Improving Biopharmaceutical Manufacturing Yield using Neural Network Classification

Will Fahey and Paula Carroll

Centre for Business Analytics, Smurfit School of Business, University College Dublin, Ireland

Abstract: Traditionally, the Six Sigma framework has underpinned quality improvement and assurance in biopharmaceutical manufacturing process management. This paper proposes a Neural Network (NN) approach to vaccine yield classification. The NN is compared to an existing Multiple Linear regression approach. This paper shows how a Data Mining framework can be used to extract further value and insight from the data gathered during the manufacturing process as part of the Six Sigma process. Insights to yield classification can be used in the quality improvement process.

Keywords: Neural Network Classification Systems; Manufacturing Management Improvement Processes; Six Sigma; CRISP-DM; Biopharmaceutical; Yield Improvement.

1. Introduction

The World Health Organization states that pneumococcal disease is the world's number one vaccine-preventable cause of death among infants and children under age five [21]. Vaccines are a crucial resource in the fight to lower infant mortality rates for a developing country [16], coming second only to clean drinking water. However, the vaccine manufacturing sector is quite fragile due to strict regulatory licensing and cost concerns [19]. Approaches to public health policies that could contribute to the sustainability of the vaccine manufacturing sector are outlined in [19]. The focus in [16] is on price bundle determination for combination vaccines which maximise social good (i.e. ensure sufficient vaccine is produced) while also ensuring minimum profit levels for manufacturers to ensure their long term viability.

This article focuses on a different idea that may contribute to ensuring the long term viability of the sector: exploiting value from the data gathered for licensing compliance and operations management.

1.1. The vaccine manufacturing process

A vaccine is typically made up of a number of individual polysaccharide components called serotypes – each of which immunise against a particular strain of the targeted disease [3]. Manufacturing pneumococcal vaccine is a complicated procedure, involving the use of bioreactors to manipulate cells to produce the various active biological substances [10]. A bioreactor is a vessel used to replicate the conditions found in a mammalian body to promote the creation of the biological components that combine to form the vaccine product. These components are passed through various ultrafiltration steps which purify the product and diafiltration steps which concentrate the product to desired levels. The process is categorised by long lead times of up to 30 days. To manufacture one batch may involve between 40 and 50 process steps. These process steps may be characterised by explanatory variables (EVs) such as measured temperature, pressure and flow rates at various points during the production process.

The vaccine manufacturing process involves the combination of a saccharide component which elicits the immune response to a carrier. The success of this combination is measured as the yield. *Yield* is defined in this study as the amount of vaccine created in a production batch as a percentage of the expected amount based on the quantity of raw materials used. The dynamic nature of biological components used in vaccine manufacturing renders static methods of measurement only indicative. This adds to the complexity of identifying root causes of yield fluctuation.

Traditionally, Six Sigma (6σ) approaches such as Design of Experiments (DOE) and Statistical Process Control (SPC) techniques have been used to improve yield and decrease variability [18]. The method utilised by 6σ to achieve Quality Improvement (QI) is to generate hypotheses about which EVs impact on quality through brainstorming. Then statistical methods such as regression or hypothesis testing are used to confirm or disprove these hypotheses. Such analyses provide the necessary feedback to product/process design (or re-design) and other corrective QI actions [13].
6σ is the incumbent approach used by company Z to identifying the root cause of poor yield. The vaccine manufacturing process has multiple biological inputs, each with multiple quality characteristics that may potentially explain yield fluctuation. There are many combinations of measurement equipment settings and possible EVs so it is often unclear how the inputs interact and how the multiple measurement settings affect process outputs. However, if the 6σ process is taken to its theoretical conclusion, a potentially exponential number of hypotheses could be generated during the measure phase of a complex issue such as yield variability. There are also considerations of inherent human bias that might influence the identification of a possible root cause in brainstorming sessions during the 6σ measure phase.
Large amounts of data are generated and collected by automated manufacturing processes, most of which are used for process control rather than process improvement [6]. This paper proposes a novel approach to extracting real business value from the wealth of data that has already been gathered. The method focuses on an evaluation of Neural Networks (NN) to generate and test hypotheses about which process parameters (or combinations of parameters) lead to a high or low yield. A hypothesis in this case is that a process parameter setting contributes to yield fluctuation. Two serotypes, referred to as Serotype X and Serotype Y, are the subject of this study which was undertaken using manufacturing data from a company given the pseudonym Z.

2. Vaccine Manufacturing Challenges and Opportunities

Biopharmaceutical manufacturing is one of the most heavily regulated industries in the world today. Regulatory bodies such as the FDA (American Food and Drug Administration) have been relentless in driving higher levels of process control and understanding in the biopharmaceutical sector. These bodies recognise the significance and untapped potential of data mining methods to enable more robust biological manufacturing processes through increased process knowledge. Analysis of manufacturing data using Multivariate Data Analysis (MVDA) was stimulated by the FDA's landmark guidance on ''Process Analytical Technology'' in 2004 [8]. Regulatory authorities are demanding a greater level of process characterisation and robustness in the biopharmaceutical industry as a means of ensuring consistent supply of safe, efficacious medicines to patients. However there remains a gap between the huge quantities of manufacturing data available and how much knowledge the industry derives from this data [15].
Regular changes to the production processes are inevitable in a manufacturing industry particularly when the strong culture of continuous improvement inherent in 6σ exists. However every change involves risk. Quantified risk assessment can only be effective in mitigating this risk when the process is sufficiently well understood. Data Mining then becomes an essential tool in assessing the impacts of changes to critical process parameters to downstream operations [23].
Some of the challenges faced by the vaccine manufacturing industry are outlined next. These challenges can also be interpreted as data mining opportunities [20], the challenges include:

- A high number of possible EVs including many statistical measures associated with input components and process stage metrics such as temperature and pressure: for complex

processes it is natural to have a large number of EVs to ensure adequate description of the process. This is especially true for biological manufacturing processes.

