Reprinted from PHARMACEUTICAL ENGINEERING®
The Official Magazine of ISPE
September/October 2012, Vol. 32 No. 5
www.PharmaceuticalEngineering.org ©Copyright ISPE 2012

This article is based on a webcast presented by the authors in cooperation with ISPE.

Traditional Lot Traceability Approaches Are Not Sufficient to Enable Upstream/Downstream Correlation Analysis for Quality by Design (QbD)

by Victor Shashilov, PhD and Justin Neway, PhD

Introduction

As Quality by Design (QbD) initiatives gain acceptance across life sciences process development and manufacturing organizations, there is an increased need to understand the dynamics of the critical portions of the process stream more fully. For example, detailed accounting of the fractional contributions of upstream process steps to downstream process steps is required at points of splitting and pooling in the process stream to conduct statistical correlations between upstream process parameters and downstream process outcomes. These fractional contributions are compounded in processes that have multiple points of splitting and pooling.

Splitting and pooling of batches is common in both process development and full scale commercial manufacturing. In commercial manufacturing, batches are often split and pooled to increase throughput and optimize equipment usage where downstream steps are slower than upstream steps or where upstream equipment has lower capacity than downstream equipment. For example, several upstream lots can be mixed into a single granulation batch in order to fully utilize the capacity of the granulator. Each granulation batch can then be split between several tablet presses, such that the total throughput of the tableting step is high enough to maintain the required load of the coating equipment and keep the hold time for the material coming from the granulator within the specified limits. In process development situations, batches are typically split into subsequent unit operations where sub-batches are run through different sets of experimental conditions and then recombined into downstream steps where the downstream equipment has higher capacity than the upstream equipment. The complexity of lot genealogy in the PD environment rapidly increases with the number of unit operations and the number of experimental conditions.

Traditional lot traceability tools are often used to track the linkages between process inputs and outputs. They are intended primarily for recall management and do not provide sufficient detailed information or flexibility on their own to allow correct calculation of such correlations. Spreadsheets are sometimes pressed into service to help calculate compounded fractional contributions across multiple process steps, but for this application, they are error-prone and hard to manage, and they become impractical very rapidly as the number of splitting and pooling points in a process grows.

Better, more automated and flexible tools are needed to perform these calculations so that useful process models can be built to link upstream Critical Process Parameters (CPPs) to downstream Critical Quality Attributes (CQAs) in processes where splitting and pooling occurs, an important requirement for achieving the goals of QbD. This article will explore the following three approaches for making such correlations and highlight the advantages and limitations of each:

1. The traditional spreadsheet-based approach
2. The manual SQL approach
3. A new on-demand SQL genealogy approach

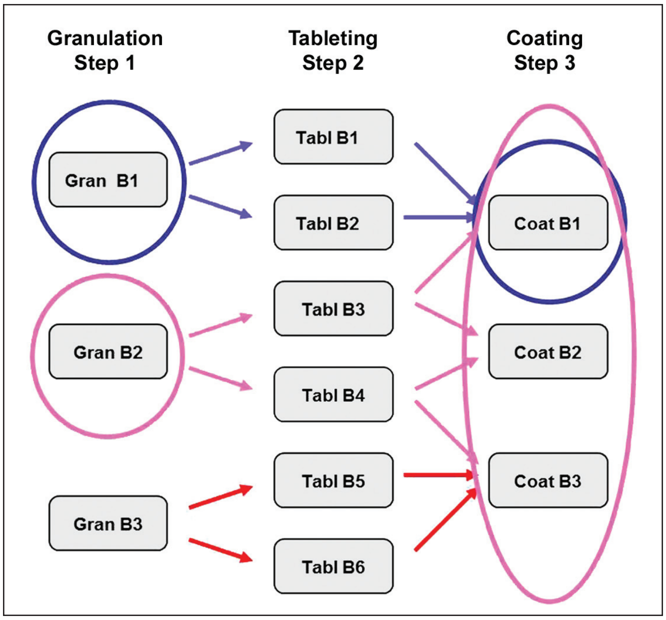


Figure 1. A simple process genealogy context shown as batches are mixed and split from left to right.

that offers opportunities for automation and the ability to handle complex process genealogies with comparative ease and simplicity

The Difference between Traditional Lot Traceability and a Full Accounting of the Process Stream Genealogy

Traditional lot traceability is typically used to manage situations where a defect has been identified in an incoming material or upstream in-process material that could jeopardize the quality of the final downstream product. By using a traceability matrix, a manufacturer can determine which final batches contain any amount of the defective upstream material. This information is used to support decisions about which lots of final product to quarantine for further testing or which shipped product lots to recall. Figure 1 illustrates an example of this type of traceability matrix. In traditional lot traceability situations, the focus is in knowing which downstream lots contained any amount of the defective upstream material, as opposed to knowing the exact amount of the defective upstream material that ended up in each downstream lot.

In many situations, it is desirable to correlate variations in upstream process conditions or materials to variations in downstream process outcomes to establish whether or not there is a relationship between them (e.g., whether or not the upstream process parameter can be considered for further evaluation as a CPP driving a downstream CQA). To enable such correlations, it is first necessary to calculate the fractional contributions of each upstream step to each downstream step across all the process steps between the upstream variable and the downstream outcome. This reveals the amount of the upstream material or condition that is associated with each downstream outcome instance. These upstream amounts or conditions are then used as #1 parameter values in the correlation calculation using the corresponding values for each downstream outcome as #2 parameters values. Furthermore, if the number of #1 parameter values does not match the number of #2 parameter values as shown in the more complex example in Figure 2, additional steps must be taken so that there are equal numbers of upstream and downstream parameter values available for the correlation calculation. The fractional contribution calculations can be very cumbersome and error-prone when using spreadsheet methods. Instead, methods based on the use of well-designed Structured Query Language (SQL) queries are a more practical way of accounting for the splitting and pooling genealogy during such investigations because they allow users to make these nested calculations more easily and reuse their work with less potential for introducing errors.

