

DATA READINESS SURVEY REPORT:

Early Days for AI but Scientific Data Management Gains Momentum

Authors

Paul Denny-Gouldson, Chief Scientific Officer, Zifo

Sakthi Prasad T, Content Director, Zifo

Data Compilation: Karthick Suresh, Analyst

Introduction

The scientific world is in the grips of AI fever, so to speak. Science driven companies are now starting to deeply appreciate the power of data for research and innovation; however, it's still a long road ahead when it comes to data standardization and ensuring ease of accessibility to support the use of AI.

Zifo's recently concluded "Data Readiness Survey," which saw participation of scientists and informaticians from over 30 global companies, shows that scientifically driven companies (Pharma, Biotech, Chemicals, F&B, FMCG, Agrotech etc.) are beginning to seriously invest in AI and ML across the entire R&D, Manufacturing and Trials value chain, and a lot of the initial effort is going into the basics -- making sure data is of good enough quality i.e. there are proper standards, and that everyone's on the same page with things like metadata, ontologies, master data etc.

However, the survey results show that both AI and ML initiatives are at very early stages and have a long road ahead before reaching maturity.

One of the big hurdles facing scientists and informaticians is the issue of data silos -- to fix this would require proper standards and alignment on elements such as metadata, ontologies, and master data. AI/ML initiatives will suffer from a 'garbage in, garbage out' problem without fixing the issues plaguing "data silos". Additionally, scientists and informaticians are concerned about measuring the value of their work, specifically how to quantify the impact of their endeavours on scientific outcomes.

The second big issue facing scientists and informaticians is getting all their lab instruments to talk to each other and exchange data seamlessly. There is a plethora of instruments, each with its own way of connecting, and sometimes the ageing infrastructure doesn't permit seamless data exchange. Fixing that often means a big investment and making sure everything is secure, which encompasses a whole set of other challenges.

While everyone is aiming for better products impacts in the end, the immediate wins Scientists and Informaticians are seeing with AI are aspects such as efficiency, cost savings, potential to speed up discoveries – "innovation", and getting a deeper understanding of the science.

But there are worries about privacy, especially with all the new generative AI tools popping up. Scientists and Informaticians are understandably cautious about where their data end up and how it's being used; that's why more



organizations seek to bring these technologies in-house to keep a tighter grip on proprietary data.

Ultimately, getting the most out of AI and scientific data will depend on how well companies have managed their data in the past, how they tackle these integration issues, and whether they focus on smaller, targeted AI applications that make a real difference.

The collected responses offer a valuable perspective on the current state of technological integration within research and development, manufacturing and trials environments, highlighting the varying degrees of preparedness across different organisations.

