Paper AD06

An Integrated J&J Application module to Enable Sustainable Clinical Trial Design

Archana Nayak, Johnson & Johnson R&D, Great Britain
LaRoche, Jason, Leiden, The Netherlands
Bhattacharjee, Eshanika, ZS Associates India
Md, Umarfaruque, ZS Associates India
Bhargava Reddy, Janssen R & D, PA, USA

ABSTRACT

Pharmaceutical companies have partnered with the Sustainable Healthcare Coalition (SHC) to develop standardized methods for measuring greenhouse gas emissions in clinical trials. However, the faster adoption of these methods into everyday trial design remains a challenge. To address this, we have developed our internal J&J application module, which includes a front-end application leveraging industry-standard methodologies to embed sustainability into trial design. The application module enables secure, large-scale data pooling from multiple systems, consolidating inputs such as country site distribution, drug product data, and other relevant information, serving as proxies for baseline protocol data elements. This unified approach allows users to design baseline emissions in a single platform, accessible for ongoing refinement, scenario testing, and reuse. Anonymized hash encoding ensures data privacy, while surfacing emission hotspots, with over 80% of trial-related emissions traced to specific components. With its user-centric interface and embedded optimization tools, the application module empowers teams to iterate on sustainable design choices and accelerate the clinical trial community's shift toward net-zero, climate-resilient operations.

INTRODUCTION

Climate change is a global crisis that affects all sectors, including the pharmaceutical industry. Clinical trials, which are necessary for the development of new therapies, often have a substantial environmental footprint. Emissions from travel, logistics, energy use, and product manufacturing contribute significantly to greenhouse gas output.
In response, industry initiatives such as those by the Sustainable Healthcare Coalition (SHC) and Johnson & Johnson (J&J) have developed emission measurement frameworks. However, integrating these frameworks into active protocol planning remains a challenge. To bridge this gap, we designed an internal J&J application module that embeds sustainability metrics into trial design workflows through a secure, intelligent application.

Background and Industry Context

Traditional trial planning does not systematically incorporate emission data, making it difficult to assess and reduce environmental impacts during the design stage. Logistics such as patient travel, monitoring visits, drug shipping, and cold-chain requirements are typically considered only from cost and operational perspectives.
Recognizing this limitation, J&J collaborated with the SHC to define sustainable practices and emission modeling strategies. The result was a multi-functional application module designed to centralize sustainability considerations during protocol planning.

Application Design and Architecture

The J&J Sustainable Trial Design Application module consists of four key components:
- Data Aggregation Layer: Integrates data across planning systems (e.g., sites, logistics, drug product configurations).
- Emissions Calculation Engine: sends only numerical data via encrypted API to the SHC-compliant models to estimate carbon footprints with only country and logistic information are sent after encoding strings

- Score is received and is decoded within our J&J firewall and mapped to relevant study specific elements and displayed.
- Visualization and Scenario Testing UI: Empowers users to model changes and simulate emission outcomes in real time.

