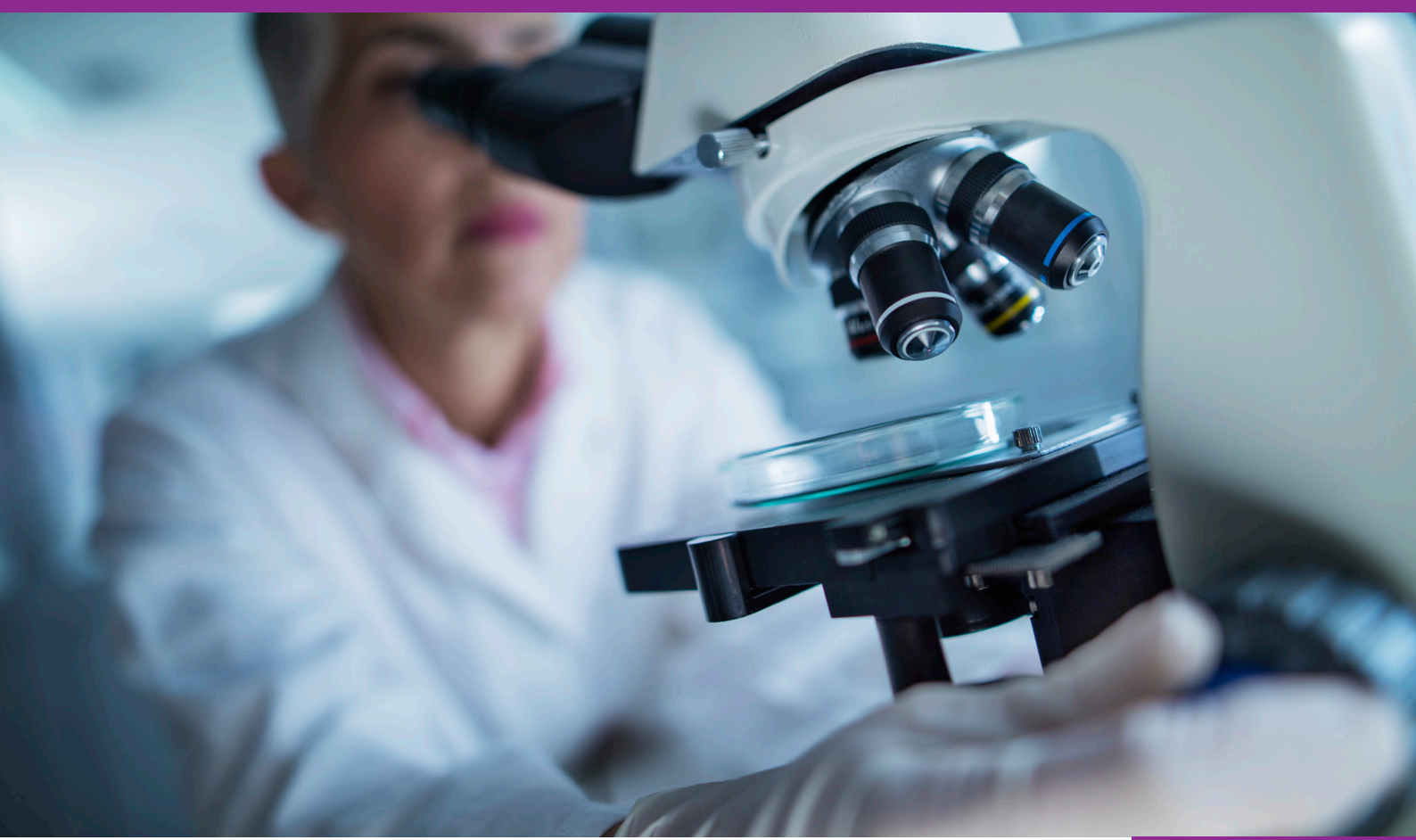
CONCLUSIONS PAPER

Institutionalizing Real-World Evidence

Best practices for establishing the essential data and analytics foundation



sas
THE POWER TO KNOW.

