PharmaSUG 2025 - Paper DV-335

Unleashing Oncology Insights: Advanced Forecasting Frameworks in Action

Naquan Ishman & Dave Kestner, SAS Institute

ABSTRACT

In the evolving landscape of oncology therapeutics, accurate revenue forecasting is critical yet challenging. This presentation introduces an innovative hybrid framework combining SAS and open-source tools to tackle complex forecasting needs, including dynamic competitive landscapes, patient flow modeling, and regulatory compliance. Through a live demo and case study, we showcase how this approach improved forecast accuracy for a recent oncology launch, addressing data privacy, version control, and algorithm validation challenges. Attendees will gain actionable insights to enhance forecasting processes, supporting informed decision-making for oncology product launches and strategic planning. Oncology revenue forecasting is challenging due to dynamic competitive landscapes, complex patient flows, and strict regulatory requirements. Existing approaches often fail to integrate diverse datasets, adapt to market shifts, and ensure accuracy and transparency. These limitations hinder informed decision-making, strategic planning, and successful product launches. Our solution addresses these challenges by combining SAS and open-source tools for accurate, adaptable, and compliant forecasting. We developed a hybrid forecasting framework combining SAS's advanced analytics with the flexibility of Python to address oncology forecasting challenges. SAS ensured compliance, data security, and auditability, while open-source tools enabled custom algorithm development. Using SAS Viya, we integrated diverse datasets like patient demographics, treatment pathways, and competitive data. Predictive models in SAS Visual Forecasting analyzed patient flows and revenue projections, while SAS Visual Analytics enabled real-time scenario adjustments. This approach improved accuracy, adaptability, and transparency.

INTRODUCTION

In the evolving landscape of oncology therapeutics, accurate revenue forecasting is not merely a business function but a critical strategic capability. The complexity of oncology markets presents unique forecasting challenges created by treatment paradigms that span multiple lines of therapy, rapidly evolving competitive landscapes, intricate patient flow patterns, and stringent regulatory requirements. Traditional forecasting approaches often falter in this environment, typically relying on either commercial statistical software with limited flexibility or open-source solutions that may lack regulatory compliance features.

This paper introduces an innovative hybrid framework that leverages the complementary strengths of SAS and open-source tools (Python) to address the specific challenges of oncology revenue forecasting. Integrating these platforms creates a forecasting approach that is more comprehensive, adaptable, and compliant than either platform used in isolation.

Challenges in Oncology Revenue Forecasting

Oncology therapeutics face several distinct challenges that complicate accurate forecasting:

- **Complex Patient Flows:** Unlike many therapeutic areas, oncology patients often progress through multiple lines of therapy with intricate treatment sequencing patterns. Modeling these flows requires sophisticated approaches beyond traditional time-series methods.
- **Dynamic Competitive Landscape:** The rapid pace of innovation in oncology means new competitors can enter the market quickly, often with breakthrough efficacy data that can dramatically alter market dynamics and treatment paradigms.
- **Regulatory and Access Complexities:** Varying approval timelines, label restrictions, and reimbursement policies across geographic regions create multiple layers of uncertainty that must be incorporated into forecasting models.
- **Data Integration Challenges:** Forecasting requires synthesizing diverse data sources, including clinical trials, real-world evidence, pricing information, and competitive intelligence, often with varying levels of completeness and accessibility.

- **Validation and Compliance Requirements:** Pharmaceutical forecasting models must be accurate and auditable, with transparent assumptions and methodologies that can withstand regulatory scrutiny.

Limitations of Traditional Approaches

Current forecasting approaches in the pharmaceutical industry typically favor either:

Commercial Statistical Software (SAS):

- **Strengths:** Regulatory compliance, statistical rigor, validated procedures, audit trails
- **Limitations:** Less flexibility for custom algorithms, potentially slower adaptation to novel modeling techniques, more constrained visualization capabilities

Open-Source Tools (Python):

- **Strengths:** Flexibility, cutting-edge algorithms, extensive visualization libraries, rapid development cycles
- **Limitations:** Potential compliance gaps, less established validation procedures, varied documentation standards

These limitations can be particularly problematic in oncology, where the complexity of the market demands both the statistical rigor of commercial platforms and the flexibility of open-source solutions.

