
SMART SIMULATION AND MODELLING FOR COMPLEX CANCER SYSTEMS



Emma Louise Aspland
School of Mathematics
Cardiff University

A thesis submitted for the degree of
Doctor of Philosophy

July 2021

Abstract

Clinical pathways are an effective and efficient approach in standardising the progression of treatment, to support patient care and facilitate clinical decision making. This research project was funded by KESS2 in collaboration with a company partner - Velindre Cancer Centre (VCC).

This thesis develops efficient and sustainable methods for pathway mapping, modelling and improving, within the context of developing a state-of-the-art decision support tool. A particular focus on lung cancer is considered for method construction and investigations.

The clinical pathways are mapped through representing each pathway as a string of letters. This enabled the development of the modified Needleman-Wunsch metric, to allow for consideration of both data and medical expert information, for the use with k-medoids clustering.

The key contribution of automating the simulation build and necessary input parameters, is developed. Models can be constructed for four routing procedures, namely *Raw Pathways*, *Full Transitions*, *Cluster Transitions* and *Process Medoids*, that explore progressively less complex and varied interpretations of the clinical pathways. Improvements can then be investigated for aligning capacity and demand.

Combining this amounted to the development of Sim.Pro.Flow, an open access decision support tool, that contains all methods discussed in this thesis. The generalised approach allows for these methods, and Sim.Pro.Flow, to be suitably flexible for application with process data from both healthcare and other industries.

Acknowledgements

I would like to acknowledge KESS2 for funding this research, along with Velindre Cancer Centre for their contributions as company partner, including my company supervisors - Phil Webb, Peter Barrett-Lee and Mark Briggs. I would like to thank everyone from all parts of Velindre and the NHS who gave me their time, particularly Mick Button whose expertise, enthusiasm and encouragement was instrumental.

To my supervisors, Paul Harper and Daniel Gartner - I am so fortunate to have had you as my supervisors. All other supervisory teams should aspire to achieve your example. Thank you for always being kind, supportive and understanding.

Thank you to the CRUK team, including Sarie Brice, Tracey England and Edilson Arruda. Acknowledging the collaboration with Edilson Arruda, thank you for taking the time, and having the patience, to share your work with me.

I am thankful to Vince Knight for introducing me to Python. To Nikoletta Glynatsi and Henry Wilde for their generous willingness to answer all my coding questions.

Special thanks to Geraint Palmer for initiating and supporting the collaboration to extend Ciw. His help has been invaluable and his work an inspiration to me.

Thank you to all the PGR students during my time at Cardiff University for creating such a warm and friendly environment.

I am extremely thankful to everyone who initiated and attended WiM (Women in Mathematics). The opportunity to be a part of a group with such a positive and passionate message was an honour.

On a personal note, thank you to Emily Williams for being my research ‘twin’, for travelling the world with me and always being willing to talk no matter the time, topic or place. This friendship is the most valuable bi-product of this endeavour.

To my family, Mum and Chloe, for everything, always, thank you.

Finally, to Matt, there are few words in this thesis more important, yet elusive, than these ones to say thank you. I am forever grateful.

Dedicated to Michael Aspland.

Dissemination

Publications

- Clinical Pathway Modelling: A Literature Review. E. Aspland, D. Gartner and P. Harper. Health Systems, 2021 [15]
- Resource Optimization for Cancer Pathways with Aggregate Diagnostic Demand: A Perishable Inventory Approach. E.F. Arruda, P. Harper, T. England, D. Gartner, E. Aspland, F.O. Ourique and T. Crosby. IMA Journal of Management Mathematics, 2020 [12]
- Modified Needleman-Wunsch Algorithm for Clinical Pathway Clustering. E. Aspland, P.R. Harper, D. Gartner, P. Webb and P. Barrett-Lee. Journal of Biomedical Informatics, 2021 [16]
- Factors Influencing the Delivery of Cancer Pathways: A Summary of the Literature. S. Brice, P. Harper, T. Crosby, D. Gartner, E. Arruda, T. England, E. Aspland and K. Foley. Journal of Health Organization and Management, 2021 [40]

Presentations

- Smart Simulation and Modelling of Complex Cancer Systems, (Poster). ORAHS, Oslo (Norway), August 2018.
- Smart Simulation and Modelling of Complex Cancer Systems, (Poster). OR60, Lancaster (England), September 2018. *Awarded Poster Prize*
- Lung Cancer Clinical Pathway Mining. SIAM Student Chapter National Conference, Manchester (England), June 2019. *NBrown Prize for Best Talk*
- Modelling Lung Cancer Clinical Pathways. EURO, Dublin (Ireland), June 2019.
- Modelling Lung Cancer Clinical Pathways. Single Cancer Pathway Workshop, Cardiff (Wales), July 2019.
- Modelling Lung Cancer Clinical Pathways. ORAHS, Karlsruhe (Germany), August 2019.
- Modelling Lung Cancer Clinical Pathways. EURO PhD Summer School, ‘Operational Research for Value Based Health Care’. Lisbon (Portugal), September 2019.
- Automate Simulation Build. ORAHS, Online via Video Conferencing, July 2020.

Contents

Abstract	i
Acknowledgements	iii
Dissemination	vii
List of Figures	xii
List of Tables	xv
Glossary	xxi
1 Introduction	1
1.1 Cancer Services	2
1.2 Clinical Pathways	3
1.2.1 Formal Definition	4
1.2.2 Cancer Clinical Pathways in the UK	5
1.2.3 Performance Targets	6
1.2.4 Patient Pathways Interpretation - Input Data	7
1.3 Problem Description	9
1.3.1 Preliminary Investigation	11
1.3.2 Decision Support Tool - Sim.Pro.Flow	12
1.4 Research Questions and Structure	13
1.5 Collaborations	15
2 Literature Review	17
2.1 Search Criteria	18
2.2 Previous Research	20
2.3 Exploration of a Sample of Papers	21
2.4 Classification of Literature	23
2.4.1 General Characteristics	24
2.4.2 Medical Context	27
2.4.3 Technical Context	32

2.4.4	Planning Decisions	39
2.5	Conclusions	42
3	Modified Needleman-Wunsch Metric	45
3.1	Introduction	46
3.2	Previous Research	47
3.3	Working Dataset - ‘Dataframe’	48
3.4	Description of Metrics	50
3.5	Properties of Metric	61
3.6	Modified Needleman-Wunsch Algorithm	62
3.7	Case Studies	72
3.7.1	Sample 1: 10 Pathways	73
3.7.2	Sample 2: 16 Pathways	75
3.7.3	Full Data	78
3.8	Sensitivity Analysis	80
3.8.1	Initial Medoids	80
3.8.2	Penalty Values	82
3.8.3	Rankings and Groupings	85
3.9	Conclusions	87
4	Automating the Simulation Build	89
4.1	Introduction	89
4.2	Considerations	90
4.3	Extensions to Ciw - Process Based Routing	91
4.3.1	Introduction to Ciw	92
4.3.2	Initial Investigation	95
4.3.3	Development	98
4.3.4	Custom Ciw	99
4.4	Working Dataset	102
4.5	Validating Input Parameters	104
4.5.1	General Arrivals	104
4.5.2	<i>Raw Pathways</i> Routing Procedure	105
4.5.3	Service	110
4.5.4	Automated Capacity	114
4.5.5	Warm Up	118
4.6	Custom Parameters	121
4.7	Capacity Investigation	123
4.7.1	Method Extension	129
4.7.2	Example	131
4.8	Conclusion and Further Work	134

5	Routing Procedures	137
5.1	Introduction	137
5.2	<i>Full Transitions</i>	139
5.3	<i>Cluster Transitions</i>	142
5.4	<i>Process Medoids</i>	147
5.4.1	Medoids Selection	148
5.4.2	Parameters	154
5.5	Routing Procedures Exploration	159
5.5.1	Overview	159
5.5.2	Results and Discussion	161
5.6	Conclusion and Further Work	166
6	Developing a Decision Support Tool - Sim.Pro.Flow	169
6.1	Introduction	169
6.2	Structure	174
6.3	Sim.Pro.Flow Key Features	176
6.3.1	Outputs Produced	176
6.3.2	Graphical User Interface	178
6.4	Conclusion	182
7	Case Study	185
7.1	Introduction	185
7.2	Individual Adjustment Investigation	187
7.3	Basic Investigation	190
7.4	End Activity Investigation	193
7.5	Target Investigation	195
7.6	Excessive Top Down Investigation	199
7.7	Demand Investigation	204
7.8	Medoids Capacity Investigation	206
7.9	Conclusion	208
8	Conclusion	211
8.1	Summary	211
8.2	Contributions	214
8.3	Further Work	215
	Bibliography	216
	Appendices	243
A	Preliminary Investigation	245

B Literature Review Tables	249
C Simulation	255
D Sim.Pro.Flow	267
E Case Study	279

List of Figures

1.1	Early “Care Pathway” - Extracted from The NHS Cancer Plan [209].	5
1.2	Example of Data Type 1.	8
1.3	Example of Data Type 2.	8
2.1	Diagram Detailing the Search Process.	20
2.2	Frequency of Publications Over Time.	24
2.3	Frequency of Publications in JCR Category.	25
2.4	Frequency of Papers Applying Collection Method.	26
2.5	Frequency of Papers in Each Condition Area.	28
2.6	Frequency of Papers in Each Care Level.	29
2.7	Frequency of Multiple Care Levels.	30
2.8	Frequency Cross Analysis Between Condition Area and Care Level.	30
2.9	Frequency of Papers by Scope.	31
2.10	Frequency of Papers Applying Method Type.	33
2.11	Frequency of Papers Applying Multiple Methods.	33
2.12	Frequency Cross Analysis Between Method and Condition Area. . .	34
2.13	Graph of the Interaction Between Mapping, Modelling and Improving the Pathway.	35
2.14	Frequency Cross Analysis Between Method and Investigation Area.	36
2.15	Frequency of Papers Considering Outcome Measure.	37
2.16	Frequency of Considering Multiple Outcomes.	37
2.17	Frequency Cross Analysis Between Outcome and Method.	38
2.18	Frequency Cross Analysis Between Outcome and Scope.	38
2.19	Frequency of Papers Considering Decision Level.	39
2.20	Frequency Cross Analysis Between Decision Level and JCR Category.	40
2.21	Frequency Cross Analysis Between Decision Level and Scope.	41
2.22	Frequency Cross Analysis Between Decision Level and Method. . .	41
2.23	Frequency Cross Analysis Between Decision Level and Outcome. . .	42
3.1	All Unique Pathways Displayed as a Heatmap.	50
3.2	Example of the Calculation for the Levenshtein Distance.	52

3.3	Example of Dynamic Programming Using the Levenshtein Distance.	53
3.4	Example of the Calculation for the Damerau-Levenshtein Distance.	53
3.5	Example of the Calculation for the Jaro Distance.	54
3.6	Example of the Calculation for the Jaro-Winkler Distance.	55
3.7	Example of the Calculation for the Needleman-Wunsch Distance. . .	56
3.8	Example of the Needleman-Wunsch Algorithm.	57
3.9	Example of bi-gram.	58
3.10	Example of Longest Common Subsequence.	60
3.11	Example of Longest Common Subsequence Dynamic Programming.	60
3.12	Example of Modified Needleman-Wunsch Algorithm Dynamic Programming.	68
3.13	Example of Modified Needleman-Wunsch Algorithm Traceback. . .	68
3.14	Example of Feature Five.	70
3.15	Comparing Prioritising in Traceback.	70
3.16	Modified Needleman-Wunsch Distance Matrix for Sample 2.	71
3.17	Comparison of the Ten Metrics Applied to Sample 1.	74
3.18	Comparison of the Ten Metrics Applied to Sample 2.	76
4.1	Ciw Network Example Extracted From Jupyter Notebook.	93
4.2	Ciw Run Network Example Extracted From Jupyter Notebook. . .	94
4.3	Ciw Run Results Example Extracted From Jupyter Notebook. . . .	95
4.4	Raw Top Level Results of Varying Pathway Orderings.	107
4.5	Raw Activity Waiting Time Results of Varying Pathway Orderings.	108
4.6	Probability of Individuals per Day.	124
4.7	Probability of Number of Test Requests.	126
4.8	Cumulative Probability of Number of Test Requests.	126
4.9	Steady State probability of Queued Number of Test Requests. . . .	128
4.10	Probability of Waiting Time in Days.	129
5.1	Network Graph for <i>Full Transitions</i>	141
5.2	Network Graph for <i>Cluster Transitions</i> Cluster 3.	146
5.3	Process Based Selection Violin Plot for $k = [2, 30]$	153
5.4	Network Graph of the Reduced Transition Probabilities.	155
5.5	Network Graph of the 21 <i>Process Medoids</i> Transition Probabilities.	156
5.6	Network Graph of the Raw 21 <i>Process Medoids</i>	157
5.7	Network Graph of the Linked 21 <i>Process Medoids</i>	158
6.1	Sim.Pro.Flow Logo.	169
6.2	Usage Map for Sim.Pro.Flow.	175
6.3	Sim.Pro.Flow - Data Panel.	178
6.4	Sim.Pro.Flow - Clustering Panel.	179

6.5	Sim.Pro.Flow - Simulation Panel.	180
6.6	Sim.Pro.Flow - Visualisation Panel.	181
A.1	Simplified National Optimal Lung Cancer Pathway.	246
A.2	Pathway Mapping Exercise: Original St Woolos.	247
A.3	Pathway Mapping Exercise: St Woolos Interpretation.	247
A.4	Pathway Mapping Exercise: Velindre Interpretation.	247
C.1	Histogram of the Original Data Total Time in System.	255
C.2	Activity Waiting Time Histograms for Original Data.	256
C.3	Line Plot for Capacity Scenario 4.	257
C.4	Network Graph for <i>Cluster Transitions</i> Cluster 1.	258
C.5	Network Graph for <i>Cluster Transitions</i> Cluster 2.	259
C.6	Network Graph for <i>Cluster Transitions</i> Cluster 4.	260
D.1	Sim.Pro.Flow - Clustering Panel Results Pop Out Window.	267
D.2	Sim.Pro.Flow - Clustering Panel Subtabs for Inputting Rankings and Groupings.	268
D.3	Sim.Pro.Flow - Simulation Panel Input Parameters.	268
D.4	Sim.Pro.Flow - Simulation Panel Utilisation Pop Out Window.	269
D.5	Sim.Pro.Flow - Capacity Panel for Capacity Investigation.	269
D.6	Sim.Pro.Flow - Simulation Panel Containing Simulation Results.	269
D.7	Capacity Utilisation Percentage Plot for <i>Raw Pathways</i> - seed 0.	273
D.8	Capacity Utilisation Queue Plot for <i>Raw Pathways</i> - seed 0.	274

List of Tables

1.1	UK and Ireland Cancer Pathway Guidelines.	5
2.1	Summary of Categories for Sample of Papers.	22
2.2	Number of Articles by Geographical Area.	24
3.1	Publications Categorised as Data Mining or Machine Learning Method.	47
3.2	Pathway Activity Names and Assigned Letters.	49
3.3	Grouping Assignments for Each Activity.	63
3.4	Ranking and Weighting Results for Each Activity.	64
3.5	Clustering of Sample 1, for All Ten Distances.	75
3.6	Clustering of Sample 2, for All Ten Distances.	77
3.7	Results of Full Data Clustering for $k = 2$	79
3.8	Results of Full Data Clustering for Best k (excluding $k = 2$).	79
3.9	Least Distance Initial Starting Medoids.	81
3.10	Most Occurred Initial Starting Medoids.	81
3.11	Random Initial Starting Medoids for Two Clusters.	81
3.12	Random Initial Starting Medoids for Three Clusters.	81
3.13	Sensitivity Analysis of Penalty Value Selection for Two Clusters.	83
3.14	Sensitivity Analysis of Penalty Value Selection for Three Clusters.	84
3.15	Sensitivity Analysis of Rankings.	86
3.16	Sensitivity Analysis of Groupings.	86
4.1	Results of Routes Performed for Cases Used in Ciw Initial Investigation.	97
4.2	Pathway Activities New Letter Assignments.	102
4.3	Top Level Results for Original Data.	103
4.4	Activity Specific Results for Original Data.	103
4.5	Top Level Results of Varying Pathway Orderings.	106
4.6	Activity Waiting Time Results of Varying Pathway Orderings.	106
4.7	Top Level Results of Pathway Scenarios.	109
4.8	Activity Waiting Time Results of Pathway Scenarios.	109
4.9	Top level Results of Varying Deterministic Service Time [0.1,1].	111
4.10	Top level Results of Varying Exponential Service Time [0.1,1].	111

4.11	Activity Waiting Time Results of Varying Deterministic Service Time [0.1,1].	112
4.12	Activity Waiting Time Results of Varying Exponential Service Time [0.1,1].	113
4.13	Automated Seven Days Capacity Patterns.	116
4.14	Top level Results of Seven Days Capacity Method.	116
4.15	Activity Waiting Time Results of Seven Days Capacity Method. . .	116
4.16	Automated Five Days Capacity Patterns.	117
4.17	Top level Results of Five Days Capacity Method.	117
4.18	Activity Waiting Time Results of Five Days Capacity Method. . . .	117
4.19	Top Level Results for Warm Up Comparisons, 7 days.	119
4.20	Top Level Results for Warm Up Comparisons, 5 days.	119
4.21	Activity Waiting Time Results for Warm Up Comparisons, 7 days. .	120
4.22	Activity Waiting Time Results for Warm Up Comparisons, 5 days. .	120
4.23	Scenarios for Capacity Investigation.	131
4.24	Capacity Patterns for the Four Scenarios.	133
4.25	Top Level Results for the Four Capacity Scenarios.	133
4.26	Activity Waiting Time Results for the Four Capacity Scenarios. . .	133
5.1	Raw Transition Matrix for the <i>Full Transitions</i>	139
5.2	Arrival Lambda for <i>Full Transitions</i>	139
5.3	Transition Probability Matrix for <i>Full Transitions</i>	140
5.4	Arrival Lambda for <i>Cluster Transitions</i>	143
5.5	Raw Transitions for <i>Cluster Transitions</i> , Cluster 1, 2, 3 and 4 Respectively.	144
5.6	Transition Probability Matrix for <i>Cluster Transitions</i> , Cluster 1, 2, 3 and 4 Respectively.	145
5.7	Process Based Clustering Results for the 21 <i>Process Medoids</i>	149
5.8	Raw Transition Matrix for the 21 <i>Process Medoids</i>	150
5.9	Reduced Transition Probability Matrix (5%).	152
5.10	Medoids Transition Probability Matrix - for the 21 <i>Process Medoids</i> . .	152
5.11	Difference Matrix Between Reduced and Medoids Transition Probability Matrices.	152
5.12	Top Level Results for Routing Procedures with Automated Capacity. .	163
5.13	Frequency for Activity for Routing Procedures with Automated Capacity.	164
5.14	Activity Mean Waiting Time Results for Routing Procedures with Automated Capacity.	165
6.1	Examples of Supported Movement Through Sim.Pro.Flow.	174

6.2	Evaluation of the Development of Sim.Pro.Flow.	182
7.1	Original and Standard Activity Waiting Times.	188
7.2	Capacity Values Used for Individual Adjustment Investigation. . . .	188
7.3	Top Level Results for Individual Adjustment Investigation.	188
7.4	Activity Waiting Time Results for Individual Adjustment Investigation.	189
7.5	Capacity Values for Basic Adjustment Investigation.	192
7.6	Top Level Results for Basic Adjustment Investigation.	192
7.7	Activity Waiting Time Results for Basic Adjustment Investigation.	192
7.8	Capacity Values for End Activity Investigation.	194
7.9	Top Level Results for End Activity Investigation.	194
7.10	Activity Waiting Time Results for End Activity Investigation. . . .	194
7.11	Capacity Values for Target Investigation.	197
7.12	Top Level Results for Target Investigation.	197
7.13	Activity Waiting Time Results for Target Investigation.	198
7.14	Top Level Results for Excessive Top Down Investigation.	201
7.15	Capacity Slots Per Day for Excessive Top Down Investigation. . . .	202
7.16	Activity Waiting Time Results for Excessive Top Down Investigation.	203
7.17	Top Level Results for Demand Investigation.	205
7.18	Activity Waiting Time Results for Demand Investigation.	205
7.19	Capacity Adjustment for Medoids Capacity Investigation.	206
7.20	Capacity Patterns for Medoids Capacity Investigation.	207
7.21	Top Level Results for Medoids Capacity Investigation.	207
7.22	Activity Waiting Time Results for Medoids Capacity Investigation.	207
7.23	Capacity Results Patterns for All Investigations.	210
B.1	Literature Review Table For: Papers of Notable Contribution. . . .	250
B.2	Literature Review Table For: Summary of Previous Literature Reviews.	251
B.3	Literature Review Table For: Frequency of Publications in JCR Cat- egory.	252
B.4	Literature Review Table For: Frequency of Papers Applying Collec- tion Method.	252
B.5	Literature Review Table For: Frequency of Papers in Each Condition Area.	252
B.6	Literature Review Table For: Frequency of Papers in Each Care Level.	252
B.7	Literature Review Table For: Frequency of Multiple Care Levels. . .	253
B.8	Literature Review Table For: Frequency of Papers by Scope.	253
B.9	Literature Review Table For: Frequency of Papers Applying Method Type.	253

B.10 Literature Review Table For: Frequency of Papers Applying Multiple Methods.	253
B.11 Literature Review Table For: Graph of the Interaction Between Mapping, Modelling and Improving the Pathway.	253
B.12 Literature Review Table For: Frequency of Papers Considering Outcome Measure.	254
B.13 Literature Review Table For: Frequency of Considering Multiple Outcomes.	254
B.14 Literature Review Table For: Frequency of Papers Considering Decision Level.	254
D.1 Output Files for Sim.Pro.Flow.	270
E.1 Top Level Results for Individual Adjustment Investigation.	279
E.2 Activity Waiting Time Results for Individual Adjustment Investigation for D, G, and I.	280
E.3 Activity Waiting Time Results for Individual Adjustment Investigation for K, C and A.	281
E.4 Activity Frequency Results for Medoids Capacity Investigation. . .	281

Glossary

Term	Description
General	
Data Type	The format of the data, where three options (DT1, DT2 and DT3) are supported in Sim.Pro.Flow as described in subsection 1.2.4.
Data	The selected input data.
I	The number of individuals in the data.
Dataframe	The dataset generated of the unique pathways within the data. This will be sorted by decreasing count then alphabetically decreasing.
i	The number of entries in the dataframe which corresponds to the number of unique pathways in the data.
Pathway	String of single character codes representing the chronological order of activities performed by an individual.
Pathways Indexes	The number corresponding to the position of the pathway in the dataframe, starting at 0.
Day	The time unit of a day.
Named Day	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday. Day 0 will be Monday in the simulation.
N	Number of activities i.e. total number of activity nodes.
n	Activity node where $n \in [1, \dots, N]$.
A_n	Number of arrivals at node n i.e. the number of times node n was the first activity in the pathway.
P	Number of days in the period covered in the data (Overall Period).
Floor	Formally the largest integer no larger than x [305] i.e. the integer.
Ceiling	Formally the smallest integer no smaller than x [305] i.e. the integer + 1. Written as <i>ceil</i> in equations.
Networks	Depiction of the relationship between nodes in the form of a directed graph.
Time in System	Integer of the last service start date minus the integer of the arrival date.
Routing Procedure	Interpretation of the clinical pathway which defines the network and the possible pathways that are available to perform. The following four routing procedures are discussed: <i>Raw Pathways</i> (for validation), <i>Full Transitions</i> , <i>Cluster Transitions</i> and <i>Process Medoids</i>

Term	Description
Clustering	
Distance Metric	Method used to calculate a numerical value to represent how different two strings are. The Python library Textdistance [274] is used to calculate all metrics except for the Modified Needleman-Wunsch.
Transition Matrix	Matrix where position (i, j) represents the exact number of times activity i followed by activity j was performed.
Reduced Transition Matrix	Same as Transition Matrix but where any entries with values less than the Adjust Percentage are reduced to 0.
Adjust Percentage	Percentage applied to the number of unique pathways e.g. 5% of 200 pathways = 10.
Transition Probability Matrix	Matrix where position (i, j) represents the probability of activity i followed by activity j was performed. Calculated by each value in the Transition Matrix divided by the sum of its row.
Difference Matrix	The total, mean and largest transitions are calculated by taking the absolute difference between the reduced transition matrix and the simulation transition matrix, not including the start and end activities.
Distance Matrix	Matrix where position (i, j) represents the distance between string i and j .
Propagated Values	As the clustering is performed for the dataframe (only unique pathways), when using the clustering results for the simulation one must consider 'propagating' the results to reflect the full data. For example, if pathway 'ABC' was assigned to cluster 0 and repeated 4 times in the original data, then the list of pathways for cluster 0 would contain ['ABC', 'ABC', 'ABC', 'ABC'].
k-Medoids	Clustering method to separate the set of pathways into k clusters with one of the pathways as the centre point for each cluster. Points are assigned to the cluster with the smallest distance to the centre point (medoid). [309]
Medoids	The pathway selected as the centre point for a cluster.
k	The number of clusters.
Max k	The maximum number of clusters to calculate.
Initial Centroids	Initial starting centres for the k-medoids clustering.
Silhouette Score	A metric to represent how 'good' the clustering is - used to suggest the number of clusters (k) to select. Describes how close a point is to its assigned cluster centre compared to the centres of the other clusters [314]. Values range between -1 and 1 where a larger values indicates a better clustering. The Python library scikit-learn is used to calculate the silhouette score [248].
Network Connections	Number of non-zero entries in the Transition Matrix
Tolerance	Value to define the results to highlight i.e. if the number of network connections are within the selected tolerance of the number of network connections of the original data then results will be highlighted for that value of k .

Term	Description
Simulation	
Seed	As Ciw [56] uses random number generators, a seed is set to ensure that the same set of random numbers are used. This allows the simulation to be run multiples times and not have the results change due to the random numbers generated.
Node	A node is an object where individuals are serviced. Each activity will have a corresponding node. Node count starts at 1.
Dummy Node	A dummy node is used when the service at the node is unimportant but the node is required to structure the system.
Class	A class is used to denote groups of individuals [58].
Arrival Rate	The rate at which individuals will arrive into the system through the arrival node/s. An Exponential Distribution is use where the general arrival rate used is calculated by $\lambda = \frac{A_n}{P} \quad (1)$
Exponential Distribution	The exponential distribution is used to sample the time between events which occur at an average rate λ [59,304]. The exponential distribution is commonly used for the arrival rate.
Deterministic Distribution	The deterministic distribution will always sample the same value [59].
Server Schedule	Takes the chosen capacity and produces a schedule to define the amount of capacity required per day. The length of the schedule will be defined for the number of weeks in 1.5* Overall Period for the original data. If warm up Iterative selected then defined for the number of weeks in w*Overall Period for the original data. Schedule is cyclical, thus if the end is reached the schedule will loop. <i>*Ensure for warm up Warm Start that the schedule is not required to loop*</i>
Warm Up	A warm up time is the time where the simulation will run without collecting results. This is to allow the simulation to 'fill up' before observation of the system. [61]
w	For the Iterative warm up, w is the number of times to simulate I individuals where only the final I individuals will record results e.g. with w = 3 and I = 200, the simulation will run for 600 individuals where only the final 200 (from id number 401 onwards) will be recorded.
Week Type	Number of working days. If 5 days selected simulation capacity will be set to 0 for Saturday and Sunday for all activities (including the dummy node).
Percentage Util 100%	$\text{Percentage Util 100\%} = \frac{\text{No. days recorded all capacity used}}{\text{No. days activity had capacity}} \quad (2)$
Overall Period	For the original data this is calculated by latest date recorded - earliest date recorded. For the Simulation results this is calculated by latest exit date - earliest exit date recorded.

Term	Description
Capacity	
P	Probability distribution for number of individuals arriving into the system each day.
λ	Arrival rate for Poisson pmf.
α	Probability of extra capacity.
A	Individual
n	Number of tests
LB	Lower bound
UB	Upper bound
P_n	Row vector containing the probability per possible number of test
C	Capacity
r	Row
t	days (i.e. time)
p	Probability
p sequence	Sequence of probability (α) values, where pseq is for daily values and PSEQ is weekly values.
D	Number of days in the week considered.
E	Expected value for number of arrivals per day.

Term	Abbreviation
Single Cancer Pathway	SCP
National Optimal Lung Cancer Pathway	NOLCP
Point of Suspicion	POS
Non-Small Cell Lung Cancer	NSCLC
Small Cell Lung Cancer	SCLC
Urgent Suspected Cancer	USC
Non Urgent Suspected Cancer	nUSC
Multi Disciplinary Team	MDT
Anesthesiology	AN
Health Policy and Services	HPS
Industrial Engineering	IE
Medical Informatics	MI
Operations Research and the Management Sciences	OR/MS
Business Process Modeling Notation	BPMN
Diagnosis-Related Groups	DRG
Quality Adjusted Life Years	QALY
Information Technology	IT
Electronic Health Records	EHRs
Discrete Event Simulation	DES
Time in System	TiS
Key Performance Indicators	KPIs
Probability Mass Function	pmf
Percent Point Function	ppf
Visual Interactive Simulation	VIS
Graphical User Interface	GUI
Data Type 1/2/3	DT1/DT2/DT3
Levenshtein	Lev
Damerau-Levenshtein	DLev
Jaro	Jaro
Jaro-Winkler	JaroW
Needleman-Wunsch	NW
Jaccard	Jac
Cosine	Cos
Longest Common Subsequence	LCS
Modified Needleman-Wunsch	MNW

Chapter 1

Introduction

Knowledge Economy Skills Scholarships (KESS) is a pan-Wales higher level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government’s European Social Fund (ESF) convergence programme for West Wales and the Valleys. This research project was funded by KESS2 [152], in collaboration with a company¹ partner – Velindre Cancer Centre (VCC) [286]. VCC is the largest cancer treatment centre in Wales, who provide specialist non-surgical solid tumour cancer services to South East Wales and farther.

This thesis investigates three main areas of developing a clinical pathway, namely, mapping, modelling and improving. Each stage of development brought with it its own set of methodological challenges resulting in interesting research developments.

This chapter provides background motivation and initial exploration as follows:

- Section 1.1 gives an overview of cancer services and discusses the motivation of focussing on lung cancer.
- Section 1.2 introduces clinical pathways through formal definition, application to cancer in the UK, performance targets and interpretation within the thesis.
- Section 1.3 provides the problem description, discusses preliminary investiga-

¹“Company Partner” is standard terminology used by KESS2.

tions and introduces the decision support tool.

- Section 1.4 formalises the research questions and outlines the thesis structure.
- Section 1.5 outlines my contribution to collaborations within the thesis.

1.1 Cancer Services

Cancer patients can present in the system in many different ways for example self reporting either through a GP appointment or Accident and Emergency (A&E) department at a hospital, screening programme, incidental finding during a routine operation or inpatient hospital stay. From this point their progression through the system can take numerous routes in the endeavour to diagnose the cancer and then diverge even further when considering treatment protocol. This leads to high variation of interactions with these services, adding a base level of complexity.

A cancer patient is likely to interact with all three levels of primary, secondary and tertiary care (e.g. General Practitioner, hospital and specialist centre respectively [210]). These are often run as individual entities that organise their operations and data collection as isolated segments. For context, Velindre provides tertiary care as the main hub for non-surgical solid tumour cancer services, where outpatient services are also offered at other secondary care sites. This spanning network of services is one of the main factors in requiring an aligned clinical pathway, to ensure that patients are receiving the same standard of care across all sites. Therefore the scope for producing a clinical pathway for cancer patients spans a complex system.

It is important that cancer patients receive individualised treatment that is suitable to them, equally it is necessary for a service to be able to plan for this high level of variation accordingly. This leads to the interesting opposing objectives of mapping out a clear structured pathway whilst also allowing for personalised care.

A pathway that encompasses a holistic view of the system for all cancer types would be an inconceivably large undertaking. Alternatively, producing a specialised clinical pathway for each of the 93 different cancer types listed by Cancer Research

UK [45] would also be a very difficult task. Therefore, it was clear that for this research to have the largest benefit to Velindre, the methods used would have to be able to adapt to various cancer sites. However, for the purpose of development and investigation one cancer site would be chosen for in depth analysis.

Lung Cancer

The World Health Organisation [317] notes lung cancer as (one of) the most common cancers and the cancer with the highest number of cancer deaths. In general it is known that cancer mortality can be reduced with early treatment and detection. For lung cancer specifically, in Wales (reported 2015 for 2010-2012 time period), one-year survival for stage 1 was 78% compared to 14% for stage 4 [297]. Comparing the lung cancer survival in Wales to the rest of Europe finds that for men and women, one-year and five-year lung cancer survival, Wales was the 28th lowest relative survival out of 29 countries [297]. These statistics indicate that there is a vast amount of room for improvement for lung cancer survival, where [297] concludes that early diagnosis is only one route for improvement, noting that key issues in a patient's pathway i.e. waiting times for x-rays, and rapid treatment could all be considered [297]. Therefore, lung cancer was the chosen cancer site for in depth analysis.

There are two main types of lung cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Approximately 80-85% of lung cancers in the UK are non-small cell (NSCLC) [46] and thus was the main focus of investigation.

1.2 Clinical Pathways

This section introduces clinical pathways by displaying the formal definition (subsection 1.2.1), demonstrating the vast number of definitions in existence. Therefore, an original formal definition is not presented. Alternatively, subsection 1.2.2 and 1.2.3 provide context for cancer clinical pathways in the UK, and subsection 1.2.4 discusses the interpretation of patient pathways applied within the thesis.

1.2.1 Formal Definition

The term “clinical pathway” was first used in 1985 by Zander et al., [330] at the New England Medical Center. Since then, the term has become more frequently used and mutated into multiple terms. For instance, de Luc et al., [80] found 17 different terms which denoted the concept of “clinical pathways”, and discussed that the most common terms were “care pathway”, “critical pathway”, “integrated care pathway” and “care map”.

Bleser et al., [32] conducted a literature review with the aim to “*survey the definitions used in describing the concept and to derive key characteristics of clinical pathways*”. The authors found 84 different definitions of a clinical pathway between 2000 and 2003.

Kinsman et al., [153] conducted a literature review and developed detailed criteria for what should be classified as a clinical pathway and tested this against 260 papers. They developed the following criteria:

- “*The intervention was a structured multidisciplinary plan of care*”.
- “*The intervention was used to channel the translation of guidelines or evidence into local structures*”.
- “*The intervention detailed the steps in a course of treatment or care in a plan, pathway, algorithm, guideline, protocol or other ‘inventory of actions’*”.
- “*The intervention had timeframes or criteria-based progression (that is, steps were taken if designated criteria were met)*”.
- “*The intervention aimed to standardise care for a specific clinical problem, procedure or episode of healthcare in a specific population*”.

If an intervention satisfied the first, and then any three of the remaining four criteria, then it was classified as a clinical pathway.

1.2.2 Cancer Clinical Pathways in the UK

Identifying the initial use and evolution of a clinical pathway for cancer in the UK is not trivial. The earliest example appears to be The NHS Cancer Plan [209] in 2000. This discusses using a “Care Pathway” accompanied by a very simplistic diagram seen in Figure 1.1.

This image has been removed by the author for copyright reasons.

Figure 1.1: Early “Care Pathway” - Extracted from The NHS Cancer Plan [209].

Presently, within the UK there are different guidelines of how to conduct the lung cancer pathway, which are summarised in Table 1.1.

Table 1.1: UK and Ireland Cancer Pathway Guidelines.

Country	Guideline	Provider
<i>England</i>	National Optimal Lung Cancer Pathway	Cancer Research UK [47]
<i>Wales</i>	Single Cancer Pathway, National Optimal Pathway for Lung Cancer	Wales Cancer Network [290], NHS Wales [211]
<i>Scotland</i>	Management of lung cancer	Healthcare Improvement Scotland [119]
<i>Northern Ireland</i>	Lung Pathway	Northern Ireland Cancer Network [213]
<i>Ireland</i>	Lung Cancer Action Plan	Irish Cancer Society [142]

The outset of this research project coincided with when Cancer Research UK published a “National Optimal Lung Cancer Pathway” (NOLCP) in 2017 (recently updated to v3.0 in 2020) which is assumed as the first official formal pathway [47]. The NOLCP v3.0 states *“This optimal pathway is primarily designed to improve outcomes in lung cancer by encouraging best practice, reducing variation, and reducing delays in diagnosis, staging and treatment”* [47]. For ease of understanding, the NOLCP (v2.0) has been converted into a simplified version just containing the

activities and maximum time frames for completion (Appendix A Figure A.1). The converted NOLCP is presented in a similar format to a traditional depiction of a discrete event simulation model (DES).

The Cancer Delivery Plan for Wales 2016-2020 [288] notes the development of the ‘Single Cancer Pathway’ (SCP) [290, 291], with the aim to “*better reflect patient experience*” and “*overcome system delays*” [288]. The SCP was implemented in June 2019, as announced by the Minister for Health and Social Services, Vaughan Gething [298]. The key development of the SCP was replacing the previous performance targets with recording patients waiting times from point of suspicion (POS) [290]. This was supported through the development of National Optimal Pathways (NOPs) [211].

1.2.3 Performance Targets

Since 2004 the Welsh Government measured cancer waiting times through two targets [292]:

- Urgent Suspected Cancer (USC) - suspected cancer referred from primary care. Target: 95% start treatment within 62 days of original receipt of referral.
- Non Urgent Suspected Cancer (nUSC) - all other referral routes e.g. through incidental finding. Target: 98% start treatment within 31 days of MDT discusses treatment plan that is accepted by the patient.

The SCP measures cancer waiting times from point of suspicion, with a target of no patient waiting longer than 62 days for treatment. The SCP defined POS as “*when a clinician refers a patient or requests a test concerned a patient may have cancer*” or abnormal report from screening [292].

The Welsh Government publicly report the statistics on cancer waiting times [266], where statistics can be found for all of the above targets (USC and nUSC from 2009 and SCP from June 2019 onwards).

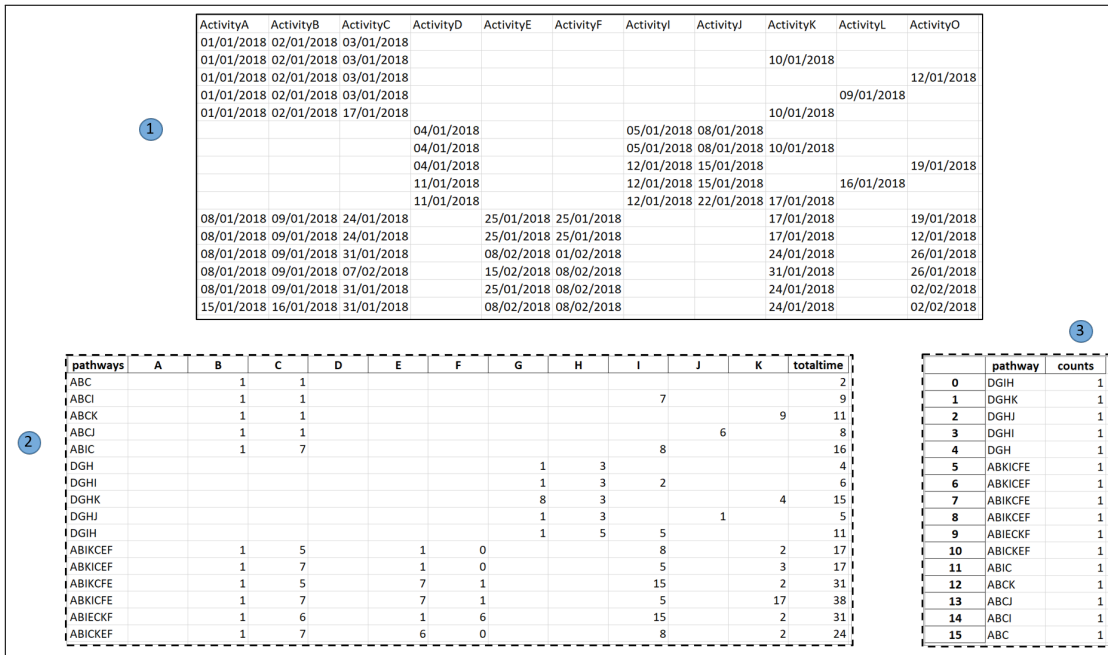
1.2.4 Patient Pathways Interpretation - Input Data

The pathway for cancer diagnosis starts at referral and ends at start of treatment, and contains many steps in between which detect the stage of the cancer. As previously stated, one of the main areas which is explored in this thesis is mapping the clinical pathway. To be able to construct a clinical pathway, it is likely that analysing previous individual patient data recording the activities that they performed (named patient pathway), will be required. Noting that in the literature, a popular method of representing a patient pathway was as a string of letters.

Two formats for the general input data were identified, denoted as data type 1 (DT1) and data type 2 (DT2) (examples in Figure 1.2 and 1.3 respectively). In DT1 each row represents a patient, with a column for each activity available, and the cell records the date stamp of when the patient performed the activity. In DT2, each row records an occurrence across three columns, where “Id” records the patient, “Activity” records the activity name and “Date” records the date stamp.

It is key to note that the main difference between the two data types is that DT1 only allows for each activity to be recorded at most once per patient (regardless of if the patient had performed the activity multiple times), where as in DT2 a patient can perform the same activity multiple times. Therefore, DT1 puts a hard constraint on not allowing multiple attendances of an activity, which will be furthermore referred to as the ‘at most once’ constraint. The main dataset provided (described in Section 3.3) is of format DT1.

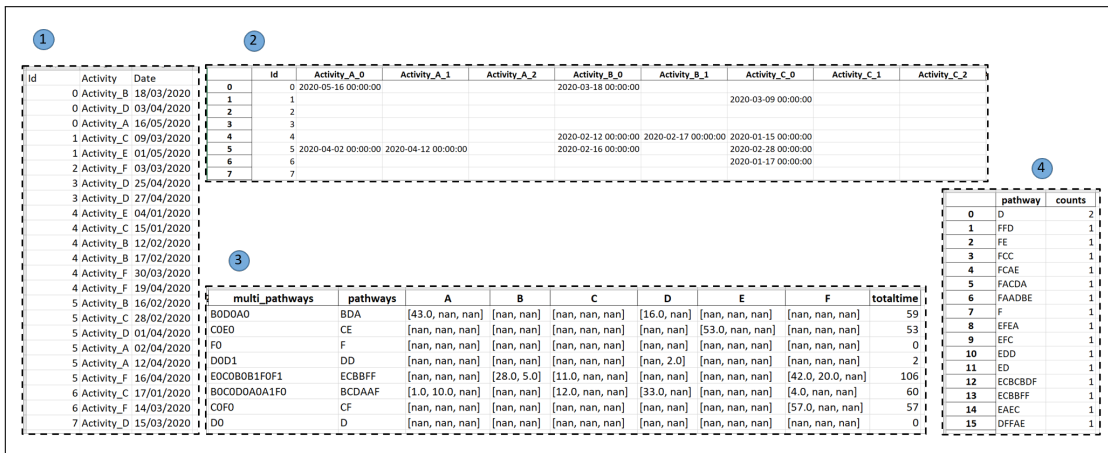
The main advantage of the format of DT1, is that all the information for each patient is on one row. Therefore, it is desirable to convert DT2 into the format of DT1, with retaining the recording of multiple occurrences. This was possible through considering the maximum number of times (R) a single patient performed an activity (n), and constructing R columns for activity n where the activity name appends an occurrence identifier number denoted by “ $_r$ ” where $r = [0, \dots, R]$ (See Figure 1.3 for example).



Extracted from Sim.Pro.Flow help document [252] where:

- (1) Input data where row = patient, column = activity, cell = date stamp.
- (2) Columns added to input data of: constructed patient pathways, activity waiting times and total waiting time (in days).
- (3) Separate constructed dataset (denoted as “dataframe”) containing the set of unique pathways and their corresponding count.

Figure 1.2: Example of Data Type 1.



Extracted from Sim.Pro.Flow help document [252] where:

- (1) Input data with three columns - Id = patient ID, Activity = activity name and Date = date stamp.
- (2) Converted into data type 1 format.
- (3) Columns added to input data of: constructed patient pathways, listed activity waiting times and total waiting time (in days).
- (4) Separate constructed dataset (denoted as “dataframe”) containing the set of unique pathways and their corresponding count.

Figure 1.3: Example of Data Type 2.

As both data types² are now in the same format (row per patient) the pathway strings can be extracted, by assigning each activity a letter code, then for each patient constructing a string of letters representing the chronological order of the activities that they performed³. Then the data will be extended through appending additional columns for the pathways, along with activity and total waiting times.

²An additional data type (DT3) is supported in Sim.Pro.Flow which takes in data in the converted format of DT2 ((2) in Figure 1.3) and groups together activities with the same name.

³Regarding DT2, the “multi_pathways” strings also contain the activity occurrence identifier, which is removed to construct the pathway string.

1.3 Problem Description

The timeline of this research coincides with the initiation of the ‘Single Cancer Pathway’ (Section 1.2), and as such led to great interest into how to map out the pathway, support changes and monitor progress.

The initial proposed aim of the collaboration was to produce a state-of-the-art decision support tool to align capacity and demand in an effective and efficient manner. Thus the initial proposed goal was to improve patient care and outcomes by reducing time to diagnosis and treatment times for those with cancer.

Preliminary investigation into the proposed aim and goal found that clinical pathways were a key focus for reducing time to diagnosis (Section 1.2). Combining this with the desire from the company partner (VCC) for this research to produce a state-of-the-art decision support tool that utilised simulation methods, and the vast number of cancer sites to consider (Section 1.1), indicates a need to produce methods that are efficient and sustainable for pathway mapping, modelling and improving.

Efficient and Sustainable

In general, for the methods produced to be efficient and sustainable for VCC, they have to satisfy the following points:

1. Generalisable: As previously stated, to be of greatest use to VCC the methods must be able to be applied to multiple cancer sites. Therefore, a general approach must be developed. (research question 4)
2. Satisfy data types: It was noted in subsection 1.2.4 that there are two types of data: DT1 has ‘at most once’ constraint, whereas DT2 can support multiple occurrences of an activity. Pre-empting the considerations for constructing a simulation model, it is traditional to use probabilities in the form of Markov chains. This would be suitable for DT2, however, to adhere to DT1’s ‘at most once’ constraint, all of the past performed activities would need to be considered when choosing the next activity to avoid duplication, which would

not align with the memory-less property of Markov chains. Therefore, to be sustainable, the chosen methods would need to be considerate of this (Chapter 3). Ultimately, there is a need to support multiple interpretations for the clinical pathway (See “routing procedure”). (research question 3)

3. Quick and easy to produce: To enable VCC to apply the methods developed to multiple cancer sites they must be relatively non-tedious to produce. There are two key areas to address this:

- Mapping the pathway: It is traditional for a researcher to explore the system they are investigating and interact with the experts who are part of that system. This method for mapping the clinical pathway was initially investigated in subsection 1.3.1 to determine if it would be sufficient. (research question 1)
- Building the simulation: It is well known in the Operational Research community that producing a discrete event simulation (DES) is a time consuming process. This is commented on by Monks et al, [202] who note other references where this is discussed. Therefore, it became a requirement to address this through the decision support tool. This is discussed further in subsection 1.3.2. (research question 2)

Mapping, Modelling and Improving

Therefore, the three key areas will need to be addressed as follows:

- Mapping: To convert the many patient pathways into a suitable interpretation of a clinical pathway that satisfies Kinsman et al.’s [153] definition (subsection 1.2.1), ensuring to reduce the complexity through minimising variations.
- Modelling: Build a DES of the clinical pathway.
- Improving: Support capacity and demand analysis.

1.3.1 Preliminary Investigation

Preliminary efforts were made to explore the lung cancer clinical pathway, through interacting with the system. The aim was to understand ‘what is the pathway?’ with the intention of mapping out what would need to be built in a model.

The converted NOLCP (Appendix A Figure A.1) was presented to experts, from different care levels, who interact with the pathway (at various stages) day to day. They were asked to annotate the converted NOLCP with how they see the pathway in practice. Simultaneously, the experts were verbally communicating additional description of their annotations and the system, which were noted for retrospective consultation. The results of this preliminary investigation are provided in Appendix A.

These interactions highlighted three key difficulties with this method. Firstly, these meetings took between 1-2 hours each, which replicating on a larger scale would be a significant amount of time. Secondly, there is scope for human error e.g. misinterpretation, subjectivity and revelations on post-reflection. Finally, the vast differences in results led to ambiguity and questions of how to combine the information.

Therefore, this exercise displayed that interpretation of these pathways will be different depending on the person asked, and also that the researcher trying to build a model may also interpret the experts interpretation differently than intended. Thus, typically this would be an iterative method, where the verbal communication noted for retrospective consultation would be combined into the interpretation, along with a series of meetings in the pathway development process.

After observing the inefficiency of the process, this investigation was retired in favour of exploring how to make the pathway mapping process more efficient and sustainable. This will be addressed through research question 1, which will be explored in Chapter 3.

1.3.2 Decision Support Tool - Sim.Pro.Flow

To enable concurrent discussion of developing the methods alongside producing the decision support tool, this section introduces the subsequently produced tool called Sim.Pro.Flow, henceforth referred to as either Sim.Pro.Flow or the ‘tool’.

Coming back to the discussion on developing an efficient and sustainable method for building the simulation (Section 1.3). The notion of developing a generic model has become popular [114, 184]. It is an interesting area to explore alternative methods for reducing complexity of the simulation build process itself. Considering this with the findings from the literature review (Chapter 2) which suggests that a novel contribution would bridge the gap between data mining and simulation, initiated the direction to automatically extract the pathways from the data to construct the simulation network.

This essentially developed into automating the simulation build process. The general idea is to allow VCC to input data (of DT1 or DT2) and automatically extract all input parameters for the DES, including arrivals, service, capacity and warm up.

In the pursuit of developing automation, it was not the aim to fully automate every aspect of the process. Instead the aim was to enhance interaction ability through automating the time consuming process, whilst still ensuring that the final decisions are in the user’s control. This would allow the user to analyse the system and interact with elements that would support typical scenario analysis investigations. As such, developing a Graphical User Interface (GUI) was key to support this.

Sim.Pro.Flow is a basic, easy to use GUI which hosts the methods developed in Chapters 3, 4 and 5. Sim.Pro.Flow was built using the Python package wxPython [318] and is hosted on GitHub as open source software [251]. Chapter 6 discusses the development of Sim.Pro.Flow, along with a description of its structure, images of the GUI and presents the key features.

1.4 Research Questions and Structure

In summary, this introduction has established the main goal to produce a decision support tool for efficient and sustainable clinical pathway analysis. There is a need to develop a generic approach to allow the methods developed to be used for multiple cancer sites. Therefore, the idea of automating the simulation build will be perused, however, consideration will ensure the end decisions are with the user. Furthermore, a suitable method for mapping the clinical pathways needs to be supported.

To refine the research questions, an extensive literature review was performed (Chapter 2). Two major findings were highlighted. Firstly, the need to combine both data and expert information harmoniously in the pursuit of clinical pathway mapping. Secondly, the need to integrate data mining and OR techniques.

The problem description (subsection 1.3) supported by the literature review led to the following four research questions:

1. Can both data and expert information integrate to inform clinical pathway mapping?
2. Is it feasible to automate the simulation build process?
3. Is it viable to support multiple interpretations of clinical pathways through combining a mixture of data mining and OR?
4. Can the development of a decision support tool provide a general method of analysing clinical pathway mapping, modelling and improving?

The structure of this thesis is as follows:

- Chapter 2 contains a literature review, comprised mainly of [15], of 175 papers surrounding clinical pathways in Information Systems (IS), Operational Research (OR) and Industrial Engineering.
- Chapter 3, comprised mainly of [16], answers research question 1 through the development of a new distance metric, modified Needleman-Wunsch algo-

rithm, applied to k-medoids clustering. This effectively reduces the complexity of the clinical pathways into more manageable groups whilst combining expert user input with the data mining method.

- Chapter 4 addresses the idea automating the build of the DES, motivated by research question 2. This chapter explores the methodological considerations required to allow for automation both theoretically - exploring how to extract and validate input parameters, and practically - presenting supporting methods for process based routing and capacity extraction. The routing procedure *Raw Pathways* is discussed throughout as it is intended for validation.
- Chapter 5 considers research question 3 through the introduction of three additional routing procedures - *Full Transitions*, *Cluster Transitions* and *Process Medoids*, that explore progressively less complex and varied interpretations of the clinical pathways, with the last two procedures making use of the clustering described in Chapter 3.
- Chapter 6 expands upon the brief introduction to Sim.Pro.Flow from subsection 1.3.2 and begins to address research question 4 by discussing the development of the tool, along with a description of its structure, images of the GUI and outlining the key features.
- Chapter 7 combines the research from the previous four chapters and uses Sim.Pro.Flow in the form of a case study (along with Chapter 6 addresses research question 4). The motivation is two fold, firstly, to explore if Sim.Pro.Flow is able to support typical exploration of the simulation, and secondly to gain deeper insights into the specific lung cancer pathways explored.
- Chapter 8 concludes this thesis with a summary of the research, discussion of satisfying the research questions and outlining further research.

Throughout the thesis the analysis follows the working example of lung cancer, where the working dataset is outlined in Sections 3.3 and 4.4.

1.5 Collaborations

Throughout this thesis there are multiple references to collaborations where work was undertaken with other researchers. This section outlines my contribution to those collaborations as follows:

- Chapter 2 comprises mainly of the paper “*Clinical Pathway Modelling: A Literature Review*” [15], for which I am the lead author. For my contribution, I undertook the search process, performed the classification, drew conclusions from each classification, and wrote the original draft.
- Chapter 3 comprises mainly of the research paper “*Modified Needleman-Wunsch Algorithm for Clinical Pathway Mining*” [16], for which I am the lead author. This publication has a “CRediT authorship contribution statement” [72] on which my contribution is listed as “*Methodology, Formal analysis, Software, Writing - original draft*”. The publication acknowledgements further note “*The authors would like to specifically acknowledge Nikoleta Glynatsi, Geraint Palmer and Henry Wilde for their support with coding.*”.
- Section 4.3 ‘Extensions to Ciw - Process Based Routing’ discusses the development of process based routing, in collaboration with Dr Geraint Palmer, which was included in Ciw v2.0.0 onwards. Dr Palmer is the original creator of Ciw and as such possesses in depth knowledge of the vast expanse, intricacies and needs of Ciw’s base code, which was invaluable to this collaboration. My contribution consisted of describing the requirement, discussing the solutions and collaboratively writing/editing the code.
- Subection 4.3.4 discusses two additional customisations for the use of Ciw:
 - ‘Restricting Capacity’ - this discusses the idea of allowing a set number of customers to be served at a node each day. This collaboration was performed with Dr Palmer similarly to the previous, where my contribution was describing the requirement, discussing the solutions and

collaboratively writing/editing the code.

- ‘Blocking Arrivals’ - this concerns blocking arrivals after a set number of customers have arrived. Consulting with Dr Palmer pointed me towards the solution of an issue logged in Ciw [62] (Issue 171). My contribution was to implement this solution and edit it to be compatible with other distributions.
- Section 4.7 details an alternative method for calculating capacity. This work was originally developed by Dr Edilson Arruda for the publication “*Resource Optimization for Cancer Pathways with Aggregate Diagnostic Demand: A Perishable Inventory Approach*” [12]. The original work was performed in R and Matlab by Arruda et al., [12]. For the purpose of integrating this into Sim.Pro.Flow, my contribution was to convert the methods into Python in a general format to allow use with compatible input data, with query support from Dr Arruda.
- Chapter 6 discusses the design and development of the decision support tool Sim.Pro.Flow [251]. The acknowledgements of Sim.Pro.Flow’s about section lists “*Paul Harper, Daniel Gartner, Geraint Palmer, Edilson Arruda*” as all the previous collaborations discussed here are incorporated into Sim.Pro.Flow. For my contribution, I took the lead on the design and coding of the tool. Furthermore, every effort was made to account for all resources that were utilised to develop the tool (along with noting licence agreements), and are listed on the Sim.Pro.Flow Github page within the file References.txt [254].

Chapter 2

Literature Review

This chapter comprises mainly of the paper ‘*Clinical Pathway Modelling: A Literature Review* [15] which discusses the literature surrounding clinical pathways, providing a general overview of the publications surrounding clinical pathways in Information Systems (IS), Operational Research (OR) and Industrial Engineering. A rigorous taxonomy to characterise an abundant work of literature around clinical pathways is presented. Subsequently, the applicability of the taxonomy is demonstrated by classifying the research papers into the different categories.

There is a vast scope for what can encompass the term clinical pathway (subsection 1.2.1), with numerous ways of formulating, approaching, and modelling them. At the time of the initial search, the only awareness known of literature reviews to explore specifically “clinical pathways” in relation to OR is Elbattah and Molloy [92] and Erdogan and Tarhan [94].

The search criteria (discussed in Section 2.1) covers the period of 1998-2018 (November). This search has not been updated to include more recently published work, as this literature review accurately represents the information present at the time that influenced the research questions. However, for completeness, more recent results were investigated and briefly discussed in Section 2.1.

For each figure that displays a key classification result i.e. caption does not start ‘Frequency Cross Analysis Between’ (excluding ‘publications over time’ and ‘geographical area’), there is a respective table within Appendix B which fully details the reference number for each paper within the category.

The remainder of the chapter is structured as follows: Section 2.1 describes the search criteria and Section 2.2 discusses previous literature reviews. Section 2.3 then explores a sample of the selected papers to aid understanding of the taxonomy, Section 2.4 displays the taxonomy results for the literature and Section 2.5 closes the chapter with a discussion and conclusion.

2.1 Search Criteria

A structured search was conducted using the Scopus search engine restricting to years 1998-2018 (November). The keywords were specified to focus on clinical pathway and its main alternative terms as indicated by de Luc et al., [80]: “care pathway”, “critical pathway” and “care map”. Two further terms were also included, namely “anticipated recovery pathways” and “patient pathway”.

The term “patient flow” was not included in the search terms as this review is specifically interested in the structure of well-defined pathways, and patient flow typically relates to the general movement of patients.

The search is focussed to journal publications from five categories in the Thomson-Reuters Journal Citation Report (JCR), and as a result only journal articles were returned by the search. The five subject categories are Anesthesiology (AN), Health Policy and Services (HPS), Industrial Engineering (IE), Medical Informatics (MI) and Operations Research and the Management Sciences (OR/MS), each of which have an impact factor.

These categories were chosen to provide an overview of papers within the Operational Research area, in addition to highlighting the type of information that is being presented to other areas on this topic. Specifically, these five categories were chosen for the following reasons:

- Anesthesiology (AN): captures a subgroup of medical journals in which quantitative methods have been published more frequently than in other medical disciplines (e.g. Anesthesia & Analgesia).¹
- Health Policy and Services (HPS): captures those journals covering impact on policy decisions and service improvement (e.g. Health Care Management Science and Health Services Research).
- Industrial Engineering (IE): is a quantitative category that covers engineering journals (e.g. Computers and Industrial Engineering or Computers and OR) in which, for example, patient scheduling papers have been published.
- Medical Informatics (MI): to include data mining and healthcare information systems topics (e.g. Journal of Medical Systems).
- Operations Research and the Management Sciences (OR/MS): covers quantitative modelling and journals surrounding OR in healthcare (e.g. Journal of the Operational Research Society).

Figure 2.1 shows a diagram detailing the search process.

The screening stage, as displayed in Figure 2.1, consisted of analysing abstracts of the resulting papers from the search. Any papers that did not refer to a pathway or used only qualitative or statistical methods e.g. interviews or regression respectively, were excluded. The screening stage also excluded papers not available in English. The diagram highlights the use of a backward search, for which the same screening criteria as described above was applied. The final number of records included in the analysis is 175.

¹Anesthesiology was included in lieu of oncology in pursuit of returning quantitative methods and justified by the former being numerical based.

