Clinical quality strategy to assess data integrity of germline variants inferred from tumor-only testing sequencing data

Timothé Ménard (timothe.menard@roche.com)
F. Hoffmann-La Roche https://orcid.org/0000-0003-4545-6944
Donato Rolo
Roche Products Ltd
Björn Koneswarakantha
F. Hoffmann-La Roche https://orcid.org/0000-0003-4585-7799

Short Report

Keywords: clinical quality, data accuracy, data integrity, cancer genetics, germline variants

Posted Date: April 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-418086/v1

Version of Record: A version of this preprint was published at Pharmaceutical Medicine on August 26th, 2021. See the published version at https://doi.org/10.1007/s40290-021-00399-4.

Abstract

In the majority of cancers, pathogenic variants are only found at the level of the tumor; however, an unusual number of cancers and/or diagnoses at an early age in a single family may suggest a genetic predisposition. Predisposition plays a major role in about 5 to 10% of adult cancers and in certain childhood tumors. As access to genomic testing for cancer patients continues to expand, the identification of Potential Germline Pathogenic Variants (PGPV) through tumor-DNA sequencing is also increasing. Statistical methods have been developed to infer the presence of a PGPV without the need of a matched normal sample. These methods are mainly used for exploratory research, for example in real-world Clinico-Genomic Databases/platforms (CGDB). These databases are being developed to support many applications such as targeted drug development, clinical trial optimization and post marketing studies. To ensure the integrity of data used for research, a Quality Management System (QMS) should be established, and Quality Oversight Activities (QOAs) should be conducted to assess and mitigate clinical quality risks (for patient safety and data integrity). As opposed to well-defined GxP areas such as Good Clinical Practice (GCP), there are no comprehensive instructions on how to assess the clinical quality of statistically derived variables from sequencing data such as PGPV. In this report, we aim to share our strategy and propose a possible set of tactics to assess the PGPV quality and to ensure data integrity in exploratory research.

1. Background

Cancer is a genetic disease. In the majority of cancers, pathogenic variants are only found at the level of the tumor (somatic variants); however, an unusual number of cancers and/or diagnosis at an early age in a single family may suggest a genetic predisposition. Predispositions play a major role in about 5 to 10% of adult cancer and in certain childhood tumors [1]. For example, BRCA1 and BRCA2 are involved in homologous recombination (HR), DNA repair and are germline Cancer Predisposition Genes (CPG) that result in a syndrome of hereditary breast and ovarian cancer (HBOC) [2].

Access to genomic testing in oncology continues to expand for treatment recommendation, disease monitoring and early detection [2]. Identification of Potential Germline Pathogenic Variants (PGPV) through tumor-DNA sequencing is also increasing with both the European Society of Medical Oncology (ESMO) [3] and the American College of Medical Genetics (ACMG) [4] having issued recommendations on how to report and analyze PGPV derived from tumor-only sequencing data.

In routine care, testing for PGPV is triggered by patient medical and family history and follows local guidelines and recommendations [5]. Testing is performed on two independent normal tissue samples (whole blood and buccal swab). Of note, there is a clear regulatory framework around testing (e.g., informed consent requirements), patient and patient family follow-up, as well as patient care. For cancer patients that undergo tumor testing, there are different strategies: tumor-only testing, tumor-normal paired testing with germline variant subtraction and tumor normal paired testing with analysis of genes associated with germline cancer predisposition [4].

In tumor-only testing, sequencing data may be used to infer germline presence of a specific variant from the Variant Allele Frequency (VAF). Germline variants on average have higher VAFs than somatic variants, for example a typical heterozygous germline variant should have a VAF of 50%. VAF in tumors is highly dependent on tumor purity and heterogeneity, and experimental thresholds need to be determined to classify germline variants [3]. Recently generalizable statistical modelling techniques have been developed to classify variants from tumor-only testing that do not require to set individual VAF thresholds for each gene [6, 7, 8, 9]. When a PGPV is inferred by tumor-only testing, recommendations are provided to the patient and their physicians, the results must be confirmed on a matched normal tissue sample, and genetic counseling is advised. Of note, these might not always be reimbursed and a PGPV could generate anxiety and stress for the patient waiting for confirmation [4].

In clinical practice, gatekeepers (e.g., normal match testing, genetic counseling, review by molecular tumor boards) must be in place to mitigate the risks to a patient's well-being and on data integrity. PGPV inferred from tumor-only testing can also be used for exploratory research, for example in real-world clinico-genomic databases/platforms (CGDB). These databases are being developed by pharmaceutical and diagnostics companies [10, 11] to support many applications such as targeted drug

development, clinical trial optimization and post marketing studies. To ensure the integrity of data used for research, a Quality Management System (QMS) should be established, where Quality Oversight Activities (QOAs) should be conducted to assess and mitigate clinical quality risks (for patient's safety and to data integrity) [12].