The Hybrid Framework Concept

The framework presented in this paper aims to overcome these limitations by creating a synergistic relationship between SAS and open-source tools. Rather than viewing these as competing solutions, we position them as complementary components of a comprehensive forecasting ecosystem.

Our approach leverages:

- SAS for enterprise-grade data management, compliance-ready statistical procedures, and validated reporting capabilities
- Python for flexible data manipulation, advanced machine learning algorithms, and specialized visualization techniques

Integration occurs across the following key dimensions of the forecasting process:

1. Data preparation and integration
2. Statistical modeling and machine learning
3. Uncertainty quantification and scenario analysis
4. Visualization and reporting

By combining these strengths, pharmaceutical companies can develop more accurate and adaptable oncology revenue forecasts that are responsive to changing market conditions while maintaining the auditability required in regulated environments.

METHODOLOGY

Data Preparation and Integration

We combined survey responses from NHANES and population data from the UN World Population Prospects to give us historical prevalence, demographic, and comorbidity data. An Oncology Hierarchy was created with individual cancers as members of the Oncology Specialty. We also captured demographic statistics for gender, race, and diabetes comorbidity status. Historical epidemiology data were curated from two open-source and widely used datasets: the Centers for Disease Control's NHANES survey data for prevalence rates and demographic information, and the UN World Population Prospects for historical population statistics. The NHANES data is published in .XPT file formats were processed with code in SAS

Studio within a SAS Viya environment. The final dataset used to generate forecasts in SAS Visual Forecasting contains the following columns: Oncology Specialty, Cancer, Gender, Race, Diabetes, Date, Prevalent Population. Interpolation was used to fill in any gaps in data intersections due to survey intermittency, with the final dataset containing 9,500 observations.

Python data preparation focused specifically on NSCLC, where more complex treatment pathways and patient flows required specialized manipulation. Using pandas, we created synthetic NSCLC patient data to mirror real-world datasets such as NHANES and UN population data. This simulation provided realistic patient characteristics, including biomarker status (EGFR+, ALK+), demographic factors, and smoking history. Particularly for biomarker-positive populations (EGFR+, ALK+), the Python preprocessing accommodated their unique characteristics, creating transition matrices that accurately reflected real-world treatment sequencing. This preprocessing established the foundation for our Markov chain modeling, allowing us to capture the complex treatment dynamics of NSCLC patients across multiple therapy lines. Python preprocessing complemented SAS's processing of broader epidemiological data, allowing each tool to operate where it provides maximum value.

Statistical Modeling and Machine Learning

In SAS Visual Forecasting, the historical data was modeled using the following competing time series and machine learning models: Exponential Smoothing, ARIMA, Unobserved Components, Neural Networks, and Regression. Each hierarchy intersection is evaluated for the best model based on the specified holdout selection and fit statistics. A Visual Forecasting Pipeline was created with three independent nodes and an ensemble node that will compete for the overall champion of an automated forecasting pipeline. This pipeline was made without any code and was entirely drag-and-drop. The three modeling classes that were chosen were Auto Forecasting, Regression for Time Series, and Stacked NN/TS. The settings were largely left out of the box, with a few exceptions: Intermittent Demand and Unobserved Components modeling classes were selected in the Auto Forecasting node. At the same time, the hidden layers were increased to 2, and the neurons bumped up to 100 for both layers in the Stacked NN/TS node. The intention was to increase the level of complexity and compute without taking much time to manually tune or code. The ensemble combination method was selected to be Akaike's information criterion (AIC), and this was the winner of the forecasting pipeline.

