DETERMINING AND VALIDATING GERMLINE VARIANT CONFIRMATION CRITERIA

A Companion White Paper to *Recommendations for Next-Generation Sequencing Germline Variant Confirmation: A Joint Report of the Association for Molecular Pathology and National Society of Genetic Counselors* (Crooks *et. al*., 2023) https://doi.org/10.1016/j.jmoldx.2023.03.012

By Members of the AMP Germline Variant Confirmation Workgroup
Stephen E. Lincoln, Kelly D. Farwell Hagman, Diana Mandelker, Avni Santani,
Ryan J. Schmidt, Robyn L. Temple-Smolkin, Kristy R. Crooks

Standard of practice is not defined by this article, and there may be alternatives. See Disclaimer for further details.

INTRODUCTION

The *Recommendations for Next-Generation Sequencing Germline Variant Confirmation: A Joint Report of the Association for Molecular Pathology and National Society of Genetic Counselors* (Crooks *et. al*., 2023) provides laboratories with guidance on a variety of topics related to the use of orthogonal methods to confirm germline variants uncovered using next-generation sequencing (NGS). Recommendation 5 in that guideline suggests that laboratories may establish technical and medical criteria that determine which variant calls should be subject to confirmation and which calls may not require confirmation. This document describes additional detailed considerations to aid clinical laboratories in determining and validating technical criteria consistent with this recommendation. For this white paper, the authors make the assumption that the reader is familiar with the Crooks *et. al*., 2023 publication, and specifically with its Recommendation 5.

In this document we provide:

1. General guidance for studies that (a) establish and then (b) validate technical criteria
2. Details on recommended calculations
3. Details regarding recommended specimens
4. Examples

This document includes Table 1 embedded. Additional tables (2-5) are provided in the accompanying Microsoft Excel file available at https://www.amp.org/resources/validation-resources/ (last accessed 5/15/2023).

1. GENERAL GUIDANCE

A study to establish technical criteria for orthogonal confirmation, and to set thresholds used in those criteria (*e.g.*, minimum quality scores), is typically performed during test development or test optimization phases of implementing a clinical test. Such studies should:

- Examine NGS variant calls made by the clinical laboratory's process in a set of specimens.
- Use high-quality orthogonal data to determine which of these variant calls are true positives (TPs) and which are false positives (FPs)

- Examine various quality metrics and determine thresholds that identify (or “flag”) all false positive variants in this data set. These criteria will likely also capture some true positives, and a goal may be to minimize the number of true positives flagged along with all the false positives. Optimization algorithms (including, potentially, machine learning methods) may be used in this step.
- These indicators and thresholds become the proposed criteria for determining which variant calls do or do not require confirmation on technical grounds.

To validate such proposed criteria in a following study, one should:

- Examine a separate set of NGS variant calls made by the laboratory in a set of specimens.
- Analyze just those NGS variant calls that meet or exceed the proposed technical criteria for not requiring confirmation.
- Use orthogonal data to determine which of these variant calls are true positives (TPs) and which are false positives (FPs)
- If the proposed criteria correctly identify the false positives as requiring confirmation, with an appropriate level of statistical significance, the criteria would be considered valid.

We make the following recommendations regarding such studies:

1. An appropriate **study design** should be used. Most importantly, to avoid **overfitting**, separate data sets should be used to (a) establish criteria (selecting quality metrics and thresholds), and (b) later validate those criteria. Alternatively, a study design using repeated training/test splits may be applied to accomplish both steps together.
2. Studies should separately consider **variants by class**. This includes separating (for example) SNVs from indels, which often have quite different error rates and root causes of errors in NGS. Also, variants in repetitive genomic regions (*i.e.*, low-complexity or homologous sequences) may be considered separate classes, as they too have different error rates and error causes compared to variants outside such regions.
3. The variants examined in the study should be **representative** of the full spectrum of pathogenic variation expected for the genes in the clinical test. Confirmation should always be required for any variant class where insufficient validation data are available to achieve statistical significance. For example, a validation study may have enough data (*i.e.*, NGS variant calls with orthogonal data) to demonstrate that SNVs meeting certain criteria do not require confirmation. But the same study may not have enough data to demonstrate that particular indels or CNVs do not require confirmation. Similarly, adequate data may not be available to validate whether variant calls in repetitive regions do not require confirmation.
4. The NGS variant calls used in the study should be generated using the same methodology that is used in diagnostic testing. Among other methodological aspects, the laboratory’s usual technical **filtering criteria**, used to remove highly likely false positives (see Crooks *et. al.*, 2023 Figure 1)

should be used. It is not appropriate to lower filtering criteria in order to admit more variants into a study -- such data will not be representative of the variants in clinical practice.

