AN EXPERIMENT TO COMPARE ALTERNATIVE VERIFICATION PROCEDURES FOR MORTALITY MEDICAL CODING

Kenneth W. Harris and Dwight K. French
National Center for Health Statistics

I. Background

One of the most difficult data processing jobs for the National Center for Health Statistics (NCHS) is the coding of medical conditions listed as causes of death in the annual file of almost two million death records. The medical portion of the certificate consists of three lines on which the attending physician or other official is instructed to enter the sequence of medical conditions that led to death, and another line for listing other significant conditions. A nosologist (medical coder) assigns numerical codes to the medical conditions according to the Eighth Revision of the International Classification of Diseases, Adapted for Use in the United States (ICDA). These codes serve as input to a computer program that assigns one condition, called the "underlying cause of death," to represent all conditions on a certificate.

The assignment of underlying causes, either by hand or by machine, is not subject to ongoing verification because the process of assignment is so accurate (less than one-half of one percent error) that a formal verification system is neither cost nor quality effective. However, the original condition codes, which are assigned by a large staff of coders with varying degrees of proficiency, are subject to sample verification. After the original (production) coder has completed a work lot, two other coders independently code a ten percent systematic sample of records. The two sets of sample records and the corresponding production records are matched by computer, line by line and position by position. If two coders have entered the same code in the same position on the same line of a record, that code is placed into a "correct" or "preferred" set of codes for the record. The third coder matches the code if she codes it anywhere on the same line, regardless of position; otherwise, she is charged with an error.

After the matching procedure is completed for an entire lot, estimates of lot error rates for all three coders are produced by dividing the number of errors charged to each coder by the sum of the numbers of preferred codes for the sample records. These error rates serve as input into the production standards system, which is used to evaluate coder performance. Also, the production coder's error rate (the estimate of the quality of the outgoing product) determines whether her work lot is acceptable for underlying cause processing. If her sample error rate is 5 percent or less, her work lot is accepted as she coded it; otherwise, the entire lot is recoded by a fourth distinct coder and rematched against the two original sample coders.

The three-way independent verification system for mortality medical coding was instituted in 1968 because it was considered a more reliable method of measuring the level of coding error in the data than the two-way dependent system previously in use. Studies on other types of data[1] [2] [3] have shown that independent verification yields truer estimates of the amount of error in the data than does dependent verification, because a dependent verifier is biased toward the work of the original coder. However, no thorough study has ever been conducted to test the accuracy of mortality medical coding error rates based on the three-way system.

This accuracy is open to question for the following reasons:

(1) Poor handwriting, incorrect or confusing placement of medical entities on the death certificate, or poor quality microfilm can make it impossible to unquestionably determine the correct code for one or more of the conditions on a certificate.

(2) Even if the certificate is clear and properly filled out, the coding instructions may be sufficiently vague to allow two or more acceptable codes for a particular entity. The appropriateness of the three-way independent system is based on the assumption that a medical condition leads to only one code, so that when two or three of three coders with comparable ability independently arrive at the same code, there is a high probability that the code is correct. If this assumption is not valid then a coder with an acceptable code will be charged with an error when the other two coders match codes.

(3) When two or three coders have nonmatching codes all three coders are charged with an error, although it is quite likely that at least one coder has an acceptable code.

The primary purpose of the experiment was to measure the accuracy of the error rates produced by three-way independent verification, and compare them with error rates produced by two other commonly used methods of verification: two-way dependent and two-way independent coding with adjudication of differences.

II. Design of the Experiment

A. Constraints and Their Effect on the Sample Size

The goals of the experiment, along with limitations on coding time and money available, imposed the following set of constraints on the design:

(1) It was necessary to conduct the experiment using death certificates that had already been processed by the three-way verification system.

(2) The number of work lots represented in the sample needed to be sufficiently large

to include both production and sample verification work for most of the coders on the staff.

(3) In order to compare the accuracy of error rates produced by different verification systems, some measure of the "true" amount of error in the data was needed. "Truth" would have to be determined by having a small group of "experts" code the data.