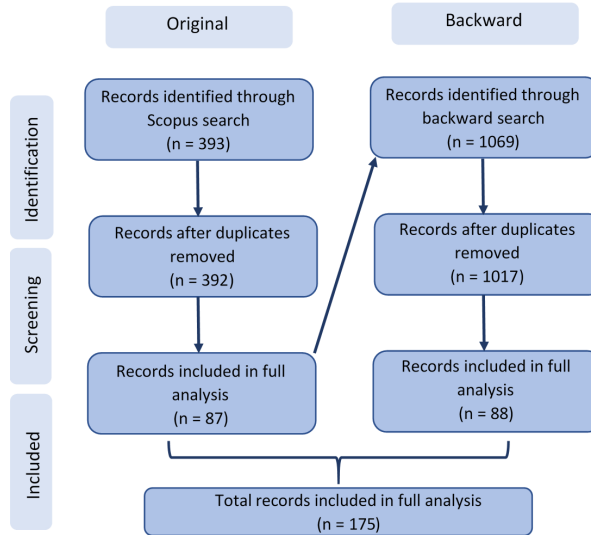


Figure 2.1: Diagram Detailing the Search Process.

Although not included in the following analysis, the initial search string was investigated for 2019-2021 (February) which returned 127 additional results. Initial inspection of the titles suggested that 33 of those results may be of relevance, with 7 of particular note from their abstract [4, 13, 14, 200, 244, 276, 316], although these were selected with retrospective knowledge of the following research.

2.2 Previous Research

Additionally to the 175 papers that were selected, the research revealed 11 papers of notable contribution and 27 literature reviews. All of these papers discuss clinical pathways and the techniques surrounding them, in some form. These are summarised and can be found in Appendix B, Table B.1 and Table B.2, respectively.

Concerning the 11 papers of notable contribution, these discussed guidelines, frameworks, case studies and I.T. artifacts that support clinical pathways.

Of the 27 review papers in Table B.2 (Appendix B), six consider process mining or data mining [55, 94, 110, 157, 188, 243], seven consider simulation [2, 151, 184, 262, 264, 265, 331] and three consider stochastic modelling [90, 175, 337].

There are seven papers that use the term “clinical pathway”, or its synonyms, in the search terms [92, 94, 184, 226, 262, 264, 283] and two papers use “patient flow” in their search terms [157, 262]. These papers all consider clinical pathways, but most focus on a different primary topic.

There are only two reviews that concern clinical pathways specifically: Elbattah and Molloy [92] and Erdogan and Tarhan [94].

Elbattah and Molloy [92] provide a comprehensive discussion of 22 papers concerning modelling and simulating clinical pathways. This literature review provides a different perspective from [92] through developing a rigorous taxonomic approach to classifying many papers.

Although Erdogan and Tarhan [94] consider process mining as a primary topic, the amount of consideration and discussion around clinical pathways is vast, indicating 59 papers concerning clinical pathways. The systematic mapping method used is reflective of this method, however as clinical pathways are not the primary consideration, it does not fully consider a discussion of clinical pathways post-discovery. With clinical pathways being the main focus of this literature review, it differs from that of Erdogan and Tarhan [94] as a more holistic view on clinical pathways is considered here.

2.3 Exploration of a Sample of Papers

This section explores a sample of the papers for the purpose to aid understanding of the classifications in Section 2.4. These papers have been chosen so as to discuss the widest range of categories, using the smallest sample of papers.

The sample of papers are discussed briefly and their relevant categories are indicated in Table 2.1

Table 2.1: Summary of Categories for Sample of Papers.

No.	Condition	Method	Outcome	Scope	Decision Level
[3]	None	Simulation	Resource & Time	Clinical	Strategic
[22]	None	Optimisation & Heuristics	Cost & Resource	Department	Strategic & Tactical
[29]	Chronic Focus	Simulation	Cost	Clinical	Strategic
[52]	Acute Focus	Simulation	Resource & Time	Clinical	Tactical
[86]	Chronic Applied	Optimisation & Heuristics	Pathway Mapping	Clinical	NA
[135]	Accute Applied	Data Mining or Machine Learning	Pathway Mapping	Clinical	NA
[155]	Surgical Applied	Data Mining or Machine Learning	Pathway Mapping & Patient Progression	Clinical	NA
[164]	Chronic Focus	Simulation	Patient Progression	Hospital	Strategic
[168]	None	Stochastic Modelling	Resource	Disease	Tactical & Operational
[199]	Surgical Focus	Data Mining or Machine Learning	Time	Disease	Operational
[325]	None	Stochastic Modelling & Data Mining or Machine Learning	Legal	Hospital	NA

Ajmi et al., [3] used Business Process Modeling Notation (BPMN) to model the workflows of the patient journey in a Pediatric Emergency Department. The aim was to identify bottlenecks and crowded situation indicators, with noting that delay occurs in the waiting time from the health care request. The study was integrated into the French National Research Agency (ANR) project, titled: “Hospital: Optimization, Simulation and avoidance of strain (HOST)”.

Barbagallo et al., [22] used BPMN 2.0 to schedule operating room activity, by room and day through a waiting list database, and applied stochastic modelling to allow optimisation.

Bending et al., [29] used Monte Carlo sampling techniques to estimate the direct cost of bowel cancer services.

Chemweno et al., [52] developed a discrete event simulation on the stay of stroke patients in a stroke unit, specifically diagnosis, to investigate capacity and waiting times.

Du et al., [86] develop a new method of handling clinical pathway variances in Takagi-Sugeno (T-S) fuzzy neural networks (FNNs). Two cases concerning osteosarcoma preoperative chemotherapy are used to validate this method.

Huang et al., [135] used Latent Dirichlet Allocation (LDA) for the purpose of discovering the treatment patterns as a probabilistic combination of clinical activities. The method was then applied as part of experiments to careflow logs concerning intracranial hemorrhage and cerebral infarction.

Konrad et al., [155] developed a method to use message exchanges to automatically establish and compare a patient's path against a clinical pathway. The method has been applied to a case study in major joint replacement.

Langley et al., [164] developed a discrete event simulation model for the diagnosis of Tuberculosis (TB) to help provide policy makers with the information to decide which tools, and where, they should be implemented for maximum effectiveness.

Lanzarone et al., [168] modelled the home care pathway using a Markov chain, where the future workload of each operator was of interest to support medium and short term resource planning. The model was developed as a simple software application, integrated into the current software used, which supports patient to operator assignment.

Michowski et al., [199] used a Bayesian Belief Network (BBN) to model the radical prostatectomy clinical pathway with an interest in patients length of stay being categorised as "met" or "delayed" given the patient's outcomes and activities. The research was implemented as an application.

Yang and Hwang [325] utilised clinical pathways, through data-mining using a Markov blanket filter, to facilitate automatic and systematic construction of an adaptable and extensible detection model of fraudulent and system abusive behaviour.

2.4 Classification of Literature

This section discusses the taxonomy which classifies the literature, and provides summary statistics.

2.4.1 General Characteristics

Figure 2.2 displays the distribution of the papers across 21 years and shows that the number of publications considering clinical pathways has rapidly increased. This may reflect the growing demand for the use of clinical pathways in practice and thus the need for more in-depth research.

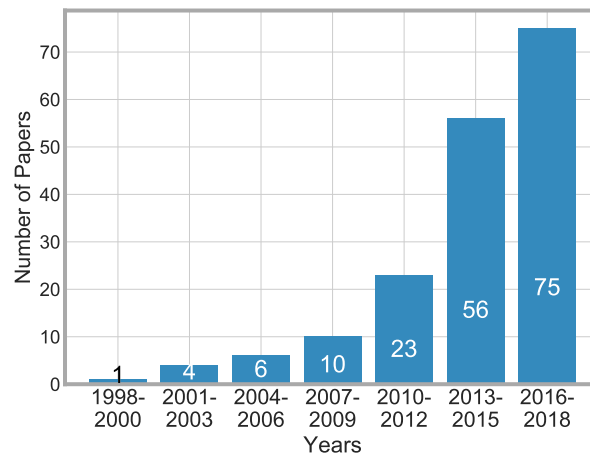


Figure 2.2: Frequency of Publications Over Time.

Table 2.2 shows how the papers were distributed across the world. A paper was classified within a geographical area if it specifically stated that the data or hospital was within that area, or failing that, through any acknowledgements of a hospital in a specific area or the country of the first author was recorded.

Table 2.2: Number of Articles by Geographical Area.

Continent	America	Asia	Europe	Other
Total	32	41	90	12

Table 2.2 shows how Europe has the greatest number of publications relating to clinical pathways, followed by Asia then America. This highlights that research into clinical pathways is of global interest.

This section concludes that research into clinical pathways is growing in popularity across the globe, year on year.

Publication Area

Figure 2.3 breaks down the publications by the JCR category which each paper was published under. Again, the five subject categories are Anesthesiology (AN), Health Policy and Services (HPS), Industrial Engineering (IE), Medical Informatics (MI) and Operations Research and the Management Sciences (OR/MS).

There are 64 papers identified in the backward search, whose ISSN numbers do not relate to any of the five JCR categories, plus a further 16 papers in the backward search that appear to have no ISSN number - and thus also no JCR group. These are not included in Figure 2.3.

Two journals not included in the JCR categories published multiple papers identified in the search, these are Lecture Notes in Computer Science [30, 78, 120, 124, 136, 160, 269] and Studies in Health Technology and Informatics [6, 106, 108, 132, 172, 189, 333].

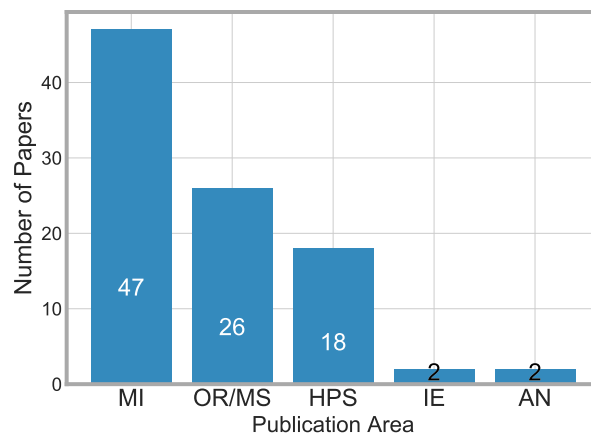


Figure 2.3: Frequency of Publications in JCR Category.

It is apparent here that MI is the most popular JCR group followed by OR/MS. Although there were only a few papers in AN and IE, it is beneficial to capture these as they provide another perspective on clinical pathways.

This highlights the need to bridge the gap between MI with OR and IS methods.

Obtaining The Pathway

Obtaining the pathway is arguably one of the most important aspects of analysing clinical pathways. As presented in the selected papers, there are two common ways of obtaining this information: either data driven or through collaboration with those who interact with the pathway.

There were many different ways of obtaining data described, including historic [199], billing [120], messages [155] and Electronic Medical Records [125]. Similarly, collaboration took on a number of different forms including, consulting with experts [27], staff [18], patients [189] and through observations [147].

Figure 2.4 explores how the information on the pathway was obtained. Forty-seven papers did not clearly specify how they obtained the data and have been classified as unspecified.

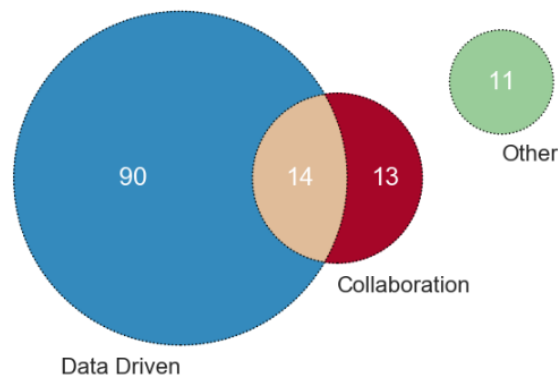


Figure 2.4: Frequency of Papers Applying Collection Method.

Eleven papers specifically stated other methods of collecting the information on the pathway, which are as follows - [29,33,65,116,164,245] stated that the information was provided to them in some way, [111] through previous work (also consulted with experts and stakeholders), [180] online user input, and [31,35,91] used national guidelines.

The advantage of using data to inform the pathway is that the pathway is derived factually and objectively from actual occurrences of the pathway. The advantage of collaboration with staff and experts is that more information can be gathered about

why certain decisions and possible variances from the pathway would occur. Therefore it is recommended to consider both data driven and collaboration with staff when deriving the pathway, although only 14 papers (8%) in this survey considered both aspects.

It is important to note that only 12 papers [3, 33, 83, 91, 105, 131, 134, 167, 168, 176, 199, 258] state that their research/product was implemented/informed policy - this is only 6.9% of the papers surveyed. Previous reviews have found similar results in regards to implementation (e.g. Brailsford et al., [39]), and therefore this finding highlights the need for more implementation and evaluation. However caution needs to be considered here as it is possible that some proposed recommendations were/will be eventually implemented but was outside the timeline of the publication.

2.4.2 Medical Context

Condition Area

The papers selected consider a variety of medical conditions which is either the main focus of the paper, or applied as case study/validation/explanation etc.

There are three condition categories: Acute, Chronic and Surgical, which have been adapted from Zhang et al. [334]. A description of the condition categories are as follows:

- **Acute** - *“Acute conditions are severe and sudden in onset. This could describe anything from a broken bone to an asthma attack.”* [195], stroke has been categorised as acute.
- **Chronic** - *“A chronic condition, is a long-developing syndrome, such as osteoporosis or asthma.”* [195].
- **Surgical** - Papers where the main condition was a specific surgical procedure are classified here.

The three condition areas have been further categorised as either focus or applied:

- **Focus** - The system surrounding the medical condition was the main motivation for the paper.
- **Applied** - The medical condition was considered as a case study or for validation purposes.

Figure 2.5 shows the frequency of papers within each condition area.

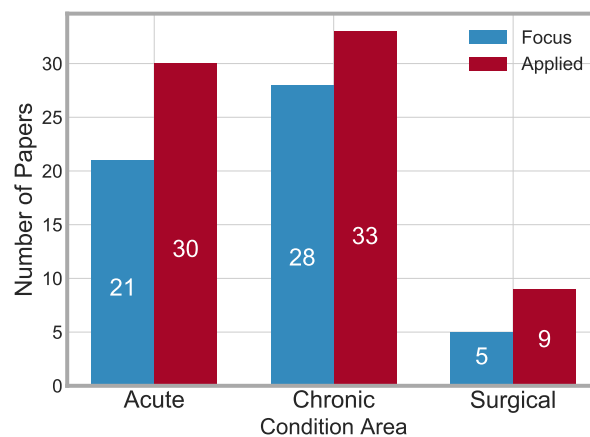


Figure 2.5: Frequency of Papers in Each Condition Area.

Forty-nine papers are not included, as they did not specify a particular condition or considered multiple diagnosis-related groups (DRG).

Chronic conditions are slightly more frequent than acute conditions, and in all three categories it is more frequent for the condition to be applied rather than the focus of the paper.

Care Level

The medical care system is typically split into three sections, Primary, Secondary and Tertiary [210], which are as follows:

- **Primary** - First point of contact e.g. General Practitioner or dentist.
- **Secondary** - Can either be elective or emergency care, also known as “hospital and community care”.

- **Tertiary** - Highly specialised treatment.

Two other levels can be considered - Home Care and Disease:

- **Home Care** - This is when care is provided to the patient at their own home.
- **Disease** - This concerns understanding how the disease progresses and the care provided progresses alongside.

Figure 2.6 shows the frequency at which each of the five care levels are considered, and displays that secondary care is considered most frequently.

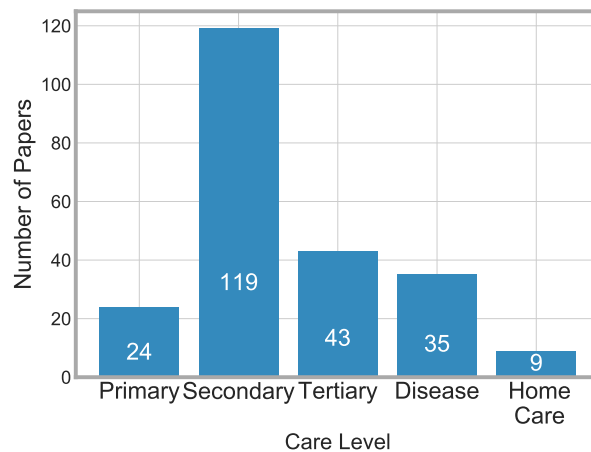


Figure 2.6: Frequency of Papers in Each Care Level.

Seven papers [19, 34, 70, 148, 163, 164, 198] consider when the patient is at home and then gets reintroduced into the system in some capacity.

It is important for these systems to work together to allow the patient a smoother journey on the pathway. Within Figure 2.6 there are 42 papers that consider more than one care level - 31 papers consider two levels, nine consider three levels. The interactions between these levels are displayed in Figure 2.7.

There are also two papers that consider four levels [24, 203] (primary, secondary, tertiary and home care) which are not displayed in Figure 2.7.

From Figure 2.7 it can be concluded that only a few papers consider three or more care levels, and therefore research is not providing the full holistic view of the pathway. It is recommended that, when appropriate, future work should make

every effort to consider multiple care levels.

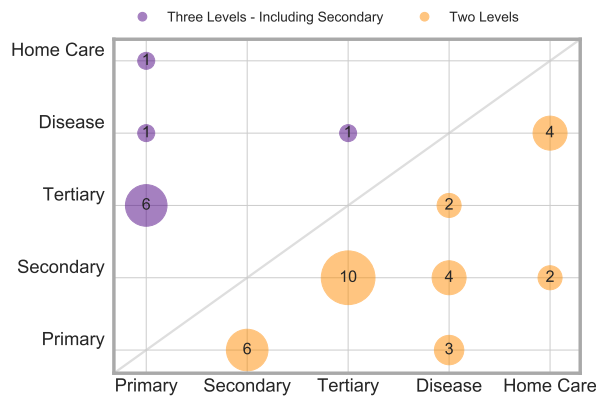


Figure 2.7: Frequency of Multiple Care Levels.

The interaction between condition area and care level can be considered, and is displayed in Figure 2.8.

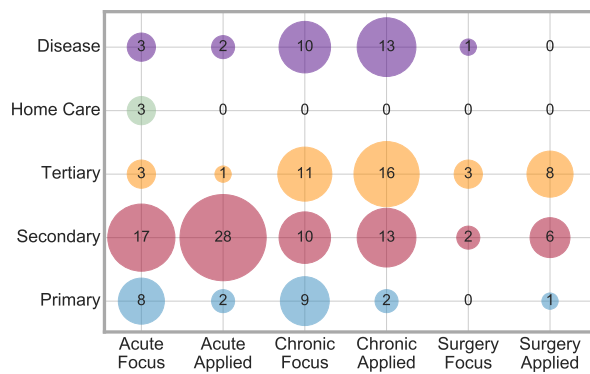


Figure 2.8: Frequency Cross Analysis Between Condition Area and Care Level.

Figure 2.8 shows that acute conditions are mainly considered at a secondary care level, whereas chronic conditions are roughly equally divided between secondary, tertiary or home care levels. This implies that chronic conditions have more range to consider different care levels.

Scope

Although a pathway always has a patient in mind, the scope of the focus on the pathway varied greatly from clinical, disease, department and hospital. This scope considers that although it is typical for the activities of concern to revolve around the patients, they may either not be required to be present, or it is the system

around the patient that is of interest and not the patient movements themselves.

To explain this further, an example for each type of scope is now discussed.

- Bayer et al., [24] is categorised as “Clinical”, as they produce a simulation of the stroke care pathway where, although some activities do not require the patient to be present, the overall focus is on the patient themselves.
- Michalowski et al., [199] is categorised as “Disease”, as the activities are related to the patients’ health, e.g. temperature, pain at rest, vital signs etc.
- van de Klundert et al., [281] is categorised as “Department”, as they define an activity as *“an atomic unit of care delivered to the patient, as meaningful to execute or record the care.”* They also state that *“Although we will not explicitly model it, patient need not be present for each of the activities (consider e.g. lab tests).”*
- Arnolds and Gartner [11] is categorised as “Hospital”, as they focus on improving hospital layout planning by using clinical pathway mining.

Figure 2.9 displays the frequency of papers in each scope category.

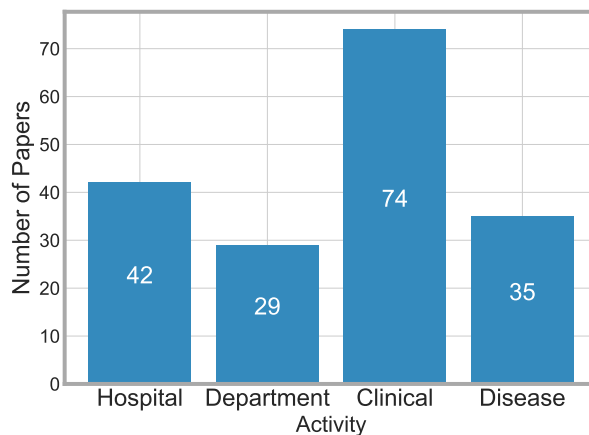


Figure 2.9: Frequency of Papers by Scope.

There were only five cases where more than one scope was considered [21, 67, 148, 180, 335]. In all of these cases both clinical and disease scope was considered. This was due to the clinical activities being dependent on the progress of disease at particular points e.g. Liu et al., [180] investigate the readmission risk percentage

based on the patient activities which differ depending on the diagnosis of the disease.

From the selected literature, it appears that considering more than one scope area is difficult to carry out in a realistic format, which is not in the form of dummy or pseudo activities. It is believed that this is a limitation of the types of methods (Figure 2.10 and Table B.9) that are considered and thus suggests an opportunity for further work.

2.4.3 Technical Context

Method

There are many methods that can be used for clinical pathways, which will now be categorised into four groups: Stochastic Modelling, Data Mining or Machine Learning, Simulation, and Optimisation and Heuristics. Further description of what methods are included, but not limited to, in each group are as follows:

- **Stochastic Modelling** - Includes Markov [64] and queueing [284] methods.
- **Data Mining or Machine Learning** - Includes Bayesian techniques and Bayesian Belief Networks [199], machine learning [108] and visualisation [30].
- **Simulation** - Includes discrete-event [111], agent based [181], Monte Carlo [10] and system dynamics [198, 285].
- **Optimisation and Heuristics** - Includes genetic algorithm [88], and mathematical programming, including dynamic [281], mixed-integer [107], mixed-integer linear [44] and goal [241].

Figure 2.10 displays the frequency of papers in each method group, and indicates that data mining or machine learning was the most popular method to be applied, closely followed by simulation.

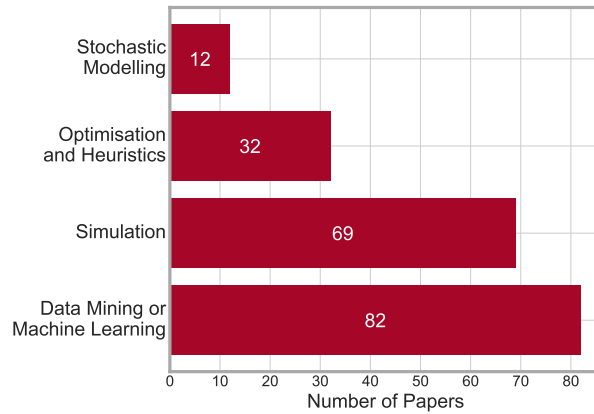


Figure 2.10: Frequency of Papers Applying Method Type.

Eighteen papers were identified as using multiple methods, 16 of those papers applied two methods and two papers applied three methods. This is just 10% of the total selected papers. The combinations of methods applied are displayed in Figure 2.11.

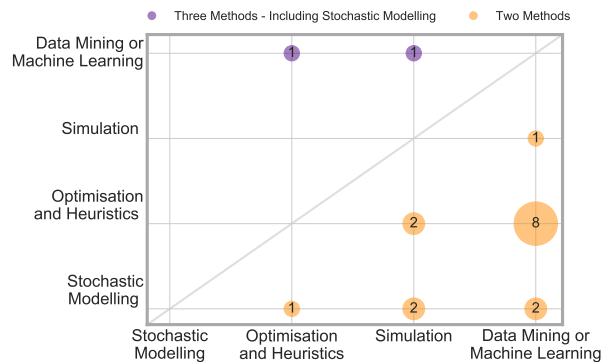


Figure 2.11: Frequency of Papers Applying Multiple Methods.

The majority of these papers use data mining or machine learning along with one other method, and thus shows that those papers using multiple techniques are bridging the gap between OR, IS and Industrial Engineering.

The interaction between method and condition can be considered, and is displayed in Figure 2.12.

Figure 2.12 shows that data mining or machine learning more frequently considers applied conditions, and simulation more frequently has the condition as the focus of the paper, in all three condition areas.

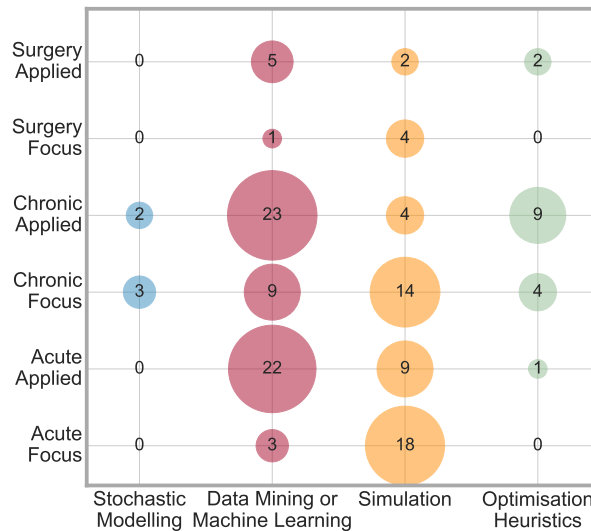


Figure 2.12: Frequency Cross Analysis Between Method and Condition Area.

Furthermore, six papers [3, 22, 35, 215, 238, 324] discuss the use of ‘Business Process Modeling Notation’ (BPMN). BPMN is the use of graphical notation for the purpose of illustrating business processes.

Fourteen papers [1, 19, 21, 28, 66, 115, 124, 141, 168, 179, 180, 201, 228, 333] indicate that they develop a type of IT artifact that can be implemented to support the clinical pathways under consideration. These papers are also bridging the gap between OR, IS and Industrial Engineering.

This highlights that to continue bridging the gap between OR, IS and Industrial Engineering future work should consider Data Mining and Machine Learning alongside OR techniques, and integrate them whenever possible.

Investigating Area

The literature discusses three ways of investigating the pathway: mapped, modelled, and improved. A paper is classed as mapping a pathway if it provides some information and process of initially defining the pathway, modelling if it created a model of that pathway, and improved if some scenario analysis, recommendation or support for improvement was made. It is possible for a paper to consider more than one of these investigation areas.

Figure 2.13 displays the frequency of papers considering each investigation area.

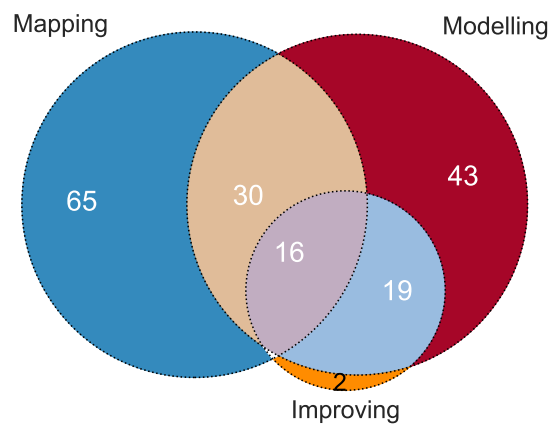


Figure 2.13: Graph of the Interaction Between Mapping, Modelling and Improving the Pathway.

The two papers that were categorised exclusively as improving discussed the development of a web-based tool to aid with clinical pathway usage, and thus did not map or model the pathway. There are no papers that both map and improve the pathway, without also modelling it. This is intuitive, as a model cannot be improved if it was not modelled.

Figure 2.13 concludes that all three investigation areas are important when considering clinical pathways, and applying all three provides a more complete picture. It is suggested that future work place more focus/importance on improving the pathway and its related outcomes, as this is one of the key advantages of using an OR technique, and can aid decision-making.

The investigation area that is considered is related to the type of method used, as displayed in Figure 2.14.

Figure 2.14 displays that the most frequently used techniques to map a pathway are data mining and machine learning, whereas simulation is the most popular technique for considering modelling or improving a pathway.

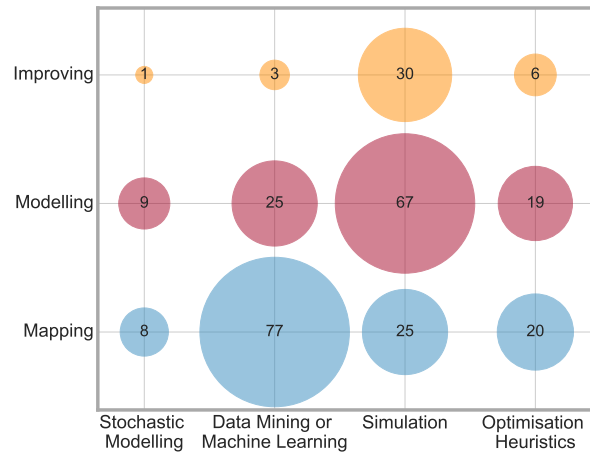


Figure 2.14: Frequency Cross Analysis Between Method and Investigation Area.

Outcome

The outcome, main decision variable or indication factor for performance of interest can lead the whole direction of research. The outcomes considered in the literature can be grouped into six categories. A description of the categories are as follows:

- **Legal:** Papers including factors of a legal matter, such as fraud or medical negligence.
- **Patient Progression:** Any factor related to the patient specifically e.g. Quality Adjusted Life Years (QALY), survival, disease progression/management.
- **Cost:** This category includes any paper related to cost.
- **Resource:** Any factor considered to be a resource e.g. MRI Scanner, capacity, staffing levels.
- **Time:** Any factor related to time is included in this category e.g. length of stay, scheduling, waiting times and travel times.
- **Pathway Mapping:** Papers that aimed to establish and map the pathway, including pathway variances are included here.

Figure 2.15 shows the frequency of papers amongst these outcomes. Pathway mapping is the most frequent category, whilst (excluding legal) patient progression is the

least frequent. This may be concerning as the patients are those whose health and lives are effected by all of the outcome factors, and thus should be at the forefront of any outcome considered. Therefore it is recommended that more emphasis should be placed on patient outcomes in a more direct manner.

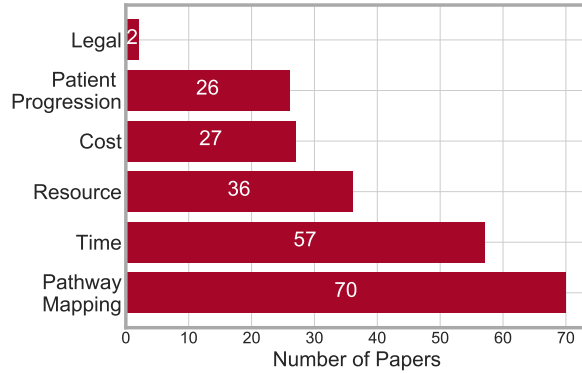


Figure 2.15: Frequency of Papers Considering Outcome Measure.

Thirty-seven papers considered multiple outcomes, where 32 considered two outcomes, and four considered three outcomes (Figure 2.16).

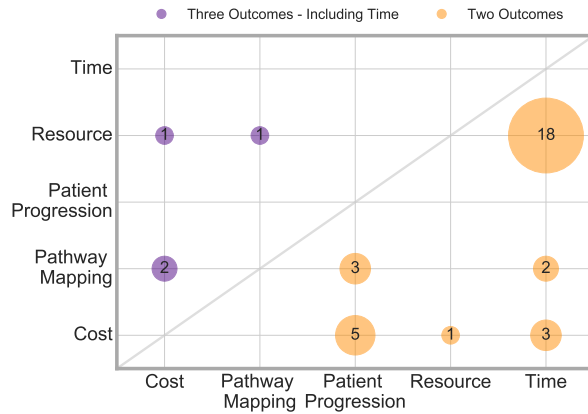


Figure 2.16: Frequency of Considering Multiple Outcomes.

Figure 2.16 shows that time and resource is most frequently considered together, and it is rare to find papers considering more than two outcome measures.

Only one paper considered four outcomes [203] (time, resource cost and patient progression), which is not displayed on Figure 2.16.

Although an outcome is often regarded as the final result of any research, this also has an impact on the areas surrounding constructing the approach, such as the

method or scope considered.

Figure 2.17 shows the frequencies of the cross analysis between outcome and method. This displays that data mining or machine learning is most frequently used for pathway mapping, whereas simulation is most frequently used to measure cost, resource or time outcome measures.

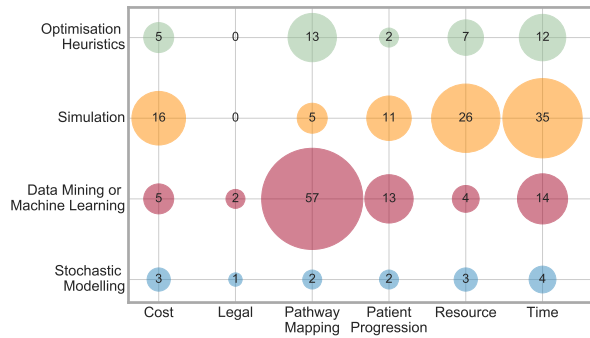


Figure 2.17: Frequency Cross Analysis Between Outcome and Method.

Figure 2.18 shows the frequencies of the cross analysis between outcome and scope. It displays that a clinical scope is most frequently used for pathway mapping, whereas resource and time are approximately equally split between hospital and departmental scope.

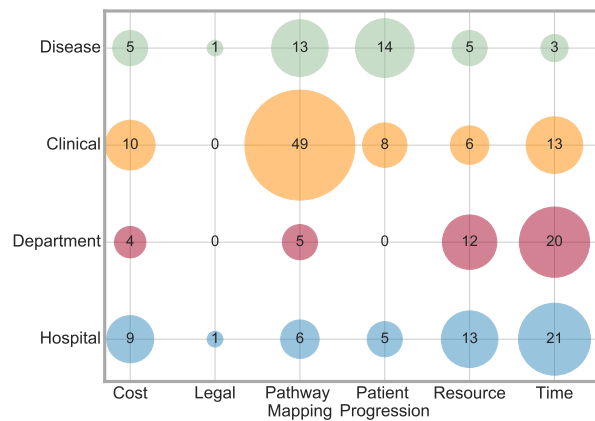


Figure 2.18: Frequency Cross Analysis Between Outcome and Scope.

As an example of the interaction between outcome factor and scope, Barone et al., [23] considered departmental scope in relation to time, resource and cost outcome factors through simulation to plan daily nurse requirements in a stroke unit. In contrast, Uzun Jacobson et al., [280] considered a clinical scope in relation to patient

progression outcomes, through discrete-event simulation of hyper-acute stroke care, concerning the percentage of patients receiving thrombolysis.

2.4.4 Planning Decisions

Hulshof et al., [138] describes a taxonomic classification of planning decisions in health care in OR/MS. This taxonomy separates the papers into three decision levels: Strategic, Tactical and Operational. A brief description of the decision levels are as follows, however a formal definition of the three decision levels can be found in Hulshof et al., [138].

- Strategic planning involves structural decision making of the design, dimensioning and development of healthcare. This typically has a long planning horizon e.g. location planning and staffing levels.
- Tactical planning organises the operation of the healthcare delivery system, typically on a mid-term planning horizon, e.g. staff shift scheduling.
- Operational planning executes the routine planning of the healthcare delivery system on a short-term planning horizon e.g. patient-to-appointment scheduling.

Figure 2.19 shows the frequency of the papers in each decision level. This highlights that strategic decisions are considered most frequently out of the three decision levels, however more often than not, there is no decision to consider.

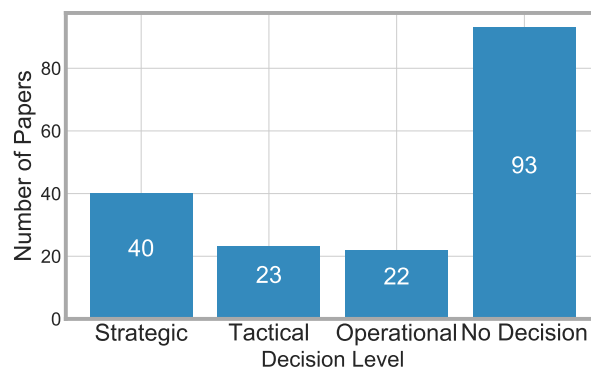


Figure 2.19: Frequency of Papers Considering Decision Level.

Three papers state that they consider more than one decision level. Barbagallo et al., [22] states that it considers both strategic and tactical decisions, Landa et al., [161] considers tactical and operational decisions and Burdett et al., [44] consider strategic and operational decisions.

This shows that the use of clinical pathways can be used across all the decision levels, from day-to-day decisions to wider policy decisions.

Hulshof et al., [138] applies the taxonomy for those papers in the OR/MS JCR category, however this is extended here to consider five JCR groups. The cross analysis between decision level and JCR category can be seen in Figure 2.20.

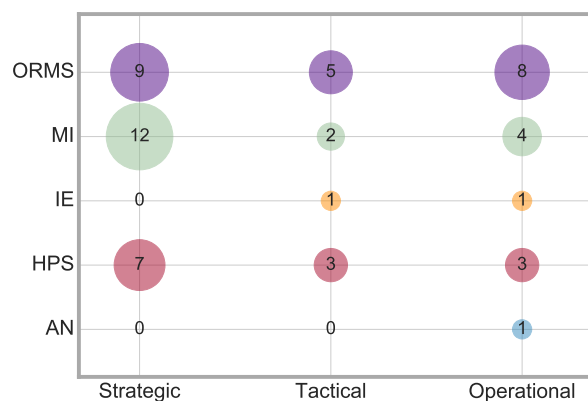


Figure 2.20: Frequency Cross Analysis Between Decision Level and JCR Category.

Figure 2.20 shows that the decision levels are in fact spread across the five JCR groups, which shows that the Hulshof et al., [138] decision level taxonomy can be applied to more than just the OR/MS JCR group.

The decision level does impact other aspects of the research that have been previously discussed. Therefore, a cross analysis between the decision level with scope, method, and outcome will now be considered. The cross analyses between decision level and method, and decision level and outcome, both help to explain why such a high number of papers refrain from considering a decision level.

Firstly, Figure 2.21 considers the interaction between decision level and scope of the research.

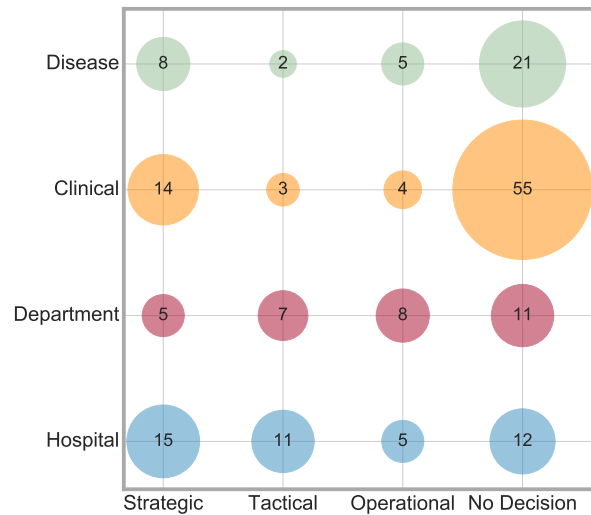


Figure 2.21: Frequency Cross Analysis Between Decision Level and Scope.

There appears to be an even dispersion of scope across the three decision levels, with strategic being most popular in clinical and disease scope than the other two decision levels. Considering the papers that had no decision level, these are most often concerning clinical scope, but there is also an equal spread between the three remaining scope areas.

Secondly, Figure 2.22 considers the interaction between decision level and method.

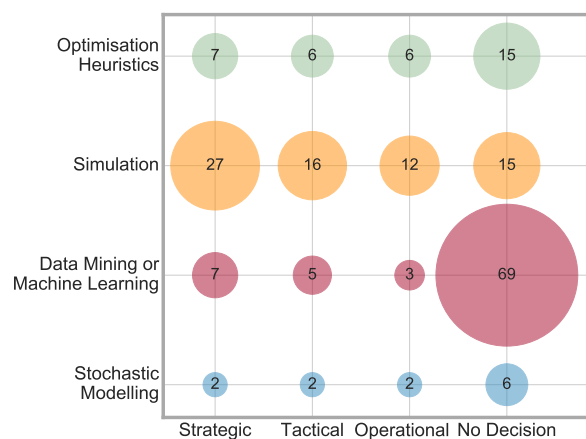


Figure 2.22: Frequency Cross Analysis Between Decision Level and Method.

This shows that simulation is most frequently used across all three decision levels. The interaction between data mining or machine learning and no decision was most commonly observed. This can be explained as this method is most frequently used for mapping a pathway (Figure 2.14) for reasons such as defining the pathway, and

therefore would have no decision associated with this.

The conclusion drawn from the above analyses can be supported when considering the interaction between decision level and outcome measure (Figure 2.23).

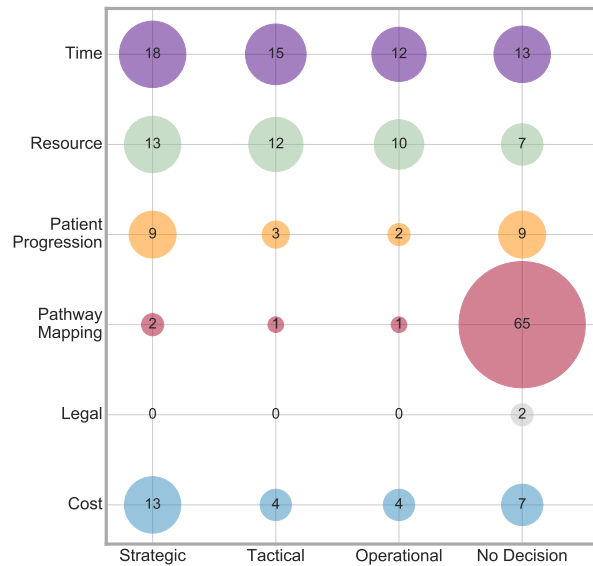


Figure 2.23: Frequency Cross Analysis Between Decision Level and Outcome.

Figure 2.23 shows that no decision most frequently occurs when the outcome is pathway mapping. Again, all of the outcome measures have an even distribution across all three decision levels, with strategic decisions being slightly more prominent.

2.5 Conclusions

There is a vast scope for what can encompass the term clinical pathway, with numerous ways of formulating, approaching, and modelling these. This literature review itself provides a novel contribution through the development of a number of taxonomies, providing a detailed classification of the publications. This enables clarity for any future publications surrounding clinical pathways to identify the current themes and methods used in the literature, and thus identify gaps.

Section 2.4 discussed the taxonomy, analysed frequencies, and provided cross anal-

ysis where appropriate. Some areas of recommended focus for future work were highlighted in the discussion, and are summarised as follows:

- Careful consideration of publication area, to ensure the information is reaching all communities involved.
- Derive the pathway from both data and collaboration with staff.
- Consider a medical condition, whether in focus or applied application, as this better fits with clinical pathways specifically.
- Include as many care levels as possible (when appropriate) to encourage communication and awareness between them.
- Improve the methods used to allow for multiple types of scope to be considered together.
- Continue to bridge the gap between OR, IS and Industrial Engineering by considering data mining and machine learning alongside OR techniques, and integrate whenever possible.
- Incorporate all three areas of mapping, modelling and improving the pathway, with particular focus on improving, as this reflects the specialities of OR techniques.
- Greater emphasis on patient outcomes in a more direct manner.
- Specify the decision planning level of focus when appropriate.

Following these recommendations should lead to a more thorough study of the whole clinical pathway. The paper from Monks et al., [203] presents a methodology for simulation modelling of stroke care systems, and captures many of the same recommendations as discussed above, and thus is a notable example of what future research should aim to aspire to.

The inclusion of the cross analysis between the identified taxonomy areas allows

those who are considering research to more carefully consider the combinations of these areas, both for quality, appropriateness and discovering areas in which there is a lack of research.

In conclusion, future work should consider Industrial Engineering and IS integrated with OR techniques, with an aim to improve the handling of multiple scope within one model, whilst encouraging interaction between the previously disjoint care levels, with a more direct focus on patient outcomes. Achieving this would continue to bridge the gap between OR, IS and Industrial Engineering, whilst improving methods for clinical pathways to aid in supporting decisions.

Chapter 3

Modified Needleman-Wunsch

Metric

Research Question 1

Can both data and expert information integrate to inform clinical pathway mapping?

This chapter mainly comprises of the research paper ‘*Modified Needleman-Wunsch Algorithm for Clinical Pathway Mining*’ [16], which discusses the development of a new distance metric, to allow for consideration of both data and medical expert information, for the use with k-medoids clustering. The Modified Needleman-Wunsch (MNW) algorithm has been specifically designed for clustering and allows for expert interaction through the use of groupings and weightings of activities, to provide context to the pathway strings. Eight other popular distance metrics are discussed and used as reference for benchmarking the performance of the modified metric.

From the findings of Chapter 2, this chapter aims to address research question 1, and is structured as follows: Section 3.1 provides a brief introduction, Section 3.2 contains a discussion of previous research and Section 3.3 gives a description of the working dataset. Then Section 3.4 discusses some current metrics and Section 3.5

their properties. Section 3.6 presents the development of the modified algorithm and Section 3.7 applies the method to case studies. In addition to source paper [16], Section 3.8 provides sensitivity analysis of the key variables. Section 3.9 closes the chapter with a conclusion and recommendations for further work.

3.1 Introduction

In the age of digital health, the organisation of health information into interactive clusters and other novel methods for stratifying health data will complement existing approaches and potentially lead to improvements in health care [261]. As health information technology (IT), such as electronic health records (EHRs), gain widespread adoption and use in healthcare industry, thereby accumulating vast amounts of real-time patient care data, there is tremendous opportunity to develop data-driven models, methods and tools to facilitate review of practice workflows and improve evidence-based care delivery by learning practice-based pathways of care [96, 335], henceforth denoted as clinical pathways.

When considering clinical pathway modelling, a primary question is often to consider what is the pathway. Chapter 2 highlighted that there are many data mining and machine learning methods available for answering such questions. However, it was clear that most of these techniques only consider the pathways discoverable from data, and do not consider the wealth of information available from the experts that interact with the pathway day to day. The benefit of consulting with experts is that they may be able to explain some obscure or outlier information that can be picked up within the data. It is speculated that the lack of interaction between using both data and expert knowledge is due to the time consuming nature of such a process.

Exploring the literature (Chapter 2) further, clustering techniques were highlighted as the most popular method for pathway discovery. Thus, this chapter focuses on distance measures applied to string data for the purpose of k-medoids clustering [214]. This method was chosen as firstly the data used is similar to that of Vogt

et al. [287], and secondly using an existing pathway as the medoid reinforces the medical experts' confidence in the pathway chosen as being realistic. Clustering methods would be suitable for both DT1 and DT2, making it more versatile and applicable.

3.2 Previous Research

From the literature in Chapter 2, there were 82 papers which stated using data mining or machine learning, for mapping, modelling or improving the clinical pathway. Table 3.1 further categorises these papers into specific method areas.

Table 3.1: Publications Categorised as Data Mining or Machine Learning Method.

Method	
<i>Clustering</i>	[54, 79, 102, 103, 115, 156, 160, 178, 207, 258, 277, 287, 335]
<i>Categorised</i>	[120, 131, 132, 199, 332]
<i>Classified</i>	[122, 327]
<i>Topic Modelling</i>	[321, 322]
<i>Probabilistic</i>	[125, 130, 131]
<i>Latent Dirichlet Allocation</i>	[127–129, 135, 136, 323, 327, 328]
<i>Pattern Mining</i>	[132, 160, 325]
<i>Sequential Pattern Mining</i>	[11, 74, 106, 228, 258, 260, 269, 275, 279]
<i>Temporal Pattern Mining</i>	[74, 78, 172]
<i>Process Mining</i>	[49, 93, 124, 134, 156, 160, 178, 189, 221, 241, 267, 320, 323]
<i>Bayesian</i>	[10, 98, 108, 180, 199]
<i>Markov</i>	[6, 19, 193, 194, 335]
<i>Heuristics</i>	[85, 103, 137, 150, 234]
<i>Semantic Web Rule Language</i>	[126, 201]
<i>Artefact</i>	[30, 66, 117, 180, 333]
<i>Business Process Model and Notation (BPMN)</i>	[35, 238, 324]
<i>Other</i>	[42, 104, 116, 155, 176, 179, 185, 212, 295, 319, 332]

It can be seen that clustering was the most popular method. On closer inspection there are multiple methods of clustering used, for example, Funkner et al., [103] use K-means, Vogt et al., use K-medoids [287] and Zhang et al., use hierarchi-

cal [335]. Furthermore, the differences go deeper when considering the distance measures used during clustering, as Funkner et al., [103] uses Levenshtein distance, Syed and Dias [269] modify the Needleman-Wunsch Algorithm, whereas Vogt et al., [287] and Zhang et al., [335] use Longest Common Subsequence (LCS).

Chapter 2 also highlighted that there are two common ways of obtaining the pathway: either data-driven or through collaboration with experts who regularly interact with the pathway. Data-driven pathway discovery was most popular, containing 90 papers, compared to 13 papers that considered collaboration only.

Chapter 2 state that there are 14 papers that considered information from both of these sources [20, 30, 52, 64, 87, 147, 149, 181, 189, 204, 221, 240, 259, 280]. All of these papers consider data alongside expert opinion, interviews or literature, and do so in a way that they enhance or fill in for missing information.

None of the papers integrate the two sets of information in a simple and direct manner. Furthermore, considering just one of these methods leaves a wealth of knowledge that is not considered.

3.3 Working Dataset - ‘Dataframe’

The main dataset contains 2,350 non-small cell lung cancer referrals provided by Velindre Cancer Centre (VCC) and 13 activities (presented in Table 3.2). The data set used is of DT1 and as such has the ‘at most once’ constraint. This must be considered when choosing the method to apply.

After extracting the patient pathways (as described in subsection 1.2.4), this resulted in 1,019 unique pathways, which are contained within the dataset ‘dataframe’. It is a key technical restriction of constructing the distance matrix that there are no repeated pathways, and thus the ‘dataframe’ is used for the remainder of this chapter.

To aid with visualization of this, Figure 3.1 shows a heatmap displaying the pathways in the ‘dataframe’, where the data has been ordered alphabetically. Along the x-axis is the position of the activity, and the y-axis is the number of unique pathways, where each integer represents one pathway. Furthermore, each activity code has been assigned a colour, and thus the heatmap represents the unique pathways as a line of various colours.

Figure 3.1 shows that there is a large amount of variation in the position, number and sequence of the activities performed. This indicates that condensing this large variation into a simple clinical pathway to be used as guideline is a difficult task.

Table 3.2: Pathway Activity Names and Assigned Letters.

Activity Name	Letter
First Seen	A
Diagnosis	B
MDT Discussion	C
Procedure	D
Decision to Treat Chemo	E
Chemotherapy Start	F
Decision to Treat Tele	G
Teletherapy Start Date	H
Decision to Treat Brachytherapy	I
Brachytherapy Start Date	J
CT Scan	K
PET/PET CT Scan	L
Bronchoscopy	M
CT Guided Biopsy	N
Specialist Nurse Seen	O

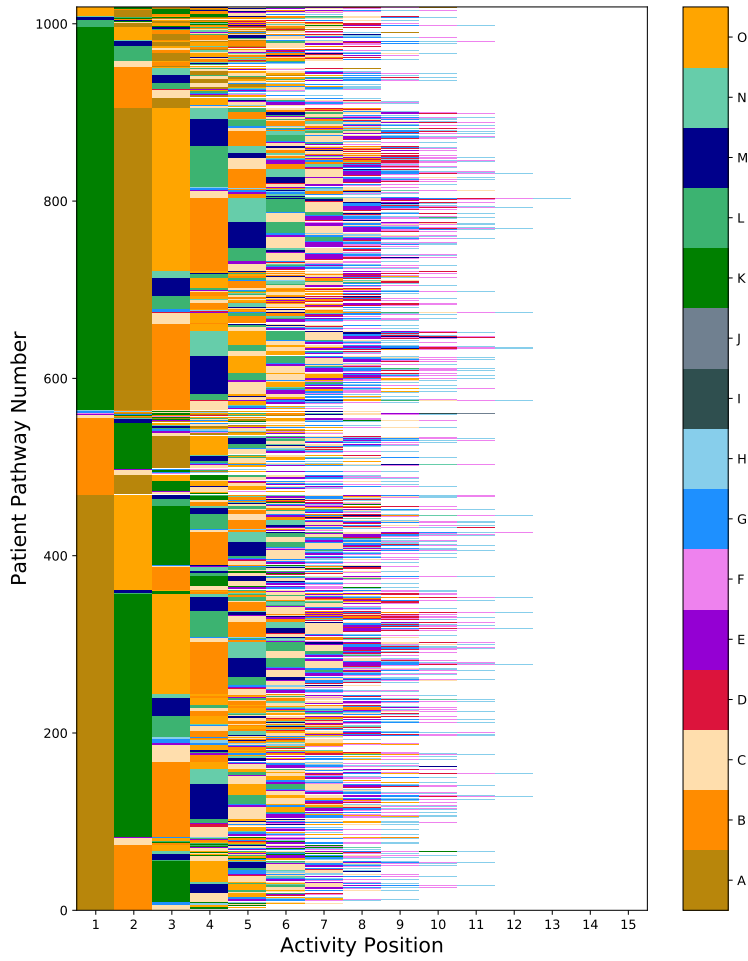


Figure 3.1: All Unique Pathways Displayed as a Heatmap.

3.4 Description of Metrics

There are many different possible metrics that can be used to compare two strings, given that the Python library `textdistance` [274] (a library to compare distance between two or more sequences) hosts over 30 algorithms for this purpose. Eight different metrics were considered to use as comparison and benchmarking for the modified algorithm, which cover edit distances, token based and sequence based distances. These eight metrics were chosen as they most appropriately fit the purpose, reflect the literature and show a variety of techniques.

Dynamic Programming

A few of the distance metrics utilise dynamic programming for their calculation. This section briefly describes the general process of the calculation, where specific details are discussed in the algorithms of the relevant metrics.

Wolsey and Nemhauser define dynamic programming as: “*Dynamic programming provides a framework for decomposing certain optimization problems into a nested family of subproblems. This nested structure suggests a recursive approach for solving the original problem from the solutions of the subproblems.*” [303]

In general the calculation recursively builds a matrix (referred throughout as dynamic programming matrix or X) that compares two strings. Instead of comparing the string as a whole, the dynamic program compares a pair of characters (one from each string) at a time. Each pair is evaluated and given a score based on their relationship, adjusting a previous value in the matrix, as defined by the metric used. It is typical for the pairs to be evaluated as either ‘aligned’ - receiving a value corresponding to either a) the characters are the same (match m) or b) different (swap s), or not ‘aligned’ and thus forcing in a blank space (gap g). The general values m , s and g are referred to as penalty values.

There are two main steps, ‘Initialise’ and ‘Fill Matrix’. These will generally be explained using two strings, P1 and P2, where P1 = ‘BKAOC’ and P2 = ‘KABCO’.

- Initialise: To construct the matrix, insert a blank space at the start of each string, then place the first string (P1) in the initial column and the second string (P2) in the initial row (see Figure 3.3 for example). The initialisation then fills the first column and first row as if each character in the string was aligned with the blank space (placing an initial 0 in the upper left corner). Essentially, this will be the position of the character multiplied by the gap penalty value (g) for that metric.
- Fill Matrix: Then working in rows, a character of P1 is evaluated against each

character of P2 in turn and given a score based on the rules of the metric being used, before moving on to the next row e.g. P1('B') will be compared with each character of P2 in turn, then repeat with P1('K') etc. Finally, the value in the bottom right hand corner gives the total score for the comparison.

Edit distances

Here there are five edit distances considered: Levenshtein, Damerau-Levenshtein, Jaro, Jaro-Winkler and Needleman-Wunsch.

Levenshtein

The Levenshtein distance was developed in 1965 for the use of correcting deletions, insertions and reversals of binary codes [171]. The general idea is to evaluate the distance between two strings as the number of single-character edits required to change one string into the other. There are many current uses for the Levenshtein distance, e.g. spell checkers, optimal character recognition correction systems and linguistic distance, to name a few.

The Levenshtein distance can easily be calculated by hand, by giving a penalty of one to each insertion, deletion or substitution, as demonstrated in Figure 3.2.

The Levenshtein distance can be translated into a dynamic programming algorithm displayed in Algorithm 1. The dynamic programming matrix X for the example from Figure 3.2 can be seen in Figure 3.3.

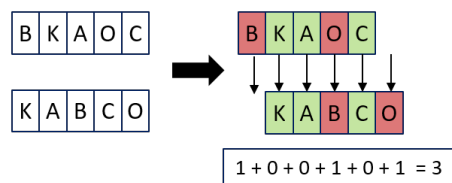


Figure 3.2: Example of the Calculation for the Levenshtein Distance.

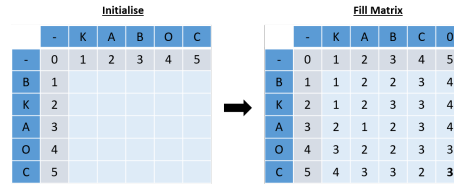


Figure 3.3: Example of Dynamic Programming Using the Levenshtein Distance.

Algorithm 1 Levenshtein Distance.

```

1: procedure LEVENSHTTEIN ▷ Initialise
2:   Insert a blank space at the start of each string
3:   for  $i \leftarrow 0, \text{len}(P1)$  do
4:      $X[i][0] = i$ 
5:   end for
6:   for  $j \leftarrow 0, \text{len}(P2)$  do
7:      $X[0][j] = j$ 
8:   end for
▷ Fill Matrix
9:   for  $i \leftarrow 1, \text{len}(P1)$  do
10:    for  $j \leftarrow 1, \text{len}(P2)$  do
11:      if  $P[i] == P[j]$  then
12:         $X[i][j] = X[i - 1][j - 1]$ 
13:      else
14:         $\min(X[i - 1][j - 1], X[i][j - 1], X[i - 1][j]) + 1$ 
15:      end if
16:    end for
17:  end for
18:  return  $X[\text{len}(P1)][\text{len}(P2)]$ 
19: end procedure

```

Damerau-Levenshtein

The Damerau-Levenshtein is an extension of the Levenshtein distance, where transpositions (swapping positions of adjacent characters) are also allowed [76].

An example of the hand calculation for the Damerau-Levenshtein distance can be seen in Figure 3.4. Again, this can also be performed using dynamic programming.

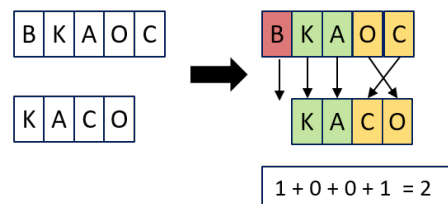


Figure 3.4: Example of the Calculation for the Damerau-Levenshtein Distance.

Jaro

The Jaro similarity was first developed for the purpose of record linkage [144, 145]. The formula considers four variables: the length of both strings (a, b), the number of matching characters (m) within position tolerance (T), and the number of transpositions of those matching characters (t). The formula for Jaro similarity is:

$$sim_{jaro} = \begin{cases} 0, & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{a} + \frac{m}{b} + \frac{m-t}{m} \right), & \text{otherwise} \end{cases} \quad (3.1)$$

where the tolerance (T) for m is calculated by

$$\left[\frac{\max(a,b)}{2} \right] - 1,$$

and only the integer-part is used. For further clarity, the match variable is an injection, where a character from one string can be mapped to at most one other character in the other string.

This will produce a value between 0 and 1, where 1 indicates that the strings are identical, and therefore a larger value is desired. To calculate the distance instead of similarity, the metric needs to be adjusted by performing $1 - sim_{jaro}$.

For example, in Figure 3.5 there are 4 matches within the tolerance of 1 (see below), shown in green, however 'C' and 'O' need to be transposed. The calculations for the example in Figure 3.5 are: $a = 5, b = 5, T = 5/2 - 1 = 1, m = 4, t = 1$

$$sim_{jaro} = \frac{1}{3} \left(\frac{4}{5} + \frac{4}{5} + \frac{4-1}{4} \right) = 0.783 \quad (3.2)$$

$$1 - sim_{jaro} = 0.216 \quad (3.3)$$

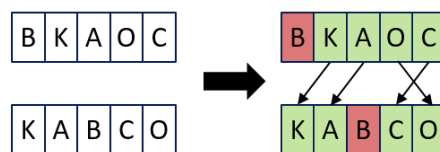


Figure 3.5: Example of the Calculation for the Jaro Distance.