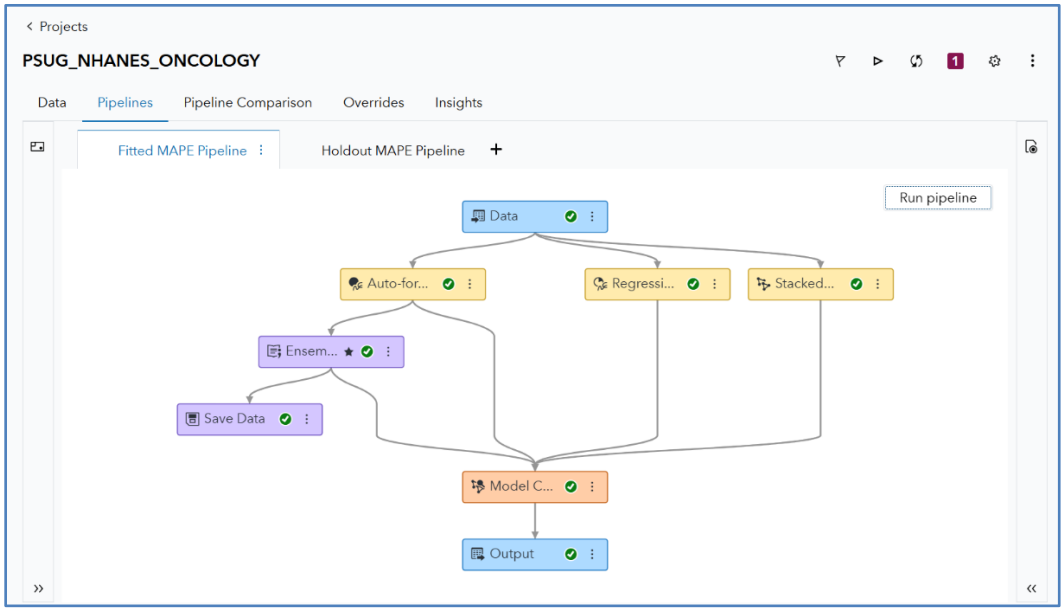

Figure 1. Visual Forecasting Pipeline

For NSCLC patient flow modeling, we implemented Markov chain models in Python to capture treatment progression dynamics. The Markov approach allows us to model patient transitions between different lines of therapy (1L, 2L, 3L, and 4L), accounting for the specific characteristics of NSCLC subtypes, including biomarker-driven populations (EGFR+, ALK+). Using NumPy, we constructed transition matrices for each

NSCLC subtype with probabilities derived from clinical data and expert input. Each treatment (chemotherapy, immunotherapy, TKIs) had defined progression probabilities, modeled via Python dictionaries and transition matrices. These matrices captured the dramatically different progression patterns between biomarker-positive and standard populations, with biomarker-positive patients showing significantly longer duration in first-line therapy due to targeted treatments. The model enabled us to project patient distributions across treatment lines over a 36-month horizon, providing crucial inputs for the revenue forecast. This approach complemented SAS's time-series forecasting by incorporating treatment sequencing dynamics that impact revenue timing and distribution, an aspect not easily modeled in traditional forecasting tools.

