Pharmaceutical Manufacturing Dataset: Comprehensive Documentation and Analysis Report

Executive Summary

This report provides a detailed analysis of a comprehensive pharmaceutical manufacturing dataset containing real production data from 1,005 batches of cholesterol-lowering tablets manufactured between November 2018 and April 2021. The dataset encompasses the complete manufacturing pipeline from raw material analysis through process monitoring to final product quality testing, making it invaluable for developing predictive models and optimizing pharmaceutical manufacturing processes.

Dataset Overview

Dataset Composition

The pharmaceutical manufacturing dataset consists of four interconnected components:

Dataset	Dimensions	Description	Primary Use
Process Dataset	1,005 × 35	Engineered features from time series data	Main dataset for modeling
Laboratory Dataset	1,005 × 55	Raw material and quality analysis results	Comprehensive process understanding
Normalization Dataset	25 × 3	Batch size normalization factors	Cross-batch comparisons
Time Series Dataset	25 files	Raw sensor data (every 10 seconds)	Advanced time series analysis

Product Scope and Manufacturing Context

The dataset covers a family of cholesterol-lowering film-coated tablets manufactured using direct compression technology. The product family includes four different strength formulations and nine different batch sizes, ranging from 240,000 to 2,400,000 tablets per batch. This diversity ensures the dataset captures various manufacturing scenarios and operational conditions.

Detailed Dataset Analysis

Process Dataset - Primary Analysis Dataset

The Process Dataset represents the most refined and analysis-ready component, containing 35 carefully engineered features derived from expert knowledge of tablet compression processes.

Key Feature Categories

Manufacturing Process Parameters:

- `batch`: Unique identifier for each production batch
- `code`: Product sub-family classification (1-25)
- `tbl_speed_mean`: Average tablet press speed (tablets/hour) excluding downtime
- `tbl_speed_change`: Number of speed adjustments normalized by batch size
- `total_waste`: Rejected tablets per batch (normalized)
- `startup_waste`: Tablets rejected during equipment setup

Process Control Metrics:

- `main_CompForce_mean`: Main compression force during production (kN)
- `main_CompForce_sd`: Standard deviation of compression force
- `tbl_fill_mean`: Average tablet fill depth (mm)
- `SREL_production_mean`: Standard relative deviation of compression force

Quality Target Variables:

- `Drug release average (%)`: Primary quality indicator
- `Drug release min (%)`: Minimum dissolution performance
- `Total impurities`: Impurity content in final product
- `Residual solvent`: Solvent residue levels

Data Quality Assessment

The Process Dataset demonstrates high completeness with 98.2% complete records. Missing values are concentrated in quality parameters (18 missing values per quality target, representing 1.8% of the

dataset). All process parameters show complete data coverage, indicating robust sensor performance and data collection procedures.

Laboratory Dataset - Comprehensive Analysis Repository

The Laboratory Dataset provides the most comprehensive view of the manufacturing process with 55 parameters covering the entire production chain.

Raw Material Analysis Parameters

Active Pharmaceutical Ingredient (API) Characteristics:

- Water content: Mean 1.5% (±0.4%, RSD 29.7%)
- Total impurities: Mean 0.2% (±0.1%, RSD 50.1%)
- Content assay: Mean 94.4% (±0.4%, RSD 0.4%)
- Particle size distribution: Three measurement points (10%, 50%, 90% cumulative volume)

Excipient Properties:

- Lactose water content: Highly controlled (Mean 0.1%, RSD 14.6%)
- SMCC (Silicified Microcrystalline Cellulose) properties including density and particle size
- Starch characteristics including pH and moisture content

Intermediate and Final Product Testing

Tablet Core Properties:

- Physical dimensions (thickness, diameter, weight)
- Mechanical properties (hardness, tensile strength)
- Manufacturing yield metrics

Final Product Quality:

- Drug release performance: Mean 90.6% (±3.4%, RSD 3.7%)
- Impurity profiles with specific tracking of critical impurities
- Residual solvent analysis

Time Series Dataset - High-Resolution Process Monitoring

The Time Series Dataset contains 25 separate files (1.csv through 25.csv), each corresponding to a specific product code and containing all batches manufactured under that configuration.

Time Series Parameters

Each time series file captures 16 critical process parameters recorded every 10 seconds:

Primary Process Indicators:

- `tbl_speed`: Tablet press speed (tablets/hour)
- `main_comp`: Main compression force (kN)
- `tbl_fill`: Tablet fill depth (mm)
- `SREL`: Standard relative deviation of compression force

Production Monitoring:

- `produced`: Count of acceptable tablets at each timestamp
- `waste`: Cumulative count of rejected tablets
- `ejection`: Tablet ejection force indicating potential sticking issues

Equipment Parameters:

- `stiffness`: Bottom punch stiffness (N)
- `cyl_main`: Cylindrical height at main compression station
- `fom`: Filling device rotational speed (rpm)

Time Series Characteristics

Individual time series datasets can contain over 160,000 data points, representing manufacturing processes lasting 2-20 hours depending on batch size. The high temporal resolution provides unprecedented insight into process dynamics and enables advanced analytics including anomaly detection and process optimization.