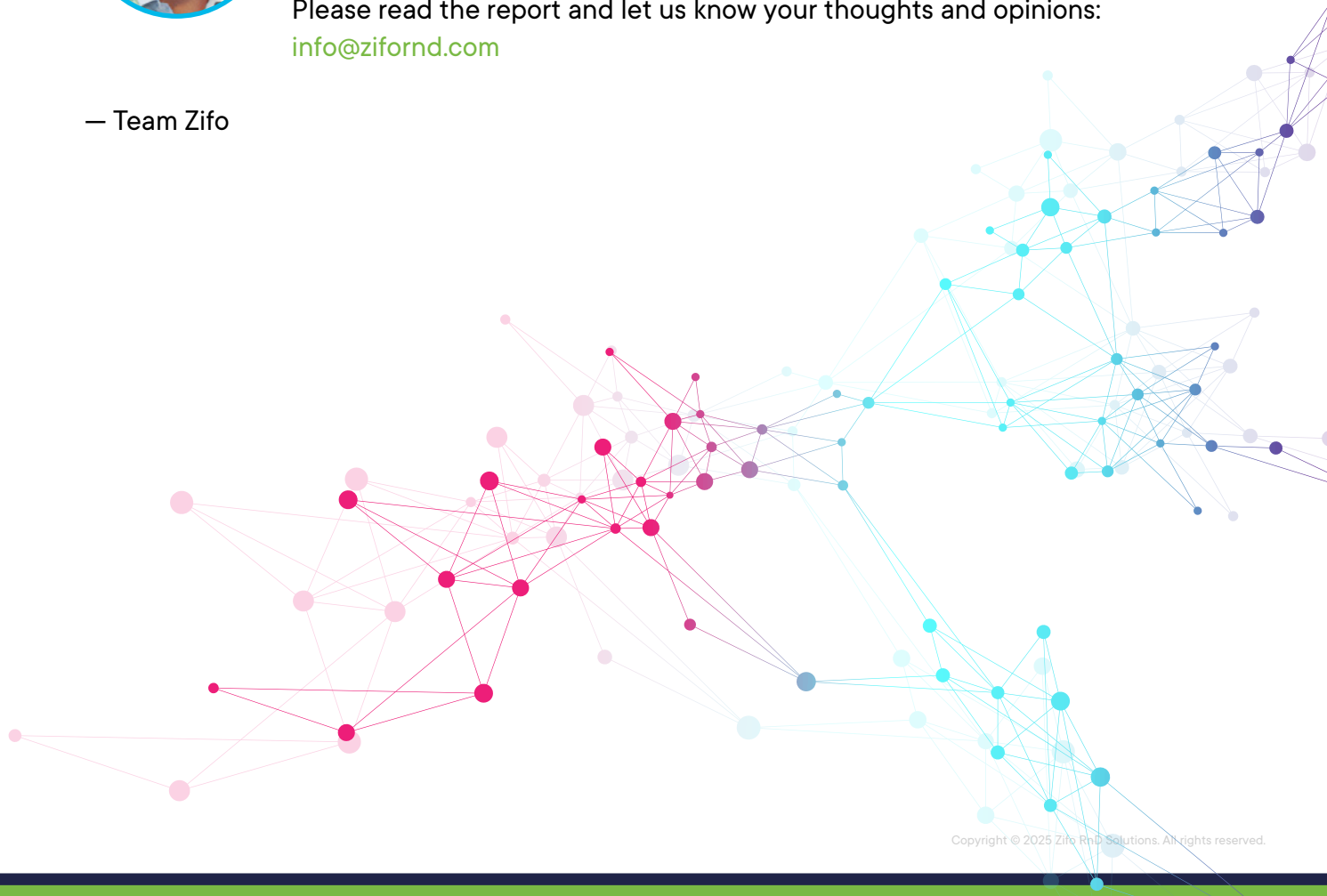
The survey data illuminates the obstacles currently encountered in the seamless integration of data to support artificial intelligence, while also identifying the key perceived benefits and concerns associated with these advancements. Ultimately, this overview provides a concise snapshot of the current landscape, underscoring both the progress achieved and the considerable work remaining to fully leverage the transformative potential of data and artificial intelligence within the scientific community.



Zifo's Chief Scientific Officer (CSO), Paul Denny-Gouldson, has offered his insights on each of the findings. We have also included select quotes from survey respondents.

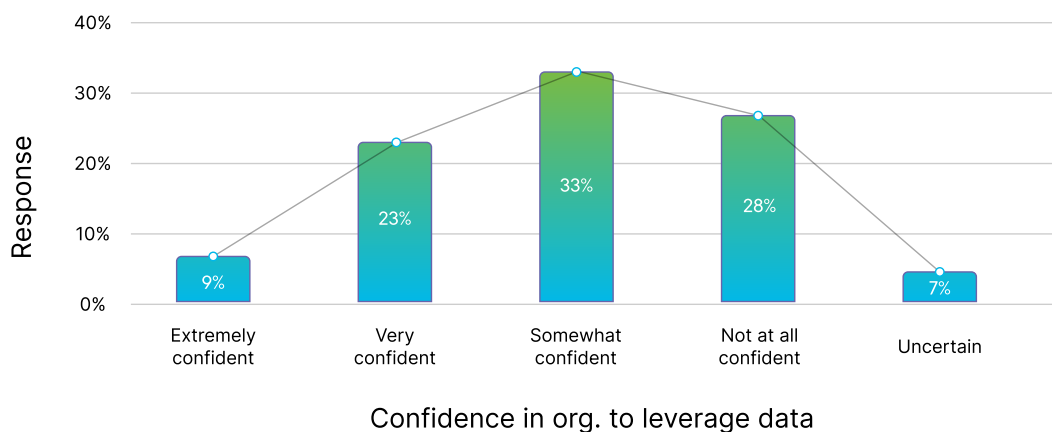
Please read the report and let us know your thoughts and opinions:
info@zifornd.com

— Team Zifo



Survey Findings

1. Given the vast amounts of scientific data generated daily within your organization, how confident are you in your organization's ability to effectively leverage this data to drive AI programs and projects?



