INT J TUBERC LUNG DIS 15(3):296–304
©2011 The Union

STATE OF THE ART SERIES
Operational Research, *Edited by* Donald A. Enarson
NUMBER 3 IN THE SERIES

Quality assurance of data: ensuring that numbers reflect operational definitions and contain real measurements

H. L. Rieder,[*][†] J. M. Lauritsen[‡][§]

[*] Tuberculosis Department, International Union Against Tuberculosis and Lung Disease, Paris, France; [†] Institute of Social and Preventive Medicine, University of Zurich, Zurich, Switzerland; [‡] Institute of Public Health, Biostatistics Unit, University of Southern Denmark, Odense, [§] EpiData Association, Odense, Denmark

SUMMARY

Any analysis is only as convincing as the quality of the underlying data. In this article, the role of data quality is exemplified by its impact on the interpretation of surveillance data, by operations research projects conducted in the training courses of the International Union Against Tuberculosis and Lung Disease, and the lessons learnt through them. It provides information why double-entry and validation of data are part of 'good clinical practice'. It is suggested how the efficiency of data entry can be maximized to reduce data entry time and data entry errors, so that psychological and physical barriers to double-entry are reduced.

KEY WORDS: research; data quality; good clinical practice

IF A STUDY reports superiority of drug A over drug B for a given indication, we may quibble over the interpretation of the data. If we were to rightfully challenge the quality of the underlying data, the study would be of no value. For the conduct of clinical trials, rules, recommendations, and indeed regulations have been elaborated to ensure the correctness of information and data quality in the United States by the Food and Drug Administration, for example, and in the European Union by a working group on data management.[1] Requirements for data documentation and data quality assurance are rigid, and rightly so. We expect impeccable information and data quality from clinical trials but often seem to be less concerned when it comes to other research. Indeed, if one peruses this *Journal*, assurance of impeccable data quality seems to be the exception rather than the rule.[2] We rarely seem to see the need to document efforts made to ensure data quality.

We suggest that data quality always matters in research, irrespective of whether it is a clinical trial, surveillance, or an operations research project. We attempt to demonstrate here how data are generated, how they should be generated and how to ensure that health events recorded on primary data sources, such as paper records, can be efficiently and accurately captured electronically to reflect the primary source. We use examples from our experience and the literature to demonstrate why data quality matters and how it can be guaranteed.

The exposé provided here should help researchers reflect on the rationale for quality-assured data capture and assist field epidemiologists in selecting key issues that increase the quality of electronic data capture. It provides them with practical recommendations on how to ensure efficiency in the tedious and repetitive task of data entry that is the backbone of any credible analysis.

SURVEILLANCE: TRANSMITTING IMPORTANT DATA IN A TIMELY FASHION

Relevant data may never enter the records in the first place, or not in a timely fashion, as is essential in surveillance. It seems a simple imperative that a good surveillance system be based on a simple, proper and timely enumeration of incident cases that become known to the health care system. How this can embarrassingly and fatally fail is illustrated by the following example.

Previous articles in this series **Editorial:** Enarson D A. Operational research, a State of the Art series in the *Journal.* Int J Tuberc Lung Dis 2011; 15(1): 3. **No 1:** Lienhardt C, Cobelens F G J. Operational research for improved tuberculosis control: the scope, the needs and the way forward. Int J Tuberc Lung Dis 2011; 15(1): 6–13. **No 2:** Harries A D, Rusen I D, Reid T, et al. The Union and Médecins Sans Frontières approach to operational research. Int J Tuberc Lung Dis 2011; 15(2): 144–154.

Correspondence to: H L Rieder, Tuberculosis Department, International Union Against Tuberculosis and Lung Disease, Jetzikofenstrasse 12, 3038 Kirchlindach, Switzerland. Tel: (+41) 31 829 4577. Fax: (+41) 31 829 4576. e-mail: TBRieder@tbrieder.org

[A version in French of this article is available from the Editorial Office in Paris and from the Union website www.theunion.org]

On 9 March 1963, a first case of *Salmonella typhi* was reported in England, and confirmed 3 days later at the Enteric Reference Laboratory in Colindale.[3] Within a few days, reports of cases from various parts of England reached the Ministry of Health. On 14 March, the Director of the Colindale Laboratory alerted the Swiss authorities to a probable typhoid epidemic, apparently waterborne, originating in Zermatt, one of Switzerland's most famous tourist resorts. A local practitioner had alerted the Mayor of Zermatt and the cantonal authorities on 10 March. The scientific report discreetly avoids mentioning whether the information was forwarded to the national authorities or whether they were taken by unpleasant surprise by the alert from the United Kingdom.[3] In any case, in England, medical officers were alerted to the outbreak on 13 March, while the Swiss national authorities officially informed the public 5 days later, on 18 March. In the United Kingdom, surveillance accomplished its purpose, while in Switzerland it failed. Following this outbreak (437 cases), Switzerland recognized that its surveillance system of communicable diseases was sub-standard and subsequently moved to mandatory federal notification by laboratories for a set of well-defined organisms, to supplement compulsory notification by physicians.

Most countries probably have compulsory notification for specified pathogens, which is often extended to include certain non-communicable diseases. However, it is known that clinicians do not consistently report notifiable diseases, so that without an additional system, such as laboratories, in the case of communicable diseases, physician-based notifications alone are often haphazard.

In a study in two hospitals in London, an inventory was made of all new diagnoses of tuberculosis (TB) during a given period of time, and compared to records in the notification system.[4,5] In the first survey, depending on the specialty, between 52% and 82% of newly diagnosed cases were reported. On repeating the survey, the apparent sensitization brought about by the first survey showed an improvement of 80% to 95% among the same specialties.