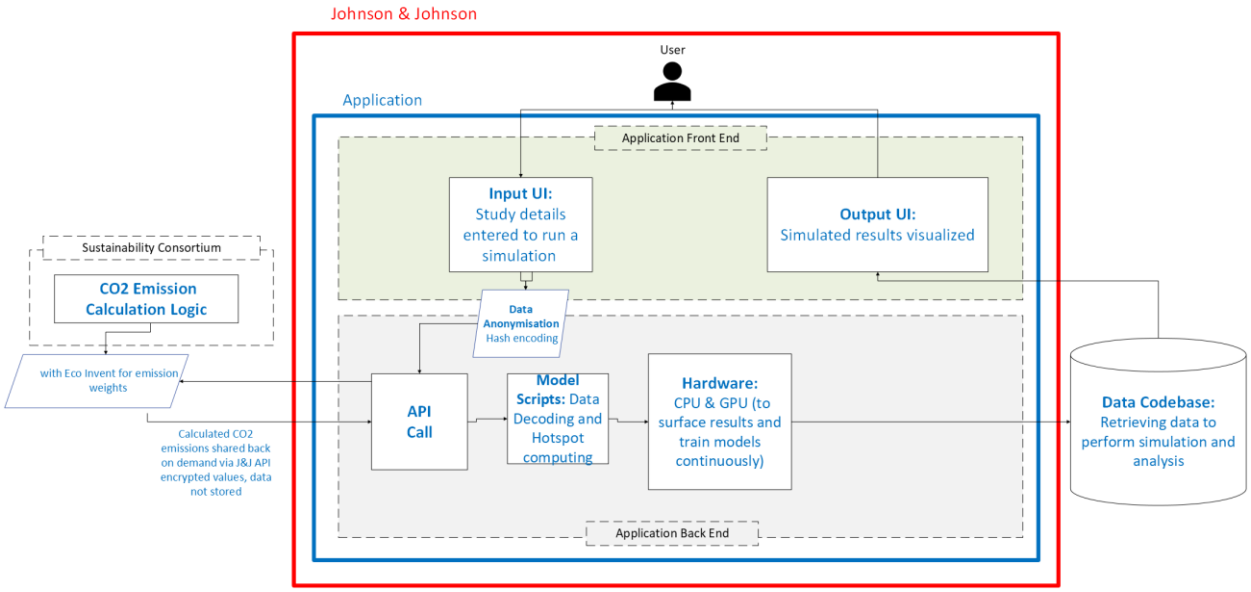

Figure: Flow diagram for automation of data selection, pre-processing and transformation steps highlighted.

Authentication and Security: Login Page Authentication

In our application, access to the front-end is restricted to authorized users through a robust login authentication mechanism. Before API calls are made in the back end, all sensitive information undergoes encoding and decoding processes to maintain data confidentiality. To ensure security, all endpoints handling sensitive data incorporate comprehensive measures such as authentication, authorization, and anonymization via hash encoding. Specifically, input data and tokens are protected using a hashing algorithm that converts data into a fixed-length string, safeguarding sensitive information. API tokens are generated and securely stored, and all API communications utilize HTTPS to encrypt data in transit, ensuring secure data exchange for both open-source and J&J-specific APIs. Additionally, rate limiting is implemented to restrict the number of requests each token can make per second or minute, preventing misuse and ensuring fair usage of API resources.

Application front-end

On the home tab we have a table of all the studies, their associated details such as phase, TA and the carbon emission score of each.

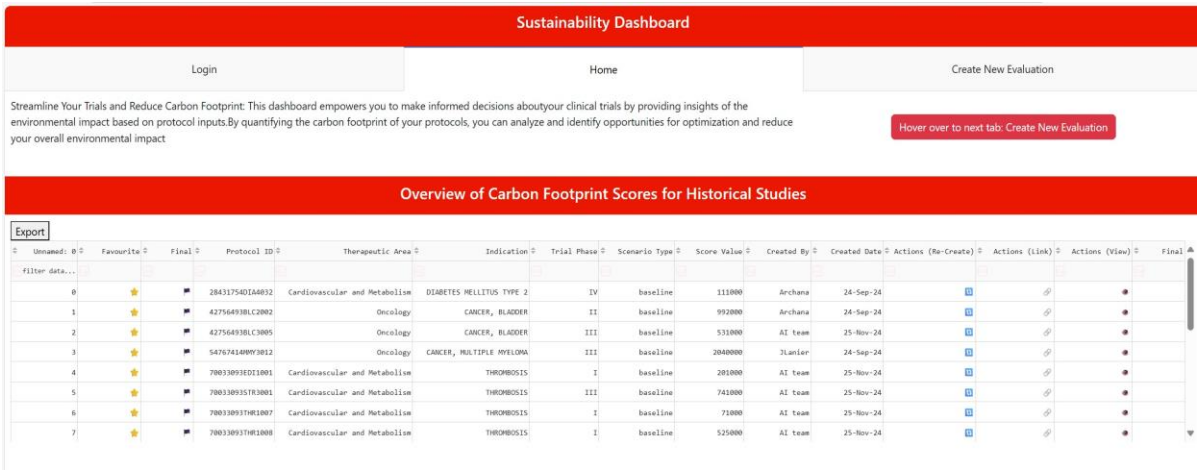

Figure: Landing page of sustainability impact tool

In the Create New Evaluation Tab, we first select a study and then generate the score. Then, we can create a new scenario, save it and load the same scenario.

