ORNL/TM-2019/1302

Process Mining in Healthcare—A Case Study for the Corporate Data Warehouse of the Veterans Affairs Office

Hilda B. Klasky
Ozgur Ozmen

September 2019

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2019/1302

Data Driven Modeling and Analysis

PROCESS MINING IN HEALTHCARE—A CASE STUDY FOR THE CORPORATE DATA WAREHOUSE OF THE VETERANS' AFFAIRS OFFICE

Hilda B. Klasky
Ozgur Ozmen

September 2019

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The authors wish to thank the following people for their technical analysis, technical coordination, and/or feedback during all aspects of this project:

Jonathan R. Nebeker, MD, MS
Chief Informatics Patient Safety
Clinical Informatics and Data Management Office
US Department of Veterans Affairs
Washington, DC

Jeanie Scott, MS
Director, Informatics Patient Safety
Clinical Informatics and Data Management Office
US Department of Veterans Affairs
Washington, DC

Merry Ward, PhD
Research and Innovation Manager
Clinical Informatics and Data Management Office
US Department of Veterans Affairs
Washington, DC

Angela L. Laurio, DPH
Statistician
Informatics Patient Safety
Clinical Informatics and Data Management Office
US Department of Veterans Affairs
Washington, DC

Frank Drews, PhD
Cognitive Scientist
Clinical Informatics and Data Manaement Office
US Department of Veterans Affairs
Washington, DC

Makoto Jones, MD
VA Infectious Disease Physician
US Department of Veterans Affairs
Washington, DC

In addition, the authors wish to thank the extended ORNL team for their contributions to all aspects of this project, including key members of the team: Olufemi Omitaomu, Laura Pullum, Olama Mohammed, Mark Martin, Rajasekar Karthik and our PI, Teja Kuruganti. Also, we acknowledge support of staff in the Information Technology Division, as well as our managers and administrative support staff.

ABSTRACT

Researchers at Oak Ridge National Laboratory are working with the Veteran's Affairs Administration Office (VA) on a series of studies having the objective of detecting harmful anomalies (i.e., hazards) in VA health information technology systems using its Corporate Data Warehouse (CDW). This progress report describes an ORNL study focused on applying a process mining methodology to that CDW database. In the approach presented herein, process mining methodology is combined with additional software tools, metrics, and filters to permit a quick examination of large volumes of data to address specific research questions. In this work, a case study is presented in which the performance of the combined ORNL approach is evaluated by applying it to the CDW database. We performed an evidence-based study to effectively identify process models and to define metrics of frequency and performance for four health care domains: *Consults*, *Radiology*, *Laboratory*, and *Outpatient Medication* (*RxOut*) orders. Additionally, we classified the termination classes of the different cases by mapping to the OASIS[1] Human Task Specification standard. We demonstrated, via process mining, that extracted raw data can aid the understanding of the flow of information in different clinical order processes. We showed that using the step-by-step approach described herein to discover processes in raw electronic health record data can be extremely effective in revealing irregular state transitions in the data and understanding clinical order information flows that are not apparent in analyzing the CDW as it is, without feature extraction.

1. INTRODUCTION

Researchers at Oak Ridge National Laboratory (ORNL) are working with the US Veteran's Affairs Administration Office (VA) on a series of studies [1] with the objective of detecting harmful anomalies (i.e., hazards) in VA health information technology (HIT) systems using its Corporate Data Warehouse (CDW). The VA's CDW contains electronic health records (EHR) systems. (EHRS). An EHR is a digital version of a patient's paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. While an EHR does contain the medical and treatment histories of patients, an EHRS is built to go beyond standard clinical data collected in a provider's office and can be inclusive of a broader view of a patient's care.

During our study, we realized that the preparation and analysis of healthcare data are far more complex than similar processes in other fields. Some reasons for the complexity are that (1) the cost of creating datasets for mining is expensive, (2) data documentation is often missing, (3) updating of data documentation is neglected, and (4) there is a lack of subject matter experts to oversee the end-to-end process. We elaborate on these reasons in the following paragraphs.

The creation of datasets for EHR analysis is expensive for several reasons, including the following two: (1) Special security resources (humans, infrastructure, and IT) must be allocated to maintain the safety and integrity of the data. Particularly, in our case, special security resources such as infrastructure policies; training for employees having access to the data; secure working environments; and even physical working rooms, such as labs, and dedicated hardware and software equipment were allocated. (2) Much time is required to become familiar with the data and to identify the appropriate subset of the data for study and analysis. As this study uses data from the VA's CDW, we spent several months just getting familiar with the entity relationship diagrams and the database itself, and writing code to extract the data for our study. Thus, data preparation is an expensive effort.

[1] **OASIS** is a nonprofit consortium that drives the development, convergence and adoption of open standards for the global information society. https://www.oasis-open.org/org

Another issue with healthcare data analysis is that data documentation is often missing. In a large CDW such as the VA CDW, health care documentation related to process flow may not be as available as it is in other fields. We found that there was no available manual describing how data flows from the beginning to the end in each data domain. In fact, before we started our study, we did not even know what the beginning or the end of the process was. Consequently, a sequence of events needed to be developed as a trace to study, analyze [2], and discover process models and deviations from those process models. Those deviations to the process models signal possible anomalies that could become HIT hazards. In addition, deviations to the process models may show other areas in the process that cause performance bottle necks and need improvement.

Neglect of updating of healthcare data documentation is another challenge to analyzing the data. Moreover, we found that the documentation available for the healthcare data domain process was not up to date. As database changes are common in HIT, fields sometimes are deprecated and no longer used; and new data may be included only as agreed upon by the different organizations that participate in the CDW. Documentation is generally the last item to be updated, and communicating changes to all those who would be affected is difficult.

Finally, finding a subject matter expert (SME) who can oversee the end-to-end process in a healthcare data domain can be difficult. Often several data systems converge into a larger repository. Finding a specialist with an understanding of a specific aspect of a data domain is not difficult. And during this study, we were lucky to receive the valuable support of SMEs. But these individuals had knowledge of only one small part of the entire healthcare system. With no one available to guide us through the entire process, we had to investigate the path that most orders take for the selected data domains by creating sequences of events from data available in the CDW database.

In recognition of the issues presented above, we developed an approach to apply process mining and to generate event sequences from the CDW. In the absence of well-defined event logs in the VA CDW during the study, we needed to extract relevant information scattered among numerous tables. We applied a feature engineering procedure to extract all relevant dates and status update records of clinical provider orders from a CDW that stores EHRs. During the data extraction, preparation, and analysis, we created event logs from the healthcare data in CDW. We also applied a modified version of a process mining methodology developed by van Eck, Mans, and van der Aalst [3-5].

This progress report describes the advances achieved by researchers at ORNL in applying a process mining methodology to the data in the VA CDW. The objectives of that effort are the following:

- Apply process mining to the following data domains in CDW: Consults, Radiology, Labs, and RxOut, and their correspondent data in the CPRSOrders domain.
- Identify process model maps.
- Identify metrics of frequency and performance.
- Map the OASIS Human Task State Transition Diagram to the process models maps to identify cases that complete successfully and those that do not complete successfully.

Our hope is that understanding model cases and unsuccessful cases will aid in eventually developing procedures for identifying harmful anomalies (i.e., hazards) in VA HIT systems.

Section 2 of this study presents related work. Section 3 describes the data and methods. Section 4 presents a case study results designed to assess the performance of the combined Oak Ridge National Laboratory

(ORNL) approach through an application to the CDW database. Section 5 summarizes the work progress achieved in developing the step-by-step ORNL approach, presenting a discussion along with the limitations, lessons learned and conclusions. Finally, Section 6 outlines recommendations for future work.

2. RELATED WORK

The application of process mining to healthcare data is largely described in the book *Process Mining in Healthcare* by Mans et al. [5], which aims to explain how process mining techniques can be used in a threefold manner: (1) to improve processes, (2) to analyze process conformance, and (3) to reduce costs. Rojas et al. [6] present a literature review of process mining in health care.

Lowry et al. [7] presented a case study of integrating EHR into a clinical workflow. They developed recommendations for EHR developers and ambulatory (outpatient) care centers to improve workflow integration with EHRs to increase efficiency, allow for better eye contact between physician and patient, improve physicians' information workflow, and reduce alert fatigue. Although Lowry et al. is not a process mining study, we found the outline of the several clinical workflows extremely helpful.

Specifically, in published studies of health care data, process mining has been applied successfully to identify and improve the understanding of workflows in health care [5, 8-13].

Mans et al. [8] applied process mining techniques to obtain meaningful knowledge about healthcare workflows to discover typical paths followed by particular groups of patients. They found this to be is a "non-trivial task given the dynamic nature of healthcare processes," which is consistent with our observations during our study. Their paper [8] demonstrates the applicability of process mining using a real case of a gynecological oncology process in a Dutch hospital. Using a variety of process mining techniques, they analyzed the healthcare process from "three different perspectives: (1) the control flow perspective, (2) the organizational perspective and (3) the performance perspective." To do so, they extracted relevant event logs from the hospital's information system and analyzed them using the ProM framework,[2] a tool that integrates the main stages of process analysis. Their results showed "that process mining can be used to provide new insights that facilitate the improvement of existing careflows."[8]

A study by Mans et al. [9] described the different types of event data found in current hospital information systems, performed a classification based on that available data, and discussed open problems that need to be solved to increase the uptake of process mining in healthcare. Their findings are consistent with our observations of the problems we found, as described in the introduction of this report.

A case study by Rovani et al. [10] showed how process mining techniques can be used to mediate between event data reflecting the clinical reality and clinical guidelines describing best practices in medicine. They used declarative models that allow for increased flexibility and thus are more suitable for describing healthcare processes that are highly unpredictable and unstable; our study also encountered unpredictability and instability. Their techniques were applied in the urology department of the Isala hospital in the Netherlands, [10] and their results demonstrated that the techniques based on ProM and the Declare tool are able to provide valuable insights related to process conformance.

Rebuge et al. [11] introduced a methodology for applying process mining techniques that leads to the identification of regular behavior, process variants, and exceptional medical cases. Their approach [11] is demonstrated in a case study conducted at a hospital emergency service for radiology. The methodology was implemented in ProM. In their approach, sequence clustering plays a key role in identifying regular

[2] http://www.promtools.org/

behavior, process variants, and infrequent behaviors using a cluster diagram and a minimum spanning tree, which provide a systematic way to analyze the results.

A study by Keymak et al. [12] assessed the applicability of process mining methods to the discovery of clinical process models for a well-defined medical procedure of moderate complexity and extent. They studied the anesthesia procedure during endoscopic retrograde dholangiopancreatography.

Work by [13] Gupta et al. used association rules in combination with the clustering technique to generate process models specific to a group of patients sharing some similar characteristics. They used data from about 20,000 patients, obtained from the Catharina Hospital, Eindhoven, to describe the various activities that took place in the Catharina intensive care unit (ICU), including information about the control-flow and the personnel (e.g., doctors, specialists) performing several clinical tasks. They applied the Heuristics Miner (HM) algorithm on real-life logs in the form of healthcare data. They discovered that such data are not suitable for mining flexible and less structured processes and also that the HM is unable to deal with unclear AND/XOR information because the dependency graph generated is inappropriate for mining process models for less structured processes. Consequently, they proposed to apply the concept of association rules to the healthcare domain to gain insights into dynamic and flexible healthcare processes; and, in combination with clustering, they enabled the partitioning of an event log into homogeneous groups of patients. They implemented their algorithms in a ProM framework: the Apriori and the PredictiveApriori in a plug-in. This plug-in not only provided insights into the process underlying an event log through association rules but also gave the option to cluster the event log based on the criteria of association rules and frequent item sets.

To our knowledge, at the time of writing this study, there were no published studies of applying process mining to the VA CDW data. The main contributions of our work to the problem we are exploring are the following:

1. Developed a customized process mining approach to apply process mining to VA CDW data.

2. Demonstrated the feasibility of applying process mining to the following data domains in CDW: Consults, Radiology, Lab Services and RxOut, and their related data in the CPRSOrders domain using large datasets.

3. Identified process model maps for to the data in item 2.

4. Outlined metrics of frequency and performance for the process maps in item 3.

5. Developed an approach to mapping the OASIS Human Task State Transition Diagram to the process models maps to identify cases that complete successfully and those that do not complete successfully.

6. Generated activities sequences by data domain rather than by clinical pathway or organization department.

In Section 3, we describe the data employed in this study and the implementation approach.

3. APPROACH

3.1 PROCESS MINING

Process Mining is a novel field of knowledge that comprises many algorithms to identify process models from information technology data [3-5, 20].

We investigated different engineering methodologies used in other studies, or combinations thereof, for possible applications to our healthcare data study [4-7, 14-19]. This study is based on the methodology and work of van Eck, Mans, and van der Aalst [3-5, 20] with some modifications made to suit our study. The following are the steps for process mining:

1. Data collection and preparation
2. Process discovery (what happened)
3. Process monitoring (what else is happening)
4. Process analysis (why it happened)
5. Repetition of steps 1 to 4 as needed

Datasets were refined with each iteration by identifying (1) new potential anomalies and areas of improvement in our own analysis process and (2) potential anomalies in the data, which were reported and discussed with the SMEs at the VA.

Data collection and preparation methods are explained in detail in Ozman et al. [2]. However, we used a slightly modified approach following the methodology presented in van Eck et al. [3], *Process Mining in Healthcare* [4], and, and Rozinat et al. [21], as is discussed later. The SMEs, the project team, and the extended VA team participated in data analysis, discussions, and independent self-study.

During the fourth step of the process mining, i.e., Process Analysis, Ozman et al. suggest the following.

- Observe processes as rendered by the event sequences.
- Identify paths most events follow.
- Identify the process deviations.
- Identify root cause of deviations.
- Suggest potential process improvements.

In the following sections, we explain how we applied this methodology. Finally, we present a case study as an example, along with the metrics obtained from our analysis.

3.2 DATA EXTRACTION AND PREPARATION

The analysis was focused on a cohort (approximately 808,000 patients) who were diagnosed with ischemic heart disease (IHD) between January 1, 2017, and January 1, 2018. This cohort was based on VA CDW data. Thus, we extracted cohort data from four health care domains: *Consults*, *Radiology*, *Laboratory Services (Labs)*, and *RxOut* or *Prescription* medication orders. We also included the CPRSOrder (CPRS stands for Computerized Patient Record System) domain date columns for records associated with packages in the four health care domains. The CPRSOrder domain is basically the provider clinical orders system that communicates with the VA's Vista system. Because of the abundance of data, it was necessary to reduce the datasets. Consequently, the Consults, Labs, and RxOut orders datasets had 6 months of data. However, the Radiology orders dataset included 12 months of data, as there were found to be fewer Radiology orders than Consults, Labs, and RxOut orders.

Our study frames CDW data from the clinical orders perspective because a clinical order is identified as an atomic process in the overall health care delivery. Therefore, our approach starts with the identification of activities, status updates, and transaction dates in clinical order processes throughout their data domains. We understand an activity as a pursuit, an undertaking, something being done.

We constructed sequence of activities for each data domain using a three-column table format: Case ID, Activity, Date, as shown in Figure 1. The data was saved in comma-separated value (csv) files. An

example event sequence or case is shown in Figure 1, which shows three columns of data separated by commas. The first column is the case ID, which in this case is 8314090, a de-identified sequence number. The second column is the activity name, as given by the column name or the event type from the database. The third column is the time stamp, which uses the following format: YYY-MM-DD HH:MM:SS. Each case is sorted by timestamp and activity in ascending order.

8314090,DesiredNotGuaranteedDateTime,2017-04-06 00:00:00
8314090,ReportedDateTime,2017-04-06 00:00:00
8314090,RequestedDateTime,2017-04-06 00:00:00
8314090,OrderStartDateTime,2017-04-06 00:00:00
8314090,EnteredDateTime,2017-04-06 15:37:00

Figure 1. Case sample of radiology orders event log fragment.

Notice in Figure 1 that some rows have only zeros in the time part of the timestamp. It is not certain whether this timestamp results from the use of a legacy or newer system, or rather from the actual definition of the date field as being a short date (i.e., day only) vs. a long date (i.e. day +time). The missing time component indicates the event occurred early in the day, before other events, when in fact it may have occurred later in the sequence. In other cases, several activities are recorded as if they happened at the same time, in clusters. Consequently, because the data did not provide a complete understanding of the data systems that feed the CDW, we examined the data and the sequences of events on a case-by-case bases.

3.3 OASIS HUMAN TASK STATE TRANSITION DIAGRAM

To further simplify the identification of successful and unsuccessful cases, we used the OASIS human task state transition diagram to map the different activities and date columns to the OASIS human task state transitions. This mapping facilitated (1) the addition of another layer of abstraction over raw details, (2) identification of successful and unsuccessful process paths (and the rules or common behaviors), and (3) identification of outliers and conformance to the process model map. In addition, it helped to define the rules for each data domain for successful vs. unsuccessful paths, according to our interpretation of unsuccessful termination states types in OASIS and how the interpretation applies to the datasets. More information about OASIS can be found in the OASIS specification [22] and in Appendix A of this report.

In our experience, it is not a trivial task to associate the different events to the state transition because, although activities happen in clusters that have the same timestamp in the log files, sorting by activity name and timestamp was not enough to clearly identify and differentiate the created, ready and reserved states in the sequences of events. Based in our observations, those three initial states occurred at the same time in most of the cases, in all the data domains studied. This complexity was exacerbated by the possibility of having several cases for the same state, for example when a state goes to suspended.

3.4 ANALYSIS PROCESS

The algorithm used to analyze each dataset in this study is as follows:

1. Identify events from the date columns and columns that contain states type, status type, and activities type data from the entity relationship diagrams of the data domain.

2. Generate datasets for each domain by writing SQL queries to extract the data.

3. Map OASIS state transition to the events in step 1.

4. Generate process maps for both raw data and OASIS mapping.

5. Identify main process path.

6. Identify key performance indicators (KPIs)*.

7. Identify outliers (data, duration, paths, loops).

8. Identify high-impact areas of performance.

9. Identify cases that go to unsuccessful termination states.

Note: Based on ORNL experience gained in applying this methodology, the identification of KPIs suggested by the methodology in *Process Mining in Healthcare* [4] was polemic from the beginning. Even though the use of KPIs is a well-known practice in business process improvement [23], we decided to exclude that step from the current study and to focus on informing results from the analysis.

3.5 TOOLS

The software tool Disco [21], from www.fluxicon.com/disco, was used for process mining and to visualize and filter the datasets. It proved to be extremely helpful, fast, and reliable. Disco can generate not only high-quality reports and images but also animations of the datasets. In addition, Disco provides output in CSV, FXL, XES[24] (ProM 6 [25]) and MXML (ProM 7) file formats. This tool provided useful information in a timely manner for the large volume of data we needed to identify the data flow of state transitions that would lead to more specific research questions. Visualizing the data in this way has been helpful in our search of HIT anomalies. Without a robust process visualization tool, we would have been unable to observe as many details in the many terabytes of data. With Disco, we could observe the paths and the flow of most sequences in each dataset. Having the capability of zooming in and out is beneficial as well in process maps and high-quality images. Examining event sequences with a powerful visualization tool provided additional insight that would have been hard to achieve otherwise.

3.6 METRICS AND FILTERS

In this section presents definitions of the metrics performed, defines the types of filtering applied to the datasets and discusses the different descriptive statistics results tables.

3.6.1 Definitions of Metrics Performed

From the process maps generated with Disco, we obtained two mayor types of metrics: (1) frequency metrics and (2) performance metrics.

Frequency metrics are absolute counts of how many times an activity was executed. These metrics can be seen in the process maps by following the arrows, in the summary tables in the next section, and in the tables in Appendix B that show frequencies and relative frequencies for each activity in the different datasets. Other frequency metrics for activities are total number of activities, minimal frequency, median frequency, mean frequency, maximal frequency, and frequency standard deviation.

* A key performance indicator (KPI) is a list of events and its associated metrics that we suggest tracking as plans and process improvements are implemented.

Performance metrics are time durations between activities, or durations for the whole process map. The following are other performance metrics used in this document and their definitions:

- Case duration: The time of the process from the start or first activity to the end, i.e., to the last activity's completion.

- Mean: The average (summing up all numbers and then dividing by the number of entries). The mean duration is the average time (arithmetic mean) spent within and between activities.

- Median duration: The value in the middle in the list of numbers (see "*median duration*" in Rozinat et al. from [21]). It is the median time spent within and between activities. In many situations, the median (also known as the 50th percentile) provides a much better idea of the typical performance characteristics of a process than the arithmetic mean, especially for datasets that contain extreme outliers.

- Mode: The most common value, i.e., the value that occurs most often.

- Standard deviation: Disco calculates the standard deviation via the Apache Commons math library[3] (see Equation 1]). This is based on the sample variance (see Equation 2]) (not the population variance).[4]

Equation 1. Standard deviation formula based on sample variance used in this study.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Equation 2. Sample variance formula used in this study.

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{N - 1}$$

The following section presents the results of these duration metrics.

3.6.2 Filtering

A very important part of process mining is filtering. Filtering is the act of selecting those paths, cases, or attributes that help answer questions. We observed that, before filtering, our datasets contained cases that began/completed in the selected timeframe or observation window. We considered those to be *complete cases*. Specifically, that term refers to the cases that contain start and end events in the OASIS Human Task State Transition mapping, as shown in Appendix C. We observed that the datasets also contained cases that intersected the timeframe, i.e., cases that started in the timeframe but did not end in the timeframe and cases that ended in the timeframe but did not start in the timeframe. To accurately calculate the statistics, we excluded the cases that did not start/end within the timeframe, as they were considered incomplete cases. Consequently, datasets were filtered to include only the most common paths and complete cases that started and ended within the period from January 1, 2017, to June 30, 2019.

[3] See https://commons.apache.org/proper/commons-math/javadocs/api-3.3/org/apache/commons/math3/stat/descriptive/SummaryStatistics.html#getStandardDeviation().

[4] Further information about sample and population variance and standard deviation can be found at https://en.wikibooks.org/wiki/Statistics/Summary/Variance.

Datasets were also filtered using a time warp to include duration statistics based on events taking place during regular business hours and excluding US federal holidays and weekends. Another reason to apply filters to our datasets was the problem called *zero timestamps* [21], i.e., faulty timestamps that are far in the past (1900) or far in the future (2100). Those dates in the data are problematic because they affect the study and analysis in general and, in particular, they impact case durations, variants, and process maps [21].

4. CASE STUDY AND RESULTS

Section 4 describes and deliberates in the study results and analysis per dataset, the summary descriptive statistics, some cases that do not conform to the process model case, and outliers to the duration mean times. We applied process mining to the VA CDW data domains Consults, Radiology, Labs, and RxOut and their related CPRS clinical orders.

4.1 DESCRIPTIVE STATISTICS

Descriptive statistics of the study and overview process mining metrics are found in Table 1, Table 2, Table 3, and Table 4.

Table 1 presents a summary of the four datasets before filtering. In the first column, c*ases* are the sequences of events in each data domain. The v*ariant* row refers to a distinct sequence of events. An *activity* is each condition in which things are happening or being done. An *event* is an occurrence of an *activity* in a given *time*. A set of activities is identified for each data domain. Most activities are date columns in tables of each domain. However, we were able to identify and include data from columns that save values that trigger changes in the dates. Finally, the *case duration* is the time span from the recorded time of the first event to the recorded time of the last event, as found in the database tables. Notice that in Table 1, the number of variants was reduced considerably when the event sequences were sorted by date and event in a subsequent recreation of the event sequences.

Table 1. Descriptive statistics summary of datasets before filtering.

	Consults	Radiology	Lab	RxOut (prescriptions)
Cases	2,204,522	865,021	9’138,290	5,406,800
Variants sorting by date only	1,990,146	851,627	1,214,792	4,508,774
Variants sorting by date and event	561,660	335,190	228,022	1,428,332
Events	51,179,519	18,536,242	104,811,071	130,044,618
Distinct activities	44	117	36	57
Events per variant	Min: 12 Max: 715	Min: 7 Max: 87	Min: 5 Max: 335	Min: 5 Max: 136
Median case duration	~12.9 days	11.2 days	~71.3 days	~51.8 weeks
Mean case duration	46.5 days	19.7 weeks	23 weeks	48.3 weeks

Table 2 presents a summary of the datasets after applying filters to identify the model cases per data domain. The filters were applied to include only complete cases: those that started and ended on events as mapped in the OASIS–WS human task state transition in Appendix A. Notice that the 80–20 rule applied

for the Consults and Radiology datasets: about 80% of the events followed similar patterns; however, Labs and RxOut data results were not as smooth. Consequently, we needed further study to properly analyze the Labs and RxOut datasets.

Table 2. Descriptive statistics summary of datasets after filtering.

	Consults	Radiology	Labs	RxOut
Events	38,916,948 76% of the events	14,762,464 79% of the events	59,853,781 57% of the events	18,033,671, 13% of the events
Cases	1,778,369 80% of the cases	656,176 75%of the cases	4,888,169 53% of the cases	933,001 only 17% of the cases
Activities	44	114	34	52
Mean case duration	4.9 days	4.3 days	4 days	13.9 days
Median case duration	~41.5 hours	18.8 hours	11.6 hours	9.6 days

After mapping the OASIS WS human task state transitions to the datasets, we obtained the data in Table 3, which presents the termination states breakdown per dataset. Notice that there is no error state defined for Consults. Also, notice that Consults, Radiology, and a fragment of one million samples of Labs data completed successfully (close to ~80%). In addition, the percentage of failed cases remained very low. The RxOut dataset was not included in this table because of its complexity and the time needed to implement this task.