(4) The number of records in the sample needed to be small enough so that the "expert" coder would not be overburdened. Each expert was limited to a maximum of two weeks of coding time to finish her assignments.

(5) Relative standard errors (RSE's) of certain key estimates should be within fixed bounds.

The sampling frame for the experiment consisted of the 472 work lots of 1974 data that passed through the three-way verification system during the months of July 1974 through March 1975. From this universe, a sample of work lots was selected. The 10 percent quality control sample of certificates within each lot represented a second stage of sampling. The RSE constraint made it necessary to have at least 25 lots in the sample, and the constraint on expert coding time limited the sample size to about 30 lots. Therefore, the first stage sample size was set at 30 lots.

B. Selection of Sample Lots

Prior to selection of the first-stage sample, the 472 work lots were sorted into 10 production error rate strata. The lots in each stratum were ordered randomly, the strata were ordered from smallest error rate upward, and the lots were temporarily renumbered from 001 to 472. The sample lots were selected systematically, so that the first-stage was essentially equivalent to a proportionate stratified sample. This procedure assured that the sample lots would be representative of the 472 lots in terms of coding difficulty (and for that matter, representative of all mortality medical coding, because the content of death certificates changes very little from year to year). In addition, the sample lots encompassed a good cross section of the 1974 Mortality Medical Coding staff. The coders who had completed the most assignments during the data year were the ones represented most often in the experiment. All coders on the staff were represented at least once as a production coder or sample coder.

C. Coding Assignments

Within each sample lot the medical codes of six distinct coders for the 10 percent quality control sample served as input data for the experiment. The first three coders were the original production coder (hereafter referred to as coder 1), and the two independent sample coders (hereafter referred to as coders 2 and 3). Because these numerical designations correspond to the coder numbers on lot-by-lot quality control reports, it was possible to distinguish coders 2 and 3 for each lot. The original production coder was always coder 1, even if her work was rejected and the lot was recoded.

A fourth coder (referred to as coder 5) was assigned to code the sample records while having access to the work of the production coder. This assignment was intended to correspond to the dependent verification procedure that was used during the 1973 data year. Because dependent verification assignments for 1973 data were given to the best available coders, most of the verification was handled by coders with relatively low error rates. In order to follow this system as closely as possible, coder 5 assignments for the experiment were allocated to 10 of the 12 coders, exclusive of the top two, with the lowest average error rates for their work during the 1974 data year. Each of the 10 coders was given 3 randomly selected lots that she had not previously worked on as coder 1, 2, or 3.

A fifth coder, hereafter called E6, was assigned to code the sample records with the codes of the production coder and one of the sample coders available to her. Her role corresponded to that of a dependent adjudicator in a two-way independent verification system. If the work lot number was odd, E6 was given access to the work of coders 1 and 2; if it was even, she had access to the work of coders 1 and 3.

The coding instructions for E6 were somewhat different from the instructions for the previous four coders. Whereas the other coders entered only one set of codes to represent the causes of death listed on a certificate, E6 was instructed to list all sets of codes that she considered acceptable to represent the certificate. If she thought that each condition on a certificate had a single correct code, she entered one set of codes. However, if one or more conditions could be coded more than one way, she entered all possible acceptable sets, changing only the code(s) in question.

Whenever E6 considered more than one set of codes acceptable, she listed the sets based on the following rules:

(1) If two or more sets are acceptable, but one is clearly preferable, list that set first and write "P" next to that set.

(2) If two or more sets are equally acceptable, the first set listed is the one she would choose if she had to decide on one of the sets. Write "D" next to that set.

The set with the P or D code following it will be referred to in future discussion as the "set of first choice."

A final coder, hereafter called E7, was assigned to code the sample records without having access to anyone else's work. Her coding instructions were exactly the same as those of E6.

The work of E6 and E7 was treated as two measures of "truth" for the purposes of this experiment. It was important, therefore, that the coders assigned to these roles be the very best coders available.