The number of variables requested on notification forms is frequently in excess of what is needed. This also extends to research,[6] possibly alienating busy practitioners. A simple but impeccable case count of cases with *S. typhi* is by far more important than asking numerous questions which can (and must) always be elicited in the outbreak investigation. Similarly, in the case of TB, a simple set of questions is sufficient for surveillance,[7] whereas a comprehensive electronic database system for case management and surveillance often fails on both counts.

When the overall incidence of a communicable disease is high, the multitude of discrete outbreaks making up the overall incidence are no longer easily discernible. This is particularly the case for conditions with long or ill-defined incubation periods such as TB, where temporal changes are expected to be gradual, rather than abrupt. In Figure 1, the year-to-year percentage changes in notifications of incident TB cases are shown for two countries: Country A in East Africa and Country B in South-East Asia. Country A had an apparently established notification system at the time the graph begins, with relatively small fluctuations in the number of cases from one year to the next. The amplitude fluctuations in case notifications in Country B are very wide until about 2002, where an approximate magnitude that we might reasonably expect is observed. Clearly, Country A had established a regular surveillance system earlier, while Country B was still approaching consolidation over several years until it became successful. Of course, such a simple graph cannot establish to what extent reported cases reflect the true number of cases known to the system, but it is evident that, until about 2002, case counts in Country B have no other value than highlighting an apparently malfunctional or deficient system.

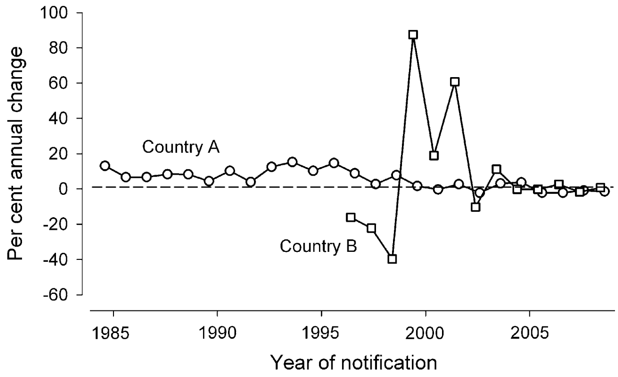


Figure 1 Percentage change in one year compared to the previous year of notified tuberculosis cases in two countries.

In contrast to the abovementioned two countries is the surveillance system in the United States. Notifiable diseases are reported weekly to the Centers for Disease Control and Prevention (CDC). While these counts are not final and verified (which happens after closure of a reporting year), the system is laid out to be sensitive and timely and collects only the case count. Through weekly comparisons of the relative change in cumulatively reported TB cases, it was definitively noted by week 39 in 1985 that cumulative TB notifications had changed the expected behavior (Figure 2), and the American public was promptly alerted to the possible impact of human immunodeficiency virus infection on TB in the United States,[8] the first national report ever to do so. The leveling off was subsequently confirmed with the provisional data for the entire year 1985,[9] and consolidated with the final data for the year.[10] This example demonstrates the overriding role of timely and impeccable case count in surveillance. The case count alone triggered future investigations specifically targeted at evaluating the

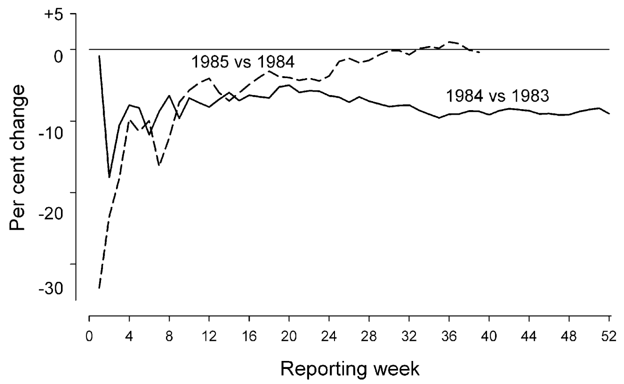


Figure 2 Percentage change of cumulatively reported incident tuberculosis cases in 1984 compared to 1983 (solid line) and the first 39 weeks of 1985 compared to the first 39 weeks in 1984. (Reprinted from Centers for Disease Control,[8] original raw data courtesy Alan B Bloch, 11 October 1994.)

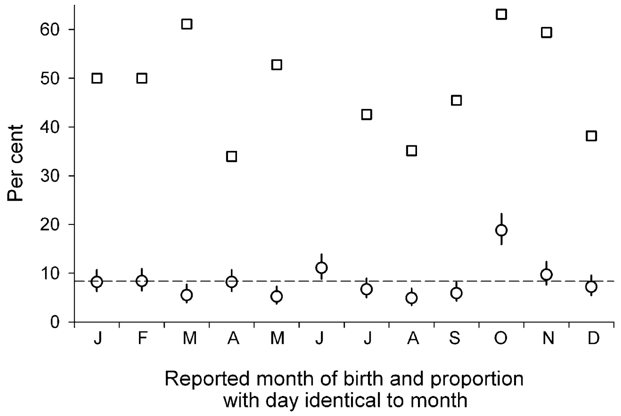


Figure 3 Percentage distribution of birth months (circles with 95% confidence intervals) recorded from parents' information regarding children with tuberculosis, and proportion in each month giving the same day as the month number in the birth date (squares). (Unpublished data courtesy Kurt Schopfer, Institute of Infectious Diseases, University of Berne, Switzerland).

hypothesis.[11,12] The first priority in surveillance is a simple and accurate case count, and not details on poorly enumerated cases.

In surveillance, the imperative is timeliness, and the greater priority this is given, the less information must be asked for, stripped down to a simple case count if outbreak intervention is at the forefront of consideration, such as in the case of *S. typhi,* meningococcal meningitis, or indeed as shown above for the re-emergence of TB in the United States. Requesting too much routine data may result in too little targeted information.

