

**DEVELOPMENT AND IMPLEMENTATION OF AN  
ENTERPRISE-WIDE DATA QUALITY  
IMPROVEMENT (EDQI) SYSTEM  
FOR BIOBANKING SOFTWARE**

By

James S. Morgan, MS

A Dissertation Submitted to Rutgers University – School of Health Related  
Professions in Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics

Spring 2015



**Final Dissertation Defense Approval Form**

Development and Implementation of an Enterprise-Wide Data  
Quality Improvement (EDQI) System for Biobanking Software

**BY**

James S. Morgan

**Dissertation Committee:**

Shankar Srinivasan, PhD, Committee Chair

Frederick Coffman, PhD, Committee Member

Sandip Patil, PhD, Committee Member

**Approved by the Dissertation Committee:**

\_\_\_\_\_

Date: \_\_\_\_\_

\_\_\_\_\_

Date: \_\_\_\_\_

\_\_\_\_\_

Date: \_\_\_\_\_

## **ABSTRACT**

To obtain a high level of data quality “high level” must be defined in detail, the current quality level needs to be objectively assessed and a system to eliminate any discrepancy between the two levels will need to be developed, implemented and measured to determine effectiveness. This study focuses on the development and implementation of an enterprise-wide data quality improvement (EDQI) system, for the biobank of Cincinnati Children’s Hospital Medical Center (CCHMC). The EDQI system was designed as a two-part system. One part is customized per the data quality needs of each functional group. The second part is generic and is applied across the entire CCHMC biobanking data system. Results of the two parts are combined and delivered as a seamless data set for end users to analyze and correct non-compliant data elements.

The EDQI system was implemented alongside a vendor developed biobanking software system, BTM (Biomaterial Tracking and Management), written by DSI (Daedalus Software Inc.). A uniform intake process was developed to gather data quality requirements per biobanking unit along with a uniform SQL Server Reporting Services (SSRS) results output process.

## ACKNOWLEDGEMENTS

I am very thankful for so many academic and professional advisors that have helped shape this dissertation. This is not an all-inclusive list by any means, but just a few individuals I would like to acknowledge.

- Dr. Shankar Srinivasan, Rutgers University - deserves a medal for literally answering thousands of my questions over the past few years and particularly for his guidance in a dissertation that will continue to improve data quality for major pediatric medical center for many years to come.
- Dr. Frederick Coffman, Rutgers University – provided contributions and expertise in pathology that helped to shape the dissertation committee from both data and biobanking perspectives.
- Dr. Sandip Patil, Cincinnati Children’s Hospital – personally exemplifies continuous quality improvement and brought a wealth of academic and professional experience to the dissertation committee.
- Dr. Keith Marsolo, Cincinnati Children’s Hospital – has been a mentor to me from the time he hired me into Cincinnati Children’s Hospital. He is a ‘trail blazer’ in pediatric medical informatics and I have learned a great deal from him.

## DEDICATION

I am thankful that God has blessed me with both the opportunity and ability to complete this dissertation. The individuals below have been special blessings in my life.

- Holly – my wife of 19 years. Together we have been through infertility issues, an emotional adoption, compromised health and financial challenges. Her strength, character and example through everything have served as an inspiration to me to continually strive to improve in all areas of life.
- Abby – my daughter. Her conduct and character make me an extremely proud father. While her childhood seemed to slip by in the blink of an eye, I could not be happier with the young lady she has become.
- Zachary – my son. He made the heart-ache of the adoption process disappear instantly and brought an element of exhilaration to our lives that we could not have previously imagined. He also makes me an extremely proud father.
- James and Lucille – my parents. In word and in deed they exemplify the meaning of a loving family and are my biggest ‘fans’. They

instilled in me both the value of higher education and the principle of loving one's neighbor as oneself. I am honored to be their son.

- Zelma – my grandmother. Educated via reading her children's college text books, in her mid-nineties, she remains as sharp as anyone I know.

## TABLE OF CONTENTS

ABSTRACT.....	iii
ACKNOWLEDGEMENTS.....	iv
DEDICATION.....	v
LIST OF FIGURES.....	x
LIST OF TABLES.....	xiv
CHAPTER 1 INTRODUCTION.....	1
1.1 Background Terms and Application.....	1
1.2 Technological Terms and Application.....	4
1.3 The Need for High Data Quality .....	9
1.4 Basic EDQI Functions.....	10
1.4.1 Identifying Non-Compliant Data Elements.....	11
1.4.2 Notifying the Data Owner of Non-Compliant Data Elements.....	13
1.5 Hypothesis and Goals.....	15
CHAPTER 2 LITERATURE REVIEW and RESEARCH GAP.....	16
2.1 Challenges.....	17

2.2	Solutions.....	18
2.3	Literature Review Themes.....	19
2.3.1	Summary of Themes.....	26
2.4	Research Gaps.....	27
CHAPTER 3 METHODOLOGY.....		27
3.1	Algorithm and System Design.....	27
3.2	EDQI Work Flow Algorithm.....	30
3.3	CCHMC Biobanking Data System and EDQI Components.....	36
3.3.1	Component #1, Epic and Cerner Systems.....	39
3.3.2	Component #2, Clarity and Red Light/Green Light.....	41
3.3.3	Component #3, Better Outcomes for Children Project.....	45
3.3.4	Components #4 - 7; PAH, CAGE, PRTR, Neonatology.....	51
3.3.5	Component #8, BTM Application.....	62
3.3.6	Component #9, i2b2 ETLs from BTM Data Warehouse.....	71
3.3.7	Component #10, Enterprise vs. Bank Level Queries.....	78
3.3.8	Component #11, Non-Compliant Data Process.....	82
CHAPTER 4 RESULTS.....		86
4.1	Initial Results.....	86
4.2	Initial Bank Level Results.....	87

4.3 Initial Enterprise Level Results .....	93
4.4 Initial Results Analysis.....	94
4.5 Post Implementation Results.....	96
CHAPTER 5 CONCLUSIONS and DISCUSSION.....	100
CHAPTER 6 FUTURE OPPORTUNITIES.....	101
REFERENCES.....	104

## LIST OF FIGURES

Figure 1-1: Percentage of Data per Standard Deviation .....	12
Figure 1-2: Sample EDQI Email.....	13
Figure 1-3: Sample EDQI Report.....	14
Figure 3-1: EDQI Algorithm Design .....	29
Figure 3-2: EDQI Work Flow Diagram.....	35
Figure 3-3: Component Level Overview of the EDQI System.....	37
Figure 3-4: Component Level Overview of the EDQI System Identified by Functional Relationship.....	38
Figure 3-5: Process and High-Level Data Flows from Inception to Completion.....	41
Figure 3-6: Process Location of RL/GL Application.....	42
Figure 3-7: BOFC Upload Components.....	46
Figure 3-8: Timing and Sequence of BOFC SQL Server Jobs.....	50
Figure 3-9: BOFC Folder Structure.....	51
Figure 3-10: Generic Upload Shared Drive Folder Structure.....	52

Figure 3-11: Generic Upload Processing Sequence.....	53
Figure 3-12: Generic Upload Error Detection Process.....	54
Figure 3-13: Generic Upload Post Import Data Placement.....	55
Figure 3-14: ProcessMigrationSet Properties.....	56
Figure 3-15: BTM Insertion Code.....	57
Figure 3-16: Migration Set Error Handling Code.....	58
Figure 3-17: Migration Set SSRS Automated Error Email.....	59
Figure 3-18: Migration Set SSRS Error Report.....	59
Figure 3-19: Sample Debiting Code Snippet.....	61
Figure 3-20: ETL Job Properties.....	61
Figure 3-21: Biobanks in CCHMC's Production Environment.....	62
Figure 3-22: Expanded Views of BTM's Navigation Panel.....	63
Figure 3-23: General Tab of BTM's Accession Biofluid Sample Screen...	64
Figure 3-24: Sample Description Tab of BTM's Accession Biofluid Sample Screen.....	65
Figure 3-25: BTM Annotation Form – Juvenile Idiopathic Arthritis.....	66
Figure 3-26: Expanded View of BTM's Storage Hierarchy.....	67

Figure 3-27: Holder Level View of BTM's Storage Hierarchy .....	68
Figure 3-28: General Tab of BTM's Create/Edit Patient Screen .....	69
Figure 3-29: Project Tab of BTM's Create/Edit Patient Screen.....	69
Figure 3-30: Family Membership Tab of BTM's Create/Edit Patient Screen.....	69
Figure 3-31: BTM Relationship Overview.....	71
Figure 3-32: i2b2 Production Workbench.....	72
Figure 3-33: BTM FactSample Table.....	74
Figure 3-34: BTM FactSubject Table.....	74
Figure 3-35: BTM to i2b2 Data Flow Diagram.....	75
Figure 3-36: BTM Data Warehouse Entity Relationship Diagram (Macro View).....	76
Figure 3-37: Detailed View of Section #1 of the BTM Data Warehouse.....	77
Figure 3-38: Detailed View of Section #2 of the BTM Data Warehouse...	77
Figure 3-39: Detailed View of Section #3 of the BTM Data Warehouse....	78
Figure 3-40: BTM Help Homepage.....	79
Figure 3-41: BTM Help Website Data Quality Improvement Report Page.	80
Figure 3-42: EDQI and Bank Level Query Data Flow.....	81
Figure 3-43: BTM Data Quality Check Report.....	83
Figure 3-44: BTM Accession Page with Verification Complete Field.....	85

Figure 4-1: Non-Compliant Data Results for PAH Bank.....	87
Figure 4-2: Non-Compliant Data Results for HIBR Bank.....	90
Figure 4-3: Non-Compliant Data Results for CCHMC Bank.....	92
Figure 4-4: Enterprise Level Non-Compliant Results for All CCHMC Banks.....	93
Figure 4-5: PAH Non-Compliant Data Elements Before EDQI Implementation.....	96
Figure 4-6: PAH Non-Compliant Data Elements After EDQI Implementation.....	96
Figure 4-7: HIBR Non-Compliant Data Elements Before EDQI Implementation.....	98
Figure 4-8: HIBR Non-Compliant Data Elements After EDQI Implementation.....	98
Figure 4-9: CCHMC Non-Compliant Data Elements Before EDQI Implementation.....	98
Figure 4-10: CCHMC Non-Compliant Data Elements After EDQI Implementation.....	99

## **LIST OF TABLES**

Table 1-1: CCHMC Pediatric Biosamples by Bank.....	3
Table 4-1: Non-Compliant Data Results for PAH Bank.....	89
Table 4-2: Non-Compliant Data Results for HIBR Bank.....	91
Table 4-3: Non-Compliant Data Results for CCHMC Bank.....	92
Table 4-4: Non-Compliant Data Results for All CCHMC BTM Banks.....	93

# **CHAPTER 1 INTRODUCTION**

This study describes the need, development, composition, implementation and initial results of an Enterprise Data Quality Improvement (EDQI) system for biobanking software at Cincinnati Children's Hospital Medical Center (CCHMC). At its highest level, the EDQI system is responsible for two primary functions:

1. Identify non-compliant data
2. Notify users of the identified non-compliant data

There are numerous inter-related components of the EDQI system that allow it to perform these primary functions which will be described in this study, but ultimately, the system succeeds or fails its objective based on its ability to successfully carry out these two functions.

## **1.1 Background Terms and Application**

The following are high-level definitions of some key terms that will be referenced throughout this study. After each term there is a description of the application of the term specific to this study within CCHMC. Both this section and the subsequent one are included as many of these terms are

exclusive to either the biobanking industry or a proprietary technology accompanying the biobanking industry.

Biobanking – the process of storing biological samples for future use in research or clinical application; and capturing, retaining and maintaining data associated with those samples.<sup>20</sup> Specific to this study, the storage process references preparation for a biofluid, bio-tissue, or nucleic acid to be placed in a freezer that ranges from -4 ° to -80° Celsius. CCHMC has approximately 45 biobanking freezers and over 500,000 pediatric biosamples.

CCHMC (Cincinnati Children's Hospital Medical Center) – is a 598-bed pediatric hospital located in Cincinnati, Ohio. CCHMC operates under the University of Cincinnati's Department of Pediatrics and is one of the United States' leading pediatric research and teaching institutions. CCHMC is ranked third among all Honor Roll hospitals in the 2014 U.S. News & World Report survey of best children's hospitals.<sup>28</sup> This study will focus on CCHMC's biobanking data system. With over 500,000 pediatric biosamples, CCHMC may have the largest pediatric biobank in the world. The following table provides a list of each CCHMC biobanking group and the total number of samples each group has in BTM:

Bank Name	Sample Count
Bariatric Surgery Bank	10,162
CCHMC CCED BioBank	2,955
CCHMC Emergency Department Bank	942
CCHMC Genomic Control Cohort (GCC)	44,892
CCHMC Heart Institute Biorepository (HIBR)	8,361
CCHMC PAH Biobank	75,110
CCHMC Pediatric Rheumatology Tissue Repository	127,892
CCHMC Perinatal Institute/CIRHML	4,660
Cincinnati Childrens Hospital Medical Center	230,754
<b>TOTAL</b>	<b>505,728</b>

Table 1-1: CCHMC Pediatric Biosamples by Bank

Data Quality Improvement - systematic and continuous actions that lead to measurable improvement in data quality.<sup>28</sup> Data quality represents the accuracy of data compared to the data source. If the data entry system is the data source other accuracy verification measures can be used. Such accuracy verification might include statistical analysis with a review of outliers, comparison of inter-related data elements, calculation verifications or other appropriate verification based on a specific data set.<sup>28</sup> Specific to this study, the data set is the BTM data that resides in a SQL database. The verification processes consists of checking the accuracy of the biobanking data against a source outside of BTM.

Enterprise-wide Data System - a large-scale software application and underlying database that supports a business process, data reporting and data analytics in a complex organization.<sup>3</sup> Specific to this study, the large-scale

application software is BTM (Biomaterial Tracking and Management), produced by DSI (Daedalus Software Inc.). The complex organization that BTM supports is CCHMC (Cincinnati Children's Hospital Medical Center). The business process that is supported is biobanking.

FTA Card – FTA is an acronym for Fast Technology for Analysis of nucleic acids. Biological samples, such as blood and saliva, adhere to the FTA card (paper) through the mechanism of entanglement, while the mixture of chemicals lyses cells and denatures proteins. Because nucleases are inactivated, the DNA is essentially stable when the sample is properly dried and stored. Nucleic acid damage from nucleases, oxidation, ultraviolet light (UV) damage, microbes, and fungus is reduced when samples are stored on an FTA card.<sup>15</sup> Multiple CCHMC banks and projects use FTA cards, usually with a blood spot on them.

## **1.2 Technological Terms and Application**

BLOB (Binary Large Object) Text – data values treated as binary strings.

They have no character set, and sorting and comparison are based on the numeric values of the bytes in column values. Text values are treated as non-binary strings (character strings). They have a character set, and values are

sorted and compared based on the collation of the character set.<sup>17</sup> This term shows up twice in this study, once referencing the data format used for laboratory test result data passed from CCHMC's laboratory system, Cerner, and another referencing BTM's annotation data.

BTM (Biomaterial Tracking and Management) Application - software application specifically developed to help biobanks maximize the use of available samples, and to encourage collecting samples prospectively, for biomedical research. BTM lets biobanks store large numbers of samples and track each sample's movement through the research center. It lets researchers search for samples by criteria relevant to a given research project, and it lets biobank workers quickly find appropriate samples and distribute them to researchers.<sup>31</sup> In this study, BTM will be referenced frequently as it is CCHMC's biobanking data system.

BTM Bank – a subunit within the BTM application that represents a department or division of people within an enterprise whose purpose is collecting, organizing, documenting, storing, and tracking biological samples.<sup>31</sup> The specific BTM Banks referenced in this study are the following:

- CAGE (Center for Autoimmune Genomics and Etiology) Bank

- Cincinnati Children's Hospital Medical Center Bank
  - BOFC (Better Outcomes for Children) Project – BOFC is a project within the Cincinnati Children's Hospital Medical Center Bank.
- NEO (Neonatology) Bank
- PAH (Pulmonary Arterial Hypertension) Bank
- PRTR (Pediatric Rheumatoid Tissue Repository) Bank

\* The list above is not a complete list of BTM Banks currently in CCHMC's production environment. These banks are specifically mentioned in this study because each has a custom ETL process associated with it and those processes significantly affected the development of the EDQI system. A complete list of BTM Banks is provided in Figure 3-21 (page 49).

BTM GUID – BTM assigned Global Unique Identification number provided to each sample in BTM. This number is 37 digits long and is guaranteed to be unique for two generations.<sup>31</sup>

Drupal - an open source, website development software maintained and developed by a community of over 1,000,000 users and software developers. Drupal is distributed under the terms of the General Public License.<sup>6</sup> This technology was used to develop the Help-BTM website displayed in Figure 4-43.

i2b2 (Informatics for Integrating Biology and the Bedside) – a

comprehensive software and methodological framework to enable clinical researchers to accelerate the translation of genomic and “traditional” clinical findings into novel diagnostics, prognostics, and therapeutics.<sup>19</sup> For this study, CCHMC has a custom version of i2b2 which researchers use for de-identified cohort identification and analysis.

Java - a platform independent programming language expressly designed for use in the distributed environment of the Internet. It was designed to have the feature similar to the C++ language, but it is simpler to use than C++ and enforces an object-oriented programming model. Java can be used to create complete applications that may run on a single computer or be distributed among servers and clients in a network. It can also be used to build a small application module or applet for use as part of a Web page. Applets make it possible for a Web page user to interact with the page.<sup>1</sup> Specific to this study, Java is the technology that was used to develop the BTM application and the RL/GL (Red Light/Green Light) application.

RL/GL (Red Light/Green Light) Application – CCHMC custom developed traffic light application that accepts a biosample accession number and returns a green light if the sample can be retained in the biorepository and a red light if the sample should be discarded based on consent status. The underlying database for this application receives consent data from CCHMC's EHR System, Epic.<sup>18</sup>

SQL Server - a relational database management system from Microsoft that is designed for the enterprise environment. SQL Server runs on T-SQL (Transact-SQL), a set of programming extensions from Sybase and Microsoft that add several features to standard SQL, including transaction control, exception and error handling, row processing, and declared variables.<sup>2</sup> Specific to this study, all of the ETL processes described are written in MS SQL.

SSIS (SQL Server Integration Services) – a tool used to build packages to merge data from heterogeneous data sources into SQL Server. They can also be used to populate data warehouses, to clean and standardize data, and to automate administrative tasks. Integration Services provides a platform to build data integration and workflow applications. The primary use for SSIS

is data warehousing as the product features a fast and flexible tool for data ETLs. The tool may also be used to automate maintenance of SQL Server databases, update multidimensional cube data, and perform other functions.<sup>5</sup> Specific to this study, SSIS was used to develop the BTM data warehouse and ETL jobs associated with it.

SSRS (SQL Server Reporting Services) - a server-based report generation software system from Microsoft. Administered via a web interface, it can be used to prepare and deliver a variety of interactive and printed reports. Reports are defined in RDL (Report Definition Language), an XML markup language. Reports can be designed using recent versions of Microsoft Visual Studio, with the included Business Intelligence Projects plug-in installed or with the included Report Builder, a simplified tool that does not offer all the functionality of Visual Studio.<sup>4</sup> Specific to this study, SSRS was used to setup the subscription email services for the EDQI and develop the reports displaying any non-compliant data elements discovered.

### **1.3 The Need for High Data Quality**

There are several reasons that necessitate high data quality from CCHMC's biobanking data:

1. Research discoveries are based on this data (particularly genomic research).
2. Biobanking data is used for numerous publications both within and outside of the biobanking industry. Erroneous data would be problematic for any healthcare specialty dependent on data from biobanking services.
3. There is some clinical overlap in biobanking data. While it may be indirect, some biobanking data contributes to clinical care decisions.

Due to the crucial need for high quality data in CCHMC's biobanking system, a system design approach was necessary for enterprise level quality improvement. The design had to be independent of any individual employee or commercial component to be effective as a long-term quality improvement solution.

## **1.4 Basic EDQI Functions**

At its highest level, the two basic functions of the EDQI system are:

1. Identify non-compliant data elements
2. Notify the data owner of the non-compliant data elements identified

### **1.4.1 Identifying Non-Compliant Data Elements**

Identifying non-compliant data elements first requires gathering appropriate requirements at both the enterprise and individual bank levels. Part of the requirements process is defining “non-compliant” data. At the enterprise level, these are standard requirements that are the same for all the banks in BTM. One example of this is one MRN should not be assigned to multiple patients within one medical facility. This is a standard requirement and a query can be developed to search for this condition. At the bank level, however, the definition will not be standard, but rather specific to a particular bank’s operations. Bank level requirements may involve specific sample, subject or patient characteristics. These characteristics may or may not be conditional. Some of these may be statistically driven, for example: a particular bank desires to know the mean of the volume of all their samples and they wish to identify all samples that fall beyond three standard deviations beyond the mean. These data elements would compose .3% of the particular data set examined see the figure below.

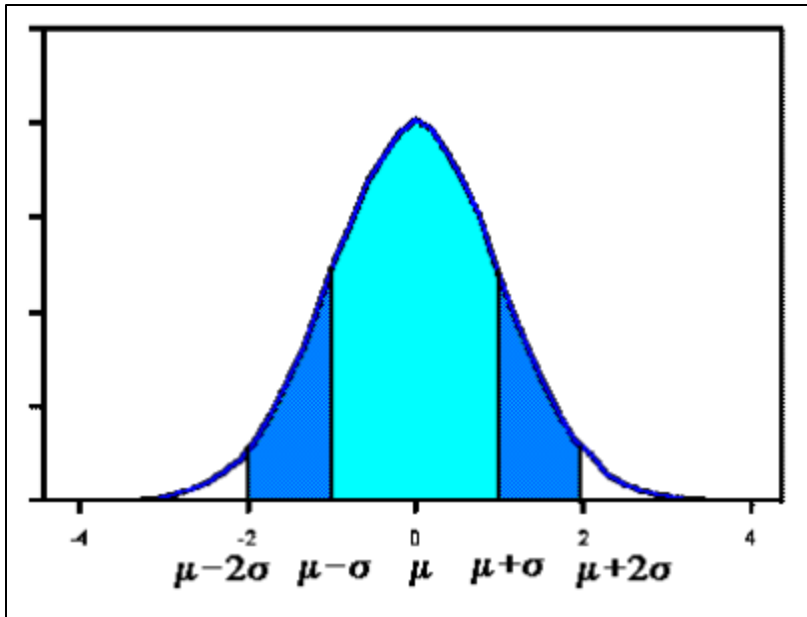


Figure 1-1: Percentage of Data per Standard Deviation<sup>21</sup>

The non-compliant identification process is agnostic of data entry method: manual, automatic (bulk upload) or migration from another data source.

Although this data check occurs subsequent to the actual data entry, and thus is not performed “real time”, the sequence is appropriate as both the automatic (bulk upload) and migration data entry methods by-pass any application level data protection mechanisms. The objective of the bulk upload ETL processes is to enter large amounts of data quickly. The frequency options of the non-compliant quality report developed via SSRS can generally minimize any time interval of data quality risk to an acceptable level. For this study, the risk is minimal enough that the EDQI system can be run against the production database. However, the EDQI

system could be implemented in a test environment and any bulk upload ETL process or migration data could be extracted into this environment and the non-compliant quality report could be run against this data prior to be pushed into the production environment. This would allow any non-compliant data to be addressed before entering the production environment.

