BioCelerate

TOXICOLOGY DATA SHARING:

Leveraging Data Sharing To Enable Better Decision Making In Research And Development

William Houser & Raja Mangipudy
Drug Safety Evaluation
Bristol-Myers Squibb, NJ, USA

ABSTRACT

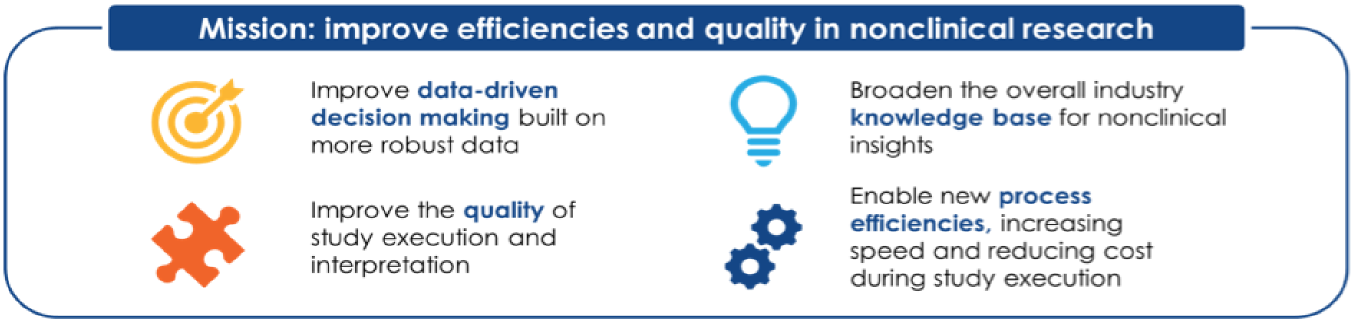
BioCelerate's Toxicology Data Sharing (TDS) initiative enables participating companies to share non-clinical toxicology and background control data using the proprietary DataCelerate technology platform. The platform is intended to enable better decision making on compound progression enabling companies to focus resources towards therapies that have the maximum benefit for patients. Since its launch in May 2018, a considerable effort has been expended on developing and enhancing the infrastructure to facilitate data upload and future collaboration. The platform has the potential to connect toxicology data with other data types (including clinical) across a compound/therapeutic target/indication. This will also enable us to derive better analytics for connectivity between preclinical and clinical data. The initiative currently focuses on first in human enabling toxicology study data and the technology platform offers the flexibility of data sharing using structured (Standard for Exchange of Nonclinical Data - SEND) and unstructured (pdf) formats. The platform also enables sharing of compound structures on a voluntary basis. As of July 1, 2019, the initiative had data that covers 8 targets across multiple modalities and includes studies across the toxicology and background control modules.

INTRODUCTION

BioCelerate was launched in 2016 out of growing support from the Member Companies of TransCelerate BioPharma, Inc., a non-profit consortium dedicated to improving the health of people around the world by streamlining and accelerating the research and development (R&D) of innovative new therapies. As a subsidiary of TransCelerate, BioCelerate focuses on the identification and development of pragmatic and tangible solutions to improve efficiencies in nonclinical research. Initiatives are approved and delivered by participating companies, comprised of a sub-set of TransCelerate membership. Today, the BioCelerate portfolio of initiatives focuses on four key aims or drivers of value in nonclinical R&D (Figure 1) and has grown to seven Member Companies.

At its founding, BioCelerate was tasked with selecting a single area of focus on which to deliver a new initiative in the nonclinical or preclinical discovery field. Participating companies prioritized the objective to engage in toxicology data sharing to enable better nonclinical decision-making during compound progression. Furthermore, the collaborative nature of the consortium supported the ambition

Figure 1: BioCelerate Mission and Aims



Mission: improve efficiencies and quality in nonclinical research

Improve **data-driven decision making** built on more robust data

Broaden the overall industry **knowledge base** for nonclinical insights

Improve the **quality** of study execution and interpretation

Enable new **process efficiencies**, increasing speed and reducing cost during study execution to share blinded information among participating companies to which individual companies would not ordinarily have access.

