6: Data Completeness

Definitions

Term	Definition
Commercial surveillance software	A web-based surveillance system developed by a private company.
Country of origin	A calculated variable that combines the responses for RVCT item 12, Country of Birth, to determine U.S.-born or foreign-born status. The reason for this calculation is to obtain rates using the only available population estimates from the U.S. Census Bureau's American Community Survey.
Data completeness	A measure that indicates whether the information submitted contains the complete set of mandatory data items.
Electronic Report of Verified Case of Tuberculosis (eRVCT)	A web-based surveillance system for reporting TB cases developed by CDC's DTBE and available to all reporting jurisdictions. The system is based on the RVCT form.
Invalid, Missing and Unknown (MUNK)	RVCT variables that are either invalid, missing, or unknown.
National Electronic Disease Surveillance System (NEDSS)	A web-based surveillance system with an infrastructure developed by CDC that uses specific Public Health Information Network (PHIN) and NEDSS messaging standards.
National TB Indicators Project (NTIP)	A monitoring system using standardized definitions, indicators, and calculations to track progress toward attaining national TB program objectives.

Term	Definition
National Tuberculosis Surveillance System (NTSS)	The only national repository of TB surveillance data in the United States. NTSS receives data on TB cases from reporting jurisdictions' web-based systems through a standardized data collection form, the Report of Verified Case of Tuberculosis (RVCT).
NTSS reporting jurisdictions	All 50 U.S. states, the District of Columbia, New York City, American Samoa, the Federated States of Micronesia, Guam, the Republic of the Marshall Islands, the Commonwealth of the Northern Mariana Islands, Puerto Rico, the Republic of Palau, and the U.S. Virgin Islands.
Public Health Information Network (PHIN) code	A standardized code used by computer programmers to assign TB data to a specified RVCT variables. These variable codes are essential in transmitting data to CDC. Several data issues have been attributed to errors on data system programming involving PHIN codes. For example, if a code is incorrect, the data can disappear. If the data are all missing, check the PHIN Variable ID.
Report of Verified Case of Tuberculosis (RVCT)	The NTSS standardized data collection form. Data are collected by 60 reporting jurisdictions and submitted electronically to CDC. Data are used to monitor national TB trends, identify priority needs, and create the DTBE annual surveillance report, Reported Tuberculosis in the United States.
Secure Access Management Services (SAMS)	A federal information technology system that gives authorized personnel secure, external access to non-public CDC applications.
State-built surveillance system	A web-based surveillance system developed by a reporting jurisdiction.
Tuberculosis Genotyping System (TB GIMS)	A secure web-based system designed to improve access, management, and application of genotyping data at the state and local level. TB GIMS contains tools to detect and prioritize TB outbreaks.
Tuberculosis Information Management System (TIMS)	TIMS was a Windows-based, client-server application that helped health departments and other facilities manage TB patients, conduct TB surveillance activities, and manage TB programs overall. TIMS replaced former DTBE software (SURVS-TB and TBDS) and provided for electronic transmission of TB surveillance data and program management reports. TIMS was replaced by web-based surveillance systems in 2009.

Quality Assurance Process for Data Completeness

Primary Purpose

This section provides a quality assurance (QA) process that captures all the relevant data on TB patients on the RVCT.

QA Process for Data Completeness

Data completeness is an important QA component because it supports and improves the function of the National Tuberculosis Surveillance System (NTSS). The process for conducting data completeness in the Cooperative Agreements **(**CoAg) includes maintaining the completeness for all RVCT variables and matching TB and HIV/AIDS registries.

Chapter 9: Quality Assurance Cross-cutting Systems and Process provides additional tools and systems (i.e., the National Tuberculosis Indicators Project [NTIP]; Tuberculosis Genotyping System [TB GIMS]; and Cohort Review that can be used for improving data completeness.

Table 6.1 includes the CoAg requirements for maintaining data completeness and possible data sources.

