A Machine Learning Approach for Predicting Clinical Trial Patient Enrollment in Drug Development Portfolio Demand Planning

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Ahmed Shoieb
May 2023

DEDICATION

I dedicate this work to my daughter, Ahd Ahmed Ahmed Shoieb.

To my parents, Dr. Ahmed Mohamed Shoieb and Dr. Mona Elgayyar.

To my wife, Yassmein Elgemaei.

I would not be the person I am today without you all, I owe you everything.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis Committee Dr. Andrew Yu, Dr. John Kobza, and Dr. James Simonton for being world-class educators, their inspirational leadership, and their continuous support throughout this wonderful program.

I am profoundly thankful for my advisor, Dr. Andrew J. Yu, Department of Industrial and Systems Engineering at the Tickle College of Engineering for his valuable suggestions, continuous guidance, and encouragement in completing this research. I am especially grateful for his vast knowledge, empathy, and willingness to collaborate on such a rare research topic. Thank you for everything, Dr. Yu. I would like to thank the faculty members in the Industrial and Systems Engineering Department for their support throughout the program, and profound knowledge.

I would also like to thank my family for their blessings, encouragement, and support throughout this program: My father Dr. Ahmed Mohamed Shoieb, my mother Dr. Mona Elgayyar, my sister Dr. Zienab Shoieb and her husband Ahmed Rizk, my brother Eng. Mohamed Shoieb, and my brother Yousef Shoieb.

Finally, I would like to thank my wife Yassmein Elgemaei for her patience and constant support during this challenging program, and my daughter Ahd Ahmed Shoieb for being the light of our life.

ABSTRACT

One of the biggest challenges the clinical research industry currently faces is the accurate forecasting of patient enrollment (namely if and when a clinical trial will achieve full enrollment), as the stochastic behavior of enrollment can significantly contribute to delays in the development of new drugs, increases in duration and costs of clinical trials, and the over- or under- estimation of clinical supply. This study proposes a Machine Learning model using a Fully Convolutional Network (FCN) that is trained on a dataset of 100,000 patient enrollment data points including patient age, patient gender, patient disease, investigational product, study phase, blinded vs. unblinded, sponsor CRO selection, enrollment quarter, and enrollment country values to predict patient enrollment characteristics in clinical trials. The model was tested using a dataset consisting of 5,000 data points and yielded a high level of accuracy. This development in patient enrollment prediction will optimize portfolio demand planning and help avoid costs associated with inaccurate patient enrollment forecasting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 – Introduction

1.1 Pharmaceutical Clinical Trial Supply Chain

Given the current competition in the pharmaceutical industry, and the need to accelerate drug development to be first to market, a critical component of delivering therapies to market has been to optimize the Clinical Supply Chain. Clinical supply disruptions can lead to serious economic impact on a program, and operational inefficiencies. These challenges cause various supply chain disruptions, which lead to delays in meeting milestones and completion of a study and could potentially delay the time to market. Figure 1 illustrates the complex nature of a clinical supply chain.

All parameters of a clinical trial, and the potential factors that influence a clinical supply chain must be considered, to develop an appropriate strategy ensuring the timely delivery of clinical supplies (Rodgers et al. 2019). Clinical supply chains are already governed under very strict regulatory guidance from global agencies; therefore, there are many limitations and factors that restrict the flexibility in a clinical supply chain. These restrictions, along with the fact that these drugs are in drug development, create complexity in furnishing to delayed recruitment in clinical trials. This brings several unique qualities to this supply chain:

1) Due to a long production lead times (sometimes up to one year or even longer), and expensive changeover cost in trial production, the resupply of these trials is not always feasible.
2) The demand for clinical drugs are stochastic, and their arrival rates at clinics vary with time (i.e., non-stationary) (Zheng et. al)[2].

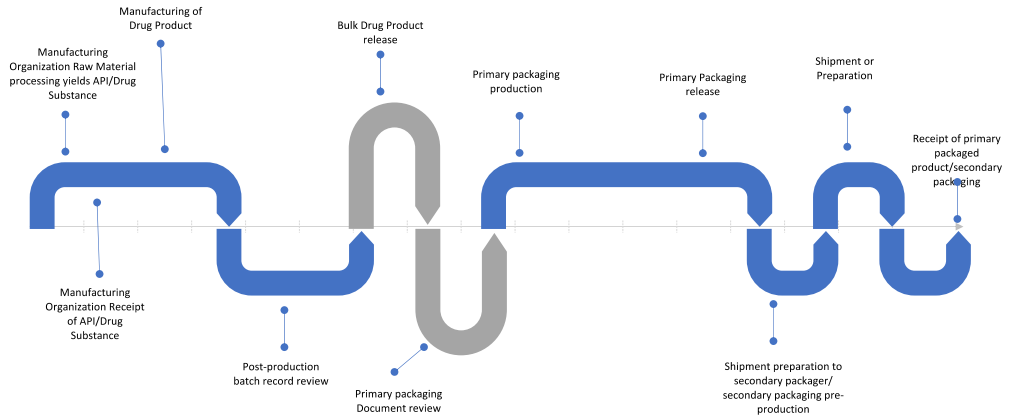

Figure 1, Clinical Supply Chain End-To-End Process Overview

In order to take a drug from discovery to commercialization, this takes many years, and the fulfillment of several regulatory requirements (Mohs et al. 2017). One of these requirements is the successful completion of studies spanning from Phase I to Phase IV. Each phased study consists of unique characteristic, making the demand planning process different between the study types. For example, in early phase studies (Phase I), safety and dose ranging are studied. These studies contain anywhere between 20-100 subjects, depending on the therapeutic area, indication, and endpoints of the study. Phase II studies study the drug efficacy of the drug, and evaluate 80-300 patients. Phase III evaluate the therapeutic effects of the drug on 100-2000 patients. Lastly, Phase IV studies are designed to study the long-term effects of the investigational product and are usually available for anyone seeking treatment (Mahan 2014). Given the intricacies related to drug development, regulation, and stochastic nature of enrollment in clinical trials, it is complex to precisely demand plan for Clinical Trials. According to research from the Tufts Center for the Study of Drug Development (Tufts), while 9 out of 10 clinical trials worldwide meet their patient enrollment goals, reaching those targets means that drug developers need to nearly double their original timelines. Clinical trial supply enrollment can be extended which impact production. Also, to support Pharmaceutical company strategies, an increase in indications and to support market entry strategies is almost always experienced. To understand what kind of demand is required to support multiple clinical trials, and expanding portfolios, common practice in industry, and in the literature on clinical trial demand planning, Monte Carlo simulations are the go-to; however, the programs and software that depend on these types of simulations require an input based on the clinical teams site feasibility data, which is not always reliable as they are mostly assumptions. The parameter of patient enrollment rate is usually input as a flat percentage and is not considered to be an essential input. The main point is that if poor

parameters and constraints are input into the model then poor results will be given as outputs, which is not a model that could be viewed as reliable in such a competitive and high stakes environment such as clinical supply management in the pharmaceutical industry, as patients' lives are at stake. Monte Carlo simulations are often useful on how to allocate a specific batch that has been already allocated to a trial, and distribution strategies to global depots; however, program and portfolio level demand planning using solely a Monte Carlo simulation requires significant improvement as drug product manufacturing lead times can be long (9-12 months in some cases), and finished product may not be available in time to meet patient dosing. Portfolio demand level planning is planning that occurs with a multi-year outlook for the product, which constitutes target indications in which the product will be investigated, as well as strategies for which global markets the product will target to enter. This requires careful planning, as the decisions made as far as production volume will impact the portfolio plan downstream. It is critical to understand the demand many years in advance due to the long production lead times, especially due to supply constraints, and complexity of the clinical supply chain. The inputs into the Monte Carlo simulation can be refined at the trial level, but this is not beneficial in understanding the program demands. With Pharma companies outsourcing more than 90% of their activities (Solem Global 2021), better insight into the future as to how much supply is needed and when is critical, which will create the framework in orchestrating batch productions, while taking into account the short shelf-life of drugs in development. The issue at hand is not just a matter of fine tuning the current demand of one trial, this is program level optimization. The pharmaceutical clinical trial supply chain is a complex process that involves various stages and stakeholders (Chen et al. 2012). The first step is planning and Study Design where the sponsor, a pharmaceutical company, develops a study protocol and design with the help of

researchers and regulatory agencies. As seen in Figure 1, this step feeds into Investigation Medicinal Product (IMP) Manufacturing, where the IMP is manufactured in accordance with Good Manufacturing Practices (GMPs) and is subject to quality control measures. After this is Investigation Medicinal Product Packaging and Labeling, where IMP is packaged, labeled, and distributed to global depots then to the clinical trial sites where it is stored and administered to patients. In parallel to these operations, clinical trial sponsors must obtain regulatory approval from regulatory bodies to carry out the trial, from agencies such as the US FDA or the European Medicines Agency (EMA). The main goal for any pharmaceutical company is to achieve commercialization, so that the drug may be commercialized and made available to patients. Given the complex nature of the drug development, throughout the clinical trial supply chain, there are various stakeholders involved, including the sponsor, investigators, clinical research organizations (CROs), regulatory agencies, and contract manufacturing organizations (CMOs). It is important to maintain close communication and collaboration between these stakeholders to ensure the success of the clinical trial.

1.1.1 Challenges Faced by Clinical Supply Managers

Due to the complex nature of clinical supply chains, clinical supply managers often face several challenges in the clinical supply management process. One of the main challenges is the accurate prediction of patient enrollment, as enrollment rate predictions can be affected by many factors, including patient demographics, study design, and feasibility strategies. What is patient enrollment in a clinical trial? According to the FDA, “patient enrollment is the process of registering or entering a patient into a clinical trial. Once a patient has been enrolled, the participant would then follow the clinical trial protocol. Clinical investigations are designed to

enroll a set number of participants to increase the likelihood of answering the trial questions." Enrollment occurs in a set duration of the clinical trial, and once defined, the study team must not close enrollment until the last patient has enrolled. Based on this stipulation, accurate enrollment predictions are essential for ensuring the appropriate supply of drugs and other materials needed for a clinical trial, and that sufficient batches of IMP are produced with sufficient shelf-life. If enrollment is extended, then additional batches will require to be produced to support the clinical trial further, as IMPs in the drug development process contain limited stability data to support shelf-life which is why they have a short shelf-life. The clinical trial will also be extended if the enrollment period is also extended. Figure 2 below shows an example schematic of impact of clinical trial enrollment delays on batch allocation and clinical trial duration. Inventory management is also a challenge that clinical supply manager face, as managing the inventory of IMP required for a clinical requires precise balance in order to minimize waste and avoid stockouts.

Clinical trials often involve multiple sites in different locations, making logistics and distribution a significant challenge for clinical supply managers. Ensuring that IMP is delivered to sites on time and in the correct quantities is essential for the success of a clinical trial. Not only is having sufficient quantities a requirement for the success of a trial, but the quality and integrity of the IMP is also key. Maintaining quality of IMP is achieved by ensuring that IMP is stored and shipped within the labeled storage conditions, as defined by the stability data. Clinical supply managers must ensure that all aspects of the clinical supply management process comply with regulatory requirements and Good Clinical Practices (GCP). This includes ensuring that IMP is process at approved suppliers, and that proper documentation is maintained throughout the clinical trial process.