Jaro-Winkler

The Jaro-Winkler distance is an extension of the Jaro distance [146] through the following formula:

$$sim_{winkler} = sim_{jaro} + (l \cdot p(1 - sim_{jaro})) \quad (3.4)$$

where l is the number of common prefix i.e. the number of characters that match before the first non-match occurs (up to a maximum of 4), and p is a scaling factor which should not exceed 0.25. Typically p is chosen to be 0.1. Again, to calculate the distance instead of similarity, the metric needs to be adjusted by performing $1 - sim_{winkler}$.

Applying this calculation to the example, as l is 0 (because the first position is a non-match), we would get the same result as the Jaro distance (0.2166666667). Therefore, the example is changed slightly as shown in Figure 3.6.

Then first calculating the Jaro distance to allow for calculating the Jaro-Winkler distance is as follows: $a = 4, b = 5, T = 5/2 - 1 = 1, m = 3, t = 0$

$$sim_{jaro} = \frac{1}{3} \left(\frac{3}{4} + \frac{3}{5} + \frac{3-0}{3} \right) = 0.78\dot{3} \quad (3.5)$$

$$1 - sim_{jaro} = 0.21\dot{6}7 \quad (3.6)$$

$$sim_{winkler} = 0.78\dot{3} + (2 \cdot 0.1(0.21\dot{6})) \quad (3.7)$$

$$= 0.82\dot{6} \quad (3.8)$$

$$1 - sim_{winkler} = 1 - 0.82\dot{6} = 0.17\dot{3} \quad (3.9)$$

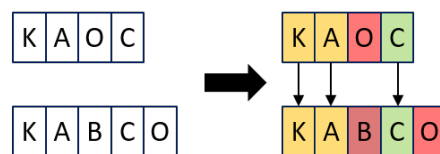


Figure 3.6: Example of the Calculation for the Jaro-Winkler Distance.

Needleman-Wunsch

The Needleman-Wunsch algorithm was first used in bio-informatics to align protein or nucleotide sequences, and makes use of dynamic programming [208]. It may also be referred to as the optimal matching algorithm or the global alignment technique.

This is a generalised variant of the Levenshtein distance, where values for match, swap and gap are chosen by the user. The most common values chosen for these variables are: Match (m) = 1, Swap (s) = -1 and Gap (g) = -1. Again, this can easily be checked by hand, as shown in Figure 3.7.

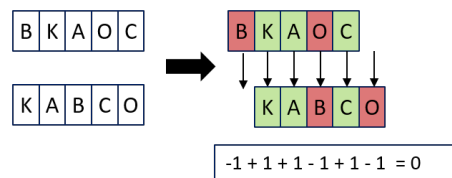


Figure 3.7: Example of the Calculation for the Needleman-Wunsch Distance.

Using these values for m , s and g would result in a larger number indicating a closer match, and thus for clustering the final value would need to be adjusted.

The Needleman-Wunsch algorithm also makes use of dynamic programming to computationally calculate the distance. The pseudo-code for which can be seen in Algorithm 2. Conversely to standard dynamic programming, for each pair of characters three values are calculated: D will use the previous diagonal value to consider an alignment (match or swap), L will use the value to the left to align the character from string P2 with a gap, and T will use the upper value to align the character from string P1 with a gap. An example of the matrix produced using the Needleman-Wunsch dynamic programming algorithm can be seen in Figure 3.8.

Once the matrix such as that in Figure 3.8 has been produced, we can perform traceback to find the alignment. This means, starting at the bottom-right corner of the matrix, and working back through the matrix to the top-left 0, and noting the direction that the value came from. This is highlighted in Figure 3.8 by the black arrows.

Algorithm 2 Needleman-Wunsch Algorithm.

```

1: procedure NEEDLEMAN-WUNSCH ▷ Initialise
   Insert a blank space at the start of each string
2:    $m = 1, g = -1, s = -1$ 
3:   for  $i \leftarrow 0, \text{len}(P1)$  do
4:      $X[i][0] = i \cdot g$ 
5:   end for
6:   for  $j \leftarrow 0, \text{len}(P2)$  do
7:      $X[0][j] = j \cdot g$ 
8:   end for ▷ Fill Matrix

9:   for  $i \leftarrow 1, \text{len}(P1)$  do
10:    for  $j \leftarrow 1, \text{len}(P2)$  do
11:     if  $P[i] == P[j]$  then
12:        $D = X[i - 1][j - 1] + m$ 
13:     else
14:        $D = X[i - 1][j - 1] + s$ 
15:     end if
16:      $L = X[i][j - 1] + g$ 
17:      $T = X[i - 1][j] + g$ 
18:      $X[i][j] = \max(D, L, T)$ 
19:   end for
20: end for
21: return  $X[\text{len}(P1)][\text{len}(P2)]$ 
22: end procedure

```

	-	K	A	B	C	O
-	0	-1	-2	-3	-4	-5
B	-1	-1 -2	-2 -2	-3 -1	-4 -2	-5 -2
K	-2	0 -2	-1 -1	-2 -2	-3 -2	-4 -3
A	-3	-3 -1	1 -2	0 0	-1 -1	-2 -2
O	-4	-4 -2	-2 0	0 -1	-1 -1	0 -3
C	-5	-5 -3	-3 -1	-1 -1	1 -2	-2 -1

Figure 3.8: Example of the Needleman-Wunsch Algorithm.

A diagonal arrow indicates an alignment, an arrow to the left indicates that the character in the top string is aligned with a gap, and an arrow straight up indicates that the character in the left string is aligned with a gap.

The `textdistance` library [274] does not easily allow for alternative values of m , s and g to be used. Furthermore, the online demo [174] is a useful resource.

Token Based

Here are discussed two token based distances, Jaccard and Cosine respectively. In this context token means a partition of the string, and in both of these distances this relates to n-grams. Furthermore, an n-gram is defined as a continuous sequence of n items i.e. split the string into substrings of length n, starting at the beginning of the string and shifting over one place until you reach the end of the string.

An example of n-grams, where $n = 2$ (bi-gram), for the two strings ‘BKAOC’ and ‘KABCO’ can be seen in Figure 3.9. Here the bi-grams are displayed in a table, where all of the bi-grams are listed in columns, and a 1 indicates that the string contains the bi-gram and 0 otherwise.

		BK	KA	AO	OC	AB	BC	CO
P1	BKAOC	1	1	1	1	0	0	0
P2	KABCO	0	1	0	0	1	1	1

Figure 3.9: Example of bi-gram.

Jaccard Distance

The Jaccard distance [143] is calculated using the following equation.

$$\frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (3.10)$$

where A is the set of n-grams in P1, and B is the set of n-grams in P2. Essentially $|A \cup B|$ is the total number of n-grams, and $|A \cap B|$ is the number of shared n-grams.

Applying Equation 3.10 using the bi-grams from the example in Figure 3.9 yields (to 2.d.p):

$$\frac{7 - 1}{7} = \frac{6}{7} = 0.86 \quad (3.11)$$

Cosine Distance

The Cosine distance is typically used to compare the number of similar words in a document and also in data mining to measure cohesions in clusters [68].

Firstly, calculating the Cosine similarity using the equation:

$$\frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (3.12)$$

where A_i and B_i are the values for each n-gram (column) i for string P1 and P2 respectively, and n is the total number of n-grams.

Then, as previously, 1 minus the similarity needs to be performed to obtain the distance. Applying this calculation to the example with previously calculated the bi-grams in Figure 3.9. This results in a cosine distance of $1 - 0.25 = 0.75$.

$$\begin{aligned} \sum_{i=1}^n A_i B_i &= (1 \cdot 0) + (1 \cdot 1) + (1 \cdot 0) + (1 \cdot 0) + (0 \cdot 1) + (0 \cdot 1) + (0 \cdot 1) = 1 \\ \sum_{i=1}^n A_i^2 &= 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 = 4 \\ \sum_{i=1}^n B_i^2 &= 0^2 + 1^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2 = 4 \\ \frac{1}{\sqrt{4}\sqrt{4}} &= 0.25 \end{aligned}$$

Sequence Based***Longest Common Subsequence***

The longest common subsequence (LCS) refers to the longest subsequence common to both sequences, where the subsequences do not have to occupy consecutive positions, but do have to be in sequence. Figure 3.10 displays that the LCS for the example is 3, where Figure 3.11 illustrates the dynamic programming calculation.

To consider LCS as a distance, we need to consider what remains when you remove the LCS. In order to calculate the distance, subtract the LCS value from the

maximum length of the two strings. Essentially in Figure 3.10 this would be what remains in white, and therefore would give a LCS distance of 2.

It has been shown that this is an NP-hard problem [186], and as such dynamic programming has been utilised to allow for computation. The pseudo-code for the dynamic programming of the LCS can be seen in Algorithm 3.

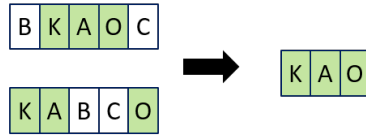


Figure 3.10: Example of Longest Common Subsequence.

Initialise							Fill Matrix						
	-	K	A	B	O	C		-	K	A	B	C	O
-	0	0	0	0	0	0	-	0	0	0	0	0	0
B	0						B	0	0	0	1	1	1
K	0						K	0	1	1	1	1	1
A	0						A	0	1	2	2	2	2
O	0						O	0	1	2	2	2	3
C	0						C	0	1	2	2	3	3

Figure 3.11: Example of Longest Common Subsequence Dynamic Programming.

Algorithm 3 Longest Common Subsequence.

```

1: procedure LCS ▷ Initialise
2:   Insert a blank space at the start of each string
3:   for  $i \leftarrow 0, \text{len}(P1)$  do
4:      $X[i][0] = 0$ 
5:   end for
6:   for  $i \leftarrow 0, \text{len}(P2)$  do
7:      $X[0][j] = 0$ 
8:   end for ▷ Fill Matrix
9:   for  $i \leftarrow 1, \text{len}(P1)$  do
10:    for  $j \leftarrow 1, \text{len}(P2)$  do
11:     if  $P[i] == P[j]$  then
12:        $X[i][j] = X[i - 1][j - 1] + 1$ 
13:     else
14:        $\max(X[i][j - 1], X[i - 1][j])$ 
15:     end if
16:   end for
17: end for
18: return  $X[\text{len}(P1)][\text{len}(P2)]$ 
19: end procedure

```

3.5 Properties of Metric

When selecting an appropriate distance metric, it is important to consider which properties are important when calculating similarity. There are three key properties that can be considered with string metrics, namely length, sequence and position.

Length: It is apparent that considering strings of differing length is a common occurrence in process data, in particular with medical diagnosis, as it is a process of discovery and one that may need different activities based on the results of a previous one. Therefore, the algorithm needs to consider the differing length of two strings.

Sequence: The sequence in which activities occur is important and must be considered, especially when considering the previous statement that the results of one activity may change the course of the pathway.

Position: The position that the activity, and the sequence of activities, occurs within the pathway is vitally important to consider when developing an algorithm for process data.

All of these properties are considered in varying degrees in each of the eight metrics considered in the previous section. For example, *length* is evidently considered in the Jaro calculation, as it is a main variable in the formula, whereas in the Levenshtein distance *length* is indirectly considered via the upper and lower bounds for the possible values (upper bound = length of the longer string, lower bound = the difference between the lengths of the strings). Furthermore, *sequence* is evidently considered in LCS, as it is in the name, whereas *sequence* is considered in an alternative way in the Jaccard distance through the use of n-grams.

This shows that string distance metrics do possess the correct qualities to be applied to process data.

The string distances are currently underperforming when considering small differences between strings. The addition of an extra character will be considered, but

it does not make a difference what character it is or what it represents. This leads to many string comparisons resulting in the same value (as seen in Figure 3.17 and Figure 3.18). It is theorised that this will lead to poor cluster distinction and modifications to achieve uniqueness will improve upon this.

In an attempt to address this it was evident that complete uniqueness was difficult to achieve as it violated some fundamental basic relationships (such as symmetry). However, adding more distinction than is currently displayed in the distance metrics was successful.

Addressing the previous property, allowed for the ability to include some meaning to the strings. As discussed previously, there is no consideration in the metrics for *which* character has been added and what that might represent. This is likely due to the origins of the metrics typically being for spell checkers etc. where there is no need to consider this. However, in terms of process data, it can cause quite a difference when considering the addition of character 'A' or character 'B' depending on what activities they represent.

In summary, we aim to modify the Needleman-Wunsch algorithm to allow for more distinction in the values to achieve better clustering results, through the addition of context provided by experts. The process for this is explained in more detail in context in Section 3.6.

3.6 Modified Needleman-Wunsch Algorithm

This section discusses the development of the Modified Needleman-Wunsch algorithm to achieve adding distinction and context to the comparisons.

The Needleman-Wunsh metric was chosen as the base for this modification, as it had the greatest potential to modify the calculation in a meaningful way. As the intention for this modified metric was to be applied to clustering process data, three fundamental properties need to be preserved: 1) a point to itself receives a score

of 0, 2) symmetry must hold and 3) a smaller value is indicative of a closer match. This will be addressed in the discussion concerning penalty values.

Variables

The first modification considered is the idea that not all activities should be allowed to swap with each other. This is because, considering the pathway from a resource planning perspective and the interaction between multiple care centres, allowing all activities to swap could lead to very different pathways being considered similar. For example, allowing an X-Ray under primary care supervision, is very different from an MDT meeting consisting of multiple personnel from the secondary and tertiary centres, from a resource perspective.

To allow for this, a no-swap penalty value (ns) needs to be defined. Furthermore, the algorithm needs to be able to decipher which activities are allowed to swap with each other. This leads to the introduction of groups of activities, where essentially, if activities are in the same group then they are allowed to swap.

Groupings

The experts will be asked to group activities that happen at similar points in the pathway into the same group. It should be explained that the purpose of these groups is that if two patients performed different activities at the same point in their pathway, but these activities are in the same group, then they would be seen as more similar to each other than if the activities were in different group. An example of the groupings used for the case study are provided in Table 3.3.

Table 3.3: Grouping Assignments for Each Activity.

Group	Activities
0	A,B,C,O
1	D
2	E,F
3	G,H
4	I,J
5	K,L,N
6	M

This permits greater meaning to be given to the pathways, however this does not lead to the values being more distinct. This is addressed by using weightings and is discussed in the next section.

Weightings

The inclusion of weightings into the algorithm increased the complexity, and as such now becomes more difficult to calculate by hand.

We first discuss how to assign the weightings to the activities, and then follow with combining these into the algorithm.

Assume that domain experts (e.g. consultants in cancer services) are asked to rank the activities from most to least important (0 to $N - 1$, where N is the number of activities). This can be thought of as, the activity that occurs most often is seen as most important, and thus ranked 0, and those activities that are more rarely occurring should be ranked as lesser important. From these rankings, they will then be converted into weightings where the least important activity will be assigned a weight of 1, and each activity will receive an incremental addition of $1/(N - 1)$. This subsequently gives the most important activity a weight of 2.

For example, Table 3.4 shows the rankings and resulting weightings (rounded to 3 d.p.) that were applied to the case study activities.

Table 3.4: Ranking and Weighting Results for Each Activity.

Activity	Rank	Weighting
A	2	1.857
B	0	2.0
C	1	1.929
D	12	1.143
E	10	1.286
F	9	1.357
G	7	1.5
H	6	1.571
I	13	1.071
J	14	1.0
K	3	1.786
L	5	1.643
M	8	1.429
N	11	1.214
O	4	1.714

Equations

As we have now defined both the groupings and weightings, we can combine these into the algorithm. We will first methodically work through the equations, including explanations, and then provide the pseudo-code. Throughout, m , s , g and ns refer to the match, swap, gap and no-swap penalty values respectively, and w_i and w_j refer to the weights of characters in position i and j respectively.

Firstly, the match equation is as follows:

$$D = X[i - 1][j - 1] \left(m + \frac{1}{X[i - 1][j - 1] + w_i} \right) \quad (3.13)$$

The match equation had to be modified using multiplication of the m parameter, to allow the initial 0 to propagate through. This is the main element that allows for a point to itself to be 0 (as required by the fundamental properties of metrics introduced in the beginning of Section 3.6).

The inclusion of the previous matrix value ($X[i-1][j-1]$) is required in the denominator to control the magnitude, and ensure that the penalty value for a match will not exceed 1.

Furthermore, as a match is a positive event, we needed to ensure that in this case, a more important activity has a smaller impact than a less important activity. This is the reason for the inverse.

Moving on to the swap equation:

$$D = X[i - 1][j - 1] + s + abs(w_i - w_j) \quad (3.14)$$

This is more intuitive, as the modification is the addition of the absolute difference of the two weightings. This results in activities that are allowed to swap, but are ranked further apart will have a larger value than those that are ranked closer.

Now considering the no-swap equation:

$$D = X[i-1][j-1] + ns + (w_i + w_j) \quad (3.15)$$

This ensures that the no-swap value is large enough to never get chosen.

The gap equations are only slightly modified through the addition of the corresponding weighting of that direction:

$$L = X[i][j-1] + g + w_j \quad (3.16)$$

$$T = X[i-1][j] + g + w_i \quad (3.17)$$

The final modification from the Needleman-Wunsch algorithm is that now we select the minimum of D, L, T opposed to the maximum. Algorithm 4 displays the pseudocode for the modified Needleman-Wunsch algorithm.

Algorithm 4 Modified Needleman-Wunsch Algorithm.

```

1: procedure MODIFIED ▷ Initialise
   Insert a blank space at the start of each string
2:   Input  $m, g, s, ns$ 
3:   for  $i \leftarrow 0, \text{len}(P1)$  do
4:      $X[i][0] = X[i-1][0] + g + w_i$ 
5:   end for
6:   for  $j \leftarrow 0, \text{len}(P2)$  do
7:      $X[0][j] = X[0][j-1] + g + w_j$ 
8:   end for ▷ Fill Matrix

9:   for  $i \leftarrow 1, \text{len}(P1)$  do
10:    for  $j \leftarrow 1, \text{len}(P2)$  do
11:     if  $P[i] == P[j]$  then
12:        $D = X[i-1][j-1] \left( m + \frac{1}{X[i-1][j-1] + w_i} \right)$ 
13:     else if  $P[i]$  and  $P[j] \in \text{Group}$  then
14:        $D = X[i-1][j-1] + s + \text{abs}(w_i - w_j)$ 
15:     else
16:        $D = X[i-1][j-1] + ns + (w_i + w_j)$ 
17:     end if
18:      $L = X[i][j-1] + g + w_j$ 
19:      $T = X[i-1][j] + g + w_i$ 
20:      $X[i][j] = \min(D, L, T)$ 
21:   end for
22: end for
23:   return  $X[\text{len}(P1)][\text{len}(P2)]$ 
24: end procedure

```

Penalty Values

In the literature surrounding the Needleman-Wunsch algorithm, it is often discussed that the user can specify the values for the match, swap and gap penalty, however there are no guidelines surrounding these.

We developed the following equations as guidelines for choosing the penalty values, to ensure that in general the preference of, match < swap < gap < no-swap, holds (as a smaller value indicates ‘better’ and in general a match is better than a swap, which is better than a gap, which is better than a no-swap).

$$1 < g$$

$$1 < s \leq g$$

$$ns = 2g + 1$$

$$m = 1$$

For further clarification, m must be set to 1 as the match equation considers a multiplication, and otherwise the factor is not consistently less than 1 (more clarification below). Moreover, it is unnecessary for ns to be larger than $2g + 1$, as this is sufficient to consistently force gaps when a no-swap is necessary.

As a result, the smallest possible penalty values are: $m = 1$, $g = 2$, $s = 2$, $ns = 5$.

As with the standard Needleman-Wunsch algorithm, changes to the penalty values will result in different distances calculated, which will propagate through to the clustering. Advice to the user when selecting the values of s and g in particular, is to select values with a larger difference between s and g to ensure a more distinct separation of these two actions.

Example

Figure 3.12 calculates the modified Needleman-Wunsch distance between the two pathways ‘ABKOGNCH’ and ‘ABC’, using the values $m = 1$, $g = 2$, $s = 2$ and $ns = 5$, with the groupings and weightings from Table 3.3 and Table 3.4 respectively. Figure 3.13 shows the resulting alignment from following the traceback.

	-	A	B	K	O	G	N	C	H
-	0.000	3.857	7.857	11.643	15.357	18.857	22.071	26.000	29.571
A	3.857	0.000 7.714 7.714 0.000	6.000 11.714 4.000 4.000	16.500 15.500 7.786 7.786	13.786 19.214 11.500 11.500	23.714 22.714 15.000 15.000	26.929 25.929 18.214 18.214	24.143 29.857 22.143 22.143	34.429 33.429 25.714 25.714
B	7.857	6.000 4.000 11.714 4.000	0.000 8.000 8.000 0.000	12.786 11.786 3.786 3.786	10.071 15.500 7.500 7.500	20.000 19.000 11.000 11.000	23.214 22.214 14.214 14.214	20.286 26.143 18.143 18.143	30.714 29.714 21.714 21.714
C	11.786	9.929 7.929 15.643 7.929	6.071 3.929 11.929 3.929	8.714 7.714 7.714 7.714	6.000 11.429 11.429 6.000	15.929 14.929 9.500 9.500	19.143 18.143 12.714 12.714	15.095 22.071 16.643 15.095	26.643 25.643 18.666 18.666

Figure 3.12: Example of Modified Needleman-Wunsch Algorithm Dynamic Programming.

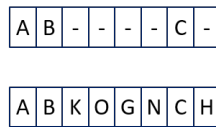


Figure 3.13: Example of Modified Needleman-Wunsch Algorithm Traceback.

Consider that intuitively it should always be better to take a swap over a gap. However, looking at the interaction between ‘B’ and ‘O’, it can be seen that this is not the case, as the value from the gap is smaller than that of the allowed swap. At first glance, this may seem incorrect, until further inspection when it is clear that this is necessary to allow the alignment of ‘B’ with itself two steps later.

This demonstrates the intelligence of the algorithm, and the consideration for the string as a whole during traceback.

Features

The modified algorithm allows for many features to be considered, which are:

1. Point to itself is 0.
2. The distance score for the string is 0 until the first non-match (similar to the common prefix idea in the Jaro-Winkler distance).
3. Distances between two pathways are commutative.
4. Matches between higher importance activities produce a smaller distance.
5. A match earlier in the string will result in a smaller value than that appearing later.

6. Gaps with higher importance activities result in a larger value than that of lower importance.
7. Swaps of activities that are closer in terms of rankings will produce a smaller value.

Figure 3.16 displays all the features described above for Sample 2 (explained below) using penalty values $m = 1, g = 2, s = 2, ns = 5$.

To add commentary to Figure 3.16, feature 1 is displayed along the diagonal of the matrix, and feature 3 (commutativity) is displayed, and thus one can ignore the bottom diagonal of the matrix, and just examine the top diagonal (highlighted in green).

Feature 2 can be confirmed by matrix locations (1,2), (1,3) and (1,4), as the value corresponds to g with the addition of the weight for the additional character as displayed in Table 3.4. These three values also confirm feature 6.

Features, 4 and 7, are displayed amongst Figure 3.16, but can easily be checked manually by combining the weightings in Table 3.4 with the equations for the match and swap (Equation 3.13 and 3.14) respectively.

Feature 5 is the most complex and a by-product of feature 1. This feature arises due to the match penalty calculation being a factor of the previous value (as previously discussed in the context of Equation 3.13). This feature can be seen in matrix locations (1,2) compared to (1,5), where (1,2) is smaller than (1,5) as the match of 'C' happens earlier in (1,2) than in (1,5). To further display this feature, consider the string 'C' compared with the following three strings: 1) 'DC', 2) 'HC', and 3) 'DHC'.

Figure 3.14 shows the full calculation matrix of each of the three scenarios. If we calculate the impact of matching 'C' in each scenario by observing the difference between the two values (indicated by the diagonal arrow in Figure 3.14), as follows:

$$\begin{aligned}
 1) & 3.763 - 3.143 = 0.62 \\
 2) & 4.221 - 3.571 = 0.65 \\
 3) & 7.491 - 6.714 = 0.777
 \end{aligned}
 \tag{3.18}$$

Equation 3.18 shows that the penalty for matching ‘C’ is different in all three scenarios. Simplified, if the previous value is larger then the effect of matching ‘C’ is also larger. Hence, the later a match appears in the string, the larger the value.

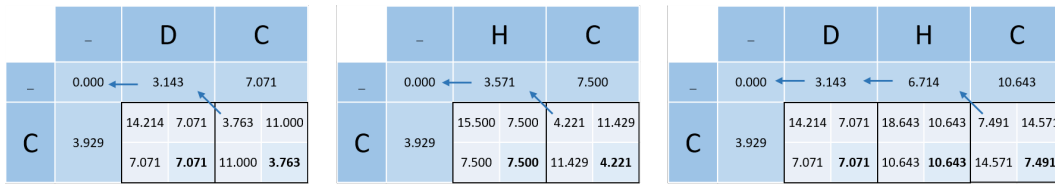


Figure 3.14: Example of Feature Five.

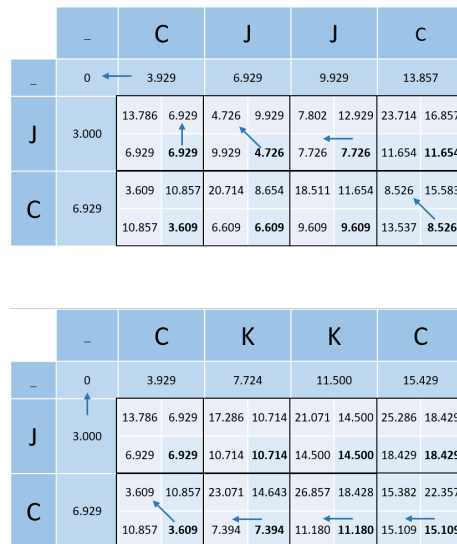


Figure 3.15: Comparing Prioritising in Traceback.

Consider feature 5 when a character appears more than once in a string. The modified algorithm will prioritise a match at the earliest occurrence of the character, unless there are characters in between that allow for a better alignment. For example in Figure 3.15, where when ‘JJ’ appears between the two ‘C’s, the later ‘C’ is aligned, compared to when the ‘KK’ appears in the middle, then the first ‘C’ is aligned.

In conclusion, the modified Needleman-Wunsch algorithm does produce a more specific value for distance, considering length, position, and sequence, whilst also considering the weightings and groupings of the activities.

	ABC	ABCK	ABCO	ABCL	ABKC		DIJ	DJK	DIJO	DIJL	DIKJ		ABKCOEF	ABKOCFE	ABKOCF	ABKCOEF	ABKCOEF
ABC	0	3.786	3.714	3.643	4.448		21.000	24.786	19.286	24.643	24.786		14.938	14.938	14.938	14.929	14.805
ABCK	3.786	0	7.500	2.143	8.234		24.786	21.922	23.071	23.143	21.910		18.724	13.339	18.724	18.714	18.591
ABCO	3.714	7.500	0	7.357	8.163		24.714	28.500	21.925	28.357	28.500		14.857	14.857	14.857	12.035	11.813
ABCL	3.643	2.143	7.357	0	8.091		24.643	23.143	22.929	21.927	23.143		18.581	14.929	18.581	18.571	18.448
ABKC	4.448	8.234	8.163	8.091	0		24.786	21.905	23.071	23.143	21.887		11.015	11.727	11.015	10.987	10.357
DIJ	21.000	24.786	24.714	24.643	24.786		0	3.786	3.714	3.643	4.577		35.143	35.143	35.143	35.143	35.143
DIJK	24.786	21.922	28.500	23.143	21.905		3.786	0	7.500	2.143	6.627		32.262	32.278	32.262	32.262	32.262
DIJO	19.286	23.071	21.925	22.929	23.071		3.714	7.500	0	7.357	8.291		32.353	32.337	32.353	32.371	32.364
DIJL	24.643	23.143	28.357	21.927	23.143		3.643	2.143	7.357	0	8.143		33.500	33.500	33.500	33.500	33.500
DIKJ	24.786	21.910	28.500	23.143	21.887		4.577	6.627	8.291	8.143	0		32.245	32.266	32.245	32.245	32.245
ABKCOEF	14.938	18.724	14.857	18.581	11.015		35.143	32.262	32.353	33.500	32.245		0	10.664	4.143	13.055	11.890
ABKOCFE	14.938	13.339	14.857	14.929	11.727		35.143	32.278	32.337	33.500	32.266		10.664	0	13.055	4.143	16.394
ABKOCF	14.938	18.724	14.857	18.581	11.015		35.143	32.262	32.353	33.500	32.245		4.143	13.055	0	10.663	11.850
ABOKCFE	14.938	13.339	14.857	14.929	11.727		35.143	32.278	32.337	33.500	32.266		13.055	4.143	10.663	0	16.374
ABKECOF	14.929	18.714	12.035	18.571	10.987		35.143	32.262	32.371	33.500	32.245		11.890	16.394	11.850	16.374	0
ABKCOEF	14.805	18.591	11.813	18.448	10.357		35.143	32.262	32.364	33.500	32.245		5.997	10.548	8.571	12.942	8.750
																	0

Figure 3.16: Modified Needleman-Wunsch Distance Matrix for Sample 2.

3.7 Case Studies

This case study applies the eight previously discussed metrics and the modified algorithm to two small samples and the full case study dataframe. These samples are very basic to allow the reader to closely examine the intricate differences that appear due to the inclusion of the weighting and rankings. Furthermore, sample 1 and sample 2 are easily assigned to two and three groups respectively, to display that the obvious solution is found in a simple example, and to provide the reader with confidence when applying this to more complex data. Although these samples are artificially constructed, they reflect the small differences between strings seen in practice.

Sample 1 consists of 10 pathways: ‘ABC’, ‘ABCK’, ‘ABCL’, ‘ABCO’, ‘ABKC’, ‘DIJ’, ‘DIJK’, ‘DIJL’, ‘DIJO’, and ‘DIKJ’. These were chosen as ‘A’, ‘B’, ‘C’ and ‘D’, ‘I’, ‘J’ are the highest and lowest ranked activities respectively.

Sample 2 consists of 16 pathways, the same 10 as in sample 1, plus a further six which display the complexity of allowed swaps between slight differences within the pathway. These are:

‘ABKOCEF’, ‘ABOKCEF’, ‘ABKOCFE’, ‘ABOKCFE’, ‘ABKECOF’, ‘ABKCOEF’.

Two examples of the modification are included in the analysis using penalty values $g = 2, s = 2, ns = 5$ and $g = 9, s = 2, ns = 19$, which will be referred to as MNW_1225 and MNW_19219 respectively.

The analysis for the two samples is as follows: Firstly, the distances between all the points are calculated using the ten previously discussed metrics, and then plotted to demonstrate how the modified algorithm allows for more separation in the data. Secondly, the k-medoids clustering is run for $k = [2, 8]$, where the use of the silhouette scores both confirms point one and displays that the modified algorithm outperforms most of the other metrics. The findings are displayed in a table, which contains the results for $k = 2$ and then the best performing k (if $k = 2$ was best,

then the second best is displayed), which includes the number of iterations.

The following Python libraries were used: `textdistance` [274] was used for calculations of the eight other distance metrics, `pyclustering` [214] was used for the k-medoids clustering and `scikit-learn` was used for the silhouette score [246].

3.7.1 Sample 1: 10 Pathways

Figure 3.17 displays a comparison of the distances between the pathways in sample 1 for each of the eight measures discussed in Section 3.4 and the two examples of the modified algorithm (MNW_1225 and MNW_19219).

To aid understanding of Figure 3.17, firstly the distance from each point to itself is 0, and therefore the colour of the dot at $x = 0$ for each pathway on y is the colour that represents that pathway e.g. pathway ‘DIJ’ is represented by the red dot. Furthermore, all pathways beginning with ‘A’ are from the blue colour pallet, and those beginning with ‘D’ are from the red colour pallet.

The y-axis displays the pathway which all others are being compared to and the x-axis displays the distance from that pathway. For example, in the top left graph considering the Levenshtein distance, the distance from ‘ABC’ (light blue) to ‘ABKC’ (dark green) is 1.

In all eight of these graphs in Figure 3.17, if you split the graph horizontally between ‘ABKC’ and ‘DIJ’, and overlaid the two halves, you can see that the distances are exactly the same, and reflects the lack of distinction. There is also little separation between the blue and red groups, with the exception of the Jaro and Jaro-Winkler graphs, where this is more clear.

Now considering the bottom two graphs in Figure 3.17, which display the modified algorithm (penalty values $g = 2, s = 2$ and $ns = 5$ on the left and $g = 9s = 2$ and $ns = 19$ on the right). It can clearly be seen that this algorithm allows for more distinction and greater separation between the colour groups, as desired.

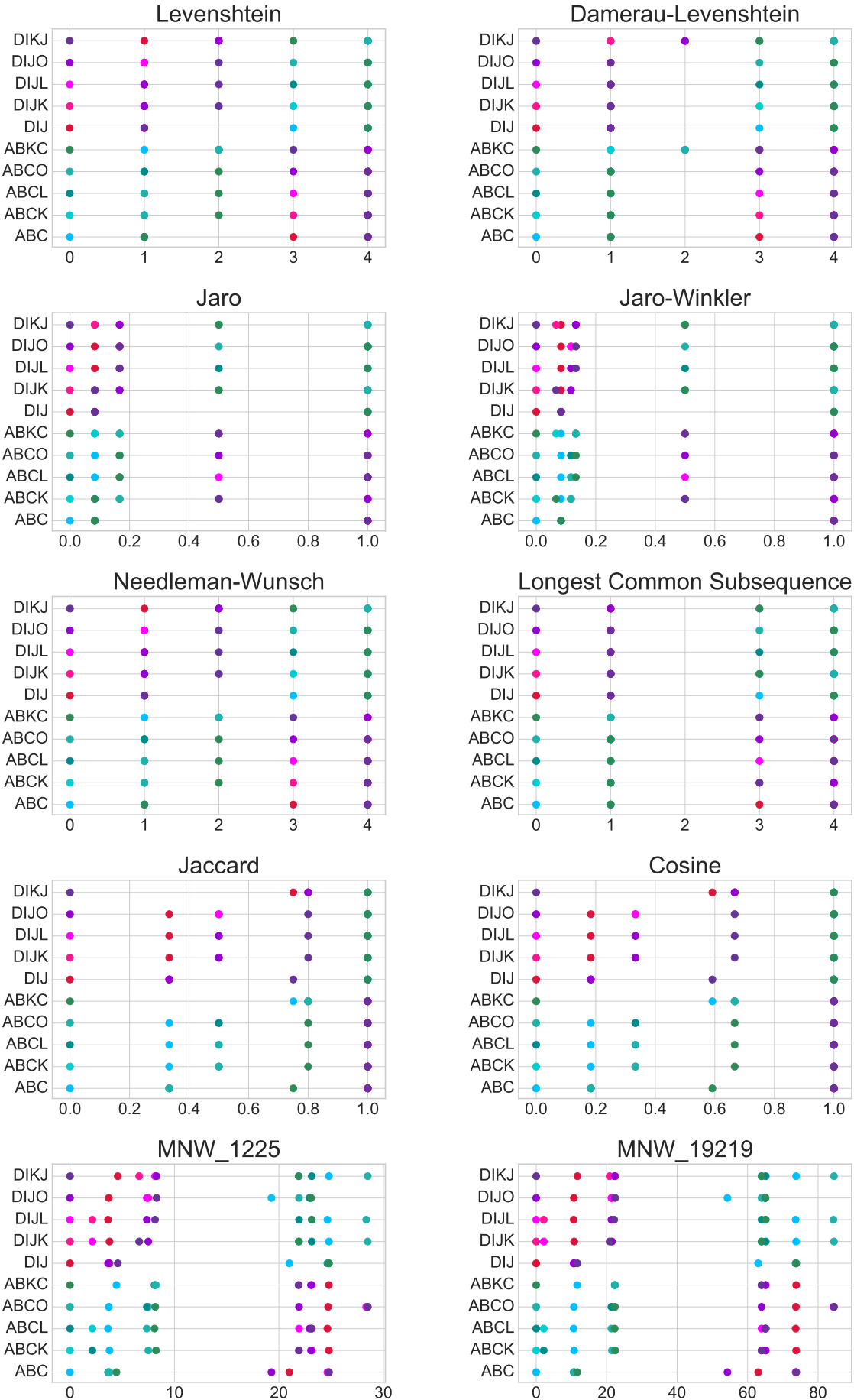


Figure 3.17: Comparison of the Ten Metrics Applied to Sample 1.

To confirm that this is reflected in the clustering, k-medoids clustering was performed for all ten metrics, the results for which are displayed in Table 3.5. The initial medoids were chosen as 0: ‘ABC’ and 5: ‘DIJ’. It is expected that the clustering algorithm should keep ‘ABC’ and ‘DIJ’ as the medoids.

Table 3.5: Clustering of Sample 1, for All Ten Distances.

Name	Medoids	Number Cluster	per	Silhouette Score
Levenshtein	0, 5	5, 5		0.65789
Damerau-Levenshtein	0, 5	5, 5		0.70614
Jaro	0, 5	5, 5		0.85602
Jaro-Winkler	0, 5	5, 5		0.88333
Needleman-Wunsch	0, 5	5, 5		0.65789
Jaccard	0, 5	5, 5		0.43500
Cosine	0, 5	5, 5		0.58577
LCS	0, 5	5, 5		0.73099
MNW_1225	0, 5	5, 5		0.76128
MNW_19219	0, 5	5, 5		0.77464

Table 3.5 displays the expected results, with the only measures that surpass the modified Needleman-Wunsch in silhouette score is the Jaro and Jaro-Winkler.

3.7.2 Sample 2: 16 Pathways

Similarly to subsection 3.7.1, Figure 3.18 displays a comparison of the distances between the pathways in sample 2 for each of the eight measures discussed in Section 3.4 and the two examples of the modified algorithm (MNW_1225 and MNW_19219).

In this sample, it is logical to assume that three clusters would be appropriate, the same two as in sample 1 and a further one containing the extra six pathways. Therefore Figure 3.18 should be examined for the appearance of three distinct groups.

This is actually not as clear cut as it was with sample 1 (in relation to two groups). In the majority of the metrics, it is difficult to find the clear groups one is expecting (one group of red, one group of blue and another of yellow). Again the distinction is more clear in the modified algorithm, especially with the penalty values $g = 9$, $s = 2$ and $ns = 19$ (as previously stated). This further confirms that the modified algorithm allows for better distinction between pathways.

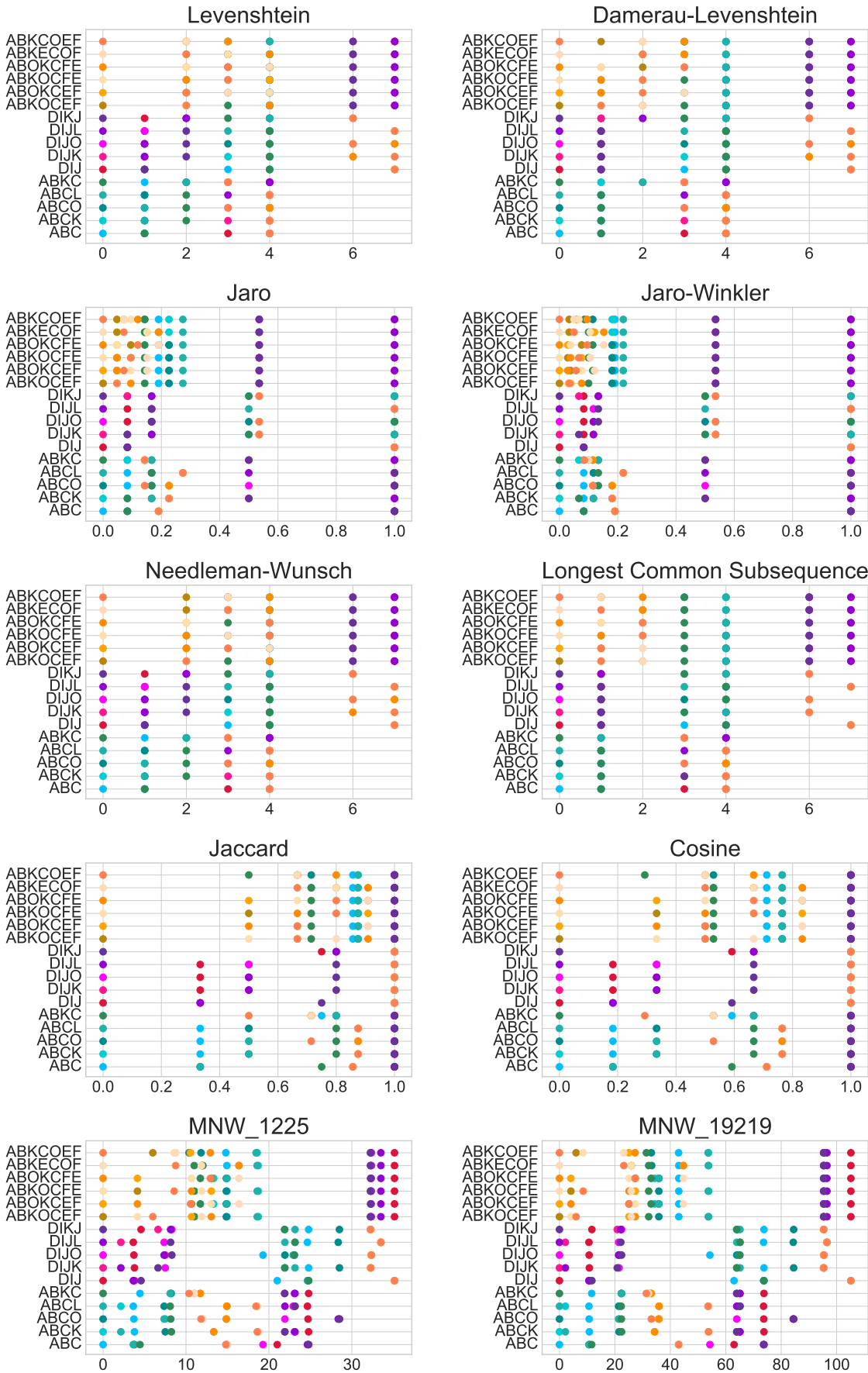


Figure 3.18: Comparison of the Ten Metrics Applied to Sample 2.

To confirm if this is reflected in the clustering, the same analysis was run as that described for sample 1, where the initial medoids were chosen as 0: ‘ABC’, 5: ‘DIJ’ and 10: ‘ABKOCEF’, and for $k = [2, 3]$. It is expected that the clustering algorithm should keep the same medoids, and that three clusters would be chosen.

Table 3.6 confirms that the modified algorithm produces silhouette scores similar to the other metrics, whilst also selecting the expected medoids, which is not the case with some of the other metrics.

It was expected that three clusters should be chosen, however, examining the silhouette scores it appears that in most cases the score for $k = 2$ is closer to 1 than in $k=3$, suggesting that two clusters is better. This indicates that possibly the silhouette score is not the most appropriate measure to use, and care is needed when selecting the appropriate number of clusters.

In conclusion both samples display that the modified algorithm does enhance the differences between strings based on user specific characteristics, and performs equally well, if not better, than the currently used metrics.

Table 3.6: Clustering of Sample 2, for All Ten Distances.

Name	Medoids $k = 2$	Number per Clus- ter $k = 2$	Silhouette Score $k = 2$	Medoids $k = 3$	Number per Clus- ter $k = 3$	Silhouette Score $k = 3$
Levenshtein	4, 5	11, 5	0.51433	0, 5, 10	6, 5, 5	0.45234
Damerau- Levenshtein	4, 5	11, 5	0.54800	0, 5, 10	5, 5, 6	0.62230
Jaro	5, 10	5, 11	0.80971	0, 5, 10	5, 5, 6	0.58120
Jaro- Winkler	4, 5	11, 5	0.84600	0, 5, 10	5, 5, 6	0.60252
Needleman- Wunsch	4, 5	11, 5	0.50676	0, 5, 10	6, 5, 5	0.43148
Jaccard	0, 5	11, 5	0.30516	0, 4, 5	4, 7, 5	0.32543
Cosine	0, 5	11, 5	0.44807	0, 4, 5	4, 7, 5	0.43025
LCS	4, 5	11, 5	0.56353	0, 5, 10	5, 5, 6	0.67356
MNW_1225	4, 5	11, 5	0.64700	0, 5, 10	5, 5, 6	0.53059
MNW_19219	4, 5	11, 5	0.67874	0, 5, 10	5, 5, 6	0.59195

3.7.3 Full Data

This section applies the eight measures discussed in Section 3.4 and the two examples of the modified algorithm (MNW_1225 and MNW_19219) to the full dataframe which was discussed in Section 3.3. As a recap, there are 2,350 patients and 1,019 unique pathways considering the 15 activities. We have applied k-medoids clustering to the dataframe, considering values of $k = [2, 8]$ and initial medoids as $[0, 1, 2, 3, 4, 5, 6, 7]$.

Table 3.7 shows the results for $k=2$ and Table 3.8 for the (next) best value of k (in terms of silhouette score). Both tables also include the medoids that were chosen and the number of pathways assigned to each of those cluster medoids. The run time for each distance matrix was under 10 minutes, where the modified Needleman-Wunsch algorithm performed within the range of the other metrics.

Both Table 3.7 and 3.8 shows that the silhouette scores for all 10 measure are quite poor. However, the silhouette score for the Needleman-Wunsch modification, with both sets of penalty values, is on par with the other metrics for $k=2$ (with the exception of LCS), and surpass most of the other measure, with the exception of the Jaro metrics when considering the second best value for k . This shows that for a full dataset, the modification performs equally as well, if not better, than the frequently used metrics when considering the silhouette score.

Furthermore, the metrics as a whole do not come to a consensus on a solution for the clustering as each of the metrics produce different results when considering the medoids selected and the number of pathways assigned to each cluster. Even when the same medoids are selected, the number of pathways assigned to those medoids clusters are not the same. This confirms that careful consideration is needed when selecting the distance metric, and what differences are to be highlighted.

Table 3.7: Results of Full Data Clustering for $k = 2$.

Name	Iter	Medoids	Pathways per Cluster	Score
Levenshtein	3	KAOBC, AKBMCEGFH	663, 356	0.15604
Damerau-Levenshtein	2	KAOB CD, AKBMCEGFH	676, 343	0.17549
Jaro	3	KAOBLCD, AKOBMCEGFH	409, 610	0.18343
Jaro-Winkler	3	KAOB CD, AKOBMCEGFH	445, 574	0.17542
Needleman-Wunsch	2	AOBC, AOBCEGFH	727, 292	0.16743
Jaccard	2	KAOB NL CGH, KAOBMCEGFH	650, 369	0.04297
Cosine*	2	KAOB NL CGDH, KAOBMCEFGH	649, 369	0.06854
LCS	2	KAOB CD, KAOBCEGFH	510, 509	0.24305
MNW_1225	2	KABC, AOBCEGFH	715, 304	0.14303
MNW_19219	2	AOBC, AKOBCEGFH	676, 343	0.17976

*For cosine, the pathway consisting of just activity B had to be removed, as it caused division by 0.

Table 3.8: Results of Full Data Clustering for Best k (excluding $k = 2$).

Name	k	Iter	Medoids	Pathways per Cluster	Score
Levenshtein	3	4	KAOBC, AKBMCEGFH, ABCO	541, 348, 130	0.06964
Damerau-Levenshtein	3	4	KAOB CD, AKBMCEGFH, ABKOC	519, 333, 167	0.09724
Jaro	3	3	KAOBLCD, KAOBMCEGFH, ABKOC	315, 503, 201	0.16252
Jaro-Winkler	3	3	KAOBLCD, KAOBMCEGFH, ABKOC	308, 487, 224	0.16254
Needleman-Wunsch	3	2	AOBC, AOBCEGFH, ABCO	582, 279, 158	0.06689
Jaccard	7	3	KAOB NLC, KAOBMCEGFH, KAOBC, AKOBNC, KABNCOEF, AOKBMC, BKAOCGH	137, 229, 117, 194, 84, 113, 145	0.05322
Cosine*	7	4	KANOMBCEFD, KAOBMCEFGH, ABC, AKOBC, KABMCO, AOKBC, BKAOC EGFH	172, 219, 45, 207, 122, 89, 164	0.08812
LCS	3	3	KAOB CD, KAOBCEGFH, ABKC	408, 509, 102	0.14132
MNW_1225	3	4	KAOBC, AOBCEGFH, AKBC	384, 199, 436	0.13354
MNW_19219	3	4	AKOBC, AOKBCEGFH, KAOBC	403, 229, 387	0.14860

*For cosine, the pathway consisting of just activity B had to be removed, as it caused division by 0.

3.8 Sensitivity Analysis

This section explores the sensitivity in regards to each area where the users can specify variables, namely, initial medoids, penalty values, rankings and groupings.

This section was additional to the source paper [16]

3.8.1 Initial Medoids

It is well documented that the k-means clustering algorithm is highly sensitive to the initial starting positions [50], and as k-medoids is very closely related to k-means, it is assumed to also be sensitive to the initial starting positions. The `pyclustering` [214] k-medoids class allows the user to specify the initial start points, and thus allows for the user to choose any method of initial medoids selection. There are a variety of initialisation methods available to k-means clustering, for example Peña and Larrañaga (1999) [227] compare random, Forgy, MacQueen and Kaufman methods. Furthermore, the `scikit-learn` [246] k-medoids algorithm offers random, heuristic and k-medoids++ as selection methods (where heuristic is default, and selects the points with the smallest sum distance).

Although this sensitivity is well documented for k-means and there are methods available for k-medoids, the comparison of such methods for k-medoids clustering are not as obviously documented. Given the initial selection methods provided by `scikit-learn` [246], we will also offer the user the following options: random, least distance and most occurred. Most occurred was not an option discussed above, however using the pathways that appeared most often in the data is a logical option to offer.

To allow for analysis, Table 3.9 and 3.10 display the clustering results for the least distance and most occurred initial medoids, for $k = [2, 3]$. Furthermore, Table 3.11 and Table 3.12 displays the results of using 10 different random initial medoids for $k = [2, 3]$ respectively. Here the silhouette score is referred to as ‘Score’.

Table 3.9: Least Distance Initial Starting Medoids.

k	Medoids	Frequency	Score
2	AKBC, KAIBC	561, 458	0.147126
3	AKBC, KABC, KAIBC	505, 183, 331	0.080843

Table 3.10: Most Occurred Initial Starting Medoids.

k	Medoids	Frequency	Score
2	AKBC, KAIBC	561, 458	0.147126
3	KABC, KAIBLC, AKBC	252, 272, 495	0.094811

Table 3.11: Random Initial Starting Medoids for Two Clusters.

Initial Index	Initial Pathways	Medoids	Frequency	Score
286, 961	AKOBNCEF, KLAMCONBD	AKBC, KAIBC	561, 458	0.147126
80, 312	ACOKBML, AKOLBDC	KAIBC, AKBC	458, 561	0.147126
1002, 225	LKABD, AKMCOBNEF	KAIBC, AKBC	458, 561	0.147126
914, 287	KBAONC, AKOBNCELF	KAIBC, AKBC	458, 561	0.147126
387, 902	AOCEGHBMF, KAONLBDC	AKOBCEGFH, KAIBC	320, 699	0.141554
580, 720	KABLOC, KANOMBCEFD	AIBC, AKOBCEF	653, 366	0.157202
419, 908	AOKBNCLGH, KBAOC	AKBCGH, KAIBC	369, 650	0.121541
82, 47	AEGHKFCOB, ABKOLCGH	AKBC, KAIBC	561, 458	0.147126
112, 452	AKBMCOEF, AOKMCBLEGFH	KAIBC, AKBCGH	650, 369	0.121541
938, 822	KBMCAOGH, KAOLBMCEDF	AKBCGH, KAIBC	369, 650	0.121541

Table 3.12: Random Initial Starting Medoids for Three Clusters.

Initial Index	Initial Pathways*	Medoids	Frequency	Score
286, 961, 981	+ KNBCAGH	AKIBC, KAIBC, KABCGH	431, 274, 314	0.123287
80, 312, 980	+ KMOABLCEGHF	AIBLC, AKBC, KAIBCEGFH	243, 479, 297	0.080809
1002, 225, 161	+ AKBOCEF	KAIBC, AOKBC, AKIBC	456, 226, 337	0.138924
914, 287, 713	+ KANLBCEGHF	KAIBC, AKIBC, KAOLBCEGFH	368, 469, 182	0.121244
387, 902, 335	+ AKOLMCEGBDFH	AIBCEF, KAIBC, AKBCGH	186, 517, 316	0.098833
580, 720, 450	+ AOKMBEGCFH	KABC, KAIBCEF, AKBMCEGFH	568, 229, 222	0.137245
419, 908, 414	+ AOKBMLGCH	AKBCGH, KAIBC, AKBC	256, 419, 344	0.133141
82, 47, 904	+ KAONLCDB	AKBC, AKBCGH, KAIBC	344, 256, 419	0.133141
112, 452, 51	+ ABKOMCGH	AKBC, AOKBCEGFH, KAIBC	415, 204, 400	0.130351
938, 822, 456	+ AOLBCEF	KABCGH, KAIBC, AKBC	270, 287, 462	0.114578

* The first two initial pathways are the same as those for $k=2$.

It is clear that across Tables 3.9, 3.10 and 3.11, there appears to be a most frequently found solution for $k = 2$ of medoids ‘AKBC’ and ‘KAIBC’. At first glance this might appear to be an ‘optimal’ solution, however in table 3.11 initial medoids of [580, 720] produce a higher silhouette score. Furthermore, when considering $k = 3$, the results in Table 3.12 display more differences, and in fact the initial medoids

[1002, 225, 161], produce the best silhouette score for $k = 3$ with medoids solution of ‘KAOBC’, ‘AOKBC’, ‘AKOBC’.

This demonstrates that the results are sensitive to the initial starting points. Furthermore, more precise selection methods do not necessarily produce better results in accordance to the silhouette score.

Although the silhouette score is a good indication of the theoretical best solution, it is still open to expert interpretation of what clustering solution they select. Therefore, care needs to be taken when selecting initial medoids.

3.8.2 Penalty Values

The selection of penalty values for gap, swap and no-swap (g , s and ns) variables also require sensitivity analysis. As previously discussed in subsection 3.6 the user is able to select various penalty values for the match, swap and gap penalties but there is no guidance on this for the standard Needlaman-Wunsch algorithm. After discussing in subsection 3.6 how the user should select the gap, swap and no-swap values, it is necessary to compare the effects. As the penalty values are guided by the value for g , we test values for $g = [2, 9]$ and all values of s available. The results for $k = [2, 3]$ are shows in Table 3.13 and 3.14 respectively.

Both Table 3.13 and Table 3.14 show that varying the penalty values does appear to produce different results. In Table 3.13 the score is *always* higher than in the basic case of $g = 2$, $s = 2$ and $ns = 5$. Therefore, it could be suggested that running multiple values and finding the solution most robust to change may be a good indicator of which solution is best. Furthermore, in both tables (3.13 and 3.14) the score is often larger when the gap between g and s is larger, as suggested in subsection 3.6.

In conclusion, the results are sensitive to penalty values selected. However, this is not a negative, as allowing this level of control to the user is a further method of allowing control and adding context to the comparisons.

g	s	ns	Medoids	Frequency	Score
2	2	5	KABC, AOBCEGFH	715, 304	0.143026
3	2	7	AOBC, AOBCEGFH	808, 211	0.172362
3	3	7	KABC, AOBCEGFH	700, 319	0.147568
4	2	9	KAOBC, AKOBCEGFH	685, 334	0.155974
4	3	9	KABC, AOBCEGFH	695, 324	0.153521
4	4	9	KABC, AOBCEGFH	702, 317	0.148849
5	2	11	KAOBC, AKOBCEGFH	681, 338	0.160720
5	3	11	KAOBC, AKOBCEGFH	680, 339	0.154746
5	4	11	KABC, AOBCEGFH	695, 324	0.153706
5	5	11	KABC, AOBCEGFH	701, 318	0.149900
6	2	13	AOBC, AKOBCEGFH	677, 342	0.172716
6	3	13	KAOBC, AKOBCEGFH	674, 345	0.160022
6	4	13	KAOBC, AKOBCEGFH	679, 340	0.154380
6	5	13	KABC, AOBCEGFH	695, 324	0.153953
6	6	13	KABC, AOBCEGFH	697, 322	0.151785
7	2	15	AOBC, AKOBCEGFH	676, 343	0.175672
7	3	15	KAOBC, AKOBCEGFH	674, 345	0.162501
7	4	15	KAOBC, AKOBCEGFH	677, 342	0.157869
7	5	15	KABC, AKOBCEGFH	674, 345	0.162233
7	6	15	KABC, AOBCEGFH	695, 324	0.155380
7	7	15	KABC, AOBCEGFH	692, 327	0.152378
8	2	17	AOBC, AKOBCEGFH	676, 343	0.177912
8	3	17	AOBC, AOKBCEGFH	670, 349	0.169386
8	4	17	KAOBC, AKOBCEGFH	673, 346	0.161693
8	5	17	KAOBC, AKOBCEGFH	676, 343	0.157081
8	6	17	KABC, AKOBCEGFH	674, 345	0.162274
8	7	17	KABC, AOBCEGFH	695, 324	0.155611
8	8	17	KABC, AKOBCEGFH	675, 344	0.157572
9	2	19	AOBC, AKOBCEGFH	676, 343	0.179763
9	3	19	AOBC, AOKBCEGFH	670, 349	0.171603
9	4	19	KAOBC, AKOBCEGFH	673, 346	0.163615
9	5	19	KAOBC, AKOBCEGFH	670, 349	0.161615
9	6	19	KABC, AKOBCEGFH	668, 351	0.165646
9	7	19	KABC, AKOBCEGFH	674, 345	0.162322
9	8	19	KABC, AOBCEGFH	695, 324	0.155923
9	9	19	KABC, AKOBCEGFH	674, 345	0.158377

Table 3.13: Sensitivity Analysis of Penalty Value Selection for Two Clusters.

g	s	ns	Medoids	Frequency	Score
2	2	5	KAOBC, AOBCEGFH, AKBC	384, 199, 436	0.133538
3	2	7	AOBC, AOBCEGFH, ABOKC	675, 210, 134	0.062149
3	3	7	KABC, AOBCEGFH, KABCN	616, 313, 90	0.078795
4	2	9	AKOBC, AKOBCEGFH, KAOLBC	440, 194, 385	0.127571
4	3	9	KAOBC, AKOBCEGFH, AKBC	387, 251, 381	0.128682
4	4	9	KABC, AOBCEGFH, KABCN	616, 314, 89	0.081155
5	2	11	AKOBC, AOKBCEGFH, KAOLBC	411, 244, 364	0.126153
5	3	11	KABC, AOBCEGFH, AOBLC	445, 261, 313	0.091549
5	4	11	KABC, AOBCEGFH, KABCN	612, 317, 90	0.084784
5	5	11	KABC, AOBCEGFH, KABCN	612, 315, 92	0.083108
6	2	13	AKOBC, AOKBCEGFH, KAOLBC	401, 229, 389	0.145826
6	3	13	KAOBC, AKOBCEGFH, AKBC	395, 265, 359	0.130803
6	4	13	KAOBC, AKOBCEGFH, AKBC	387, 254, 378	0.129436
6	5	13	KABC, AOBCEGFH, KABCN	612, 319, 88	0.084809
6	6	13	KABC, AOBCEGFH, KABCN	608, 319, 92	0.085237
7	2	15	AKOBC, AOKBCEGFH, KAOLBC	402, 229, 388	0.147083
7	3	15	AKOBC, AOKBCEGFH, KAOLBC	390, 243, 386	0.142946
7	4	15	KAOBC, AKOBCEGFH, AOKBC	416, 283, 320	0.120786
7	5	15	KABC, AKOBCEGFH, KABCN	591, 342, 86	0.090407
7	6	15	KABC, AOBCEGFH, KABCN	606, 321, 92	0.087531
7	7	15	KABC, AOBCEGFH, KABCN	605, 323, 91	0.085199
8	2	17	AKOBC, AOKBCEGFH, KAOLBC	402, 229, 388	0.147742
8	3	17	AKOBC, AOKBCEGFH, KAOLBC	390, 243, 386	0.144037
8	4	17	KAOBC, AKOBCEGFH, ABKOC	366, 295, 358	0.125987
8	5	17	KAOBC, AKOBCEGFH, AKBC	384, 257, 378	0.129628
8	6	17	KABC, AKOBCEGFH, KABCN	591, 343, 85	0.089875
8	7	17	KABC, AKOBCEGFH, KABCN	589, 342, 88	0.089923
8	8	17	KABC, AOBCEGFH, KABCN	603, 325, 91	0.086392
9	2	19	AKOBC, AOKBCEGFH, KAOLBC	403, 229, 387	0.148601
9	3	19	AKOBC, AOKBCEGFH, KAOLBC	402, 270, 347	0.128966
9	4	19	KAOBC, AKOBCEGFH, ABKOC	366, 295, 358	0.127205
9	5	19	KAOBC, AKOBCEGFH, AOKBC	410, 292, 317	0.121076
9	6	19	KABC, AKOBCEGFH, KABCN	585, 349, 85	0.092270
9	7	19	KABC, AKOBCEGFH, KABCN	587, 343, 89	0.092171
9	8	19	KABC, AKOBCEGFH, KABCN	589, 343, 87	0.089933
9	9	19	KABC, AOBCEGFH, KABCN	603, 325, 91	0.087039

Table 3.14: Sensitivity Analysis of Penalty Value Selection for Three Clusters.