5. Validation studies may include variants **regardless of clinical interpretation**: Benign and pathogenic variants of the same class can provide equal information about the analytic performance of the NGS methods for that variant class. Studies may also incorporate **off-target variant calls**, which would not normally be clinically reported because of their location but still can be informative regarding the accuracy of NGS methods. An important caveat however is that off-target variants need to meet the same quality standards [read depth, quality scores, etc.] as on-target variants in order for these data to be representative. Because repetitive sequences are common outside of exons, considering repeat-associated variants as a separate class is recommended when using off-target variant calls.
6. Informative calculations of validation results should be used, including the **analytic PPV** (**aPPV**) of variants for which confirmation may not be needed. **Confidence intervals** (CI) must be calculated for these metrics, and the **lower bound** of the confidence interval should be used as an appropriate statistical indicator of the study's results. Details are provided below.
7. The **level of performance** that should be demonstrated depends on a variety of factors, described below. One notable factor is the size of the test's target regions (*i.e.*, single gene, panel, exome, etc.), because the chance of a false positive in any patient increases with the size of the test. Another factor is the expected patient true positive rate (PTPR), which links analytic PPV to clinical PPV, the chance that a positive test report is truly, and not falsely, positive.
8. An **adequately sized** validation study (measured in number of variants, not necessarily the number of samples) is needed given these factors. Of note, more variants can be required to rigorously demonstrate aPPV than are typically required to demonstrate sensitivity. Resources such as the Genome in a Bottle specimens (described below) can help make such studies practical even for smaller laboratories.

2. DETAILS ON RESULT CALCULATIONS

Validation study results can be represented as a standard 2x2 contingency table, where each cell contains a count of variant calls that have both NGS and reliable orthogonal data. This 2x2 table should be constructed separately for variants in each class, as defined above:

Table 1: Contingency table for use in validation studies

Variant calls that...	NGS True Positives (TP)	NGS False Positives (FP)
Meet or exceed the proposed technical criteria ("**unflagged**" variants). These calls would potentially not require confirmation.	A	B
Fail one or more technical criteria ("**flagged**" variants). These calls would require confirmation if reported.	C	D

Two metrics are recommended to summarize these results, which may be used independently or together:

aPPV, the analytic PPV of unflagged variant calls = A / (A + B)

FPsens, the sensitivity of the proposed criteria to flag FPs = D / (D + B)

A spreadsheet that performs these calculations is provided as Table 2 (in accompanying Excel file).

Ideally, both aPPV and FPsens will be 100% (*i.e.*, cell B in the above table should ideally be zero, indicating that no NGS FPs would escape confirmation using the proposed criteria). We recommend against using measures such as specificity or FPR (false positive rate) in studies of confirmation, for similar reasons as are discussed elsewhere[1] . Blended metrics that combine FP and FN errors, such as F0 and overall accuracy (OA), are also not recommended for validation studies of confirmation criteria.

A challenge with the FPsens metric is that few false positives are typically present in NGS data that have been properly filtered. Thus, the FPsens calculation can require a very large study data set in order to achieve statistical significance (*i.e.*, to have a narrow confidence interval). For this reason, aPPV may be the more commonly used metric, although studies with hundreds or thousands of variant calls can still be required in order to demonstrate adequate performance to an appropriate degree of statistical significance (see below).

For both aPPV and FPsens, we strongly recommend using the **lower bound of the 95% confidence interval** when evaluating and presenting study results. This recommendation is similar to that of other AMP guidelines[1,2] which recommend that a minimum number of variants be used in measuring sensitivity such that detecting 100% of these variants during validation demonstrates at least 95% sensitivity at p=0.05.

Because well-performing technical criteria will have aPPV and FPsens values at or near 100%, the CI should be computed using a statistical method which produces robust results for extreme rates. This is true for the **Wilson Score**, **Jeffreys,** and **Tolerance Interval** methods, for example (which all produce roughly similar bounds[3]). This is not true however for the traditional (Wald) method often computed using the built-in functions in Microsoft Excel.

Table 3 in the accompanying Excel file provides the Tolerance Interval calculation[1,2] for various numbers of variants when the observed (point estimate) aPPV is 100%. This table may be useful for planning a validation study, particularly for determining the number of variants needed to rigorously demonstrate a particular aPPV level. Calculators for the Wilson Score and Jeffreys Cis, which do not require an aPPV of 100%, are available on the internet.

The calculation of aPPV or its lower bound does not depend on parameters other than those in Tables 1 and 2 above, although determining the aPPV level that should be required for any test does. We recommend setting performance requirements using the **clinical PPV (cPPV) lower bound**, which can be estimated from the aPPV lower bound using three additional parameters:

1. The **size** of the test's reportable target regions. A larger target gives many more opportunities for an FP to occur in each tested patient, thus a higher aPPV is required for larger test targets in order to limit the expected number of FPs per patient.