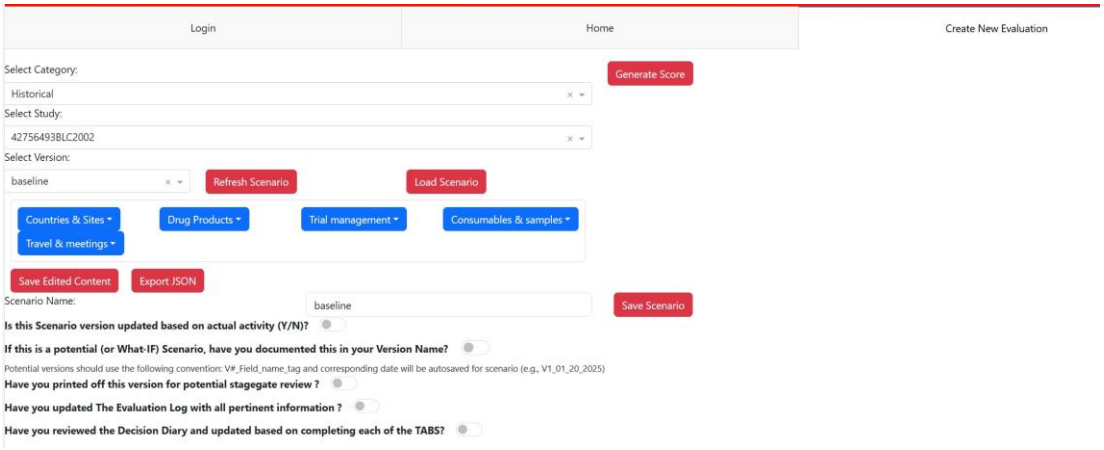

Figure: New Scenario creation page

After generating the score, we get the following graphical representations:

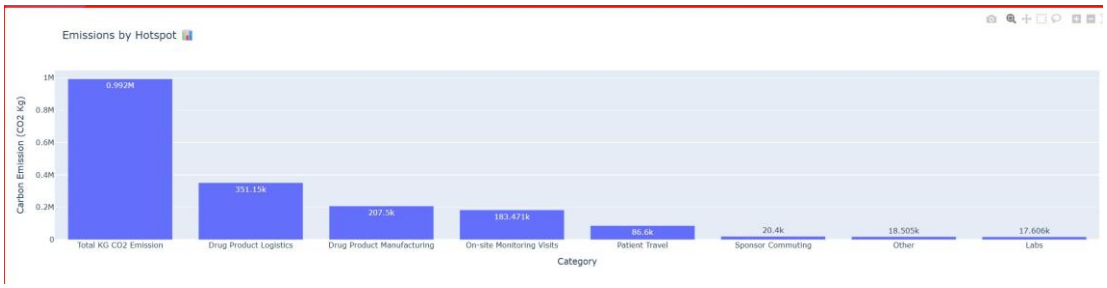

Figure: Plot of carbon emission by hotspot

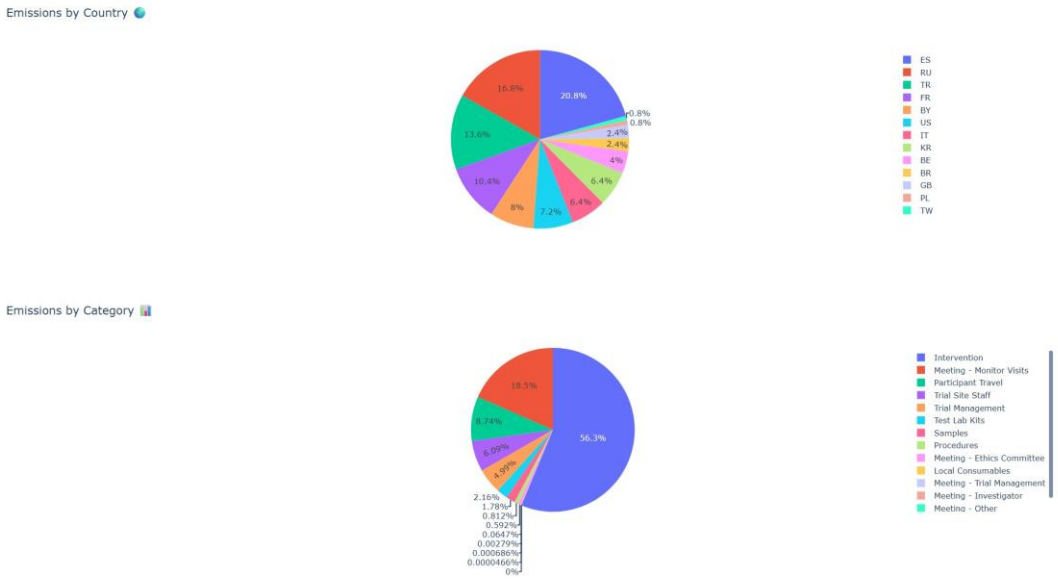

Figure: Plot of carbon emission by country and targeted area

AI-Powered Design Pattern Recognition and Pre-Fill Engine

A key feature of the application module is its XG-Boost based pre-fill engine. This AI model is trained on historical clinical trial data, including design parameters and emissions outcomes. When initiating a new trial, users are presented with intelligent defaults derived from prior trials with similar characteristics.