As opposed to well defined GxP areas such as Good Clinical Practice (GCP) [12], there are no comprehensive instructions on how to assess the clinical quality of derived variables from sequencing data such as PGPV. In this report, we share our strategy and a set of tactics to assess the quality of PGPV and to ensure data integrity in exploratory research. Our approach serves as a framework for quality professionals to partner with researchers and can be applied to statistical inference methods that analyse clinico-genomics data.

2. Methods

Prerequisites

The primary objective of this research is to provide a strategy and a set of tactics for clinical quality professionals to assess the quality (of the assessment of) PGPV used for exploratory research. They are not guidelines for analytical and/or clinical validation of methods used to infer PGPV from sequencing data, nor do they relate to clinical Quality Assurance (QA) regulatory requirements. The scope was tumor-only testing, where academic institutions and/or diagnostic companies do not use a normal match to verify that the variants found at the somatic level are also present in the germlines. Several methods have been developed, we selected four [6, 7, 8, 9] which had been published in peer-reviewed journals (see Table 1), and we used them as examples to illustrate our strategy. Of note, we did not review further inference methods that set individual VAF thresholds, as ACMG and ESMO [3, 4] guidelines already provided recommendations on how to assess their performance and limitations.

Table 1
Examples of methods for tumor-only testing PGPV inference

Authors	Method
Khiabanian *et al.* [6]	LOHGIC (statistical model)
Hiltemann *et al.* [7]	Virtual Normal (virtual dataset as normal match)
Park *et al.* [8]	ALFRED (statistical model)
Sun *et al.* [9]	SGZ (machine learning model)

Problem statement

To design a fit-for-purpose quality review, the problem statement and the scope should be clearly defined: getting assurance that PGPV data can be used with confidence to address the respective research questions, while being transparent on the data limitations and their impact on the analysis.

The proposed quality strategy to address the above problem statement followed a 2-step approach. First, the methods used should be assessed to understand the potential impact on the quality of the PGPV data, how the methods were built, and how they were evaluated. In a second phase, quality checks can be established if PGPV data are available with other clinico-genomics data.

Assessing the method used to infer PGPV

Details on methods used in exploratory research are sometimes available through peer-reviewed scientific publications [6, 7, 8, 9]. After an initial review step, we rejected the ALFRED method by Park *et al.* [8]. This method was designed to discover new CPG (i.e., never reported as such before) from tumor-only sequencing data, and used a statistical model to test the Knudson two hits hypothesis [13].

In order to assess the methods used to infer PGPV from tumor-only testing, the following areas should be considered for review.

a. Any model used for PGPV classification needs to undergo an expert review that addresses the model applicability to the current dataset considering the original sample population and model validation techniques. The published methodologies for PGPV assessment have been evaluated on a limited selection of tumor types and in restricted population samples. It is important that these limitations are disclosed and that documented expert opinion permits these methodologies to be applied to the specific tumor type and patient population at hand. The expert review should include a brief tabular summary as exemplified in Table 2.
b. Any PGPV classification modelling technique needs to be adequately validated. The validation strategy is dependent on the statistical method and should be included in the expert review. Classification metrics such as accuracy can be misleading for imbalanced data sets [14], we therefore recommend that the following performance metrics should be disclosed: true positive rate (TPR), false positive rate (FPR), true negative rate (TNR) and false negative rate (FNR) which allow to reconstruct the entire confusion matrix [14]. It also needs to be clear whether the classification estimate addresses the pathogenicity or the germline presence or both of these PGPV characteristics.
c. Requirements to ensure optimal performance of the inference method. For example, which breadth and depth of sequencing coverage are required, under which tumor purity level does the model perform best [3, 4].
d. Potential pitfalls of the inference method and how it could impact the quality of PGPV data. What risk mitigations could be taken (e.g., discard PGPV identified outside of the inference method specifications).

Quality checks using clinico-genomics data

For clinical QA professionals and researchers having access to clinico-genomics data (e.g., through CGDBs [10, 11]), there are a number of tactics that can be applied to further assess the quality of PGPV. For the purpose of this report, we considered alterations, histopathology and biomarker data. These quality checks can also be implemented for inference methods that set individual VAF thresholds [3, 4].

When biomarker data are available for the same patient, a concordance analysis can be performed. For example, data on germline testing of a normal blood sample for BRCA1/2 can be compared to the inferred PGPV for BRCA1/2.