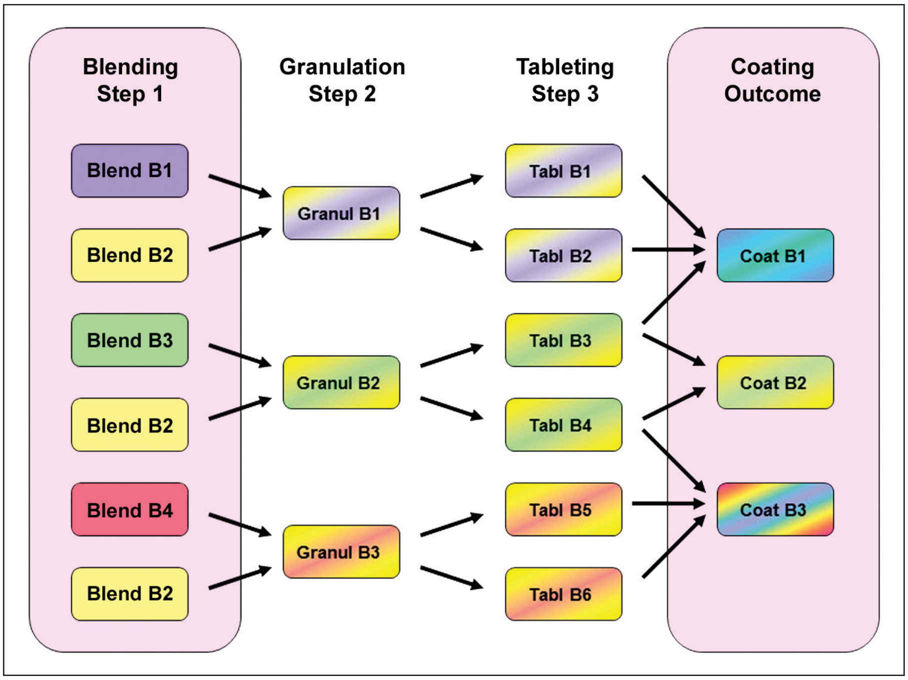


Figure 2. A more complex process genealogy context showing how batches can be split and pooled from left to right with fixed and varying cardinality, and how the number of upstream input batches can be unequal to the number of downstream outcome batches.

Upstream/Downstream Analysis

The following two examples illustrate some of the complexities inherent in upstream/downstream data analysis:

In the first example, a nutrient supplement is added to a seed fermentor and there is a need to determine whether a significant difference exists between the resulting process yields from batches

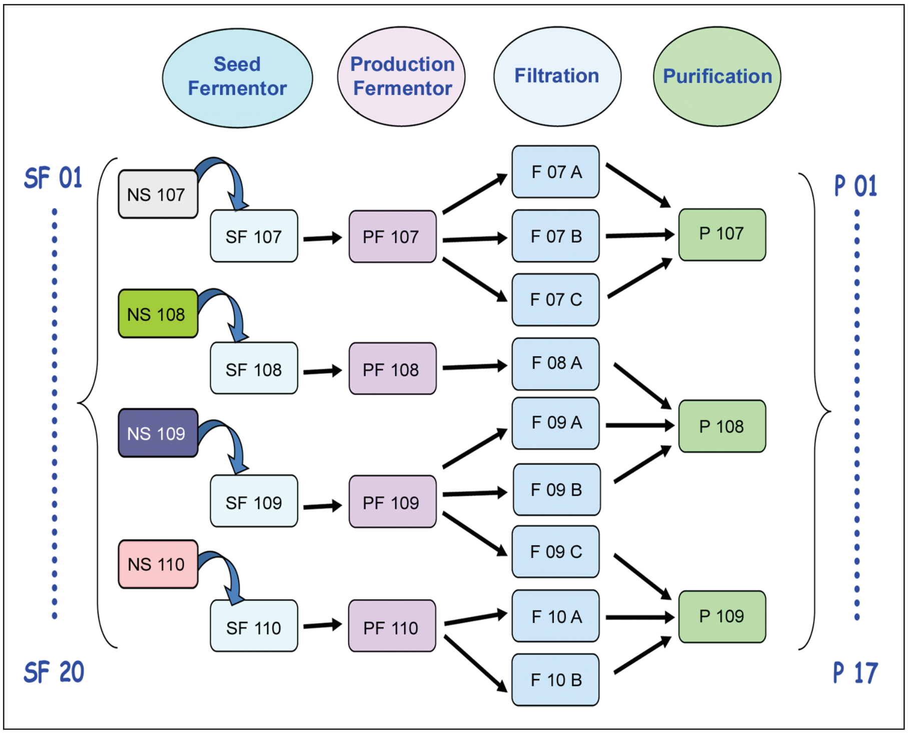


Figure 3. A nutrient supplement supplied by four different vendors is added at the seed fermentation step with four batches split into eight by the filtration step and then combined into three batches at the purification step.

produced using supplement from four separate vendors. To make this determination, lots need to be traced across four steps, including: (1) the seed fermentor step, (2) the production fermentor step, (3) the filtration step, and (4) the downstream purification step. Nutrient supplements supplied by four different vendors are added at the seed fermentation step. To reduce the total hold time for this degradation-prone protein, the four batches are split into nine at the filtration steps for parallel processing and then combined into three batches at the purification steps as shown in Figure 3.

To determine whether the specific vendor of the nutrient supplement has an effect on the yield of the product, the mixing at the purification step needs to be taken into account and included in calculations of the fractional contributions of each vendor's supplement into each of the final purification batches.

The second example illustrates additional important considerations for correlating CPPs to CQAs in a process with a more complex lot genealogy as shown in Figure 4. In this case, there is a need to determine whether a correlation exists between the osmolality of the media that is fed into each seed fermentation step and an impurity found in the final product lots. Splitting and pooling in the process stream needs to be taken into account along with the number of incoming media lots and the number of final purification lots, which have different total numbers. Using a manual spreadsheet-based approach for such calculations would require a significant amount of time and effort, but *could* be done using the following steps.

Three Approaches for Upstream/Downstream Correlations Analysis

The Manual Spreadsheet Approach

To manually correlate a parameter from Step A (Par A) to a parameter from Step X (Par X), the following steps are required:

1. Construct the lot genealogy chart.
2. Starting with Step A, for each pair of adjacent steps, perform calculations as in #3.
3. These calculations need to be repeated for each step and each batch within the step.
   a. Calculate fractional contributions of the material from all the batches from the previous step feeding into the given batch.
   b. Calculate the weighted average of the weighted averages of parameter A, computed for batches in the previous step.

While it is possible to use this spreadsheet-based method for upstream/downstream analysis, calculations in each subsequent iteration use weighted averages obtained in the previous iteration; therefore, errors can accumulate rapidly. In a typical situation where the cardinality varies among batches, the number of fractional contribution calculations required to support the calculation of correlations between any two steps of Ns steps is roughly proportional to the square of the number of steps and the number of batches as illustrated in the following formula:

$$\text{Effort} \sim N_S^2 \bullet N_{\text{batches}}$$

Table A illustrates the relative effort involved in the manual spreadsheet approach as a function of the number of batches, process steps, and parameters at each step. As seen from the table, the effort rapidly grows with the number of steps and batches, which can make this approach impractical, even for processes of moderate complexity.