In addition, the same group of coders had to be assigned the roles of E6 and E7 so that variation between coders would not cause variation between the work of the E6 and E7 groups. To satisfy these requirements, a group of 6 "experts" was designated for these two roles. Four of the six were supervisors of the Mortality Medical Coding Unit, and the other two were the top rated coders in the Unit. The E6 and E7 assignments were randomly distributed so that each expert had ten assignments, five as E6 and five as E7. No expert was given E6 and E7 assignments for the same lot, and no expert was given an assignment in a lot she had previously coded as a production coder, sample coder, or recoder. Since the average number of sample records per lot (300) was less than the minimum daily production standard for Mortality Medical Coders (425), each expert could be expected to finish the 10 assignments within 10 coding days (2 weeks), and thus satisfy the time constraint for expert coding.

III. Analysis

A. Record Match and Assignment of Errors

After coders 5, E6, and E7 completed their coding assignments, their codes were fed into a computer program along with the codes of the original three coders. The program matched each of the original coders (1, 2, 3) with each of the coders used in the experiment (5, E6, E7), assigning errors to the original coders when their codes did not match. Whenever E6 and/or E7 coded more than one set of codes, the program observed the following rules:

(1) The number of code comparisons (the denominator in computing the error rate) between each original coder and the expert was taken from the expert's set of first choice. However,

(2) The number of errors charged to each original coder was taken from the expert's set that minimized the number of errors. The number of codes in the expert's set of first choice was used so that the denominator for computing error rates would be the same for each original coder.

B. "Expert" Agreement

In order to compare the merits of various verification systems used in this experiment, it was first necessary to determine the "true" value of the statistic being estimated, i.e., the error rate of mortality production coding (coder 1). Since the E6 and E7 assignments were completed by the very best available nosologists, these two assignment groups provided us with the "true" error rate in the sample of thirty lots. With the recognition that the "truth" from two sources might not necessarily be the same (but hopefully would be very close), we measured the agreement between E6 and E7 for all cases where at least one of them entered a code. For records where E6 and/or E7 entered more than one set of codes, the following rules were established to determine which sets should be used to measure the agreement between them:

(1) Select the comparison that minimizes the number of differences between the two experts. If two or more comparisons yield the same number of differences,

(2) Select from that group the comparison that maximizes the number of agreements between the two experts. If two or more comparisons yield the same minimum number of differences and maximum number of agreements,

(3) Select one of those comparisons using the following priority order:

a) The comparison involving the set of first choice by both experts.

b) A comparison involving E7's set of first choice.

c) A comparison involving E6's set of first choice.

d) Any other comparison.

Whenever a difference between E6 and E7 occurred, the correct code was credited to one of the experts if at least three of the other coders (1, 2, 3, 5) matched her code. The other expert was charged with an error. If no expert had a 3-1 or 4-0 majority match, neither expert was charged with an error. If at least three coders agreed on a code different from the codes of E6 and E7, both were charged with an error. This procedure enabled us to estimate the error rates of the experts. These error rates could then be used to adjust the production error rates as determined by the experts, thus leading to our best measure of the "true" production error rate. We recognize this expert error determination may be somewhat biased in favor of E6 because she had access to the codes of two coders before listing her codes; however, no more suitable measuring procedure was as easily adaptable.

As can be seen in the table which follows, E6 and E7 coded 8,973 records, generating 27,752 code comparisons. These comparisons resulted in the following rates:

(1) Agreement rate = 97.76%

(2) Difference rate = 2.24%

a) error rate of E6 = 0.56%

b) error rate of E7 = 1.26%

c) unresolved = 0.49%

The sum of a, b, and c is greater than the difference rate because some differences resulted in errors being charged to both experts.