#### 1.4.2 Notifying the Data Owner of Non-Compliant Data Elements

The notification process takes place via an email to the data owner to conduct an investigation per the data elements identified. The email is produced through an SSRS subscription. The email contains a report, this can be in the body of the email or as an attachment, which lists each non-compliant data requirement and the total number of data elements that were identified via the query developed from the user's requirements.

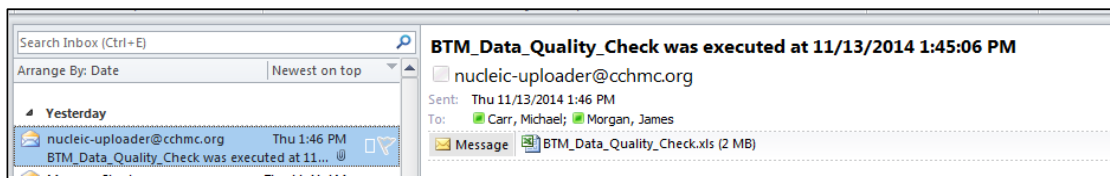


Figure 1-2: Sample EDQI Email

1	2	A	B	C	D	E	F	GH	I	J	K	L	M	O	P	Q	R	T
	1	BTM Data Quality Check Report																
	2	Bank: CCHMC PAH Biobank										Date: 11/13/2014						
	3																	
	4	Missing CollectionSiteEnumID										Count: 100						
+	5	Sample GUID										Bank Specific Sample Id						
	108																	
	111	Missing CollectionDateFrom										Count: 0						
	112	Sample GUID										Bank Specific Sample Id						
	115																	
	117	Collection/Arrival Difference										Count: 3392						
+	118	Sample GUID										Bank Specific Sample Id			Collection Arrival Dates Difference			
	3513																	
	3516	Incorrect MostRecentMedicalFacilityID										Count: 376						
+	3517	Sample GUID										Bank Specific Sample Id			Medical Facility Name			
	3895																	
	3897	Incorrect Project Name										Count: 7						
+	3898	Sample GUID										Bank Specific Sample Id			Project Name			
	3908																	
	3911	Space In ProjectSampleID										Count: 1						
+	3912	Sample GUID										Bank Specific Sample Id			Project Sample ID			

Figure 1-3: Sample EDQI Report

The user (or designated compliance agent) then conducts an investigation into the identified data elements. Outcomes from the investigation can be:

- 1) The non-compliant data element is correct and no change should be made.
- 2) The non-compliant data element is incorrect and the data is corrected.
  - a) One common example of this is a data field has a null value and per the requirements this field should always possess a value, thus it is non-compliant.

Once the data owner's investigation is complete, a Verification Complete checkbox field is selected within the BTM application at the appropriate record level (patient, sample or subject) to indicate that the non-compliant data has been addressed. Activation of the indicator checkbox will trigger the EDQI system to skip this record to avoid causing the data owner to investigate the same data element multiple times. The only exception to this rule occurs when data in the record (patient, sample or subject) is modified, once this happens, the value of the Verification Complete checkbox is reset to null. This is a protective mechanism that ensures that the EDQI system check always post cedes data entry. The Verification Complete checkbox was a data element that the vendor, DSI, added to the BTM software specifically for this study.

## **1.5 Hypothesis and Goals**

The hypothesis for this project is: if an EDQI system is developed and implemented for CCHMC's biobanking data system, then the percentage of non-compliant data will decrease.

The Primary end point of the hypothesis is:

- 1) Identification of non-compliant data.

Secondary end points of the hypothesis are:

- 1) Conversion of non-compliant data requirements into SQL statements.
- 2) Automated execution of non-compliant detection queries against the production enterprise database.
- 3) Development of an SSRS report displaying non-compliant data categorized per requirement.
- 4) Implementation of an SSRS email subscription to distribute non-compliant results.

## **CHAPTER 2 LITERATURE REVIEW and RESEARCH GAP**

There has been significant work done regarding data entry error detection at the application level, both programmatic and manual. Examples of each include:

- 1) Programmatic – implementation of a field mask on a date field to limit acceptable data entries to values equal to or less than current date.
- 2) Manual – implementation of a data save restriction that prevents a transaction from being pushed to the database until a second user logs into the system and verifies the pending transaction.

For this study, the following keywords were used for a literature review search: Data Entry Errors, Software Business Rules, Software Validation and Improve Data Quality. Sources for these keywords were PubMed and Google Scholar. Over 200 papers were reviewed and this study shares the most pertinent ones regarding the targeted research. The following challenges and solutions section provides a summation of the information reviewed.

## **2.1 Challenges**

Accurate data entry is essential for carrying out high-quality research, and increasing data quality is an ongoing concern in medical research. The consequences of making invalid conclusions as a result of incorrect data could have dire consequences for the researcher, the researcher's institution and possibly a medical patient.<sup>14</sup> Data entry errors are very common and originate in several different ways such as:

- Migration (source) errors including duplicate controls during migration, spelling errors and integrity control on certain values<sup>10</sup>
- Incorrect or incomplete data transfer from an EHR or other data source
  - This could be the result of an inaccurately configured Infobutton or other ETL process<sup>8</sup>

- Typographical error by free-form data entry<sup>12, 14</sup>

## 2.2 Solutions

Each challenge listed above can be mitigated by implementation of an appropriate quality improvement mechanism for a particular system. The following is a list of possible solutions to help improve data quality:

- Checklists (programmed into a CPOE/EHR with mandatory task completion acknowledgment) seem to be a key instrument to reduce individual variations between staff members doing the same tasks, reduce human error and ensure quality standards are maintained.<sup>13</sup>
- Alert systems provide valuable checks/balances and early alerts can result in quick identification of erroneous data.<sup>7</sup>
- Although time consuming and costly, double data entry is still more accurate than single.<sup>14</sup> A hybrid form via a programmatic mandatory quality check by a second user prior to transaction execution also achieves this objective.
- When possible, restrict response options to predefined, validated responses<sup>14</sup>
  - Regarding ETL processes, data filters may need to be applied as a pre-transmission screening or pre-table load screen post transmission
    - Some of these filters may duplicate those at the application layer, but are necessary to accommodate a bulk uploading process. Some of the filters will be unique to bulk data processing, such as a data element that is entered once but distributed to many records.

## 2.3 Literature Review Themes

A general theme in the literature review regarding data quality improvement system was the need for institutions to create site-wide data warehouses to support quality audits and longitudinal research. This need was even implied in articles where smaller one-off type quality improvement measures were implemented. The more robust and complex the quality improvement system was, especially those with some type of integration into EHRs such as Epic or Cerner, the greater the need existed for a data warehouse.<sup>32</sup>

Another theme was simply that of ‘designed focus’. This phrase constitutes the necessity of a certain level of customization. Quality improvement systems that include repeated assessments, feedback, and training appear to improve data quality in a range of practices. Per the literature review, there is simply not a ‘plug and play’ quality improvement system that can be purchased off the shelf, installed and be effective. The nature of the composition of a quality improvement system requires both initial and continual analyses of the systems components and the results achieved against a set of pre-defined performance metrics. The individual components of a quality improvement system can be exchanged; however the quality improvement system itself has to continually be analyzed for, and

modified to accommodate, continual improvement. The implications of this are cost and resources, this is particularly true regarding the initial design of the quality improvement system.<sup>33</sup>

In the literature review, the two most frequently cited data quality attributes were ‘accuracy’ and ‘completeness’. In a review of quality improvement practices to improve the data quality of medical registries, a direct comparison is made regarding data collected automatically (computer system to computer system) versus data collected manually. Automatic collection had a higher accuracy rate versus manually collected data. However, manually collected data was slightly more complete than automatically collected data. The higher level of completeness is a result of situations requiring human discretion and adaptation. It is challenging to program for every contingency and automatic data entry requires specific instructions for all possible scenarios, however the majority of scenarios left unaccounted during the requirements gathering phase are typically discovered in the initial phase after implementation. One mitigation strategy for this is to conduct parallel systems for a short time after initial implementation. Some procedures designed to minimize inaccurate and incomplete data include the following:

- 1) Adequate training of data collection personnel

- 2) Proper design of a data collection protocol
- 3) Monitoring data to detect data errors
- 4) Correction of discovered data errors
- 5) Root cause analysis of discovered data errors
- 6) Double data entry and subsequent comparison

While the literature review revealed that definitions of data quality and data quality attributes are often non-specific, there was a reoccurring emphasis that for a quality improvement system to be effective, it is necessary to determine what attributes constitute data quality before implementation.<sup>34</sup>

Per the literature review, terms describing ‘data quality’ included: accuracy, accessibility, comprehensiveness, consistency, currency, definition, granularity, precision, relevancy and timeliness. Poor data quality was associated with unreliability, increased work time, increased errors and increased liability. While computer based documentation has improved data quality over paper-based documentation, there still seems to be viability in coding due to variations in training and experience. Aside from training and experience variations, there is also some challenge with eliminating paper signatures from clinical practices. Anytime a traditional signature is required, an opportunity for error is introduced because an electronic process has to migrate back to a paper process and often subsequently migrate back to an electronic process. Each migration introduces risk to data quality.

Traditional signatures also introduce another quality risk which is a delay in time between data transfer from the data source to the next data process.

Time delays introduce risk because the data system now has to account for the gap in time, other related data elements that may have changed during that gap in time and additional related data elements that may have been added during that gap in time.<sup>35</sup>

A general observation from the literature review is the potential for EHR use in medical research is tremendous. Unfortunately, the data quality of many EHR systems tempers that potential a bit. There is a lack of consistency in EHR data quality assessment. This places the burden on researchers, or their respective organizations, using EHR data to develop: systemic, empirically driven, statistically based methods for data quality assessment. Some strategies used for EHR data assessment include the following:

1. Data element agreement – comparing multiple elements within an EHR
2. Data element presence – seeking data elements expected to be in the EHR
3. Data source agreement – comparing data from an EHR with another source
4. General validation – comparing EHR data elements with logical data values

The last strategy listed, general validation, deals with plausibility. Part of this determination may include looking for data elements with values outside a generally accepted biological range or comparing data elements against a pre-determined acceptable range per the subject matter experts of a particular process.

Another tool for data quality assessment and data quality improvement is the use of a 'gold standard'. This is generally defined as a dataset drawn from another source. This could include concurrently kept paper records, a separate electronic dataset, information supplied by patients, a data quality check performed by patients or separate data sets which contributed to the complete EHR file. As the name, 'gold standard', implies; this is used as a trusted source of truth for quality improvement systems to reference as appropriate.<sup>36</sup>

Part of the literature review included a data quality assessment of a multicenter registries study. This included a coordinated effort between scientists and data managers to develop a data governance infrastructure consisting of both organizational and technical solutions. An early quality assessment of the data in the central database used in this study revealed

numerous data quality problems including: inconsistencies, missing values, and errors in baseline epidemiology. Reasons for the poor data quality included: multiple data sources, by-passed input rules, and distributed heterogeneous systems. A steering committee was established to direct the organizational needs of the system and a technical working group was established to translate those needs into technical requirements for development of a quality improvement system. The technical working group was charged with ensuring data accuracy and completeness. A validation system was established for error detection and correction for data centers submitting data to a central informatics database. The validation system was developed on allowable values and a crosscheck of related database elements for logical and scientific consistency. The validation system was a two-tiered system consisting of a vendor-developed quality control tool which runs as a Java WebStart application to download the current release of the XML data dictionaries and perform all of the validation checks on the registries hardware prior to data submission. The quality control tool outputs a list of the validation errors and warnings to assist the data submitters in correcting their data submission file to match the data dictionary. This enables sites to prepare and fix many data problems weeks before submitting the file to the informatics support center. This

‘pre-validation’ is valuable in mitigating risks prior to the data being transferred or migrated to another data system. The absence of an enterprise level quality improvement system for each member institution contributed to the need for the pre-validation.<sup>37</sup>

In the literature review, there was a recurring theme that any time EHR data is used for a purpose other than providing patient care, there is a necessity to have a data validation system in place in order to use that data for surveillance or research purposes. The Canadian Primary Care Sentinel Surveillance Network developed a validation system composed of both manual and automated processes. The validation system initially discovered the following data quality issues:

1. Missing data
2. Variation in terminology
3. Misclassification of coding
4. Significant variation between diseases

The missing data issue was primarily missing diagnoses. Some diagnosis had ICD-9 codes listed but no other diagnosis documentation in the medical chart. When this was the case, algorithms in the automated processes found the diagnoses but the diagnoses were missed in the manual chart reviews.

The opposite also occurred as some diagnoses were recorded in free text only. Those were found in the manual chart review but were missed by the algorithms. A separate issue with this network, but a common theme per the literature review was many data discrepancies discovered via the validation process had occurred during data extraction or data migration processes. Also, similar to the registries study, the absence of an enterprise level quality improvement system for each member institution contributed to the need for a pre-validation system.<sup>38</sup>

### **2.3.1 Summary of Themes**

The culmination of literature reviewed emphasized the need for quality improvement systems to have a level of customization to be truly effective. Several articles from the literature review presented a trade-off in the balance between quality improvement mechanisms that can be automated with technology and those that require subjective ‘human’ intervention. The literature also emphasized that while certain mechanisms may be exchanged between systems and processes, to truly be effective every ‘system’ seems to require a certain level of customization. Cumulatively, the literature indicated a direct correlation between the level of customization of a quality

improvement system and the corresponding level of effectiveness derived from that system.

## **2.4 Research Gaps**

The major research gaps identified were:

- 1) Data quality improvement of the entire enterprise
- 2) Data quality improvement specific to biobanking data
- 3) Using combinations of SSIS, SSRS along with web building tools to develop a customized quality improvement system

## **CHAPTER 3 METHODOLOGY**

### **3.1 Algorithm and System Design**

Figure 3-1 provides a visual display of development algorithm process. This is an overly simplistic view of the process; however, it is worth noting that an agile approach was able to be utilized because an exact replication of the BTM database was readily available to test solutions without placing production data at any risk. This replication database could be, and often was, updated in real time to match the production environment. This proved valuable in the development of the EDQI system with regard to efficiency in

development. A second valuable resource was an ample supply of end-user participation in both providing development requirements and testing solutions. The availability of these resources permitted a development process that was both iterative and incremental. Figure 3-1 below provides a high-level overview of the EDQI development algorithm.

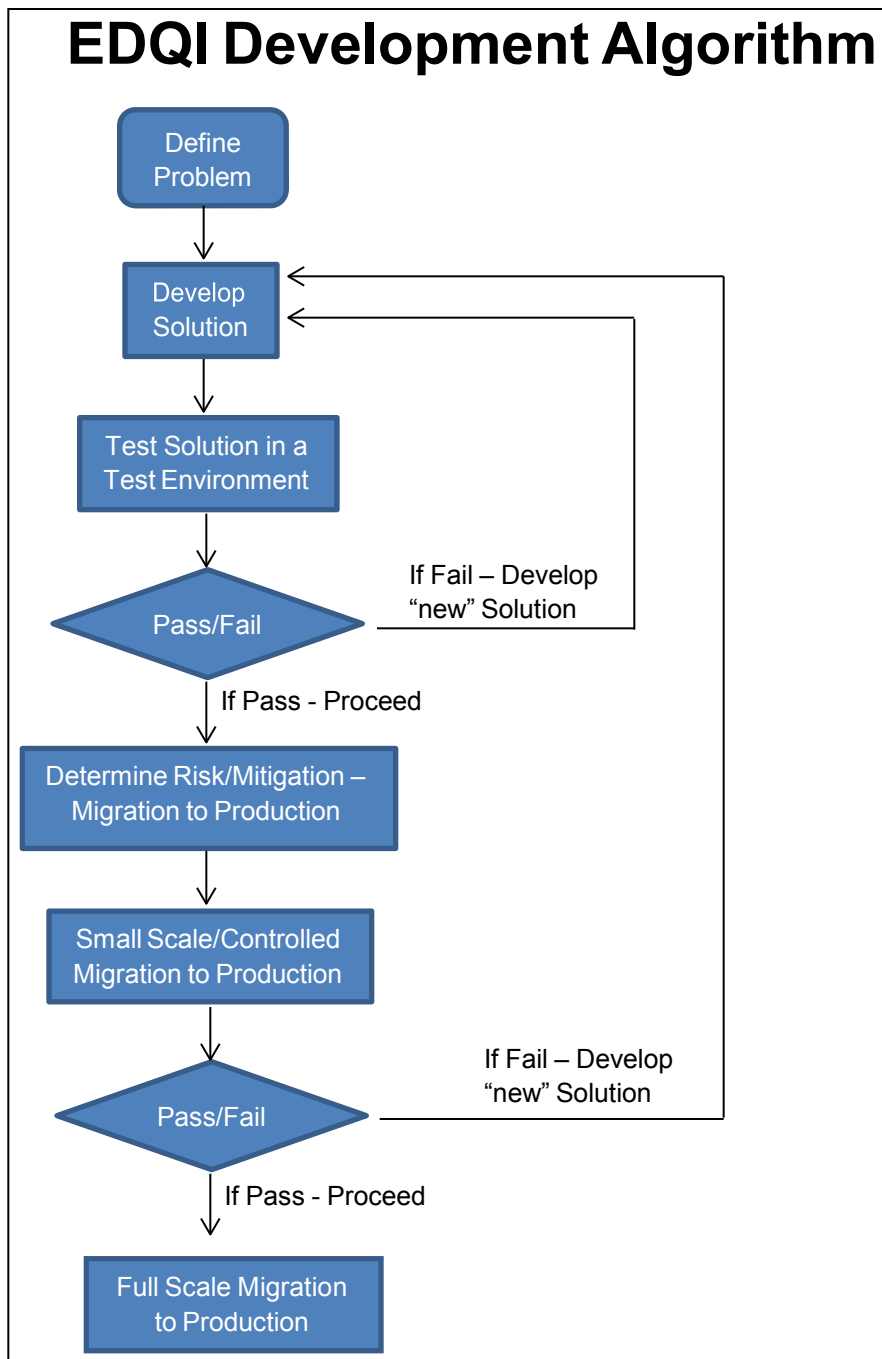


Figure 3-1: EDQI Algorithm Design

### **3.2 EDQI Work Flow Algorithm**

A visual workflow diagram of the EDQI algorithm is depicted in Figure 3-2 below. The first two process symbols represent enterprise and bank level queries which are very unique in objective, however they are combined and executed in a manner that provides a uniform output to the end user.

Enterprise level queries were developed with criteria applicable to all CCHMC biobanks in BTM. One example of an enterprise level query is a check for patients having multiple MRNs (Medical Record Numbers) within the same medical facility. Another example of an enterprise level query is a check for samples having multiple GUIDs (Global Unique Identifiers).

Neither of the aforementioned examples should occur for any bank within the BTM biobank software, so these enterprise level queries are developed and run agnostic of any bank level considerations.

Bank level queries are specific to each bank in the BTM software. As part of this study, an additional page in the BTM Help website was developed, see Figure 3-41, to capture the initial requirements needed to write the bank specific queries. This process was applied retroactively to pre-existing banks and the process has been implemented as part of a larger set of processes that are applied to all new banks.

The following are two examples of bank level queries:

- 7) If a biosample type is blood, the additive should be EDTA (Ethylenediaminetetraacetic acid) and the tube type should be a 'pink top tube'.
- 8) If a biosample type is EBV (Epstein–Barr virus) cells, the additive should be acid-citrate-dextrose and the tube type should be a cryovial – 'yellow tube'.

While the specifics elements of bank level queries differ, the basic structure is similar which helps to streamline the EDQI implementation process for each new bank.

The following is an example of a bank level requirement verifying appropriate biofluid volume amounts:

For total volume, capture any of the following conditions:

- 1) For all samples if a value is not selected
- 2) For blood tubes if a value is less than 3 mls.
- 3) For plasma/serum if a value is over 1.2 mls.
- 4) For plasma/serum if a value is under .2 mls.

And here is the query produced based on the previous requirement:

```
SELECT

    Sample.GUID

    , BankSample.BankSpecificSampleId

    , CASE

        WHEN SampleType.Name IS NULL THEN 'No Sample Type'

            WHEN SampleType.Name = 'blood' AND
SampleDescription.Volume < 3 THEN 'Blood Less Than 3mls'

            WHEN SampleType.Name IN ('plasma', 'Serum') AND
(SampleDescription.Volume < .2 OR
SampleDescription.Volume > 1.2) THEN 'Plasma/Serum
Outside .2 - 1.2 Range'

            END AS VolumeIssues

    , SampleDescription.Volume

    , SampleType.Name AS SampleTypeName

FROM Sample

INNER JOIN BankSample

    ON Sample.GUID = BankSample.SampleID

INNER JOIN Bank

    ON BankSample.BankID = Bank.GUID
```

*INNER JOIN SampleDescription*

*ON Sample.GUID = SampleDescription.SampleID*

*INNER JOIN SampleType*

*ON SampleDescription.SampleTypeID =  
SampleType.GUID*

*WHERE*

*Bank.NAME = 'CCHMC PAH Biobank' AND*

*CASE*

*WHEN SampleType.Name IS NULL THEN 'No Sample  
Type'*

*WHEN SampleType.Name = 'blood' AND  
SampleDescription.Volume < 3 THEN 'Blood Less Than 3mls'*

*WHEN SampleType.Name IN ('plasma', 'Serum') AND  
(SampleDescription.Volume < .2 OR  
SampleDescription.Volume > 1.2) THEN 'Plasma/Serum  
Outside .2 - 1.2 Range'*

*END IS NOT NULL*

*ORDER BY VolumeIssues, SampleType.Name,  
SampleDescription.Volume*

The following is an example of an enterprise level requirement identifying patients with multiple MRNs:

1. Capture patients with multiple patient records and MRNs for the same medical facility.

And here is the query produced based on the previous requirement:

```
USE BTMResearch

SELECT pp.FirstName, pp.LastName, pp.DateOfBirth,
pmrn.MedicalFacilityId, COUNT28

FROM PatientMrn pmrn

INNER JOIN PatientProfile pp ON pmrn.PatientId = pp.PatientGuid

WHERE MedicalFacilityId = '29b59d5b-05fc-4297-9a56-
6e201d1fd727'

GROUP BY pp.FirstName, pp.LastName, pp.DateOfBirth,
pmrn.MedicalFacilityId

HAVING COUNT28 > 1

ORDER BY pp.LastName, pp.FirstName
```

Both enterprise level and bank level queries are bundled and distributed as one set of results at the frequency desired by the bank owners. The EDQI algorithm work flow below provides a high level display of the sequence of

steps from query design to acknowledgment of investigation by the end user within the application.

