CLINICAL AND LABORATORY STANDARDS INSTITUTE®

3rd Edition

EP12

Evaluation of Qualitative, Binary Output Examination Performance

This guideline includes descriptions of the types of qualitative, binary output examinations and procedures for evaluating their performance.

A guideline for global application developed through the Clinical and Laboratory Standards Institute consensus process.

EP12-Ed3
March 2023
Replaces EP12-A2

Evaluation of Qualitative, Binary Output Examination Performance

Jeffrey R. Budd, PhD
Karl De Vore, BA, SSBB
Ralf C. Bollhagen, PhD
Abdel-Baset Halim, PharmD, PhD, DABCC
Paul S. Horn, PhD
Marina V. Kondratovich, PhD
Qin Li, PhD
Kristen Meier, PhD
Gene Pennello, PhD
Diane M. Ward, PhD

Abstract

Clinical and Laboratory Standards Institute guideline EP12—*Evaluation of Qualitative, Binary Output Examination Performance* describes the categories of qualitative, binary output examinations and covers their performance evaluations for imprecision, including estimating C5 and C95, clinical performance (sensitivity and specificity), and stability and interferences.

Clinical and Laboratory Standards Institute (CLSI). *Evaluation of Qualitative, Binary Output Examination Performance.* 3rd ed. CLSI guideline EP12 (ISBN 978-1-68440-176-5 [Print]; ISBN 978-1-68440-177-2 [Electronic]). Clinical and Laboratory Standards Institute, USA, 2023.

CLINICAL AND LABORATORY STANDARDS INSTITUTE®

Suggested Citation

CLSI. *Evaluation of Qualitative, Binary Output Examination Performance*. 3rd ed. CLSI guideline EP12. Clinical and Laboratory Standards Institute; 2023.

Previous Editions:
August 2002, January 2008

EP12-Ed3
ISBN 978-1-68440-176-5 (Print)
ISBN 978-1-68440-177-2 (Electronic)
ISSN 1558-6502 (Print)
ISSN 2162-2914 (Electronic)
Volume 43, Number 4

Contents

Contents (Continued)

Foreword

In vitro diagnostic (IVD) examinations report either the value or the characteristics of a clinical property. Quantitative examinations (measurement procedures) measure and report the value of a property of clinical samples. Other examinations of clinical samples report characteristics of a property by placing them into two binary categories: unordered (nominal) and ordered (ordinal). This guideline covers qualitative examinations that, in the user's hands, provide binary (eg, yes or no, positive or negative), nominal outputs (see Appendix A for examples). These examinations span a wide range of medical laboratory specialties, medical purposes, measurement technologies, and types of reported results. A binary response is created based on:

- A device's internal continuous response and a cutoff to provide binary results
- Algorithmic decision-making techniques that detect whether an analyte is present
- In some cases, a yes or no output without the aid of instrumentation

The performance of an IVD examination should be assessed during and after its development, and its performance should be validated before any examination results are used to make clinical decisions. This guideline covers the development of qualitative, binary, results-reporting or output examinations (referred to as qualitative, binary examinations throughout) and is intended to promote uniformity in performance assessment among:

- Developers of qualitative, binary examinations for:
  - Designing and developing examinations
  - Establishing and validating examination performance based on how an examination is designed
- Laboratories that verify qualitative, binary examinations before they are placed into service
- Laboratories that develop their own qualitative, binary examination(s)

Many quantitative examinations provide measurand values in units plus a decision threshold that can be applied to obtain a binary interpretation (see examination examples in Appendix A). Although the binary output performance of these examinations can be determined using the methods described in this guideline, performance evaluations designed for quantitative methods provide more flexibility and have more power to detect differences in performance. Therefore, performance assessment of quantitative examinations should be based on the guidelines described in CLSI document EP19.[1]

Overview of Changes

This guideline replaces the previous edition of the approved guideline, EP12-A2, published in 2008. Several changes were made in this edition, including:

- Expanding the types of procedures covered to reflect ongoing advances in laboratory medicine
- Adding protocols to be used by developers, including commercial manufacturers or medical laboratories, during examination procedure design as well as for validation and verification
- Adding topics such as stability and interferences to the existing coverage of the assessment of precision and clinical performance (or examination agreement)
- Moving most of the statistical details, including equations, to the appendixes

NOTE: The content of this guideline is supported by the CLSI consensus process and does not necessarily reflect the views of any single individual or organization.