- A high number of dependencies: the number of dependencies to be modelled increases when several components are integrated into one system. However it is not only the high number of statistically proven dependencies that require significant resources to model but also potential dependencies that have to be accepted or rejected as contributing to an improved model. This calls for an efficient way of pruning hypothesised relations. Inherent yield variability (referred to in 6σ as Common Cause Issues) is very rarely attributable to a single input value or process setting. It is much more likely that interdependencies between EVs conspire together to produce a low yield.
- Uncertainty of measurement data: While the proportion of manually recorded data is getting smaller it is still present, meaning the possibility of transcription error still exists. Methods to capture the uncertainty associated with autocapture of other manufacturing data also aim to quantify doubt about the validity of the result of a measurement including sampling, precision and possible calibration errors.
- Incomplete information: This is a common problem when using raw manufacturing data. Values are sometimes deemed unimportant to the process outputs and due to resource constraints are not gathered fully. DM has an advantage over traditional statistical methods as it offers intelligent ways of replacing missing values like *k*-means clustering [1].

With missing data, statistical tests can lose power, results can be biased, or analysis may not be feasible at all. With missing value imputation, missing values are replaced with estimated values according to an imputation method or model. In the *k*-Nearest Neighbour (*k*-NN) method, a case is imputed using values from the *k* most similar cases. *K*-NN is a non-parametric lazy learning algorithm. Nonparametric means that it does not make any assumptions about the underlying data distribution. This property is useful in this case study as the data does not necessarily allow typical theoretical assumptions, such as following a typical distribution such as normal or exponential.
Lazy refers to the fact that the algorithm does not use the training data points to do any generalisation. In other words, there is no explicit training phase. This speeds up the algorithm, making it practical to use in one of the nested operators in the DM process.
This allows a stronger model than simply replacing each value with the mean of the other values. The approach to missing data is used successfully by [12]. This approach illustrates another advantage DM techniques have over 6σ - regression cannot be performed with missing values and there is no direction on how to deal with missing values.

2.1. The Six Sigma and *Cross Industry Standard Process for Data Mining* methodologies

Quality Improvement (QI) programmes aim for improvements in manufacturing yield using the define-measure-analyse-improve-control (DMAIC) approach to reach 6σ quality levels (less than 3.4 defects per million opportunities). Each project in the 6σ methodology has five phases, represented by the initials DMAIC. An overview of each phase is given next:

- *Define* the nature of the problem and frame the problem statement. Make sure this aligns with the project sponsor's outlook on the issue. Map the process to ensure consensus.
- *Measure* key aspects of the current process and collect relevant data. This involves visualising and investigating the data to provide insight and potential root causes of the issue. Use these as a benchmark for brainstorming all potential root causes of the issue.

- *Analyse* the data to investigate and verify cause-and-effect relationships. Use statistical techniques to rule in or rule out the potential root causes. Techniques include regression and hypothesis testing.
- *Improve* the confirmed root causes by error proofing out the issue. Set up pilot runs to establish process capability.
- *Control* by piloting the future state process to ensure that any deviations from target are corrected before they result in defects. Implement control systems such as statistical process control and monitor the process to make sure the improvements are effective.

The 6σ process has many advantages: the goals are clear and defined from the outset and the structure and sequential nature provides a common language so that stakeholders from every level can understand the problem and how it will be solved.
DMAIC also provides a data-driven structure to a diverse team of subject matter experts (SMEs) who each bring expert but possibly biased understanding to the root cause identification process. In the absence of the DMAIC structure, SMEs may jump to premature conclusions based on their own process experience.
Wu [27] points out that classical methods such as control charts aim to monitor the process and not to infer the relationship between the target attribute, input attributes and most importantly outputs. Büchner [5] elaborates on the shortcomings of retrospective statistical methods, stating that they considerably limit the potential for continuous process improvement.
CRISP DM (*Cross Industry Standard Process for Data Mining)* is the de facto industry standard process methodology for Data Mining. The process was inspired by the 6σ DMAIC methodology and a need identified by practitioners to allow DM be adopted as a key part of business processes [25]. It is an iterative adaptive hierarchical process based on real-world experience of how people conduct DM projects and provides an overview of the life cycle of a DM project. The CRISP DM process framework defines six phases of a DM project, their respective tasks, the relationships between these tasks and deliverables of each phase. A brief outline of the phases is given next:

1. *Business understanding:* This initial phase focuses on understanding the project objectives and requirements from a business perspective. This phase is comparable to the *Define* phase of a Six Sigma project, where a plan is formed and the project goals are reviewed by the project sponsor.
2. *Data understanding**:*** The data understanding phase starts with initial data collection and proceeds with identification of data quality problems. Some early Exploratory Data Analysis is also carried out in order to gain an initial impression of the possible relationships present in the data. This can be compared to the preliminary stage of the *Measure* phase of a 6σ project.
3. *Data preparation*: The data preparation phase covers all activities needed to construct the final data set from the initial raw data. This includes dimensionality reduction, dealing with missing values, data normalisation and dealing with outliers. This phase is not usually required during a Six Sigma project.
4. *Modelling*: In this phase, modelling techniques are selected and applied, and their parameters are calibrated to optimal values.
5. *Evaluation*: The practical applications of the model are evaluated. Before proceeding to final deployment of the model, it is important to thoroughly evaluate and review the steps executed to create it, to be certain the model properly achieves the business objectives. Any risk to applying the model must also be assessed.
6. *Deployment*: Creation of the model is not the end of the project. The knowledge gained will need to be translated to a format that the customer can use and understand.

Figure 1 shows the make-up of the type of team required to complete an analytics manufacturing project [5]. Three skillsets are essential in building a team for a data mining project in the manufacturing domain: a domain expert, a data expert and a DM expert.
The domain expert in this study belonged to the technical operations group and had significant experience with the manufacturing process and 6σ statistical techniques.
Ideally the data expert should belong to the IT department and have experience with relational databases. This proved to be the case in this study. The data expert was an automation engineer with experience in querying databases using SQL. This point is expanded in the discussion section under opportunities for future work.
Data mining is usually carried out in large organisations, however a domain expert who is also an expert in the data stored by the organisation is rare. Often the data mining expert is a consultant with no knowledge of the manufacturing process (which is a distinct disadvantage).The team for this project was fortunate in that the data mining expert had experience in 6σ techniques and data retrieval using SQL, and was also familiar with the manufacturing process at a high level.