The limitations and risks of the manual spreadsheet approach include:

1. Prone to errors – mistakes are easy to make and difficult to find.
2. Error propagation – errors made in the beginning are carried over and accumulated.
3. Complexity – calculations become unmanageable for large numbers of steps and parameters and/or complex genealogies.
4. Time consuming – spreadsheets are, by nature, extremely inefficient.

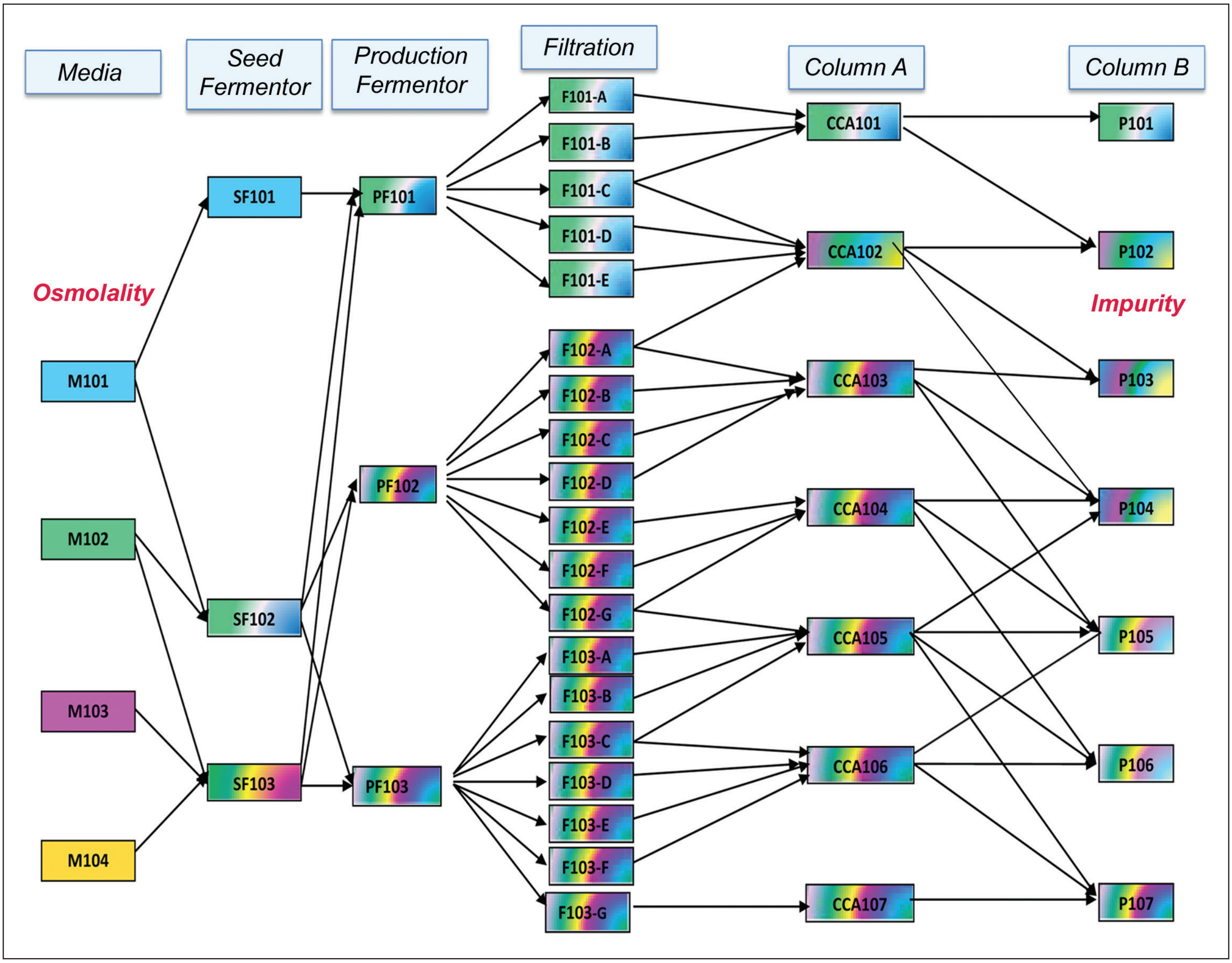


Figure 4. A more complex genealogy for which CPPs and CQAs need to be correlated during the fermentation process to determine whether there is a correlation between the osmolality of the media that is fed into the seed fermentation step to the impurity of the final product.

5. Difficult to validate – requires at least two people doing the same calculation independently and coming up with the same results every time.
6. Difficult to automate – all calculations have to be repeated from scratch for new batches and parameters.

The Manual SQL Approach

Another more useful method of dealing with the complexity of upstream/downstream analysis when there is splitting and pooling in the process stream is by taking a manual SQL query approach to model the lot genealogy using applications with data modeling capabilities, such as Oracle, Microsoft Access, or other enterprise applications that support SQL queries.[1] The approach considered here relies on the assumption that the end user is able to select any parameters from any two process steps for a correlation analysis without the need to write any SQL statements. To accomplish this, all the necessary data views need to be created in advance and then properly maintained when the structure of the source data or the manufacturing process changes. An example of a data view is shown in Table B. This data view maps media osmolality values ("Osmolality" field) to the impurity values in the production fermentor ("Impurity" fields). "Media," "Seed F," and "Prod F" columns compose the "lot tree." The lot tree in this example links media lots to production fermentor batches through batch IDs at the intermediate step – the Seed Fermentor ("Seed F") step.

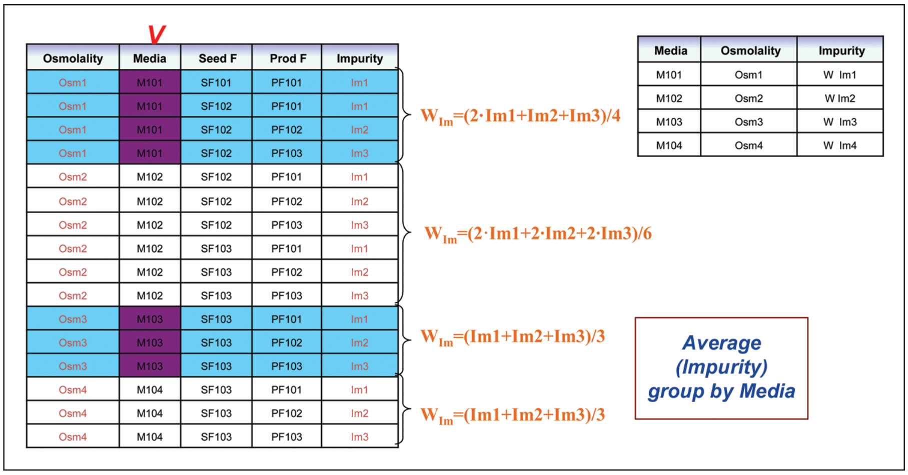
The complete mapping of the lot genealogy in this manual SQL approach relies on the creation of a set of genealogy "lot

# Steps	# Batches	Relative Effort
2	2	1
2	10	5
10	10	125
30	10	1,125

Table A. Relative effort encountered using the manual spreadsheet approach as a function of process complexity.

trees" as the first step. Lot trees require a set of views[1] called "mapping views." These views map the upstream steps to the downstream steps in all pairs of adjacent steps in the process. For each pair of non-adjacent steps A and B where a lot number change occurs, a pathway map or lot tree is created by sequential joining of mapping views.

