	Transfer		Return		Duplicate		Relapse		Inconclusive		Missing data		Total
	N	%	N	%	N	%	N	%	N	%	N	%	N
Mid-West	298	23.1	137	10.6	753	58.4	85	6.6	15	1.2	1	0.1	1,289
DF	39	47.0	16	19.3	8	9.6	16	19.3	3	3.6	1	1.2	83
GO	164	16.8	63	6.4	725	74.0	23	2.4	4	0.4	0	0.0	979
MS	46	46.0	24	24.0	14	14.0	14	14.0	2	2.0	0	0.0	100
MT	49	38.6	34	26.8	6	4.7	32	25.2	6	4.7	0	0.0	127
Northeast	2,187	47.7	1,021	22.3	602	13.1	607	13.3	155	3.4	10	0.2	4,582
AL	90	51.4	51	29.2	4	2.3	27	15.4	3	1.7	0	0.0	175
BA	780	56.1	291	20.9	110	7.9	137	9.9	65	4.7	7	0.5	1,39
CE	324	36.6	199	22.4	207	23.3	129	14.6	27	3.1	0	0.0	886
MA	276	58.7	90	19.2	23	4.9	72	15.3	9	1.9	0	0.0	470
PB	125	43.7	40	14.0	100	35.0	19	6.6	2	0.7	0	0.0	286
PE	365	42.1	227	26.2	104	12.0	136	15.7	32	3.7	3	0.3	867
PI	132	59.7	29	13.1	11	5.0	43	19.5	6	2.7	0	0.0	221
RN	71	32.7	70	32.3	39	18.0	27	12.4	10	4.6	0	0.0	217
SE	24	34.3	24	34.3	4	5.7	17	24.3	1	1.4	0	0.0	70
North	629	59.6	240	22.7	59	5.6	111	10.5	17	1.6	0	0.0	1,056
AC	2	8.0	10	40.0	0	0.0	12	48.0	1	4.0	0	0.0	25
AM	228	68.9	46	13.9	15	4.5	40	12.1	2	0.6	0	0.0	331
AP	30	63.8	13	27.7	1	2.1	2	4.3	1	2.1	0	0.0	47
PA	268	54.3	141	28.5	32	6.5	46	9.3	7	1.4	0	0.0	494
RO	38	46.9	23	28.4	11	13.6	6	7.4	3	3.7	0	0.0	81
RR	13	76.4	2	11.8	0	0.0	2	11.8	0	0.0	0	0.0	17
TO	50	82.0	5	8.2	0	0.0	3	4.9	3	4.9	0	0.0	61
Southeast	4,051	48.4	2,242	26.7	962	11.5	880	10.5	233	2.8	4	0.1	8,372
ES	51	39.2	43	33.1	6	4.6	27	20.8	3	2.3	0	0.0	130
MG	315	48.2	158	24.1	104	15.9	51	7.8	22	3.4	4	0.6	654
RJ	954	38.4	937	37.8	228	9.2	258	10.4	105	4.2	0	0.0	2,482
SP	2,731	53.5	1,104	21.6	624	12.2	544	10.7	103	2.0	0	0.0	5,106
South	760	43.3	501	28.5	111	6.3	308	17.5	73	4.2	3	0.2	1,756
PR	183	37.4	140	28.6	56	11.5	85	17.4	22	4.5	3	0.6	489
RS	433	44.3	294	30.0	38	3.9	177	18.1	36	3.7	0	0.0	978
SC	144	49.8	67	23.2	17	5.9	46	15.9	15	5.2	0	0.0	289
Brazil	7,925	46.4	4,141	24.3	2,487	14.6	1,991	11.7	493	2.9	18	0.1	17,055

Source: SINAN/SVS/MS

Table 5. Tuberculosis incidence rates by state and year of notification in complete and lean databases and percent rate difference in both databases. Brazil, 2000–2004.