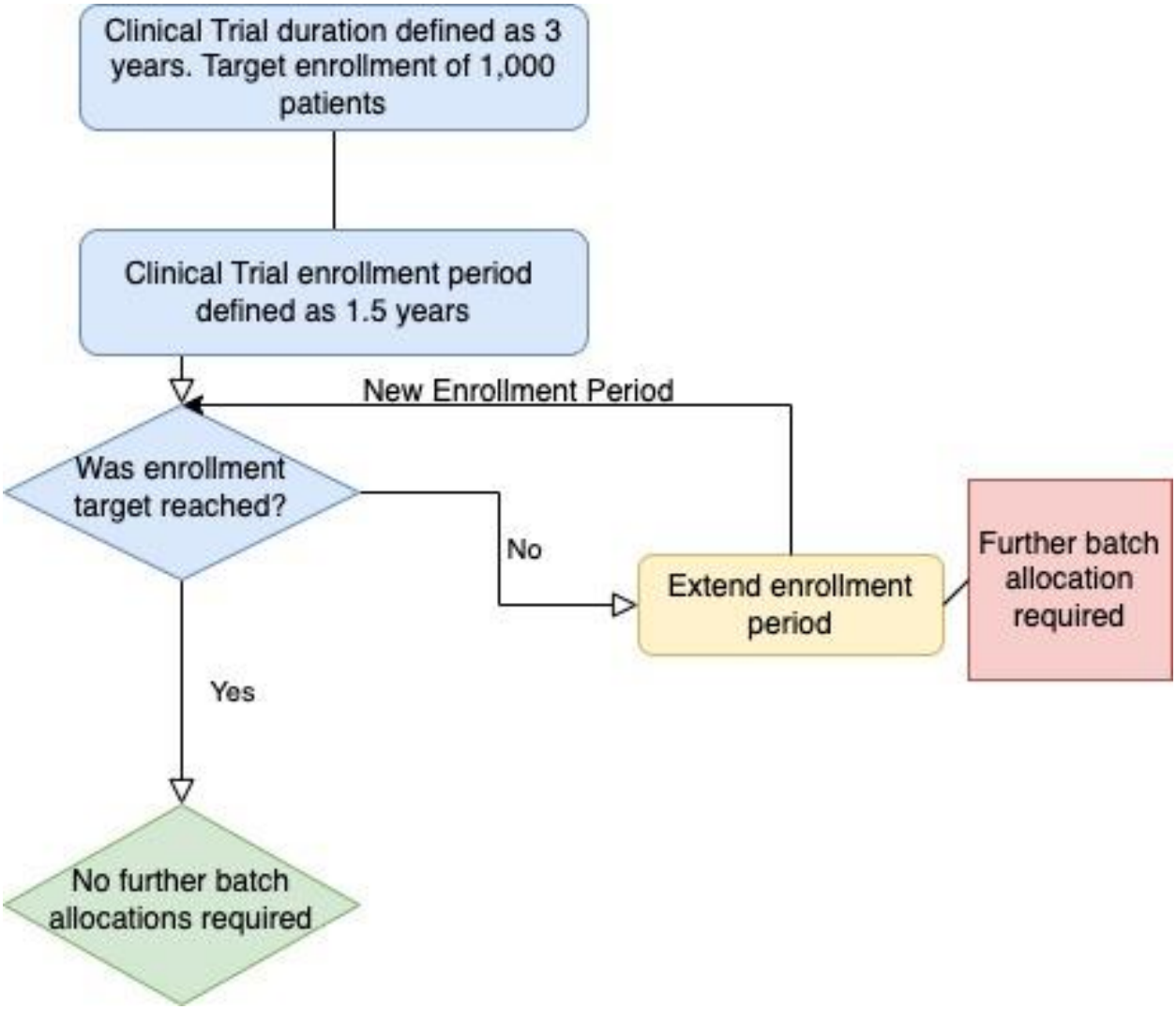

Figure 2, Patient Enrollment Duration Impact on Batch Allocation and Clinical Trial Duration

1.1.2 Technological Solutions for Challenges

Addressing these challenges requires a combination of careful planning, robust processes, and innovative technologies such as machine learning and AI algorithms. Examples of potential innovative technological solutions to these challenges include, machine learning algorithms, which can be used to improve the accuracy of enrollment rate predictions. These algorithms can take into account multiple factors, and provide more accurate predictions than traditional statistical methods. For inventory management, real-time data and analytics can help clinical supply managers optimize their inventory management. Predictive analytics can help identify trends, and patterns in drug usage and provide insight into future demand. This information can be used to optimize inventory levels, and reduce waste. These systems can provide real-time visibility into the supply chain, allowing clinical supply managers to identify potential problems and address them before they impact the clinical trial. To assist in quality control, the use of automated temperature monitoring systems, and cold chain management solutions can help ensure that IMP is stored and transported at the correct temperature. This helps to maintain the quality, and integrity of IMP throughout the clinical trial process. Machine learning can be a clinical supply manager's partner if utilized correctly.

It is widely recognized that wasted clinical supply can be a significant cost for pharmaceutical companies. In clinical trials, wasted clinical supply can occur when there is an overproduction of clinical supplies, which leads to expired or unused materials. This can result in significant financial losses, as the materials and the resources required to produce and store them are essentially wasted. Additionally, the storage and transportation of expired or unused materials can also result in additional costs. It is estimated that the cost of wasted clinical supply can range

from 10-30% of the total clinical supply budget for a clinical trial. For large, multi-national pharmaceutical companies, this can represent millions of dollars in wasted resources and lost profits, as a clinical trial can cost an average of $80MM to $150 billion. Therefore, pharmaceutical companies are actively looking for ways to minimize the amount of wasted clinical supply in their clinical trials. This includes using advanced forecasting and optimization techniques, such as machine learning and artificial intelligence algorithms, to improve the accuracy of their clinical supply forecasts, and reduce the amount of overproduction, by predicting more accurate patient enrollment.

1.2 Pharmaceutical Clinical Trials and Patient Enrollment

Pharmaceutical clinical trials are an essential step in the development of new treatments and are used to assess the efficacy and safety of new medications before they are made available to the general population. To find the optimal treatment choice, these studies are carried out in stages, with each step building on the findings of the one before it. (Mahan 2014)

According to CT.gov and the Food and Drug Administration (FDA), there are four phases of clinical trials. These trials are conducted in phases, with each phase building on the results of previous trials to determine the best possible treatment option. Phase I trials focus on determining the safety of a new drug and involve a small number of healthy volunteers. Phase II trials are designed to evaluate the efficacy of the drug, and are conducted in a larger patient population. Phase III trials are larger, randomized, controlled trials that provide the most definitive evidence of a drug's safety and efficacy. Finally, Phase IV trials are conducted after the drug has been approved for public use, and are designed to monitor the long-term effects and

risks of the drug. Clinical trials are a critical step in the drug development process, and can take several years to complete. The process begins with the development of a study protocol, which outlines the study design, patient population, and endpoints. The protocol is reviewed by regulatory agencies such as the FDA, and the trial is only approved if it meets rigorous scientific and ethical standards. Once the trial has been approved, the clinical trial material is manufactured and shipped to clinical trial sites. Patients are then recruited, and the trial begins. Clinical trial sites collect data on patient outcomes and side effects, which is then analyzed by the sponsor and reported to regulatory agencies. If the results of the trial are favorable, the sponsor may seek regulatory approval for the drug, and it may be made available to patients.

Given the high level of regulation, and the complexity of clinical trial designs, clinical supply chains are in turn very complex in nature. It is important for clinical trials to be supported by a strong supply chain in order to be able to support global patient populations, and deliver supply just in time for patient dosing. A strong supply chain is stemmed from accurate forecasting, and prediction of patient enrollment, which gives the Clinical Supply managers the essential information required to accurately predict how many batches to produce, the quantities in each batch, and when to produce and release these batches. Many parameters are taken into consideration in these forecasts, such as the clinical trial design, where information about the design of the clinical trial, including the objectives, study population, number of study sites, and trial duration, is required to determine the necessary supply levels and plan the clinical supply management process. Another piece of critical information to accurately demand plan is the clinical trial site feasibility assessment, where the CRO predicts patient recruitment, percentage of patients endorsing treatment consent, and primary investigators willing to take on patients.

The limitations with this information on predicting patient enrollment is that CROs depend on their past experiences with specific sites, inability to onboard a site due to various reasons, and high turnover at the CROs where knowledge is also lost. This limits the ability to provide accurate inputs into forecasting tools to best understand where patients will enroll and when, this information is not reliable.

Clinical trial enrollment refers to the process of recruiting and selecting patients to participate in a clinical trial. Enrolling patients in a clinical trial is a complex process that involves several steps and multiple stakeholders. The process begins with the selection of eligible study participants, followed by informed consent, randomization, and initiation of treatment. The first step in the process of enrolling patients in a clinical trial is to identify and select individuals who meet the inclusion and exclusion criteria for the study. The enrollment process is based on a set of predetermined criteria, such as age, sex, medical history, and current health status (Cui et al., 2016). Once eligible participants have been identified, they must provide informed consent to participate in the study. Informed consent is a process by which study participants are provided with information about the study, including its purpose, risks, benefits, and alternatives, and are given the opportunity to ask questions and make an informed decision about participation (FDA, 2020). After informed consent is obtained, eligible participants are randomized into different treatment groups, and randomization is a process that is used to allocate study participants to different treatment groups in a way that is random and unbiased. Once patients have been randomized, they are initiated on treatment. This includes scheduling visits, and administering the study drug or intervention according to the study protocol. Patients are monitored throughout

the trial and followed up regularly to assess the safety and efficacy of the study drug or intervention.

A global clinical trial implies a diverse population with various cultural and social backgrounds; therefore, it is critical to consider these parameters when developing, and accessing enrollment predictions. The enrolment rates in clinical trials can be impacted by several factors. The complexity of the study design can make it more difficult for patients to understand and participate in the trial, which could lead to a low enrollment rate, as they will be reluctant to enroll into a trial that they do not understand. Rigorous eligibility criteria can also impact the number of patients who are able to participate in the trial, as this will disqualify a higher subset of patients. The location of the trial can impact enrollment rates also, as patients may be more likely to participate in a trial that is closer to their home, which has been observed in many global trials, especially with patients who are located in rural areas. Long study duration could also receive push back from a patient, as this would mean they would need to commit to a longer time. Clinical indications can also impact patient enrollment, for example, a trial that targets a rare disease may have fewer potential participants than a trial that targets a more common condition, as rare disease patients are less prevalent. The characteristics of a clinical trial that may impact the enrollment rate are many; therefore, it is important to consider these factors when predicting patient enrollment rates. All of these are contributing factors to a longer clinical trial duration, which significantly increases the cost of a clinical trial.

1.3 Research Objectives

The primary objective of this research is to project a study's enrollment timeline at the portfolio demand planning phase, when there is very little details about the anticipated studies known. This indicates that the studies have not yet seen any actual in-trial enrollment of patients and that no comprehensive patient enrollment planning information is available. Very basic details about the intended studies, such as the total number of participants to be enrolled, and the disease indication, are provided for enrollment prediction purpose. To accomplish the primary objective, it is important to use enrollment data from prior clinical studies to model the enrollment projections as the in trial enrollment data is not yet accessible. For the portfolio demand planning to be successful, it is essential to estimate which specific countries the enrolled subjects will come from as well as when the patients will enroll in each country. Projected trial enrollment and associated costs are crucial feasibility factors that senior management must consider before deciding whether to invest in an asset. A predictive modeling approach capable of offering accurate enough enrollment forecast across all portfolios to aid in the endorsement of management's decision is hugely valuable in any pharmaceutical company.