3.8.3 Rankings and Groupings

As the user is encouraged to select rankings and groupings for activities, it will be beneficial if there is a degree of sensitivity with various configurations. To investigate the effects of various rankings, the groupings were fixed as in Table 3.3 and assigned random rank order, the results for $k = [2, 3]$ can be seen in Table 3.15.

To investigate groupings, the rankings were fixed as in Table 3.4 and the groupings were randomly selected, by assigning each activity a value between 2 and 7, in an effort to minimise the number of groups with only one activity. As above, the results for $k = [2, 3]$ can be seen in Table 3.16.

Both Table 3.15 and Table 3.16 show that the values assigned to the rankings and groupings possesses the same amount of variation in results as the penalty value choices. Again, this displays that the results are sensitive towards the rankings and groupings selected.

To help guide the user on selecting these values in Sim.Pro.Flow, there is an option presented to the user of ‘default values’, where the activities are ranked in order of frequency (where a smaller rank number indicates higher frequency). If the user wishes to obtain the objectively ‘best’ clustering i.e. good cluster definition, then using the silhouette score as a guide would be beneficial. Furthermore, the user could examine the chosen medoids and cluster assignment frequency to inform the choice of values.

It should be noted that these modifications were intended to increase user interaction, however, it could be an area of further work to explore presenting the user with a few options for values.

Rankings	k	Medoids	Frequency	Score	k	Medoids	Frequency	Score
[7, 8, 2, 0, 4, 9, 10, 1, 14, 12, 3, 6, 11, 5, 13]	2	KABC, AKOBCEGFH	690, 329	0.150343	3	KABC, AKOBCEGFH, KABCN	607, 322, 90	0.082529
[3, 10, 0, 5, 12, 13, 8, 1, 7, 4, 11, 2, 6, 14, 9]	2	KABC, AOBCEGFH	695, 324	0.145901	3	KAOBC, AOBCEGFH, AKBC	394, 216, 409	0.117079
[11, 5, 7, 13, 9, 4, 8, 10, 14, 3, 2, 6, 12, 0, 1]	2	KABC, AKOBCEGFH	681, 338	0.150200	3	KAOBC, AKOBCEGFH, AKBC	378, 232, 409	0.123999
[7, 8, 12, 4, 0, 9, 6, 10, 3, 14, 5, 2, 1, 13, 11]	2	KABC, AKOBCEGFH	681, 338	0.151720	3	KAOBC, AKOBCEGFH, AKBC	381, 240, 398	0.125440
[6, 3, 10, 7, 9, 11, 12, 1, 0, 8, 5, 14, 2, 4, 13]	2	KABC, AKOBCEGFH	690, 329	0.144982	3	KAOBC, AKOBCEGFH, AKBC	384, 236, 399	0.126615
[6, 10, 4, 0, 13, 8, 14, 3, 11, 12, 5, 1, 7, 2, 9]	2	KABC, AKOBCEGFH	694, 325	0.140625	3	KABC, AKOBCEGFH, KABCN	584, 320, 115	0.078299
[7, 6, 1, 0, 10, 8, 13, 12, 5, 4, 11, 2, 3, 9, 14]	2	KABC, AOBCEGFH	718, 301	0.137896	3	KABC, AOBCEGFH, KABCN	634, 297, 88	0.074783
[12, 8, 4, 2, 1, 3, 6, 14, 13, 0, 9, 7, 11, 5, 10]	2	KABC, AKOBCEGFH	686, 333	0.155987	3	KABC, AKOBCEGFH, KABCN	602, 324, 93	0.083032
[9, 7, 4, 1, 14, 0, 6, 8, 2, 11, 10, 12, 5, 13, 3]	2	AOBC, AOBCEGFH	803, 216	0.169143	3	AKBC, AOBCEGFH, KAOBC	442, 207, 370	0.127518
[4, 13, 9, 3, 2, 0, 12, 7, 8, 14, 1, 11, 5, 10, 6]	2	KABC, AKOBCEGFH	680, 339	0.156321	3	KAOBC, AKOBCEGFH, AKBC	404, 240, 37	0.129861

Table 3.15: Sensitivity Analysis of Rankings.

Groupings	k	Medoids	Frequency	Score	k	Medoids	Frequency	Score
[A, B, L][C, F, J, N][D, G][E, I, O][H][K, M]	2	KAOBC, AKOBCEGFH	693, 326	0.120623	3	KAOBC, AKOBCEGFH, ABKC	484, 276, 259	0.089406
[A, J][B, O][C, E, F, G, L][D, N][H, M][I][K]	2	ABC, AOBCEGFH	638, 381	0.143741	3	KAOBC, AKOBCEGFH, AKBC	443, 213, 363	0.136208
[A, I][B, F][C, H, O][D, G][E, L][J, N][K, M]	2	AOBC, AOBCEGFH	741, 278	0.133041	3	KAOBC, AKOBCEGFH, ABOC	478, 239, 302	0.083123
[A, C, O][B, N][D, E, K, L][F, M][G][H, J][I]	2	KABC, AKOBCEGFH	670, 349	0.129042	3	KABC, AKOBCEGFH, KAOB CD	417, 298, 304	0.101415
[A, E, F, H, N][B, C][D][G][I, K, O][J, L][M]	2	AOBC, AOBCEGFH	769, 250	0.166322	3	AKBC, AOBCEGFH, KAOBC	411, 228, 380	0.139426
[A][B, F, L, M, O][C, J, K, N][D, H, I][E][G]	2	AOBC, AOBCEGFH	763, 256	0.145183	3	KAOBC, AKOBCEGFH, AKBC	432, 242, 345	0.129746
[A, E, J][B, G, K, M][C, D, H, L, N][F][I][O]	2	KAOBC, AKOBCEGFH	694, 325	0.125991	3	KAOBC, AKOBCEGFH, AKBCGH	497, 180, 342	0.111152
[A, B, E, J, M][C, N][D][F][G, I][H, K, L][O]	2	AKBC, KA OBCEGFH	654, 365	0.128011	3	KAOBC, AKBMCEGFH, KA OB CGH	575, 232, 212	0.115357
[A, K, M][B, F, O][C, J][D, L][E, I][G][H][N]	2	KABC, KA OBCEGFH	682, 337	0.134682	3	KABC, KA OBCEGFH, ABKC	402, 314, 303	0.077893
[A][B, C, H, J][D, K][E, G, M][F][I][L, N, O]	2	AOBC, AOBCEGFH	806, 213	0.154813	3	AKBC, AOBCEGFH, KAOBC	412, 179, 428	0.146364

Table 3.16: Sensitivity Analysis of Groupings.

3.9 Conclusions

In response to the findings of Chapter 2, this chapter discusses the development of a new distance metric, modified from the Needleman-Wunsch dynamic programming algorithm, that is specifically designed for clustering, and allows for expert interaction through the use of groupings and rankings of activities.

The modified metric was compared against eight other popular metrics, where it performed equally well, if not better, when used with k-medoids clustering. This comparison further highlight that each of the metrics produce different results and as such, confirms the hypothesis that careful consideration is needed when selecting a string metric.

Sensitivity analysis has shown that the clustering is sensitive to the initial medoids selected, however, these are issues that are common amongst these metrics and clustering. Methods for improving upon these factors have been discussed as further work, but it is anticipated that any gains here are likely to be small. Furthermore, the penalty values along with rankings and groupings selection were also sensitive to the users choices of values. This is as expected, as this is the main function that allows the user to interact with the method. It is advised to the user to take care when selecting these values and analyse the results to ensure that they are in agreeance with the results for the chosen values. Providing more guidance to users on selecting these values is an important area of further work.

This method can support clinical pathway redesign or optimisation by initially providing a more time efficient process for mapping clinical pathways through combining both data and expert knowledge. As a result of combining both data and expert knowledge the clusters should be more clinically relevant using the modified Needleman-Wunsch metric due to the rankings and groupings feature.

From a clinical perspective, the resulting clusters enable deeper examination of the activity interactions which can help to highlight patterns that were previously

undetectable when looking at the data as a whole. This can support decision makers in the pathway redesign process which could lead to reducing delays to diagnosis and improved outcomes. This can also allow decision makers to prospectively consider the capacity required at activities due to a awareness of preceding activity demand. Overall, the modified metric paves the way to adding more context to string distances, and bridges the gap between data and expert interaction.

Further Work

The following areas have been identified as further work:

- Smart selection of penalty values: Machine learning techniques could be utilised to select penalty values which highlight various relationships as appropriate.
- Modify the Jaro distance metric [144, 145] using the same idea for modifications, as it produces good silhouette scores.
- Consider a final adjustment to the modified value to account for the total number of characters that appear in both strings i.e. divide final value by the number of characters appearing in both.
- Further sensitivity analysis to aid guidance in selecting penalty values, rankings and groups, possibly providing the user with a small set of options to choose from.
- Investigation of the impact of allowing groupings of singular activities, and how this could be used effectively.
- Consider the use of value selection in practice, possibly applying the DELPHI method [75].

Chapter 4

Automating the Simulation Build

4.1 Introduction

Research Question 2

Is it feasible to automate the simulation build process?

Chapter 1 introduced the idea and motivation for automating the simulation build. This chapter explores how to automate and validate the input parameters.

The general idea for the model is to allow for capacity and demand analysis. Discussion with staff at VCC concluded that capacity could be reflected through how many individuals could be seen each day. Expanding on this, capacity can be thought of in terms of a finite number of perishable slots, i.e. once all slots are used, no more individuals can be seen until the slots reset the next day, and any excess is not carried over.

The chapter is structured as follows:

- Section 4.2 expresses the considerations required for supporting automation and introduces the routing procedures.
- Section 4.3 describes the extensions to Ciw [56] required to support the sim-

ulation model (see Section 1.5).

- Section 4.4 outlines the working dataset used henceforth.
- Section 4.5 explores and validates the general approach to automatically extracting the input parameters i.e. arrivals, service, capacity levels and warm up. The routing procedure *Raw Pathways* is discussed throughout, as it is intended to be used for validation.
- Section 4.6 discusses how flexibility will be build into Sim.Pro.Flow to support the custom parameters.
- Section 4.7 presents a method of calculating capacity based on percentage time targets (see Section 1.5).
- Section 4.8 closes the chapter with a summary of the overall simulation parameters and highlights areas for further work.

4.2 Considerations

The feasibility of automating the simulation build heavily depended on selecting a simulation software that could support this. The Python library Ciw [56] was chosen as it has a open source nature and code driven structure, which was ideal.

When building a DES, as previously discussed, the first consideration is the network structure for the model i.e. activities and relations between activities. This network will provide the framework that individuals can move through, henceforth denoted as pathways. Therefore, this network and the pathways need to be automatically extracted from the data, being aware of the general nature for application that is required (Section 1.3).

Recall from Section 1.3 that multiple interpretations of the clinical pathway will be required. Due to this, four routing procedures were identified namely *Raw Pathways*, *Full Transitions*, *Cluster Transitions* and *Process Medoids*. The general intention

is to use the *Raw Pathways* as a method of validating the input parameters. The remaining routing procedures then provide different interpretations of constructing the network and pathways, which progressively aim to reduce the complexity and minimise variations for the produced clinical pathway.

As the four routing procedures provide differing methods for constructing the networks and exploring the pathways, and as such will need to be extracted accordingly. For note, this chapter will focus solely on the *Raw Pathways* as a method for validation, and the specific considerations for the remaining routing procedures are explored in detail in Chapter 5.

For context, these four routing procedures can be split into two areas: *Full Transitions* and *Cluster Transitions* make use of a probabilistic approach to constructing the pathways from a transition matrix (not compatible with DT1), alternatively the *Raw Pathways* and *Process Medoids* construct exact pathways that need to utilise a process based routing approach. Process based routing is where the individual's path is generated as it enters the model and not probabilistically as it progresses through the system. This was not possible in Ciw. As a consequence Ciw was extended to incorporate the feature of process based routing (Section 4.3).

4.3 Extensions to Ciw - Process Based Routing

Ciw is a Python library for DES of open queuing networks [57], developed by Dr Palmer and has had multiple contributors/authors [57]. Ciw uses an event scheduling approach which reflects the three-phase simulation approach discussed by Robinson (2014) [242] [57]. Palmer et al., [219] discusses the development of Ciw and the mission statement of allowing reproducibility and best practices. It was this aspect that lent itself well to automation and gave Ciw the advantage over competing software e.g. SimPy [256] when choosing the software. In addition, Ciw has detailed documentation available [57] which provides information of how to use Ciw, providing examples of standard use, and how the user can stretch the capabilities of the

library further through custom behaviour [57].

Prior to this collaboration Ciw utilised only a transition probability matrix, which takes in an individual at activity i and sends them to activity j with probability at position (i, j) of the matrix. The matrix is independent of the individual and is memoryless i.e. only considers the individual's current position to determine the next.

This would be suitable for the probability based routing procedures (*Full Transitions* and *Cluster Transitions* reflecting DT2) but would not be able to support the routing procedures that required a fixed assigned pathway (*Raw Pathways* and *Process Medoids* reflecting DT1).

Taking inspiration from process based simulation systems [242], which are used in alternative Python simulation packages such as SimPy [256], where an individual's entire route is determined when the individual is generated. The idea was to allow the user to either use the transition probability matrix (already possible) by stating the matrix, or to assign each arrival activity a routing function. The function could be written by the user and take some information available from the individual and return their route through the network, i.e. pathway.

The following subsections will introduce Ciw in more detail before discussing the initial investigation comparing scenarios applying both methods (transition matrix and process based) and ultimately integrating the development into Ciw v2.0.0. Furthermore, additional customisations of Ciw were required to support other aspects of the simulations will also be discussed.

4.3.1 Introduction to Ciw

Essentially, Ciw is made up of a collection of nodes where events take place. To set up a system in Ciw a network has to be defined including four main parameters: Arrivals, Service, Servers and Routing. Another element that is important to highlight is customer class; this allows for different groups of individuals to be handled

differently in the network. It is also possible to include a parameter in the network that allows an individual to change customer class (this will be required during the initial investigation).

Ciw uses general time units which means that 1 can represent either 1 second, minute, hour, day, week, month, year etc., depending on the specific time units required by the user. Therefore, care is required when defining a network to ensure that consistent units are used throughout. The implementation of Ciw in Sim.Pro.Flow always defines 1 as 1 day and as such this shall be used throughout the report.

A minimal working example, using version v2.0.1 of Ciw, shall now be explored supported with explanation. For detailed information and tutorials the reader should explore Ciw’s documentation [57].

Consider a system with four activities, (node 1: ‘A’, node 2: ‘B’, node 3: ‘C’, node 4: ‘D’). There are two classes of individuals, where class 0 and class 1 will proceed through the network performing tasks in the order ‘ABC’ and ‘BCAD’ respectively. Therefore, class 0 individuals will arrive at node 1 (activity ‘A’) and class 1 individuals will arrive at node 2 (activity ‘B’), with arrival distribution exponential with $\lambda = 1$ (1 per day). Each node will have service rates sampled from exponential distribution with $\mu = 2$ (2 per day) and one server. This network can be seen in Figure 4.1.

```

N = ciw.create_network(
  arrival_distributions={
    'Class 0': [ciw.dists.Exponential(1), ciw.dists.NoArrivals(), ciw.dists.NoArrivals(),ciw.dists.NoArrivals()],
    'Class 1': [ciw.dists.NoArrivals(), ciw.dists.Exponential(1), ciw.dists.NoArrivals(),ciw.dists.NoArrivals()],
  },
  service_distributions={
    'Class 0': [ciw.dists.Exponential(2), ciw.dists.Exponential(2), ciw.dists.Exponential(2), ciw.dists.Exponential(2)],
    'Class 1': [ciw.dists.Exponential(2), ciw.dists.Exponential(2), ciw.dists.Exponential(2), ciw.dists.Exponential(2)],
  },
  number_of_servers= [1,1,1,1],
  routing={
    'Class 0': [[0,1,0,0],[0,0,1,0],[0,0,0,0],[0,0,0,0]],
    'Class 1': [[0,0,0,1],[0,0,1,0],[1,0,0,0],[0,0,0,0]],
  }
)

```

Figure 4.1: Ciw Network Example Extracted From Jupyter Notebook.

Note in Figure 4.1 that routing has been defined in terms of a transition probability

matrix. Typically in a transition matrix the rows sum to 1, however in C_{iw} , any remaining probability is assigned to exit the system e.g. class 0 row 3 has all 0's meaning that from node 3 individuals will exit the system with probability 1.

Furthermore, the arrivals, service and routing are defined by class but servers are universal.

The simulation seed can then be defined and the network can be passed to the simulation as required by C_{iw} . Finally there are two methods of running the simulation - until maximum time or maximum number of customers has been reached [57]. There are three methods available for specifying the stop criteria in relation to the specified number of customers: 'Finish' - hit the exit node, 'Arrive' - created in the arrival node, and 'Accept' - created and been accepted into the system [57]. For this example simulate until maximum customer of 250 was chosen with method specified as 'Finish' as shown in Figure 4.2

```

ciw.seed(0)
Q = ciw.Simulation(N)
Q.simulate_until_max_customers(250, method='Finish', progress_bar=True)

```

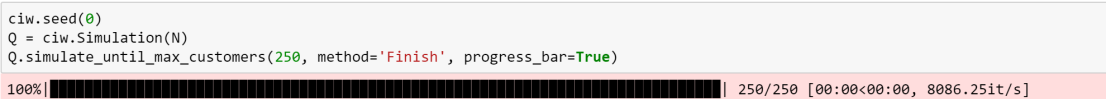


Figure 4.2: Ciw Run Network Example Extracted From Jupyter Notebook.

Finally, once the simulation has been run, C_{iw} can produce a table of results [57] where each row logs an event at a node. In other words, for each unique customer ID there can be multiple rows which log the individual's interaction with a node, including information such as arrival date, wait time, service time, destination and queue size at arrival/departure, to name a few (for full list see 'List of Available Results' in C_{iw} 's documentation [57]). These records provide a solid base from which to extract further results if the user requires, such as percentage waiting less than a specified time per node, the average waiting time at a node etc.

Figure 4.3 shows an example of how to check that the expected number of individuals per class were generated, which in this case should be approximately evenly spread amongst the two classes as their arrival rates were the same.


```
recs = pd.DataFrame(Q.get_all_records())
len(set(recs.id_number))
257
recs_c0 = recs[recs.customer_class == 0]
len(set(recs_c0.id_number))
137
recs_c1 = recs[recs.customer_class == 1]
len(set(recs_c1.id_number))
120
```

Figure 4.3: Ciw Run Results Example Extracted From Jupyter Notebook.

4.3.2 Initial Investigation

The purpose of the initial investigation was to identify what situations could not be implemented in Ciw in its current state and as proof of concept that process based routing would perform correctly and satisfy the gaps not covered by the transition matrix method. Version 1.1.6 of Ciw was used for this investigation.

This initial investigation enabled us to explore how to adapt Ciw to accommodate process based routing, where it was necessary to make use of Ciw's *CustomArrivalNode* and *CustomNode* classes to enable the exploration, as described below.

The *have_event* function of the *CustomArrivalNode* was edited to allow for the individual to have a *route* attribute consisting of a *generate_route* function. The *generate_route* function changes for each case explored, but fundamentally takes the customer class as input and returns an ordered list of nodes.

Within the *CustomNode* class, the *next_node* function was completely rewritten to perform the following steps: 1) remove the first entry in the route list (corresponding to the current node) 2) if there are no more nodes in the route, then exit the system, otherwise set the next node number to be the new first entry in the route list (which will be the next destination).

For the initial investigation three routing cases were explored for both event scheduling and process based methods each satisfying a different goal.

- Case 1 investigated how to implement process based routing for a basic network. The chosen network was as described in the previous subsection, with two classes each performing a different route.
- Case 2 considered repeating an activity a set number of times, specifically the route ‘AAA’.
- Case 3 considered specific routes being performed based on set probability by individuals of the same class. Specifically, individuals would perform the route ‘AABCA’ 40% of the time and route ‘BCAB’ the remaining 60% of the time.

Each case required specific technical accommodations to perform the intended routing. Discussing each case in turn;

In case 1 (previously described in detail) the methods differ in routing, where the event based used a transition matrix, whereas, the process based routing function take in the customer class and returns [1, 2, 3] for class 0 and [2, 3, 1, 4] for class 1.

In case 2, for event scheduling approach it would have not been possible here to use the typical transition matrix with A to itself with probability 1, as the individual would never exit the system. Therefore, case 2 utilised Ciw’s class change matrix. The network was constructed of only the one node with four classes of customers. Only individuals of class 0 arrived at node 1 (‘A’) with all other classes having *NoArrivals*. Each time the individual performed the activity, they changed class according to the class change matrix, cycling through class 0 to 3 in order. Each class had a routing transition matrix which kept it at activity ‘A’ with probability 1 until it was required to leave the system at class 3. For clarification the order of events would be as follows: Arrive at node 1 as class 0, perform event, change to class 1, transition matrix sends class 1 to node 1, perform event, change to class 2, transition matrix sends class 2 to node 1, perform event, change to class 3, transition matrix sends class 3 to exit node. Alternatively, the process based allowed for a simple system consisting of one node and class, where the routing function was always [1, 1, 1].

Case 3 was the most complex, expanding the idea of case 2 with an exact number of repeats, whilst assigning one route 40% of the time and another the remaining 60%. This was deemed not possible for the event scheduling method. Whereas the process based method allowed for a simple and elegant solution where the individuals arrived at a dummy node (1), assigned their route, where the routing function was: generate a random number in the range $[0,1]$ denoted rnd , if $rnd < 0.4$ then perform $[1, 2, 2, 3, 4, 2]$, otherwise perform $[1, 3, 4, 2, 3]$. Note the node numbers corresponding to the letter are increased by 1 to accommodate the dummy node.

The basic set up was similar to the previously worked example, where the arrival distributions were exponential with $\lambda = 1$, the service distributions were exponential with $\mu = 2$, each node had one server. A difference to note is that the simulation was run until max time of 250 time units (days) were reached. Furthermore, in case 3 the dummy node required a service time, which was set at deterministic with $\mu = 1$.

From the results produced by Ciw it was possible to extract the routes that were performed by the individuals which are displayed in Table 4.1. This confirms that the correct routes were executed in all three cases, including approximately the correct proportion of routes performed in case 3 (246 individuals, class 0 104 times and class 1 142 times).

Table 4.1: Results of Routes Performed for Cases Used in Ciw Initial Investigation.

Case	Event Scheduling	Process Based
1	Class 0: $\{(1, 2), (1,), (1, 2, 3)\}$ Class 1: $\{(2,), (2, 3, 1, 4), (2, 3)\}$	Class 0: $\{(1, 2), (1,), (1, 2, 3)\}$ Class 1: $\{(2,), (2, 3, 1, 4), (2, 3), (2, 3, 1)\}$
2	$\{(1,), (1, 1), (1, 1, 1)\}$	$\{(1,), (1, 1), (1, 1, 1)\}$
3	Not Possible	$\{(1,):2, (1, 2, 2):1, (1, 2, 2, 3, 4):1, (1, 2, 2, 3, 4, 2):100, (1, 3, 4, 2):5, (1, 3):1, (1, 3, 4, 2, 3):136\}$

**format set{(routes performed): frequency}*

This initial result concludes that it is possible and appropriate to use process based, whilst also satisfying a gap in Ciw's abilities.

The main consideration to take forward into development was when to assign the route to the individual. There were two options available, 1) have the individuals all arrive at the same node (dummy node), then assign the route, then send it to the first node (like case 3) or 2) assign the route at the correct arrival node, ensuring that the first node in the list matches that of the arrival node. On discussion, option 2 was superior as it aligned with Ciw's structure of allowing different arrival distributions for each node and could also accommodate option 1. Therefore, we proceeded with option 2, taking care to satisfy the restriction mentioned.

4.3.3 Development

The development stage consisted of progressing the initial investigation to make a contribution to Ciw's base code. Technically this involved forking the Ciw repository on Github [57] to obtain a copy of Ciw's code to work on. Development could then take place which would not affect the main parent repository of Ciw and commits could record the changes made. To adhere to the best practices that Ciw promotes, testing files using the unittest library [278] were required to ensure the expected outcomes are produced (a *test_process_based* file was created). Once all necessary changes were complete, a pull request was made which allowed Dr Palmer to review the contribution and merge them into the main branch.

There were many elements that were required to support the process based routing (including progressing/implementing those discussed in the previous section). However, the critical development undertaken was as follows:

- The keyword *routing* can take in an ordered list of length N . This will detect that process based routing is requested.
- Each position of the list that corresponds to an arrival node should contain the relevant routing function, otherwise contain *ciw.no_routing*.
- The routing function should take in the individual, and return an ordered list of node numbers, which has to start at the current node.

A page describing the process based routing is included in Ciw's documentation. This both explains to the user how to use the routing and the corresponding node restrictions to be noted [57]. Furthermore, the pull request was accepted by Dr Palmer and the process based routing contribution was included in the Ciw from version 2.0.0 onwards.

4.3.4 Custom Ciw

Ciw has the ability to allow users to define custom behaviour for nodes and arrival nodes. This was extended to also include servers and individuals in Ciw v2.1.2 onwards. Furthermore, custom distributions can be used for arrivals and service.

Two custom behaviours were utilised as they were necessary to accommodate specific features required for the simulation. The development of these will be discussed in this section, whereas the necessity and implementation of these will be discussed and referenced when required.

Restricting Capacity

It was required that a specific number of individuals to be serviced at each node each day. Thus a custom node was developed that would no longer accept individuals to be serviced once a limit was reached (see Section 1.5). Therefore, the limit needs to be specified and a count of the individuals needs to be recorded. These will be referred to as *slot_capacity* and *current_count* respectively, Furthermore the limit needs to be reset occasionally.

The requirement of setting a limit and resetting that limit lent itself well to utilising the server schedule. Ciw's servers schedule feature allows users to specify the number of servers available at a node until a specified time unit, where servers can service an individual on a one-to-one basis. The server schedule is cyclical and as such once the maximum defined time unit for the schedule has been reached, the schedule will repeat. For example, defining six slots on day one, seven slots on day two would be

$$\text{number_of_servers} = [[6, 1], [7, 2]]. \quad (4.1)$$

Users should be aware that Equation 4.1 describes having six servers until time unit 1, where 1 time unit is a day. Therefore day one occurs between the time units 0 and 1, where 1 is not included.

Implementing this in the custom node required three main considerations: define *current_count*, establish if service can begin and how reset the define *current_count*. This involved changing methods that already existed as part of the node class, where the original docstrings are retained for information and comments highlighted the added or indented lines of code. The three main considerations will now be discussed.

Firstly, simply defining *current_count* as a new attribute and setting to 0 on initialisation satisfied the first consideration. Secondly, there are three methods defined which consider if an individual can begin service, namely:

begin_service_if_possible_accept, *begin_service_if_possible_release* and *begin_service_if_possible_change_shift*.

In each of these methods the same three changes were made: set the slot capacity to the number of servers (defined as c in C_{iw}), accept the individual if the *current_count* is less than the *slot_capacity* and if the service goes ahead then increment the *current_count* by one. Each of these methods had various prior conditions determining if they would accept individuals which have all been retained. Thirdly, when a shift is changed, the current count needs to be reset to 0.

This custom node will be used in conjunction with a server schedule which will make use of a weekly schedule as described in subsection 4.5.4. The cyclical nature of the servers schedule was not utilised, which allowed for the development of the warm start (subsection 4.5.5). As such the schedule was defined for every day over a period of 1.5 times the ‘overall period’, where the ‘overall period’ is the total number of days covered in the original data. This was chosen as it was deemed large enough

to cover the number of days covered by the simulation, whilst not being too much of a hindrance on run time.

Reducing the schedule down to a seven day defined cycle will be suggested as further work along with validation to ensure that the features of the custom node still hold.

Blocking Arrivals

When simulating the *Raw Pathways* there were only a finite number of pathways that could be assigned. This finite number related to the number of individuals in the original dataset, which will be referred to as I . This caused an issue as once all the pathways had been assigned the simulation would stop due to the error that an individual was not assigned a node to move on to. Therefore, blocking arrivals after I number of individuals has been reached was explored as a possible solution. It is noted that this would result in the final few individuals not having to compete for resources (discussed further in Section 5.5).

Utilising the solution to an Issue logged in Ciw (Issue 171), which concerned requiring a fixed number of customers for each class (see Section 1.5). The solution describes a custom exponential distribution where after a defined number of individuals have arrived, the next arrival occurs at time infinity. This custom exponential distribution is called *LimitedExponential*, and takes in two parameters rate and limit, where rate is the arrival rate λ and limit is the number at which the next arrival will occur at time infinity i.e. for $I = 1865$, once individual with ID number 1865 has arrived the next arrival will occur at time infinity, essentially blocking any more arrivals. This elegant solution proved to be sufficient.

For validation purposes, the custom distribution was extended to consider the deterministic distribution (*LimitedDeterministic*) to allow for consistent arrivals, which allowed for the simulation results to be more easily assessed. This was not implemented in any final model, however it remains in the *custom_ciw* file for potential future use.

4.4 Working Dataset

This section outlines the worked example dataset used throughout this chapter. From the initial dataset presented in Section 3.3, early investigation of the simulation noted that the overall period completed in the simulation was abnormally long. This was due to the fact that the initial dataset covered 1,069 days due to some records containing dates outside of a sensible period. Therefore, a sub-dataset was selected by taking the 12 month interval from the 1st of January, and any individuals whose record contained a date outside of this period were excluded. This interval was chosen as it retained the highest number of individuals at 1,865. This resulted in 783 unique pathways, which are contained within the dataset ‘dataframe’, and is used when performing clustering. Although the most popular pathway (‘BIAMC’) was performed 105 times. A summary of the data can be seen in Appendix D.

In doing this, there were no longer any records of brachytherapy - activity ‘I’ or activity ‘J’. Therefore, these activities were removed and as such the letters of all activities post ‘I’ and ‘J’ have been renamed as in Table 4.2. In regards to the rankings and groupings for clustering: activities ‘I’ and ‘J’ were the last two ranks, thus the rankings remain the same (as they are presented as ‘chosen by an expert’). Furthermore, activities ‘I’ and ‘J’ were in a group of their own, thus the group assignments remain the same with an updated number. These are included in Table 4.2 for clarity.

Table 4.2: Pathway Activities New Letter Assignments.

Name	Old Letter	New Letter	Rankings	Groupings
First Seen	A	A	2	5
Diagnosis	B	B	0	5
MDT Discussion	C	C	1	5
Procedure	D	D	12	3
Decision to Treat Chemotherapy	E	E	10	0
Chemotherapy Start	F	F	9	0
Decision to Treat Teletherapy	G	G	7	1
Teletherapy Start Date	H	H	6	1
CT Scan	K	I	3	2
PET/PET CT Scan	L	J	5	2
Bronchoscopy	M	K	8	4
CT Guided Biopsy	N	L	11	2
Specialist Nurse Seen	O	M	4	5

There are two main tables of results reported throughout this chapter.

- Top Level Results - This contains the following results:
 - Mean TiS - Mean average time in system
 - Median TiS - Median average time in system
 - Target - Percentage of total waiting times within target of 62 days
 - Overall Period - time period covered
- Activity Waiting Time - Displays the mean waiting times for each activity.

Table 4.3 presents the top levels results for the original data (the sub-dataset) while Table 4.4 contains the activity waiting times, to allow comparison throughout the chapter. Additionally, Table 4.4 also contains the activity frequency, along with median, 25th and 75th percentiles for more detail.

Table 4.3: Top Level Results for Original Data.

Mean TiS	Median TiS	Target	Overall Period*
60.0	41.0	64.40	362.0

**Overall periods not 365 as expected, speculated due to no activity recorded on the first/last few days in the period.*

Table 4.4: Activity Specific Results for Original Data.

Activity	Frequency	Mean	25 th Percentile	Median	75 th Percentile
A	1797	12.52	3.0	7.0	15.00
B	1865	11.86	2.0	7.0	14.00
C	1855	9.40	4.0	7.0	11.00
D	232	21.91	0.0	12.0	38.00
E	473	11.19	4.0	7.0	15.00
F	475	20.20	5.0	12.0	24.50
G	536	6.41	0.0	1.0	7.00
H	537	40.21	6.0	18.0	70.25
I	1797	4.05	0.0	0.0	2.00
J	537	13.66	6.0	12.0	17.00
K	589	3.58	0.0	0.0	2.00
L	364	3.85	0.0	0.0	0.00
M	1577	3.20	0.0	0.0	2.00

4.5 Validating Input Parameters

This section investigates how to automatically extract the main input parameters (arrivals, service, capacity and warm up) and validates the chosen methods. As the *Raw Pathways* are the routing procedure intended for validating the model, it will be discussed here and used with the worked example dataset for lung cancer throughout to enable investigation.

The simulation will run until I individuals have exited (Ciw method "Finish"), where for this working example $I = 1865$. Each trial will consist of 25 runs unless otherwise stated. Twenty-five runs were deemed sufficient after initial investigation showed that at this point the standard deviation becomes consistent.

The key performance indicators (KPI's) are reported in the format of the two tables (top level results and activity waiting times) as described in Section 4.4.

4.5.1 General Arrivals

The arrival rate is the rate at which individuals will arrive into the system through the arrival node/s. Typically the exponential distribution is used for the arrival process, with mean arrival rate λ . A general approach to calculating λ is shown in Equation 4.2.

$$\lambda = \frac{A_n}{P} \quad (4.2)$$

Here A_n is the number of arrivals at an arrival node, which will be defined specifically for each routing procedure (see 4.5.2 for *Raw Pathways* and Chapter 5 otherwise). Furthermore, P is the overall period where in these models, as the simulation clock is continuous and does not pause for weekends, the total time period is the number of days covered by the data, regardless of if 5 or 7 working days a week will be used in the simulation.

4.5.2 *Raw Pathways* Routing Procedure

The *Raw Pathways* routing procedure will simulate the exact pathways from the input data using process based routing. Therefore, as the pathways are limited to only what has previously been performed, this would make it difficult to use these pathways to predict the future. As such the *Raw Pathways* routing procedure is intended to be used for validation of the input parameters.

The visualisation of the *Raw Pathways* can be displayed as a heat map, like that seen in the Summary Sheet (Appendix D). This displays the variety of pathways that are performed and could provide a useful aid to examine if the *Raw Pathways* are in a particular order e.g. if alphabetical, then blocks of colour in ascending alphabetical order could be seen.

In general, arrivals for the *Raw Pathways* model will be through a dummy node (node 1) where individual's will be assigned a pathway. As such all individual's arrive at node 1, thus $A_1 = I$.

For example, if the an individual in the data performed the pathway 'ABC', the corresponding individual in the simulation would perform the route [1,2,3,4], as the letters are mapped to their position in the alphabet plus one (as the dummy arrival node is $n = 1$). This then raises the question of: Does the initial order that the pathways appear in the data affect the simulation?

The initial experiment explores fifty random orderings of the pathways. The random ordering was chosen through using Python's random library and the sample function. One run for each ordering was performed with the simulation seed fixed at 0. This was to allow for isolating the effect of the orderings.

For the purpose of the experiment, the input parameters (arrivals, service and capacity) for the *Raw Pathways* routing procedure need to be set.

For the arrivals, Equation 4.2 is applied in Equation 4.3, where the overall period was taken from Table 4.3. This λ is then used with the custom distribution *LimitedExponential* with limit of $I = 1865$ (see **Blocking Arrivals** on page 101 for details).

$$\lambda = \frac{A_n}{P} = \frac{I}{P} = \frac{1865}{362} = 5.15 \quad (2 \text{ d.p}) \quad (4.3)$$

As the service and capacity values are yet to be explored, they will be set as follows:

- Service rate (see subsection 4.5.3): dummy node = deterministic with $\mu = 0$, activity node = deterministic with $\mu = 0.1$.
- Capacity (see subsection 4.5.4): dummy node = 543 slots a day¹, activity node = 4 slots a day (as this is just below the arrival rate and should allow for queues to form), for 7 working days a week.

The results were collected and confidence intervals across the fifty orderings were calculated. The top level results are displayed in Table 4.5 and the activity waiting times are displayed in Table 4.6.

Table 4.5: Top Level Results of Varying Pathway Orderings.

Mean TiS	Median TiS	Target	Overall Period
66.4 (65.05, 67.75)	66.94 (65.4, 68.48)	46.48 (45.43, 47.53)	486.63 (486.12, 487.13)

Table 4.6: Activity Waiting Time Results of Varying Pathway Orderings.

Activity	Waiting Time
A	29.74 (28.97, 30.51)
B	8.95 (8.78, 9.13)
C	2.45, (2.37, 2.52)
D	0.0 (0.0, 0.0)
E	0.0 (0.0, 0.0)
F	0.0 (0.0, 0.0)
G	0.0 (0.0, 0.0)
H	0.0 (0.0, 0.0)
I	27.12 (26.41, 27.83)
J	0.0 (0.0, 0.0)
K	0.0 (0.0, 0.0)
L	0.0 (0.0, 0.0)
M	0.91 (0.86, 0.97)

¹Arbitrarily large to ensure no queues. Set using equation $1.5P = 1.5 \times 362 = 543$.

The confidence intervals displayed indicated little variation, suggesting that the ordering does not matter. However Figure 4.4 and Figure 4.5 graph the raw results shown in Table 4.5 and Table 4.6 respectively. On closer inspection of the raw results, there is more variation displayed in Figure 4.4 and Figure 4.5. In particular, Mean time in system (Figure 4.4) varies between 54 and 75 days and Activity ‘A’ waiting time in (Figure 4.5) varies between 25 and 35 days.

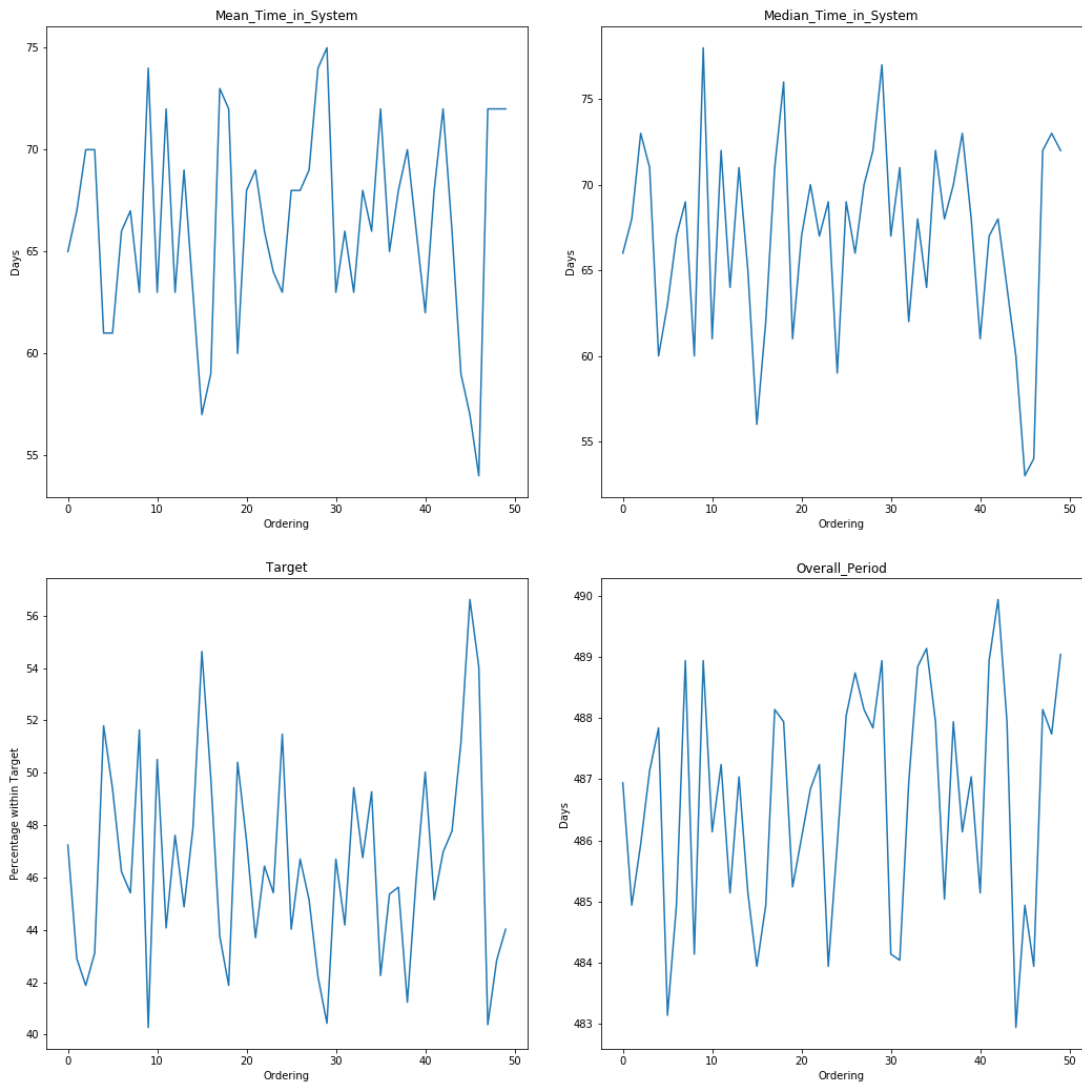


Figure 4.4: Raw Top Level Results of Varying Pathway Orderings.

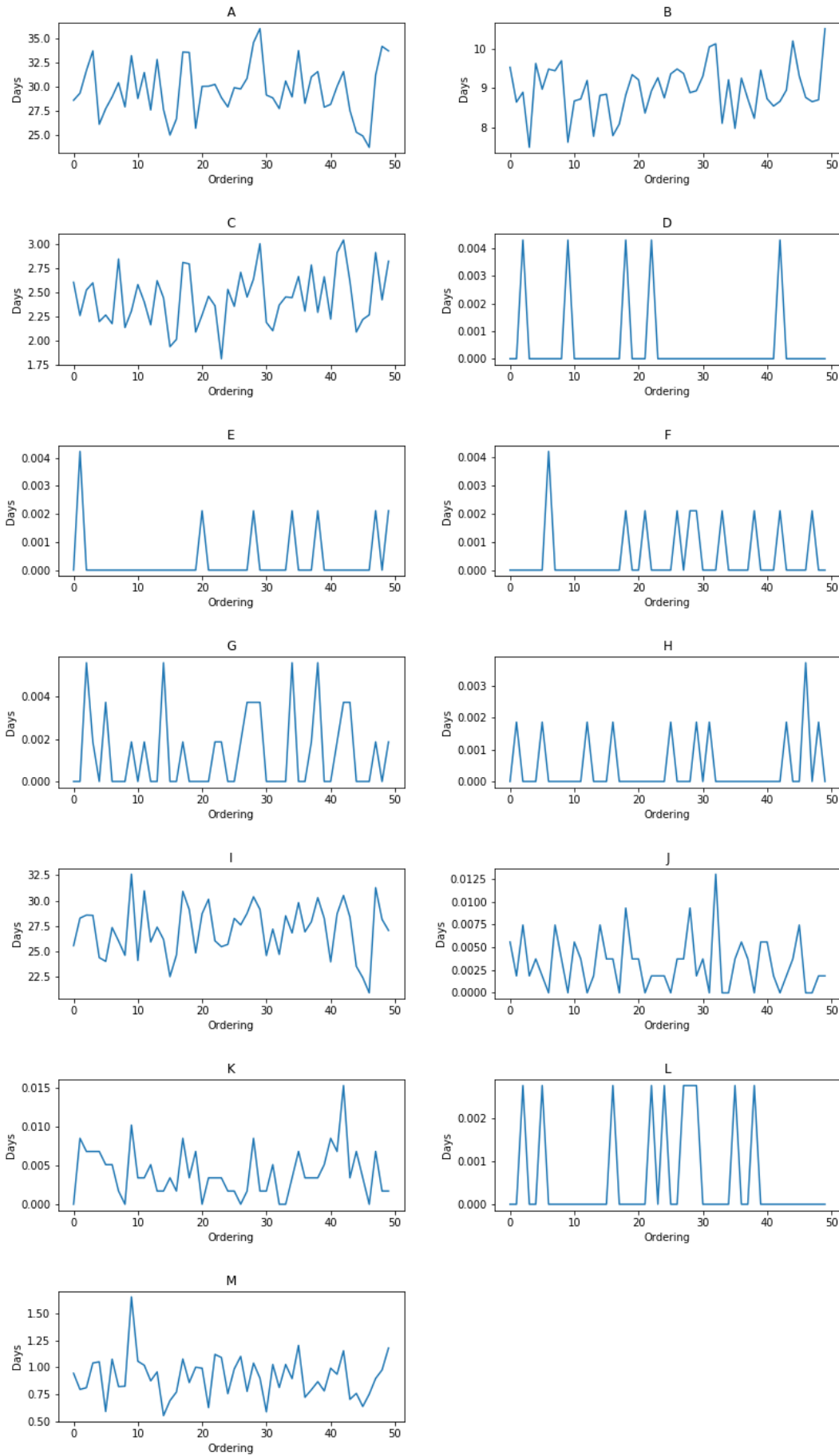


Figure 4.5: Raw Activity Waiting Time Results of Varying Pathway Orderings.

This suggests that order does matter and as such the automated orderings will be taken exactly as they appear in the data. However, in this case one needs to consider if the data is already in a specific order and if some particular orderings require extra care. This led further investigation of 4 specific scenarios: default ordering, reverse default ordering, alphabetically ascending (A to Z), alphabetically descending (Z to A). Using the same variables set up as previous, Table 4.7 and Table 4.8 display the top level results and activity wait time results of this investigation respectively.

Table 4.7: Top Level Results of Pathway Scenarios.

Method	Mean TiS	Median TiS	Target	Overall Period
Default	72.0	65.0	46.434316	501.738853
Reverse	61.0	63.0	47.292225	480.838853
Ascending	58.0	58.0	52.761394	486.838853
Descending	76.0	81.0	40.268097	496.238853

Table 4.8: Activity Waiting Time Results of Pathway Scenarios.

Activity	Original	Reverse	Ascending	Descending
A	30.036171	27.810239	10.677240	39.030607
B	16.775871	5.232172	10.192493	11.363003
C	4.916981	2.525067	3.258760	12.328302
D	0.000000	0.004310	0.000000	0.000000
E	0.002114	0.006342	0.000000	0.000000
F	0.000000	0.004211	0.000000	0.000000
G	0.005597	0.001866	0.042910	0.013060
H	0.005587	0.000000	0.018622	0.000000
I	22.396216	26.785754	29.584307	11.698943
J	0.007449	0.003724	0.432030	0.061453
K	0.010187	0.018676	0.122241	0.056027
L	0.000000	0.000000	0.019231	0.000000
M	0.606848	1.227647	6.983513	5.024096

Observing the results in Table 4.7 it is clear that using specific orderings will have an effect on the top level results. The explanation for this lies in Table 4.8. In the original and reverse orderings, the results are not too dissimilar until considering activity 'B'. The wait time for activity 'B' is a lot higher in the original ordering than in the reverse ordering. Considering the ascending and descending orderings allow for more clear explanation.

In the descending ordering, pathways starting with 'M' would occur first, and pathways starting with 'A' would appear last. Therefore, by the time the pathways

starting with ‘A’ would appear in the system, the queue for activity ‘A’ would already be quite large. Comparing this to the ascending ordering, where the opposite would occur, as those beginning at activity ‘A’ would appear first, there would be a lot less time for a queue to build up.

Overall it can be concluded that specific orderings will have an effect on the wait time of an activity, and as such the ordering will be taken exactly as it appears in the data. This has a further benefit as the users may want to investigate that specific ordering which this method would support. Users should be aware of particular ordering scenarios that could unintentionally impact the waiting times of the activities.

4.5.3 Service

As the number of individuals able to perform an activity each day is restricted by the number of slots available, this raises the question of if service time is trivial - does it make a difference how long an individual occupies that slot, and if so, how much of an impact?

To investigate service time, an experiment was run varying service time values (μ) between 0.1 and 1 at 0.1 intervals. Both deterministic and exponential distributions were tested and reported separately. For the purpose of this investigation the arrivals and capacity levels are as set on page 106, and 10 runs were included in each trial.

The top level results are reported in Table 4.9 and Table 4.10 and the activity waiting time results are reported in Table 4.11 and Table 4.12, for deterministic and exponential distributions respectively.

All four tables show that varying the service time does not significantly effect the KPI's. As a result, service time will be automatically set to deterministic with $\mu = 0.1$ for each activity node (for dummy nodes $\mu = 0$).

Table 4.9: Top level Results of Varying Deterministic Service Time [0.1,1].

μ	Mean TiS	Median TiS	Target	Overall Period
0.1	74.0, (70.44, 77.56)	68.0, (64.18, 71.82)	45.23, (42.72, 47.75)	504.21, (502.37, 506.00)
0.2	71.0, (67.26, 74.74)	66.0, (62.01, 69.99)	46.99, (44.35, 49.64)	503.12, (501.05, 505.18)
0.3	73.0, (69.68, 76.32)	67.0, (63.46, 70.54)	46.04, (43.76, 48.33)	504.08, (502.32, 505.83)
0.4	73.0, (70.28, 75.72)	67.0, (64.12, 69.87)	45.92, (44.05, 47.80)	504.36, (502.93, 505.78)
0.5	74.0, (71.58, 76.42)	68.0, (65.49, 70.51)	45.40, (43.70, 47.10)	504.70, (503.41, 505.98)
0.6	74.0, (71.90, 76.10)	68.0, (65.82, 70.18)	45.41, (43.94, 46.89)	504.83, (503.71, 505.95)
0.7	74.0, (72.07, 75.93)	68.0, (66.01, 69.99)	45.01, (43.64, 46.37)	505.01, (503.98, 506.04)
0.8	75.0, (73.19, 76.81)	69.0, (67.13, 70.87)	44.56, (43.27, 45.85)	505.27, (504.30, 506.24)
0.9	75.0, (73.32, 76.68)	69.0, (67.26, 70.74)	44.48, (43.27, 45.69)	505.48, (504.58, 506.39)
1	76.0, (74.44, 77.56)	69.0, (67.37, 70.63)	44.03, (42.91, 45.16)	506.00, (505.12, 506.87)

Table 4.10: Top level Results of Varying Exponential Service Time [0.1,1].

μ	Mean TiS	Median TiS	Target	Overall Period
0.1	79.0, (75.61, 82.39)	73.0, (69.55, 76.45)	40.62, (38.43, 42.81)	511.50, (509.60, 513.40)
0.2	79.0, (76.85, 81.15)	72.0, (69.44, 74.56)	41.56, (39.81, 43.32)	509.77, (508.19, 511.35)
0.3	76.0, (73.39, 78.61)	70.0, (67.14, 72.86)	43.24, (41.33, 45.15)	508.01, (506.27, 509.74)
0.4	76.0, (73.64, 78.36)	69.0, (66.54, 71.46)	44.00, (42.32, 45.68)	507.22, (505.62, 508.80)
0.5	75.0, (72.96, 77.04)	69.0, (66.85, 71.15)	44.17, (42.71, 45.64)	506.96, (505.62, 508.30)
0.6	76.0, (74.14, 77.86)	69.0, (67.09, 70.91)	44.08, (42.80, 45.36)	506.81, (505.63, 508.00)
0.7	76.0, (74.26, 77.74)	69.0, (67.22, 70.78)	44.10, (42.91, 45.29)	506.74, (505.67, 507.80)
0.8	75.0, (73.37, 76.63)	69.0, (67.29, 70.71)	44.50, (43.34, 45.66)	506.44, (505.45, 507.42)
0.9	75.0, (73.48, 76.52)	68.0, (66.41, 69.59)	44.70, (43.59, 45.74)	505.99, (505.04, 506.94)
1	75.0, (73.60, 76.40)	68.0, (66.55, 69.45)	44.76, (43.76, 45.76)	505.77, (504.89, 506.60)

Table 4.11: Activity Waiting Time Results of Varying Deterministic Service Time [0.1,1].

Activity	0.1	0.2	0.3	0.4	0.5
A	31.15, (29.42, 32.87)	30.1, (28.44, 31.75)	30.69, (29.27, 32.12)	30.66, (29.49, 31.84)	30.75, (29.71, 31.79)
B	17.23, (16.79, 17.67)	16.74, (16.23, 17.24)	16.93, (16.48, 17.37)	16.95, (16.56, 17.33)	17.11, (16.76, 17.46)
C	5.36, (5.1, 5.62)	5.26, (4.97, 5.55)	5.4, (5.15, 5.64)	5.48, (5.28, 5.68)	5.48, (5.29, 5.67)
D	0.0, (-0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)
E	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.01)	0.01, (0.0, 0.01)
F	0.0, (0.0, 0.0)	0.0, (0.0, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)
G	0.0, (0.0, 0.01)	0.0, (0.0, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)
H	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.01)	0.0, (0.0, 0.01)
I	22.62, (21.23, 24.0)	21.74, (20.27, 23.21)	22.34, (21.02, 23.66)	22.33, (21.25, 23.4)	22.39, (21.45, 23.34)
J	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
K	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
L	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)
M	0.6, (0.53, 0.66)	0.55, (0.49, 0.6)	0.58, (0.53, 0.63)	0.59, (0.54, 0.63)	0.59, (0.55, 0.63)
Activity	0.6	0.7	0.8	0.9	1
A	30.65, (29.75, 31.55)	30.73, (29.9, 31.56)	30.8, (30.03, 31.57)	30.84, (30.14, 31.55)	30.86, (30.22, 31.5)
B	17.1, (16.79, 17.4)	17.16, (16.88, 17.44)	17.25, (16.98, 17.52)	17.29, (17.04, 17.53)	17.36, (17.13, 17.58)
C	5.47, (5.31, 5.64)	5.46, (5.31, 5.62)	5.49, (5.34, 5.64)	5.51, (5.37, 5.65)	5.53, (5.4, 5.65)
D	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)
E	0.01, (0.0, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
F	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
G	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
H	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
I	22.29, (21.47, 23.12)	22.36, (21.62, 23.11)	22.46, (21.77, 23.16)	22.51, (21.87, 23.16)	22.58, (22.0, 23.17)
J	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
K	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
L	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)
M	0.6, (0.56, 0.64)	0.6, (0.57, 0.63)	0.59, (0.57, 0.62)	0.6, (0.57, 0.63)	0.6, (0.58, 0.63)

Table 4.12: Activity Waiting Time Results of Varying Exponential Service Time [0.1,1].

Activity	0.1	0.2	0.3	0.4	0.5
A	30.92, (29.42, 32.42)	31.35, (30.38, 32.32)	30.86, (29.79, 31.93)	30.88, (29.89, 31.88)	30.92, (30.06, 31.79)
B	18.05, (17.5, 18.6)	17.87, (17.53, 18.2)	17.44, (17.03, 17.84)	17.31, (16.96, 17.66)	17.35, (17.04, 17.66)
C	5.85, (5.5, 6.21)	5.78, (5.58, 5.99)	5.64, (5.42, 5.86)	5.6, (5.38, 5.82)	5.59, (5.41, 5.77)
D	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)
E	0.02, (0.01, 0.02)	0.01, (0.01, 0.02)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
F	0.02, (0.01, 0.02)	0.01, (0.01, 0.02)	0.01, (0.01, 0.02)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
G	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
H	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
I	22.75, (21.36, 24.14)	23.01, (22.13, 23.89)	22.56, (21.61, 23.51)	22.56, (21.7, 23.41)	22.53, (21.79, 23.27)
J	0.02, (0.01, 0.02)	0.01, (0.01, 0.02)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
K	0.03, (0.03, 0.04)	0.02, (0.02, 0.03)	0.02, (0.02, 0.03)	0.02, (0.02, 0.02)	0.02, (0.01, 0.02)
L	0.01, (0.0, 0.01)	0.0, (0.0, 0.01)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)
M	0.64, (0.61, 0.67)	0.63, (0.6, 0.66)	0.61, (0.58, 0.64)	0.61, (0.58, 0.63)	0.61, (0.58, 0.64)
Activity	0.6	0.7	0.8	0.9	1
A	31.13, (30.34, 31.92)	31.31, (30.57, 32.06)	31.16, (30.47, 31.85)	31.05, (30.41, 31.69)	31.03, (30.44, 31.62)
B	17.37, (17.09, 17.65)	17.37, (17.11, 17.63)	17.33, (17.08, 17.57)	17.28, (17.05, 17.51)	17.26, (17.04, 17.47)
C	5.59, (5.43, 5.75)	5.61, (5.46, 5.75)	5.58, (5.45, 5.71)	5.51, (5.38, 5.64)	5.49, (5.37, 5.61)
D	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)
E	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
F	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
G	0.01, (0.01, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)
H	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)	0.01, (0.0, 0.01)
I	22.76, (22.07, 23.46)	22.89, (22.21, 23.56)	22.75, (22.13, 23.38)	22.64, (22.06, 23.22)	22.62, (22.09, 23.15)
J	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)	0.01, (0.01, 0.01)
K	0.02, (0.01, 0.02)	0.01, (0.01, 0.02)	0.01, (0.01, 0.02)	0.01, (0.01, 0.02)	0.01, (0.01, 0.01)
L	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)	0.0, (0.0, 0.0)
M	0.61, (0.59, 0.64)	0.61, (0.59, 0.64)	0.61, (0.59, 0.63)	0.61, (0.59, 0.63)	0.6, (0.58, 0.62)

4.5.4 Automated Capacity

When considering how to automate capacity it is important to acknowledge that there are many factors that are linked to capacity i.e. staff schedules, shared resources etc., making it unrealistic to have different levels of capacity every single day. Furthermore, it is typical to encounter activities that happen on a specific day of the week (every week). Therefore it was decided that capacity would be defined as a weekly schedule, where the amount of capacity each named day (Monday, Tuesday, etc) could be different if necessary.

As the records for when an activity was performed are logged as a date stamp, it is possible to extract what day of the week each date occurred using Python datetime library [77]² The average number of times an activity was performed for each named day can then be calculated.

For example, for the working dataset for lung cancer, the ‘Pattern’ column in Table 4.13 shows the capacity extracted using this method, where the first value is the weekly total followed by a list of the named day capacity, where the first entry corresponds to Monday etc. In this case, the capacity was quite often different for each named day. This could be due to the method only being able to extract the capacity that was used, which may not be the same as what was available.

Although the ‘Pattern’ shows the capacity as often different every named day, this might be difficult to implement in reality. Alternatively, taking an average of the capacity across the named days could be considered to allow for a consistent value on every day in the week. The average is calculated and the integer value taken for each activity, as is shown in the ‘Average’ column of Table 4.13.

However, the ‘Average’ method no longer has the same total amount of capacity per week due to rounding. A solution is to take the total capacity per week divided by the number of days in the week (7 in this case), using the integer value as the base

²Dates are converted into a string in the format: named day, week number of the year, year (in Python `strftime('%A, %U, %y')` - see [77] for details).

consistent capacity, and then adding any remainder in increments of one to each day starting at Monday, as shown in the ‘Smoothed’ column of Table 4.13. This allows for a more consistent spread without loss of capacity.

It should also be noted that some activities also have capacity recorded on the weekend. However, it is more typical for a 5 day working week. Therefore, each of these capacity calculation methods can be investigated in the simulation, where the arrivals and service levels are as set on page 106. The top level results and the activity waiting time results can be seen in Table 4.14 and Table 4.15 for seven working days, and Table 4.17 and 4.18 five working days a week (with five days capacity in Table 4.16) respectively.