The Toxicology Data Sharing Initiative (TDS) is designed to accelerate in vivo toxicology through precompetitive sharing of information among participating companies. Motivated in part by the Food and Drug Administration's (FDA) 2011 Strategic Plan for Regulatory Science, which includes objectives to modernize toxicology to enhance product safety, TDS is focused on enabling access to a broader cross-company set of toxicology and background control data. The knowledge gathered through TDS has the potential to enable companies to avoid unnecessary animal use by leveraging existing animal data (following the principles of Replacement, Reduction and Refinement or 3Rs). A longer-term vision is to build a translational bridge with TransCelerate via the common data sharing tool DataCelerate which could potentially help with effective correlation of nonclinical toxicology findings to clinical safety data.

The dual aims of sharing both toxicology study data and background control data contribute to several benefits. Access to and analysis of toxicology studies submitted across multiple sources may help to increase drug development efficiency (and therefore reduce cost) by identifying fatally flawed compounds earlier in the process. Using a data driven approach, an organization may approach early decision making with greater confidence through improved understanding of on-target and off-target toxicity. Using this information, Member Companies can make informed choices on where to invest development resources towards making safer and more effective compounds. Furthermore, a larger set of background control data consolidated in one platform allows companies to better determine the significance of rare and incidental findings, improving ability to respond to regulators. The convenience of having access to control data from a variety of sources consolidated in one venue will improve speed of decision making on the relevance of such findings. Finally, an additional value proposition for TDS is aided by the implementation of the Standard for the Exchange of Nonclinical Data (SEND), which aligns with the FDA's new nonclinical submission requirements. The structured data format represents a significant opportunity to apply analytics and modelling to data across studies and sources. Usage of SEND for TDS enables direct comparability among studies contributed to the initiative while also offering a side benefit of amassing lessons learned among participating companies who are becoming more familiar with the implementation of SEND.

Since its inception, the TDS initiative followed a principle of "start small and think big" to guide the development of a data sharing platform. This approach facilitates agile and rapid delivery of an accessible, ready-to-use platform to build momentum, followed by a progression of capability advancements over time. This paper examines the experiences of the BioCelerate team in designing, building and using DataCelerate, a scalable platform used to initially share de-identified toxicology and background control data. It also examines the long-term vision for nonclinical data sharing, the evolution of use cases beyond TDS, and the readiness of industry and other stakeholder groups to contribute to this holistic value proposition.

FEATURES

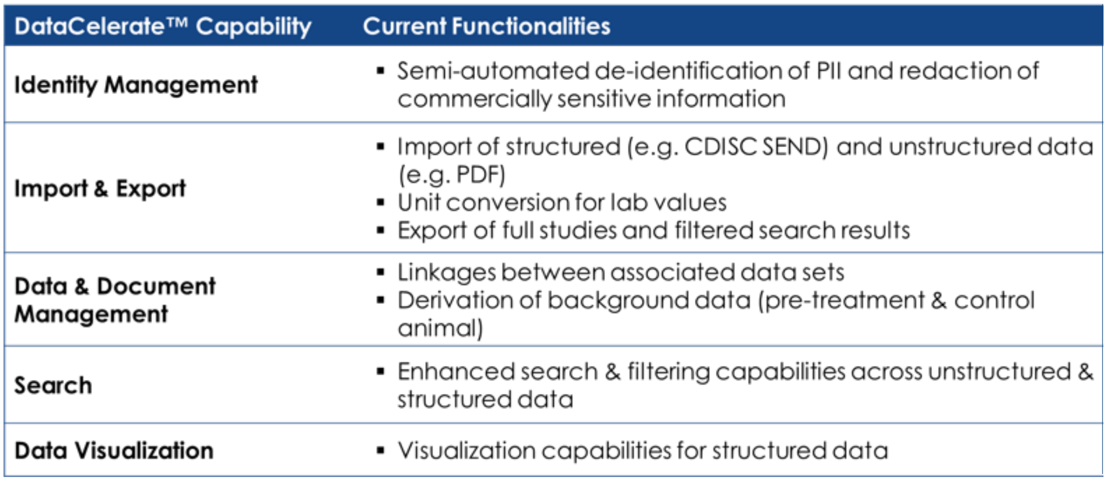
Using the “start small and think big” principle, three decisions were taken at the inception of the TDS initiative to guide early development and launch of a platform.