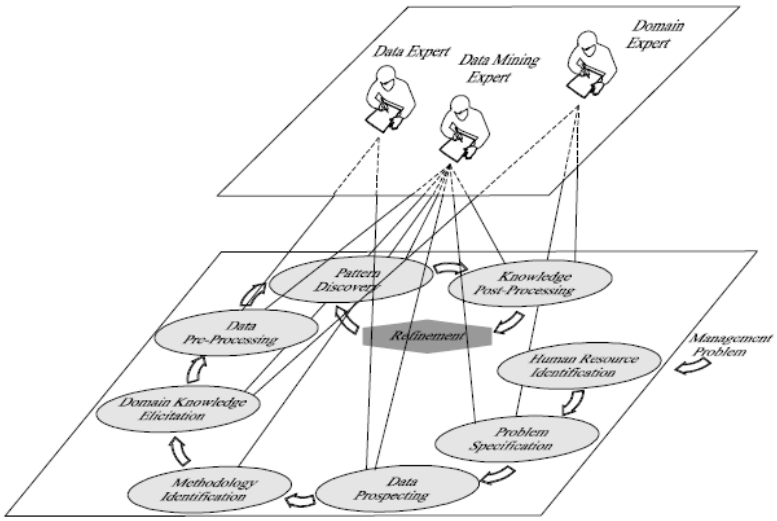

Figure 1: Team skillsets, [5]

2.2. Data Mining and Analytics Opportunities

Machine Learning extends the use of DM through the use of algorithms that learn patterns from the data. Machine learning approaches, such as NNs, are well equipped to deal with the range of problems outlined in Section 2. In many cases NNs are used for modelling complex non-linear relations with a large number of EVs, see for example [11]. Chien et al [7] show how NNs can also adapt dynamically to changes occurring in the modelling system in real time. This is essential for manufacturing applications. Even though the initial training results may not be accurate, the NN performance improves with time as more training data samples are provided.
One advantage of using a NN is that they can be fitted to any kind of data set, they do not require the relationships in the model to be explicitly stated. NNs are particularly suited when data may be noisy and relationships may be non-linear such as the data set in this study. Because of the complexity and non-linearity involved in vaccine manufacturing systems, such systems lend themselves well to the use of NNs, benefiting from the NN online learning and adaptive abilities. NNs are criticised for being a black box but have demonstrated their usefulness in many practical applications within the manufacturing sector [13].

NNs are a supervised learning approach designed to model the method by which human brains accomplish a certain task. Tetko et al gives some characteristics of NNs that have led to their widespread use [22]. A NN can learn by adjusting the topology (also called architecture or structure) and edge weights of a network connecting certain input signals to a desired output response. Such a training process is an iterative process which is run until no further adjustment is required. Once a NN has been designed based on training data set, it can then be tested and evaluated on a test data set. NNs can be used for classification or prediction tasks.
In this paper, a NN is used to classify a production batch as high or low yield depending on the values of the manufacturing production process EVs.

2.2.1. Cross Validation of the model

The cross validation method involves repeated training of the neural network using a number of partially-overlapping arbitrary large portions of data as the training sets, with the remainder of the data in each case being used as the independent test set. In this way, all data will eventually be used in the test set, and errors due to the inclusion of non-representative data in either set are avoided. This is effective but computationally expensive.
A validation set is used either to refine the topology of the network or to serve as a stopping criterion. NN topology design parameters such as the number of units in a hidden layer or the number of hidden layers determine the structure of the network.
In the first methodology, the network designer assesses the performance of different trained networks by evaluating an objective function using the validation set. The network with the smallest error is selected. In the second approach, training and validating take place concurrently. The network stops learning once the sum of residuals, based on the validation set, starts to increase beyond a user-specified number of iterations. A testing set is later used to avoid overfitting where the network has learned the noise in the training data and is no longer useful to generalise to unseen data [9].
The accuracy measure for evaluating the performance of classifiers is defined as:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

In this study which predicts high or low yield, a true positive is a high yield production batch that is correctly identified by the classifier as high yield. Accuracy is then defined as the number of correct (or true) high and correct low yield predictions divided by the total number of tests.
Precision is another measure used in ML quality assessment to measure the ability of the classifier to make positive predictions correctly. Precision of the classifier to identify high yield is defined for this study as the number of correct high yield predictions divided by the total number of high yield predictions (both correct and incorrect). A similar low yield precision measure is also defined.
Recall is a measure used to quantify the sensitivity of an ML classification system. In the case of high yield classification, it is defined as the number of correct (true) high yield predictions divided by the sum of the correct high and incorrect low yield predictions. A similar low yield recall measure is also defined.
The design of a NN is more of an art than a science. There is no unified approach on setting the design parameters of a NN. Zobel gives a good overview of selecting the design parameters of NNs [28]. The general approach is one of trial and error to change the design parameters and note if it has an effect on the performance of the model. The NN design parameters include:

- Hidden Layers: This parameter describes the number and size of all hidden layers. The user can define the structure (network topology) of the neural network with this parameter. The

hidden layer links the input layer to the output layer. Within each node in the hidden layer a weighted sum calculation is carried out relating the input layer to the output using a predefined function.

- Training Cycles: This parameter specifies the number of training cycles used for the neural network training. In a back-propagation approach, the output values are compared with the correct answer to compute the value of some predefined error-function. The error is then fed back through the network. Using this information, the back-propagation algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. This training process is repeated a number of times.
- Learning Rate: This parameter determines how much the weights are changed at each step.
- Momentum: The momentum adds a fraction of the previous weight update to the current one. This prevents local maxima and smooths optimisation directions.

