


Demand Forecasting with Supply-Chain Information and machine learning: Evidence in the Pharmaceutical Industry

Xiaodan Zhu

Computer Science Department, William & Mary, Williamsburg, Virginia 23185, USA, xzhu08@email.wm.edu

Anh Ninh*

Mathematics Department, William & Mary, Williamsburg, Virginia 23185, USA, atninh@wm.edu

Hui Zhao 

Smeal College of Business, The Pennsylvania State University, University Park, Pennsylvania 16802, USA, huz10@psu.edu

Zhenming Liu

Computer Science Department, William & Mary, Williamsburg, Virginia 23185, USA, zliu@cs.wm.edu

Accurate demand forecasting is critical for supply chain efficiency, especially for the pharmaceutical (pharma) supply chain due to its unique characteristics. However, limited data have prevented forecasters from pursuing advanced models. Such problems exist even when long history of demand data is available because historical data in the distant past may bring little value as market situation changes. In the meantime, demands are also affected by many hidden factors that again require a large amount of data and more sophisticated models to capture. We propose to overcome these challenges by a novel demand forecasting framework which “borrows” time series data from many other products (cross-series training) and trains the data with advanced machine learning models (known for detecting patterns). We further improve performance of the cross-series models through various “grouping” schemes, and learning from non-demand features such as downstream inventory data across different products, information of supply chain structure, and relevant domain knowledge. We test our proposed framework with many modeling possibilities on two large datasets from major pharma manufacturers and our results show superior performance. Our work also provides empirical evidence of the value of downstream inventory information in the context of demand forecasting. We conduct prior and post-hoc field work to ensure the applicability of the proposed forecasting approach.

Key words: demand forecasting; machine learning; cross-time-series training; supply chain data; pharmaceutical industry

History: Received: March 2019; Accepted: March 2021 by Nitin Joglekar, after 4 revisions.

1. Introduction

Accurate demand forecasting is the basis for supply chain efficiency since it essentially drives all important operational decisions, from raw material supply planning, production planning, and inventory management, to financial goals. For a pharmaceutical (pharma) manufacturer, demand forecasting can be even more critical because (1) any mismatch between demand and supply could ripple through the drug distribution channel and impact the patients, sometimes even causing life-threatening situations; and (2) any demand that is not fulfilled could potentially lead to permanent lost sales from a patient, because patients who cannot afford the uncertainty in their

order fulfillment may switch to an alternative drug. For drugs treating chronic illnesses, this could mean huge financial losses for the drug manufacturer.

Below, we describe the current state of demand forecasting in the pharma industry (section 1.1), provide details of our interactions with top pharma companies (section 1.2), and propose a new forecasting approach to address the many challenges faced in demand forecasting of pharma products (section 1.3).

1.1. Current State of Pharma Demand Forecast

Based on surveys, reports, and literature, this section describes the current situation of pharma demand forecasting in terms of methods, data used, forecast horizon, and the use of forecasts.

Forecasting Methods

In 2018, the global market for pharmaceuticals reached \$1.2 trillion, up by \$100 billion from 2017 (IQVIA Institute for Human Data Science) and the United States alone holds over 45% of the global pharma market. Due to the high profit margin—the top 10 pharma companies in the United States had a median profit margin of 17% (Angell 2004)—there was a low need for supply chain efficiency and the pharma industry did not pay much attention to demand forecasting methods until more recently (Kiely 2004, Merkurueva et al. 2019). This seemingly explains the dominance of simple demand forecasting methods used in the industry. Jain (2003), based on a pharma industry survey, listed the most popular forecasting models as the basic exponential smoothing, moving averages, and regression. More recently, Weller and Crone (2012) surveyed 200 companies (14 of which are pharma companies) and confirmed that univariate statistical methods have maintained their dominant position in pharma and other industries. In particular, basic exponential smoothing, moving average, and naive methods account for 82.1% of all statistical forecasts. This is true even in the era of using software. Analyzing the results from a joint research initiative of IndustryWeek and SAS, Chase (2016) summarized that while companies might be using various software to help with demand forecasting, moving average, basic exponential smoothing, and simple regression models are still the most popular forecasting methods.

Cook (2016) outlined the typical procedure of demand forecast for in-market pharma products as (i) trending historical data, (ii) applying the effects of ex-trend events (i.e., external or internal events that may affect demand but not reflected in the historical data), and (iii) converting trended data into forecast outputs based on the first two steps. The challenges to the forecasters are to identify these ex-trend events and quantify the effects of these events on the forecast. While these could be done by human experts' judgments, this is atypical because most pharma companies deal with a large number of national drug codes (NDCs), ranging anywhere from hundreds to thousands of marketed products with different therapeutic characteristics (Cook 2016). Each National Drug Code (NDC) is a unique 10-digit or 11-digit, 3-segment identifier for drug products in the United States. Even for drugs with the same active pharmaceutical ingredient (API), they may have different dosages, delivery methods (tablets vs. injection, etc.), corresponding to different NDCs. Since each NDC has its own demand, the manufacturers forecast the demand for each of them. In practice, forecasting is typically done using software (e.g., SAP, Oracle, R, Excel). One other concern of human judgment is its quality,

consistency, and dependence on experiences; hence, human judgment is typically only incorporated for special cases such as new product launch and promotion (Arvan et al. 2019). In these cases human judgment is incorporated on top of the algorithm-generated forecasts (Jain 2003). Hence, accuracy of algorithm-generated demand forecasting is particularly important.

As pharma companies face increasing pressure from government and the public to reduce drug prices, they are open to more advanced demand forecasting technologies which, as mentioned, are critical for many aspects of supply chain efficiency. However, there are also many challenges in this area as we will lay out in section 1.3, which this work is aimed to help resolve.

Data Used

Currently, for statistical forecasting methods, historical sales are the most commonly used data for forecasting (Weller and Crone 2012). Benchmarking studies reported in Merkurueva et al. (2019) point out that "although there is plenty of data useful for more accurate demand forecasting, data usage is limited due to various aspects (e.g., different data formats; lack of data integration tools)." Chase (2016) also observes that despite all the improvements with data collection, downstream data have not been utilized for supply chain demand forecasting and planning. The value of downstream data has been overlooked, even after *supply chain visibility* is made available. Since mid-2000s, to streamline ordering and purchasing processes in the drug supply chain, electronic data interchange (EDI) have been adopted in the pharma industry. As part of the fee-for-service (FFS) arrangements with the manufacturer, wholesalers must provide inventory data to the manufacturer, typically via the EDI interface known by their numerical designations such as 867, 852, 180 among many others. For instance, EDI 852 contains inventory, product stocking, and product movement records from the trade partners' distribution centers to the manufacturer (Xu et al. 2018). However, there has been a "lack of new models for increasing forecasting intelligence" (Merkurueva et al. 2019) and "minimal investment in the analytic skills of demand planners" (Chase 2016). Our paper is the first to explore the value of some of these data to pharma demand forecasting tasks.

Forecast Horizon and the Use of Demand Forecast

According to Cook (2016), the forecasting horizon within the pharma industry can range from short-term, medium-term, to long-term forecasts. The long-term forecasts (> 5 years ahead) are used for strategic planning. For example, to launch a new product 3 years from now, a 10-year forecast is often used. The medium-term forecasts (> 1 year) are produced for financial forecasting and budget planning. The short-

term forecasts (daily, weekly, monthly) involve with operations such as inventory decisions and manufacturing decisions (e.g., procurement of raw materials, scheduling). Our paper primarily focuses on forecasting roles in supply chain operations, hence the forecasting horizon considered is 1–8 weeks. However, the proposed approach can be adapted for longer horizon as well.

1.2. Interactions with Pharma Companies

To further confirm the current state of demand forecasting in the pharma industry, we interacted with five top pharma companies based on a focused questionnaire directed at the points of interest in this study. Specifically, we designed a list of questions in the three aforementioned categories: forecasting methods, data used, and forecast horizon and the use of demand forecast, to which the companies provided answers. Appendix D includes the list of specific questions. When in doubt, we had further interactions with the companies through interviews or additional correspondences.

All pharma companies we interacted with confirmed that they use software to forecast demands with simple models such as exponential smoothing (two companies), moving average (two companies), and linear regression models (three companies). Some companies use more than one of the three. In terms of human judgment, one company estimated that human judgment is involved in <10% of the cases, while the remaining companies use human judgments in 10–30% of the cases. In a recent review paper, Arvan et al. (2019) reported that human judgments are primarily used for new product launches and promotions. Our survey confirmed this observation; none of the companies mentioned using human judgment for any other purposes. Furthermore, all companies verified that only historical demand is used to generate the statistical forecasts. In terms of forecasting horizon, it varies from 3 months (three companies), 1 month (one company), to 1 week (one company). Results of demand forecast are used in a wide variety of activities. All companies utilize demand forecast for inventory decisions. In addition, it is also used for cash flow and workforce planning (four companies), capacity planning (three companies), production planning (three companies), and promotion planning (three companies) and setting sales targets (one company).

Finally, none of the companies surveyed are currently using machine learning models, although one is in the early stage of investigating it. All companies mentioned that they are open to machine learning models as long as these models can bring sufficient improvement, even if such models are less interpretable than the currently used models, where

sufficient improvement was defined in our survey as >10%.

While our interactions with five companies should not be generalized beyond what they should be, they do seem to be fairly consistent with the current state of pharma demand forecasting practices described in section 1.1 based on literature and benchmarking reports.

1.3. Current Challenges and Innovations of Our Proposed Approach

There are a few major challenges in pharma demand forecasting. First, as discussed in sections 1.1 and 1.2, the current demand forecasting in the pharma industry focuses on using simple statistical methods with historical demand to extract future demand patterns. However, simple models cannot capture complex patterns. We know that demand is under the influences of many factors, usually hidden factors, in addition to historical trends. Some factors are across industries, such as the economic environment, while others are unique to the pharma industry. These include distinctive demand patterns due to special pharma situations (e.g., investment buying), the change of regulations (pharma industry is highly regulated), market competition between brands as well as between brand and generics, special contracts, and media effects (high public attention to pharma). Capturing such factors requires more sophisticated models built upon a large amount of data and domain knowledge of the industry. Indeed, the emerging concept of demand sensing, which focuses on identifying and including various factors affecting demand aside from historical demand, has attracted much attention (Chase 2013, Richard 2014). Simple time series models often ignore these hidden factors or assume that these factors manifest themselves in the *individual* drug demand time series so that the future demand for a pharma product is simply a function of its own previous demands.

At the same time, in the pharma industry, FFS and EDI have generated a significant amount of data. These data, however, have not, in general, been utilized for demand forecasting or production planning (Schwarz and Zhao 2011). While theoretical work from the operations management literature has shown the potential value of the downstream's demand and inventory data when being incorporated in the upstream's optimal production/inventory decisions (e.g., Cachon and Fisher 2000, Zhao et al. 2012), it remains to be shown empirically whether the aforementioned EDI data can provide any additional value in the upstream's demand forecasting that we focus on. The next immediate challenge is how to effectively mine this additional information and adequately capture the hidden factors mentioned above in improving demand forecast.

Machine learning is known to be an effective method to detect unknown patterns in structured and unstructured data. Recently, there have been more and more applications of machine learning in supply chain and operations management literature. It is well-known that effectively training and testing these machine learning models requires a large amount of data for accurate estimation of the model parameters (see, e.g., Halevy et al. 2009). Yet, this condition is not readily met in most forecasting context since there exists a temporal constraint, that is, old data from the distant past may have little value to the current prediction task (i.e., later confirmed in our data analysis). Hence, lack of data becomes another challenge that negatively impacts the performance of the more advanced models like machine learning in forecasting (Van Belle et al. 2021).

