University of West Florida

Capston Project In Data Science IDC-6940 | Professor Dr. Shusen Pu

Edwin Quijano

Tuesday, September 16, 2025

Literature Review

The study looks at how pharmaceutical manufacturers can better predict short-term demand by combining traditional sales history with "downstream" information such as inventory held by distributors and by using modern, pattern-finding computer models. The authors show that when manufacturers learn from many products at once and include fresh information about what is happening in their distribution network, forecasts improve meaningfully over common methods used in practice. This has clear implications for planning production, purchasing materials, and meeting patient needs reliably.

Background and Purpose

Pharmaceutical companies must match what they make with what patients and pharmacies actually need. If they produce too little, people can face shortages; if they produce too much, they waste money and storage space. Traditionally, many firms forecast demand by looking only at their own past sales and applying simple formulas. But the pharma supply chain generates a lot of additional information especially from wholesalers and distribution centers that can hint at what orders will look like in the next few weeks.

The study by Zhu et al. (2021) focuses on two common problems. First, each individual product may not have enough recent, reliable data to support a strong forecast on its own. Second, simple forecasts miss outside signals like how much inventory partners are currently holding that can push orders up or down. The authors ask: Can we (a) learn across many products at the same time to make better use of limited data, and (b) fold in fresh, downstream inventory information to sharpen short-term predictions?

1

This question matters because short-horizon forecasts (the next one to eight weeks) drive day-to-day operational decisions: how much raw material to buy, which production lines to schedule, and how much finished product to stage for shipment. Getting these predictions right improves service levels and reduces costly swings in inventory.

Methods Used

The authors use large, real-world datasets from two pharmaceutical manufacturers, covering hundreds of drug products over multiple years. For each product, they assemble weekly records that include orders placed on the manufacturer and the inventory levels reported by partners in the distribution network.

Rather than building a separate forecast for each product, they train models that learn from all products together. Think of it as pooling experience: lessons from one product can help forecast another, especially when their demand patterns are similar. To keep "apples with apples," the researchers also test ways to group products by sales level and variability, by therapeutic class, or by similarity in their historical patterns—before training these pooled models within each group.

They compare these pooled approaches to familiar methods like moving averages and exponential smoothing. They also test several types of modern models. When the paper refers to a "recurrent neural network," we can think of it as an AI method designed to recognize patterns over time—like how demand rises, falls, and spikes—and to use those patterns to predict the near future.

A key part of the design is adding downstream inventory (what wholesalers and distribution centers currently have on hand) as an extra input. The team examines whether adding this information—especially the most recent week—improves accuracy for short-term forecasts.

Moving Average (MA) & Exponential Smoothing (SES/Holt–Winters)

Role in the study: These were the practice-style baselines. For every product (SKU) and forecast horizon (1–8 weeks ahead), the authors fit standard univariate models to that SKU's own history and produced forecasts.

How they were applied:

MA: For a chosen window k, they computed $\hat{y} = \sum_{i=0}^{k-1} y_t - i$ and rolled this forward for horizons 1–8 with refits/updates in a rolling-origin scheme.

SES/Holt–Winters: They estimated smoothing parameters on the training window α for SES, Plus β, τ if trend/seasonality were used, and generated 1–8 week-ahead forecasts.

Why: These are widely used by planners, so they serve as conservative, real-world baselines to judge whether sophisticated methods are truly better.

Per-SKU Linear Regression with Lags

Role: A transparent baseline that allows adding exogenous variables, downstream inventory q_t
How applied: For each SKU, the study fit

$$\hat{y} = \beta_0 + \sum_{i=0}^{p-1} \beta_i + 1 y_t - i + \sum_{j=0}^{q-1} \gamma_j + 1 q t - j + \epsilon_t + h$$

choosing (demand lags) and (inventory lags) via validation. This showed the incremental value of inventory even in a simple linear setup.

Vector Autoregression (VAR)

Role: A multivariate time-series benchmark to test whether jointly modeling many series with linear cross-lag terms helps.

How applied: The authors specified a VAR( ) across a chosen set of series (e.g. multiple SKUs or node-level series), estimated coefficient matrices, and produced multi-step forecasts.

Finding: With many series/lags, VAR can become over-parameterized, often underperforming the more flexible ML approach that shares parameters more efficiently.

The results

Learning across products helps a lot. When the model is trained on many products at once, it can spot general patterns that a single-product model would miss. This is especially helpful for products with sparse or noisy histories.

Simple, fresh downstream signals matter. Including current inventory from wholesalers and distribution centers makes forecasts better, particularly for the next few weeks. Interestingly, the most recent inventory snapshot tends to carry the most predictive power; adding long histories of inventory does not add much.

Grouping similar products strengthens results—especially for low-volume items. Training pooled models within sensible groups (for example, by sales level/volatility or by therapeutic class) further improves accuracy. The biggest benefits appear for low-volume products, which are usually the hardest to forecast.

Together, these findings show that manufacturers can raise forecast accuracy—often by a meaningful margin—without reinventing their entire process. They can keep using weekly

cycles, but with models that learn from many products at once and that pay attention to what is happening in the downstream network right now.

Why does this matter beyond the study? Better forecasts translate to steadier production schedules, fewer emergency expedites, and improved product availability. In pharmaceuticals, that also supports patient safety and trust.

How It Fits with Other Studies

The results connect two long-running themes in operations and supply-chain research:

Information sharing helps. Prior theory has argued that sharing downstream data (like inventory and sales-through) should improve upstream decisions. Zhu et al. (2021) provide concrete, large-scale evidence of that effect for forecasting at the manufacturer level.

Modern models can unlock value when framed correctly. While there is a growing body of work using advanced analytics for retail and e-commerce, this paper shows that, in a manufacturing context, the way you structure the problem—learning across many related products and adding timely external signals—matters as much as the specific algorithm.

The authors also caution that pushing models to very granular nodes like building separate forecasts for every distribution center and then adding them up is not always helpful for the aggregate manufacturer forecast. The noise introduced at that level of detail can outweigh the benefits unless the decisions are truly made at that node.

Relevance for tne Capstone Project

My capstone project focuses on intelligent inventory optimization that depends on accurate short-term forecasts. This study offers a practical blueprint:

Forecasting backbone. Use a pooled, cross-product model that can learn from all items together. Start globally, then test sensible product groupings to further improve accuracy, with special attention to low-volume items.

Right-now signals. Add the latest distributor and DC inventory as an input to the forecast. Prioritize the most recent week to capture near-term shifts in ordering pressure.

Keep granularity aligned with decisions. If ordering and production decisions are made at the manufacturer level, don't over complicate the forecast by going too deep into node-level detail unless there is a clear payoff.

Bridge to optimization. Feed the improved short-term forecasts directly into inventory policies or decision tools, including reinforcement learning, if used, so the planning system responds to current network conditions rather than just historical averages.

Conclusion

Zhu et al. (2021) show that pharmaceutical manufacturers can significantly improve short-term demand forecasts by (1) learning across many products at once, (2) organizing products into sensible groups, and (3) adding fresh downstream inventory information. These steps are practical, data-driven, and aligned with how operations teams already work on weekly planning cycles. For organizations seeking steadier production, lower costs, and better service—and for projects like mine that rely on strong near-term predictions—this approach provides a clear, actionable path forward.

References

Zhu, X., Ninh, A., Zhao, H., & Liu, Z. (2021). Demand Forecasting with Supply-Chain Information and Machine Learning: Evidence in the Pharmaceutical Industry. Production and Operations Management, 30(9), 3231–3252. https://doi.org/10.1111/poms.13426