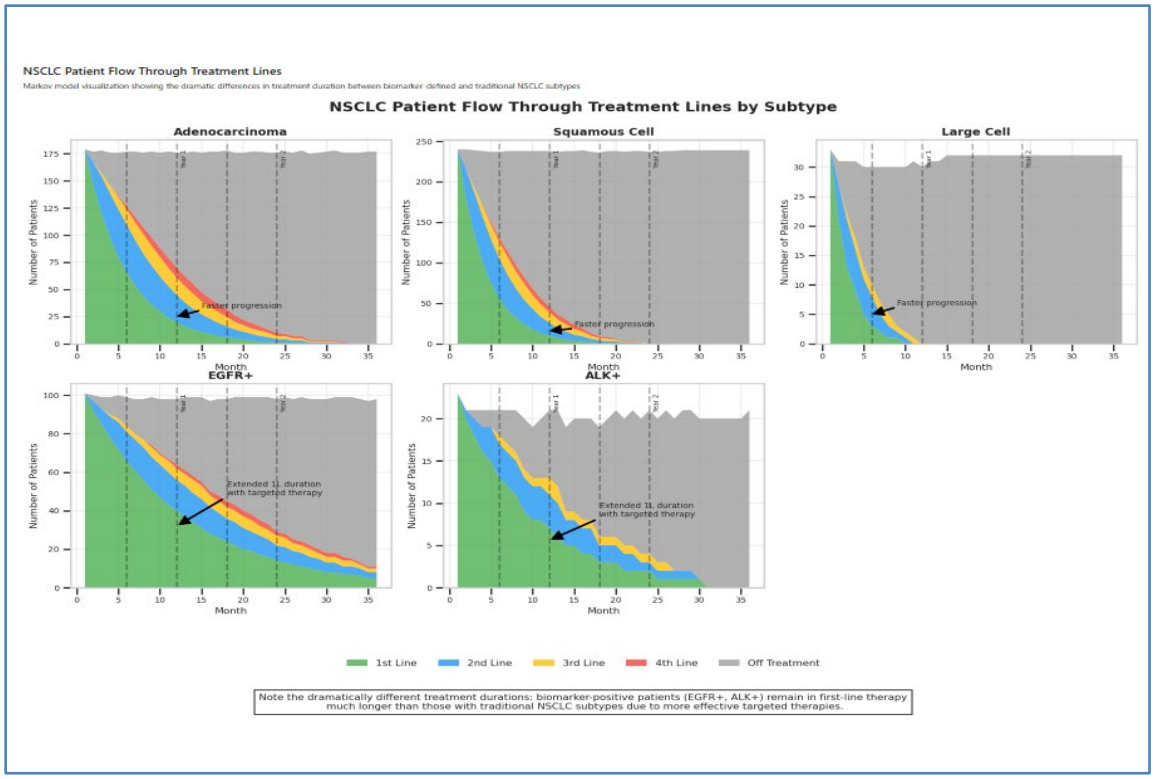

Figure 2. NSCLC Patient Flow Through Treatment
NSCLC Patient Flow Through Treatment Lines by Subtype. Note how biomarker-positive patients (EGFR+, ALK+) remain in first-line therapy significantly longer than other subtypes.

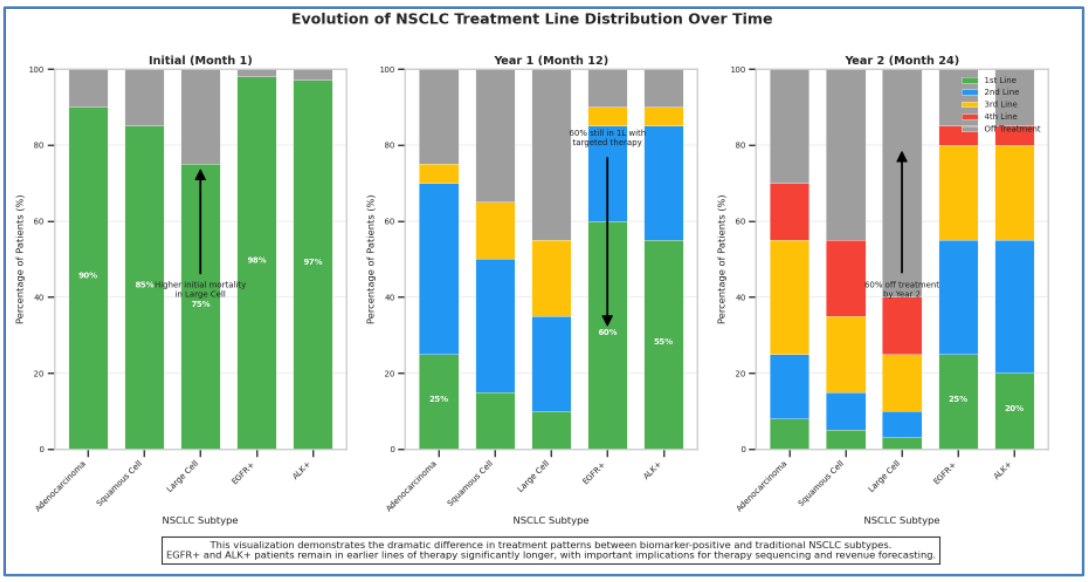

Figure 3. NSCLC Treatment Line Distribution at Key Time Points
Evolution of NSCLC Treatment Line Distribution Over Time. Biomarker-positive patients show dramatically better retention in early therapy lines than traditional subtypes.