1. Early data contributions were to focus on First-In-Human-enabling studies due to uniformity of design, abundance of toxicity findings for assessing on- and off-target effects, and capacity to be correlated more directly with early clinical safety data (Roberts et al 2014; Blomme & Will 2016; Alderton et al, 2014).
2. Study contributions would be expected to have the same quality as those submitted to health authorities and would preferentially be in structured (SEND) data formats to enable cross-study analysis use cases on individual animal data. SEND was chosen for this purpose as a basis to facilitate efficient data collation and minimize manual curation. We feel that this is critical for future automation and extraction of data to leverage analysis across multiple data sources. To enable easier contribution of legacy studies and build database volume, participating companies also have the option to submit unstructured data (PDF).
3. Involvement in the initiative would be limited to Member Companies of TransCelerate who additionally contribute to the vision and mission of the BioCelerate subsidiary. This requirement promotes active and meaningful involvement on the part of participating companies to deliver the initiative at pace and according to a shared industry value proposition narrowly focused on pragmatic goals.

Platform development and capabilities

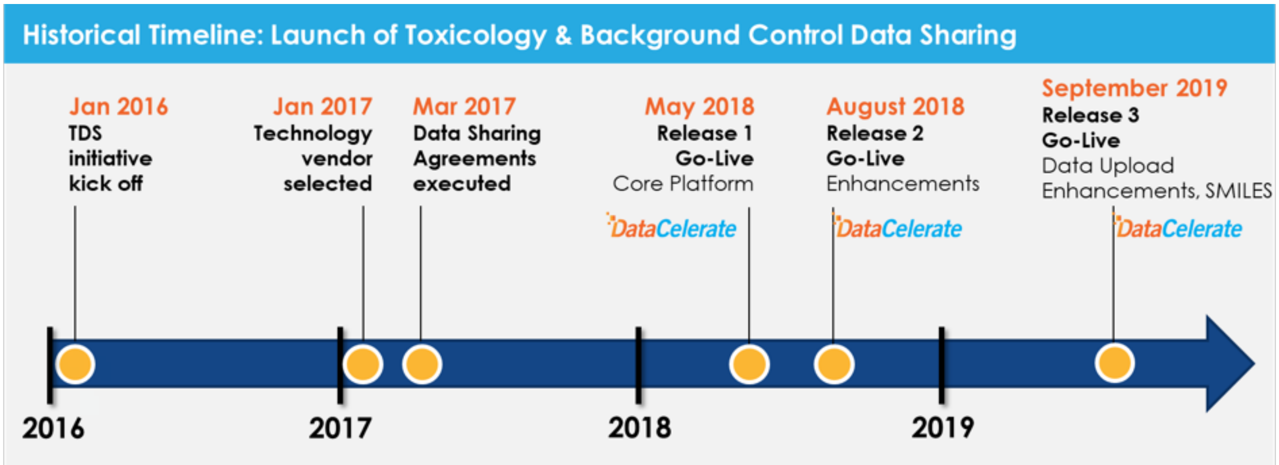
With such decisions in mind, the TDS initiative team proceeded to define use cases and business requirements for data sharing. This first step was critical to understanding both the model required to govern inter-company data sharing as well as the expectations of the technology platform that would enable data sharing. Two parallel efforts were undertaken to define Data Sharing Agreements across multiple participating companies and to execute a vendor evaluation and selection process. One of the fundamental goals of the platform selection was to implement a scalable solution that would not only satisfy the needs of the TDS initiative but also the broader data sharing needs of all TransCelerate and BioCelerate projects (Table 1). The platform was designed on the Accenture Insights Platform (AIP) and takes into consideration the CDISC SEND v3.1 Implementation Guide and the FDA Study Data Specifications v2.0 public documents. On May 7, 2018, the DataCelerate platform was released to production with the TDS module, allowing participant users to upload, search for, visualize, and download de-identified data from respective toxicology and background control databases. Enhanced features were subsequently released to production in August 2018. In September 2019, the platform was further enhanced to facilitate the study data upload process (Toxicology and Background Control) and to add the feature of SMILES (Simplified molecular-input line-entry system). The SMILES feature allows for a line notation to describe the structure of chemical species using short ASCII strings. The timeline for the launch and enhancement of the DataCelerate platform is shown in Figure 2.