Table 6.1
Data Completeness Quality Assurance Process
CoAg Requirements

Note: The requirements are based on Fiscal Year 2014 CoAg and may need to be updated when the CoAg is updated. The CoAg is reformatted into the following table with an addition of possible data sources and activities.

CoAg Requirements	Description	Possible Data Sources and Activities
Maintain completeness for all RVCT variables.	**Report TB case data to CDC using the Revised RVCT form via an electronic format that conforms to** • Public Health Information Network (PHIN), **and/or** • National Electronic Disease Surveillance System (NEDSS) messaging standards.	Complete and submit the RVCT form via an electronic format.
	Report the HIV status • For at least 95% of all newly reported TB cases.	Review HIV reports.
	Report a valid genotype accession number (generated by the CDC-sponsored genotyping laboratory) • For at least 85% of all reported culture-positive cases.	Complete genotyping reports via TB GIMS.
	Maintain at least 95% reporting completeness • For all variables existing on the pre-2009 RVCT.	Complete pre-2009 RVCT report.
	Achieve 95% completeness of all variables in the revised RVCT.	Complete post-2009 RVCT report.
Match TB and AIDS registries.	**Collaborate with the HIV/AIDS program to conduct at least annually** • TB and AIDS registry matches to ensure completeness of reporting of HIV and TB coinfected patients to both surveillance systems.	Examine • TB database and • HIV/AIDS registries.
	Investigate and verify all TB cases reported to the HIV/AIDS program and not reported to the TB program. • Update the TB registry and report to CDC as needed.	

CoAg Requirements	Description	Possible Data Sources and Activities
	At least annually • Assess reasons for incomplete HIV results on the RVCT for each verified case of TB.	
	Determine whether patients • Were **not tested** for HIV, or • **Were tested but results not reported to the TB program**.	
	Develop and implement plans to improve • HIV testing and • Reporting of HIV test results to patients and TB programs.	

Example: Invalid, Missing, and Unknown (MUNK)

Primary Purpose

This section describes Invalid, Missing and Unknown (MUNK) reports and explains how jurisdictions can correct their missing and unknown RVCT variables.

Description

MUNK reports contribute to the QA process by determining the completeness and accuracy of RVCT data reported to CDC from the jurisdictions. MUNK reports reflect 4 years' data (including the current year's data plus the previous 3 years). **Only verified and countable cases are included in the MUNK reports.**

Benefits of MUNK Reports

The MUNK reports enable jurisdictions to

- Perform data validation checks.
- Identify and correct inaccuracies and missing information in their data.
  For example, in the National TB Indicators Project (NTIP):
  - If a case is missing information for Status at Diagnosis, the case will fail to enter NTIP.
  - If the record does not include status at diagnosis–alive or dead–the entire case will be ignored in NTIP.

    - If positive culture is missing or unknown and the record has drug susceptibility results, it will not be included in the NTIP denominator for calculation of completion of therapy.
    - Other common variables that will result in failure to accurately enter NTIP if they are missing or unknown include start and stop therapy dates and the date the case was counted.
- Correct identified problems in
    - TB GIMS and
    - NTIP. The MUNK reports improve the link between variables and NTIP indicators.