Comparing the results between the seven and five day experiments, the top level results (Table 4.14 and Table 4.17) show that the ‘Pattern’ and ‘Smoothed’ mean time in system increases by approximately 8 days, which is expected when the capacity is reduced slightly. Furthermore, the results show that the ‘Average’ method performs quite different to the other two methods. This is likely due to the lower amount of capacity each week which can be evidenced by the long waiting times for activities ‘E’, ‘J’ and ‘L’ where the capacity was much lower using the average method. This suggests that the ‘Average’ method is not suitable.

In conclusion, the ‘Pattern’ method will be used as the automated capacity, as this will be able to detect if activities only run on certain days (as previously discussed).

Additionally, for the particular working example, considering that the *Raw Pathways* model is suggested to be used for validation for the input parameters, it is noticeable that the reported results here are not a good reflection of those seen in the original data (on page 103). Therefore, as the ‘Pattern’ model is using the capacity levels extracted from the data, then there must be another factor which needs to be considered. This could be due to the time it takes for initial queues to build as the simulation starts from an empty system. The idea of setting a warm up period will be explored in subsection 4.5.5 as a means to address this.

Table 4.13: Automated Seven Days Capacity Patterns.

Activity	Pattern	Average	Smoothed
A	38, (6, 7, 10, 6, 6, 1, 2)	35, (5, 5, 5, 5, 5, 5, 5)	38, (6, 6, 6, 5, 5, 5, 5)
B	37, (6, 7, 8, 7, 7, 1, 1)	35, (5, 5, 5, 5, 5, 5, 5)	37, (6, 6, 5, 5, 5, 5, 5)
C	38, (5, 7, 9, 16, 1, 0, 0)	35, (5, 5, 5, 5, 5, 5, 5)	38, (6, 6, 6, 5, 5, 5, 5)
D	10, (2, 2, 1, 2, 2, 0, 1)	7, (1, 1, 1, 1, 1, 1, 1)	10, (2, 2, 2, 1, 1, 1, 1)
E	13, (2, 3, 2, 3, 2, 1, 0)	7, (1, 1, 1, 1, 1, 1, 1)	13, (2, 2, 2, 2, 2, 2, 1)
F	12, (2, 2, 2, 2, 3, 1, 0)	7, (1, 1, 1, 1, 1, 1, 1)	12, (2, 2, 2, 2, 2, 1, 1)
G	15, (2, 3, 2, 3, 3, 1, 1)	14, (2, 2, 2, 2, 2, 2, 2)	15, (3, 2, 2, 2, 2, 2, 2)
H	15, (5, 2, 2, 2, 2, 1, 1)	14, (2, 2, 2, 2, 2, 2, 2)	15, (3, 2, 2, 2, 2, 2, 2)
I	40, (7, 7, 7, 7, 8, 2, 2)	35, (5, 5, 5, 5, 5, 5, 5)	40, (6, 6, 6, 6, 6, 5, 5)
J	14, (2, 4, 2, 3, 2, 1, 0)	14, (2, 2, 2, 2, 2, 2, 2)	14, (2, 2, 2, 2, 2, 2, 2)
K	15, (2, 4, 4, 3, 2, 0, 0)	14, (2, 2, 2, 2, 2, 2, 2)	15, (3, 2, 2, 2, 2, 2, 2)
L	10, (2, 2, 2, 2, 2, 0, 0)	7, (1, 1, 1, 1, 1, 1, 1)	10, (2, 2, 2, 1, 1, 1, 1)
M	33, (5, 7, 9, 5, 5, 1, 1)	28, (4, 4, 4, 4, 4, 4, 4)	33, (5, 5, 5, 5, 5, 4, 4)

Table 4.14: Top level Results of Seven Days Capacity Method.

Method	Mean TiS	Median TiS	Target	Overall Period
Pattern	21.0 (20.04, 21.96)	17.0 (15.53, 18.47)	98.08 (97.68, 98.49)	376.68 (375.74, 377.62)
Average	64.0 (62.16, 65.84)	47.0 (44.4, 49.6)	67.87 (65.99, 69.74)	479.2 (478.77, 479.63)
Smoothed	20.0 (18.76, 21.24)	17.0 (15.04, 18.96)	98.51 (98.14, 98.87)	376.53 (375.53, 377.53)

Table 4.15: Activity Waiting Time Results of Seven Days Capacity Method.

Activity	Pattern	Average	Smoothed
A	0.92 (0.84, 0.99)	2.81 (2.31, 3.31)	0.73 (0.63, 0.83)
B	2.6 (2.11, 3.09)	6.76 (6.2, 7.32)	2.99 (2.35, 3.63)
C	3.82 (3.58, 4.06)	3.21 (3.11, 3.31)	2.9 (2.57, 3.24)
D	0.69 (0.65, 0.74)	1.1 (1.02, 1.19)	0.64 (0.59, 0.69)
E	4.76 (4.59, 4.93)	88.19 (87.89, 88.48)	4.87 (4.7, 5.04)
F	4.1 (3.99, 4.2)	3.03 (2.91, 3.15)	4.21 (4.04, 4.37)
G	0.59 (0.56, 0.62)	0.27 (0.25, 0.3)	0.48 (0.46, 0.49)
H	0.8 (0.77, 0.83)	0.16 (0.15, 0.17)	0.38 (0.36, 0.41)
I	0.34 (0.31, 0.37)	1.98 (1.54, 2.42)	0.29 (0.25, 0.32)
J	8.8 (8.29, 9.31)	1.21 (1.15, 1.27)	8.75 (8.27, 9.24)
K	18.88 (18.1, 19.65)	19.78 (19.38, 20.17)	17.54 (16.83, 18.25)
L	10.21 (9.73, 10.69)	46.86 (46.55, 47.18)	10.56 (10.0, 11.12)
M	1.06 (0.98, 1.14)	12.79 (12.02, 13.56)	1.09 (0.97, 1.2)

Table 4.16: Automated Five Days Capacity Patterns.

Activity	Pattern	Average	Smoothed
A	35, (6, 7, 10, 6, 6)	35, (7, 7, 7, 7, 7)	35, (7, 7, 7, 7, 7)
B	35, (6, 7, 8, 7, 7)	35, (7, 7, 7, 7, 7)	35, (7, 7, 7, 7, 7)
C	38, (5, 7, 9, 16, 1)	35, (7, 7, 7, 7, 7)	38, (8, 8, 8, 7, 7)
D	9, (2, 2, 1, 2, 2)	5, (1, 1, 1, 1, 1)	9, (2, 2, 2, 2, 1)
E	12, (2, 3, 2, 3, 2)	10, (2, 2, 2, 2, 2)	12, (3, 3, 2, 2, 2)
F	11, (2, 2, 2, 2, 3)	10, (2, 2, 2, 2, 2)	11, (3, 2, 2, 2, 2)
G	13, (2, 3, 2, 3, 3)	10, (2, 2, 2, 2, 2)	13, (3, 3, 3, 2, 2)
H	13, (5, 2, 2, 2, 2)	10, (2, 2, 2, 2, 2)	13, (3, 3, 3, 2, 2)
I	36, (7, 7, 7, 7, 8)	35, (7, 7, 7, 7, 7)	36, (8, 7, 7, 7, 7)
J	13, (2, 4, 2, 3, 2)	10, (2, 2, 2, 2, 2)	13, (3, 3, 3, 2, 2)
K	15, (2, 4, 4, 3, 2)	15, (3, 3, 3, 3, 3)	15, (3, 3, 3, 3, 3)
L	10, (2, 2, 2, 2, 2)	10, (2, 2, 2, 2, 2)	10, (2, 2, 2, 2, 2)
M	31, (5, 7, 9, 5, 5)	30, (6, 6, 6, 6, 6)	31, (7, 6, 6, 6, 6)

Table 4.17: Top level Results of Five Days Capacity Method.

Method	Mean TiS	Median TiS	Target	Overall Period
Pattern	29.0 (27.48, 30.52)	27.0 (24.85, 29.15)	93.73 (92.72, 94.74)	386.68 (385.73, 387.63)
Average	48.0 (46.12, 49.88)	37.0 (34.65, 39.35)	73.67 (72.64, 74.7)	411.02 (410.11, 411.93)
Smoothed	28.0 (26.04, 29.96)	25.0 (22.2, 27.8)	94.04 (92.8, 95.27)	385.32 (384.06, 386.59)

Table 4.18: Activity Waiting Time Results of Five Days Capacity Method.

Activity	Pattern	Average	Smoothed
A	3.4 (2.7, 4.1)	2.97 (2.31, 3.63)	3.32 (2.44, 4.21)
B	7.2 (6.58, 7.81)	7.96 (7.2, 8.71)	6.6 (5.91, 7.29)
C	1.66 (1.57, 1.74)	2.64 (2.55, 2.74)	0.92 (0.83, 1.01)
D	0.71 (0.67, 0.76)	9.59 (9.03, 10.15)	0.74 (0.69, 0.78)
E	6.79 (6.56, 7.03)	20.79 (20.35, 21.22)	7.13 (6.89, 7.37)
F	7.55 (7.24, 7.87)	3.16 (3.06, 3.26)	7.94 (7.6, 8.29)
G	1.25 (1.17, 1.33)	13.31 (13.14, 13.47)	1.33 (1.24, 1.42)
H	1.25 (1.19, 1.31)	1.04 (0.99, 1.09)	0.81 (0.77, 0.85)
I	1.43 (1.27, 1.6)	2.27 (1.73, 2.8)	1.39 (1.13, 1.65)
J	14.92 (14.61, 15.23)	51.21 (50.85, 51.57)	14.93 (14.58, 15.27)
K	14.93 (14.35, 15.5)	13.1 (12.52, 13.68)	15.24 (14.73, 15.76)
L	6.38 (6.08, 6.69)	3.12 (3.05, 3.2)	6.36 (6.08, 6.64)
M	2.26 (2.08, 2.44)	3.21 (2.87, 3.55)	2.19 (1.98, 2.4)

4.5.5 Warm Up

There is a lot of discussion considering how long to set the warm up period and what method to use, where Hoad et al., [123] found 44 warm up methods through a literature search. Furthermore, Hoad et al., [123] investigates “automating warm-up length estimator” i.e. how long to run the warm up. For this model, a warm up time is the time where the simulation will run without collecting results, to allow the simulation to ‘fill up’ before observing the system [61]. This is done to ensure that collecting results from an empty system does not bias the results [73]. For more information on warm up see [169].

Addressing the question of how long to run the warm up for is restricted when considering the *Raw Pathways* requirement to keep the pathways in the same order (see subsection 4.5.2). Therefore, the investigation explores the idea to repeat the pathways in the exact same order for an integer (w) number of times, and then selecting the last segment of exactly I individuals. This will be referred to as the Iterative approach.

After initially investigating the results of applying the Iterative approach for the working example of lung cancer (discussed further below), the waiting times for the activities did not satisfactorily reflect the original data. This is because some of the activities had very long waiting times in the original data which just appeared to be unachievable using this method.

A second method was considered where, for each activity service was blocked for an initial set number of days. Therefore, even though individuals continue to enter the system they cannot undertake service at activities until the capacity becomes unblocked. The motivation for this was to consider that some activities have a long waiting list and could be the cause of these unaligned waiting times. The automated method takes the ceiling of the mean waiting time for each activity (as in Table 4.4) as the number of days to block capacity. This method shall be referred to as warm

start going forward.

To investigate these warm up methods, both seven and five working days a week, where the arrivals and service levels are as set on page 106, and the automated capacity ‘Pattern’ from Table 4.13 and Table 4.16 respectively.

The investigation firstly considered the Iterative approach with a total of two iterations (Iter 2) and then three iterations (Iter 3) meaning that there was one round and two rounds before collecting results respectively. Secondly considering the Warm Start approach with the automated number of days blocked (previously described), and also doubling the number of days blocked³. These will be referred to as Warm and Warm 2 respectively. The top level and activity waiting time results are displayed in Table 4.19 and Table 4.21 respectively for seven days a week and Table 4.20 and Table 4.22 for five days a week. The results for no warm up (None) are also included for comparison.

Table 4.19: Top Level Results for Warm Up Comparisons, 7 days.

Type	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
None	21.16 (20.15, 22.17)	17.8 (16.29, 19.31)	98.08 (97.68, 98.49)	376.47 (375.54, 377.39)
Iter 2	27.76 (25.19, 30.33)	24.16 (21.39, 26.93)	94.67 (92.67, 96.68)	379.94 (378.76, 381.12)
Iter 3	29.88 (26.7, 33.06)	24.8 (21.71, 27.89)	92.05 (89.36, 94.73)	383.07 (380.89, 385.25)
Warm	33.56 (31.79, 35.33)	28.64 (26.64, 30.64)	90.56 (88.87, 92.25)	384.13 (383.39, 384.86)
Warm 2	51.88 (50.11, 53.65)	41.64 (39.64, 43.64)	66.69 (65.33, 68.05)	394.99 (394.91, 395.07)

Table 4.20: Top Level Results for Warm Up Comparisons, 5 days.

Type	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
None	28.56 (27.1, 30.02)	25.52 (23.64, 27.4)	94.31 (93.4, 95.22)	387.16 (385.93, 388.38)
Iter 2	50.12 (47.59, 52.65)	45.92 (43.26, 48.58)	74.36 (70.6, 78.12)	399.7 (397.5, 401.91)
Iter 3	63.8 (60.42, 67.18)	58.4 (54.65, 62.15)	52.95 (45.94, 59.97)	414.84 (412.02, 417.65)
Warm	42.88 (41.02, 44.74)	37.88 (35.4, 40.36)	80.83 (79.32, 82.33)	394.02 (393.86, 394.19)
Warm 2	65.08 (63.34, 66.82)	56.72 (54.4, 59.04)	53.12 (50.34, 55.9)	407.66 (407.46, 407.86)

³Number of days blocked for the respective activities were: Warm = [0, 13, 12, 10, 22, 12, 21, 7, 41, 5, 14, 4, 4, 4] and Warm 2 = [0, 26, 24, 20, 44, 24, 42, 14, 82, 10, 28, 8, 8, 8], where the initial 0 is for the Dummy Arrival node.

Table 4.21: Activity Waiting Time Results for Warm Up Comparisons, 7 days.

Activity	Original	None	Iter 2	Iter 3	Warm	Warm 2
A	12.52	0.92 (0.84, 0.99)	1.12 (0.98, 1.26)	1.03 (0.92, 1.15)	2.99 (2.39, 3.6)	10.89 (9.8, 11.98)
B	11.86	2.6 (2.11, 3.09)	4.64 (3.7, 5.58)	4.8 (3.33, 6.28)	4.65 (4.01, 5.3)	7.07 (6.43, 7.71)
C	9.40	3.82 (3.58, 4.06)	5.04 (4.51, 5.57)	5.65 (4.95, 6.34)	5.59 (5.24, 5.93)	6.21 (6.11, 6.31)
D	21.91	0.69 (0.65, 0.74)	0.72 (0.66, 0.78)	0.78 (0.71, 0.85)	0.72 (0.67, 0.76)	0.7 (0.67, 0.73)
E	11.19	4.76 (4.59, 4.93)	6.02 (5.76, 6.28)	6.37 (6.06, 6.67)	5.12 (5.02, 5.21)	5.31 (5.19, 5.42)
F	20.20	4.1 (3.99, 4.2)	6.93 (6.29, 7.57)	8.53 (7.78, 9.28)	9.56 (9.39, 9.73)	15.21 (15.07, 15.35)
G	6.41	0.59 (0.56, 0.62)	0.71 (0.66, 0.76)	0.78 (0.73, 0.83)	0.61 (0.58, 0.64)	0.6 (0.57, 0.63)
H	40.21	0.8 (0.77, 0.83)	0.86 (0.82, 0.89)	0.88 (0.85, 0.92)	5.68 (5.62, 5.73)	24.62 (24.52, 24.72)
I	4.05	0.34 (0.31, 0.37)	0.66 (0.45, 0.88)	0.8 (0.58, 1.02)	0.5 (0.45, 0.56)	0.63 (0.58, 0.68)
J	13.66	8.8 (8.29, 9.31)	9.6 (8.85, 10.34)	10.33 (9.57, 11.09)	10.95 (10.55, 11.36)	10.69 (10.58, 10.8)
K	3.58	18.88 (18.1, 19.65)	21.09 (19.69, 22.49)	22.37 (20.83, 23.92)	24.13 (23.61, 24.65)	24.1 (23.92, 24.27)
L	3.85	10.21 (9.73, 10.69)	11.22 (10.28, 12.17)	12.24 (11.15, 13.33)	12.14 (11.63, 12.66)	11.0 (10.85, 11.15)
M	3.20	1.06 (0.98, 1.14)	1.56 (1.27, 1.86)	1.59 (1.36, 1.83)	1.98 (1.72, 2.23)	3.1 (2.99, 3.22)

Table 4.22: Activity Waiting Time Results for Warm Up Comparisons, 5 days.

Activity	Original	None	Iter 2	Iter 3	Warm	Warm 2
A	12.52	2.51 (2.1, 2.92)	5.31 (4.54, 6.09)	5.62 (4.72, 6.53)	8.82 (7.55, 10.09)	23.25 (21.61, 24.9)
B	11.86	6.72 (6.18, 7.26)	14.1 (12.91, 15.29)	23.54 (21.65, 25.44)	8.44 (7.88, 9.01)	9.64 (9.39, 9.88)
C	9.40	1.69 (1.59, 1.79)	3.42 (3.28, 3.56)	3.99 (3.88, 4.11)	2.09 (2.01, 2.17)	1.97 (1.9, 2.05)
D	21.91	0.72 (0.67, 0.78)	0.64 (0.6, 0.68)	0.65 (0.6, 0.71)	0.73 (0.68, 0.78)	0.73 (0.69, 0.77)
E	11.19	6.72 (6.48, 6.96)	6.77 (6.61, 6.93)	6.45 (6.22, 6.68)	7.21 (7.05, 7.37)	7.53 (7.41, 7.64)
F	20.20	7.41 (7.14, 7.69)	9.91 (9.72, 10.1)	10.55 (10.34, 10.76)	13.15 (13.0, 13.29)	18.29 (18.11, 18.47)
G	6.41	1.21 (1.13, 1.29)	1.6 (1.51, 1.7)	1.53 (1.43, 1.64)	1.52 (1.45, 1.6)	1.42 (1.36, 1.48)
H	40.21	1.31 (1.25, 1.37)	1.23 (1.19, 1.27)	1.19 (1.14, 1.23)	9.4 (9.29, 9.5)	34.04 (33.89, 34.19)
I	4.05	1.1 (0.97, 1.23)	6.9 (6.0, 7.79)	10.27 (9.1, 11.44)	1.62 (1.42, 1.83)	1.51 (1.38, 1.65)
J	13.66	15.0 (14.68, 15.33)	16.8 (16.58, 17.02)	17.31 (17.08, 17.54)	15.49 (15.32, 15.67)	14.02 (13.83, 14.21)
K	3.58	14.64 (13.93, 15.36)	19.96 (19.67, 20.25)	19.53 (19.23, 19.83)	18.46 (18.23, 18.69)	16.58 (16.38, 16.77)
L	3.85	6.62 (6.38, 6.86)	9.35 (9.01, 9.7)	8.75 (8.52, 8.97)	6.36 (6.17, 6.54)	5.36 (5.19, 5.52)
M	3.20	2.33 (2.13, 2.53)	3.33 (3.05, 3.61)	3.58 (3.09, 4.07)	2.9 (2.78, 3.02)	3.02 (2.93, 3.11)

Comparing the results for seven and five days a week, the warm up (regardless of method) has a greater effect on the five days a week than seven. Considering the top level results, Warm 2 produces results closest to the original data in mean, median and target. Despite the top level results improving, it is still difficult to obtain accurate waiting times for some activities. For activity ‘H’, Warm 2 is clearly the most accurate method, but still not extremely close in the seven days a week results. Conversely, the wait time for activity ‘D’ does not move closer to the original.

In conclusion, no warm up will be automatically applied.

4.6 Custom Parameters

As discussed in Chapter 1 it was key to ensure that, to support the automation, the end decisions are with the user. Therefore, this section describes the flexibility surrounding the input parameters that will be included in Sim.Pro.Flow by selecting to use the ‘Custom’ option (see Appendix Figure D.3 for implementation).

Pathways

The supported routing procedures should be used as the method for exploring different pathways. However, specifically considering the *Raw Pathways*, as the pathways are sampled from the data in the exact order that they appear, it is possible for the user to change the ordering in the input data before loading it into Sim.Pro.Flow.

Arrivals

The situation may arise when a user wants to explore the results of an increased demand e.g. would the capacity be sufficient if a 10% increase in demand occurs? To observe this, the number of individuals simulated and the corresponding arrival rate will need to be adjusted⁴.

⁴This feature is not available in Sim.Pro.Flow v2.1 for the *Raw Pathways* but is performed manually for the Demand Investigation (Section 7.7) and is listed as further work.

Service

As both the deterministic situation and exponential distributed service durations are explored in subsection 4.5.3, these are both supported where the user can specify the value of μ for each activity node.

Capacity

For the automated capacity, the ‘Pattern’ method will be calculated and presented to the user. For the custom options, the ‘Smoothed’ method discussed in subsection 4.5.4 will also be available to select. Furthermore, the capacity for each named day for each activity can be explicitly defined if the user knows the capacity values, or wishes to explore alternative values.

It was noted in subsection 4.5.4 that the automatically extracted capacity can only observe the capacity that was used. Therefore, an alternative method of defining capacity will be explored in Section 4.7, which will be available to select as a ‘Custom’ parameter.

Warm up

It was concluded that no warm up would automatically be applied to the simulation, however both methods (Iterative and Warm Start) are available in Sim.Pro.Flow, which each have an automated or custom option.

- Iterative: The custom option to choose how many iterations (w) of I individuals to run, where the automated value is $w = 2$ (one iteration before collecting results of the second).
- Warm Start: The user can choose how many days to block each activity, where the automated value is the ceiling of the average waiting time for that activity.

4.7 Capacity Investigation

The previous method of calculating capacity uses a relatively simple calculation. This method can only reflect what has previously happened in the data. It would be more beneficial to consider the demand and suggest what the capacity should be to achieve desired targets. Arruda et al., [12] describes a method which considers the aggregate demand and suggests appropriate capacity based on desired percentage time targets. It is suggested that this method be applied more for a scenario testing instance than for validation. For note, the notation used in this section is as in Glossary Capacity Table.

The general idea is to consider capacity as a perishable inventory - once the inventory has expired it cannot be used. The calculation is split into two parts. Part one considers demand through the number of individuals, and the number of times they perform the activity in consideration. These are combined to form the aggregated demand for that activity and the expected number of arrivals in a day. Part two considers the amount of capacity required to satisfy the demand, considering the daily amount of capacity and how often (α) this would need to increase by one to satisfy the demand. The method makes use of Markov Chains and Steady State to test various values for α and returns the probability for number of queued tests. Finally, a simple conversion allows for the number of queued tests to be converted into the percentage of number of wait days. This allows decision makers to consider what percentage of individuals they require to be seen within a target number of days and select the corresponding capacity value.

The following discussion replicates the main case study in Arruda et al., [12] for Python (see Section 1.5). The case study considers C.T. Scan as the activity, which was performed 393 times by 341 individuals, over a 26 week period of 5 days a week.

Part One

Part one firstly calculates the probability distribution for number of individuals arriving into the system each day (P). The Poisson probability mass function (pmf) is used with arrival rate (λ) and maximum number arriving. The arrival rate λ is calculated considering the total number of individuals arriving a day (regardless of activity). The Poisson percent point function (ppf) is used with λ and 0.99 to obtain the maximum number of individuals a day. To satisfy that the distribution sums to 1, any remaining probability is crudely added on to that for the maximum value. Figure 4.6 shows the distribution of incoming referrals per day, with a maximum of 7 and average rate of 2.623.

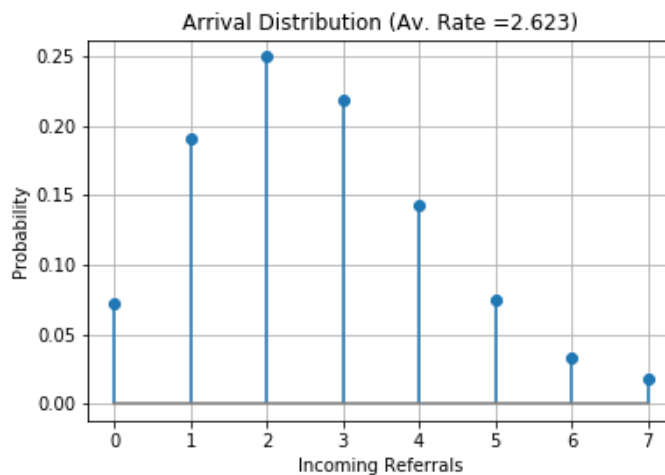


Figure 4.6: Probability of Individuals per Day.

Secondly, consider how many times each individuals perform the same activity. This can be extracted from the data where the probability that an individual will need one test is calculated by the number of individuals having one test/total number of individuals, and the same for two tests etc. Arruda et al., [12] stated that 52 individuals had two tests ($393 - 341 = 52$) and thus the probability of one test and two tests are $P(1) = 0.848$ and $P(2) = 0.152$ respectively.

To calculate the aggregated demand, Arruda et al., theoretically describes the method and applies Equation 2.3 (from [12]), for which the practical execution is described below.

The aim is, for each number of incoming individuals, calculate the probability of the number of tests being required. For example, if two individuals arrive, what is the probability that they would each need 1 test each (total of 2), one need 1 test and the other need 2 (total of 3) or both need two tests (total of 4). There are some obvious statements to make, including if 0 individuals arrive then there is 100% probability of 0 tests required. Furthermore, the bounds of tests to consider can be calculated by multiplying the bound of number of tests (n) with the bound of incoming individuals (A) i.e. $[A_{LB} * n_{LB}, A_{UB} * n_{UB}]$, which for this example is $[1 * 0, 2 * 7] = [0, 14]$.

To obtain this distribution the binomial expansion of $(x+y)^A$, where A is the number of incoming individuals and substituting x and y for $P(1)$ and $P(2)$ respectively. Therefore the first segment of the equation would equate to a total of one test, the second a total of two tests etc. For clarification, the matrix for $P(1) = 0.848$ and $P(2) = 0.152$ is shown below truncated at $A = 4$, and the calculation for $A = 4$ has been explicitly demonstrated.

Matrix showing probability of total number of tests per number of individuals

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \dots \\ 0 & 0.847507 & 0.152493 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \dots \\ 0 & 0 & 0.718269 & 0.258477 & 0.023254 & 0 & 0 & 0 & 0 & 0 \dots \\ 0 & 0 & 0 & 0.608738 & 0.328592 & 0.059124 & 0.003546 & 0 & 0 & 0 \dots \\ 0 & 0 & 0 & 0 & 0.515910 & 0.371312 & 0.100216 & 0.012021 & 0.000541 & 0 \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots \end{pmatrix}$$

$$\text{Binomial expansion of } (x + y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

Taking each element and substituting $x = P(1) = 0.848$ and $y = P(2) = 0.152$

$$(0.848^4, 4(0.848^3 * 0.152), 6(0.848^2 * 0.152^2), 4(0.848 * 0.152^3), 0.152^4)$$

$$\Rightarrow (0.515910, 0.371312, 0.100216, 0.012021, 0.000541)$$

The final step is to multiply the probability distribution P by the transpose of the matrix. This will result in a row vector containing the probability per possible number of test (P_n) as shown in Figure 4.7. Taking the expected value of the distribution will give the average rate of requests, which can be used as a basis for capacity.

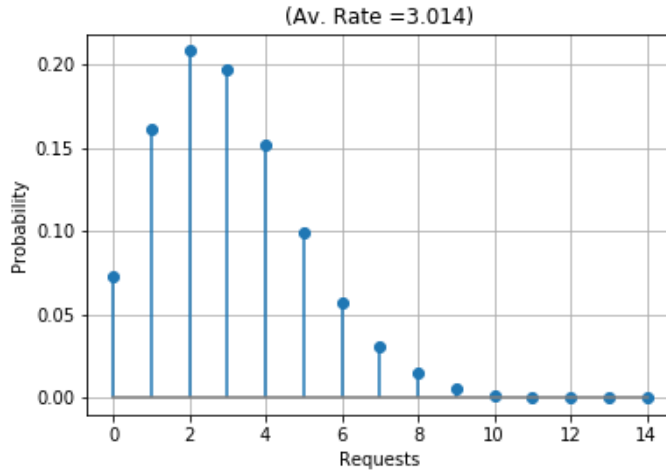


Figure 4.7: Probability of Number of Test Requests.

Furthermore, taking the cumulative distribution of Figure 4.7 allows to suggest the capacity based on desired percentage service rate. For example, in Figure 4.8 to achieve a service rate of greater than 80%, a capacity of five is required.

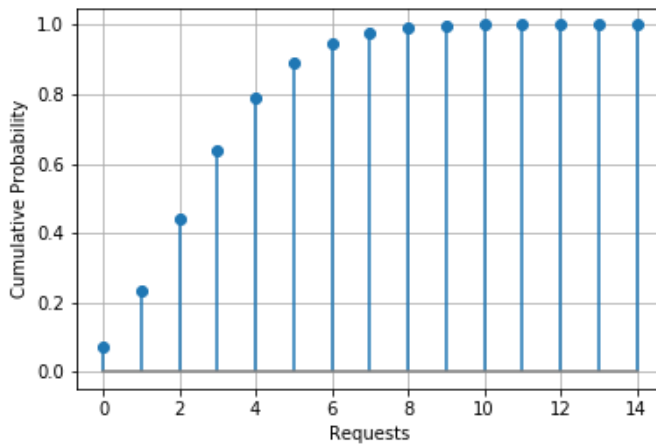


Figure 4.8: Cumulative Probability of Number of Test Requests.

Part Two

Part two describes a system using a Markov chain with the assumption that the demand each day will be the current demand, plus the demand left over from the previous day. There is also an assumption that all requests wait one period to be processed. The steady state behaviour of this system is then investigated by finding the limiting distribution, i.e. find the probability distribution that will be the same today, as tomorrow, as the next day and so on. The method is discussed below, however for technical information see Arruda et al., [12].

The Markov chain system which Arruda et al., [12] calls P^C is limited to a 1000 square matrix, with rows representing how many requests for tests there are today, and columns representing that for tomorrow. In all rows, r , where r is less than or equal to the capacity (C) all requests will be able to be processed, and thus these rows are occupied by P_n followed by trailing 0's i.e. first value of P_n in space $P^C(r, 0)$. For all rows, where r is greater than the capacity some test requests will carry over to tomorrow, and thus there is an initial shift in these rows of $r - C$ i.e. if $C = 3$ and $r = 4$, first value of P_n in space $P^C(4, 1)$.

The method investigates two values for capacity (C), namely the integer bounds of the expected value calculated from Figure 4.7 (which will be [3, 4] in this example). The lower bound will act as the standard capacity and the upper bound as extra capacity occurring with probability α . This gives the transition matrix

$$P^C = \alpha P^{UB} + (1 - \alpha) P^{LB} \quad (4.4)$$

where LB and UB are the lower and upper bounds respectively as described above.

Now having obtained the matrix we need to investigate the steady state of this system by solving the system (Equation 3.11 in [12])

$$\pi P^C = \pi,$$

$$\sum_{i=0}^{\infty} \pi(i) = 1$$

This is an iterative method where this system is continuously calculated until it achieves a solution within a tolerance, which is set to be $1e^{-5}$. A simplistic approach is taken for the the initial guess for π which is a row vector of length 1000 with each values as 0.001. The solution can be plotted (Figure 4.9) for each value of α , where the values of α investigated (referred to as p sequence) were, pseq = [1, 0.7, 0.5, 0.25, 0.05, 0.142].

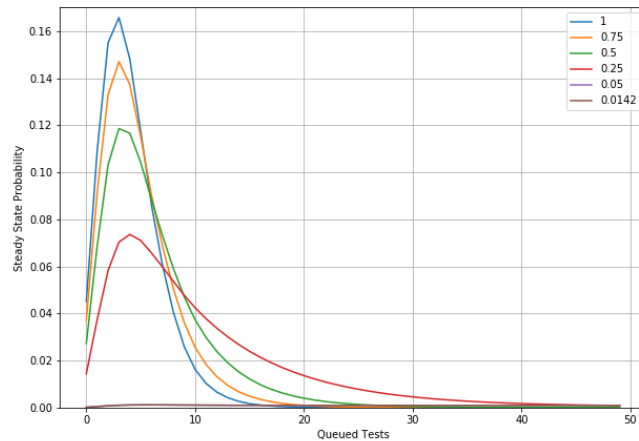


Figure 4.9: Steady State probability of Queued Number of Test Requests.

The final step is to convert the queued test into wait days so that a value of α can be selected based on a desired level of waiting time, where this is in terms of waiting more than t days.

To do this, for each day (t), firstly take 1 - the cumulative sum of the solution as plotted in Figure 4.9. Then for each day calculate the ceiling of t/C to get the number of wait days. This will produce multiple probabilities for each waiting day. Taking the minimum of the values for each waiting day (as this represents waiting more than t days), a plot can be produced as shown in Figure 4.10

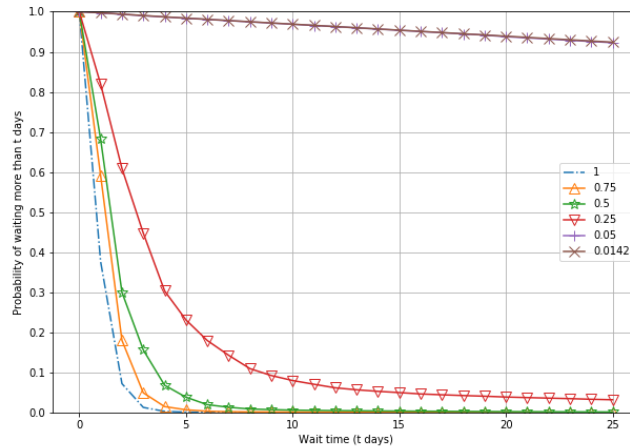


Figure 4.10: Probability of Waiting Time in Days.

Interpreting Figure 4.10, it can be said that if no more than 10% of individuals should wait more than 10 days then a value of $\alpha = 0.25$ will suffice. This means that one in every four times that this service is provided there should be a capacity of four and the remaining time should have a capacity of three. Furthermore, if the target was no more than 10% of individuals should wait 5 days then a value of $\alpha = 0.5$ would be necessary, meaning the capacity would need to be 4 50% of the time and 3 the remaining 50%.

4.7.1 Method Extension

After publication Dr Arruda considered that presenting the amount of capacity required each week is more beneficial than presenting the amount per day. This allows for larger increments between values of α considered and greater flexibility overall. To make a clear distinction, pseq refers to values of α for daily capacity and PSEQ for weekly capacity. The method extension is described below.

Firstly, calculating the lower bound for the weekly capacity values to test using Equation 4.5

$$LB = \text{ceil} \left(D \frac{E}{0.99} \right) \quad (4.5)$$

where D is the number of days in the week considered and E is the expected value for number of arrivals per day. The division by 0.99 ensures enough capacity is being considered by increasing the expected value slightly.

The weekly sequence is then calculated by incrementing from the lower bound (LB) by the specified amount (h) and number of times (UB) as in Equation 4.6

$$\text{PSEQ} = [LB + bh \mid b \in [0, UB)] \quad (4.6)$$

i.e. to investigate 10 increments of 5 (capacity), $h = 5, UB = 10$

To perform the calculation, these need to be decreased back down to daily amount by dividing by D and rounding to two decimal places, as in Equation 4.7

$$\text{pseq} = [\text{round} \left(\frac{p}{D}, 2 \right), \forall p \text{ in PSEQ}]. \quad (4.7)$$

The calculation then proceeds as previously using the pseq values, however the results are reported in terms of the corresponding PSEQ value. The previously discussed smoothed technique to obtain the daily capacity pattern is then applied.

Careful consideration is required when calculating the input arrival rate. Contrary to the arrival rate into the system (as discussed surrounding Equation 4.2), here the arrival rate into the activity is being calculated. Therefore, we do need to account for the number of days worked in a week. For example, considering applying Equation 4.2 for the working lung cancer example, with the total number of individuals (1865) and total number of days (362). This gives a daily demand of approximately 5.15. To get weekly demand we multiply the result by the number of working days, which would give a demand of approximately 36.05 and 25.75 individuals a week for seven and five working days respectively. This is not correct as our demand should be the same. Therefore consider the adjustment in Equation 4.8

$$\lambda = \frac{I}{T \left(\frac{D}{7} \right)} \quad (4.8)$$

where I is the total number of individuals, T is the time period and D is the number of working days. Applying seven and five working days to Equation 4.8 gives a daily demand of 5.15 and 7.21 and a weekly demand of 36.05 and 36.05 respectively. This ensures our demand across the week is the same.

The above method has been integrated into Sim.Pro.Flow (see Figure D.5 in Appendix D) input information, such as arrival rate and probability of number of tests etc, is extracted from the data. The GUI also allows users to select the percentage for the target and the target days itself, how many increments for the p sequence and the amount to increment (h), along with the number of wait days to plot and whether to run the calculation for the activity or not. For example, for activity ‘A’ we wish to investigate 90% of individuals seen within 4 days with 5 increments of 5 and plotted over 5 days would be represented as A: [90, 4, 5, 5, 5, ‘Yes’], where ‘Yes’ or ‘No’ are the options to specify if to perform the calculation or not. Furthermore, in Sim.Pro.Flow the values for target days are automatically extracted from the mean waiting time for each activity (as in Table 4.4). This value plus an additional 5 (to allow for spacing) is used as the automated values for the wait days to plot.

4.7.2 Example

Applying the above method to achieve the target 90% of TiS within 62 days. Four scenarios were investigated by varying the days in a week, the step amount and number of increments as shown in Table 4.23.

Table 4.23: Scenarios for Capacity Investigation.

Scenario	Days	h	UB
1	7	5	5
2	5	5	5
3	7	10	2
4	5	10	2

Investigating the increment amount allows for examination of how the results change as a finer level of detail is considered. The four scenarios generate solutions of total weekly capacity and patterns as shown in Table 4.24. Furthermore, the line plot similar to that seen in Figure 4.10 can be produced for each activity and each scenario. An example of this graph can be seen for scenario 4 can be seen in Appendix C (Figure C.3).

Note how as the increment amount decreased from 5 to 2, the capacity levels per week either stayed the same or decreased slightly. This decrease is due to values in between the original increments of 5 can be considered and thus produces a tighter distinction. Applying these capacity patterns to the *Raw Pathway* simulation (where arrivals and service times are from page 106 and without a warm up) gives the top level and activity waiting time results as in Table 4.25 and Table 4.26 respectively.

Comparing the top level results of seven and five days a week (Scenario 1 with 2 and 3 with 4) it is unsurprising that the seven days a week have a smaller average time in system in comparison to five days a week due to not pausing service due to the weekend. Considering the different increment amounts for the same number of days a week (Scenario 1 with 3 and 2 with 4), the smaller increments (Scenario 3 and 4) have longer mean time in system than those with the larger increments (Scenario 1 and 2) as there was occasionally less capacity available. Furthermore, the overall period is at most 22 days longer than that in the original data, which considering the various times in system between the scenarios is not too dissimilar.

Overall, the target values in Table 4.25 are getting close to 90%, which was the aim of the investigation, and thus this method can indicate what levels of capacity can achieve the target waiting times.

Table 4.24: Capacity Patterns for the Four Scenarios.

Activity	Scenario 1	Scenario 2	Scenario 3	Scenario 4
A	41 (6, 6, 6, 6, 6, 6, 5)	36 (8, 7, 7, 7, 7)	38 (6, 6, 6, 5, 5, 5, 5)	36 (8, 7, 7, 7, 7)
B	42 (6, 6, 6, 6, 6, 6, 6)	42 (9, 9, 8, 8, 8)	39 (6, 6, 6, 6, 5, 5, 5)	39 (8, 8, 8, 8, 7)
C	42 (6, 6, 6, 6, 6, 6, 6)	42 (9, 9, 8, 8, 8)	39 (6, 6, 6, 6, 5, 5, 5)	39 (8, 8, 8, 8, 7)
D	10 (2, 2, 2, 1, 1, 1, 1)	5 (1, 1, 1, 1, 1)	7 (1, 1, 1, 1, 1, 1, 1)	5 (1, 1, 1, 1, 1)
E	15 (3, 2, 2, 2, 2, 2, 2)	10 (2, 2, 2, 2, 2)	12 (2, 2, 2, 2, 2, 1, 1)	10 (2, 2, 2, 2, 2)
F	10 (2, 2, 2, 1, 1, 1, 1)	10 (2, 2, 2, 2, 2)	10 (2, 2, 2, 1, 1, 1, 1)	10 (2, 2, 2, 2, 2)
G	16 (3, 3, 2, 2, 2, 2, 2)	16 (4, 3, 3, 3, 3)	13 (2, 2, 2, 2, 2, 2, 1)	13 (3, 3, 3, 2, 2)
H	16 (3, 3, 2, 2, 2, 2, 2)	16 (4, 3, 3, 3, 3)	13 (2, 2, 2, 2, 2, 2, 1)	13 (3, 3, 3, 2, 2)
I	41 (6, 6, 6, 6, 6, 6, 5)	41 (9, 8, 8, 8, 8)	38 (6, 6, 6, 5, 5, 5, 5)	38 (8, 8, 8, 7, 7)
J	16 (3, 3, 2, 2, 2, 2, 2)	16 (4, 3, 3, 3, 3)	13 (2, 2, 2, 2, 2, 2, 1)	13 (3, 3, 3, 2, 2)
K	17 (3, 3, 3, 2, 2, 2, 2)	17 (4, 4, 3, 3, 3)	16 (3, 3, 2, 2, 2, 2, 2)	14 (3, 3, 3, 3, 2)
L	13 (2, 2, 2, 2, 2, 2, 1)	13 (3, 3, 3, 2, 2)	10 (2, 2, 2, 1, 1, 1, 1)	10 (2, 2, 2, 2, 2)
M	36 (6, 5, 5, 5, 5, 5, 5)	36 (8, 7, 7, 7, 7)	35 (5, 5, 5, 5, 5, 5, 5)	35 (7, 7, 7, 7, 7)

**Format: total, (pattern)*

Table 4.25: Top Level Results for the Four Capacity Scenarios.

Scenario	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
1	16.8 (16.17, 17.43)	3.96 (3.43, 4.49)	89.58 (89.12, 90.04)	363.19 (359.57, 366.81)
2	24.28 (23.39, 25.17)	6.88 (5.87, 7.89)	83.88 (83.31, 84.45)	377.03 (376.03, 378.03)
3	27.32 (26.19, 28.45)	15.6 (13.86, 17.34)	88.4 (87.33, 89.47)	370.04 (369.01, 371.07)
4	34.76 (33.61, 35.91)	22.76 (21.08, 24.44)	82.88 (82.27, 83.5)	383.8 (382.96, 384.63)

Table 4.26: Activity Waiting Time Results for the Four Capacity Scenarios.

Activity	Original	Scenario 1	Scenario 2	Scenario 3	Scenario 4
A	12.52	0.25(0.22, 0.27)	2.29(1.68, 2.9)	0.68(0.59, 0.77)	1.62(1.3, 1.94)
B	11.86	0.26(0.22, 0.3)	0.19(0.16, 0.22)	1.67(1.28, 2.06)	1.45(1.16, 1.74)
C	9.40	0.34(0.25, 0.43)	0.2(0.17, 0.23)	2.79(2.46, 3.13)	4.92(4.62, 5.22)
D	21.91	1.35(1.26, 1.43)	49.54(49.15, 49.93)	5.05(4.8, 5.31)	35.33(34.98, 35.68)
E	11.19	2.94(2.78, 3.1)	44.31(43.54, 45.08)	9.82(9.45, 10.19)	29.01(28.45, 29.56)
F	20.20	42.68(42.21, 43.16)	1.21(1.13, 1.28)	25.5(25.31, 25.69)	1.16(1.11, 1.22)
G	6.41	0.49(0.46, 0.52)	0.4(0.38, 0.42)	2.47(2.4, 2.54)	1.05(1.0, 1.11)
H	40.21	0.38(0.36, 0.4)	0.18(0.17, 0.2)	0.73(0.69, 0.77)	0.4(0.38, 0.43)
I	4.05	0.22(0.19, 0.24)	0.24(0.23, 0.26)	0.58(0.5, 0.65)	0.44(0.4, 0.49)
J	13.66	3.14(2.7, 3.57)	2.58(2.36, 2.8)	18.42(17.88, 18.96)	14.38(13.86, 14.9)
K	3.58	8.08(7.45, 8.71)	7.5(6.97, 8.04)	12.18(11.35, 13.01)	25.4(24.49, 26.32)
L	3.85	1.43(1.33, 1.54)	1.28(1.19, 1.38)	9.5(8.9, 10.11)	7.79(7.28, 8.31)
M	3.20	0.47(0.39, 0.54)	0.25(0.22, 0.28)	0.6(0.52, 0.69)	0.39(0.36, 0.42)

4.8 Conclusion and Further Work

This chapter answers research question 2 by demonstrating that it is feasible to automate the simulation build process. The simulation software Ciw was introduced, along with developing supporting customisations. The general method for automatically extracting the input parameters (arrivals, service, capacity and warm up) were discussed and validated using the *Raw Pathways* routing procedure for the working example of lung cancer.

A summary of the automated approach for the input parameters are as follows:

- Arrivals: Exponential distribution with arrival rate λ applied for arrival node/s, with *LimitedExponential* used for the *Raw Pathways* dummy node.
- Service: Deterministic at 0.1 for activity nodes and 0 for dummy nodes.
- Capacity: ‘Pattern’ method applied - seven day weekly pattern will be formed from the average number of individuals seen on a named day (e.g Monday, Tuesday etc).
- Warm up: No warm up will be automatically applied.

Additionally, each of the input parameters have ‘Custom’ options in Sim.Pro.Flow (as discussed in Section 4.6), where in particular capacity can be calculated considering the amount required to achieve waiting time targets (see Section 4.7 and 1.5).

Further Work

The suggestions for further work consider firstly, how to improve upon the general model presented and secondly how to progress the idea forward.

Some technical considerations for the general model should be made as follows:

1. Utilise the cyclical nature of Ciw’s server schedule, and define the capacity for a seven day period only. This will need to consider the warm up method used and model validation will be required to ensure the cyclical feature will

support the capacity definition. If possible, this should further reduce the run time of the simulation.

2. Consider service times of greater than one day for scenarios including inpatient stays.
3. Explore the feasibility of randomising the sampling order for the *Raw Pathways* to develop an additional routing procedure to explore use with predicting the future.
4. Investigate methods for calculating the values for the Warm Start method.

To progress the idea of automating the simulation build, this would required development of software to support the methods and enable accessibility (addressed in Chapter 6). Any future software produced would need to consider that automation should not replace the need for detailed analysis or reduce flexibility for exploration. The automation should enable users to more time efficiently build a model.

Chapter 5

Routing Procedures

5.1 Introduction

Research Question 3

Is it viable to support multiple interpretations of clinical pathways through combining a mixture of data mining and OR?

Chapter 1 introduced the requirement that to support multiple data types and minimise the complexity of the produced clinical pathway, multiple interpretations of constructing the network and pathways would be required. As such, Section 4.2 introduced the idea of routing procedures to address this. The first of the four routing procedures, *Raw Pathways*, is intended for validation and was explored in Chapter 4. The remaining routing procedures, *Full Transitions*, *Cluster Transitions* and *Process Medoids* are now introduced in more detail.

Full Transitions: The transition matrix can be extracted from the *Raw Pathways*. This allows for some variation to be introduced into system, which can be more reflective of future events. However, this can lead to unrealistic and impossible pathways, as the pathways are formed using probabilities, it could lead to individuals bouncing around within the system and performing really long path-

ways. Furthermore, the ‘at most once’ constraint (see subsection 1.2.4) cannot be enforced here and as such this is not suitable for DT1, but is suitable for DT2.

Cluster Transitions: Following on from the *Full Transitions*, the aim here is to minimise the variations observed and reduce the complexity of the produced clinical pathway, using the transition matrix method. From the results of the clustering we can extract the transition matrix for each cluster, which should produce less variation as the pathways have been clustered with those that they are most similar to. This method is also not suitable for DT1, but is suitable for DT2.

Process Medoids: The second piece of usable information that is produced from the clustering are the pathways that are selected as the medoids. These exact pathways can be performed, and as now only previous pathways can be performed - if the input data is of DT1 and has the ‘at most once’ constraint, so will the *Process Medoids* routing procedure. This is a more restricted view for the clinical pathway as the amount of variation is minimised. However, this does allow exploration of what could be the impact of restricting movement to a set of exact pathways.

As the *Full Transitions* and *Cluster Transitions* are not suitable for DT1 (the data type of the working example) these will be discussed here for the purpose of producing general methods, however these are not intended to produce pathways reflective of the working example data (as they do not hold the ‘at most once’ constraint).

The structure of this chapter explores each of the routing procedures in turn in Section 5.2, 5.3 and 5.4 respectively. Each section will discuss the specific application of the general arrivals (from subsection 4.5.1) and pathways interpretation for the routing. Furthermore, a visualisation of the network and pathways for each routing procedure is possible to produce through Sim.Pro.Flow using Graphviz [113]. The general method of how the routing procedures are extracted are discussed, with the working example of lung cancer applied throughout.

For note, the majority of this chapter was produced through using the code from Sim.Pro.Flow through a Jupyter notebook to support the reporting of the tables.

Penultimately, Section 5.5 explores the four routing procedures for the purpose of displaying the capabilities of automating the simulation build for multiple interpretations of the clinical pathway. The model build is recapped applying the STRESS guidelines to support reporting [202,271] (completed checklist in Appendix C). Section 5.6 closes the chapter with conclusion and considerations for further work.

5.2 Full Transitions

The *Full Transitions* use the transition matrix extracted from the *Raw Pathways*.

Table 5.1 shows the raw counts for the transition matrix.

Table 5.1: Raw Transition Matrix for the *Full Transitions*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	882	259	1	0	0	0	2	1	703	4	1	0	12	0
A	0	438	51	0	3	0	7	0	484	19	28	6	759	2
B	41	0	377	96	19	4	8	7	461	75	395	295	60	25
C	21	91	0	62	339	34	209	35	14	27	10	8	219	786
D	0	3	12	0	0	57	1	5	0	2	1	1	4	146
E	0	14	20	29	0	182	221	0	1	2	0	0	4	0
F	1	0	4	1	0	0	4	155	4	1	1	0	5	299
G	1	14	36	14	0	131	0	301	2	3	4	1	26	3
H	0	15	19	1	0	44	3	0	5	0	4	0	3	443
I	811	330	190	3	4	3	4	6	0	52	52	9	323	10
J	8	98	313	8	16	3	12	2	2	0	23	15	30	7
K	9	51	287	4	19	2	14	1	5	93	0	5	92	7
L	7	19	170	3	12	1	9	5	2	81	5	0	40	10
M	16	533	375	11	61	14	42	19	114	178	65	24	0	125

Arrivals

To construct the arrivals for the *Full Transitions* model, each activity that was an arrival activity will be assigned an arrival distribution, whilst the remaining activities will be assigned *NoArrivals*. Applying Equation 4.2 to the first line of the raw transition matrix in Table 5.1 and returns the values of λ in Table 5.2 (rounded to 3.d.p for reporting). These are then used with Ciw's built in exponential distribution.

Table 5.2: Arrival Lambda for *Full Transitions*.

	A	B	C	D	E	F	G	H	I	J	K	L	M
λ	2.436	0.715	0.003	0.0	0.0	0.0	0.006	0.003	1.942	0.011	0.003	0.0	0.033

Routing

To convert Table 5.1 into the transition matrix (excluding ‘Start’ row), each value is divided by the sum of its row. The last column (‘End’) is then removed as in Ciw if the row does not sum to 1, any remaining probability is used to exit the system. Finally making a small adjustment to account for any rounding error where the probability then exceeds 1, the excess is subtracted from the largest value. The resulting transition probability matrix can be seen in Table 5.3 (rounded to 3.d.p for reporting). This transition matrix is used as the routing in the simulation.

Table 5.3: Transition Probability Matrix for *Full Transitions*.

	A	B	C	D	E	F	G	H	I	J	K	L	M
A	0.000	0.244	0.028	0.000	0.002	0.000	0.004	0.000	0.269	0.011	0.016	0.003	0.422
B	0.022	0.000	0.202	0.052	0.010	0.002	0.004	0.004	0.247	0.040	0.212	0.158	0.032
C	0.011	0.049	0.000	0.033	0.183	0.018	0.113	0.019	0.008	0.015	0.005	0.004	0.118
D	0.000	0.013	0.052	0.000	0.000	0.246	0.004	0.022	0.000	0.009	0.004	0.004	0.017
E	0.000	0.030	0.042	0.061	0.000	0.385	0.467	0.000	0.002	0.004	0.000	0.000	0.008
F	0.002	0.000	0.008	0.002	0.000	0.000	0.008	0.326	0.008	0.002	0.002	0.000	0.011
G	0.002	0.026	0.067	0.026	0.000	0.244	0.000	0.562	0.004	0.006	0.007	0.002	0.049
H	0.000	0.028	0.035	0.002	0.000	0.082	0.006	0.000	0.009	0.000	0.007	0.000	0.006
I	0.451	0.184	0.106	0.002	0.002	0.002	0.002	0.003	0.000	0.029	0.029	0.005	0.180
J	0.015	0.182	0.583	0.015	0.030	0.006	0.022	0.004	0.004	0.000	0.043	0.028	0.056
K	0.015	0.087	0.487	0.007	0.032	0.003	0.024	0.002	0.008	0.158	0.000	0.008	0.156
L	0.019	0.052	0.467	0.008	0.033	0.003	0.025	0.014	0.005	0.223	0.014	0.000	0.110
M	0.010	0.338	0.238	0.007	0.039	0.009	0.027	0.012	0.072	0.113	0.041	0.015	0.000

Visualisation

The full transition probability matrix (including start and end) can be represented in a network graph as seen in Figure 5.1. Note the start probabilities are calculated in the same method as the transition probability (dividing by the total in the row).

Observing Figure 5.1, this evidences how complex the pathways within the working example are, as there is a high level of variation captured. Considering presenting this as the clinical pathway, for a case this complex, would not provide a clear representation. This provides strong justification for exploring how to reduce the variation to present a more understandable and useable clinical pathway. However, it is possible in less complex cases this model may prove to be sufficient.

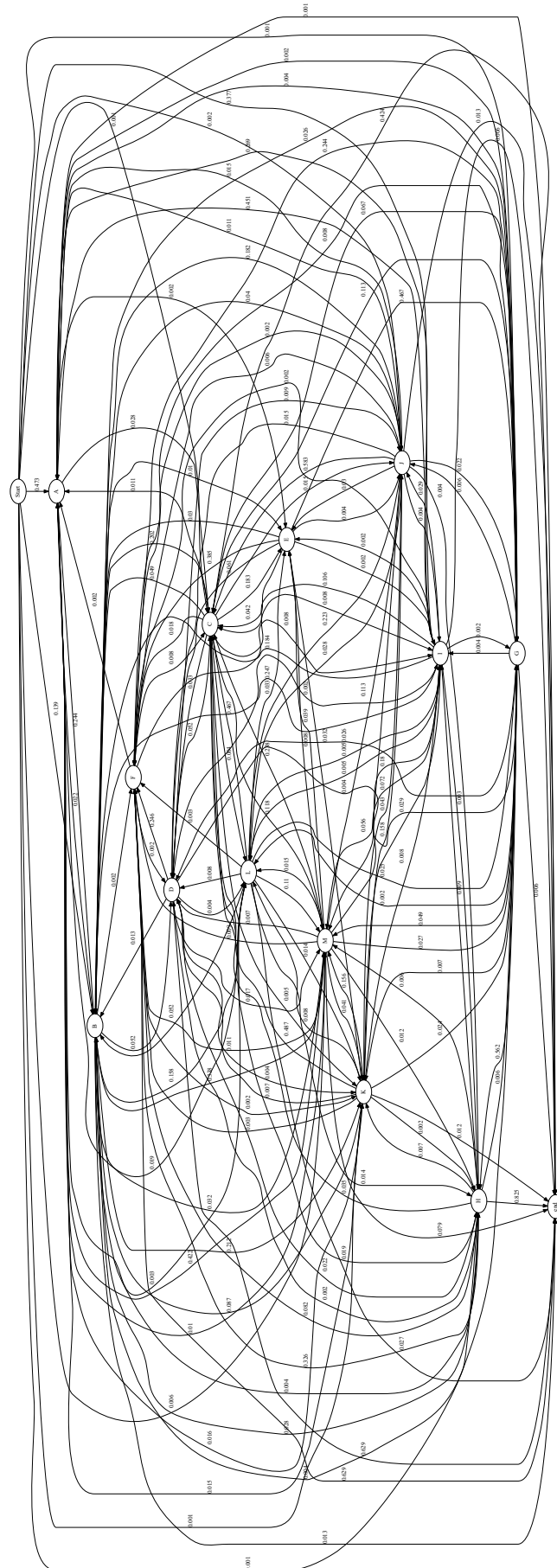


Figure 5.1: Network Graph for *Full Transitions*.

5.3 Cluster Transitions

The process for calculating the arrivals and pathways for the *Cluster Transitions* follows the same process as that for the *Full Transitions* with the exception that the process is performed for each cluster. Implementing this in Ciw is relatively easy as each cluster can correspond to a *customer class*.

For the purpose of this discussion, the classic k-medoids clustering was applied to the example working dataframe with rankings and groupings as in Table 4.2. Furthermore the centroids were chosen as the most occurring pathways, and values of $k = [2, 10]$ were investigated. Four clusters were chosen as this yielded the best silhouette score of 0.118 (rounded to 3 decimal places). It is important to note that as the clustering is performed on the dataframe (dataset of unique pathways as noted in Section 3.3 and 4.4), the raw transitions need to be ‘propagated’ to account for repeats. This is done by listing each pathway that is assigned to a cluster the number of times it is repeated in the original data. For example, if pathway ‘ABC’ was assigned to cluster 1 and repeated 4 times in the original data, then the list of pathways for cluster 1 would contain [‘ABC’, ‘ABC’, ‘ABC’, ‘ABC’]. The raw transitions are then calculated from these ‘propagated’ lists.

The total number of ‘propagated’ pathways that were assigned to each cluster were: Cluster 1 = 651, Cluster 2 = 367, Cluster 3 = 149 and Cluster 4 = 698. Note that these sum to 1865, which is the total number of pathways in the original data.

The four raw transition matrices can be seen in Table 5.5. Here the transitions can be compared to evaluate the clustering. For example, in cluster 1 the majority of arrivals are at activity ‘A’ or ‘B’, and the activity following ‘A’ is ‘I’ with a clear majority. Cluster 2 has a majority of arrivals at activity ‘A’ with a clear majority going on to perform activity ‘B’ and from ‘B’ a clear majority to activity ‘I’. Cluster 3 is smaller than the others and mainly contained pathways beginning ‘AMI...’. Finally, in cluster 4 almost all the pathways begin with activity ‘I’ and following on to activity ‘A’. Deeper relationships can also be considered, such as in

cluster 4 activity ‘I’ goes to activity ‘A’ 631 times, where as activity ‘A’ never went to activity ‘I’. Examining these clusters in detail could help the user gain insight when choosing the groupings for the activities during clustering.

Arrivals

Similarly to the *Full Transitions*, applying Equation 4.2 to the first line of each of the raw transition matrices, the resulting values of λ for each cluster can be seen in Table 5.4. Each cluster has a corresponding *customer class*, where C_{iw} ’s built in exponential distribution is used for each arrival activity and *NoArrivals* otherwise.

Table 5.4: Arrival Lambda for *Cluster Transitions*.

Cluster	A	B	C	D	E	F	G	H	I	J	K	L	M
1	0.687	0.309	0.000	0.0	0.0	0.0	0.000	0.000	0.000	0.003	0.000	0.0	0.002
2	0.842	0.109	0.000	0.0	0.0	0.0	0.003	0.003	0.022	0.000	0.003	0.0	0.019
3	0.846	0.121	0.007	0.0	0.0	0.0	0.000	0.000	0.000	0.000	0.000	0.0	0.027
4	0.000	0.000	0.000	0.0	0.0	0.0	0.001	0.000	0.996	0.003	0.000	0.0	0.000

Routing

Using the same calculation method as the *Full Transitions*, the transition probability matrix for each cluster can be seen in Table 5.6. The routing is defined as a transition matrix for each *customer class* corresponding to a cluster.

