

# Quality Control in Nucleic Acid Amplification Methods: Use of Elementary Probability Theory

DANIEL S. SHAPIRO\*

Department of Medicine and Department of Pathology and Laboratory Medicine, Boston University  
School of Medicine and Boston Medical Center, Boston, Massachusetts

Received 7 May 1998/Returned for modification 17 November 1998/Accepted 25 November 1998

**Since it is not possible to state with certainty that contamination has occurred during a nucleic acid amplification assay in the absence of a positive result for a negative control, methods of elementary probability theory are used to illustrate how to identify those runs in which the possibility of contamination should be considered. The use of binomial and Poisson distributions and an analysis of clusters are presented with illustrative examples to demonstrate their use.**

In working with target amplification methodologies such as PCR, ligase chain reaction (LCR), transcription-mediated amplification, and nucleic acid sequence-based amplification, the clinical laboratory must strictly adhere to a number of procedures to minimize the risk of contamination (2). These include such safeguards as employing enzymatic inactivation with uracil-*N*-glycosylase when performing PCR, performing the amplification part of the assay in a different location from the specimen processing, and using positive displacement pipettes. Despite the use of such practices, the possibility of contamination of specimens with amplified products remains. Contamination, producing false-positive assay results, can have important clinical consequences.

Although it is not possible to determine with absolute certainty whether or not contamination has occurred during a given amplification run in the absence of a negative control that yields a positive result, it is possible, by using elementary probability theory, to identify those runs that arouse suspicion of contamination. In order to use these methods, data regarding the tested population should be known. It is important to note that the analyses used in the examples in this work apply only to qualitative assay systems and not to quantitative systems, with the exception of the use of the Poisson approximation of the copy number.

Several examples will be used to illustrate the use of elementary probability theory. In these examples, there are several terms that will be used. These include the following: (i) events, which will be represented by A, B, etc.; (ii) probabilities of events, represented by P(A), P(B), etc. (example: in the toss of a fair coin, if the event H is obtaining heads and the event T is obtaining tails, then P(H) = P(T) = 0.5; (iii)  $P(X = k)$  is the probability of obtaining  $k$  successes in a number of trials; (iv) the number of ways to choose an unordered subset of size  $k$  out of  $n$  elements is  $\binom{n}{k} = n!/k!(n - k)!$ , where  $n!$  = the product of all positive integers less than or equal to  $n$  (thus,  $1! = 1$ ,  $2! = 1 \cdot 2$ ,  $3! = 1 \cdot 2 \cdot 3$ , and in general  $n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot n$ ;  $0!$  is defined as equaling 1;  $\binom{n}{k}$  is also the number of combinations of  $n$  items taken  $k$  at a time); (v)  $e = 2.71828 \dots$

**Example 1: the binomial distribution.** On the basis of historical data, 10% of all specimens submitted for LCR for detection of *Chlamydia trachomatis* are positive in a given

laboratory. Thus, the probability that an individual specimen is positive for *C. trachomatis* is 0.1 in this laboratory.

An individual run consists of 19 patient specimens and five controls. A run is performed in which all control specimens are within range. Of the 19 patient specimens, 14 are negative and 5 are positive. Is this the result of contamination?

**Analysis of example 1.** In a case in which there are  $n$  independent repetitions of an experiment in which there can be only two outcomes, one with probability  $p$  and the other with probability  $(1 - p)$ , the resulting density is known as the binomial density, with parameters  $n$  and  $p$ .

Let  $X$  = the number of successes in  $n$  independent trials, with probability  $p$  of success on each trial. Then,  $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$  for  $k = 0, 1, 2, \dots, n$ . Let  $q = 1 - p$ . Then,

$$P(X = k) = \binom{n}{k} p^k (q)^{n-k} \text{ for } k = 0, 1, 2, \dots, n \quad (1)$$

(Note: The sum of the probabilities of all the possibilities

$$\sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = (p + q)^n = 1, \text{ since } p + q = 1.)$$

Example 1 corresponds to a binomial density with parameters  $n = 19$  and  $p = 0.1$ . The solution can be calculated where  $n = 19$ ,  $p = 0.1$ , and  $q = 0.9$  as follows. Let A be the event where five or more positive results occur among the 19 patient specimens in the run; then, P(A) = the probability of obtaining five or more positive results in the run. Since the probability of obtaining at least five positive results is  $1 - (\text{probability of obtaining zero or one or two or three or four positive results})$ :  $P(A) = 1 - [P(0) + P(1) + P(2) + P(3) + P(4)]$ . Substituting from equation 1, for  $n = 19$ :

$$P(A) = 1 - [q^{19} + 19q^{18}p + 171q^{17}p^2 + 969q^{16}p^3 + 3,876q^{15}p^4]$$

Since  $p = 0.1$  (the probability of obtaining a positive result for any single specimen) and  $q = 1 - p = 0.9$ :

$$P(A) = 1 - [0.9^{19} + 19(0.9)^{18}(0.1) + 171(0.9)^{17}(0.1)^2 + 969(0.9)^{16}(0.1)^3 + 3,876(0.9)^{15}(0.1)^4] = 0.035$$

Therefore, given the assumption that each sample has a probability of 0.1 of a positive result, five or more positive results would be anticipated to occur in a run of 19 samples 3.5% of the time.