The objective of this research will contribute to the current literature by accurately predicting patient enrollment at the portfolio demand planning level in clinical trials based on historical patient enrollment data including patient age, patient gender, patient disease, investigational product, study phase, blinded vs. unblinded, sponsor CRO selection, enrollment quarter, and enrollment country. Portfolio demand level planning occurs with a multi-year outlook for the product, which constitutes target indications in which the product will be investigated, as well as strategies for which global markets the product will target to enter. This requires careful planning, as the decisions made as far as production volume will impact the portfolio plan

downstream, since production lead times are long, and procurement of Active Pharmaceutical Ingredients (API) and excipients is challenging due to global supply chain constraints in the current global environment due to COVID-19, and manufacturing personnel shortages. This enhancement of patient enrollment prediction will allow for the optimization of clinical trial supply strategies at the portfolio demand level, and also in a clinical trial, which will minimize clinical trial duration, and will allow for the minimization of total clinical trial production costs, and reduce the overall clinical trial cost.

This will be achieved by the utilization of a Machine Learning model using a Fully Convolutional Network (FCN) to predict the values of enrollment Quarter and enrollment Country. This novel approach to the inputs of patient enrollment will more accurately predict patient enrollment, which will allow for more accurate batch production planning, as the output data will allow demand planning to have insight into when the patients will arrive, and in which country. This will avoid an industry wide issue of over- and under- estimation of clinical trial supply demand.

Chapter 2 – Literature Review

2.1 Analysis of Previous Studies

Several studies have been conducted to investigate the current state of clinical trial supply chain management and identify challenges and opportunities for improvement, especially with that of patient enrollment predictions. The overall theme in the literature with regards to patient enrollment is that, patient enrollment follows an independent non-stationary Poisson process. The forecast demand at each clinical site can be approximated from the record of patients visiting the site and/or the previous data of clinical trials for the same disease. This type of distribution for patients is not reliable, as not all clinical trials experience the same exact constraints. The data can give us an idea; however, there is opportunity to optimize and improve this view of patient enrollment data. Since patient enrollment behavior differs between indications, programs, companies, regions, and products, it cannot be a one size fits all approach; therefore, patient enrollment prediction will need to be viewed exclusively for a specific product, specifically for individual companies (sponsors), in each program. Patient enrollment prediction will be considered on a case-by-case basis, based on the historical data of each program.

A study conducted by Chen et al. (2012) focused on the importance of accurate forecasting in clinical trial supply chain management. The study found that accurate forecasting of patient enrollment, and demand for IMP is critical for effective supply chain management and can help to minimize waste and ensure that the IMP is available when and where it is required. The study also found that improved forecasting techniques, such as simulation modeling, can help to improve the accuracy of patient enrollment predictions and support effective clinical trial supply chain management. This study; however, assumes that patient enrollment demand profile rate is low at the beginning of the trial, and then slows down towards the end of the trial. While this is

the likely scenario, this is not always the case in clinical trial enrollment, as patient enrollment could occur at any time, even at the tail end of the study.

Zhao et al. 2019 also researched this matter, and addressed the challenges of managing time and cost in clinical trials. Clinical trials are essential in the development of new drugs and medical devices, but they are expensive and time-consuming, with many interrelated tasks that need to be coordinated. The authors propose a production planning model that optimizes both time and cost in clinical trials by incorporating multiple objectives. The model aims to minimize the overall completion time and cost of a clinical trial while meeting quality requirements and constraints such as resource availability, patient recruitment, and regulatory compliance. The proposed model uses a mixed-integer linear programming (MILP) approach to optimize the production planning process. The MILP approach allows the model to consider multiple objectives, constraints, and uncertainties simultaneously. The authors use real-world data from a clinical trial to demonstrate the effectiveness of the proposed model. Patient enrollment is predicted by using a non-stationary Poisson distribution, which is the state of the art in predicting patient enrollment, but is not the most accurate, as it does not take into account that each program and company are constrained to specific resources and regions.

Kasenda et al. 2020, study the rationale and design of an international collaborative study, called RECRUIT-IT, which aims to develop, and validate a prediction model for the recruitment of participants in randomized clinical trials. The RECRUIT-IT study aims to address this issue by developing and validating a prediction model that can accurately estimate the time and number of participants needed to recruit in a clinical trial. The objective of the model is to investigate

participant recruitment patterns and study site recruitment patterns and their association with the overall recruitment process. While this information is helpful for site level optimization of supply inventory, this does not address the issue of pooled program supply demand, which has long lead times. The model can aid in allocating supply to specific sites once the program supply is released; however, this will not benefit program/portfolio level demand planning and will not address the long lead times of batch production.

Zhong, Sheng et al. propose a novel method for enrollment forecasting that uses site-level historical data to estimate patient enrollment rates, where their statistical framework is based on generalized linear mixed-effects models (GLMM) and the use of non-homogeneous Poisson processes through Bayesian hierarchical framework to model and predict the country initiation, site activation and subject enrollment sequentially in a systematic fashion, utilizing historical site-level enrollment related data. The method involves three steps: 1. collecting historical data from sites that have participated in similar trials, 2. identifying site-level factors that are associated with patient enrollment rates, such as site location, patient population, and trial complexity, and 3. using a machine learning algorithm to model the relationship between site-level factors and patient accrual rates and make enrollment forecasts. The method is applied to a portfolio of clinical trials in oncology, and compare the forecasts with actual enrollment data. The results show that the method can accurately predict patient accrual rates, with a mean absolute error of 12.6% and a predicted enrollment curve with 95% confidence bands. It is demonstrated that the method can be used to optimize trial portfolios by selecting sites with high enrollment potential and adjusting trial designs to improve patient recruitment. This method is more efficient than the traditional statistical approach which utilizes the simple Poisson-Gamma

model. This study will be beneficial to understanding country start up timelines, site initiations, and more clinical operation as it was based on the historical data of these parameters.

Liu et al. propose a novel machine learning framework for predicting recruitment in clinical trials during the design phase. The goal of this comparison is to predict the number of patients enrolled per month at a clinical trial site over the course of a trial's enrollment duration. They evaluate three approaches to this prediction problem: LightGBM with a tweedie loss function, Zero-Inflated Poisson (ZIP) regression, and a family of hurdle models with Poisson, truncated Poisson, or negative binomial count distributions. The model uses historical trial data and other relevant factors to predict recruitment with greater accuracy. The limitation of the paper is that it does not predict portfolio demand, which would not help support an entire program; however, the model could be used to optimize trial supply after it has been released, and optimize distribution plans to the site at the micro level.

2.2 Research Gap

Overall, there have been many attempts in the literature to use statistical models to predict patient enrollment, and little to no attempts in the literature to use a Machine Learning Neural network for enrollment forecast in portfolio demand planning. The objective of the statistical analyses currently in the literature have been to improve the prediction of patient enrollment based on site level historical data, clinical trial characteristics, and clinical operations data. While these attempts portray a very promising future in utilizing machine learning in predicting patient enrollment and patient arrivals in a clinical trial, it appears that improvement is still needed in the field to better understand when patients will be projected to enroll in a trial, and in which

countries in order to optimize batch production. These two paramters, when and where, and the most critical pieces of information rquired for portfolio planning, as this will define the production plan, and batch allocation.

Since there is little to no published literature for enrollment prediction in portfolio demand planning, and to fill this gap, we propose a novel a novel Machine Learning framework based on a Fully Convolutional Network (FCN) to model and predict patient enrollment by quarter, and country of enrollment. This Machine Learning FCN will predict when patients will enroll and where.

Chapter 3 – Methods

3.1 Problem Definition

The problem which will be solved by this research is the lack of accurate prediction of patient enrollment in the portfolio demand planning stage. The primary objective of this research is to solve this challenge by projecting a study's enrollment timeline at the portfolio demand planning phase, when there is very little details about the anticipated studies known. We will use technology to solve this problem, with a more data driven approach. This will be solved by the development of a Machine Learning Fully Convolutional Network (FCN) that will learn and train on historical data collected from multiple clinical trials collected over a period of three years, and to predict when patients will enroll and where, and as a result be able to forecast when patients will enroll and in which countries for a given trial. Figure 3 presents an overview of the methodology of this research.

3.2 Machine Learning and Neural Networks

Machine learning neural networks, also known as artificial neural networks (ANNs), are a subset of machine learning that model the structure and function of biological neural networks in the brain to solve complex problems. Neural networks consist of interconnected layers of nodes or artificial neurons that process and transmit information. These networks can be trained using large amounts of data to recognize patterns and make predictions. The basic architecture of a neural network consists of input, hidden, and output layers. The input layer receives data from external sources, and the hidden layers process this data to generate output in the output layer. Each node in the hidden layer receives input from nodes in the previous layer and applies a

Figure 3, Methodology

mathematical transformation to this input before passing it on to the next layer. The output layer produces a prediction or classification based on the input data.

The most commonly used neural network architectures are feedforward neural networks, convolutional neural networks, and recurrent neural networks. Feedforward neural networks are the simplest type of neural network, where data flows in only one direction from input to output. Convolutional neural networks are designed to process data with a grid-like structure, such as images or audio, and use convolutional filters to extract features from the input data. Recurrent neural networks are designed to process sequences of data, such as text or speech, and use a loop to pass information from one time step to the next (Goodfellow et al., 2016).

In this study, we develop a Fully Convolutional Network (FCN), which is a type of neural network used for segmentation, which involves dividing a dataset into different regions or segments based on the properties of the data. FCNs were first introduced by Long et al. in 2015, and have since become a popular method for semantic segmentation, object detection, and other segmentation tasks. The main difference between FCNs and other neural networks, such as feedforward neural networks, is that FCNs only use convolutional layers and do not include fully connected layers. This allows the network to take inputs of any size and produce outputs of the same size, which is important for any segmentation tasks. The architecture of an FCN typically includes an encoder and a decoder. The encoder consists of multiple convolutional layers that reduce the spatial dimension of the input image and extract high-level features. The decoder consists of multiple deconvolutional or up sampling layers that increase the spatial dimension of the output and generate the final segmentation map (Ronneberger et al., 2015). Given that this

problem is a segmentation problem, and not a classification problem, FCN's ability to partition a dataset into multiple parts or regions, based on the characteristics of the inputs in the dataset. This is why an FCN was selected for this research, as the variables are of various sizes, and not necessarily connected. Given the binary output of the final output, the FCN was most appropriate. Refer to Figure 4 below for an overall schematic of the neural network design framework.

Between the algorithm's input and output, a hidden layer is present in neural networks. In this layer, the function gives the inputs weights and sends them through an activation function as the algorithm's output. The network's inputs are transformed nonlinearly by the hidden layers. The neural network's hidden layers can vary based on how it performs, and the layers themselves can vary based on the weights that go with them. Hidden layers are a collection of mathematical operations, each of which is meant to produce a certain output that corresponds to a desired outcome. Squashing functions, for instance, are a name for some types of hidden layers. These functions are especially helpful when the algorithm's desired result is a probability because they take an input and output a value between 0 and 1, which is the range used to define probability (Long et al., 2015).