Contents

From clinical trial to the real world

Clinical trials are intended to demonstrate the performance of a therapy in a human population, but they are constrained. Studies recruit patients with very specific characteristics and conditions – certain lab parameters, age ranges, health conditions, previous treatments and so on. In the quest to demonstrate safety and efficacy for regulatory approval, studies are targeted to a certain type of patient, using populations and inclusion criteria that don't necessarily represent the full spectrum of use in actual medical practice.

That's why we're seeing a lot more discussion around *real-world data* and *pragmatic clinical trials* with much larger sample sizes and broader eligibility criteria. Real-world data is also collected from a wide variety of external sources. According to the US Food and Drug Administration:

> These [data] sources may include large simple trials, or pragmatic clinical trials, prospective observational or registry studies, retrospective database studies, case reports, administrative and healthcare insurance claims, electronic health records, data obtained as part of a public health investigation or routine public health surveillance, and registries (e.g., device, procedural or disease registries).
>
> The data is typically derived from electronic systems used in health care delivery, data contained within medical devices, and/or in tracking patient experience during care, including in home-use settings.

In short, real-world data comes from a multitude of novel data sources. *Real-world evidence* is derived from aggregating and analyzing that real-world data – as much of it as is relevant, available and consumable.

The ability to quickly transform real-world data sources such as claims data or electronic medical records (EMR) into evidence can improve health outcomes for patients by helping life sciences firms be more efficient in drug development and smarter in commercialization.

However, the move toward real-world evidence is putting new pressures on researchers, data scientists, and the data and analytical processes they rely on.

What falls short, what needs to change, and why now?

As life sciences organizations move toward greater use of real-world data, they face several challenges.

Massive data volumes. "The volume and variety of data for real-world evidence generation are growing exponentially," said Ari Yacobi, Chief Data Scientist at the data and analytics consultancy Knowledgent. Whereas traditional clinical trials generated finite and carefully controlled types of data, the real world has neither constraints nor conformity.

In many organizations, the sheer volume of available real-world data is outstripping the capacity to digest it.

"Electronic health records provide patient data collected at point of care in near-real time, blurring the line between retrospective and prospective research," said Yacobi. "Add previously untapped sources of data, such as imaging data and free-form text from physician notes. Beyond health systems, there's also a wealth of data generated by individuals' mobile health devices and social media." Vast, diverse and disparate data sources are coming at us faster than ever.

The need for speed. In the face of a data tsunami, there's the pressure for rapid discovery. When analyst teams grapple with a backlog of projects, or a safety issue is looming, who can wait 10 or 12 days for results of analysis – then another week or more for the next round of an iterative query process?

"Jobs have typically taken a very long time to run," said Robert Collins, Senior Life Sciences Industry Consultant at SAS. "The queries are difficult and complex to write. It's hard to find individuals who know the intricacies of different data formats and the caveats to working with each of them."

Limited adoption. There's a lot of merit in getting more groups across the organization to adopt real-world evidence. What may have been localized to the health economics group should be expanded to R&D and commercial groups, if it isn't already. However, these groups have unique needs and are looking for their own breed of analysis and reporting from the data. The result has been redundant effort, duplicated data and little collaboration.

Lack of cohesion. "In most organizations, we see analytics implemented as a bit of a patchwork, where there's a combination of commercially available products and internally developed analytic tools, most of them running in silos," said Yacobi. "Different groups use different tools to do their own analyses. Many data providers provide query tools to analyze their own proprietary data. Some platforms target specific types of analysis, such as for pharmacovigilance, while some focus on a particular type of data.

"In most cases we have seen all the above within the same organization. Gaps that exist within each tool are addressed manually, which results in other gaps, overlaps and inconsistent workflows." When data and analytics processes grow and change in an uncontrolled way, they become unwieldy to use, manage and maintain.

