Clinical Chemistry

Universal Presence of Gene/Variant Nomenclature Errors in Journal Manuscript Submissions

Article:
Lisa A Lansdon, Binu Porath, Mari Mori, David T Miller, Diane Dunham Drexler, Catherine Wattenberg, Morgan Richardson, Robert D Steiner, Peter J Freeman. *Universal Presence of Gene/Variant Nomenclature Errors in Journal Manuscript Submissions.*
Clin Chem 2026; 72(6): 652–61. https://doi.org/10.1093/clinchem/hvag010

Guest: Dr. Peter Freeman is a lecturer in Healthcare Sciences at the University of Manchester in the United Kingdom.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I am Bob Barrett.

Next-generation sequencing has revolutionized the care of individuals with inherited disorders by allowing the rapid identification of pathogenic variants, however many of these conditions are extremely rare and it's challenging to find enough individuals with the same presenting features and the same genetic variant to conclusively characterize that variant as pathogenic. Rapidly expanding access to next-generation sequencing should only help address this issue. Testing of more individuals should result in more frequent detection of rare variants and accelerate their characterization in variant databases.

Unfortunately, there's a problem. Researchers often use non-standardized language in their publications. This means that although multiple research teams may have independently described patients with the same inherited disorder and identified the same suspected variant causing the disorder, they can't find each other's work because they're using different terms to describe the same variant.

To address this, professional societies have developed standardized nomenclature and even released software to help authors adhere to these standards, but use of other terminology is still widespread. What else can be done? How can we ensure uniform language to help to find variants of uncertain significance as either clearly benign or clearly pathogenic? A research article in the June 2026 issue of *Clinical Chemistry* surveys the published literature to determine the scope of the problem and suggest ways for journals to help authors adhere to nomenclature standards.

Today, we welcome the article's senior author. Dr. Peter Freeman is a lecturer in Healthcare Sciences at the University of Manchester. His research focuses on variant nomenclature, data standardization and improving the accuracy of genetic reporting through his software VariantValidator. And Dr. Freeman, your paper shows that

variant nomenclature errors are essentially universal in submitted manuscripts. What kinds of errors are most common and why do they persist?

Peter Freeman: Well, what I'd say about this is that sadly the majority of the variant description errors are actually not complex errors, and it seems to reflect a kind of failure to implement basic reporting standards. For example, in the manuscript we found that 50% of manuscripts didn't provide a reference sequence for the variant description and you need that reference sequence in order -- it's like your map. If you describe the position and say that this DNA base is changed to that DNA base and provide a position and don't provide the map, it's like looking for a position in London when you're actually in Helsinki. So we need that.

We also found that there were other areas outside of just standard, you know, the Human Genome Variation Society [HGVS] nomenclature, just simple basic errors such as 41% of manuscripts that have been submitted to *Genetics in Medicine* over the time period that we were studying failed to report a genome build. And if you don't have the genome build, it's hard to replicate the study. 85% failed to provide stable gene identifiers. Okay, many provide the gene symbol but gene symbols do change over time whereas the ID doesn't.

So, if you're doing a retrospective look back, you know for literature, you have this problem where if the gene symbol has changed, you might miss the actual publication if you're searching at the gene level. And we also had had cases where 98% didn't provide complete variant descriptions, so with the complete range at the level of the genome, the transcript, and if necessary, the protein. And beyond these omissions there were lots of very common syntax errors. For example, combining your description types from the VCF description merged together with an HGVS description to this horrible mess of a description that's neither one nor the other, and the persistent use of deprecated or historic naming systems.

And as I said, this isn't necessarily to do with complex issues. This is all to do very often with very simple, even single nucleotide variant descriptions. In terms of why they persist, I think there's several layers to this and we're still looking into this, we're trying to understand it, but I think one thing is that there's a tendency for variant nomenclature to be treated as more of a descriptive label rather than what its intended to use is, is a structured data format for evidence discovery. So it's kind of treated not as a data source, but as optional formatting. And I think that also, I mean from my use of software and others use of software, both in research and clinical, reporting pipelines and software use in clinical and research settings often don't adhere to the basic reporting

standards. They're not fully compliant, even when producing things like a formal clinical report, errors appear even in those clinical reports.

So, the non-standard outputs are systematically and routinely copied into manuscripts by authors, and as a result the inaccurate or complete descriptions become normalized and they propagate downstream into databases and the final publications.

Bob Barrett: Are the mistakes mainly author mistakes, tooling limitations, or issues with journal workflows?