An electronic surveillance system should not be confused with an electronic case management system, which has an entirely different dimension of complexity, is hugely expensive and must remain the domain of technically highly advanced countries with a highly computer-literate user base.[13]

In research, where timeliness is less paramount, more data may be required than for surveillance; however, economy must nevertheless drive the approach, as less quantity often provides better quality in terms of completeness of data (low levels of missing data) and certainly more efficiency.

HOW WRONG DATA MAY END UP ON PAPER RECORDS

Certain characteristics are easily recognized, such as the patient's sex, and we thus hopefully record this correctly. It gets trickier with 'age', as patients may know neither their age in years nor their birth date. It is not unusual that the level of precision in information elicited by the educated health care worker is higher than the patient is actually able to supply. Precise but inaccurate information may be provided by interviewees not wishing to disappoint the interviewer. An example of this is when the birth date was asked from parents of 600 children with TB in a setting where the question about the birth date seemingly had a different meaning for parents than for the interviewer. Not only was October given as the birth month way in excess above the expected 8.3%, but the day was the same as the month number in 30–60%, rather than the expected 3% of cases (Figure 3). A particular favorite was 10 October, given for 13% of all children (unpublished data courtesy Kurt Schopfer, Institute of Infectious Diseases, University of Berne, Switzerland, 9 August 2010). The question regarding the birth date as we understand it was clearly inappropriate to the cultural context.

Misclassification is a recognized problem, for example, in the case of race/ethnicity of certain minority populations in the United States.[14] In this article, we will abstain further from discussing primary misclassifications and focus on how to prevent additional errors of omission and commission when transferring paper-based data to electronic files.

TRANSFERRING PAPER-BASED DATA TO AN ELECTRONIC DATA FILE

If it is considered difficult in surveillance to count cases correctly, it will be all the more complex to transfer more than one variable from paper to computer. Computers allow rapid and reproducible analysis, and the amount of analysis work is the same whether the database comprises 100 or 100 000 cases. Computers also allow complex data analysis that is simply not possible manually. The appeal of computer-based analysis is so strong that it is often forgotten that the number of potential errors per electronically captured record increases with an increase in the number of variables. This is regardless of the method of electronic capture. While there is a direct relationship between the number of variables and the proportion of records with at least one erroneous entry, there might be more, but less apparent, problems in the frequency of errors at the variable level. Although it is often agreed that only data that will be published should be collected,[6] this basic principle is commonly violated in routine systems.

DATA CAPTURE: THE NEED FOR SIMPLE, FAST AND ACCURATE DATA COLLECTION

Data entry is tedious, repetitive and should not pose an intellectual challenge. It is thus commonly relegated to 'data entry clerks' who often have little stake in ownership or understanding of the content. In a Union (International Union Against Tuberculosis and Lung Disease) collaborative study on the TB case register,[15] researchers entering data themselves were slower than data entry professionals, but they also made fewer errors (N B Hoa, National Tuberculosis Program, Viet Nam, personal communication, 12 June 2010).

In the following section, we illustrate with an example from the collaborative work of The Union how the answer to a relevant operational question was

Table Process of defining variables for data capture, designing the data entry form, putting restrictions on data entry, double-entry and validation

Step	Explanation
Research hypothesis	Formulate a testable research hypothesis
Minimally required variables	Define the absolute minimum number of variables that are required to test the research hypothesis
Key explanatory variables	Define essential explanatory variables that will be analyzed. Abstain from adding 'nice to know' variables and stick to 'need to know' variables
Code book	A 'code' book, also called a 'data documentation sheet', lists all the variables that will be captured. The following aspects should be defined for each variable: Field name: A short, single-word, intuitive name or sequential number for the variable, e.g., 'age' or 'V1' Field label: An explanatory label for the field, preferably exactly as it appears on the primary data source, e.g., 'age in years at last birth date'—often this is the 'question text', when in a questionnaire Field type: Type of field, e.g., 'I' (integer), 'F' (float or real number), 'D' (date), 'T' (text), 'U' (upper-case text), etc. Avoid Boolean fields (yes-no) and give preference to integer fields Field values: Legal values for the field, i.e., ranges and legal values for continuous numeric variables, date (and text fields); integers for categorical variables, e.g., 1, 2, 9 for a field 'sex', etc. Value labels: For categorical variables, explanatory labels that will be paired with the field values are defined, e.g., 'female sex' for the value 1, 'male sex' for the value 2, 'sex not recorded' for the value 9 Missing value: A defined value (or values) used to inform that this particular value represents 'no information available', either missing (not obtained), e.g., the value 9, or irrelevant for this case (e.g., the value 8) Explanatory remarks: Specifications for continuous variables and dates are written into the data form and instruct the user what to enter if information is not recorded to prevent entering first an erroneous value and only then being alerted to the constraints imposed by the legal values
Data entry form	The proof that the code book is fully explanatory is assured if an independent person can independently produce the data entry form on its basis. In EpiData, it provides the structure of the dataset which is then inherited by the actual data file used for data entry
Data entry controls	Data entry controls restrict what can and what cannot be entered. In Epi Info 6 and the current version of EpiData Entry, a separate, so-called 'check' file serves this purpose. For numerically coded categorical variables, it also provides 'pop-up' windows from which the correct value associated with the explanatory label is picked. Controls must also ensure that critical information is actually entered (e.g., preventing a record from being saved without a valid identifier), etc.
Pilot testing	Entering a series of test records will show how sturdy, user-friendly, and efficient the performance of a data entry form is in actual practice and to determine whether or not changes in the data entry form are warranted
Data entry	If a data entry form suffices the criteria of user-friendliness and efficiency data entry, while tedious, is swift
Double-entry	The same data are entered a second time into an empty copy of the data entry form with precisely the same structure, supported by an identical check file
Data validation	The two putatively identical data files are compared to produce a list of records with any discordance in one or more fields
Data correction and finalization	To keep a permanent record allowing reproducibility, one of the two files is exported to a final file in which the corrections are made by looking up the correct value in the original record for a given field. The finalized file is now ready for analysis on quality-assured data

approached using an efficient data collection instrument and validating the data to ensure high accuracy. The Table summarizes the generic steps from formulating a research hypothesis to a finalized, quality-assured dataset that is applicable to any situation and also specifically to the example presented here.