This is what happens when organizations rely on one-off processes, disconnected tools, costly and redundant data stores, and ad hoc discovery methods. It's time for life sciences organizations to standardize real-world data and analytics platforms – to establish much-needed consistency, governance, sharing and reuse. Organizations that succeed at this will formalize their knowledge base and make it scalable, while significantly reducing costs.

"Today we see a lot of repeat processes within different groups, where one group is doing the exact same work that the other group is doing. They're reinventing the wheel because they're not aware of the work that's being done, and since it is probably done using a different analytic tool or data model, it's not reusable across functions."

Ari Yacobi, Chief Data Scientist, Knowledgent

Exponential growth in the velocity and variety of data sources – and the need for greater adoption across functional business units – is requiring life sciences organizations to standardize their real-world data and analytics platforms.

Key components for institutionalizing real-world evidence

To formalize the practice of managing real-world data and generating real-world evidence from it, organizations need to have several foundational capabilities.

Unified data architecture

"With a unified data architecture, the IT team has one cohesive set of tools to manage and maintain," said Yacobi. "Tools can be used across functional groups, so epidemiology, commercial or R&D groups are all working off the same set of tools and analytic models. And third, as you build more predictive analytic models on this architecture, you're compounding the work that has already been done."

Platform for the rapid ingestion of data

"With real-world data sources growing exponentially, speed to desktop becomes another important theme," said Yacobi. "The real-world evidence platform needs to get internal and external data to researchers quickly, in days rather than weeks.

"One of themes we see is migration of data from real-world data sources from purpose-built storage appliances into the Hadoop environment, because it's cost-effective and it scales nicely."

An informal poll conducted during a SAS webinar showed file-based systems (such as SAS data sets or CBV files) being currently the most widely used storage mechanism, followed by relational databases, then Hadoop and cloud-based solutions. Dedicated storage appliances were least popular.

Common data standards

Databases vary in purpose and design. For instance, EMRs are designed to support clinical practice at the point of care, while administrative claims databases are designed for the insurance reimbursement processes. Each has its own logical organization, physical format, terminology and coding schemes.

A common data model transforms data from these disparate databases into a consistent format for analysis. Several common data models have been developed to promote interoperability, such as:

- Observational Health Data Sciences and Informatics (OHDSI)/Observational Medical Outcomes Partnership (OMOP).
- FDA Sentinel.
- National Patient-Centered Clinical Research Network.
- Innovation in Medical Evidence Development and Surveillance.

"Some standards are more mature than others," said Yacobi. "For example OHDSI/OMOP is very commonly used, but there is no one common standard. Even within an organization we see different functional groups comply to different standards. For

> "Real-world data is astoundingly diverse, but it can be tamed to be useful and credible for analysis. With a standardized data model and analytics tools, everyone is singing off the same music sheet."
>
> Ari Yacobi, Knowledgent

example, health economics may prefer OMOP while the safety group prefers Mini-Sentinel. We see a lot of healthy debate to move toward one standard, but there will always be a group that wants its own data model, and that limits the ability to standardize analytics."

Yacobi said SAS and Knowledgent are "tackling this issue with an intermediate data model, which is an overlay on top of the commonly used data models. Together we are building ETL (extract, transform and load) capabilities that can automatically transform data from our intermediate data model to OMOP, Mini-Sentinel, and to SAS submission. We believe researchers should be able to download or access the study population in the data model of their liking."

Well-governed data management process

"We have always been able to take different data sources, such as claims, EMR, registry, safety and unstructured data, and transform them into a standard structure," said Collins. "It's important that all of these transformations occur in some sort of reproducible mechanism, and with governance." There are three elements to this:

- Governing the transformations, so you know they occur the same way each time.
- Governing access to make sure the right people are accessing the right information.
- Providing a structured way to maintain data processes to ensure a consistently behaving environment.