Approximately one third (32%) express “high confidence” (Extremely Confident + Very Confident), while a similar portion (35%) are either not confident at all or uncertain. The third group (33%) feels “somewhat confident,” indicating a mixed perspective on the organization’s capabilities. A notable 28% lack confidence entirely, highlighting potential challenges in data utilization for AI initiatives.

Overall, the responses suggest a need to address barriers and build greater confidence in effectively leveraging scientific data for AI.

Respondent Comments

- “Data is not harmonized in terms of storage or metadata”.
- “Since the organization is large and diverse, and each department is driven with a focus to crank out results with available technologies, it is difficult to gauge the ability of each individual department’s ability to use AI driven programs to enhance productivity and efficiency”.
- “Our company has an internal AI program, but it is very limited”.

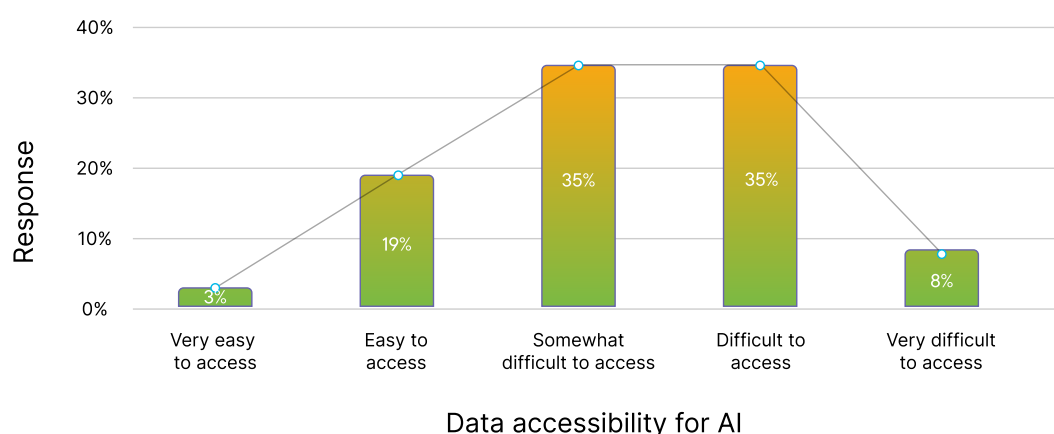
Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

“Scientists are increasingly using data to drive their experimental design, instrument setup, and overall scientific workflow. They not only generate data within their own labs but also frequently need to share it with or receive it from other parts of the scientific process. This data exchange can involve instruments, master data systems, registration systems, data warehouses, or

other data repositories, which are then used by others. Therefore, ensuring data quality, trustworthiness, and adherence to FAIR principles (Findable, Accessible, Interoperable, Reusable) is crucial. The level of “fairness” required can vary depending on the data’s value and intended use, such as for advanced analytics or AI. Data readiness is paramount, along with understanding how the data is consumed. Digital literacy among scientists is also important for them to grasp the importance and usage of their data.

The informatics team’s role is to facilitate data availability, ensure quality, and make it consumable across the organization. The challenge lies in establishing the necessary data foundations without overwhelming scientists with excessive data preparation tasks, balancing this with the costs and time of implementation. IT serves as a support function to enable scientific progress and must adapt quickly to evolving scientific needs. While established, routine lab work is easier to manage, the dynamic nature of scientific research requires flexibility”.

2. In your experience, how easily accessible is the scientific data required for AI initiatives within your organization?



A substantial majority (70%) of respondents report that accessing scientific data for AI initiatives is either difficult or somewhat difficult within the organization. Conversely, only a small fraction (22%) find the required data to be easily accessible. The equal percentages for “Difficult” and “Somewhat difficult” access (both 35%) indicate a widespread and significant challenge.