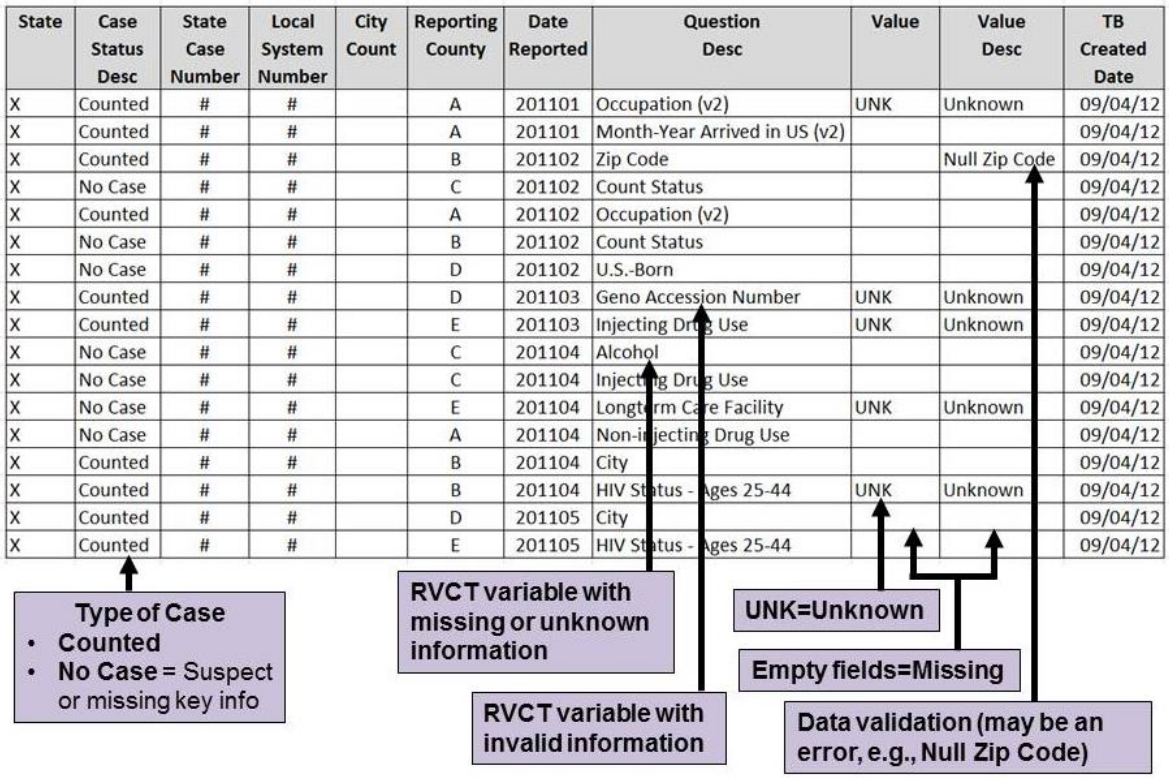
The MUNK reports are calculated for the RVCT variables listed in (Table 6.2).

Table 6.2
RVCT Variables Included in MUNK Reports

RVCT Variables	
Additional TB Risk Factors	Linking State Case Number
Alcohol	Long-Term Care Facility at DX
Chest X-Ray (v1)	Microscopic Exam
Chest X-Ray / CT (v2)	Month-Year Arrived in US (v1)
City	Month-Year Arrived in US (v2)
Correctional Facility at DX	Moved
Correctional Facility ICE	NAA Date Collected
Count status	NAA Result
Country of Birth (v2)	Non-Injecting Drug Use
Country of Origin (v1)	Occupation (v1)
County	Occupation (v2)
Culture of Tissue	Pediatric TB Patient
Date Counted	Previous TB
Date of Birth	Previous TB Year
Date Reported	Provider Type
Date Therapy Stopped	Race
DOT	Reason Evaluated for TB
Ethnicity	Reason Therapy Extended > 12 months
Final Susceptibility Testing to IR	Reason Therapy Stopped
Final Susceptibility Testing Done	Sex
Genotyping Number	Site of TB Disease
Genotyping Submitted	Skin Test (v1)
Healthcare Provider	Skin Test / IGRA (v2)
HIV Status	Sputum Conversion
HIV Status - Ages 25-44	Sputum Culture
Homeless	Sputum Smear
IGRA results	Start Therapy Date
Initial Drug Susceptibility	State Case ID
Initial Chest CT Scan or Other Chest Imaging Study	Status of TB Diagnosis
Injecting Drug Use	Susceptibility to INH

The MUNK report provides a list of RVCT variables that have missing or unknown data. Figure 6.1 is an example of a MUNK report and illustrates the type of information that can be used to determine completeness of data.