To discard possible False Positive/False Negative (FP/FN) results and to identify other anomalies, data quality checks can be implemented. These checks should be based on the latest medical theory and clinical practice, reflecting empirical results. The tactics suggested below are examples and they can be complemented (with additional data and variables if available):

- Retrieving all data reported as PGPV across all tumor types, and comparing the data with a list of known CPG [1, 15]. PGPV associated with an unknown CPG would likely be a FP.
- If a founder variant (e.g., for BRCA1/2, APC, MSH2, MSH6, CHEK2, or MUTYH) is identified at the somatic level, it is likely to be found in the germlines [16]. This can help identify potential FN.
- When a pathogenic BRCA1/2 variant is present at somatic level, it is expected to be found also at the germline level in ~ 70–80% of the cases in Breast Cancer (BC) and in ~ 60% in Ovarian Cancer (OC) [17, 18]. The proportion of somatic BRCA1/2 variants also flagged as PGPV can be calculated and compared against these ratios.
- Germline CDH1 pathogenic variants are associated with an increased risk of lobular BC. In CDH1 associated BC, malignant cells show loss of adhesion (as CDH1 codes for E-cadherin) [19]. Hence by querying the associated histopathological data, we can identify false positive PGPV, i.e., CDH1 inferred PGPV associated with ductal BC.

3. Discussion

Inference methods

To ensure integrity of PGPV data used for research, it is important to understand how the method to infer PGPV was designed, i.e., what type of model was fitted and on which dataset, the requirements to ensure the model optimal performance (e.g., tumor specimen quality requirements) and its validation. Furthermore, reviewing available performance metrics and disclosed limitations can help tailoring the analysis further (See Table 2). For example, the method published by Hiltemann *et al.* [7] cannot adequately account for rare germline variants that are private to a family or small population; hence should these variants be flagged as PGPV, they would likely be FPs. Also, to avoid a FP and FN, the analysis should be adjusted on the method parameters and the correct filters should be applied, for example:

- If the model performs best on a defined range of tumor purity (e.g., for the SGZ method by Sun *et al.*, a high level of accuracy is maintained from 10% through 75% tumor purity [9], PGPV inferred from sequencing of specimens which have a tumor purity outside of the range can be excluded, to reduce FP and FN)
- If the model requires minimum sequencing depth (e.g., the method published by *Khiabanian et al.* [6] high depth sequencing coverage, at least over 500)
- Filter for specimen that passed quality checks (so called "qualified or valid" specimen)

Table 2
Tumor-only PGPV Inference methods review and assessment

	LOHGIC [6]	SGZ [9]	VN [7]
Model set-up and validation approach	Statistical model - two-steps validation approach: (a) Model was tested on tumor-only sequencing data from 64 cancer patients with candidate pathogenic BRCA1/2 mutations - LOHGIC's predictions were ambiguous for 12% of the patients (eight of 64 patients) (b) concordance analysis on 28/64 patients that had genetic testing reports (on normal blood sample)	Statistical model, three validation approaches: (a) 87 specimens from 30 non-small cell lung and colon cancer patients with matched normal where the true origin of all alterations was known (b) To assess the robustness of the method to different levels of tumor purity, three cancer cell lines were examined, which were titrated with their matched lymphoblastoid normal to six levels of tumor purity (10%, 20%, 30%, 40%, 50%, 75%) (c) 20182 clinical FFPE specimens with known somatic drivers where real-world somatic variant recovery was assessed	Not a statistical model. In absence of matched normal samples to confirm PGPV, this method was built using a set of 931 samples from healthy, unrelated individuals, originating from two different sequencing platforms, to serve as a Virtual Normal (VN). The goal is to reduce the FPR when a tumor-normal method (here Illumina's tumor-normal sequencing service) is used. In that, this method differs from LOHGIC and SGZ. For validation, comparisons were performed to public databases, e.g., COSMIC validated variants data. Comparisons were made separately and analyzed on different datasets for Structural Variations (SV), Single Nucleotide Variants (SNV) and insertions-deletions (indels)
Performance metrics	Accuracy = 93% Precision = 100% Recall = 96% Note: These metrics were obtained against a small sample composed of 28 cancer patients with concordance analysis.	Accuracy = 95–99% Confusion matrix (TP, TN, FP, FN) and derived metrics (e.g., precision, recall) not disclosed	Limited metrics disclosed - due to the set-up of the methods, primarily FPRs are available The method identifies 10–30% false-positive SVs and 20–30% FP SNVs and indels (vs. the tumor-normal method has 20–50% and 40–45% false positives for SVs and SNVs and indels, respectively)