Table 3. Dataset cases termination states breakdown.

% Cases	Consults	Radiology	Labs (1 Million cases sample)[5]	RxOut
Completed	1,800,138 (80%)	657,189 (76%)	740,874 (74%)	TBD
Exited	445,046 (19%)	189,434 (22%)	75,695 (7.5%)	TBD
Error	--	16,551 (~2%)	72,166 (7.2%)	TBD
Failed	20,019 (~1%)	96 (~0.01%)	218 (0.02%)	TBD

The frequencies of the types of activities included in the datasets are presented in Table 4.

Table 4. Types of activities included in datasets.

	Consults	Radiology	Labs	RxOut
Dates	20	36	34	34
Order action	3	5	5	5
Status updates	21	76	16	18

[5] The remining cases are not *complete* cases, as this sample dataset was not filtered.

Total	44	117	55	57

4.2 CONSULTS

Our clinical Consults dataset was generated from the CDW Consults data domain. The types of activities included 20 dates, 3 order actions, and 21 status updates for a total of 44 distinct activities (see Table 4). The complete process map for Consults, before filtering, can be found in Figure 2. It includes all cases and activities. Because this map is extremely complex and impossible to analyze as is, process mining was applied to understand patterns and identify any possible anomalies.

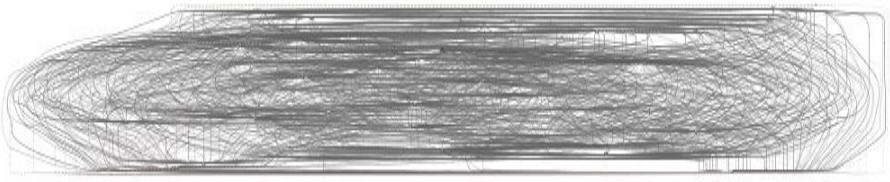

Figure 2. Complete consults process map before filtering.

4.2.1 Process Model Map

The raw process map for Consults consists of ~51 million events, ~2.2 million cases, 561,660 case variants, and 44 activities (see Table 1). The process model map for the Consults dataset is shown in Figure 3. In Figure 3, the path that most cases take can be seen by following the darker blue boxes and the bold and thick arrows. The numbers near boxes and arrows are the absolute frequencies, i.e., how many times an activity was performed in total.

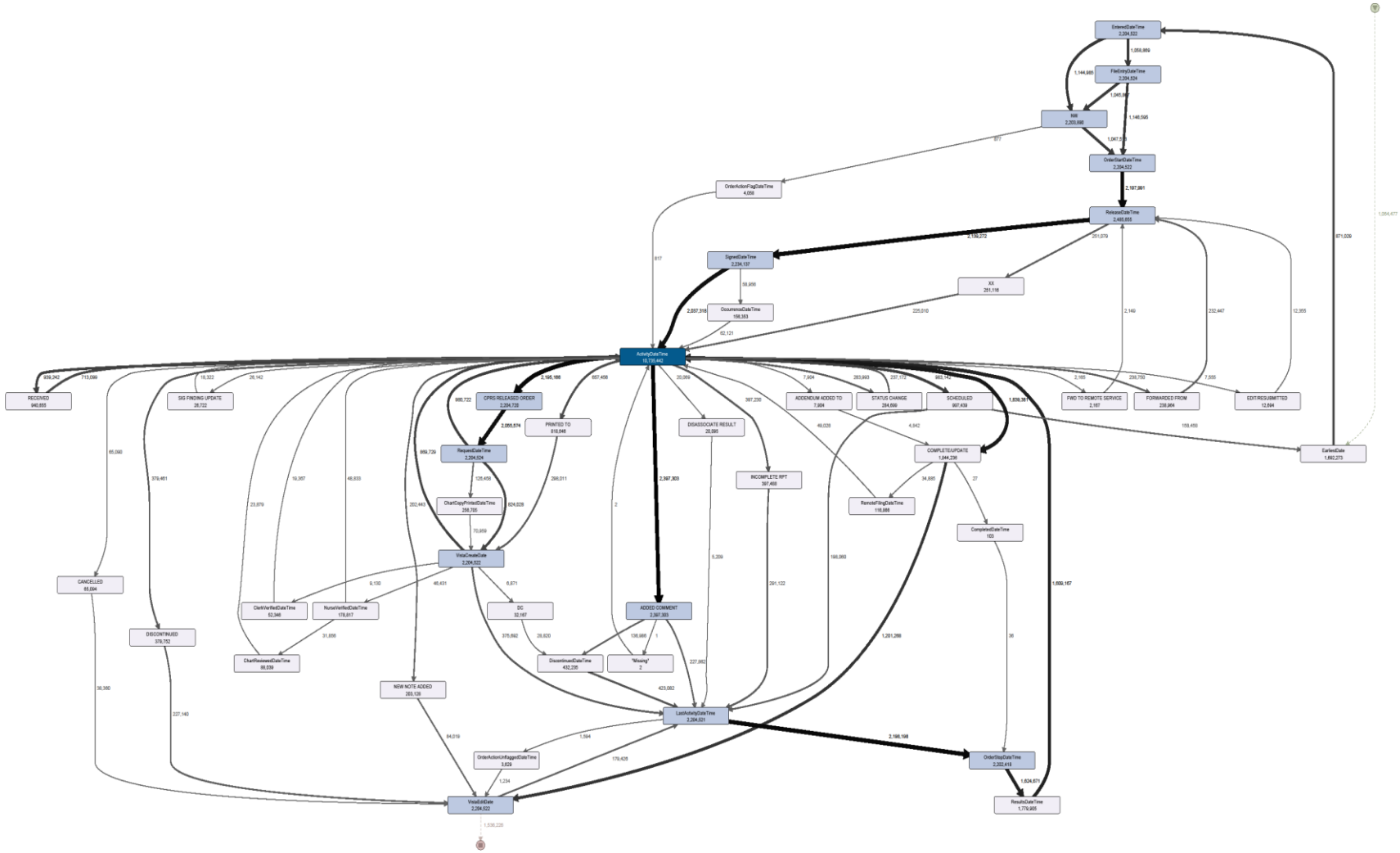

Figure 3. Consults process model map before filtering.

The process map Figure 3, in is consistent with an earlier frequency process map with fewer activities shown in Figure 4. In Figure 4 the consult is the first set. It occurs, and then, is closed as a three-part process from top to bottom.

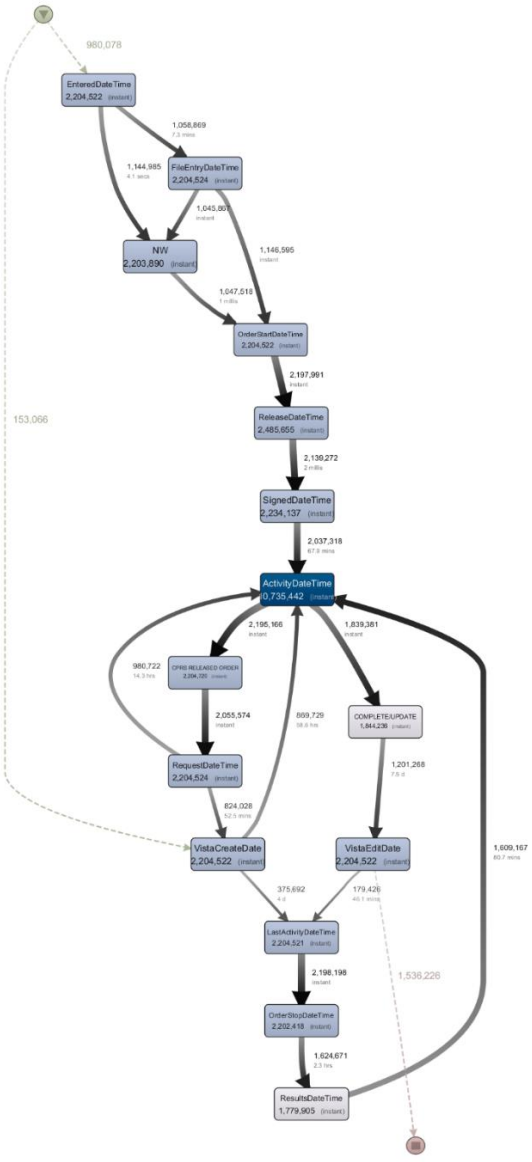

Figure 4. Consults most frequent path.

4.2.2 Frequency Metrics

The list of activities in the Consults dataset is found in Table 5 which also presents the frequencies, and the relative frequency percentage.

Table 5. Consult dataset activities, frequency and relative frequency.

Activity	Frequency	Relative frequency

ActivityDateTime	10,735,442	20.98%
ReleaseDateTime	2,485,655	4.86%
ADDED COMMENT	2,397,303	4.68%
SignedDateTime	2,234,137	4.37%
CPRS RELEASED ORDER	2,204,720	4.31%
FileEntryDateTime	2,204,524	4.31%
RequestDateTime	2,204,524	4.31%
EnteredDateTime	2,204,522	4.31%
OrderStartDateTime	2,204,522	4.31%
VistaCreateDate	2,204,522	4.31%
VistaEditDate	2,204,522	4.31%
LastActivityDateTime	2,204,521	4.31%
NW	2,203,890	4.31%
OrderStopDateTime	2,202,418	4.30%
COMPLETE/UPDATE	1,844,236	3.60%
ResultsDateTime	1,779,905	3.48%
EarliestDate	1,692,273	3.31%
SCHEDULED	997,439	1.95%
RECEIVED	940,655	1.84%
PRINTED TO	818,646	1.60%
DiscontinuedDateTime	432,235	0.84%
INCOMPLETE RPT	397,488	0.78%
DISCONTINUED	379,752	0.74%
STATUS CHANGE	284,699	0.56%
ChartCopyPrintedDateTime	256,705	0.50%
XX	251,116	0.49%
FORWARDED FROM	238,964	0.47%
NEW NOTE ADDED	203,128	0.40%
NurseVerifiedDateTime	178,817	0.35%
OccurrenceDateTime	156,353	0.31%
RemoteFilingDateTime	116,866	0.23%
ChartReviewedDateTime	88,039	0.17%
CANCELLED	65,094	0.13%
ClerkVerifiedDateTime	52,346	0.10%
DC	32,167	0.06%
SIG FINDING UPDATE	26,722	0.05%
DISASSOCIATE RESULT	20,095	0.04%
EDIT/RESUBMITTED	12,694	0.02%
ADDENDUM ADDED TO	7,904	0.02%
OrderActionFlagDateTime	4,058	0.01%
OrderActionUnflaggedDateTime	3,629	0.01%
FWD TO REMOTE SERVICE	2,167	0%
CompletedDateTime	103	0%

Missing	2	0%

Figure 5 presents a Pareto chart of frequency by activities for Consults. There are 44 total activities in Consults. The minimal frequency is 1 and the median frequency is 303,379. The mean frequency is 884,476.09. The maximal frequency is 7,540,480 and the frequency standard deviation is 1,282,429.63.

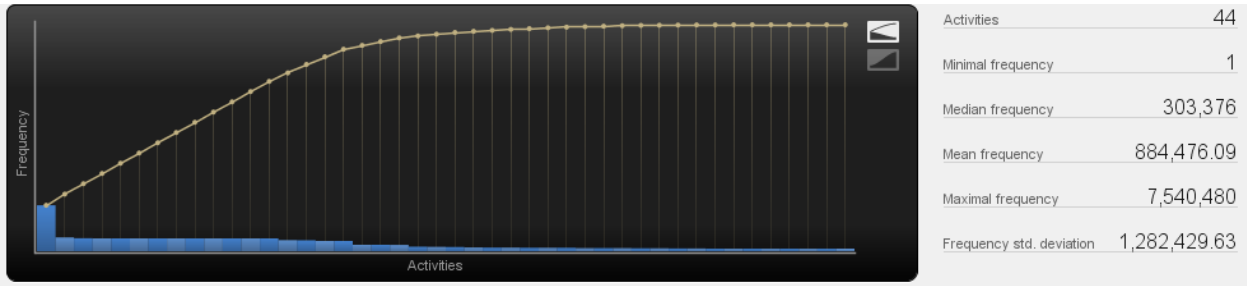

Figure 5. Consults Pareto chart—frequency by activities.

In addition, Figure 6 shows Consult frequency metrics after filtering.

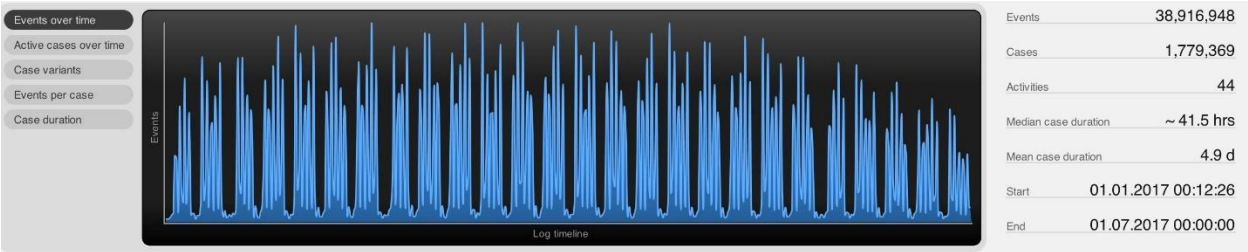

Figure 6. Consults stats—after filtering.

4.2.3 Performance Metrics

We applied filters to the Consults dataset to identify the duration of most common path and other duration statistics, as shown in Figure 7. Notice that the process map in Figure 7 includes all activities.

After filtering, the duration of the most common path can be found in Table 2, where we can see that the median case duration for Consults is ~41.5 hours and the mean case duration is 4.9 days. Information in Table 2, Figure 6 and Figure 7 are taken as a model case and as a base for identifying outliers and cases that do not conform with the model case, in order to identify possible anomalies.

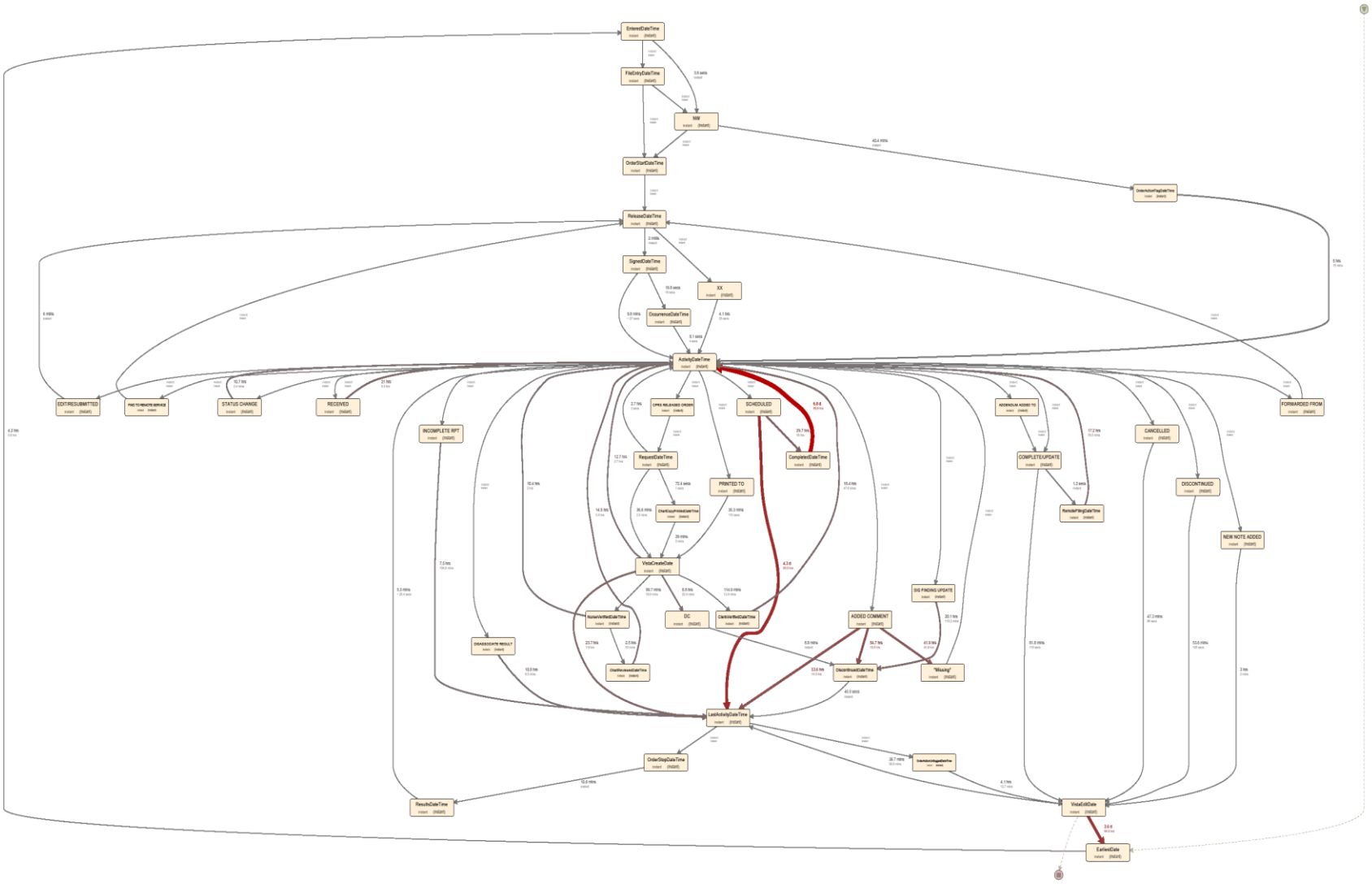

Figure 7. Consults performance map after filtering.

4.2.4 Mapping to OASIS Human Task State Transition and Termination States for Consults

Table 6 presents the association of the different Consult activity names to the OASIS human task state transitions.

Table 6. Consults dataset mapping to OASIS Human Task State Transition.

OASIS Human Task State Transitions	Consult date columns	Consults activities and other events
CREATED	EarliestDate **EnteredDateTime**	
READY	FileEntryDatetime OrderStartDateTime ActivityDateTime ReleaseDateTime	**NW**
RESERVED	SignedDateTime RequestDateTime VistaCreateDate [PatientAppointmentDateTime]	 **RECEIVED** CPRS RELEASED ORDER
IN_PROGRESS	ActivityDateTime ChartCopyPrintedDateTime OccurrenceDateTime LapsedDateTime VistaEditDate OrderActionDateTime NurseVerifiedDateTime ClerkVerifiedDateTime ChartReviewedDateTime DescontinuedHoldUntilDateTime OrderActionFlagDateTime OrderActionUnflaggedDateTime InterventionDate PharmacistCommentsDateTime ActivityEntryDateTime RemoteFilingDateTime	**SCHEDULED** PRINTED TO ADDED COMMENT INCOMPLETE RPT NEW NOTE ADDED FORWARDED FROM XX STATUS CHANGE SIG FINDING UPDATE EDIT/RESUBMITTED ADDENDUM ADDED TO DISASSOCIATED RESULT FWD TO REMOTE SERVICE ADMIN. CORRECTION DC Missing
COMPLETED	 LastActivityDateTime OrderStopDateTime CompletedDateTime **ResultsDateTime**	COMPLETE/UPDATE
SUSPENDED		WAITING STATES with time interval!
FAILED	Completed and there is Discontinued Date Time recorded,	DISASSOCIATE RESULT
ERROR	DiscontinuedDateTime	DISCONTINUED

OASIS Human Task State Transitions	Consult date columns	Consults activities and other events
		CANCELED without package number
EXITED	DiscontinuedDateTime (else: not at Completed or Ready)	DISCONTINUED CANCELED with package number
CLOSED	VistaEditDate	

The process map with OASIS mapping for Consults is found in Figure 8. As shown, most cases completed successfully; however, ~445,000 cases, or 19% of the cases, were in the error state *excited*, and about 20,000 were in the *failed* state, which was about 1% of the Consults dataset cases in a 6-month period.

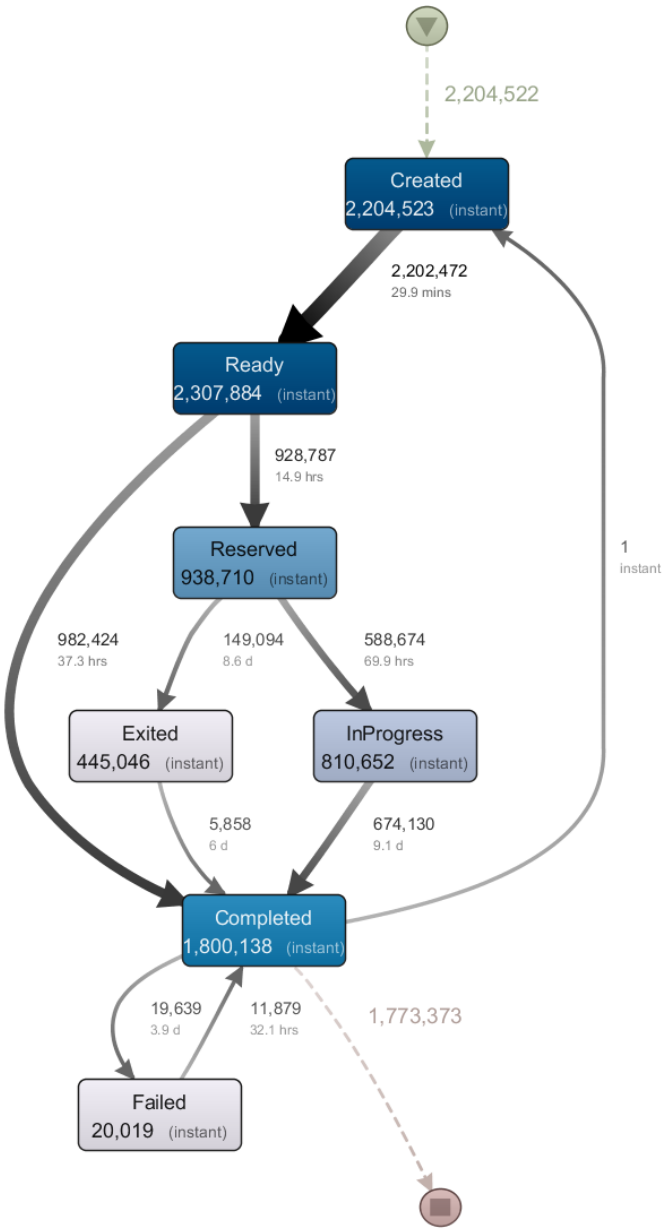

Figure 8. Consults dataset mapped to OASIS Human Task state transitions.

We found that the mapping to the OASIS Human Task State Transition diagram was helpful in simplifying the Consults most common path, as shown in Figure 3 and in Figure 8. For example, identifying *excited* and *failed* cases is difficult in the process map shown in Figure 3 and in Figure 4, but those cases are easily recognizable in Figure 8. Thus, OASIS mapping allowed us to quickly visualize how many cases go to error states. In the Consults case, 80% of the paths were *completed* successfully, while 19% *exited* and 1% *failed*, as shown in Table 3. Figure 9 presents the Consults process model map with state transitions.

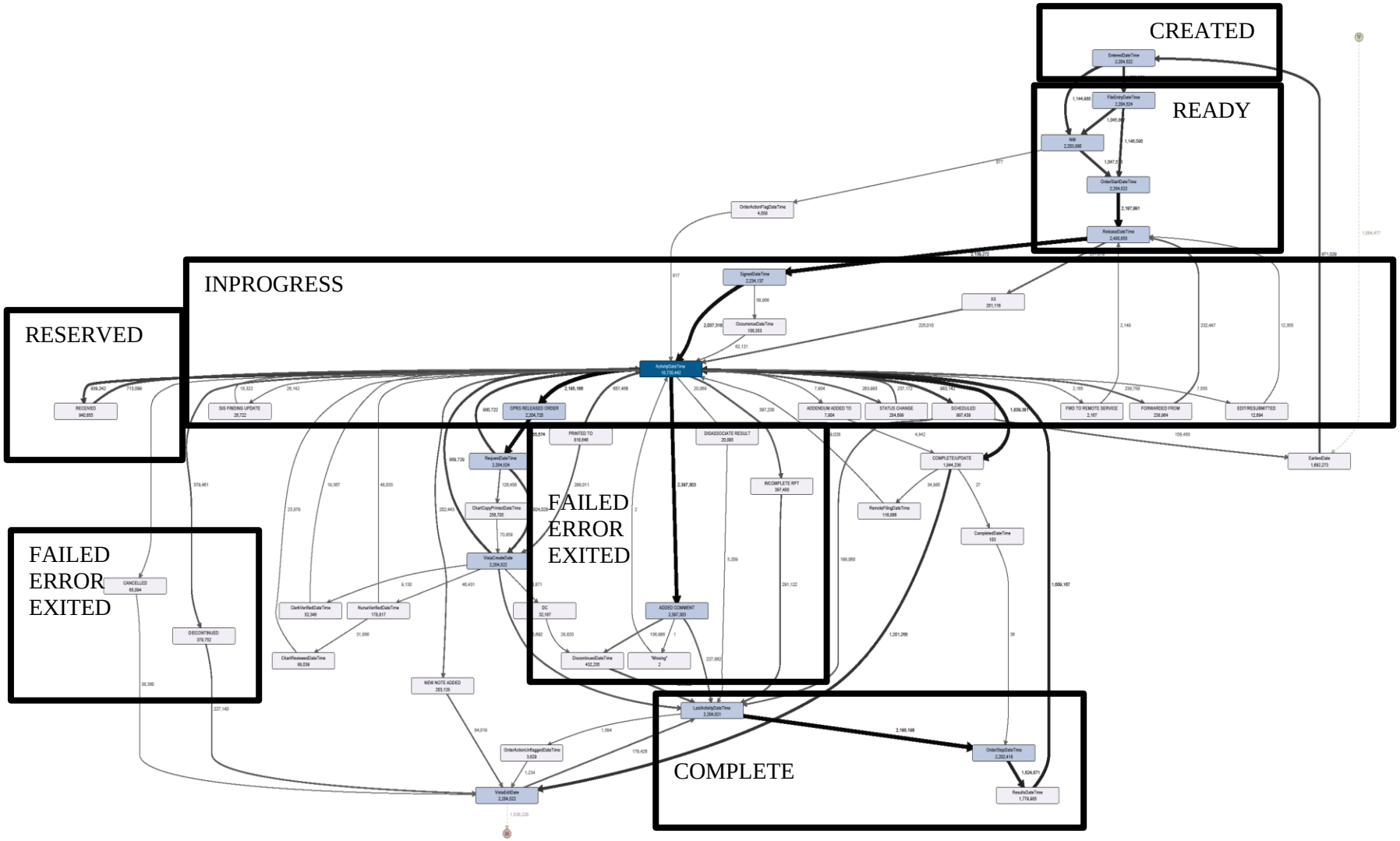

Figure 9. Consults Process Model Map presenting state transitions between activities.

