JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, January 2018, Vol. 63, No. 1
doi: 10.1111/1556-4029.13687
Available online at: onlinelibrary.wiley.com

Letter to the Editor—Appropriate Standards for Verification and Validation of Probabilistic Genotyping Systems*

Sir,

As the President's Council of Advisors on Science and Technology (PCAST) recently noted, the interpretation of mixed DNA samples can be extremely difficult, particularly when there is uncertainty about the number of contributors and whether all of the contributors' alleles have been detected (1). One promising way to address this problem is the development of automated systems for probabilistic genotyping (PG), (2) but how will we know that these automated systems work properly? If conventional approaches that rely on human judgment are not up to the task, how will we assess the validity of PG systems? These are essential questions for forensic scientists to consider as standards for DNA testing are developed through groups such as OSAC and SWGDAM. They will also be important for courts to consider when they assess the admissibility of evidence generated by PG systems.

We urge forensic scientists interested in these questions to pay close attention to the standards for software validation that have been developed by the Institute of Electrical and Electronics Engineers (IEEE). Software-based PG approaches are necessarily rooted in collaboration between experts in the areas of molecular biology, population genetics, statistics, forensic science, computer science, and software engineering. While it is important to consider the perspectives of all of these disciplines on the validation issue, we think that the perspectives of software engineers are particularly important. Decades of experience with software failures have led to established practices for what is commonly known as verification and validation (V&V) of software. We urge that those practices be followed when evaluating PG systems.

In the world of software engineering, verification and validation entail "evaluating a system or component during or at the end of the development process to determine whether it satisfies specified requirements" (3). IEEE Standard 1012-2012, *IEEE Standard for System and Software Verification and Validation* (4) spells out universally applicable and broadly accepted software V&V standards. It requires that each software component be assigned an integrity level that increases from 1 to 4 depending on the consequences of a failure. Consequences are categorized as "negligible," "minor," "critical" (causing "major and permanent injury, partial loss of mission, major system damage, or major financial or social loss") or "catastrophic" (causing "loss of human life, complete mission failure, loss of system security and safety, or extensive financial or social loss"). As the "criticality" of software increases across these integrity levels, the intensity and rigor of the V&V tasks required by the standards also increase. A PG system used to generate evidence in criminal cases should clearly be assigned a very high integrity level and hence would warrant intense and rigorous V&V under these standards.

As the integrity level increases, it also becomes more important that V&V be independent of software development. IEEE Standard 1012-2012 describes three distinct dimensions of independence in the V&V process. Technical independence involves utilizing "personnel who are not involved in the development of the software" (5). Managerial independence is accomplished when V&V responsibilities are administrated by an organization that is separate from the organizations that develop and manage the software system. Financial independence requires that "control of the V&V budget be vested in an organization independent of the development organization" (5). As a general rule of thumb, software developers working in areas as varied as word processing and satellite communications expect that 10–50% of their budget is set aside for V&V processes; for example, 35% of IBM's budget for developing the Space Shuttle's flight software was allocated to its V&V team (6).

Software defects can be caused by flaws in both the design and the implementation (coding) of a system or component. For PG systems, there is a multitude of potential points of failure that warrant independent evaluation. For example, is the modeling of stutter artifacts used the best available, coded as designed, or even appropriate to the problem? Does the PG algorithm systematically favor inclusions? How likely are false negatives and positives? Would outside experts agree with the software's decisions at each stage of analysis? And so on. It is important to know both the conceptual model being used *and* whether the software implements that model reliably (2). Without independent verification and validation, we can only hope and trust. In any scientific endeavor, it is better to verify and validate.