	LOHGIC [6]	SGZ [9]	VN [7]
Requirements for optimal performance	High sequencing depth (> 500x) is required Strongly rely on the statistical confidence in assessing VAF (which is also dependant on the sequencing depth) and tumor purity Moreover, because accuracy in VAF estimates depends on sequencing depth, inferring Loss of Heterozygosity (LOH) is statistically robust at depths that provide sufficient confidence in measuring VAF within 1–5%	High sequencing depth (> 500x) and large coverage, using Massive Parallel Sequencing (MPS) are required High level of accuracy with 10–75% tumor purity; accuracy to drop if tumor content > 90% Adequate admixture of the surrounding normal tissue (at least 10% normal tissue, i.e., tumor content under 90%) Major misfit of the copy number model can lead to misclassification of somatic versus germline status, especially when tumor content is high, where the expected difference between germline and somatic allele frequency is reduced	The VN size should consist of minimum 200–400 genomes No other information regarding requirements for optimal performance is available

	LOHGIC [6]	SGZ [9]	VN [7]
Methods limitations, Potential impact on PGPV data, possible risk mitigation	BRCA1/2 PGPV only, not applicable to other CPG Presence of reversion mutations in BRCA1/2, which occur under acquired resistance to platinum and Poly(ADP-Ribose) Polymérase (PARP) inhibitors can also lead to an incorrect or ambiguous inference May be confounded by low tumor purity and insufficient sequencing data Risk of FP/FN PGPV if: - Low coverage; sequencing depth < 500x - Low tumor purity Consider to discard PGPV inferred from data sequenced outside the required specifications for optimal performance Additional quality data checks are recommended if clinico-genomics data are available	Risk of FP/FN PGPV if: - Low coverage; sequencing depth < 500x - and/or tumor purity outside of 10–75% range - and/or copy number model not accurate High accuracy ≠ good performance of the model if the data are imbalanced [14] Consider to discard PGPV inferred from data sequenced outside the required specifications for optimal performance Additional quality data checks are recommended if clinico-genomics data are available	Cannot adequately account for rare germline variants that are private to a family or small population, hence there is a risk of False Negative for rare germline variants Ensure that the risk of FN is taken into account / disclosed in the research analysis Additional quality data checks are recommended if clinico-genomics data are available

Data quality checks

Descriptive analytics can be used to facilitate the interpretation of the quality review, for example for reviewing the PGPV classification estimates by plotting them against expected population frequencies, which makes discrepancies easy to spot (Fig. 1).

The possible outcomes and their interpretations for the quality checks proposed in the Methods section are detailed in Table 3.

Table 3
Examples of data quality checks with clinico-genomics data

Quality checks	Possible outcome(s)	Interpretation(s)	Action(s)
Compare all flagged PGPVs to list of CPG	Variants flagged as PGPVs that have never been identified as CPG	FP	Root cause investigation Data should not be considered
Check list of selected founder variants (e.g. for BRCA1/2, APC, MSH2, MSH6, CHEK2, or MUTYH) that were not flagged as PGPV	Founder variant usually found in the germline but not flagged as PGPV	FN	Root cause investigation Account for potential bias Disclosure of potential bias
Ratio Germline/Somatic for BRCA1/2 in HBOC (BC ~ 80% / 20%) (OC ~ 66 % / 33 %)	Ratio not on par to what is expected in BC/OC population	Possible selection bias Small sample size, cannot be interpreted FP / FN	Root cause investigation Account for potential bias Disclosure of potential bias
Cross-checks (e.g. CDH1 in BC patients vs. histopathology data)	CDH1 PGPV associated to ductal BC	FP	Root cause investigation Data should not be considered
Concordance analysis on biomarker data (e.g., BRCA1/2 on normal match)	Discrepancies between normal match and inferred PGPV from tumor-only	FP FN	Root cause investigation Document decision on which data should be considered as correct

Quality strategy

The quality strategy is summarized in a flow chart (Fig. 2) that can be used by clinical quality professionals to partner with researchers for reviewing and documenting the quality of PGPV data.

It is up to the researchers, depending on the questions they need to answer, to follow-up on the proposed actions. In certain situations, some PGPV can be excluded from the analysis, but a thorough root cause investigation is recommended. It could help fixing the underlying issue that generates incorrect PGPV flagging but also avoids the recurrence of similar quality issues moving forward. As an alternative, if the root cause cannot be identified and/or the quality issues can be solved, we advise that the risks of using potentially biased data are taken into account, and mitigated where possible. We also advise full disclosure on the data limitations for transparency.

There are currently no binding regulatory requirements for clinical quality data management in exploratory research, but implementing this strategy could help accelerate research, by addressing some of the root causes of the data quality issues. Furthermore, should the decision be made to use PGPV data inferred from tumor-only testing for a regulatory filing, having assessed the method and applied quality checks could avoid having to implement QA measures retrospectively, and therefore save significant resources. Of note, if PGPV data are intended to be used for filing, all regulatory and quality requirements pertaining to process, systems and Computer System Validation (CSV) must be met [20].