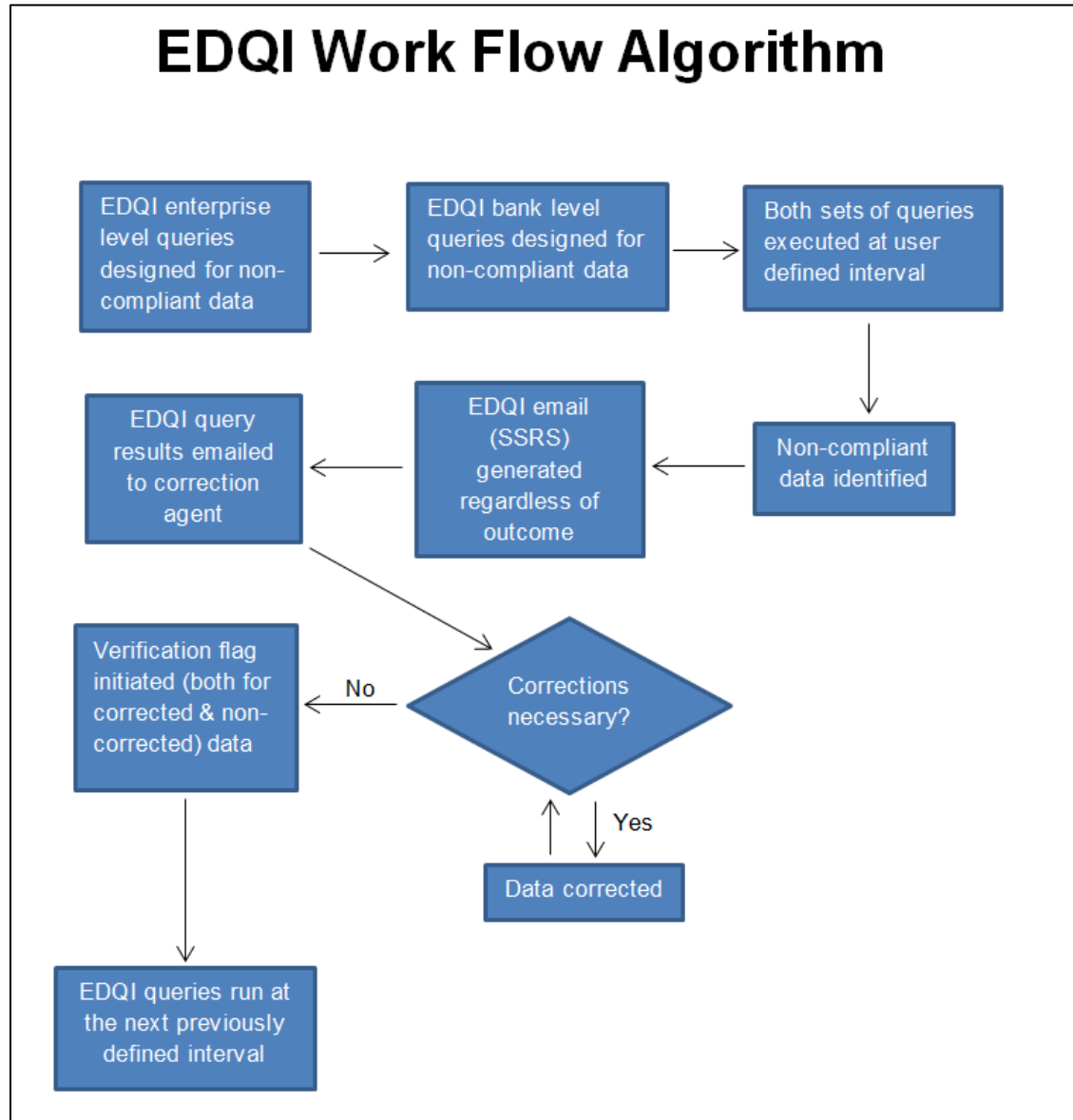


Figure 3-2: EDQI Work Flow Diagram

### **3.3 CCHMC Biobanking Data System and EDQI Components**

The following sections provide descriptions of each of the biobanking data systems and EDQI components. Figures 3-3 and 3-4 provide a high level illustration of these components. The simple numerical scheme used in Figure 3-4 will be referenced throughout the remainder of this report. The initial section of each EDQI component is a non-technical “Component Overview” which provides some background information about that particular EDQI component. After the “Component Overview”, is a “Component Description” section which provides technical details of each EDQI component.

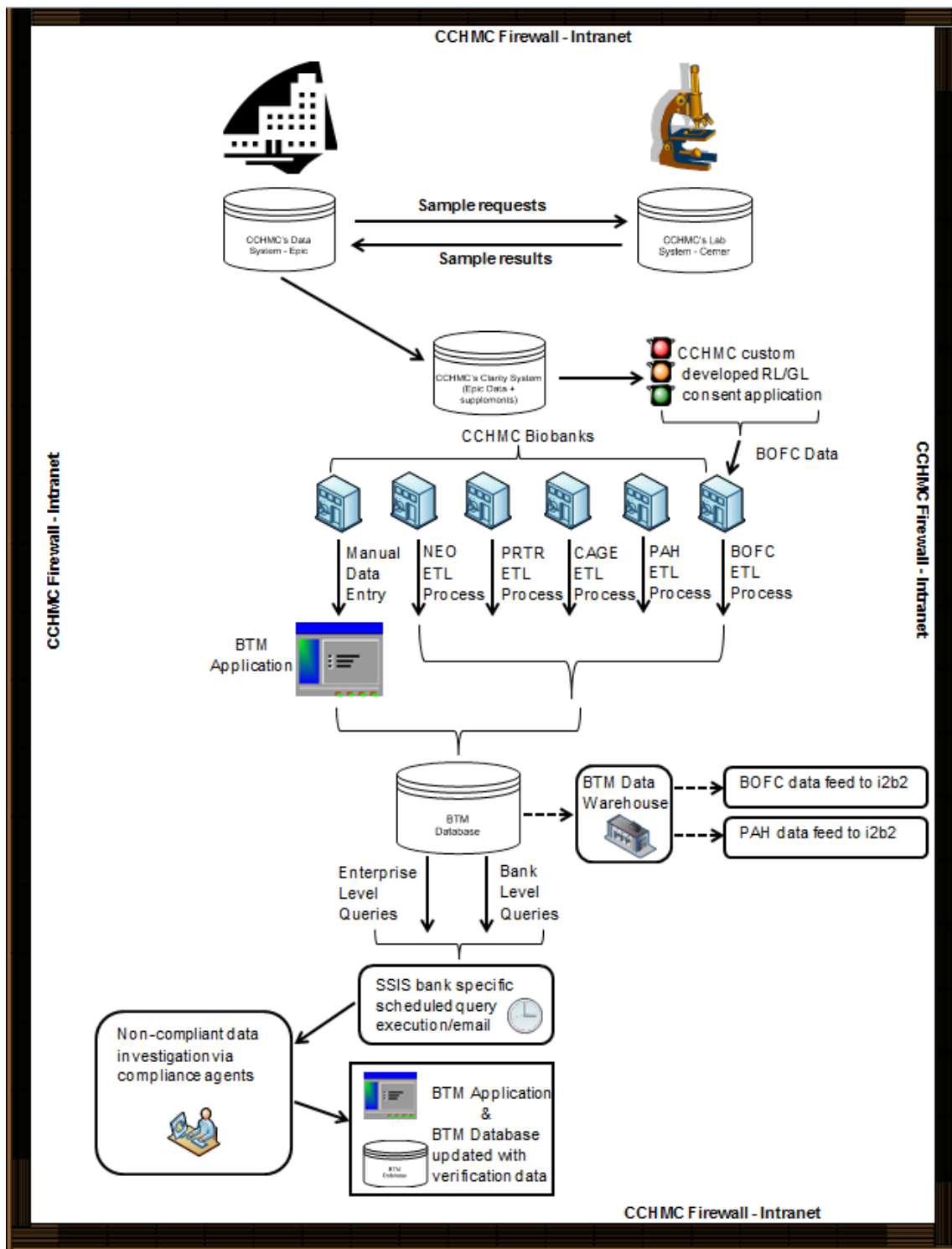


Figure 3-3: Component level overview of EDQI system.

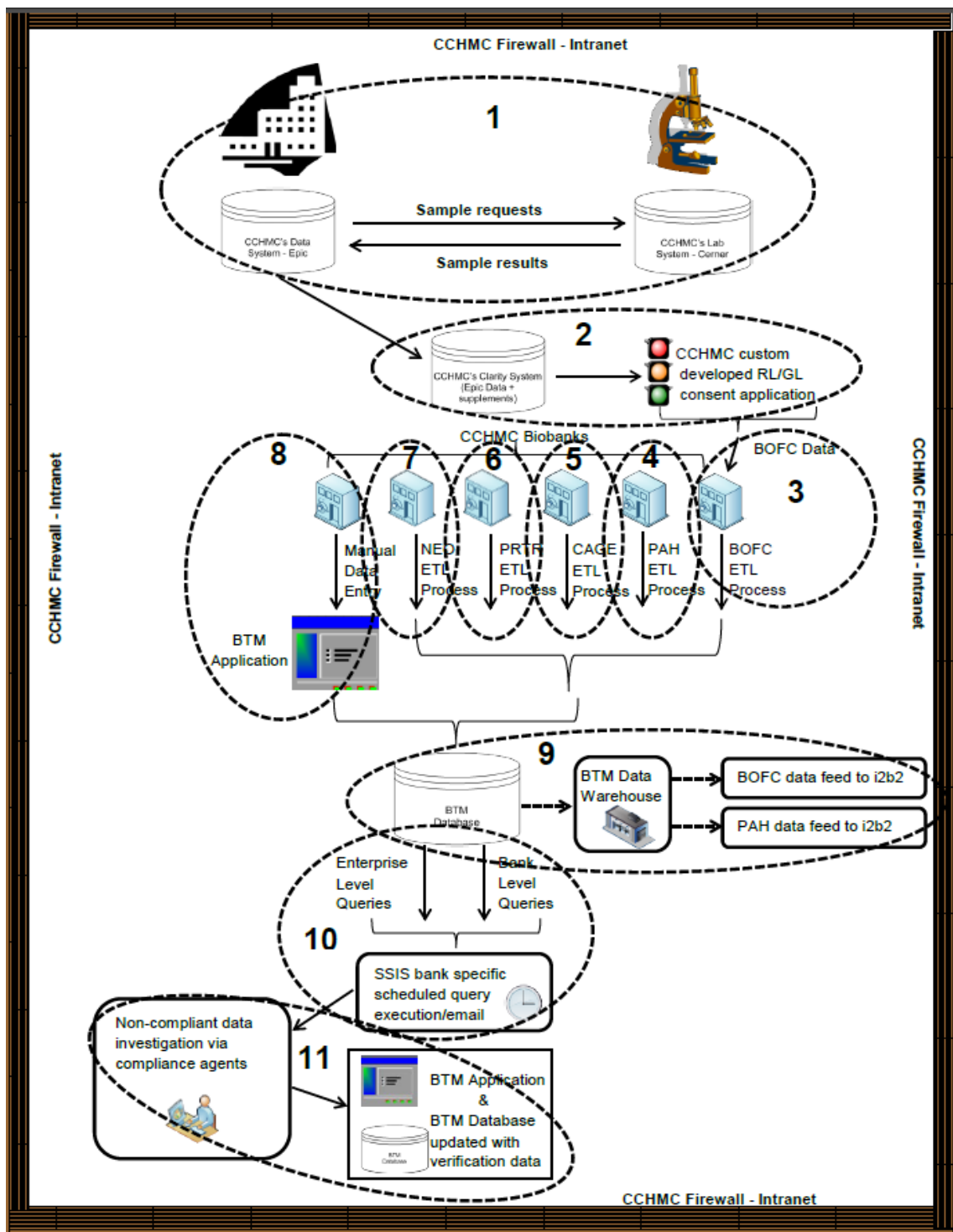


Figure 3-4: Component level overview of the EDQI system numbered by functional relationships.

### **3.3.1 Component #1, Epic and Cerner Systems**

A significant portion of the data residing in BTM starts in CCHMC's EHR system, Epic. Smaller portions come from other medical institutions and direct data entry into BTM. CCHMC's laboratory operations data resides in a separate system called Cerner. The Cerner system is mainly used to process laboratory tests. The requests for these tests and the results of these tests are both stored in Epic.<sup>24</sup>

Figure 3-5 below illustrates some of the physical processes and the accompanying data flows. The process begins with a patient visit and registration data entered into Epic. It is during the registration process that the patient consent process also takes place. This process will be covered fully as part of Component #2, but to understand the complete process it is important to know that consent is obtained at the time of registration and the consent decision is stored in Epic. Lab tests are also defined and ordered via Epic; however, the lab test request is passed via an HL7 interface to Cerner. The Cerner system assigns each sample a unique Accession identification number. The Cerner system stores the tests performed along with test results and passes this information back to Epic via the HL7 interface referenced earlier. The test results passed from Cerner to Epic may be composed of

numeric data or BLOB text. Any further data entry or data management takes place in Epic.<sup>24</sup>

On a daily basis, Epic's underlying Chronicles database is copied into Clarity, an Oracle database. It is from the Clarity database that the RL/GL (subject of the Component #2 section of Figure 3-4) database is populated. It is from the RL/GL database that data is pulled into the BTM database (subject of the Component #3 section of Figure 3-4). All of these transfers are represented in the high level process and data flow diagrams in the figure below.<sup>24</sup>

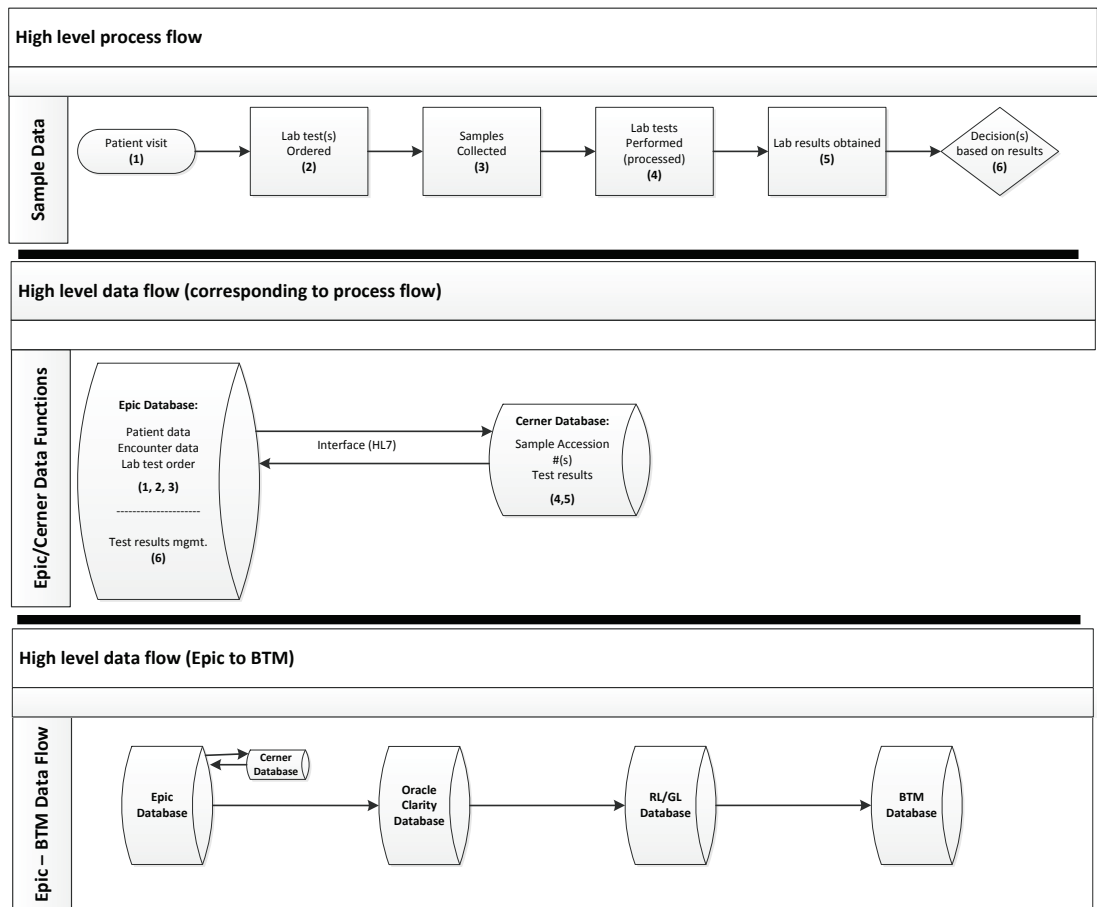


Figure 3-5: Process and High-Level Data Flow, Inception to Completion

### 3.3.2 Component #2, Clarity and Red Light/Green Light (RL/GL)

This component comprises the data process depicted in the last three database figures which are illustrated in the last portion of Figure 3-5 above:

the Oracle Clarity database, the RL/GL database and the BTM database.

The one subcomponent that is not illustrated in Figure 3-5 above is the RL/GL application. This application is obviously connected to the RL/GL database, but from a data flow perspective, application utilization takes place

between the RL/GL database and the BTM database as illustrated in Figure 3-6 below.<sup>23</sup>

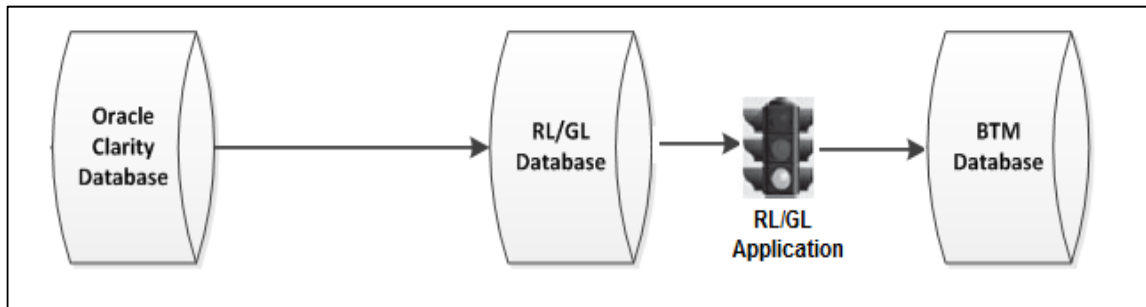


Figure 3-6: Process location of RL/GL application

The RL/GL database is populated via an ETL job that runs daily and obtains a pre-defined data set from Clarity. Basic patient demographic data is obtained in this ETL along with basic sample data such as department (requestor), date, accession number, and sample type. The most important data in the RL/GL database, and the primary purpose of both the RL/GL application and database, is the consent status. From a work flow perspective, the end-user (usually a laboratory worker) uses a bar code scanner to scan the bar code on a biosample into the RL/GL application in order to see the consent status. If the patient, or legally authorized representative, has provided consent, the application displays a green light illuminated on a traffic light display. In this scenario, the user will retain the sample for research purposes and it will go through the proper preparation

methods for storage. On the data side, after a batch of samples is scanned in RL/GL, the data is exported to a scan log spreadsheet. This spreadsheet is first compared against a ‘frequent flyer’ list to determine if samples from that particular individual are already present in BTM (as only a certain quantity of biosamples are kept on any one patient). After that, the scan log spreadsheet is used to populate the BOFC upload spreadsheet (this process is described in the next section).<sup>23</sup>

As referenced above, the most important data in the RL/GL database is patient consent status, specifically, consent status regarding authorization to use residual clinical samples for research.

The patient’s BOFC consent decision is recorded in Epic. It is recorded during the registration process along with the patient’s decisions on documents such as the ‘consent to treat’ and ‘notification of privacy practices.’ By design, the BOFC project is an opt-in option for patients to provide one of the following answers:

1. Refuse consent
2. Consent with notification
  - a. Allows the patient/family providing the sample to receive information on incidental research findings.
3. Consent without notification

- a. This is essentially anonymous; the patient's residual clinical samples are made available in the repository with no mechanism for re-contact.

#### 4. Consent deferred

- a. This option is used in cases where the patient/family needs more time to consider their decision, where an appropriate parent or legal guardian does not accompany the minor patient, or when clinical circumstances are not conducive to obtaining a meaningful consent.
- b. This option will prevent the patient from being asked for consent again for seven days.<sup>18</sup>

Separate from the options above is also an option for the patient to withdraw consent after previously providing consent. If a patient exercises this option; all existing samples in the biorepository will be destroyed. If a patient initially provides consent and later decides to refuse participation, but does not withdraw; any sample collected during the initial period of BOFC consent can still be used for research.<sup>18</sup>

Finally, if a patient is under 18 years of age at the time when the original BOFC consent is obtained, a parent will provide consent and the child 'assent'. After turning 18, the patient will be asked to provide consent at the first visit after their 18th birthday. This BOFC consent does not expire unless the patient specifically withdraws from the study.<sup>18</sup>

All of the conditions and business rules listed above, plus several others, are built into the RL/GL application's logic. When a sample accession number is entered into the application (usually by scanning a bar code label affixed to the sample), the application provides a fast and easily interpreted result. This result provides the instruction for the end user to retain the biosample for long-term storage or to discard the sample.<sup>18</sup>

CCHMC's EDQI system relies on data from RL/GL to provide an on-going consent status comparison. If there is a discrepancy between the consent expiration date in RL/GL and BTM, a non-compliant data element is placed on the BOFC Data Quality Report for investigation. The expiration date field was chosen as the field of comparison because upon notification of a consent withdrawal, the compliance agent changes the consent expiration date in Epic to the date of the withdrawal.

### **3.3.3 Component #3, Better Outcomes for Children Project**

The BOFC (Better Outcomes for Children) bulk upload into BTM is the most complex of all the upload processes listed on the EDQI Components diagram (See Figure 3-4). The BOFC project is one of forty-two projects in the Cincinnati Children's Hospital Medical Center Bank. This is the last BTM Bank in the screen shot of all the BTM Banks in Figure 3-21. It was

for this project that the RL/GL application (see Component #2 of Figure 3-4,) was created. Figure 3-7 below provides a macro-level view of the BOFC upload components.