Visualisation

Each of the transition probability matrices for the clusters can be displayed as a network graph using the same method as for the *Full Transitions*.

Figure 5.2 shows the network graph for cluster 3 (where the remaining cluster networks can be found in Appendix C). Comparing the network graph displayed for the *Full Transitions* (Figure 5.1) and the *Cluster Transitions*, does display that the individual diagrams for the *Cluster Transitions* are less complex, however there are multiple diagrams.

Table 5.5: Raw Transitions for *Cluster Transitions*, Cluster 1, 2, 3 and 4 Respectively.

	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	447	201	0	0	0	0	0	0	0	2	0	0	1	0
A	0	1	28	0	2	0	1	0	445	0	10	1	157	0
B	10	0	133	34	4	2	3	2	190	16	155	69	25	8
C	9	31	0	13	100	11	73	9	1	6	3	2	94	297
D	0	2	3	0	0	21	1	0	0	1	1	1	2	42
E	0	4	9	11	0	58	72	0	0	1	0	0	1	0
F	1	0	1	0	0	0	2	51	0	0	0	0	3	99
G	0	2	10	3	0	39	0	109	0	0	1	0	16	1
H	0	5	6	1	0	16	1	0	1	0	4	0	0	147
I	175	186	19	1	2	1	3	2	0	26	19	3	202	1
J	2	24	91	4	5	1	2	0	0	0	9	8	18	1
K	1	18	114	1	7	0	5	0	2	43	0	1	35	2
L	0	9	39	0	3	0	0	1	0	18	1	0	15	5
M	0	168	196	6	33	8	18	7	1	52	26	6	0	48

	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	309	40	0	0	0	0	1	1	8	0	1	0	7	0
A	0	265	2	0	0	0	0	0	19	1	1	0	36	0
B	4	0	34	1	1	0	2	0	271	5	10	29	8	0
C	0	1	0	9	30	7	22	6	10	2	2	2	44	225
D	0	0	1	0	0	2	0	0	0	0	0	0	0	13
E	0	0	3	1	0	15	20	0	1	1	0	0	0	0
F	0	0	0	0	0	0	0	14	2	0	0	0	1	25
G	1	2	4	1	0	12	0	31	2	0	1	1	2	1
H	0	2	6	0	0	2	1	0	2	0	0	0	1	45
I	3	18	159	1	2	2	1	3	0	7	17	3	101	7
J	0	4	27	0	0	1	2	0	0	0	2	0	2	1
K	1	2	15	0	2	0	1	0	2	5	0	0	10	3
L	0	0	21	1	2	0	0	1	1	8	0	0	2	2
M	6	33	88	2	4	1	8	3	6	10	7	3	0	43

	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	126	18	1	0	0	0	0	0	0	0	0	0	4	0
A	0	0	1	0	0	0	0	0	20	1	0	0	127	0
B	15	0	51	7	6	0	1	0	0	3	34	27	0	5
C	1	10	0	5	35	5	24	4	3	1	0	0	6	55
D	0	0	3	0	0	1	0	1	0	0	0	0	0	9
E	0	1	2	0	0	15	30	0	0	0	0	0	0	0
F	0	0	1	0	0	0	0	20	2	0	0	0	0	25
G	0	1	6	0	0	18	0	32	0	1	1	0	0	0
H	0	2	1	0	0	6	0	0	1	0	0	0	0	49
I	2	86	8	1	0	0	0	1	0	12	12	2	10	2
J	0	16	25	0	2	0	2	1	1	0	0	0	0	0
K	0	8	23	0	2	2	0	0	1	12	0	0	2	0
L	1	1	16	0	0	1	1	0	1	8	1	0	0	0
M	4	6	11	1	3	0	1	0	107	9	2	1	0	4

	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	0	0	0	0	0	0	1	0	695	2	0	0	0	0
A	0	172	20	0	1	0	6	0	0	17	17	5	439	2
B	12	0	159	54	8	2	2	5	0	51	196	170	27	12
C	11	49	0	35	174	11	90	16	0	18	5	4	75	209
D	0	1	5	0	0	33	0	4	0	1	0	0	2	82
E	0	9	6	17	0	94	99	0	0	0	0	0	3	0
F	0	0	2	1	0	0	2	70	0	1	1	0	1	150
G	0	9	16	10	0	62	0	129	0	2	1	0	8	1
H	0	6	6	0	0	20	1	0	1	0	0	0	2	202
I	631	40	4	0	0	0	0	0	0	7	4	1	10	0
J	6	54	170	4	9	1	6	1	1	0	12	7	10	5
K	7	23	135	3	8	0	8	1	0	33	0	4	45	2
L	6	9	94	2	7	0	8	3	0	47	3	0	23	3
M	6	326	80	2	21	5	15	9	0	107	30	14	0	30

Table 5.6: Transition Probability Matrix for *Cluster Transitions*, Cluster 1, 2, 3 and 4 Respectively.

	A	B	C	D	E	F	G	H	I	J	K	L	M
A	0.000	0.002	0.043	0.000	0.003	0.000	0.002	0.000	0.690	0.000	0.016	0.002	0.243
B	0.015	0.000	0.204	0.052	0.006	0.003	0.005	0.003	0.292	0.025	0.238	0.106	0.038
C	0.014	0.048	0.000	0.020	0.154	0.017	0.112	0.014	0.002	0.009	0.005	0.003	0.145
D	0.000	0.027	0.041	0.000	0.000	0.284	0.014	0.000	0.000	0.014	0.014	0.014	0.027
E	0.000	0.026	0.058	0.071	0.000	0.372	0.462	0.000	0.000	0.006	0.000	0.000	0.006
F	0.006	0.000	0.006	0.000	0.000	0.000	0.013	0.325	0.000	0.000	0.000	0.000	0.019
G	0.000	0.011	0.055	0.017	0.000	0.215	0.000	0.602	0.000	0.000	0.006	0.000	0.088
H	0.000	0.028	0.033	0.006	0.000	0.088	0.006	0.000	0.006	0.000	0.022	0.000	0.000
I	0.273	0.291	0.030	0.002	0.003	0.002	0.005	0.003	0.000	0.041	0.030	0.005	0.316
J	0.012	0.145	0.552	0.024	0.030	0.006	0.012	0.000	0.000	0.000	0.055	0.048	0.109
K	0.004	0.079	0.498	0.004	0.031	0.000	0.022	0.000	0.009	0.188	0.000	0.004	0.153
L	0.000	0.099	0.429	0.000	0.033	0.000	0.000	0.011	0.000	0.198	0.011	0.000	0.165
M	0.000	0.295	0.344	0.011	0.058	0.014	0.032	0.012	0.002	0.091	0.046	0.011	0.000

	A	B	C	D	E	F	G	H	I	J	K	L	M
A	0.000	0.818	0.006	0.000	0.000	0.000	0.000	0.000	0.059	0.003	0.003	0.000	0.111
B	0.011	0.000	0.093	0.003	0.003	0.000	0.005	0.000	0.742	0.014	0.027	0.079	0.022
C	0.000	0.003	0.000	0.025	0.083	0.019	0.061	0.017	0.028	0.006	0.006	0.006	0.122
D	0.000	0.000	0.062	0.000	0.000	0.125	0.000	0.000	0.000	0.000	0.000	0.000	0.000
E	0.000	0.000	0.073	0.024	0.000	0.366	0.488	0.000	0.024	0.024	0.000	0.000	0.000
F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.333	0.048	0.000	0.000	0.000	0.024
G	0.017	0.034	0.069	0.017	0.000	0.207	0.000	0.534	0.034	0.000	0.017	0.017	0.034
H	0.000	0.034	0.102	0.000	0.000	0.034	0.017	0.000	0.034	0.000	0.000	0.000	0.017
I	0.009	0.056	0.491	0.003	0.006	0.006	0.003	0.009	0.000	0.022	0.052	0.009	0.312
J	0.000	0.103	0.692	0.000	0.000	0.026	0.051	0.000	0.000	0.000	0.051	0.000	0.051
K	0.024	0.049	0.366	0.000	0.049	0.000	0.024	0.000	0.049	0.122	0.000	0.000	0.244
L	0.000	0.000	0.553	0.026	0.053	0.000	0.000	0.026	0.026	0.211	0.000	0.000	0.053
M	0.028	0.154	0.411	0.009	0.019	0.005	0.037	0.014	0.028	0.047	0.033	0.014	0.000

	A	B	C	D	E	F	G	H	I	J	K	L	M
A	0.000	0.000	0.007	0.000	0.000	0.000	0.000	0.000	0.134	0.007	0.000	0.000	0.852
B	0.101	0.000	0.342	0.047	0.040	0.000	0.007	0.000	0.000	0.020	0.228	0.181	0.000
C	0.007	0.067	0.000	0.034	0.235	0.034	0.161	0.027	0.020	0.007	0.000	0.000	0.040
D	0.000	0.000	0.214	0.000	0.000	0.071	0.000	0.071	0.000	0.000	0.000	0.000	0.000
E	0.000	0.021	0.042	0.000	0.000	0.312	0.625	0.000	0.000	0.000	0.000	0.000	0.000
F	0.000	0.000	0.021	0.000	0.000	0.000	0.000	0.417	0.042	0.000	0.000	0.000	0.000
G	0.000	0.017	0.102	0.000	0.000	0.305	0.000	0.542	0.000	0.017	0.017	0.000	0.000
H	0.000	0.034	0.017	0.000	0.000	0.102	0.000	0.000	0.017	0.000	0.000	0.000	0.000
I	0.015	0.632	0.059	0.007	0.000	0.000	0.000	0.007	0.000	0.088	0.088	0.015	0.074
J	0.000	0.340	0.532	0.000	0.043	0.000	0.043	0.021	0.021	0.000	0.000	0.000	0.000
K	0.000	0.160	0.460	0.000	0.040	0.040	0.000	0.000	0.020	0.240	0.000	0.000	0.040
L	0.033	0.033	0.533	0.000	0.000	0.033	0.033	0.000	0.033	0.267	0.033	0.000	0.000
M	0.027	0.040	0.074	0.007	0.020	0.000	0.007	0.000	0.718	0.060	0.013	0.007	0.000

	A	B	C	D	E	F	G	H	I	J	K	L	M
A	0.000	0.253	0.029	0.000	0.001	0.000	0.009	0.000	0.000	0.025	0.025	0.007	0.647
B	0.017	0.000	0.228	0.077	0.011	0.003	0.003	0.007	0.000	0.073	0.281	0.244	0.039
C	0.016	0.070	0.000	0.050	0.250	0.016	0.129	0.023	0.000	0.026	0.007	0.006	0.108
D	0.000	0.008	0.039	0.000	0.000	0.258	0.000	0.031	0.000	0.008	0.000	0.000	0.016
E	0.000	0.039	0.026	0.075	0.000	0.412	0.434	0.000	0.000	0.000	0.000	0.000	0.013
F	0.000	0.000	0.009	0.004	0.000	0.000	0.009	0.307	0.000	0.004	0.004	0.000	0.004
G	0.000	0.038	0.067	0.042	0.000	0.261	0.000	0.542	0.000	0.008	0.004	0.000	0.034
H	0.000	0.025	0.025	0.000	0.000	0.084	0.004	0.000	0.004	0.000	0.000	0.000	0.008
I	0.905	0.057	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.006	0.001	0.014
J	0.021	0.189	0.594	0.014	0.031	0.003	0.021	0.003	0.003	0.000	0.042	0.024	0.035
K	0.026	0.086	0.502	0.011	0.030	0.000	0.030	0.004	0.000	0.123	0.000	0.015	0.167
L	0.029	0.044	0.459	0.010	0.034	0.000	0.039	0.015	0.000	0.229	0.015	0.000	0.112
M	0.009	0.505	0.124	0.003	0.033	0.008	0.023	0.014	0.000	0.166	0.047	0.022	0.000

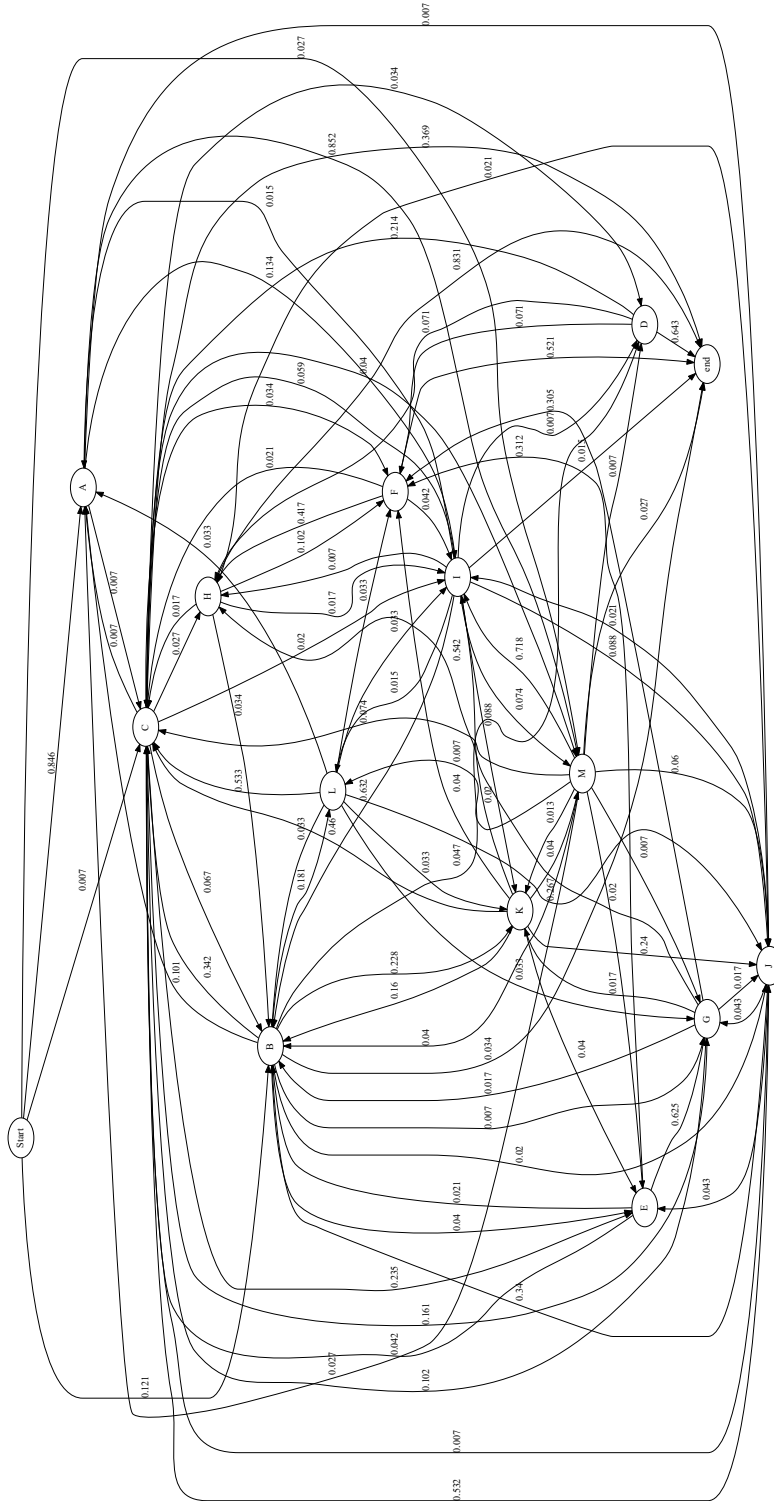


Figure 5.2: Network Graph for *Cluster Transitions* Cluster 3.

5.4 *Process Medoids*

The *Process Medoids* routing procedure applies the previously discussed clustering method but then only simulates each cluster medoid as a representative pathway for that cluster, using process based routing.

However, considering how many clusters (k) to select when only the medoids are going to be performed poses a challenge. The silhouette score that was previously used as the main selection indicator resulted in small values of k to be selected (less than ten). It may not be appropriate to consider so few pathways for this system, as they may not capture enough of the variation and the resulting clinical pathway would be too simplistic. Therefore additional information needs to be provided to the user to enable them to make a more informed decision on the value of k , whilst considering that the goal is to produce a clinical pathway that is understandable and usable.

There are many evaluation metrics that can be used to select the number of clusters, such as Silhouette Score and Davies-Bouldin Index to name a few [170] [229] [249]. Furthermore, the Scikit-Learn Library [246] discuss that the metrics should compare if members of the same cluster are more similar to each other than those of another cluster. The Scikit-Learn library offers many metrics, namely, Adjusted Rand index, Mutual Information based scores, Homogeneity, completeness and V-measure, Fowlkes-Mallows scores, Silhouette Coefficient, Calinski-Harabasz index, Davies-Bouldin index, and Contingency Matrix.

As k-medoids is an unsupervised approach the only available appropriate metrics are the Silhouette Coefficient, Calinski-Harabasz index and the Davies-Bouldin index.

However, alongside clear cluster definition, it is vital to produce a good selection of medoids that reflect the original data as close as possible. To achieve both of these standards, it is logical to report more than one evaluation measure. Therefore, we report on the silhouette score alongside additional information that can allow the

user to make an informed decision on the number of clusters.

When considering the additional information to report, this must reflect the intentions for use and what ‘good’ means. Two key features were identified as ease of understanding and close reflection of the original data. Putting this in terms of the network, this would translate to minimise the number of edges whilst also minimising the differences in the edge values in comparison to the full network.

Note the analysis and figure show in this subsection were produced using Sim.Pro.Flow.

5.4.1 Medoids Selection

Considering the two key features identified above (minimising the number of edges and differences in the values), a summary of the calculation process is as follows:

1. From the original data raw transition matrix, reduce transitions occurring less than a specified number of times (selected as a percentage of the total pathways referred to as adjust percentage) to 0 to produce the reduced matrix, and count the number of non-zero connections (referred to as reduced connections),
2. Calculate the reduced transition probability matrix,
3. Construct the medoids transition probability matrix, accounting for connections representing the number of pathways in the cluster (‘propagated’),
4. Count the number of non-zero connections in the medoids transition probability matrix,
5. Calculate the absolute difference matrix (from the reduced matrix and medoids transition probability matrix),
6. Calculate the average percentage points different in the difference matrix,
7. Plot all non-zero differences of the difference matrix.

Using this method for values of k from 2 to 30 for the working example of lung

cancer produced a solution of 21 clusters. This solution is displayed in Table 5.7 and will be used as example to aid understanding while each step is now discussed in detail.

Table 5.7: Process Based Clustering Results for the 21 *Process Medoids*.

Cluster	Medoids	Counts	Propagated Counts	Arrivals λ
0	BIAMC	27	159	0.439
1	ABIC	15	102	0.282
2	ABIMC	45	124	0.343
3	IAMBC	31	87	0.240
4	ABICM	16	53	0.146
5	AIBC	66	161	0.445
6	AIMBC	74	135	0.373
7	IAMBKC	16	49	0.135
8	AMIBC	54	113	0.312
9	IAMBLC	24	73	0.202
10	IAMBCEGFH	57	90	0.249
11	IAMBCGH	69	132	0.365
12	IAMBCEF	36	71	0.196
13	BIAC	16	43	0.119
14	AIBMC	48	84	0.232
15	BC	22	58	0.160
16	AMBIC	24	38	0.105
17	IAMJCBD	38	63	0.174
18	IAMBKCEF	20	39	0.108
19	IABMC	48	106	0.293
20	AIMBKC	37	85	0.235

Step one considers that the aim is to minimise the number of edges. The idea is to remove transitions that occur less than a specified frequency, in the original raw transition matrix, before calculating the probability. This should allow the medoids transitions to be more likely to achieve a similar transition matrix to the original, accounting for a reduced amount of variation. Using the raw transition matrix from Table 5.1, as there are 1,865 pathways considered here, even removing transitions that occur less than (a modest) 5% of the time, would be those that occurred less than 93.25 times. Nevertheless, for the purpose of this example, 5% was selected. To obtain the reduced connections, the number of non-zero connections are counted, which in this example is 36 (original number of connections was 159).

Step two calculates the reduced transition probability matrix, in the same way as previously explained, but this time keeping the start and end transitions. The reduced transition probability matrix can be seen in Table 5.9

Step three builds the medoids transition probability matrix. Similarly to the *Clustered Transitions*, consideration is needed to account for the fact that each of the medoids represent all of the pathways that are assigned to that cluster. Therefore each connection is not counted as 1 but as the ‘propagated’ value (see page 142 for explanation). For example, from Table 5.7 cluster 0 has medoid ‘BIAMC’ and 27 pathways were assigned to this cluster, which when propagated to consider the repeats of the pathways is 159¹. Therefore each transition for cluster 0 would count as 159 transitions. The raw transition matrix for 21 medoids can be seen in Table 5.8. This medoids transition probability matrix can then be calculated (in the same way as previously explained) and is shown in Table 5.10

Table 5.8: Raw Transition Matrix for the 21 *Process Medoids*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	895	260	0	0	0	0	0	0	710	0	0	0	0	0
A	0	385	43	0	0	0	0	0	465	0	0	0	914	0
B	0	0	847	63	0	0	0	0	519	0	173	73	190	0
C	0	63	0	0	200	0	132	0	0	0	0	0	53	1417
D	0	0	0	0	0	0	0	0	0	0	0	0	0	63
E	0	0	0	0	0	110	90	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	90	0	0	0	0	0	110
G	0	0	0	0	0	90	0	132	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	222
I	912	358	193	0	0	0	0	0	0	0	0	0	344	0
J	0	0	63	0	0	0	0	0	0	0	0	0	0	0
K	0	0	173	0	0	0	0	0	0	0	0	0	0	0
L	0	0	73	0	0	0	0	0	0	0	0	0	0	0
M	0	799	473	0	0	0	0	0	113	63	0	0	0	53

Step four counts the number of non-zero connections in the medoids transition probability matrix. This allows for comparison with the number of connections found in step 1 (36). For note, Sim.Pro.Flow allows the user to specify a tolerance surrounding the number of connections where values of k with number of connections within the tolerance of the reduced connections are displayed. For this example, choosing a tolerance of 2 highlighted results for $k = 14, 15, 16, 17, 18, 19, 20$, and the 21 medoids, as these solutions have connections in the range [34,38].

¹For further explanation, in Table 5.8 the transition from the start to activity A sums the propagated counts for clusters 1, 2, 4, 5, 6, 8, 14, 16, and 20 with values 102, 124, 53, 161, 135, 113, 84, 38 and 85 respectively (totalling 895).

Step five creates the difference matrix by calculating the absolute difference between the corresponding values in the reduced and medoids transition probability matrices. The values within the difference matrix can be described as the absolute number of percentage points different between the transition probabilities of the reduced and medoids transition probability matrices. The difference matrix of Table 5.9 and Table 5.10 can be seen in Table 5.11.

Step six calculates the average percentage points different, by calculating the average of the non-zero values in the difference matrix. For the 21 medoids solutions this was 0.081 (rounded to 3.d.p), meaning on average there were 8.1 percentage points different between the reduced transition probability matrix (Table 5.9) and 21 medoids transition probability matrix (Table 5.10).

Finally, step seven plots all the non-zero values of the difference matrix as a violin plot, with the title containing the information: k , number of connections and average percentage points different respectively separated by an underscore, as seen in Figure 5.3. Observing the violin plots provide the user with a visual representation of the spread of the non-zero values in the difference matrix. On each violin plot, the upper and lower lines are the maximum and minimum values respectively and the middle line indicates the mean. This allows the user to compare multiple values of k to aid the selection. Note, in Sim.Pro.Flow it is also possible to save the solution as presented in Table 5.7 for each value of k (not including the Arrivals λ column), allowing the user to inspect the pathways chosen as the medoids for a deeper evaluation of appropriateness.

Analysing Figure 5.3, plots $k = [2, 11]$ (the top two rows) shows a maximum value of 1.0 (meaning there was 100% difference for some values in the difference matrix) and there were a few values above 50%, indicated by the wider volume between 0.5 and 1.0. Comparing these to $k = [12, 19]$, the volume between 0.5 and 1.0 is virtually non-existent, however there still remains a record with a value of 1.0 (indicated by the maximum). For $k = [20, 30]$ the maximum is now below 0.5, and

the vertical spread of the distribution is getting smaller, indicated by the plot and the reduction in the mean. However, the number of connections are increasing and thus the user should find a balance between minimising the number of connections and minimising the mean. For this example, from the violin plot 21 medoids was selected as the solution, as the number of connections were within the tolerance. Although the solution of 20 medoids also had 38 connections, the average difference was slightly lower for the 21 medoids (0.081 compared to 0.084).

Table 5.9: Reduced Transition Probability Matrix (5%).

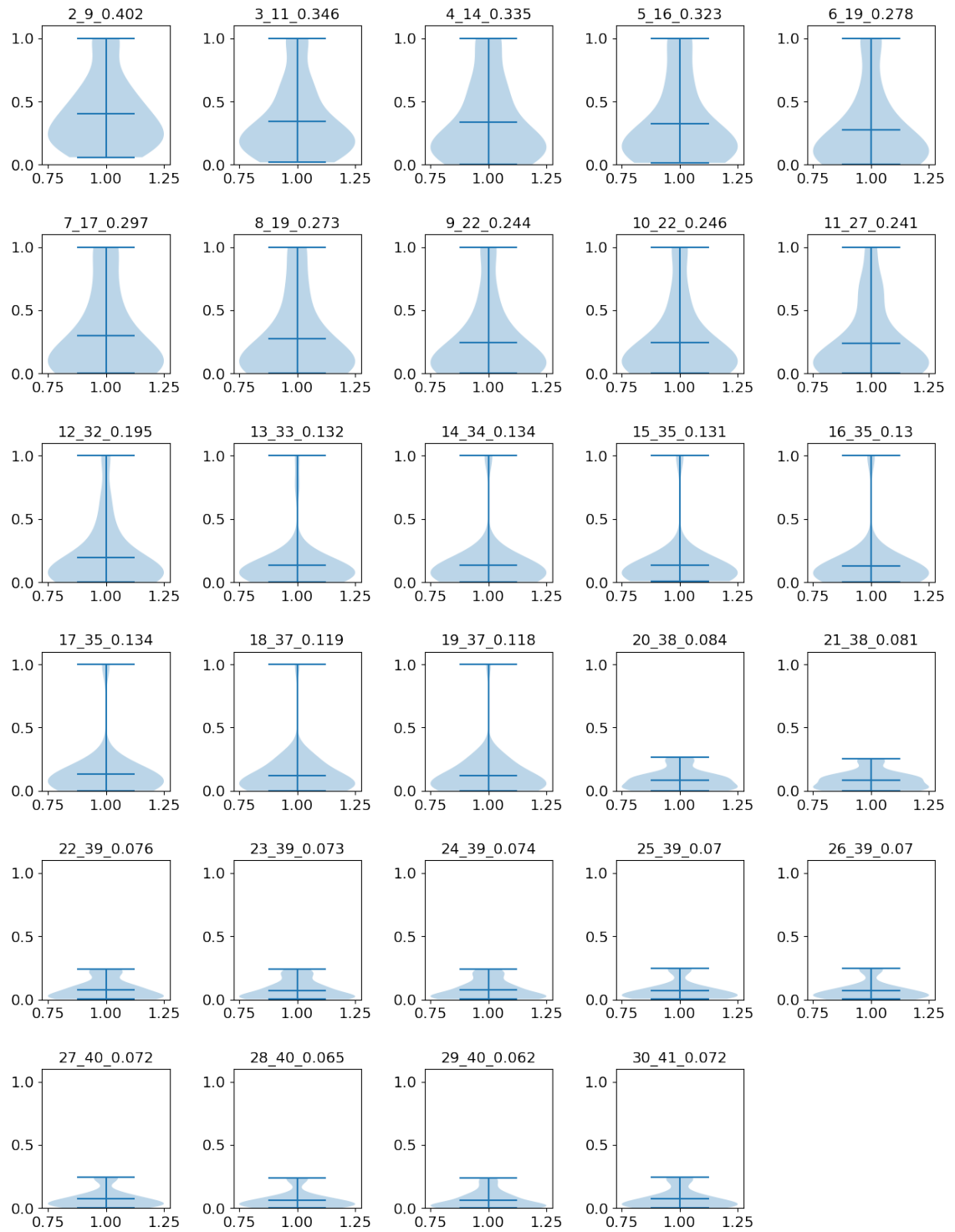
	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	0.478	0.140	0.000	0.000	0.000	0.000	0.000	0.000	0.381	0.000	0.000	0.000	0.000	0.000
A	0.000	0.261	0.000	0.000	0.000	0.000	0.000	0.000	0.288	0.000	0.000	0.000	0.452	0.000
B	0.000	0.000	0.232	0.059	0.000	0.000	0.000	0.000	0.284	0.000	0.243	0.182	0.000	0.000
C	0.000	0.000	0.000	0.000	0.218	0.000	0.135	0.000	0.000	0.000	0.000	0.000	0.141	0.506
D	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
E	0.000	0.000	0.000	0.000	0.000	0.452	0.548	0.000	0.000	0.000	0.000	0.000	0.000	0.000
F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.341	0.000	0.000	0.000	0.000	0.000	0.659
G	0.000	0.000	0.000	0.000	0.000	0.303	0.000	0.697	0.000	0.000	0.000	0.000	0.000	0.000
H	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
I	0.490	0.200	0.115	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.195	0.000
J	0.000	0.238	0.762	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
K	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
L	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
M	0.000	0.402	0.283	0.000	0.000	0.000	0.000	0.000	0.086	0.134	0.000	0.000	0.000	0.094

Table 5.10: Medoids Transition Probability Matrix - for the 21 *Process Medoids*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	0.480	0.139	0.000	0.000	0.000	0.000	0.000	0.000	0.381	0.000	0.000	0.000	0.000	0.000
A	0.000	0.213	0.024	0.000	0.000	0.000	0.000	0.000	0.257	0.000	0.000	0.000	0.506	0.000
B	0.000	0.000	0.454	0.034	0.000	0.000	0.000	0.000	0.278	0.000	0.093	0.039	0.102	0.000
C	0.000	0.034	0.000	0.000	0.107	0.000	0.071	0.000	0.000	0.000	0.000	0.000	0.028	0.760
D	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
E	0.000	0.000	0.000	0.000	0.000	0.550	0.450	0.000	0.000	0.000	0.000	0.000	0.000	0.000
F	0.000	0.000	0.000	0.000	0.000	0.000	0.405	0.000	0.595	0.000	0.000	0.000	0.000	0.550
G	0.000	0.000	0.000	0.000	0.000	0.405	0.000	0.595	0.000	0.000	0.000	0.000	0.000	0.000
H	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
I	0.505	0.198	0.107	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.190	0.000
J	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
K	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
L	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
M	0.000	0.532	0.315	0.000	0.000	0.000	0.000	0.000	0.075	0.042	0.000	0.000	0.000	0.035

Table 5.11: Difference Matrix Between Reduced and Medoids Transition Probability Matrices.

	A	B	C	D	E	F	G	H	I	J	K	L	M	End
Start	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.00	0.000	0.000	0.000
A	0.000	0.047	0.024	0.000	0.000	0.000	0.000	0.000	0.031	0.000	0.00	0.000	0.054	0.000
B	0.000	0.000	0.222	0.025	0.000	0.000	0.000	0.000	0.006	0.000	0.15	0.143	0.102	0.000
C	0.000	0.034	0.000	0.000	0.111	0.000	0.064	0.000	0.000	0.000	0.00	0.000	0.113	0.254
D	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.000	0.000	0.000
E	0.000	0.000	0.000	0.000	0.000	0.098	0.098	0.000	0.000	0.000	0.00	0.000	0.000	0.000
F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.109	0.000	0.000	0.00	0.000	0.000	0.109
G	0.000	0.000	0.000	0.000	0.000	0.102	0.000	0.102	0.000	0.000	0.00	0.000	0.000	0.000
H	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.000	0.000	0.000
I	0.014	0.001	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.000	0.005	0.000
J	0.000	0.238	0.238	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.000	0.000	0.000
K	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.000	0.000	0.000
L	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.000	0.000	0.000
M	0.000	0.130	0.032	0.000	0.000	0.000	0.000	0.000	0.011	0.092	0.00	0.000	0.000	0.059

Figure 5.3: Process Based Selection Violin Plot for $k = [2, 30]$.

5.4.2 Parameters

Arrivals

Following on from the selection of 21 medoids, similarly to the *Cluster Transitions*, each cluster is represented by a *customer class*. Arrivals for each *customer class* are only generated for the start activity of the respective medoid, with NoArrivals listed otherwise. Equation 4.2 is used to calculate the arrival rate for the propagated counts for each cluster, as shown in Table 5.7. Again Ciw's exponential distribution is used with the values of λ from Table 5.7.

Pathways

Process based routing is applied, where the routing function take in the individual's *customer class* and returns the route corresponding to the medoid. The routing function only needs to be defined once per node as it is universal to all classes.

As each cluster is a *customer class*, and the pathways contain no repeated activities, the traditional transition matrix could have been applied. However, the process based routing was applied to represent the generalised approach.

Visualisation

Four diagrams, each displaying a different depth perspective are produced. Firstly, to aid comparison of the reduced transition probability matrix and the medoids transition probability matrix, both network graphs are produced and can be seen in Figure 5.4 and 5.5 respectively.

However, this is not technically an accurate depiction of the pathways that are being simulated, therefore Figure 5.6 displays the raw medoids for each cluster. Furthermore, Figure 5.7 represents the raw medoids by grouping activities by position. The connections between the activities in the groups represent the number of times that connection occurred in the raw medoids².

²Note in all four diagrams a light grey line with no label represents a value of 1, whether that be one connection (Figure 5.7) or a probability of 1 (Figure 5.5).

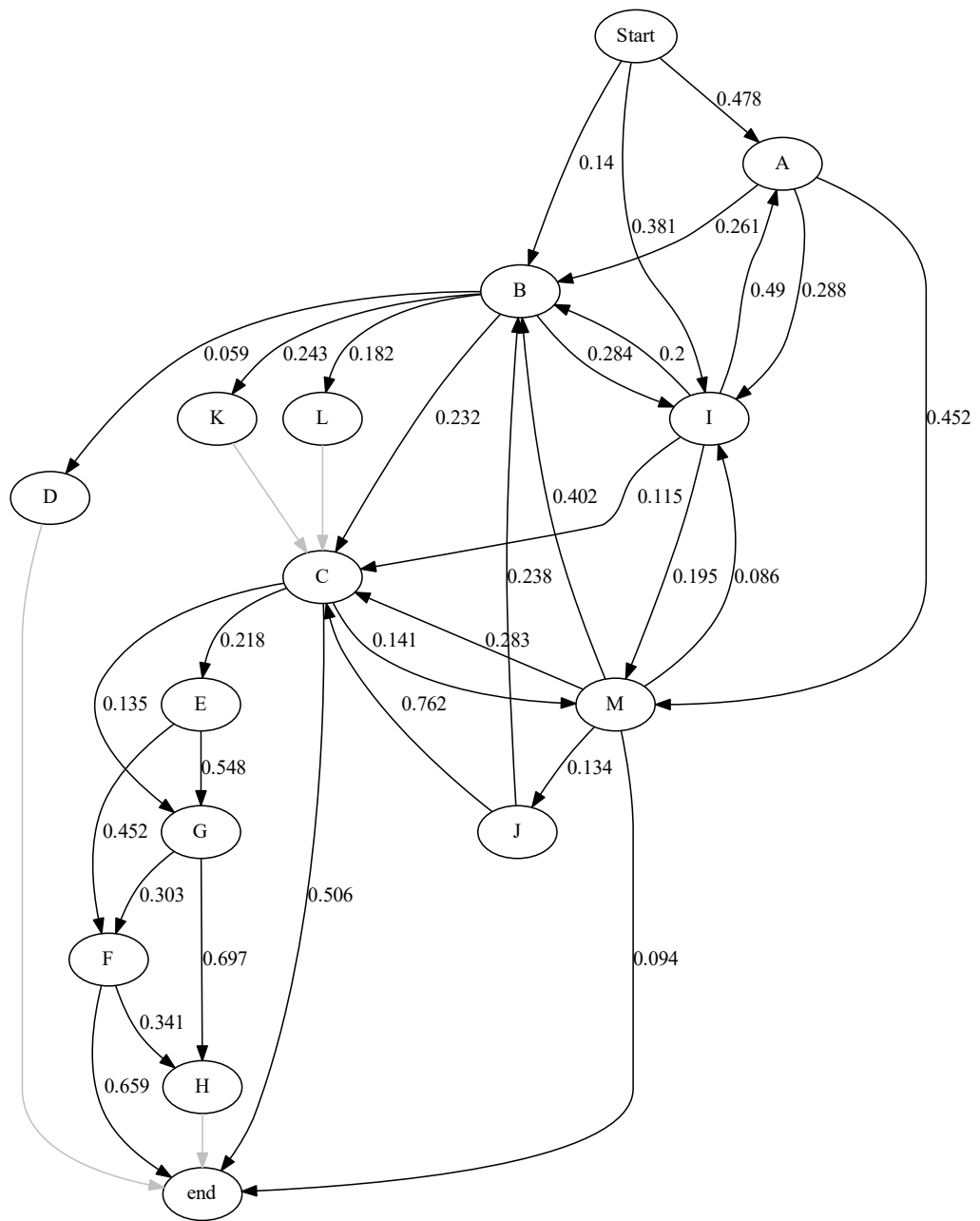


Figure 5.4: Network Graph of the Reduced Transition Probabilities.

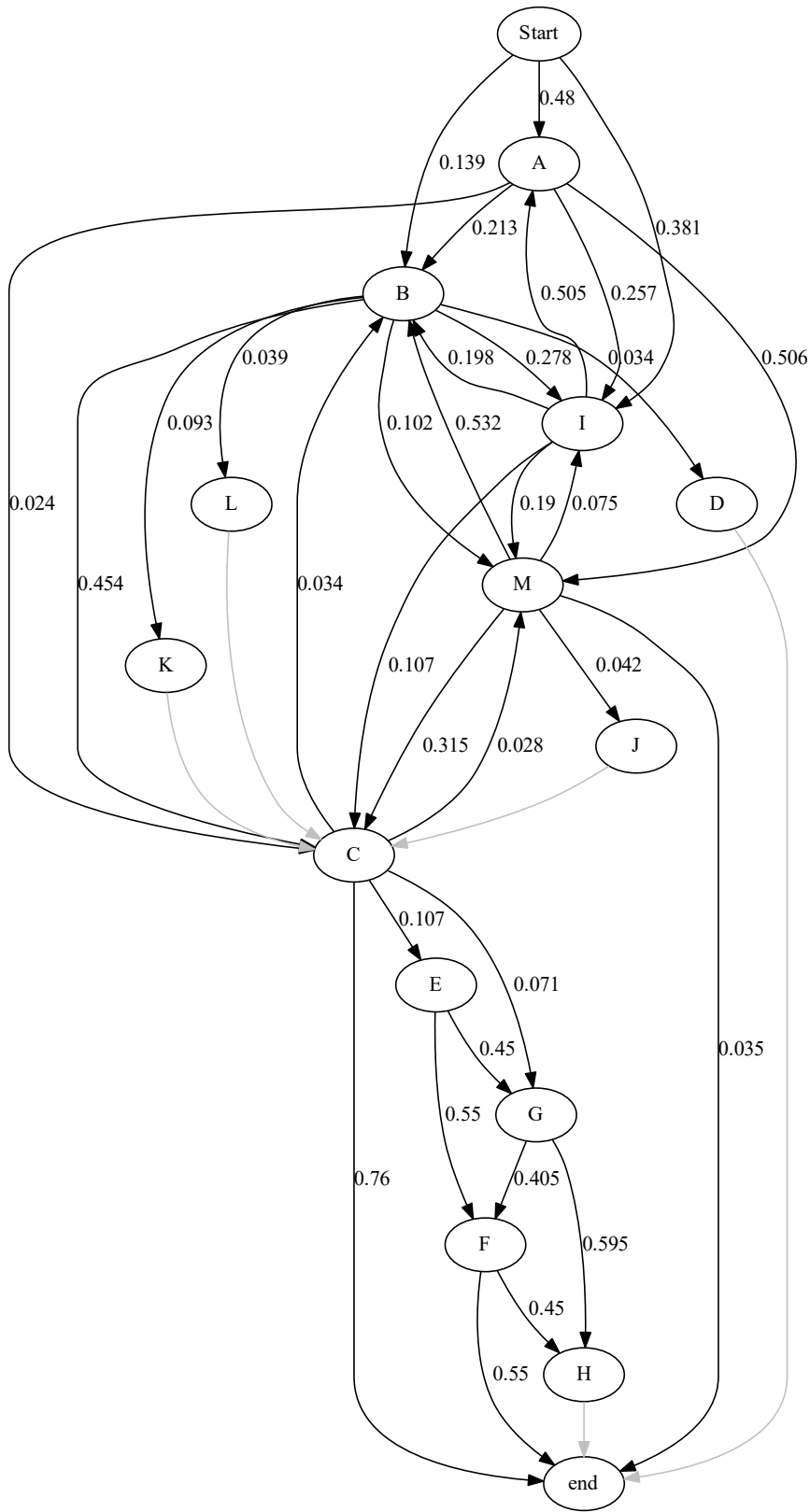


Figure 5.5: Network Graph of the 21 *Process Medoids* Transition Probabilities.

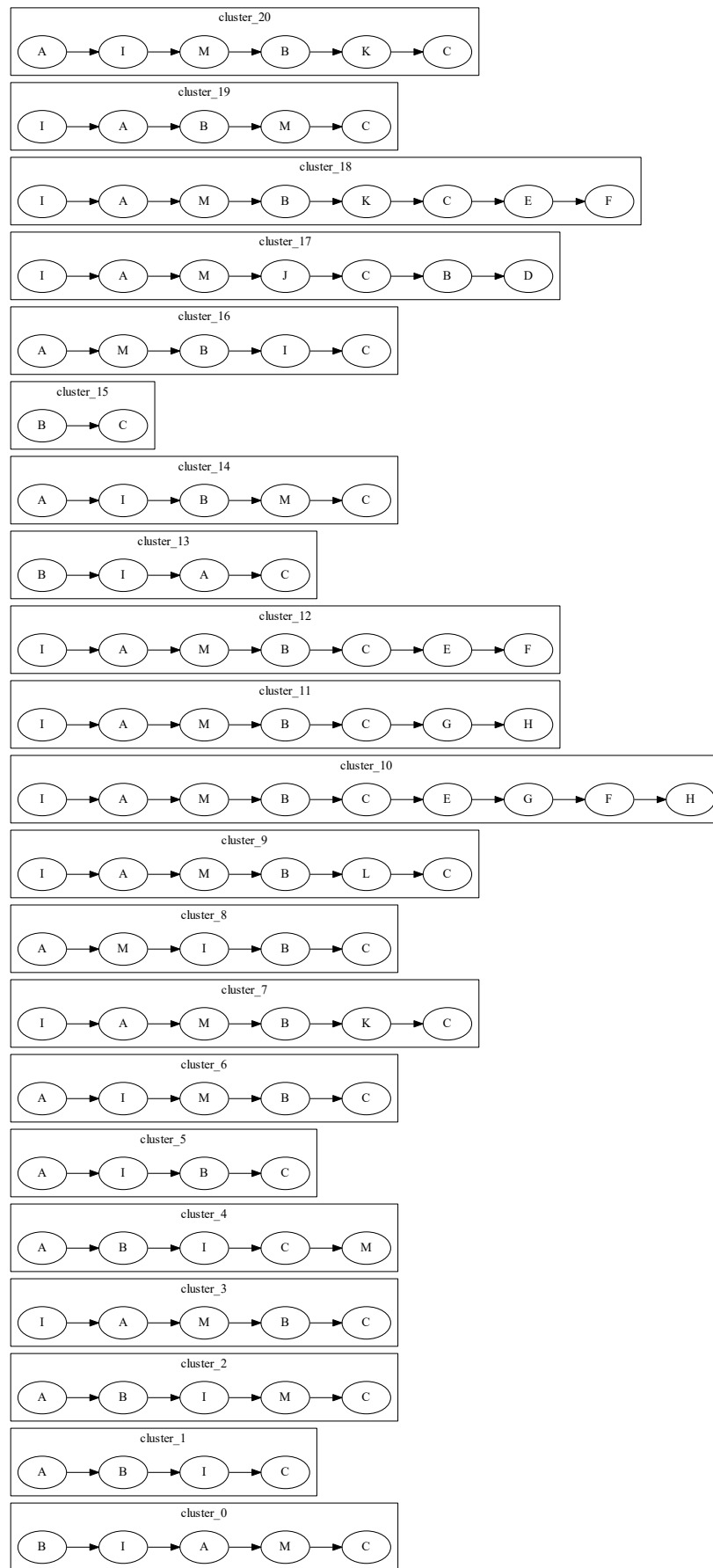


Figure 5.6: Network Graph of the Raw 21 *Process Medoids*.

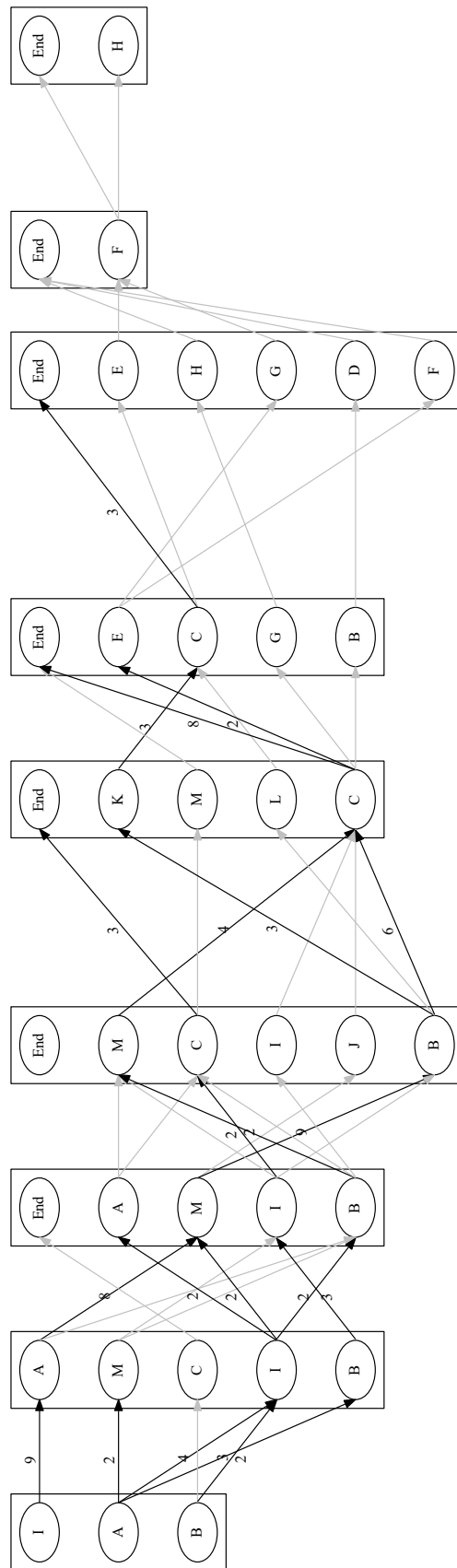


Figure 5.7: Network Graph of the Linked 21 *Process Medoids*.

5.5 Routing Procedures Exploration

As all inputs have now been discussed, the simulation for each routing procedure can be built, run and compared. This section will recap all the information regarding the simulation build and give an overview of each model applying the STRESS guidelines to support reporting [202, 271].

5.5.1 Overview

Background

The clinical pathway for lung cancer is explored for 1,865 individuals covering a period of 362 days. The input data was of DT1, and therefore the ‘at most once’ constraint is in place. The motivation is to define the clinical pathway and explore how capacity aligns with demand. The aim is to automatically build the simulation model from the input data. The objective is to explore the various definitions of the clinical pathway.

Reported Results

Four definitions for the clinical pathway are explored, *Raw Pathways*, *Full Transitions*, *Cluster Transitions* and *Process Medoids*, which have previously been described in detail. To allow for comparison, three main tables are used for reporting as follows:

- Top Level Results - In addition to the previously discussed results (from page 103), results supporting pathway comparison are reported. Namely, number of unique pathways (Unique), number of pathways performed once (Once), number performed more than once (Once More), total transitions, mean transitions, largest transition, day last arrival into system. The total, mean and largest transitions are calculated by comparing the original transition matrix with the simulation transition matrix, not including the start and end activities.

- Frequency of Activity - This counts the number of times each activity was performed.
- Activity Waiting time - The mean waiting times for each activity is reported (as on page 103).

Base Model and Scenarios

The *Raw Pathways* model was used for validation of the input parameters. The remaining three routing procedures can be considered scenarios, where different interpretations for the clinical pathways with the intention to reduce variation are explored. Furthermore, due to the ‘at most once’ constraint the *Full Transitions* and *Cluster Transitions* routing procedures are not suitable for this data, however they have been included for comparison.

Parameters

The general parameters were summarised in Section 4.8, where the specifications for each routing procedure have been discussed in Sections 5.2 – 5.4. Specifically Table 4.16 ‘Pattern’ column describes the capacity applied, and additionally a Warm Start using the Warm 2 procedure was applied (see subsection 4.5.5).

A First In First Out (FIFO) prioritisation was used for the queuing discipline, as default in Ciw. Furthermore, in these models the time unit definition in Ciw is that 1 represents one day.

Trials were run with 25 runs for each model, where the model runs until 1865 customers have exited the system³. The seed for each run was changed each time and was representative of the run number (starting at 0). This was implemented for each routing procedure (scenario) and as such the first run for each of the models had seed 0 etc. The definition of the seed was supported by Ciw’s *ciw.seed()* function [60].

³Recall for the *Raw Pathways* that additional individuals will not enter the system after 1865 individuals, resulting in the final few individuals not having to compete for resources.

Implementation

The models were run on a Windows 10 operating system. The open source code for Sim.Pro.Flow was applied through a Jupyter Notebook to support more in depth reporting of the results tables. Python v3.6. was used and all other libraries and versions are as stated in the requirements file for Sim.Pro.Flow [255].

Each run took at most 5 seconds to complete, however the conversion applied to format the results is time consuming. The overall time taken to run all the models was not recorded.

Although the Sim.Pro.Flow interface was not used, the methods and code from within Sim.Pro.Flow was applied. If the same dataset was available, the models should be able to be recreated in Sim.Pro.Flow with ease due to the automated build feature of the software. The only adjustment required by the user would be to include the Warm 2 Warm Start.

5.5.2 Results and Discussion

The results for the models, as discussed above, are reported in Table 5.12 – 5.14. The *Raw Pathways* model was used to validate the input parameters and as such produces results similar to the original. The following discussion answers two main questions to analyse the results of the remaining models, noting the purpose of the comparison is to display the capabilities of the models and note considerations for when using each routing procedure.

1. *How do the activity waiting times change with the different routing procedures?*

In general, the activity waiting times are lower for the other routing procedures compared to the *Raw Pathways*, which also results in lower mean and median total waiting times. This indicates that with differing levels of variation the capacity should be sufficient, although capacity could be investigated further (see Chapter 7 for *Raw Pathways* and *Process Medoids*). Furthermore, it is important to consider how reflective the pathways are of the real life system.

2. *Are the pathways produced reflective of the real life system?*

Recall for this working example that the *Full Transitions* and *Cluster Transitions* are not compatible routing procedures for this data type as they do not hold the ‘at most once’ constraint. Considering the frequency of activities (Table 5.13) these are similar to the *Raw Pathways*, which indicates similar pathways are produced. However, in the top level results (Table 5.12) the number of pathways that occur more than once are relatively similar to the original data, however the number that appear once are much higher. This suggest that there is a larger number of different pathways being produced here than in the original data. This could be investigated further through observing the top 10 most occurring pathways. However running trials increases the difficulty of reporting the results, as these will change on each run for some of the routing procedures. The results for the first run (seed 0) could be compared to allow for deeper analysis. Furthermore, comparing the total transitions, this is lower for the *Cluster Transitions* indicating that they display less variation than the *Full Transitions*, which was the aim.

Alternatively, the purpose of the *Process Medoids* routing procedure was to address the ‘at most once’ constraint and reduce the variation whilst still reflecting the *Raw Pathways*. Again, considering the frequency of activities (Table 5.13), the *Process Medoids* have a large decrease in frequencies for activities ‘D’, ‘J’ and ‘L’, however the most frequently occurring activities (‘A’, ‘B’, ‘C’, ‘I’ and ‘M’) all have values close to the original data. This displays that the variation has been reduced, whilst considering that the medoids only select pathways that have previously happened, also displays that they are reflective of the real life system, and can be used to investigate scenarios for a clinical pathway with reduced variation.

Table 5.12: Top Level Results for Routing Procedures with Automated Capacity.

Level	Mean TiS	Median TiS	Target	Unique	Once	Once More
Original	60	41	64.4	783	576	207
Raw	65.08 (63.34, 66.82)	56.72 (54.4, 59.04)	53.12 (50.34, 55.9)	783.0 (783.0, 783.0)	576.0 (576.0, 576.0)	207.0 (207.0, 207.0)
Full	34.96 (33.05, 36.87)	29.68 (27.64, 31.72)	85.39 (83.71, 87.08)	1312.56 (1306.27, 1318.85)	1150.04 (1143.09, 1156.99)	162.52 (159.91, 165.13)
Cluster	36.92 (35.1, 38.74)	33.28 (31.37, 35.19)	86.86 (85.17, 88.55)	1227.28 (1220.14, 1234.42)	1049.56 (1040.46, 1058.66)	177.72 (174.77, 180.67)
Medoids	35.12 (33.42, 36.82)	35.68 (33.97, 37.39)	98.71 (98.51, 98.91)	21.0 (21.0, 21.0)	0.0 (0.0, 0.0)	21.0 (21.0, 21.0)
Level	Total Transitions	Mean Transitions	Largest Transition	Day Last Arrival	Overall Period	
Original	0	0	0	354	362	
Raw	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	360.89 (358.09, 363.68)	407.66 (407.46, 407.86)	
Full	3142.8 (3108.9, 3176.7)	17.27 (17.08, 17.45)	287.64 (280.17, 295.11)	360.24 (356.23, 364.25)	387.07 (384.53, 389.61)	
Cluster	2498.64 (2471.65, 2525.63)	13.73 (13.58, 13.88)	169.8 (164.83, 174.77)	357.22 (354.06, 360.37)	387.44 (385.8, 389.09)	
Medoids	6512.52 (6480.79, 6544.25)	35.78 (35.61, 35.96)	633.96 (628.54, 639.38)	360.94 (358.26, 363.61)	391.5 (391.05, 391.95)	

Table 5.13: Frequency for Activity for Routing Procedures with Automated Capacity.

Activity	Original	Raw Pathways	Full Transitions	Cluster Transitions	Process Medoids
A	1797	1797.0 (1797.0, 1797.0)	1719.4 (1703.87, 1734.93)	1741.88 (1731.92, 1751.84)	1802.08 (1799.18, 1804.98)
B	1865	1865.0 (1865.0, 1865.0)	1769.0 (1754.61, 1783.39)	1791.92 (1781.79, 1802.05)	1865.0 (1865.0, 1865.0)
C	1855	1855.0 (1855.0, 1855.0)	1819.52 (1809.11, 1829.93)	1823.96 (1811.29, 1836.63)	1865.0 (1865.0, 1865.0)
D	232	232.0 (232.0, 232.0)	229.12 (225.27, 232.97)	227.64 (220.41, 234.87)	64.56 (61.5, 67.62)
E	473	473.0 (473.0, 473.0)	463.72 (457.77, 469.67)	469.72 (461.67, 477.77)	194.16 (189.32, 199.0)
F	475	475.0 (475.0, 475.0)	470.72 (463.07, 478.37)	471.04 (464.8, 477.28)	194.16 (189.32, 199.0)
G	536	536.0 (536.0, 536.0)	527.2 (517.42, 536.98)	525.12 (517.58, 532.66)	222.8 (217.55, 228.05)
H	537	537.0 (537.0, 537.0)	531.08 (521.4, 540.76)	527.88 (521.07, 534.69)	222.8 (217.55, 228.05)
I	1797	1797.0 (1797.0, 1797.0)	1727.04 (1711.07, 1743.01)	1738.4 (1727.17, 1749.63)	1802.08 (1799.18, 1804.98)
J	537	537.0 (537.0, 537.0)	525.64 (517.42, 533.86)	528.52 (518.39, 538.65)	64.56 (61.5, 67.62)
K	589	589.0 (589.0, 589.0)	559.96 (552.08, 567.84)	564.32 (555.98, 572.66)	169.76 (165.11, 174.41)
L	364	364.0 (364.0, 364.0)	347.92 (342.18, 353.66)	351.72 (344.41, 359.03)	72.6 (69.46, 75.74)
M	1577	1577.0 (1577.0, 1577.0)	1512.2 (1496.15, 1528.25)	1523.84 (1513.55, 1534.13)	1496.4 (1488.59, 1504.21)

Table 5.14: Activity Mean Waiting Time Results for Routing Procedures with Automated Capacity.

Activity	Original				Raw Pathways	Full Transitions	Cluster Transitions	Process Medoids
	Mean	25 th Percentile	Median	75 th Percentile				
A	12.52	3.0	7.0	15.00	23.25 (21.61, 24.9)	16.9 (14.71, 19.09)	21.07 (19.36, 22.78)	24.99 (23.37, 26.61)
B	11.86	2.0	7.0	14.00	9.64 (9.39, 9.88)	8.56 (6.93, 10.19)	7.14 (5.93, 8.35)	7.4 (6.97, 7.83)
C	9.40	4.0	7.0	11.00	1.97 (1.9, 2.05)	1.32 (1.23, 1.41)	1.34 (1.24, 1.45)	1.2 (1.18, 1.22)
D	21.91	0.0	12.0	38.00	0.73 (0.69, 0.77)	1.28 (1.09, 1.47)	1.25 (1.1, 1.4)	0.22 (0.15, 0.29)
E	11.19	4.0	7.0	15.00	7.53 (7.41, 7.64)	0.64 (0.57, 0.71)	0.67 (0.59, 0.76)	0.06 (0.05, 0.07)
F	20.20	5.0	12.0	24.50	18.29 (18.11, 18.47)	2.15 (1.83, 2.47)	1.88 (1.68, 2.09)	0.22 (0.17, 0.27)
G	6.41	0.0	1.0	7.00	1.42 (1.36, 1.48)	0.67 (0.56, 0.79)	0.62 (0.54, 0.7)	0.06 (0.05, 0.06)
H	40.21	6.0	18.0	70.25	34.04 (33.89, 34.19)	18.81 (17.14, 20.49)	17.97 (16.52, 19.42)	5.89 (5.4, 6.38)
I	4.05	0.0	0.0	2.00	1.51 (1.38, 1.65)	1.87 (1.59, 2.15)	1.52 (1.35, 1.69)	1.65 (1.5, 1.8)
J	13.66	6.0	12.0	17.00	14.02 (13.83, 14.21)	1.03 (0.92, 1.14)	0.92 (0.76, 1.07)	0.01 (0.0, 0.02)
K	3.58	0.0	0.0	2.00	16.58 (16.38, 16.77)	0.48 (0.43, 0.53)	0.49 (0.45, 0.52)	0.03 (0.02, 0.04)
L	3.85	0.0	0.0	0.00	5.36 (5.19, 5.52)	0.46 (0.41, 0.51)	0.5 (0.45, 0.56)	0.01 (0.01, 0.02)
M	3.20	0.0	0.0	2.00	3.02 (2.93, 3.11)	2.05 (1.65, 2.45)	1.9 (1.53, 2.28)	0.66 (0.6, 0.71)

5.6 Conclusion and Further Work

In conclusion, this section combines the resulting methods from Chapter 3 and 4 by simulating various definitions of the clinical pathway. This chapter describes three routing procedures (*Full Transitions*, *Cluster Transitions* and *Process Medoids*) and discusses their adaptations of the simulation build. The simulation for each routing procedure is run and the results are discussed.

Overall the results show that the *Raw Pathways* can be used for validation of the input parameters, as the model is reflective of the original data. Although not suitable for this data type and not producing pathways similar to the real life system, the *Full Transitions* and *Cluster Transitions* do produce similar frequency of activities as the original data. Furthermore, the *Cluster Transitions* do display less variation than the *Full Transitions* as intended. Finally, the *Process Medoids* do display less variation than the *Raw Pathways*, whilst producing pathways reflective of the real life system.