Standard cohort definitions

"Standardizing the data sets has two parts to it," said Yacobi. "One is having a common model that everyone agrees and accepts, and second is the ability to build and define cohorts in a consistent and repeatable way. Using a common model, your cohort [definition] can be run against any data source, because all the data sources are mapped to that common data model. Being able to identify a cohort independent of a data source – to standardize the data to one model – will play an important role as you institutionalize real-world evidence."

Templated analytics use cases

"Basic use cases that are common across the organization, such as signal detection, should be templated," said Yacobi. "Having a common set of use cases that are ready to go helps accelerate adoption, while more sophisticated use cases can be built on top of them. When foundational use cases have been tackled up front, less time is spent building the basic use cases that are repeatable, and more time spent doing predictive analytics."

Self-service capabilities for analytics

The goal is to make basic descriptive and diagnostic analytics available to a self-service interface for two purposes: (A) to enable analysts to refine their queries without having to go back and forth writing manual code, and (B) to empower nontechnical users to run their own basic analyses, thereby freeing analysts for more high-value activities such as predictive analytics.

> "The intermediate data model enables us to standardize the cohort-building process or commonly used analytics use cases to one model that can be scaled across the organization, while giving researchers the ability to use any model of their choice."
>
> Ari Yacobi, Knowledgent

Bring big data together for analysis without stifling individual choice of analytic tools, then deliver analytic results in a self-service fashion that suits both data-savvy and nontechnical users.

Take real-world evidence to the next level

Beyond the foundation components, SAS and Knowledgent are bringing advanced capabilities that make real-world evidence faster, more intuitive, more powerful, and more consistent and repeatable, including:

- Close coupling of SAS and Hadoop.
- Standard, customizable cohort builder.
- Traditional and visual analysis.
- Advanced analytics for deeper insights.

Close coupling of SAS® and Hadoop

Hadoop is a software framework for running applications on large cluster of commodity hardware – hundreds or thousands of servers. Bringing a high-tech twist to the adage "Many hands make light work," Hadoop stores data on local disks, and processing is pushed to the data rather than vice versa for faster processing speeds. Based on the average cost per terabyte of compute capacity, Hadoop is easily 10 times cheaper for comparable computing capacity compared to higher-cost, specialized hardware.

SAS offers capabilities to work from, in and with Hadoop, said Collins.

- **From Hadoop** – "If you want to treat Hadoop as just another data source for SAS, we have mechanisms for doing that with SAS/ACCESS® to Hadoop and Cloudera engines."
- **In Hadoop** – "Our data loader platform for Hadoop allows you to embed SAS processes – such as for data quality, data transformation and manipulation, and scoring inside Hadoop, and they actually execute there."
- **With Hadoop** – "Customers are using our asymmetric load capabilities to very quickly load large volumes of data from a Hadoop environment into SAS in-memory tools."

Standard, customizable cohort builder

For organizations that use a cohort builder, the choices have traditionally been at two ends of the spectrum:

- Simple and easy to use but with limited capabilities, which hindered adoption.
- Sophisticated in analytics capabilities but so difficult to use that technical expertise was required.

SAS and Knowledgent have created a cohort builder that strikes a better balance of sophistication and ease of use, with the following characteristics:

- **Intuitive user interface** – "We think the best way to build a cohort is through a graphical interactive tool," said Collins. "It should support those who understand the population but don't necessarily know how to code, and guide them through the process of selecting the criteria of interest."
- **Support for complex queries** – "It's not enough to just do some simple subsetting," said Collins. "That graphical tool needs to support the complex query logic often required for these types of analysis, such as multiple events and temporal relationships between activities or events in a patient's history."

Analytic tools already in place (such as SAS® Enterprise Miner™), tools relating to data management (such as SAS® Data Integration Studio) and foundation tools (such as Base SAS®) can be used to work with Hadoop data. What you have now becomes even more valuable.