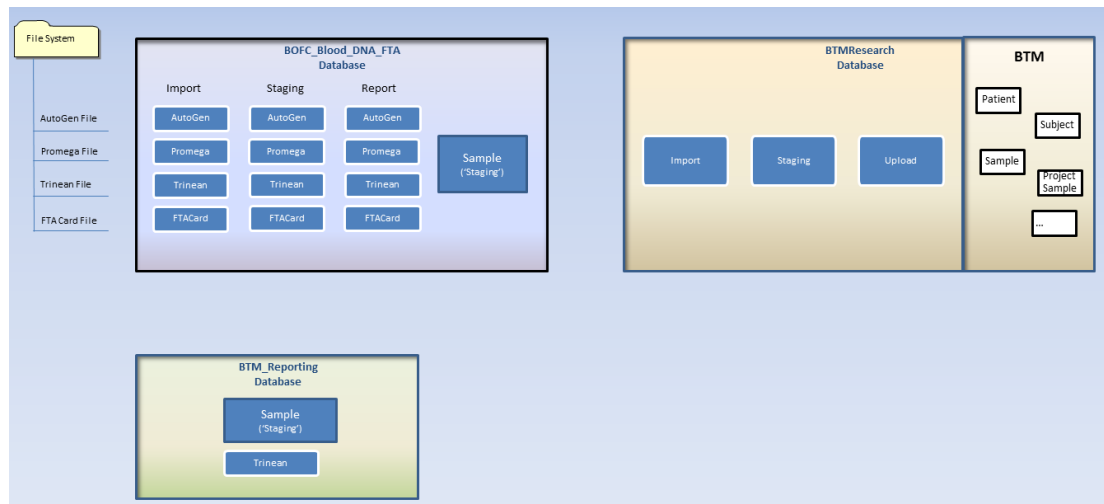


Figure 3-7: BOFC Upload Components<sup>27</sup>

Beginning in the upper left corner of Figure 3-7, only one file (AutoGen, Promega, Trinean or FTA) is processed at a time. The ETL script loads the file name and associated file data into the import table. For this load, there are no data checks, but incorrectly formatted files cause critical errors and the upload fails. When one of these errors occurs, an email, via SSRS, is sent and a file moved to the \_error folder. There is a dedicated transitional table which always remains empty except when processing a file. Data is

copied from import table to the staging table and initial data checks are completed here for things like:

1. Field format (SampleID, DNA1 ID, etc.)
2. Storage location (ensure existence, check for occupancy)
3. Duplicate SampleID

Any failed data checks will reject the file and it will go into the `_error` folder and an SSRS email will be sent with error info. If data passes these checks, the file will go into the `_report` table. The transitional table referenced earlier, only contains last file processed. Data is copied from the staging table to the report table. The report table is a permanent table, data for all samples always resides in this table and it is used as the source for the sample ('staging') table. Data is merged from the report table into the sample ('staging') table. All samples are deleted and re-created every time a file is processed.<sup>27</sup>

Per Figure 3-7 above, as data moves from the `BOFC_Blood_DNA_FTA` database to the BTM Research database, the ETL process selects all samples from the sample ('Staging') table where `InBTM = 0` and other minimal eligibility criteria is met. Samples are copied into the transitional `_genericupload_import` table with 'BOFC DNA FTA Loader' as the `MigrationSet` and `MigrationDate` of current date and time is assigned. At

this point, all fields are checked for correct data types and possible truncation errors. If there is an error, an SSRS email is sent with error information, the upload to BTM is stopped and the import table is cleared. If there are no errors, all records from the Import table are inserted into the Staging table. Key lookup values are updated in the staging table with their associated BTM GUID. This could be a GUID for: Bank, Project, MedicalFacility, SampleCategory, TissueCategory, ParentSampleCategory or ParentTissueCategory. There are also additional updates at this point in the process such as:

1. Patient\_LDAP\_MRN
2. SortOrder - the sample parent/child relationship is established via a combination of data from both the BTM database and the file.
3. ParentSampleGUID - this will be used in later processes by insert procedures.
4. AliquotOrdinal - the parent/child aliquot order is established from a combination of data from both the BTM database and the file.<sup>27</sup>

At this point, all validation checks are performed on the data in the staging table. If there is a validation error; an SSRS email is sent with the error information, the upload to BTM is stopped and the import and staging tables are cleared. If there are no errors in the staging table, all records are copied to the upload table. From the upload table, data is inserted into BTM tables shown on the right side of Figure 3-7 (patient, subject, sample, project

sample). Also at this point in the process, when applicable, the following additions and/or update are made:

1. Clinical Finding Record (Patient)
2. Patient Consents
3. Patient Diseases
4. Storage (Sample)

Again, referencing Figure 3-7, the updates described in the previous section are now pushed from BTM (BTM database listed on the right side of Figure 3-7) back to the Sample ('Staging') table in the BOFC\_Blood\_DNA\_FTA database.<sup>27</sup>

Finally, the Sample ('Staging') and Trinean tables are truncated and data is copied from the BOFC\_Blood\_DNA\_FTA database to the BTM\_Reporting database (shown in the lower left portion of Figure 3-7). There are two triggers that initiate this process:

1. Total sample count differs
2. Count of samples with HasTrinean = 1 differs

The timing and sequence of the SQL Server jobs are illustrated in the figure below.

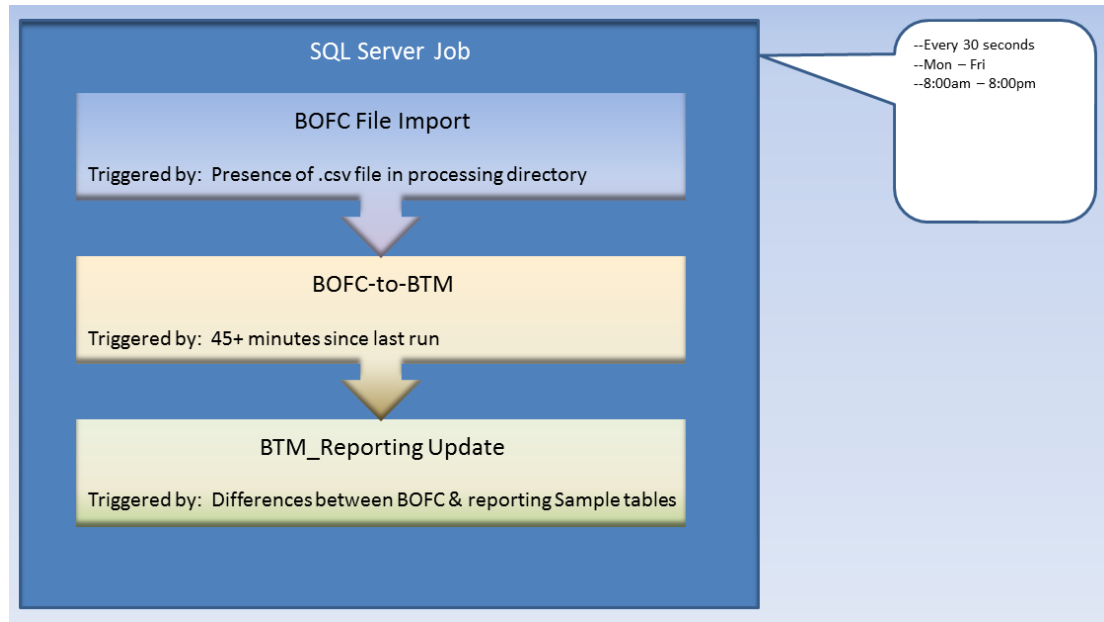


Figure 3-8: Timing and sequence of BOFC SQL Server jobs<sup>27</sup>

The BOFC folder structure below illustrates a list of files in the history & error folders. The filenames with 'Main' and 'Reading' in them correspond to different branches of the Cincinnati Children's Hospital Medical Center biobank. 'Champ' corresponds to a project within the Cincinnati Children's Hospital Medical Center biobank.<sup>27</sup>

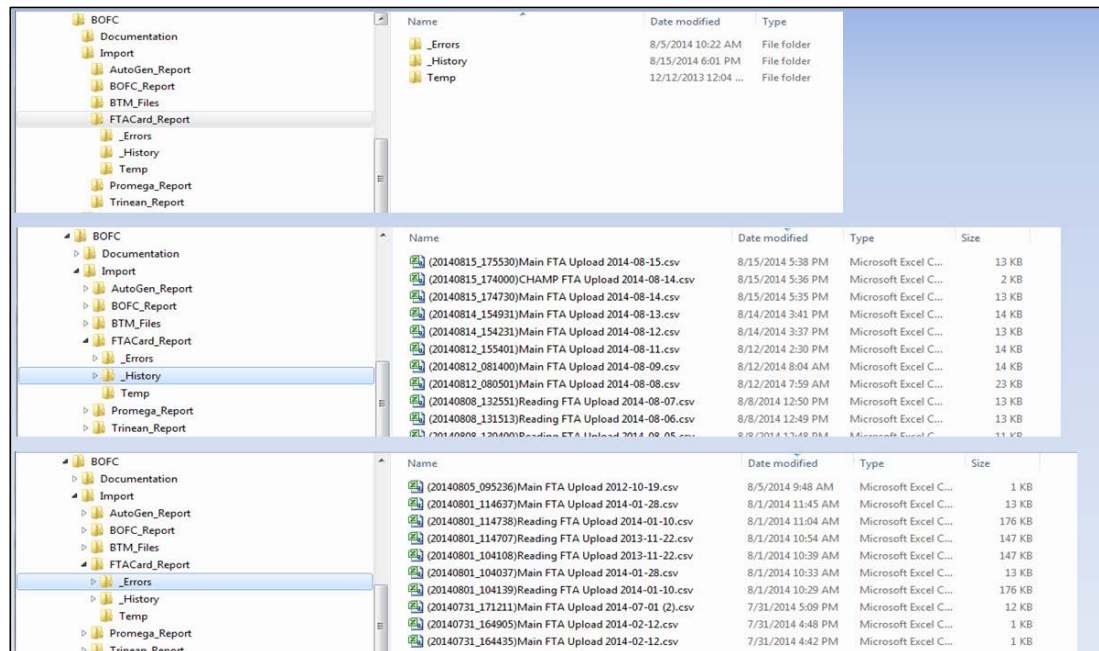


Figure 3-9: BOFC Folder Structure

### 3.3.4 Components #4 - 7; PAH, CAGE, PRTR, Neonatology

The components for: PAH, CAGE, PRTR and Neonatology are grouped together due to their similar composition and function. Unlike the BOFC ETL, which represents one project in a large bank, each of these ETL processes represents an entire bank.<sup>26</sup>

From the end-user perspective, the processes are identical. The end-user places a file in folder on shared drive. See the figure below.

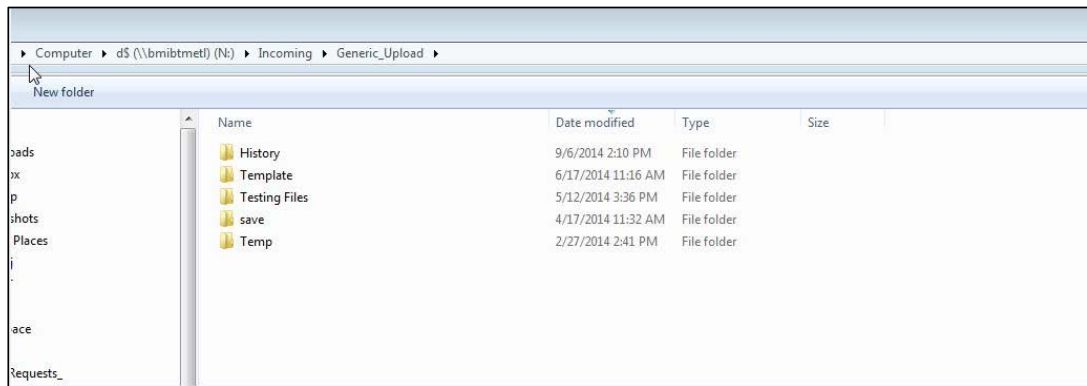


Figure 3-10: Generic Upload Shared Drive Folder Structure<sup>26</sup>

A text connector looks to see if the shared folder is occupied. This check is performed every ten minutes. Files that end in “txt” are read out of the shared folder. When a file is present, it is moved to a cache area of memory set up to store the data from the file. Work tables in the database are cleared and truncated to prepare for data entry. Each row in the spreadsheet gets a unique ID and each row has a migration set name which dictates which migration set the generic uploader utilizes. There are some coding variations per migration set based on the business rules of the bank associated with the migration set. For each run, there is a migration date and time which goes down to the millisecond. Finally, data is extracted out of stage table one row at a time via the Foreach Loop as seen in the third box of the figure below.<sup>26</sup>

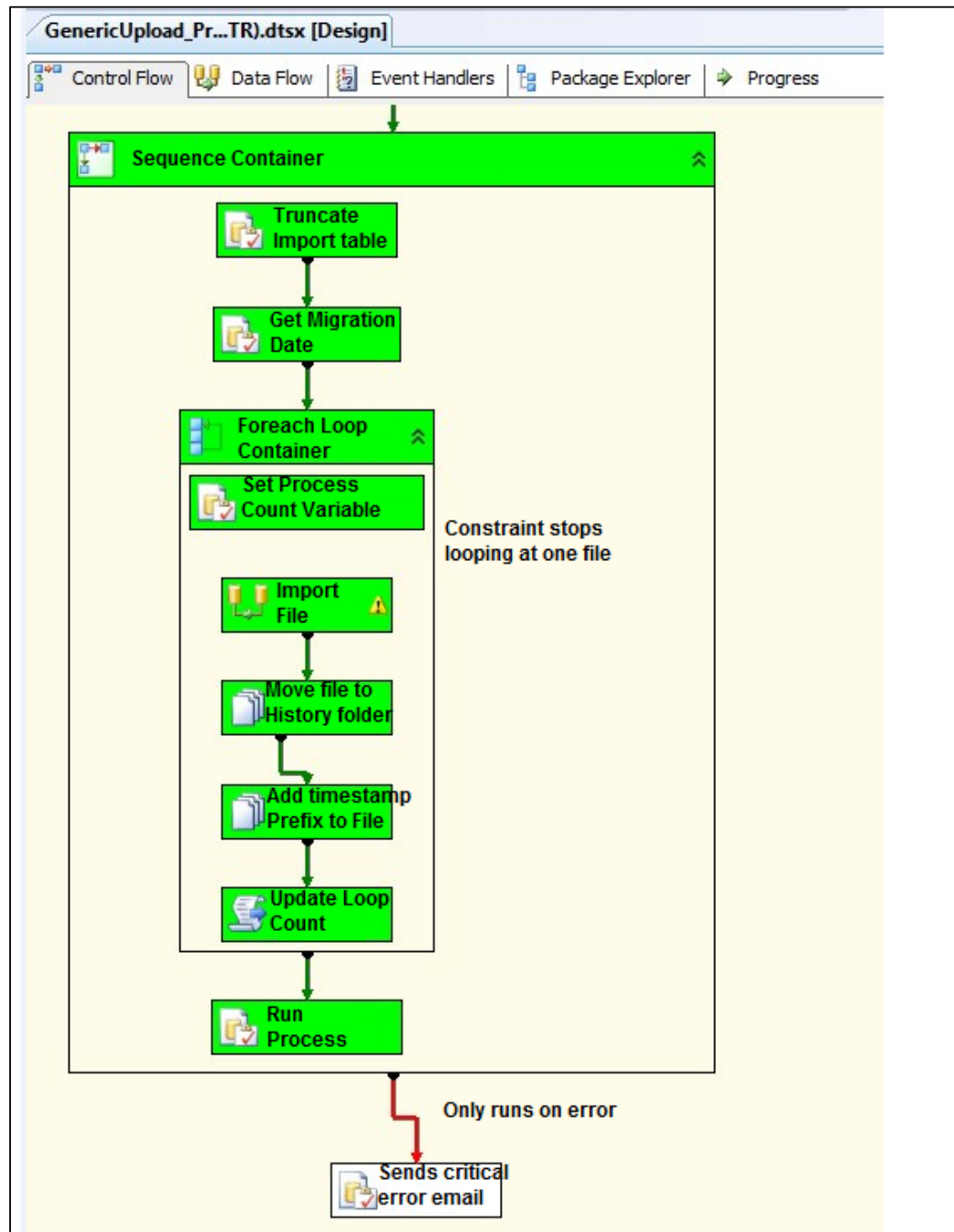


Figure 3-11: Generic upload processing sequence.<sup>26</sup>

Next the process checks for errors or conditions that will prevent a successful upload. One such example is a sample is assigned a storage location that is already in use.

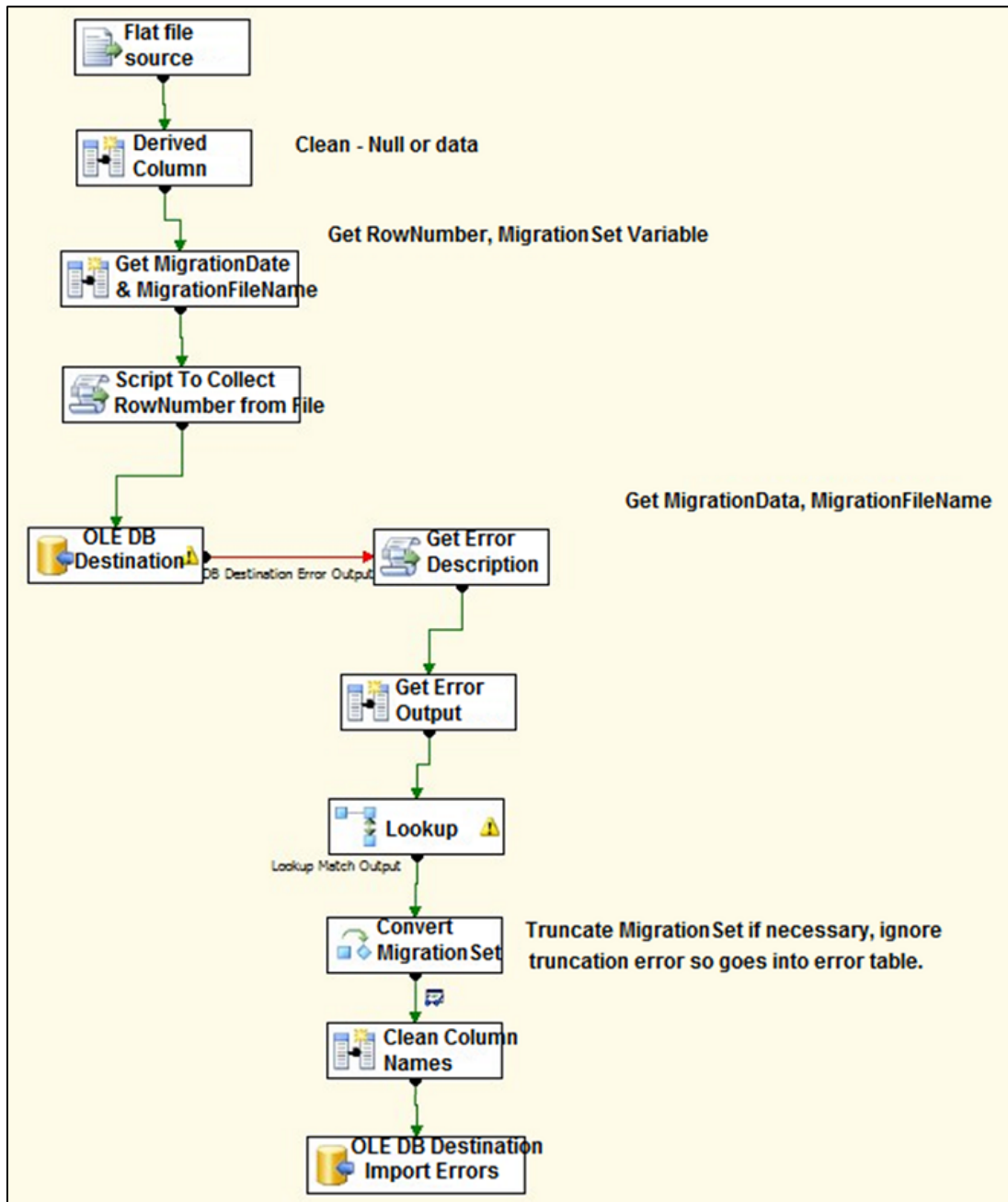


Figure 3-12: Generic Upload Error Detection Process<sup>26</sup>

The data is initially read into the genericimportpackage table. Validation checks are made for valid sample type, sample category, bank sample identification number (must be unique), project sample identification number (must be unique), and the storage assignment must be available. Once all validation tests have passed, the data is imported into the staging table.<sup>26</sup>

After the data is imported into the staging table, the folder is moved to the history file with a date and time stamp placed in front of the file name, see the figure below.

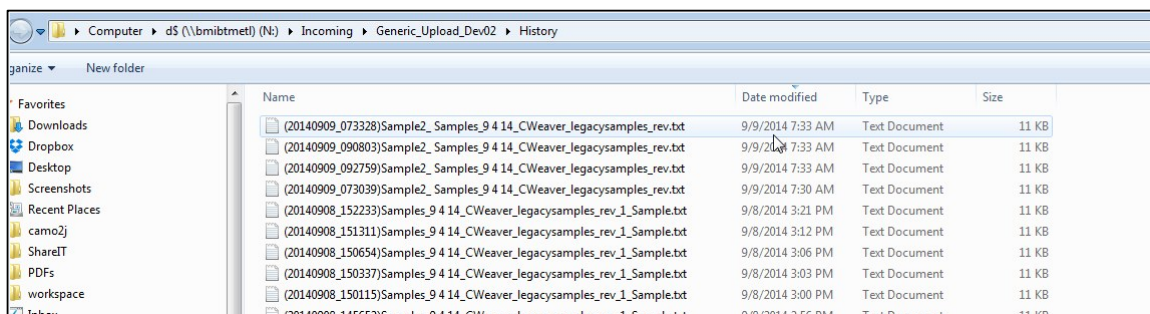


Figure 3-13: Generic Upload Post Import Data Placement<sup>26</sup>

It is at this point that the stored procedure, ProcessMigrationSet, is executed, see the figure below.

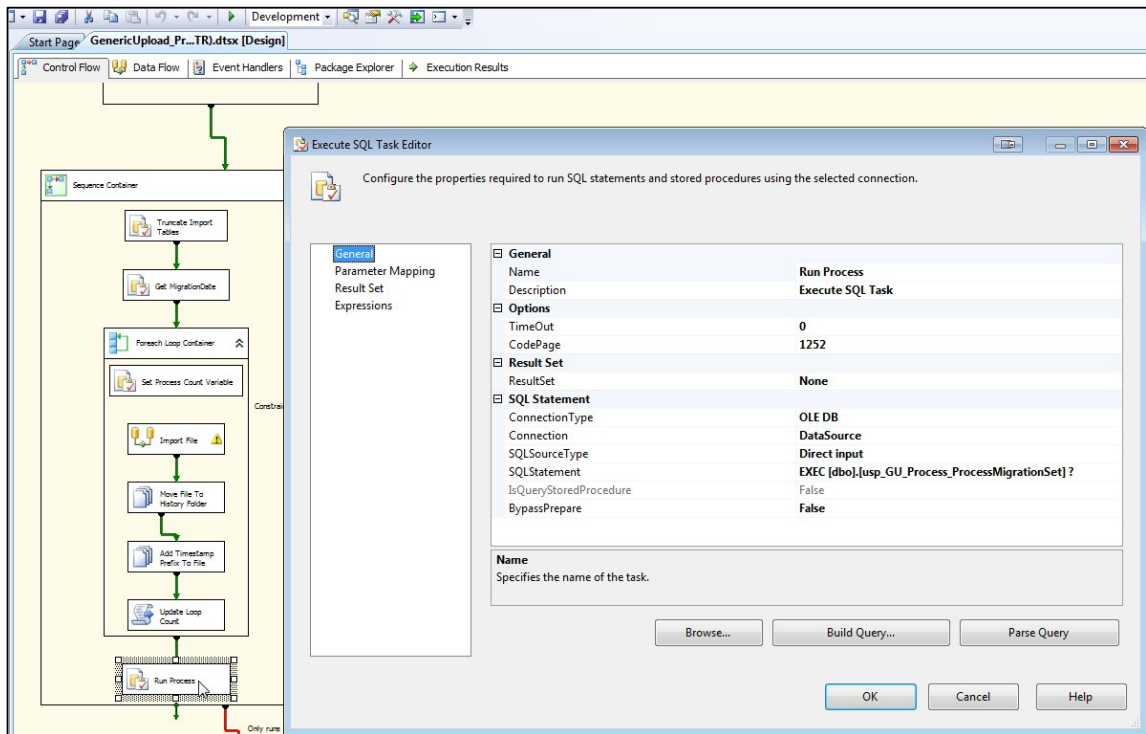


Figure 3-14: ProcessMigrationSet Properties<sup>26</sup>

Next the process checks to ensure that the generic uploader is not active.