Table 1: Capabilities and Functionalities of the DataCelerate Platform



DataCelerate™ Capability	Current Functionalities
Identity Management	▪ Semi-automated de-identification of PII and redaction of commercially sensitive information
Import & Export	▪ Import of structured (e.g. CDISC SEND) and unstructured data (e.g. PDF) ▪ Unit conversion for lab values ▪ Export of full studies and filtered search results
Data & Document Management	▪ Linkages between associated data sets ▪ Derivation of background data (pre-treatment & control animal)
Search	▪ Enhanced search & filtering capabilities across unstructured & structured data
Data Visualization	▪ Visualization capabilities for structured data

Figure 2: Development Milestones for TDS


Historical Timeline: Launch of Toxicology & Background Control Data Sharing
Jan 2016
TDS initiative kick off
Jan 2017
Technology vendor selected
Mar 2017
Data Sharing Agreements executed
May 2018
Release 1 Go-Live Core Platform
DataCelerate
August 2018
Release 2 Go-Live Enhancements
DataCelerate
September 2019
Release 3 Go-Live Data Upload Enhancements, SMILES
DataCelerate
2016
2017
2018
2019

Partially Automated De-identification and Publishing Process Flow

To maintain the confidentiality of sensitive data, these functionalities are divided across three separate locations: the Local Machine, the Member Company Private Area, and the TDS Shared Area (Figure 3). For unstructured data (e.g. PDF files), users supply text-searchable PDFs and complete manual redaction of text and images outside the system. Structured data (e.g. SEND-formatted XPT files) is automatically de-identified within the company Private Area using a tool developed by Accenture. The user also confirms compliance with CDISC SEND format by utilizing a validation tool (such as Pinnacle 21) prior to upload. In addition, automatic lab unit conversion occurs for structured data in the company Private Area.

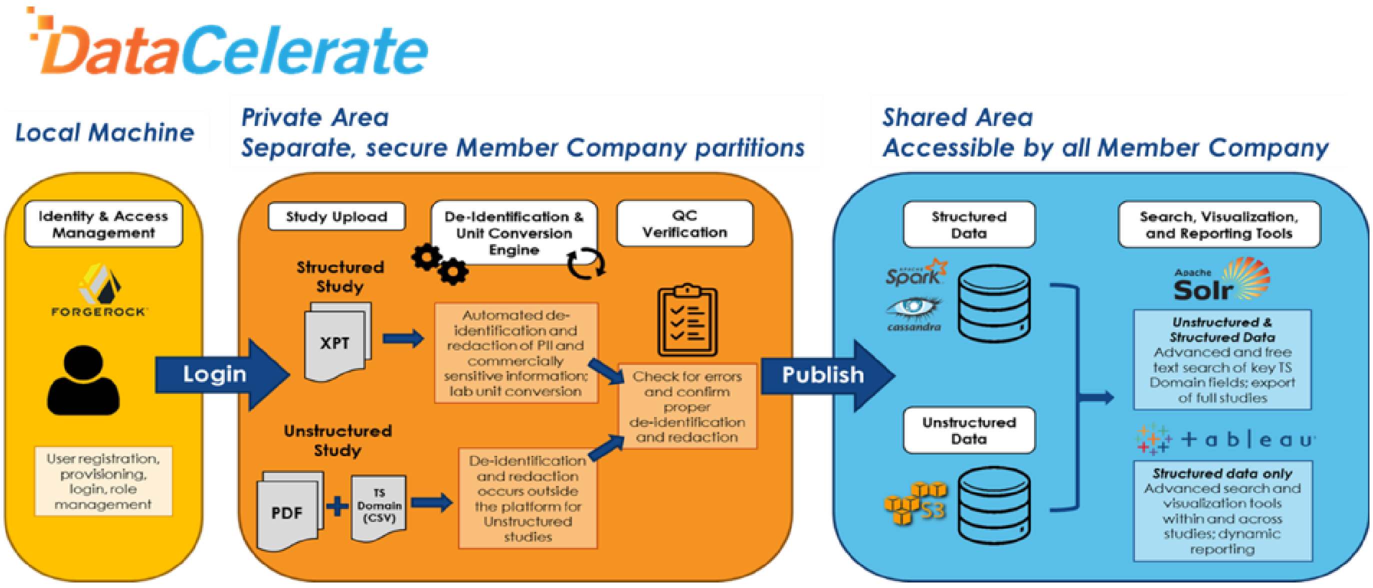
Following completion of these steps, the de-identified study remains visible within the company Private Area only to users within the contributing company; at this point, the de-identified study is not yet searchable by another company's users. The final step requires a user to perform a quality control check on the de-identified study prior to publishing to the TDS Shared Area where it will be searchable by users from all participating companies. From within the TDS Shared Area, users may search for, view, visualize, and export published unstructured and structured data.