2.2.2. Principal Component Analysis and Data Reduction

Bellmans "curse of dimensionality" indicates that a large number of explanatory variables and a small number of batches (or samples) can lead to a poor model. In this study this problem occurs, many possible explanatory variables are available for a number of production runs. A Time Series Plot (TSP) shows a sequence of observations of variables of interest such as temperature, pressure and flowrate during the manufacturing process steps. TSP data can have extremely high dimensionality because each time point can be viewed as a single dimension (giving a tuple of values for the EVs). High dimensionality can lead to an over-fitted ML model and the raw time series may be computationally too expensive to process during the NN training stage, so the number of dimensions must be reduced [25].

One of the challenges faced early in the data preparation phase of this study was deciding how to deal with time series (TS) data from the pivotal filtration/concentration steps of the manufacturing process. It was essential to add this data to the model for consideration, however distributions of the values over time were erratic and did not fall into a recognised pattern (e.g. binomial, log, normal). To achieve data dimension reduction, the distribution of each EV (for example, temperature) was represented by a number of descriptive statistics values such as the "moments" of the EV (for example, the mean is the first moment). These statistics were then passed as the inputs to the NN. The descriptive statistics method of data reduction was motivated by Bickel [2] as a means to summarise a non parametric model. Bickel recommends the following group of statistics:

- Mean: The average of the values.
- Standard Error: The standard deviation of the sampling distribution of a distribution. Standard Error of the Mean (SEM) is calculated by dividing the standard deviation by the root of the number of observations.
- Median: The median is the number in the middle of a set of numbers: half the numbers have values that are greater than the median, and half have values that are less.
- Mode: The most frequently occurring or repetitive value in an array or range of data.
- Standard Deviation: The standard deviation tells us how much variation is present in a distribution.
- Trimmed Mean (20%): Trimmed mean discards the top 10% and lowest 10% of values. This was included to account for a large number of outliers. A significant number of outliers can be identified by comparing this value to the mean.

- Kurtosis: Kurtosis characterises the relative “peakedness” or flatness of a distribution compared with the normal distribution. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution.
- Skewness: Skewness characterises the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.
- Maximum: The maximum value recorded.
- Minimum: The minimum value recorded.
- Quartile 1: Q1 is the "middle" value in the first half of the rank-ordered data set.
- Quartile 3: Q3 is the "middle" value in the second half of the rank-ordered data set.
- Interquartile Range: The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles and are denoted by Q1, Q2, and Q3 respectively.

Principal Component Analysis (PCA) [26] is a multivariate dimension reduction technique applicable to large datasets. The set of possibly correlated EVs is reduced to a set of linearly uncorrelated principal components values. PCA identifies a vector similar to a basis that uncovers the underlying structure in the data. They are mathematical constructs that point in the direction where there is the most variance. PCA takes into account the combined contribution to the variation of a number of vectors as opposed to the univariate point of view represented by a correlation matrix.

3. Methodology

The incumbent baseline MLR used by company Z to identify yield improvements models yield based on a correlation factor for univariate relationships within the dataset. The NN model is compared with the incumbent MLR model using root mean squared as the error metric.

3.1. CRISP_DM implementation

For the NN method, TSP data is first summarised using the measures suggested in [2]. PCA is then used (for dimension reduction) to identify a minimal set of prioritised EVs. A NN model is then created to relate the identified EVs to manufacturing yield output. The CRISP_DM framework is used to support the implementation of the study, a summary of the CRISP_DM phases of this study follows:

1. *Business understanding:* As this methodology was a new approach for company Z, significant ground work had to be completed to ensure management buy in. This involved presentation of the methods as win-win as there was little or no capital required and no new data was required - historical data was re-used. Manpower resources whose skillsets were complementary to the projects requirements were identified – in this case a mix of process and statistical knowledge with an interest in process modelling were required. Due to the company Z’s unfamiliarity with the CRISP_DM methods it was difficult to assign a realistic goal for yield improvement. Company Z agreed to an initial study as proof of concept of the CRISP-DM process under the banner of a company-wide innovation initiative.

2. *Data understanding:* The data was extracted from multiple data sources and assembled into a format that could be read by the DM software. This was the most time consuming part of the process as most of the data preparation tasks were completed manually. Missing values were imputed using the *k*-means clustering algorithm.
3. *Data preparation*: The data from each database was combined into one data set, TSP data was summarised and PCA applied to reduce the data to the most significant EVs. The RapidMiner software platform was used for these tasks. RapidMiner is an integrated software tool for ML and DM applications.

   Each Principal Component (PC) explains a certain proportion of the variation in the target variable (yield). Each PC is a mathematical construct, for the dimension reduction filtering effect to be useful from a practical point of view the PCs are not used in the NN model but are related back to original EVs [14]. Each PC is correlated with a number of EVs which the RapidMiner software ranks in order from highest to lowest. The grading for the study data showed there was a natural drop off in correlation at 40 EVs. This process produces a prioritised dataset which is then passed to the NN for the modelling phase. For example in this study, a PC that explains 50% of the yield variation then contributes 20 EVs to the pool of 40. This is discussed further in Section 4.
4. *Modelling*: The 40 EVs identified during the data preparation step are used to create a NN model.
5. *Evaluation*: The output of the model shows an equation that gives each variable a correlation coefficient to illustrate how it relates to the other variables when yield is at the optimum. These findings are validated by the domain expert.
6. *Deployment*: Creation of the model is not the end of the project. The settings for EVs are implemented on the manufacturing floor, normally under protocol to validate the findings before they are committed to standard operating procedures.

Date Preparation Phase: Time Series Representation and Data Reduction

The Data Preparation phase of the CRISP_DM methodology required careful consideration. As noted in Section 2.2.2 large volumes of data need to be reduced to a manageable representation to allow a tractable model. Due to the high number of variables – 800-900 for each vaccine batch in this study, it was necessary to distil the number down to a more manageable size before it is passed to the NN model. In the Data Preparation phase of the CRISP_DM process, each of the statistics proposed in [2] as described in Section 2.2.2 was calculated for each EV time series, creating a set of values that represent the TS, like the components of a fingerprint.