2. The assumed likelihood that a FP variant call will appear **reportable** (**FPrep**). Note that FP variant calls are typically not filtered out by processes that remove common variants based on population allele frequency. Thus, many FPs (including many FP SNVs) can be reported as VUS and some will appear pathogenic. The fraction of variant calls that are reported may be much higher for FPs than it is for TPs.

3. The expected **patient true positive rate (PTPR)** in the population undergoing testing. The specific metric used is the expected number of analytic true positive reportable variants in each patient.

An important nuance is that the parameters 2 (FPrep) and 3 (PTPR) should be calculated on a similar basis. For example, if VUS are reported by the test and (per lab policy) would potentially be subject to confirmation, then both parameters should include VUS. If only pathogenic variants are reported (*e.g.*, in a screening test) then both parameters should include only pathogenic variants.

As a general recommendation, demonstrating observed (point estimate) aPPV and cPPV levels of 100%, and demonstrating cPPV lower bounds of at least 90 – 95% is often appropriate. (Recall that a cPPV of 90%, for example, means that one out of every 10 positive clinical reports will be false). Depending on the test, achieving this can require demonstrating a much higher aPPV lower bound, which requires a large study (see section IV below for examples).

Demonstrating even higher cPPV bounds may be important, depending on the clinical consequences of an analytic false positive appearing in a test report. Specific cPPV thresholds may be best determined by professional guidelines that focus on individual disease areas.

Exome- and genome-based tests present a special challenge, as multiple unfiltered FPs will likely arise in each patient. In typical use however, only a small fraction of genes (a virtual panel) are analyzed for clinical reporting in any patient's exome or genome sequence. Thus, it may be appropriate to calculate cPPV bounds using the expected size of a virtual panel, not the full assay target size for exomes and genomes.

3. DETAILS REGARDING SPECIMENS

Validation studies of confirmation criteria may use clinical data (*i.e*., variant calls in patient specimens with orthogonal confirmatory data) and/or data from reference specimens that have high accuracy, independent variant call sets available.

The Genome in a Bottle (GIAB) samples can be a particularly useful component of such studies, as sequencing a small number of GIAB samples can contribute many variants to the study. As of this writing, seven GIAB samples are available (https://www.nist.gov/programs-projects/genome-bottle, last accessed 05/15/2023)[4–6]. Each GIAB specimen has high accuracy variant calls available for most (approximately 90%) of the genome.

Importantly the GIAB data sets provide "reference calls," indicating which regions in each sample are confidently known to *not* contain any variants (said differently, these regions are known to be homozygous for the allele present in the GRCh reference human genome sequence). This information allows comparisons between the lab's NGS variant calls and the comparison data set to identify NGS calls that are likely false positives, and to distinguish these FPs from NGS calls where the reference specimen's genotype is uncertain. In the GIAB data, this information is available in the "benchmark region" BED format files (referred to as "high confidence regions" in the older GIAB publications). GIAB specimen analysis should always be limited to these regions.

An important caveat to using the GIAB samples is that the vast majority of variants present in protein coding exons of these samples are SNVs (this is particularly true in protein coding exons of clinically important genes). Few indels, CNVs or SVs are present in these exons, as is expected for individuals without a clinical indication. Studies that rely on the GIAB data alone may be unable to validate performance of confirmation criteria on indels, CNVs or complex variant types without the use of additional samples.

Another important caveat is that many clinically important genes are not yet fully characterized by the GIAB consortium[3,7,8] in these samples. Unfortunately, these uncharacterized regions tend to have the highest NGS error rates and present unique challenges for NGS sequencing. Supplementing a validation study with additional clinical specimens is likely required if the laboratory wishes to exclude variants in these regions from requiring confirmation.

Table 4 in the accompanying Excel file estimates the contribution of the GIAB specimens to a validation study. For smaller tests, supplementing the GIAB samples with additional data will be required to achieve an adequate aPPV bound, particularly if only on-target variant calls are used. For larger tests, more GIAB variants are available, and supplementing GIAB will be required mostly for variant types that are poorly represented in the GIAB data (*e.g.*, indels, CNVs).

Specimens may be included in such a study in replicate. However, we recommend against using any one specimen more than twice in the same study, as diversity of variants is required to have confidence in the study's results.

4. EXAMPLE VALIDATION STUDIES

Table 5 below shows example study sizes and results (in number of variants) for a range of tests. A copy of this table, with additional details and the formulas embedded, is provided as Table 8 in accompanying Excel file.

