

Are your MRI contrast agents cost-effective?

Learn more about generic Gadolinium-Based Contrast Agents.



FRESENIUS  
KABI

caring for life

# AJNR

## The Z-Shift: A Need for Quality Management System Level Testing and Standardization in Neuroimaging Pipelines

N.B. Dadario, P. Nicholas, A. Henkin, B. Sin, K. Dyer, M.E. Sughrue and S. Doyen








This information is current as of May 2, 2024.

*AJNR Am J Neuroradiol* 2022, 43 (3) 320-323

doi: <https://doi.org/10.3174/ajnr.A7435>

<http://www.ajnr.org/content/43/3/320>

# The Z-Shift: A Need for Quality Management System Level Testing and Standardization in Neuroimaging Pipelines

 N.B. Dadario,  P. Nicholas,  A. Henkin,  B. Sin,  K. Dyer,  M.E. Sughrue, and  S. Doyen

The field of neuroimaging has considerably improved in mapping the human brain largely due to massive advancements in machine learning (ML) capabilities and big data approaches.<sup>1,2</sup> Growth in collaboration among software developers, researchers, and clinicians is leading to more advanced (and complex) brain analytics software that can guide neurosurgical treatment, such as surgical navigation. Any complex system may contain inherent problems that threaten the viability of the system for its intended use. Medical devices used in clinical settings are no exception. Because the exact number of such problems is difficult to estimate, it is notable that for cleared and approved medical devices in the United States (ie, those that have gone through regulatory review), software issues represented the top cause for recalls (removal from the field due to an issue) for 19 of 20 consecutive quarters according to the 2021 Recall Index compiled across all industries in the United States by Sedgwick.<sup>3</sup> However, these recall statistics do not consider software used in clinical practice that is not developed under a controlled development process (ie, those subject to regulatory review), putting patients at an even higher risk despite good intentions.

In this editorial, we describe how the implementation of a planned approach to verification and validation, under the umbrella of an effective Quality Management System (QMS), allowed us to identify and resolve a fundamental computer science fault found in a commonly used data science asset in medical device development. We suggest that companies (developing any clinically used pipelines, whether as a medical device or not) should invest in building a fit-for-purpose verification and validation framework under a QMS that would facilitate the discovery and elimination of faults in a systematic manner.

## Z-Shift: One of Many Possible Problems

We start by describing an anecdote demonstrating how deep-seated errors in code can arise within even the most used data science packages for medical devices. Most important, the error we identified was with a package common to most neuroimaging pipelines.

**Standard Method for Shipping Code: Docker.** Effective neuroimaging analyses for tractography require alignment of diffusion-weighted scans with anatomic (eg, T1, T2) images. A number of

calculations are performed to merge the underlying anatomic image with overlaid tractography to achieve this. One method to accomplish this includes computing the space in which the anatomic image and tractography are in by using Python programming language. A transformation function is then applied to overlay the tractographic map onto the anatomic image. The series of steps performed to achieve such alignment can be referred to as a “neuroimaging pipeline” (or part of one). Once a successful neuroimaging pipeline is created and demonstrates reproducibility in a single computational environment (ie, on a developer’s computer), software engineers and researchers alike most commonly use the open-source platform Docker (<https://www.docker.com/>) to encase their pipeline in a container and then deliver it to other users and consumers. These second parties can then pull the software from Docker and run it in other computational environments (eg, their own computers). Therefore, once a neuroimaging pipeline has been “Dockerized,” it is generally thought to be “crystalized,” in that any other individual who re-executes the workflow should be able to exactly replicate the original results, irrespective of the operating system. This property is particularly desirable when working in a cloud-based environment, in which underlying virtual computer specifications may vary. Such environments are commonly used, especially in medical image-processing and analysis.

Docker is the industry standard for deploying code to production in many sectors, including the medical device industry. The National Institutes of Health uses Docker technology, recently using it to facilitate their mission of delivering machine learning-based image analysis software to hospitals to guide medical diagnoses.<sup>4,5</sup>

**Our Case Detailing the Z-Shift Problem.** In accordance with the practices described above, a company Data Scientist created updated code for aligning tractography in the human brain. This code was created specifically on the developer’s machine, which runs the Mac Operating System (MacOS), but because of common knowledge of production environments with Docker containers, it was generally believed that this code was appropriate to run reproducibly on any other operating system. However, while the new software and code were undergoing QMS verification and validation processes, in the final steps of validation, which included review by qualified neurosurgeons, the acceptance criteria failed. In particular, a number of scans were being shifted along the z-axis. In Fig 1, we present an obvious case in which the corticospinal tract is seen shifted incorrectly along the z-axis. Similar obvious cases were not always commonplace, and when many scans and analyses are running at large computing scales, such small discrepancies can become incredibly easy to miss. With neurosurgical treatment in particular, there are obvious concerns of misalignment and inaccuracy in neuroimaging, such as with tracts or ROIs not being perfectly aligned with anatomic scans. Had a staged quality-management process not been implemented, with predefined acceptance criteria that remove subjectivity as far as possible, we may not have identified this issue.