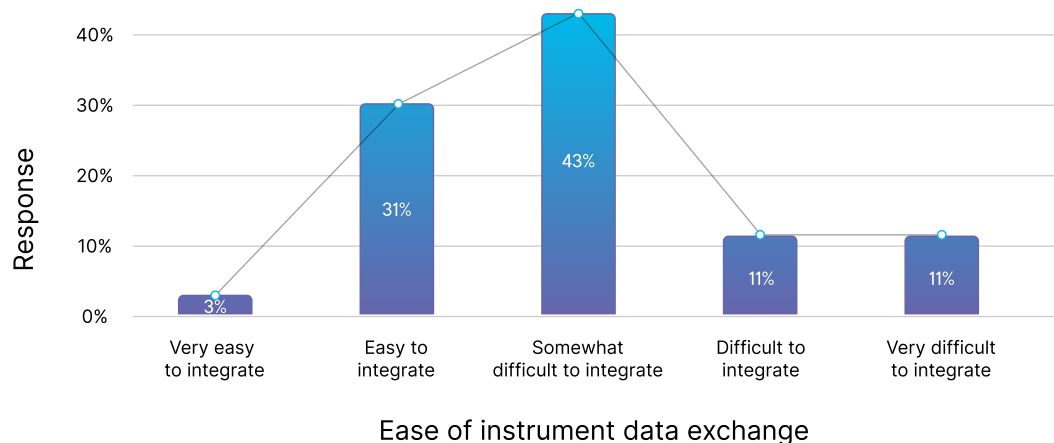
Overall, the data strongly suggests that improving data accessibility is a critical area for supporting AI programs within the organization.

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

“The current state of AI and machine learning adoption and integration within scientific organizations is still in its early stages. A key focus is on understanding and educating scientists about the necessity of high-quality data that is well-documented with metadata. This ensures the data can be effectively utilized for algorithm development and machine learning initiatives. While AI can assist in data cleaning, the primary challenge lies in establishing a foundation of clean, well-described data. The speed at which this landscape changes over time will be a key indicator of progress. If minimal improvement in data accessibility for AI is observed in the near future, it might suggest that current strategies are insufficient or that this is a more protracted issue.

Given the ongoing nature of scientific research, solutions must be implemented alongside active projects. Unlike a “greenfield” scenario, the existing complexities and change management aspects within established organizations (“brownfield”) make this a significant change management and value demonstration challenge. Organizations and IT departments need to collaborate with the scientific community to evolve existing systems, rather than implementing radical overhauls. This evolution requires substantial investment”.

3. How would you describe the ease of data pipelining, integration and exchange of lab instruments data in your organization?



Most organizations find data pipelining, integration, and exchange of lab instrument data to be challenging. Specifically, 43% find it “Somewhat difficult to integrate,” and a further 11% report it as “Difficult to integrate,” while another 11% find it “Very difficult to integrate.” In contrast, a smaller portion find it “Easy to integrate” (31%), and a very small fraction consider it “Very easy to integrate” (3%).