- **Data-agnosticism** – Some cohort builders were created by data providers and are specific to that provider's data format(s). "Being able to define one cohort and run it across multiple data sources – and bring it to a place where you can work on it within minutes versus hours – has huge value," said Yacobi.
- **Access to the data for analysis your way** – "With some cohort builders we've seen out there, you can specify your cohort, but at the end of the day, the only thing you get out of the product is summary analyses," said Collins. "It's a dead end; you don't get access to the data. We think you should be able to take the data and do work downstream from the query tool itself.

  "We also think it should be more than just a black box. Some tools allow you to do certain analyses, but only the way they've defined it. If you want to do something different, or use a different algorithm, you can't. A tool should adapt to your needs and allow you to implement the analytics you'd like."
- **Rapid feedback** – "As a user is selecting inclusion/exclusion criteria, it's important to instantly see how that affects the final population," said Collins. "Have I just applied criteria that drops me down to 10 patients, so I should go back and relax some criteria a bit?" Rapid feedback, coupled with the user's domain knowledge, makes quick work of specifying a cohort.
- **Flexibility** – "What comes out of the box may not work 100 percent for all the functional groups in the organization," said Yacobi, "You need to be able to expand and adapt the cohort builder, to customize it into something that's fitting and unique for your organization."
- **Reusability** – "In many cases, you can expect to have very similar cohorts to deal with again, for instance, for similar work across a therapeutic area," said Collins. "You can reduce time and cost by being able to save, reuse and manage cohort definitions for future use."

Traditional and visual analysis

SAS gives users access to real-world data with tools that fit their needs and capabilities – traditional, code-based interfaces or intuitive graphical tools with one-click selection of data sources, drop-down menus, sliders, filters and auto-generation of a wide variety of diagrams.

For users who have some understanding of the data, the population and what they're trying to do (but are not trained statisticians), a visual interface enables ad hoc exploration. But the visual analytics interface benefits technical users as well, Collins explained. "As a statistician, I can apply models, perform regression analysis and so forth in a visual tool that's working in memory. So now my large data is being fit to my models very rapidly, and I can hone in on which variables are important or of most interest. Then I can take that insight and do a more detailed, maybe hand-coded type of analysis of the data."

Rapid results from visual analysis show researchers where to dive into the details, which can lead to better understanding of a population, better model fit, or seeing new indicators or safety signals that might otherwise have been missed.

"We think the best way to build a cohort is through a graphical interactive tool. It should support those who understand the population but don't necessarily know how to code, and guide them through the process of selecting the criteria of interest."

Robert Collins, Senior Life Sciences Industry Consultant, SAS

"The right balance is having a cohort builder that can offer diagnostic analytics such as comparative analytics and treatment pathway analyses, with an easy-to-use interface."

Ari Yacobi, Knowledgent

"The visual interface empowers nontechnical users to do their own analysis, but it serves statisticians as well. Imagine the value of being able to do rapid what-if analyses without having to go back and submit code each time you want to change a parameter."

Robert Collins, SAS

Advanced analytics for deeper insights

Advanced analytics changes the story from hindsight to insight and then foresight. "This is where the next generation of innovation is going to come from," said Yacobi. "Organizations need to spend more time and energy bringing predictive analytics and machine learning into the R&D practice."

Unlike hypothesis-driven research, machine learning uses automated model building to adapt to what's happening in population and finds things the human might not have thought to search for. With every iteration, the algorithms get smarter and deliver more accurate results. It's easy to see the value of machine learning for understanding which patients will benefit the most – or potentially be harmed by – a therapy.

"With machine learning, you can find correlations between variables that would otherwise be missed," said Yacobi. "Humans can define certain inclusion and exclusion criteria and find patterns, but machine learning can analyze large, multidimensional data sources holistically and find correlations that the human eye cannot see."

"SAS has had machine learning capabilities in its products since the 1980s," said Collins. "SAS/STAT® and SAS® Enterprise Miner™ have had it. Last year we introduced our SAS® Viya™ architecture, a distributed in-memory platform focused on advanced machine learning techniques with very large volumes of data."

Machine learning has significant promise for real-world evidence, but few are taking advantage so far. In an informal webinar poll, only 16 percent of respondents said their organizations use it, and 28 percent are considering it. That means machine learning can provide a real competitive edge.