The Underlying Problem is Near a Half-Century Old: Floating-Point Decimals. After QMS processes flagged the issue, our team had to peel back many layers of code to find what specifically caused this discrepancy. Numerous tests were completed, such as controlling for different operating systems by testing the code in Dockerized environments, but the discrepancy still existed. Eventually, we discovered that the root cause of our Z-shift problem was related to some of the very foundational concepts used in software engineering. Docker technology is commonly used with the belief that it provides a reliable production environment agnostic to the system in which you run it and isolates and containerizes the platform-independent file system and libraries that are required to make that software run. However, despite that property, Docker still relies on the host operating system on which it is being run, which can introduce discrepancies. In other words, the same Docker image yields different results on MacOS compared with Linux due to a platform-dependent variance between the 2 systems (linked to an older version of the LibM math library), which, at a very low level, impacted the way floating-points are stored but had the

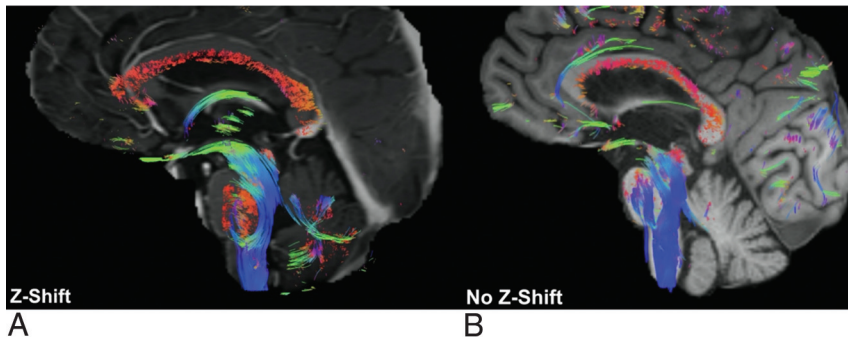
effect of snowballing into considerable differences over multiple iterations.

Thus, one must conclude that the docker run-time for different platforms can and does introduce portability issues for computation. Such a problem, so many layers down, rests at the very basics of data science techniques.<sup>6</sup> We learned that when you have a function, despite attempting to control for some variance between the test system and the deployment system, there may still be differences sufficient to throw off the results in a way that a clinician would say was unacceptable (but an automated test may not). To manage these problems, among other possible ones, one must first identify them correctly in a systematic way, such as by having implemented a QMS as in our example.<sup>7</sup>