Operating Environment and System Security

The DataCelerate platform utilizes a multi-tiered publication approach to make data available for use by each participating company (Figure 3). Apache Spark is used to automatically de-identify the data contributed to the system based upon rules specified by the core team and designed to comply with the legal requirements for confidentiality in data sharing in the precompetitive space. As mentioned earlier, after automatic de-identification is complete, the user has an opportunity to further manually redact the data, if necessary. After the data have been reviewed a final time, the data are moved to an Apache Cassandra instance and indexed by Apache SOLR. Springboot is used to parse and display all functions and data stored in the TDS Shared Area. The Tableau visualization server provides the ability to search, visualize, and download datasets. The datasets displayed in the Tableau visualization server are sent via data extract each day from the Cassandra server.

All user interaction with the DataCelerate platform is secured via username and password credentials. All communication between the web browser and TDS User Interfaces is secured with Secure HTTP (https). The DataCelerate platform is not open to the Internet at-large. Each participating company may gain access to the DataCelerate platform by using their web browser to navigate to the secured instance on the Amazon Web Services (AWS) Cloud. The confidentiality and security model described above has been thoroughly assessed and accepted by each of the Member Companies.

Figure 3: Platform Architecture & Process Flow


DataCelerate
Local Machine
Private Area
Separate, secure Member Company partitions
Shared Area
Accessible by all Member Company
Identity & Access Management
FORGEROCK
User registration, provisioning, login, role management
Login
Study Upload
Structured Study
XPT
De-Identification & Unit Conversion Engine
Automated de-identification and redaction of PII and commercially sensitive information; lab unit conversion
Unstructured Study
PDF
TS Domain (CSV)
De-identification and redaction occurs outside the platform for Unstructured studies
QC Verification
Check for errors and confirm proper de-identification and redaction
Publish
Structured Data
Spark
cassandra
Unstructured Data
S3
Search, Visualization, and Reporting Tools
Apache Solr
Unstructured & Structured Data
Advanced and free text search of key TS Domain fields; export of full studies
tableau
Structured data only
Advanced search and visualization tools within and across studies; dynamic reporting

Search and Visualization of Data

One of the first challenges the team anticipated was analyzing and comparing datasets with different units for clinical pathology measurements. For this, the team decided to adopt the International System of Units (Le Système International d'Unités or SI) for laboratory values as published in the AMA Manual of Style (Fontanarosa PB and Christiansen S, 2009) and build the conversion functionality directly into the tool for ease of use. The platform automatically applies unit conversions for the measurements we frequently collect and stores the result in a new variable to ensure the supplied values are not obscured.

