

Harnessing Data Curation To Unlock a **\$330 Million** **Opportunity**

A US-based global biopharma company is committed to addressing complex health issues in developing countries. They collaborated with Elucidata to repurpose publicly available data to validate the effectiveness of their probiotic formulation for treating malnourished children.

Problem Statement

The organization's research team aimed to leverage pre-existing studies within the public domain to assess and validate the effectiveness of their probiotic treatments. However, the team experienced significant delays in locating pertinent studies. Effectively deriving value from these studies was time-consuming. The company reached a juncture to either find methods to expedite the process or abandon the project entirely. At this point, they approached Elucidata for support in identifying, comparing, and analyzing relevant publicly available data to accelerate their validation pipeline.

Challenges

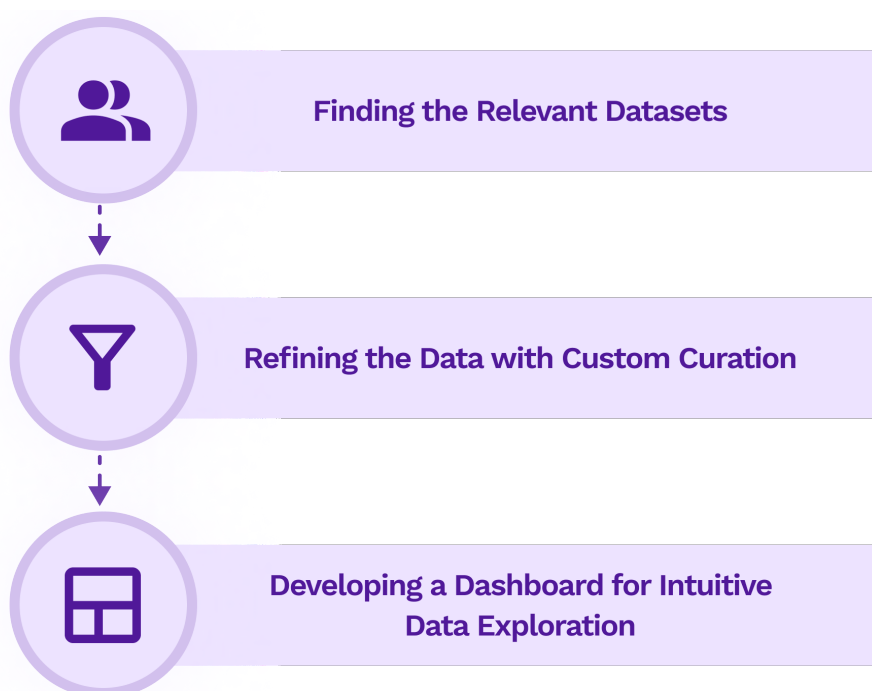
The biopharma company encountered several roadblocks which impeded research progress and led it to seek assistance. The major challenges were as follows:

1. **Relevant Control Datasets Were Hard To Find:** A substantial proportion of the children within the study locations suffered from malnutrition. This presented a notable challenge in sourcing studies with data from healthy control subjects. Without this data, they could not determine the efficacy of their treatment.
2. **Available Data Was Not Readily Usable:** Even when relevant datasets were found, they were present in a semi-structured form with incomplete metadata and varied data formats depending on the repository they were housed in. The researchers spent significant time manually cleaning and standardizing the data before they could use it.

3. **Data Visualization and Analysis Were Difficult:** Out-of-the-box data visualization tools couldn't be used for data visualization and analysis due to the project's unique nature. Their specific research questions required customization. Moreover, sharing the data with the team and obtaining their feedback posed additional challenges in this collaborative project.

Elucidata's Solution

With [Elucidata's](#) support, the company successfully overcame challenges in utilizing publicly available data, enabling it to continue the project. Our solution used a three-tiered approach (shown below) to find and reuse public data effectively. Let's delve into the specifics.