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

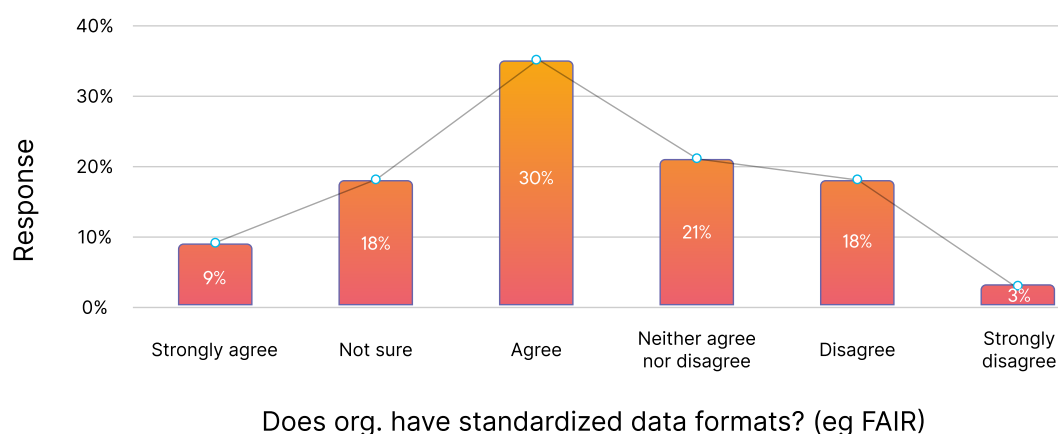
"The difficulties in integrating lab instrument data stem from several factors. These include the need for configuration and maintenance, the complexities of connecting to lab execution environments like ELNs or LIMS, and the intricate, often multidimensional, nature of the data itself. The diverse and heterogeneous instrument landscape, while adhering to connectivity standards like RS 232 or whatever it is, historically lacked the necessary physical infrastructure (cabling) for seamless data transfer. Addressing this often requires significant investment in infrastructure upgrades and ensuring network security, especially in regulated environments.

Furthermore, the cost-benefit analysis of full integration can sometimes favor manual or paper-based systems, though this can impact scientist satisfaction.

Establishing a 'single view of truth' for data is fundamental for progress, requiring instruments to communicate and data to be centrally stored in a standardized format. The importance of this varies depending on the stage of research, development, or manufacturing, with raw data being crucial in regulated environments. While archiving is essential for regulatory compliance, transforming archived data into a usable data source for analysis requires additional effort to meet FAIR principles.

A key challenge is determining which data will hold future value, particularly in research. Leaders often question how to identify and prioritize data for long-term investment when future relevance is uncertain. This highlights the need to balance current data integration efforts with the potential future value of the data being captured".

4. Does your organization have standardized data formats and ontologies (for eg. FAIR principles) that enable ease of data retrieval?



A significant portion of respondents (39%) agree or strongly agree that their organization has standardized data formats and ontologies facilitating data retrieval. However, a considerable 39% are either unsure or disagree, indicating a lack of standardization in some organizations. About 21% remain neutral on this aspect. This suggests a mixed landscape regarding the adoption of standardized data practices for easier data access.

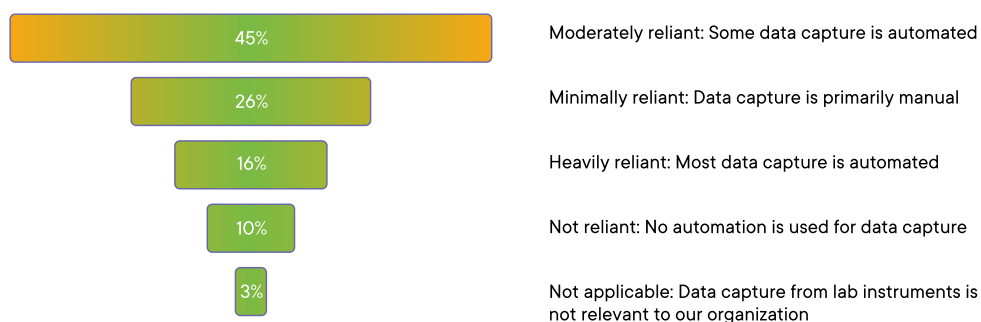
Respondent Comments

- “Some departments work with FAIR principle, while others are trying to implement it”.
- “Work in progress”

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

“Standardized data format does depend on the organization and how focused they have been in the last 10 years or so when it comes to putting up infrastructure to capture data and data reuse. So, you see a big spread of organizations where some have not been spending a lot on data infrastructure, while a few did. And for those who haven’t invested in harnessing their data, it is now becoming a problem because of the requirement to do all the AI machine learning, which is dependent on data reuse, ontologies and so on. Data Management is fundamental to ensuring data reuse and data retrieval, because that is the lifeblood of what enables FAIR data.”

5. To what extent does your organization rely on automation for data capture from lab instruments?



Based on the survey results, a significant portion of organizations (45%) are moderately reliant on automation for data capture from lab instruments, indicating a blend of automated and manual processes. A substantial number (26%) still primarily rely on manual data capture, highlighting a lower level of automation in these organizations. Conversely, a smaller but notable percentage (16%) heavily automates their data capture, while a minority (10%) uses no automation at all. Finally, a small fraction (3%) indicates that data capture from lab instruments is not relevant to their operations.

