J. Hyg., Camb. (1976), **77**, 359
Printed in Great Britain

Quality control of the isolation rate of pathogens in medical microbiology laboratories

By HILARY E. TILLETT
Epidemiological Research Laboratory, Central Public Health Laboratory, London NW9 5HT
AND P. B. CRONE
Microbiology Quality Control Laboratory, London NW10 8EY

(*Received* 24 *May* 1976)

SUMMARY

Two statistical analyses are suggested to compare the success rates in isolating and identifying pathogenic organisms from specimens achieved by different laboratories participating in a quality control scheme. An example is given in which the analyses are applied to 25 laboratories that received 30 simulated specimens.

INTRODUCTION

The principles of quality control have long been applied in industry where the standard of production is required to fall within certain limits. Obviously, quality control is only needed if variation of the standard is possible. In extending the concept of quality control to laboratory work the idea is to detect whether variability in performance does, in fact, exist and if so to try to find out why. A single laboratory may monitor its own results to check its consistency, but what is of particular interest is for several laboratories to collaborate to see whether they achieve comparable results when examining replicates of clinical specimens.

The variability between biochemical laboratories in determining the content of calcium, sugar and similar substances in serum has been studied in Britain (Whitehead, 1974). Each laboratory in this quality control scheme receives a replicate from the same batch of serum and reports the content of these substances. Thus the analysis involves comparison of continuous measures, and differences between laboratories can readily be described. This paper is concerned with the rather different problems which are encountered in microbiological laboratories where results are usually qualitative, such as isolation of pathogens. The aim is to detect laboratories which are less or more successful than the rest. Such information can be used to investigate merits of different laboratory methods, or simply to alert microbiologists to the potential for improvement.

DESCRIPTION OF THE QUALITY CONTROL TRIALS

The Microbiology Quality Control Laboratory (M.Q.C.L.) prepares large batches of simulated specimens inoculated with a pathogen and appropriate non-pathogenic organisms. The latter are included to deny to the examiner of the specimen the opportunity to use a sensitive but impracticable method of isolation. The batches are then divided into portions which are posted to participating laboratories for culture, together with a brief hypothetical clinical history such as the laboratory might receive from the physician in charge of the supposed patient. Much work has been done on the preparation of simulated specimens so that there is every chance that each specimen will contain a realistic number of viable organisms which survive the postal journey to the participating laboratory. The special techniques needed for the preparation of simulated water samples have been reported (Gray & Lowe, 1976). As a check extra specimens are sent out to a few laboratories at every distribution with a request that these should be posted straight back. If the M.Q.C.L. fails to find the pathogens from these specimens the distribution is regarded as unsatisfactory and the results are not formally analysed.

Distribution of the micro-organisms throughout the batch is made as random as possible by thorough mixing. It is known that some bacteria tend to clump but the batches are inoculated to obtain a density of 10^4 to 10^6 microbes per ml. Each specimen has a volume of 5 ml. and so it is highly unlikely that any should be devoid of organisms. The greater hazard is failure to survive. The quality control distributions are confined to micro-organisms which the laboratory should be able to identify in routine work.

For the purpose of our analysis the results are regarded as binary, being either correct or incorrect. A decision has to be made for each distribution as to what will be a correct result since anomalies may arise. For example, identification of the species of pathogen may be sufficient for clinical purposes, but if a laboratory does go on to sub-type the organism and gets the sub-type wrong then there needs to be a ruling as to whether the laboratory is correct or incorrect. But most errors are of the type where the laboratory has obviously failed to find or identify even crudely the relevant pathogen.

ANALYSIS OF RESULTS

Cochran (1950) proposed a test for use with binary data whereby several 'methods' (here interpreted as participants) are judged against each other on the basis of several trials (distributions of quality control specimens). These trials may be of varying difficulty. Cochran's method is effectively an extension of McNemar's test for matched pairs which would be applicable if two laboratories examined a series of N specimens with the following results:

		Laboratory L_1		
		Correct	Incorrect	Total
Laboratory L_2	Correct	e	f	$e+f$
	Incorrect	g	h	$g+h$
	Total	$e+g$	$f+h$	N

https://doi.org/10.1017/S002217240005573X Published online by Cambridge University Press

The null hypothesis of no difference between the two laboratories is judged by a chi-squared statistic, considering the discrepant results g and f, which should be similar under the null hypothesis

$$\chi_1^2 = \frac{(|g-f|-1)^2}{g+f}.$$

Cochran extends this to the situation where c laboratories examine N matched specimens and as a result score 1 for correct and 0 for wrong.

