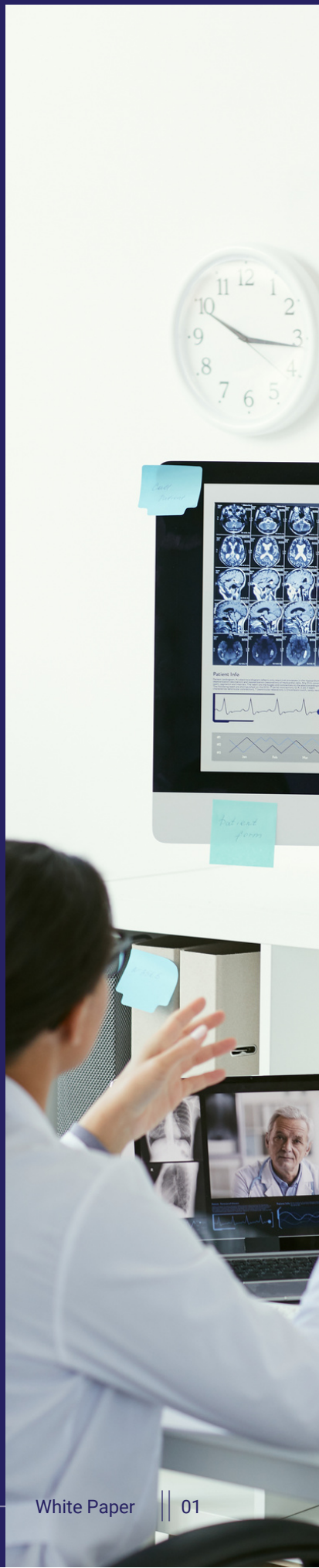
SYMPATIC

WHITE PAPER

How Sympatic helps Universities comply with NIH Policy for Data Management and Sharing.

TL;DR

The National Institutes of Health (NIH) has established a new requirement for data management and sharing (DMS Policy) to promote the management and sharing of scientific data generated from NIH-funded or conducted research. This is of major concern for Universities and Academic Medical Centers as failing to comply can result in losing your grant funding. In contention is keeping your data and IP well protected. Sympatic helps institutions maintain control over their data while also complying with NIH data sharing rules by allowing zero-copy access to the data, providing audit controls, and tearing down environments after use.



Overview

In today's academic community, data sharing is becoming a requirement. The National Institutes of Health (NIH) has established a set of guidelines for data sharing that institutions must follow in order to comply with regulations and maintain funding. However, as custodians of health data, institutions also have a prior obligation to protect patient confidentiality, protect institutional IP, and preserve data assets. Balancing these two obligations can be challenging, but it is crucial for institutions to have a clear path through the rules and regulations surrounding data sharing.

In this white paper, we will provide an overview of the key points to be considered when working with NIH data-sharing rules. We will discuss different metadata standards that can be used to balance patient confidentiality with data-sharing requirements. We will also introduce the use of Sympatic VirtualVaults®, a tool that can help institutions comply with NIH data-sharing rules while maintaining their IP and data assets, and how Sympatic can assist institutions with services and integrations to develop optimal data sharing and compliance without loss of data integrity.

The NIH Data Sharing Policy

The National Institutes of Health (NIH) is issuing a new policy to promote the management and sharing of scientific data generated from NIH-funded or conducted research. The NIH Policy for Data Management and Sharing (DMS) establishes the requirements for submission of Data Management and Sharing Plans and compliance with NIH Institute, Center, or Office (ICO)-approved Plans.

One of the key points of the policy is the emphasis on good data management practices. The policy establishes the expectation for maximizing the appropriate sharing of scientific data generated from NIH-funded or conducted research, with justified limitations or exceptions. This is intended to accelerate biomedical research discovery, as it enables validation of research results, provides accessibility to high-value datasets, and promotes data reuse for future research studies.

Furthermore, DMS is an important step in promoting transparency and accountability of research funded by government resources. NIH has long championed policies that make research available to the public to achieve these goals. The 2003 NIH Data Sharing Policy, the 2014 Genomic Data Sharing Policy, and the 2016 Policy on the Dissemination of NIH-Funded Clinical Trial Information are the precursors to DMS.