Respondent Comment

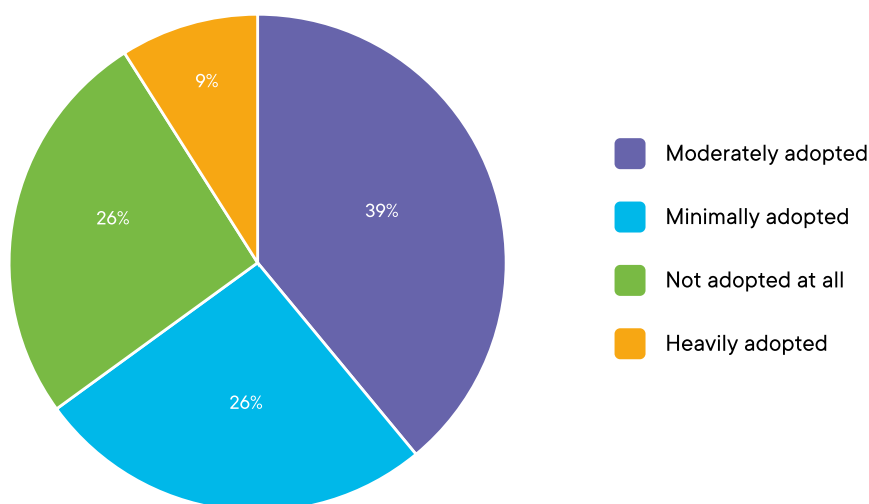
- “Would be nice to have more automation”

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

“The readiness for data reuse and AI/ML adoption varies significantly across organizations, largely reflecting their historical investment in data management over the past decade. Organizations that haven’t prioritized data infrastructure are now facing challenges due to the increasing need for AI, machine learning, and broader data utilization. Essential data management practices, including standardized data, master data management, and ontologies, are fundamental for ensuring data’s ‘fairness’ (Findable, Accessible, Interoperable, Reusable) and facilitating effective data retrieval.

While some organizations report having these crucial elements in place, a substantial either express uncertainty or disagree, indicating a widespread deficit in foundational data management capabilities. This suggests that the primary factor influencing data readiness is likely the individual company’s approach to data management rather than the specific sector (research, development, manufacturing or trials) it operates within. Many organizations still lack the core data management infrastructure necessary to fully leverage current data-driven technologies.”

6. To what extent has your organization adopted AI in its R&D operations?



The survey reveals that most organizations have adopted AI to some extent in their R&D operations. A significant 39% have moderately adopted AI, indicating its integration into some aspects of their R&D processes. Furthermore, a smaller but still considerable 26% have minimally adopted AI, suggesting initial or limited implementation. Notably, an equal percentage (26%) have not adopted AI at all in their R&D operations, while a smaller 10% have heavily adopted AI, signifying its widespread use.

Respondent Comments

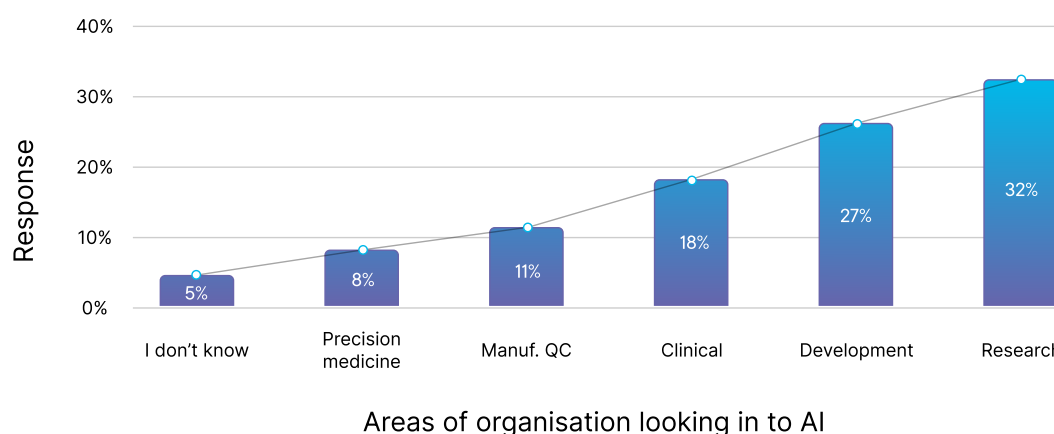
- “Getting more and more adopted”.
- “Work in progress”
- “Just starting with artificial intelligence”