Agreement Rates Between E6 and E7 and Conversion of Difference Rates to Error Rates, for all Combinations of Single and Multiple Sets Coded by E6 and E7

Coding Set Combination	Number of Records	Number of Codes	Agreement Rate (%)	Difference Rate (%) Charge Error to: E6	E7	E6 and E7	Unresolved
Total	8,973	27,752	97.76	0.47	1.17	0.09	0.49
E6 & E7 one set each	8,289	24,743	98.14	0.39	1.04	0.04	0.38
E6 ≥ 2 sets; E7 = 1 set	280	1,215	95.23	0.91	1.32	0.66	1.89
E6 = 1 set; E7 ≥ 2 sets	290	1,244	93.81	1.45	2.89	0.48	1.37
E6 ≥ 2 sets; E7 ≥ 2 sets	114	550	95.45	0.91	2.91	0.18	0.55

C. Estimate of "True" Error Rate

Although the error rate of E6 is much lower than the error rate of E7, we feel that the estimate of the production error rate as measured by E7 is a closer measure of the "truth." Our rationale is based on the fact that E7 operates independently in arriving at her code selections, while E6, because of her access to the work of two coders, is, minimally at least, subject to their influence. A valid counter argument, of course, is that access to the work of two coders gives the expert a broader perspective on different, acceptable coding strategies, although this argument is not supported by the number of multiple sets coded by the two experts (394 for E6 and 404 for E7). However, with the goal of obtaining an expert truth as completely free of other influences as possible, we decided on E7 as the ultimate determinant.

After the typing and punching errors made by coders 5, E6, and E7 were corrected, the "true" error rate of production coding in the sample was measured at 4.10 percent by E6 and 5.36 percent by E7. These rates, however, include the error rates of the two experts, 0.56 percent and 1.26 percent, respectively. In order to determine what proportion of these error rates should be subtracted from the production error rates in order to get "truth," the errors charged to the experts were reviewed. 8.33 percent of the time that E6 was charged with an error, the production coder agreed with her code. This percentage of E6's error rate was not subtracted from the production coder's error rate. For the same reason, 5.13 percent of E7's error rate was not subtracted from the production coder's error rate. The true production error rate, then, as measured by the two experts, is:

By E6 = $4.10 - (1 - .0833)0.56 = 3.59$ percent

By E7 = $5.36 - (1 - .0513)1.26 = 4.16$ percent--best measure

D. Comparison of Production Error Rate as Measured by Different Verification Systems

After determining the best measure of the true production error rate in the sample, we were interested in determining which of the following systems,

(1) Three-way independent coding

(2) Two-way independent coding with dependent adjudication of differences

(3) Dependent verification

provides the best estimate of the true error rate.

The three-way independent coding system estimated the production error rate at 3.75 percent. E6 provides the best measure of the error rate that would be obtained under two-way independent coding with dependent adjudication of differences. However, the 4.10 percent referred to above would not be applicable because E6 would review only those code situations in which the production coder and one independent sample coder disagreed. There were 27,952 comparisons between the two independent coders whose work E6 had access to. Of this total, they agreed on 25,897 codes, so there would be no adjudication of these codes. In the difference cases remaining, E6 charged the production coder with 852 errors. Then the production coder's error rate can be estimated by

$$\frac{852}{27,952} = 3.05 \text{ percent.}$$

We encountered a major problem in trying to use the work of coder 5 to estimate the production coder's error rate based on two-way dependent verification. After coder 5's typing and punching errors were removed from the file, she charged the production coder with 1,168 errors out of 27,575 codes, for an error rate of 4.24 percent. This estimate is higher than the estimate based on the three-way system, and is very close to the "true" error rate based on the work of E7. Such a result is, of course, quite surprising because error rates based on dependent verification are expected to be smaller than error rates based on indepen-

dent verification. In fact, the estimated production error rate for 1973 Mortality Medical Coding, based on a dependent verification system, was 0.7 percent less than the production error rate for 1974 data, measured by independent verification. This deterioration in the error rate occurred despite the fact that the coding staff and the coding instructions were virtually unchanged from 1973 to 1974.