To address the above challenges, we propose the following novel demand forecasting framework. First, while limited to expand vertically to the distant past, we address the lack of data issue by looking horizontally across drugs and jointly training the machine learning model using the time series of different products (referred to as *cross-series training*). However, the properties of different product time series vary greatly. The model trained on global time series (that uses all products available) may perform poorly for certain drugs. Therefore, we develop three grouping schemes and further enhance the performance of the cross-series training by dividing the global time series into subgroups utilizing these grouping schemes to balance the tradeoff between sample size and sample quality. While two schemes use product demand and product characteristics based on pharma domain knowledge, the third scheme requires no knowledge of the data and uses a time series clustering algorithm to group the drugs. Note that our cross-series training with grouping is different from forecasting by product segmentation (cluster-then-predict) approach adopted in previous literature/industry practice, where no cross-series training is used and clustering is only used to select the forecasting methods for different product segments; in each cluster, a product is still forecasted using its own data. In our approach, time series data from other products in the same group are “borrowed” to forecast demands of these products together to resolve the lack of data issue when using the advanced models. Such innovation makes possible the high performance of the advanced models. Finally, in our new forecasting framework, we also include two non-demand features in training the forecasting model, that is, the downstream inventory information and the supply chain structure information. We explore whether and how much such information would help improve the model performance.

To execute our model framework, we work with two large datasets from two top drug manufacturers whose names are hidden for confidentiality. The first dataset includes weekly demand and inventory information extracted from EDI 852 over a 10-year period (2007–2017) for 133 unique products represented as 133 NDCs, that are sold to 28 trade partners (TPs) (e.g., wholesalers), through their respective 247 distribution centers (DCs). The second dataset includes similar data over a 7-year period (2011–2017) with 112 unique products from five TPs through their respective 73 DCs. We use the first dataset for our analysis throughout the paper; and run through the analysis on the second dataset in section 7 to validate our important findings. The numerous design considerations pertaining to our framework (e.g., different machine learning algorithms, various grouping schemes, different levels of aggregation of the data based on supply chain structure, and different number of time lags of historical demand and inventory data to use) result in an extensive set of numerical experiments. In addition to the numerical results which show superior performance of our proposed forecasting framework, we also provide potential reasons and conduct corresponding analysis to explain the effectiveness of our method over other approaches. Finally, we conducted post-hoc interviews to confirm the applicability and identify potential areas of attention in implementing our forecasting method in the pharma industry.

Our paper makes the following contributions. First, we propose a novel demand forecasting model framework for pharma products with the following new aspects: (a) Cross-series training to resolve the lack of data issue and further enhanced performance by various grouping schemes to balance the tradeoff between sample size and sample quality; (b) Including two key non-demand features in the demand forecasting framework, that is, downstream inventory levels and supply chain structure information, and designing how to effectively include these features; (c) Identifying that recurrent neural network (RNN) works best with the cross-series learning framework and providing potential explanation of its superior performance using domain knowledge and numerical analysis. Second, using two unique and large datasets with hundreds of NDCs, we validate the superior performance of our proposed model framework and provide important empirical evidence of the value of the downstream inventory information, which has been discussed in theoretical operations literature (e.g., Cachon and Fisher 2000, Zhao et al. 2012). Third, our cross-series forecasting model framework with grouping schemes and non-demand features demonstrates the value of machine learning in demand forecasting and can be potentially applied to other industries.

The remainder of the paper is organized as follows. In section 2, we review the literature and position our paper among the related works. In section 3, we present the research setting with a description of data and identify the new non-demand features to be included in the model. In section 4, we develop the cross-series forecasting framework, describe the grouping schemes, introduce benchmarks, and discuss the suitable forecasting models to be used and the implementation details to be used with the data. In section 5, we report our results on the benefit of cross-series training, the benefit of grouping, the value of downstream inventory information, and the value of supply chain structure information. We also include robustness checks in terms of models used, forecasting horizons, and benefits of our model framework on inventory performance. In section 6, we provide possible explanation of the evident effectiveness of the best performing model based on domain knowledge and additional numerical analysis. Section 7 validates our findings on a second dataset, which provides some evidence for generalizability of our results. We discuss the implications of our post-hoc interviews and conclude the paper in Section 8 with a summary of important insights. More detailed results that cannot be presented in the paper due to space limitations are provided in Appendix.

2. Literature Review

In this section, we review demand forecasting literature, particularly the interface between machine learning and forecasting in the operations management literature, with a special focus on the recent research motivated by real industry problems using data-driven approach.

Due to its crucial role in production and inventory control (Gardner 1990), demand forecasting has been extensively studied in the past decades (e.g., Boone et al. 2018, Brown and Meyer 1961, Kurawarwala and Matsuo 1996, Sastri 1985, Schmittlein et al. 1990, Winters 1960). Practical considerations such as collaborative forecasting partnerships between retailers and manufacturers (Aviv 2007), performance of hierarchical forecasting at different levels of aggregation in the supply chain (Kremer et al. 2015), and combining forecasts from multiple models (Grushka-Cockayne et al. 2016) have also been studied. Most of the studies focus on traditional time series methodologies.

Recent years have seen a great development of machine learning applications across many disciplines due to their remarkable abilities to capture hidden patterns. In forecasting domain, Hill et al. (1996) successfully applied neural network models to time series and achieved much better performance as compared to that from traditional statistical forecasting

methods. Recently, the similar type of research has appeared in the operations management literature. While limited in numbers, there is an upward trend in this data-driven research. For example, Carbonneau et al. (2008) studied the effectiveness of both machine learning and traditional forecasting methods on simulated and real sales data. They reported that traditional methods work well on simulated data, but are less competitive against more advanced machine learning techniques on real data. A more recent paper leveraging the power of machine learning in demand forecasting is Cui et al. (2018), which uses both the operational data (sales and marketing data) and the social media information to improve the accuracy of daily sales forecasts.

Notice that all demand forecasting models we have discussed so far predict the future demands for a product using its own data, where there may be problems in model parameters estimation when the amount of data is limited. When using the more advanced machine learning models like RNN, proper training calls for vast amount of data. While RNN is known for its use on text, images, and video data, which are abundant (see Karpathy 2015), lack of data is fairly common in pharma demand forecasting because, as mentioned, distant data in the past may not provide much information as the market condition constantly changes. Therefore, to address this problem as well as to help find the common hidden factors, we propose to leverage information from other products. This idea of cross-learning from other products has been used in new product forecasting in which future sales are predicted from a set of features such as price, brand, style based on comparable products (see Baardman et al. 2018, Ferreira et al. 2015.) However, in those settings, the dataset is limited to only similar products and the data are not time series as mentioned. In dealing with the more challenging time series data, which and how many products to cross-train together, that is, balancing between sample size and sample quality, requires careful design and analysis.

Vector autoregression (VAR) is a well-established econometric method for learning from related time series when making forecasts (Sims 1980). However, VAR is very different from what we propose and is not applicable to our problem due to overfitting. Specifically, VAR allows all variables to interact linearly with their own and each other's current and past values (lags). Therefore, when there are many time series involved with many lags, as in our problem for cross-series learning, the number of VAR coefficients to be estimated is very large, leading to severe overfitting, and larger forecast errors even with regularization. As confirmed by our numerical results, VAR performance is significantly worse than even those from the baselines (see Section 5.1). Hyndman and

Athanasopoulos (2018) suggests using VAR only for a small number of time series which are known to be correlated with each other. In addition, VAR only captures linear relationships. As confirmed by our results, there exists significant nonlinear relationships in our data.

Recent papers have proposed using non-demand information, such as social media data, to improve the performance of forecasting, for example, Liu et al. (2016), Cui et al. (2018), Boone et al. (2018), Lau et al. (2018), and See-To and Ngai (2018). For further details, we refer the readers to Choi et al. (2018) for a review. In our paper, we identify other suitable non-demand features and use them to predict demand across products. In particular, our idea of leveraging inventory and supply chain structure information within the cross-series training framework is based on existing theoretical operations management literature (Cachon and Olivares 2010). Indeed, the benefit of using downstream's inventory data to enhance the upstream's optimal production/inventory decisions has been studied in the analytical models (e.g., Cachon and Fisher 2000, Zhao et al. 2012), but has not been well-studied empirically in the academic literature. Furthermore, among the possible drivers for different inventory levels in the system, the impact of distribution network structure is significant (Olivares and Cachon 2009). In our paper, we empirically explore the value of downstream inventory information and supply chain structure (DC-level data) in demand forecasting using real data in a cross-series training setting.

There is limited and relatively primitive academic literature on demand forecasting for pharma products compared to other industries (e.g., tourism, energy,

etc.). Besides the monograph by Cook (2016), prior studies on pharma demand forecasting are summarized and compared to our study in Table 1. The first row specifies which tier's demand is forecasted in the respective research because different data might be available and used for that specific tier. The pharma supply chain has multiple tiers, including manufacturers, TPs/wholesaler, DCs, and point-of-care.

Table 1 also compares our paper to other research works in terms of benchmarks, models proposed, data utilized, metrics used, and forecasting horizon. Their benchmarks (moving average, basic exponential smoothing, and regressions) are consistent with those reported in the industry practices described in section 1.1. In terms of data used, there seems to be a lack of distinction between demand and sales in the pharma forecasting literature (with sales being the right-censored demand by available inventory). All papers in the table except ours used historical sales (instead of actual historical demand) to predict demand since typically sales instead of demand is observed. Fortunately, our datasets provide the original orders from the wholesalers to the manufacturers, hence we are able to capture the actual demand and use it to predict future demand. Besides historical sales, little other information is used for forecasting. The ones that used other information include Kim et al. (2015) that used customers' response collected in blog documents to help improve drug demand forecasts for a retailer, Merkurieva et al. (2019) used discounted prices in a causal forecasting model to forecast demand at a distributor and Van Belle et al. (2021) that used sell-through data (product-related data available to the wholesaler) to forecast demand at a manufacturer. In measuring forecasting accuracy,

Table 1 Comparison of Pharmaceutical Demand/Sale Forecasting Papers in the Literature

Tier	Kim et al. (2015) Retailer	Merkuryeva et al. (2019) Distributor	Nikolopoulos et al. (2016) Point-of-care	Van Belle et al. (2020) Manufacturer	Khalil Zadeh et al. (2014) Distributor	Our Model Manufacturer
Benchmark	AR	Moving average Linear regression	Diffusion models ARIMA Exp. smoothing Linear regression	ARIMA ETS MLP, SVR, RF	ARIMA	Exp. state-space models Moving Average Linear Regression
Proposed model	VARX	Symbolic regression	None	Linear Regression	Graph-based analysis and ANN	Clustering and RNN
Utilized data	Historical sales Social network	Historical sales Price	Historical sales (prescription data)	Historical sales Sell-through data	Historical sales	Historical demand Downstream inv. Supply chain info.
Forecast frequency	Monthly	Weekly	Yearly	Weekly	Monthly	Weekly
Forecasting horizon	1 month ahead	1–weeks ahead	1–5 years ahead	1–5 weeks ahead	1 month ahead	1–2 months ahead
Metrics	Prediction error rate	R^2 MAD	R^2 ME, MAE, MSE	RMSE	R^2 MSE, MAE	NME, NMAE, NMSE
Number of NDCs	4	1	11	50	217, but only 21 were analyzed	133 (First dataset) 112 (Second dataset)
Time series cross-validation	No	No	No	Yes	No	Yes

cross-validation is the standard approach and should be used to ensure the generalizability of the forecasting model to new data. The basic cross-validation procedure involves separating the data into training and test sets, where the training data are used to estimate a forecasting model's parameter, and the test set ($\approx 20\%$ of the observations) is used to evaluate its accuracy (see Makridakis et al. 2008). However, this reporting standard is seldom used in the pharma forecasting literature. Furthermore, in practice, for accuracy evaluation of forecasting methods, it is recommended to use the more sophisticated time series cross-validation, where there are a series of test sets and the forecasting accuracy is computed by averaging over these test sets (Hyndman and Athanasopoulos 2018). However, most of these papers did not implement time series cross-validation and only used a small number of drugs. None of the papers considered cross-series training nor clustering for forecasting—a product was only forecasted using its own data. It is worth noting that Van Belle et al. (2020) considers basic machine learning models and their results show that the machine learning models do not perform well (no better than linear regression models). This shows the value of our proposed cross-series training with clustering in making machine learning forecast models reach their peak performance for forecasting.

2.1. Pharma Distribution Network

The pharma supply chain is a complex system in which drugs are delivered from manufacturers to patients through the distribution networks. An oversimplified pharma distribution network highlighting what is pertaining to our work is depicted in Figure 1. In particular, the flow of pharma products originates from manufacturers to multiple TPs, who then

distribute these products via their network of DCs, to downstream point-of-cares (POCs), such as clinics, hospitals, or retail pharmacies. Trade partners can be categorized as “traditional wholesalers” and “specialty distributors.” The former typically have large networks, carry a large variety of drugs, and more often distribute to hospitals, retail pharmacies, and homecare providers, while the latter typically have more controlled networks, specialize in specialty drugs and more often distribute to physician offices, clinics, and independent specialty pharmacies (Xu et al. 2018). Regardless of its type, each TP has its own network of DCs, through which POCs receive the drugs.