Peter Freeman: As I said before, I think it's a combination of various things. I think that tooling that's widely used that produces inconsistencies with the nomenclature, that then propagates downstream. And as I said I've observed and others that I know have observed these failures and this does apply even to tools that I would class myself and others would class as world-leading and absolutely excellent, we still see that they make consistent and persistent mistakes. Researchers and even clinicians trust these outputs and assume that they're correct and publication-ready. In parallel, the authors like I said, you use that in outdated terminology that they recognized in their specific field and they want to try and hold on to it, even though there is a designated standard, which is the HGVS standard, there's still this push to try and use the outdated nomenclature systems. This reflects the fact that variant naming is not treated as a data standard, it's kind of treated as a familiar label. And many authors don't realize that validation may be required even if you've got the descriptions off of trusted software. And as I said, if there's no validation of that final output, then they flow directly into clinical reports and publications allowing those errors to persist.

In terms of journal workflows, I mean I will potentially talk about more about this later, but it's very difficult for journals to actually pick up on this unless there is an availability of trained reviewers. We are very lucky at *Genetics in Medicine* that the chief editors decided to assemble this team of technical experts, like myself and the lovely people who wrote the publication and who did this study. But I think we're, I don't know of any other journals that have got specific technical editors whose sole role is to actually pick up on these nomenclature issues and feed back to the authors, and effectively enforce policing, saying "Stop, we won't publish until the data is accurate."

Bob Barrett: So, doctor what are the real world consequences of getting variant nomenclature wrong?

Peter Freeman: It's really difficult to pin this down precisely because the diagnostic process is based on a massive accumulation of evidence and it's governed by standards such as the American College for Medical Genetics Variant Classification Guidelines plus lots of, you know, local guidelines that have stemmed from that. For example, we have our own versions of this in the UK, but within these standards they explicitly rely on searching literature and databases for supporting evidence. And if the naming of the variants in that literature or in the database is incorrect, it's going to become difficult or even impossible to retrieve.

So you are trying to search a database such as LOVD or ClinVar for a specific variant but if it's propagated from the clinic to the journal incorrectly, and then into their records incorrectly, it can be unfindable. And we also have amazing tools that try to do retrospective searches of literature such as LitVar2 and Mastermind and others. And because of the kind of like variation of different formats, they're often tripping up and unable to find the evidence.

So as a result, there may be some really strong published evidence out there such as a functional study which could be used as diagnostic evidence or a case control study which could be used as diagnostic evidence, but if you can't find it, it cannot be used as evidence. That evidence that could support or resolve the diagnosis, if it's not findable, this can directly affect variant classification, diagnostic certainty, and ultimately patient outcomes.

Bob Barrett: Why do you think these problems persist despite long-standing standards like HGVS nomenclature?

Peter Freeman: We've actually been discussing this widely in the *Genetics in Medicine* team and we think that there are a few central areas where there's a problem. We think one central area could be around education. There remains a clear and unmet educational need around understanding that variant nomenclature is a prerequisite of strong evidence and discovery and clinical interpretation and is not just treated as optional formatting and needs to be treated as a core data infrastructure.

So, often in many clinical training pathways, you'll be taught about the HGVS nomenclature, how to use it, but you're not necessarily told that it's absolutely essential data that we need to use to make our evidence discoverable. It's a direct communication between a format that humans can interpret and something that can then be picked up and converted into more machine readable formats, but it is that entry point from the clinic to the coding. So as a result, it's not clearly understood that HGVS is used as a data standard that enables the interoperability, findability, and the reuse of evidence, but

there are positive examples in education where this is actually being addressed. For example, in the UK, we have an NHS, or National Health Service, Scientist Training Program, where we directly train NHS healthcare scientists, and in particular, the clinical bioinformatics, genomic scientists are taught very clearly that the HGVS is a kind of data reference. It's not just a way of communicating, it's actually used to standardize the data. It needs to be findable and needs to be correct.

And the reason we're talking and specifically training these particular scientists is that they're developing the workflows and toolings that will present data to the clinicians that then go on and make the diagnostic decisions. And also within the U.S. Laboratory Genetics and Genomics Fellowship, there's similar emphasis on correct variant reporting. However, on top of that, there's no widely recognized professional standard that governs the quality of reported genetic data.

So, throughout the clinical diagnostic field, there's no consistency mandated in clinical reports or in database, and historically, not in published literature. Many journals are now beginning to require what would be considered a basic level of reporting and this is improving. I think at *Genetics in Medicine*, we're really leading the way on this and it was really encouraging that since the publication of this, there's been some communication with *Clinical Chemistry*, we've seen that there's actually been improvements at *Clinical Chem*, or certainly conversation at *Clinical Chemistry*, to provide more detailed guidance to authors on how the data should be structured correctly.

Bob Barrett: Dr. Freeman, you've been developing VariantValidator for ten years now. How does this address the problems that you've identified in this study?