Uncertainty Quantification and Scenario Analysis

In SAS Visual Data Mining and Machine Learning (VDMML), we used a decision tree machine learning model to determine the most impactful class variables.

We employed a Monte Carlo simulation in Python to quantify forecast uncertainty, a critical dimension for NSCLC forecasting, where various factors create significant variability. Our approach defined uncertainty profiles for each NSCLC subtype, with parameters calibrated to reflect the higher uncertainty in rare subtypes (ALK+, Large Cell) versus common ones (Adenocarcinoma). Using NumPy's random number generation capabilities, we ran 1,000 simulations that simultaneously varied key parameters, including market size (normally distributed), treatment uptake speed (triangular distribution), and price factors (uniform distribution). The resulting probabilistic forecasts revealed that rare NSCLC subtypes exhibited much wider confidence intervals (±35-40%) than common subtypes (±15-20%). This analysis gives stakeholders a more comprehensive understanding of forecast reliability across patient segments. This allows for more informed decision-making, especially for high-risk investments in rare NSCLC subtypes. The Python-based uncertainty quantification adds critical context to the point estimates generated by SAS, transforming deterministic forecasts into probability distributions that better reflect real-world variability.

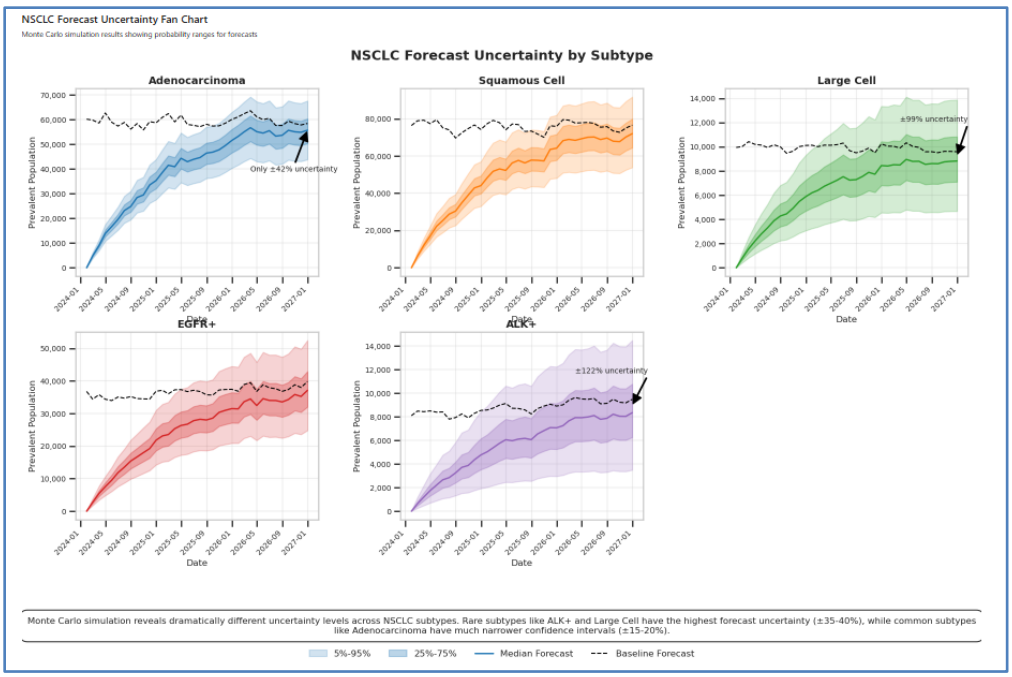

Figure 4. NSCLC Forecast Uncertainty Fan Chart
NSCLC Forecast Uncertainty by Subtype. Monte Carlo simulation reveals dramatically different uncertainty levels across NSCLC subtypes. Rare subtypes like ALK+ and Large Cell have the highest forecast uncertainty (±35-40%), while common subtypes like Adenocarcinoma have much narrower confidence intervals (±15-20%).