\* Mailing address: Clinical Microbiology and Molecular Diagnostics Laboratories, ENC H-406, Boston Medical Center, 88 East Newton St., Boston, MA 02118. Phone: (617) 638-8705. Fax: (617) 638-7878. E-mail: dshapiro@bu.edu.

**Interpretation of results for example 1.** Since the probability of obtaining such a result due to chance alone is 3.5%, a number of possibilities have to be considered. (i) This could, in fact, be due to chance alone and should be expected to occur approximately one time in each 29 runs. If this assay is run daily, such a result would be expected to occur, on average, approximately once per month. (ii) Contamination may have occurred. If no explanation is apparent, this must be evaluated. Such an evaluation should include performing a run with known negative specimens as well as performing the assay with swab specimens from areas of the instrument and the laboratory to see if a positive result occurs, indicative of contamination. (iii) At least one of the assumptions used to analyze this event may be incorrect. A number of possibilities exist, including the two following possibilities. (a) The probability of a positive result in the population may have increased since it was initially determined. Even a rather modest change, such as a change in the probability of a positive test from 10 to 12%, can have a major impact on the probabilities calculated from equation 1. The resulting calculation with  $n = 19, p = 0.12, q = 0.88$  would be

$$P(A) = 1 - [0.88^{19} + 19(0.88)^{18}(0.12) + 171(0.88)^{17}(0.12)^2 + 969(0.88)^{16}(0.12)^3 + 3,876(0.88)^{15}(0.12)^4] = 0.069 \text{ or } 6.9\%$$

(b) The specimens, which in aggregate have a probability of 0.1 of being positive for *C. trachomatis* by LCR, are, in fact, made up of subpopulations that differ in the probability of a positive specimen (see example 2). Those specimens that come, for example, from the population of patients in the sexually transmitted disease (STD) clinic may have a probability of >0.1 of being positive, while those specimens from outpatient areas in which screening is performed on asymptomatic patients may have a probability of <0.1 of being positive. If, as occasionally happens, the number of specimens from the STD clinic in a given run is higher than expected, the population that has been sampled does not, in fact have a probability of 0.1 of having a positive result from the assay, but a probability of >0.1.

**Example 2: the binomial distribution for specimens that come from two different populations.** On the basis of historical data, 20% of specimens submitted for LCR for detection of *C. trachomatis* from the STD clinic are positive and 1% of specimens from other patient locations are positive. In a given run, 9 specimens from the STD clinic and 10 specimens from other sites are tested. Five of these specimens are positive. What is the probability that five or more specimens would be positive on the basis of chance alone?

**Analysis of example 2.** The site-specific probabilities will often not be known by laboratories, and patient demographics will not be supplied with the specimens. This is particularly likely to be the case in reference laboratories. In this example, by knowing the site-specific probabilities of a positive assay, it is possible to refine the model that was used in example 1. Here, if a total of  $n$  specimens are tested, with  $r$  specimens from population 1 and  $(n - r)$  from population 2, we can calculate the probability of obtaining a stated number,  $k$ , of positive results. In population 1, let the probability of a positive result from a specimen be  $p_1$ . Then, since  $p_1 + q_1 = 1$ , the probability of a negative result is  $1 - p_1 = q_1$ . Since  $r$  specimens are obtained from this population then, from equation 1 we obtain the following:

$$P(X = k_1) = \binom{r}{k_1} p_1^{k_1} (q_1)^{r-k_1} \tag{2}$$

TABLE 1. Example 2. Possible ways to have no more than four positive results

No. of positive results from STD clinic	No. of positive results from other sites
0.....	0 or 1 or 2 or 3 or 4
1.....	0 or 1 or 2 or 3
2.....	0 or 1 or 2
3.....	0 or 1
4.....	0

Similarly, in population 2, the probability of a positive result from a specimen is  $p_2$  and the probability of a negative result is  $1 - p_2 = q_2$ . If  $(n - r)$  specimens are obtained from this population then the probability of obtaining  $k_2$  positive results is as follows:

$$P(X = k_2) = \binom{n-r}{k_2} p_2^{k_2} (q_2)^{n-r-k_2} \tag{3}$$

Based on equations 2 and 3, we can calculate the solution to the problem where  $r = 9, n = 19, p_1 = 0.2$ , and  $p_2 = 0.01$ . Let A be the event where five or more positive results occur among the 19 patient specimens in the run; then,  $P(A)$  = the probability of obtaining five or more positive results in the run. Since the probability of obtaining at least five positive results is  $1 -$  (probability of obtaining zero or one or two or three or four positive results),  $P(A) = 1 - [P(0) + P(1) + P(2) + P(3) + P(4)]$ .