A zoom into the *exited* and *failed* cases process map for Consults is shown in Figure 10. Noticed that *failed* cases follow the *complete* state. However, cases that end in the *exited* termination state can happen after *ready*, *reserved* or *inprogress* states. Also notice that some cases can go from *exited* to *ready*, these cases need further study.

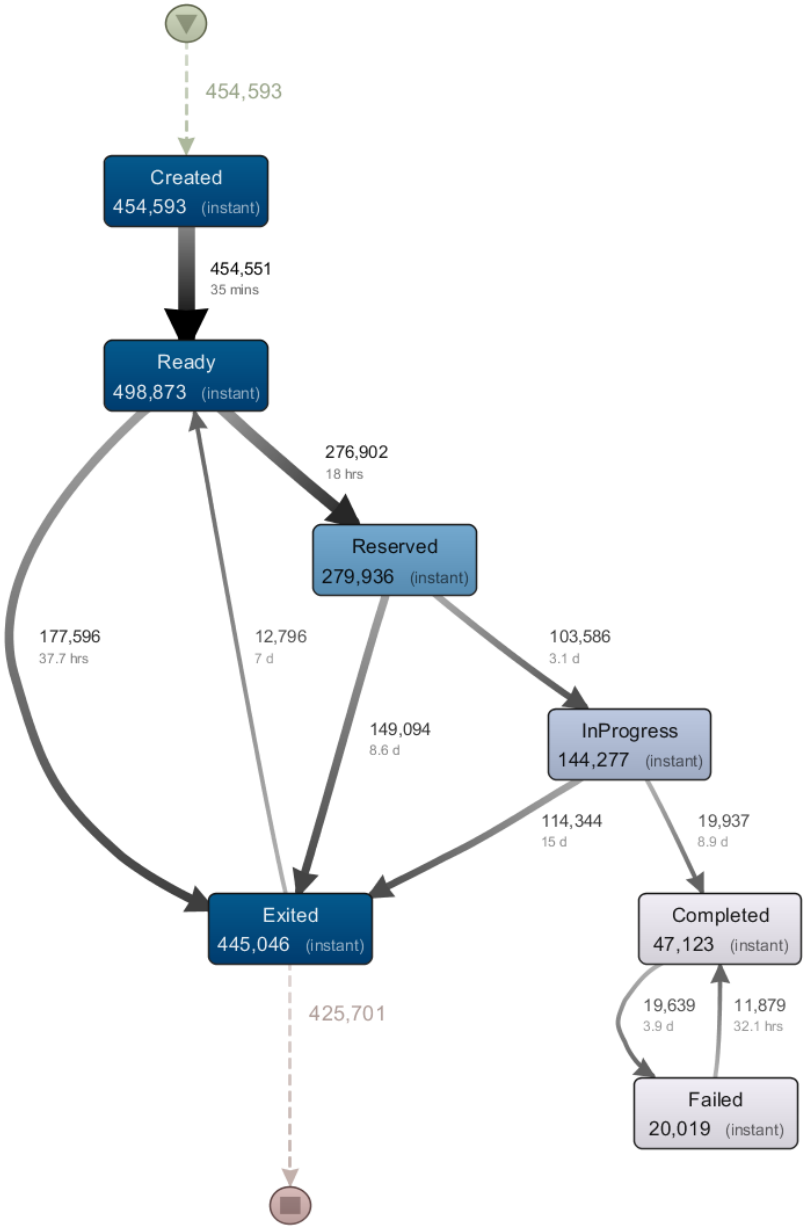

Figure 10. Exited and failed cases in Consults.

A sample failed case is shown in Table 7. Notice the Activity INCOMPLETE RPT on row 18 and the DISASSOCIATE RESULT on row 20 that signal that something went wrong and consequently this order fails.

Table 7. Consults sample failed case - fragment (16525684).

	Activity	Date	Time
1	EarliestDate	07.04.2017	00:00:00
2	EnteredDateTime	07.04.2017	15:26:00
3	FileEntryDateTime	07.04.2017	15:26:00
4	NW	07.04.2017	15:26:00
5	OrderStartDateTime	07.04.2017	15:26:00
6	ReleaseDateTime	07.04.2017	15:26:00
7	SignedDateTime	07.04.2017	15:26:00
8	ActivityDateTime	07.04.2017	15:26:47
9	CPRS RELEASED ORDER	07.04.2017	15:26:47
10	PRINTED TO	07.04.2017	15:26:47
11	RequestDateTime	07.04.2017	15:26:47
12	VistaCreateDate	07.04.2017	16:27:03
13	ActivityDateTime	10.04.2017	07:24:43
14	RECEIVED	10.04.2017	07:24:43
15	ActivityDateTime	10.04.2017	08:44:40
16	SCHEDULED	10.04.2017	08:44:40
17	ActivityDateTime	11.04.2017	13:18:00
18	INCOMPLETE RPT	11.04.2017	13:18:00
19	ActivityDateTime	11.04.2017	14:26:05
20	DISASSOCIATE RESULT	11.04.2017	14:26:05
21	ActivityDateTime	11.04.2017	15:36:11
22	INCOMPLETE RPT	11.04.2017	15:36:11
23	LastActivityDateTime	11.04.2017	16:00:00
24	OrderStopDateTime	11.04.2017	16:00:00
25	ResultsDateTime	11.04.2017	16:00:00
26	ActivityDateTime	11.04.2017	16:00:15
27	COMPLETE/UPDATE	11.04.2017	16:00:15
28	VistaEditDate	11.04.2017	17:04:03

4.2.5 Possible Anomalies

This section describes observed outliers and possible anomalies for Consults.

4.2.5.1 Missing data

The following two events do not have data in the Consults dataset.

1. CPRSOrder.PatientAppointmentDateTime
2. Order actions: Hold (HD) and Release Hold (RL)

4.2.5.2 Duration outliers

In the Consults dataset, four cases span unreasonable dates, as shown in Table 8. Notice that the dates are completely out of the reasonable range.

Table 8. Consult dataset duration outliers.

Case ID	Event	Outlier
4686387	VistaCreateDate	01/01/1900
22860331	ActivityDateTime	01/01/1935
22860331	CANCELED	01/01/1935
22863931	CANCELED	01/01/1935
4165399	ActivityDateTime	18/05/1948
4165399	RECEIVED	18/05/1948
3699106	EarliestDate*	10/05/2030

* In general, Consults has the EarliestDate event many years in the future, as in this case, and in general about 10 years after the LastActivityDateTime. We found 280 cases in this variant.

4.2.5.3 Performance high-impact areas

Table 9 presents Consults high-impact areas of performance, after filtering. The data in Table 9 comes from Figure 7.

Table 9. Consults high-impact areas of performance—after filtering (from Figure 7).

From	To	Elapsed time
CompletedDateTime	ActivityDateTime	6.8 days
SCHEDULED	LastActivityDateTime	4.3 days
ADDED COMMENT	LastActivityDateTime	53.6 hours
ADDED COMMENT	DiscontinuedDateTime	54.7 hours
ADDED COMMENT	Missing	41.9 hours

In addition to Table 9, there seems to be a duration bottleneck in two places in Consults, part of which is shown in Figure 11, which does not present all activities but only the ones that present more delays:

1. Time from VistaCreateDate to ActivityDateTime
2. Time from VistaCreateDate to LastActivityDateTime

Our understanding is that item 1 above may refer to the day when the Consult takes place, and item 2 above may correspond to when the Consult sequence is completed. The largest bottlenecks in the list of items for the Consults process map are shown in Figure 11. As in previous process maps, the large number that follows the arrows is the mean duration, and the small number is the median duration. Figure 7 and Figure 11 both relate to events ActivityDateTime and LastActivityDateTime. Figure 7 takes a closer look at the paths.

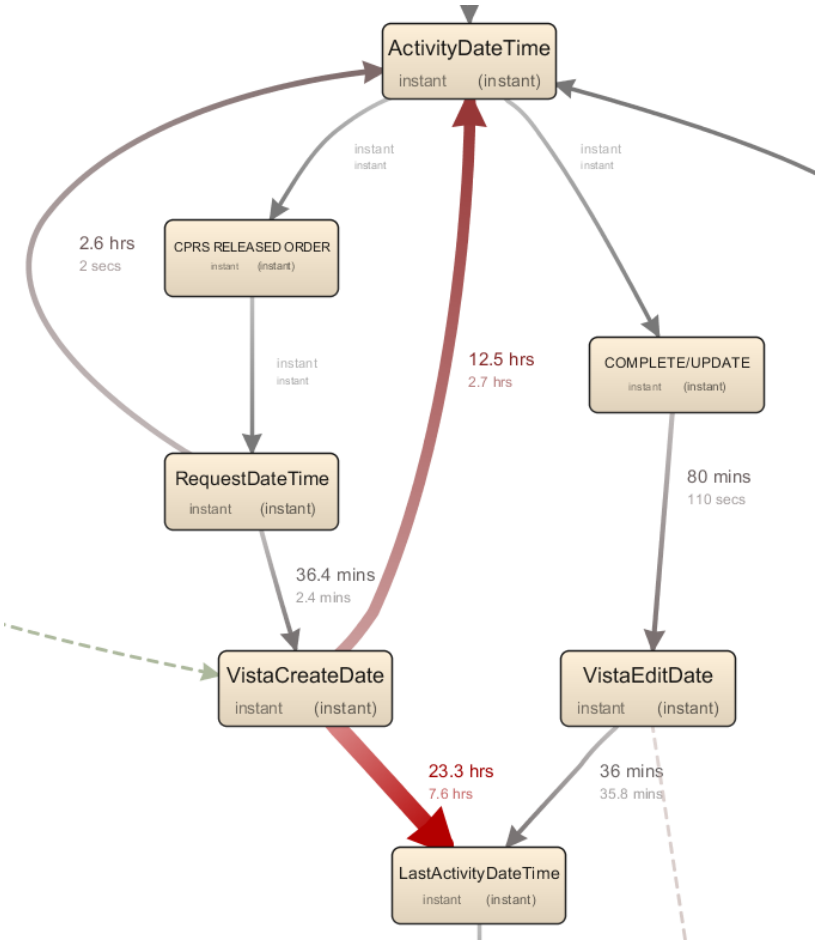

Figure 11. Consults process map duration outliers.

4.2.5.4 Loops in the Consults process

Because of the volume of the data and the complexity of the map process graph, only the most dominant paths and activities were analyzed and are presented herein. Figure 12 shows a portion of the clinical Consults data flow. The *ActivityDateTime* node is the most visited node of the Consults process map with almost 11,000 visits. Also note the cycles from *RECEIVED* to *ActivityDateTime* and back to *RECEIVED*, from *ActivityDateTime* to *Scheduled* and back to *ActivityDateTime*, and from *ActivityDateTime* to *Added Comment* and back to *ActivityDateTime*. This last cycle is particularly of interest, as we found about 50 cases that loop between 200 times and more than 700 times, which is completely out of the norm.

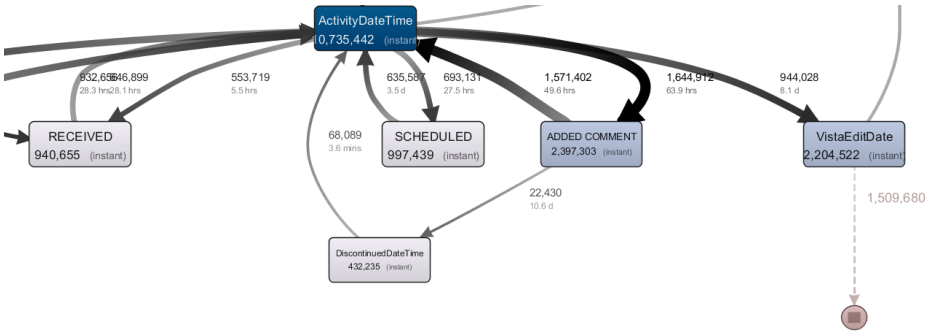

Figure 12. Consults loops in data flow - fragment.

4.3 RADIOLOGY

The complete process map for Radiology before filtering, as shown in Figure 13, is extremely complex and impossible to analyze, making process mining necessary to understand the patterns and identify possible anomalies.

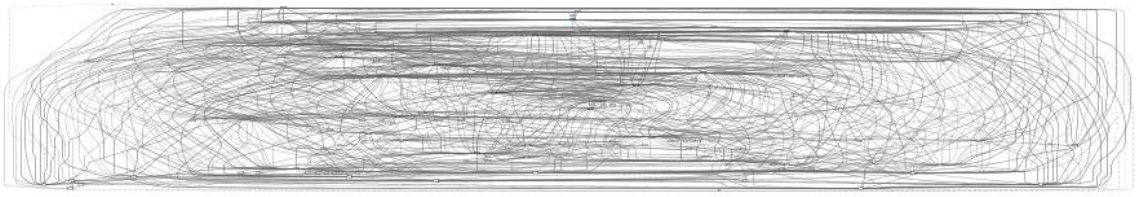

Figure 13. Radiology process map before filtering. Includes all activities and only 25% of the paths.

In our original application of process mining, we applied filters to the Radiology dataset to identify the most common path and descriptive statistics, as presented in the following Figure 14. In this figure, the activities move in three main cycles joined by a COMPLETE status event. The first part consists of eight steps related to the creation of the Radiology consult; the second part is a loop in the lower left side of the image, which consists of the radiology exam itself. Finally, in the lower right-side loop, the third step completes and closes the Radiology workflow. However, we realized that in this case, we can see that the process map has multiple 'COMPLETE' activities. This repeating activity in multiple contexts throughout the process creates false loops that are not actual loops but here a lack of specificity in the activity names creates spider like activities [21]. Initially, we didn't know that there were multiple activities' values with the same name in different contexts and with different timestamps. We realized that it is important to clearly identify each activity with a unique name to avoid the creation of false loops.

The updated Radiology process maps with distinct activity names can be found on Figure 15.

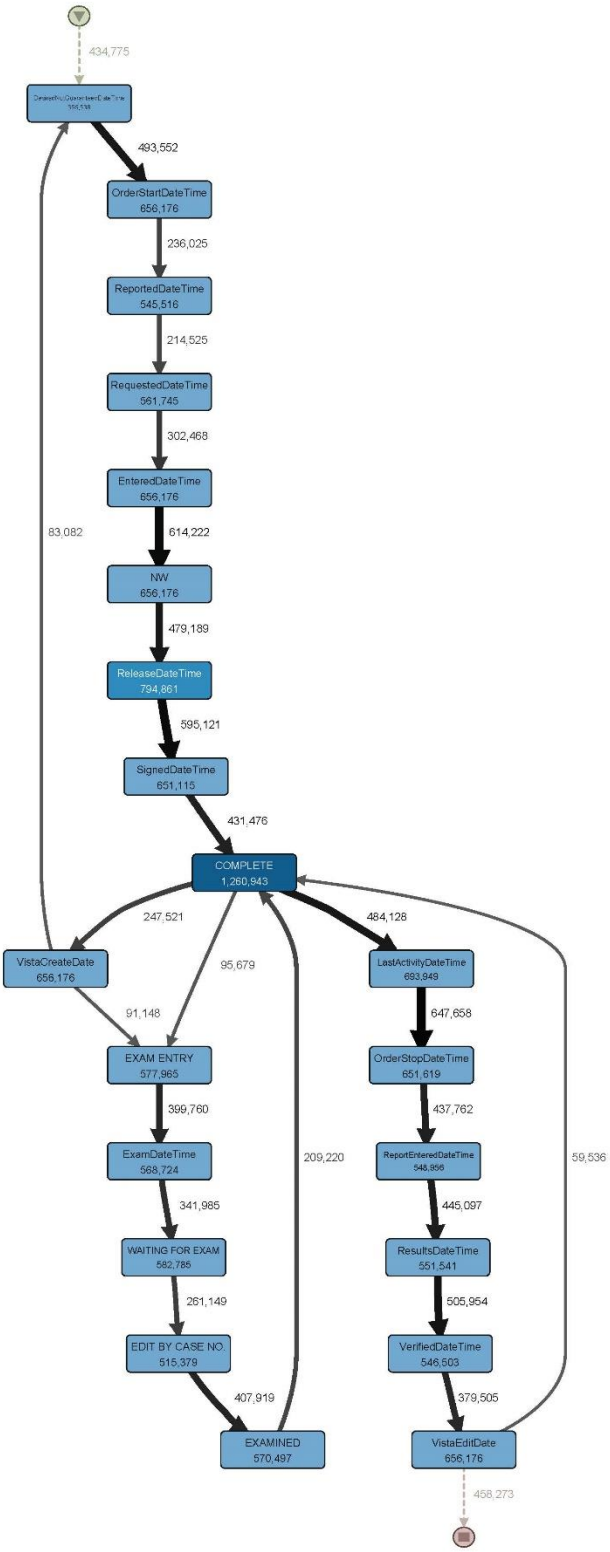

Figure 14. Radiology process map – notice the false loops due to repeated activity on multiple contexts.

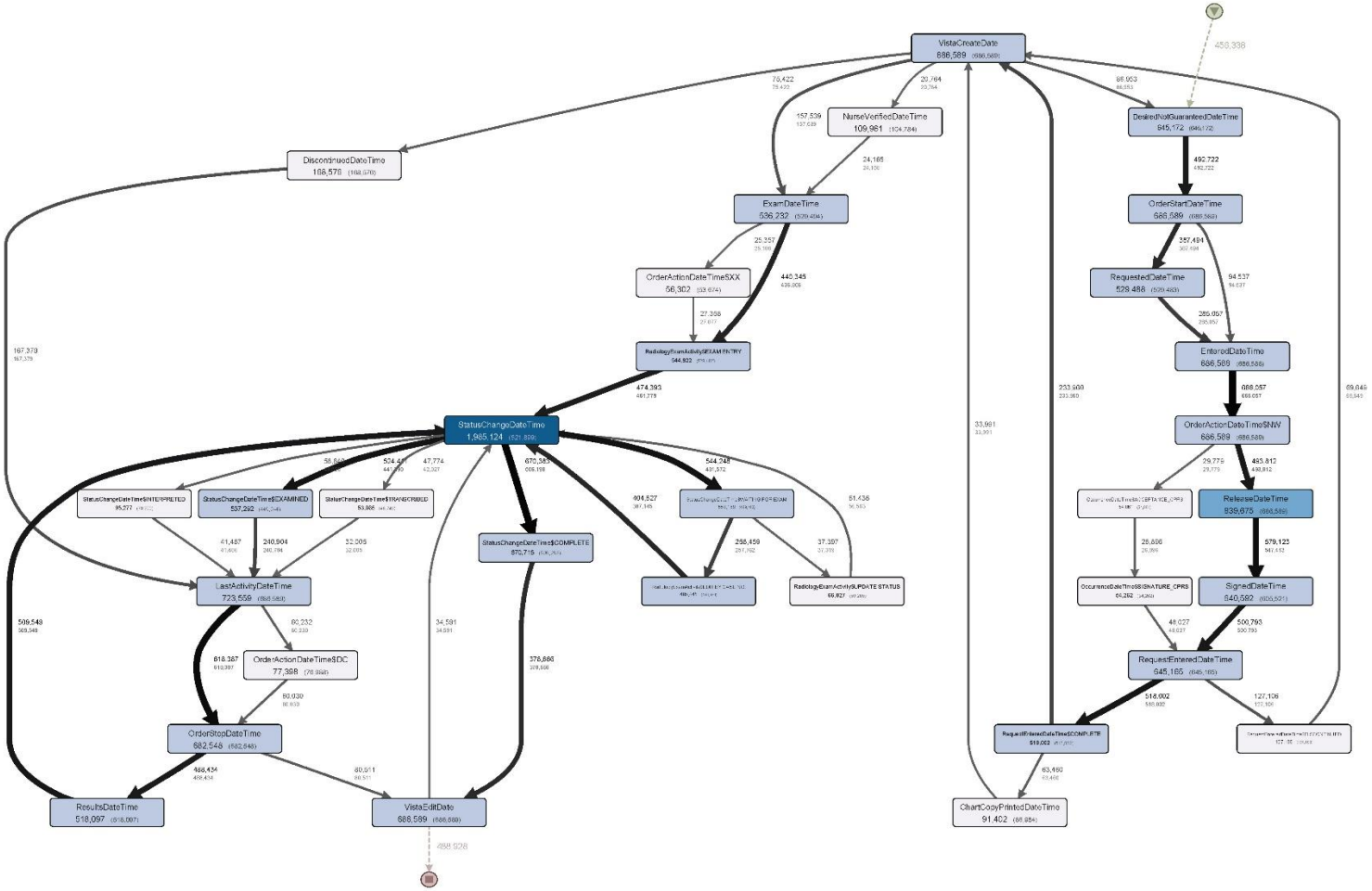

Figure 15. Radiology process model map with frequencies after filtering—activity names have been corrected to be unique.

4.3.1 Frequency Metrics

The list of activities in the Radiology dataset is found in Table 10 which also presents the frequency and the relative frequency percentage.

Table 10. Radiology dataset activities, frequency and relative frequency.

Activity	Frequency	Relative frequency
COMPLETE	1,498,775	8.09%
ReleaseDateTime	1,008,262	5.44%
LastActivityDateTime	913,163	4.93%
OrderStartDateTime	865,020	4.67%
VistaCreateDate	865,020	4.67%
VistaEditDate	865,020	4.67%
EnteredDateTime	865,019	4.67%
DesiredNotGuaranteedDateTime	864,369	4.66%
OrderStopDateTime	858,459	4.63%
NW	810,611	4.37%
SignedDateTime	764,857	4.13%
WAITING FOR EXAM	700,283	3.78%
EXAM ENTRY	693,362	3.74%
ExamDateTime	682,095	3.68%
EXAMINED	675,598	3.64%
RequestedDateTime	673,021	3.63%
ResultsDateTime	658,059	3.55%
ReportEnteredDateTime	653,688	3.53%
ReportedDateTime	650,257	3.51%
VerifiedDateTime	649,877	3.51%
EDIT BY CASE NO.	617,870	3.33%
DiscontinuedDateTime	206,606	1.11%
DISCONTINUED	206,172	1.11%
OccurrenceDateTime	142,757	0.77%
NurseVerifiedDateTime	131,742	0.71%
ChartCopyPrintedDateTime	114,944	0.62%
INTERPRETED	113,518	0.61%
DC	97,266	0.52%
UPDATE STATUS	81,882	0.44%
XX	73,684	0.40%
TRANSCRIBED	68,618	0.37%
ChartReviewedDateTime	42,034	0.23%
OutsideFilmsRegisterDateTime	41,489	0.22%
ClerkVerifiedDateTime	37,012	0.20%
DiagnosticPrintDateTime	36,240	0.20%
CANCELLED	35,136	0.19%

EDIT BY PATIENT	20,287	0.11%
HD	20,029	0.11%
CALLED FOR EXAM	19,712	0.11%
REGISTERED FOR EXAM	18,070	0.10%
PastVisitDateTime	17,945	0.10%
IMAGING COMPLETE	12,887	0.07%
examined	11,703	0.06%
RL	10,294	0.06%
DiscontinuedHoldUntilDateTime	9,864	0.05%
InitialOutsideReportEntryDateTime	8,353	0.05%
TRANSCRIPTION	8,219	0.04%
DICTATED/INTERPRETED	6,927	0.04%
OrderActionFlagDateTime	6,922	0.04%
Released	6,502	0.04%
ACTIVE	6,214	0.03%
OrderActionUnflaggedDateTime	6,034	0.03%
REQUESTED	5,637	0.03%
PT REGISTERED	5,574	0.03%
ReturnedDateTime	5,064	0.03%
AdministeredDoseDateTime	5,013	0.03%
WAITING	4,795	0.03%
ORDERED	4,015	0.02%
VERIFIED REPORT	3,879	0.02%
NeededBackDateTime	3,742	0.02%
Interpreted	3,352	0.02%
EXAM IN PROGRESS	3,198	0.02%
SecondaryDiagnosisPrintDateTime	3,077	0.02%
COMPLETED	2,848	0.02%
EXAM COMPLETED	2,829	0.02%
SCHEDULED EXAM	2,762	0.01%
UNVERIFIED/RELEASED	2,438	0.01%
completed	2,286	0.01%
COM	2,174	0.01%
PreOpScheduledDateTime	2,128	0.01%
BEING EXAMINED	2,106	0.01%
COMPLETE STATUS OVERRIDE	1,909	0.01%
interpreted	1,831	0.01%
REGISTERED FOR EXAMS	1,645	0.01%
VERIFY	1,447	0.01%
EXAMINED(NO REPORT)	1,284	0.01%
AWAITING TRANSCRIPTION	828	0%
Imaging Complete	785	0%
ReportPrintedDateTime	708	0%

SCHEDULED	567	0%
complete	544	0%
RELEASED/NOT VERIFIED	533	0%
exam	479	0%
REGISTERED	475	0%
imaging complete	356	0%
PENDING	347	0%
INTERRPRETED	327	0%
RELEASED/UNVERIFIED	327	0%
DrawnDateTime	322	0%
ScheduledOptionalDateTime	299	0%
MedicationAdministeredDateTime	265	0%
REPORTED	233	0%
Complete	209	0%
Scheduled	192	0%
VERIFIED	181	0%
EXAMINE	176	0%
IMAGED	176	0%
CALLED FOR	163	0%
X-RAYED	148	0%
Missing	147	0%
lower	139	0%
PRELIMINARY	125	0%
MISSING TECH	104	0%
VERIFICATION	78	0%
CompletedDateTime	64	0%
TEST IN PROGRESS	43	0%
HOLD	42	0%
UNVERIFIED	36	0%
RELEASED NOT VERIFIED	25	0%
INTEPRETED	6	0%
PROCEDURE COMPLETE	6	0%
DIAGNOSIS ENTRY BY CASE NO.	2	0%
INTERPERTED	1	0%
mammo	1	0%
RELEASED	1	0%
Test in Progress	1	0%
waiting for exam	1	0%

The raw process map for Radiology consists of ~18.5 million events, 865,000 cases, and 117 activities (see Figure 14). Because the Radiology dataset has so many activities, we decided to present only the most visited activities. Figure 16 presents the Pareto chart for Radiology for frequency by activities, after filtering, there were 114 total activities in Radiology. The minimal frequency is 1, the median frequency

is 3,178. Figure 16 also shows the metrics of mean frequency, maximal frequency, and frequency standard deviation, which were obtained after filtering.