It's important to note that the DMS policy is not only a requirement but also an opportunity for researchers, to be able to more widely participate in new discoveries, and collaborations, and accelerate the pace of scientific advancement. Therefore, NIH encourages researchers to take advantage of this opportunity to share their data.

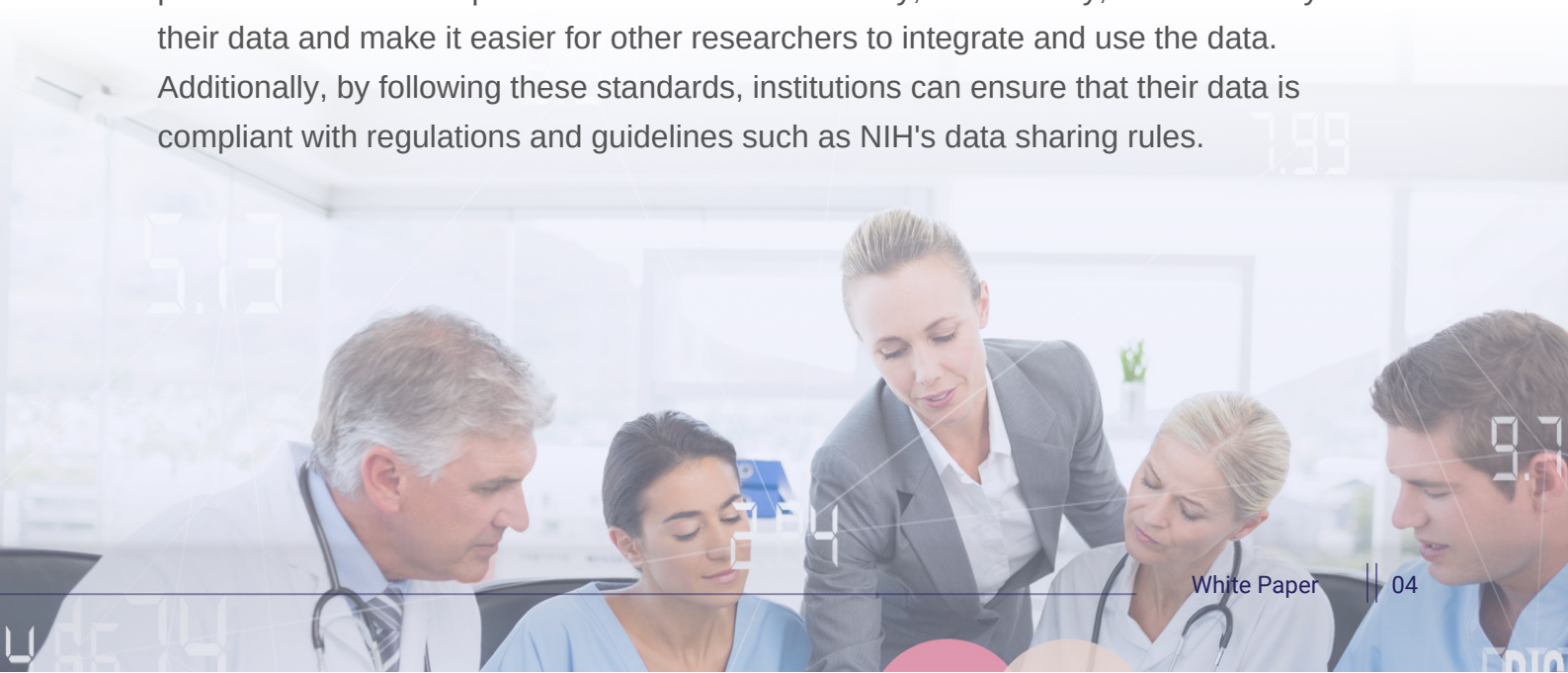
Standards such as FAIR and OMAP lead the way

In addition to data structures, institutions should also consider using metadata standards such as FAIR and common data models such as OMOP. These standards not only help with compliance but also increase internal usability, leading to future opportunities from past work. Standards like FAIR and OMOP provide guidelines for how data should be structured, organized, and shared.