Elucidata's three-tiered solution

Tier 1: Finding the Relevant Datasets

The first step to accurately identifying the relevant datasets was to define the criteria needed to shortlist datasets for the entire study. Curating and harmonizing datasets take time; therefore, defining all the specifications (experiments, metadata, etc.) needed in a dataset in the initial stages was crucial. In collaboration with the research team, we formulated a stringent set of criteria (shown below) for their target datasets to provide a clear focus for the search process.

No:	Criteria
1	The data must be from human gut microbiome samples from children aged 0 to 2 years.

2	The samples must be either 'treatment' (probiotic), 'disease' (malnutrition), or 'control' (healthy) .
3	The 'control' samples must be from similar geographical locations as those of the 'disease' and 'treatment' populations.
4	Metadata such as 'breastfeeding status,' 'current or past disease,' and 'antibiotic history' should be available.
5	The data must be linked to the original publication or source .

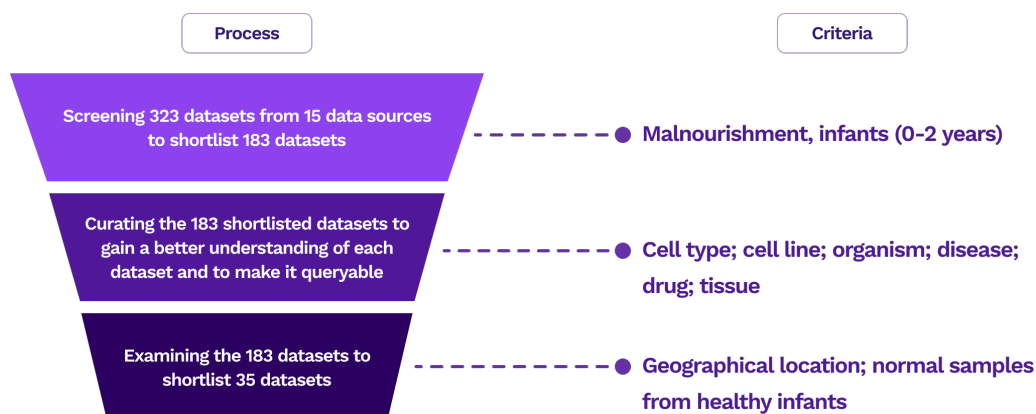
The process to identify the relevant datasets that fit their criteria was carried out in 3 consecutive steps:

Step 1: The relevant microbiome and metatranscriptomics datasets were identified using two criteria: 'malnourishment' and 'infants (0–2 years)'. We screened more than 323 datasets from 15 public resources, ranging from the [Microbiome DB](#) and literature surveys to the [SILVA Ribosomal RNA Gene Database](#). Based on the abovementioned keywords' criteria and their relevance to the problem statement (malnourishment in developing nations), we selected 183 datasets for further processing.

Step 2: The shortlisted datasets were readily curated for six standard fields using Elucidata's proprietary data curation pipelines. The curation process for the six standard fields- defined by Elucidata as cell type, cell line, organism, disease, drug, and tissue- is swift due to the available predefined pipelines and processes. This initial curation round offers an overall understanding of each dataset. This curated data is ingested and stored in an organized, queryable manner within the [Polly](#) platform.

Polly creates a unique, centralized ecosystem that enables a diverse team of biologists, bioinformaticians, and scientific leaders to generate and share insights. It has enabled the detection of multiple validated drug targets across immunology, oncology, and metabolic disorders through ML-ready data and a scalable data infrastructure that enables easy downstream analysis.