The research question: incremental yield of serial sputum smear examinations in routine clinical practice

The number of serial specimens that need to be collected has long concerned laboratory specialists and clinicians alike. Perhaps one of the largest series was reported by Hunter in 1940.[16] A routine was implemented in a sanatorium laboratory to examine up to 14 serial smears subsequent to admission of a new TB patient or until the first examination became positive. Of 1103 pulmonary TB cases examined, 825 (74.8%) were confirmed by direct sputum smear microscopy. In this setting, paying careful attention to obtaining the highest possible efficacy, 71% of all ever-positives were detected on the first examination; however, only 88% of all positives were detected with the first three examinations. With diminishing return, each sequential additional examination yielded a cumulative additional 12% of cases up to the fourteenth examination.

A relevant consideration is therefore why the international community reached the recommendation of conducting up to three smear examinations before declaring a suspect to be sputum smear-negative.[17,18] The decision was most likely based on some notion of effectiveness, balancing what was conceived to be an acceptable yield with an acceptable workload for the technicians. Data on incremental yield commonly originated from laboratories that were not as overburdened as some laboratories in high-burden countries, and findings are thus not simply transferable. For at least 80 years, the importance of allotting sufficient time for examination to find rare bacilli has been recognized.[19] Despite recommendations to the contrary, the number of fields examined, reflected in examination time, often remains too short, frequently resulting in missing paucibacillary specimens.[20]

Initial (insufficient) approaches to answer the research question

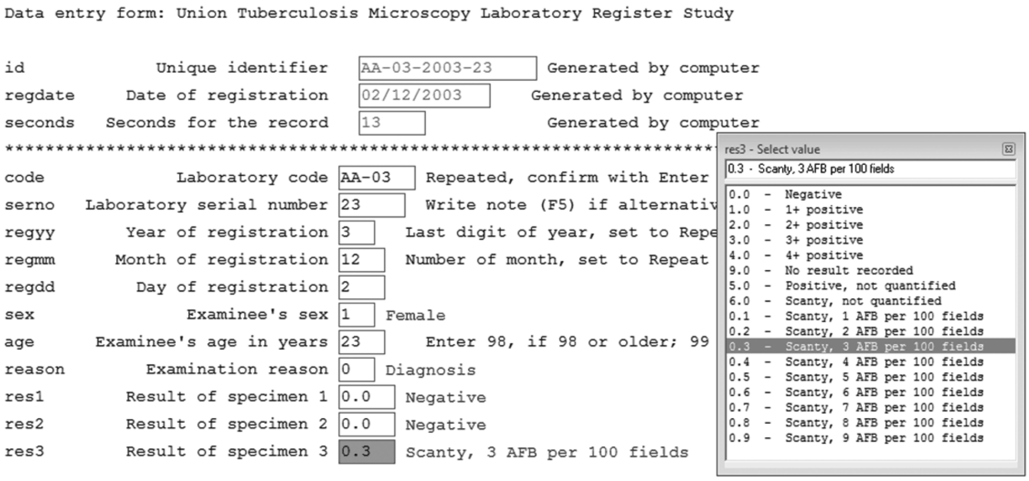
A study in rural Tanzania using manually aggregated data had shown that the incremental yield from a third serial examination was very low in routine work.[21] This motivated The Union in the mid-1990s to supplement its operations research training with courses on data collection. In-class courses were followed by collaborative field work on an operationally relevant question under close mentoring throughout the project. The first project was to evaluate the incremental yield from serial smears in Benin, Malawi, Nicaragua and Senegal.[22] Individual data were electronically captured but not validated. The only data quality assurance consisted in estimating data error frequency from a 10% sample, without any attempt at error correction. One country with errors in 15% of the re-checked records was discreetly removed from the study and was not mentioned in the final publication. There were other study deficiencies, most importantly that random sampling of registers was not rigorously enforced. Nevertheless, this study showed again that the incremental gain from a third serial smear examination was only 0.7–3%, except in Nicaragua (7.2%). Important from the research perspective was the recognition that quality assurance of data was paramount. However, the approach to design and, in particular, to data quality assurance was poor: the decision to take a 10% sample was arbitrary, as was the decision about which error frequency made a given dataset unusable. As a result, data credibility was poor, and the conclusions were perhaps correct but not sufficiently sturdy to lead to policy change.

Professionalizing operations research by designing an efficient data entry form

By 2003, The Union's operations research training concept had sufficiently matured for The Union to insist on a technically detailed research protocol that had to be strictly adhered to. A research hypothesis was formulated, according to which if more than x smears (the number x was defined by the program management of the country and differed between collaborating study countries) are required to find one additional case of sputum smear-positive TB on the third serial examination that had been missed by the first and second, then the requirement to routinely examine three sputum smears to exclude sputum smear-positive TB would be abolished in the country. A random sample from an exhaustive list of all laboratories in each country was drawn up and the data from at least one full calendar year had to be captured from each selected register. Thus, the design was representative of the public sector.