Due to the high number of EVs, *p*, in comparison to *n* (the number of batches) it was necessary to further reduce the number of possible EVs that are presented to the NN in order to obtain a tractable model. The dimension reduction was necessary as the software either could not handle the number of EVs or, in the cases where it could, the NN classification model was poor. This ratio of *n*:*p* was 24: 180 for *Serotype X* and 21: 344 for *Serotype Y*. A number of NN training runs were attempted without reducing the dimension of the data, the NN training process was stopped after 60 hours without having yielded a result.

Multi-group modelling is based on the assumption that a common eigenvector subspace exists for the individual variance/covariance matrix. Through the pooled sample variance/covariance matrix of the batches relating to different yields, the principal component loading is calculated. The EVs that are most strongly correlated to yield in isolation are identified and these 15-20 EVs (as there is a natural drop off in correlation coefficient after this point) are used in the MLR model. Table 1 shows a comparison of the data reduction techniques for the MLR and NN approaches.

	Incumbent MLR Model	**Proposed NN Model**
Dimension reduction technique	Correlation matrix	Principal Component Analysis
Modelling technique	Multiple Linear Regression	Neural Network

Table 1: Comparison of approaches

3.2 Modelling using Neural Networks

3.2.1 Neural Network Design Parameter optimisation

Having prepared the data, the next phase of the CRISP_DM method focuses on building an appropriate model. The RapidMiner software platform was used to develop the NN model and find good settings for the NN design parameters described in Section 2.2.1. A summary of the impact of changes in the NN design parameters for the Serotype X NN is shown in Table 2. This information shows that the number of hidden layers and the momentum are significant factors. These NN design parameters were fed to the "Optimise Parameters" operator in Rapidminer to ensure high accuracy is reached in the final NN model.

Model Parameter adjusted	Values	Effect on accuracy
Training cycles	100, 200, 300, 500 and 1000	None
Learning rates	0.1, 0.3, 0.5, 0.8, 1.0	None
Hidden Layers	1,2,3	2 hidden layers increased accuracy
Momentum	0.1, 0.3, 0.5, 1.0	1.0 decreased accuracy

Table 2: Neural network model parameters for Serotype X

Table 3 shows the impact on accuracy for changes in the Serotype Y NN design parameters.

Model Parameter adjusted	Values	Effect on accuracy
Training cycles	100, 200, 300, 500 and 1000	None
Learning rates	0.1, 0.3, 0.5, 0.8, 1.0	None
Hidden Layers	1,2,3	More than 1 hidden layer reduces accuracy
Momentum	0.1, 0.3, 0.5, 1.0	1.0 decreased accuracy

Table 3: Neural network model parameters for Serotype Y

The number of hidden layers and momentum are also significant in the Serotype Y NN model.

4. Results and Analysis

4.1 Model Performance Comparisons

Table 4 shows the comparison root mean squared error for each model. The NN model offers a significant improvement over the MLR model, in both *Serotype X* and *Serotype Y*. Not only do both NN models have a better root mean squared error, but the variance is also considerably smaller. This indicates that the distance from the residuals to the fitted model does not vary significantly from point to point.

Model	**Root Mean Squared Error (RMSE)**	***p*-Value**
Serotype X NN	0.244 +/- 0.279	NA
Serotype X MLR	11.892 +/- 5.751	0.05
Serotype Y NN	0.464 +/- 0.301	NA
Serotype Y MLR	3.724 +/- 2.827	0.06

Table 4 Performance measure for model comparison

In the *Serotype X* MLR model there is a possible additional error of +/- 5.751 on top of the already large RMSE. The *p*-values indicate that the MLR model is unsuitable. The residuals are large and not normally distributed so the resulting outputs would be susceptible to misinterpretation. The *Serotype Y* MLR model shows similar findings, but on a smaller scale.

The *Proportion Cumulative* column of Table 5 shows that the top 12 Principal Components are responsible for almost 90% of the variability of the target yield variable for Serotype X. The standard deviation column indicates how far the variables are dispersed from the principal component vector. As noted in Section 3, PCA is used to identify a set of significant EVs. The PCs themselves are not passed to the NN, rather the set of prioritised EVs identified by PCA. The number of prioritised EVs passed to the NN model is forty for each of the serotypes. This number was selected as there is a natural drop off in the correlation of the EVs to the principal components from this point onwards. From this pool of forty, the number taken from each PC vector is proportional to its cumulative contribution, as shown in Table 5. For example, if a Principal Component contributes 40% to the variability of the target yield variable, then the top 16 EVs constituting that Principal Component are passed to the NN model. Table 5 shows that PC1 explains 20% of the yield variation so it identifies 8 of the 40 EVs, this is shown in the Explanatory Variable Entitlement column.

PCA Component	Std Deviation	Proportion	Cumulative Proportion	Explanatory Variable Entitlement
PC 1	3.492	0.2	0.2	8
PC 2	3.315	0.18	0.38	7
PC 3	2.697	0.119	0.499	5
PC 4	2.273	0.085	0.584	3
PC 5	1.839	0.055	0.639	2
PC 6	1.81	0.054	0.693	2
PC 7	1.653	0.045	0.738	2
PC 8	1.623	0.043	0.781	2
PC 9	1.448	0.034	0.815	1
PC 10	1.3	0.028	0.843	1
PC 11	1.25	0.026	0.868	1
PC 12	1.168	0.022	0.891	1
PC 13	1.093	0.02	0.91	1

Table 5 PCA for *Serotype X*

Table 6 shows a similar summary for Serotype Y of the EVs contribution to each principal component (based on how much they contribute to yield variation). The top 12 Principal components explain 90% of the variability in Serotype Y, as shown in Table 6. PC1 explains 30.2% of Serotype Y yield variation and so identifies 12 of the 40 EVs.