There is a job status flag that is set to “on” when this procedure starts to run.

Then the records are inserted into the BTM database, see the figure below.

```

-- =====
-- Insert Records Into BTM
-- =====

IF @IsSuccessful = 0
BEGIN
    PRINT '-----'
    PRINT ' Upload Subject/Sample Records '
    PRINT '-----'

    -- Upload Subject/Sample Records
    -- =====
    SET @Error_Comment = 'Insert Records Into BTM'
    SET @StartDate = GETDATE()
    EXEC @IsSuccessful = [usp_GU_Process_UploadSubjectSample]
        @par_MigrationSet = @par_MigrationSet,
        @par_MigrationDate = @par_MigrationDate
    PRINT 'Total Time: ' + CAST(DATEDIFF(s,@StartDate,GETDATE()) AS VARCHAR(5))
    IF @IsSuccessful <> 0 GOTO HANDLE_ERROR
END

```

Figure 3-15: BTM Insertion Code<sup>26</sup>

If there were no errors, the data is uploaded into BTM and the process is completed. If errors were detected with any pre-insert validation, then nothing goes beyond the staging table to the BTM database and the error is written as illustrated in the figure below.

```

SQLQuery25.sql | SQLQuery24.sql | SQLQuery22.sql | SQLQuery20.sql | SQLQuery19.sql - not connected | SQLQuery18.sql - not connected | SQLQuery11.sql* | Upload_Assignment...
472      PM.GUID AS [ProcessingMethodEnumID],
473      SU.Sample_ProcessingMethod AS [ProcessingMethodValue],
474      SU.Sample_Diagnostic AS [Sample_Diagnostic],
475      SU.Sample_ProcuredNotBanked AS [ProcuredNotBanked]
476  FROM _GenericUpload_Upload AS SU
477      INNER JOIN @tbl_SampleCategory AS SC ON SC.ID = SU.ImportID
478      INNER JOIN @tbl_SampleSource AS SS ON SS.ID = SU.ImportID
479      INNER JOIN @tbl_SampleStatus AS SST ON SST.ID = SU.ImportID
480      INNER JOIN @tbl_CollectionSite AS CS ON CS.ID = SU.ImportID
481      INNER JOIN @tbl_ProcedureLaterality AS PC ON PC.ID = SU.ImportID
482      INNER JOIN @tbl_SampleLaterality AS SL ON SL.ID = SU.ImportID
483      INNER JOIN @tbl_CollectionProcedure AS CP ON CP.ID = SU.ImportID
484      INNER JOIN @tbl_HealthStatus AS HS ON HS.ID = SU.ImportID
485      INNER JOIN @tbl_HemolysisScore AS HSC ON HSC.ID = SU.ImportID
486      INNER JOIN @tbl_Characteristics AS C ON C.ID = SU.ImportID
487      INNER JOIN @tbl_SubCharacteristics AS SCA ON SCA.ID = SU.ImportID
488      INNER JOIN @tbl_ProcessingMethods AS PM ON PM.ID = SU.ImportID
489  WHERE (Global_MigrationSet = @par_MigrationSet AND Global_MigrationDate = @par_MigrationDate)
490
491  END TRY
492  BEGIN CATCH
493      GOTO HANDLE_ERROR
494  END CATCH
495
496  --Exit and error handling
497  PROC_EXIT:
498      RETURN @IsSuccessful
499
500  HANDLE_ERROR:
501
502      -- Log error if this procedure errored
503      SET @IsSuccessful = -1
504
505      SELECT
506          @Error_ServerName = @@SERVERNAME,
507          @Error_DatabaseName = DB_NAME(),
508          @Error_Procedure = ERROR_PROCEDURE(),
509          @Error_Number = ERROR_NUMBER(),
510          @Error_Line = ERROR_LINE(),
511          @Error_Message = ERROR_MESSAGE(),
512          @Error_State = ERROR_STATE(),
513          @Error_Severity = ERROR_SEVERITY(),
514          @Error_NestedLevel = @@NESTLEVEL,
515          @Error_User = SUSER_SNAME(),
516          @Error_Host = Host_NAME()
517
518      -- Return failure
519      SET @IsSuccessful = -1
520      GOTO PROC_EXIT
521
522  END

```

Figure 3-16: Migration Set Error Handling Code<sup>26</sup>

Any errors detected will prevent the upload set from completing and will generate an email similar to the one illustrated in the figure below.



spreadsheet corrected, the user will proceed to restart the process by placing the spreadsheet in the shared folder displayed in the figure above.<sup>26</sup>

This BTM ETL process has been written in SQL and in Perl. Both processes operate as all-or-nothing uploads as described above. The SQL/SSIS validation is more thorough than the Perl version as it validates each column (field) against the database before writing anything to the database, although both versions have validation checks encoded. Both processes contain code which debits a sample's volume when that sample has aliquots created from it. When this happens the original sample is henceforth referred to as a 'parent' and the aliquots are referred to as 'children'. Creating aliquots is a common practice in medical research for numerous reasons. One important reason is to mitigate the risk of losing a sample (particularly a rare disease sample) via mechanical failure to a freezer. By creating aliquots, samples can be stored in separate freezers even separate freezers in separate buildings when appropriate. The figure below displays is a segment of Perl code which debits the volume of a parent sample.

```

1974 }
1975
1976 ##### update_parent_sample_amount #####
1977 sub update_parent_sample_amount {
1978     my %hash_in;
1979     my ($hash) = @_;
1980     %hash_in = %{$hash};
1981
1982     #0.75 is ml (cc in the interface) for saliva. I think they use % of that for their preps,
1983     #so 0.35. Then 0.75-0.35=0.4 ml remaining. The DNA would then be in ul and I think is 100 ul
1984     #(maybe 200 or 300 depending on how they run the instrument). This is why it is important to have
1985     #units in the headings. If you want them to always use ul, put that in the headings and make sure to tell them to do so.
1986     #CAST(PostalCode AS INT)
1987
1988     # Gets the name of this procedure, here = 'create_new_sampleevent_record' so it can be saved to log record if needed
1989     my $subs_name = ( caller(0) )[3];
1990     print "SUB NAME: $subs_name\n";
1991
1992     my $QUERY = '';
1993     if ( $parentsamplero == 0 ) {
1994         $QUERY = "BEGIN TRY
1995             BEGIN TRANSACTION UPDTPARENTSAMPLEVOL
1996             UPDATE [$db_env].[dbo].[SampleDescription]
1997             SET [Volume]= " . $hash_in{'parent_sample_new_vol'} . "
1998             WHERE [SampleID]= " . $hash_in{'parentsamplero'} . "
1999
2000             COMMIT TRANSACTION UPDTPARENTSAMPLEVOL
2001         END TRY
2002         BEGIN CATCH
2003             IF @@TRANCOUNT > 0 ROLLBACK TRANSACTION UPDTPARENTSAMPLEVOL
2004             exec [$db_env].[dbo].[usp_Error_LogError] \@Error_LogTable = '_BTM_ErrorLog', \@Error_Procedure = '$subs_name', \@Error_Comment = 'Perl:$program_name:
2005             END CATCH";
2006     }
2007     else {
2008         $QUERY = "BEGIN TRY
2009             BEGIN TRANSACTION UPDTPARENTSAMPLEVOL
2010             UPDATE [$db_env].[dbo].[SampleDescription]
2011             SET [Volume]= NULL
2012             WHERE [SampleID]= " . $hash_in{'parentsamplero'} . "
2013
2014             COMMIT TRANSACTION UPDTPARENTSAMPLEVOL
2015         END TRY

```

Figure 3-19: Sample Debiting Code Snippet<sup>26</sup>

The ETL processes that compose components four through seven run every ten minutes as displayed in the figure below.

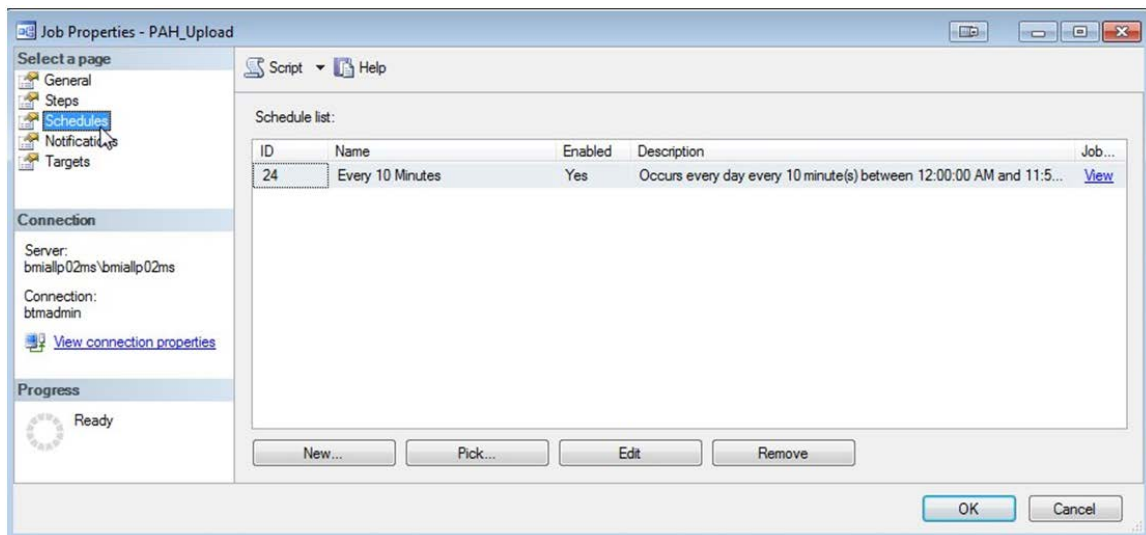


Figure 3-20: ETL Job Properties<sup>26</sup>

### 3.3.5 Component #8, BTM Application

Aside from the ETL processes described in the previous component sections, there is a significant portion of data that is manually entered into the BTM application and saved to the BTM database. This section only will cover the basic data entry operations performed by a BTM end-user; this is by no means intended to be an exhaustive description of BTM application functionality, but rather will provide an application overview to facilitate the understanding and connection to CCHMC's EDQI designed around BTM. The figure below provides a screenshot of the home screen of the BTM application. This figure also displays all of the banks currently in CCHMC's production environment.

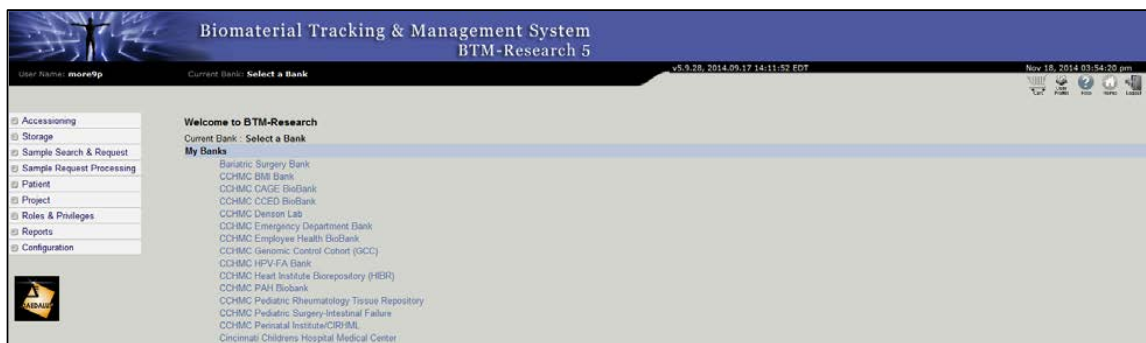


Figure 3-21: Biobanks in BTM's Production Environment<sup>30</sup>

The figure below provides expanded views of the BTM navigation panel, displayed above in Figure 3-21 on the left hand side.

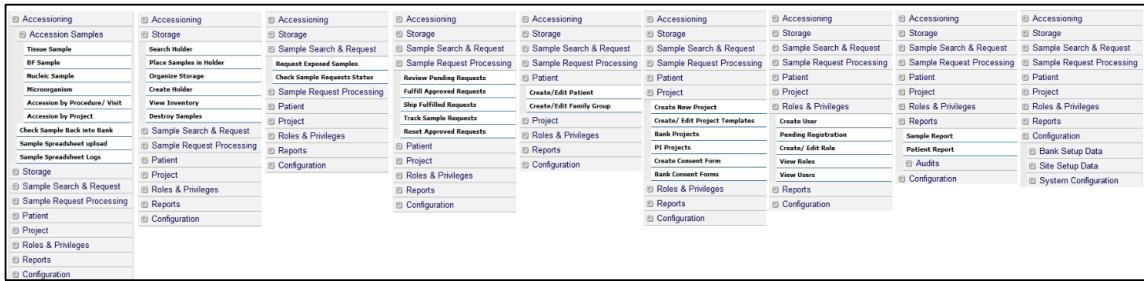


Figure 3-22: Expanded Views of BTM's Navigation Panel<sup>30</sup>

The general starting point for manual data entry is the accessioning page.

While the figure below displays a biofluid sample, there are similar screens for tissue, nucleic and isolate microorganism samples. Noteworthy on this screen is the assignment of the BTM GUID. If a user hits the save button without entering any data, BTM will assign a BTM GUID number. After a BTM GUID has been assigned, the user is presented with links to create aliquots, derivatives or isolate microorganism. Clicking any of those links provides the user with an opportunity to create a 'child' sample where the appropriate characteristics, dependent upon the particular link selected, are inherited for the user to view upon creation.<sup>31</sup>

Figure 3-23: General Tab of BTM's Accession Biofluid Sample Screen<sup>30</sup>

Users enter all sample characteristics on the Sample Description tab. As is the case throughout the BTM application, all the drop-down lists on this tab are configurable. Another key feature of the application is the configurability of user-level permissions. This is especially true with compliance to the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPPA) of 1996. BTM users can be added with or without access to protected health information (PHI) of research subjects. At CCHMC, most users have access to the PHI of the subjects within the bank they have permission to add or edit data; however they cannot view PHI, or other data, from any bank in BTM that they have not been granted permission to access.<sup>31</sup>

Accession Sample Module > Accession Biofluid Sample

### Accession Biofluid Sample

Bank Sample ID :      Project Sample ID :      External Sample ID :

Location :

General	Sample Description	Sample Comments	Oncology	Lab	Dates and Times
Total Volume	<input type="text"/>	cc	Visit Type	Select	Visit Date 04/01/2014
Number Of Containers	1				
Sample Type	Select				Cell Count: <input type="text"/> x 10E <input type="text"/>
Source	Select				
Tube Type	Select				
Additive	Select				
Collection Procedure	Select				
Hemolysis Score	Select				
Health Status at Collection	Select				
Processing Method	Select				
<b>Sample Characteristics</b>					
Add More Characteristics					
<b>Specify Individual Containers</b>					
<b>index</b>	<b>Sample Type</b>	<b>Container Type</b>			

Figure 3-24: Sample Description Tab of BTM's Accession Biofluid Sample Screen<sup>30</sup>

The annotation form is a supplemental part of BTM that can be used in lieu of or in addition to the core components of the application. The information from the annotation form is stored as BLOB text, so it is more difficult to organize or report off of these data versus data housed in other portions of the application which are stored in a SQL database. CCHMC used the annotation form function of BTM in a migration project for the PRTR Bank which is number six on the numbered components diagram listed in Figure 3-4. There were certain fields in the legacy data system that were not part of BTM. These fields were not deemed to be of significant

importance, however, the users of the legacy system wanted to retain them in some format just in case they were needed for reference at some point in the future. The figure below is a blank version of the annotation form created for that project.

Patient Module > Create/Edit Patient

### Edit Patient

Juvenile Idiopathic Arthritis ▸

#### Juvenile Idiopathic Arthritis Information

Diseased subject

PRTR disease

Family Type

#### JIA Information

Date of Onset

Date of Diagnosis

Onset (First six months)

Course

ANA: Date

ANA: Result

B27: Date

B27: Result

Rheumatoid Factor: Date

Rheumatoid Factor: Result

Rheumatoid Factor No. 2: Date

Rheumatoid Factor No. 2: Result

Evidence of Cartilage Erosion by X-ray

Joint With Erosion

Evidence of space narrowing by X-ray

Joint with Narrowing

Iritis (Uveitis/iridocyclitis)

Date of most recent exam

Active arthritis at most recent visit?

Figure 3-25: BTM annotation form example – Juvenile Idiopathic Arthritis<sup>30</sup>

The hierarchal sample storage in the BTM application is generally setup in the following format:

- 1) Facility
  - a) Bank
    - 1. Freezer
      - (1) Shelf
        - (a) Rack
          - (i) Row (optional)
            - 1. Box
              - a. Slot<sup>31</sup>

The figure below provides an illustration of BTM's storage hierarchy from facility to box.

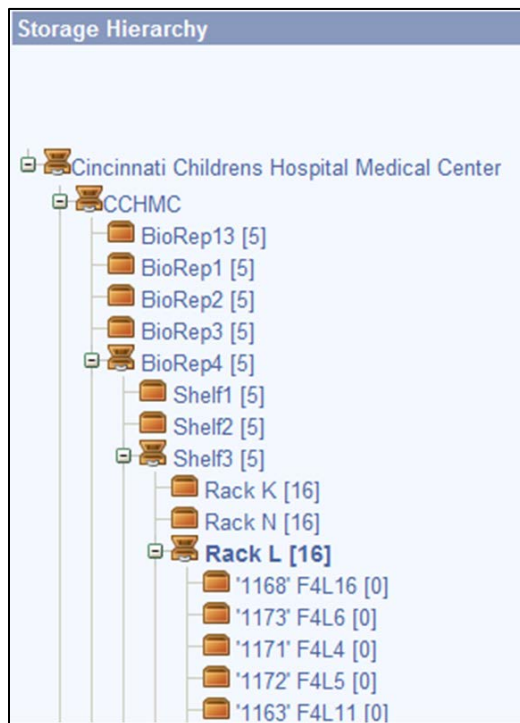


Figure 3-26: Expanded view of BTM's storage hierarchy<sup>30</sup>

The figure below provides a visual illustration of a box (or holder) configuration with all the slot (or well) identification numbers.

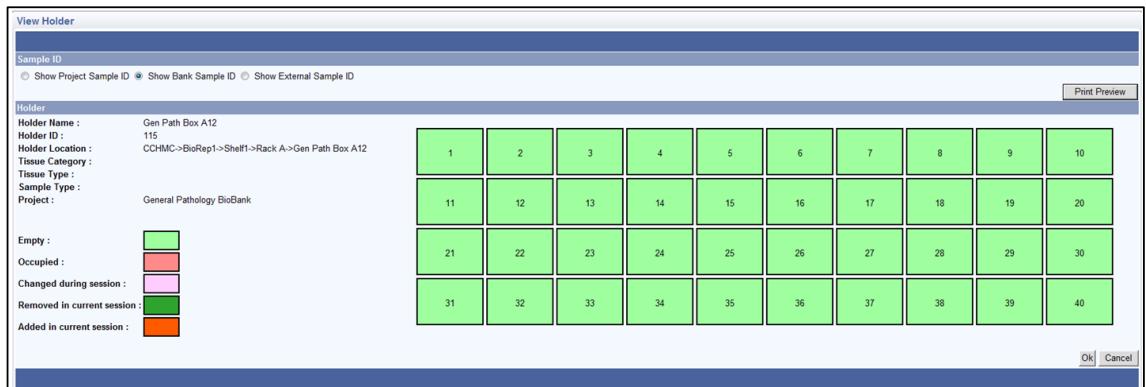


Figure 3-27: Holder Level View of BTM's Storage Hierarchy<sup>30</sup>

BTM captures patient data in a similar fashion to sample data. The following three figures below display three different tabs within the Create/Edit Patient module. An important piece of functionality within this module is the Family Group concept shown on the Project tab of the Create/Edit Patient module; displayed in subsequent figures below. The Family Group unique identification number provides a link between research subjects (patients) and their family members. This functionality is particularly important for genetic research projects when tracing the genetic linkage from the subject forward or backward.<sup>31</sup>

Subject Module > Create/Edit Patient

Create/Edit Patient

Medical Facility/ MRN

Verification Complete

General Demographics Project Oncology Soc Hx Fam Hx Procedure/Observation Annotations Clinical Patient Comments File Upload

First Name

Middle Name

Last Name

Verification Complete

MRN/ Unique ID

Medical Facility

Select

add more

Figure 3-28: General Tab of BTM's Create/Edit Patient Screen<sup>30</sup>

Subject Module > Create/Edit Patient

Create/Edit Patient

Medical Facility/ MRN

Verification Complete

General Demographics Project Oncology Soc Hx Fam Hx Procedure/Observation Annotations Clinical Patient Comments File Upload

Project Display Name

Research Subject ID

Family Group

#Samples

Anonymous control - P01

Figure 3-29: Project Tab of BTM's Create/Edit Patient Screen<sup>30</sup>

Subject Module > Create/Edit Subject

Create/Edit Subject

Project Integrative Genomics Research Subject ID Family/ Group Name Verification Complete

General Demographics Demographics Consents Family Membership

Family/ Group Name Last Name First Name Medical Facility MRN/ UID Subject ID Project Gender Relationship with Main Subject Afflicted

gr granddaughter (biological)

gr granddaughter (non-biological)

gr grandson (biological)

gr grandson (non-biological)

Husband

Maternal Aunt

Maternal Aunt (consanguine)

Maternal Aunt (non-biological)

Maternal Cousin (Female)

Maternal Cousin (Male)

Maternal Grand Father

Maternal Grand Mother

Maternal Grandfather

Maternal Grandmother

Maternal Great-Aunt

Maternal Great-Grandmother

Maternal Uncle

Maternal Uncle (by marriage)

Maternal Uncle (consanguine)

Monozygotic Twin

Mother (biological)

Mother (non-biological)

Nephew

Nephew (by marriage)

Nephew (Consanguine)

Niece (by marriage)

Niece (Consanguine)

Older Brother

Older Sister

Other

Figure 3-30: Family Membership Tab of BTM's Create/Edit Patient Screen<sup>30</sup>

Two of the three preceding figures were from the Create/Edit Patient screen and one was from the Create/Edit Subject screen. The following bullet

points provide information on the fundamental data relationships and connections in BTM:

- A 'Site' contains multiple 'Banks'.
- A 'Bank' contains 'Medical Facilities' which make 'Patients' available to the 'Bank'.
- A 'Bank' contains 'Projects' which contains 'Patients' as 'Subjects'.
  - 'Subjects' are unique on a project level.
- A 'Project' contains 'Samples', which are linked to the 'Subjects'.
  - 'Banks' can also contain 'Samples', which can be linked directly to 'Patients', however, this usually, happens via 'Projects' and 'Subjects'.<sup>31</sup>

The figure below provides a graphical illustration of the BTM relationships.