Example 1. By way of illustration, consider a simple process with three steps (A, B, and C) and assume that a lot number change and pooling occurs between steps A and B and again between steps B and C along with splitting as shown below.



Osmolality	Media	Seed F	Prod F	Impurity
Osm1	M101	SF101	PF101	Im1
Osm1	M101	SF102	PF101	Im1
Osm1	M101	SF102	PF102	Im2
Osm1	M101	SF102	PF103	Im3
Osm2	M102	SF102	PF101	Im1
Osm2	M102	SF102	PF102	Im2
Osm2	M102	SF102	PF103	Im3
Osm2	M102	SF103	PF101	Im1
Osm2	M102	SF103	PF102	Im2
Osm2	M102	SF103	PF103	Im3
Osm3	M103	SF103	PF101	Im1
Osm3	M103	SF103	PF102	Im2
Osm3	M103	SF103	PF103	Im3
Osm4	M104	SF103	PF101	Im1
Osm4	M104	SF103	PF102	Im2
Osm4	M104	SF103	PF103	Im3

$W_{Im}=(2\cdot Im1+Im2+Im3)/4$

$W_{Im}=(2\cdot Im1+2\cdot Im2+2\cdot Im3)/6$

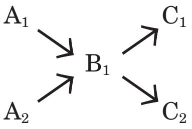
$W_{Im}=(Im1+Im2+Im3)/3$

$W_{Im}=(Im1+Im2+Im3)/3$

Media	Osmolality	Impurity
M101	Osm1	W Im1
M102	Osm2	W Im2
M103	Osm3	W Im3
M104	Osm4	W Im4

Average (Impurity) group by Media

Table B. Data view for calculating weighted averages using the production fermentor as the universe.



First, mapping views are constructed to link lot IDs of the adjacent steps. In this case, there are two mapping views: AB and BC. Each mapping view contains only two columns : lot IDs of step A and lot IDs of step B in view AB, and lot IDs of step B and lot IDs of step C in view BC, as shown below.

$$AB = \begin{matrix} A_1\,B_1 \\ A_2\,B_1 \end{matrix} \quad \text{and} \quad BC = \begin{matrix} B_1\,C_1 \\ B_1\,C_2 \end{matrix}$$

Now that the mapping views are available, non-adjacent steps can be mapped by constructing lot trees. In this case, there is only one pair of non-adjacent steps: A and C. These steps (A and C) can be linked in either a forward manner (i.e., from step A to step C) or backward manner (i.e., from step C to step A). In either case, the lot tree contains three fields: Lot IDs of step A, Lot IDs of step B, and Lot IDs of Step C as shown below:

$$AC = \begin{matrix} A_1\,B_1\,C_1 \\ A_1\,B_1\,C_2 \\ A_2\,B_1\,C_1 \\ A_2\,B_1\,C_2 \end{matrix} \text{ and } CA = \begin{matrix} C_1\,B_1\,A_1 \\ C_1\,B_1\,A_2 \\ C_2\,B_1\,A_1 \\ C_2\,B_1\,A_2 \end{matrix}$$

Lot trees AC and CA link lot IDs at step A to lot IDs at step C through lot IDs of the intermediate step (step B).

In the case of a three-step process, four views are needed (AB, BC, AC, and CA) to be able to map all pairs of adjacent and non-adjacent steps. If a similar approach is applied to a four step process (A-B-C-D) then three mapping views: AB, BC, and CD and six lot trees: AC, AD, BD, CA, DA, DB, need be created, which gives a total of nine views. Using the combinatorial theory[2] it can be shown that to allow for correlations between any two steps in an N-step process, the total number of views to be created is defined by formula 1 and is shown in Table C.

Formula 1 takes into account the fact that views obtained by forward joins (i.e., from step A to step B) and backward joins (i.e., from step B to A) will in general not be identical.

Formula 1 $\quad N_{Views} = N_{steps} \bullet (N_{steps} - 2) + 1$

The major limitation and risk of the manual SQL approach is the rapidly increasing complexity as the number of steps and parameters increases. The equation from combination theory below shows that the total number of data views is proportional to the square of the number of steps and the square of the number of parameters, assuming for simplicity the same number of parameters in each step. (Note: Equation 2 is used to calculate the maximum number of required views in situations where each data view contains only two parameters.) The number of data views will be smaller if several data views are combined to contain multiple parameters from both steps mapped in the view.)

Equation 2 $\quad N_{Views} = \{N_{steps} \bullet (N_{steps} - 2) + 1\} \bullet N^2_{parameters}$

Thus, for a 30-step process, 84,100 views need to be created and maintained as shown in Table D.

Because the calculations involved in both the manual spreadsheet approach and the manual SQL approach are cumbersome and time consuming, the ideal solution lies in creating views and performing analyses on-demand, based on an easier method of selecting steps and parameters that maximizes flexibility and reuse while at the same time reducing the potential for errors. Thus, a tool is needed that provides the ability to refresh data and re-execute each analysis in a more automated fashion.

Due to the large amount of the data modeling effort required to enable ad-hoc correlations between any parameters and process steps, business users usually have to limit the number of parameters and process steps included in the data model and request changes to the data model each time the need for more data becomes obvious. Therefore, in most real time situations, the manual SQL approach doesn't deliver the ability to perform ad-hoc correlations between any parameters of a user's choice across complex lot genealogies. However, when such a capability is required, end users need to write fairly complex SQL queries against mapping views and data views,

# Steps	# Lot Trees
2	1
5	16
10	81
30	841

Table C. Number of lot trees as a function of the number of process steps.

which makes this approach error prone, time consuming, and difficult to validate.

The On-Demand SQL Genealogy Approach

A more practical alternative to the manual SQL query approach described above is the on-demand SQL genealogy approach, which eliminates the need to create and store large numbers of views, allowing the user to more easily perform correlations between any parameters across any number of steps at any time, and instead uses queries and genealogies that are automatically created on-demand. This approach saves time and minimizes opportunities for miscalculations and error propagation due to human error.