The only ways of obtaining 0 or 1 or 2 or 3 or 4 positive results are summarized in Table 1. Table 2 summarizes the probability of each of the events, calculated from equations 2 and 3, by using the same method of calculation as that used in example 1, and rounded to three decimal places.

Since the events—obtaining a given number of positive results from the STD clinic specimens and obtaining a given number of positive results from the specimens from other sites—are independent events, the probability that both events occur is the product of their respective probabilities. Thus,

$$P(A) = 1 - [(0.134)(1.000) + (0.302)(1.000) + (0.302)(1.000) + (0.176)(0.996) + (0.066)(0.904)] = 0.027$$

TABLE 2. Probabilities (by site of specimen origin) of obtaining four or fewer positive results

No. of positive results from site	Probability of obtaining this result from specimens <sup>a</sup>
<b>STD clinic</b>	
0.....	0.134
1.....	0.302
2.....	0.302
3.....	0.176
4.....	0.066
<b>Other</b>	
0 or 1 or 2 or 3 or 4.....	1.000
0 or 1 or 2 or 3.....	1.000
0 or 1 or 2.....	1.000
0 or 1.....	0.996
0.....	0.904

<sup>a</sup> For positive results from samples collected at an STD clinic, the probability refers to that of obtaining the same result from nine STD clinic specimens. For results from samples collected at other sites, the probability refers to that of obtaining the same result from 10 specimens from other sites.

Therefore, the probability of obtaining a total of five or more positive results from the two groups of specimens due to chance alone is 0.027, or 2.7%.

**Example 3: the use of the Poisson approximation to the binomial distribution.** A laboratory performs PCR for *Mycobacterium tuberculosis* on samples of cerebrospinal fluid. Historically, positive results have constituted 3% of the samples tested. In reviewing the results of 100 samples subjected to PCR, what is the probability that there will be at most 2 positive samples?

**Analysis of example 3.** We have seen, as in example 1, that an exact calculation can be performed by using the binomial expansion. Since this case corresponds to a binomial density with  $n = 100$  and  $p = 0.03$ , we can calculate the probability of this directly, using equation 1, where A represents the event that there are zero, one, or two positive results and  $P(A)$  is the probability of obtaining zero, one, or two positive results:  $P(A) = q^{100} + 100q^{99}p + 9900q^{98}p^2 = 0.42$ . For large numbers of samples, if the probability of a positive result is small, it may be more convenient to use the Poisson approximation to the binomial distribution, where  $\lambda = np$ .

Then, the probability of having  $k$  positive results in  $n$  samples is approximately

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Here,  $\lambda = np = (100)(0.03) = 3$ . Therefore, the probability of obtaining no positive results is found when  $k = 0$ :  $P(X = 0) = 3^0 e^{-3}/0! = e^{-3}$ . The probability of obtaining one positive result is found when  $k = 1$ :  $P(X = 1) = 3^1 e^{-3}/1! = 3e^{-3}$ . The probability of obtaining two positive results is found when  $k = 2$ :  $P(X = 2) = 3^2 e^{-3}/2! = 9e^{-3}/2$ . Therefore, the probability of obtaining zero, one, or two positive results is

$$\begin{aligned} P(X = 0) + P(X = 1) + P(X = 2) &= e^{-3} + 3e^{-3} \\ &+ \frac{9e^{-3}}{2} = \frac{17e^{-3}}{2} = 0.42 \end{aligned}$$

**Example 4: further use of the Poisson distribution.** A positive control for a PCR assay was made by diluting a known positive sample and then using aliquots of this dilution to make individual positive controls. In testing the positive control by PCR amplification, it was positive (yielded the appropriate PCR product) 90% of the time but was negative (did not yield the PCR product) 10% of the time. Assuming that the amplification was performed properly, that the efficiency of the amplification assay is 100%, and that the efficiency does not vary from run to run, on average how many copies of the target nucleic acid sequence are present in each aliquot of the positive control?