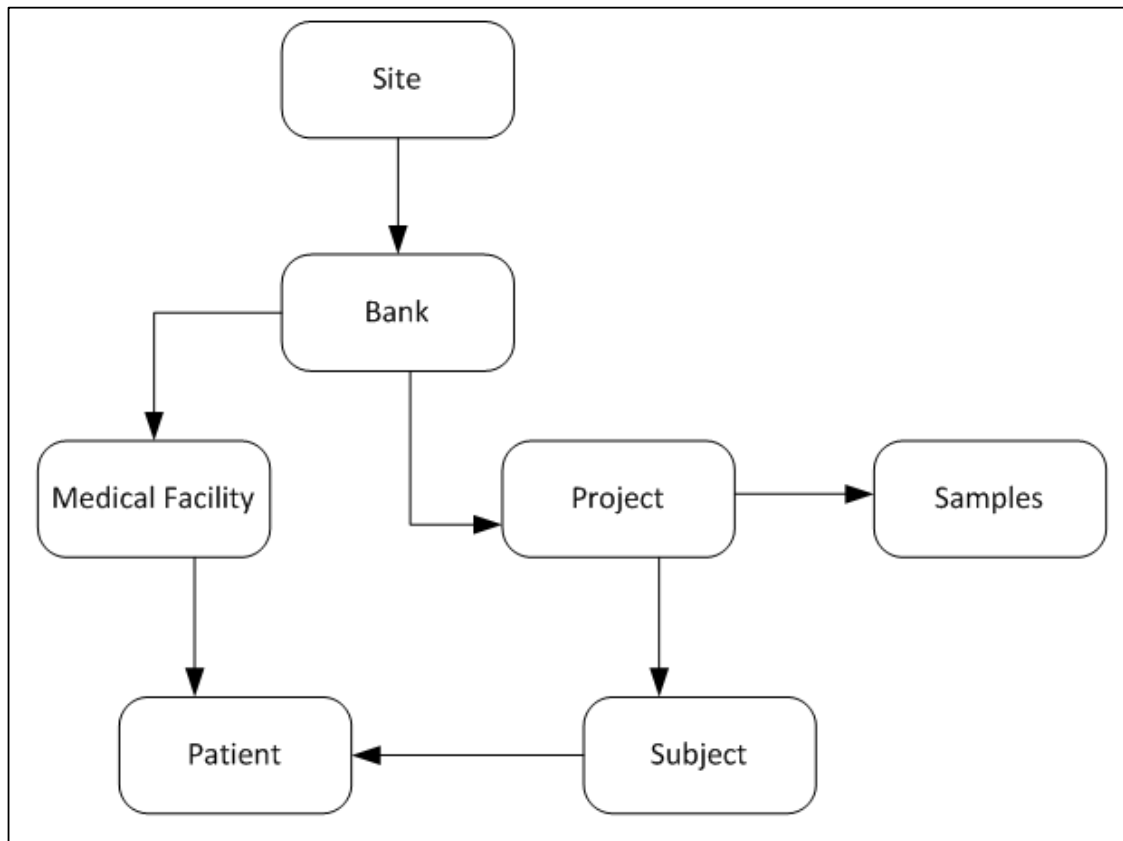


Figure 3-31: BTM Relationship Overview<sup>25</sup>

### 3.3.6 Component #9, i2b2 ETLs from BTM Data Warehouse

This component describes two of CCHMC's most important ETL processes and emphasizes the need for an EDQI system as poor data quality in BTM will migrate directly to i2b2 and will adversely affect decisions made by CCHMC medical researchers. CCHMC researchers perform de-identified cohort identification and analysis via CCHMC's custom version of i2b2. CCHMC uses i2b2 in this capacity for several data sources in addition to

BTM. The figure below is an example of one i2b2 Workbench where a de-identified cohort would be obtained.<sup>24, 25</sup>

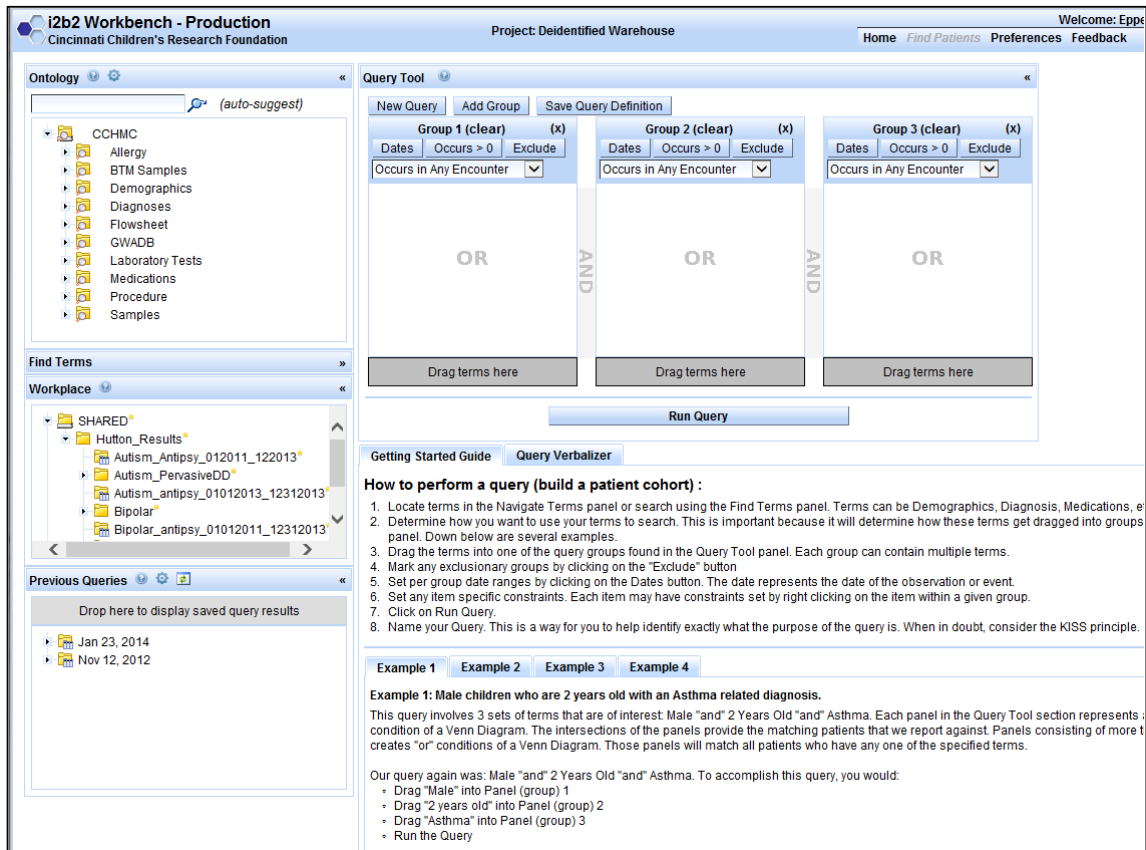


Figure 3-32: i2b2 Production Workbench<sup>19</sup>

End users combine cohort elements as desired to see options available in the biorepository. This concept is explained at the bottom of the figure above with the example of two year old male patients with asthma.<sup>19</sup>

As depicted in the Component Diagram, Figure 3-4, the i2b2 data feeds from BTM actually come from the BTM data warehouse, not the BTM production database. The BTM data warehouse is a copy of numerous data fields from the BTM production database. It is refreshed nightly which is more than satisfactory for research needs. The BTM data warehouse was constructed for three basic reasons:

1. Reduce the complexity of the production database
  - a. The BTM data warehouse only has three fact tables and approximately 40 total tables versus the production database which has over three times that many.
2. Security
  - a. The BTM data warehouse provides a mechanism for limited views to be provided to researchers versus providing researchers with access to the production database.
3. Reporting
  - a. For analytics, the BTM data warehouse provides an alternative to running reports against the production database which can negatively affect database performance.<sup>25</sup>

Per the fact table reference in number one above, in BTM, projects can belong to multiple banks (although CCHMC has not done this to date).

Therefore, all queries including 'Bank' need to go through the FactSample table. Also, in order to get the correct counts by sample, the FactSample.SampleID needs to be counted distinctly, see the figure below.

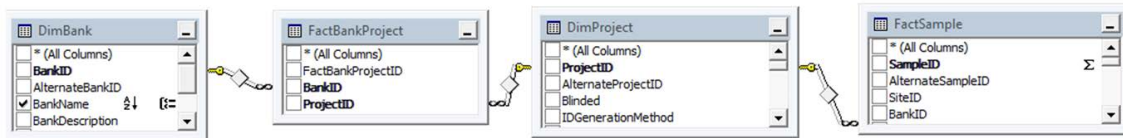


Figure 3-33: BTM FactSample Table<sup>25</sup>

In BTM, a subject is a patient that belongs to a project. A patient can belong to multiple medical facilities and projects. Meaning, a patient potentially has multiple projects, medical facilities, subjectids, and MRNs. In order to get correct counts by Patient/Subject, the FactSubject.PatientID needs to be counted distinctly, see the figure below.<sup>25, 31</sup>

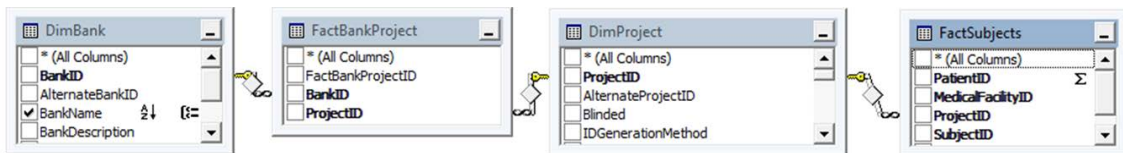


Figure 3-34: BTM FactSubject table

The figure below illustrates the basic data flow from the BTM database to the BTM data warehouse then to the i2b2 data warehouse. The final data feed to i2b2 provides data from the BOFC and PAH banks in BTM banks. This is Component #9 on the Component Diagram in Figure 3-4.

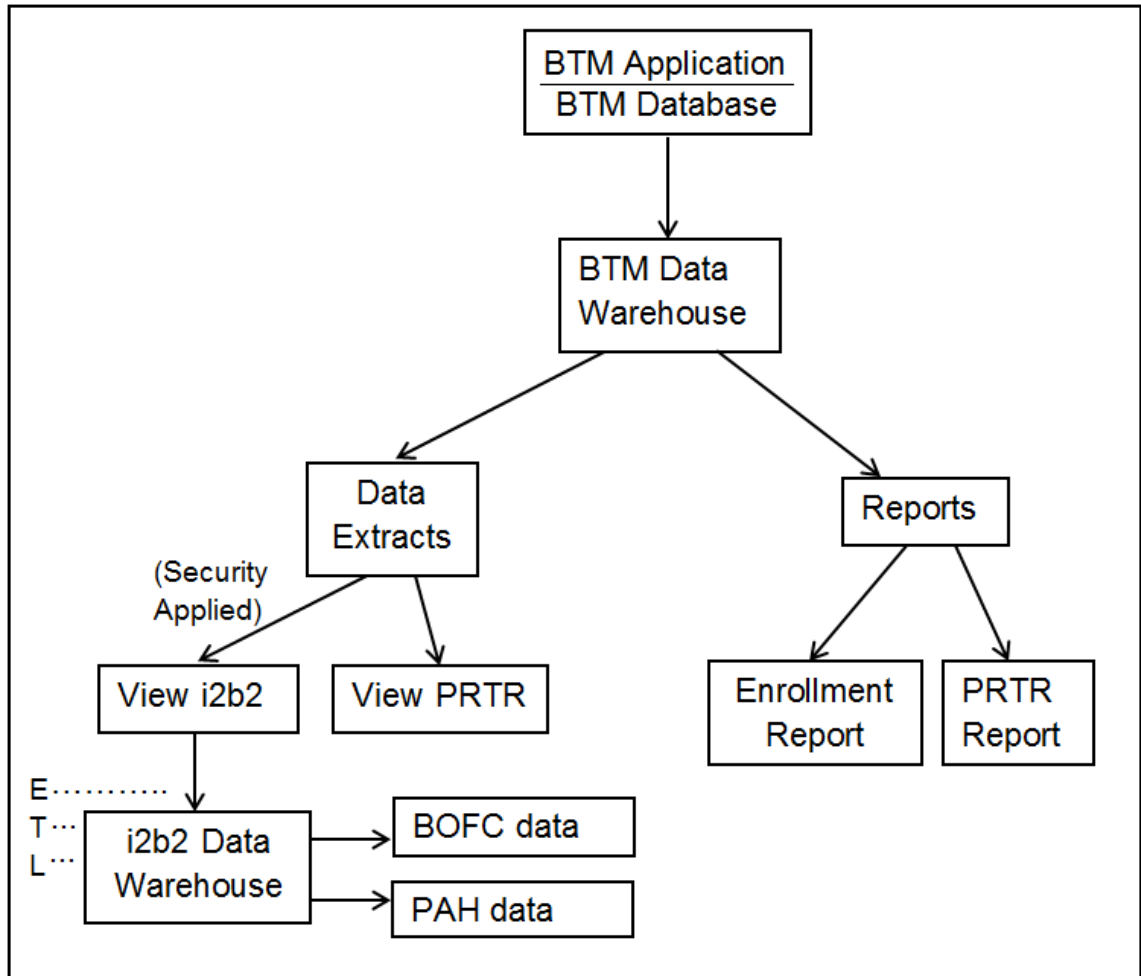


Figure 3-35: BTM to i2b2 Data Flow Diagram<sup>25</sup>

The BTM data warehouse nightly refresh process consists of the following:

1. SQL Server Agent Job “BTM\_DW\_ETL\_Process”.
2. Completion of the previous step initiates an SSIS Package “BTM\_DW Version02\_From\_Prod.dtsx”
  - a. This package moves all relevant tables from Production to [BTM\_DW\_ETL].
  - b. This package also executes procedure [BTM\_DW\_ETL].[dbo].[usp\_ETL\_Version02]

3. Completion of the previous step initiates the truncation and rebuilding of [DW]
  - a. [DW] is then copied to the production database [BTM\_DW]

The figure below is a macro-level view of the BTM Data Warehouse Entity Relationship Diagram (ERD). This view is strictly to see the general design characteristics of the warehouse, the subsequent figures provide more detailed views and legible column names used in the warehouse construction.<sup>25</sup>

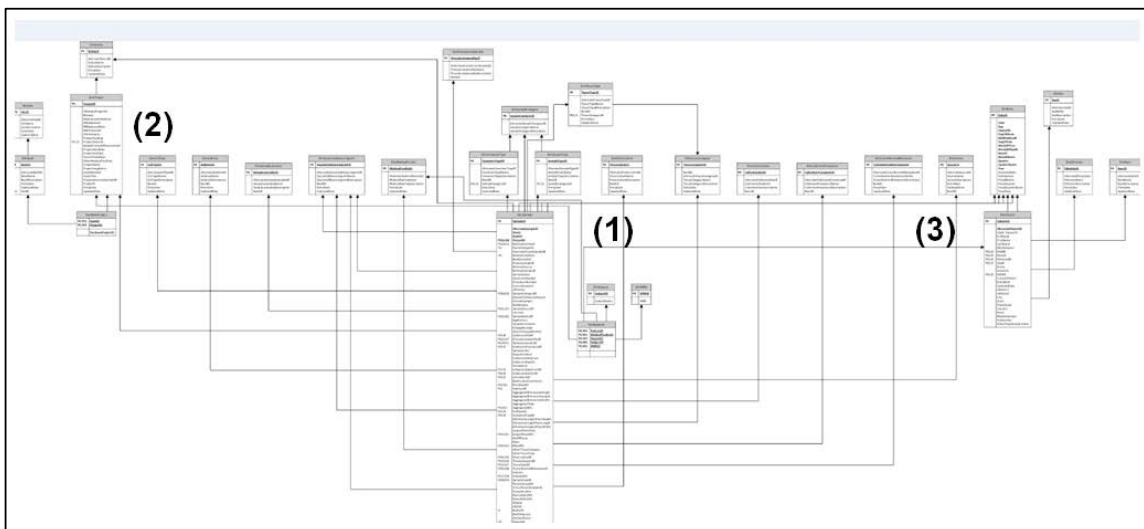


Figure 3-36: BTM Data Warehouse Entity Relationship Diagram (Macro View)<sup>25</sup>

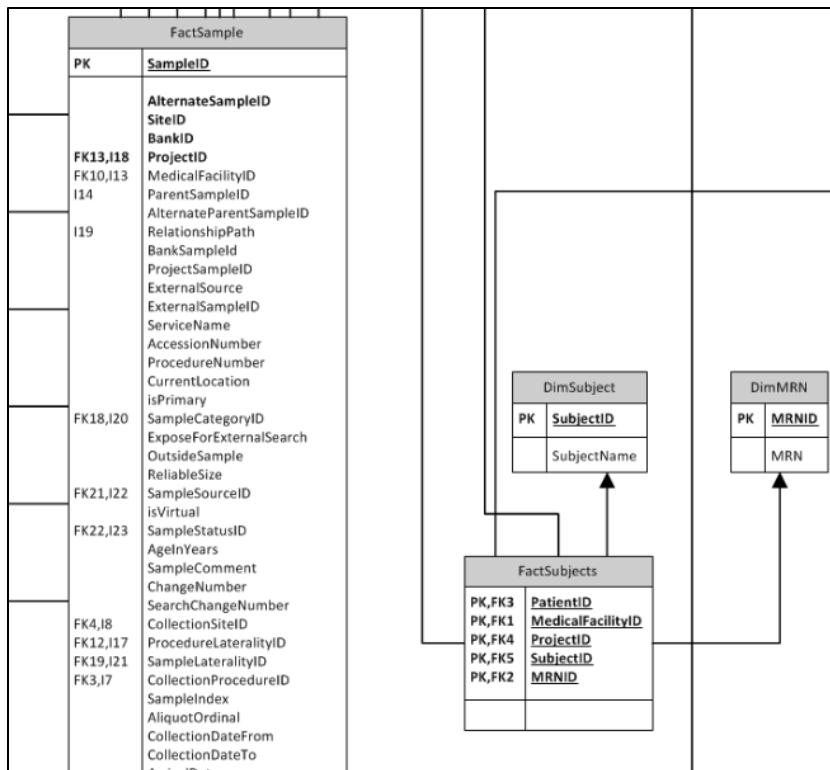


Figure 3-37: Detailed View of Section #1 of the BTM Data Warehouse<sup>25</sup>

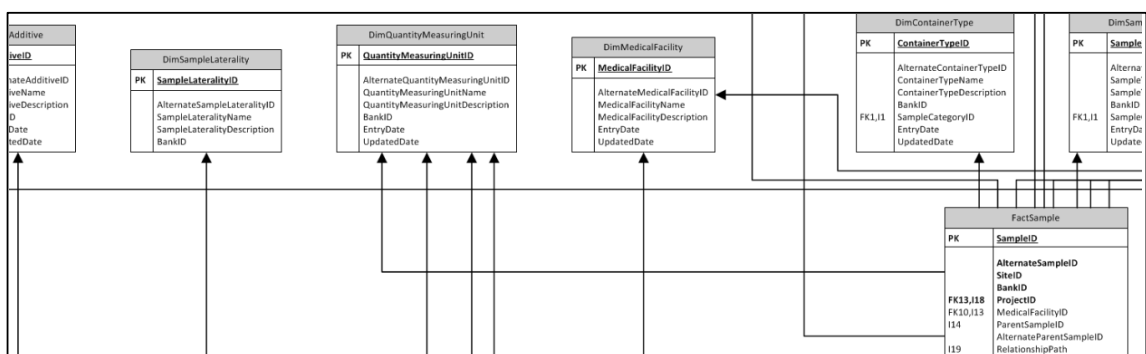


Figure 3-38: Detailed View of Section #2 of the BTM Data Warehouse<sup>25</sup>

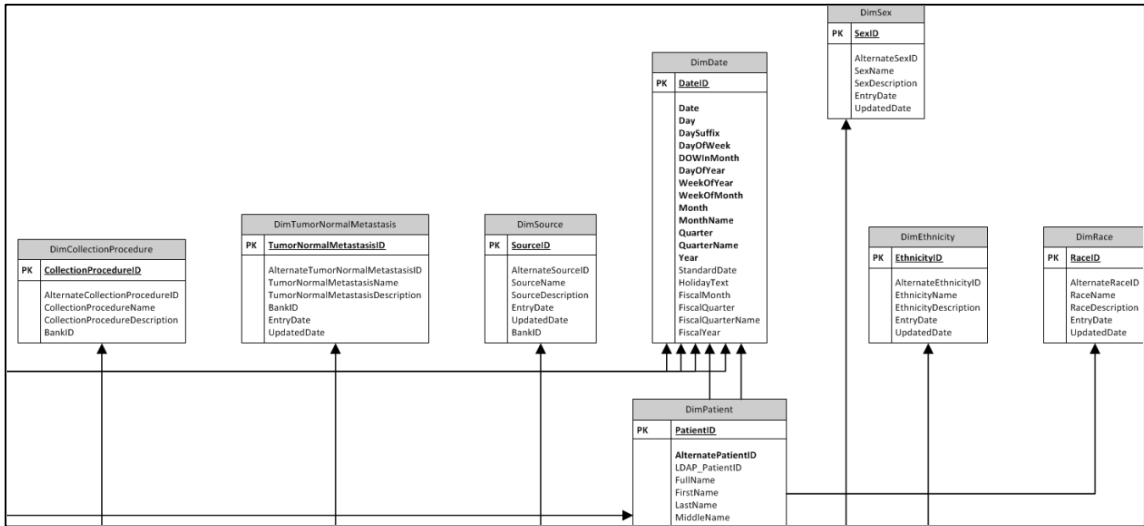


Figure 3-39: Detailed View of Section #3 of the BTM Data Warehouse<sup>25</sup>

### 3.3.7 Component #10, Enterprise vs. Bank Level Queries

At a high level, the EDQI system consists of two major components that continually identify non-compliant data elements, enterprise and bank level quality queries. As depicted in the figure below, the enterprise level queries go across all the banks in BTM and are coded independent of specific BTM bank data quality requirements. ‘Non-compliant’ at the enterprise level describes a violation of data rules that are consistent amongst all banks in BTM. The results are grouped per the BTM bank to which they belong.

Bank level queries are coded per the specific requirements provided by the BTM bank owner or designee. The process starts when the user goes to the BTM Help website, displayed in the subsequent figures below, and input

their requirements for non-compliant data. As displayed in the figure below, as part of this study, the ‘Data QI’ tab represents the implementation of the EDQI system for CCHMC’s BTM software. Moving forward, this process will be one of a group of processes that are activated with each new CCHMC bank added to BTM.