As mentioned, we obtained from our industry collaborator two large datasets of EDI 852 of two top pharma manufacturers. The datasets consist of supply chain channel data on all pharma products that the respective manufacturer has at the time of data collection. Each product is determined by a unique, three-segment identifier, called NDC. We will focus on analyzing the first dataset and then use the second dataset to confirm our insights in section 6. We next present more detailed description of the first dataset.

2.2. Data

Our data include all transactions (quantity sold, TP, distributor, and DC's inventory level) between a drug manufacturer and its TPs' DCs, collected weekly for 133 unique NDCs from July 2007 to August 2017. Quantity sold and inventory level for each transaction are measured in pack unit (PU) or extended unit (EU) for each NDC. We choose to use extended unit (i.e., one capsule or tablet for solid dosage forms or one milliliter for liquid drug products) because this measurement helps to normalize different package sizes, which allows comparisons across NDCs.

Figure 2 illustrates a typical time series of the weekly order quantities of a drug from 2007 to 2017. The figure shows many demand spikes occurring throughout the years. As will be discussed later, existence of such spikes is due to the prevalent investment buying in the pharma industry. Hence, being able to capture such spikes is important to the performance of forecasting models.

In addition to the past demand information, we also collect the following product information (ATC code) which is used in one of the grouping schemes and use the following non-demand features in our cross-series forecasting model.

Product Information (ATC code). Aside from the EDI 852 data, we also collected additional information of each drug's ATC code from public databases. The ATC code classifies drugs into distinct groups based on their chemical, pharmacological, and therapeutic properties (see World Health Organization). The ATC

Figure 1 Illustration of a Pharma Distribution Network

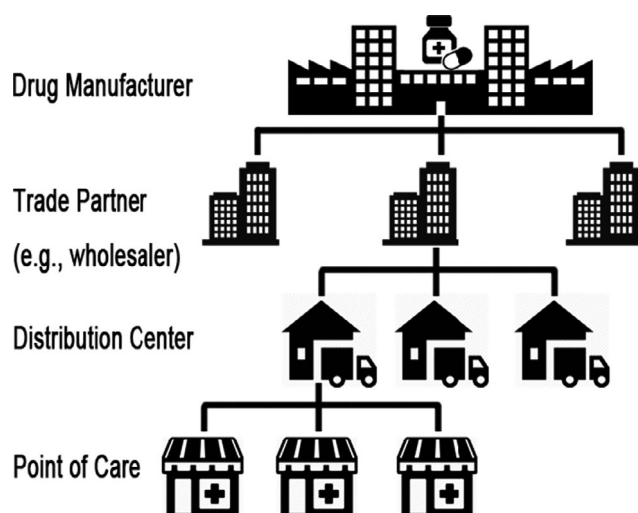
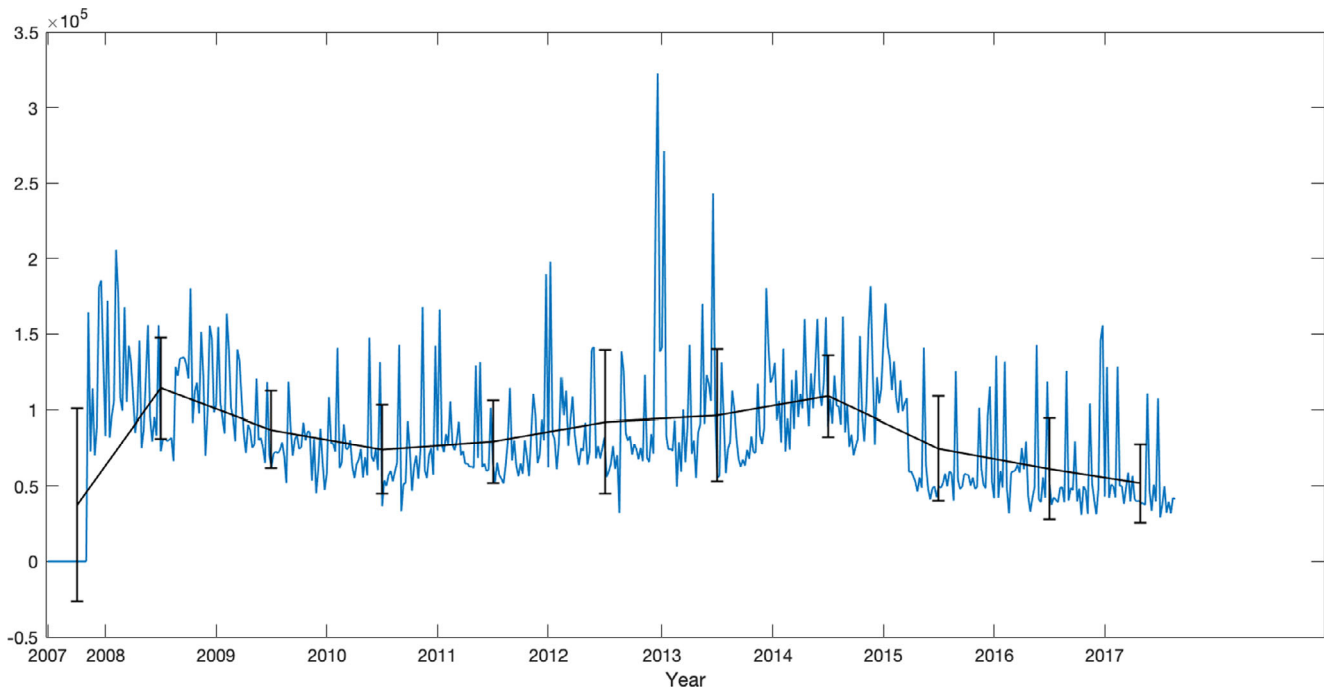


Figure 2 Illustration of a Drug's Order Quantities Over Time [Color figure can be viewed at wileyonlinelibrary.com]



code has five levels of progressively more detailed information. We use the first level, also known as the main anatomical group, which divides drugs into 14 main groups. Specifically, we first extract the non-proprietary name for each NDC from the National Drug Code Directory, and then search for the respective ATC code from the database at the World Health Organization's collaborating center for drug statistic methodology. Since the ATC code is widely used in pharma industry to classify drugs based on their characteristics, we later use the ATC code as one of the grouping schemes in our cross-series forecasting models. We also obtain the wholesale acquisition cost (WAC), which is the list price for each NDC. Real prices typically include discounts and rebates, but the WAC is a widely accepted reference price (Zhao et al. 2012).

Inventory Information. The fee-for-service (FFS) contracts prevalent in the pharma industry require downstream wholesalers to share inventory information with manufacturers via the EDI 852 interface. We collect weekly historical order quantity and inventory at the DCs to help predict future order quantity. This approach derives from the theoretical OM literature that a supplier's order quantity decisions can be improved when demand and inventory data are shared within the supply chain (see Cachon and Fisher 2000, Zhao et al. 2012, and the references therein).

Supply Chain Structure Information. Since the structure of our two datasets is the same, our

introduction of the data and analysis/results (including all the tables) will be based on the first dataset. In section 7, we will introduce and run through the main analysis on the second dataset to validate important findings obtained from the first dataset. The 133 NDCs are sold to 28 TP (including the top 3 wholesale distributors which represent 85% of the total annual US sales) through their respective 247 DCs. We later tested at the TP level to see whether such supply chain information benefits the forecasts. On average, each TP purchases 64 NDCs from the drug manufacturer, while each DC receives roughly 34 NDCs every week. We provide the descriptive statistics of our dataset below Table 2.

3. Model Development

Exploratory analysis (such as the demand spikes in Figure 2) suggests that different drugs may have

Table 2 Descriptive Statistics of Our Dataset

	Mean	Std.	Min	Median	Max
# of dist. centers per trade partner	9	15	1	2	50
# of NDCs per trade partner	64	41	2	68	127
# of Obser. per trade partner	123,031	272,775	240	25,749	970,393
Avg. Order Qty per trade partner	21,616	63,999	39	903	285,866

some similar demand patterns. Therefore, we design a new model framework to leverage data across different drugs to train the model to capture these patterns. In what follows, we develop our model framework. Section 3.1 introduces notations for our cross-series training models. Section 3.2 proposes grouping schemes to further optimize the cross-series forecast model. Section 3.3 discusses the choice of model (VAR and various machine learning models) and proposes to use recurrent neural network model. Section 3.4 presents baseline models and section 3.5 provides the implementation details.

3.1. Cross-series Training

The objective of our model is to predict the *total* demand at the manufacturer from all DCs for each of the 133 NDCs. We first introduce the following important notations for the forecasting models.

- \mathcal{I} : The set of all NDCs in the dataset, that is, $\mathcal{I} = \{1, \dots, I\}$.
- \mathcal{J} : The set of all DCs in the dataset, that is, $\mathcal{J} = \{1, \dots, J\}$.
- \mathcal{K} : The set of all TPs in the dataset, that is, $\mathcal{K} = \{1, \dots, K\}$. Furthermore, the set of all DCs from TP k will be denoted as $\mathcal{J}_k, k \in \mathcal{K}$, and we have $\sum_{k \in \mathcal{K}} \mathcal{J}_k = \mathcal{J}$.
- $x_{i,j,t}$: The order quantity of drug $i \in \mathcal{I}$ from DC $j \in \mathcal{J}$ to manufacturer at time t .
- $y_{i,j,t}$: The inventory of drug $i \in \mathcal{I}$ in DC $j \in \mathcal{J}$ at time t .
- X_{it} : The cumulative order quantity of drug $i \in \mathcal{I}$ from all DCs to manufacturer at time t , that is, $X_{it} = \sum_{j \in \mathcal{J}} x_{i,j,t}$.
- Y_{it} : The cumulative inventory of drug $i \in \mathcal{I}$ across all DCs at time t , that is, $Y_{it} = \sum_{j \in \mathcal{J}} y_{i,j,t}$.
- p, q : The number of time period lags for order quantity and inventory, respectively, that is, the most recent p weeks of order quantity and q weeks of inventory will be used in the forecasting model.
- h : The forecast horizon, measured in weeks.
- $\hat{X}_{i,t+h}$: The predicted cumulative order quantity for drug $i \in \mathcal{I}$ at time $t+h$, made at time t , that is, predict at time t the demand in h periods in the future.

When including the past p weeks of order quantities of all drugs in set \mathcal{I} and q weeks of the

corresponding downstream inventory information, we design the global cross-series training model (cross-training using all drugs available) as follows:

$$\hat{X}_{i,t+h} = f(X_{i,t}, \dots, X_{i,t-p+1}, Y_{i,t}, \dots, Y_{i,t-q+1}), \forall i \in \mathcal{I} \quad (1)$$

where the forecast of drug i is obtained from drug i 's features only, but the mapping f is learned from data of all drugs. While learning across drugs greatly increases the sample size hence provides a solution to the lack of data issue, we must also consider the tradeoff between sample size and sample quality for advanced models: learning and training across more “similar” drugs would potentially bring better sample quality. Therefore, a more advanced cross-series training model would first group drugs according to some schemes, and build a forecasting model for these drugs cross-trained within each group. This idea will be further explored below.

3.2. Grouping Schemes

In this subsection, we propose three grouping schemes based on different rationales including domain knowledge of the pharma industry.

Grouping by Demand Volume and Volatility This scheme is to group the drugs based on their demand volume (average order quantity per NDC) and demand volatility (measured by coefficient of variation—CV). Industry has also used these two criteria to group drugs in various situations (e.g., for product segmentation). Using the medians of demand volume and volatility, we partition all NDCs into four non-overlapping groups, namely, high volume-low volatility (HL), high volume-high volatility (HH), low volume-low volatility (LL), and low volume-high volatility (LH). Summary statistics regarding order quantities (EU) of all four groups are provided in Table 3. Since drugs are not evenly distributed in the volume/volatility space, the four groups are not necessarily equal in sizes.