Challenges

One of the key challenges with Real World Data (RWD) is that Source Data Verification (SDV) cannot always be performed. Therefore when a potential data quality issue is identified, it might be challenging to understand the root cause. Implementing methods for reviewing and assessing the quality of RWD, together with embedding automated quality checks, can reduce the risk of integrity of the data being compromised. A thorough documentation about the quality checks implemented, their review and decisions made to keep or discard RWD for analysis is highly recommended.

When using clinico-genomics data for verifying the quality of PGPV data, researchers should first assess and understand how the database they use has been designed and set-up. It is very likely that the data do not represent the general cancer population. For example:

- The database might be composed mainly of patients with late-stage disease and/or older patients.
- The samples of a particular cancer type might also be small, and therefore not representative.

Furthermore, as genomic testing is not widely accessible to all cancer patients, it is highly probable that there is an underlying selection bias. For example, comparing the demographics and the frequency of PGPV in BC, Ovarian Cancer (CR) and Colorectal Cancer (CRC) (which are cancer most frequently associated with CPG) [1], to a cohort selected from a CGDB might not be representative of what is found in real world. Also, younger-than-typical age at onset may suggest an underlying cancer predisposition, such as very early onset BC [21], but this might not be reflected in a cohort built from a CGDB.

Last but not least, should new methods be developed to infer PGPV from tumor-only testing, we encourage researchers to clearly describe its specifications and limitations to avoid false expectations/understanding. For example, in the four methods available in scientific publications [6, 7, 8, 9], not all relevant performance metrics were explicitly disclosed, and some additional features (e.g., confidence intervals, thresholds for accuracy) could have been useful.

4. Conclusion

In this report we defined a strategy and a set of tactics to assess the quality of PGPV data inferred from tumor-only testing to ensure the data integrity in exploratory research activities. Our method can be easily tailored to the specifics of the models used for PGPV inference. It can help researchers get further insights on the quality of PGPV data and implement quality by design for RWD [22]. Our strategy could be expanded to other machine-learning derived biomarkers from sequencing data (e.g., MSI status), although a deep understanding of the topic is always required in order to design the quality review appropriately. Our strategy will continue to be improved and further tactics for quality checks can be performed by using additional variables/data, such as patient family history [23]. Finally, this project is part of a broader effort to deliver QA effectively, by leveraging analytics and implement quality by design to accelerate patient access to innovative healthcare solutions [24, 25, 26]. Application of a tailored data quality strategy to identify and address quality issues will aid researchers in reproducing results, demonstrating integrity of the data collected/generated, and ultimately giving stakeholders confidence in the research outcome.

Declarations

Acknowledgements

Content review was performed by Christopher Ganter and Alaina Barros who were employed by Roche.

Conflicts of interest/Competing Interests

Timothé Ménard, Donato Rolo and Björn Koneswarakantha were employees of Roche at the time this research was completed.

Funding

Funding for this research was provided by Roche.

Authors contributions

TM proposed the research, designed the method and drafted the outline. All authors wrote the manuscript and produced the figures.