**Analysis of example 4.** This is an example of a limiting dilution problem. The methods used to solve this have been employed in limiting dilution assays to estimate the number of bacteria or viruses as well as the number of specific cell types in cell culture. When there are several dilutions with which PCR is performed multiple times, it is possible to obtain an estimate of the number of copies of the nucleic acid in an individual aliquot by using a computer program that can be downloaded or used with a Java-capable browser from the World Wide Web (3). Assume that an individual aliquot of the positive control can have any integral number of copies of the target sequence: zero, one, two, three. . . .

The probability that zero copies of the target sequence are present,  $P(0)$ , is 0.1, since in 10% of the controls the amplification did not yield the amplified PCR product.

In solving this problem, we can use the Poisson distribution:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$P(X = 0) = 0.1 = \frac{\lambda^0 e^{-\lambda}}{0!}$$

Thus,  $0.1 = e^{-\lambda}$ , meaning that  $\ln(0.1) = -\lambda$ ; therefore,  $\lambda = 2.3$ .

**Interpretation of results for example 4.** Since  $\lambda$  is the mean of a Poisson distribution, there are an average of 2.3 copies of the target nucleic acid sequence in an aliquot of the positive control.

**Example 5: analysis of clusters.** In a run of 19 specimens, there are two positive results. Both the positive results, of note, occur in consecutive specimens, forming a cluster. Is this the result of contamination? What if this occurs with three consecutive specimens instead of two?

**Analysis of example 5.** This is analogous to the problem, frequently presented in elementary probability texts, of removing  $n$  balls from an urn (without replacement) in which  $r$  balls are white and  $(n - r)$  balls are black. What is the probability that the  $r$  white balls are all removed consecutively in such a setting?

The general solution for this problem can be calculated by determining the number of ways that the  $r$  consecutive positive specimens can occur during a run of  $n$  total specimens and dividing this by the number of ways that  $r$  positive specimens can be selected from  $n$  total specimens.

The general solution to this problem is

$$P = \frac{(n - r + 1)}{n!/r!(n - r)!}$$

For example, in an LCR run of 19 samples, the probability that both of two positive samples will occur consecutively is

$$P = \frac{(19 - 2 + 1)}{19!/2!(19 - 2)!} = \frac{18}{171}$$

Therefore,  $P = 0.105$  or 10.5%.

By contrast, the probability that all of the three positive results will occur consecutively would be

$$P = \frac{(19 - 3 + 1)}{19!/3!(19 - 3)!} = 0.018 \text{ or } 1.8\%$$

**Interpretation of results for example 5.** A cluster in which both positive results occur consecutively could certainly be due to chance alone and would be anticipated to occur approximately one time in nine. A cluster in which all positive results occur consecutively is unlikely to be due to chance alone if there are three positive results among 19 samples, occurring less than 1 time in 50. Rather than immediately discarding the results of the assay as due to contamination, which is certainly a possible cause of the results, the underlying assumptions must be evaluated. Such an evaluation may demonstrate an alternative explanation.

Examples of alternative explanations include (i) a lack of independence of the three positive samples, such as would be the case for three specimens originating from the same patient or from a patient and two sexual contacts of the patient and (ii) the three positive specimens' being from a population at very high risk of chlamydial infection while the remainder of the specimens are from a low-risk patient population.

**Summary.** The use of principles of elementary probability theory can help identify incidents of contamination in nucleic acid amplification assays. In practice, a simple table can be constructed based upon the methods used here and the particular probabilities in the given laboratory, which can be incorporated into the amplification procedure. Such a table would give those situations which, due to the low calculated probability of the situation occurring, require additional review.

The methods described to mathematically evaluate these incidents require knowledge of the binomial and Poisson distributions and simple combinatorics, topics that are typically included in the first portion of a standard college-level probability textbook (1) and should, therefore, be easily accessible to microbiologists and molecular biologists. In addition, knowl-

edge of the underlying assumptions such as the population(s) from which the specimens are obtained and independence of the results of the assay for different specimens are essential to further evaluate those assay runs that arouse suspicion of contamination.

#### REFERENCES

1. **Hoel, P. G., S. C. Port, and C. J. Stone.** 1971. Introduction to probability theory. Houghton Mifflin Company, Boston, Mass.
2. **National Committee for Clinical Laboratory Standards.** 1995. Molecular diagnostic methods for infectious diseases. Approved Guideline NCCLS document MM3-A. National Committee for Clinical Laboratory Standards, Wayne, Pa.
3. **Rodrigo, A. G., P. C. Goracke, K. Rowhanian, and J. I. Mullins.** 1997. Quantitation of target molecules from polymerase chain reaction-based limiting dilution assays. *AIDS Res. Hum. Retroviruses* **13**:737-742.