We build one model for each of the four group. Each model is trained using only the information from the drugs belonging to that specific group. The model is shown in Equation (2) as follows:

$$\hat{X}_{i,t+h}^{(g)} = f^{VV}(X_{i,t}^{(g)}, X_{i,t-1}^{(g)}, \dots, X_{i,t-p+1}^{(g)}, Y_{i,t}^{(g)}, Y_{i,t-1}^{(g)}, \dots, Y_{i,t-q+1}^{(g)}), g \in \{\text{HL}, \text{HH}, \text{LL}, \text{LH}\}, \quad (2)$$

Table 3 Order Quantities in Four Groups Based on Volume/Volatility

Group name	Mean	Median	C.V.	Min	Max	# of NDCs	# of Obser.
HL	393,084	29,160	3.13	0	25,086,100	55	2,434,628
HH	179,289	3,784.04	3.22	0	9,064,080	11	167,734
LL	3,124	1,200	1.91	0	145,920	11	354,526
LH	1,704	0	5.18	0	514,967	56	487,987

where $X_{i,t}^{(g)}$ and $Y_{i,t}^{(g)}$ stand for the order quantity and inventory information of drug i in group g at time t , respectively, that our model will include. The forecast of drug i is still obtained using drug i 's features but the mapping f^{VV} (with “VV” for volume and volatility) is now learned using the data of the drugs in the same demand volume and volatility group.

Grouping by ATC code This scheme groups the drugs based on ATC code. Recall that drugs in the same ATC code have similar therapeutic, pharmacological, and chemical properties. Thus, we can think of ATC Code as a product-based criterion while volume and volatility as a demand-based criterion to group the drugs. The drugs in our dataset belong to six major ATC code groups, namely, A, B, C, G, J, and N. Summary statistics of order quantities from NDCs in each of the six ATC groups are provided in Table 4. Observe that some ATC groups have more NDCs and observations than the others. The cross-series training model by ATC code can be expressed as:

$$\hat{X}_{i,t+h}^{(\phi)} = f^A(X_{i,t}^{(\phi)}, X_{i,t-1}^{(\phi)}, \dots, X_{i,t-p+1}^{(\phi)}, Y_{i,t}^{(\phi)}, Y_{i,t-1}^{(\phi)}, \dots, Y_{i,t-q+1}^{(\phi)}),$$

$$\phi \in \{A, B, C, G, J, N\}, \quad (3)$$

where $X_{i,t}^{(\phi)}$ and $Y_{i,t}^{(\phi)}$ stand for the order quantity and inventory information of drug i at time t , respectively, with ATC code. Now, the mapping f^A (with “A” for ATC codes) is learned using the data of drugs in the same ATC code group.

Grouping by clustering algorithm So far, we have proposed a demand-based approach recognized by industry and a product-based domain knowledge approach to group NDCs for cross-series training. If such domain knowledge is not readily available, we propose to use clustering algorithms. Here, we adopt an unsupervised machine learning technique called

K-means clustering on the drugs historical demand to group the NDCs into K different clusters so that drugs in the same group are more similar to each other than to those in other groups. In particular, we adopt the dynamic time warping (DTW) algorithm (Berndt and Clifford 1994) to measure the similarity between two demand time series. DTW essentially finds the optimal alignment between two drugs' demand time series, and thereby measures their shape similarity accordingly. Due to the magnitude differences in the order quantities of different drugs, we normalize each time series (by subtracting it by the respective mean and then dividing by the standard deviation) prior to DTW computation.

Selecting the optimal number of clusters of drugs, K , requires balancing between the clustering quality and the number of observations in each cluster. To do so, we use the Davies–Bouldin index (DBI) as the clustering evaluation metric, defined as

$$DBI = \frac{1}{K} \sum_{i=1}^K \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right), \quad (4)$$

where σ_i and σ_j are the intra-cluster distances of clusters c_i and c_j , respectively. Note that an intra-cluster distance measures the average DTW distances of all pairs of drugs' demand time series within the same cluster. In contrast, the inter-cluster distance $d(c_i, c_j)$ measures the distance between the two clusters c_i and c_j , that is, the average DTW distances between all pairs of drugs' demand time series in which one is selected from cluster c_i and the other is selected from c_j . The number of clusters K is selected so that it has the lowest value of DBI , which indicates better quality for the respective clustering performance. For our data, our approach results in five clusters. Table 5 provides the

Table 4 Order Quantities in ATC Code Groups

Group name	Mean	Median	C.V.	Min	Max	# of NDCs	# of Obser.
A	72,223	7,200	2.66	0	2,930,385	44	568,582
B	416,259	4,881	2.07	0	6,558,360	12	344,521
C	137,728	6,443	2.31	0	4,741,848	18	840,984
G	2,351,915	130,375	2.07	0	25,086,100	7	234,391
J	31,472	3,600	2.59	0	857,400	26	64,229
N	88,855	2,100	4.08	0	9,064,080	17	768,342

Table 5 Order Quantities in the Generated Clusters Under Grouping by Clustering Algorithm

Cluster index	Mean	Median	C.V.	Min	Max	# of NDCs	# of Obser.
1	225,470	4,000	6.42	0	25,086,100	20	851,298
2	961	0	11.19	0	514,967	18	86,908
3	161,985	10,274	2.52	0	5,469,480	32	1,137,336
4	3,744	46	3.91	0	453,570	27	392,932
5	529,288	61,560	2.17	0	13,720,620	36	976,401

summary statistics of order quantities of drugs in each cluster.

The cross-series training model by clustering can be expressed as

$$\hat{X}_{i,t+h}^{(\kappa)} = f^C(X_{i,t}^{(\kappa)}, X_{i,t-1}^{(\kappa)}, \dots, X_{i,t-p+1}^{(\kappa)}, Y_{i,t}^{(\kappa)}, Y_{i,t-1}^{(\kappa)}, \dots, Y_{i,t-q+1}^{(\kappa)}),$$

$$\kappa \in \{1, 2, \dots, K\}, \quad (5)$$

where $X_{i,t}^{(\kappa)}$ and $Y_{i,t}^{(\kappa)}$ stand for the order quantity and inventory information of drug i at time t in cluster κ , respectively. The mapping f^C (with “C” for clustering) is learned using the data of drugs in the same cluster.

3.3. Machine Learning (ML) Models

With the groups obtained from the preceding schemes, we next develop the learning model to predict future drugs’ demand in each group. As discussed, vector auto-regression (VAR), a method used for multiple time series forecasting, is not suitable in our problem because in practice, it is recommended to use on a small number of time series that are correlated with each other (Hyndman and Athanasopoulos 2018). This practical guideline clearly limits the applicability of VAR to our dataset. Indeed, if we build a model to predict demand for 10 drugs using 8 lags (2 months of data), there are 81 coefficients per VAR equation, giving a total of 810 coefficients to be estimated. Furthermore, as noted before, VAR only captures linear relationships.

We propose to use a ML algorithm to forecast drugs’ demand in a group. Since the data exhibit nonlinear patterns, we focus the discussion on nonlinear methods, while linear methods will be used as one of the baseline models (see section 3.4). Among a plethora of nonlinear methods, there are three widely used classes in the literature: support vector regression (SVR), random forest (RF), and neural networks. A basic structure for neural networks is the fully connected neural network (FC). For a short description of these methods, see Appendix A. To learn demand patterns such as the demand spikes with SVR and RF, the timing of the spikes needs to be treated as categorical variables. However, mixed data of numerical and categorical data usually hurt the performance of both SVM and RF (Ke et al. 2019). In contrast, a *recurrent neural networks* (RNN), a special type of neural network with memory cells to enable tracking of short- and long-term dependencies in the input, can potentially capture the hidden patterns well in our data. Note that RNN is particularly suitable for processing sequential data, for example, time series data (Hochreiter and Schmidhuber 1997), hence making it a great candidate. See Appendix A (figure 8) for a schematic of the RNN structure.

Each computation unit of RNN is called a cell. Different RNNs may have different cell structures (LSTM, GRU, etc.), but their most crucial feature is that each cell’s outputs have connections backward. Therefore, at each time step, the cell receives inputs as well as its own output from the previous time step. As a result, a cell’s outputs are influenced by not only the most recent input, but also the entire history of past inputs. Furthermore, each cell implements a series of gates in which information can be passed on or forgotten. This particular architecture makes it possible for RNN to explore the temporal dynamic information from a time series (Karpathy 2015). In our demand forecasting context, the input of the cell at period t is the order quantity and inventory at period t as well as the output of the cell at period $t-1$. The output is the prediction of order quantity at time t .

3.4. Baseline/benchmark Models

While we do not have the manufacturer’s internal forecast (it is missing from the data shared with us), we fortunately obtained the forecasts provided by the downstream distributors to the manufacturer from January 2012 to March 2016 for the first dataset. Specifically, since downstream distributors are closer to the customers, they have more accurate demand forecasts (confirmed in our post-hoc interviews as well) than the manufacturer. The aggregate forecast from all the downstream distributors can readily serve as the company’s internal forecast. Such aggregate forecast also took into considerations of all human adjustments, if any, because they represent the distributors’ best forecast (orders to the manufacturer deviating from these forecasts require explanations).

Since one may argue that the internal forecast can only represent the forecasting capability of a specific company, we also select another baseline representing the state-of-the-art industry practice, based on the reported benchmarks in the literature (Table 1), the current practice in the industry (section 1.1), and the results from our questionnaire about the current demand forecasting practices (section 1.2). These sources converge to the forecasting methods of moving average (MA), linear regression (LR), and basic exponential smoothing (ES). Furthermore, to make sure that the state-of-the-art baseline models are used, we improve the aforementioned simple ES model by utilizing the innovative exponential state space model (ETS) developed by Hyndman and Athanasopoulos (2018), which include 30 separate models. Note that these sophisticated models were developed to automatically forecast demands for thousands of drugs. Each model has an observation equation and transition equations, one for each state (level, trend, seasonal) with additive or multiplicative errors. The ETS

forecast package in R, implementing the above models, automatically chooses the most appropriate ETS method as well as the optimal parameters for the forecasting task. The ETS model represents state-of-the-art industry practice and is expected to exceed the performance of most pharma companies' software. The ETS model empirically provides slightly more accurate forecasts than ARIMA (Hyndman and Athanasopoulos 2018), so we do not use ARIMA as a benchmark. To be inclusive and conservative, we include all three models discussed above (i.e., MA, LR, ETS) and choose the best of the three as our baseline models. Note that all baseline models do not implement our proposed cross-series training idea.

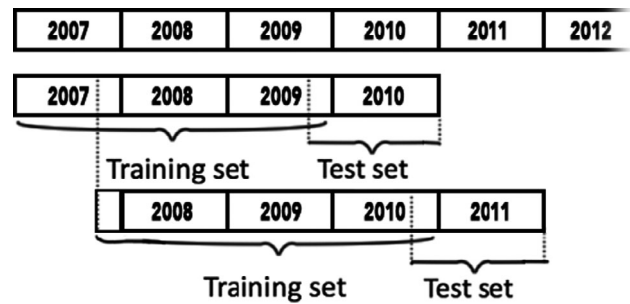
While the aforementioned company internal forecasts and the baseline models help us understand the value of our proposed cross-series forecasting approach compared to the current demand forecasting practices in the pharma industry, we also validate the performance and robustness of RNN, compared to other machine learning algorithms, that is, LR, SVR, and RF, all implemented with cross-series training. Hence, we report the performances of these models as well for the completeness of the study, as robustness checks of our model choices (section 5.2).

3.5. Implementation Details

Time lags. To estimate the future order quantity, we use information from past weeks (e.g., the previous p weeks of historical demand to the manufacturer and/or q weeks of inventory information). We test different time lags $p, q = 1, 2, \dots$ and choose the best performing combination of p and q values for each model via cross-validation. For instance, whenever a model uses inventory information, we test all combinations of time lags for order quantity ($p = 1, 2, \dots, 10$ weeks) and inventory information ($q = 0, 1, \dots, 10$ weeks) and report the best one. Thus, in total, we test 110 combinations for each model.