References

1. Hiltemann,S.,Jenster,G.,Trapman,J.,van derSpek,P.,&Stubbs,A.(2015).Discriminating somatic and germline mutations in tumor DNA samples without matching normals.Genome Research,25(9),1382–1390.doi:10.1101/gr.183053.114
2. Morganti,S.,Tarantino,P.,Ferraro,E.,D'Amico,P.,Duso,B.,&Curigliano,G.(2019).Next Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer. Translational Research And Onco-Omics ApplicationsInThe Era Of Cancer Personal Genomics,9–30.doi:10.1007/978-3-030-24100-1_2
3. Mandelker,D.,Donoghue,M.,Talukdar,S.,Bandlamudi,C.,Srinivasan,P.,&Vivek,M.etal.(2019).Germline-focussed analysis of tumour-only sequencing: recommendations from the ESMO Precision Medicine Working Group.Annals Of Oncology,30(8),1221–1231.doi:10.1093/annonc/mdz136
4. Li,M.,Chao,E.,Esplin,E.,Miller,D.,Nathanson,K.,&Plon,S.etal.(2020).Points to consider for reporting of germline variation in patients undergoing tumor testing: a statement of the American College of Medical Genetics and Genomics (ACMG).Genetics In Medicine,22(7),1142–1148.doi:10.1038/s41436-020-0783-8
5. Forbes,C.,Fayter,D.,deKock,S.,&Quek,R.(2019).A systematic review of international guidelines and recommendations for the genetic screening, diagnosis, genetic counseling, and treatment of BRCA mutated breast cancer.Cancer Management And Research,Volume11,2321–2337.doi:10.2147/cmar.s189627
6. Khiabanian,H.,Hirshfield,K.,Goldfinger,M.,Bird,S.,Stein,M.,&Aisner,J.etal.(2018).Inference of Germline Mutational Status and Evaluation of Loss of Heterozygosity in High-Depth, Tumor-Only Sequencing Data.JCO Precision Oncology,(2),1–15.doi:10.1200/po.17.00148.
7. Hiltemann,S.,Jenster,G.,Trapman,J.,van derSpek,P.,&Stubbs,A.(2015).Discriminating somatic and germline mutations in tumor DNA samples without matching normals.Genome Research,25(9),1382–1390.doi:10.1101/gr.183053.114
8. Park,S.,Supek,F.,&Lehner,B.(2018).Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits.Nature Communications,9(1).doi:10.1038/s41467-018-04900-7
9. Sun,J.,He,Y.,Sanford,E.,Montesion,M.,Frampton,G.,&Vignot,S.etal.(2018).A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal.PLOS Computational Biology,14(2),e1005965.doi:10.1371/journal.pcbi.1005965
10. Report Linker.Real-WorldEvidenceSolutionsMarket-Growth,Trends,andForecasts(2020–2025).Retrieved09April2021,from<background-color:#FF3300;udirection:rtl;>https://www.reportlinker.com/p05974125/Real-World-Evidence-Solutions-Market-Growth-Trends-and-Forecasts.html</background-color:#FF3300;udirection:rtl;>
11. Research and Market.NorthAmericaDigitalGenomeMarket2021–2028.Retrieved09April2021,from<background-color:#FF3300;udirection:rtl;>https://www.researchandmarkets.com/reports/5264345/north-america-digital-genome-market-2021-2028</background-color:#FF3300;udirection:rtl;>
12. Guideline for Good Clinical Practices - International Conference of Harmonization.Retrieved09April2021,from<background-color:#FF3300;udirection:rtl;>https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf</background-color:#FF3300;udirection:rtl;><udirection:rtl;>.</udirection:rtl;>
13. Knudson,A. (1971).MutationandCancer:StatisticalStudyofRetinoblastoma.ProceedingsOfTheNationalAcademyOfSciences,68(4),820–

823.doi:10.1073/pnas.68.4.820

14. James,G.,Witten,D.,Hastie,T.,&Tibshirani,R.(2013).An introduction to statistical learning(1sted.).New-York, NY:Springer.doi:10.1007/978-1-4614-7138-7
15. Castellanos,E.,Gel,B.,Rosas,I.,Tornero,E.,Santín,S.,&Pluvinet,R.etal.(2017).A comprehensive custom panel design for routine hereditary cancer testing: preserving control, improving diagnostics and revealing a complex variation landscape.Scientific Reports,7(1).doi:10.1038/srep39348
16. Mandelker,D.,Zhang,L.,Kemel,Y.,Stadler,Z.,Joseph,V.,&Zehir,A.etal.(2017).Mutation Detection in Patients With Advanced Cancer by Universal Sequencing of Cancer-Related Genes in Tumor and Normal DNA vs Guideline-Based Germline Testing.JAMA,318(9),825.doi:10.1001/jama.2017.11137
17. Meric-Bernstam,F.,Brusco,L.,Daniels,M.,Wathoo,C.,Bailey,A.,&Strong,L.etal.(2016).Incidental germline variants in 1000 advanced cancers on a prospective somatic genomic profiling protocol.Annals Of Oncology,27(5),795–800.doi:10.1093/annonc/mdw018
18. Winter,C.,Nilsson,M.,Olsson,E.,George,A.,Chen,Y.,&Kvist,A.etal.(2016).Targeted sequencing of BRCA1 and BRCA2 across a large unselected breast cancer cohort suggests that one-third of mutations are somatic.Annals Of Oncology,27(8),1532–1538.doi:10.1093/annonc/mdw209
19. Corso,G.,Intra,M.,Trentin,C.,Veronesi,P.,&Galimberti,V.(2016).CDH1 germline mutations and hereditary lobular breast cancer.Familial Cancer,15(2),215–219.doi:10.1007/s10689-016-9869-5
20. Ménard,T.,Barros,A.,&Ganter,C.(2021).Clinical quality considerations when using Next-Generation Sequencing (NGS) in clinical drug development.ResearchGate (Preprint).doi:10.13140/RG.2.2.27345.86885/1
21. Aloraifi,F.,Alshehhi,M.,McDevitt,T.,Cody,N.,Meany,M.,&O'Doherty,A.etal.(2015).Phenotypic analysis of familial breast cancer: Comparison of BRCAx tumors with BRCA1-, BRCA2-carriers and non-familial breast cancer.European Journal Of Surgical Oncology (EJSO),41(5),641–646.doi:10.1016/j.ejso.2015.01.021
22. Gliklich,R.,&Leavy,M.(2020).Assessing Real-World Data Quality: The Application of Patient Registry Quality Criteria to Real-World Data and Real-World Evidence.Therapeutic Innovation & Regulatory Science,54(2),303–307.doi:10.1007/s43441-019-00058-6
23. Leon,P.,Cancel-Tassin,G.,Bourdon,V.,Buecher,B.,Oudard,S.,&Brureau,L.etal.(2021).Bayesian predictive model to assess BRCA2 mutational status according to clinical history: Early onset, metastatic phenotype or family history of breast/ovary cancer.The Prostate.doi:10.1002/pros.24109
24. Ménard,T.,Barmaz,Y.,Koneswarakantha,B.,Bowling,R.,&Popko,L.(2019).Enabling Data-Driven Clinical Quality Assurance: Predicting Adverse Event Reporting in Clinical Trials Using Machine Learning.Drug Safety,42(9),1045–1053.doi:10.1007/s40264-019-00831-4
25. Ménard,T.,Bowling,R.,Mehta,P.,Koneswarakantha,B.,&Magruder,E.(2020).Leveraging analytics to assure quality during the Covid-19 pandemic - The COVACTA clinical study example.Contemporary Clinical Trials Communications,20,100662.doi:10.1016/j.conctc.2020.100662
26. Ménard,T.(2021).Letter to the Editor: New Approaches to Regulatory Innovation Emerging During the Crucible of COVID-19. Therapeutic Innovation & Regulatory Science.doi:10.1007/s43441-021-00281-0