It can be seen that all three alternative routing procedures reduce the time in system and the activity waiting times⁴. Therefore, further investigation should explore how to adjust the capacity appropriately.

Further work

There are a few potential areas for further work as follows:

- Exploring other routing procedures using the ideas from the *Full Transitions* and *Cluster Transitions* but continuing to reduce complexity and minimise variations i.e.:
 1. Defining a routing function that uses the transition matrix, but forces the length of the pathway to be between a lower and upper bound, could result in more realistic pathways, and possibly produce less variation.

⁴Care needs to be taken here, as the *Full Transitions* and *Cluster Transitions* should not be considered for this model (and were just included for demonstration).

2. Exploring a transition matrix that considers the transitions between the groups of activities (as defined during the clustering), where once the individual moves to a group they are either randomly or proportionally assigned an activity. This could be taken further with the specific activity performed suggesting the probability of the next group to move to.
- Explore a more optimal level of capacity for each of the routing procedures, by allowing the target capacity method (Section 4.7) to be used with the pathways resulting from the simulation.

Chapter 6

Developing a Decision Support Tool - Sim.Pro.Flow



SIMULATE PROCESS FLOW

Figure 6.1: Sim.Pro.Flow Logo.

6.1 Introduction

Research Question 4

Can the development of a decision support tool provide a general method of analysing clinical pathway mapping, modelling and improving?

One of the main desired outcomes from this research for the company partner (Velindre Cancer Centre), and research question 4, was the development of a decision

support tool. Throughout the programme of research, when selecting the methods to use, they not only had to be technically appropriate but also suitable to be integrated into the eventual decision support tool. This thesis has so far described how mapping, modelling and improving the clinical pathways have been addressed:

- Mapping: Two methods have been discussed, firstly a simple direct extraction of pathways from data (*Raw Pathways* and *Full Transitions*), and secondly through the application of clustering (*Clustered Transitions* and *Process Medoids*).
- Modelling: Automated simulation build has been developed to further explore the mapped pathways.
- Improving: Capacity investigations, through either the technical method or basic principles.

The decision support tool named Sim.Pro.Flow was designed and developed, which is described further through answering the four main questions:

What? Why? Who? and How?.

After this initial discussion, the remainder of the chapter is structured as follows:

- Section 6.2 introduces the structure of version 2.1 of Sim.Pro.Flow.
- Section 6.3 discusses the key features of Sim.Pro.Flow v2.1.
- Section 6.4 concludes the chapter through evaluating the build process and highlighting further work.

What - is Sim.Pro.Flow?

Sim.Pro.Flow (Simulate Process Flow) is a decision support tool that automates the build of a discrete event simulation and allows for mapping, modelling and improving of system pathways. There are three main particular novel contributions of Sim.Pro.Flow:

1. To allow the methods to seamlessly interact with each other (Chapter 2)

2. To allow for the input of user information and interaction (Chapter 3)
3. To support timely production of pathway analysis (Chapter 4 and 5)

Why - develop a decision support tool?

There are popular tools and software available that could have been utilised for the basic ideas applied in this research. For example, ProM [235] is a popular choice for pathway mining, whereas Simul8 [257], Arena [9] and AnyLogic [8] all support discrete event simulation. However, these all host individual aspects of the clinical pathway mapping, modelling and improving processes, and cannot be integrated together to form one cohesive system. This was a key element identified in the literature review (Chapter 2) and as such developed into one of Sim.Pro.Flow's novel contributions.

Furthermore, more recently (2020) the tool PathSimR [222] was developed, which provides a user interface in R [236] for discrete event simulation for healthcare pathways [223]. PathSimR has the user manually enter information of the system, such as service points, exit points, transitions and input parameters (arrivals, service and capacity), however it does appear that the depiction of the network is then generated by the tool [224].

Palmer et al., [219] presents Ciw's debut paper which displays a comparison with some of the aforementioned packages. Furthermore, Palmer et al., [219] discusses how simulation packages with GUI's are "*more accessible, easily modifiable and easier to communicate with non-specialists*". Conversely, it is also stated that those packages also have disadvantages, noting low model reusability, and can lead to bad simulation practice through bias by "*building models that represent how a system should work instead of how they actually work*". As Ciw was the simulation method chosen, it was important to respect and address these statements.

Palmer et al., [219] directs the reader towards Bell and O'Keefe [26] who in 1986 discussed Visual Interactive Simulation (VIS) in depth, noting that VIS was first

presented by Hurrion in 1976 [139,140]. Bell and O’Keefe [26] discuss how guidelines for building VIS began to develop, and summarised the work of Bell [25] into guidelines for good practice as follows:

- *“Get the user involved as early as possible”*
- *“Get the picture up as soon as possible”*
- *“Make the interaction as general as possible”*
- *“Try to transfer the simulation to the end user”*

These guidelines, along with their further explanation, will be discussed in the conclusions (Section 6.4) to retrospectively evaluate Sim.Pro.Flow’s functionality.

Who - are the intended users?

Motivated by research question 1 and the problem description (Section 1.3), integrating expert knowledge and increasing user interaction became one of the main focuses for the tool. Therefore, the tool was built with the intention of allowing experts to impart their knowledge whilst interacting through the tool. This could be through individual direct usage, or a collaborative effort with a technical professional on either a one-on-one or workshop basis.

It was intended to trial all three of these potential methods of usage, however due to the COVID-19 global pandemic these could not go ahead.

The intended methods of use were recognised in design decisions, to allow the interaction to be as easy as possible. For example, the preprocessing of the input data allows for a smooth transition and creating the summary sheet was included to allow the user to become familiar with the data before delving deeper.

How - was Sim.Pro.Flow developed?

The emergence of the field of Design Science is credited to Herbert Simon in 1969 [250], and has become a point for best practices when developing an artefact. Specifically Hevner et al., [121] develop seven guidelines for design science

research, and ultimately stresses that there must be a balance between the abilities of technological artefacts and theoretical foundation. Although this was not a formal design science project, the interested reader may wish to consult [121,217,250] for more information.

The main consideration for the technical development of Sim.Pro.Flow was accessibility. Palmer et al., [219] notes how Ciw presents a main quality of open accessible code with a MIT license, noting this as an advantage over other software, as the user can inspect and modify the code as necessary without requiring license fees. Therefore, Sim.Pro.Flow and all its code is also open access available on Github [251] with an MIT license.

To support this consideration, the tool itself had to be built using components which are also open access and have suitable licences. Therefore the main library providing the GUI is wxPython [318] as it satisfies this consideration. Furthermore, every effort was made to account for all resources utilised to develop the tool (along with noting licence agreements), and are listed within the file References.txt [254].

The tool was built alongside discovering and learning both Python and the art of developing open source software. Therefore, best practices may not have always been adhered to and the code might not be as efficient as possible. However, given the desire for the code to be open source, access considerations were given to the file structure and modularising the code as follows:

- The main file, *App.py*, hosts the GUI code.
- The file *Functions.py* acts as a management system to direct the main file to the source code in the src folder.
- The src folder contains nine topic specific code files e.g. *clustering.py* contains the clustering code, *simulation.py* contains the simulation code etc.

The tool evolved along side the theoretical development, along with informal feedback from a variety of potential users was collected and considered.

6.2 Structure

The Sim.Pro.Flow v2.1¹ GUI is set out to reflect the natural method of reading a book, moving top to bottom then turning the page, moving towards the right. This is intended so that the progression would feel logical and need little explanation. This allowed for the development of the four main tabs: ‘Data’, ‘Clustering’, ‘Simulation’ and ‘Visualisation’ (see Section 6.3). Figure 6.2 displays the four main tabs, in hierarchical progression of the intended full usage (top down).

In Figure 6.2 each box contains the components contained on the main tab, and the movement to subsequent sub-tabs. The arrows between the main tabs and sub-tabs shows the movement of information allowed. The only strict usage is that the user must enter through the data tab and complete the first three points (‘Save Location’, ‘Select Original Data’ and ‘Choose Activities’) before progressing (with the exception of setting the target days 6.3.2). Table 6.1 describes some of the possible movements supported.

Table 6.1: Examples of Supported Movement Through Sim.Pro.Flow.

Analysis Description	Path
View raw data results only	‘Data’ >> ‘Simulation Results’ subtab
View graphics for raw data only	‘Data’ >> ‘Visualisation’
Run capacity analysis	‘Data’ >> ‘Capacity’ subtab
Perform clustering only	‘Data’ >> ‘Clustering’ <i>*Optional ‘Rankings’ and ‘Groupings’ subtabs if MNW selected*</i>
Explore <i>Raw Pathways</i> simulation	‘Data’ >> ‘Simulation’ <i>*Optional additional ‘Simulation Results’ subtab and/or ‘Visualisation’ tab*</i>
Explore <i>Cluster Transitions</i> simulation	‘Data’ >> ‘Clustering’ >> ‘Simulation’ <i>*Both previous optional additions available*</i>

If the user wishes to perform a complete analysis, then using all four tabs in the logical progression would suffice. However some users may only want to use ‘Data’ and ‘Simulation’ (*Raw Pathways* or *Full Transitions*), or ‘Data’ and ‘Visualisation’. This was considered during the tool development and thus the user can easily switch between tabs/subtabs at various stages, as shown within the usage map (Figure 6.2).

¹Note that v2.1 is a prototype phase.

This flexibility makes the tool multi-functional as it can be used to support pathway analysis in multiple ways. Furthermore, although the tool itself can be fully used as a stand-alone tool, it can also be used as part of a wider ‘toolkit’. For example, after initial analysis of the pathways, the user could explore the data manually, or taking the produced simulation and replicating it in other packages to explore features that Sim.Pro.Flow may not support.

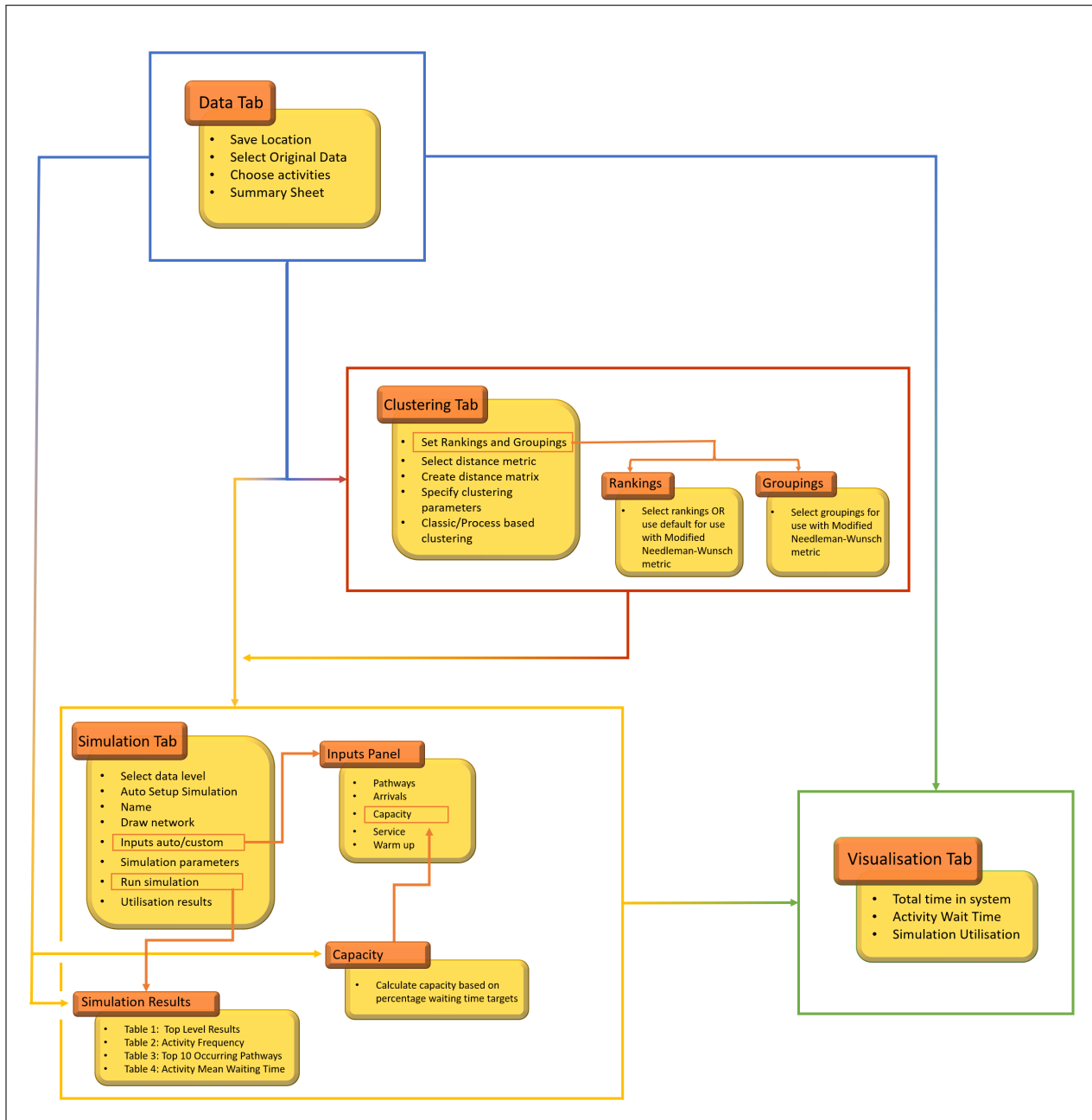


Figure 6.2: Usage Map for Sim.Pro.Flow.

6.3 Sim.Pro.Flow Key Features

6.3.1 Outputs Produced

Throughout the development of Sim.Pro.Flow it became apparent that to fully understand and analyse such complex pathways a variety of information would be required. This was also noted by the initial informal feedback from VCC staff. Whilst building and using the tool, the question ‘what does the user need to know - both fundamentally and for deeper exploration?’ was continually asked. Once the need identified, the question turned to ‘how can this be automatically produced?’ and ‘how can it be displayed?’.

The outputs are either automatically or instructively (clicking a button) generated and are all saved within the selected save location and its subsequent folders. All of the outputs can be viewed from the file explorer save location and are saved in popular formats either Excel, Word, PDF or PNG. Some of the outputs can be accessed via Sim.Pro.Flow, e.g. specific plots can be viewed in the plot viewer panels, or will open the external file in its default software. The five main types of outputs are discussed below. Furthermore, a detailed log of how the output is generated, where it is saved and what it contains is displayed in Appendix D Table D.1.

1. Summary Sheet: The summary sheet is a Word document comprised of general summary information about the data (example in Appendix D). The document includes a reference list for the activity to letter assignments, along with top level information about the number of pathways and waiting time statistics. This also contains four figures displaying: 1) the frequency of each activity 2) heatmap of all pathways 3) histogram of the total time in system and 4) boxplot of activity waiting times.
2. Plots: All plots generated are saved as PNG files. There are three points at which plots are automatically generated as follows:

- Capacity investigation: A combined plot of line graphs (e.g. Figure C.3) with subplots per activity that the calculation was performed for.
 - *Process Medoids* Violin Plots: A combined plot with subplots in the range $[2, \text{max_k}]$ containing violin plots displaying the non-zero values from the difference matrix (e.g. Figure 5.3).
 - Simulation: Multiple plots are produced which will be viewable in the visualisation tab (see subsection 6.3.2). Furthermore, multiple plots containing the standard deviation of the values within the key results tables, to support the selection of number of runs are produced.
3. Excel: There are many points where Excel files are created e.g. simulation results, clustering results etc, (See Appendix D Table D.1 for more detail). The purpose of generating the Excel documents were three fold.
- (a) Avoid unnecessarily creating this interface within Sim.Pro.Flow itself.
 - (b) Allow further exploration for the user to gain a deeper understanding.
 - (c) To act as a recording mechanism.
4. Raw Variables: The raw variables for the simulations can be saved as Python code. This can also allow further exploration through Python directly if the user wishes.
5. Network Diagrams: These visualisations of the networks were discussed and displayed in Chapter 5. Producing these visualisation were considered a key design feature, as it allows the user to interact with a visual depiction of the network (consider the guidelines from Section 6.1). This also brings Sim.Pro.Flow more inline with alternative software that produce animation. To align with research question 4, it was necessary for the diagrams to be automatically generated, without the need for user interaction to either ‘draw’ or ‘rearrange’ the diagram. This was possible through using the Python library Graphviz [113] and specifically digraph was used to produce these diagrams.

6.3.2 Graphical User Interface

Taking the four main panels in turn, this section notes the key features available. For note, Figures 6.3 – 6.6 (and those in Appendix D) are extracted from the Sim.Pro.Flow help document [252], and the key features are edited from the discussion within the help document (which contains more detail and user considerations).

Data Panel

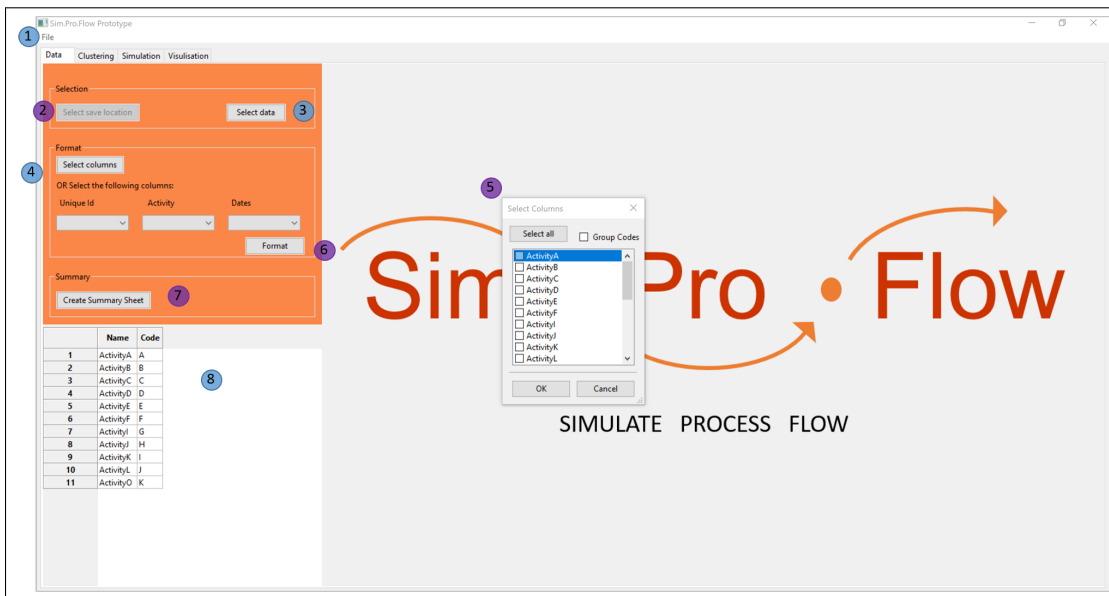


Figure 6.3: Sim.Pro.Flow - Data Panel.

- **Selection:** The user can choose the folder to save the information for their session and select the dataset to use.
- **Format:** This supports DT1, DT2 and DT3, where the user can either select the columns containing the activities they want to explore (DT1) with additionally selecting to group the codes (DT3) or select the three columns required for DT2 (see subsection 1.2.4).
- **Summary:** When selected, a Word document containing summary information on the input data is created (example in Appendix D).
- A table displays the activity name and corresponding code for the activities used in this session, as a point of reference for the user.

Clustering Panel



Figure 6.4: Sim.Pro.Flow - Clustering Panel.

- Modified Needleman-Wunsch: Supports the use of the MNW, through allowing selection of the penalty values and additional subtabs where the user can enter the rankings and groupings (Appendix Figure D.2).
- Distance Matrix: Allows for selection of the distance metric to use, where the eight metrics (plus MNW) discussed in Chapter 3 are supported. On creating the distance matrix a progress bar is displayed to keep the user informed.
- Cluster: Contains options available for clustering:
 - Centroids: Five methods supported for selecting the initial centroids.
 - Results: Selects the level of results which will be displayed in a pop out window (Appendix Figure D.1), and select the value for k .
 - Process based: Allows for selection of the adjust percentage and tolerance (see Section 5.4).
 - Save check box: Saves information on the clustering for further analysis
- Visualisation panel: Hosts the process based violin plots (e.g. Figure 5.3).

Simulation Panel



Figure 6.5: Sim.Pro.Flow - Simulation Panel.

- Setup: Select the routing procedure, automatically build the simulation and allow specification of the simulation name.
- Network: Draws the network for the selected routing procedure and allow them to be viewed in an external PDF viewer.
- Simulation Inputs: Select either the 'Auto' or 'Custom' options for the simulation inputs. The right hand panel will be populated with five tabs displaying the input parameters ('Pathways', 'Arrivals', 'Service', 'Capacity' and 'Warm up' - see Appendix Figure D.3), enabling the user to interact with the simulation through custom values as described in Section 4.6.
- Time: Contains options available concerning simulation times, such as number of days a week, the target number of days for total waiting time, number of trials and simulation seed.
- Utilisation: Pop out window containing capacity utilisation results, including the percentage of days that used 100% of the capacity, and the mean average percentage of capacity used on each named day (Appendix Figure D.4).
- Capacity (Figure D.5) and Simulation Results (Figure D.6) sub-tabs available.

Visualisation Panel

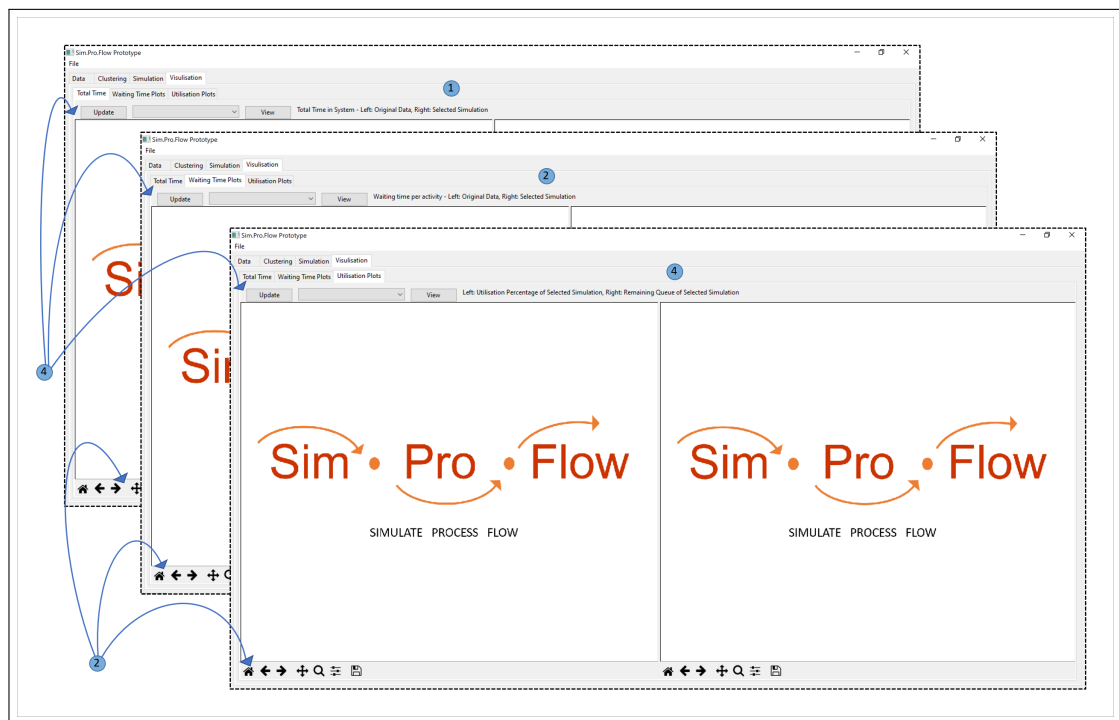


Figure 6.6: Sim.Pro.Flow - Visualisation Panel.

- Total Time Sub-tab: Displays a histogram for total time in system, where the left canvas is for the original data (e.g. Figure C.1) and the right canvas is for the selected simulation.
- Waiting Time Sub-tab: Displays a plot containing multiple histograms displaying the waiting time for each activity, where the left canvas is for the original data (e.g. Figure C.2) and the right canvas is for the selected simulation.
- Utilisation Sub-tab: Displays two line charts of the capacity utilisation, where the left canvas plots the percentage of the capacity utilised each day (e.g. Figure D.7) and the right canvas plots the number of individuals remaining in the queue at the end of each day (e.g. Figure D.8).

6.4 Conclusion

Returning to the points made in the introduction regarding the guidelines for good practice by Bell [25] and summarised by Bell and O’Keefe [26], these were initially established for the design process. However, they also reflect the intentions for the interactions of use, and can be used to evaluate the decision support tool as discussed in Table 6.2.

Table 6.2: Evaluation of the Development of Sim.Pro.Flow.

Guidelines for good practice by Bell [25] and summarised by Bell and O’Keefe [26]

Guideline	Development	Interactions
<i>“Get the user involved as early as possible”</i>	Initial informal feedback on the tool was collected as noted in Section 6.1. Intentions to more fully progress this were cancelled due to the COVID-19 global pandemic.	The initial creation of the data files and the summary sheet reflect this guideline by providing places of interaction for the user.
<i>“Get the picture up as soon as possible”</i>	The tool was developed alongside the mathematical model (guidelines suggest design before model), although the intentions for interaction were considered in the mathematical model.	The generation of supporting materials, in particular the network diagrams reflects this guideline.
<i>“Make the interaction as general as possible”</i>	The cognitive progression through the tool was a major design decision - to allow the user’s natural instincts i.e. the way to read a book, guide design placement. General use was another major consideration as the specifics of the input data are an unknown factor and had to be considered.	Some effort was directed towards predicting the user’s interactions, through efforts such error and information messages to inform the user of actions that were not desired.
<i>“Try to transfer the simulation to the end user”</i>	This guideline emphasises a VIS that can be used by the user on a regular basis. This was another major aspect that was desired, and motivated the automation process. Practically this is also reflected by the tool being hosted open access on Github [251].	This was reflected through allowing the user flexibility using the custom simulation inputs. Allowing the user to guide aspects of the simulation as required.

It should be commented that the general interpretation of the data and tool design is a major advantage. For the company partner (Velindre Cancer Centre) this means they now have a decision support tool that they can use for any of their cancer sites (e.g. lung, breast etc) rather than one in-depth model. Furthermore, this generalisation feature also widens the applicability and usability of the tool to not just cancer, not just healthcare even, but any process data that satisfies one of the input data types.

Further work

As the tool is still in prototype phase (with v2.1) there are two specific areas of further work that could be considered:

- Usability: Ensuring the user cannot perform ‘undesirable’ actions and that the code runs error free. There is a list of current issues listed on the Sim.Pro.Flow Github page [253].
- Accessibility: Developing a formal testing module to ensure the tool downloads and runs smoothly and correctly.

The current help document available is comprehensive, however this could be developed further with user feedback of missing/inadequate information. Furthermore, although a single video has been made to overview the functionality of the tool, a series of shorter ‘How to?’ videos could be produced to help modernise the support materials and guide the user on specific functions.

Finally, to further enable ease of use, an execution button could be created so that Sim.Pro.Flow can be launched through ‘double-click’ to run the application, like any other software, making the use of the command line redundant. This would help with accessibility for anyone not proficient in command line use.

In conclusion, this chapter begins to address research question 4, by combining all the methods developed in the previous chapters into one decision support tool. The demonstration to support research question 4 is continued in Chapter 7.

Chapter 7

Case Study

7.1 Introduction

Research Question 4

Can the development of a decision support tool provide a general method of analysing clinical pathway mapping, modelling and improving?

This chapter presents a case study focussing on the lung cancer pathway. The purpose is two fold, firstly, to explore if Sim.Pro.Flow is able to support typical exploration of the simulation (research question 4), and secondly to gain deeper insights into the specific lung cancer pathways explored.

There are seven investigations discussed in this section, where the first six explore the *Raw Pathways* model and the final explores the *Process Medoids* solution. A brief description of each investigation is as follows:

1. **Individual Adjustment Investigation:** The aim of this investigation is to explore the capacity levels to obtain more accurate waiting times for activities which displayed discrepancies with the original data.
2. **Basic Investigation:** Progressing previous findings, this section investigates

an instinctive¹ initial exploration of how to improve the top level and waiting time results through adjusting the capacity levels.

3. **End Activity Investigation:** A systematic approach is used to focus on improving the waiting times for end of pathway activities.
4. **Target Investigation:** This investigation explores what levels of capacity are required to achieve the 95% within 62 days target, including whether it is even achievable. The method applied uses the previous selected capacity levels and then increases the capacity to achieve the target.
5. **Excessive Top Down Investigation:** The aim here is the same as the previous investigation, however, an alternative method is applied. This explores starting with extreme excessive levels of capacity with systematically and incrementally decreasing the capacity to achieve the target.
6. **Demand Investigation:** Conversely, this examines the consequences of increased demand on the capacity levels of a chosen capacity pattern.
7. **Medoids Capacity Investigation:** Considering the *Process Medoids* solution (Section 5.4). This explores what levels of capacity would be required if a fixed set of pathways were implemented.

The standard simulation models used were those chosen as a result of Section 5.5, and thus the input parameters (pathways, arrivals, service, capacity and warm up) are as summarised in Section 5.5.

For each investigation the results tables are the same as described in Section 4.4. Furthermore, ‘Original’ refers to the summary results of the original data and ‘Standard’ refers to the basic model as previously described. Colour coding of table cells is used to highlight the attention of the reader, where **orange** indicates a progressive change, **teal** indicates a regressive change and **purple** indicates an unintended resulting change.

¹This is reflective of a casual reactive exploration.

In general the process for each investigation followed the same format. Initial experimentation was performed using the interface of Sim.Pro.Flow, where one run of each change was performed. This was then replicated in a Jupyter notebook using the code of Sim.Pro.Flow, and one run was performed to ensure that the front (interface) and back end (code) of Sim.Pro.Flow were producing the same results. Finally, twenty-five runs of each change were performed in the Jupyter notebook to gain confidence interval results. This was done to allow multiple investigations to run simultaneously for efficient use of time (see Section 7.9 for explanation). There are two exceptions to this methodology, namely the Individual Adjustment and Demand Investigations, where only the Jupyter notebook was performed (see respective sections for details).

This chapter concludes with a summary of the results of each investigation to allow for comparisons and discussion.

7.2 Individual Adjustment Investigation

Subsection 4.5.5 concluded that using the standard capacity levels along with the Warm Start approach (Warm 2) was sufficient to achieve satisfactory top level results (Table 4.20). However, the specific activity waiting times (Table 4.22) did display some large discrepancies when comparing the standard model with the original data. Table 7.1 displays the activity waiting times and highlights the activities with the largest discrepancies. Therefore, the purpose of this investigation was to achieve more accurate activity waiting times and observe the effects on the top level results.

As highlighted in Table 7.1 activities A², C, D, G, I and K were chosen to examine, where C, D, G, I were producing results too small and A, K were producing results too large. Multiple capacity levels were tested for each activity individually as shown in Table 7.2 - detailed results for which are included in Appendix E (Table

²Note that throughout the thesis activity letters and pathways have been highlighted by single quotation marks. This format is not adhered to in this chapter to allow for easier reading.

E.1 – E.3). In each case a capacity level (highlighted in green in Table 7.2) was chosen that improved upon the discrepancy.

Table 7.1: Original and Standard Activity Waiting Times.

Activity	Original	Standard
A	12.52	23.25, (21.61, 24.9)
B	11.86	9.64, (9.39, 9.88)
C	9.40	1.97, (1.9, 2.05)
D	21.91	0.73, (0.69, 0.77)
E	11.19	7.53, (7.41, 7.64)
F	20.20	18.29, (18.11, 18.47)
G	6.41	1.42, (1.36, 1.48)
H	40.21	34.04, (33.89, 34.19)
I	4.05	1.51, (1.38, 1.65)
J	13.66	14.02, (13.83, 14.21)
K	3.58	16.58, (16.38, 16.77)
L	3.85	5.36, (5.19, 5.52)
M	3.20	3.02, (2.93, 3.11)

Table 7.2: Capacity Values Used for Individual Adjustment Investigation.

Activity	Standard	Direction	1	2	3	4
D	(2, 2, 1, 2, 2)	Decrease	(1, 1, 1, 1, 1)	(2, 2, 1, 1, 1)	(2, 1, 1, 1, 1)	
G	(2, 3, 2, 3, 3)	Decrease	(2, 2, 2, 2, 2)	(3, 3, 2, 2, 2)	(3, 2, 2, 2, 2)	
I	(7, 7, 7, 7, 8)	Decrease	(7, 7, 7, 7, 7)	(7, 7, 7, 7, 6)	(7, 7, 7, 6, 6)	(6, 7, 7, 7, 7)
K	(2, 4, 4, 3, 2)	Increase	(3, 4, 4, 3, 3)	(4, 4, 4, 4, 3)	(4, 4, 4, 3, 3)	
C	(5, 7, 9, 16, 1)	Decrease	(5, 7, 9, 9, 1)	(5, 7, 9, 12, 1)	(5, 7, 9, 10, 1)	(5, 7, 9, 11, 1)
A	(6, 7, 10, 6, 6)	Increase	(7, 7, 10, 7, 7)	(7, 8, 10, 7, 7)	(7, 7, 10, 7, 6)	(7, 7, 10, 6, 6)

A final model was explored which incrementally applied the chosen capacity levels from Table 7.2 in the order D, G, I, K, C, A. At points where the capacity was decreased it is expected that the top level results will perform worse and vice versa. Table 7.3 and Table 7.4 show the top level and waiting time results respectively.

Table 7.3: Top Level Results for Individual Adjustment Investigation.

Name	Mean TiS	Median Tis	Target	Overall Period
Original	60	41	64.4	362
Standard	65.08, (63.34, 66.82)	56.72, (54.4, 59.04)	62, (53.12, (50.34, 55.9)	407.66, (407.46, 407.86)
D	66.52, (64.29, 68.75)	57.6, (54.3, 60.9)	62, (50.53, (46.55, 54.51)	407.89, (407.62, 408.15)
DG	66.32, (64.53, 68.11)	57.28, (54.85, 59.71)	62, (52.28, (49.93, 54.63)	407.9, (407.69, 408.11)
DGI	64.04, (61.91, 66.17)	53.68, (50.87, 56.49)	62, (54.59, (51.43, 57.76)	408.02, (407.77, 408.26)
DGIK	65.88, (63.79, 67.97)	57.2, (54.13, 60.27)	62, (51.05, (47.17, 54.92)	408.02, (407.8, 408.23)
DGIKC	69.6, (67.74, 71.46)	66.44, (63.72, 69.16)	62, (39.89, (36.0, 43.78)	428.88, (428.76, 429.01)
DGIKCA	69.56, (67.72, 71.4)	67.56, (65.17, 69.95)	62, (39.48, (35.9, 43.07)	428.62, (428.38, 428.86)

It is evident that the expected result (described above) did not come to fruition, due to the highlighted unintended changes (purple). For example, in the addition of I (model DGI) the target in Table 7.3 improved despite the decreasing of capacity, which appears to be the result of a decrease in the waiting time for activity A.

Table 7.4: Activity Waiting Time Results for Individual Adjustment Investigation.

Activity	D	DG	DGI
A	23.24, (21.2, 25.27)	23.11, (21.38, 24.84)	20.48, (18.61, 22.36)
B	9.76, (9.43, 10.09)	9.76, (9.55, 9.97)	9.48, (9.19, 9.76)
C	2.01, (1.93, 2.08)	2.03, (1.95, 2.11)	1.9, (1.84, 1.97)
D	14.75, (14.36, 15.13)	14.73, (14.4, 15.07)	14.65, (14.29, 15.01)
E	7.57, (7.45, 7.7)	7.64, (7.5, 7.78)	7.73, (7.55, 7.91)
F	16.78, (16.61, 16.95)	15.49, (15.3, 15.68)	15.36, (15.09, 15.62)
G	1.45, (1.38, 1.53)	4.75, (4.6, 4.89)	4.69, (4.52, 4.87)
H	33.86, (33.7, 34.01)	30.83, (30.66, 31.01)	30.84, (30.64, 31.04)
I	1.57, (1.39, 1.75)	1.61, (1.49, 1.72)	2.31, (2.02, 2.61)
J	13.97, (13.81, 14.13)	14.02, (13.83, 14.21)	14.2, (13.95, 14.45)
K	16.55, (16.32, 16.77)	16.52, (16.32, 16.71)	16.38, (16.17, 16.6)
L	5.34, (5.2, 5.49)	5.55, (5.39, 5.7)	5.41, (5.26, 5.56)
M	2.97, (2.86, 3.08)	2.98, (2.88, 3.07)	2.96, (2.85, 3.06)

Activity	DGIK	DGIKC	DGIKCA
A	23.04, (21.26, 24.82)	22.58, (20.91, 24.25)	17.1, (15.56, 18.64)
B	10.0, (9.66, 10.34)	9.69, (9.33, 10.04)	11.5, (11.16, 11.84)
C	1.71, (1.66, 1.76)	7.13, (6.92, 7.33)	8.22, (8.09, 8.36)
D	15.04, (14.7, 15.37)	13.9, (13.66, 14.14)	12.47, (12.18, 12.77)
E	10.28, (10.01, 10.54)	9.38, (9.16, 9.59)	9.14, (8.89, 9.39)
F	16.12, (15.85, 16.4)	16.03, (15.76, 16.3)	15.89, (15.68, 16.11)
G	7.56, (7.38, 7.74)	7.35, (7.21, 7.5)	7.1, (6.91, 7.29)
H	31.27, (31.11, 31.43)	30.28, (30.14, 30.43)	30.07, (29.94, 30.19)
I	2.67, (2.34, 3.0)	2.44, (2.14, 2.73)	3.17, (2.77, 3.57)
J	15.77, (15.56, 15.99)	16.07, (15.81, 16.32)	16.68, (16.44, 16.92)
K	2.58, (2.48, 2.69)	2.49, (2.39, 2.59)	3.01, (2.88, 3.15)
L	6.33, (6.12, 6.55)	6.25, (6.08, 6.43)	6.4, (6.17, 6.62)
M	3.67, (3.51, 3.83)	3.66, (3.55, 3.77)	5.72, (5.54, 5.9)

Exploring this further utilising the information from the transition probability matrix (Table 5.3) activity I goes to activity A 45% of the time. Therefore, decreasing the capacity at I would as a result lessen the pressure on activity A by moving the waiting time to activity I instead. Although this is then counter-acted in the next change (model DGIK).

Furthermore, the results take a considerable decline in model DGIKC, noting that achieving a more accurate waiting time for activity C worsens the general overall performance. Lastly with the final change (model DGIKCA) there are lots of unintended resulting changes here, some positive and some negative, with most notably the increase in waiting time for activity M. Again noting the transition probability matrix (Table 5.3) activity A goes to activity M 42% of the time. Conversely to previously, with the increase in capacity at activity A this would increase the pressure on activity M, causing the increased waiting time.

In conclusion, this investigation highlighted that concentrating the focus of improvement of a particular activity will result in unintended changes. This further highlights the complexities of working with a highly complex pathway system.

7.3 Basic Investigation

It is typical after obtaining an initial model to explore how to improve upon the key KPI's. Therefore, the purpose of this investigation is to improve the percentage of individuals seen within 62 days (Target) using an instinctive method. Table 7.5 displays the capacities that were changed throughout this investigation, Table 7.6 shows the top level results (specifically noting the target column), and Table 7.7 shows the mean waiting time for each activity.

A logical first step would be to keep the two capacity increases from the previous investigation (without the capacity decreases) to observe the effect. This can be seen in the first two models, namely K and KA. This does improve the target slightly, however, the waiting time for A is still higher than the original data.

Instinctively, the next step is to further increase the capacity of A, using the information from the previous investigation for model KAA. This change does not substantially impact the target despite the waiting time for activity A drastically decreasing. As seen previously, the waiting time for activity M has increased as a result of this change, and subsequently guides the next alteration. It is also noticeable that the waiting time for activity B is also increasing slowly compared to the previous models, but is coming in line with the original data.

Focussing on decreasing the waiting time for activity M in model KAAM. The results reveal that the waiting time for activity M does come nicely in line with the original data and the target has started to increase. There were three unintended changes here, namely waiting time for B continued to increase and activity E and J took a sizeable increase in waiting time. Although B does seem in control it is a very high frequency activity, with occurring 1865 times indicates that it was present in every pathway. Therefore, it would be important to keep a close eye on this activity and also gains here should improve the experience for every individual. Alternatively, E and J are lower frequency activities (occurring 473 and 537 times respectively).

Progressing with investigating decreasing the waiting time for activity B in model KAAMB. The target does not notably increase despite the large drop in the waiting time for activity B. The reason for this is speculated that this is due to activity B having a low exit rate and mid-range arrival rate, and as such is a key mid pathway activity that can lead to any other activity (Table 5.1).

Whilst keeping the capacity change for activity B, it is now appropriate to address the continuously increasing waiting time for activity J in model KAAMBJ. This does hugely decrease the waiting time and activity J along with improving the target. Finally, it is questionable whether the increase in capacity for activity B was necessary. Thus reverting the capacity for activity B to its original state whilst keeping the increase in capacity for activity J is seen in model KAAMJ. Notably the waiting time for activity B drastically increases greater than it was before its reduction state, and the target only slightly decreased, indicating that changes for B and J need to be considered together. Although, interestingly they do not have a strong relationship within the raw transition matrix (Table 5.1).

This investigation concludes that an instinctive approach proves difficult to obtain large improvement as the target only achieves a maximum of 63%, well below the desirable 95%. This investigation showed that it might not be beneficial to focus on improvements on mid pathway activities (such as activity B) as this appears to just move the waiting time to other points in the pathway. Therefore it could be suggested that finding gains in end pathway activities could have a larger impact on the target results. Furthermore, the increase of capacity for K and A will subsequently form the first step of improvements for subsequent models.

Table 7.5: Capacity Values for Basic Adjustment Investigation.

Activity	Standard	First			Second		
		C*	D	P	C	D	P
K	15, (2, 4, 4, 3, 2)	18, (4, 4, 4, 3, 3)	+3	+20%	38, (7, 7, 10, 7, 7)	+3	+9%
A	35, (6, 7, 10, 6, 6)	36, (7, 7, 10, 6, 6)	+1	+3%			
M	31, (5, 7, 9, 5, 5)	33, (7, 7, 9, 5, 5)	+2	+6%			
B	35, (6, 7, 8, 7, 7)	38, (8, 8, 8, 7, 7)	+3	+9%			
J	13, (2, 4, 2, 3, 2)	16, (4, 4, 3, 3, 2)	+3	+23%			

*C = total slots, (weekly pattern), D = comparative difference with the standard, and P = comparative adjusted percentage with the standard.

Table 7.6: Top Level Results for Basic Adjustment Investigation.

Name	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
Standard	65.08, (63.34, 66.82)	56.72, (54.4, 59.04)	53.12, (50.34, 55.9)	407.66, (407.46, 407.86)
K	63.44, (61.24, 65.64)	54.72, (51.58, 57.86)	55.09, (52.35, 57.84)	407.61, (407.4, 407.82)
KA	62.44, (60.12, 64.76)	52.08, (48.77, 55.39)	55.46, (52.47, 58.45)	405.31, (404.64, 405.99)
KAAM	62.48, (60.48, 64.48)	53.12, (50.21, 56.03)	55.18, (52.74, 57.62)	401.77, (401.34, 402.21)
KAAMB	59.28, (57.43, 61.13)	45.48, (43.22, 47.74)	58.29, (57.08, 59.5)	399.5, (399.3, 399.7)
KAAMBJ	57.08, (55.13, 59.03)	43.4, (41.44, 45.36)	59.81, (58.28, 61.34)	394.14, (393.96, 394.31)
KAAMJ	53.4, (51.24, 55.56)	38.04, (35.98, 40.1)	62.59, (61.76, 63.41)	386.65, (386.38, 386.93)
KAAMJ	59.4, (58.01, 60.79)	45.32, (43.4, 47.24)	61.28, (60.72, 61.84)	399.39, (399.19, 399.59)

Table 7.7: Activity Waiting Time Results for Basic Adjustment Investigation.

Activity	Original	Standard	K	KA	KAAM
A	12.52	23.25, (21.61, 24.9)	22.62, (20.64, 24.6)	17.18, (15.24, 19.12)	10.4, (9.31, 11.49)
B	11.86	9.64, (9.39, 9.88)	9.57, (9.27, 9.86)	10.87, (10.48, 11.25)	12.13, (11.45, 12.81)
C	9.40	1.97, (1.9, 2.05)	1.71, (1.64, 1.77)	2.01, (1.89, 2.12)	1.87, (1.77, 1.97)
D	21.91	0.73, (0.69, 0.77)	0.99, (0.92, 1.06)	1.01, (0.96, 1.05)	1.0, (0.93, 1.08)
E	11.19	7.53, (7.41, 7.64)	10.32, (10.14, 10.5)	10.07, (9.87, 10.27)	9.58, (9.39, 9.78)
F	20.20	18.29, (18.11, 18.47)	19.5, (19.28, 19.71)	19.23, (19.04, 19.41)	18.76, (18.57, 18.95)
G	6.41	1.42, (1.36, 1.48)	2.67, (2.57, 2.77)	2.64, (2.56, 2.72)	2.71, (2.63, 2.8)
H	40.21	34.04, (33.89, 34.19)	36.05, (35.87, 36.23)	35.81, (35.66, 35.97)	35.3, (35.09, 35.52)
I	4.05	1.51, (1.38, 1.65)	1.53, (1.39, 1.66)	1.79, (1.6, 1.98)	2.56, (2.12, 3.0)
J	13.66	14.02, (13.83, 14.21)	16.15, (15.92, 16.38)	16.89, (16.73, 17.05)	17.57, (17.37, 17.76)
K	3.58	16.58, (16.38, 16.77)	2.78, (2.66, 2.89)	3.21, (3.11, 3.3)	4.13, (4.01, 4.25)
L	3.85	5.36, (5.19, 5.52)	6.28, (6.08, 6.49)	6.56, (6.4, 6.72)	7.16, (6.99, 7.33)
M	3.20	3.02, (2.93, 3.11)	3.97, (3.88, 4.06)	6.64, (6.55, 6.73)	11.92, (11.62, 12.21)
Activity		KAAM	KAAMB	KAAMBJ	KAAMJ
A		10.13, (9.22, 11.03)	9.96, (8.95, 10.96)	10.22, (9.17, 11.27)	10.77, (10.03, 11.5)
B		12.41, (11.58, 13.23)	6.62, (5.95, 7.29)	5.88, (4.98, 6.78)	15.25, (14.55, 15.95)
C		2.19, (2.06, 2.32)	4.55, (4.42, 4.68)	3.28, (3.17, 3.4)	1.5, (1.47, 1.54)
D		1.03, (0.95, 1.11)	0.99, (0.93, 1.05)	2.87, (2.74, 2.99)	2.47, (2.36, 2.58)
E		12.5, (12.26, 12.74)	13.97, (13.79, 14.15)	18.64, (18.44, 18.83)	16.69, (16.48, 16.91)
F		19.03, (18.84, 19.21)	19.11, (18.92, 19.29)	18.54, (18.33, 18.75)	18.55, (18.35, 18.76)
G		3.22, (3.14, 3.31)	3.43, (3.28, 3.57)	4.89, (4.72, 5.06)	4.26, (4.14, 4.38)
H		36.87, (36.72, 37.02)	37.53, (37.36, 37.69)	40.75, (40.57, 40.93)	39.47, (39.32, 39.62)
I		2.76, (2.3, 3.22)	2.78, (2.43, 3.14)	2.61, (2.15, 3.08)	3.11, (2.74, 3.48)
J		21.43, (21.26, 21.6)	21.48, (21.23, 21.73)	2.52, (2.44, 2.59)	2.6, (2.5, 2.69)
K		5.52, (5.43, 5.61)	5.9, (5.72, 6.08)	6.18, (6.0, 6.36)	5.17, (5.08, 5.26)
L		8.4, (8.24, 8.57)	9.46, (9.17, 9.74)	12.47, (12.26, 12.67)	9.52, (9.31, 9.73)
M		3.79, (3.64, 3.93)	4.13, (3.91, 4.36)	4.72, (4.59, 4.85)	3.91, (3.82, 4.0)

7.4 End Activity Investigation

Investigating the potential improvements for considering the waiting times for end of pathway activities, it is first necessary to identify these activities. Noting any activity that exits the system more than 100 times using the raw transition matrix (Table 5.1) returns activities C, D, F, H and M (values 786, 146, 299, 443 and 125 respectively). Considering that activity C and D were used in the Individual Adjustment Investigation (Section 7.2) as their waiting times were considerably smaller in the standard model than the original data, it is not logical to consider these for improvements. Furthermore, the increase of capacity K and A is the standard first step (previously model KAA now denoted as KA henceforth).

Taking the approach of making improvements for the largest to smallest exit activity results in the order of H, F and M. Examining the improvements in turn:

- H: Due to its initial large waiting time, a large increase of capacity was considered (+8 a week). Although this is a relatively low frequency activity (537 occurrences) with its initial large waiting time and being an exit activity on 82% of its occurrences, and can be seen as a ‘quick win’ in terms of improvements. This is reflected in the target improving to 63% - the same value as found with changing three activities in the previous investigation.
- F: Increasing the capacity by 4 a week allows for there to be an equal number of slots each day. This has a visibly positive effect for the target, achieving the largest value so far at 71%. Furthermore, this had little effect on the other activities.
- M: As this activity was required to be considered in the previous investigation where it was noted that a small increase of just 2 slots a week made an improvement, that same capacity was applied. This further improves the target, achieving 75%, however does have the same unintended consequences of increasing the waiting times of activities B and J.

In conclusion, improving the end activities were easier to navigate as they had little effect on other activities. Activity M was the exception to this finding, although this is unsurprising when further examining the raw transition matrix (Table 5.1) and that activity M is a high frequency activity, occurring 1577 times.

Table 7.8: Capacity Values for End Activity Investigation.

Activity	Standard	First C*	D	P
K	15, (2, 4, 4, 3, 2)	18, (4, 4, 4, 3, 3)	+3	+20%
A	35, (6, 7, 10, 6, 6)	38, (7, 7, 10, 7, 7)	+3	+9%
H	13, (5, 2, 2, 2, 2)	21, (5, 4, 4, 4, 4)	+8	+62%
F	11, (2, 2, 2, 2, 3)	15, (3, 3, 3, 3, 3)	+4	+36%
M	31, (5, 7, 9, 5, 5)	33, (7, 7, 9, 5, 5)	+2	+6%

*C = total slots, (weekly pattern), D = comparative difference with the standard, and P = comparative adjusted percentage with the standard.

Table 7.9: Top Level Results for End Activity Investigation.

Name	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
Standard	65.08, (63.34, 66.82)	56.72, (54.4, 59.04)	53.12, (50.34, 55.9)	407.66, (407.46, 407.86)
KA	62.48, (60.48, 64.48)	53.12, (50.21, 56.03)	55.18, (52.74, 57.62)	401.77, (401.34, 402.21)
KA_H	55.76, (53.71, 57.81)	48.92, (46.18, 51.66)	62.86, (60.65, 65.06)	401.22, (401.04, 401.41)
KA_HF	51.56, (49.67, 53.45)	47.68, (45.26, 50.1)	70.86, (68.3, 73.42)	401.18, (400.97, 401.4)
KA_HFM	47.92, (46.46, 49.38)	42.24, (40.24, 44.24)	75.01, (74.05, 75.98)	399.45, (399.25, 399.66)

Table 7.10: Activity Waiting Time Results for End Activity Investigation.

Activity	Original	Standard	KA
A	12.52	23.25, (21.61, 24.9)	10.4, (9.31, 11.49)
B	11.86	9.64, (9.39, 9.88)	12.13, (11.45, 12.81)
C	9.40	1.97, (1.9, 2.05)	1.87, (1.77, 1.97)
D	21.91	0.73, (0.69, 0.77)	1.0, (0.93, 1.08)
E	11.19	7.53, (7.41, 7.64)	9.58, (9.39, 9.78)
F	20.20	18.29, (18.11, 18.47)	18.76, (18.57, 18.95)
G	6.41	1.42, (1.36, 1.48)	2.71, (2.63, 2.8)
H	40.21	34.04, (33.89, 34.19)	35.3, (35.09, 35.52)
I	4.05	1.51, (1.38, 1.65)	2.56, (2.12, 3.0)
J	13.66	14.02, (13.83, 14.21)	17.57, (17.37, 17.76)
K	3.58	16.58, (16.38, 16.77)	4.13, (4.01, 4.25)
L	3.85	5.36, (5.19, 5.52)	7.16, (6.99, 7.33)
M	3.20	3.02, (2.93, 3.11)	11.92, (11.62, 12.21)
Activity	KA_H	KA_HF	KA_HFM
A	10.55, (9.47, 11.63)	10.98, (9.91, 12.04)	10.0, (9.23, 10.77)
B	11.72, (11.0, 12.44)	11.88, (11.23, 12.53)	13.42, (12.79, 14.04)
C	1.82, (1.77, 1.87)	1.82, (1.77, 1.88)	1.98, (1.88, 2.08)
D	1.05, (0.97, 1.12)	1.05, (0.99, 1.11)	1.01, (0.95, 1.08)
E	9.51, (9.33, 9.69)	9.56, (9.37, 9.74)	12.29, (12.06, 12.52)
F	22.37, (22.16, 22.58)	1.23, (1.19, 1.26)	1.25, (1.21, 1.29)
G	2.63, (2.55, 2.7)	2.63, (2.55, 2.71)	3.21, (3.11, 3.32)
H	8.84, (8.79, 8.89)	10.69, (10.63, 10.76)	10.88, (10.79, 10.96)
I	2.42, (2.04, 2.81)	2.74, (2.36, 3.12)	2.71, (2.38, 3.03)
J	17.41, (17.21, 17.6)	17.44, (17.14, 17.75)	21.29, (21.11, 21.48)
K	4.2, (4.06, 4.34)	4.03, (3.9, 4.17)	5.44, (5.33, 5.56)
L	7.93, (7.68, 8.18)	7.8, (7.61, 7.99)	8.99, (8.78, 9.2)
M	12.15, (11.83, 12.46)	12.14, (11.88, 12.39)	3.75, (3.61, 3.89)

7.5 Target Investigation

So far there has been progress with improving upon the target results, now the purpose of this investigation is to achieve the 95% target. For note of the naming convention, an `_` indicates a progression that will remain for the subsequent models. Occasionally there is a reversion of a change to the previous state highlighted in `teal` which will be reflected by the removal of the letter in the previous `_` segment in the model name. The improvements were made in three main segments described in the following four paragraphs with the middle two paragraphs relating to the second segment. The model progressed into the next segment is highlighted in `green` in Table 7.12 and Table 7.13. The capacity changes are recorded in Table 7.11.

Following on from the findings of the previous investigations, and observing the results in Table 7.12, the first two models (KA and KA_HFM) implement the improvements from the previous investigations. The initial improvement for this investigation is applied to activity J, achieving a target of 82%. This negatively impacts the waiting time of activities E and L. Therefore the next set of models explores these activities, initially increasing capacity at activity E and L individually by 3 slots a week, followed by both improvements (models KA_HFM_J_E, KA_HFM_J_L and KA_HFM_J_EL respectively). Interestingly, individually E improves the target (86%) whereas L slightly decreases the target (81%), but the inclusion of both increases the target to the largest so far at 87%. It was considered that the improvement in activity L was slightly unnecessarily large (model KA_HFM_J_EL2), thus an increase of 2 instead of 3 was implemented, which further improved the target to a new maximum of 88%, possibly attributed to the unintended decreasing the waiting times at activity A and B.

Alternatively to adjusting activity L to observe an improvement in activity B, the alternative is to adjust activity B directly in model KA_HFM_JEL_B. Although this did not increase the target, the mean and median TiS did decrease. Model KA_HFM_JEL_BH considers the previous ‘quick win’ from increasing capacity at

activity H by increasing by an additional 1 slot a week, whilst persevering with the change to B. This only resulted in a small improvement however, reverting B to its previous capacity (retaining the change in H) results in a larger overall improvement of the target to a new maximum of 90%. Exploring reverting the capacity of activity L to its original levels (model KA_HFM_JE_H) decreases the target, then further increasing the capacity at H again (model KA_HFM_JE_H2) does increase the target but not past the previous maximum.

This fluctuation makes the decision of which model to progress quite difficult as no model is notably different. From the first segment (excluding J as it was present in all models discussed) there was a total increase of 6 slots a week. In the second segment, the total increase of slots a week across the models respectively were 9, 18, 15, 12, 13. Considering the two least increases leave model KA_HFM_JEL_B and KA_HFM_JE_H, activity H initially has the larger waiting time improvement, and thus this model was progressed to the next segment (despite not having the maximum target).

Progressing into the third segment sees many attempts to improve upon the results, with a variety of activity improvements, including smoothing the spread of the capacity through the week (model KA_HFM_JEH_CS). Again here there are some fluctuations but no vast improvements.

In conclusion, no model in this investigation could reach above 90%, with the final chosen model taking into account balancing the amount of capacity increase (KA_HFM_JEH) achieving 87%. With this still being below the desired 95% it raises the question if this target is actually achievable? This will be further explored in the next investigation.

Table 7.11: Capacity Values for Target Investigation.

Activity	Standard	First			Second		
		C*	D	P	C	D	P
K	15, (2, 4, 4, 3, 2)	18, (4, 4, 4, 3, 3)	+3	+20%			
A	35, (6, 7, 10, 6, 6)	38, (7, 7, 10, 7, 7)	+3	+9%			
H	13, (5, 2, 2, 2, 2)	21, (5, 4, 4, 4, 4)	+8	+62%			
F	11, (2, 2, 2, 2, 3)	15, (3, 3, 3, 3, 3)	+4	+36%			
M	31, (5, 7, 9, 5, 5)	33, (7, 7, 9, 5, 5)	+2	+6%			
J	13, (2, 4, 2, 3, 2)	16, (4, 4, 3, 3, 2)	+3	+23%			
E	12, (2, 3, 2, 3, 2)	15, (3, 3, 3, 3, 3)	+3	+25%			
L	10, (2, 2, 2, 2, 2)	13, (3, 3, 3, 2, 2)	+3	+30%	12, (3, 3, 2, 2, 2)	+2	+20%
B	35, (6, 7, 8, 7, 7)	38, (8, 8, 8, 7, 7)	+3	+9%			
H**	13, (5, 2, 2, 2, 2)	22, (5, 5, 4, 4, 4)	+9	+69%	23, (5, 5, 5, 4, 4)	+10	77%
A	35, (6, 7, 10, 6, 6)	40, (8, 8, 10, 7, 7)	+5	+14%			
C	38, (5, 7, 9, 16, 1)	40, (7, 7, 9, 16, 1)	+2	5%	38, (5, 7, 9, 9, 8)	+0	0%
F	11, (2, 2, 2, 2, 3)	18, (4, 4, 4, 3, 3)	+7	+63%			
I	36, (7, 7, 7, 7, 8)	38, (8, 8, 7, 7, 8)	+2	+6%			

*C = total slots, (weekly pattern), D = comparative difference with the standard, and P = comparative adjusted percentage with the standard.

**Note: H first and second improvements after the initial improvement shown above

Table 7.12: Top Level Results for Target Investigation.