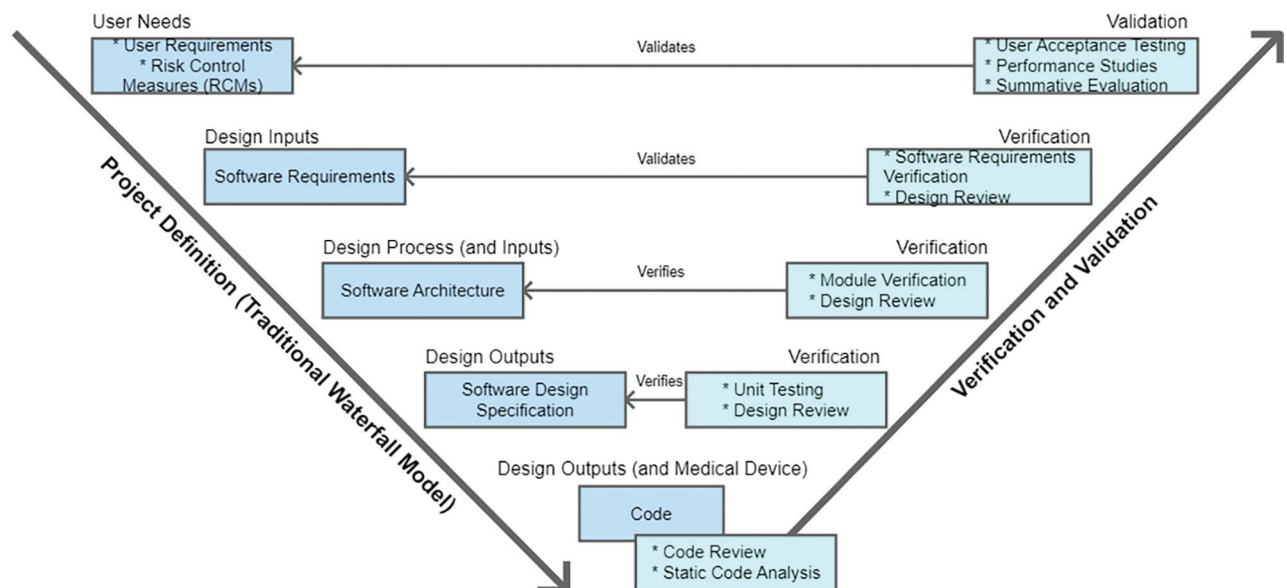
### QMSs

Subtle and unforeseen problems will inevitably arise with most software solutions that require dependable results. To combat this issue in the medical device space, regulatory agencies (such as the US Food and Drug Administration [FDA]) require companies that market medical devices for clinical use, such as brain-analysis software, to implement QMSs that are effective for the purpose of the organization and are continually assessed for this purpose.

What is a QMS? A QMS is a series of documented interconnected processes (such as purchasing, development, testing, deployment, and support) that are created, implemented, and maintained per the requirements of the organization (according to product risk) and international standards and regulations.<sup>8</sup> For medical devices, which include analytics software, the QMS is



**FIG 1.** The Z-Shift. This figure demonstrates a shift in the alignment of the tracts in a human brain along the z-axis. For instance, in A, one can see that the corticospinal tract is shifted incorrectly along the z-axis compared with what is seen in B, with correct alignment of the tracts.



**FIG 2.** Traditional Waterfall development method mapped to software verification and validation activities.

the foundation for maintaining regulatory compliance, reducing risk, driving improvement, and ultimately meeting customer expectations. Most important, a QMS is only as good as the individuals and/or company implementing the processes.<sup>9</sup> While each process has its own requirements, steps, and process flow and may be very clearly documented in a procedure, if team members in an organization choose not to follow the documented processes, the complexity and dependability of those processes are irrelevant and problems can go into production unnoticed.

**QMS Processes for New Products.** The process at the heart of this commentary is called the design and development process. This process governs how new products are brought from user need to market introduction. The FDA describes this process as shown in Fig 2 (adapted from Design Control Guidance for Medical Device Manufacturers).<sup>10</sup> This is commonly referred to as the “Waterfall Design Process.” The process describes a development effort, in which user needs are translated to design inputs, which are then used for development, culminating in a set of outputs at various levels of abstraction, which then come together to form a final medical device. Embedded in this process is the need to verify (ensure that outputs meet inputs) and validate (ensure that the medical device ultimately meets the needs of the user). It is out of scope of this editorial to discuss the merits and deficiencies of this model in light of more modern agile methods; however, the concept of applying testing (verification and validation) at a number of different levels of development abstraction (for example, review at code level, unit testing, module interface testing, system functional testing, unstructured testing, and user acceptance testing) to root out as many defects as possible is what we suggest is required for any company developing clinically used pipelines.

**From Code to QMS to Clinical Practice or Back to Code.** For the rest of this piece, we will explain the current QMS practices used in a specific neurotechnology company, Omniscient Neurotechnology, to contextualize how an issue may arise in a neuroimaging pipeline and how it can be safely identified under an effective QMS, as detailed in the anecdote discussed in the section “Z-Shift: One of Many Possible Problems.” Neurotechnology companies, like all software companies, use data scientists who create digital products that require recurrent updates for efficiency and accuracy. When a data scientist creates a new line of code to introduce into production to update or create a product, a series of quality steps according to preoutlined QMS processes must be met before reaching clinical practice. As part of validation (described above), a medical device company often uses clinicians, or neurosurgeons in our case, to help validate the product on the basis of the new update. They are the ultimate user and so can make a determination of whether the product meets their needs. Some aspects of the product, such as tractography generation, do not have ground truth to be checked against. Therefore, it is ultimately up to clinicians and their clinical understanding to

determine whether the tractography that is presented (among other product features) is anatomically plausible.

As part of these validations, subject-qualified clinicians are requested to review the alignment of tractography against underlying anatomic scans and the presentation of tractography in a number of areas with which they have experience. They are presented with a rubric to score a series of scans along the aforementioned dimensions (as well as others), a process that ultimately determines the acceptability of any changes made to the product. Omniscient Neurotechnology maintains predetermined cutoff limits. If said limits are not reached, the evaluation fails and an investigation is performed to determine the root causes and potential corrective actions for the issues. Most important, this procedure is different from simply following a development process that involves clinicians as users (for example, incorporating their feedback). In this process, a predefined and agreed-upon rubric is applied by clinicians using medical images that have not been previously used for development, thereby eliminating subjectivity, which would otherwise contaminate the evaluation.