Figure 6.1
Example of a MUNK Report



State	Case Status Desc	State Case Number	Local System Number	City Count	Reporting County	Date Reported	Question Desc	Value	Value Desc	TB Created Date
X	Counted	#	#		A	201101	Occupation (v2)	UNK	Unknown	09/04/12
X	Counted	#	#		A	201101	Month-Year Arrived in US (v2)			09/04/12
X	Counted	#	#		B	201102	Zip Code		Null Zip Code	09/04/12
X	No Case	#	#		C	201102	Count Status			09/04/12
X	Counted	#	#		A	201102	Occupation (v2)			09/04/12
X	No Case	#	#		B	201102	Count Status			09/04/12
X	No Case	#	#		D	201102	U.S.-Born			09/04/12
X	Counted	#	#		D	201103	Geno Accession Number	UNK	Unknown	09/04/12
X	Counted	#	#		E	201103	Injecting Drug Use	UNK	Unknown	09/04/12
X	No Case	#	#		C	201104	Alcohol			09/04/12
X	No Case	#	#		C	201104	Injecting Drug Use			09/04/12
X	No Case	#	#		E	201104	Longterm Care Facility	UNK	Unknown	09/04/12
X	No Case	#	#		A	201104	Non-injecting Drug Use			09/04/12
X	Counted	#	#		B	201104	City			09/04/12
X	Counted	#	#		B	201104	HIV Status - Ages 25-44	UNK	Unknown	09/04/12
X	Counted	#	#		D	201105	City			09/04/12
X	Counted	#	#		E	201105	HIV Status - Ages 25-44			09/04/12

For an explanation of the variables in the MUNK Report see Chapter 10: Toolkit for Quality Assurance, Completeness Tool–6: Explanation of Invalid, Missing, and Unknown (MUNK) Variables.

Exercise for MUNK

See Chapter 9, Exercise 9.1: Identifying NTIP and MUNK Missing Data for Country of Origin. This exercise illustrates how to use an NTIP Report to identify invalid, missing or unknown data in the MUNK Report.

Access to MUNK Reports

MUNK reports are updated weekly on each Monday.

MUNK reports can be accessed through the Secure Access Management System (SAMS) portal using the NTSS reports application.
https://sams.cdc.gov/

Additional Information

Contact the Tuberculosis Applications Support (TAPS)
Phone: 678-460-7828
Phone: 404-639-8444
ntss@cdc.gov

Example: Completeness and Accuracy of 2009 Data from the National Tuberculosis Surveillance System (NTSS)

Purpose

This study provides an example of how to evaluate completeness and accuracy of TB surveillance data. The process used may be helpful to jurisdictions in evaluating their data.

Introduction

In 2009, NTSS underwent surveillance and reporting revisions. Eleven new surveillance items were added, and 25 of the 38 existing surveillance items were modified. Simultaneously, reporting jurisdictions transitioned from the Tuberculosis Information Management System (TIMS) to one of four types of systems to electronically transmit data to CDC:

- National Electronic Disease Surveillance System (NEDSS)
- Electronic Report of Verified Case of Tuberculosis (eRVCT)
- State-built surveillance software
- Commercial surveillance software

The goals of this project were to

- Evaluate the completion and accuracy of 2009 TB data reported to CDC.
- Determine the impact of surveillance and reporting changes.
- Identify areas for data quality improvement.

Methods

Data from three different datasets were examined:

- Cleaned and finalized 2008 dataset
- Cleaned and finalized 2009 dataset
- Preliminary 2009 dataset (prior to cleaning and finalizing)

TB data reported to CDC are subject to a data cleaning routine. This data cleaning routine is performed after the data are received by DTBE, Data Management and Statistics Branch. In order to better understand the nature of the data that are received by CDC, the preliminary dataset was evaluated. Responses from pre-existing NTSS variables from years 2008 and 2009 and new variables from 2009 were examined. Data were received at CDC and reported either in original form (such as sex and birth date) or used to calculate additional variables (such as site of disease and age).