FAIR stands for "Findable, Accessible, Interoperable, and Reusable" and it is a set of principles that provide guidance on how to make scientific data findable, accessible, interoperable, and reusable. While there are good examples of how to maintain this ethos, the standard is less of a formulation. Adhering to the FAIR principles, researchers can ensure that their data is discoverable, accessible, and usable by other researchers, which can increase the impact and reuse of their data.

OMOP stands for "Observational Medical Outcomes Partnership" and it's a common data model that helps to standardize the structure of observational healthcare data. OMOP defines a set of standard data elements, concepts, and relationships that can be used to represent observational medical data in a consistent way across different studies and institutions. This allows for easier integration, comparison, and sharing of observational healthcare data.

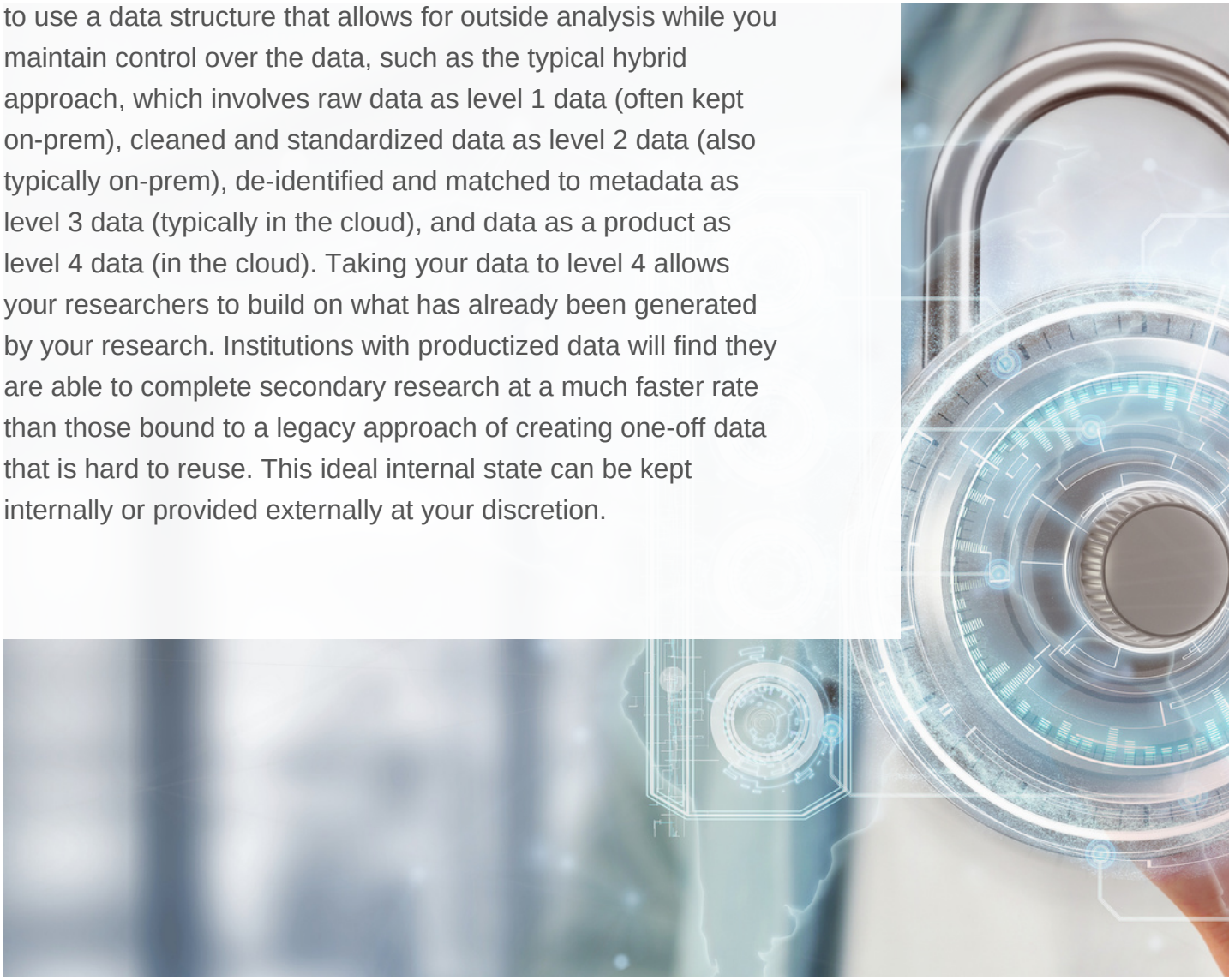
By using standards like FAIR and OMOP, institutions can ensure that their data is structured, organized, and shared in a way that is consistent with established best practices. This can help to increase the discoverability, accessibility, and reusability of their data and make it easier for other researchers to integrate and use the data. Additionally, by following these standards, institutions can ensure that their data is compliant with regulations and guidelines such as NIH's data sharing rules.