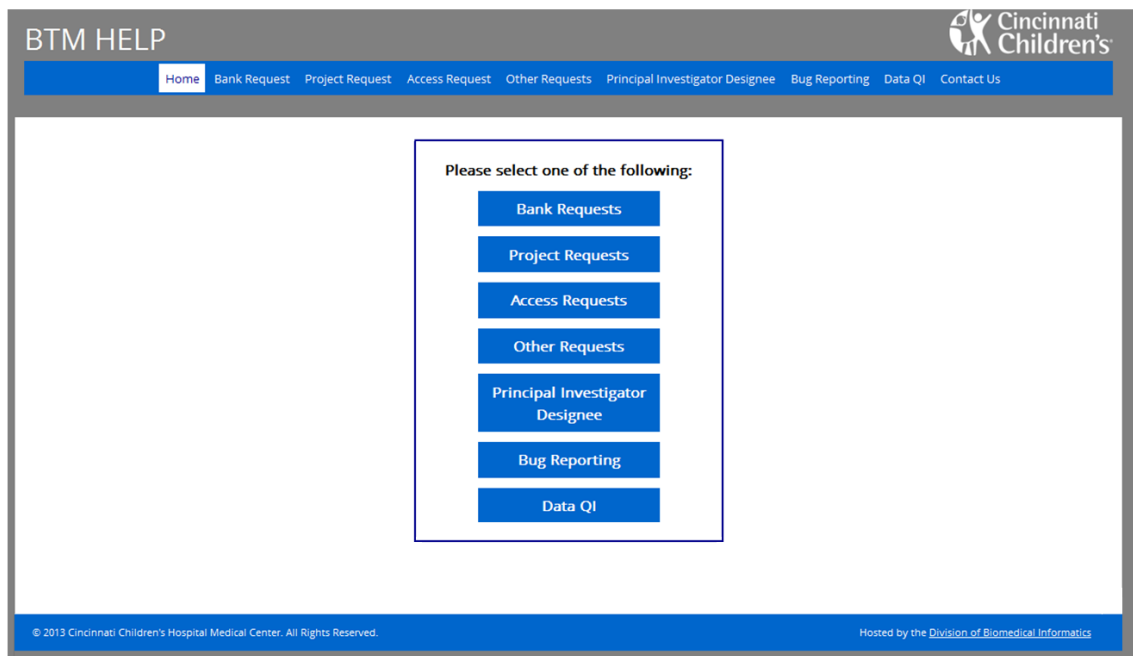



Figure 3-40: BTM Help Homepage<sup>29</sup>

Upon clicking the Data QI button, users are taken to the screen displayed in the figure below where they input the specific bank level requirements.

Although the website is setup to submit all entries directly to CCHMC’s RT HelpDesk System, which creates a ticket and sends a copy of the ticket to all members of CCHMC’s BTM support team, some minimal translation

between user and developer is usually required by an Applications Specialist or Business Analyst.

BTM HELP



HomeBank RequestProject RequestAccess RequestOther RequestsPrincipal Investigator DesigneeBug ReportingData QIContact Us

## Data Quality Improvement Report

\* denotes required field

Contact Information

Name \*

Email \*

Confirm Email \*

Additional email addresses that should receive the Quality Report

Email

Email

Email

Bank Name \*

Project Name(s)

Please list the BTM field(s) as they appear in the application along with the condition(s) you would like checked.  
Example #1: Collection Site should never be null (blank).  
Example #2: If the Sample type is blood and the Additive is EDTA then the Tube Type should be Pink Top Tube.

\*

Example #1: Collection Site should never be null (blank). Example #2: If the Sample type is blood and the Additive is EDTA, then the Tube Type should be Pink Top Tube.

Frequency of Quality Report Results Distribution \*

☐ Daily

☐ Weekly

☐ Monthly

☐ Quarterly

Submit

© 2013 Cincinnati Children's Hospital Medical Center. All Rights Reserved.

Hosted by the [Division of Biomedical Informatics](#)

Figure 3-41: BTM Help Website Data Quality Improvement Report Page<sup>29</sup>

After the bank level requirements are clarified, the bank level queries are developed and implemented. Both the bank level and the enterprise level queries are executed at the interval requested by the bank owner. The query results are bundled per BTM bank and distributed via an SSRS email to the bank's designated recipients. The end user receives a single list of non-compliant data elements for further investigation. Per the end user's perspective, it is a unified process. In reality, however, the only common elements in this process are the enterprise level queries. The SSRS method of delivering the results of the queries is virtually the same; the only customized piece is the frequency of delivery. The figure below displays the relationship between the queries, distribution tools and the banks.

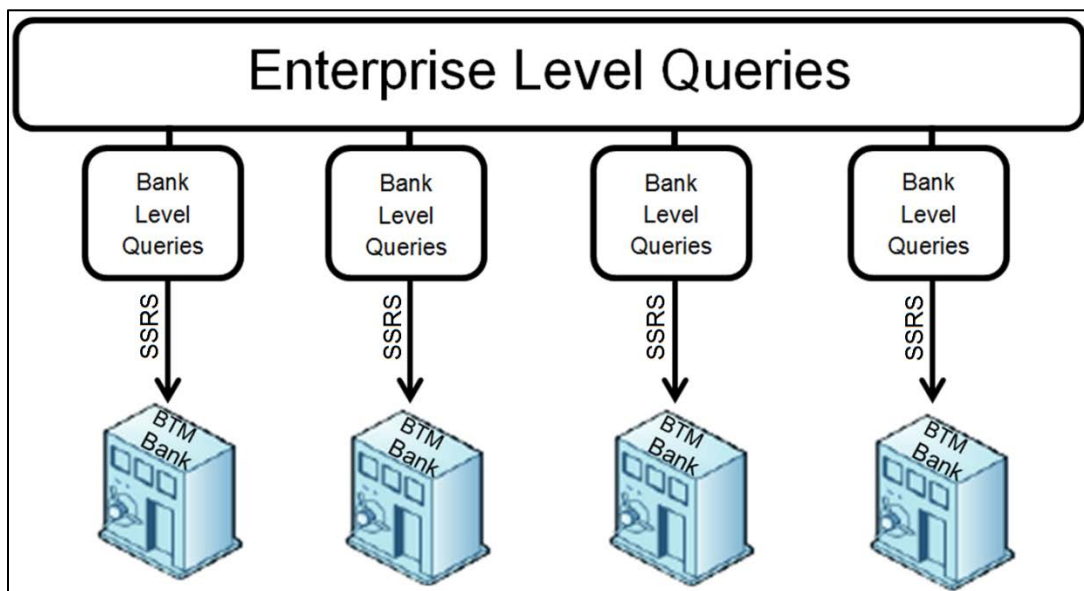


Figure 3-42: EDQI and Bank Level Query Data Flow

As with any requirements gathering process, sometimes the end result is not what the end-user expects and the requirements are adjusted post implementation to better accommodate user expectations. Developing the bank level queries are not exempt from this tendency as users often receive greater or fewer non-compliant data returns than originally anticipated. Generally, after a few cycles of receiving the EDQI results and performing subsequent investigations, the requirements are modified to meet the needs of the lab and the process is generally maintenance free. The only exceptions to this statement are intentional changes that occur as changes take place with a particular lab's workflow. Such a change would prompt the EDQI process to be modified and to evolve in sync with the lab's operations.

### **3.3.8 Component #11, Non-Compliant Data Process**

The data correction workflow for this component begins when the designated BTM end-users receive automated emails via SSRS listing any patients, samples or subjects that were identified as non-compliant via the quality scripts that were described in the previous section. The figure below is an example of such an email:

From: nudeic-uploader@cchmc.org											
To: Carr, Michael; Morgan, James											
Cc:											
Subject: BTM_Data_Quality_Check was executed at 11/5/2014 2:25:07 PM											
<p style="text-align: center;"><b>BTM Data Quality Check Report</b></p> <p><b>Bank: CCHMC PAH Biobank</b> <span style="float: right;"><b>Date: 11/5/2014</b></span></p> <table border="1"> <thead> <tr> <th colspan="2">Missing CollectionSiteEnumID: 100 Records</th></tr> <tr> <th>Sample GUID</th><th>Bank Specific Sample Id</th></tr> </thead> <tbody> <tr> <td>967adcd9-872e-4ab0-bebb-c079e610edfd</td><td>BS 123</td></tr> <tr> <td>0fa274be-e4f8-441b-9bb6-d01cc9b3b2f9</td><td>BS 456</td></tr> <tr> <td>714cc39a-6b42-4e3d-925c-909ffd6ac4e9</td><td>BS 789</td></tr> </tbody> </table>		Missing CollectionSiteEnumID: 100 Records		Sample GUID	Bank Specific Sample Id	967adcd9-872e-4ab0-bebb-c079e610edfd	BS 123	0fa274be-e4f8-441b-9bb6-d01cc9b3b2f9	BS 456	714cc39a-6b42-4e3d-925c-909ffd6ac4e9	BS 789
Missing CollectionSiteEnumID: 100 Records											
Sample GUID	Bank Specific Sample Id										
967adcd9-872e-4ab0-bebb-c079e610edfd	BS 123										
0fa274be-e4f8-441b-9bb6-d01cc9b3b2f9	BS 456										
714cc39a-6b42-4e3d-925c-909ffd6ac4e9	BS 789										

Figure 3-43: BTM Data Quality Check Report<sup>27</sup>

As described in the previous section, the emails will combine both enterprise and bank level results. The responsibility for investigating each non-compliant data element identified via the quality scripts lies with the compliance agent. “Compliance agent” generally refers to any member of a particular lab who is assigned the responsibility of investigating the data elements received via the SSRS email. This may or may not be an actual Compliance Specialist, but it is generally the lab’s liaison with the IRB. After receiving the non-compliant data list, the compliance agent will proceed to check the identified data elements against the data source for those data elements. The data source will vary from lab to lab and from data element to data element. Some examples of data sources include CCHMC’s EHR system - Epic, CCHMC’s laboratory information system – Cerner,

EHR systems for other medical organizations, laboratory paperwork sent from other medical organizations and data located on the sample itself. Once the compliance agent completes the investigation and either corrects the data or acknowledges that it is a legitimate outlier, they will then select the “Verification Complete” checkbox. As a part of this study, Daedalus Software Inc. added the “Verification Complete” checkbox to the application. The “Verification Complete” checkbox has been added at the sample, subject and patient levels. The EDQI queries perform an initial status check of this checkbox prior to any other data comparison. If the checkbox has been selected, the EDQI queries will by-pass the associated data elements at that record level (sample, subject or patient). This reduces the number of false positives after the initial investigation and checkbox selection. The figure below displays the Verification Complete checkbox at the sample level. The patient and subject level functionality is the same as the sample level.

**Biomaterial Tracking & Management System**  
BTM-Research 5

User Name: **moretp** Current Bank: **Cincinnati Childrens Hospital Medical Center** v5.9.27, 2014.09.09 10:14:07 EDT Sep 17, 2014 09:21:13 am

**Accessioning**  
Accession Records Module > Accession Biofluid Sample

**Accession Biofluid Sample**

Bank Sample ID : Project Sample ID : External Sample ID : Verification Complete :  
Patient First Name : Patient Last Name : Family Group : Diagnosis :

Location :

**General** **Sample Description** **Sample Comments** **Oncology** **Lab** **Date and Times** **Consent** **File Upload** **All Samples** **Annotations** **Clinical**

BTM GUID :  Scan barcode, if available

External Source :

External Sample ID :

Collection Site :

Collection Date : 09/17/2014 15:31

Surgery Date :

Arrival Date : 09/17/2014 15:31

Current Medical Facility :

MRN Unique ID :  Search Create/Edit Patient

Current Clinical Lab Service :

Clinical Lab Accession Number :  Scan barcode, if available

Outside Sample : ☐

Diagnostic : ☐

Sample Procured but Not Banked : ☐

**Verification Complete** : ☒

Project Name :

Project Sample ID :

Research Subject ID :

Clear All Fields Back Add New Sample Save Cancel

Figure 3-44: BTM Accession Page with Verification Complete Field<sup>30</sup>

This checkbox is never a mandatory field at the application level. Other than visual display to the end user, the only purpose of the checkbox is to serve as an indicator for the EDQI query to ignore a particular record. The ‘Verification Complete’ visual display could also be useful to some BTM users in audit scenarios, when accompanied by appropriate policies or procedures referencing the checkbox.

After the compliance agent completes their investigation and performs the appropriate subsequent action, their participation in the EDQI system is complete until the next pre-determined SSRS email arrives.

## **CHAPTER 4 RESULTS**

### **4.1 Initial Results**

Sections 4.2 and 4.3 below provide total non-compliant data elements identified at the bank level and at the enterprise level. The percentage of non-compliant data elements versus total samples is displayed simply to provide an indication of how often a non-compliant data element could be encountered for researchers, lab workers or administrative staff utilizing that particular bank's data set. Most of the non-compliant data requirements have targeted sample data versus patient data; however one results set below addresses patient data, specifically consent data derived from the RL/GL database previously described in section 3.3.2.

## 4.2 Initial Bank Level Results

### PAH Bank

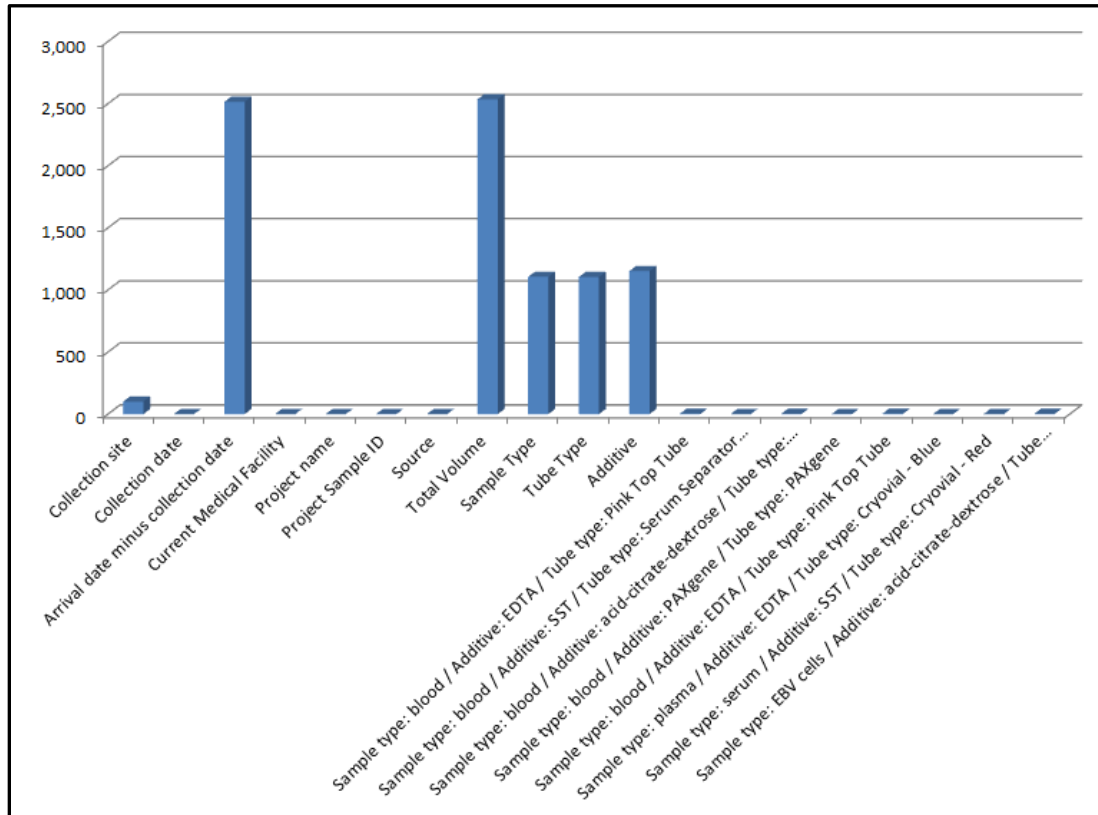


Figure 4-1: Non-Compliant Data Results for PAH Bank

The table below provides the total number of non-compliant data elements discovered via the query and the percentage that number represents against the total number of samples in the PAH Bank.

<b>Query:</b>	<b># of Non-compliant Data Elements Identified:</b>	<b>% of Non-compliant Data Elements vs. Total PAH Samples</b>
Collection site	101	0.13%
Collection date	0	0.00%
Arrival date minus collection date	2,515	3.35%
Current Medical Facility	0	0.00%
Project name	0	0.00%
Project Sample ID	1	0.00%
Source	0	0.00%
Total Volume	2,535	3.38%
Sample Type	1,105	1.47%
Tube Type	1,103	1.47%
Additive	1,151	1.53%
Sample type: blood / Additive: EDTA / Tube type: Pink Top Tube	3	0.00%
Sample type: blood / Additive: SST / Tube type: Serum Separator Tube (SST)	0	0.00%
Sample type: blood / Additive: acid-citrate-dextrose / Tube type: Yellow Top Tube	2	0.00%
Sample type: blood / Additive: PAXgene / Tube type: PAXgene	1	0.00%

Sample type: blood / Additive: EDTA / Tube type: Pink Top Tube	3	0.00%
Sample type: plasma / Additive: EDTA / Tube type: Cryovial - Blue	0	0.00%
Sample type: serum / Additive: SST / Tube type: Cryovial - Red	0	0.00%
Sample type: EBV cells / Additive: acid- citrate-dextrose / Tube type: Cryovial - Yellow	2	0.00%
<b>Total:</b>	<b>8,522</b>	<b>11.35%</b>

Table 4-1: Non-Compliant Data Results for PAH Bank

## HIBR Bank

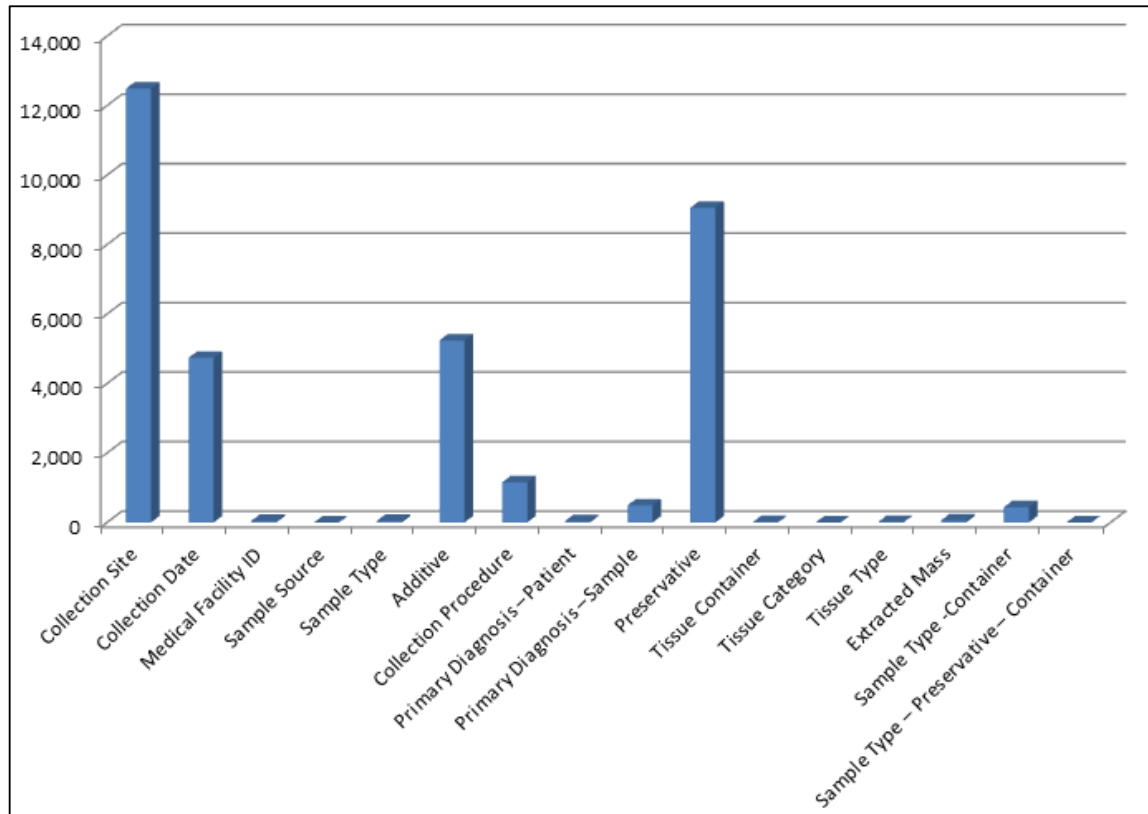


Figure 4-2: Non-Compliant Data Results for HIBR Bank

The table below provides the total number of non-compliant data elements discovered via the query and the percentage that number represents against the total number of samples in the HIBR Bank.

<b>Query:</b>	<b># of Non-compliant Data Elements Identified:</b>	<b>% of Non-compliant Data Elements vs. Total HIBR Samples</b>
Collection Site	12,486	139.43%
Collection Date	4,740	52.93%
Medical Facility	47	0.52%
Sample Source	0	0.00%
Sample Type	44	0.49%
Additive	5,240	58.51%
Collection Procedure	1,155	12.90%
Primary Diagnosis – Patient	38	0.42%
Primary Diagnosis – Sample	492	5.49%
Preservative	9,057	101.14%
Tissue Container	16	0.18%
Tissue Category	0	0.00%
Tissue Type	18	0.20%
Extracted Mass	61	0.68%
Sample Type - Container	438	4.89%
Sample Type – Preservative – Container	1	0.01%
<b>Total:</b>	<b>33,833</b>	<b>377.80%</b>

Table 4-2: Non-Compliant Data Results for HIBR Bank

## CCHMC Bank

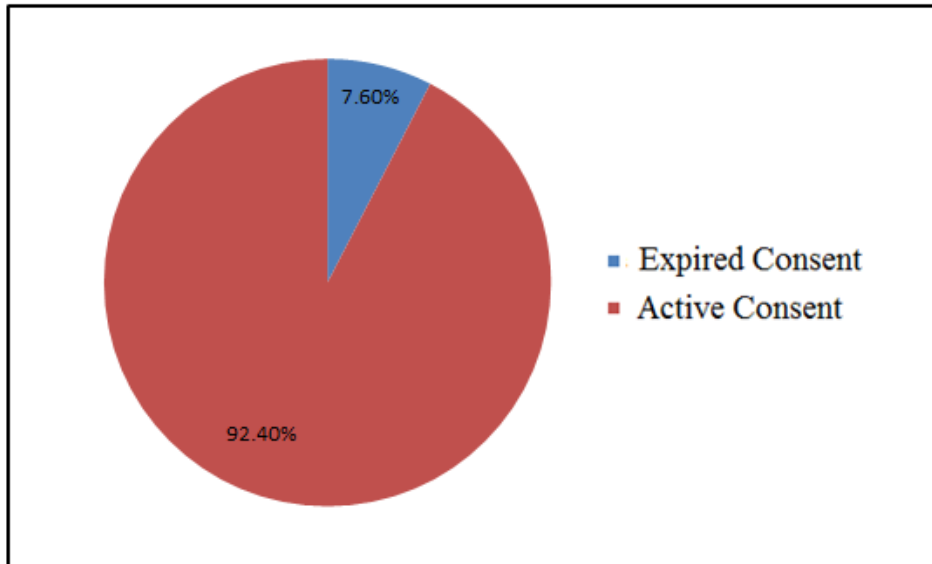


Figure 4-3: Non-Compliant Data Results for CCHMC Bank

The table below lists the total number of non-compliant data elements and the percentage that number represents against the whole. This query was slightly different than the two above as it targeted subject (patient) data versus sample data. To this particular bank, the most important quality check was that of active consent.