There are two reasons that might explain this higher-than-expected error rate. The first is purely speculative: It is possible that the coders who worked as coder 5, knowing they were working on a special project, were more careful than they would have been if they had been working on a regular data file. The second reason involves a procedural difference between dependent verification of the 1973 file and dependent verification during the experiment. During the experiment, the dependent verifiers were forced to enter one set of codes to represent each certificate. When coding the regular file, dependent verifiers could change the codes of the production coder, yet not charge her with any errors (a process similar to coding two sets of codes, and assigning a "P" to one set). This difference unquestionably caused the production coders in the experiment to be assigned more errors than they would have been if the previous dependent verification procedure had been followed exactly.

In order to estimate the extent of errors charged to the production coder by coder 5 because coder 5 was not permitted to code multiple sets, we examined the multiple sets produced by E6 and E7. The coding of multiple sets by E6 and E7 resulted from ambiguity and confusion as to the correct code for certain medical conditions described and transcribed on the death certificate. It seems reasonable that coder 5, by design less knowledgeable than E6 and E7, would have at least as many ambiguous coding situations as an expert coder. Under this premise, then, coder 5, had she been permitted, would have entered multiple sets for approximately 400 records, 60 percent of which would have had a preferred set, i.e., been given a "P" at the end of the first set. (E6 and E7 designated a preferred set for 59.1 percent and 59.4 percent of their multiple sets, respectively.) These sets represent instances where coder 5 could have overruled the production coder, yet would not have charged her with an error for an acceptable code that she did not feel was the preferred code. Since the procedure followed by E6 is a closer approximation of the one followed by coder 5, the multiple sets of E6 were analyzed. It was determined that 156 codes by the production coder did not match the codes in E6's preferred sets, but were not counted as errors because they were matched in secondary sets. A comparable reduction in the number of errors charged to the production coder by coder 5 yields what we consider to be the best estimate of the expected production error rate based on two-way dependent verification, i.e., $\frac{1168 - 156}{27{,}575} = 3.67$ percent.

For the thirty work lots included in the experiment, then, we determined that the "true" production error rate is 4.16 percent. The estimate of this "true" error rate from three different coding-verification systems is as follows:

(1) Three-way independent coding--3.75 percent

(2) Two-way independent coding with dependent adjudication of differences--3.05 percent

(3) Dependent verification--3.67 percent

As the above figures indicate, the error rates based on the three-way independent system and the dependent verification system are very similar. This finding contradicts the results of other studies that compared independent verification with dependent verification. While acknowledging the apparent comparability between the two systems under experimental conditions, it should be re-emphasized that the awareness of this special study probably influenced the work of coder 5, the dependent coder. Ordinarily the work of a dependent verifier is minimally reviewed. Knowing that all of her work was going to be analyzed may have led coder 5 to perform more diligently than she would have in a normal coding situation. This possible deviation from the normal or expected coding pattern is given additional credence when we consider the work environment of the NCHS coding units. A production standard system exists that places a premium on productivity. That is, the more work that is produced, the more cash remuneration the coder can qualify for. It seems reasonable that a dependent verifier, competing against the clock, would be more inclined to agree with most of the codes listed by the production coder rather than repeat the coding process to determine if she agrees with the listed codes. This, of course, would lead to an underestimate of the production error rate.

Based on these preliminary findings, it appears that dependent verification could estimate the error in Mortality Medical Coding as accurately as three-way independent verification if

(1) The dependent verifier's work was systematically reviewed, and

(2) The current production standards system was revised to place more emphasis on the quality of coding.

Otherwise, the three-way independent system probably provides the best estimate of the quality of the coding.

References

1. Gilford, L., Estimating the Error of the Verifying Clerk when His Work is Not Independent of the Original Clerk's. Sept. 1956. Unpublished paper.
2. McKean, A., Estimating the "True" Process Average by Means of Controller Errors. Unpublished paper.
3. Minton, G., Inspection and Correction Error in Data Processing, Journal of the American Statistical Association, Vol. 64, pp. 1256-1275. Dec. 1969.