Balancing patient privacy/confidentiality with data sharing

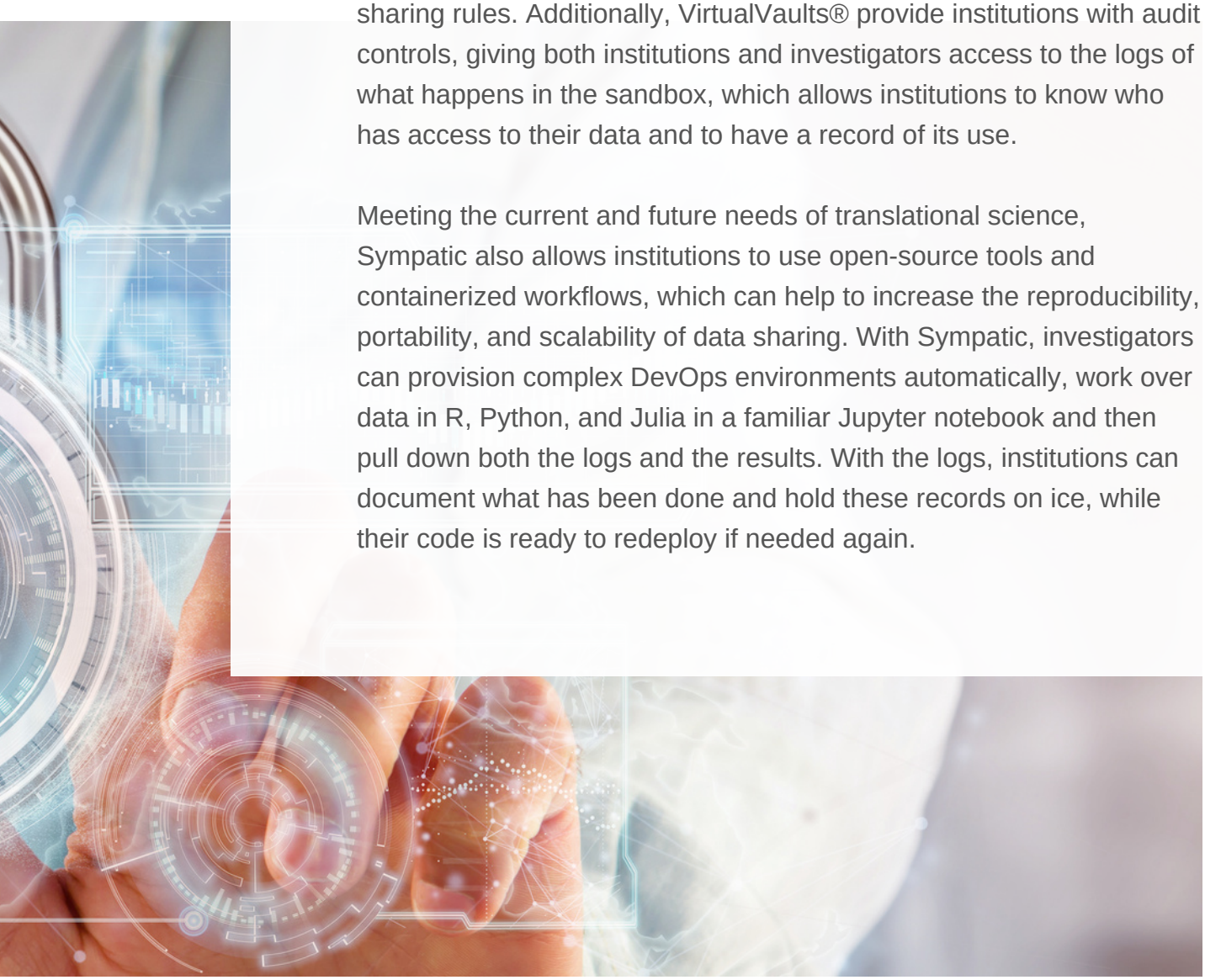
We all recognize the delicate balancing of patient confidentiality with data-sharing requirements is crucial for ensuring the ethical and responsible use of healthcare data. Patient confidentiality refers to the ethical principle of protecting personal information and maintaining patient privacy. Data sharing refers to the practice of making data available to other researchers or organizations for research or other purposes.