Figure 5 depicts a typical genealogy data modeling workflow, which is the same for the manual and on-demand SQL genealogy approaches.

- Step A is to create views (called *mapping views*) that map the relationships between all pairs of adjacent steps in the process. The lot genealogy information required to perform step A is typically available from such systems as Enterprise Resource Planning (ERP) systems, paper record systems, Manufacturing Execution Systems (MES), batch record systems, etc.
- Step B is to create *lot trees* by joining individual mapping views to map the relationships between any two non-adjacent steps in the process.
- Finally, lot trees are joined to data to produce the data views that are used to perform cross-step correlations (Step C).

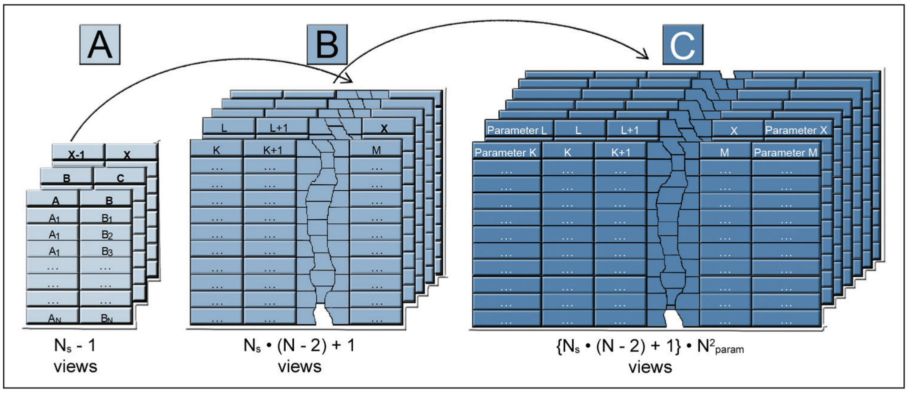


Figure 5. A typical genealogy data modeling workflow, which is the same for the manual and on-demand SQL genealogy approaches.

# Steps	# Parameters at each Step	# Views to Create and Maintain
2	1	1
2	2	4
5	3	144
10	5	2,025
30	10	84,100

Table D. Number of data views as a function of the number of process steps and parameters.

In the manual SQL approach described above, all three steps (A, B, and C) must be performed manually in advance, which results in the need to create a total of 84,100 views to support calculations between all pairs of parameters in the 30-step process as seen in Table D. In the on-demand SQL approach, only mapping views (step A) need to be constructed manually, while a corresponding set of re-usable lot trees (step B) and data views (step C) can be constructed by the software at the time of a user request. This is done by joining only those mapping views necessary for the query at the time the query is executed to map the step-to-step relationships for the portion of the process between the start and end points requested by the user. For example, in the 30-step process referred to in Table D, only 29 mapping views would need to be constructed as compared to the construction in advance of 84,100 views that would be required to support calculations between all pairs of parameters in the 30-step process as seen in Table D in the manual SQL approach.

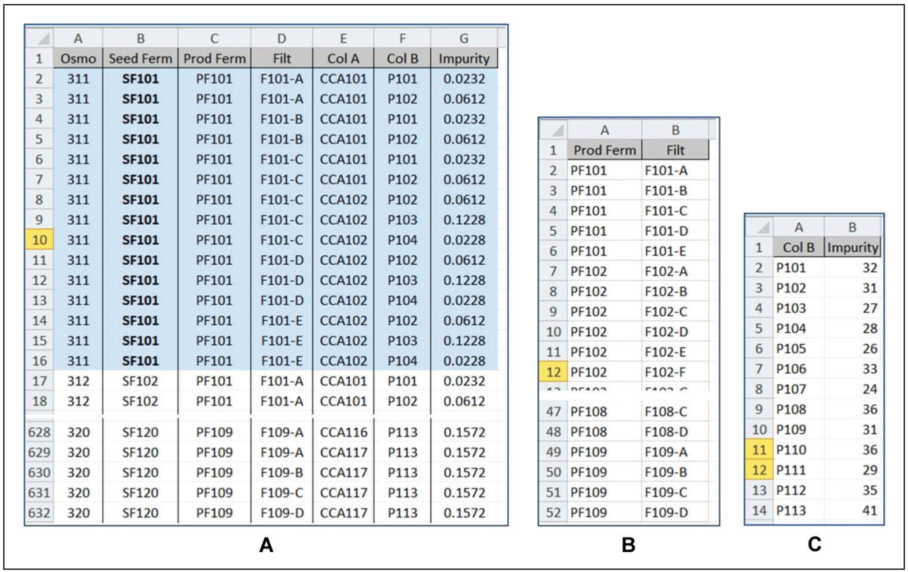
Example 2. To compare directly the effort involved in the three approaches described here (manual spreadsheet, manual SQL, and on-demand SQL), consider the following example of a fermentation process shown in Figure 4. Batches in this process are frequently split and pooled between the unit operations (seed fermentor, production fermentor, filtration, and two column purification steps) such that the starting material from 20-seed fermentor batches ends up in 13 column B purification batches (note that Figure 4 shows only a subset of all the batches). The goal is to prepare all the required data tables needed to correlate parameters between any two of the five steps of this process by using the three methods described in this article.

A. Manual spreadsheet approach.

1. Starting with the Seed Fermentor batch SF101 and looking at the lot genealogy map in Figure 4, all the sequences of batches that link batch SF101 to the Column B purification batches are manually entered (Figure 6A, rows for batch SF101 are highlighted in blue).
2. Repeating step 1 for all 20-seed fermentor batches produces an Excel spreadsheet containing 631 rows (Figure 6A).

3. The spreadsheet created in step 2 can now be used to perform correlations between parameters from the Seed Fermentor and column B purification steps where the seed fermentor step defines the degrees of freedom (i.e., the seed fermentor step is the universe). To allow for correlations between all the other two out of five steps, 15 more spreadsheets would be needed, each containing a subset of the table shown in Figure 6A.

To summarize this manual spreadsheet approach, the total number of rows in the 16 spreadsheets in this example was 2,200 and each of the spreadsheets are created by manual manipulation of data in Excel. Any addition of new batches would require updating each of these 16 spreadsheets.