3.3 Study Design and Data Sources

In this study, a Machine Learning model using Fully Convolutional Network (FCN) was trained to predict patient enrollment in multi-centered Global Clinical Trials. These clinical trials are in support of early and late phase portfolio assets in Pharmaceutical Drug Development. The outputs of this model will predict where patients arrive and when, which addresses a massive

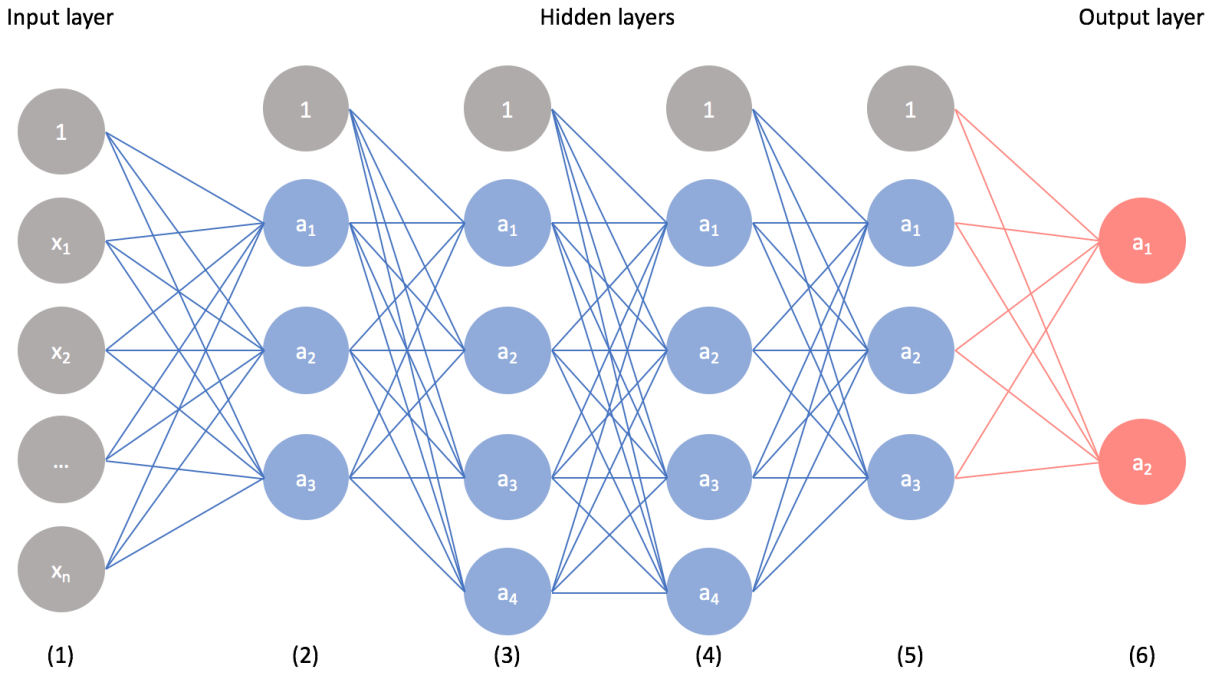

Figure 4, Convolutional Neural Network Schematic

issue faced by Pharmaceutical and Biotechnology companies globally. Figure 3 above describes the overall high-level methodology for this research paper.

The study aims to train a Machine Learning model using a Fully Convolutional Network (FCN) to predict the values of Quarter and Country of a patient based on their Age, Gender, Disease, Product, Phase, Blinded, and CRO values. The model is trained on a dataset of subject records that includes these attributes as well as the Quarter and Country values. The code developed can be adapted to work with other datasets with similar attributes.

This project requires the following dependencies:

Python 3.x

NumPy

PyTorch

Keras

Installation:

Install the required dependencies using pip install -r requirements.txt

Ensure that your dataset is in a CSV format with the following columns: Age, Gender, Disease, Product, Phase, Blinded, CRO, Quarter, and Country.

Update the train_df and test_df variables in train.py to point to the location of your training and testing datasets.

Run train.py to train the model. The trained model will be saved to a file named model_new.pth.

Once the model is trained, you can use predict.py to predict Quarter and Country values for new patients. Update the input_data variable to include the patient attributes for which you want to make predictions.

Run predict.py to generate predictions. The output will be a dictionary with keys "Quarter" and "Country" and corresponding predicted values.

The first step was to prepare the data. The data in this study was comprised of 105,000 data points collected from several multi-centered global clinical trials over a period of three years, or 12 quarters. The data was partitioned into training, and testing/validation sets. 100,000 data points were used for training, and 5,000 were used for testing/validation. This data collected from each patient was Age, Gender, Disease, Product, Trial Phase, enrolled in a Blinded vs. Unblinded study, and CRO (Contract Research Organization). The model is trained on a dataset of patient information that includes these attributes as well as the Quarter and Country values, which are also the target output values. Refer to Table 1 for a hypothetical example of the dataset inputs, and Table 2 for the data input key.

Refer to Table 3 for all variables used in all equations in this section. The dataset is then converted into one hot encoding, which is the conversion of categorical information into a format that may be fed into machine learning algorithms to improve prediction accuracy, mathematically represented as:

$$A\,(x) := \begin{cases} 1, & if\ x \in A \\ 0, & if\ x \notin A \end{cases} \tag{1}$$

Table 1, Hypothetical Example of Dataset Inputs

ID	Age	Country	Gender	Disease	Product	CRO	Blinded	Phase	Date Enrolled	Quarter
90001	65	12	0	1	1	1	0	1	2021-05-25	10

Table 2, Data Input Key

Key	
Age	(18-100)
Gender	(1=male, 0=female)
Disease	(1=BNB, 0=CFH)
Product	(1=ZX-44, 0=CX-55)
CRO	(1,2,3),
Blinded	(0=false, 1=true)
Phase	(1,2,3)
Country	(0-38)
Quarter	(0-11)

Table 3, Variables

Variables	Definition
A	A set
$A\ (x)$	A function of A
n	The number of samples
R	Input layers
$R^{n\times x}$	Input or output layer to the sample size multiplied by the number of inputs
$X \in R^{n\times x}$	Input matrix
$Yc \in R^{n\times x}$	Country output matrix
$Yq \in R^{n\times x}$	Quarter output matrix
Yc	Country output value
Yq	Quarter output value
$[Yc\ Yq\ ]$	Output matrix of country and quarter values
P	The ground truth segmentation map
q	predicted segmentation map
L	Cross-entropy loss
log	The natural logarithm

where A is a set. If x is an element of A, return 1. Else return 0. Thus A is the set of cases that is assigned a 1 to in the encoding vector; therefore, one-hot encoding is a vector form of this indicator function that applies component wise. The data is then converted to tensors, to be further evaluated algebraically as matrices, and segmented more efficiently as it will embedding high-dimensional data into a multi-dimensional array.

Next was defining the model, where the architecture typically consists of an encoder and an activation function. The encoder consists of convolutional layers that extract high-level features from the input image, and the activation function is responsible for processing weighted inputs and helping to deliver an output, virtually decoding layers that reconstruct the segmentation map (Long et al., 2015). The skip connections between the encoder and activation function layers help to preserve spatial information and improve the accuracy of the segmentation results, which is why the FCN is more of an efficient model. The encoder utilized was the 'nn.linear' function, and for the activation function, 'nn.Relu' or rectified linear unit was utilized, to introduce non-linearity into the network. FCNs are models to do nonlinear regression, and they are built up from linear models. They take one of the mainstays of scientific analysis, linear regression, and generalize it to handle complex, nonlinear relationships among data. Figure 5 expresses the framework function on nonlinear regressive neural networks.

This FCN can be expressed mathematically as follows:

Let $X \in R^{n\times 7}$ be the input matrix, where n is the number of samples and 7 is the number of input features, including Age, Gender, Disease, Product, Phase, Blinded, and CRO. Let $Yc \in R^{n\times 1}$ and $Yq \in R^{n\times 1}$ be the output matrices for the predicted country and quarter values, respectively.

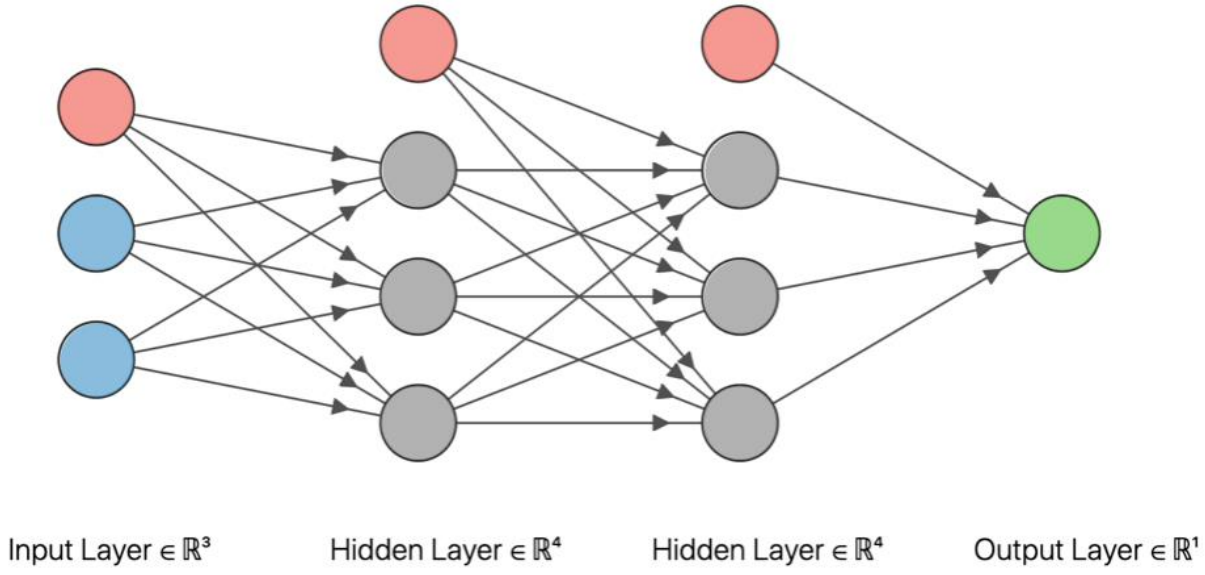

Figure 5, Non-linear Regression Mathematical Framework (Goings 2020)

We can define the FCN-based model as a function $f : R^{n\times 7} \rightarrow R^{n\times 4}$, where the output matrix is comprised of the predicted country and quarter values, represented as *[Yc Yq]*.

The FCN-based model can be represented mathematically as (Goodfellow et al., 2016):

$$[Yc\ Yq] = f(X);\ \text{where } f \text{ is the function implemented by the FCN-based model.} \quad (2)$$

Explanation of the FCN model:

The brackets "[…]" are the concatenation of the predicted country and quarter values, which are represented as separate matrices. This means that the output of the neural network is a matrix with two columns, where the first column contains the predicted country values and the second column contains the predicted quarter values.

In the equation *f(X) = [Yc Yq]*, *f* is the function implemented by the Neural Network, *X* is the input matrix containing the patient features, and *[Yc Yq]* represents the output matrix containing the predicted country and quarter values. Again, we take patient features and outputs a matrix with the predicted country and quarter

Functions can output matrices, just like numbers or other data types. In our model, the Neural Network is designed to output a matrix containing two columns of predicted values, country and quarter. As a result, the output could be a matrix, even though the input is a matrix as well.