PCA Component	Std Deviation	Proportion	Cumulative	Explanatory Variable Entitlement
PC 1	5.928	0.302	0.302	12
PC 2	3.965	0.135	0.437	5
PC 3	3.669	0.116	0.553	5
PC 4	2.769	0.066	0.619	3
PC 5	2.555	0.056	0.675	2
PC 6	2.433	0.051	0.726	2
PC 7	2.387	0.049	0.775	2
PC 8	1.99	0.034	0.809	1
PC 9	1.85	0.029	0.838	1
PC 10	1.732	0.026	0.864	1
PC 11	1.674	0.024	0.888	1
PC 12	1.479	0.019	0.907	1

Table 6 PCA for *Serotype Y* s

4.2 Serotype X Neural Network results

The RapidMiner software tool was used to create a NN for *Serotype X* using the prioritised EVs from the PCA as inputs. An image of the best NN is shown in Figure 2. The forty EVs are seen as inputs on the left feeding in to a small number of hidden nodes. The outputs on the right are the yield classification (high or low).

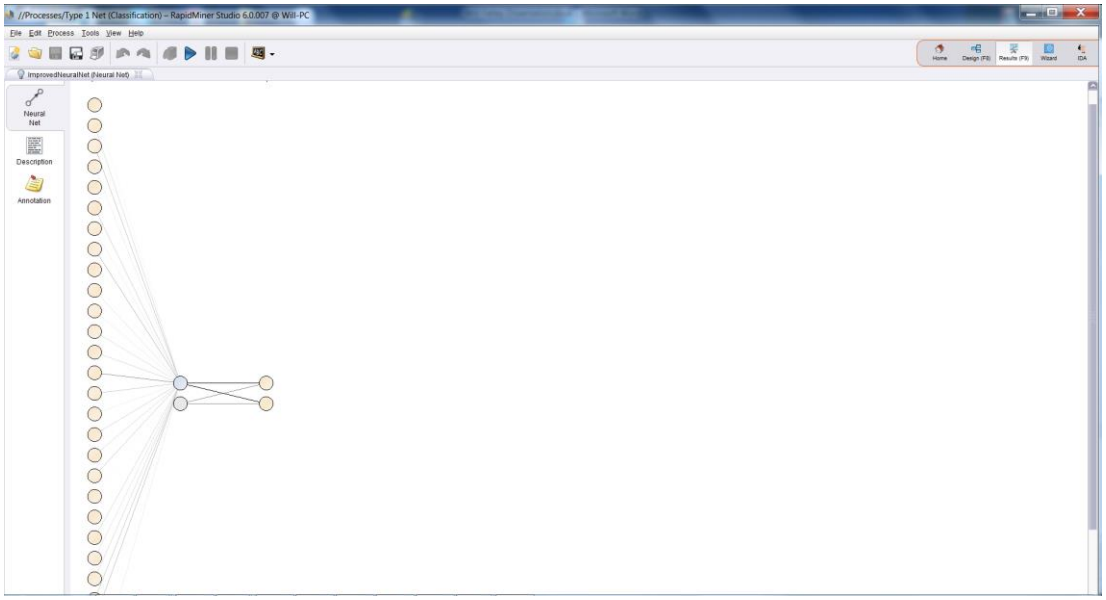

Figure 2 *Serotype X* Neural Network

Sample output of the Serotype X NN model is shown in Table 7. The model performs very well – predicting *Serotype X* yield with an accuracy of 87.5%. The confusion table can be read as the true values in the columns are compared with the predicted values in the rows to calculate the accuracy. For example, in this case the model correctly predicted 16 batches as having a high yield but predicted one batch as having a high yield when in fact it was low. The model has extremely high class precision identifying EVs that predict high yield (94%).

	true High	true Low	class precision
pred. High	16	1	94.12%
pred. Low	2	5	71.43%
class recall	88.89%	83.33%	

Table 7 Confusion table for *Serotype X* Neural network

The precision results show that the NN has a high capability of correctly identifying Serotype X high yield batches (94.12%). Precision and recall measure often have an inverse relationship but we see that the NN also has a high recall value (88.89%). The NN shows good performance in prediction low yield production batches.

4.3 Serotype Y Neural Network results

Figure 3 shows the topology of the NN created using the RapidMiner software for Serotype Y. The forty prioritised EVs identified by the PCA are fed in as inputs to produce the yield output classification.