If u_i = number of laboratories successful with the ith specimen, T_j = total number of successes by the jth laboratory, $\bar{T}$ = average number of successes per laboratory, then

$$\chi_{c-1}^2 = \frac{c(c-1)\sum_{j=1}^{N}(T_j-\bar{T})^2}{c\sum_{i=1}^{N}u_i-\sum_{i=1}^{N}u_i^2}. \quad (1)$$

This tests whether there is significant difference between the participating laboratories in their success rates by looking at discrepant results. The closer the results of trials are to 50% of laboratories being correct the fewer trials will be needed before differences between laboratories become apparent, if they exist. But the M.Q.C.L. is interested in testing routine diagnostic work with which it is hoped that participating laboratories have a high success rate. Therefore, many of the distributions result in 90% or more of laboratories being correct. Occasionally a more difficult or unusual pathogen is distributed and the rate drops. Because many of the trials are relatively easy a long series of them is required before different standards are detected and thus the quality control exercise justified. It has been possible (Hart, 1975) to detect differences between laboratories with as few as 8 virological specimens sent to 33 laboratories for which the average number of successes was 6·6.

Cochran's test requires that all participants should complete every trial. When a long series is involved inevitably some laboratories have missing results. These can arise for several reasons such as pressure of work or absenteeism at the laboratory so that this extra non-essential work is not done, failure of the specimen to arrive in adequate condition, or the laboratory may not deal routinely with that type of specimen. Also this quality control scheme has been growing so that new laboratories join the scheme during each series of trials. Initially about 60 laboratories were involved. There is a possibility of all PHLS and NHS hospital laboratories, which total approximately 500, joining eventually.

Analysis of a selected series of trials can be confined to those laboratories which have no missing results in which case Cochran's test is applied. But an assessment of results from each individual laboratory is also required and is important in keeping participants interested in the scheme. Thus the following analysis is applied.

Suppose that a particular laboratory has completed n quality control trials and these are any subset of the series N being analysed. Let x_i be the result of the ith trial so that $x_i = 1$ for a correct and $x_i = 0$ for an incorrect report.

https://doi.org/10.1017/S002217240005573X Published online by Cambridge University Press

The total score,

$$r = \sum_{i=1}^{n} x_i,$$

has expected value

$$E(r) = \sum_{i=1}^{n} p_i, \tag{2}$$

where p_i is the probability of achieving the correct result with the ith specimen, under the null hypothesis that all laboratories are equal.

$$\text{Variance,} \quad \text{var}\,(r) = E\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} p_i\right)^2.$$

It can be assumed that p_i and p_j are independent since they involve different specimens, therefore

$$\text{var}\,(r) = \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} p_i^2,$$

which is the sum of the variances $p_i(1-p_i)$ from a set of n independent binomial trials.

The x's are a series of independent observations and their aggregate r, by the central limit theorem, tends to normality for long series. As with the simple binomial distribution the convergence is quickest for p values close to 0·5. In this quality control series the p's tend to be high and therefore convergence will be slower.

In practice the p values are estimated by $\hat{p}_i$ = (total successes for ith specimen)/(total laboratories reporting for ith specimen). For the n specimens examined by the laboratory in question the following range is calculated:

$$\sum_{i=1}^{n} \hat{p}_i \pm 2\left\{\left|\left(\sum_{i=1}^{n} \hat{p}_i - \sum_{i=1}^{n} \hat{p}_i^2\right)^{\frac{1}{2}}\right| + 0{\cdot}5\right\}. \tag{3}$$

The continuity correction of 0·5 is included to compensate for the fact that we observe only integer values of r, whereas the range is based on a Normal distribution which assumes a continuous variable.

If this range does not include r, the observed number correct, the laboratory is designated better or worse than the general standard. This range will not be exactly 95% confidence interval because the distribution of r will tend to be skewed left unless the series is very long, but in the following examples will be shown to give adequate approximations.