KEY WORDS

binary reporting
clinical sensitivity
clinical specificity
interferences
negative likelihood ratio
negative predictive value
positive likelihood ratio
positive predictive value
precision
qualitative examination
stability

Evaluation of Qualitative, Binary Output Examination Performance

1 Introduction

1.1 Scope

EP12 provides product design guidance and protocols for performance evaluation of the Establishment and Implementation Stages of the Test Life Phases Model of examinations (see CLSI document EP19[1]). EP12 characterizes a target condition (TC) with only two possible outputs (eg, positive or negative, present or absent, reactive or nonreactive). EP12 is written for both manufacturers of qualitative, binary, results-reporting or output examinations (referred to as qualitative, binary examinations throughout) and medical laboratories that create laboratory-developed, binary examinations (both termed developers). These protocols are also intended to help users verify examination performance in their own testing environment. Performance evaluation of examinations that provide outputs with more than two possible categories in an unordered (nominal) set or that report ordinal categories are outside the scope of this guideline.

1.2 Background

It is often necessary to provide test method results that have binary outputs (eg, yes or no). Health care providers may want to order an examination to help determine whether a disease is present in a patient, an analyte is present in a sample, a woman is pregnant, or the results of a drug test are positive. The two primary evaluations used for such examinations are clinical performance (sensitivity and specificity) and precision.

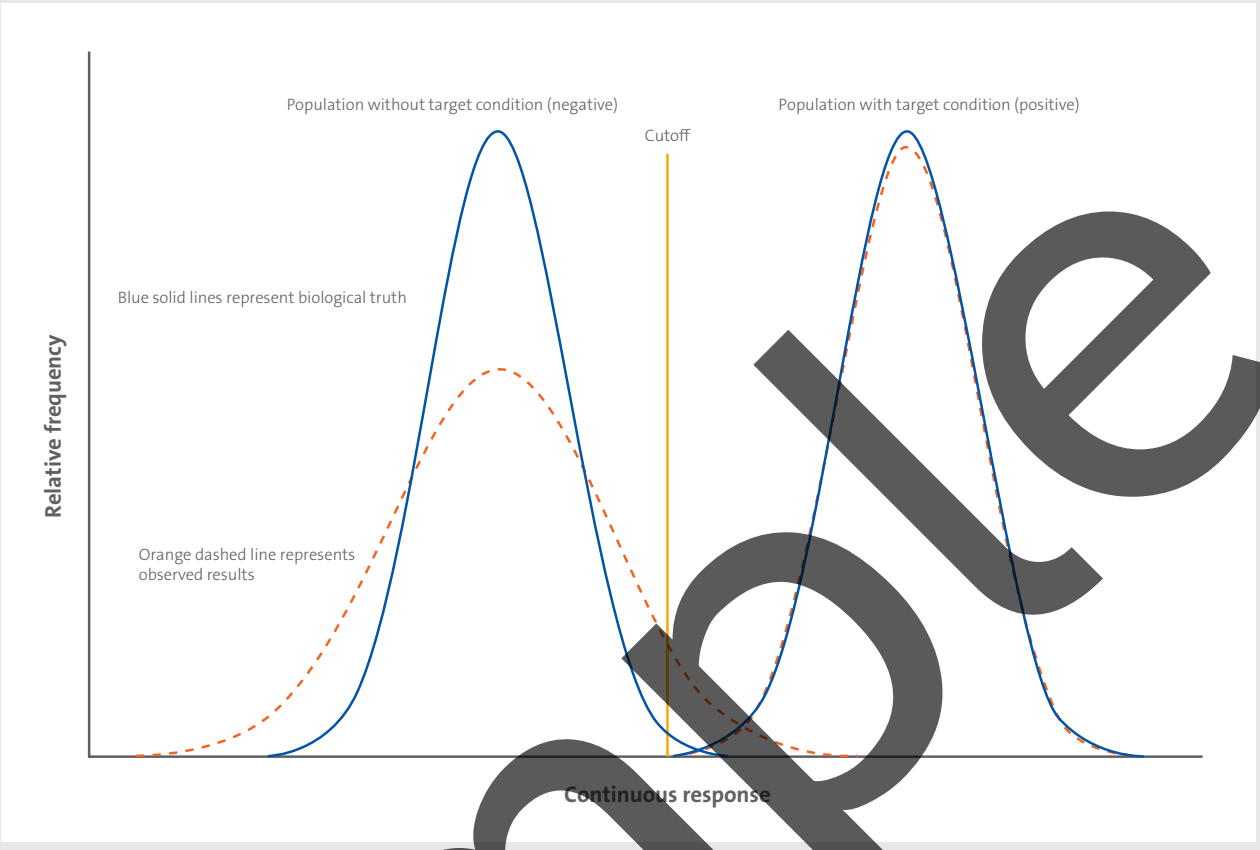
Assessments of clinical performance during the establishment and validation of a binary categorization process are performed by examining samples from subjects with a known category (ie, TC present vs TC absent). Ideally this analysis involves the comparison of binary results for an intended-use population of samples from the candidate examination with an independent procedure that provides the best available assessment of TC. However, in many cases, this comparison of binary results is an evaluation of agreement to positive or negative results from a comparative examination that is not such an assessment procedure.