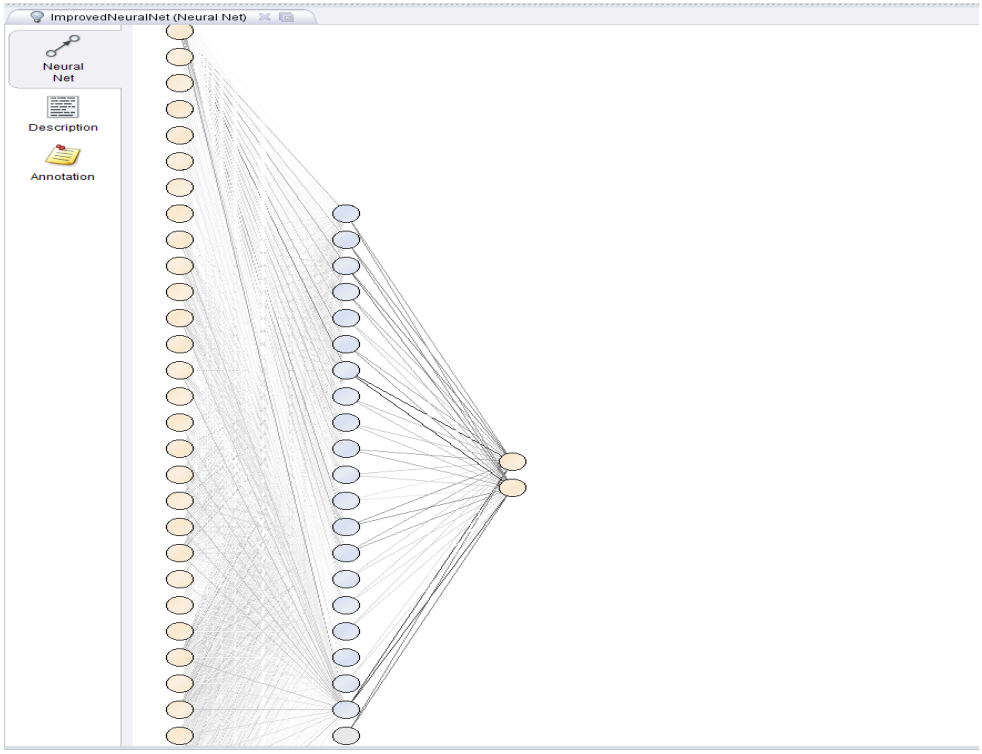

Figure 3 *Serotype Y* Neural Network topology

	true High	true Low	class precision
pred. High	12	4	75.00%
pred. Low	3	2	40.00%
class recall	80.00%	33.33%	

Table 8 Confusion table for *Serotype Y*

The accuracy for the *Serotype Y* model is given as 66.7%. This is not as good as the *Serotype X* model and the reasons for this are explored in the Section 5. Again the NN is better able to predict high yield production runs than to correctly predict low yield production runs for Serotype Y. Unbalanced

datasets are a common problem when data mining manufacturing data [17]. Failure rates tend to be low so that the data are unbalanced with only a low percentage of failed items. It is difficult for the ML technique to distinguish a failure as they occur so infrequently.

4.4 MLR results

The incumbent MLR method is used to identify EVs with high correlation to yield in a univariate manner. These EVs are then investigated under the Statistical Process Control frame work. However, as noted in Section 4.1, the MLR model results in a poor fit to the data. Direct comparison to the multivariate NN results is not possible. The NN approach allows a combination of parameters to be identified and adjusted in unison to improve yield.

4.5 SME interpretation of results

Throughout the CRISP DM process it was necessary to work closely with the domain expert to validate the modelling process. At each stage of the process the model outputs were reviewed by the domain expert to make sure that redundant variables that could not affect the yield of the batch were removed. In addition, batches that were not representative due to process changes were excluded from the data set.

Perhaps the most crucial consideration (while at the same time being the most nebulous) were "effects" that were removed from the data set at the domain expert recommendation. Correlation is not to be confused with causality. The domain process expert reviewed the data and removed what were perceived as effects rather than causes of yield fluctuations. With the high number of variables there is a chance that some of these effects were missed and used as input data to the model. This would mean that they would be correlated with yield and be prioritised as significant during data pre-processing.

5. Recommendations and Discussion

5.1. Challenges identified during the study

A number of challenges remain in applying an analytics model to a complex manufacturing system such as Vaccines Conjugation. It is true to say there is a wealth of data accumulated from modern day manufacturing, but it is also true to say that it is not stored with ease of access or extraction of value in mind. A large amount of time for this project was spent gathering data from disparate locations and converting it to a useable format for the Rapidminer program. There are significant challenges to aggregate and clean data from several different sources. The number of batches available and eligible for this study was smaller than the ideal (<25). This was a result of changes in the ongoing production process during the study period. Changes to the process were, for example, procuring a raw material from a different vendor. Because the components of the system are biological these changes may not just affect the sub system the change is applied to, but could also have unforeseen knock on consequences further downstream. Changes in the production configuration of the system and sub-systems constantly occur over time in contrast to traditional Design of Experiment controlled settings. The control charts used in the 6σ approach identified some batches that were excluded from the study as they were deemed to be too different due to incremental process improvement changes.

The approach to the data in this paper aligns with the 6σ philosophy of continuous process improvement. As changes to the process occur in small incremental steps, sufficient data is available for the NN approach on a rolling basis in line with incremental process improvement changes but care needs to be taken in identifying batch data that is representative of the process as it currently stands.

Although an expansive data set was gathered, it was not exhaustive. The problem of a lurking or hidden variable is ever present i.e. a variable that is significant to yield but has not been analysed. It is

important to recognise there are limits to what we can capture and explain due to the sheer number of possible permutations. In the words of George E. P. Box: “All models are wrong, but some are useful.”

With so much data produced by a modern manufacturing process, analytics has a distinct advantage in that it is exploratory rather than ruling in or out a particular hypothesis. This is a very important quality when analysing manufacturing data as the investigator may not always know what they are looking for. In the case of this study the results were a significant improvement on the incumbent Six Sigma method and the NN approach has been adopted by company Z on a trial basis for other serotypes.

Currently statistics are viewed as the domain of experts. Analytics has the potential to be a more widely accessible toolkit because of the availability of DM tools with GUI interfaces. Importantly, coaching on the statistical significance of results and a grounding in the limitations of the models are a prerequisite for the appropriate application of analytics.

With the advent of Electronic Batch Records and Manufacturing Execution Systems the raw materials required for the application of analytics are readily available. There is an abundance of real-time shop-floor data - however, the skillsets to translate this into knowledge using analytics are scarce. It is an opportune time to start combining the two most valuable resources a manufacturing company has – its data and its people. The challenge is to invest in the systems and skillsets that will allow companies to optimise their use of existing process information — the first step being the commissioning of a dedicated analytics server which combines all the disparate pockets of data into a format that is easily and quickly analysed by a data mining packages. The true power of these techniques lies in their accessibility, an ideal scenario being that the domain expert becomes proficient in the use of these tools.

Analytics techniques have the potential, with very little outlay, to significantly increase profit margins particularly in the fragile vaccine manufacturing domain [19]. The success of this project has lead company Z to extend the methods to the remaining of the vaccines serotypes that make up the product. But this is a secondary consideration compared with the effect these vaccine products have on the patients that receive them. The vaccine that is the subject of this study is predicted to save 1.5 million lives by 2020.