Balancing patient confidentiality with data-sharing requirements can be challenging.One way to balance these two obligations is to use a data structure that allows for outside analysis while you maintain control over the data, such as the typical hybrid approach, which involves raw data as level 1 data (often kept on-prem), cleaned and standardized data as level 2 data (also typically on-prem), de-identified and matched to metadata as level 3 data (typically in the cloud), and data as a product as level 4 data (in the cloud). Taking your data to level 4 allows your researchers to build on what has already been generated by your research. Institutions with productized data will find they are able to complete secondary research at a much faster rate than those bound to a legacy approach of creating one-off data that is hard to reuse. This ideal internal state can be kept internally or provided externally at your discretion.



Balancing patient confidentiality with data-sharing requirements is crucial for protecting individuals' rights and privacy while also allowing for the advancement of science. Sympatic can help institutions maintain ownership while meeting data-sharing requirements. Its zero-copy sharing tools enable institutions to maintain control over their data while also complying with NIH data-sharing rules.

One of the key features of Sympatic is the VirtualVault®, which provides institutions with a way to manage data sharing while maintaining control over the data. Unlike traditional data-sharing methods, investigators are given secure access to zero-copy, project based sandboxes called VirtualVaults®. This allows institutions to maintain control over their data and ensure compliance with NIH data-sharing rules. Additionally, VirtualVaults® provide institutions with audit controls, giving both institutions and investigators access to the logs of what happens in the sandbox, which allows institutions to know who has access to their data and to have a record of its use.



Meeting the current and future needs of translational science, Sympatic also allows institutions to use open-source tools and containerized workflows, which can help to increase the reproducibility, portability, and scalability of data sharing. With Sympatic, investigators can provision complex DevOps environments automatically, work over data in R, Python, and Julia in a familiar Jupyter notebook and then pull down both the logs and the results. With the logs, institutions can document what has been done and hold these records on ice, while their code is ready to redeploy if needed again.

The Four Levels of Data

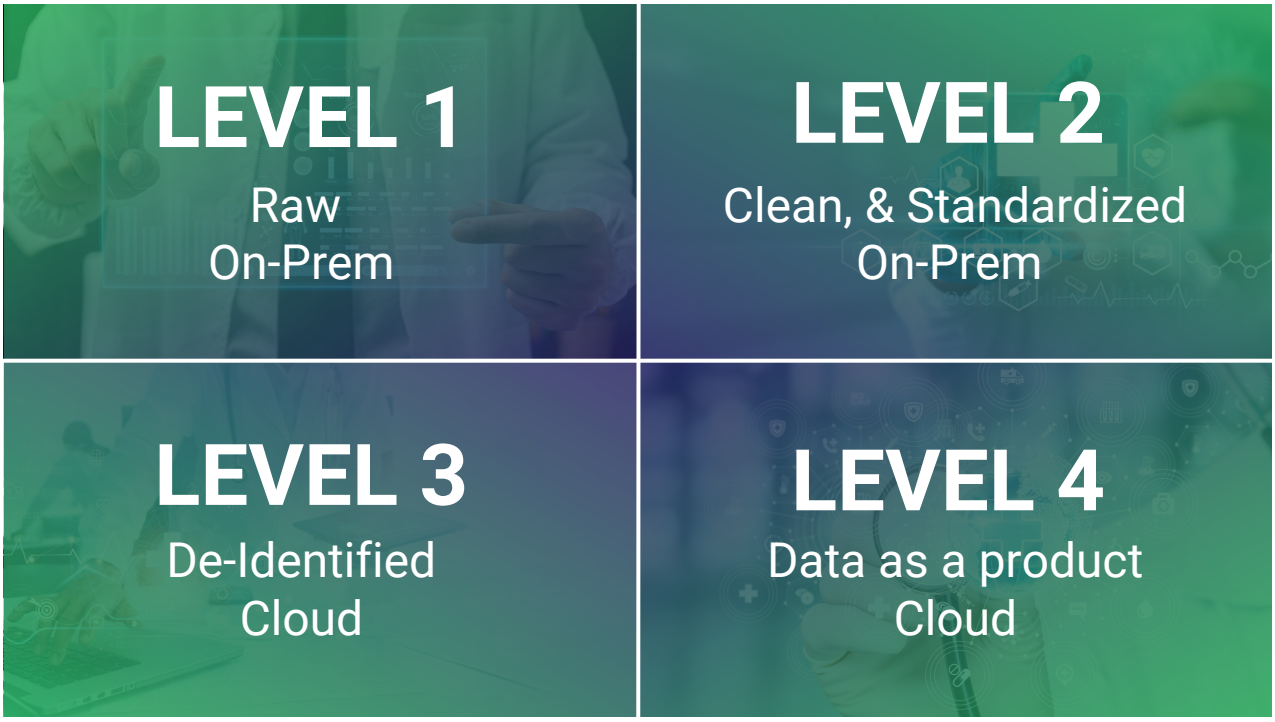
At Sympatic, we think of different levels of data in the data collection, management, and analysis life cycle. We refer to these as the Four Levels of Data. The different levels of data refer to stages of data management and processing. These levels are used to understand data readiness and where to protect patient confidentiality. An explanation of each level and how it relates to the NIH data sharing rules follows:

Level 1 data: This is the raw data that is collected from patients. It is considered the most sensitive level of data and is typically kept on-premises in order to maintain control over the data and protect patient confidentiality. It is not usually shared with external parties, except in rare cases where explicit consent has been obtained and the data have been de-identified.

Level 2 data: This data is cleaned and standardized, meaning that it has been checked for errors, inconsistencies, and outliers. This process ensures that the data is accurate and can be used for research purposes. It is still kept on-premises and is not shared with external parties.

Level 3 data: This data is de-identified and matched to its metadata. Any personal identification information has been removed, and the data is linked to information about how it was collected, who collected it and needed contextual information. This level of data has been considered safe to share with external parties and is often shared with internal researchers, under certain conditions and agreements.

Level 4 data: This is data that is considered "data as a product," meaning that it has been cleaned, standardized, de-identified, and matched to its metadata, but it has also been further manicured with intentional research purposes in mind. For example, data that has been aggregated, anonymized or transformed to make it more suitable for studies in cardio-oncology, with a mix of structured and unstructured data may have subcomponents that would be reusable for a heart disease study. Providing data products as composable building blocks makes these parts easy to rearrange, allowing for internal new uses of the same data with great ease. This drives the velocity of research for your organization.



While level 1 and 2 data may be best served with local storage, level 3 and 4 data should be hosted in a cloud resource to allow for dynamic environments, scaling with use, and minimizing persistence costs. This approach acknowledges that some institutions may not be ready to fully commit to cloud, but allows for the benefits of scalability, accessibility, and performance found with cloud.

What this means for institutions and hybrid cloud

Another key area that institutions must address when working with NIH data-sharing rules is data structures. The NIH guidelines outline three main areas that institutions need to address: data, methods, and outcomes.

One of the key features of Sympatic is its ability to manage data across different levels, such as raw data (level 1), cleaned and standardized data (level 2), de-identified and matched to metadata (level 3), and data as a product (level 4). By using a hybrid approach, institutions can keep sensitive, raw data on-premises while taking advantage of the scalability and accessibility of cloud-based resources for levels 3 and 4 data.

Sympatic also allows institutions to use cloud resources while also acknowledging that some institutions are not ready to fully commit to cloud. With Sympatic, institutions can convert dev-ops infrastructure costs (NIH does not allow funds for this) into egress and storage charges, both of which are allowable under the new regulations. This allows institutions to maintain control over their data while also taking advantage of the scalability and accessibility of cloud-based resources.