Secondly, the TDS initiative team tackled the challenge of creating a search functionality to allow users to determine which data is relevant to their current needs. We enabled it to perform cross-study searches to see and download the complete packages or as a set of values for particular measurements. Search results for unstructured data (redacted pdf files) is more limited that for SEND contributions; however, even for these contributions the tool enables searches for text strings and field-specific searches on specific data fields supplied in the Trial Summary domain.

The visualization functionality enables a user to take study data identified on the TDS Shared Area and conduct additional analysis within the TDS Module using Tableau software. If desired, users may also download the data and conduct analysis using their own in-house tools. The Tableau implementation enables many analyses to be conducted quickly and conveniently while the ability to download the studies permit more extensive analysis to be performed without increasing the cost of the platform for elements that would be costly to build and rarely used.

Introduction of additional SEND parameter codes

To achieve the goals of sharing many studies with interesting data we realized we would need to redact commercially sensitive information like compound names. This would have eliminated the ability to determine toxicity related to pharmacological targets unless additional information were supplied. As a result, we introduced several new SEND trial summary parameter codes:

- Planned Pharm Target Common Name
- Planned Pharm Target Entrez Gene ID
- Planned Pharm Target Entrez Gene Symbol
- Planned Pharm Target Mode of Action

We also identified that the planned dose frequency was not easily determined without analyzing the dosing records in the EX domain. Since this is an important element of a study design, we proposed that a parameter for this also be added to the Trial Summary domain:

- Planned Dose Frequency

CDISC also agreed that all of these were important enough concepts to be added to the standard and they were included in the 2018-03-30 release and subsequent releases of the controlled terminology. This was even adopted by the FDA as a parameter they request all submitters to supply.

OUTCOMES & FUTURE STATE

Figure 6 shows the contributions to the Toxicology Data Sharing Initiative and Background Control Initiative as of July 2019.

Each of the 8 targets is assigned a unique color:

▲ ▲ ▲ ▲ ▲ ▲ ▲ ▲

Contributions without a target specified are empty triangles. △

The first compound for a target is shown as a triangle. ▲

The second compound for the same target is shown as a square (e.g. ■)

Each contributed study is represented as either a triangle or a square.

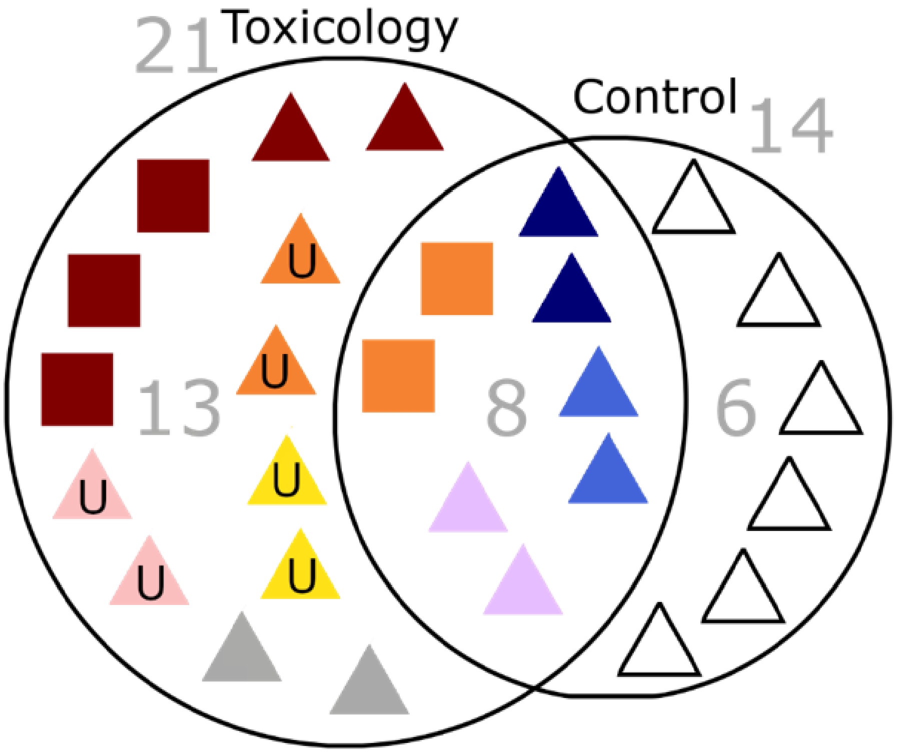
All studies are in SEND format except those with a "U" in the diagram to indicate unstructured contributions