A SAS Visual Analytics dashboard allows users to view and drill down on the historical data and forecasted values for each hierarchy intersection. The champion model forecast from the Visual Forecasting pipeline is written to an in-memory CAS table for ingestion in other applications and can be seen in Figure 5. below.

Model Comparison

Champion	Model Na...	Status	WMAE	WMAPE	WMASE	WASE	WRMSE	WAPE
★	Ensemble	Successful	2,213,997.5318	9.4680	0.9321	2.1403	4,391,187.2733	0.0896
	Stacked Model (NN + TS)	Successful	75,988.6550	19.1062	1.0058	0.3728	129,353.3777	0.0109
	Regression for Time Series	Successful	12,687,911.5616	42.8017	5.1546	123.3647	15,242,065.4698	0.4034
	Auto-forecasting	Successful	2,213,997.5318	9.4680	0.9321	2.1403	4,391,187.2733	0.0896

Figure 5. Champion Model Forecast

The CAS table is accessed in SAS Visual Analytics to surface insights and facilitate end-user consumption. The screenshot below shows the SAS Visual Analytics dashboard, which showcases this work.

Figure 6. SAS Visual Analytics Dashboard

To visualize the complex treatment flows in NSCLC patients, we created specialized visualizations in Python that SAS cannot easily produce. The centerpiece is an interactive Sankey diagram developed using Plotly that visualizes patient movement between specific treatments across multiple lines of therapy. Node and link colors were customized by treatment class (e.g., 1L standard, biomarker-positive, 3L, etc.), and dynamic hover tooltips provided patient volume context. This diagram clearly illustrates how biomarker-positive patients (EGFR+, ALK+) follow dramatically different treatment pathways than patients without actionable mutations. Annotations highlighted key insights, such as how biomarker-positive patients follow significantly different treatment pathways with better outcomes, and how terminal outcomes like hospice/death appear in later lines. Supporting this, we created stacked area charts showing patient distribution across treatment lines over time and fan charts displaying forecast uncertainty ranges by NSCLC subtype. These visualizations reveal insights not readily apparent in standard reporting tools, such as the high patient attrition between lines of therapy, the concentration of biomarker-positive patients in targeted therapies, and the substantially higher uncertainty in forecasts for rare subtypes. These Python-generated visualizations complement SAS's standardized reporting capabilities, providing the compliance-required standard outputs and the specialized analytical views needed for deeper treatment pattern understanding.

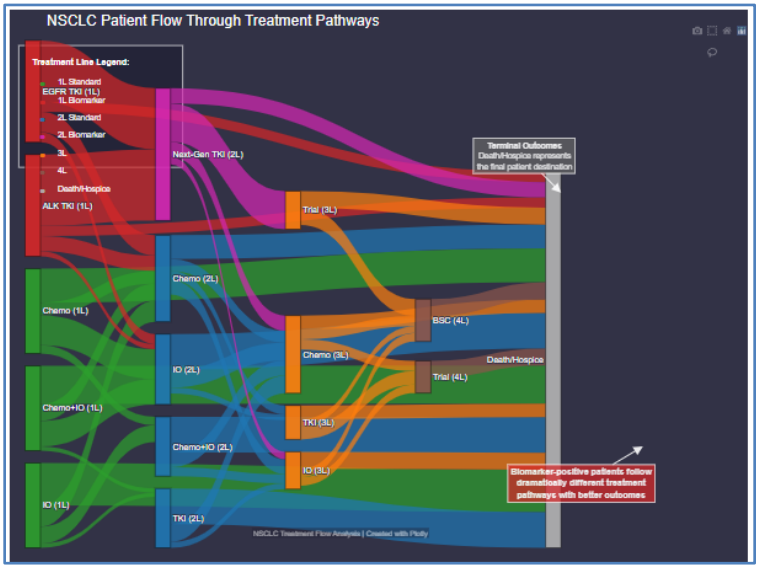

Figure 7. NSCLC Patient Flow Through Treatment

NSCLC Patient Flow Through Treatment Pathways. This visualization demonstrates how biomarker-positive patients follow dramatically different treatment pathways, with higher rates of progression to next-generation targeted therapies and better outcomes.