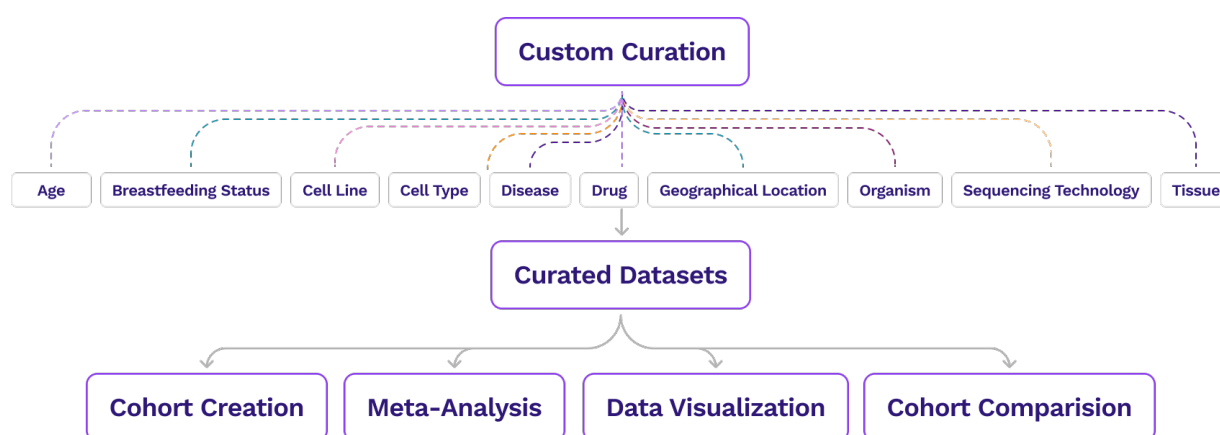
Step 3: Out of the initial pool of 183 curated datasets, we refined our selection to 35 datasets by applying more specific filters. These filters considered factors like 'geographical location' and the presence of 'normal samples from healthy infants' in the dataset.



Process flow to shortlist 35 datasets that cover all inclusion/exclusion criteria

Tier 2: Refining Cohorts With Custom-Curated Data

Tier 1 of the solution presented a collection of 35 curated datasets highly relevant to their area of interest. To further streamline the downstream data analysis process and generate cohorts for effective comparative studies, the research team wanted to curate the 35 shortlisted datasets for 4 more fields critical to their research – **age**, **geographical location**, **breastfeeding status**, and **sequencing technology**. This curation process was done using Elucidata's ML-powered '[Curation app](#),' built to curate data at lightning speed.



Custom curating fields for various downstream applications

"I definitely think this is a good pipeline for the bioinformatics or computation biologist team who can actually rely on you guys not just to collect the data but also to curate the data. I think that was the most important part. It greatly reduced our burden when working on data collection."

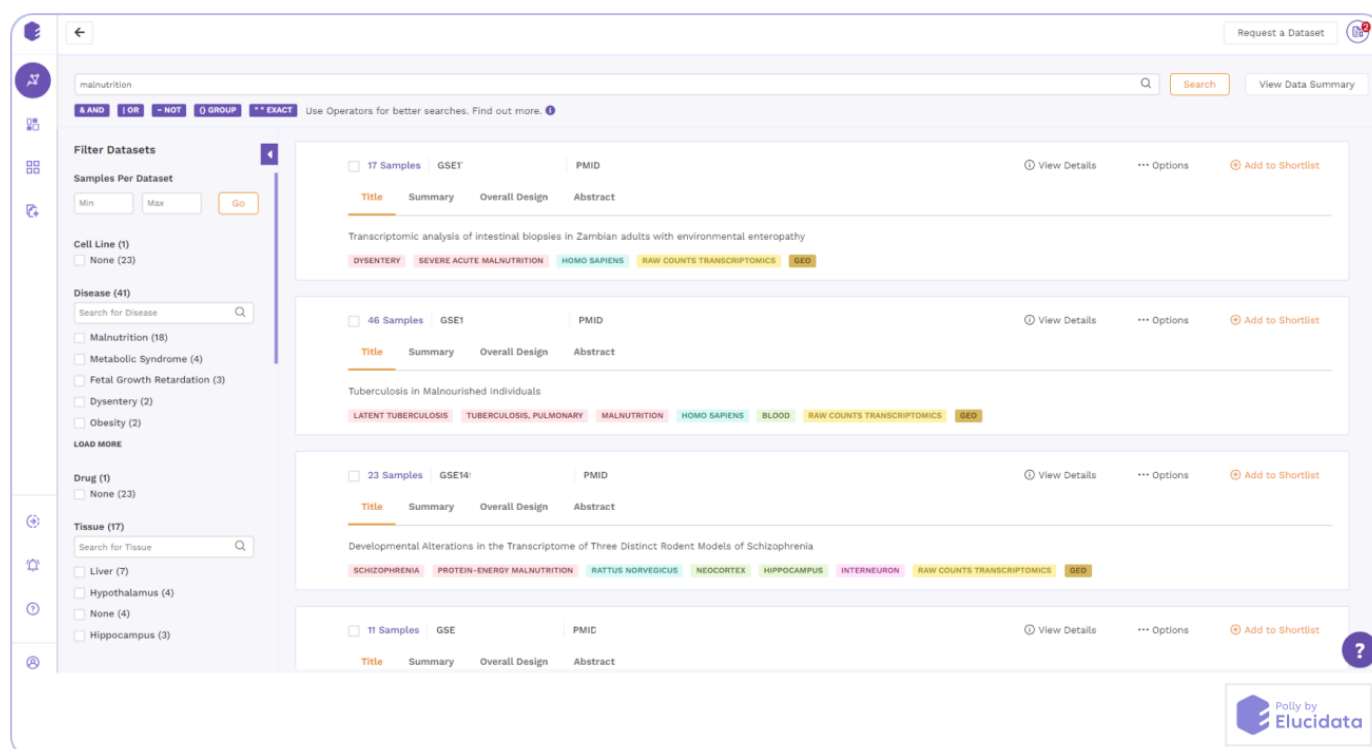
~ Senior Scientist, The Biopharma Company.