After this was defining the loss function and optimizer for training. FCN is trained using the prepared dataset of 100,000 patient data points. The training process involves optimizing the model parameters to minimize the loss function, which measures the difference between the

predicted and ground truth segmentation maps. The optimization is typically done using stochastic gradient descent or one of its variants. ‘optim.Adam’ was selected as the optimizer. Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data, which is critical in validation which occurs internally during training, as it returns training loss and accuracy over time; therefore, so it is performing an iteration of training and then validates, then an additional iteration and validation, and so on and picks the best accuracies of all.

The loss function used for training an FCN depends on the specific application. For segmentation tasks, the commonly used loss function is the cross-entropy loss, which measures the difference between the predicted segmentation map and the ground truth segmentation map. ‘nn.CrossEntropyLoss’ was selected as the loss function. The cross-entropy loss can be written as:

$$L = -sum(p \times log(q) + (1 - p) \times log(1 - q)) \quad (3)$$

where p is the ground truth segmentation map, q is the predicted segmentation map, and log is the natural logarithm. FCNs can be modified in various ways to improve their performance. For example, skip connections can be added to connect the output of an earlier layer to the output of a later layer. This helps to preserve spatial information and improves the accuracy of the segmentation map.

The training accuracy is defined as follows:

Training accuracy = (number of correct predictions on training set) / (total number of training examples); 87,000/100,000 = 0.87

During training the FCN's parameters (i.e., the weight matrices) are adjusted to maximize the training accuracy over a set of training examples, Adam, which iteratively updates the parameters in the direction that improves the accuracy.

After collecting the data, defining the mode, and training parameters, then comes the actual training of the Machine Leaning FCN. The FCN was trained using 100,000 data points of patient and clinical trial data (input values in Table 2).

The outputs of the training are portrayed in Figure 6, Training Loss over time, and Figure 7, Training Accuracy. The training loss shows how the model improves performance over time, and how often it misses the right prediction. Training accuracy is a metric that shows how accurate the training model was.

While training accuracy is useful for monitoring the progress of the model during training, it can be misleading if the model is overfitting the training data. Overfitting occurs when the model becomes too specialized to the training data and does not generalize well to new, unseen data. In such cases, the training accuracy may be high, but the model's performance on a separate validation or test set may be poor; therefore, a test run using a new subset of data is performed to run the model on data the FCN had not seen in the past. For the test, 5,000 data points were utilized, and the model successfully trained on the data, with no error. Figure 8 shows patient

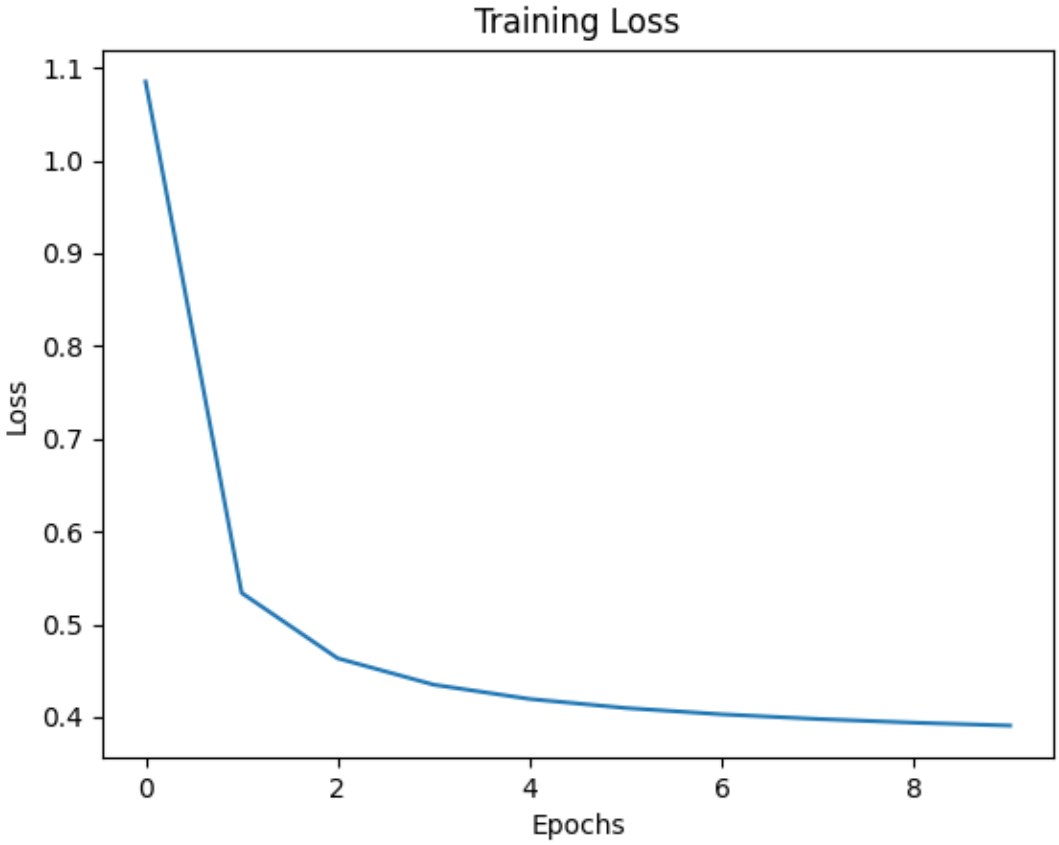

Figure 6, Training Loss Over Time

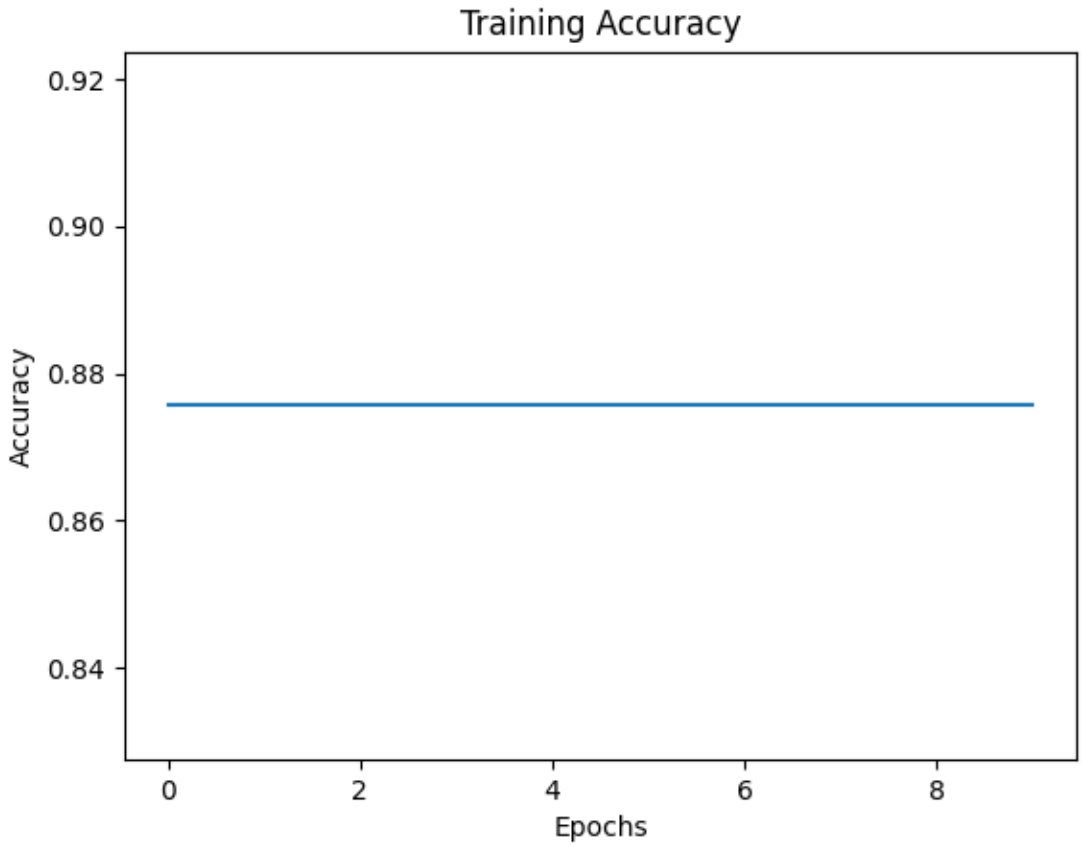

Figure 7, Training Accuracy

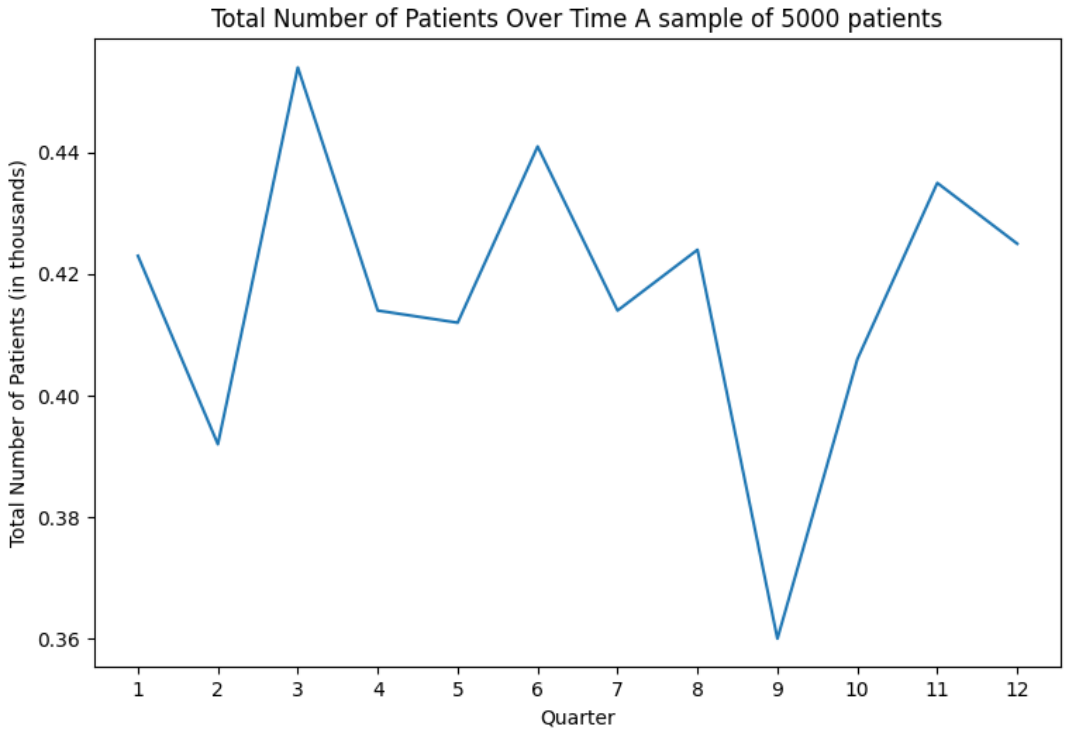

Figure 8, Patient Enrollment Over Time for Data Subset

enrollment over time for the 5,000 patient subset, which shows that the data trained, and predicted the instance successfully.

Chapter 4 – Results and Discussion

4.1 Results

As mentioned in previous sections, this Machine Learning FCN was developed to model patient enrollment in Global Clinical Trials, which will support portfolio demand planning. Portfolio demand planning is planning that occurs with a multi-year outlook for the product, which constitutes target indications in which the product will be investigated, as well as strategies for which global markets the product will target to enter. This requires careful planning, as the decisions made as far as production volume will impact the portfolio plan downstream, since production lead times are long, and procurement of Active Pharmaceutical Ingredients (API) and excipients is challenging due to global supply chain constraints in the current global environment due to COVID-19, and manufacturing personnel shortages. This enhancement of patient enrollment prediction will allow for the optimization of clinical trial supply strategies at the portfolio demand level, and also in a clinical trial, which will minimize clinical trial duration, and will allow for the minimization of total clinical trial production costs, and reduce the overall clinical trial cost.