AN EXAMPLE

A sample of results is shown in Table 1. The 30 specimens were simulated specimens from patients except for three public health specimens which were of contaminated drinking water or milk. The 25 laboratories in this example included two which joined the quality control scheme part way through the series of specimens and one which did not examine urine specimens. Thus there are 22 laboratories with complete results and to which Cochran's test can be applied.

https://doi.org/10.1017/S002217240005573X Published online by Cambridge University Press

Table 1. *Example of result with* 30 *simulated specimens*

Specimen no. ... 01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Totals per laboratory of specimens	
Type of specimen* WS	F	U	WS	F	F	WS	U	F	TS	WS	U	F	F	TS	TS	W	U	CSF	F	F	U	W	F	U	TS	F	M	WS	F		
Code no. of laboratory	*Staph. aureus*	*Sal. agona*	*Klebsiella*	*Staph. aureus*	*Sh. sonnei*	*Sh. flexneri*	Tubercle bacilli	Coliform	*Vib. cholerae*	*C. diphtheriae*	Haem. strep. (group A)	Enterococcus	*Sal. typhimurium*	*Sh. sonnei*	*C. ulcerans*	Haem. strep. (group C)	Coliforms	*Pseudomonas* sp.	Pneumococci	*Sal. agona*	*Sh. flexneri*	*Klebsiella* sp.	Coliforms	*Sh. sonnei*	*Proteus* sp.	*C. diphtheriae*	*Sh. sonnei*	*Br. abortus*	Haem. strep. (group A)	*Sal. enteritidis*	Examined	Correct
01	2	2	2	2	2	2	1	2	2	1	2	2	2	2	2	2	2	2	2	2	2	1	1	2	2	2	2	1	2	2	30	25
02	2	2	2	2	2	2	2	2	1	2	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	30	28
03	1	2	2	2	2	2	2	2	1	1	2	2	2	2	1	1	2	2	2	2	2	2	1	2	2	2	2	1	1	2	30	20
04	2	2	2	1	2	2	1	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	1	2	2	30	25
05	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	30	29
06	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	1	2	30	26
07	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	1	2	2	2	2	2	1	2	2	2	1	2	2	30	25
08	2	2	1	2	2	2	1	2	2	2	2	2	2	2	2	1	2	2	2	2	1	2	2	1	2	2	2	1	2	2	30	24
09	2	2	—	2	2	2	1	—	2	2	—	2	2	2	1	2	2	—	2	2	2	—	2	1	—	2	1	2	2	1	24	19
10	2	2	2	2	2	2	1	2	2	2	2	2	2	1	2	2	1	2	2	2	2	1	2	2	1	2	2	2	1	2	30	24
11	1	2	2	2	2	2	1	2	2	2	2	2	2	2	1	2	2	2	2	2	1	2	1	2	1	2	2	2	2	2	30	24
12	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	2	1	2	2	2	2	2	2	1	2	2	2	2	2	30	25
13	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	30	28
14	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	30	28
15	2	2	2	2	2	2	1	1	2	2	2	2	2	1	1	2	2	2	2	2	1	2	1	1	2	2	2	1	2	2	30	22
16	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	2	2	2	2	2	2	30	26
17	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	1	2	1	2	2	2	2	1	2	2	30	26
18	2	2	2	2	2	2	1	2	2	2	2	2	2	1	1	2	2	2	2	2	2	2	1	1	2	2	2	1	2	2	30	24
19	2	2	2	2	2	2	1	2	2	2	2	2	2	1	1	2	2	2	2	1	1	2	1	2	2	2	2	2	1	2	30	23
20	1	2	2	2	1	2	1	2	2	2	1	2	2	1	1	2	2	1	2	2	2	2	1	2	2	2	2	1	2	2	30	21
21	1	2	1	2	2	1	1	2	2	1	2	2	2	1	1	1	2	2	2	1	2	2	2	2	2	1	2	1	2	2	30	19
22	1	1	2	2	2	2	1	2	2	1	2	2	2	1	1	2	2	2	2	1	2	2	1	2	2	2	2	1	2	2	30	20
23	1	2	2	2	2	2	1	2	2	1	1	2	2	2	2	1	2	2	2	2	2	2	1	2	2	2	2	2	1	2	30	23
24	—	—	—	—	—	—	—	—	1	2	—	2	2	2	2	1	—	2	2	2	1	2	2	—	2	2	2	2	1	2	20	16
25	—	—	—	—	—	—	—	—	—	—	—	2	2	2	1	2	—	2	1	2	2	2	—	1	2	2	2	2	1	2	17	13
% correct for each specimen	74	96	86	96	96	96	30	95	92	71	79	92	96	96	52	44	87	96	88	100	68	92	61	40	92	96	96	52	76	96	Mean % correct = 81	

Key to result: 1 = incorrect; 2 = correct; — = not received or not attempted.