The **Curation app** is a data annotation tool for labeling multiple data types for metadata fields. It has an integrated active learning system that uses machine learning to provide annotation suggestions to increase the speed of curation. This app is driven by Polly's AI-powered curation engine, which has been shown to have human-level accuracy in several named entity recognition (NER) tasks. Read more [here](#).

This rigorous process yielded 35 custom-curated datasets containing over 25,000 cells with complete and accurate metadata. These 35 custom-curated, ready-to-use datasets were then deposited into an OmixAtlas created for the customer. Once the data was on their OmixAtlas, they could explore it in depth.

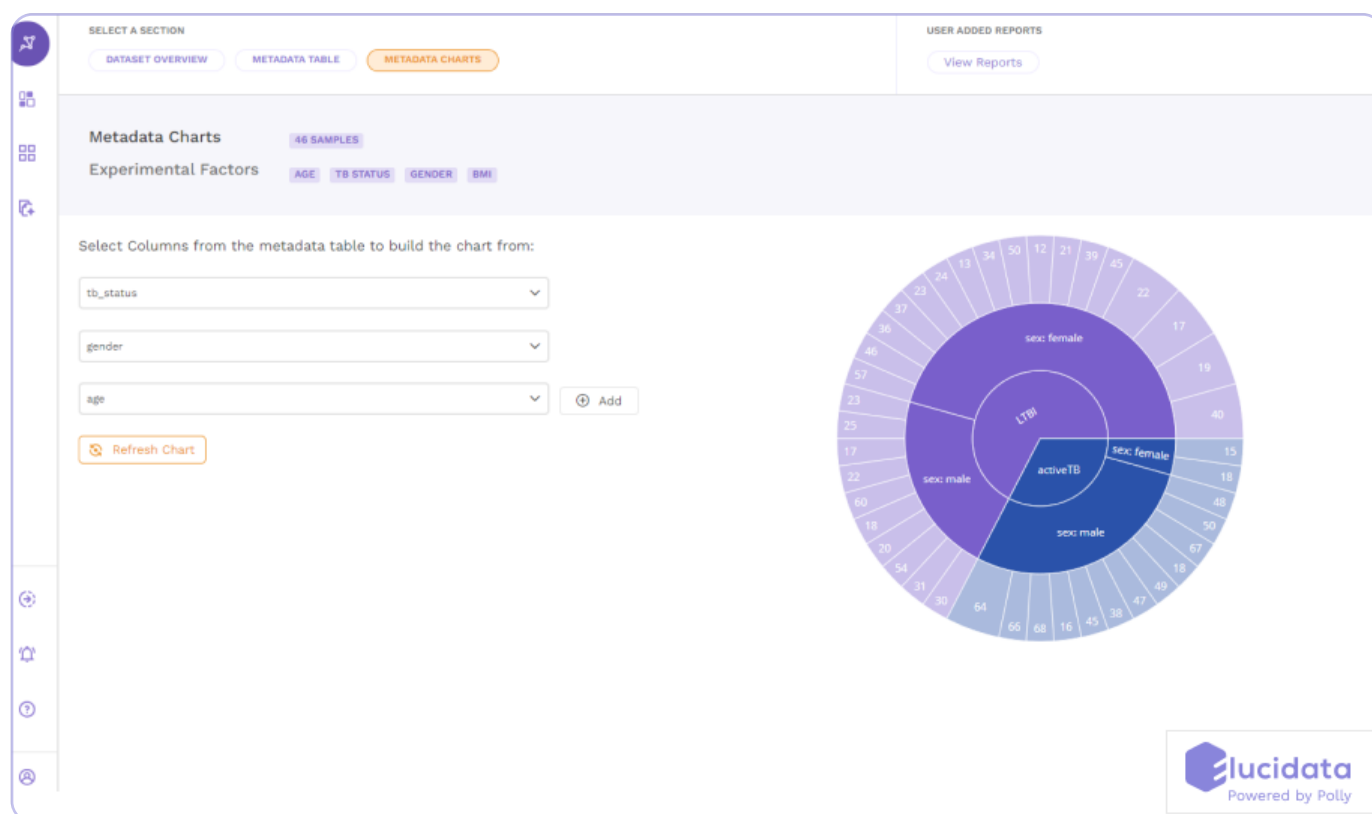