Reporting areas were categorized into one of five software types to examine differences by reporting system:

- NEDSS
- eRVCT
- TIMS
- State-built surveillance software
- Commercial surveillance software

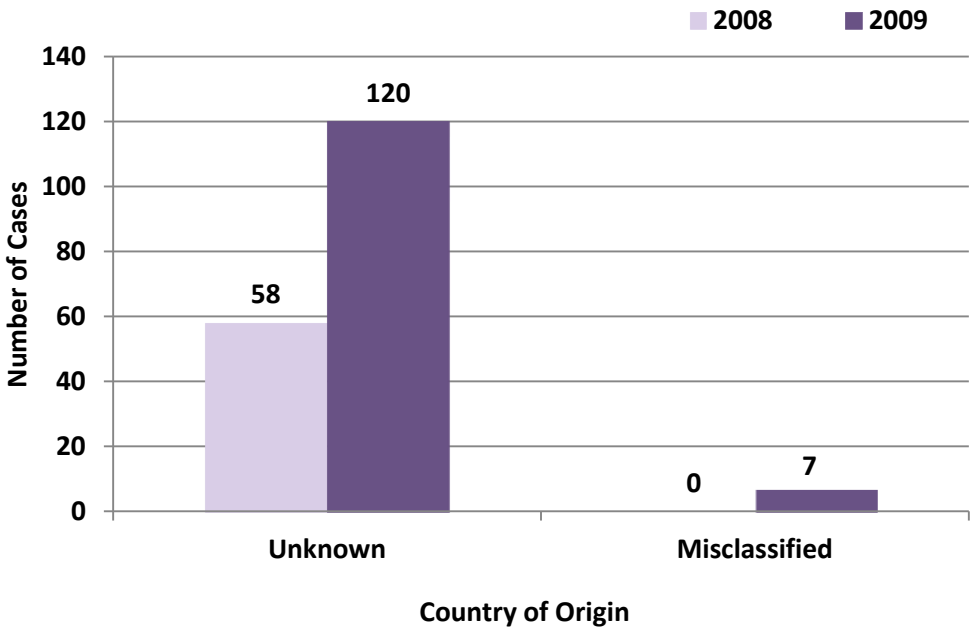
Completeness and accuracy assessment was conducted for 62 variables (among 224 total variables in the national TB dataset). Results for select variables are presented.

Results

In 2008, there were 58 records with an unknown country of birth (Figure 6.2). In comparison, there were 120 records in 2009 with an unknown country of birth. Eighty-four (70%) records that reported unknown country of birth were from a single state (this state did not report any records of unknown country of origin in 2008).

Country of origin, meaning whether the TB case was U.S.-born or foreign-born, is a variable that is calculated at CDC using the country of birth and U.S.-born variables from the RVCT. Seven records in 2009 had country of origin reported as United States, but U.S.-born was reported as 'No,' causing these records to be misclassified as foreign-born rather than U.S.-born origin (Figure 6.2). This did not occur in 2008.

Figure 6.2
Unknown and Misclassified Country of Origin, NTSS
2008 (N=12,904) and 2009 (N=11,545)



For complete data, when indicating that a TB case had a previous history of TB, the year of previous TB should also be reported if possible. There were 488 records in 2009 that indicated a previous history of TB (Table 6.2). Of these, 26 (5%) records did not report a year of previous TB. In 2008, there were 572 records that indicated a previous history of TB, and 8 (1%) of these records did not report a year of previous TB.

For accurate data, if a TB case does not have a history of previous TB, a previous year of TB should not be reported. There were 10,934 records in 2009 that indicated no previous history of TB (Table 6.2). Of these, 1 (0.01%) record was reported with a year of previous TB. In 2008, there were 12,213 records that indicated no previous history of TB, and none of these records were reported with a year of previous TB.