Peter Freeman: Back in 2016, in fact yeah it was at start 2016, VariantValidator was the brainchild of a Professor Raymond Dalgleish at the University of Leicester, and he was maintaining several databases to do with collagen disorders for such as osteogenesis imperfecta and Ehlers-Danlos syndrome. Well, it was this problem was, he was identifying this problem like back in 2016 that there were many reported cases in clinical literature with some sort of nomenclature but the nomenclature wasn't well formatted, wasn't standardized, and he needed tooling in order to be able to standardize the nomenclature and apply the correct HGVS to go into the LOVD databases that he was maintaining.

So back in 2016, I was loaned to Raymond as a postdoc for a period of three months and now in 2026, we are still working on this project together. We have both been members of the HGVS Variant Nomenclature Committee. We have a good relationship with Ivo Fokkema at LOVD, he is a current

member of the HGVS Nomenclature Committee. Raymond, although he's retired, he is now still working on the project as a retiree.

So that's how VariantValidator came to be and what we've done over these ten years or so has been actively and continuously developing the tooling, making sure that evolves alongside the evolving standard, making sure that we are as robust as possible on the standards and on the nomenclature, and that we provide all of the detail that would be needed to go into an accurate clinical report or text or database submission.

Through close collaboration with the LOVD team, we've co-created tooling, that in my view, when these two tools are combined, their interface, which is a HGVS variant syntax checker and the power of VariantValidator as a sequence level checker, we've got probably the world's leading tool in terms of accuracy for checking nomenclature. The tooling is available in multiple formats. We've got a graphical web interface for general users such as clinicians. We've also got programmatic interfaces that can communicate directly with computers and that could be embedded directly into pipelines and workflows or into software that other people want to build.

So this allows the tooling to be used within clinical and research workflows and also during manuscript preparation and peer review, and to get the data ready for database submission.

However, as highlighted in the Lansdon paper, where we actually used VariantValidator in the peer review process, so the idea was that at the point of first submission, authors also validated their data using VariantValidator and had to send along with their submission documented evidence they've been used with the correct variant descriptions provided by VariantValidator. But even when the authors were presented with the accurate data, the errors within the manuscript were not updated and corrected in line with what they were given from the tooling. So that was a disappointing gap. And we really think that, you know, we can have this -- as I said, I mentioned this before, we've got this unaligned software ecosystem where multiple tools used in clinical workflows are trusted but are not necessarily providing the right outputs, but we do have tools like VariantValidator that correct this. But we have this persistent educational gap where accurate reporting standards are not fully appreciated, understood, and prioritized. So even when you're presenting the accurate data, you may not be going back and making sure that what you're actually outputting from the lab or into a manuscript or into a database is actually fully correct.

Bob Barrett: Well, finally doctor, what practical steps could journals, reviewers, and researchers take right now to reduce these errors?

Peter Freeman: Well, I think that a really big groundbreaking change in the field is going to be that later in this year, later in 2026 or maybe in early 2027, the American College for Medical Genetics, in association with the vast majority of major professional standards bodies and quality assurance bodies across the U.S., EU, UK, and Canada will release a professional standard defining the minimum requirements for reporting of interpreted genetic data in publications, clinical reports, and databases. The *Genetics in Medicine* authorship guidelines and the related materials we developed alongside the Lansdon paper have actually gone towards informing this standard and we've worked very closely with the team putting this together. Through organizations such as the Human Genome Organization Reporting of Sequence Variants Committee, there's a clear opportunity to engage with publishing houses, with specific journals, and try to align the journal publishing standards so the authorship guidelines and the requirement for accurate data across the field and clearly define what must be reported with the publication to ensure that data are findable and reusable.

The key challenge that comes after that though is really one of policing enforcement. There really does need to be well, as things stand currently let me say, this does fall on peer reviewers. And I guess one of the things we allude to in the paper is that there may be a wide pool of peer reviewers that could actually have the expertise to police those standards to make sure that the submitted manuscripts are up to spec just as we do at *Genetics in Medicine*.

Again, this is open to potentially human error as well, and we have to find those reviewers that really understand the standards. So ideally there will be development of software tools in this field that can take a manuscript text, or any text, and review it automatically and flag nomenclature errors and provide user-friendly feedback so that this can be updated, almost given a kind of green flag or a certificate of quality at the end to say this is now distribution ready.

So, I think combining professional standards, aligning the journals, making sure there's either human or automated validation, and getting the tooling in place, and basically the education and everything else around this area is the most likely way that we're going to reduce errors at scale in the future.

Bob Barrett: That was Dr. Peter Freeman from the University of Manchester in the United Kingdom. He wrote a research article in the June 2026 issue of *Clinical Chemistry* describing

errors in genetic variant nomenclature, and he has been our guest in this podcast on that topic. I'm Bob Barrett. Thanks for listening.