OmixAtlas is a data warehouse that can store biomedical data in a structured format ready for downstream machine learning and analytical applications.

The OmixAtlas user interface allows deep exploration of the metadata in the datasets using the elastic search bar, customizable filters, and easy-to-view tags.



Organized data on OmixAtlas

Further, easy visualization of relationships between different cohorts is also facilitated with interactive sunburst plots. At a glance, the researchers could discern patterns, overlaps, and distinctions among cohorts, which made it easier to draw insights and understand the broader landscape of the data they were dealing with.



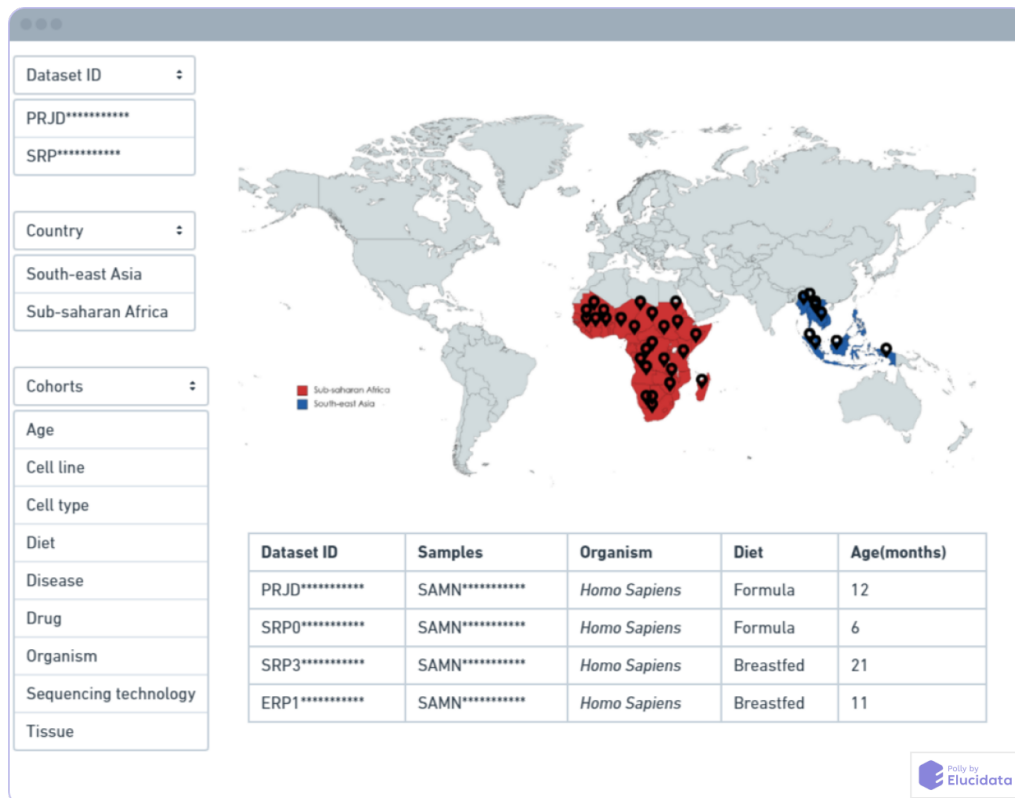
The study shows the effect of malnourishment on an associated disease.