All 21 studies contributed to the toxicology initiative are first-in-human enabling studies for these compounds.

All 14 studies contributed to the control initiative are in SEND format.

The 8 studies, represented by the intersection of the circles, have been contributed to both the Toxicology and Control initiatives.

Figure 6: Data Contributions to the Toxicology and Background Control Initiatives as of July 2019



CONCLUSION

This initiative has highlighted both key opportunities and challenges.

One major highlight was the recognition of the importance of data sharing amongst the Member Companies and the cooperation in enabling the effort. This was supportive of the premise upon which this initiative was built. This augurs well for the longevity and value of this platform. Unlike other data sharing initiatives with a similar purview of improving predictivity, the TDS platform is aligned with the FDA's SEND format. This avoids duplicity of platforms, data repositories, and data structures. The TDS platform provides an ecosystem that is focused and consistent with current electronic data submission standards. However, this journey has also highlighted some challenges. It is important to acknowledge that the volume of data collected since launch is low. This is perhaps an outcome of the initial scope that was envisioned (FIH enabling toxicology data), initial data upload challenges, and the small cohort of Member Companies.

On a positive note, each one of these issues is being addressed. We anticipate that as SEND continues to be adopted more broadly across the globe, the ease of submitting studies in this format will increase amongst the Member Companies ultimately allowing for cross-study analysis to be streamlined, without the need to back-convert datasets. We anticipate expanding the scope of the contributions from earlier in drug development to later stage toxicity studies. Platform enhancements have been to not only facilitate data upload but also enhanced features such as SMILES. Finally, the consortium has new members. These factors collectively are expected to enhance the data volume in both the toxicology data sharing and background control initiatives. In addition, with the impending implementation of Placebo and Standard of Care (PSoC) datasets using the DataCelerate platform, we anticipate having the benefit of a single modular sign-on repository for all preclinical and clinical data sharing initiatives across BioCelerate and TransCelerate. Over a period of time, this is anticipated to provide a powerful means to connect datasets across the preclinical and clinical space for compounds in the system and enable translational insights. Finally, the platform realizes significant systems efficiency by leveraging a software model that has the potential to capture economies of scale, cost savings, and enhanced functionality.

The platform is already in use and we anticipate further growth based on the steps that have been outlined above. These steps are strategically intended to enhance volume and functionality. With these enhancements, the platform is expected to provide an integrated venue for bringing together toxicology and background control data from a variety of study types and durations across multiple modalities and indications in both unstructured and SEND formats.

REFERENCES

[1] Roberts RA, Kavanagh SL, Mellor HR, Pollard CE, Robinson S, Platz SJ, Reducing Attrition in Drug Development: Smart Loading Preclinical Safety Assessment. Drug Discov Today. 2014 Mar;19(3):341-7

[2] Blomme EA, Will Y, Toxicology Strategies for Drug Discovery: Present and Future., Chem Res Toxicol. 2016 Apr 18;29(4):473-504

[3] Alderton W, Holder J, Lock R, Pryde D, Reducing Attrition Through Early Assessment of Drug Safety—Highlights From The Society of Medicines Research Symposium held on March 13th, 2014, National Heart and Lung Institute, Kensington, U.K., Published in Drugs of the Future, 2014 39(5): 373-377.

[4] Fontanarosa PB and Christiansen S, Laboratory Values Section in AMA Manual of Style: A Guide for Authors and Editors (10th edition), AMA Manual of Style Committee, © American Medical Association, Oxford University, 2009