Time series Cross-validation. To ensure model robustness across different time periods, its performance is measured using the time series cross-validation technique (Hyndman and Athanasopoulos 2018). In particular, we use three consecutive years as the training set and the fourth year as the test set. Within each training set, we perform cross-validation by using a rolling forecasting origin. That is, we use the observations in a rolling window to train the model and the observation outside the window to validate. In each cross-validation round, we compute the accuracy metrics and select the model parameters that achieve the best performance on the test set. This time series cross-validation procedure for model evaluation is suited to time series data since serial correlation and potential non-stationarity may exist (see, e.g., Bergmeir et al. 2018). When available, we also include

Figure 3 Training Set and Test Set



p weeks of data before the start of the training set time to ensure all the training samples will have the right inputs. For example, suppose the training set is from 2008 to 2010. We include in the training set the last p weeks' order quantities at the end of 2007 to predict the order quantity for the first week in 2008. Generation of the training and test set is demonstrated in Figure 3.

Using the processing methods described above, we create a data matrix for each training and test set. Since we have records of I drugs in a given 3 years, and there are k weeks in these 3 years, we want to use the past p weeks of order quantity and q weeks of inventory data to predict the order quantity for the next week. As a result, we obtain an $I k \times (p + q + 1)$ matrix. The first column of this matrix contains the order quantity to be predicted, labeled as the response variable.

Parameter Tuning. In RNN, there are several hyperparameters to be tuned, for example, the number of neurons, the number of layers, learning rate, batch size, etc. We also need to decide the suitable number of lags used in the model. This can be done using the standard grid-search and cross-validation methods outlined above. At the end of this process, we will obtain the optimal set of hyperparameters to forecast demand. In particular, p and q can vary from 1 to 10, the number of neurons ranging from 100 to 1000 in each cell, the initial learning rate ranging from $1e-4$ to $1e-1$, the batch size ranging from 2^4 to 2^6 , and the number of epochs ranging from 100 to 300. Regarding the optimizer for RNN, we use the adaptive moment estimation (Adam) optimization algorithm (known to work well in most practical applications), with the exponential decay rate for the first moment estimates ranging from 0.9 to 0.999, and the second moment estimate set at the default value of 0.999. The L_2 regularization parameter ranges from 10^{-3} to 10^{-1} .

Evaluation Metrics. Note that the popular mean absolute percentage error (MAPE) cannot be used for our dataset due to the possibility of zero demands. Thus, we use normalized mean square error (NMSE),

normalized mean absolute error (NMAE), and bias to evaluate our models' performance. Normalization is used to facilitate comparison across different models and different NDCs. For a thorough review of forecast accuracy measures, see Hyndman and Koehler (2006). In reporting the forecasting accuracy, for a given forecasting horizon (e.g., 1 week), we evaluate the forecasting accuracy for each group by taking the average across all drugs in that group. For most of the models we report the results for the forecasting horizon of 1 week, with the robustness check of the forecasting horizon varying from 1 to 8 weeks in Section 5.6.

4. Results and Discussion

In this section, we report results of our proposed framework in terms of the benefit of cross-series forecasting, the benefit of grouping in cross-series training, the value of inventory information, the value of supply chain structure information, and their implications in the pharma demand forecasting context. We then conclude with a few robustness checks.

4.1. No Cross-series Models: Internal Forecasts and Baseline

Recall that the internal forecast reflects the forecasting capability of the company and the baseline models represent the state-of-the-art forecasting approach used in the pharma industry. Neither of them uses the cross-series information. Table 6 presents the performance measures, along with the respective 95% confidence intervals, of the company's internal forecasts in comparison with the baseline models, that is, MA, ETS, and LR.

The numerical results in Table 6 clearly show that the baseline models outperform the company's internal forecasts across all different accuracy measures,

which confirms our conservative choice of the baseline models. Therefore, in the following sections, we will only compare the performance of our cross-series models against the baseline (the best of MA, ETS, and LR without cross-series training).

4.2. Benefit of Cross-series Forecasting

Table 7 presents the performance measures of the global cross-series training models in comparison with the baseline models (none of the baseline models use cross-series training).

The numerical results show that cross-series training is beneficial across all performance metrics, and RNN performs the best. In general, cross-series training with more advanced ML models gives better results than cross-series training models with just simple models like LR, or even SVR. RNN's superior performance to that of the linear models indicates that RNN can pick up some nonlinear patterns. The other nonlinear models do not perform well on the dataset, possibly because RNN is better suited for time series data (Hochreiter and Schmidhuber 1997). Indeed, we show in section 7 that RNN is the only method that can efficiently capture the demand spikes. As expected, VAR's performance is inferior to even those of the baseline models; hence, we will not further consider it in the remaining analysis.

Our results show that the global cross-series forecasting increases the sample size and brings improvements in the forecast accuracy. We next see how grouping the drugs and building a separate model for each group (i.e., decrease the sample size for each model while increasing the sample quality) could further improve the forecasting performances using ML models.

4.3. Benefit of Grouping Drugs in Cross-series Training

In reporting the results in the remainder of the paper, the suffix "4" indicates grouping by the four volume/volatility categories, "ATC" by ATC code, and "DTW" by clustering—no suffix indicates global cross-series training. To facilitate comparison between competing methods, we only report the performance measures of the best performing baseline models and the percentage improvement of the cross-series training models over those (see Appendix B for the raw values

Table 6 Forecasting Bias and Accuracy Measures of the Baseline Models vs. Internal Forecasts*

	MA	ETS	LR	Internal forecast
NMSE	1.58 ± 0.15	2.08 ± 0.18	1.63 ± 0.16	4.10 ± 0.24
NMAE	0.29 ± 0.05	0.34 ± 0.06	0.30 ± 0.05	0.45 ± 0.06
Bias	0.00 ± 0.01	0.02 ± 0.02	−0.04 ± 0.00	0.09 ± 0.04

*: Internal forecasts are obtained for Jan. 2012–Mar. 2016.

Table 7 Forecasting Bias and Accuracy Measures of the Global Cross-series Training Models

	Baselines (no cross-series)			Global cross-series training					
	MA	ETS	LR	VAR	LR	SVR	RF	FC	RNN
NMSE	1.71 ± 0.09	2.06 ± 0.15	1.67 ± 0.10	2.74 ± 0.23	1.58 ± 0.09	1.83 ± 0.11	1.60 ± 0.10	1.50 ± 0.09	0.98 ± 0.12
NMAE	0.31 ± 0.03	0.34 ± 0.04	0.30 ± 0.03	0.43 ± 0.08	0.30 ± 0.03	0.32 ± 0.05	0.29 ± 0.04	0.29 ± 0.03	0.25 ± 0.04
Bias	0.03 ± 0.02	−0.02 ± 0.03	−0.03 ± 0.02	−0.06 ± 0.06	−0.04 ± 0.02	0.03 ± 0.06	0.01 ± 0.04	0.02 ± 0.02	0.01 ± 0.04

and confidence intervals). Furthermore, in order to see the performance over different categories of drugs (e.g., high volume vs. low volume), we report the aggregate performance metrics computed over all drugs (All Drugs) and specific categories (HL, HH, LL, LH). We use a dash when a model is worse than the baseline. For easy exposition, we only report the performance of LR and RNN. See Appendix B for the performance of other ML models (all inferior to RNN for cross-series training).

Tables 8 and 9 show the benefit of grouping by demand volume/volatility and by ATC code, respectively. The tables show that the performance of RNN is significantly better than that of LR and in general, both LR and RNN benefit from grouping. For example, improvement in NMSE over the baseline for all drugs has increased from 5.4% to 7.2% for linear models, and from 41.3% to 46.7% for RNN when grouping by demand volume/volatility. Furthermore, only RNN achieves consistent improvement for all drug groups, especially for the low volume drugs (Table 8). For different ATC code groups, RNN has significant

improvement for most of the groups. Even for the ATC group with smaller sample size (group J), training within ATC groups shows improvement of 10.8%.

Table 10 reports and compares different models' performances under all three grouping schemes. While both LR and RNN show benefits from using any grouping schemes, RNN has superior performance to LR. In addition, grouping helps RNN to particularly improve the typically more difficult to forecast products (low-volume drugs), with the best performance seen in grouping by demand volume/variance, followed by grouping by clustering, then ATC code. On the other hand, the performances of different grouping schemes are generally comparable to each other for demand forecast for the high-volume products.

In summary, the reported results demonstrate that grouping further optimizes the cross-series training models. While grouping drugs by volume/volatility or by ATC codes has better interpretability, grouping by clustering, which does not require any knowledge of the data, also performs well.

Table 8 Baseline and Percentage Improvement Over Baseline of Cross-series Training Models with Grouping by Volume/Volatility

		Best baseline	LR	LR_4	RNN	RNN_4
NMSE	All drugs	1.67	5.4%	7.2%	41.3%	46.7%
	HL	1.05	3.8%	3.8%	39.0%	44.8%
	HH	1.44	16.7%	5.6%	44.4%	47.2%
	LL	2.69	9.3%	10.0%	–	29.4%
	LH	17.37	–	–	–	49.8%
NMAE	All Drugs	0.30	0.0%	0.0%	16.7%	16.7%
	HL	0.29	3.4%	3.4%	10.3%	17.2%
	HH	0.37	10.8%	2.7%	27.0%	24.3%
	LL	0.69	4.3%	5.8%	–	18.8%
	LH	0.81	–	–	–	19.8%

Table 9 Baseline and Percentage Improvement Over Baseline of Cross-series Training Models with Grouping by ATC Code

		Best baseline	LR	LR_ATC	RNN	RNN_ATC
NMSE	All drugs	1.67	5.4%	6.0%	41.3%	47.3%
	A	1.46	6.2%	4.8%	2.7%	24.0%
	B	0.55	14.5%	12.7%	38.2%	52.7%
	C	1.24	–	–	–	4.0%
	G	1.45	–	–	35.9%	37.9%
	J	1.20	10.0%	8.3%	–	10.8%
	N	7.53	15.1%	3.6%	39.2%	42.1%
NMAE	All drugs	0.30	0.0%	3.3%	16.7%	16.7%
	A	0.39	5.1%	5.1%	2.6%	5.1%
	B	0.29	10.3%	10.3%	17.2%	27.6%
	C	0.35	–	0.0%	–	0.0%
	G	0.39	–	–	12.8%	15.4%
	J	0.36	5.6%	2.8%	–	13.9%
	N	0.42	9.5%	2.4%	19.0%	16.7%

4.4. Value of Downstream Inventory Information

This section reports the value of downstream inventory information in the manufacturer's demand forecasting. Table 11 shows RNN's performance for models using and not using inventory information for various grouping schemes (see table 23 in Appendix for the performance of LR and the other nonlinear models). Generally, adding inventory information almost always improves RNN's performance on all groups regardless of the grouping schemes, except for the low volume and high volatility drugs when using clustering. RNN is particularly helpful when grouping by demand volume and volatility. In addition, we find that among the many lags we tried for the inventory information, $q=1$, typically provides the best results. In other words, inventory information in the distant past does not help, that is, including the inventory information in the most recent period is sufficient to garner most benefit.