Figures

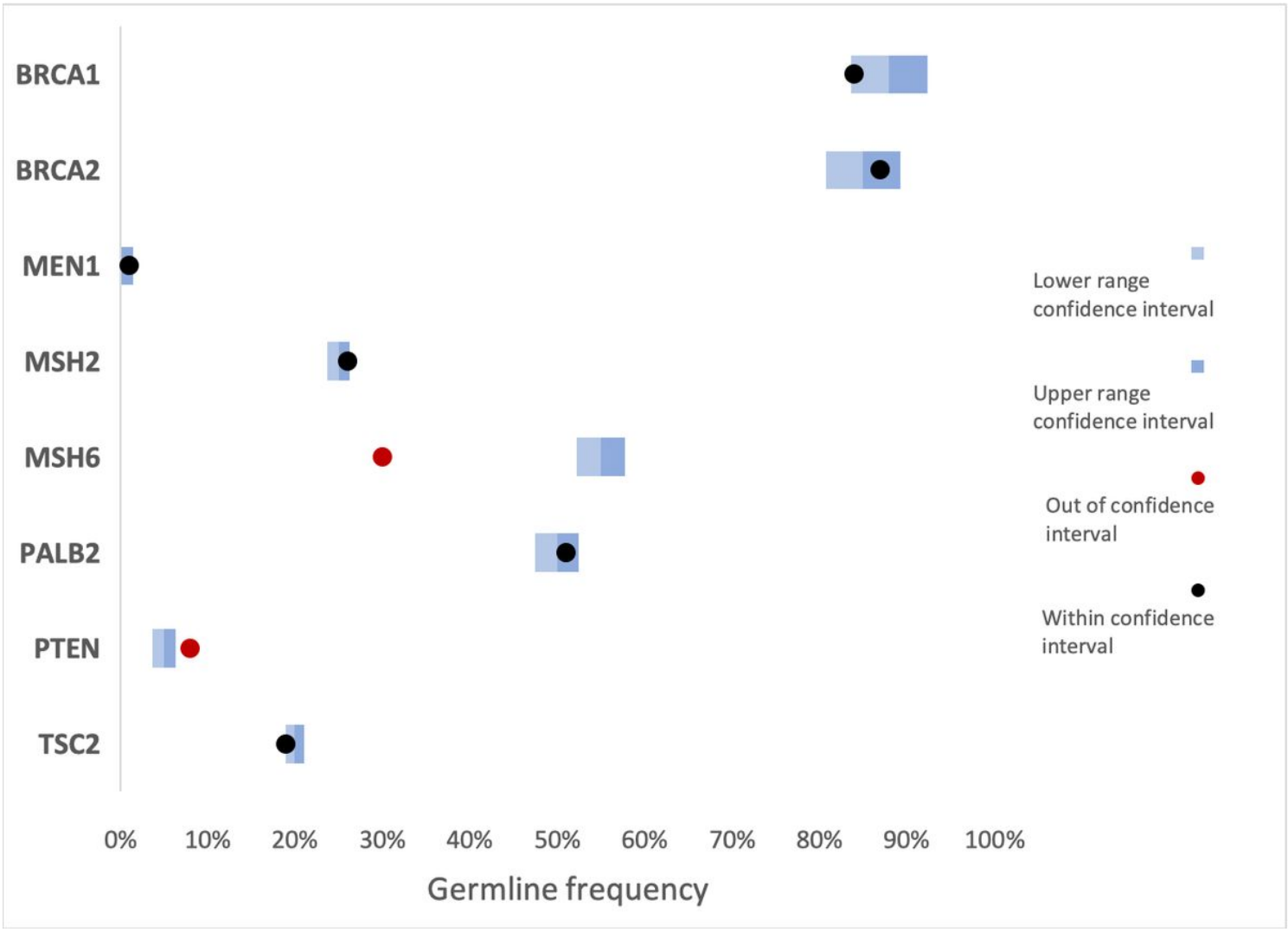

Figure 1

PGPV frequencies (dots) against expected population germline variant frequencies (bars). Example