Qualitative examination determinants of imprecision depend on the idea of the values in a relevant scale where a binary examination declares a sample to be positive 5% of the time (C5) and 95% of the time (C95). Unlike clinical performance, the means of determining C5 and C95 are different depending on how the binary categorization is performed (see Subchapters 3.1.3 and 3.2).

Reagent stability and examination interferences are two other performance attributes of examination systems. Although both topics are covered in CLSI documents EP07[2] (interferences) and EP25[3] (stability), there are several considerations unique to qualitative examinations that warrant additional coverage in EP12. See Subchapters 3.4 (stability) and 4.3 (interferences) for more information on these topics.

1.3 Standard Precautions

Because it is often impossible to know what isolates or specimens might be infectious, all patient and laboratory specimens are treated as infectious and handled according to "standard precautions." Standard precautions are guidelines that combine the major features of "universal precautions and body substance isolation" practices. Standard precautions cover the transmission of all known infectious agents and thus are more comprehensive than universal precautions, which are intended to apply only to transmission of bloodborne pathogens. Published



Abbreviation: TC, target condition.

Figure 4. Comparison of Misclassification Incidence for a Qualitative Examination With Subpopulations With and Without the TC

Because it is impossible to evaluate a sample from every individual in the intended-use population, the measurement procedure's total variability around the cutoff can be compared with the allowable variability that has been calculated. The concepts of C5 and C95, which are described in Subchapter 3.1.2, can be useful for this comparison. The C5 and C95 reflect the variability or SD of the examination under stipulated precision conditions (study design) and can be used to determine whether the variability will result in an unacceptable rate of misclassification for the examination. Knowing the allowable variability for samples both from individuals with and without the TC (as derived from the acceptable rate of misclassification), the allowable C5 to C95 interval can be calculated for any cutoff. If the measured C5 to C95 interval is wider than this allowable interval, the rate of misclassification may not be acceptable, and the imprecision of the ICR should be reduced to improve examination performance.

During examination development, the cutoff on the ICR scale should be determined using samples representative of the intended-use population. A range of cutoffs should be explored and the cutoff that satisfies clinical performance requirements (see Figure 3) should be chosen; this period of exploration is often referred to as "training." For more information, see CLSI document EP24.[15] Then, after the cutoff is locked, the examination's clinical performance should be assessed in a separate study with a different, typically larger sample set from the intended-use population, as described in Subchapter 4.2.

4.1.3 Binary Report Example

All types of qualitative examinations can report study results as the frequency at which a sample is declared positive, negative, or invalid. For example, in a study design that evaluates three reagent lots with two operators and two instruments over five days (see Table 3 for daily logistics), the binary results over all days for samples at different levels (eg, C5, C95, and C100) are summarized in Table 4. Although not shown in Table 4, the analyte-detection examination goal for a sample with no analyte is to never be declared positive. Examples of within-laboratory and reproducibility study designs are provided in Appendix F.

Table 3. Number of Replicates for a Within-Laboratory Precision Study Design

	Day 1											
	Reagent Lot 1				Reagent Lot 2				Reagent Lot 3			
	Operator 1		Operator 2		Operator 1		Operator 2		Operator 1		Operator 2	
	Inst 1	Inst 2	Inst 1	Inst 2	Inst 1	Inst 2	Inst 1	Inst 2	Inst 1	Inst 2	Inst 1	Inst 2
Reps	2	2	2	2	2	2	2	2	2	2	2	2

Abbreviations: inst, instrument; reps, replicates.

Table 4. Example Count of Binary Study Results

Source	N	Level 1, C5			Level 2, C95			Level 3, C100		
		Positive	Negative	Invalid	Positive	Negative	Invalid	Positive	Negative	Invalid
Combined	120	6	114	0	113	7	0	119	0	1
Day 1	24	2	22	0	23	1	0	23	0	1
Day 2	24	1	23	0	22	2	0	24	0	0
Day 3	24	1	23	0	22	2	0	24	0	0
Day 4	24	1	23	0	23	1	0	24	0	0
Day 5	24	1	23	0	23	1	0	24	0	0
Reagent lot 1	40	2	38	0	37	3	0	40	0	0
Reagent lot 2	40	2	38	0	38	2	0	40	0	0
Reagent lot 3	40	2	38	0	38	2	0	39	0	1
Operator 1	60	3	57	0	57	3	0	60	0	0
Operator 2	60	3	57	0	56	4	0	59	0	1
Instrument 1	60	3	57	0	56	4	0	60	0	0
Instrument 2	60	3	57	0	57	3	0	59	0	1

Abbreviations: C5, the value in a relevant scale where a binary examination declares a sample to be positive 5% of the time; C95, the value in a relevant scale where a binary examination declares a sample to be positive 95% of the time; C100, the value in a relevant scale where a binary examination declares a sample to be positive 100% of the time; N, number of total replicates per level.