### ***Advice and Solutions Moving Forward***

We provide just 1 example of a very complicated error that happened using well-tested packages that all independently worked well but together had a roundoff error that went deep into how decimals are represented in the code. Such an issue was only identified with systematic, preplanned verification and validation under established procedures as part of a QMS. The floating-point precision problem is not the only issue that will inevitably arise in many neuroimaging pipelines,<sup>11</sup> and these issues are only safely addressed with good development practices in place. Unfortunately, while regulated medical devices require an implemented QMS, which (as demonstrated in this case) may help in catching faults, de novo home-grown neuroimaging pipelines that can still be legally used in the operating room (due to regulatory agencies not having the authority to regulate medical professions) can represent an unmitigated risk to patients. Unforeseen bursts in public machine learning understanding, capability, and availability have allowed independent researchers to develop advanced neuroimaging technologies; however, such pipelines may not always be working the way they believe and may potentially lead to patient harm if not assessed systematically. Whereas pipelines are not claimed to be medical devices but are used clinically, we encourage end users to explicitly seek evidence of a fit-for-purpose verification and validation framework that aims to root out faults. Most important, such a framework should be available for any pipeline, whether neuroimaging or other.

An additional concern is that similar unforeseen problems to our Z-shift example may be a large contributor to the problems of reproducibility, which are so commonplace in the field of neuroimaging.<sup>6</sup> For instance, complex software programs can often include hard-to-discover bugs in software code, and these may lead to inflated false-positive rates that go unseen in many peer-reviewed journals.<sup>12</sup> When such a bug is placed in a commonly used open-source software for fMRI analysis for instance, this bug may be perpetuated throughout the neuroimaging

community over multiple iterations and eventually lead to decreased reproducibility of results, inefficient use of scientific funding, and ultimately limit our scientific advancement.<sup>12,13</sup> If we are to advance the field of emerging data-driven technologies such as ML and artificial intelligence in general, techniques implemented under the umbrella of a QMS will be imperative to ensure the safety and effectiveness in clinical practice.

Disclosure forms provided by the authors are available with the full text and PDF of this article at [www.ajnr.org](http://www.ajnr.org).

## REFERENCES

1. Glasser MF, Coalson TS, Robinson EC, et al. **A multi-modal parcellation of human cerebral cortex.** *Nature* 2016;536:171–78 CrossRef Medline
2. Doyen S, Nicholas P, Poologaindran A, et al. **Connectivity-based parcellation of normal and anatomically distorted human cerebral cortex.** *Human Brain Mapping* 2021 Nov 2 [Epub ahead of print] CrossRef Medline
3. Sedgwick. **First Edition 2021 Recall Index. US Edition.** 2021. <https://marketing.sedgwick.com/acton/fs/blocks/showLandingPage/a/4952/p/p-036b/t/page/fm/0>. Accessed November 27, 2021
4. Hansen MS, Sorensen TS. **Gadgetron: an open source framework for medical image reconstruction.** *Magn Reson Med* 2013;69:1768–76 CrossRef Medline
5. Matelsky J, Kiar G, Johnson E, et al. **Container-based clinical solutions for portable and reproducible image analysis.** *J Digit Imaging* 2018;31:315–20 CrossRef Medline
6. Wilkinson JH. **Error analysis of floating-point computation.** *Numer Math* 1960;2:319–40 CrossRef
7. Goldberg D. **What every computer scientist should know about floating-point arithmetic.** *ACM Comput Surv* 1991;23:5–48 CrossRef
8. US FDA. **Quality System (QS) Regulation/Medical Device Good Manufacturing Practices.** <https://www.fda.gov/medical-devices/postmarket-requirements-devices/quality-system-qs-regulationmedical-device-good-manufacturing-practices>. Accessed July 1, 2021
9. Eisner R, Patel R. **Strengthening the regulatory system through the implementation and use of a quality management system.** *Rev Panam Salud Publica* 2017;41:e12 CrossRef Medline
10. FDA CDRH. **Design Control Guidance for Medical Device Manufacturers.** 1997. <https://www.fda.gov/media/116573/download>. Accessed 17 February, 2022
11. Bhagwat N, Barry A, Dickie EW, et al. **Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses.** *GigaScience* 2021;10:giaa155 CrossRef Medline
12. Butler RW, Finelli GB. **The infeasibility of quantifying the reliability of life-critical real-time software.** *EEE Transactions on Software Engineering* 1993;19:3–12 CrossRef
13. Poldrack RA, Baker CI, Durnez J, et al. **Scanning the horizon: towards transparent and reproducible neuroimaging research.** *Nat Rev Neurosci* 2017;18:115–26 CrossRef Medline