To address the primary research hypothesis, it would have sufficed to obtain five variables, i.e., a unique identifier, the type of examination (diagnostic or follow-up) and the three possible examination results. To allow orientation by time, place and person[23] for subsequent analyses,[24–27] the date of registration, laboratory code, and age and sex of the examinee were also captured, at a small additional cost. The first two of these additional variables were also utilized to construct a unique identifier.

The key variable in any dataset is a unique identifier, and EpiData Entry, the software used in the study (freely available from the EpiData Association, Odense, Denmark, http://www.epidata.dk), provided a user-friendly interface that constructed the composite identifier and checked discreetly in the background to ensure that all identifiers entered were indeed unique and that no record could be saved without it. The TB



```
Data entry form: Union Tuberculosis Microscopy Laboratory Register Study

id            Unique identifier  AA-03-2003-23   Generated by computer
regdate     Date of registration  02/12/2003   Generated by computer
seconds   Seconds for the record  13             Generated by computer
******************************************************************
code             Laboratory code AA-03  Repeated, confirm with Enter
serno   Laboratory serial number 23   Write note (F5) if alternativ
regyy       Year of registration 3   Last digit of year, set to Repe
regmm      Month of registration 12  Number of month, set to Repeat
regdd        Day of registration 2
sex              Examinee's sex  1  Female
age     Examinee's age in years  23   Enter 98, if 98 or older; 99
reason        Examination reason 0  Diagnosis
res1       Result of specimen 1  0.0  Negative
res2       Result of specimen 2  0.0  Negative
res3       Result of specimen 3  0.3  Scanty, 3 AFB per 100 fields
```

```
res3 - Select value
0.3 - Scanty, 3 AFB per 100 fields

0.0  -  Negative
1.0  -  1+ positive
2.0  -  2+ positive
3.0  -  3+ positive
4.0  -  4+ positive
9.0  -  No result recorded
5.0  -  Positive, not quantified
6.0  -  Scanty, not quantified
0.1  -  Scanty, 1 AFB per 100 fields
0.2  -  Scanty, 2 AFB per 100 fields
0.3  -  Scanty, 3 AFB per 100 fields
0.4  -  Scanty, 4 AFB per 100 fields
0.5  -  Scanty, 5 AFB per 100 fields
0.6  -  Scanty, 6 AFB per 100 fields
0.7  -  Scanty, 7 AFB per 100 fields
0.8  -  Scanty, 8 AFB per 100 fields
0.9  -  Scanty, 9 AFB per 100 fields
```

Figure 4 Data entry form in EpiData for data capture of the tuberculosis microscopy laboratory register.

microscopy laboratory register uses a sequential serial number starting with 1 at the beginning of each calendar year for each examinee. The software combined this number with the laboratory code and registration year, and thus ensured that each examinee in a given country was uniquely identifiable. To permit the fastest possibly entry and to minimize data entry errors, field values were coded numerically and supplemented with metadata in a pop-up menu with fully explanatory labels, which were also displayed after entering the numeric value as a visual control (Figure 4). All fields required a value to prevent confusion between missing and 'forgot-to-enter' information.

Compulsory stringency in data quality assurance can be attained

The US CDC implores its newly joining epidemiology trainees at the outset that an epidemiologist should never find him- or herself in the position to be forced to defend the quality of the data,[2] offering the following advice: 'Consider where you want to "do battle": on the quality of the data, or on their analysis and interpretation'.[23] As we would expect from any clinical trial, the accuracy of the data is of such paramount importance that there cannot be any compromise in this regard, however small or large the study may be.

While not an absolute requirement in the European recommendations, double-entry of data is defined as the definitive gold standard of good clinical practice.[1] If double-entry and validation are not used, complex inbuilt checks for data plausibility are essential. The need for double-entry has been challenged, albeit only on a model with simulated data.[28] Research on actual data has consistently revealed that there is a huge range in the quality of data entry—in some settings there might be minor errors, in others a large proportion of erroneous entries.[29–32] While it is predictable that complex entries will cause more errors, the performance of a given data entry person is not. Double-entry of data will not prevent all errors. It cannot address the issue of a primarily poor data source (e.g., wrongly recorded data or illegible handwriting), nor will it uncover an instance where the same erroneous entry is made twice. Data entry controls should therefore always contain in-built plausibility checks, such as issuing an alert if a legal but unusual value is entered that conflicts with values in other fields.

In the TB laboratory register study, the choice was to satisfy good clinical practice. The electronic files had to be as exact a copy as possible of the relevant components of the paper registers to accurately reflect what was actually done in the routine microscopy services of the country. It was therefore imperative that all data were entered twice and validated by comparing the files, resolving uncovered discordances by referring to the original paper record and correcting every error. Insisting on such rigorous methodology and meticulous attention to data accuracy was not acceptable to all course graduates, and attrition during the field work was, unsurprisingly, high. Nevertheless, research graduates from four countries, Moldova, Mongolia, Uganda, and Zimbabwe, had the required stamina and brought the study to an end with publication,[33,34] which suggests that this stringency can be attained.

Twinning entry efficiency with data accuracy

The most efficient approach for ensuring data validity will often be a combination of the following features: careful limitation of the number of variables, simplicity of data entry, in-built checks,[35] and finally, double-entry and validation with the necessary corrections.

To lighten the chore of double-entry of data, each entry has to be designed for speed and reduction of possible erroneous entries to reduce the number of records that need rechecking after validation. The number of records with at least one error increases proportionally with the number of variables, and the number of fields with an error will increase with the number of key strokes per variable.

Certain techniques can be adapted in the software to minimize errors. Examples are auto-completion of dates, using integer fields of length 1 with only allowed defined entries (e.g., 1, 2 and 9) rather than string fields, and moving automatically to the next field upon completion of entry.