In essence, OmixAtlas didn't just provide a reservoir of highly curated data; it presented the data in an accessible, shareable, intuitive, and visually engaging manner. All the data on OmixAtlas could be easily downloaded and fed into additional proprietary pipelines and applications for further analysis using ML models.

Tier 3: Developing a Dashboard for Intuitive Data Visualization

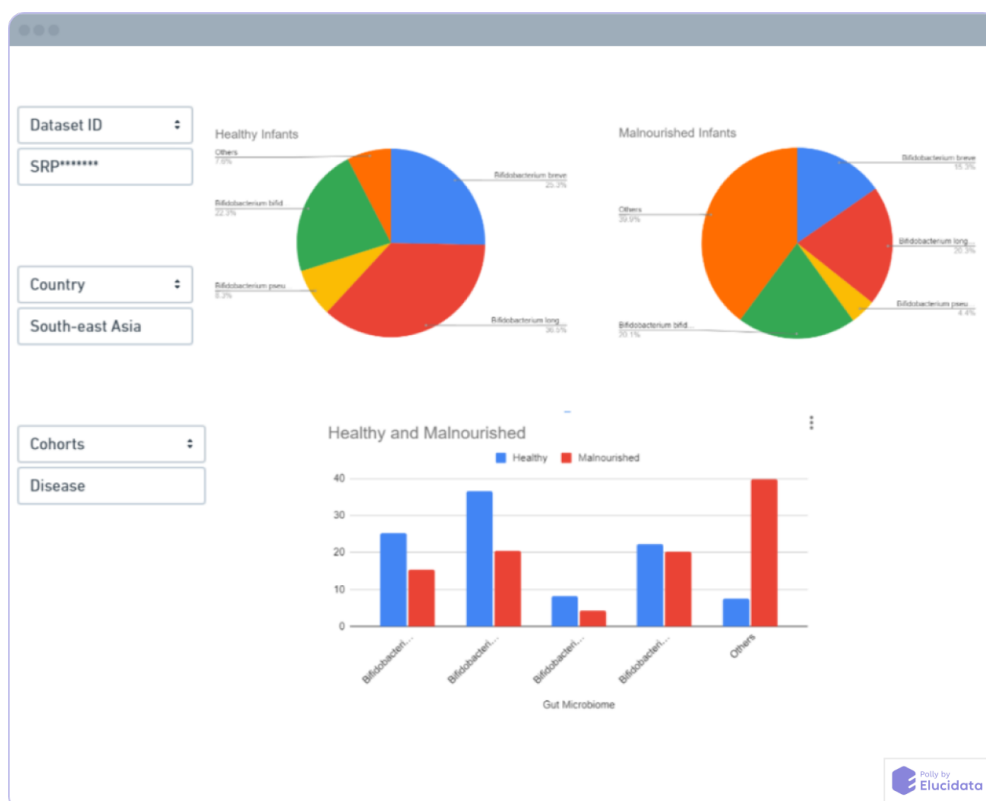
Omixatlas efficiently tackled the challenge of exploring and visualizing the meticulously curated datasets. However, they also needed a comprehensive overview of the available data to serve as a central source of information for research decision-making. To address this need, we developed a user-friendly data dashboard that streamlined data interaction and comprehension.