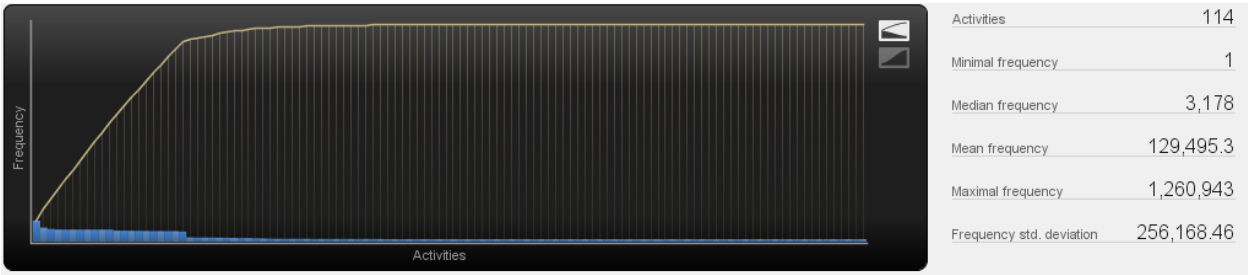

Figure 16. Radiology Pareto chart—frequency by activities.

Figure 17 shows Radiology events, cases and activities.

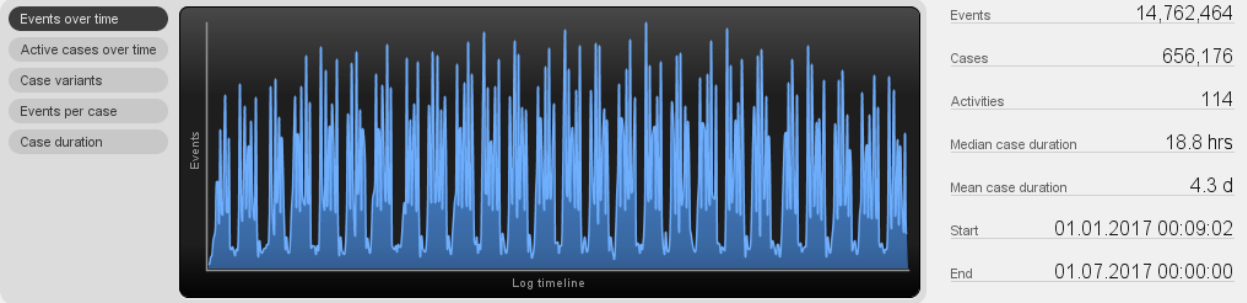

Figure 17. Radiology stats - after filtering.

4.3.2 Performance Metrics

We applied filters to the Radiology dataset to identify the duration of the most common path and other duration statistics, as shown in Figure 17. In more details, the median and mean duration of the most common path can be found in Table 2 and Figure 17, where we can see that the median case duration for Radiology is 18.8 hours and the mean case duration is 4.3 days. Table 2 and Figure 17 are considered the process model case to identify outliers and cases that do not conform with the average case in order to identify possible anomalies.

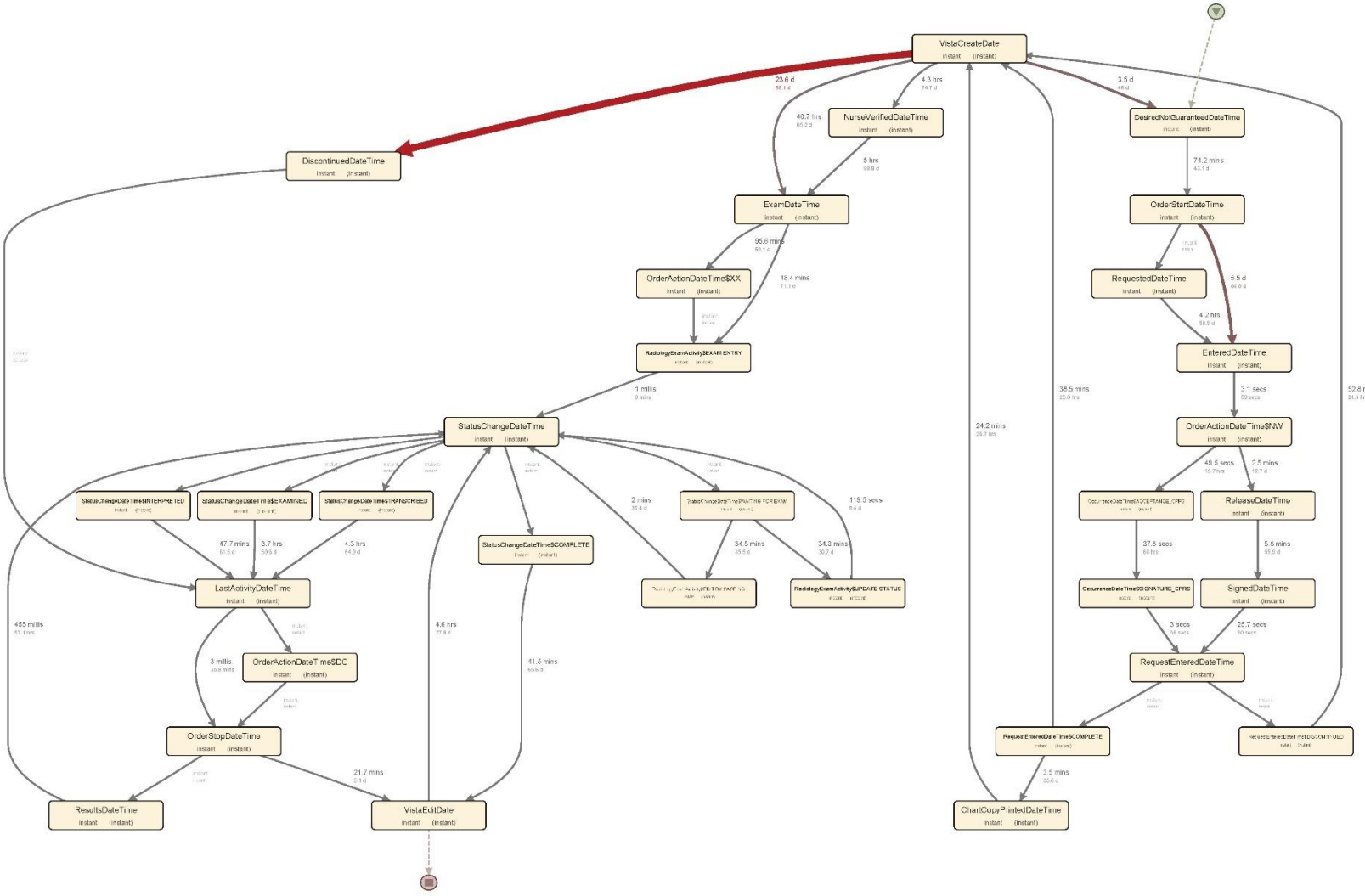

Figure 18. Radiology process map with frequencies after filtering activity names have been corrected to be unique.

4.3.3 Mapping to OASIS Human Task State Transition and termination States for Radiology

Table 11 presents the mapping proposal of the CDW Radiology data domain to the OASIS Human Task State Transitions.

Table 11. Radiology dataset mapping to OASIS Human Task State Transition.

OASIS Human Task State Transitions	Radiology date columns	Radiology other events
CREATED	**EnteredDateTime** DesiredNotWaraneedDateTime RequestedDateTime	
READY	OrderStartDateTime OrderActionDateTime ReportedDateTime ReleaseDateTime	OrderAction: NW RadiologyOrderAction: NEW ExamStatus: ACTIVE
RESERVED	SignedDateTime VistaCreateDate ReportedDateTime	RequestStatus: SCHEDULED ExamStatus: CALLED FOR CALLED FOR EXAM ORDERED PT REGISTERED REGISTERED REGISTERED FOR EXAM REGISTERED FOR EXAMS REQUESTED SCHEDULED SCHEDULED EXAM
IN_PROGRESS	OccurrenceDateTime ExamDateTime LogDateTime StatusChangeDateTime PatientAppointmentDateTime LapsedDateTime VistaEditDate OrderActionDateTime NurseVerifiedDateTime ClerkVerifiedDateTime ChartReviewedDateTime OrderActionFlagDateTime OrderActionUnflaggedDateTime ChartCopyPrintedDateTime InterventionDate SecondaryDiagnosisPrintDateTime MedicationAdministeredDateTime	RadiologyActionType: EXAM ENTRY EDIT BY CASE NO. DIAGNOSIS ENTRY BY CASE NO. NO PURGING SPECIFIED COMPLETE STATUS OVERRIDE EDIT BY PATIENT EXAM STATUS TRACKING UPDATE STATUS RadiologyOrderAction: RELEASE HOLD CHANGE RequestStatus: ACTIVE UNRELEASED ExamStatus: BEING EXAMINED DICTATED/INTERPRETED exam

OASIS Human Task State Transitions	Radiology date columns	Radiology other events
	OutsideFilmsRegisterDateTime NeededBackDateTime ReturnedDateTime ReportPrintedDateTime PreVerificationDateTime DiagnosticPrintDateTime PreOpScheduledDateTime RequestEnteredDateTime PastVisitDateTime LastActivityDateTime ScheduledOptionalDateTime DrawnDateTime AdministeredDoseDateTime	EXAM IN PROGRESS EXAMINE examined EXAMINED(NO REPORT) IMAGED IMAGING COMPLETE INTEPRETED INTERPERTED INTERPRETED INTERRPRETED MISSING TECH PRELIMINARY Released RELEASED NOT VERIFIED RELEASED/NOT VERIFIED RELEASED/UNVERIFIED REPORTED TEST IN PROGRESS TRANSCRIBED TRANSCRIPTION UNVERIFIED UNVERIFIED/RELEASED VERIFICATION VERIFIED VERIFIED REPORT VERIFY WAITING WAITING FOR EXAM X-RAYED lower mammo
COMPLETED	***ResultsDateTime*** LastActivityDateTime OrderStopDateTime ReportEnteredDateTime VerifiedDateTime CompletedDateTime	RequestStatus: COMPLETE ExamStatus: COM COMPLETE COMPLETED EXAM COMPLETED PROCEDURE COMPLETE
SUSPENDED	DescontinuedHoldUntilDateTime	RadiologyOrderAction: ***HOLD*** RequestStatus: HOLD PENDING ExamStatus:

OASIS Human Task State Transitions	Radiology date columns	Radiology other events
		HOLD AWAITING TRANSCRIPTION PENDING
FAILED	***In state Completed and there is Discontinued Date Time recorded***	ExamStatus: *Missing*
ERROR	***In state Ready and there is Discontinued Date Time recorded***	
EXITED	***DiscontinuedDateTime, not at Completed or Ready***	RadiologyActionType: CANCELLED RadiologyOrderAction: DISCONTINUE RequestStatus: DISCONTINUED ExamStatus: CANCELLED
CLOSED	VistaEditDate	

Table 3 presents the breakdown of successful/failed cases for Radiology: 865,020 of the cases *complete* successfully (76%), while 189,439 cases *exited* (22%), 16,552 cases ended in *error* (about 2%), and 96 *failed* (~0.01%). The process map with OASIS mapping for Radiology filtered is found in Figure 19. In it, most cases (81%) complete successfully in the common path. Thus, to identify the failed states, we need to remove the filters.

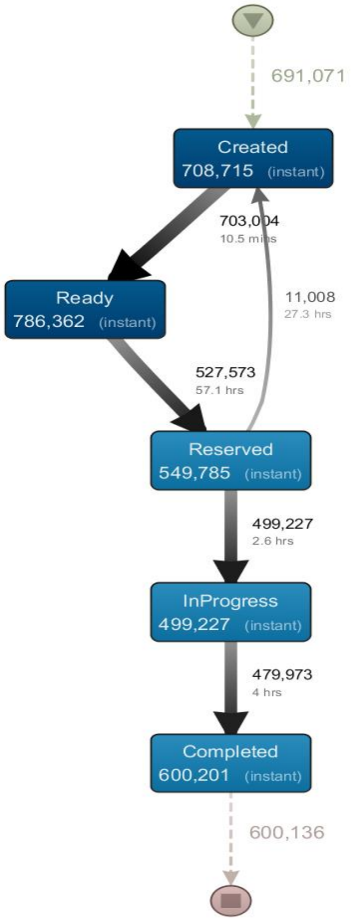

Figure 19. OASIS human task process map for Radiology.

Notice that on Figure 20, the OASIS state human task state transitions have been identified on top of the raw process map to identify the cluster of activities associated to each state.

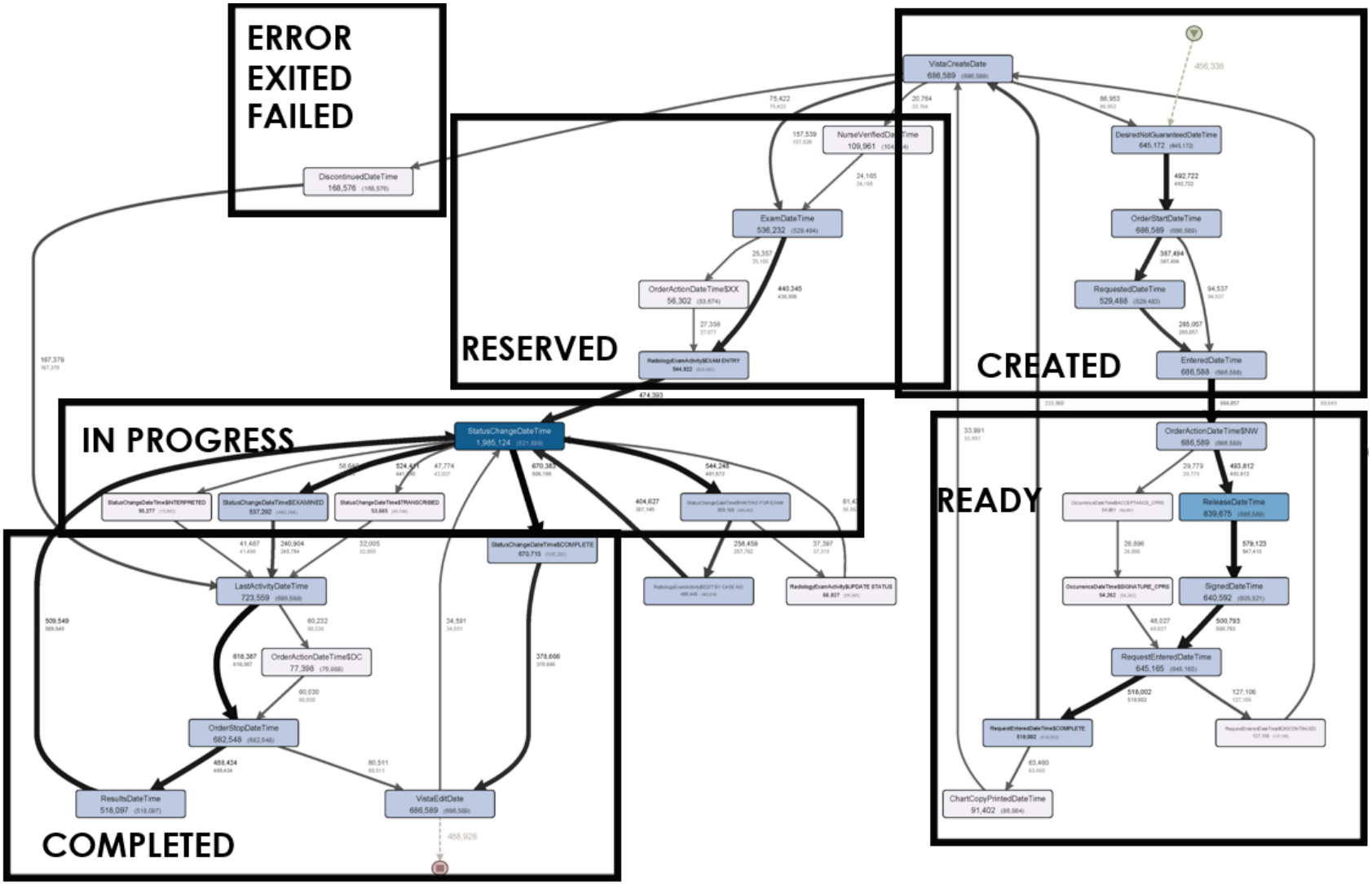

Figure 20. Radiology Process Map presenting state transitions between activities.

Figure 21 presents a subset of Radiology cases that went to an error state. Recall that this dataset was filtered to 6 months (January to June 2017). In this subset, we can see that more than 106,000 cases *exited*, 2,257 were *error* cases, and 61 were *failed* cases.

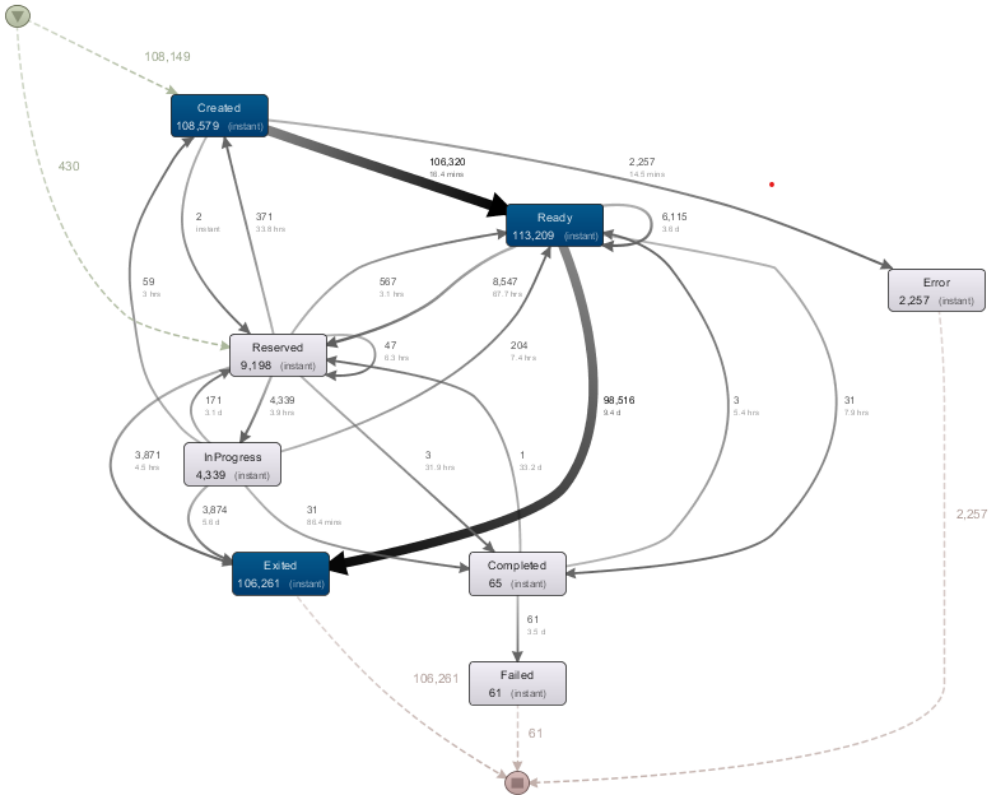

Figure 21. Radiology dataset after filtering cases that included error states.

After examining Figure 21, a process map was generated of all Radiology dataset cases that terminated in an error state and removed the timeframe filter, as shown in Figure 22. There, we can see that 96 cases *failed*, ~190,000 cases *exited*, and 16,551 cases transitioned to the *error* state.

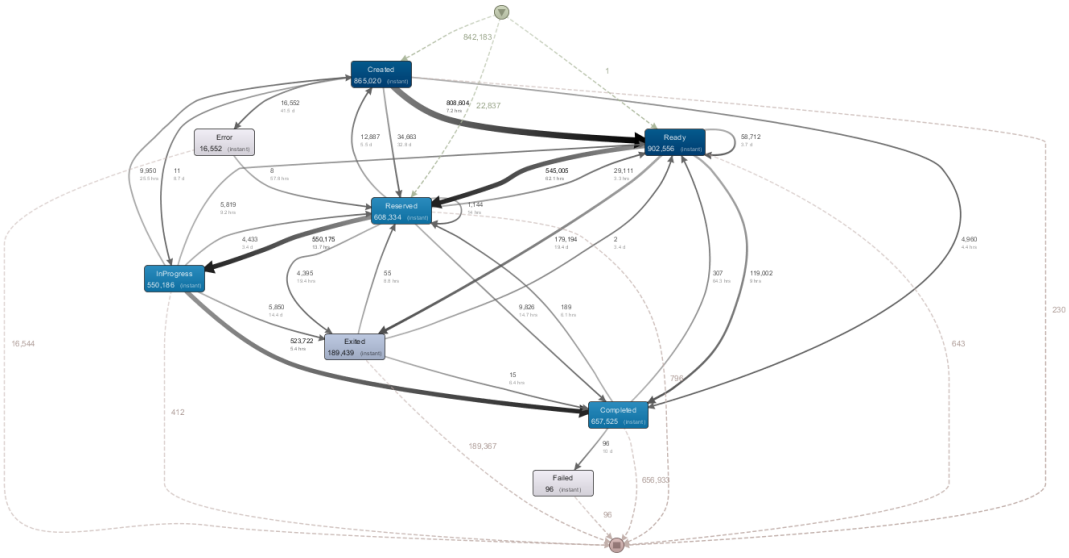

Figure 22. OASIS state transition process map for Radiology dataset.

4.3.4 Possible Anomalies

This section describes observed outliers and possible anomalies in Radiology.

4.3.4.1 Duration outliers

The Radiology dataset includes five cases with VistaCreateDate with a date of 1/1/1900. In addition, there are 593 cases with values in the NeededBackDateTime column of January 2023 and later. An extreme case has the date 12/12/2090, as shown in Table 12.

Table 12. Radiology dataset duration outliers

Case ID	Activity	Outlier
10994986, 4686379, 20305825, 4686383, 10994982	VistaCreateDate	01.01.1900
3589584, 4361857, 3868089	NeededBackDateTime	18.09.2036 09.06.2096 12.12.2090

* This is a similar case to the one observed in Consults in which the EarliestDate event has dates in the future.