STUDY RESULTS AND IMPACT ON POLICY

The objective in the Union TB laboratory register study comprising information on 130 000 individuals was to ascertain the reality of the actual incremental yield of serial sputum smear examinations within the context of a national TB program. The sheer size of the dataset made it all the more important that there should never be doubts about the data quality, as large studies are almost intrinsically lent more credibility, a trust that must be honored by quality-assured data.

The results were sobering (Figure 5). It has been suggested that a full-time technician using bright field microscopy should not process more than 25 smear examinations per day.[36] If the country concerned not only had the luxury of such full-time workers in the peripheral laboratories but also allowed them to spend a whole week's work (125 slides) to find one additional case on a third serial sputum smear examination that had been missed on the two earlier examinations, only one of the four countries would have been able to remain within the limits. Indeed, in one country, the yield was so poor as to challenge the notion that even a second examination was within reasonable requirements.

Never had such a large database been brought together with representative sampling among all the laboratories within each of the four national TB programs. The conclusion was inescapable: if evidence was required that there was no point in continuing to insist on systematically examining three serial smears, irrespective of the setting, before declaring a suspect to be negative, here it was. In parallel and independently, similar information accumulated from other low-income countries[37] and a formal systematic review further summarized the issue.[38] The World Health Organization (WHO) subsequently adapted its recommendation to state that, as a routine, two negative serial smears will suffice to exclude sputum smear-positive TB.[39]

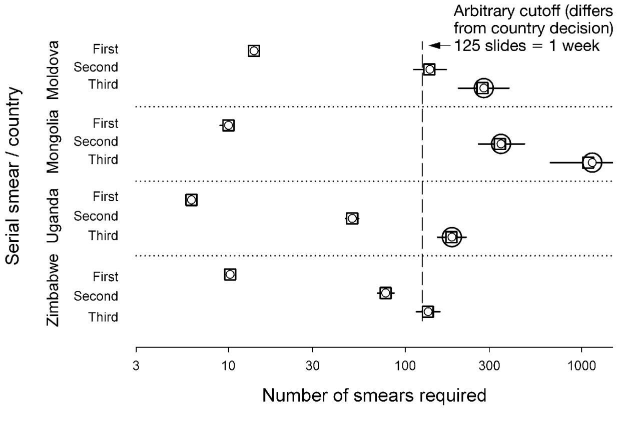


Figure 5 Incremental yield of serial sputum smear examinations in Moldova, Mongolia, Uganda and Zimbabwe, expressed as the number of smears to be examined to find one additional case or smear-positive tuberculosis not found in the earlier examination(s) with the median (circles), and the point estimate of the mean (squares) with 95% Bayesian credibility intervals (lines).

HANDHELD COMPUTERS AND DIGITAL ASSISTANTS

Digital assistants are becoming ubiquitous and provide a large range of applications for various professions, including managers, clinicians, and epidemiologists. In a trial in Peru, the time required to collect and process laboratory data was considerably reduced by the use of digital assistants, and user acceptance was high.[40] Although the timeliness of information processing was thoroughly studied, data quality was not the subject of the study. In a clinical trial, the quality of data entry on paper records was compared with data entry on handheld computers.[41] While staff found handheld computers easy to use and liked them, they voiced discomfort using them for data collection, and the frequency of data entry errors was judged to be excessive. In a study from Kenya, missing records were a huge problem with digital assistants, and missing fields abounded.[42] Thus although digital assistants generally enjoy high user acceptability, missing records and error frequency would seem to preclude their use for serious research, in addition to concerns about the lack of possibility of rechecking paper forms.

APPROPRIATE SOFTWARE FOR COMPUTERS FOR QUALITY-ASSURED DATA CAPTURE

The first software designed for the needs of epidemiologists was Epi Info at the CDC, which became available in a usable format in 1985 (Atlanta, GA, USA, http://www.cdc.gov/epiinfo/background.htm). From Version 4 onwards, the CDC began collaborating with the WHO to further develop it until it became fully mature, with Version 6, in 1992.

Epi Info 6 covered every need of the epidemiologist. The global public health community throughout the world recognized this: the software was free and legal to distribute, it ran on the slowest computers, it had a very small file size and it was highly efficient in verifying and validating data entry. With the development of the Windows™ platform (Microsoft, Redmonds, WA, USA), the original DOS™ interface became increasingly annoying, and continued functionality became at risk. There was debate and disagreement about how to accomplish the necessary move forward. The EpiData Association (http://www.

epidata.dk) took the course to retain all the assets of Epi Info 6, in particular its speed, text-based architecture, and the policy of using but not interfering with the Windows™ operating system. The result was a simple yet pleasant interface that accommodated all of the needs and expectations of both older and emerging Windows™-only generations of epidemiologists alike, yet retained the tiny file size suitable for e-mail exchange on even the slowest connections. It is notably independent of complex and non-transparent proprietary software file specifications.

The comparison of two putatively identical datasets is called 'validation'. A powerful feature of Epi Info and EpiData is that the validation process is a simple matter of pressing a few buttons, and the two files are compared record by record, identifying any discordance between the two files in any of the fields in a given record. Data validation with commonly available proprietary software often requires algorithms to be written to compare the values in a given field between the two putatively identical files, and the result is a variable-by-variable rather than a record-by-record comparison, rendering the process more complex and inefficient.

It is still quite common to see that spreadsheets are used for data entry. Spreadsheets are a superb tool for calculations, but they are not suitable as a data entry base. More sophisticated proprietary software is available at a cost beyond the salary constraints of colleagues in low-income countries, yet the basic problem of efficient data entry and validation has never been as elegantly solved as in the case of Epi Info and EpiData software. Powerful analysis software such as Stata™ (StataCorp, College Station, TX, USA) or R (R Foundation for Statistical Computing, Vienna, Austria) are not designed for data entry; these are tools for sophisticated analysis requirements that exceed those offered by EpiData software, mostly quite unnecessary for the majority of operationally relevant research. Whatever the preference in analytical software, even the most sophisticated epidemiologist still depends critically on prior quality-assured data entry; otherwise any analysis remains questionable and conclusions potentially misleading.