The dashboard highlighted the geographical locations from which the selected studies were conducted.



The dashboard homepage displayed the geographical study locations of each shortlisted dataset.

It also featured multiple user-friendly visual representations, like pie charts and bar graphs, to drive insight generation.



The dashboard features various graphs for intuitive visualization.

The dashboard offered clear visibility at a glance, which resulted in time and resource savings while facilitating informed decision-making.

"I am excited to get access to all this data, and I find the platform and the API easy to use. So, it's been good!"

~ Senior Scientist, The Biopharma Company.

Outcome

When the company approached Elucidata, this project was at a point where the company had to significantly expedite its data exploration process or abandon the project. With the help of our unique approach, their research team could navigate publicly accessible data to locate datasets pertinent to their research, reviving the possibility of moving forward with the project.

- 1. Identified Datasets Closely Aligned With Their Research Objectives:** We identified **35** datasets with 23,000+ samples from 15 public databases, **matching all their inclusion/exclusion criteria.**
- 2. Made 35 Relevant Datasets Readily Available for Downstream Applications:** The shortlisted datasets, curated with 10 pertinent fields, allowed the team to navigate the data landscape efficiently, create cohorts, and easily perform **relevant downstream and machine learning applications.**
- 3. Facilitated Intuitive Data Exploration and Collaborative Research:** The team was provided with an intuitive, user-friendly dashboard for ease of collaboration and data analysis. As a result, research team members could effortlessly **comprehend their findings and communicate them to their peers and senior management.**
- 4. Significant Acceleration in the Data Exploration Process:** With the assistance of our three-tiered solution, the research team significantly accelerated their data discovery process, **enabling them to advance the project.**

The estimation of opportunity costs was calculated based on:

1. An attainable market size of approximately 11 million (Sub-saharan Africa and South-east Asia), assuming a 10% market penetration rate.
2. The consumption of probiotic drugs, priced at an average of \$5 per strip, taken twice daily.

Impact

Opened the doors for a million-dollar opportunity: We supported the company in reinitiating its stagnant project, thereby opening up the possibility of a sizable opportunity of **\$330 million per month**, impacting more than **13 million patients**.

>2X Acceleration: The whole project was completed in 5 months. It would have generally taken at least a year for the company with a dedicated bioinformatician to do the same.

~70% Cost Reduction: The process of finding, curating, and analyzing the relevant data was accomplished at a significantly lower cost, thanks to the efficient Polly platform and pipelines.

About Elucidata

Elucidata transforms biological discovery by supporting programs at top pharma companies and having 35+ research partners from premier biopharma companies and research labs. Their FAIR biomedical data platform, Polly, makes data easily findable and more reusable, thus enabling scientists to focus on insight derivation via data analysis and visualization instead of data wrangling and engineering. Elucidata has enabled 10x faster identification of therapeutic assets with high odds of success in the clinic. Experience the versatility of our platform, infrastructure, and team firsthand!

[Join us](#) on the journey to customized solutions and accelerated discovery programs.

[Book a Demo](#)

For more information, visit Elucidata | Home or reach out to us at info@elucidata.io.



San Francisco | 114 Sansome St, 250, San Francisco, CA 94104, USA

Cambridge | 625 Massachusetts Ave 2nd Floor, Cambridge, MA 02139, USA

Delhi | Avanta Business Center, Saket, New Delhi, Delhi 110017, India

Bangalore | UrbanVault HSR Layout 761, Bengaluru, Karnataka 560102, India