4.3.4.2 Performance high-impact areas

Table 13 presents Radiology high-impact areas of performance, after filtering. The data from Table 13 is generated from Figure 23. Radiology high-impact areas map (all activities) after filtering

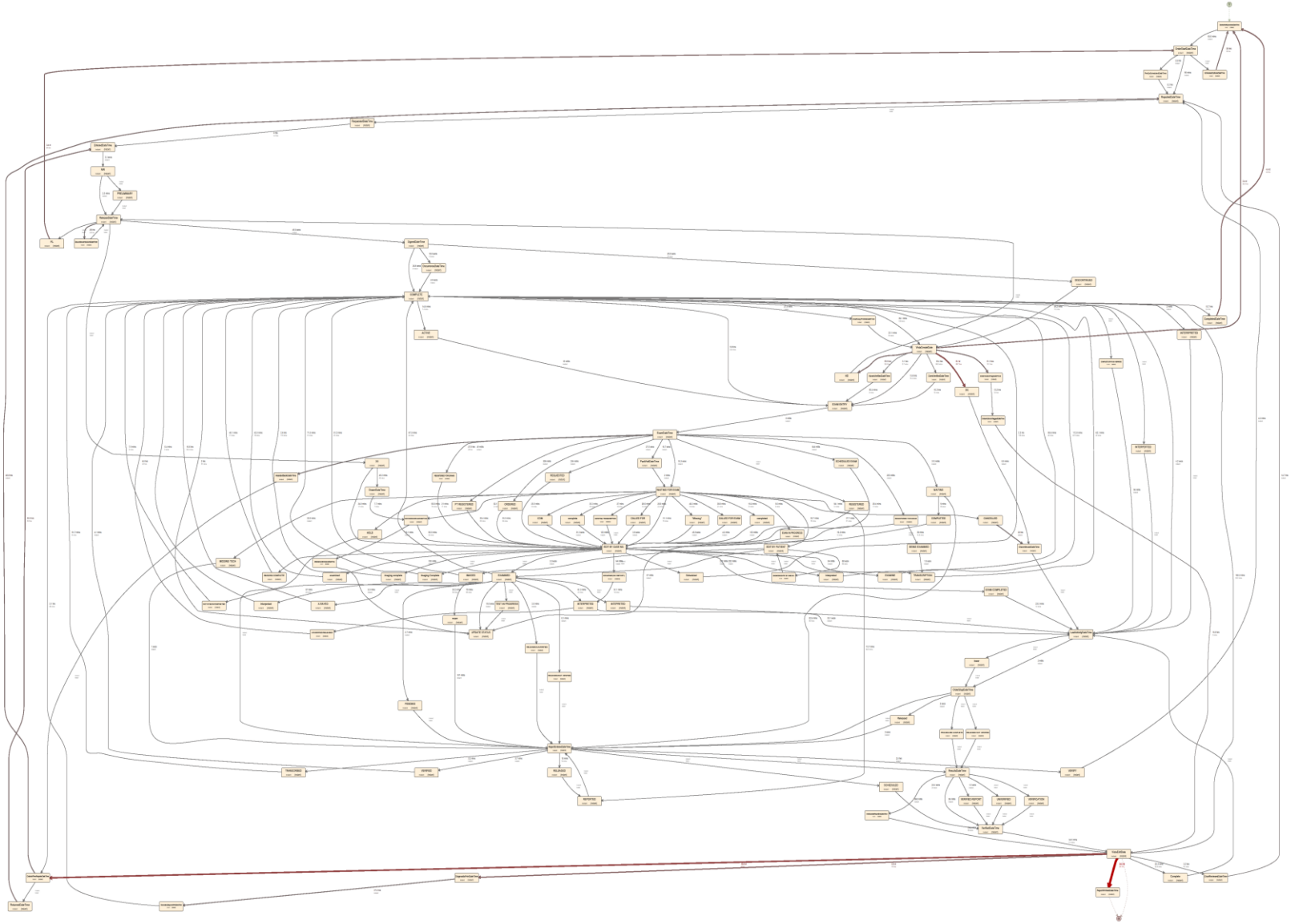

Figure 23. Radiology high-impact areas map (all activities) after filtering.

Table 13. Radiology high-impact areas of performance—after filtering (from Figure 23)

From	To	Elapsed time
VistaCreateDate	DesiredNotGuaranteedDateTime	3.4 days
VistaCreateDate	DC	8.3 days
VistaEditDate	OutsideFilmsRegisterDateTime	10.8 days
VistaEditDate	ReportPrintedDateTime	24.3 days

4.3.4.3 Loops in the Radiology process

There are several loops in the Radiology process map. An initial version of the Radiology process model map is shown in Figure 14, in which the resulting process map is shown with zero paths, which is complex. Activities move in three main cycles joined by a COMPLETE status. In this case, we can see that the process map has multiple "COMPLETE" activities. This repeating activity in multiple contexts throughout the process creates false loops that are not actual loops, but a lack of specificity in the activity names creates spider-like activities (see "Spider activity" explanation in Rozinat et al. [21] https://flux-icon.com/book/read/simplification/#strategy-8-removing-spider-activities). Therefore, we wondered whether this is the complete of the radiology exam status column, or the complete from the RadiologyNuclearMedicineOrder table RequestStatus (RequestEnteredDateTime replaced by RequestStatus) Initially, we did not know that there were multiple activities values with the same name in different contexts and with different timestamps. We realized that it is important to clearly identify each activity with a unique name to avoid the creation of false loops. Thus, we renamed those activities that had the same name in different contexts by adding the column name to the name of the activity. The corrected Radiology process map with frequencies is shown in Figure 15, and the corrected Radiology process model map with performance values is shown in Figure 18.

As another example of cycles in Radiology, a sample portion of an earlier analysis shown in Figure 24. With several columns holding values of actions with a field code and a description of the action type, such as NEW, HOLD, COMPLETE, DISCONTINUED, we considered how date columns are related to actions or statuses and then added descriptions of those actions as an event in the sequence. As the figure shows, the data flows from ReportedDateTime to EXAMDATETIME to EXAMENTRY to WAITING FOR EXAM TO COMPLETE, to IMAGING COMPLETE, or EDITBYPATIENT, or EXAMINED or UPDATE STATUS or CANCELED to COMPLETE.

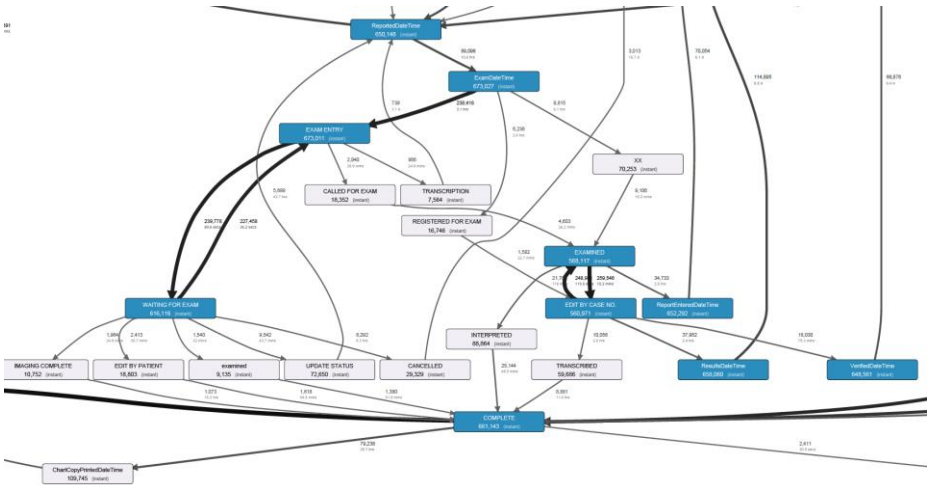

Figure 24. Radiology consults process fragment.

Another illustration is a case having 87 events (case ID: 8182577), as shown in Figure 25, which loops between the activities Examined and Transcribed an unusually high number of times.

	Activity	Date	Time
1	EnteredDateTime	27.03.2017	08:24:00
2	NW	27.03.2017	08:24:30
3	ReleaseDateTime	27.03.2017	08:26:00
4	SignedDateTime	27.03.2017	08:26:00
5	COMPLETE	27.03.2017	08:26:05
6	VistaCreateDate	27.03.2017	08:27:05
7	DesiredNotGuaranteedDateTime	28.03.2017	00:00:00
8	RequestedDateTime	28.03.2017	00:00:00
9	OrderStartDateTime	12.04.2017	17:00:00
10	EXAM ENTRY	12.04.2017	17:03:00
11	ExamDateTime	12.04.2017	17:03:00
12	WAITING FOR EXAM	12.04.2017	17:03:00
13	EDIT BY CASE NO.	12.04.2017	18:12:00
14	EXAMINED	12.04.2017	18:12:00
15	TRANSCRIBED	12.04.2017	22:29:00
16	EXAMINED	12.04.2017	23:06:00
17	TRANSCRIBED	12.04.2017	23:08:00
18	EXAMINED	12.04.2017	23:09:00
19	TRANSCRIBED	12.04.2017	23:10:00
20	EXAMINED	12.04.2017	23:11:00
21	TRANSCRIBED	12.04.2017	23:13:00
22	EXAMINED	12.04.2017	23:14:00
23	TRANSCRIBED	12.04.2017	23:15:00
24	EXAMINED	12.04.2017	23:17:00
25	TRANSCRIBED	12.04.2017	23:18:00
26	EXAMINED	12.04.2017	23:19:00
27	TRANSCRIBED	12.04.2017	23:20:00
28	EXAMINED	12.04.2017	23:22:00
29	TRANSCRIBED	12.04.2017	23:23:00
30	EXAMINED	12.04.2017	23:24:00
31	TRANSCRIBED	12.04.2017	23:25:00
32	EXAMINED	12.04.2017	23:27:00
33	TRANSCRIBED	12.04.2017	23:28:00
34	EXAMINED	12.04.2017	23:29:00
35	TRANSCRIBED	12.04.2017	23:30:00
36	EXAMINED	12.04.2017	23:32:00
37	TRANSCRIBED	12.04.2017	23:33:00
38	EXAMINED	12.04.2017	23:34:00
39	TRANSCRIBED	12.04.2017	23:35:00
40	EXAMINED	12.04.2017	23:37:00
41	TRANSCRIBED	12.04.2017	23:38:00
42	EXAMINED	12.04.2017	23:39:00
43	TRANSCRIBED	12.04.2017	23:40:00
44	EXAMINED	12.04.2017	23:42:00
45	TRANSCRIBED	12.04.2017	23:43:00
46	EXAMINED	12.04.2017	23:44:00
47	TRANSCRIBED	12.04.2017	23:45:00
48	EXAMINED	12.04.2017	23:47:00
49	TRANSCRIBED	12.04.2017	23:48:00
50	EXAMINED	12.04.2017	23:49:00
51	TRANSCRIBED	12.04.2017	23:50:00
52	EXAMINED	12.04.2017	23:52:00
53	TRANSCRIBED	12.04.2017	23:53:00

Figure 25. Radiology case 8182577 loops between Examined and Transcribed - fragment.

4.3.4.4 Radiology special case in daylight savings time

In working on mapping from the Radiology dataset to the OASIS human task state transition diagram, we found a case in which Disco shuffled some of the rows of data shown in one case (see Figure 26). Notice how the record sequence flows naturally from *created* to *ready* to *reserved* to *inprogress* to *completed*.

```
1920239,Created, 2017-03-12 02:58:00
1920239,Ready, 2017-03-12 02:58:00
1920239,Reserved, 2017-03-12 03:00:00
1920239,InProgress, 2017-03-12 03:09:00
1920239,Completed, 2017-03-12 09:33:00
```

Figure 26. Original data from radiology case in daylight savings time.

However, in our setup, in importing the above data to Disco, the setting of the timestamp pattern added one hour in our computer's configuration (see Figure 27). We were able to reproduce this change on several computers. Notice the change in Figure 28. After importing the record, the activities Created and Ready appeared to have happened after Reserved and InProgress because of the daylight saving times addition.

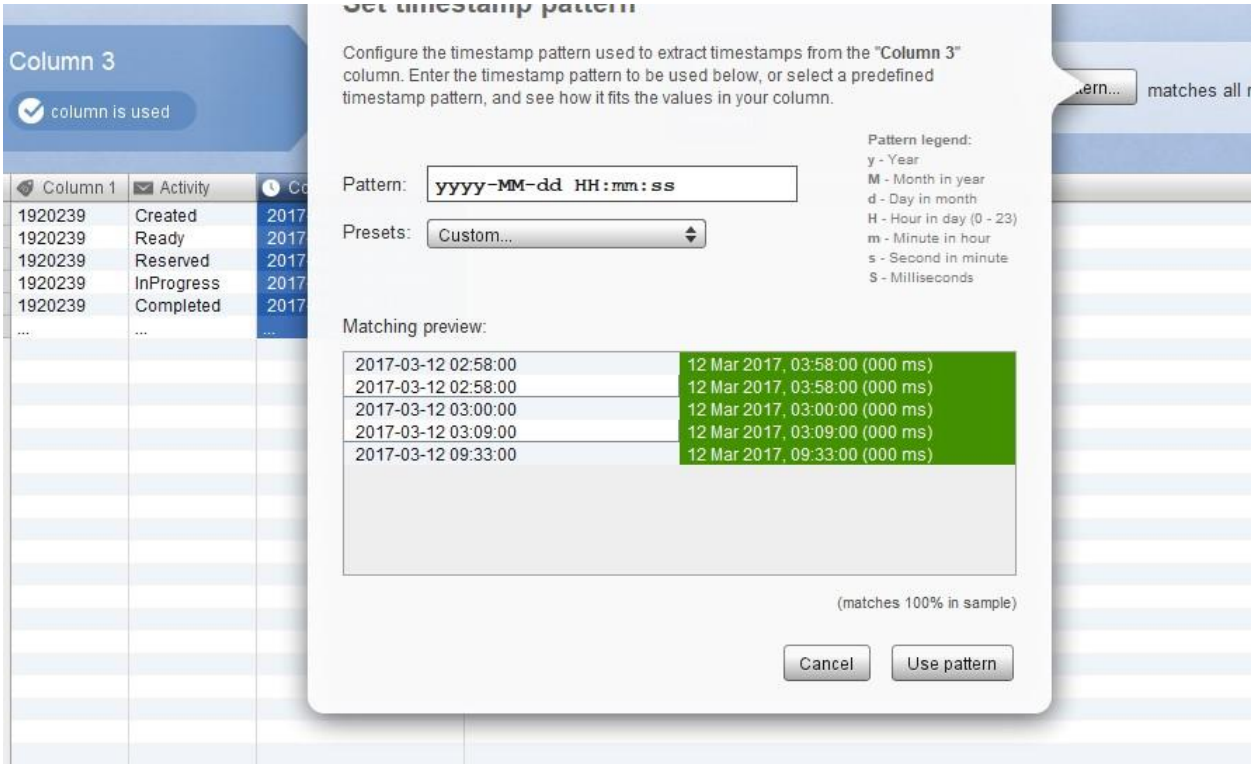

Figure 27. Daylight saving times affected import.

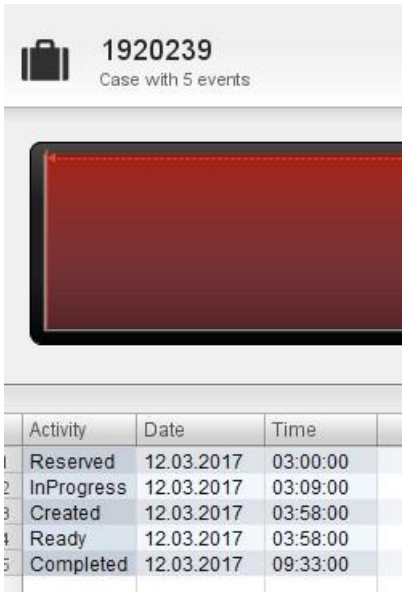

Figure 28. After importing the record, the activities Created and Ready appeared to have happened after Reserved and InProgress because of the daylight saving times addition.

We consulted the Disco support team about this issue. They were not able to reproduce it. However, they explained that the reason is that on March 12, 2017, there was a change to daylight savings time in the United States. At 02:00 in the morning, the clock was moved forward by one hour to 3:00. So, there could not have been a 2:58 timestamp on this day (Figure 28). See http://www.calendarpedia.com/when-is/daylight-saving-time.html The daylight savings time switch was on another date for Europe (the company Fluxicon, which develops Disco, is located in The Netherlands). Therefore, Disco developers could not reproduce the issue. (The Disco app is developed in the Java language. Java interprets the timestamps for the user's local time zone.) We do not know where this particular record was recorded in CDW; perhaps it was recorded in a different time zone. Our OASIS mapping marked this case as a failed case because the switch to daylight savings time swapped the activities, when in fact it was a normal case.

4.4 LABORATORY SERVICES

Because of the time it takes to generate the whole process map with all the activities and paths, it was not included in this document. Because of its complexity, it would have been impossible to analyze.

The Lab Services (Laboratory or Labs) domain generated the largest dataset for this study. Because of its size, during initial analysis, we divided the Labs dataset into three parts. The process maps for each of the three partitions of Labs dataset, before filtering, can be found in Figure 29, Figure 30, and Figure 31. Although the process map generated by the first part of the data set (shown Figure 29) may look different from the process maps generated by the remaining two-thirds of the dataset (shown in Figure 30 and Figure 31) the process path was the same and was merely shifted by the Disco system. Later, during the analysis, we were able to generate a process map for the whole Labs dataset. If we compare the process maps generated by randomly dividing the Labs dataset into three parts with the process map of the whole dataset, we can see important similarities.

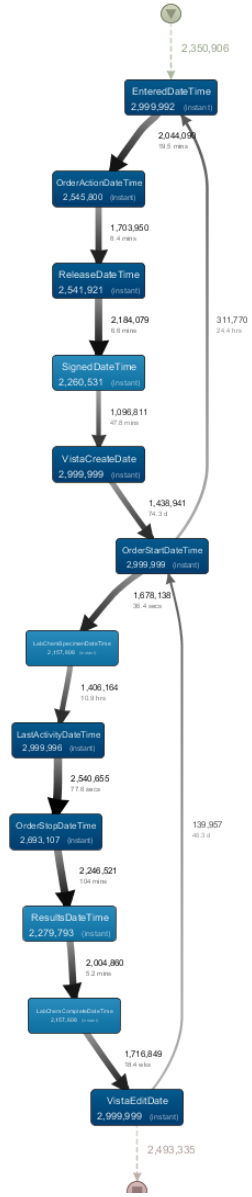

Figure 29. Lab Services most visited path part 1.

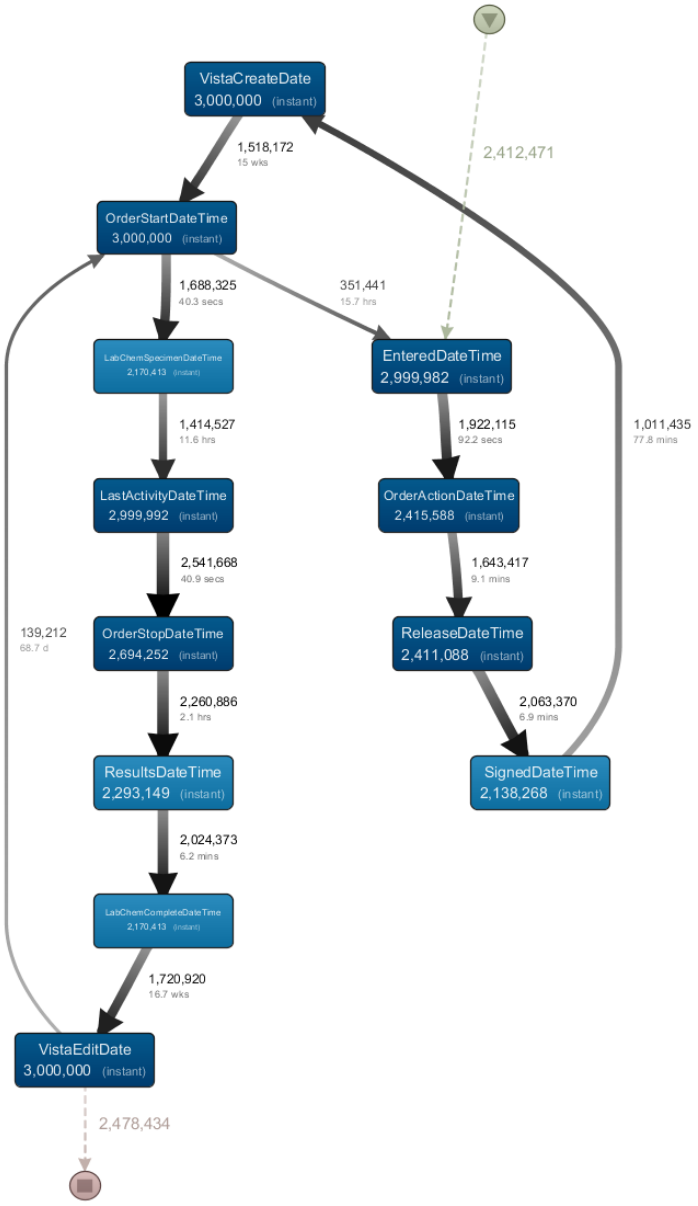

Figure 30. Lab Services most visited path part 2.

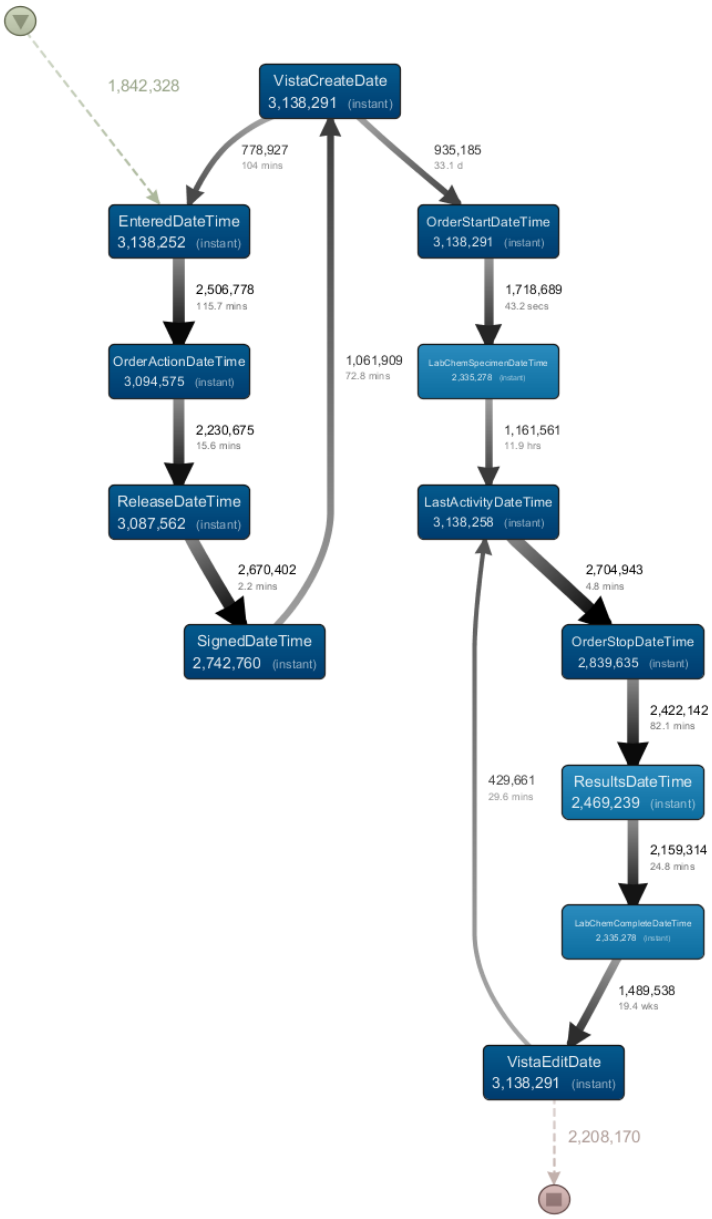

Figure 31. Lab Services most visited path part 3.

To clarify, in the Labs dataset we included not only the CDW Labs data domain but also the Microbiology (micro-domain) tables. Thus, Labs has 3 general areas—pathology, clinical, and microbiology—each with a different flow, as shown in Figure 32. We learned from our discussions with SMEs that it is difficult to apply time statistics across all Labs sections when it is expected that "clinical" is usually minutes/hours, whereas "pathology" and "microbiology" are days/weeks. Thus, we may consider more focused process mining applied to each type of Labs in the future.

4.4.1 Process Model Map

The raw process map for Labs consists of ~105 million events, ~9 million cases, and 228,022 case variants (see Table 1). The raw process map for the Labs dataset is shown in Figure 32. Figure 32 shows

the path that most cases take with the darker blue boxes and the bold, thick arrows. The numbers near boxes and arrows are the absolute frequencies, i.e., how many times an activity was performed in total.

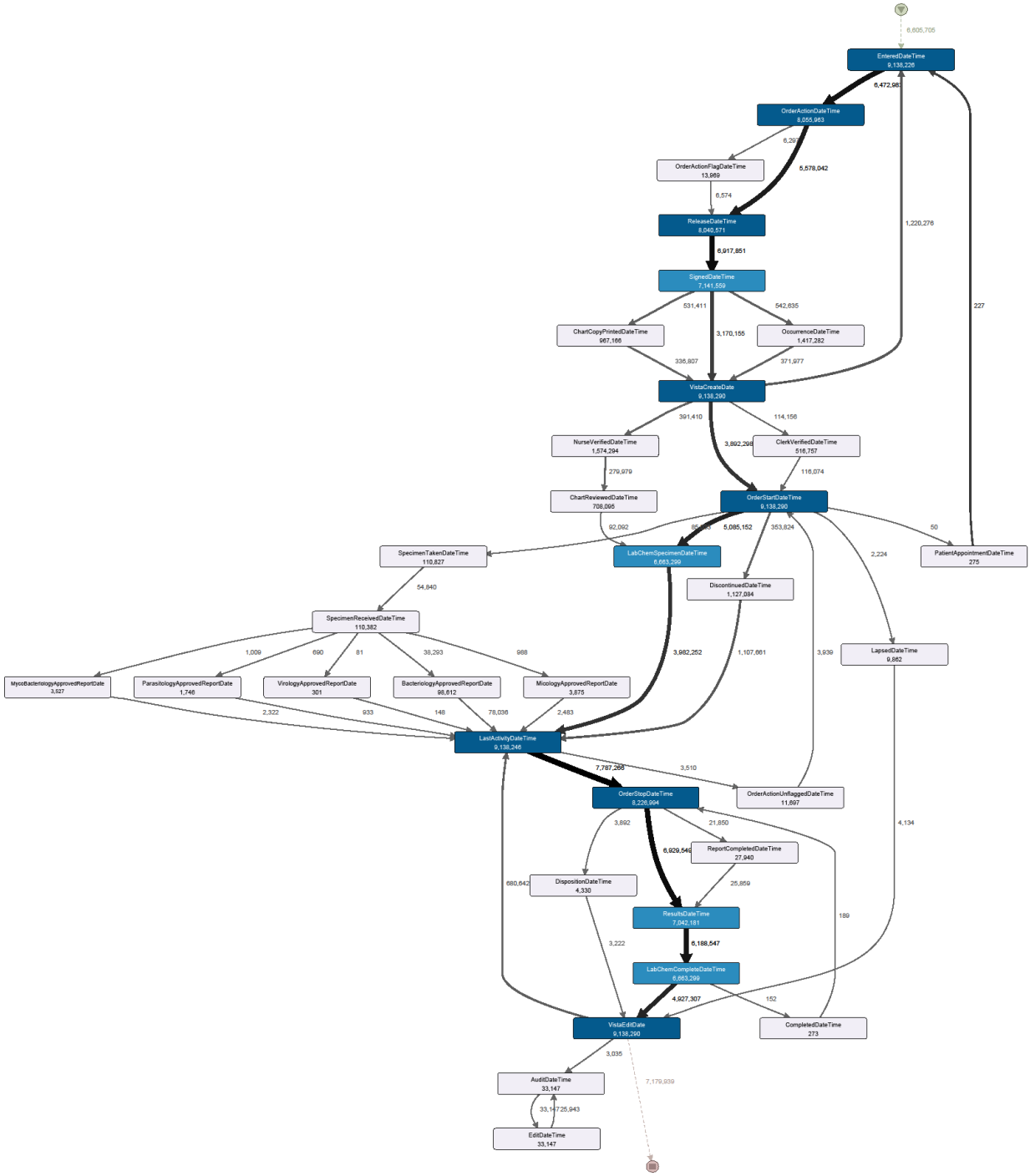

Figure 32. Lab Services process map before filtering.