Table 5: Examples of tests and validation studies

Test	Reportable Target Size	Number of Variants in Study	Observed aPPV (CI lower bound)	Clinical PPV corresponding to the aPPV lower bound at specified patient true positive rate (PTPR)
CFTR diagnostic test	3.5 Kb	60	100% (95.1%)	85% cPPV @ 25% PTPR
BRCA1/2 diagnostic test	10 Kb	300	100% (99.0%)	86% cPPV @ 15% PTPR
BRCA1/2 screening test	10 Kb	900	100% (99.7%)	86% cPPV @ 1.0% PTPR
40 Gene Panel, diagnostic test	50 Kb	450	100% (99.3%)	86% cPPV @ 50% PTPR
40 gene Panel, screening test	50 Kb	2200	100% (99.9%)	85% cPPV @ 2% PTPR
200 gene panel or phenotype driven exome slice, diagnostic test	200 Kb	2200	100% (99.9%)	85% @ 25% PTPR
“Medical exome” slice of ~3000 genes. screening test	3 Mb	2600	100% (99.9%)	85% @ 100% PTPR

Legend: See Table 8 in the accompanying Excel file for details. VUS are included in both the PTPR values and the cPPV estimates for diagnostic tests but not for screening tests. Also, PTPR values are assumed to be much lower for screening tests compared to diagnostic tests because of the lower yield in an unselected population. Study sizes were chosen to meet the minimum recommended cPPV bounds. Larger studies may be warranted depending on the clinical use of the test (see Crooks *et. al*., 2023 Technical Background).

DISCLAIMER

The Association for Molecular Pathology (AMP) Clinical Practice Guidelines and Reports are developed to be of assistance to laboratory and other health care professionals by providing guidance and recommendations for particular areas of practice. The Guidelines or Report should not be considered inclusive of all proper approaches or methods, or exclusive of others. The Guidelines or Report cannot guarantee any specific outcome, nor do they establish a standard of care. The Guidelines or Report are not intended to dictate the treatment of a particular patient. Treatment decisions must be made based on the independent judgment of health care providers and each patient’s individual circumstances. The AMP makes no warranty, express or implied, regarding the Guidelines or Report and specifically excludes any warranties of merchantability and fitness for a particular use or purpose. The AMP shall not be liable for direct, indirect, special, incidental, or consequential damages related to the use of the information contained herein.

REFERENCES

1. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding K V, Wang C, Carter AB: Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines. J Mol Diagnostics, Elsevier, 2018, 20:4–27.

2. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding K V, Nikiforova MN: Guidelines for Validation of Next-Generation Sequencing&#x2013;Based Oncology Panels. J Mol Diagnostics, Elsevier, 2017, 19:341–365.

3. Lincoln SE, Truty R, Lin C-F, Zook JM, Paul J, Ramey VH, Salit M, Rehm HL, Nussbaum RL, Lebo MS: A Rigorous Interlaboratory Examination of the Need to Confirm Next-Generation Sequencing&#x2013;Detected Variants with an Orthogonal Method&#xa0;in Clinical Genetic Testing. J Mol Diagnostics, Elsevier, 2019, 21:318–329.

4. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M: Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 2014, 32:246–251.

5. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre ABR, Chandramohan D, Chen F, Jaeger E, Moshrefi A, Pham K, Stedman W, Liang T, Saghbini M, Dzakula Z, Hastie A, Cao H, Deikus G, Schadt E, Sebra R, Bashir A, Truty RM, Chang CC, Gulbahce N, Zhao K, Ghosh S, Hyland F, Fu Y, Chaisson M, Xiao C, Trow J, Sherry ST, Zaranek AW, Ball M, Bobe J, Estep P, Church GM, Marks P, Kyriazopoulou-Panagiotopoulou S, Zheng GXY, Schnall-Levin M, Ordonez HS, Mudivarti PA, Giorda K, Sheng Y, Rypdal KB, Salit M: Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 2016, 3:160025.

6. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, De La Vega FM, Xiao C, Sherry S, Salit M: An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol 2019, 37:561–566.

7. Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, Salit M, Ashley EA: Medical implications of technical accuracy in genome sequencing. Genome Med 2016, 8:24.

8. Lincoln SE, Hambuch T, Zook JM, Bristow SL, Hatchell K, Truty R, Kennemer M, Shirts BH, Fellowes A, Chowdhury S, Klee EW, Mahamdallie S, Cleveland MH, Vallone PM, Ding Y, Seal S, DeSilva W, Tomson FL, Huang C, Garlick RK, Rahman N, Salit M, Kingsmore SF, Ferber MJ, Aradhya S, Nussbaum RL: One in seven pathogenic variants can be challenging to detect by NGS: An analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. Genet Med 2021, Online:Ahead of print.