CONCLUSIONS

The biggest hurdle for many researchers is the self-discipline required to limit the number of variables that are to be collected. The fewer the number of variables, the greater the likelihood of their actually being analyzed. Furthermore, the time of data entry is reduced and with that a reduction in the number of records needing correction; the time saved is best invested in double-entry. To reduce the number of errors in values for each variable, numeric coding with metadata to allow explicit and unambiguous assignment is the preferred approach for any data entry whenever the character of the variable allows categorization. Poor data can ruin any analysis,[43] and 'garbage in, garbage out' holds as true as ever. While an error frequency of 1 per 1000 key strokes may be achieved in some settings,[28] such a low frequency would first have to be documented rather than assumed in any project, thus requiring some form of validation in any case. Good clinical practice does not require double-entry of data as such, but appropriate verification.[1] As a minimum, researchers should therefore document in their publications the measures they took to assure the reader of the quality of their data.[2]

Conflict of interest statement

JML is the initiator and developer of EpiData software. HLR collaborates with the EpiData Association.

References

1 Transnational Working Group on Data Management. European Clinical Research Infrastructures Network: GCP-compliant data managment in multinational clinical trials. http://www.ecrin.org/fileadmin/user_upload/public_documents/About_ecrin/downloads/ECRIN_Report_D10_Vers1_final_150908.pdf Accessed September 2010.

2 Rieder H L. What knowledge did we gain through *The International Journal of Tuberculosis and Lung Disease* in 2008 on the epidemiology of tuberculosis? Int J Tuberc Lung Dis 2009; 13: 1219–1223.

3 Bernard R P. The Zermatt typhoid outbreak in 1963. J Hyg Camb 1965; 63: 537–563.

4 Sheldon C D, Cock H, King K, Wilkinson P, Barnes N C. Notification of tuberculosis: how many cases are never reported? Thorax 1992; 47: 1015–1018.

5 Brown J S, Wells F, Duckworth G, Paul E A, Barnes N C. Improving notification rates for tuberculosis. BMJ 1995; 310: 974.

6 Mahoney M R, Sargent D J, O'Connell M J, Goldberg R M, Schaefer P, Buckner J C. Dealing with a deluge of data: an assessment of adverse event data on North Central Cancer Treatment Group Trials. J Clin Oncol 2005; 23: 9275–9281.

7 Rieder H L, Watson J M, Raviglione M C, et al. Surveillance of tuberculosis in Europe. Recommendations of a working group of the World Health Organization (WHO) and the European Region of the International Union Against Tuberculosis and Lung Disease (IUATLD) for uniform reporting on tuberculosis cases. Eur Respir J 1996; 9: 1097–1104.

8 Centers for Disease Control. Tuberculosis—United States, first 39 weeks, 1985. Morb Mortal Wkly Rep 1985; 34: 625–628.

9 Centers for Disease Control. Tuberculosis—United States, 1985—and the possible impact of human T-lymphotropic virus type III/Lymphadenopathy-associated virus infection. Morb Mortal Wkly Rep 1986; 35: 74–76.

10 Centers for Disease Control. Tuberculosis—United States, 1985. Morb Mortal Wkly Rep 1986; 35: 699–703.

11 Centers for Disease Control. Tuberculosis and acquired immunodeficiency syndrome—Florida. Morb Mortal Wkly Rep 1986; 35: 587–590.

12 Centers for Disease Control. Tuberculosis and acquired immunodeficiency syndrome—New York City. Morb Mortal Wkly Rep 1987; 36: 785–796.

13 Hillestand R, Bigelow J, Bower A, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Affairs 2005; 24: 1103–1117.

14 Stehr-Green P, Bettles J, Robertsen D. Effect of racial/ethnic misclassification of American Indians and Alaskan Natives on

Washington State death certificates, 1989–1997. Am J Public Health 2002; 92: 443–444.

15 Hoa N B, Chen W, Chay S, Lauritsen J M, Rieder H L. Completeness and consistency in recording information in the tuberculosis case register, Cambodia, China and Viet Nam. Int J Tuberc Lung Dis 2010; 14: 1303–1309.

16 Hunter R A. The routine examination for tubercle bacilli in sputum. Tubercle 1940; 21: 341–359.

17 World Health Organization. Laboratory services in tuberculosis control. Part II: microscopy. WHO/TB/98.258. Geneva, Switzerland: WHO, 1998.

18 International Union Against Tuberculosis and Lung Disease. Technical guide. Sputum examination for tuberculosis by direct microscopy in low-income countries. Paris, France: The Union, 2000.

19 Pottenger J E. The importance of the time of search in examining stained preparations for rare tubercle bacilli. J Clin Lab Med 1931; 16: 985–992.

20 Cambanis A, Ramsay A, Wirkom V, Tata E, Cuevas L E. Investing time in microscopy: an opportunity to optimise smear-based case detection of tuberculosis. Int J Tuberc Lung Dis 2007; 11: 40–45.

21 Ipuge Y A I, Rieder H L, Enarson D A. The yield of acid-fast bacilli from serial smears in routine microscopy laboratories in rural Tanzania. Trans R Soc Trop Med Hyg 1996; 90: 258–261.

22 Rieder H L, Arnadottir T, Tardencilla Gutierrez A A, et al. Evaluation of a standardized recording tool for sputum smear microscopy for acid-fast bacilli under routine conditions in low income countries. Int J Tuberc Lung Dis 1997; 1: 339–345.