The failure of PG software developers to complete appropriate V&V processes can translate directly to severe financial hardships or even the loss of liberty or life. Hence, V&V of PG systems requires urgent attention from the forensic science community. Users of PG software should demand that PG developers utilize software industry standards (e.g., IEEE) for the development, documentation, verification, validation, acquisition, use, and maintenance of their systems—in the same way that they expect and demand that companies that provide their reagents and instruments adhere to rigorous quality assurance and quality control practices. Standards-setting bodies like OSAC should incorporate the IEEE standards into their standards for software validation. Accrediting bodies such as ASCLD, ANAB, and UKAS should require accredited laboratories to use only PG systems that can demonstrate compliance with appropriate software industry standards. Defense experts and attorneys should insist on complete documentation of V&V processes for any PG systems used in criminal cases in which they are involved. Criminal defendants should have access to documentation generated during V&V processes. Without independent V&V, courts will only have the self-interested assurance of the software developers themselves that the system works properly. Given the critical importance of PG systems in criminal justice, those assurances are not good enough.

History provides many examples of the failure of engineered systems, (7) which underlines the importance of validation and

*Co-authors Nathaniel Adams and Dan Krane declare a conflict of interest, in that they are employed by Forensic Bioinformatic Services, Inc., an independent forensic DNA consulting firm.

verification. A well-known example involving software defects was the Therac-25 radiation therapy machine, which was programmed in a manner that, under some circumstances, caused overdoses of radiation to be administered to patients, leading to burns and radiation poisoning (8). The manufacturer initially dismissed reports of system malfunction as impossible, claiming that the software had been checked and tested extensively and found to operate correctly. Through painstaking experimentation, a physicist eventually determined the precise circumstances under which the system failed. By then, however, several patients had received fatal overdoses of radiation due to the software defect. This disaster highlights the importance of careful, independent V&V of software that performs critical functions.

If forensic scientists ignore these lessons in a rush to embrace PG systems, they invite their own disasters. Software defects in PG system will be difficult to detect, particularly if (as with the Therac-25) critical errors occur only under some circumstances but not others. A problem of this type might persist without being noticed for a considerable period of time, while doing incalculable damage to the public interest and, ultimately, to the reputation of forensic science.

References

1. President's Council of Advisors on Science and Technology. Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods, 2016; https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (accessed May 30, 2017).
2. Haned H, Gill P, Lohmueller K, Inman K, Rudin N. Validation of probabilistic genotyping software for use in forensic DNA casework: definitions and illustrations. Sci Justice 2016;56(2):104–8.
3. Institute of Electrical and Electronics Engineers. IEEE Std 1012-2012 – IEEE standard for system and software verification and validation. Section 3.1-Definitions, 2012;7-11; http://ieeexplore.ieee.org/document/6204026/(accessed May 30, 2017).
4. Institute of Electrical and Electronics Engineers. IEEE Std 1012-2012 – IEEE standard for system and software verification and validation, 2012;209; http://ieeexplore.ieee.org/document/6204026/(accessed May 30, 2017).
5. Institute of Electrical and Electronics Engineers. IEEE Std 1012-2012 – IEEE standard for system and software verification and validation. Annex C, Definition of independent V&V (IV&V), 2012;165–7; http://ieeexplore.ieee.org/document/6204026/(accessed May 30, 2017).
6. Tomayko J. *Computers in spaceflight*. Washington, DC: NASA Contractor Report-182505, Mar. 1988. https://history.nasa.gov/computers/contents.html (accessed June 15, 2017).
7. Perrow C. Normal accidents: living with high risk technologies. Princeton, NJ: Princeton University Press, 1999.
8. Leveson N, Turner C. An investigation of the Therac-25 accidents. Computer 1993;26(7):18–41.

Nathaniel Adams,[1] B.S.; Roger Koppl,[2] Ph.D.; Dan Krane,[3] Ph.D.; William Thompson,[4] J.D., Ph.D.; and Sandy Zabell,[5] Ph.D.
[1]Forensic Bioinformatic Services, Inc., Fairborn, OH
[2]Department of Finance, Syracuse University, Syracuse, NY
[3]Department of Biological Sciences, Wright State University, Dayton, OH
[4]Department of Criminology, Law, and Society, University of California Irvine, Irvine, CA
[5]Department of Statistics, Northwestern University, Evanston, IL
E-mail: Dan.Krane@wright.edu