For purposes of this research, a Phase III, Double Blinded, Multi-Centered, Global Clinical Trial investigating the hypothetical indication “CFH” utilizing product CX-55 enrollment projections, for approximately 1,907 patients, stratified into four age strata. The input parameters into the Machine Learning FCN were as follows:

First age stratum (N = 477):

1) Age: 20
2) Gender: 0 (Female)

3) Disease: 0 (CFH)
4) Product: 0 (CX-55)
5) Phase: 2.0 (Phase III)
6) Blinded: 1.0 (True)
7) CRO: 0 (CRO 1)

Second age stratum (N = 476):

1) Age: 45
2) Gender: 1 (Male)
3) Disease: 0 (CFH)
4) Product: 0 (CX-55)
5) Phase: 2.0 (Phase III)
6) Blinded: 1.0 (True)
7) CRO: 0 (CRO 1)

Third age stratum (N = 476):

1) Age: 60
2) Gender: 0 (Female)
3) Disease: 0 (CFH)
4) Product: 0 (CX-55)
5) Phase: 2.0 (Phase III)
6) Blinded: 1.0 (True)
7) CRO: 0 (CRO 1)

Fourth age stratum (N = 477):

1) Age: 85
2) Gender: 1 (Male)
3) Disease: 0 (CFH)
4) Product: 0 (CX-55)
5) Phase: 2.0 (Phase III)
6) Blinded: 1.0 (True)
7) CRO: 0 (CRO 1)

The outputs of the model predicted where the patients would enroll (country), and during which quarter of a three-year period (which is a typical lookout window in portfolio demand planning). Refer to Figure 9 for Patient Enrollment Per Quarter Per Country (Map). Figure 9 will assist significantly in defining a more robust and optimal distribution plan for the CX-55 product, as it is evident which countries will enroll, and when.

Total patient enrollment per quarter is portrayed in Figure 10, where it is evident that there is a spike in Enrollment towards the end of the three-year defined period (12 quarters). It is typically assumed in most Clinical Trials that enrollment spikes at the start of a trial; however, this assumption is not accurate given the historical data for this product and indication. Figure 11 shows the percentage distribution of enrollment per country and quarter. As shown in the figure, Turkey has the highest rate of enrollment 50.68% of all patients, in the 12^{th} quarter, while South Korea and Singapore tie for the lowest patient enrollment percentage at 0.05%. Based on the data

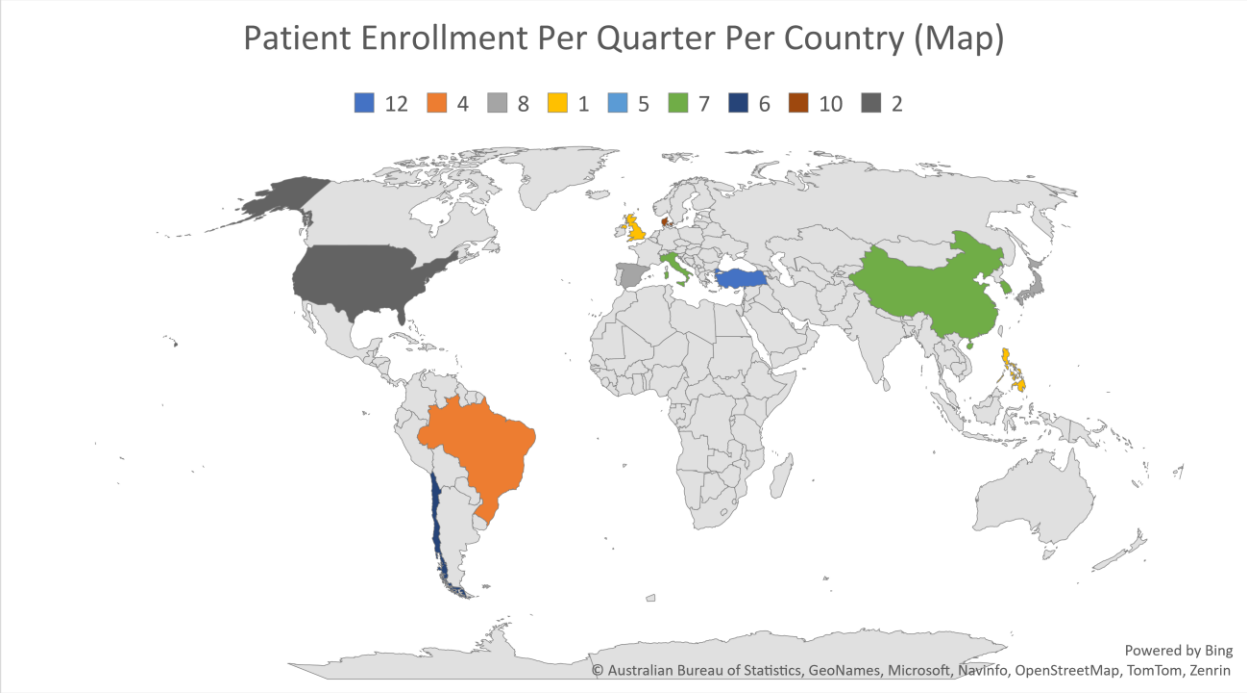

Figure 9, Patient Enrollment Per Quarter Per Country (Map)

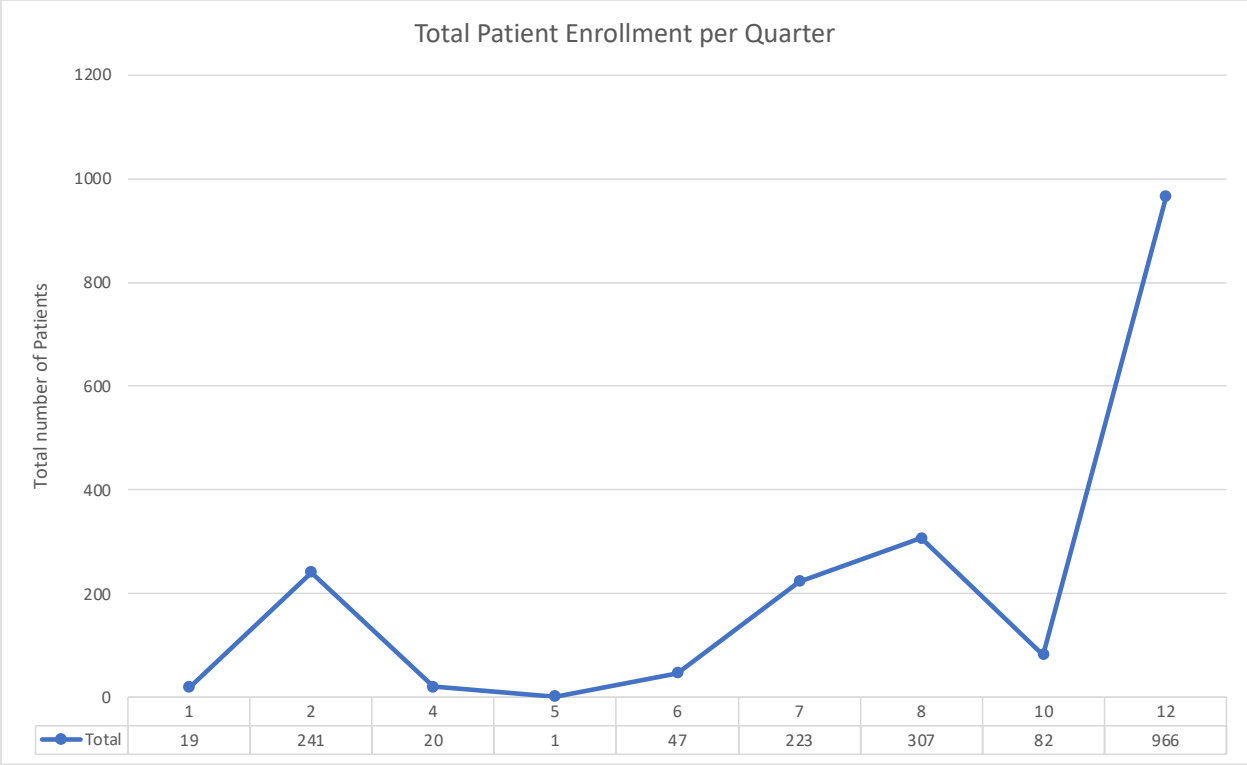

Figure 10, Total Patient Enrollment per Quarter

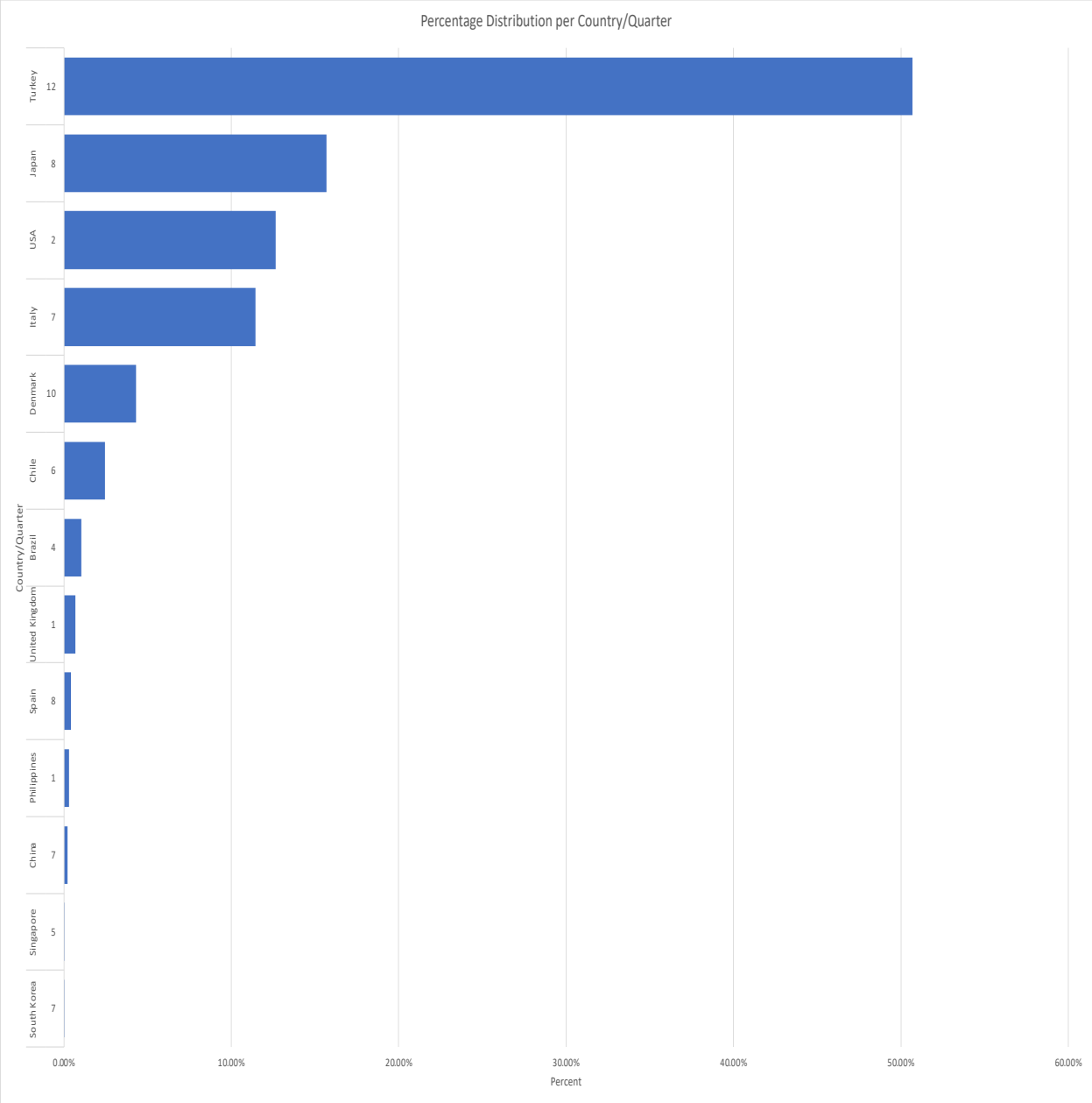

Figure 11, Percentage Distribution per Country/Quarter

from Figure 11, batch release would be prioritized for Philippines, United Kingdom, United States, and Brazil for the first year, as these countries are anticipated to enroll sooner than the rest of the countries in the distribution. Second year prioritization would include Singapore, Chile, South Korea, Japan, and Italy as these countries will enroll patients in quarters 5-8. Lastly, prioritization to batch release would be given to Denmark, and Turkey, where these countries project enrollments to occur in quarters 10 and 12.