	A	B	C	D	E	F	G
1	Osmo	Seed Ferm	Prod Ferm	Filt	Col A	Col B	Impurity
2	311	**SF101**	PF101	F101-A	CCA101	P101	0.0232
3	311	**SF101**	PF101	F101-A	CCA101	P102	0.0612
4	311	**SF101**	PF101	F101-B	CCA101	P101	0.0232
5	311	**SF101**	PF101	F101-B	CCA101	P102	0.0612
6	311	**SF101**	PF101	F101-C	CCA101	P101	0.0232
7	311	**SF101**	PF101	F101-C	CCA101	P102	0.0612
8	311	**SF101**	PF101	F101-C	CCA102	P102	0.0612
9	311	**SF101**	PF101	F101-C	CCA102	P103	0.1228
10	311	**SF101**	PF101	F101-C	CCA102	P104	0.0228
11	311	**SF101**	PF101	F101-D	CCA102	P102	0.0612
12	311	**SF101**	PF101	F101-D	CCA102	P103	0.1228
13	311	**SF101**	PF101	F101-D	CCA102	P104	0.0228
14	311	**SF101**	PF101	F101-E	CCA102	P102	0.0612
15	311	**SF101**	PF101	F101-E	CCA102	P103	0.1228
16	311	**SF101**	PF101	F101-E	CCA102	P104	0.0228
17	312	SF102	PF101	F101-A	CCA101	P101	0.0232
18	312	SF102	PF101	F101-A	CCA101	P102	0.0612
628	320	SF120	PF109	F109-A	CCA116	P113	0.1572
629	320	SF120	PF109	F109-A	CCA117	P113	0.1572
630	320	SF120	PF109	F109-B	CCA117	P113	0.1572
631	320	SF120	PF109	F109-C	CCA117	P113	0.1572
632	320	SF120	PF109	F109-D	CCA117	P113	0.1572

A

	A	B
1	Prod Ferm	Filt
2	PF101	F101-A
3	PF101	F101-B
4	PF101	F101-C
5	PF101	F101-D
6	PF101	F101-E
7	PF102	F102-A
8	PF102	F102-B
9	PF102	F102-C
10	PF102	F102-D
11	PF102	F102-E
12	PF102	F102-F
47	PF108	F108-C
48	PF108	F108-D
49	PF109	F109-A
50	PF109	F109-B
51	PF109	F109-C
52	PF109	F109-D

B

	A	B
1	Col B	Impurity
2	P101	32
3	P102	31
4	P103	27
5	P104	28
6	P105	26
7	P106	33
8	P107	24
9	P108	36
10	P109	31
11	P110	36
12	P111	29
13	P112	35
14	P113	41

C

Figure 6. A) A data view created in MS Excel as part of the manual spreadsheet approach; B) An example of the mapping view required by manual and on-demand SQL methods; C) An example of a parameter view used in manual and on-demand SQL approaches.

B. Manual SQL approach.

1. Four mapping views are created, each containing ~ 50 records. An example of the mapping view linking the production fermentor and filtration batches is shown in Figure 6B.
2. Five parameter views are created, one for each process step, each containing two fields: batch IDs of the step and parameter values. Figure 6C depicts one of the parameter views created in this example.
3. To map batch IDs from non-adjacent steps, 12 lot trees are constructed by joining mapping views as appropriate. This step requires writing complex SQL queries.
4. To allow for correlations between each two out of five steps, 16 data views are constructed by joining parameter views to 12 lot trees from step 3 and to 4 mapping views from step 1)

To summarize this manual SQL approach, nine views are constructed with a total of 280 records, and 12 lot trees and 16 data views are built using complex SQL commands.

C. On-demand SQL approach.

1. Similar to the manual SQL approach, four mapping views and five parameter views are built that are identical to those for the manual SQL approach. No creation of lot trees and data views or writing of SQL statements is required as all lot trees and data views are constructed by the software at the time the user selects steps and parameters to correlate. Furthermore, in many industrial applications, the manual step of creating step views may not be required either because these views may already be available in such systems as, for example, MES or the electronic batch record.

Table E summarizes the effort required to provide the user with the ability to perform correlations between parameters at any two out of five steps in the example process. Notice that the manual spreadsheet approach is the most effort intensive and error prone and the on-demand SQL approach is the least laborious and does not require either manual data manipulation or SQL skills. The advantage of the on-demand SQL approach becomes more pronounced as the number of steps and batches increase.

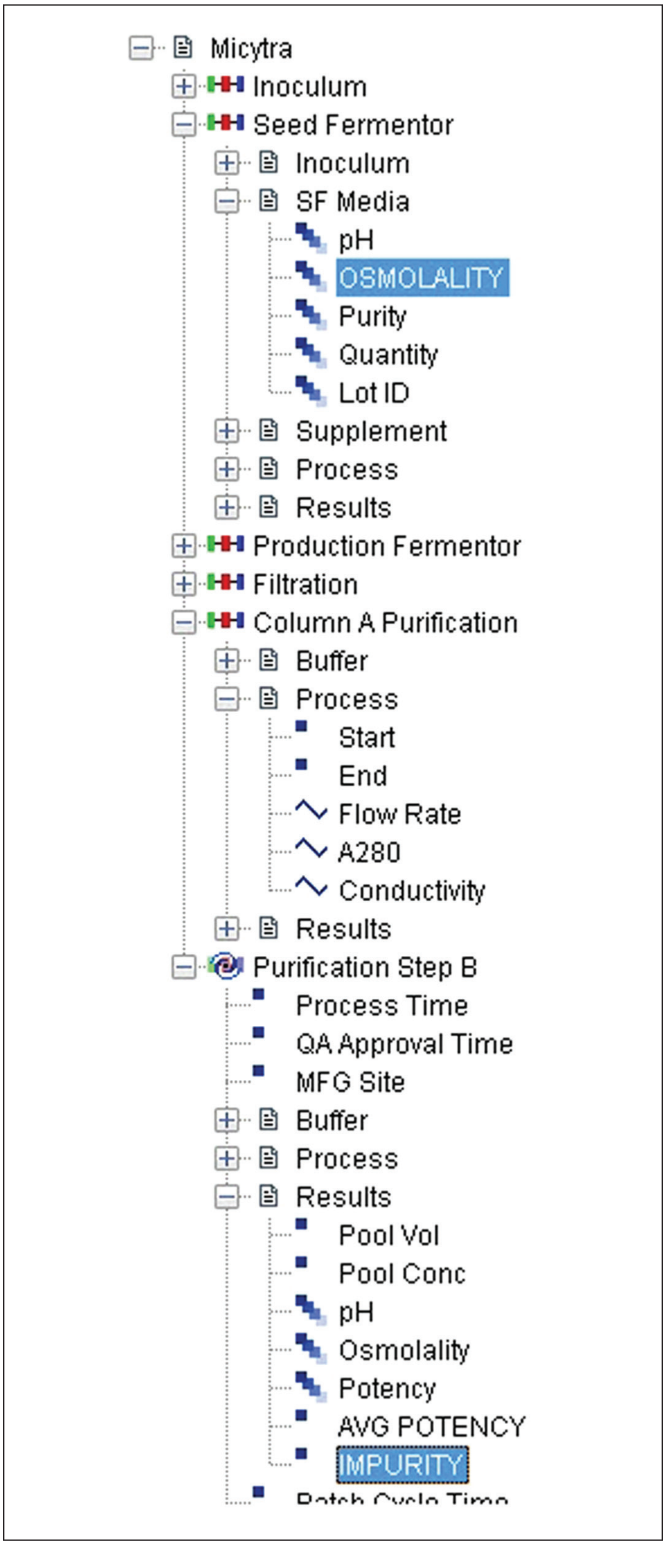
To enable an upstream/downstream correlation, data must be joined with lot trees (Figure 5, step C); therefore, all of the corresponding data must be pre-organized by batch. Translating data into the batch context in the manual SQL approach usually takes significant additional data modeling effort, thus slowing down this type of analysis and making it more complex and error prone. On the other hand, the on-demand SQL approach described here also can embody built-in data contextualization capabilities to automate all of the data modeling query generation required to support upstream/downstream calculations.