Big data technologies let you apply advanced analytics to vast data sources. Masses of genomic data, clinical trial data, electronic health records, claims data and research study data can be brought together to reveal important discoveries and support better operational and medical decisions – and ideally, to bring analytic insights right to the point of care.

"Most real-world analysis to date has been descriptive or diagnostic in nature. There has been very limited use of predictive analytics, because so much work is spent just getting the basic analytics right. Once organizations standardize the diagnostic aspects, it creates time and opportunity to focus on more predictive analytics."

Ari Yacobi, Knowledgent

The SAS architecture for real-world evidence can significantly increase speed to insight by being neutral to the data sources, allowing selection of complex patient cohorts without coding, and providing rapid analytics results while still being extensible, flexible and transparent.

The investment is worth it

The prescription for data management: Move data from traditional high-performance data warehouses to a data lake architecture based on commodity servers (Hadoop). Adopt a common data model to standardize content, tools and analytics across the organization. And establish governance over all stages of data management.

> **The outcome**: Moving data from a massively parallel processing (MPP) appliance ($20,000 to $30,000 per terabyte) to Hadoop (about $4,000 per terabyte) saves $800,000 for every 50 terabytes of data. Data is ingested in days instead of weeks or months. Researchers save five to eight hours a week because data governance and catalog capabilities make it easier to find and access data. Productivity increases about 20 percent within 12 months. ROI accelerates as real-world evidence artifacts and their reuse grow.

The prescription for cohort identification: Move from manual code writing with an average 12-day turnaround to an interactive, visual cohort builder that supports complex queries and provides near-real-time results.

> **The outcome**: Get better visibility into the data early in the formation of a study, which leads to better study structure and protocol development. Complete two to four times as many studies in a given timeframe. Watch investments in the cohort builder plus self-service descriptive and diagnostic analytics pay for itself within six to 12 months, even for a mid-sized company.

The prescription for analytics: Move from black-box solutions and renegade snippets of code to an automated process that provides fast, repeatable and high-quality results. Take advantage of built-in SAS Analytics you can customize to suit. Give users the option to work with traditional coding or interactive visual analysis – or to use their analytic tools of choice.

> **The outcome**: Users are empowered to do self-service and ad hoc analysis with immediate results. Analytic insights are enriched with SAS and third-party tools (such as R and Python). Machine learning generates predictive insights that the human eye and other analytic methods would miss.

"These improvements translate into real value for organizations, especially when taken in light of the investments already made in the data sources themselves," said Collins. "Because it's not just about a research project to wrap up. It's about gaining insights to make decisions that benefit your company."

Real-world evidence helps identify which patients will get the most benefit and value from a therapy, based on genetic, social and lifestyle attributes that might not be captured in clinical trials. It provides a clear picture of a product's safety, effectiveness, economics and value in actual use. It provides a deeper understanding of epidemiology trends and disease management. And ultimately, it provides insights that lead to better products that advance the "triple aim" – better care, better health and better costs.

Technology can speed the delivery of real-world evidence and accelerate decisions in ways that enable – and in no way compromise – the science.

If those missions sound important, it's time to bring robust governance, standardization and consistency to the platforms and processes that generate real-world evidence.

For more information about SAS solutions for real-world evidence, please visit sas.com/rwe.

About the presenters

Robert Collins has more than 25 years' experience performing and supporting life sciences research in academic and commercial environments. A significant part of that work includes analysis of longitudinal cohorts and large-scale outcomes research using Medicare and commercial claims sources.

Ari Yacobi is the Chief Data Scientist and RWE Practice Lead at Knowledgent. A seasoned real-world data scientist, he has successfully led RWE strategy and analytics engagements at large pharmaceutical and health insurance companies.

To contact your local SAS office, please visit: sas.com/offices

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2017, SAS Institute Inc. All rights reserved. 108976_G53880.0817