produced using dummy data.

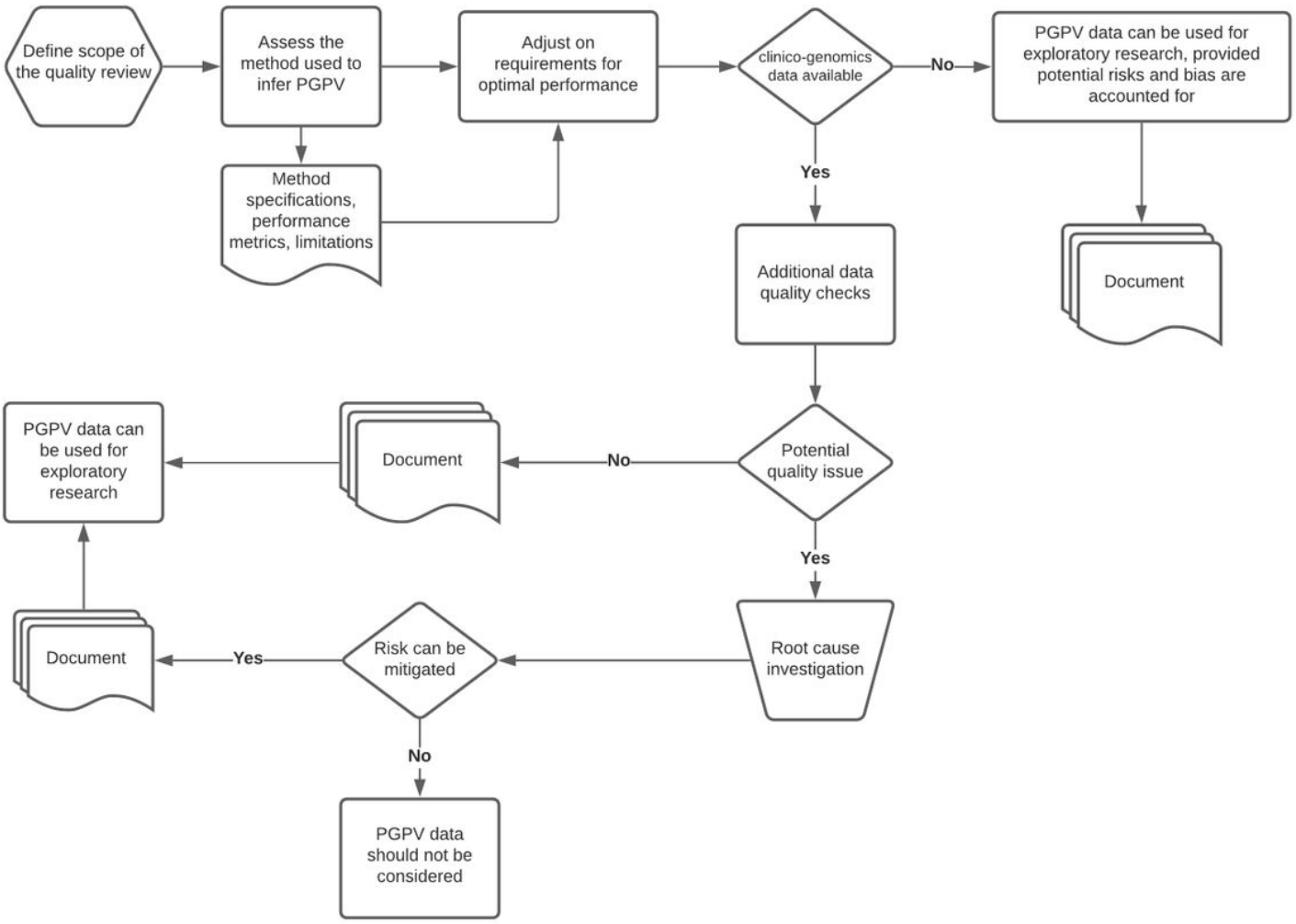

Figure 2

Strategy to evaluate quality of PGPV inferred from tumor-only data