4.4.2 Frequency Metrics

The list of activities in the Lab Services dataset is found in Table 14, which also presents the frequencies and the relative frequency percentage.

Table 14. Lab Services dataset list of activities, frequency and relative frequency

Activity	Frequency	Relative frequency
VistaCreateDate	9,138,290	8.76%
OrderStartDateTime	9,138,290	8.76%
VistaEditDate	9,138,290	8.76%
LastActivityDateTime	9,138,246	8.76%
EnteredDateTime	9,138,226	8.76%
OrderStopDateTime	8,226,994	7.89%
OrderActionDateTime	8,055,963	7.72%
ReleaseDateTime	8,040,571	7.71%
SignedDateTime	7,141,559	6.85%
ResultsDateTime	7,042,181	6.75%
LabChemSpecimenDateTime	6,663,299	6.39%
LabChemCompleteDateTime	6,663,299	6.39%
NurseVerifiedDateTime	1,574,294	1.51%
OccurrenceDateTime	1,417,282	1.36%
DiscontinuedDateTime	1,127,084	1.08%
ChartCopyPrintedDateTime	967,166	0.93%
ChartReviewedDateTime	708,095	0.68%
ClerkVerifiedDateTime	516,757	0.50%
SpecimenTakenDateTime	110,827	0.11%
SpecimenReceivedDateTime	110,382	0.11%
BacteriologyApprovedReportDate	98,612	0.09%
AuditDateTime	33,147	0.03%
EditDateTime	33,147	0.03%
ReportCompletedDateTime	27,940	0.03%
OrderActionFlagDateTime	13,969	0.01%
OrderActionUnflaggedDateTime	11,697	0.01%
LapsedDateTime	9,862	0.01%
DispositionDateTime	4,330	0%
MicologyApprovedReportDate	3,875	0%
MycoBacteriologyApprovedReportDate	3,527	0%
ParasitologyApprovedReportDate	1,746	0%
VirologyApprovedReportDate	301	0%
PatientAppointmentDateTime	275	0%
CompletedDateTime	273	0%

Before applying filters, the Labs dataset included 36 distinct activities. For these activities, the minimal frequency was 92. The median frequency was 340,238, the mean frequency 1,662,215, and the maximal frequency 5,197,514. The frequency standard deviation was 2,140,272.55 (see Figure 33. Lab Services Pareto chart—frequency by activities., which shows metrics after filtering).

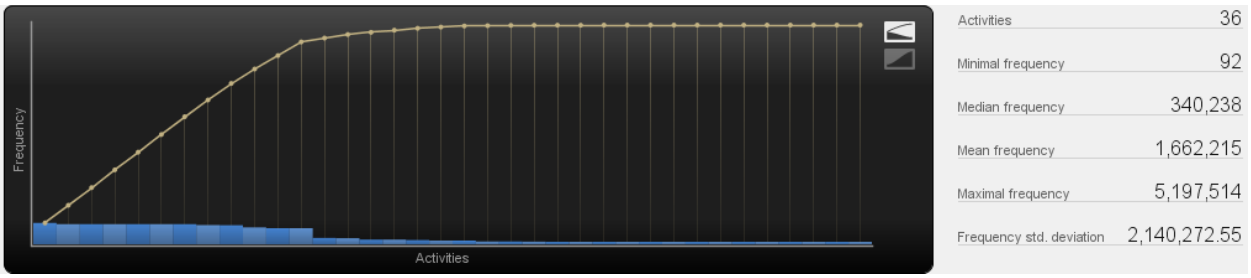

Figure 33. Lab Services Pareto chart—frequency by activities.

In addition, Figure 34 shows frequency metrics after filtering. Labs data set has close to 60 million events, close to 5 million cases, and 34 distinct activities.

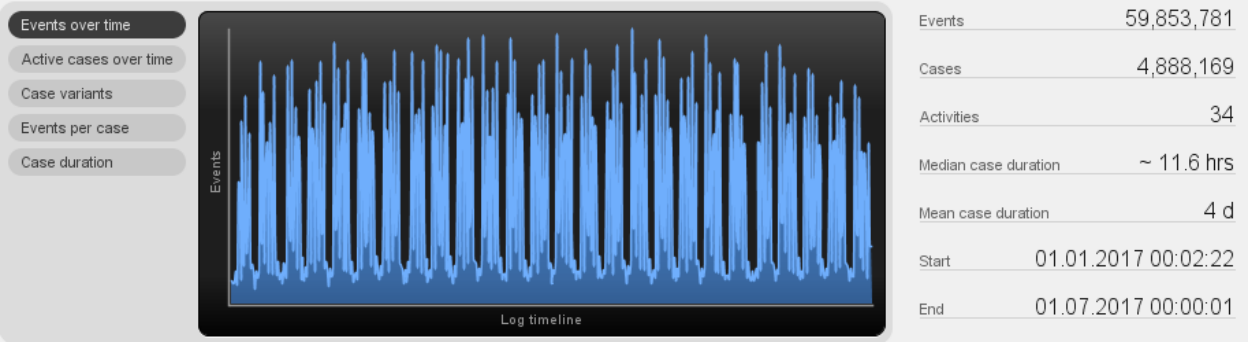

Figure 34. Lab Services stats—after filtering.

4.4.3 Performance Metrics

Labs' performance statistics, after filtering, are shown in Figure 34, which shows the median and mean case duration. Information in Table 2 and Figure 34 was taken as a model case and as a base to identify outliers and cases that do not conform with the model case in order to identify possible anomalies.

We applied time filters, as in previous datasets, to the Labs dataset to identify the most common path's other duration statistics. The duration of the process model case can be found in Table 2 and in Figure 35. Lab Services performance map after filtering. Figure 35 depicts the high-impact areas in bold red, thick arrows where performance bottlenecks are identified. The metrics in a larger font are the mean durations between activities, and the smaller numbers are the median duration (50th percentile).

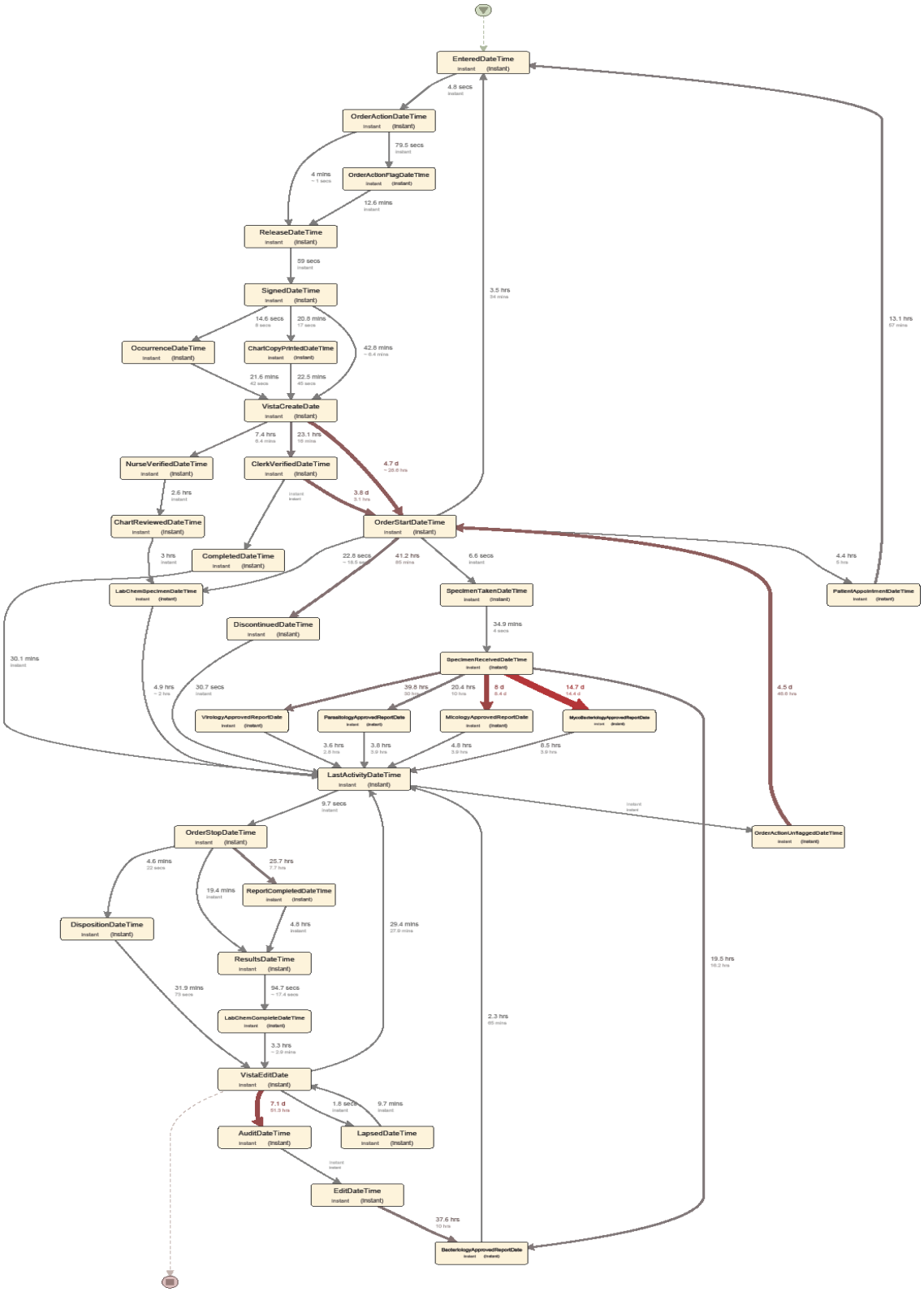

Figure 35. Lab Services performance map after filtering.

4.4.4 Mapping of the OASIS Human Task State transition and Termination States for Laboratory

Table 15 presents the mapping proposal of the CDW Lab Services data domain to the OASIS Human Task State Transitions.

Table 15. CDW Lab Services domain mapping to OASIS Human Task State Transition

OASIS Human Task State Transitions	Lab Service date columns	Lab service orders other events
CREATED	**EnteredDateTime**	
READY	OrderActionDateTime OrderActionFlagDateTime **ReleaseDateTime**	**NW**
RESERVED	**LabChemSpecimenDateTime** SignedDateTime OccurrenceDateTime PatientAppointmentDateTime ChartCopyPrintedDateTime VistaCreateDate NurseVerifiedDateTime	
IN_PROGRESS	OrderStartDateTime SpecimenTakenDateTime SpecimenReceivedDateTime BacteriologyApprovedReportDate ParasitologyApprovedReportDate MycologyApprovedReportDate MycoBacteriologyApprovedReportDate VirologyApprovedReportDate ClerkVerifiedDateTime ChartReviewedDateTime OrderActionUnflaggedDateTime InterventionDate LapsedDateTime VistaEditDate AuditDateTime EditDateTime DispositionDateTime	
COMPLETED	LastActivityDateTime OrderStopDateTime **ResultsDateTime** LabChemCompleteDateTime	

OASIS Human Task State Transitions	Lab Service date columns	Lab service orders other events
	CompletedDateTime ReportCompletedDateTime	
SUSPENDED	DescontinuedHoldUntilDateTime	
FAILED	Completed and there is Discontinued Date Time recorded	
ERROR	Ready and there is Discontinued Date Time recorded	
EXITED	DiscontinuedDateTime Discontinued Date Time recorded (else: not at Completed or Ready)	
OBSOLETE		
CLOSED	VistaEditDate	

The process map with OASIS mapping for the Laboratory dataset is found in Figure 36, showing a sample of 1 million Labs cases. There we can see that most cases completed successfully:

- 740,874 (74%) *Completed*
- 75,695 (~7.5%) *Exited*
- 72,166 (~7.2%) *Error*
- 218 (0.02%) *Failed*

The remining cases are not *complete* cases, as this sample dataset was not filtered.

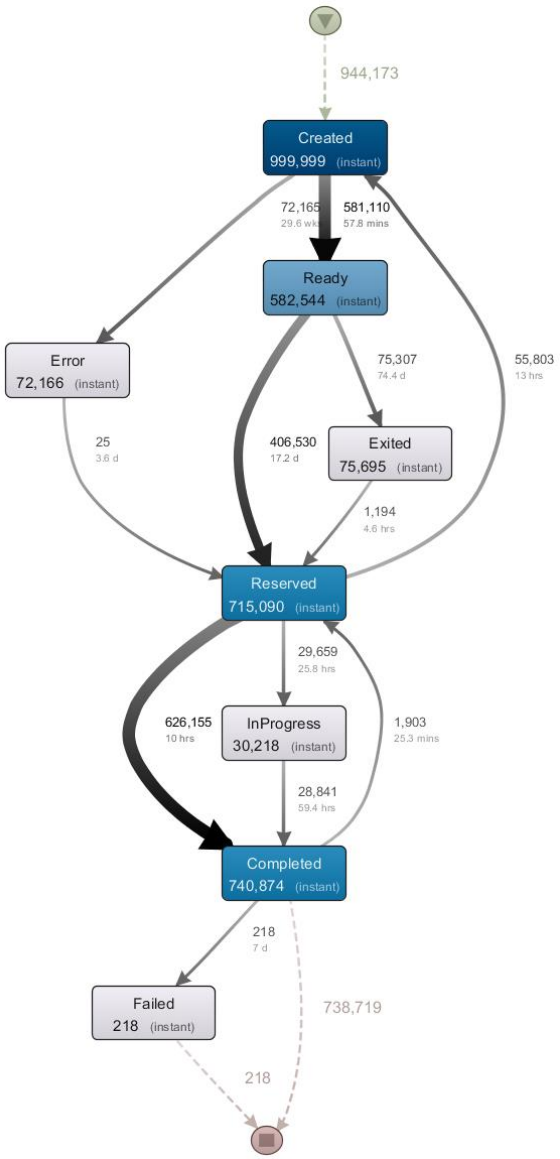

Figure 36. OASIS human task state transitions for a fragment of Lab Services data - frequencies.

Another interesting finding resulted from mapping the OASIS states to a portion of the Labs dataset with 1 million cases. In this example, after filtering endpoints (Initiation event values = *Created, Ready or Reserved*, and Termination event values = *Completed, Error, Exited and Failed*) and excluding weekends and holidays (filtering), we found that cases go from *Ready* to *Reserved* with a mean duration of 4.9 days and median duration of 10 hours, which suggest delays on scheduling. Also, notice the arrow that goes from *Ready* to *Exited* with a mean duration of 8 days and a median duration of 47.9 hours; these were the longest durations. This suggests that if an order is not reserved within a given time threshold, it is likely that it will be terminated with the unsuccessful *exited* termination state. Another duration to notice, because of the large volume of transactions, is from *Ready* to *Completed*, which had a mean of 49.7 hours and a median of 6.9 hours (see Figure 37).

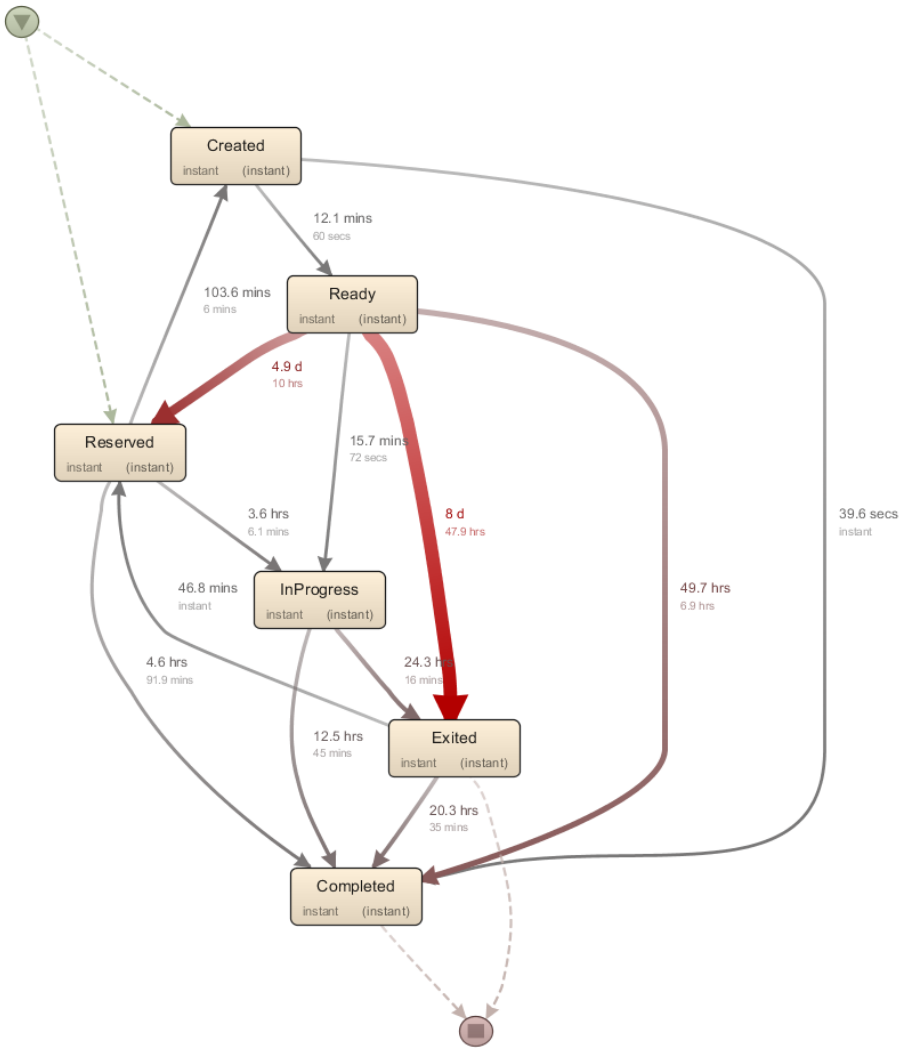

Figure 37. OASIS human task state transitions for Lab Services performance bottlenecks.

Figure 38 presents the Laboratory process model map with the state transition between activities.

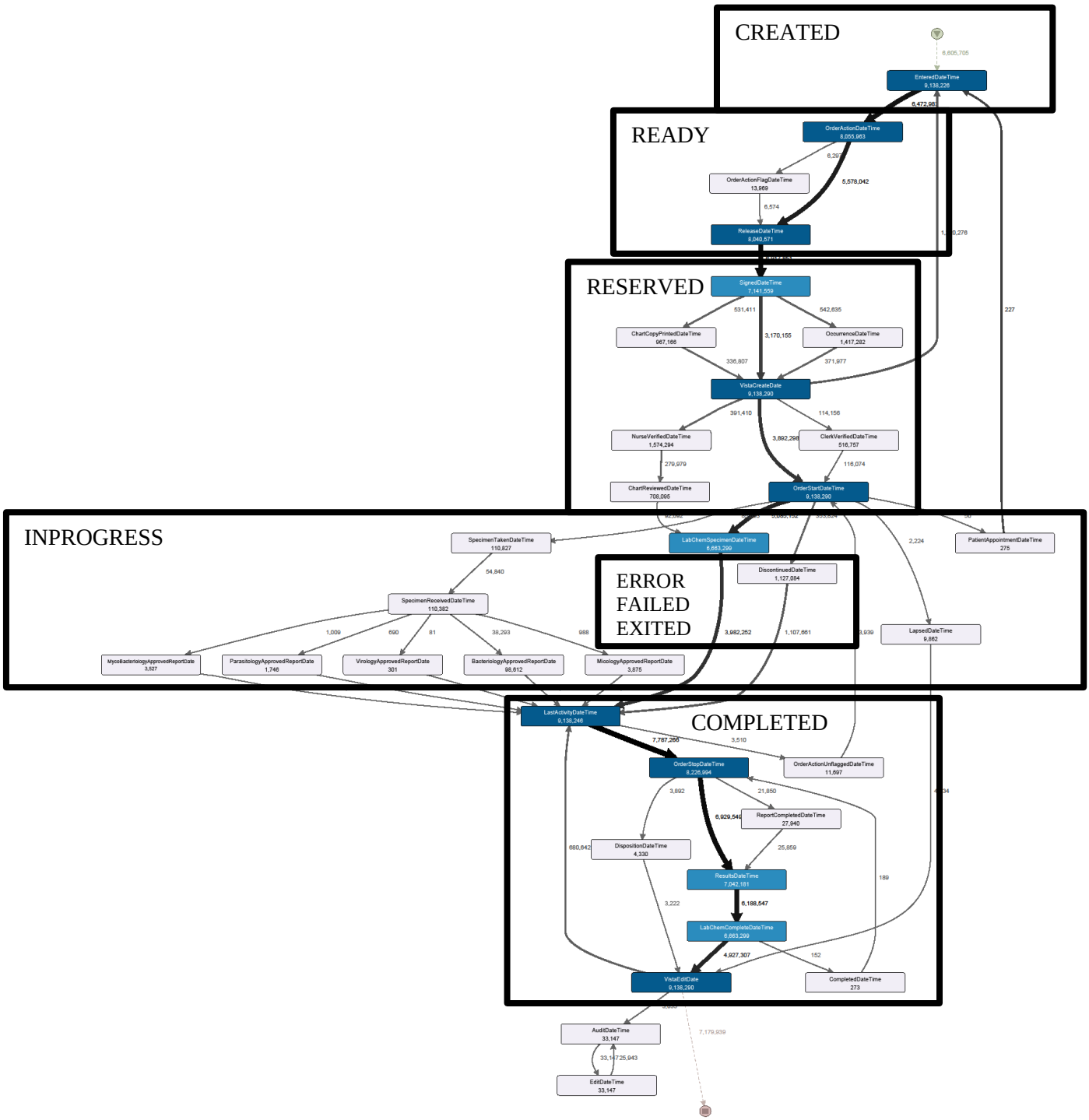

Figure 38. Laboratory Process Model Map presenting state transitions between activities.

4.4.5 Possible Anomalies

This section presents possible anomalies in the Labs dataset, it also describes cases considered outliers.

4.4.5.1 Missing data

We found no data associated with Labs orders in the *CPRSOrder*. *OrderCheckInstance* table *PharmacistCommentsDateTime* column.

4.4.5.2 Duration outliers

In the Lab Services dataset, there are six cases with VistaCreateDate with a date in 1/1/1900 (Table 16).

Table 16. Lab Services dataset duration outliers.

Case ID	Activity	Outlier
10994958, 10994962, 10994966, 10994970, 10994974, 10994978, 20305829	VistaCreateDate	01.01.1900

Table 17 presents Lab Services high-impact areas of performance, after filtering. The data in Table 17 comes from Figure 35.

Table 17. Lab Services high-impact areas of performance—after filtering (from Figure 35).

From	To	Elapsed Time
VistaCreateDate	OrderStartDateTime	4.7 days
ClerkVerifiedDateTime	OrderStartDateTime	3.8 days
SpecimenTakenDateTime	MicologyApprovedReportDate	8 days
SpecimenTakenDateTime	MycoBacteriologyApprovedRerpotDate	14.7 days
VistaEditDate	AuditDateTime	7.1 days

4.4.5.3 Loops in the Lab Services process

In the Labs dataset, there are three cases that have more than 100 events (case IDs 9346795, 8384781, and 5936126). As an example, one case is presented Figure 39, which shows a portion of the long loop between AuditDateTime and EditDateTime activities.