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

“The level of automation in data management, including the use of ontologies and master data management principles, aligns with previous observations, with a majority indicating some level of implementation. Automation becomes particularly beneficial in manufacturing and regulated environments by minimizing human intervention and reducing the risk of data quality issues associated with manual data entry.

However, the decision to automate is also influenced by user needs and the frequency of the task. For instance, in research settings with diverse activities, the effort required to automate a less frequent process might outweigh the benefits, leading to a preference for manual methods. Conversely, in high-throughput research or workflows heavily reliant on robotics, full automation is often prioritized to maximize efficiency and enable continuous operation, facilitating a higher volume of experiments than manual approaches would allow.”

7. To the best of your knowledge, which areas of your organisation are looking at using AI? (Multiple Selections Allowed)



Organizations are primarily exploring AI adoption in Research (32%) and Development (27%) functions, indicating a strong interest in leveraging AI for innovation and product creation. Clinical applications are also being considered by a notable portion (18%), while Manufacturing QC and Precision Medicine show emerging interest (11% and 8% respectively). A small percentage of respondents (5%) are unsure about specific areas of AI exploration within their organization.

Respondent Comments

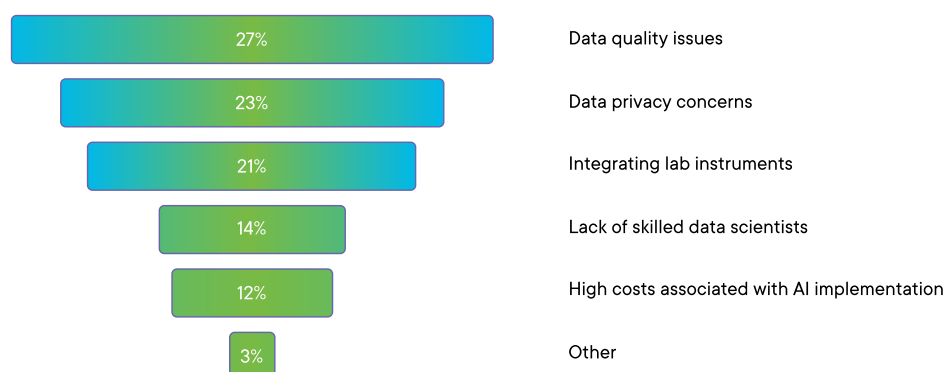
- “Barely looking”
- “Medical Affairs”
- “Interpreting huge amount chromatographic data, impurity profiling comparisons, and report generating”.
- “There’s interest across the board, but I can’t speak for some of these areas”.
- “All”

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

“The current trend in leveraging AI in R&D and related fields is shifting towards targeted, specific applications that augment human decision-making rather than fully automating it. Initially, the focus will likely be on AI assisting scientists in identifying data trends and providing insights, while human experts retain responsibility for final decisions. This approach is particularly relevant in research, where the primary need is for generating new ideas without immediate regulatory constraints. Agentic AI is beginning its journey, but this is where the decisions are made automatically based on inputs – there are various safety, regulatory and ethical concerns that need to be addressed before this is widely adopted.

As projects progress towards development and clinical stages, the influence of regulatory considerations increases. Market feedback indicates a preference for numerous small, focused, AI interventions across a given process and value chain, rather than large-scale, all-encompassing AI solutions. These smaller integrations aim to incrementally improve speed, precision, and insights, enabling faster and more informed decision-making. While large-scale AI projects garner significant media attention, the true value is increasingly seen in integrating AI and machine learning into routine workflows across the entire value chain”.