* CSF = Cerebrospinal fluid; F = faeces; M = milk; TS = throat swab; U = urine; W = water; WS = wound swab.

The numbers of laboratories with the indicated numbers of correct results were as follows:

No. of specimens correct =	30	29	28	27	26	25	24	23	22	21	20	19
No. of laboratories =	–	1	3	—	3	4	4	2	1	1	2	1

Evaluating equation (1),

$$\chi^2_{21} = 44{\cdot}3 \quad (P < 0{\cdot}005).$$

Having established that the different success rates observed between laboratories is unlikely to be due to chance the observer may look at individual laboratories including the three with incomplete series. Equations (2) and (3) should be applied and produce the following results.

For laboratories completing the series the expected number of correct results, $E(r)$, is 24·3 and the acceptable range is 20·1–28·4. Thus laboratories with 29 or more were probably performing better and laboratories with 20 or less were probably performing worse than their fellows. For laboratories with incomplete series:

Laboratory 09 $E(r) = 18{\cdot}9$ range 15–22,
Laboratory 24 $E(r) = 17{\cdot}6$ range 14–20,
Laboratory 25 $E(r) = 15{\cdot}2$ range 12–17.

Thus none of these three laboratories did significantly worse than expected.

It should be noted that the inclusion of specimen 20, for which all laboratories returned a correct report, makes no contribution to either analysis.

Calculation of probabilities

A computer program has been written to calculate exact probabilities of getting 0, 1, 2 or 3 correct or 0, 1, 2 or 3 incorrect results with up to 30 specimens. The probabilities are obtained by accumulating probability values for each possible combination of correct or incorrect results which give the score being considered. Thus there are 30 ways in which a laboratory can get one specimen in 30 wrong, 435 ways for 2 wrong and 4060 ways for 3 wrong. The calculations are made with what are assumed to be the chances of a correct result with each specimen. Again these are estimated by the observed proportions and so will be affected by complex sampling errors associated with the products of the probabilities. These will be of decreasing importance as the number of laboratories increases (in practice our analyses deal with many more laboratories than are shown in Table 1), and the calculated probabilities are meant to illustrate the adequacy of the approximate ranges of acceptable results based on Normal distribution.

We have been able to calculate complete probability distributions for series of 8 specimens. Thus considering the first 8 and then the last 8 specimens from the example in Table 1 the probabilities, using results from all participating laboratories, are as shown in Table 2. Cumulative probabilities, that is the chances of obtaining up to and including the relevant number of incorrect or correct results are also shown. Probabilities for scores of 27 and over are shown for all 30 specimens. The ranges within which the scores would be expected to fall are calculated

Table 2. *Probabilities associated with results from series of quality control specimens*

No. incorrect	Specimens 1–8 Probability	Specimens 1–8 Cumulative probability	Specimens 23–30 Probability	Specimens 23–30 Cumulative probability	Specimens 1–30 Probability	Specimens 1–30 Cumulative probability
0	0·1552		0·0780		0·0006	
						0·0065
1	0·4698		0·2808		0·0059	
						0·0344
2	0·2928		0·3634		0·0279	
3	0·0726	0·0822	0·2123		0·0792	
4	0·0090		0·0578	0·0655		
		0·0096				
5	0·0006		0·0073			
				0·0077		
6	0·0000		0·0004			
7	0·0000		0·0000			
8	0·0000		0·0000			
Estimated ranges of acceptable no. incorrect	3·5 to −0·9		4·5 to −0·7		1·6 to 9·9	

from equation 3, but translated into numbers incorrect instead of numbers correct and are shown at the foot of the table.

A 95% acceptance interval rejects results corresponding to probabilities of 0·025 at either end of the scale. Where only 8 specimens are used the probabilities of none incorrect are greater than 0·025 and, in agreement, the upper limits of the ranges are unrealistic and lie outside the top possible score. The lower limits of the two ranges for the first 8 and the last 8 specimens fall between 3 and 4 and between 4 and 5 respectively, which intervals contain the cumulative probability of 0·025. When all 30 specimens are considered the upper limit of the range falls in the interval bounded by 0·0065 and 0·0344 corresponding to between 1 and 2 incorrect.