23 Gregg M B. Field epidemiology. 2nd ed. New York, NY, USA: Oxford University Press, 2002.

24 Mabaera B, Lauritsen J M, Katamba A, Laticevschi D, Naranbat N, Rieder H L. Sputum smear-positive tuberculosis: empiric evidence challenges the need for confirmatory smears. Int J Tuberc Lung Dis 2007; 11: 959–964.

25 Mabaera B, Lauritsen J M, Katamba A, Laticevschi D, Naranbat N, Rieder H L. Making pragmatic sense of data in the tuberculosis laboratory register. Int J Tuberc Lung Dis 2008; 12: 294–300.

26 Mabaera B, Naranbat N, Katamba A, Laticevschi D, Lauritsen J M, Rieder H L. Seasonal variation among tuberculosis suspects in four countries. Int Health 2009; 1: 53–60.

27 Rieder H L, Lauritsen J M, Naranbat N, Katamba A, Laticevschi D, Mabaera B. Quantitative differences in sputum smear microscopy results for acid-fast bacilli by age and sex in four countries. Int J Tuberc Lung Dis 2009; 13: 1393–1398.

28 Day S, Fayers P, Harvey D. Double data entry: what value, what price? Contr Clin Trials 1998; 19: 15–24.

29 Caloto T, Huerta C, Moreno T, et al. Quality control and data-handling in multicentre studies: the case of the Multicentre Project for Tuberculosis Research. BMC Med Res Methodol 2001; 1: 14.

30 Goldberg S I, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. AMIA Annu Symp Proc 2008 Nov 6: 242–246.

31 Vannan E. Quality data—an improbable dream? A process for reviewing and improving data quality makes for reliable—and usable—results. Educause Quart 2001; 1: 56–58.

32 Weir C R, Hurdle J F, Felgar M A, Hoffman J M, Nebeker J R. Direct text entry in electronic progress notes. An evaluation of input errors. Methods Inf Med 2003; 42: 61–67.

33 Mabaera B, Naranbat N, Dhliwayo P, Rieder H L. Efficiency of serial smear examinations in excluding sputum smear-positive tuberculosis. Int J Tuberc Lung Dis 2006; 10: 1030–1035.

34 Katamba A, Laticevschi D, Rieder H L. Efficiency of a third serial sputum smear examination in the diagnosis of tuberculosis in Moldova and Uganda. Int J Tuberc Lung Dis 2007; 11: 659–664.

35 Needham D M, Sinopoli D J, Inglas V D, et al. Improving data quality control in quality improvement projects. Int J Qual Health Care 2009; 21: 145–150.

36 Rieder H L, Van Deun A, Kam K M, et al. Priorities for tuberculosis bacteriology services in low-income countries. 2nd ed. Paris, France: International Union Against Tuberculosis and Lung Disease, 2007.

37 Van Deun A. Optimization of smear microscopy for acid-fast bacilli in tuberculosis control programs. Thesis. Leuven, Belgium: Katholieke Universiteit te Leuven, 2008.

38 Mase S R, Ramsay A, Ng V, et al. Yield of serial sputum specimen examinations in the diagnosis of pulmonary tuberculosis: a systematic review. Int J Tuberc Lung Dis 2007; 11: 485–495.

39 World Health Organization. Implementing the WHO Stop TB Strategy. A handbook for national tuberculosis control programmes. WHO/HTM/TB/2008.401. Geneva, Switzerland: WHO, 2008: pp 1–184.

40 Blaya J A, Gomez W, Rodribuez P, Fraser H. Cost and implementation analysis of a personal digital assistant system for laboratory data collection. Int J Tuberc Lung Dis 2008; 12: 921–927.

41 Shelby-James T M, Abernethy A P, McAlindon A, Currow D C. Handheld computers for data entry: high tech has its problems too [Correspondence]. Trials 2007; 8: 5.

42 Auld A F, Wambua N, Onyango J, et al. Piloting the use of personal digital assistants for tuberculosis and human immunodeficiency virus surveillance, Kenya, 2007. Int J Tuberc Lung Dis 2010; 14: 1140–1146.

43 De Veaux R D, Hand D J. How to lie with bad data. Statist Sci 2005; 20: 231–238.

RÉSUMÉ

Toute analyse n'est convaincante qu'en fonction de la qualité des données qu'elle étudie. Dans cet article, on donne un exemple du rôle de la qualité des données sous forme de leur impact sur l'interprétation des données de surveillance, par les projets de recherche opérationnelle menés dans les cours de formation de l'Union Internationale contre la Tuberculose et les Maladies Respiratoires. On signale également les leçons que l'on peut en tirer. Ce travail fournit des informations sur la raison pour laquelle la double entrée des données et leur validation font partie d'une bonne pratique clinique, et il suggère la manière de porter au maximum l'efficience de l'entrée des données afin de réduire la durée et les erreurs d'entrée des données de telle manière qu'on puisse réduire les barrières psychologiques et physiques à la double entrée de ces données.

RESUMEN

Todo análisis es tan convincente como la calidad de los datos que lo sustentan. En el presente estudio, se pone de manifiesto el valor de la calidad de los datos, con su repercusión en la interpretación de los datos de vigilancia, al examinar los proyectos de investigación operativa llevados a cabo en los cursos de capacitación de la Unión Internacional Contra el Tuberculosis y las Enfermadades Respiratorias y las enseñanzas extraídas de los mismos. El análisis aporta información sobre la utilidad de la doble introducción de los datos y de la validación como parte de las 'prácticas clínicas óptimas' y ofrece sugerencias sobre la forma de maximizar la eficiencia de la introducción de datos, con el fin de acortar el tiempo dedicado a esta etapa y disminuir los errores inherentes a la misma. De esta manera se aminoran los obstáculos sicológicos y físicos que genera la doble entrada de los datos.