	Activity	Date	Time
1	AuditDateTime	19.01.2007	07:56:44
2	EditDateTime	19.01.2007	07:56:44
3	AuditDateTime	19.01.2007	08:36:48
4	EditDateTime	19.01.2007	08:36:48
5	AuditDateTime	20.01.2007	13:29:39
6	EditDateTime	20.01.2007	13:29:39
7	AuditDateTime	21.01.2007	12:50:02
8	EditDateTime	21.01.2007	12:50:02
9	AuditDateTime	24.01.2007	11:13:01
10	EditDateTime	24.01.2007	11:13:01
11	AuditDateTime	24.01.2007	11:13:10
12	EditDateTime	24.01.2007	11:13:10
13	AuditDateTime	25.01.2007	06:30:46
14	EditDateTime	25.01.2007	06:30:46
15	AuditDateTime	25.01.2007	08:14:32
16	EditDateTime	25.01.2007	08:14:32
17	AuditDateTime	25.01.2007	08:15:02
18	EditDateTime	25.01.2007	08:15:02
19	AuditDateTime	26.01.2007	08:48:37
20	EditDateTime	26.01.2007	08:48:37
21	AuditDateTime	26.01.2007	16:06:29
22	EditDateTime	26.01.2007	16:06:29
23	AuditDateTime	27.01.2007	08:52:37
24	EditDateTime	27.01.2007	08:52:37
25	AuditDateTime	27.01.2007	11:31:19
26	EditDateTime	27.01.2007	11:31:19
27	AuditDateTime	29.01.2007	09:26:54
28	EditDateTime	29.01.2007	09:26:54
29	AuditDateTime	09.02.2007	08:36:15
30	EditDateTime	09.02.2007	08:36:15
31	AuditDateTime	11.02.2007	11:49:46
32	EditDateTime	11.02.2007	11:49:46
33	AuditDateTime	11.02.2007	11:49:58
34	EditDateTime	11.02.2007	11:49:58
35	AuditDateTime	21.02.2007	21:02:31
36	EditDateTime	21.02.2007	21:02:31
37	AuditDateTime	22.02.2007	09:12:10
38	EditDateTime	22.02.2007	09:12:10
39	AuditDateTime	23.02.2007	14:59:33
40	EditDateTime	23.02.2007	14:59:33
41	AuditDateTime	25.02.2007	10:10:04
42	EditDateTime	25.02.2007	10:10:04
43	AuditDateTime	25.02.2007	10:11:28
44	EditDateTime	25.02.2007	10:11:28
45	AuditDateTime	25.02.2007	11:02:50
46	EditDateTime	25.02.2007	11:02:50
47	AuditDateTime	25.02.2007	11:03:28
48	EditDateTime	25.02.2007	11:03:28
49	AuditDateTime	26.02.2007	08:40:40
50	EditDateTime	26.02.2007	08:40:40
51	AuditDateTime	26.02.2007	09:04:28
52	EditDateTime	26.02.2007	09:04:28
53	AuditDateTime	26.02.2007	11:36:43
54	EditDateTime	26.02.2007	11:36:43
55	AuditDateTime	26.02.2007	12:36:33
56	EditDateTime	26.02.2007	12:36:33
57	AuditDateTime	27.02.2007	11:17:08
58	EditDateTime	27.02.2007	11:17:08

Figure 39. Lab services portion of sample looping case with 169 events (case id 5936126).

A sample portion of one of the Labs processes is shown in Figure 40. The names in the nodes are the activities or events, and the thick arrows represent times when a greater number of cases traverses the path between two events. For example, note the bold arrows from OrderStartDateTime to LabChemSpecimenDateTime to OrderStopDateTime,

An interesting part of the path is shown in Figure 40, in which paths flow in cycles. The arrows go from OrderStopDateTime to LastActivityDateTime and back to OrderStopDateTime. Looking at the event sequence cases, we see that both events have the same timestamps; thus, they appear one after the other in the event sequence.

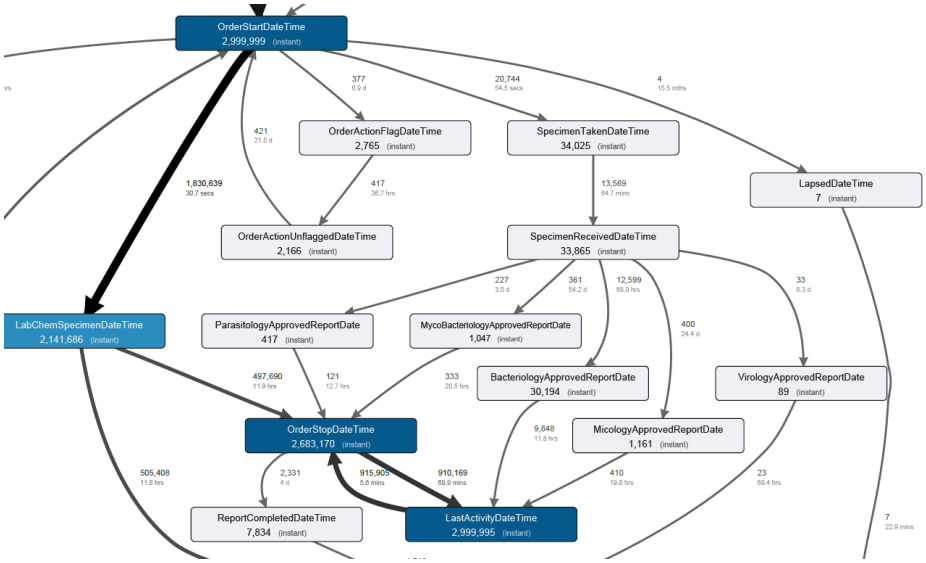

Figure 40. Laboratory Services data flow fragment.

In addition, specific cases of activities related to different lab services types can be observed by following the path from OrderStartDateTime to SpecimenTakenDateTime, to SpecimenReceivedDateTime, to ParasitologyApprovedReportDate, or MycoBacteriologyApprovedReportDate, or BacteriologyApprovedReportDate, or MicologyApprovedReportDate, or VirologyApprovedReportDate to LastActivityDateTime, or OrderStopDateTime. This fragment illustrates the microbiology data domain interacting with the CPRSOrder data domain. Further, more focused, analysis is suggested in this part of the Labs workflow.

4.5 RXOUT (OUTPATIENT PRESCRIPTION MEDICATIONS)

The RxOut domain, outpatient prescription medications, also generated a large dataset for this study (See Table 1). We found that the RxOut domain is particularly complex. Even after the dataset was sorted by date and event, there were still close to 1.5 million variants (See Table 2). During this study, the RxOut domain was generated multiple times as well; and each time, the number of variant cases was close to 1.5 million. Consequently, earlier during the analysis, we divided the RxOut dataset in two parts. For the reason of the time required to generate the process map, because of its size and all of its activities and all paths, we did not include it in this document. Those earlier generated process maps are found in Figure 41 and Figure 42.

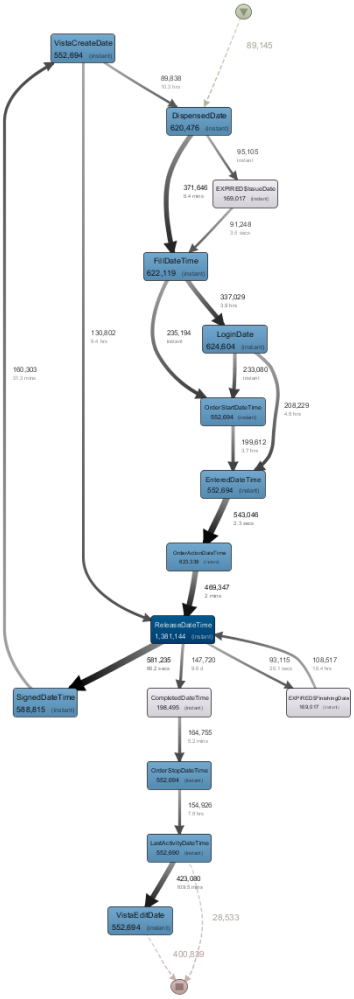

Figure 41. RxOut dataset part 1, most common path.

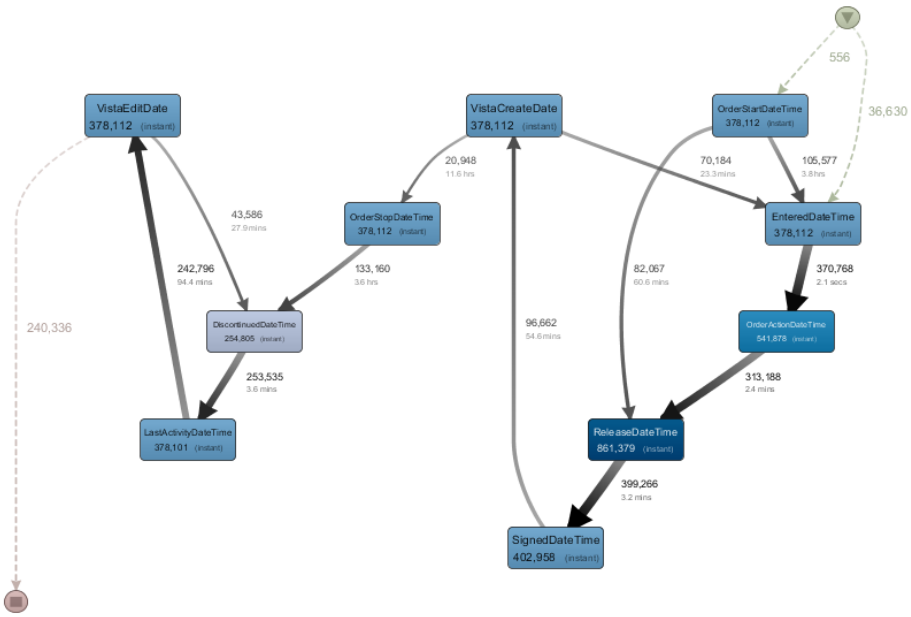

Figure 42. RxOut dataset part 2, most common path.

4.5.1 Process Model Map

We were able analyze the whole RxOut dataset. We compared the process maps generated by randomly dividing the RxOut dataset in two parts (see Figure 41 and Figure 42) with the process map of the whole dataset (see Figure 43), we can see important similarities, which validates the process map generated on each portion of the whole RxOut dataset. Further study is needed to understand the RxOut domain variants.

The raw process map for RxOut consisted of ~ 130 million events, ~ 5.5 million cases, ~1.5 million variants, and 57 distinct activities (see Table 1). The raw process map for the RxOut dataset is shown in Figure 43. Figure 43 shows the path that most cases take with darker blue boxes and bold and thick arrows. The numbers near the boxes and arrows are the absolute frequencies, i.e., how many times the activity was performed in total.

After generating the process model map, we realized how different the RxOut dataset is from the other datasets. Sorting the RxOut sequences of events by event and date decreased considerably the number of case variants, but there were still 1.5 million variants! RxOut presents so much diversity in the process paths of the orders that only 17% of the cases and 13% of the events are contained in the model case. Because of the unique complexity of the RxOut dataset, we plan to study different approaches to simplify the process map.

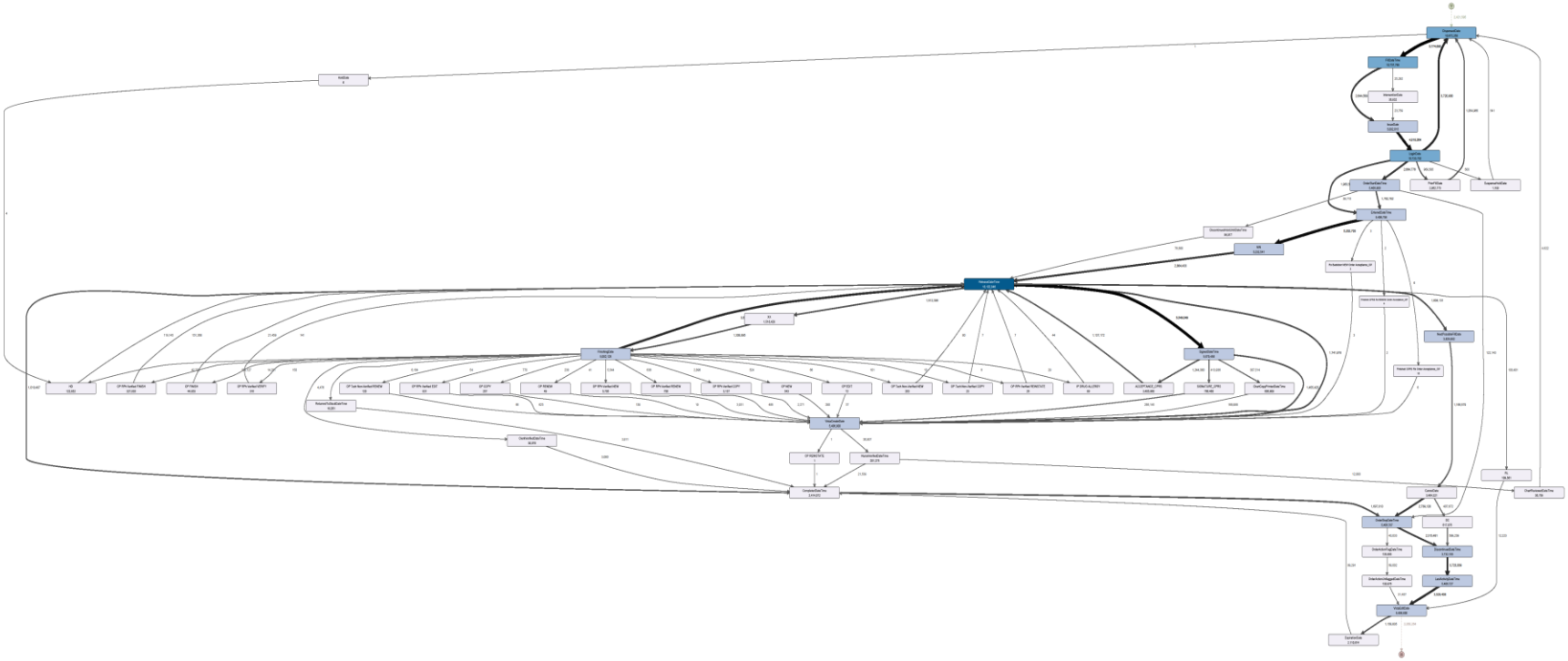

Figure 43. RxOut process map —all activities, before filtering.

4.5.2 Frequency Metrics

The list of activities in the RxOut dataset is found in Table 18, which presents also the frequencies, and the relative frequency percentage.

Table 18. RxOut dataset activities, frequency and relative frequency

Activity	Frequency	Relative frequency
ReleaseDateTime	18,102,846	13.92%
FillDateTime	10,737,765	8.25%
LoginDate	10,733,752	8.25%
DispensedDate	10,672,250	8.20%
SignedDateTime	5,579,455	4.29%
VistaCreateDate	5,406,800	4.16%
OrderStartDateTime	5,406,800	4.16%
VistaEditDate	5,406,800	4.16%
EnteredDateTime	5,406,798	4.16%
OrderStopDateTime	5,406,787	4.16%
LastActivityDateTime	5,406,727	4.16%
NW	5,232,541	4.02%
FinishingDate	5,082,125	3.91%
IssueDate	5,082,015	3.91%
DiscontinuedDateTime	3,732,155	2.87%
NextPossibleFillDate	3,625,882	2.79%
CancelDate	3,464,321	2.66%
ACCEPTANCE_CPRS	3,405,960	2.62%
PriorFillDate	2,862,773	2.20%
CompletedDateTime	2,414,072	1.86%
ExpirationDate	2,110,614	1.62%
XX	1,518,429	1.17%
SIGNATURE_CPRS	766,460	0.59%
DC	617,876	0.47%
ChartCopyPrintedDateTime	606,908	0.47%
OP RPh Verified FINISH	327,835	0.25%
NurseVerifiedDateTime	201,275	0.15%
OrderActionFlagDateTime	135,893	0.10%
OrderActionUnflaggedDateTime	130,876	0.10%
HD	122,852	0.09%
RL	109,561	0.08%
DiscontinuedHoldUntilDateTime	84,917	0.07%
OP FINISH	44,853	0.03%
InterventionDate	36,832	0.03%
ChartReviewedDateTime	36,759	0.03%
ClerkVerifiedDateTime	30,070	0.02%
ReturnedToStockDateTime	18,051	0.01%

OP RPh Verified NEW	5,705	0%
OP RPh Verified COPY	3,137	0%
SuspenseHoldDate	1,188	0%
OP RPh Verified EDIT	831	0%
OP RPh Verified RENEW	700	0%
OP NEW	543	0%
OP RPh Verified VERIFY	316	0%
OP COPY	257	0%
OP Tech Non-Verified NEW	203	0%
OP Tech Non-Verified RENEW	120	0%
IP DRUG ALLERGY	89	0%
OP EDIT	72	0%
OP RENEW	49	0%
OP RPh Verified REINSTATE	26	0%
OP Tech Non-Verified COPY	23	0%
Finished CPRS Rx Order Acceptance_OP	18	0%
HoldDate	6	0%
Finished CPRS Rx RENEW Order Acceptance_OP	6	0%
Rx Backdoor NEW Order Acceptance_OP	3	0%
OP REINSTATE	1	0%

Figure 44 presents the RxOut Pareto chart of frequency by activities after filtering. It also shows that RxOut had a total of 52 different activities. For those activities, the minimal frequency was 3 and the median frequency was 37,419. Figure 44 also shows the mean frequency, the maximal frequency, and the frequency standard deviation. Further study is needed to understand the RxOut domain variants.

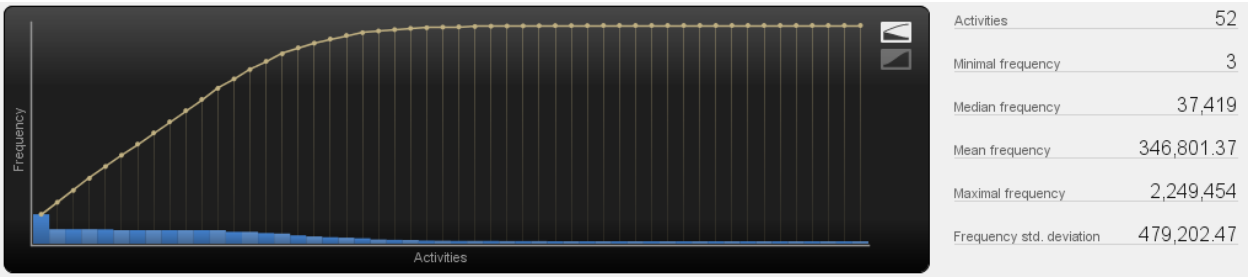

Figure 44. RxOut Pareto chart—frequency by activities.

In addition, Figure 45 shows RxOut frequency metrics after filtering. RxOut has more than 18 million events, about 933 thousand cases and 52 activities.

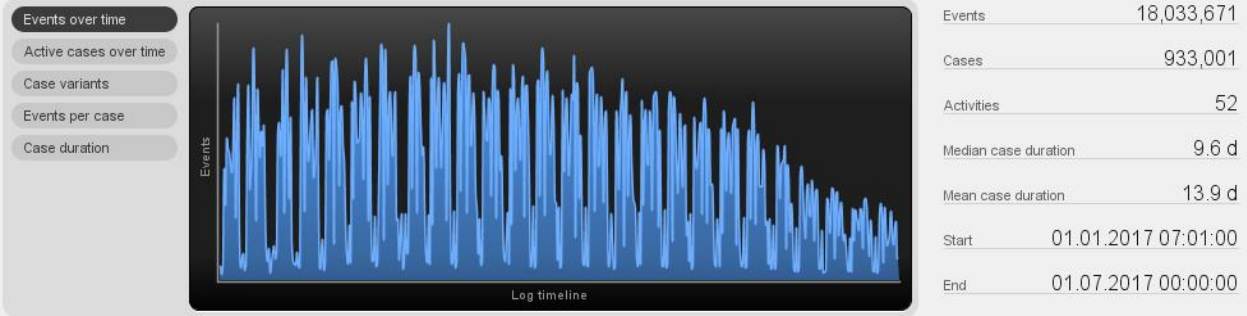

Figure 45. RxOut stats—after filtering.

4.5.3 Performance Metrics

The mean case duration in RxOut datset is 13.9 days, after filtering (the median is 9.6 days) as shown in Figure 45. RxOut stats—after filtering.

We applied time filters to the RxOut dataset, as was done for previous datasets, to identify the time duration of the most common path and other duration statistics. Figure 46 shows the performance process map after filtering. Note that in Figure 46, the bolder, thicker red arrows show the high-impact areas of performance. In Figure 46, the larger numbers that follow the arrows are the mean durations, and the smaller numbers under the mean durations are the median durations. Information in Table 2 and Figure 46 are taken as a model case and as a base to identify outliers and cases that do not conform with the model case in order to identify outliers.

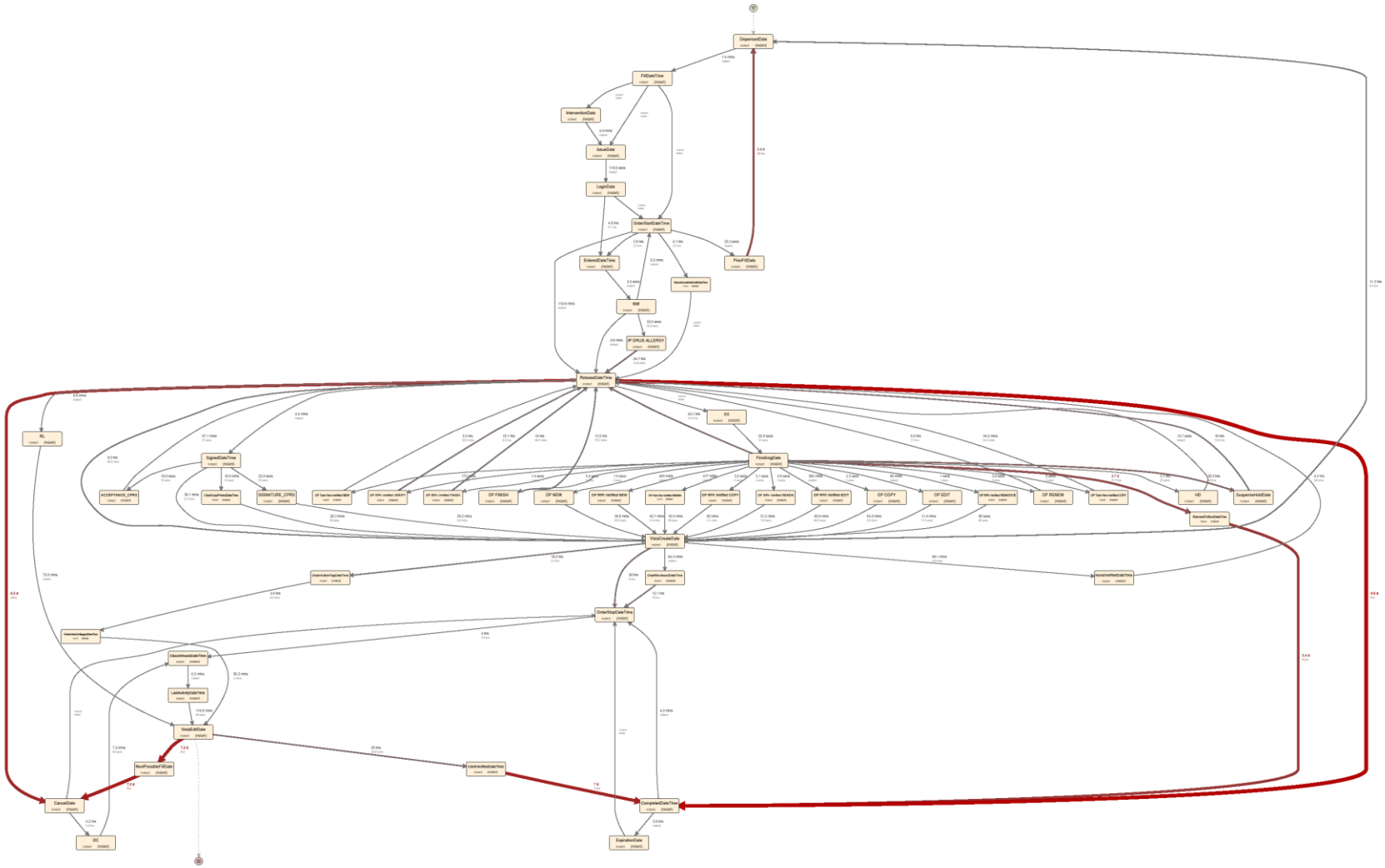

Figure 46. RxOut performance process map—after filtering.

4.5.4 Mapping to OASIS Human Task State Transition and Termination States for RxOut

Because of the abundance of the data in the RxOut dataset, we are still in the process to create process model maps for this dataset. Table 19 presents the mapping proposal of the RxOut data domain to the OASIS Human Task State Transitions.

Table 19. RxOut dataset mapping to OASIS Human Task State Transitions.

OASIS Human Task State Transitions	RxOut date columns	Rxout activities and other events
CREATED	LoginDate OrderStart EnteredDateTime	
READY	OrderActionDateTime OrderStartDateTime	NW
RESERVED	ReleaseDateTime SignedDateTime VistaCreateDate	
IN_PROGRESS	OccurrenceDateTime DispensedDate FillDateTime LastDispensedDate NextPossibleFillDate PriorFillDate ExpirationDate ReturnedToStockDateTime ValidatedDateTime PatientAppointmentDateTime LapsedDateTime VistaEditDate NurseVerifiedDateTime ClerkVerifiedDateTime ReleaseDateTime ChartReviewedDateTime OrderActionFlagDateTime OrderActionUnflaggedDateTime ChartCopyPrintedDateTime InterventionDate IssueDate	RxStatus: ACTIVE
COMPLETED	ResultsDateTime CompletedDateTime OrderStopDateTime LastActivityDateTime	
SUSPENDED	HoldDate SuspenseHoldDate FinishingDate	

OASIS Human Task State Transitions	RxOut date columns	Rxout activities and other events
	DescontinuedHoldUntilDateTime	
FAILED/ERROR	CancelDate DiscontinuedDateTime	RxStatus: DISCONTINUED DISCONTINUED (EDIT) DISCONTINUED BY PROVIDER
EXITED	DiscontinuedDateTime	RxStatus: DELETED EXPIRED DISCONTINUED
OBSOLETE	DiscontinuedDateTime	RxStatus: DISCONTINUED DISCONTINUED (EDIT) DISCONTINUED BY PROVIDER
CLOSED	VistaEditDate	

In Figure 47 the the OASIS human task state transitions has been identified on top of the RxOut process model map.