Thus if an individual laboratory's expected upper and lower limits for the first or last 8 or upper limit for all 30 specimens were calculated using the approximate range suggested in this paper, in every case the accepted range would be consistent with an approximate 95% confidence interval, assuming good estimates of the $\hat{p}_i$'s.

DISCUSSION

Biomedical quality control is unusual in that the items which form the sample cannot be items from the routine work. The specimens arrive from the M.Q.C.L. with specially designed report forms (which include a code number for that laboratory for confidentiality) and it is likely that the work is carried out with extra care. Nevertheless the need for quality control has been adequately justified by the consistent finding of significant differences between success rates. As information

https://doi.org/10.1017/S002217240005573X Published online by Cambridge University Press

accumulates from larger numbers of distributions the success rates are being related to different methods used by laboratories. Up to now analysis has kept separate only the series of bacterial and viral specimens within these series. As numbers grow it will be possible to confine analysis to one type of specimen (e.g. faeces for bacterial investigation) or one species of pathogen.

All analysis described in this paper is for binary results which are classified as correct or incorrect. The earliest quality control exercises in microbiology laboratories in this country were concerned with testing the sensitivity of bacteria to antibiotics and similar drugs. Here the participating laboratory had to isolate the correct pathogen and then report whether it was resistant, moderately resistant or sensitive to a selection of drugs. The results have given much insight into the problems facing laboratories involved in such work routinely (Stokes & Whitby, 1971; George, 1974) but as yet little statistical analysis has been applied.

Microbiology quality control schemes are under way in several countries. In U.S.A. analysis is made of scores which are modified by the success rate of selected participants (Report 1975).

A problem can arise in analysing results of distributions requesting the identification of pathogens. If a laboratory receives the specimens but does not report its findings should it be counted as not participating or as wrong? If results are to be treated as binary it is important to insist that a laboratory which decides not to participate should reply to this effect by return of post, otherwise the suspicion arises that the laboratory has failed to isolate anything and is refraining from making a report (Hart, 1975).

Our analysis is confined to single functions of a microbiology laboratory's work – usually the identification of a pathogen. Any results which are incorrect because of poor specimens are assumed to be distributed at random among the participating laboratories. Every effort is made to maintain the uniformity of specimens to the time of posting. Once the specimen arrives the way in which it is handled is the responsibility of the receiving laboratory. But the interval between dispatch and receipt is, to a large extent, not controllable. Some check is made by recording the length of time specimens take to reach each laboratory and it is hoped that systematic adverse conditions on any particular postal route would show up with the special specimens the M.Q.C.L. sends to different parts of the country for immediate return and testing.

In the second stage of our analysis we test the hypothesis 'similar to average' for each laboratory individually. As in any significance testing errors can occur since the laboratory may fall outside the range merely by chance. But the acceptable ranges have also to avoid the other type of error – that is, the inclusion of laboratories which are, in fact, different in performance. The ranges used represent a compromise.

Quality control in this field is a relatively new venture and provides opportunity for developing suitable analytical methods. This paper has described the use of two relatively simple analyses.

We are indebted to Mr W. B. Fletcher for much helpful advice and criticism. This work was stimulated by and made possible by the participating laboratories.

Request for reprints should be addressed to P. B. Crone.

REFERENCES

COCHRAN, W. G. (1950). The comparison of percentages in matched samples. *Biometrika* **37**, 256.

GEORGE, R. H. (1974). Progress towards agreement in reports of antibiotic sensitivity. *Journal of Clinical Pathology* **27**, 1001.

GRAY, R. D. & LOWE, G. H. (1976). The preparation of simulated water samples for the purpose of bacteriological quality control. *Journal of Hygiene* **76**, 49.

HART, R. J. C. (1975). Quality control in diagnostic virology. *Proceedings of the Royal Society of Medicine* **68**, 622.

REPORT (1975). *Bacteriology II; Proficiency Testing July* 1974. U.S. Department of Health, Education and Welfare, Atlanta.

STOKES, E. J. & WHITBY, J. L. (1971). Quality control in bacteriology: preliminary trials. *Journal of Clinical Pathology* **24**, 790.

WHITEHEAD, T. P. (1974). Quality control techniques in laboratory services. *British Medical Bulletin* **30**, 277.