4.5. Value of Supply Chain Structure Information

The goal of this study is to forecast the manufacturer's demand for each drug. So far, we do so by forecasting the aggregate demand from all DCs from all TPs. Another approach is to forecast demand at the TP level or DC level and then aggregate the forecasts to obtain the total demand required for each drug at the manufacturer. With this approach, not only do we learn from other drugs, we could also learn from other DCs or learn among the group of DCs belonging to the same TP. One can argue that this forecasting approach can be beneficial because all DCs from the same TP must share some similarities such as

Table 10 Baseline and Percentage Improvement Over Baseline of Cross-series Training Models with Grouping Schemes

		Best baseline	LR	LR_4	LR_ATC	LR_DTW	RNN	RNN_4	RNN_ATC	RNN_DTW
NMSE	All drugs	1.67	5.4%	7.2%	6.0%	7.8%	41.3%	46.7%	47.3%	53.3%
	HL	1.05	3.8%	3.8%	2.9%	3.8%	39.0%	44.8%	44.8%	50.5%
	HH	1.44	16.7%	5.6%	4.9%	13.2%	44.4%	47.2%	54.9%	57.6%
	LL	2.69	9.3%	10.0%	–	5.9%	–	29.4%	–	9.7%
	LH	17.37	–	–	–	–	–	49.8%	25.4%	27.1%
NMAE	All drugs	0.30	0.0%	0.0%	3.3%	3.3%	16.7%	16.7%	16.7%	20.0%
	HL	0.29	3.4%	3.4%	3.4%	3.4%	10.3%	17.2%	17.2%	20.7%
	HH	0.37	10.8%	2.7%	5.4%	8.1%	27.0%	24.3%	29.7%	27.0%
	LL	0.69	4.3%	5.8%	–	2.9%	–	18.8%	–	13.0%
	LH	0.81	–	–	–	0.0%	–	19.8%	12.3%	11.1%

Table 11 Baseline and Percentage Improvement Over Baseline of Cross-series Training Models by Using Inventory Information

		Best baseline	RNN	RNN_inv	RNN_4	RNN_4_inv	RNN_ATC	RNN_ATC_inv	RNN_DTW	RNN_DTW_inv
NMSE	All Drugs	1.67	41.3%	46.1%	46.7%	49.1%	47.3%	51.5%	53.3%	55.7%
	HL	1.05	39.0%	43.8%	44.8%	47.6%	44.8%	49.5%	50.5%	53.3%
	HH	1.44	44.4%	47.2%	47.2%	47.9%	54.9%	56.3%	57.6%	60.4%
	LL	2.69	–	–	29.4%	61.0%	–	–	9.7%	14.1%
	LH	17.37	–	–	49.8%	51.5%	25.4%	37.1%	27.1%	–
NMAE	All drugs	0.30	16.7%	16.7%	16.7%	20.0%	16.7%	20.0%	20.0%	20.0%
	HL	0.29	10.3%	17.2%	17.2%	20.7%	17.2%	20.7%	20.7%	20.7%
	HH	0.37	27.0%	18.9%	24.3%	27.0%	29.7%	32.4%	27.0%	27.0%
	LL	0.69	–	–	18.8%	26.1%	–	–	13.0%	2.9%
	LH	0.81	–	–	19.8%	23.5%	12.3%	45.7%	11.1%	–

ordering patterns. However, predicting at the downstream level also sees more volatility. Thus, it remains a question whether the benefit will overcome the drawback. This approach falls under the category of group time series forecasting, which is of significant interest to many researchers (Hyndman et al. 2011). Thus, this section explores this forecasting paradigm and provides detailed discussion on the findings.

At the DC level, we formulate the model for DC j with the input consisting of the last p weeks of order quantity and q weeks of inventory (6) as

$$\hat{x}_{ij,t+h} = f^{DC}(x_{ij,t}, \dots, x_{ij,t-p+1}, y_{ij,t}, \dots, y_{ij,t-q+1}), \quad (6)$$

where the mapping f^{DC} is learned using the information of drugs at DC j . We note that some DCs may not have sufficient training data (i.e., rank deficiency problem). For the DC-level model, given there are 247 DCs, we should have one model for each DC. Since complex machine learning models can suffer from overfitting due to the limited amount of data at the DC or TP levels, we could only build the DC-level model using LR. In particular, when building the DC-level model, we first train a model across all DCs, which we call LR_DC. If a specific DC is rank deficient, we use the predictions obtained from the LR_DC model for that DC.

To train the TP-level model, we use cross training among all DCs from the same TP. The rationale is that

DCs from the same TP share a similar supply chain management system; hence, the ordering policies and inventory control of these DCs are more likely to follow similar patterns. If DC j belongs to TP φ , we can use Equation (7) with the inputs of last p weeks' order quantity and q weeks' inventory to make predictions as follows

$$\hat{x}_{ij,t+h}^{(\varphi)} = f^{TP}(x_{ij,t}^{(\varphi)}, \dots, x_{ij,t-p+1}^{(\varphi)}, y_{ij,t}^{(\varphi)}, \dots, y_{ij,t-q+1}^{(\varphi)}), \quad (7)$$

where $x_{ij,t}^{(\varphi)}$ and $y_{ij,t}^{(\varphi)}$ are the order quantity and inventory of DC j which belongs to TP φ at time t .

Tables 20–22 in Appendix compare the performances of the three different levels of models: the aggregate level, the TP level and the DC level. The results confirm that in general, the performances of the models at the TP or DC level are no better than that from the aggregate model across all metrics (NMSE, NMAE, and bias). Given the amount of additional effort it requires to run each TP-level or DC-level model, forecasting at these levels may not be worthwhile unless each TP has a large number of DCs and each DC has a large amount of data.

4.6. Robustness Check

So far, we explored four machine learning models (LR, SVR, RF, and FC) compared to RNN, in combination with three different grouping schemes to predict future drug demands, with and without cross-series

information. All findings confirm that RNN has significantly better performance in terms of its forecast accuracy than the other ML models (see Appendix B for complete results).

Robustness of forecasting horizon. Forecast horizon refers to how far in the future we predict the demand. Figure 4 shows that, as expected, the forecast accuracy for all models decreases as we forecast further in the future (from 1 week to 8 weeks) because of higher uncertainty. This trend continues as the forecasting horizon goes beyond 8 weeks. However, regardless of the grouping schemes, cross-series training with downstream inventory information consistently leads to significant forecast improvements for all horizons.

Impact of RNN forecasting models on inventory performance. To further validate the benefits of our forecast models, we roughly estimate its respective service level (measured by the average number of weeks without stockout per year), the average annual stockout costs, and the average annual inventory cost. Specifically, we use the order-up-to policy for inventory replenishment (one of the most widely used replenishment policies). Stock levels are weekly reviewed and orders are placed to bring stock levels to the target levels. The target levels cover the weekly

drug demands with consideration to the specific accuracy of the forecasting method (Axsäter 2015). Next, using the WAC price of each NDC, we compute the inventory cost as the total value of unsold inventory across drugs. Table 12 shows that RNN with different grouping schemes (with or without inventory information) significantly outperforms the baseline in terms of service level, stockout, and inventory cost. The fact that the costs of stockout and inventory costs improve at the same time are beneficial because the two costs typically go in opposite directions. Improving demand forecasting reduces the mismatch of supply and demand, hence improving stockout and inventory costs simultaneously.

5. Explanation of the Benefits of RNN

So far, we have seen that RNN consistently outperforms other ML methods (LR is the simplest ML method). In this section, we try to provide some insights into the effectiveness of RNN over other methods for cross-series demand forecasting.

First, previous explorations of drugs' demand patterns suggest that many drugs demonstrate demand spikes in January, June, and December (see Figure 5). Specifically, a demand is considered to be a spike if it

Figure 4 NMSE of Cross Training Models with RNN Over Prediction Horizon [Color figure can be viewed at wileyonlinelibrary.com]

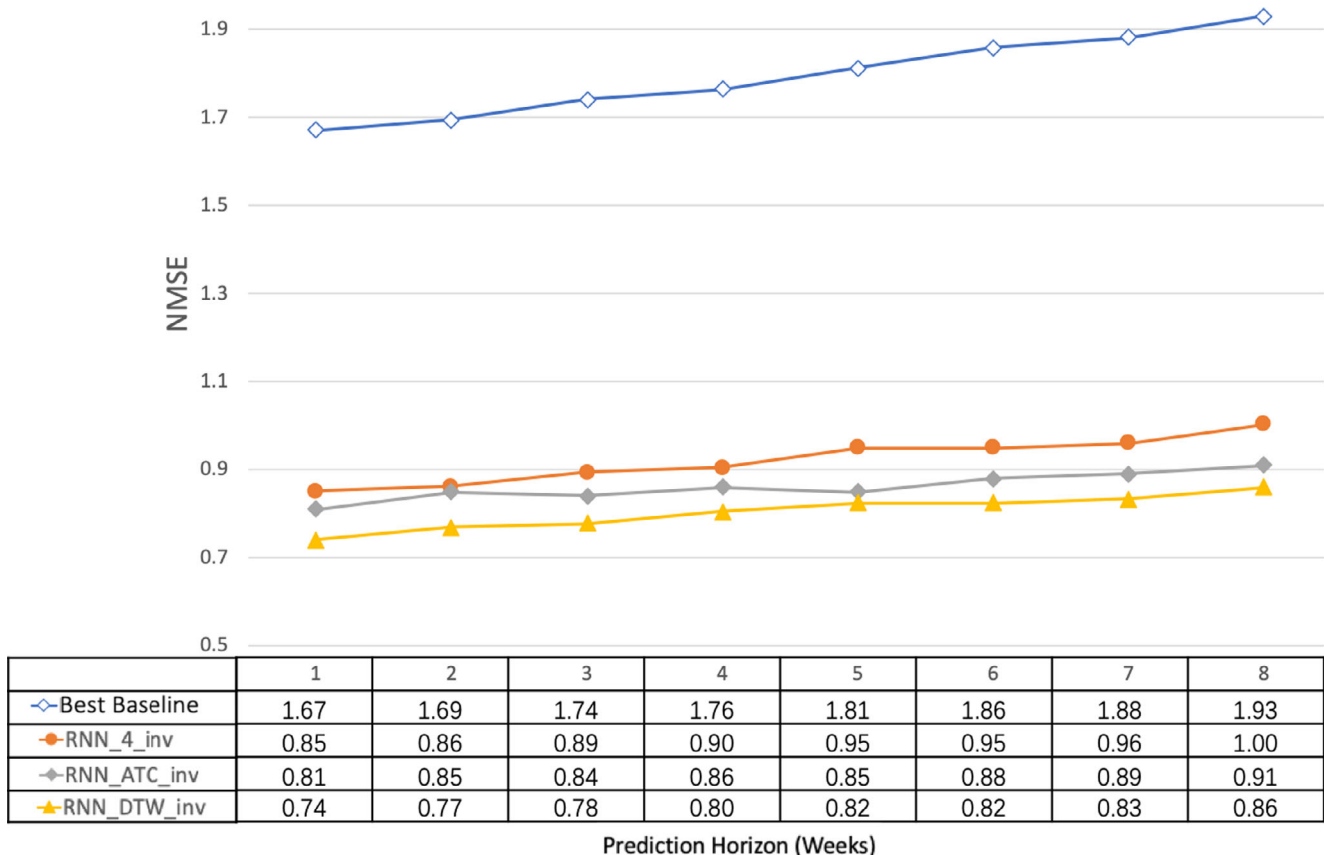
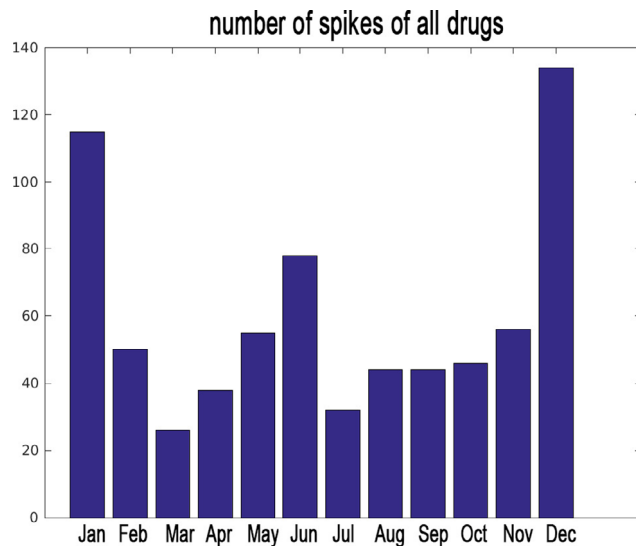


Table 12 Service Level, Inventory and Stockout Costs of RNN with Different Grouping Strategy

	Best baseline	RNN	RNN_inv	RNN_4	RNN_4_inv	RNN_ATC	RNN_ATC_inv	RNN_DTW	RNN_DTW_inv
Service level (# weeks with stockouts per year)	4.71	1.69	1.33	2.04	2.08	2.64	1.37	2.64	2.56
Inventory cost (in million \$)	3.42	3.54	3.43	2.56	2.49	2.76	3.08	2.76	2.66
Stockout cost (in million \$)	0.17	0.14	0.11	0.13	0.11	0.11	0.12	0.12	0.13

Figure 5 Timing of Demand Spikes from Our Data [Color figure can be viewed at wileyonlinelibrary.com]



is three standard deviations above the annual average demand. This phenomenon is closely related to the prevalent investment buying behavior of the pharma distributors that is well-documented and studied in the literature (Schwarz and Zhao 2011, Zhao et al. 2012). Specifically, investment buying refers to the phenomenon where distributors intentionally purchase large quantities of pharma products in anticipation of manufacturers' price increases in order to make profits by speculation on inventory. While exact price increase dates for different drugs are uncertain, these price increases often occur at the beginning, end, and/or in the middle of the year. Such timing may also reflect the manufacturer's incentive to get rid of inventory to meet financial/sales targets at certain times of the year. Given the prevalence of such phenomena, a good forecasting model should be able to capture such spikes.