This reduces manual data entry, accelerates the workflow, and encourages the reuse of low-emission design patterns. Users can modify or test alternative scenarios using dedicated tools, but the AI provides a sustainability-informed starting point. Over time, the model adapts and improves as more data becomes available, reinforcing successful emission-reduction strategies.

Methodology

We employed an advanced machine learning approach to predict the number of attendees for clinical trials using the XG-Boost regression algorithm, optimized through the Grid Search CV hyperparameter tuning framework. We experimented with two modeling techniques.

In the first model, we adopted a sequential approach where the predictions from the initial model were used to train the second model. This model leveraged a diverse set of features, including categorical variables such as 'country_name', 'therapeutic_area', and 'top_phase', alongside numerical features like 'no_of_sites', 'no_of_subjects', 'median_pages', 'days_onsite', 'days_remote', and 'num_of_subject_visits'. Prior to training, data preprocessing steps—such as encoding categorical variables and normalizing numerical features—were performed to enhance model performance.

The XG-Boost model was configured to learn complex nonlinear relationships within the data. Hyperparameters were fine-tuned via Grid Search to identify the optimal combination for minimizing prediction error. This systematic tuning included parameters such as learning rate, maximum tree depth, and subsample ratio, resulting in a robust and generalized predictive model for estimating attendee counts.

To address the challenge of features like pages_entered, which are present in the training data but missing in the test set, we adopted a median imputation strategy based on contextual levels. Specifically, for each missing feature in the test set, such as "pages," we computed the median value within relevant groups—namely, TA (Technical Area), PHASE, and COUNTRY. This approach ensures that imputed values are representative of the broader distribution within each group, thereby reducing potential bias caused by missing data.

Additionally, total emissions data were grouped and aggregated at these same levels, with their median values used during both training and testing. Incorporating these median-derived features enhanced the model's robustness by maintaining consistency across datasets and providing additional granularity, ultimately improving predictive accuracy on unseen data.

This model was used to predict the number of attendees, and subsequently, the percentage of attendees was calculated for each country in each study. The second model was then trained to predict the number of days onsite at the country level, using the predicted number of attendees from the first model. To determine the number of meetings, we used the predicted number of attendees to estimate the days onsite at the country level, which was then aggregated at the study level.

In this sequential modeling approach, the performance was not optimal due to error propagation from the first to the second model.

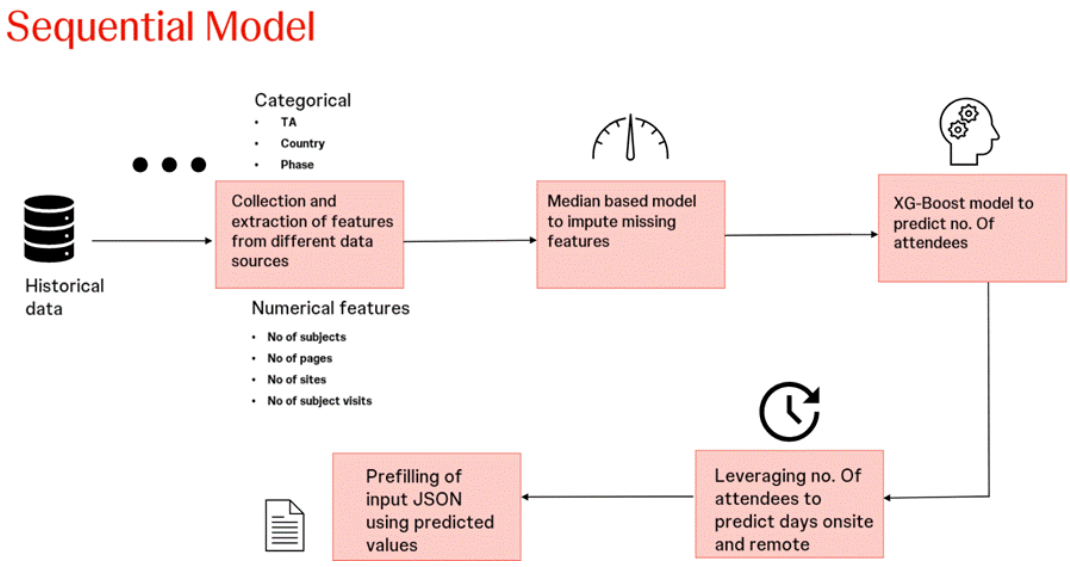

Figure: Sequential Modelling approach for prefilling

In the second modelling technique, we trained a combined model with multiple target variables. This model fared better in terms of overall RMSE (59.6 VS 33.6) and r2(0.80 VS 0.88) performance metrics as compared to the sequential model.