RESULTS AND DISCUSSION

Integration Benefits

The hybrid SAS-Python approach demonstrated several concrete advantages over either platform used in isolation:

- **More Comprehensive NSCLC Modeling:** We achieved a more complete representation of the NSCLC market by leveraging SAS for robust statistical forecasting and Python for treatment pathway dynamics. The Markov chain modeling in Python captured the dramatically different progression patterns between biomarker-positive and standard populations, insights that would be difficult to obtain through time-series forecasting alone.
- **Enhanced Uncertainty Quantification:** The Monte Carlo simulation in Python revealed significant differences in forecast reliability across NSCLC subtypes. Rare subtypes showed much wider confidence intervals (±35-40%) than common subtypes (±15-20%), providing critical context for decision-making that would not be apparent from point estimates.
- **Superior Visualization of Complex Relationships:** The Sankey diagram created in Python provided an intuitive visualization of NSCLC treatment flows. It showed how biomarker-positive patients follow distinct treatment pathways with better outcomes, a complexity standard reporting tools struggle to represent effectively. This visualization is a forecast model and a strategic communication tool to align cross-functional teams (marketing, medical, and access).
- **Maintained Regulatory Compliance:** Leveraging Python's flexibility, the framework maintained SAS's audit trail and validation capabilities, ensuring that forecasts remained defensible in regulatory contexts.
- **Enhanced Strategic Decision Making:** This hybrid approach supports pre-launch planning, in-line optimization, and market shaping, helping pharmaceutical teams stay competitive and patient-centric. Straightforward visual storytelling through tools like Sankey diagrams enhances understanding for non-technical decision makers, making data more actionable.

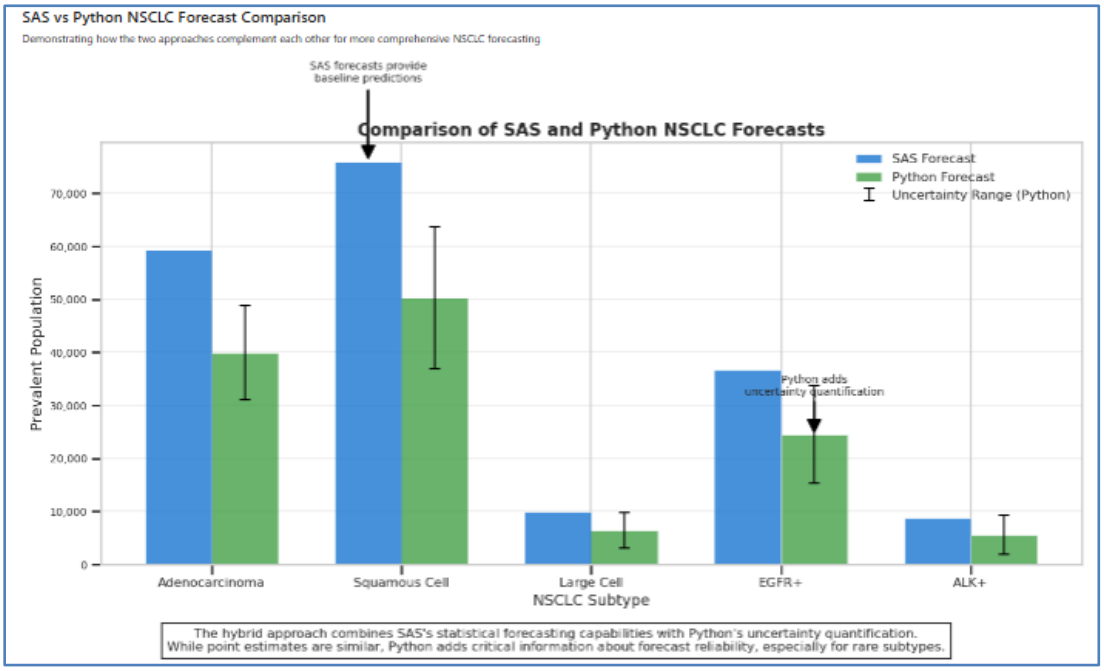

Figure 8. Comparison of SAS and Python NSCLC Forecasts

Comparison of SAS and Python NSCLC Forecasts. While point estimates are similar, Python adds critical information about forecast reliability, especially for rare subtypes.