Next, data is made available by accessing it directly from an on-screen hierarchical view of the process flow linked to an on-demand SQL generator which populates the "where clauses" in the SQL queries using the node names in the hierarchy

Method	Manual Spreadsheet	Manual SQL	On-demand SQL
# Records manipulated manually	2,200	0	0
# Views created manually or retrieved from database	0	9	9
# Views created with SQL statements	0	28	0

Table E. Effort involved to organize data for upstream/downstream correlation analysis in Example 2.

(Figure 7). In this example, the data is made available in a form that is contextualized by batch as the organizing principle of the hierarchical view (i.e., with all the parameter values organized so that they are associated with their corresponding batches to enable easy comparisons between batches). Other data organizing principles can be used for such hierarchies, such as production shifts, individual unit operations, raw materials, sample or test IDs, and other organizing identifiers used in enterprise data systems.



- Micytra
  - Inoculum
  - Seed Fermentor
    - Inoculum
    - SF Media
      - pH
      - OSMOLALITY
      - Purity
      - Quantity
      - Lot ID
    - Supplement
    - Process
    - Results
  - Production Fermentor
  - Filtration
  - Column A Purification
    - Buffer
    - Process
      - Start
      - End
      - Flow Rate
      - A280
      - Conductivity
    - Results
  - Purification Step B
    - Process Time
    - QA Approval Time
    - MFG Site
    - Buffer
    - Process
    - Results
      - Pool Vol
      - Pool Conc
      - pH
      - Osmolality
      - Potency
      - AVG POTENCY
      - IMPURITY
    - Batch Cycle Time

Figure 7. A hierarchy for a process with a complex lot genealogy between the seed fermentor and the final column purification step. To correlate the media osmolality to the impurity, the user selects two parameters shown by clicking them on the screen.

Finally, weighted averages based on the fractional contributions of upstream steps to downstream steps are automatically calculated as part of the batch contextualization function. Table B illustrates how the data should be grouped and aggregated to accomplish such batch contextualized averaging. In this case, weighted averages are calculated by using the cardinality of mixing between the media lots. Media lots are considered to be the "step universe" which is the higher-level organizing principle around which the other organizing principles are organized. The average impurity values across all production fermentations are calculated for each media lot in the "step universe" in this illustration. For each media lot in the "step universe," the weighted average impurity value is calculated by averaging impurity values across all replicate production fermentor lots in which that particular media lot was used. This type of mapping leads to:

1. Replicate parameters, defined as multiple impurity values associated with each media lot, in which each replicate value represents the endpoint of a pathway by which a given media lot contributes to the impurity outcome of each fermentor batch. (Note: These replicate values can be used to calculate the average impurity (weighted-by-cardinality) associated with each media lot.)

2. Equal numbers of input parameters and impurity outcomes as required for meaningful correlation calculations.

Similarly, data also can be grouped using the production fermentor batches as the step universe in order to calculate the weighted average osmolality values associated with each production fermentor batch. As a general rule, when correlating parameters between different process steps, the step with the smallest number of batches should be used as the step universe, to minimize the number of degrees of freedom and thus avoid the overestimating the correlation.

However, in some types of analyses, such as comparing raw material suppliers using Analysis of Variance (ANOVA), selecting the step with the smaller number of batches as the step universe can be impractical. Referring to the example of comparing vendors of nutrient supplements above (Figure 3), the seed fermentation step should be used as the step universe to avoid having to do calculations of the weighted averages of vendors that would be required if the purification step were selected. Figure 8 also illustrates that one of the vendors (shown in the middle of the figure) is associated with a significantly higher level of impurity in the filtered bulk.

The following steps can be used to correlate a parameter from step A (Par A) to a parameter from step X (Par X) using the SQL method:

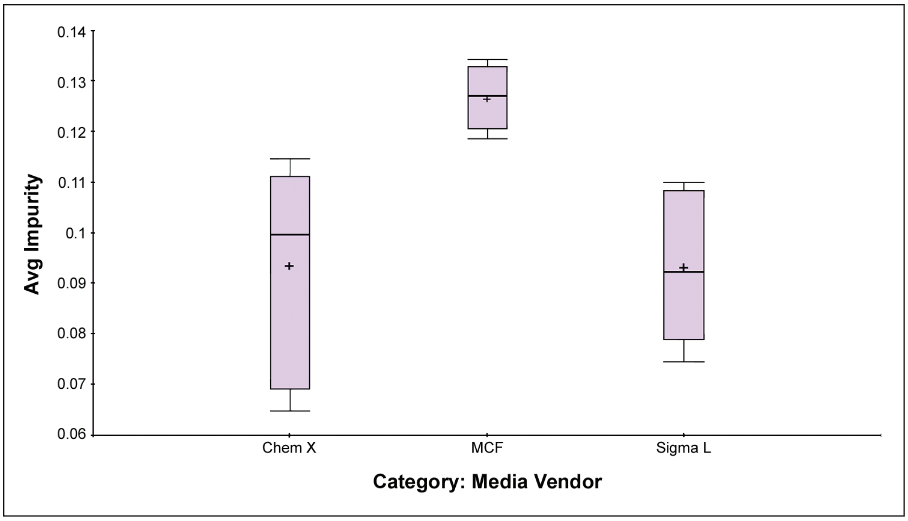


Figure 8. The nutrient supplement vendor shown in the middle of the figure is associated with a significantly higher level of contaminant in the filtered bulk using the seed fermentation step as the step universe.

1. Prepare Mapping Views so that each view maps two adjacent steps between A and X. This will require a total of X-A mapping views: [A]->[A+1], [A+1]->[A+2], … ,[X-1]->[X].