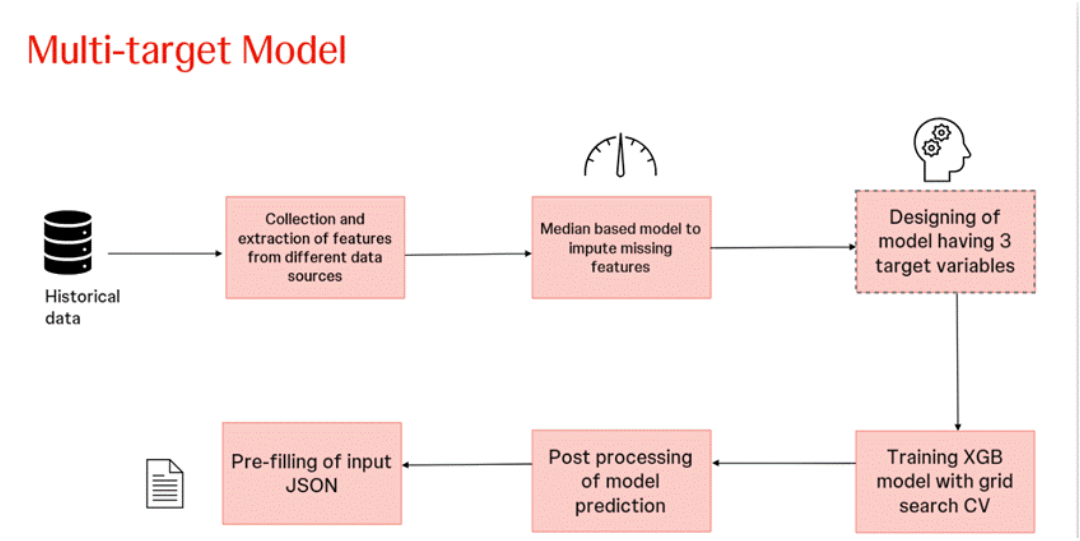

Figure: Multi-target Modelling approach for prefilling

Discussion

Our in-house deployed application represents a paradigm shift in trial design. Instead of reacting to emissions data after a trial, teams now proactively incorporate sustainability considerations from the outset. The AI engine not only enhances speed and efficiency by pre-filling values based on current data snapshots but also embeds industry-aligned green practices as thoughtful considerations throughout the study design process.

Conclusion

The J&J Sustainable Trial Design application module enables measurable, scalable, and repeatable sustainability improvements across clinical trials. With AI-powered assistance and embedded emissions modeling, it supports faster, smarter, and greener decision-making—aligning R&D efforts with climate goals.

Unlike traditional planning tools that require manual configuration of trial parameters, our application integrates an AI-powered pre-fill engine trained on historical trial design-emissions data. This feature not only accelerates protocol initiation but also provides users with a sustainability-informed design baseline—transforming passive emissions tracking into active, data-driven optimization.

References

1. Sustainable Healthcare Coalition (2024). Measuring Emissions in Clinical Trials: A Framework for Action.
2. McCoy, D., et al. (2023). The carbon footprint of clinical trials: A systematic review. BMJ Open, 13(1), e057229.
3. Johnson & Johnson (2023). Health for Humanity Report. https://healthforhumanityreport.jnj.com
4. World Health Organization (2023). Sustainable Healthcare: Key Guidelines and Practices.
5. International Council for Harmonisation (ICH). (2021). Good Clinical Practice (GCP) E6(R3)

ACKNOWLEDGMENTS

Extending our sincere thanks to the excellent leadership team and all members of Advanced insights, Data science and Digital health department, BGIA, for the dynamic Sustainability application development activities and ongoing collaboration.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Archana Nayak, PhD
DSDH, Johnson &Johnson R&D
Great Britain
Email: anayak32@its.jnj.com

Jason LaRoche
JLaRoche@its.jnj.com
BGIA, Johnson & Johnson R&D
Netherlands
Email: JLaRoche@its.jnj.com

Bhattacharjee, Eshanika
ZS Associates,
ebhatta2@its.jnj.com
India

Md Umarfaruque
ZS Associates,
umd@its.jnj.com
India

Bhargava Reddy
Janssen R & D
PA, USA
BReddy1@ITS.JNJ.com