RNN is capable of capturing temporal features, for example, demand spikes, due to its special design with memory cells that can remember distant past. Figure 6 shows the weekly spikes distribution in different months from the data and from the predictions using RNN and LR. If there are five weeks in a month, we merge the fifth week spikes into the fourth week. Observe that the patterns of demand spikes generated

from RNN's predictions matches well with the ground truth. In using LR, we tried two models: with and without indicator variables, which are added manually to capture the demand spikes in January, June, and December. While LR's predictions without indicator variables cannot capture the demand spikes in June, LR with indicator variables is too aggressive in the way that many non-spikes values are mistaken for spikes, which leads to its poor performance. On the other hand, RNN predictions "mimic" the ground truth more closely.

Second, RNN outperforms the other models due to its ability to generate rather complicated nonlinear features through its hidden layers. To empirically validate this observation, we purposely help the competing methods (LR, SVR, RF, and FC) with feature engineering. That is, we extract new features from existing data and use them as extra inputs to enhance performance of these models. These features include exponential moving averages of historical order quantities, the minimum, maximum, variance, maximum absolute deviation around the mean of order quantities, and the linear slope of past order quantities computed for respective window sizes from 2 to 10. In total, there are about 50 newly created features. However, even with the inclusion of engineered nonlinear features in the other models, RNN captures many hidden demand patterns the others miss. This helps to explain the better performance of RNN within our cross-series forecasting framework.

6. Application to a Second Dataset

To further validate the performance of our proposed cross-series forecasting framework, we test it on our second dataset from a different manufacturer using the same cross-validation procedure. The dataset includes all transactions from January 2011 to December 2017 between a drug manufacturer and the DCs of its TPs, collected weekly for 112 unique NDCs via 5 TPs and 73 DCs.

Table 13 clearly shows our framework is indeed beneficial, VAR is not a suitable method of choice, and RNN with cross-series training has the best performance across different accuracy metrics.

Table 14 confirms the superior performance of RNN across different grouping schemes (particularly

Figure 6 Spiked Patterns Captured by Different Forecasting Models [Color figure can be viewed at wileyonlinelibrary.com]

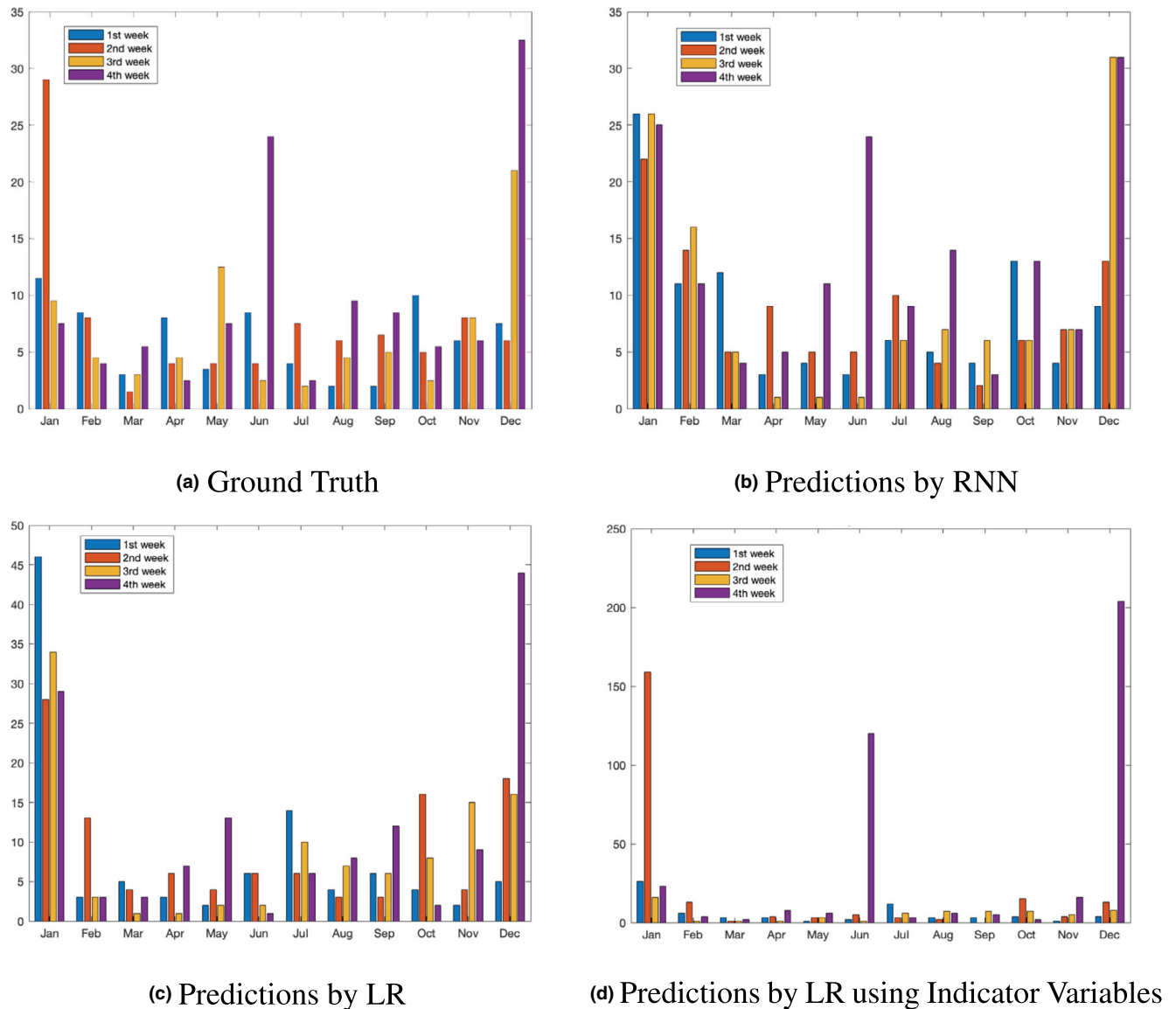


Table 13 Forecasting Bias and Accuracy Measures of the Global Cross-series Training Models (second dataset)

	Baselines (no cross-series)			Global cross-series training					
	MA	ETS	LR	VAR	LR	SVR	RF	FC	RNN
NMSE	4.67 ± 0.10	4.95 ± 0.14	4.29 ± 0.13	6.63 ± 0.22	3.98 ± 0.14	3.72 ± 0.13	3.67 ± 0.13	3.65 ± 0.11	2.69 ± 0.37
NMAE	0.42 ± 0.02	0.44 ± 0.02	0.41 ± 0.02	0.62 ± 0.05	0.40 ± 0.02	0.40 ± 0.02	0.38 ± 0.02	0.39 ± 0.02	0.30 ± 0.06
Bias	-0.01 ± 0.01	-0.14 ± 0.02	-0.11 ± 0.02	-0.20 ± 0.05	-0.07 ± 0.02	-0.05 ± 0.02	-0.07 ± 0.02	-0.04 ± 0.03	-0.01 ± 0.02

grouping by demand volume/volatility and clustering), which is consistent with our previous observations in the first dataset. For the counterparts of information in Tables 8 and 9, refer to Appendix C. Table 15 verifies the benefit of downstream inventory information in pharma demand forecasting.

The above results show that our model framework works well and major insights hold true to the second dataset. This, together with our discussion and justification of the models we choose (section 4) as well as the possible explanation of the effectiveness of the RNN models (section 6), provides evidence of the generalizability of our model framework and results.

Table 14 Baseline and Percentage Improvement Over Baseline of Cross-series Training Models with Grouping Schemes (second dataset)

		Best baseline	LR	LR_4	LR_ATC	LR_DTW	RNN	RNN_4	RNN_ATC	RNN_DTW
NMSE	All Drugs	4.29	7.23%	6.29%	1.17%	1.40%	37.30%	45.45%	43.36%	49.18%
	HL	1.34	8.21%	6.72%	1.49%	1.49%	39.55%	45.52%	44.03%	48.51%
	HH	4.05	0.49%	2.22%	–	1.98%	11.60%	46.42%	34.57%	49.14%
	LL	0.98	2.04%	2.04%	–	0.00%	–	65.31%	–	68.37%
	LH	5.12	14.06%	14.65%	15.23%	–	–	28.52%	–	28.32%
NMAE	All Drugs	0.41	2.44%	2.44%	0.00%	0.00%	26.83%	39.02%	26.83%	43.90%
	HL	0.38	2.63%	2.63%	0.00%	0.00%	34.21%	39.47%	31.58%	42.11%
	HH	0.54	0.00%	0.00%	–	3.70%	1.85%	37.04%	16.67%	40.74%
	LL	0.53	0.00%	1.89%	–	0.00%	–	52.83%	–	56.60%
	LH	0.59	0.00%	5.08%	8.47%	–	–	37.29%	–	37.29%

Table 15 Baseline and Percentage Improvement Over Baseline of Cross-series Training Models by Using Inventory Information (second dataset)

		Best baseline	RNN	RNN_inv	RNN_4	RNN_4_inv	RNN_ATC	RNN_ATC_inv	RNN_DTW	RNN_DTW_inv
NMSE	All drugs	4.29	37.30%	50.12%	45.45%	60.61%	43.36%	42.89%	49.18%	67.60%
	HL	1.34	39.55%	51.49%	45.52%	61.94%	44.03%	44.78%	48.51%	54.48%
	HH	4.05	11.60%	37.28%	46.42%	49.63%	34.57%	29.14%	49.14%	59.51%
	LL	0.98	–	–	65.31%	33.67%	–	–	68.37%	70.41%
	LH	5.12	–	–	28.52%	24.41%	–	–	28.32%	22.46%
NMAE	All Drugs	0.41	26.83%	36.59%	39.02%	41.46%	26.83%	26.83%	43.90%	46.34%
	HL	0.38	34.21%	42.11%	39.47%	42.11%	31.58%	42.11%	42.11%	44.74%
	HH	0.54	1.85%	18.52%	37.04%	37.04%	16.67%	35.19%	40.74%	46.30%
	LL	0.53	–	–	52.83%	15.09%	–	–	56.60%	58.49%
	LH	0.59	–	–	37.29%	25.42%	–	–	37.29%	35.59%

7. Discussion and Conclusion

7.1. Post-hoc Work

To examine the applicability of our proposed forecasting method to the pharma industry, we conducted post-hoc interviews with several of the major pharma companies which we previously interacted with through the forecasting questionnaires. Each of the post-hoc interviews lasted roughly 45 mins and followed a specific list of questions (see Appendix E).

The post-hoc interviews confirmed the superior performance of our proposed forecasting method over both the company and the industry forecasts. As mentioned, while we are not able to obtain the manufacturer's internal forecasts, we did obtain the more accurate forecasts shared with the manufacturer from its downstream DCs for the first dataset. All interviewees confirmed that the downstream distributors' forecast, if available, are more accurate than the manufacturer's forecast. This is also expected by the supply chain theory since the downstream is closer to the customers. Furthermore, these shared forecasts from the distributors to the manufacturer are the downstream's best forecasts (distributors must provide explanations when they place orders deviating much from their forecasts). Hence, these forecasts have already included their human adjustments, if any.

When asked if human adjustments could predict demand spikes caused by investment buying prevalent in the industry, the interviewees noted that while large demand spikes are statistically observed at the beginning, middle, or end of the year, their exact timing was typically unknown to the pharma companies' forecasting groups. Even if the cause of demand spikes (e.g., price increases) is known beforehand, those decisions are made by other departments, and usually the demand forecasting groups are not informed, so that human adjustments are rare for these situations. Correspondingly, the ability of our proposed approach in predicting such demand spikes is particular important.

All interviewees were open to the implementation of advanced forecasting methods. When asked about the potential hurdles in implementations, besides associated costs (e.g., costs to acquire/implement such technology), the interviewees indicate IT integration to facilitate transition, usage, and technical support as the primary hurdle to overcome because most companies have their supply chain planning tools/forecasting platforms. Integration into these platforms would be key for successful implementation.