2. Create a lot tree by joining all X-A mapping views created in step 1.

3. Prepare views for **Par A** organized by the batch ID at step A and **Par X** organized by the batch ID at step A.

4. Join **Par A**, **Par X** views to the lot tree created in step 2.

5. Decide which step will be the organizational unit (discussed above).

6. Calculate the weighted average of the parameters by grouping the view created in Item 4 by the step universe batch ID.

7. Perform the analysis.

Conclusion

This study considered three approaches to performing upstream/downstream data analysis, focusing on the differences between them with regard to labor intensity, complexity, ability to account for splitting and pooling in the process stream, and their ability to inadvertently propagate errors. The manual spreadsheet approach was the most labor-intensive, complex, time consuming, and error prone way to perform upstream /downstream data analysis and was severely limited in its ability to account for splitting and pooling in the process stream genealogy without significant risks of inadvertent error propagation. In a complex manufacturing process with 20 to 25 unit operations containing four or five points of splitting and pooling in a process stream, the manual spreadsheet approach would need tens of thousands of spreadsheet rows to allow for the necessary calculations to correlate upstream inputs with downstream outcomes. Such a complex spreadsheet could probably not be used without a significant number of errors, thus providing questionable functionality. The manual SQL approach was potentially less error prone than the manual spreadsheet approach, but was still too labor intensive and complex to be useful as a practical tool for complex manufacturing processes. The on-demand SQL genealogy approach required an initial investment in the development and configuration of mapping views similar to that used in the manual SQL approach. Once the initial investment was made, this approach provided a high degree of reuse of the views along with minimal potential for errors, simplicity of use, and the ability to easily perform upstream/downstream correlations in complex manufacturing processes with multiple points of splitting and pooling in the process stream.

An important benefit of being able to easily perform upstream/downstream correlations in complex manufacturing processes is that significant barriers are removed to identifying potential cause-and-effect relationships between upstream process conditions and downstream process outcomes. Such relationships drive the formation of hypotheses that can be confirmed, extended, or refuted using mechanistic knowledge and/or experimentation. The information thus gained about the relationships between upstream process parameters and downstream process outcomes is a major component of process models used for process control, and also contributes in the development of sophisticated process models for use in Real Time Adaptive Control (RTAC).

A complex manufacturing process with multiple splits and recombinations in the process stream may be operating in a state of control until a process upset occurs (e.g., an unexpected change in a raw material which threatens to produce unacceptable downstream outcomes). In this situation, the control system must be supported by a process model to determine what adjustments to make (either automatically or with the help of manual intervention) to re-establish control of the process within the design space. Such a process model would be most efficiently prepared using the on-demand SQL approach described in this study so that the quantitative relationships between upstream parameters and downstream process outcomes is available to the control system to make the appropriate adjustments.

The on-demand SQL genealogy approach described in this study can be embodied in a computer software program that allows process models to be built efficiently and with minimum potential for errors. Such a software program could provide data values for process parameters contextualized by batch and organized to include the genealogy of the process stream. This would simplify and reduce errors in the work involved in understanding upstream/downstream parameter relationships in complex processes that include splitting and pooling in the process stream, a critical success factor for building process models that link CPPs and CQAs.

In summary, lot traceability is an important capability for recall management, but it is not enough to support the development of sufficient process understanding for achieving the goals of QbD. A flexible capability for performing upstream/downstream correlations, such as the on-demand SQL approach described in this study, accounts for fractional contributions across process steps and makes it possible to draw statistically sound conclusions about the relationships between upstream process parameters and downstream process outcomes. This helps to make processes better understood and outcomes more predictable by linking CPPs with CQAs to shape useful process models that meet the goals of QbD.

References

1. Rockoff, L., *The Language of SQL*: *How to Access Data in Relational Databases*, 2011.

2. Hall, Marshall, *Combinatorial Theory*, John Wiley & Sons, 440 pages, 1998, 2nd Edition.

Acknowledgements

We would like to thank Brent Rognlie for conceiving the original ideas presented here in the “On-Demand SQL Genealogy Approach” and for critical reading of this manuscript, and Joe Ruth for developing the first software system that embodied those ideas. We also would like to thank Kate Lusczakoski for her critical reading of this manuscript.

About the Authors

Victor Shashilov, PhD is Senior Applications Specialist at Aegis Analytical Corp. He has more than nine years of experience in statistics, chemometrics, latent variable methods, bio-spectroscopy, and protein studies. As part of his PhD studies, Dr. Shashilov developed new multivariate algorithms for the analysis of continuous and spectral data. At Aegis Analytical, Dr. Shashilov assists in creating product requirements and specifications, developing and implementing new analytical capabilities, writing white papers and technical briefs, and the creation and delivery of user and statistical training. He also provides support for using Discoverant software with external statistical, reporting, and display applications. Dr. Shashilov holds a PhD in physical chemistry from State University of New York (2007) and an MS in physics from Belarus State University (2000). Dr. Shashilov has published refereed papers in chemical and biological journals and serves as a reviewer for the American Chemical Society and World Scientific Publishing. He has been a speaker at scientific conferences in the field chemistry and protein studies. He can be contacted by telephone: +1-303-625-2100 or email: vshashilov@aegiscorp.com.

Aegis Analytical Corp., 1380 Forest Park Circle, Suite 200, Lafayette, Colorado 80026, USA.

Justin O. Neway, PhD is Vice President and Chief Science Officer at Aegis Analytical Corp. He has more than 30 years of experience in pharmaceutical and biotechnology process development and manufacturing, and in the application of software solutions to quality compliance and operational issues in pharmaceutical and biotherapeutic manufacturing. During the 15 years prior to founding Aegis in 1997, Dr. Neway was with Wyeth Biosciences, Novartis, and Baxter Hemoglobin Therapeutics, where he held increasingly responsible positions, including technical manager, principal scientist, project leader for the CMC section of the IL-2 BLA, department director, and director of process sciences. During this time, Dr. Neway became intimately familiar with the practical difficulties of accessing process development and manufacturing data from separate databases and on paper records, and the significant paybacks achievable from connecting to that data in real time for investigational analysis. Dr. Neway received his BSc (microbiology, 1975) and MSc (biochemistry, 1977) from the University of Calgary (Canada), and his PhD in biochemistry from the University of Illinois (USA) in 1982. He has published refereed papers in scientific journals as well as other papers, editorials, op-ed pieces, and interviews in trade publications. He also has presented technical papers and chaired technical conferences in the field of pharmaceutical manufacturing, quality compliance, and data analysis. He can be contacted by telephone: +1-303-625-2100 or email: jneway@aegiscorp.com.

Aegis Analytical Corp., 1380 Forest Park Circle, Suite 200, Lafayette, Colorado 80026, USA.