The NN model results, while promising, have limitations. NNs are a heuristic technique so these results are empirical evidence only. Much of this document describes measures taken in order to secure management buy in, but measures must also be taken to manage management expectations. This NN classifies production settings that produce a high and low yield. It is important that management understand the model outputs and limitations of the NN and ML approaches.

The analytics era is in its infancy from a manufacturing standpoint, but the practice of advanced analytics is grounded in years of mathematical research with successful applications to the equally volatile and complex areas of banking and finance industries. While these powerful tools are easy to use, a good understanding of their statistical foundations is crucial to valid interpretation of results and to ensure that assumptions underlying the statistical techniques are not violated. This is why the company-wide initiative and use of Six Sigma at all levels of the company should provide a fertile ground for making the case for data mining and facilitating its acceptance. The Six Sigma mindset of measuring the performance of processes and analysing data promotes data-based decision making, therefore making data mining a natural extension of this methodology.

Acknowledgements

We would like to thank Gareth Thornton for his help proof reading this document. We would also like to thank the staff of the Smurfit Business School for sharing their expertise and insights in the course

of this work. Thanks also to Dermot O'Malley and Paraic Fahey for their endless patience andhelp. Thanks also to Tony Walsh for taking a keen interest in the methodology and his sponsorship of the project. Thanks to Eamonn Nixon for his flexibility and support.

References

[1] Areteaga F, Ferrer A. Dealing with missing data in MSPC: several methods, different interpretations, some examples. Journal of Chemometrics 2002; 16 (8-10): 408-418.

[2] Bickel P. J. and Lehmann E.L. Descriptive Statistics for non parametric models. The Annals of Statistics 1975; 3(5): 1038-1044.

[3] Black S, Shinefiled H, Fireman B. Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. Pediatric Infectious Disease Journal YYYY 19(3): 187-195.

[4] Baumgarten, Matthias, et al. "User-driven navigation pattern discovery from internet data." *Web Usage Analysis and User Profiling*. Springer Berlin Heidelberg, 2000. 74-91.

[5] Büchner, Alex G., and Maurice D. Mulvenna. "Discovering internet marketing intelligence through online analytical web usage mining." *ACM Sigmod Record* 27.4 (1998): 54-61.

[6] Chakravorty Satya S. Six Sigma Failures: An Escalation Model. Operations Management Research 2009; 2(1-4): 44–55.

[7] Chien, Chen-Fu, Wen-Chih Wang, and Jen-Chieh Cheng. "Data mining for yield enhancement in semiconductor manufacturing and an empirical study." *Expert Systems with Applications* 33.1 (2007): 192-198.

[8] FDA. Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance**,** 2004.

[9] Feng C.-X.J., Wang X.F. (2004) Data mining technique applied to predictive modelling of the knurling process. IIE Transactions 36: 253–263

[10] Gambillara The Conception and Production of Conjugate Vaccines Using Recombinant DNA Technology. BioPharm International 2012; 25(1): 28-32.

[11] Hickey C, Kelly S, Carroll P, O'Connor J. Prediction of Forestry Planned End Products Using Dirichlet Regression and Neural Networks. Forest Science 2014; 61(2): 289-297.

[12] Jerez, M, Ignacio Molina, Pedro J Garcı, Emilio Alba, Nuria Ribelles, Leonardo Franco, and Miguel Martı. 2010. "Artificial Intelligence in Medicine Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem." 50: 105–15.

[13] Köksal Gülser, Inci Batmaz, Murat Caner Testik. A review of data mining applications for quality improvement in manufacturing industry. Expert Systems with Applications 2012; 38: 13448–13467.

[14] Kourti Theodore, MacGregor John F. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemometrics and Intelligent Laboratory Systems 1995: 28(1): 3-21.

[15] Molcho G, Zipori Y, Schneor R,.Computer aided manufacturability analysis: Closing the knowledge gap between the designer and the manufacturer, CIRP Annals – Manufacturing Technology 2008; 57(1): 153-158.

[16] Proano R. A., S. H. Jacobson, and W. Zhang. Making combination vaccines more accessible to low-income countries: The antigen bundle pricing problem, Omega 2012; 40(1): 53 - 64.

[17] Provost, Foster. Machine learning from imbalanced data sets 101. In Proceedings of the AAAI 2000 workshop on imbalanced data sets 2000; pp 1-3.

[18] Raisinghani Mahesh S., Hugh Ette, Roger Pierce, Glory Cannon, Prathima Daripaly. Six Sigma: concepts, tools, and applications. Industrial Management & Data Systems 2005; 105(4): 491 – 505.

[19] Robbins M. J. and S. H. Jacobson. Pediatric vaccine procurement policy: The monopsonist's problem. Omega 2011: 39(6): 589 - 597, 2011.

[20] Rokach, Lior. "Decomposition methodology for classification tasks: a meta decomposer framework." Pattern Analysis and Applications 9.2-3 (2006): 257-271.

[21] Rudan I, Tomaskovic L, Boschi-Pinto C, Campbell H. Global estimate of the incidence of clinical pneumonia among children under five years of age. Bull World Health Organ 2004; 82: 895-903.

[22] Tetko, Igor V, David J Livingstone, and Alexander I Luik. 1995. "Neural Network Studies. 1." 826–33.

[23] Thomassen, Yvonne E., et al. "Multivariate data analysis on historical IPV production data for better process understanding and future improvements." *Biotechnology and bioengineering* 107.1 (2010): 96-104.

[24] Wang X., K. Smith, and R. Hyndman. Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery 2006; 13(3): 335-364.

[25] Wirth, R., & Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining 2000: pp. 29-39.

[26] Wold Svante, Esbensen Kim, Geladi Paul. Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems 1987; 2(1-3) pp 37-52 in Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[27] Wu, Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and Information Systems* 14.1 (2008): 1-37.

[28] Zobel Chris and Cook Deborah. Evaluation of neural network variable influence measures for process control. Engineering Applications of Artificial Intelligence 2011; 24: 803–812.