Query:	# of Non-compliant Data Elements Identified:	% of Non-compliant Data Elements vs. Total CCHMC Bank Subjects:
Subjects with expired consent	3,167	7.6%

Table 4-3: Non-Compliant Data Results for CCHMC Bank

## 4.3 Initial Enterprise Level Results

### All CCHMC BTM Banks

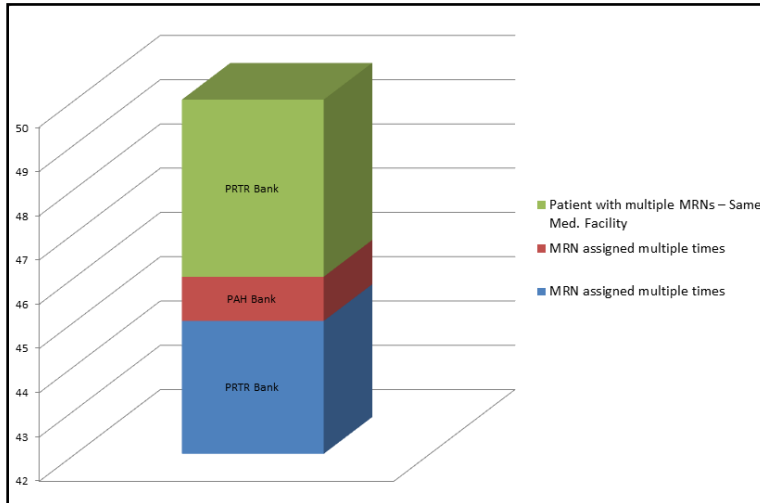


Figure 4-4: Enterprise Level Non-Compliant Results - All Banks

The table below displays the non-compliant data elements identified via enterprise level queries. These queries identified MRNs assigned multiple times and patients with multiple MRNs within the same medical facility.

Query:	# of Non-Compliant Data Elements Identified:	Bank of Non-Compliant Data
MRN assigned multiple times	45	PRTR
MRN assigned multiple times	1	PAH
Patient with multiple MRNs – Same Med. Facility	4	PRTR

Table 4-4: Non-Compliant Data Results for All CCHMC BTM Banks

## 4.4 Initial Results Analysis

One data comparison worthy of note is the total percentage of non-compliant data elements versus total number of samples for the PAH and HIBR banks. The PAH bank has a non-compliant data element in one out of ten samples, in contrast, the HIBR bank averages over three non-compliant data elements for every sample. A common finding amongst the two bank's data sets is the 'Additive' category was the third highest non-compliant data finding for both banks. Considering the data entry methods of these two banks, it is also worth noting that the PAH bank has an established ETL upload process (Component #4 of Figure 3-4) versus the HIBR bank which strictly performs manual data entry (Component #8 of Figure 3-4).

The CCHMC findings could have positive financial implications. With over 3,000 samples in the freezers with expired consent, that's a substantial amount of biobanking real estate that can be freed up and reallocated to new samples, if desired. This also empowers the biobank manager to develop new storage processes, such as replacing expired consent samples with newly consented samples. The main reason this query was developed however, is compliance. Without appropriate consent, samples should never be distributed to researchers. At the very least, this data can be combined

with the i2b2 ETL (Component #9 of Figure 3-4) to restrict the de-identified data set to only include samples with current consent. Otherwise, the check for current consent takes place after a sample request is made and this can be misleading to investigators regarding the quantity of available samples based on the specified criteria entered into i2b2 (Component #9 of Figure 3-4) .

The EDQI system will not, and should not, eliminate the need for a consent review prior to sample distribution, however, the EDQI system in essence makes this a second review which substantially decreases the probability of a sample being erroneously distributed.

## 4.5 Post Implementation Results

### PAH Data

The following charts illustrate a side-by-side data comparison of PAH data before and after implementation of the EDQI system.

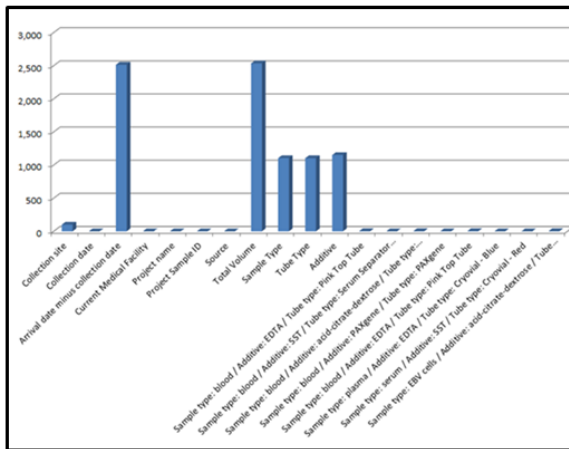


Figure 4-5: PAH Non-Compliant Data Elements Before EDQI Implementation

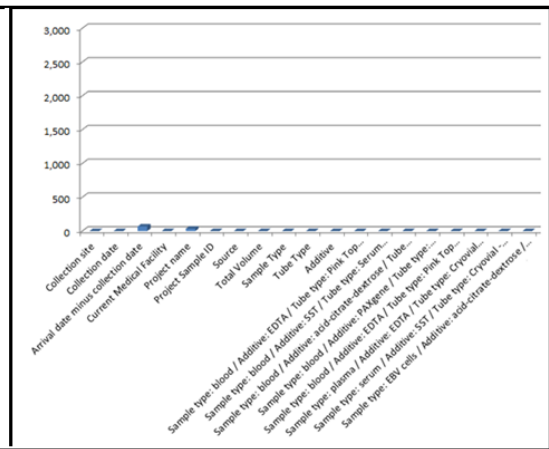


Figure 4-6: PAH Non-Compliant Data Elements After EDQI Implementation

Although visually indistinguishable in Figure 4-6, there were 100 non-compliant data elements that remained after the EDQI implementation. This was a reduction of 8,422 or 98.83% of the original non-compliant data elements. Causes of non-compliant data included:

1. Transposed dates – particularly 2024 versus 2014. This caused date range requirements to show up out-of-specification on the EDQI report.
2. Transposed fields – particularly the ‘Collection Date’ field and the ‘Arrival Date’ field were frequently interchanged. Neither field would be caught by an isolated EDQI query, however, when calculating the difference between the two dates to analyze sample degradation due to delays, the time range results were negative, a condition which is caught by an EDQI query.
3. Missing data – particularly the ‘Collection Site’ and ‘Additive’ fields were frequently omitted.
4. Incorrect drop-down selection – particularly the ‘Collection Site’, field had a high volume of non-compliant selections.

## HIBR Data

The following charts illustrate a side-by-side data comparison of HIBR data before and after implementation of the EDQI system.

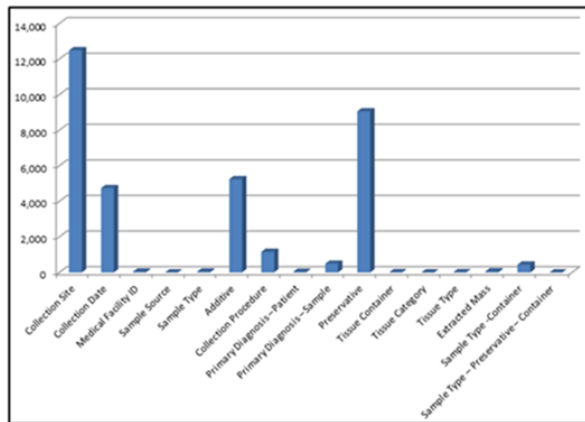


Figure 4-7: HIBR Non-Compliant Data Elements Before EDQI Implementation

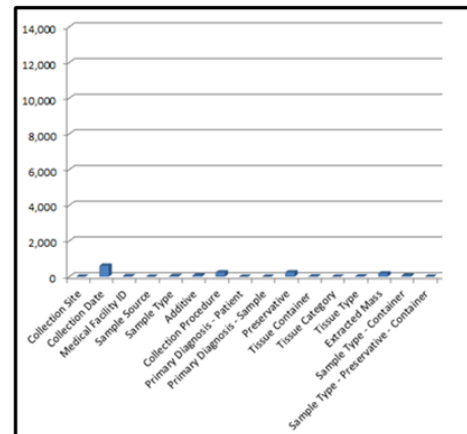


Figure 4-8: HIBR Non-Compliant Data Elements After EDQI Implementation

Although visually indistinguishable in Figure 4-8, there were 1,532 non-compliant data elements that remained after the EDQI implementation. This was a reduction of 32,301 or 95.47% of the original non-compliant data elements. Causes of non-compliant data included:

1. Omission of data fields – particularly the ‘Additive’, ‘Preservative’, ‘Primary Diagnosis for Patient’ and Primary Diagnosis for Sample’ fields.

2. Inappropriate inclusion of data fields – particularly ‘Additive’ for sample types where this field is not applicable.
3. Inappropriate configuration – the ‘Collection Date’ field existed in two tabs at the application level. One was hard-coded via the application and one was created via the configuration module.

## CCHMC Data

The following charts illustrate a side-by-side data comparison of CCHMC data before and after implementation of the EDQI system.

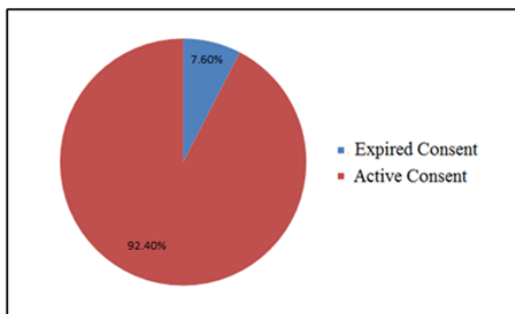


Figure 4-9: CCHMC Non-Compliant Data Elements Before EDQI Implementation

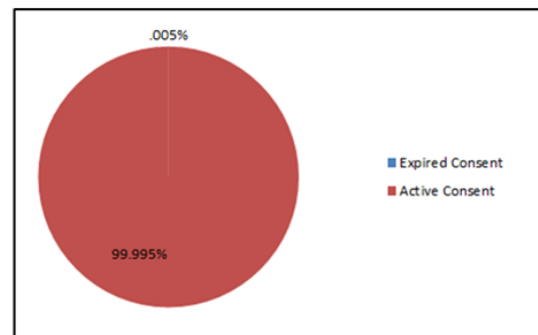


Figure 4-10: CCHMC Non-Compliant Data Elements After EDQI Implementation

Although visually indistinguishable in Figure 4-10, there were 3 non-compliant data elements that remained after the EDQI implementation. This is a reduction of 3,164 or 99.91% of the original non-compliant data elements. In contrast to the PAH and HIBR data, CCHMC’s EDQI queries

relied on a direct comparison with CCHMC's EHR system Epic which retains consent data. Comparisons were made for any non-compliant status, which included consent withdrawn and consent expired. Consents that are withdrawn are changed by the registrars to expired status, so both statuses were combined into the 'Expired' non-compliant data category.

## **CHAPTER 5 CONCLUSIONS and DISCUSSION**

Discovery of non-compliant data elements was made at both the bank and enterprise levels. This discovery is the initial, and arguably most crucial, mechanism in the EDQI system. After implementation of the EDQI system, a total of 45,572 non-compliant data elements were discovered. The non-compliant data elements were routed to the biobank owners for investigation and correction. Some correction efforts included creation and execution of scripts to correct large data sets at one time. Other correction efforts were manual efforts requiring investigation into small data sets or even individual data elements. After all correction efforts were complete, the EDQI queries were run again, and of the 45,572 non-compliant data elements discovered, 43,887 were corrected. That is a decrease in non-compliant data elements of 96.30%. Thus the hypothesis presented in section 1.5, "if an EDQI system is

developed and implemented for CCHMC's biobanking data system, then the percentage of non-compliant data will decrease" can be confirmed.

Development and implementation of an EDQI system for this dissertation combined static and flexible elements for a custom solution for CCHMC's biobanking data. Each biobanking group now has the ability to build upon the EDQI system to ensure data quality to the degree desired and in the particular data areas of concern. Concepts such as six sigma, poke yoke and kaizen, which are frequently associated with improving manufacturing processes, can be applied to each biobanking group's data process using the EDQI system.

## **CHAPTER 6 FUTURE OPPORTUNITIES**

For CCHMC, confirmation of the hypothesis will lead to a full and permanent implementation of the EDQI system for CCHMC's biobanking data. Post implementation, 'quality improvement techniques' can be implemented against the 'quality improvement system' (EDQI) itself, just like any other system that needs systemic improvement. After the system has been in place for a period of time longitudinal analysis can be completed on non-compliant data results to look for trends or patterns. Identification of

a trend or pattern may influence the data entry method, the query that led to the discovery or the workflow in the lab. On the informatics side, a macro-level view for trends or patterns across all banks will enable best practices in quality improvement techniques to be shared across the organization. The informatics side also will also have the unique collection of requirements that can be shared amongst all the CCHMC banks as best practices in quality improvement.

The focus on the EDQI system is proactively addressing any erroneous data before the data is distributed and decisions are made on that data. Another future opportunity to improve the EDQI system is performing a retrospective analysis of incidents where erroneous data was discovered and how the data got into the BTM database. Such a retrospective analysis will provide additional requirements for the EDQI system and enhance the system to prevent those incidents from occurring in the future.

Finally, some of the EDQI requirements may eventually be included in institutional policies and procedures. This would seem to be particularly appropriate with legal and compliance issues such as family group identification regarding genetic research and consent status where the stakes are high regarding erroneous data and include institutional liability versus data that might only bring limited liability to a particular study or a

particular principal investigator. The principal investigator obviously has a vested interest in filling out the bank level EDQI requirements to ensure integrity of the data at the study level. The EDQI system provides a mechanism to protect both strategic level institutional needs and tactical level study-based needs.

## REFERENCES

1. Yepes AJ, Berlanga R. *Knowledge based word-concept model estimation and refinement for biomedical text mining*. J Biomed Inform. 2015 Feb; 53:300-7.
2. Nicholas Anderson, Aaron Abend, Aaron Mandel, Estella Geraghty, Davera Gabriel, Rob Wynden, Michael Kamerick, Kent Anderson, Julie. *Implementation of a de-identified federated data network for population-based cohort discovery*. J Am Med Inform Assoc. 2012 Jun; 19(e1):e60-7.
3. Tracy Ann Sykes, Viswanath Venkatesh, Arun Rai. *Explaining physicians' use of EMR systems and performance in the shakedown phase*. J Am Med Inform Assoc. 2011 Mar-Apr; 18(2):125-30.
4. Epstein RH, St Jacques P, Stockin M, Rothman B, Ehrenfeld JM, Denny JC. *Automated identification of drug and food allergies entered using non-standard terminology*. J Am Med Inform Assoc. 2013 Sep-Oct; 20(5):962-8.
5. Wade TD, Hum RC, Murphy JR. *A Dimensional Bus model for integrating clinical and research data*. J Am Med Inform Assoc. 2011 Dec; 18 Suppl 1:i96-102.
6. Agarwal RK, Sedai A, Dhimal S, Ankita K, Clemente L, Siddique S, Yaqub N, Khalid S, Itrat F, Khan A, Gilani SK, Marwah P, Soni R, Missiry ME, Hussain MH, Uderzo C, Faulkner L. *A prospective international cooperative information technology platform built using open-source tools for improving the access to and safety of bone marrow transplantation in low- and middle-income countries*. J Am Med Inform Assoc. 2014 Nov-Dec; 21(6):1125-8

7. A. Boussadi, C. Bousquet, B. Sabatier, T. Caruba, P. Durieux, P. Degoulet. *A Business Rules Design Framework for a Pharmaceutical Validation and Alert System*. Methods Inf Med. 2011; 50(1):36-50
8. Strasberg HR, Del Fiol G, Cimino JJ. *Terminology challenges implementing the HL7 context-aware knowledge retrieval ('Infobutton') standard*. J Am Med Inform Assoc. 2013 Mar-Apr; 20(2):218-23.
9. Bogdan CM, Popovici DM. *Information system analysis of an e-learning system used for dental restorations simulation*. Comput Methods Programs Biomed. 2012 Sep; 107(3):357-66.
10. Cormont S, Vandenbussche PY, Buemi A, Delahousse J, Lepage E, Charlet J. *Implementation of a platform dedicated to the biomedical analysis terminologies management*. AMIA Annu Symp Proc. 2011; 2011:1418-27.
11. Sorani MD, Manley GT, Claude Hemphill J, Baranzini SE. *Dynamic, multi-level network models of clinical trials*. Biocomputing 2011: pp. 38-49.
12. Anchala R, Di Angelantonio E, Prabhakaran D, Franco. *Development and validation of a clinical and computerised decision support system for management of hypertension (DSS-HTN) at a primary health care (PHC) setting*. PLoS One. 2013 Nov 5; 8(11):e79638
13. Boussadi A, Caruba T, Karras A, Berdot S, Degoulet P, Durieux P, Sabatier B. *Validity of a clinical decision rule-based alert system for drug dose adjustment in patients with renal failure intended to improve pharmacists' analysis of medication orders in hospitals*. Int J Med Inform. 2013 Oct; 82(10):964-72.
14. Kevin A. Kupzyk and Marlene Z. Cohen. *Data Validation and Other Strategies for Data Entry*. West J Nurs Res. 2015 Apr; 37(4):546-56.

15. Wang SM, Hu SY, Chen F, Chen W, Zhao FH, Zhang YQ, Ma XM, Qiao YL. *Clinical evaluation of human papillomavirus detection by careHPV™ test on physician-samples and self-samples using the indicating FTA Elute® card*. APJCP Vol. 15, 2014 Issue Number 17, 7085-7090
16. Zimlichman E, Rozenblum R, Salzberg CA, Jang Y, Tamblyn M, Tamblyn R, Bates DW. *Lessons from the Canadian national health information technology plan for the United States: opinions of key Canadian experts*. J Am Med Inform Assoc. 2012 May-Jun; 19(3):453-9
17. Marsolo, Keith. *Informatics and operations--let's get integrated*. J Am Med Inform Assoc. 2013 Jan-Feb; 20(1): 122–124.
18. Keith Marsolo, Jeremy Corsmo, Michael G Barnes, Carrie Pollick, Jamie Chalfin, Jeremy Nix, Christopher Smith, Rajesh Ganta. *Challenges in creating an opt-in biobank with a registrar-based consent process and a commercial EHR*. J Am Med Inform Assoc. 2012 Nov-Dec; 19(6): 1115–1118.
19. Natter MD<sup>1</sup>, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, Marsolo K, McMurry AJ, Sandborg CI, Schanberg LE, Wallace CA, Warren RW, Weber GM, Mandl KD. *An i2b2-based, generalizable, open source, self-scaling chronic disease registry*. J Am Med Inform Assoc. 2013 Jan-Feb; 20(1): 172–179.
20. Shaw DM, Elger BS, Colledge F. *What is a biobank? Differing definitions among biobank stakeholders*. Clinical Genetics, Volume 85, Issue 3, pages 223–227, March 2014
21. Mackowiak PA<sup>1</sup>, Wasserman SS, Levine MM. *A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick*. JAMA. 1992 Sep 23-30; 268(12):1578-80.

22. U.S. Department of Health and Human Services. *Quality Improvement (QI) and the Importance of QI from U.S. Department of Health and Human Services*. HRSA Website.  
<http://www.hrsa.gov/quality/toolbox/methodology/qualityimprovement/>. Accessed June 23, 2014.
23. Jeremy Nix, Sr. Applications Developer, CCHMC. Interviewed by James Morgan, Mgr., CCHMC. 9/23/2014.
24. Parth Divekar, Sr. DBA, CCHMC. Interviewed by James Morgan, Mgr., CCHMC. 9/16/2014.
25. Todd Hoffert, Data Warehouse Analyst, CCHMC. Interviewed by James Morgan, Mgr., CCHMC. 9/23/2014.
26. Dave Campbell, Sr. DBA, CCHMC. Interviewed by James Morgan, Mgr., CCHMC. 9/9/2014.
27. Mike Carr, Sr. Systems Programmer, CCHMC. Interviewed by James Morgan, Mgr., CCHMC. 9/1/2014.
28. Cincinnati Children's Hospital Intranet Site. *CenterLink*.  
<http://centerlink.cchmc.org/templates/centerlink/2014/portal/home.aspx?pageid=130147>. Accessed 07/16/2014.
29. Cincinnati Children's Hospital Medical Center Intranet Site. *Help-BTM*.  
<https://btmhelp.cchmc.org/>. Accessed 7/20/2014.
30. Cincinnati Children's Hospital Medical Center Intranet Site. *Biomaterial Tracking & Management System, BTM-Research 5*.  
<https://research1.cchmc.org/btmresearch5/startup/Home.jsf>. Accessed 7/22/2014.

31. Daedalus Software, Inc. *BTM<sup>TM</sup> – Research Version 5.9.29 User Guide*. 2014 © Daedalus Software, Inc.
32. Benjamin P Rosenbaum, Nikolay Silkin, Randolph A Miller. *Easily configured real-time CPOE Pick Off Tool supporting focused clinical research and quality improvement*. J Am Med Inform Assoc. 2014 May-Jun; 21(3):564-8.
33. Mark Porcheret, Rhian Hughes, M Litt, Dai Evans, Kelvin Jordan, Tracy Whitehurst, Helen Ogden, Peter Croft, on behalf of the North Staffordshire General Practice Research Network. *Data Quality of General Practice Electronic Health Records: The Impact of a Program of Assessments, Feedback, and Training*. J Am Med Inform Assoc. 2004 Jan-Feb; 11(1):78-86.
34. Arts DG, De Keizer NF, Scheffer GJ. *Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework*. J Am Med Inform Assoc. 2002 Nov-Dec; 9(6):600-11.
35. Adeleke IT, Adekanye AO, Onawola KA, Okuku AG, Adefemi SA, Erinle SA, Shehu AA, Yahaya OE, Adebisi AA, James JA, AbdulGhaney OO, Ogundiran LM, Jibril AD, Atakere ME, Achinbee M, Abodunrin OA, Hassan MW. *Data quality assessment in healthcare: a 365-day chart review of inpatients' health records at a Nigerian tertiary hospital*. J Am Med Inform Assoc. 2012 Nov-Dec; 19(6):1039-42
36. Nicole Gray Weiskopf, Chunhua Weng. *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research*. J Am Med Inform Assoc. 2013 Jan 1; 20(1):144-51.
37. Peter B McGarvey, Sweta Ladwa, Mauricio Oberti, Anca Dana Dragomir, Erin K Hedlund, David Michael Tanenbaum, Baris E Suzek, Subha Madhavan. *Informatics and data quality at collaborative*

*multicenter Breast and Colon Cancer Family Registries*. J Am Med Inform Assoc. 2013 Jan 1; 20(1):144-51.

38. Nathan Coleman, Gayle Halas, William Peeler, Natalie Casaclang, Tyler Williamson, Alan Katz. *From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database*. BMC Fam Pract. 2015 Feb 5; 16(1):11.