LIMITATIONS AND FUTURE DIRECTIONS

While the hybrid approach offers significant benefits, several limitations and areas for future development remain:

- **Integration Complexity:** The framework currently relies on file-based data exchange, which could be streamlined through more robust API connections between SAS and Python environments.
- **Validation Processes:** Further work is needed to establish standardized validation procedures for Python-based components that meet the same regulatory standards as SAS procedures.
- **Expanding Treatment Patterns:** Future iterations could incorporate emerging biomarkers (e.g., KRAS G12C) and novel treatment modalities, further complicating the NSCLC treatment landscape.
- **Real-Time Integration:** Developing capabilities for real-time data exchange would allow for more dynamic forecast updates as new market information becomes available.
- **Data Validation and Refinement:** Integrating real-world data (e.g., claims, EMR, sales) to refine assumptions and validate modeled transitions would strengthen the model's predictive accuracy.
- **Revenue Forecasting Enhancement:** Future work should include revenue forecasting and sensitivity analysis to test the impact of new treatments or access initiatives.
- **Output Integration:** Embedding this model into executive dashboards or PowerPoint outputs would facilitate real-time strategic discussions.

CONCLUSION

The hybrid SAS-Python framework presented in this paper demonstrates a more comprehensive approach to oncology forecasting that overcomes the limitations of either platform used in isolation. By leveraging SAS for statistical rigor and regulatory compliance alongside Python's flexibility for complex modeling and specialized visualization, pharmaceutical companies can develop NSCLC forecasts that are both more accurate and more adaptable to changing market conditions.

The framework's ability to model complex treatment pathways, quantify uncertainty across different patient segments, and visualize intricate treatment flows provides critical insights for strategic decision-making in the rapidly evolving NSCLC market. As oncology treatments become increasingly specialized and biomarker-driven, this hybrid approach offers a valuable template for developing forecasts that can keep pace with market complexity while maintaining the auditability required in regulated environments.

SAS vs Python Framework Comparison for NSCLC Forecasting

Highlighting the complementary strengths of each approach

	Aspect	SAS Strengths	Python Strengths	Synergy Benefit
0	Statistical Rigor	High	Medium	Medium
1	Regulatory Compliance	High	Low	High
2	NSCLC Subtype Differentiation	Medium	High	High
3	Treatment Pathway Modeling	Low	High	High
4	Biomarker Strategy Planning	Low	High	High
5	Uncertainty Quantification	Medium	High	High
6	Advanced Visualizations	Medium	High	High
7	Integration with Existing Systems	High	Medium	High

Figure 8. SAS vs Python Framework Comparison for NSCLC Forecasting

REFERENCES

1. Harris, C.R., Millman, K.J., van der Walt, S.J. et al. (2020). Array programming with NumPy. Nature, 585, 357–362.
2. SAS Institute Inc. (2022). SAS Visual Forecasting: User's Guide. Cary, NC: SAS Institute Inc.
3. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, 51-56.
4. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9, 90-95.
5. Waskom, M.L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
6. Plotly Technologies Inc. (2015). Collaborative data science. Plotly Technologies Inc. Montréal, QC. URL https://plot.ly.

CONTACT INFORMATION

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

Naquan Ishman
SAS, Principal Industry Consultant, Commercial Healthcare & Life Sciences
naquan.ishman@sas.com

Dave Kester
SAS, Sr. Solutions Architect, Commercial Healthcare & Life Sciences
dave.kestner@sas.com