Based on Figure 10 data, given that 966 patients, or 50.68% of the total patient population are projected to enroll by quarter 12, this spike in enrollment will allow clinical manufacturing to stagger batch productions in order to be able to product newer batch later in the period, since short-shelf life is a constraint. Production will also need to consider production 13% of the allocated batches for quarter 1 and 2, and ensure not to overproduce to avoid wastes. The data in Figure 10 data can be used to optimize batch production runs, to support the staggering of batch production and release to support patient enrollment. The data in Figure 9 and Figure 12 would be used to optimize distribution strategies, optimizing the use of allocated supplies to specific trials.

4.2 Discussion and Business Insights

There are many benefits for using a Machine Leaning FCN to predict patient enrollment in clinical trials are the portfolio demand planning stage. The two metrics outputted are the quarter of enrollment, and the country of enrollment for the patient in a given trial. In portfolio demand planning, there is little to no data on the trial other than very high-level information such as the phase of the trial, and the target number of patients. When this information can be inputted into a

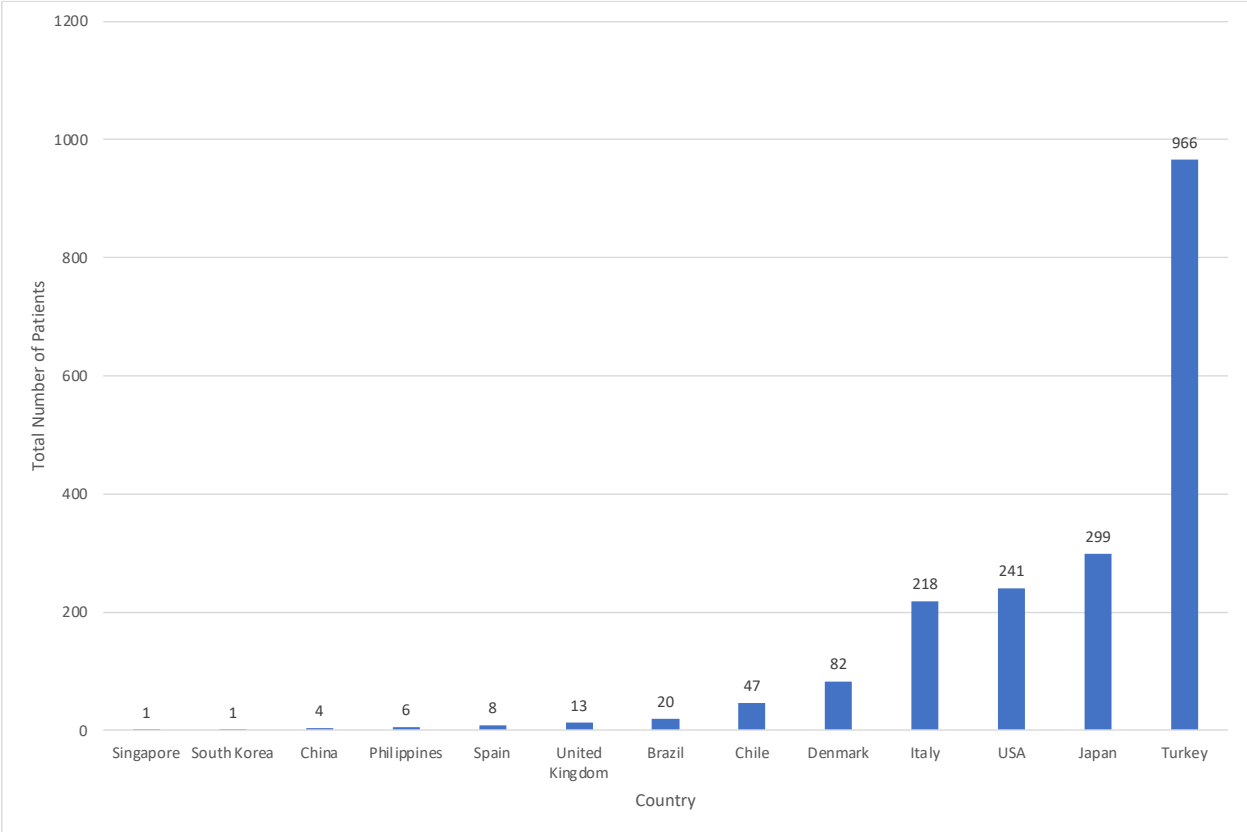

Figure 12, Total number of Patients Enrolled per Quarter

model, and output where and when the patients will enroll, the benefits of these outputs, and metrics can be utilized in multiple functions. There are several business benefits that come with improving the accuracy of this prediction.

Some of the business benefits for pharmaceutical companies of having an improved predictive model for patient enrollment are as follows:

1) **Optimizing Production and Waste Reduction:** Having an accurate prediction of when patients will enroll greatly benefits demand planning in that it optimizes batch productions on the production horizon. Given that the products in question are still in drug development, these products do have a short life, sometimes 3-6 months only. With such short shelf-life, producing batches large enough to cover the span of the full duration of a clinical trial will lead to high level of waste, as most of the drugs will expire within the first few months of the trial. Given the long lead times of 9-12 months for drug production, it is also not the most optimal to prepare drug product just-in-time (JIT), as this strategy could present its own unique challenges in lights of cold chain logistics and would not be able to absorb any delays. In drug development it is best to be proactive, and not reactive. This model will ensure there is sufficient supply is manufactured to support the inventory of the clinical trials.

   A clinical trial can cost an average of $80MM to $150MM, and the cost of wasted clinical supply can range from 10-30% of the total clinical supply budget (~15% of total trial) for a clinical trial. Taking the lower end of the range ($80MM USD), if by

optimizing production based on this prediction model reduces waste by only 1%, pharmaceutical, the potential savings is as follows:

$80MM (Total Trial Budget) × 15% = $12MM (Total Clinical Supply Chain Budget) × 30% = $3.6MM (Total Waste) × 1% = $36,000

companies will save $36,000 for every 1% improvement in enrollment predictions, which is an astronomical saving. In addition, this model will assist in streamlining the internal process for batch allocations. The demand allocations for each batch will be optimized, as patient arrival data will now be more realistic.

3) **Regulatory Submission Strategy Improvement:** Having an improved accurate prediction of where patients will enroll, will allow for regulatory teams to optimize their strategies with country regulatory submissions. The average cost of regulatory submissions to the FDA alone can cost in excess of $3MM (without taking into account head count) for phase 1, 2, and 3 clinical trials across therapeutic areas (Ledesma 2023). If a regulatory team in a pharmaceutical company is aware of which countries will enroll patients sooner than others, then the team can strategize prioritizations between countries, significantly decreasing the risk of mistakes from expediting regulatory applications globally. Mistakes in regulatory submissions could lead to applications rejections, which would prevent a clinical trial from participating in a certain country. The impact of this would cause a barrier to entry to that market, virtually costing a pharmaceutical company billions of dollars in potential missed revenue.

4) **Streaming Internal Processes and Improved Decision Making:**

Projected trial enrollment and associated costs are crucial feasibility factors that senior management must consider before deciding whether to invest further cash in an asset. A predictive modeling approach capable of offering accurate enough enrollment prediction across all portfolios to aid in the endorsement of management's decision is hugely valuable in any pharmaceutical company. This model will improve the decision-making process, as total trial costs will be reduced due to the improvement of patient enrollment projections, which reduces the waste associated, which will make the budget forecast more desirable to the decision makers.

Chapter 5 – Conclusions and Recommendations

One of the biggest challenges the clinical research industry currently faces is the accurate prediction of patient enrollment, as the stochastic behavior of enrollment can significantly contribute to delays in the development of new drugs, increases in duration and costs of clinical trials, and the over- or under- estimation of clinical supply. The primary objective of this research is to solve this challenge by projecting a study's enrollment timeline at the portfolio demand planning phase, when there is very little details about the anticipated studies known.

This was achieved by the utilization of a Machine Learning model using a Fully Convolutional Network (FCN) to predict the values of enrollment Quarter and enrollment Country. This novel approach to the inputs of patient enrollment will more accurately predict patient enrollment, which will allow for more accurate batch production planning, as the output data will allow demand planning to have insight into when the patients will arrive, and in which country. This will avoid an industry wide issue of over- and under- estimation of clinical trial supply demand, which will minimize the total clinical trial cost.

The model trained on 100,000 historical clinical trial data points in two investigational medicinal products, and two therapeutic indications. 5,000 clinical trial data points were used for validation. The primary objective was accomplished at an 87% accuracy rate in the training and validation stages in the Machine Leaning FCN model. The FCN model was able to accurately to predict patient enrollment in a Phase III, Double Blinded, Multi-Centered, Global Clinical Trial investigating the hypothetical indication "CFH" utilizing product CX-55 enrollment projections, for approximately 1,907 patients, stratified into four age strata. This study concludes that

Machine Learning Neural Networks are more accurate than the current methods in the literature and can be used to improve accuracy in projecting patient enrollment in Clinical Trials at the Portfolio Demand Planning Stage.

Recommendations for this study would be for there to be further application for Machine Leaning Models in Clinical Trial data. Pharmaceutical companies are continuously looking for ways to streamline their processes, reduce waste, and optimize their data analytics to improve decision-making; therefore, applying similar approaches to other fields would be beneficial.

List of References

[1] Rodgers, Mark & Singham, Dashi. (2019). A Framework for Assessing Disruptions in a Clinical Supply Chain Using Bayesian Belief Networks. Journal of Pharmaceutical Innovation. 10.1007/s12247-019-09396-2.

[2] Chen, Ye & Mockus, Linas & Orcun, Seza & Reklaitis, Gintaras. (2012). Simulation-optimization approach to clinical trial supply chain management with demand scenario forecast. Computers & Chemical Engineering. 40. 10.1016/j.compchemeng.2012.01.007.