State	Incidence rate - 2000			Incidence rate – 2001			Incidence rate - 2002			Incidence rate - 2003			Incidence rate - 2004		
	Complete	Lean	Difference %	Complete	Lean	Difference %	Complete	Lean	Difference %	Complete	Lean	Difference %	Complete	Lean	Difference %
AC	59.4	57.8	2.7	56.6	55.9	1.2	54.3	53.5	1.5	50.1	49.1	2.0	46.2	44.8	3.0
AL	39.6	38.2	3.5	39.7	37.7	5.0	40.7	37.9	6.9	40.9	38.9	4.9	41.3	39	5.6
AM	73	72.8	0.3	81	79.8	1.5	73.4	72	1.9	68.8	66.6	3.2	72.7	69	5.1
AP	10.1	9	10.9	38.3	37.1	3.1	49.4	45.5	7.9	40.8	38.1	6.6	40.9	37.1	9.3
BA	52.8	51.2	3.0	56.5	52.3	7.4	48	43.9	8.5	52.9	49.4	6.6	50.2	47.2	6.0
CE	45.4	43.8	3.5	43.8	41.7	4.8	45	41.9	6.9	67.5	61.3	9.2	48.7	45.5	6.6
DF	17.9	17.1	4.5	16.4	15.6	4.9	15.9	15.4	3.1	17.1	16.3	4.7	15.7	14.8	5.7
ES	41.6	40.8	1.9	42	40.4	3.8	42.6	41.5	2.6	40.2	39.5	1.7	38.5	37.8	1.8
GO	31.1	20.5	34.1	29.2	19.1	34.6	30.5	19.2	37.0	30	19.1	36.3	25.5	16.7	34.5
MA	49.7	47.2	5.0	46.7	43.7	6.4	47.7	44.7	6.3	46	43.8	4.8	46.6	43.2	7.3
MG	0.3	0.3	0.0	7.1	6.8	4.2	29.4	28	4.8	29.5	27.8	5.8	29.1	27.2	6.5
MS	41.2	39.9	3.2	39.2	38	3.1	35.7	34.1	4.5	39.8	38.3	3.8	41.8	39.1	6.5
MT	46.8	45	3.8	48.4	46.8	3.3	42.2	40.5	4.0	39.8	38.4	3.5	37.1	35.3	4.9
PA	46.4	44.5	4.1	49.4	46.2	6.5	52.2	49.2	5.7	53.6	50.9	5.0	53.6	51.2	4.5
PB	37.8	34	10.1	33.3	30.9	7.2	33.3	31	6.9	33.3	31.4	5.7	33.7	31.1	7.7
PE	46.3	43.3	6.5	47.5	43.5	8.4	51.6	47.3	8.3	53.6	49	8.6	55.3	51	7.8
PI	41.1	35.8	12.9	41.6	36.8	11.5	37.9	33.3	12.1	34.6	32	7.5	39.2	35.5	9.4
PR	25.3	24.4	3.6	26.3	25.1	4.6	26.9	25.5	5.2	27.8	26.3	5.4	26.1	24.6	5.7
RJ	95.9	90.7	5.4	95.1	87.9	7.6	94.6	87.2	7.8	89.1	82.1	7.9	85.8	79.7	7.1
RN	40.1	39.4	1.7	38.8	37.3	3.9	41.7	38.4	7.9	39.9	36.3	9.0	41.7	37.4	10.3
RO	38.3	37.5	2.1	40.5	38.9	4.0	37.6	36.5	2.9	37.8	36.7	2.9	36.2	35.4	2.2
RR	56.1	55.8	0.5	50.1	49.8	0.6	43	42.4	1.4	47.3	45.6	3.6	53.3	51.9	2.6
RS	45.3	43.2	4.6	42.6	39.2	8.0	44.9	42.4	5.6	46.5	43.8	5.8	47	44.1	6.2
SC	24.5	23.2	5.3	25.8	23.8	7.8	28.5	26.5	7.0	28.1	26.6	5.3	27.2	26.2	3.7
SE	28.5	26.5	7.0	23.6	23.1	2.1	25.5	25	2.0	29	28.1	3.1	26.6	25.7	3.4
SP	49.9	45.8	8.2	48	42.6	11.3	44.1	36.8	16.6	44.4	37.5	15.5	43.8	38.2	12.8
TO	21.2	18.2	14.2	23	20.7	10.0	23.2	21.3	8.2	19	16.4	13.7	19.9	18	9.5
Brazil	44	41.3	6.1	44.5	40.8	8.3	45.8	41.5	9.4	46.9	42.6	9.2	45.4	41.6	8.4

Source: SINAN/SVS/MS
* Per 100,000 inhabitants.

Analysis of underreporting, missing information, data inconsistence, and delayed information transmission was out of the scope of the present study but it would have been necessary if the aim of the study was a comprehensive assessment of data quality in SINAN-TB database.

Besides considerations on the study approach, it is believed that the TB annual incidence rates found in this study reflect closer estimates to the actual true rates than those obtained based on crude data both at national and state levels. TB record linkage using SINAN's core tools or other related linkage applications should be continuously promoted for improving quality of notification data.[1]

The present study is part of the *Programa Nacional de Controle da Tuberculose* (National Program for Tuberculosis Control) evaluation study, coordinated by the Department of Health Status Analysis and the Brazilian Ministry of Health Department of Epidemiological Surveillance. Data linkage using the approach here described allowed to assess baseline quality of SINAN-TB database for 2000–2004 and to develop an intervention strategy implemented in the second half of the year 2005.

REFERENCES

1. Camargo Jr KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. *Cad Saude Publica*. 2000;16(2):439-47.

2. Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc*.1969; 64(328):1183-210.

3. Laguardia J, Domingues CMA, Carvalho C, Lauerman CR, Macário E, Glatt R. Sistema de Informação de Agravos de Notificação (Sinan): desafios no desenvolvimento de um sistema de informação em saúde. *Epidemiol Serv Saude*. 2004;13(3):135-46.

Note: See the Letter to the Editor in this Supplement.