Normalization Dataset - Cross-Batch Standardization

The Normalization Dataset provides essential scaling factors to enable fair comparisons across different batch sizes[1][2]. With batch sizes varying from 240,000 to 2,400,000 tablets, normalization factors range

from 2.40 to 24.00, ensuring that process parameters can be meaningfully compared across the product family.

Statistical Analysis and Data Quality

Key Parameter Statistics

Parameter	Mean	Std Dev	RSD (%)	Min	Max
API Water Content	1.5%	0.4%	29.7%	0.0%	2.7%
API Total Impurities	0.2%	0.1%	50.1%	0.1%	0.5%
Drug Release Average	90.6%	3.4%	3.7%	82.5%	102.7%
Total Impurities	0.1%	0.1%	71.3%	0.1%	0.6%

Process Capability Analysis

The dataset demonstrates well-controlled manufacturing processes with process capability indices (Ppk) exceeding 1.0 for critical quality parameters[1][2]. This indicates that the manufacturing process operates within specification limits with adequate control margins, validating the industrial relevance and quality of the dataset.

Missing Data Analysis

Process Dataset Missing Values:

- Quality parameters: 18 missing values each (1.8%)
- All process parameters: Complete coverage

Laboratory Dataset Missing Values:

- API analysis parameters: 0.2-0.9% missing
- Generally associated with specific API material codes

The missing data patterns are systematic rather than random, primarily associated with specific raw material batches, making them manageable through targeted imputation strategies or subset analysis.

Project Applications and Use Cases

Primary Use Cases

1. Quality Prediction Modeling

- **Objective**: Predict final product quality from process parameters and raw materials
- **Target Variables**: Drug release performance, impurity levels, residual solvents
- **Business Impact**: Reduce testing time by 50-70%, enable real-time quality assessment

2. Process Optimization

- **Focus Areas**: Tablet press speed optimization, compression force tuning, waste reduction
- **Methods**: Multi-objective optimization, design of experiments, statistical process control
- **Expected Outcomes**: 10-15% reduction in manufacturing waste, improved process consistency

3. Anomaly Detection and Predictive Maintenance

- **Data Sources**: High-frequency time series process data
- **Applications**: Equipment health monitoring, quality deviation early warning
- **Methods**: Statistical control charts, machine learning anomaly detection

4. Raw Material Impact Analysis

- **Scope**: Understanding how material variations affect final product quality
- **Business Value**: Supplier qualification, specification optimization, cost reduction

Advanced Analytics Opportunities

Time Series Analysis:

- Process dynamics modeling with 160,000+ data points per product configuration
- Seasonal variation analysis across 2.5-year dataset
- Equipment performance degradation tracking

Multi-modal Data Integration:

- Combining structured laboratory data with high-frequency sensor data
- Cross-domain feature engineering from chemical, physical, and process parameters

- Holistic manufacturing system optimization

Technical Implementation Guide

Data Loading and Preprocessing

File Structure Requirements:

```
# All CSV files use semicolon separator
process_df = pd.read_csv('Process.csv', sep=';')
lab_df = pd.read_csv('Laboratory.csv', sep=';')
norm_df = pd.read_csv('Normalization.csv', sep=';')

# Time series files require datetime conversion
ts_df = pd.read_csv('2.csv', sep=';')
ts_df['timestamp'] = pd.to_datetime(ts_df['timestamp'])
```

Data Integration Strategy:

```
# Primary join on batch ID
combined_df = pd.merge(process_df, lab_df, on='batch', how='inner')

# Add normalization factors
combined_df = pd.merge(combined_df, norm_df,
                       left_on='code_x', right_on='Product code', how='left')
```

Feature Engineering Recommendations

Normalization Applications:

- Apply batch size normalization factors to waste metrics and time-dependent parameters
- Create efficiency ratios combining yield and waste metrics
- Develop composite quality scores from multiple quality indicators

Time Series Feature Extraction:

- Statistical aggregations (mean, std, min, max) for process stability assessment
- Process duration and interruption frequency analysis

- Dynamic process behavior characterization

Model Development Framework

Target Variable Selection:

- **Primary Targets**: Drug release average (%), Total impurities
- **Secondary Targets**: Residual solvent, specific impurity levels
- **Process Targets**: Batch yield, manufacturing efficiency metrics

Feature Selection Strategy:

- Remove highly correlated parameters (correlation > 0.95)
- Apply normalization factors for cross-batch comparisons
- Consider temporal relationships in time series analysis

Validation Approach:

- Time-based splits to respect temporal dependencies
- Cross-validation across product sub-families
- Hold-out validation on recent manufacturing periods

Dataset Advantages and Industry Relevance

Unique Dataset Characteristics

1. **Industrial Authenticity**: Real production data from 1,005 actual manufacturing batches over 2.5 years
2. **Comprehensive Coverage**: Complete process chain from raw materials to final product quality
3. **High Temporal Resolution**: 10-second interval process monitoring providing unprecedented detail
4. **Multi-variant Design**: Four product strengths and nine batch sizes ensuring broad applicability
5. **Regulatory Compliance**: Data collected under pharmaceutical industry quality standards