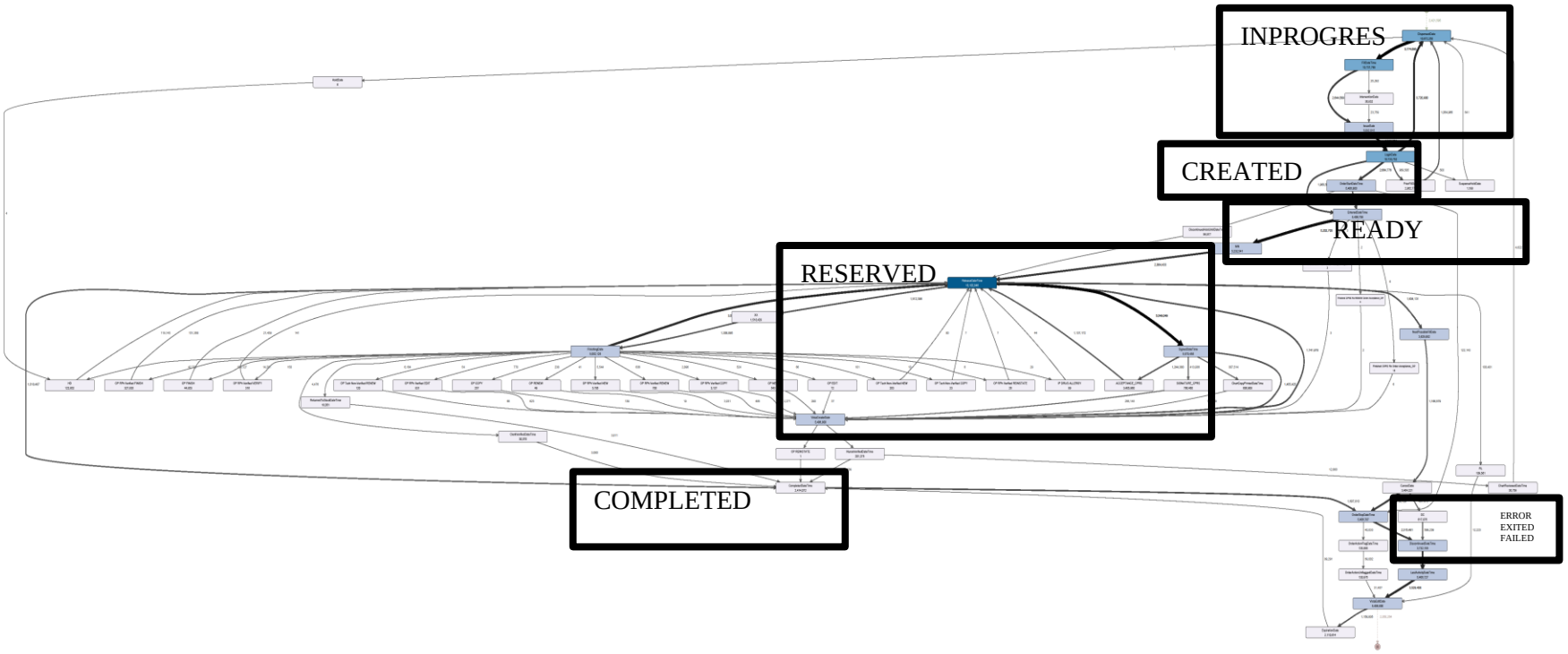

Figure 47. RxOut Process Map presenting state transitions between activities.

4.5.5 Possible Anomalies

This section presents RxOut process deviations and outliers.

4.5.5.1 Missing data

We found no data associated with RxOut orders in the CPRSOrder. OrderCheckInstance table PharmacistCommentsDateTime column.

4.5.5.2 Duration outliers

In RxOut dataset 1, there are 2,164 cases with an ExpirationDate in 31.12.1899. In addition, there are 1,156 cases that end in 25.12.2110 (Table 20).

Table 20. RxOut dataset1 duration outliers.

Case ID	Activity	Outlier
22541051 (sample – too many to list)	ExpirationDate	31.12.1899
22140698 (sample – too many to list)	ExpirationDate	25.12.2110

4.5.5.3 Performance high-impact areas

Table 21 presents RxOut high-impact areas of performance, after filtering. The data in Table 21 comes from Figure 46. RxOut performance process map—after filtering.

Table 21. RxOut high-impact areas of performance—after filtering (from Figure 46)

From	To	Elapsed time
PriorFillDate	DispensedDate	3.4 days
ReleaseDateTime	CompletedDateTime	9.8 days
FinishingDateTime	ReturnedToStockDateTime	71.7 hours
ReturnedToStockDateTime	CompletedDateTime	5.4 days
ClerkVerifiedDateTime	CompletedDateTime	7 days
ReleaseDateTime	CancelDate	6.5 days
VistaEditDate	NextPossibleFillDate	7.2 days
NextPossibleFillDate	CancelDate	7.5 days

4.5.5.4 Loops in the RxOut process

Cases with a large number of events were also found in RxOut, such as case 15455364. A portion of this case is shown in Figure 48.

	Activity	Date	Time
78	RL	26.10.2017	08:46:00
79	ReleaseDateTime	26.10.2017	08:47:18
80	HD	27.10.2017	09:20:00
81	ReleaseDateTime	27.10.2017	09:20:00
82	DispensedDate	17.11.2017	00:00:00
83	FillDateTime	17.11.2017	00:00:00
84	LoginDate	17.11.2017	00:00:00
85	ReleaseDateTime	17.11.2017	12:23:34
86	DiscontinuedHoldUntilDateTime	17.11.2017	12:27:00
87	ReleaseDateTime	17.11.2017	12:27:00
88	RL	17.11.2017	12:27:00
89	HD	17.11.2017	14:49:00
90	ReleaseDateTime	17.11.2017	14:49:00
91	DispensedDate	14.12.2017	00:00:00
92	FillDateTime	14.12.2017	00:00:00
93	LoginDate	14.12.2017	00:00:00
94	DiscontinuedHoldUntilDateTime	14.12.2017	14:28:00
95	ReleaseDateTime	14.12.2017	14:28:00
96	RL	14.12.2017	14:28:00
97	HD	14.12.2017	16:32:00
98	ReleaseDateTime	14.12.2017	16:32:00
99	DispensedDate	09.01.2018	00:00:00
100	FillDateTime	09.01.2018	00:00:00
101	LoginDate	09.01.2018	00:00:00
102	PriorFillDate	09.01.2018	00:00:00
103	OrderActionFlagDateTime	09.01.2018	13:56:00
104	DiscontinuedHoldUntilDateTime	09.01.2018	14:22:00
105	ReleaseDateTime	09.01.2018	14:22:00
106	RL	09.01.2018	14:22:00
107	ReleaseDateTime	09.01.2018	14:24:10
108	HD	09.01.2018	14:55:00
109	OrderActionUnflaggedDateTime	09.01.2018	14:55:00
110	ReleaseDateTime	09.01.2018	14:55:00
111	FillDateTime	05.02.2018	00:00:00
112	LoginDate	05.02.2018	00:00:00
113	OrderActionFlagDateTime	05.02.2018	15:00:00
114	DiscontinuedHoldUntilDateTime	05.02.2018	16:05:00
115	ReleaseDateTime	05.02.2018	16:05:00
116	RL	05.02.2018	16:05:00
117	OrderActionUnflaggedDateTime	05.02.2018	16:06:00
118	DispensedDate	07.02.2018	00:00:00
119	ReleaseDateTime	07.02.2018	22:23:59
120	NextPossibleFillDate	01.03.2018	00:00:00
121	CancelDate	05.03.2018	00:00:00
122	OrderStopDateTime	05.03.2018	00:00:00
123	VistaEditDate	05.03.2018	11:49:12
124	DiscontinuedDateTime	05.03.2018	12:37:00
125	LastActivityDateTime	05.03.2018	12:37:00
126	ExpirationDate	15.03.2018	00:00:00
127	ExpirationDate	17.04.2018	00:00:00
128	ExpirationDate	11.05.2018	00:00:00
129	ExpirationDate	12.06.2018	00:00:00
130	ExpirationDate	27.07.2018	00:00:00
131	ExpirationDate	20.08.2018	00:00:00
132	ExpirationDate	19.09.2018	00:00:00
133	ExpirationDate	25.10.2018	00:00:00
134	ExpirationDate	16.11.2018	00:00:00
135	ExpirationDate	13.12.2018	00:00:00
136	ExpirationDate	08.01.2019	00:00:00

Figure 48. RxOut large number of events case fragment (case 15455364).

5. SUMMARY AND CONCLUSIONS

5.1 SUMMARY

The following are the principal elements of the efforts in this research effort:

1. Several challenges encountered during our attempts to extract data and to format and generate event sequence logs in healthcare data were identified and are discussed.

2. The approach adopted for interrogating the CDW database focused on a process mining methodology [3-5] that was slightly modified to accommodate the needs of our study.

3. Additional software tools, metrics, and filters were employed to augment the modified process mining methodology, allowing us to quickly examine large volumes of data to identify more specific research questions. Those supporting elements included

    - Using the OASIS human task state transition diagram in data mapping to further simplify the identification of specific cases

    - Using the Disco tool [21] to visualize and filter the datasets, which proved to be extremely helpful, fast, and reliable

    - Obtaining two types of metrics from the process maps generated with Disco: frequency metrics and performance metrics.

A case study was presented in which the combined ORNL approach (e.g., process mining methodology, software tools) was applied to the CDW to evaluate the performance of that approach.

5.2 ANALYSIS RESULTS

1. We compared completed cases against failed, exited, and error cases. We observed that completed cases:

a) followed similar patterns,
b) had activities that mostly followed a sequence with few loops,
c) spent similar amounts of time between steps,
d) had regular times from start to end.

2. In contrast, the failed, exited or error cases often

i. uses unusual activities,
ii. included zero timestamps (i.e., dates far in the past or far in the future),
iii. present spider-like shapes (i.e., from one activity, there were several outgoing paths),
iv. were incomplete (i.e., did not start/end on the same activities as completed cases or missed other activities that completed cases had) presented extra loops.

3. An event sequence that presents numerous iterations within a loop may be considered an outlier. We observed that those cases that kept looping usually were discontinued and consequently, destined to fail. These are examples of possible anomalies that could become HIT hazards.

5.3 STRATEGIES TO REDUCE COMPLEXITY THAT WORKED FOR US

1. Sorting event sequences not only by date but also by activity reduced the number of variants to about a quarter of the dataset.

2. It is important to clearly identify each activity (event) with a unique name to avoid the creation of false loops (i.e., spider-like activities—repeating activity in multiple contexts throughout the process creates false loops that are not actual loops, but a lack of specificity in the activity names creates spider-like activities).

3. Mapping the dataset activities to the OASIS human task state transition diagram was extremely helpful for rapidly identifying and visualizing those cases that ended in unsuccessful termination states, such as *Failed*, *Exited*, and *Error*.

5.4 DATA QUALITY PROBLEMS ENCOUNTERED

1. Our study detected Zero Timestamps [faulty timestamps that are far in the past [e.g., 1900] or far in the future [e.g., 2100]) in the four CDW data domains. An example: an activity with date: 01.01.1900 affected the overall study and analysis by impacting durations, variants, and process maps. Those cases were filtered to identify the real durations.

2. Our study detected redundancy of data where some values refer to the same classification but are entered with different names. An example: complete vs. Complete vs. COMPLETE vs. COMPLETED vs. completed vs. COM.

3. Daylight savings time can affect the analysis and impact process maps generated in Disco. Attention should be placed to activities recorded during daylight savings time thresholds to prevent possible minor but impactful changes in the resulting process maps and metrics.

5.5 CONCLUSION

We performed an evidence-based study to effectively identify process models and to define metrics of frequency and performance for four health care domains: Consults, Radiology, Labs, and RxOut. Additionally, we classified the termination classes of the different cases by mapping to the OASIS[6] standard. We demonstrated, via process mining, that extracted raw data can aid understanding of the flow of information in different clinical order processes. We showed that application of the step-by-step approach to discover processes in raw EHR data can be extremely effective in revealing irregular state transitions in the data and understanding clinical order information flows that are not apparent in analyzing the CDW as is, without feature extraction.

Several conclusions were drawn from the results of the case study, including the following:

1. Process models, as well as defined metrics for frequency and performance, were discovered for four different CDW data domains (Consults, Radiology, Labs, and RxOut).

[6] **OASIS** is a nonprofit consortium that drives the development, convergence and adoption of open standards for the global information society. https://www.oasis-open.org/org

2. Consults and Radiology process maps had fewer case variants than Labs and RxOut.

3. An initial set of metrics and key performance indicators were outlined as baselines.

4. The frequencies are higher on those activities related to the CPRSOrder domain, which suggests higher use, because this, the VA's IT organization should give special attention to the IT resources that maintain data in this domain, including storage, network bandwidth, memory, and processor speed.

5. **Combining event sequences with a powerful visualization tool like process mining can reveal important aspects of the data that are hard to interpret otherwise!**

The data-driven exploration of the case study yielded other important information summarized in the tables presented herein, but there is still much to learn from the process maps generated by the analysis. More analysis of specific, smaller cases is needed to suggest process improvements and to determine which cycles make the most sense to the SMEs. However, being able to identify those cycles and how the data flows in cycles is important for identifying possible anomalies. Each case needs to be studied to determine whether it is a normal operation cycle, if the flow cycle needs improvement, and which cycles are not normal.

6. FUTURE WORK

There is much to be done to identify more opportunities of improvement and possible anomalies in this project. The following is a list of tasks for fiscal year 2021.

1. Event data selection, preparation and refinement: Throughout FY 2019, datasets for consults, radiology, lab services and prescriptions have been already created for this project; however, there is much to be learned and more refinement is needed to continue identifying possible anomalies.

2. Process model discovery and description: This task refers to identifying a model process, i.e., the most common path or sequence of events per data domain or per use case. It involves describing the initial, intermediate, and final events on each model process and failure path, according to the OASIS state transition diagram. Process model discovery will require the following activities:
   - Event log inspection: investigating by observing the behavior of the events to determine the path that most events follow. Focus on specific events that are meaningful to HIT for hazard detection.
   - Event log filtering: Two possible methodologies can be used for filtering:
     – Removing incomplete sequences, based in comparing sequences to the most common path and use case
     – Removing events that are not very meaningful to the use case or the data domain, or that do not follow the most common path.
   - Re-inspecting event logs: repeating the steps above as needed

3. Conformance checking (using the inductive visual miner algorithm): This algorithm refers to confronting the process model with all variants of event sequences in the data set to identify where they agree and where they disagree. This task involves the following:
   - Documenting all observations related to
     – cases of deviations of the process model
     – events skipped
     – misalignments exiting in the data.
   - Documenting the following to the extent allowed by the time and project's budget:

  - identifying root cause of deviation
  - suggesting potential process improvements
- Performing global statistics on the event logs for the following:
  - percentage of correct behavior
  - mean duration
  - mean duration between events
  - other statistics

4. KPI identification: Conformance checking will generate the identification of KPIs for each dataset or use case. This task will include documenting percentages, statistics, and other measurements of the different events, sequences, and other HIT features to track through this study. In addition, for all identified KPIs, a set of desired target values will be defined. All KPI measurements obtained that are different from the target values will be to be flagged as possible anomalies.

5. Presentation and discussion of observations, stats and possible anomalies: This task implies meetings, conference call presentations, and discussions of the work performed. In addition, possible anomalies identified can be implemented as possible hazards for the hazard detector dictionary, which can be added to the prototype tool being developed in this project.

6. Formatting datasets to be shared with the VA and other healthcare open data source communities: This task refers to preparing, formatting and negotiating to facilitate access to the datasets used in this study by the VA and a healthcare open source community selected by the VA team. Making these datasets available to replicate results and statistics and to facilitate other studies related to process mining and artificial intelligence for health science projects will provide a legacy of this project for future analysis.

7. Publications and report document: This task refers to writing abstracts and papers for conferences in healthcare data and process analytics and a report document describing the lessons learned and the execution of these tasks.

7. REFERENCES

1. Olufemi A. Omitaomu, et al., *Real-Time Automated Hazard Detection Framework for Health Information Technology Systems.* Health Systems, 2019.
2. Ozgur Ozmen and e. al, *Event Sequence Extraction to Enable Hazard Detection in Health Information Technology Systems Through Process Mining*, in *AMIA Annual Meeting* 2019.
3. van Eck, M.L., et al. *PM2: A Process Mining Project Methodology*. in *International Conference on Advanced Information Systems Engineering*. 2015. Springer.
4. *Process Mining in Healthcare*. 2019; Available from: https://www.futurelearn.com/courses/process-mining-healthcare/.
5. Mans, R.S., W.M. Van der Aalst, and R.J. Vanwersch, *Process mining in healthcare: evaluating and exploiting operational healthcare processes*. 2015: Springer.
6. Rojas, E., et al., *Process mining in healthcare: A literature review.* Journal of biomedical informatics, 2016. **61**: p. 224-236.
7. Lowry, S.Z., et al. *Integrating electronic health records into clinical workflow: An application of human factors modeling methods to ambulatory care*. in *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*. 2014. SAGE Publications Sage India: New Delhi, India.

8. Mans, R.S., et al. *Application of process mining in healthcare–a case study in a dutch hospital*. in *International joint conference on biomedical engineering systems and technologies*. 2008. Springer.
9. Mans, R.S., et al., *Process mining in healthcare: Data challenges when answering frequently posed questions*, in *Process Support and Knowledge Representation in Health Care*. 2012, Springer. p. 140-153.
10. Rovani, M., et al., *Declarative process mining in healthcare.* Expert Systems with Applications, 2015. **42**(23): p. 9236-9251.
11. Rebuge, Á. and D.R. Ferreira, *Business process analysis in healthcare environments: A methodology based on process mining.* Information systems, 2012. **37**(2): p. 99-116.
12. Kaymak, U., et al. *On process mining in health care*. in *2012 IEEE international conference on Systems, Man, and Cybernetics (SMC)*. 2012. IEEE.
13. Gupta, S., et al., *Workflow and process mining in healthcare.* Master's Thesis, Technische Universiteit Eindhoven, 2007.
14. Russell, N., et al. *Workflow resource patterns: Identification, representation and tool support*. in *International Conference on Advanced Information Systems Engineering*. 2005. Springer.
15. Van Der Aalst, W., K.M. Van Hee, and K. van Hee, *Workflow management: models, methods, and systems*. 2004: MIT press.
16. Van der Aalst, W., T. Weijters, and L. Maruster, *Workflow mining: Discovering process models from event logs.* IEEE Transactions on Knowledge and Data Engineering, 2004. **16**(9): p. 1128-1142.
17. Van der Aalst, W.M., *The application of Petri nets to workflow management.* Journal of circuits, systems, and computers, 1998. **8**(01): p. 21-66.
18. van Der Aalst, W.M., *Workflow patterns.* Encyclopedia of Database Systems, 2009: p. 3557-3558.
19. Günther, C.W. and W.M. Van Der Aalst. *Fuzzy mining–adaptive process simplification based on multi-perspective metrics*. in *International conference on business process management*. 2007. Springer.
20. Van der Aalst, W., *Data science in action*, in *Process Mining*. 2016, Springer. p. 3-23.
21. Anne Rozinat, C.W.G., and Rudi Niks, *Process Mining in Practice*. 2019.
22. OASIS, *Web Services Human Task (WS-Human Task) Specification Version 1.1 - Committee Specification Draft 12 / Public Review Draft 05*. 2012.
23. Parmenter, D., *Key performance indicators: developing, implementing, and using winning KPIs*. 2015: John Wiley & Sons.
24. Günther, C.W. and E. Verbeek, *XES standard definition.* Fluxicon Process Laboratories (November 2009), 2014.
25. Van Dongen, B.F., et al. *The ProM framework: A new era in process mining tool support*. in *International conference on application and theory of petri nets*. 2005. Springer.

APPENDIX A. OASIS WS–HUMAN TASK OVERVIEW

APPENDIX A. OASIS WS–HUMAN TASK OVERVIEW

The WS–Human Task Specification [22] is an OASIS (https://www.oasis-open.org/org) standard. OASIS is a nonprofit consortium focused on the "development, convergence, and adoption of open standards for the global information society." In the OASIS–Human Task specification, the concept of "human tasks" is used to specify work that must be accomplished by people.

The WS–Human Task specification introduces the definition of human tasks, including their properties, behavior, and a set of operations used to manipulate human tasks. The focus is on human tasks because human tasks are part of business processes. However, those tasks can also be used to design human interactions that are invoked as services, whether as part of a process or otherwise.

In addition, the specification introduces a coordination protocol to control the autonomy and life cycle of service-enabled human tasks in an interoperable manner by Web Services. For the purposes of this effort, we ignored the Web Services protocol part of the standard because the diagram in Figure A-1 presents a standard definition of human tasks that we applied to reverse-engineer clinical processes.

The human task behavior and state transitions in Figure A-1 depict a state diagram with several states and the transitions between them. A walk through the states follows.

1. Upon creation, a task goes into the *Created* state. There is no need for a task owner in this state. The task remains in the state *Created* until it is activated and has potential owners.

2. When a task has multiple potential owners or is assigned to a work queue, it transitions into the *Ready* state, indicating that it can be claimed by one of its potential owners.

3. When the task is claimed by a single owner, it transitions into the *Reserved* state, indicating that it is assigned to a single actual owner. The current actual owner of a human task can release a task to again make it available for all potential owners.

4. Once work is started on a task that is in state *Ready* or *Reserved*, it goes into the *InProgress* state, indicating that it is being worked on.

5. The task will go into *Suspended* state when a suspend operation is invoked or a suspend event is received.

6. On successful completion of the work, the task transitions into the *Completed* final state.

7. On unsuccessful completion of the work, the task transitions into the *Failed* final state.

8. A received exit event will make the task transition into the *Exit* final state.

9. A nonrecoverable error event will make the task transition into the *Error* final state.

10. A skip operation invoked (if the task is skippable) will cause a task to go to the *Obsolete* final state.

The life cycle of subtasks is the same as that of the main task. More information about this diagram can be found in OASIS [22]. We observed that the OASIS standard overlaps with the majority of the computerized physician order entry transactions in VistA (https://www.data.va.gov/dataset/veterans-health-information-systems-and-technology-architecture-vista), providing a good framework for representing CDW data as event sequences.

The work described herein is based on the OASIS WS–Human Task Specification Version 1.1, specifically, the Human Task Behavior and State Transitions diagram shown in Figure A-1.

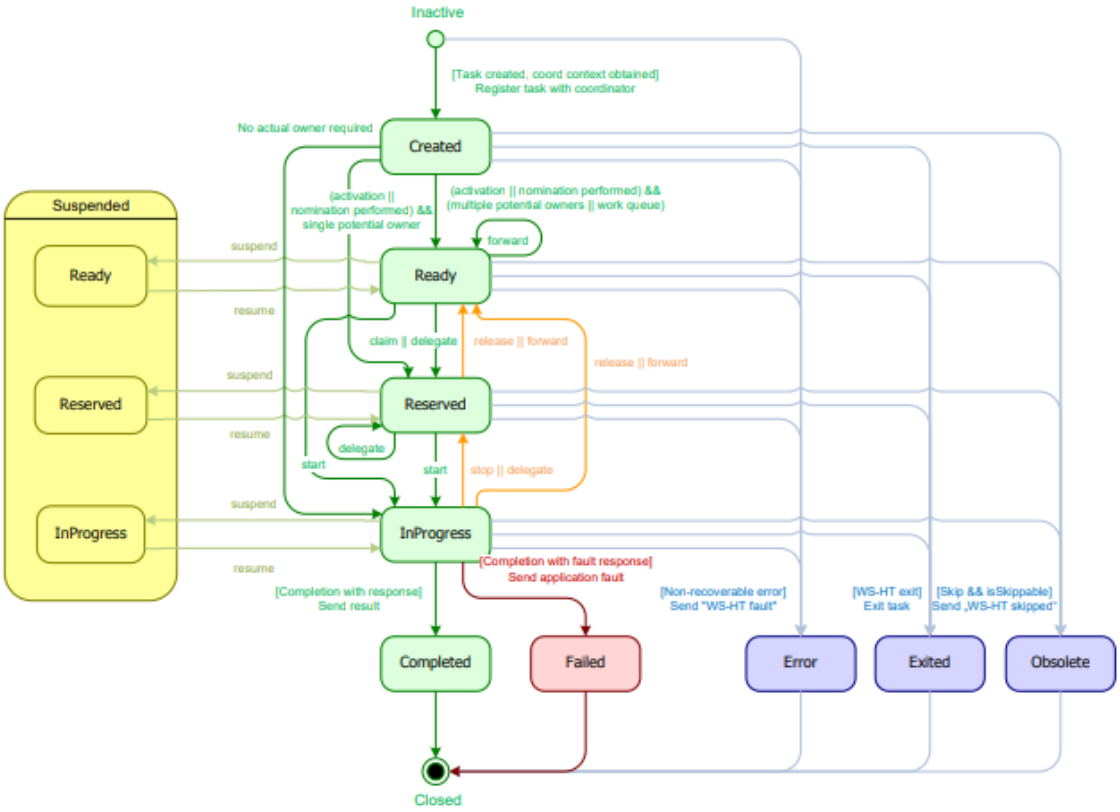

Figure A-1. OASIS WS–Human Task Behavior Version 1.1 and State Transitions Diagram from page 58 of the WS Human Task Specification document.