Table 6.2
Previous History of TB

Previous History of TB	Number of Cases		Cases without a Year of Previous TB n (%)	
	2009	2008	2009	2008
Yes	488	572	26 (5)	8 (1)
Previous History of TB	**Number of Cases**		**Cases with a Year of Previous TB n (%)**	
	2009	**2008**	**2009**	**2008**
No	10,934	12,213	1 (0.01)	0

If a sputum smear or sputum culture result is reported, then a date of specimen collection should also be reported. Sputum smear and sputum culture collection dates are new fields on the RVCT, and therefore there are no data on these variables prior to 2009 for comparison. A positive or negative sputum smear or sputum culture result indicates that a specimen was collected and should therefore have a specimen collection date. In 2009, of the 4,776 records with negative or positive results reported from a sputum smear, 128 (3%) of these records were missing collection date (Table 6.3). Of the 4,682 records with negative or positive results from a sputum culture, 116 (3%) were missing a collection date. Eighty-four (66%) records with sputum smear results reported without a collection date were from state-built systems (nearly all from a single state). Eighty-six (74%) records with sputum culture results reported without a collection date were from state-built systems (nearly all from a single state).

Date collected only applies if there are negative or positive sputum smear or sputum culture results. Therefore, if a sputum smear or sputum culture was not done, there should not be a specimen collection date reported. In 2009, of the 662 records that indicated a sputum smear test was not done, 9 (1%) reported a sputum smear collection date (Table 6.3). Of the 704 records that indicated a sputum culture test was not done, 5 (0.7%) reported a sputum culture collection date.

Table 6.3
Sputum Smear and Sputum Culture Results

Negative or Positive Results were Reported	Number of Cases	Cases without a Collection Date
	2009[1]	2009[1] n (%)
Sputum Smear[2]	4,776	128 (3)
Sputum Culture[2]	4,682	116 (3)
Tests Not Done	**Number of Cases**	**Cases with a Collection Date**
	2009[1]	**2009**[1] **n (%)**
Sputum Smear[2]	662	9 (1)
Sputum Culture[2]	704	5 (0.7)

[1]Assessment for 2009 excludes data from states that reported through TIMS.
[2]Includes cases among persons with pulmonary and extrapulmonary disease and cases of miliary TB.

For data completeness purposes, all cases reported as dead at diagnosis should also have an accompanying date of death. Date of death is a new field on the RVCT, and therefore there are no data on this variable prior to 2009 for comparison. In 2009, of the 156 cases reported dead at TB diagnosis, 17 (11%) are without a date of death reported (Table 6.4).

It is inaccurate to report a date of death for those who were alive at TB diagnosis. Date of death should only be reported for those who were dead at TB diagnosis. Of the 7,096 cases reported alive at TB diagnosis in 2009, 45 (0.6%) have a date of death reported (Table 6.4). This occurred only among reporting jurisdictions that used commercial and state-built software systems.

Table 6.4
Sputum Smear and Sputum Culture Results

Status at TB Diagnosis	Number of Cases	Cases without a Date of Death Reported
	2009[1]	2009[1] n (%)
Dead	156	17 (11)
Status at TB Diagnosis	**Number of Cases**	**Cases with a Date of Death Reported**
	2009[1]	**2009[1] n (%)**
Alive	7,096	45 (0.6)[2]

[1]Assessment for 2009 excludes data from states that reported through TIMS.
[2]Date of death should not be reported for those alive at TB diagnosis. This occurred only among reporting jurisdictions that used commercial and state-built software systems.

Further evaluation shows that the 33 (73%) who were reported as alive at diagnosis but had a date of death were also reported to have died during therapy. The date of death and the date therapy stopped was the same for 26 (79%) of the 33 who were alive at diagnosis and had a date of death reported. Therefore, it is likely that some reporting jurisdictions were using the date of death field on the RVCT to record the date a patient died during therapy. However, this is an inaccurate use of the date of death field. Currently, there is no field on the RVCT to capture the date a patient died on therapy.