[3] Tufts Center for the Study of Drug Development. 89% of Trials Meet Enrollment, but Timelines Slip, Half of Sites Under-Enroll. Impact Report, 2013, 15(1)

[4] ZHAO, HUI & HUANG, EDWARD & Dou, Runliang & Wu, Kan. (2019). A Multi-Objective Production Planning Problem with the Consideration of Time and Cost in Clinical Trials. Expert Systems with Applications. 124. 10.1016/j.eswa.2019.01.038.

[5] Cui, Zhicheng & Chen, Wenlin. (2016). Multi-Scale Convolutional Neural Networks for Time Series Classification.

[6] Kasenda, Benjamin & Liu, Junhao & Jiang, Yu & Gajewski, Byron & Wu, Cen & Elm, Erik & Schandelmaier, Stefan & Moffa, Giusi & Trelle, Sven & Schmitt, Andreas & Herbrand, Amanda & Gloy, Viktoria & Speich, Benjamin & Hopewell, Sally & Hemkens, Lars & Sluka, Constantin & Mcgill, Kris & Meade, Maureen & Cook, Deborah & Briel, Matthias. (2020). Prediction of RECRUITment In randomized clinical Trials (RECRUIT-IT)—rationale and design for an international collaborative study. Trials. 21. 10.1186/s13063-020-04666-8.

[7] Zhong, Sheng & Xing, Yunzhao & Yu, Mengjia & Wang, Li. (2023). Enrollment Forecast for Clinical Trials at the Portfolio Planning Phase Based on Site-Level Historical Data. 10.48550/arXiv.2301.01351.

[8] Liu, Jingshu & Allen, Patricia & Benz, Luke & Blickstein, Daniel & Okidi, Evon & Shi, Xiao. (2021). A Machine Learning Approach for Recruitment Prediction in Clinical Trial Design.

[9] Tufts. CSDD impact report - 89% of trials meet enrolment, but timelines slip, half of sites underenrol, Tufts Center for the Study of Drug Development, Impact report. v. 15 (1), 2013.

[10] Virtual companies and the pharmaceutical industry. Solem Global. (2021, April 6). https://solemglobal.com/articles/research_reports_articles/virtual-companies-and-the-pharmaceutical-industry/

[11] Mohs, Richard & Greig, Nigel. (2017). Drug discovery and development: Role of basic biological research. Alzheimer's & Dementia: Translational Research & Clinical Interventions. 3. 10.1016/j.trci.2017.10.005.

[12] Mahan, Vicki. (2014). Clinical Trial Phases. International Journal of Clinical Medicine. 05. 1374-1383. 10.4236/ijcm.2014.521175.

[13] National Institute of Health. Understanding Clinical Trials. https://www.clinicaltrials.gov/ct2/about-studies/learn

[14] Goodfellow, I., Bengio, Y.,, Courville, A. (2016). Deep Learning. MIT Press. ISBN: 9780262035613

[15] Schmidhuber, Juergen. (2014). Deep Learning in Neural Networks: An Overview. Neural Networks. 61. 10.1016/j.neunet.2014.09.003.

[16] Long, Jonathan & Shelhamer, Evan & Darrell, Trevor. (2015). Fully convolutional networks for semantic segmentation. 3431-3440. 10.1109/CVPR.2015.7298965.

[17] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[18] "Neural Networks by Analogy with Linear Regression." Joshua Goings, 5 May 2020, https://joshuagoings.com/2020/05/05/neural-network/#:~:text=I'm%20simplifying%20a%20lot,complex%2C%20nonlinear%20relationships%20among%20data.

[19] Ledesma, P. (2022, November 23). How much does a clinical trial cost? Sofpromed. Retrieved March 14, 2023, from https://www.sofpromed.com/how-much-does-a-clinical-trial-cost#:~:text=The%20average%20cost%20of%20phase,median%20of%20%2441%2C117%20per%20patient.

Appendix

```
import torch
import pickle
import os
import pandas as pd

device = ("cuda" if torch.cuda.is_available() else "cpu")

# import training and test data
train_df = pd.read_csv('data_new.csv')
test_df = pd.read_csv('data_new_test.csv')

#train_df = pd.DataFrame(columns=["Age", "Country", "Gender", "Disease", "Product", "CRO",
"Blinded", "Phase", "Quarter"])
#print(train.head())

# drop the ID column
train_df = train_df.drop(columns=['ID', 'Date Enrolled'])
test_df = test_df.drop(columns=['ID', 'Date Enrolled'])

# convert the data to one hot encoding
#train_df = pd.get_dummies(train_df,
columns=['Age','Country','Gender','Disease','Product','CRO', 'Blinded', 'Phase', 'Quarter'])
#test_df = pd.get_dummies(test_df, columns=['Age','Country','Gender', 'Disease', 'Product',
'CRO', 'Blinded', 'Phase', 'Quarter'])

# Split the data into features and labels
train_features = train_df.drop(columns=['Country', 'Quarter'])
train_labels = train_df[['Country', 'Quarter']]

test_features = test_df.drop(columns=['Country', 'Quarter'])
test_labels = test_df[['Country', 'Quarter']]

# convert the data to tensors to be used in the model
train_features = torch.from_numpy(train_features.to_numpy()).float()
train_labels = torch.from_numpy(train_labels.to_numpy()).float()
test_features = torch.from_numpy(test_features.to_numpy()).float()
test_labels = torch.from_numpy(test_labels.to_numpy()).float()

# define the model
import torch.nn as nn

class FCN(nn.Module):
    def __init__(self):
```

```
        super(FCN, self).__init__()
        self.fc1 = nn.Linear(7, 10)
        self.fc2 = nn.Linear(10, 52) # change the number of output nodes to 52
        self.relu = nn.ReLU()

    def forward(self, x):
        x = self.relu(self.fc1(x))
        x = self.fc2(x)
        return x


net = FCN()

# define the loss function and optimizer
import torch.optim as optim

criterion = nn.CrossEntropyLoss()
optimizer = optim.Adam(net.parameters(), lr=0.001)


train_losses = [] # to store training loss values
train_accs = [] # to store training accuracy values
"""
# train the model
for epoch in range(1000):
    # forward pass
    outputs = net(train_features)
    loss = criterion(outputs, torch.argmax(train_labels, dim=1))

    # backward pass
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()

    # print the loss every 100 epochs
    if (epoch + 1) % 100 == 0:
        acc = (torch.argmax(outputs, dim=1) == torch.argmax(train_labels, dim=1)).float().mean()
        print('Epoch [{}/{}], Loss: {:.4f}, Accuracy: {:.4f}'.format(epoch + 1, 1000, loss.item(),
acc))
        train_losses.append(loss.item())
        train_accs.append(acc.item())
"""

# save the model
torch.save(net.state_dict(), 'model_new.pth')
```

```
# test the model
with torch.no_grad():
    outputs = net(test_features)
    predicted = torch.argmax(outputs, dim=1)
    correct = predicted.eq(torch.argmax(test_labels, dim=1)).sum().item()
    #print('Accuracy: {}/{} ({:.0f}%)'.format(correct, len(test_labels), 100. * correct /
len(test_labels)))


quarters_map = {
    0: 1,
    1: 2,
    2: 3,
    3: 4,
    4: 5,
    5: 6,
    6: 7,
    7: 8,
    8: 9,
    9: 10,
    10: 11,
    11: 12
}

# predict the output
# predict the output for a single instance
def predict(model, instance):
    model = model.to(device)
    # convert instance to a tensor and reshape it
    instance = torch.Tensor(instance).reshape(1, -1).to(device)

    # pass the instance through the model
    outputs = model(instance)

    # get the predicted class probabilities
    probs = torch.nn.functional.softmax(outputs, dim=1)

    # get the predicted class labels (countries have to be mapped to the correct index)
    preds = torch.argmax(probs, dim=1)


    # get the predicted country and quarter
    country = preds[0].item()
    quarter = quarters_map[country % 12]
```

```
    return countries_map[country], quarter


model = FCN()
# example usage



# predict country and quarter for the test data
# ID,Age,Country,Gender,Disease,Product,CRO,Blinded,Phase,Date Enrolled,Quarter
#1,58,19,1,1,1,3,0,3,2021-05-21,10
# based on this list:
"""

countries = ['Argentina', 'Austria', 'Belgium', 'Brazil', 'Canada', 'Chile', 'China', 'Colombia', 'Czech
Republic',
             'Denmark', 'Finland', 'France', 'Germany', 'Greece', 'Hungary', 'India', 'Indonesia',
'Ireland', 'Italy',
             'Japan', 'Mexico', 'Netherlands', 'Norway', 'Peru', 'Philippines', 'Poland', 'Portugal',
'Russia', 'Singapore',
             'South Africa', 'South Korea', 'Spain', 'Sweden', 'Switzerland', 'Thailand', 'Turkey',
'United Kingdom', 'USA',
             'Vietnam']

"""
countries_map = {
    0: 'Argentina',
    1: 'Austria',
    2: 'Belgium',
    3: 'Brazil',
    4: 'Canada',
    5: 'Chile',
    6: 'China',
    7: 'Colombia',
    8: 'Czech Republic',
    9: 'Denmark',
    10: 'Finland',
    11: 'France',
    12: 'Germany',
    13: 'Greece',
    14: 'Hungary',
    15: 'India',
    16: 'Indonesia',
    17: 'Ireland',
    18: 'Italy',
    19: 'Japan',
```

```
    20: 'Mexico',
    21: 'Netherlands',
    22: 'Norway',
    23: 'Peru',
    24: 'Philippines',
    25: 'Poland',
    26: 'Portugal',
    27: 'Russia',
    28: 'Singapore',
    29: 'South Africa',
    30: 'South Korea',
    31: 'Spain',
    32: 'Sweden',
    33: 'Switzerland',
    34: 'Thailand',
    35: 'Turkey',
    36: 'United Kingdom',
    37: 'USA',
    38: 'Vietnam'
}

instances = [[90.0, 1.0, 0.0, 0.0, 2.0, 0.0, 0.0]]
predictions = predict(model, instances)
print(predictions[0], predictions[1])

import random
# create a file csv to store the results
file = open('results.csv', 'w')
file.write('Country,Quarter\n')
for i in range(300):
    instances = [[random.randint(0, 100), random.randint(0, 1), random.randint(0, 1),
random.randint(0, 1), random.randint(0, 2), random.randint(0, 1), random.randint(0, 1)]]
    predictions = predict(model, instances)
    print(instances, predictions[0], predictions[1])
    # make a header for the csv file, two column, first is country, second is quarter
    file.write(predictions[0] + ',' + str(predictions[1]) + '\n')
```

Vita

Ahmed Ahmed Mohamed Shoieb was in Knoxville, Tennessee to parents Ahmed Mohamed Shoieb and Mona Ahmed Elgayyar. He moved to Cairo, Egypt, in 2004 and attended elementary, middle, and part of high school until 2011, when he moved back to Tennessee. He graduated with honors from Bearden High School in Knoxville in May 2012, while already having started his undergraduate career early, in January 2012 at the University of Tennessee, Knoxville. He graduated with a Bachelor's degree in Neuroscience and Biochemistry in December of 2011. After graduation, Ahmed started his career in the Pharmaceutical industry, and then began his Master's degree in Industrial and Systems Engineering at the University of Tennessee, Knoxville in 2020. He plans to graduate this May and looks forward to pursuing his PhD while continuing a career in the Pharmaceutical industry.