8. According to you, what is the biggest challenge your organization faces in leveraging scientific data for AI? (Multiple Selections Allowed)



Data quality issues are the most significant challenge (27%), followed closely by data privacy concerns (23%). Integrating lab instruments is also a major hurdle (21%). A lack of skilled data scientists (14%) and high costs associated with AI implementation (12%) are also notable challenges. A small percentage of organizations (3%) cited other challenges.

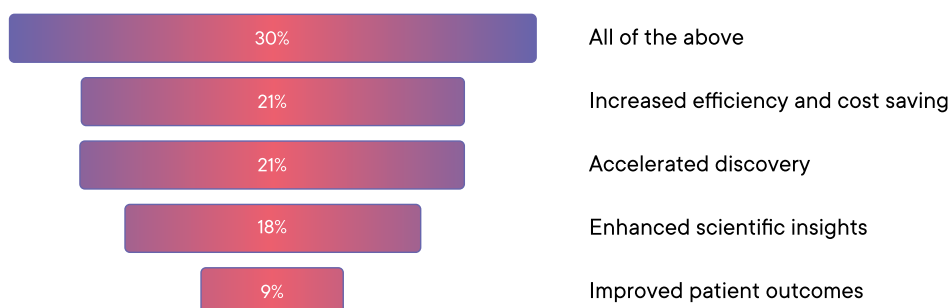
Respondent Comments

- “GxP”
- “Lack of mindset”
- “Most data management systems focus on lab notebooks. I run the HPC system, and ELNs do not help with the petabytes of unstructured data a typical HPC system has, so there’s a gap between neat capture of data from the instrument (which is well automated) and storing of fair data products of analysed data (which we have well covered). In between the processing of the instrument data and other HPC analysis is not well covered”.

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

“The prominence of data privacy concerns as a major challenge in leveraging scientific data for AI (ranking second) reflects a fundamental human instinct to protect one’s data. Concerns arise from the potential for misuse or lack of understanding regarding how data will be utilized, particularly in the context of data reuse. The example of public generative AI tools like ChatGPT highlights these anxieties, as input data can inadvertently become part of the public learning model, posing significant confidentiality risks. This has prompted organizations to rapidly adopt in-house versions of such technologies to better manage and secure sensitive information within their own networks”.

9. What do you see as the greatest potential benefit of AI in biotech and pharma? (Multiple Selections Allowed)



A significant portion of respondents (30%) believe that “All the above” options represent the greatest potential benefit, indicating a holistic view of AI’s positive impact. Among the individual benefits, “Accelerated discovery” and “Increased efficiency and cost savings” are seen as highly beneficial (21% each). “Enhanced scientific insights” is also considered a major advantage (18%), while “Improved patient outcomes” is also recognized, though slightly less frequently (9%).

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

“While improved patient outcomes should ideally be the end goal, the survey results suggest that respondents are also strongly focused on other tangible benefits of AI. These include increased efficiency and cost savings, accelerated discovery, and enhanced scientific insights. These intermediate goals are often seen as direct pathways to achieving better patient outcomes, as improvements in these areas contribute to the development of better products and processes.

This perspective highlights the practical application of AI, where measurable outcomes like optimized processes, faster discoveries, and deeper data understanding are readily apparent and contribute to the overarching objective of improved patient care. The effectiveness of AI in achieving these benefits is most evident when applied in a targeted manner with a clear understanding of the use case and the expected value”.

| Looking Ahead

Data standardization and seamless data exchange across R&D, Manufacturing and Trials value chain is critical for scientifically driven industries such as Pharma, Biotech, Chemicals, F&B, FMCG, and Agrotech. If one were to put a label on the current era, we could call it “The Age of Data Management”, which will then eventually lead to “The Age of AI”.

| About Zifo

Science led, people driven, technology centered.

Zifo is the leading global enabler of AI and data driven enterprise informatics for science driven organizations. With extensive solutions and services expertise spanning research, development, manufacturing, and clinical domains, we serve a diverse range of industries, including Pharma, Biotech, Chemicals, Food and Beverage, Oil & Gas, and FMCG. Trusted by over 190 science-focused organizations worldwide, Zifo is the partner of choice for advancing digital scientific innovation.