7.2. Conclusion

Demand forecasting drives many operational decisions and directly relates to companies' financial

goals. Demand forecasting in the pharma industry is especially critical to drug manufacturers due to the unique features of the industry. However, the performance of existing forecasting models in the pharma industry seems to have reached a bottleneck, often limited by the amount of available data. At the same time, the availability of supply chain channel data and the advances in machine learning technologies provide new opportunities. Under this situation, in this study, we propose a novel cross-series forecasting framework which leverages information *across drugs* regarding historical demand and non-demand features such as downstream inventory data and supply chain structure information. This framework resolves the lack of data issue for more advanced machine learning models and further optimizes the cross-series training through various grouping schemes. In addition, while analytical work has long shown the value of downstream inventory information, we are among the first to empirically capture such value in the cross-series demand forecasting setting.

Using two rich datasets from top drug manufacturers, we conducted extensive computational experiments to test our proposed forecasting framework. Our results provide some important insights:

- Training time series across drugs with careful grouping design shows significant improvement in forecast accuracy compared with a benchmark representing the state-of-the-art of industry and academic forecasting methods.
- The proposed cross-series forecasting framework exploits the potential of machine learning for demand forecasting.
- Cross-series training is most effective for demand forecasting of low-volume drugs, whose forecasts are the most difficult in practice, possibly because it helps the most to alleviate the lack-of-data issues for these drugs.
- Cross-series training with different grouping schemes based on product-specific information, either by demand volume/volatility or by product-based domain knowledge (ATC code in our case), is effective. On the other hand, clustering algorithms can be a great option when lacking domain knowledge in grouping.
- With cross-series training, our additional numerical analysis shows that RNN consistently performs the best, far exceeding other machine learning methods, likely because (i) it most effectively captures hidden factors such as demand spikes caused by investment buying behavior, and (b) its special architecture makes it suitable for time series data. While the latter has been reported in some studies in the

literature, we are the first to document the former in the pharma context.

- Downstream inventory information is indeed beneficial in demand forecasting; however, as expected, any distant past inventory information does not bring additional benefit. This finding empirically confirms the value of downstream inventory information shown in the theoretical operations management literature. Our work also offers some practical guidelines of what product groups' inventory information is most useful.
- While more detailed supply chain structure information such as the downstream DC level or TP level data is helpful to learn across DCs or TPs to capture possible common hidden factors, its benefits do not seem to overcome the loss of accuracy due to disaggregation. As a result, it may not be worthwhile to collect the more detailed DC-level data for the aggregate demand forecasting.

Our proposed forecasting framework (including grouping schemes, using downstream inventory and supply chain structure information, in combination with the RNN models) can be applied to other pharma manufacturers, wholesalers, and possibly other industries as well based on its robust performances. Domain knowledge is important for making modifications when adapting to other industries. This industry-specific customization could be a promising research direction since products in different industries may have unique characteristics that can be extracted and incorporated into the forecasting framework to fully boost its performance. Finally, as leading pharma companies (Pfizer, Sanofi, etc.) have already considered using AI platforms to help drug development processes, pharma companies seem very open for AI applications in other areas of the pharma supply chains. Demand forecasting using machine learning is shown to be very fruitful by this research.

Note

¹The 14 main groups include: alimentary tract and metabolism (A), blood and blood forming organs (B), cardiovascular system (C), dermatologicals (D), genito-urinary system and sex hormones (G), systemic hormonal preparations, excluding sex hormones and insulins (H), anti-infectives for systemic use (J), anti-neoplastic and immunomodulating agents (L), musculo-skeletal system (M), nervous system (N), anti-parasitic products, insecticides and repellents (P), respiratory system (R), sensory organs (S), and various (V) (see World Health Organization, <https://www.who.int/tools/atc-ddd-toolkit/atc-classification>).

References

- Angell, M. 2004. Excess in the pharmaceutical industry. *CMAJ* **171** (12): 1451–1453.
- Arvan, M., B. Fahimnia, M. Reisi, E. Siemsen. 2019. Integrating human judgement into quantitative forecasting methods: A review. *Omega* **86**: 237–252.
- Aviv, Y. 2007. On the benefits of collaborative forecasting partnerships between retailers and manufacturers. *Management Sci.* **53**(5): 777–794.
- Axsäter, S. 2015. *Inventory Control*, volume **225**. Springer, Berlin.
- Baardman, L., I. Levin, G. Perakis, D. Singhvi. 2018. Leveraging comparables for new product sales forecasting. *Prod. Oper. Manag.* **27**(12): 2339–2349.
- Bergmeir, C., R. Hyndman, B. Koo. 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Statist. Data Anal.* **120**: 70–83.
- Berndt, D., J. Clifford. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases, 1994*, pages 359–370.
- Boone, T., R. Ganeshan, R. Hicks, N. Sanders. 2018. Can google trends improve your sales forecast? *Prod. Oper. Manag.* **27**: 1770–1774.
- Brown, R., R. Meyer. 1961. The fundamental theorem of exponential smoothing. *Oper. Res.* **9**(5): 673–685.
- Cachon, G., M. Fisher. 2000. Supply chain inventory management and the value of shared information. *Management Sci.* **46**(8): 1032–1048.
- Cachon, G., M. Olivares. 2010. Drivers of finished-goods inventory in the U.S. automobile industry. *Management Sci.* **56**(1): 202–216.
- Carbonneau, R., K. Laframboise, R. Vahidov. 2008. Application of machine learning techniques for supply chain demand forecasting. *Eur. J. Oper. Res.* **184**(3): 1140–1154.
- Chase, C. 2013. *Demand-Driven Forecasting: A Structured Approach to Forecasting*. John Wiley & Sons, Hoboken.
- Chase, C. 2016. *Next Generation Demand Management: People, Process, Analytics, and Technology*. John Wiley & Sons, Hoboken.
- Choi, T., S. Wallace, Y. Wang. 2018. Big data analytics in operations management. *Prod. Oper. Manag.* **27**(10): 1868–1883.
- Cook, A. 2016. *Forecasting for the Pharmaceutical Industry: Models for New Product and In-market Forecasting and How to Use Them*. CRC Press, Boca Raton. Gower.
- Cui, R., S. Gallino, A. Moreno, D. Zhang. 2018. The operational value of social media information. *Prod. Oper. Manag.* **27**(10): 1749–1769.
- Ferreira, K., B. Lee, D. Simchi-Levi. 2015. Analytics for an online retailer: Demand forecasting and price optimization. *Manuf. Serv. Oper. Manag.* **18**(1): 69–88.
- Gardner, E. 1990. Evaluating forecast performance in an inventory control system. *Management Sci.*, **36**(4): 490–499.
- Grushka-Cockayne, Y., V. R. Jose, K. C. Lichtendahl, Jr. 2016. Ensembles of overfit and overconfident forecasts. *Management Sci.*, **63**(4): 1110–1130.
- Halevy, A., P. Norvig, F. Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**(2): 8–12.
- Hill, T., M. O'Connor, W. Remus. 1996. Neural network models for time series forecasts. *Management Sci.* **42**(7): 1082–1092.
- Hochreiter, S., J. Schmidhuber. 1997. Long short-term memory. *Neural Comput.* **9**(8): 1735–1780.
- Hyndman, R., G. Athanasopoulos. 2018. *Forecasting: Principles and Practice*, OTexts.
- Hyndman, R., A. Koehler. 2006. Another look at measures of forecast accuracy. *Int. J. Forecast.* **22**(4): 679–688.
- Hyndman, R., R. Ahmed, G. Athanasopoulos, H. Shang. 2011. Optimal combination forecasts for hierarchical time series. *Comput. Statist. Data Anal.* **55**(9): 2579–2589.
- Jain, C. 2003. Benchmarking forecasting practices in pharmaceutical industry. In *Proceedings of Pharmaceutical SAS Users Group*.
- Karpathy, A. 2015. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy blog* **21**: 23.
- Ke, G., Z. Xu, J. Zhang, J. Bian, T. Liu. 2019. Deepgbm: A deep learning framework distilled by gbdm for online prediction tasks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 384–394.
- Khalil Zadeh, N., M. Sepehri, H. Farvareh. 2014. Intelligent sales prediction for pharmaceutical distribution companies: A data mining based approach. *Math. Probl. Eng.* **2014**. <https://doi.org/10.1155/2014/420310>
- Kiely, D. 2004. The state of pharmaceutical industry supply planning and demand forecasting. *J. Bus. Forecast.* **23**(3): 20.
- Kim, W., J. Won, S. Park, J. Kang. 2015. Demand forecasting models for medicines through wireless sensor networks data and topic trend analysis. *Int. J. Distrib. Sens. Netw.* **11**(9): 907169.
- Kremer, M., E. Siemsen, D. Thomas. 2015. The sum and its parts: Judgmental hierarchical forecasting. *Management Sci.* **62**(9): 2745–2764.
- Kurawarwala, A., H. Matsuo. 1996. Forecasting and inventory management of short life-cycle products. *Oper. Res.* **44**(1): 131–150.
- Lau, R., W. Zhang, W. Xu. 2018. Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Prod. Oper. Manag.* **27**(10): 1775–1794.
- Liu, X., P. Singh, K. Srinivasan. 2016. A structured analysis of unstructured big data by leveraging cloud computing. *Market. Sci.* **35**(3): 363–388.
- Makridakis, S., S. Wheelwright, R. Hyndman. 2008. *Forecasting Methods and Applications*. John Wiley & sons, Hoboken.
- Merkuryeva, G., A. Valberga, A. Smirnov. 2019. Demand forecasting in pharmaceutical supply chains: A case study. *Procedia Comput. Sci.* **149**: 3–10.
- Nikolopoulos, K., S. Buxton, M. Khammash, P. Stern. 2016. Forecasting branded and generic pharmaceuticals. *Int. J. Forecast.* **32**(2): 344–357.
- Olivares, M., G. P. Cachon. 2009. Competing retailers and inventory: An empirical investigation of General Motors' dealerships in isolated US markets. *Management sci.* **55**(9): 1586–1604.
- Richard, V. 2014. Demand shaping: Achieving and maintaining optimal supply-and-demand alignment. *SAS - White paper*, pages 1–12.
- Sastri, T. 1985. A state space modeling approach for time series forecasting. *Management Sci.* **31**(11): 1451–1470.
- Schmittlein, D., J. Kim, D. Morrison. 1990. Combining forecasts: Operational adjustments to theoretically optimal rules. *Management Sci.* **36**(9): 1044–1056.
- Schwarz, L., H. Zhao. 2011. The unexpected impact of information sharing on us pharmaceutical supply chains. *Interfaces* **41**(4): 354–364.
- See-To, E., E. Ngai. 2018. Customer reviews for demand distribution and sales nowcasting: A big data approach. *Ann. Oper. Res.* **270**(1–2): 415–431.
- Sims, C. 1980. Macroeconomics and reality. *Econ. J. Econ. Soc.*, **48** (1): 1–48.
- Van Belle, J., T. Guns, W. Verbeke. 2021. Using shared sell-through data to forecast wholesaler demand in multi-echelon supply chains. *Eur. J. Oper. Res.* **288**(2): 466–479.

- Weller, M., S. Crone. 2012. Supply chain forecasting: Best practices & benchmarking study. *Lancast. Center Forecast.* <https://doi.org/10.1155/2014/420310>
- Winters, P. 1960. Forecasting sales by exponentially weighted moving averages. *Management Sci.* 6(3): 324–342.
- Xu, L., V. Mani, H. Zhao. 2020. 'Not a box of nuts and bolts': Distribution Channels for Specialty Drugs. Working paper. Available at SSRN: <https://ssrn.com/abstract=3831324> (accessed date April 22, 2021).
- Zhao, H., C. Xiong, S. Gavirneni, A. Fein. 2012. Fee-for-service contracts in pharmaceutical distribution supply chains: Design, analysis, and management. *Manuf. Serv. Oper. Manag.* 14(4): 685–699.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A Description of Machine Learning Models

Appendix B Performance of Cross-series Forecasting Models

Appendix C Performance of Cross-Series Forecasting Models on Our Second Dataset

Appendix D Questionnaire of Pharma Forecasting Practices

Appendix E Post-hoc interview's questions