Discussion

The majority of variables reported in 2009 followed the same completeness patterns as 2008 reporting. Summary reports generated by CDC may be useful tools for reporting jurisdictions to evaluate accuracy and completeness of their TB data. Ten states were still using TIMS in 2009 and were therefore unable to report new variables, so this is not a complete assessment of NTSS data.

CDC performs a data cleaning and validation routine to data. The cleaning routine is carried out in the following ways:

- If sputum smear or culture=NOT DONE, then the date of specimen collection is deleted.
- If previous history of TB=NO, then previous year of TB is deleted.

Similar cleaning is applied to several variables. This cleaning routine replaces validation rules but may not improve the quality of data reported to CDC. The cleaning routine uses a hierarchical strategy that creates a cleaner-looking dataset, but not necessarily a more accurate dataset (for example, the cleaning routine will delete a year of previous TB if history of previous TB=NO, however it may be that the year of previous TB is correct and history of previous TB=YES). Accurate and complete reporting from reporting jurisdictions to CDC is the best way to ensure a high-quality dataset.

Conclusion

CDC would like to work with reporting jurisdictions to amend unknown/missing data for key variables (such as country of origin), and encourage appropriate reporting of new variables (such as dates of death and sputum collection). Further training may prevent continuing errors in NTSS data and improve the completeness, accuracy, and thus the quality of national TB data.

Data Completeness Tools

The Data Completeness Tools are listed below (Table 6.5). Examples of the tools are located in Chapter 10: Toolkit for Quality Assurance. To view or download the tools, please visit: http://www.cdc.gov/tb/programs/rvct/default.htm.

Table 6.5
Data Completeness Tools

Tool #	Tool Name	Description and How to Use	Format	Source Contact
Completeness–1	**Source List for Locating RVCT Data**	Document used to locate information (i.e., location on medical chart, laboratory report) for each item on the RVCT	Word 2 pages	Adapted from Tuberculosis Control Program, Public Health–Seattle & King County

Tool #	Tool Name	Description and How to Use	Format	Source Contact
Completeness–2	**Treatment Outcome Status**	Table used to indicate therapy status by 12-month interval. This spreadsheet is used to monitor treatment progress with the goal of completing treatment within 12 months. There are built-in calculations for 3, 6, 9, and 12 months from treatment start that are populated when the Date Therapy Started is entered. This tool targets the National TB Indicators Project (NTIP) objective of treatment completion within 12 months.	Excel 1 page	Tennessee TB Elimination Program
Completeness–3	**Culture and Drug Susceptibility Status**	Table that indicates culture and drug susceptibility status by jurisdiction. This report shows the susceptibility results for isoniazid, rifampin, pyrazinamide, and ethambutol. It shows those cases that are multi drug-resistant and also those who have an unknown or blank susceptibility report. It is for all culture-positive TB cases. The tool targets the NTIP objective of drug susceptibility reporting.	Excel 1 page	Tennessee TB Elimination Program
Completeness–4	**TB Program Area Module (PAM) Process: Initiation of RVCT through Case Closure**	Flow chart that shows the TB PAM process (initiation of RVCT through case closure). This chart was created for Tennessee's use with TB PAM, from initiating the RVCT to closing the case. This flow chart also identifies the responsible person(s) for the various steps.	Word 1 page Legal size	Tennessee TB Elimination Program
Completeness–5	**Data Abstraction Instructions**	Detailed procedures for RVCT quality control queries	Word 4 pages	Tuberculosis Control Program, Public Health–Seattle & King County

Tool #	Tool Name	Description and How to Use	Format	Source Contact
Completeness–6	**Explanation of Invalid, Missing, and Unknown Variables**	A description of invalid, missing, and unknown variables in the MUNK report	Excel 14 pages Over-sized (fit all columns on one page)	CDC/DTBE