With Sympatic, your hybrid cloud approach has the flexibility to straddle legacy infrastructure with a modern application front, providing institutions with the ability to manage data across different levels and use on-prem and/or cloud resources in a way that fits their need and goal. Institutions hosting level 3 and 4 data in a cloud can dynamically change environments, scale with use, and keep persistence costs at a minimum.

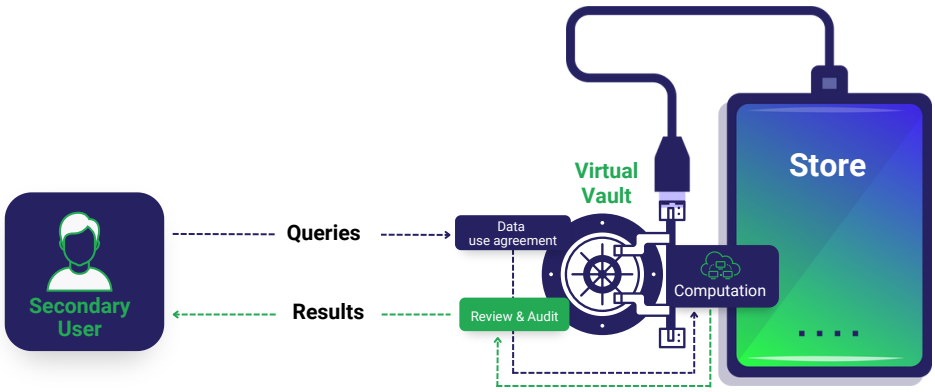
Containerized Workflows create repeatable analysis.

One way to make data sharing more efficient and reliable is by using containerized workflows. Containerized workflows allow researchers to package their entire analysis environment, including software dependencies, into a single container. This makes it easy to share their work with other researchers and replicate their analysis, ensuring reproducibility of results. Additionally, containers are platform-agnostic, which means that they can be run on any system that supports the container runtime. This makes it easy to share and run workflows across different environments, such as local workstations, clusters, or cloud-based platforms.

Another advantage of containerized workflows is the autonomy they provide. Containerized workflows can be easily scaled to meet the needs of large-scale data processing and analysis.

Another benefit of containerized workflows is their cost-effectiveness. Containerized workflows can help reduce costs by eliminating the need to purchase and maintain expensive hardware or software. By using cloud-based container orchestration tools, researchers can run their workflows on pay-as-you-go basis resources and scale up and down environments on demand, leading to an overall lower infrastructure bill.

Finally, containerized workflows provide a level of security by isolating applications from the host system and other containers. This can help to protect against malicious attacks and data breaches. Sympatic makes containerized workflow applications not only simpler but provides a zero-copy environment for the application of workflows along with attestation of use.



How Sympatic can help:

Sympatic provides consulting to assess data management landscape, outline options, develop integrations/APIs, and customize solutions.

The Sympatic team has led and built major data-sharing infrastructure initiatives and helped teams efficiently deliver workflows and applications. Sympatic provides best practices for data management and sharing, as well as assisting with the implementation of policy and procedures. Abilities include developing data management plans, creating data sharing guidelines, and providing training and education on responsible data management and sharing practices. Sympatic can help you identify potential risks with your current data handling and risks associated with providing your data to a third-party repository.

VirtualVaults are themselves based on reusable workflows and dramatically simplify the repeatable deployment and record-keeping process for deploying containerized analytic workflows over data or even simply making the deployment of coding environments simple and adding greater control and oversight.
Sympatic develops both off-the-shelf and customized zero-copy data-use solutions leveraging patented VirtualVault® cloud technology. Sympatic provides planning and integration services in addition to SaaS and patent licensing.

Sympatic provides powerful software and services that help institutions maintain control over their data while also complying with NIH data-sharing rules. Sympatic can help institutions comply with regulations while preventing the loss of Patient data, IP, and valuable data.

About the authors

Dr. Piers Nash is the Founder and CEO of Sympatic. Dr. Nash is a former University of Chicago professor of Cancer Research, a Director for the development of the NCI Genomic Data Commons, IBM Global Consultant for Genomics and Health Data with Watson and Cloud, and Managing Director for the American Medical Association innovation lab.

Krister Kroll is the head of Operations and develops strategic opportunities at Sympatic. Krister holds an MBA from Colorado University and previously scaled operations in aerospace defense.

Trademarks and Patents