Name	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
Standard	65.08, (63.34, 66.82)	56.72, (54.4, 59.04)	53.12, (50.34, 55.9)	407.66, (407.46, 407.86)
KA	62.48, (60.48, 64.48)	53.12, (50.21, 56.03)	55.18, (52.74, 57.62)	401.77, (401.34, 402.21)
KA_HFM	47.92, (46.46, 49.38)	42.24, (40.24, 44.24)	75.01, (74.05, 75.98)	399.45, (399.25, 399.66)
KA_HFM_J	45.48, (43.63, 47.33)	41.24, (38.66, 43.82)	81.55, (79.68, 83.43)	399.46, (399.25, 399.66)
KA_HFM_J_E	43.4, (41.33, 45.47)	40.36, (37.82, 42.9)	85.84, (83.92, 87.75)	399.43, (399.23, 399.63)
KA_HFM_J_L	45.2, (43.47, 46.93)	41.28, (38.93, 43.63)	81.43, (78.83, 84.04)	399.53, (399.34, 399.72)
KA_HFM_J_EL	43.08, (40.97, 45.19)	40.36, (37.72, 43.0)	87.36, (85.31, 89.42)	399.54, (399.3, 399.77)
KA_HFM_J_EL2	42.32, (40.65, 43.99)	39.6, (37.35, 41.85)	88.26, (86.82, 89.71)	399.52, (399.31, 399.74)
KA_HFM_JEL_B	37.64, (35.81, 39.47)	33.44, (31.44, 35.44)	87.44, (86.35, 88.52)	381.82, (381.2, 382.45)
KA_HFM_JEL_BH	38.12, (35.98, 40.26)	34.4, (32.1, 36.7)	88.68, (87.48, 89.88)	381.7, (381.14, 382.27)
KA_HFM_JEL_H	41.56, (39.81, 43.31)	38.72, (36.57, 40.87)	89.61, (88.3, 90.92)	399.6, (399.34, 399.85)
KA_HFM_JE_H	42.84, (40.69, 44.99)	39.64, (36.86, 42.42)	87.27, (85.46, 89.07)	399.42, (399.22, 399.61)
KA_HFM_JE_H2	42.24, (40.31, 44.17)	39.08, (36.61, 41.55)	88.16, (86.21, 90.11)	399.55, (399.29, 399.81)
KA_HFM_JEH_B	39.36, (36.78, 41.94)	34.4, (31.34, 37.46)	85.91, (83.36, 88.45)	385.93, (385.45, 386.41)
KA_HFM_JEH_BL	38.12, (35.98, 40.26)	34.4, (32.1, 36.7)	88.68, (87.48, 89.88)	381.7, (381.14, 382.27)
KA_HFM_JEH_A	42.68, (41.14, 44.22)	39.12, (37.22, 41.02)	86.61, (84.79, 88.42)	399.22, (399.03, 399.4)
KA_HFM_JEH_C	43.32, (40.86, 45.78)	40.6, (37.5, 43.7)	85.89, (82.75, 89.03)	399.3, (399.11, 399.49)
KA_HFM_JEH_F	44.08, (42.16, 46.0)	41.28, (38.97, 43.59)	85.87, (84.11, 87.64)	399.6, (399.37, 399.83)
KA_HFM_JEH_I	41.08, (39.07, 43.09)	37.08, (34.63, 39.53)	87.93, (85.96, 89.9)	399.17, (399.03, 399.32)
KA_HFM_JEH_CS	42.4, (40.11, 44.69)	39.16, (36.21, 42.11)	86.9, (84.51, 89.3)	397.24, (396.72, 397.76)

Table 7.13: Activity Waiting Time Results for Target Investigation.

	Original	Standard	KA	KA_HFM	
A	12.518	23.25, (21.61, 24.9)	10.4, (9.31, 11.49)	10.0, (9.23, 10.77)	
B	11.8618	9.64, (9.39, 9.88)	12.13, (11.45, 12.81)	13.42, (12.79, 14.04)	
C	9.40291	1.97, (1.9, 2.05)	1.87, (1.77, 1.97)	1.98, (1.88, 2.08)	
D	21.9052	0.73, (0.69, 0.77)	1.0, (0.93, 1.08)	1.01, (0.95, 1.08)	
E	11.1882	7.53, (7.41, 7.64)	9.58, (9.39, 9.78)	12.29, (12.06, 12.52)	
F	20.2021	18.29, (18.11, 18.47)	18.76, (18.57, 18.95)	1.25, (1.21, 1.29)	
G	6.41199	1.42, (1.36, 1.48)	2.71, (2.63, 2.8)	3.21, (3.11, 3.32)	
H	40.2146	34.04, (33.89, 34.19)	35.3, (35.09, 35.52)	10.88, (10.79, 10.96)	
I	4.05119	1.51, (1.38, 1.65)	2.56, (2.12, 3.0)	2.71, (2.38, 3.03)	
J	13.6567	14.02, (13.83, 14.21)	17.57, (17.37, 17.76)	21.29, (21.11, 21.48)	
K	3.58163	16.58, (16.38, 16.77)	4.13, (4.01, 4.25)	5.44, (5.33, 5.56)	
L	3.8544	5.36, (5.19, 5.52)	7.16, (6.99, 7.33)	8.99, (8.78, 9.2)	
M	3.20064	3.02, (2.93, 3.11)	11.92, (11.62, 12.21)	3.75, (3.61, 3.89)	
	KA_HFM_J	KA_HFM_J_E	KA_HFM_J_L	KA_HFM_J_EL	KA_HFM_J_EL2
A	9.99, (9.15, 10.83)	10.05, (8.98, 11.13)	10.41, (9.45, 11.38)	10.15, (9.02, 11.28)	9.59, (8.64, 10.54)
B	14.97, (14.01, 15.94)	14.91, (13.96, 15.86)	15.02, (14.25, 15.79)	15.02, (14.13, 15.9)	14.98, (14.19, 15.77)
C	1.38, (1.34, 1.41)	1.36, (1.33, 1.38)	1.4, (1.34, 1.45)	1.38, (1.35, 1.42)	1.44, (1.41, 1.48)
D	2.39, (2.26, 2.52)	2.03, (1.93, 2.13)	2.25, (2.15, 2.36)	1.93, (1.82, 2.04)	2.0, (1.89, 2.1)
E	16.54, (16.32, 16.76)	2.75, (2.67, 2.82)	17.82, (17.58, 18.06)	2.86, (2.79, 2.94)	2.84, (2.76, 2.92)
F	1.33, (1.3, 1.36)	3.72, (3.61, 3.84)	1.35, (1.31, 1.39)	3.88, (3.76, 4.0)	3.93, (3.82, 4.04)
G	4.27, (4.12, 4.43)	6.11, (5.96, 6.25)	4.97, (4.82, 5.12)	7.27, (7.12, 7.41)	7.17, (7.03, 7.32)
H	10.94, (10.84, 11.03)	11.78, (11.69, 11.86)	10.95, (10.87, 11.04)	11.82, (11.76, 11.89)	11.81, (11.74, 11.88)
I	2.79, (2.42, 3.16)	2.97, (2.6, 3.33)	2.96, (2.56, 3.36)	3.18, (2.76, 3.61)	2.68, (2.36, 3.0)
J	2.53, (2.47, 2.6)	2.55, (2.46, 2.64)	3.33, (3.24, 3.42)	3.41, (3.29, 3.54)	3.44, (3.33, 3.55)
K	5.15, (5.05, 5.25)	5.08, (4.96, 5.2)	5.15, (5.04, 5.26)	5.04, (4.89, 5.18)	5.12, (4.99, 5.24)
L	10.11, (9.88, 10.34)	9.95, (9.7, 10.19)	0.71, (0.66, 0.77)	0.79, (0.73, 0.84)	1.46, (1.38, 1.54)
M	3.96, (3.83, 4.1)	4.01, (3.85, 4.16)	4.23, (4.08, 4.37)	4.35, (4.24, 4.46)	4.49, (4.38, 4.59)
	KA_HFM_JEL_B	KA_HFM_JEL_BH	KA_HFM_JEL_H	KA_HFM_JE_H	KA_HFM_JE_H2
A	10.38, (9.35, 11.41)	10.72, (9.51, 11.93)	9.57, (8.65, 10.49)	9.78, (8.79, 10.76)	9.77, (8.77, 10.77)
B	5.65, (4.96, 6.33)	6.18, (5.38, 6.98)	14.59, (13.82, 15.36)	14.82, (13.79, 15.86)	14.54, (13.68, 15.41)
C	3.36, (3.23, 3.49)	3.09, (2.95, 3.24)	1.4, (1.37, 1.43)	1.37, (1.34, 1.4)	1.39, (1.36, 1.42)
D	2.37, (2.27, 2.47)	2.36, (2.26, 2.47)	1.99, (1.9, 2.09)	2.09, (1.98, 2.19)	2.16, (2.06, 2.27)
E	3.23, (3.16, 3.3)	3.26, (3.18, 3.33)	2.88, (2.8, 2.96)	2.73, (2.66, 2.81)	2.73, (2.65, 2.81)
F	4.14, (4.01, 4.27)	4.12, (3.98, 4.25)	3.97, (3.86, 4.09)	3.78, (3.66, 3.91)	3.78, (3.68, 3.88)
G	8.61, (8.46, 8.76)	8.65, (8.51, 8.79)	7.29, (7.14, 7.43)	6.09, (5.93, 6.25)	6.1, (5.95, 6.25)
H	11.94, (11.84, 12.03)	11.0, (10.92, 11.09)	10.89, (10.83, 10.96)	10.86, (10.78, 10.93)	10.14, (10.07, 10.2)
I	3.15, (2.63, 3.67)	3.41, (2.91, 3.9)	2.75, (2.45, 3.06)	2.78, (2.41, 3.14)	2.73, (2.41, 3.04)
J	3.71, (3.59, 3.84)	3.64, (3.53, 3.75)	3.43, (3.3, 3.56)	2.56, (2.48, 2.65)	2.58, (2.5, 2.67)
K	6.06, (5.91, 6.2)	6.03, (5.89, 6.17)	5.09, (4.96, 5.21)	5.08, (4.95, 5.21)	5.07, (4.94, 5.21)
L	1.17, (1.12, 1.22)	1.11, (1.07, 1.16)	0.8, (0.75, 0.86)	10.13, (9.92, 10.33)	10.06, (9.79, 10.34)
M	5.25, (5.07, 5.43)	5.14, (4.96, 5.33)	4.45, (4.31, 4.59)	4.03, (3.87, 4.19)	4.12, (3.97, 4.27)
	KA_HFM_JEH_B	KA_HFM_JEH_BL	KA_HFM_JEH_A	KA_HFM_JEH_C	KA_HFM_JEH_F
A	10.74, (9.19, 12.29)	10.72, (9.51, 11.93)	6.23, (5.7, 6.75)	10.25, (8.93, 11.57)	10.49, (9.52, 11.47)
B	14.59, (13.82, 15.36)	6.18, (5.38, 6.98)	15.44, (14.65, 16.24)	15.1, (14.07, 16.13)	15.62, (14.74, 16.51)
C	3.45, (3.33, 3.58)	3.09, (2.95, 3.24)	1.36, (1.34, 1.39)	1.02, (1.01, 1.04)	1.41, (1.37, 1.44)
D	2.4, (2.21, 2.59)	2.36, (2.26, 2.47)	2.2, (2.11, 2.3)	1.96, (1.84, 2.08)	2.05, (1.95, 2.15)
E	3.07, (2.99, 3.15)	3.26, (3.18, 3.33)	2.57, (2.51, 2.63)	2.71, (2.65, 2.78)	2.74, (2.65, 2.82)
F	3.95, (3.81, 4.08)	4.12, (3.98, 4.25)	3.74, (3.62, 3.86)	3.85, (3.75, 3.96)	1.25, (1.21, 1.3)
G	7.27, (7.14, 7.41)	8.65, (8.51, 8.79)	5.42, (5.26, 5.58)	6.05, (5.92, 6.19)	6.08, (5.95, 6.2)
H	10.93, (10.86, 11.01)	11.0, (10.92, 11.09)	10.88, (10.79, 10.97)	10.92, (10.84, 10.99)	11.58, (11.5, 11.67)
I	3.32, (2.64, 4.0)	3.41, (2.91, 3.9)	3.73, (3.24, 4.22)	3.13, (2.6, 3.67)	3.05, (2.68, 3.43)
J	2.73, (2.58, 2.88)	3.64, (3.53, 3.75)	2.96, (2.86, 3.05)	2.57, (2.48, 2.67)	2.5, (2.42, 2.58)
K	6.04, (5.83, 6.26)	6.03, (5.89, 6.17)	5.63, (5.47, 5.78)	4.97, (4.82, 5.13)	4.97, (4.83, 5.11)
L	13.23, (12.92, 13.54)	1.11, (1.07, 1.16)	10.33, (10.09, 10.58)	10.23, (9.88, 10.58)	9.95, (9.7, 10.2)
M	4.79, (4.56, 5.01)	5.14, (4.96, 5.33)	6.12, (5.92, 6.32)	3.95, (3.78, 4.13)	4.06, (3.95, 4.17)
	KA_HFM_JEH_I	KA_HFM_JEH_CS			
A	9.97, (8.79, 11.15)	10.15, (8.9, 11.4)			
B	14.27, (13.39, 15.15)	14.8, (13.84, 15.75)			
C	1.33, (1.3, 1.35)	0.56, (0.53, 0.6)			
D	2.12, (2.01, 2.22)	2.0, (1.87, 2.13)			
E	2.79, (2.7, 2.89)	2.69, (2.62, 2.76)			
F	3.86, (3.75, 3.97)	4.07, (3.97, 4.18)			
G	6.08, (5.95, 6.21)	6.15, (6.02, 6.29)			
H	10.95, (10.88, 11.02)	11.02, (10.93, 11.11)			
I	1.07, (0.99, 1.15)	2.81, (2.45, 3.17)			
J	2.53, (2.45, 2.61)	2.46, (2.37, 2.54)			
K	5.29, (5.2, 5.38)	5.13, (5.02, 5.23)			
L	10.33, (10.16, 10.51)	10.15, (9.91, 10.39)			
M	4.31, (4.14, 4.48)	4.03, (3.9, 4.17)			

7.6 Excessive Top Down Investigation

The previous investigation attempted to find the capacity levels required to achieve the 95% through a 'bottom up' approach. This came up fruitless and proved to be a tedious process. This investigation still aims to achieve the 95% by applying an alternative 'top down' approach. Here the initial capacity levels are set excessively high to reach 100% and then systematically the capacity is lowered to find a balance between capacity levels and achieving the 95% target. Again this approach is discussed in three segments over the following four paragraphs, with the third segment split over the last two paragraphs. For reference all mentions of target relate to Table 7.14 and waiting times relate to Table 7.16, with capacity values noted in Table 7.15.

In the first segment, the capacity for each activity was set to 15 slots a day. Surprisingly the target did not achieve 100%. Observing the waiting times shows that the waiting time for activity H was still very large (14.23 days on average). The next model increased the capacity for just activity H to 30 slots a day, where the waiting time did decrease slightly but not enough for the target to reach 100%. Investigating the Warm Start applied to activity H, which represents the idea of a backlog/waiting list, it is set to 82 days. Decreasing the Warm Start (and reverting the capacity for H to 15) did not reach the 100% target until the Warm Start was set to 40 days. This is similar to the average waiting time for activity H (originally 41 days, taking the ceiling of the mean in Table 4.4). This concludes that if the backlog/waiting list is longer than 40 days, then even extreme values of capacity will not be sufficient. Thus the remaining models in the investigation will set the Warm Start for activity H to 40 days.

In this segment, after reducing the capacity for all activities down to 10 slots a day (model All_10) where the target remains at 100%, applying a methodical approach to reducing the capacity. The activities can be split into two groups: High demand - activities A, B, C, I and M, with the remaining activities as lower demand - D, E, F,

G, H, J, K and L (Frequency in Table 4.4). Model 105 and 85 implements a decrease in daily slots from 10 to 5 for the low demand activities, followed by decreasing the high demand activities from 10 to 8 in the respective models. Both of these see the target remain at 100%, although there is an increase in the mean and median TiS. It takes dropping the capacity for the lower demand activities from 5 to 3 to see the target drop from the 100% position, achieving a 89% result. All the lower demand activities see a noticeable increase in waiting times as the number of slots a week (15) is now approaching the original values for these activities, particularly noting the large waiting times for activity J and K. This indicates that the next segment needs to consider activities on a more specific basis.

This segment alternates between grouped (high and lower demand) capacity decreases and specific activity adjustments. Thus firstly, adjusting activity J and K to 4 slots a day takes the target back up to 98%. Following the alternation, the high demand activities are dropped to 7 daily slots each in model 734, where the target reaches 96%, noting that the mean and median TiS have increased considerably. This appears to be due to the large waiting time for activity A where interestingly the capacity is now matching the original weekly amount. Alternating back to specific activity adjustments, model 8734 reverts activity A to 8 capacity a day, seeing an improvement in the target (99%) but results in a large increase in the waiting time for activity B. Counteracting this by reverting the capacity for activity B to 8 slots a day in model 88734 where the target remains around 99%. Although the previous couple of adjustments haven't had a large effect on the top level results, one should consciously consider the spread of the waiting time across activities.

In the last segment, model 887346 briefly considers dropping the capacity for activity M to 6 slots a day (as M is lowest demand of the high demand activities). Although this does counteract some of the negative unintended results of the previous model (increased waiting time at C and I) the target does drop below the desired 95%, thus this change was reverted for the subsequent models. Still following the alternating

method, model 88724 decreases the lower demand activities that were at 3 slots a day to 2 slots a day, which takes the target too low at 65%, with the waiting times for E, G, H and L taking extreme increases. Thus, reverting the capacity for these activities to 3 slots a day in model 887234 does make an improvement on the target (78%) however this is still below the 95%, as the waiting time for activity F is extremely large. Reverting the capacity for activity F brings the investigation to a close as the target achieves 99% with a relatively good spread of waiting time across the activities.

In conclusion, it is possible to achieve the 95% target with the capacity as shown in the final column of Table 7.15. Although this appears to be a good solution, there may be a more optimal solution through varying the amount of capacity on each day for the activities and bringing this inline with the pattern seen in the original weekly pattern to reflect the real life situation.

Table 7.14: Top Level Results for Excessive Top Down Investigation.

Name	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
Standard	65.08, (63.34, 66.82)	56.72, (54.4, 59.04)	53.12, (50.34, 55.9)	407.66, (407.46, 407.86)
All.15	7.08, (6.81, 7.35)	1.0, (1.0, 1.0)	96.42, (96.23, 96.62)	360.59, (357.73, 363.45)
15_H30	6.28, (6.07, 6.49)	1.0, (1.0, 1.0)	96.88, (96.71, 97.05)	363.88, (360.04, 367.72)
All.15.WarmH60	4.4, (4.18, 4.62)	1.0, (1.0, 1.0)	99.88, (99.88, 99.89)	360.1, (355.52, 364.67)
All.15.WarmH40	3.08, (2.93, 3.23)	1.0, (1.0, 1.0)	100.0, (100.0, 100.0)	361.15, (357.8, 364.5)
All.10	4.56, (4.29, 4.83)	1.48, (1.28, 1.68)	100.0, (100.0, 100.0)	361.97, (359.54, 364.41)
105	6.24, (5.9, 6.58)	1.96, (1.88, 2.04)	100.0, (100.0, 100.0)	362.4, (358.88, 365.93)
85	12.04, (10.76, 13.32)	9.68, (7.63, 11.73)	100.0, (100.0, 100.0)	363.67, (360.84, 366.5)
83	30.56, (28.61, 32.51)	22.76, (20.42, 25.1)	88.85, (86.97, 90.74)	374.18, (373.36, 374.99)
834	21.04, (19.48, 22.6)	16.8, (14.68, 18.92)	97.9, (97.3, 98.51)	364.8, (362.08, 367.53)
734	39.88, (37.43, 42.33)	40.0, (37.41, 42.59)	96.29, (92.62, 99.95)	406.7, (405.74, 407.67)
8734	38.56, (36.67, 40.45)	38.4, (36.17, 40.63)	98.66, (97.6, 99.72)	401.32, (401.04, 401.6)
88734	34.64, (32.7, 36.58)	33.68, (31.45, 35.91)	99.09, (98.76, 99.41)	397.53, (396.96, 398.1)
887346	42.0, (40.05, 43.95)	44.92, (42.83, 47.01)	93.6, (90.55, 96.66)	403.9, (403.1, 404.7)
88724	55.44, (53.06, 57.82)	36.28, (33.3, 39.26)	64.98, (64.11, 65.85)	415.89, (415.81, 415.97)
887234	46.56, (44.6, 48.52)	33.92, (31.22, 36.62)	77.71, (77.57, 77.85)	397.54, (396.96, 398.11)
8872334	35.32, (33.26, 37.38)	34.36, (31.8, 36.92)	98.99, (98.6, 99.37)	397.54, (396.97, 398.12)

Table 7.15: Capacity Slots Per Day for Excessive Top Down Investigation.

Activity (Original Weekly)	15	H30	15_WarmH60	15_WarmH40	10	105	85	83	
A (35)	15	15	15	15	10	10	8	8	
B (35)	15	15	15	15	10	10	8	8	
C (38)	15	15	15	15	10	10	8	8	
D (9)	15	15	15	15	10	5	5	3	
E (12)	15	15	15	15	10	5	5	3	
F (11)	15	15	15	15	10	5	5	3	
G (13)	15	15	15	15	10	5	5	3	
H (13)	15	30	15	15	10	5	5	3	
I (36)	15	15	15	15	10	10	8	8	
J (13)	15	15	15	15	10	5	5	3	
K (15)	15	15	15	15	10	5	5	3	
L (10)	15	15	15	15	10	5	5	3	
M (31)	15	15	15	15	10	10	8	8	

Activity (Original Weekly)	834	734	8734	88734	887346	88724	887234	8872334	C	D	P
									Final	(8872334)	
A (35)	8	7	8	8	8	8	8	8	40	+5	+14%
B (35)	8	7	7	8	8	8	8	8	40	+5	+14%
C (38)	8	7	7	7	7	7	7	7	35	-3	-8%
D (9)	3	3	3	3	3	2	2	2	12	+3	+33%
E (12)	3	3	3	3	3	2	3	3	15	+3	+25%
F (11)	3	3	3	3	3	2	2	3	15	+4	+36%
G (13)	3	3	3	3	3	2	3	3	15	+2	+15%
H (13)	3	3	3	3	3	2	3	3	15	+2	+15%
I (36)	8	7	7	7	7	7	7	7	35	-1	-3%
J (13)	4	4	4	4	4	4	4	4	20	+7	+54%
K (15)	4	4	4	4	4	4	4	4	20	+5	+33%
L (10)	3	3	3	3	3	2	3	3	15	+5	+50%
M (31)	8	7	7	7	6	7	7	7	35	+4	+13%

*C = total slots, (weekly pattern), D = comparative difference with the standard, and P = comparative adjusted percentage with the standard.

Table 7.16: Activity Waiting Time Results for Excessive Top Down Investigation.

Activity	Original	Standard		
A	12.52	23.25, (21.61, 24.9)		
B	11.86	9.64, (9.39, 9.88)		
C	9.40	1.97, (1.9, 2.05)		
D	21.91	0.73, (0.69, 0.77)		
E	11.19	7.53, (7.41, 7.64)		
F	20.20	18.29, (18.11, 18.47)		
G	6.41	1.42, (1.36, 1.48)		
H	40.21	34.04, (33.89, 34.19)		
I	4.051	1.51, (1.38, 1.65)		
J	13.66	14.02, (13.83, 14.21)		
K	3.58	16.58, (16.38, 16.77)		
L	3.85	5.36, (5.19, 5.52)		
M	3.20	3.02, (2.93, 3.11)		
Activity	All_15	15_H30	All_15_WarmH60	All_15_WarmH40
A	1.68, (1.57, 1.78)	1.67, (1.57, 1.78)	1.68, (1.57, 1.78)	1.68, (1.58, 1.79)
B	0.17, (0.16, 0.18)	0.17, (0.17, 0.18)	0.17, (0.17, 0.18)	0.19, (0.18, 0.19)
C	0.07, (0.07, 0.08)	0.07, (0.06, 0.08)	0.07, (0.06, 0.08)	0.09, (0.08, 0.1)
D	0.17, (0.16, 0.18)	0.17, (0.16, 0.17)	0.17, (0.17, 0.18)	0.19, (0.18, 0.19)
E	0.02, (0.02, 0.03)	0.02, (0.02, 0.02)	0.02, (0.02, 0.03)	0.02, (0.02, 0.02)
F	2.31, (2.29, 2.34)	2.31, (2.28, 2.34)	2.31, (2.29, 2.34)	2.45, (2.43, 2.48)
G	0.02, (0.02, 0.03)	0.02, (0.02, 0.03)	0.02, (0.02, 0.03)	0.02, (0.02, 0.02)
H	14.23, (13.8, 14.66)	12.1, (11.79, 12.41)	5.63, (5.45, 5.8)	0.54, (0.53, 0.56)
I	0.09, (0.09, 0.1)	0.09, (0.09, 0.1)	0.09, (0.09, 0.1)	0.09, (0.09, 0.1)
J	0.06, (0.06, 0.06)	0.06, (0.06, 0.06)	0.06, (0.06, 0.06)	0.06, (0.06, 0.06)
K	0.02, (0.02, 0.03)	0.02, (0.02, 0.02)	0.02, (0.02, 0.03)	0.02, (0.02, 0.02)
L	0.03, (0.02, 0.03)	0.03, (0.02, 0.03)	0.03, (0.02, 0.03)	0.03, (0.02, 0.03)
M	0.03, (0.03, 0.04)	0.03, (0.03, 0.04)	0.04, (0.03, 0.04)	0.04, (0.03, 0.04)
Activity	All_10	105	85	83
A	2.94, (2.73, 3.15)	2.96, (2.74, 3.19)	7.45, (6.59, 8.32)	7.64, (6.62, 8.67)
B	0.36, (0.34, 0.39)	0.32, (0.31, 0.34)	1.76, (1.39, 2.14)	2.21, (1.66, 2.77)
C	0.4, (0.37, 0.43)	0.15, (0.14, 0.16)	1.36, (1.23, 1.49)	4.82, (4.41, 5.23)
D	0.15, (0.14, 0.16)	0.15, (0.14, 0.17)	0.13, (0.13, 0.14)	0.18, (0.17, 0.19)
E	0.01, (0.01, 0.02)	0.62, (0.59, 0.66)	0.15, (0.13, 0.17)	2.56, (2.49, 2.64)
F	1.67, (1.64, 1.7)	4.72, (4.66, 4.79)	2.89, (2.82, 2.95)	5.54, (5.46, 5.63)
G	0.02, (0.01, 0.02)	0.1, (0.09, 0.11)	0.06, (0.05, 0.06)	0.69, (0.65, 0.72)
H	0.38, (0.37, 0.39)	0.98, (0.95, 1.02)	0.48, (0.47, 0.5)	2.13, (2.07, 2.19)
I	0.21, (0.2, 0.22)	0.21, (0.2, 0.22)	0.63, (0.55, 0.7)	0.65, (0.56, 0.74)
J	0.05, (0.05, 0.06)	0.29, (0.26, 0.31)	0.14, (0.13, 0.15)	9.89, (9.5, 10.28)
K	0.02, (0.02, 0.02)	2.43, (2.2, 2.67)	0.29, (0.27, 0.31)	30.29, (29.88, 30.7)
L	0.02, (0.02, 0.02)	0.06, (0.05, 0.07)	0.03, (0.02, 0.03)	0.4, (0.38, 0.43)
M	0.08, (0.08, 0.09)	0.08, (0.08, 0.09)	0.23, (0.21, 0.24)	0.28, (0.25, 0.31)
Activity	834	734	8734	88734
A	7.66, (6.76, 8.57)	22.39, (20.63, 24.16)	5.71, (5.21, 6.22)	5.67, (5.14, 6.19)
B	1.67, (1.3, 2.05)	9.27, (8.78, 9.77)	18.19, (17.25, 19.14)	2.39, (1.89, 2.9)
C	2.28, (1.93, 2.64)	2.24, (2.17, 2.31)	3.39, (3.3, 3.48)	12.46, (12.18, 12.74)
D	0.28, (0.26, 0.3)	0.21, (0.19, 0.22)	0.22, (0.2, 0.23)	0.22, (0.21, 0.24)
E	5.94, (5.79, 6.09)	2.91, (2.82, 3.0)	2.92, (2.85, 2.99)	3.4, (3.32, 3.48)
F	9.53, (9.33, 9.74)	6.27, (6.19, 6.36)	5.97, (5.88, 6.07)	6.58, (6.47, 6.69)
G	5.35, (5.24, 5.46)	1.57, (1.47, 1.66)	1.68, (1.62, 1.74)	2.29, (2.24, 2.34)
H	5.59, (5.46, 5.72)	3.15, (3.03, 3.28)	2.92, (2.83, 3.01)	3.33, (3.24, 3.41)
I	0.62, (0.55, 0.7)	2.54, (2.16, 2.91)	6.43, (5.7, 7.17)	8.09, (7.02, 9.17)
J	0.78, (0.73, 0.83)	0.26, (0.25, 0.28)	0.38, (0.36, 0.41)	0.44, (0.41, 0.47)
K	5.02, (4.84, 5.2)	0.88, (0.82, 0.93)	1.32, (1.27, 1.36)	2.56, (2.43, 2.7)
L	0.6, (0.55, 0.65)	0.26, (0.25, 0.28)	0.25, (0.23, 0.27)	0.55, (0.5, 0.6)
M	0.27, (0.25, 0.29)	0.37, (0.36, 0.39)	1.27, (1.18, 1.37)	1.63, (1.54, 1.72)
Activity	887346	88724	887234	8872334
A	5.8, (5.33, 6.27)	5.73, (5.09, 6.36)	5.54, (5.06, 6.01)	5.95, (5.41, 6.5)
B	4.58, (4.01, 5.14)	2.05, (1.59, 2.51)	2.34, (1.85, 2.83)	2.38, (1.91, 2.85)
C	3.56, (3.44, 3.68)	8.4, (8.09, 8.71)	12.38, (11.99, 12.76)	12.33, (12.08, 12.59)
D	0.21, (0.19, 0.22)	1.2, (1.13, 1.27)	1.16, (1.09, 1.24)	1.15, (1.08, 1.22)
E	2.61, (2.56, 2.66)	49.82, (49.61, 50.03)	3.46, (3.38, 3.54)	3.42, (3.33, 3.5)
F	5.83, (5.75, 5.91)	6.46, (6.33, 6.59)	58.08, (57.92, 58.25)	6.64, (6.5, 6.77)
G	0.62, (0.59, 0.65)	31.8, (31.57, 32.04)	2.28, (2.21, 2.34)	2.29, (2.22, 2.36)
H	2.29, (2.18, 2.4)	11.48, (11.26, 11.71)	0.94, (0.91, 0.98)	3.32, (3.21, 3.42)
I	6.19, (5.29, 7.08)	8.14, (6.72, 9.57)	7.76, (6.55, 8.97)	8.67, (7.41, 9.94)
J	0.29, (0.27, 0.32)	0.31, (0.29, 0.33)	0.42, (0.39, 0.46)	0.42, (0.39, 0.44)
K	1.01, (0.98, 1.05)	2.48, (2.32, 2.63)	2.63, (2.5, 2.77)	2.48, (2.36, 2.6)
L	0.2, (0.18, 0.22)	15.93, (15.73, 16.13)	0.54, (0.5, 0.58)	0.51, (0.48, 0.54)
M	22.36, (21.88, 22.85)	1.09, (1.02, 1.16)	1.6, (1.5, 1.7)	1.57, (1.49, 1.65)

7.7 Demand Investigation

It is common to investigate the sustainability of a chosen model through considering the impact of an increase in demand. Therefore this investigation takes the model KA_HFM_JEH from the Target Investigation and discusses the impact on the top level and activity waiting time results arising from a 5% and 10% increase in demand. This investigation was only performed in a Jupyter notebook as, at the time of writing this, increasing the demand for the *Raw Pathways* model is not supported in the Sim.Pro.Flow v2.1 GUI.

To accommodate this experiment the arrival rate for the dummy node was recalculated by inflating the original number of individuals (1865) by the desired percentage as shown in Equation 7.1 (rounded to 2.d.p for reporting).

$$\lambda_{5\%} = \frac{\text{ceil}(1.05(A_n))}{P} = \frac{\text{ceil}(1.05(1865))}{362} = \frac{1959}{362} = 5.41$$

$$\lambda_{10\%} = \frac{\text{ceil}(1.10(A_n))}{P} = \frac{\text{ceil}(1.10(1865))}{362} = \frac{2052}{362} = 5.67$$
(7.1)

This was still used with the *LimitedExponential* distribution (Section 4.3.4) - where the arrivals stop after a set number of individuals have arrived, which was kept at 1865 as the simulation was run until max customers (1865 individuals) reached the exit node.

Table 7.17 shows that the selected model is not sustainable as the target majorly decreases with each increase of demand. Exploring this in Table 7.18 shows that the main increase in waiting times are activity A, B and I. This is unsurprising as these are the main first activities in the pathways (Table 5.1). In conclusion, this suggests that if the demand increases then the capacity at the first activities needs to be considered accordingly.

Table 7.17: Top Level Results for Demand Investigation.

Name	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
Standard	65.08, (63.34, 66.82)	56.72, (54.4, 59.04)	53.12, (50.34, 55.9)	407.66, (407.46, 407.86)
Selected	42.84, (40.69, 44.99)	39.64, (36.86, 42.42)	87.27, (85.46, 89.07)	399.42, (399.22, 399.61)
Increase 5%	53.12, (51.18, 55.06)	53.92, (51.46, 56.38)	70.56, (65.14, 75.98)	399.62, (399.37, 399.88)
Increase 10%	62.04, (59.66, 64.42)	64.12, (61.71, 66.53)	45.39, (39.78, 50.99)	400.26, (399.92, 400.6)

Table 7.18: Activity Waiting Time Results for Demand Investigation.

Activity	Original	Standard	Selected	Increase 5%	Increase 10%
A	12.52	23.25, (21.61, 24.9)	9.78, (8.79, 10.76)	15.19, (13.83, 16.54)	21.06, (19.46, 22.66)
B	11.86	9.64, (9.39, 9.88)	14.82, (13.79, 15.86)	18.8, (18.27, 19.33)	20.71, (20.28, 21.13)
C	9.40	1.97, (1.9, 2.05)	1.37, (1.34, 1.4)	1.54, (1.48, 1.59)	1.79, (1.69, 1.89)
D	21.91	0.73, (0.69, 0.77)	2.09, (1.98, 2.19)	1.7, (1.57, 1.83)	1.51, (1.38, 1.63)
E	11.19	7.53, (7.41, 7.64)	2.73, (2.66, 2.81)	2.79, (2.72, 2.86)	2.86, (2.79, 2.93)
F	20.20	18.29, (18.11, 18.47)	3.78, (3.66, 3.91)	3.56, (3.46, 3.65)	3.38, (3.29, 3.48)
G	6.41	1.42, (1.36, 1.48)	6.09, (5.93, 6.25)	5.84, (5.67, 6.01)	5.66, (5.48, 5.84)
H	40.21	34.04, (33.89, 34.19)	10.86, (10.78, 10.93)	10.85, (10.78, 10.91)	10.82, (10.74, 10.9)
I	4.05	1.51, (1.38, 1.65)	2.78, (2.41, 3.14)	4.78, (4.32, 5.23)	6.79, (6.06, 7.53)
J	13.66	14.02, (13.83, 14.21)	2.56, (2.48, 2.65)	2.51, (2.43, 2.59)	2.24, (2.12, 2.36)
K	3.58	16.58, (16.38, 16.77)	5.08, (4.95, 5.21)	4.82, (4.66, 4.98)	4.29, (4.08, 4.5)
L	3.85	5.36, (5.19, 5.52)	10.13, (9.92, 10.33)	8.89, (8.65, 9.12)	8.03, (7.7, 8.36)
M	3.20	3.02, (2.93, 3.11)	4.03, (3.87, 4.19)	3.65, (3.48, 3.83)	3.18, (3.03, 3.33)

7.8 Medoids Capacity Investigation

The purpose of this final investigation considers the impact on capacity if a set of exact pathways were implemented.

Taking the *Process Medoids* solution from Section 5.4 which selected a set of 21 pathways that proportionately represented the variation displayed in the original data (Table 5.7). After running the *Process Medoids* standard automated model (as described in Section 7.1) for one run (seed 0), the demand for each activity was compared with the original values as seen in Table 7.19. The activity demand does vary in the *Process Medoids* model, displayed in Appendix E (Table E.4 for the interested reader). This showed that for the high demand activities (A, B, C, I and M) the demand levels were relatively similar to the original. The ceiling of the percentage difference was calculated by dividing the medoids demand by the original demand, which was then applied to the number of capacity slots per week (rounded to the ceiling value) to give an adjusted amount of capacity to consider. It was decided to only adjust the capacity for those activities using less than 100% of the capacity in the first instance, to observe the effect of only decreasing the capacity from its original levels.

Table 7.19: Capacity Adjustment for Medoids Capacity Investigation.

Activity	Original	Medoids	Difference	Percentage	Capacity	Adjusted	Difference
A	1797	1799	+2	101%	35	35	0 (+3*)
B	1865	1865	+0	100%	35	35	0
C	1855	1865	+10	101%	38	38	0
D	232	64	-168	28%	9	3	-6
E	473	197	-276	42%	12	6	-6
F	475	197	-278	42%	11	5	-6
G	536	225	-311	42%	13	6	-7
H	537	225	-312	42%	13	6	-7
I	1797	1799	+2	101%	36	36	0
J	537	64	-473	12%	13	2	-11
K	589	148	-441	26%	15	4	-11
L	364	73	-291	21%	10	3	-7
M	1577	1484	-93	95%	31	30	-1

*Note: Activity A was increased by a further 3 slots a week in the 'Additional' model.

There are four capacity models explored, where the first two (Medoids and Smoothed) retain the original capacity levels for the default and smoothed patterns respectively, to allow for comparison with the adjusted models.

Table 7.20: Capacity Patterns for Medoids Capacity Investigation.

Activity	Pattern	Smoothed	Adjusted	Additional
A	35, (6, 7, 10, 6, 6)	35, (7, 7, 7, 7, 7)	35, (7, 7, 7, 7, 7)	38, (8, 8, 8, 7, 7)
B	35, (6, 7, 8, 7, 7)	35, (7, 7, 7, 7, 7)	35, (7, 7, 7, 7, 7)	35, (7, 7, 7, 7, 7)
C	38, (5, 7, 9, 16, 1)	38, (8, 8, 8, 7, 7)	38, (8, 8, 8, 7, 7)	38, (8, 8, 8, 7, 7)
D	9, (2, 2, 1, 2, 2)	9, (2, 2, 2, 2, 1)	3, (1, 1, 1, 0, 0)	3, (1, 1, 1, 0, 0)
E	12, (2, 3, 2, 3, 2)	12, (3, 3, 2, 2, 2)	6, (2, 1, 1, 1, 1)	6, (2, 1, 1, 1, 1)
F	11, (2, 2, 2, 2, 3)	11, (3, 2, 2, 2, 2)	5, (1, 1, 1, 1, 1)	5, (1, 1, 1, 1, 1)
G	13, (2, 3, 2, 3, 3)	13, (3, 3, 3, 2, 2)	6, (2, 1, 1, 1, 1)	6, (2, 1, 1, 1, 1)
H	13, (5, 2, 2, 2, 2)	13, (3, 3, 3, 2, 2)	6, (2, 1, 1, 1, 1)	6, (2, 1, 1, 1, 1)
I	36, (7, 7, 7, 7, 8)	36, (8, 7, 7, 7, 7)	36, (8, 7, 7, 7, 7)	36, (8, 7, 7, 7, 7)
J	13, (2, 4, 2, 3, 2)	13, (3, 3, 3, 2, 2)	2, (1, 1, 0, 0, 0)	2, (1, 1, 0, 0, 0)
K	15, (2, 4, 4, 3, 2)	15, (3, 3, 3, 3, 3)	4, (1, 1, 1, 1, 0)	4, (1, 1, 1, 1, 0)
L	10, (2, 2, 2, 2, 2)	10, (2, 2, 2, 2, 2)	3, (1, 1, 1, 0, 0)	3, (1, 1, 1, 0, 0)
M	31, (5, 7, 9, 5, 5)	31, (7, 6, 6, 6, 6)	30, (6, 6, 6, 6, 6)	30, (6, 6, 6, 6, 6)

The Adjusted model applies the adjusted capacity from Table 7.20 where despite the large adjustment in capacity levels, the top level results still perform very well, achieving a 97% target (Table 7.21). However, the waiting time for activity A is disproportionately large compared to the other activities. Thus a further model is explored where an additional 3 slots a week are added for activity A (model Additional). This allows the waiting times for all the activities to become more equally spread, whilst not majorly negatively impacting the top level results and still allowing for the 95% target to be achieved. In conclusion, having a specific set of pathway allows, on the whole, for a considerable reduction in the capacity required, to achieve a good performance.

Table 7.21: Top Level Results for Medoids Capacity Investigation.

Name	Mean TiS	Median TiS	Target	Overall Period
Original	60	41	64.4	362
Medoids	35.12, (33.42, 36.82)	35.68, (33.97, 37.39)	98.71, (98.51, 98.91)	391.5, (391.05, 391.95)
Smoothed	33.92, (31.58, 36.26)	34.76, (32.06, 37.46)	98.1, (96.65, 99.54)	390.82, (390.46, 391.17)
Adjusted	35.52, (33.52, 37.52)	35.8, (33.46, 38.14)	97.31, (96.86, 97.75)	391.07, (390.75, 391.39)
Additional	36.52, (34.49, 38.55)	36.4, (34.14, 38.66)	96.26, (95.37, 97.15)	390.54, (390.12, 390.95)

Table 7.22: Activity Waiting Time Results for Medoids Capacity Investigation.

Activity	Original	Medoids	Smoothed	Adjusted	Additional
A	12.52	24.99, (23.37, 26.61)	24.99, (22.91, 27.07)	24.79, (23.04, 26.54)	11.44, (10.32, 12.56)
B	11.86	7.4, (6.97, 7.83)	6.95, (6.59, 7.31)	7.0, (6.56, 7.45)	15.76, (15.17, 16.35)
C	9.40	1.2, (1.18, 1.22)	0.31, (0.28, 0.33)	0.3, (0.27, 0.32)	0.33, (0.31, 0.35)
D	21.91	0.22, (0.15, 0.29)	0.22, (0.16, 0.29)	1.85, (1.68, 2.03)	1.81, (1.68, 1.94)
E	11.19	0.06, (0.05, 0.07)	0.06, (0.05, 0.07)	0.87, (0.8, 0.94)	0.81, (0.76, 0.85)
F	20.20	0.22, (0.17, 0.27)	0.2, (0.15, 0.24)	1.48, (1.26, 1.69)	1.5, (1.29, 1.71)
G	6.41	0.06, (0.05, 0.06)	0.05, (0.04, 0.07)	0.99, (0.88, 1.1)	1.09, (0.99, 1.19)
H	40.21	5.89, (5.4, 6.38)	5.73, (5.12, 6.35)	12.17, (10.54, 13.81)	13.46, (12.06, 14.85)
I	4.05	1.65, (1.5, 1.8)	1.76, (1.44, 2.08)	1.57, (1.39, 1.76)	2.66, (2.22, 3.11)
J	13.66	0.01, (0.0, 0.02)	0.0, (-0.0, 0.01)	4.56, (4.04, 5.08)	4.38, (3.94, 4.81)
K	3.58	0.03, (0.02, 0.04)	0.01, (0.0, 0.01)	4.22, (3.43, 5.01)	3.94, (3.19, 4.7)
L	3.85	0.01, (0.01, 0.02)	0.01, (0.0, 0.01)	2.32, (2.11, 2.53)	2.1, (1.93, 2.28)
M	3.20	0.66, (0.6, 0.71)	0.58, (0.55, 0.62)	0.94, (0.83, 1.04)	5.81, (4.82, 6.81)

7.9 Conclusion

This section summarises the results of the seven investigations. Table 7.23 displays the capacity patterns and comparative capacity adjustments for each investigation. Furthermore, for each investigation it was concluded that:

1. **Individual Adjustment Investigation:** Concentrating the focus of improvement of a particular activity will result in unintended changes. This further highlights the complexities of working with a highly complex pathway system.
2. **Basic Investigation:** It might not be beneficial to focus on improvements on mid pathway activities as this appears to just move the waiting time to other points in the pathway and it could be suggested that finding gains in end pathway activities could have a larger impact on the target results.
3. **End Activity Investigation:** Improving the end activities were easier to navigate as they had little effect on other activities and displayed a greater improvement in the results than the Basic Investigation.
4. **Target Investigation:** No model could reach the 95% desired target, with the final chosen model taking into account balancing the amount of capacity increase (KA_HFM_JEH) achieving 87%. This raised the question if this target is actually achievable. This is subsequently further investigated in the Excessive Top Down Investigation.
5. **Excessive Top Down Investigation:** It is possible to achieve the 95% target through increasing the capacity at a minimum of 2 and maximum of 7 slots per week. A major change from the previous investigation was reducing the time blocked in the Warm Start for activity H to 40 days.
6. **Demand Investigation:** If the demand increases then the target results quickly become unachievable and notes that the capacity at the pathway first activities needs to be considered accordingly.

7. **Medoids Capacity Investigation:** Having a specific set of pathway allows, on the whole, for a considerable reduction in the capacity required, to achieve a good performance.

Further work

There are two main areas to consider for further work - investigation specific and Sim.Pro.Flow development, as follows:

- Investigations: The Excessive Top Down Investigation could be explored further to find an optimal solution where the amount of capacity varies per day.
- Development: Extending the outputs produced for trials.
- Development: Allowing for increase in demand for the *Raw Pathways* (Demand Investigation) could be implemented in Sim.Pro.Flow.
- Development: The model run time can be improved, as it took between 8 and 25 minutes to complete 25 runs of the simulation. It should be noted that the simulation itself took less than 5 seconds to run each time, however, converting the results into the desired format took up the majority of the run time. Therefore further work should explore the results conversion to allow easier execution of trials.

Overall the aims of the chapter were satisfied as it was displayed that Sim.Pro.Flow can support typical simulation exploration.

Table 7.23: Capacity Results Patterns for All Investigations.

Activity	Standard	DGIKCA			Basic			End		
		C*	D	P	C	D	P	C	D	P
A	35, (6, 7, 10, 6, 6)	36, (7, 7, 10, 6, 6)	+1	+3%	38, (7, 7, 10, 7, 7)	+3	+9%	38, (7, 7, 10, 7, 7)	+3	+9%
B	35, (6, 7, 8, 7, 7)				38, (8, 8, 8, 7, 7)	+3	+9%			
C	38, (5, 7, 9, 16, 1)	33, (5, 7, 9, 11, 1)	-5	-13%						
D	9, (2, 2, 1, 2, 2)	6, (2, 1, 1, 1, 1)	-3	-33%						
E	12, (2, 3, 2, 3, 2)									
F	11, (2, 2, 2, 2, 3)							15, (3, 3, 3, 3, 3)	+4	+36%
G	13, (2, 3, 2, 3, 3)	12, (3, 3, 2, 2, 2)	-1	-8%						
H	13, (5, 2, 2, 2, 2)							21, (5, 4, 4, 4, 4)	+8	+62%
I	36, (7, 7, 7, 7, 8)	35, (7, 7, 7, 7, 7)	-1	-3%						
J	13, (2, 4, 2, 3, 2)				16, (4, 4, 3, 3, 2)	+3	+23%			
K	15, (2, 4, 4, 3, 2)	18, (4, 4, 4, 3, 3)	+3	+20%	18, (4, 4, 4, 3, 3)	+3	+20%	18, (4, 4, 4, 3, 3)	+3	+20%
L	10, (2, 2, 2, 2, 2)									
M	31, (5, 7, 9, 5, 5)				33, (7, 7, 9, 5, 5)	+2	+6%	33, (7, 7, 9, 5, 5)	+2	+6%

Activity	Standard	Target			Excessive			Medoids		
		C	D	P	C	D	P	C	D	P
A		38, (7, 7, 10, 7, 7)	+3	+9%	40, (8, 8, 8, 8, 8)	+5	+14%	38, (8, 8, 8, 7, 7)	+3	+9%
B					40, (8, 8, 8, 8, 8)	+5	+14%	35, (7, 7, 7, 7, 7)	0	0%
C					35, (7, 7, 7, 7, 7)	-3	-8%	38, (8, 8, 8, 7, 7)	0	0%
D					12, (2, 2, 2, 2, 2)	+3	+33%	3, (1, 1, 1, 0, 0)	-6	-33%
E		15, (3, 3, 3, 3, 3)	+3	+25%	15, (3, 3, 3, 3, 3)	+3	+25%	6, (2, 1, 1, 1, 1)	-6	-50%
F		15, (3, 3, 3, 3, 3)	+4	+36%	15, (3, 3, 3, 3, 3)	+4	+36%	5, (1, 1, 1, 1, 1)	-6	-55%
G					15, (3, 3, 3, 3, 3)	+2	+15%	6, (2, 1, 1, 1, 1)	-7	-54%
H		22, (5, 5, 4, 4, 4)	+9	+69%	15, (3, 3, 3, 3, 3)	+2	+15%	6, (2, 1, 1, 1, 1)	-7	-54%
I					35, (7, 7, 7, 7, 7)	-1	-3%	36, (8, 7, 7, 7, 7)	0	0%
J		16, (4, 4, 3, 3, 2)	+3	+23%	20, (4, 4, 4, 4, 4)	+7	+54%	2, (1, 1, 0, 0, 0)	-11	-85%
K		18, (4, 4, 4, 3, 3)	+3	+20%	20, (4, 4, 4, 4, 4)	+5	+33%	4, (1, 1, 1, 1, 0)	-11	-73%
L					15, (3, 3, 3, 3, 3)	+5	+50%	3, (1, 1, 1, 0, 0)	-7	-70%
M		33, (7, 7, 9, 5, 5)	+2	+6%	35, (7, 7, 7, 7, 7)	+4	+13%	30, (6, 6, 6, 6, 6)	-1	-3%

*C = total slots, (weekly pattern), D = comparative difference with the standard, and P = comparative adjusted percentage with the standard.

Chapter 8

Conclusion

This chapter closes the thesis with a summary of the work, presents the contributions made and suggestions for further work.

8.1 Summary

This research project was funded by KESS2 [152] in collaboration with a company partner - Velindre Cancer Centre (VCC) [286]. The desire of VCC was to produce a state-of-the-art decision support tool which ultimately supported efficient and sustainable methods of exploring a DES. This thesis investigated three main areas of developing a clinical pathway, namely, mapping, modelling and improving.

Chapter 1 introduced the idea of clinical pathways and cancer services. Investigations here found that proceeding with pathway mapping in a traditional manner would be very time consuming. Additionally, generating a model for a holistic view of cancer services would be an inconceivably large task. This identified the needs for VCC as 1) developing a process to build a model which is time efficient and sustainable to produce and 2) creating a state-of-the-art decision support tool. Furthermore, lung cancer was chosen as the specific cancer site for deeper investigation.

Chapter 2 presented a review of the vast literature covering clinical pathway mod-

elling in Information Systems (IS), Operational Research (OR) and Industrial Engineering. This chapter concluded many areas that further work should consider, three of which were addressed in this thesis:

- Derive the pathway from both data and collaboration with staff.
- Continue to bridge the gap between OR, IS and Industrial Engineering by considering data mining and machine learning alongside OR techniques, and integrate whenever possible.
- Incorporate all three areas of mapping, modelling and improving the pathway, with particular focus on improving, as this reflects the specialities of OR techniques.

The findings from Chapter 1 and 2 resulted in the research questions:

1. Can both data and expert information integrate to inform clinical pathway mapping?
2. Is it feasible to automate the simulation build process?
3. Is it viable to support multiple interpretations of clinical pathways through combining a mixture of data mining and OR?
4. Can the development of a decision support tool provide a general method of analysing clinical pathway mapping, modelling and improving?

Chapter 3 addressed research question 1 through the development of a distance metric, modified from the Needleman-Wunsch dynamic programming algorithm, that is specifically designed for clustering, and allows for expert interaction through the use of groupings and rankings of activities. The modified metric was compared against eight other popular metrics, where it performed equally well, if not better, when used with k-medoids clustering. This comparison further highlighted that each of the metrics produce different results and as such, confirms the hypothesis that careful consideration is needed when selecting a string metric.

Chapter 4 discussed the novel contribution of automatically building the discrete event simulation (DES), in response to research question 2. The chosen simulation software Ciw [56] was introduced, along with discussion of customisations (Section 1.5) required to support the model. Each of the DES input parameters (arrival, service, capacity and warm up) were explored in turn, where the *Raw Pathways* routing procedure was introduced for validating the chosen automation methods. Furthermore, a perishable inventory method of calculating capacity was discussed [12] (Section 1.5).

Chapter 5 combined the resulting methods from Chapter 3 and 4 by presenting various definitions of the clinical pathway for simulation, addressing research question 3. This chapter described three additional routing procedures, *Full Transitions*, *Cluster Transitions* and *Process Medoids*, and discusses their adaptations of the simulation build. The simulation for each routing procedure was run and the results were discussed.

Chapter 6 presented Sim.Pro.Flow, the decision support tool encompassing all of the research methods produced in Chapters 3, 4 and 5, beginning to address research question 4. The development of the tool satisfies the model sustainability required through its generic exploration of data and automated processes. Furthermore, this generalisation feature also widens the applicability and usability of the tool to not just cancer, not just healthcare even, but any process data that satisfies one of the input data types.

Finally, Chapter 7 continued to answer research question 4 through exploring a case study of seven investigations of the lung cancer pathways, evidencing that Sim.Pro.Flow is able to support typical exploration of the simulation.

8.2 Contributions

The initial broad scope of this research allowed for exploration across multiple areas of IS, OR and Industrial Engineering. As a result there are five main novel contributions produced from this work:

1. The literature review (Chapter 2) developed of a number of taxonomies, providing a detailed classification of the publications. This enables clarity for any future publications surrounding clinical pathways to identify the current themes and methods used in the literature, and thus identify gaps. The sheer number of papers included in the literature review along with lack of comparable literature reviews evidenced the need for this extensive review.
2. The development of the modified Needleman-Wunsch algorithm (Chapter 3) with the purpose of adding context to the string through combining data and expert information, whilst additionally considering the application specifically for process data combined with clustering.
3. The presentation of automating the simulation build, through both building the network and populating the input parameters. This allows for an efficient and sustainable method of building a DES.
4. Designing and developing the decision support tool Sim.Pro.Flow provides accessibility to the methods developed as part of this research. Furthermore, the built-in custom flexibility allows for the end decision to be with the user. Additionally the development process of production alongside technical/methodical evolution ensures that the methods developed are compatible with the end product.
5. Finally, the generic approach i.e. representing the pathways as strings, supporting multiple data types and routing procedures, allows for a wide application of the research in its entirety.

8.3 Further Work

Each chapter individually discussed very specific areas for further work. The aspects of further work can be categorised as either perfecting/developing the technical methods or as broader suggestions. The following areas were identified:

- Chapter 3: Careful consideration is needed when selecting a string metric as the user should be aware of the context for application. This suggests further work to apply meaning to the strings whilst considering the context, through either the development of more string metrics or through the perfecting elements within the developed modified Needleman-Wunsch algorithm.
- Chapter 4: Technically, the automated models could be expanded to consider a wider remit of applications i.e. consider service times of greater than one day for scenarios including inpatient stays. More broadly, it can be suggested to explore more ways to develop automated model build whilst retaining technical coherence and support flexibility.
- Chapter 5: The production of process based routing brings with it a wide range of possibilities to defining routing functions. Further work would be suggested to explore various methods of defining routing functions and progressing this as a more widely used method.
- Chapter 6: Specifically for Sim.Pro.Flow, further work could increase the usability and accessibility through 1) progressing out of prototype phase 2) evolving the support materials to a video platform and 3) developing an execution button which does not depend on command line knowledge.
- Chapter 7: The case study identified further technical aspects of Sim.Pro.Flow to develop, such as making the results conversion process more time efficient.

Finally, further work in general could consider applying the methods developed to other cancer sites using Sim.Pro.Flow, or expanding the remit outside of healthcare and evaluating the application to other industries using process data.

Bibliography

- [1] S.R. Abidi and S.S.R. Abidi. An ontological modeling approach to align institution-specific clinical pathways: Towards inter-institution care standardization. *Proceedings - 25th IEEE Symposium on Computer-Based Medical Systems*, pages 1–4, 2012.
- [2] H.H.A. Afzali, J. Karnon, and J. Gray. A critical review of model-based economic studies of depression: Modelling techniques, model structure and data sources. *PharmacoEconomics*, 30(6):461–482, 2012.
- [3] I. Ajmi, H. Zgaya, L. Gammoudi, S. Hammadi, A. Martinot, R. Beuscart, and J.-M. Renard. Mapping patient path in the pediatric emergency department: A workflow model driven approach. *Journal of Biomedical Informatics*, 54:315–328, 2015.
- [4] A. Alahmar, M.E. Crupi and R. Benlamri. Ontological framework for standardizing and digitizing clinical pathway in healthcare information systems. *Computer Methods and Programs in Biomedicine*, 196, 2020.
- [5] J.M. Albert and S. Nelson. Generalized causal mediation analysis. *Biometrics*, 67(3):1028–1038, 2011.
- [6] A. Alharbi, A. Bulpitt, and O.A. Johnson. Towards unsupervised detection of process models in healthcare. *Studies in Health Technology and Informatics*, 247:381–385, 2018.
- [7] M. Andellini, S. Fernandez Riesgo, F. Morolli, M. Ritrovato, P. Cosoli, S. Petruzzellis, and N. Rosso. Experimental application of business process management technology to manage clinical pathways: A pediatric kidney transplantation follow up case. *BMC Medical Informatics and Decision Making*, 17(1):151, 2017.
- [8] AnyLogic. The AnyLogic Company. Last Accessed: 17/02/2021. <https://www.anylogic.com/>
- [9] Arena. Rockwell Automation. Last Accessed: 17/02/2021. <https://www.arenasimulation.com/>
- [10] R. Argiento, A. Guglielmi, E. Lanzarone, and I. Nawajah. A bayesian framework for describing and predicting the stochastic demand of home care patients. *Flexible Services and Manufacturing Journal*, 28(1-2):254–279, 2016.
- [11] I.V. Arnolds and D. Gartner. Improving hospital layout planning through clinical pathway mining. *Annals of Operations Research*, 263(1-2):453–477, 2018.

-
- [12] E.F. Arruda, P. Harper, T. England, D. Gartner, E. Aspland, F.O. Ourique, T. Crosby. Resource optimization for cancer pathways with aggregate diagnostic demand: a perishable inventory approach. *IMA Journal of Management Mathematics*, 32(2):221-236, 2020.
- [13] M. Askari, J.L.Y.Y. Tam, M.F. Aarnoutse and M. Meulendijk. Perceived effectiveness of clinical pathway software: A before-after study in the Netherlands. *International Journal of Medical Informatics*, 135:104052, 2020.
- [14] M. Askari, J.L.Y.Y. Tam and J. Klundert. The effectiveness of clinical pathway software in inpatient settings: A systematic review. *International Journal of Medical Informatics*, 147:104374, 2021.
- [15] E.L. Aspland, D. Gartner, and P.R. Harper. Clinical pathway modelling: A literature review. *Health Systems*, 10(1):1-23, 2021.
- [16] E. Aspland, P.R. Harper, D. Gartner, P. Webb and P. Barrett-Lee. Modified Needleman-Wunsch algorithm for clinical pathway clustering. *Journal of Biomedical Informatics*, 115, 2021.
- [17] Y. Asukai, M. Baldwin, T. Fonseca, A. Gray, L. Mungapen, and D. Price. Improving clinical reality in chronic obstructive pulmonary disease economic modelling: Development and validation of a micro-simulation approach. *PharmacoEconomics*, 31(2):151-161, 2013.
- [18] N. Bahou, C. Fenwick, G. Anderson, R. van der Meer, and T. Vassalos. Modeling the critical care pathway for cardiothoracic surgery. *Health Care Management Science*, 21(2):192-203, 2018.
- [19] K. Baker, E. Dunwoodie, R.G. Jones, A. Newsham, O. Johnson, C.P. Price, J. Wolstenholme, J. Leal, P. McGinley, C. Twelves, and G. Hall. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. *International Journal of Medical Informatics*, 103:32-41, 2017.
- [20] M. Bakker and K.-L. Tsui. Dynamic resource allocation for efficient patient scheduling: A data-driven approach. *Journal of Systems Science and Systems Engineering*, 26(4):448-462, 2017.
- [21] P. Balatsoukas, R. Williams, C. Davies, J. Ainsworth, and I. Buchan. User interface requirements for web-based integrated care pathways: Evidence from the evaluation of an online care pathway investigation tool. *Journal of Medical Systems*, 39(11):183, 2015.
- [22] S. Barbagallo, L. Corradi, J. De Ville De Goyet, M. Iannucci, I. Porro, N. Rosso, E. Tanfani, and A. Testi. Optimization and planning of operating theatre activities: An original definition of pathways and process modeling. *BMC Medical Informatics and Decision Making*, 15(1):38, 2015.
- [23] P. Barone, F. Imbimbo, R. Napoletano, S. Riemma, and D. Sarno. A simulation tool to plan daily nurse requirements. *4th International Workshop on Innovative Simulation for Health Care, IWISH 2015*, pages 61-65, 2015.

- [24] S. Bayer, C. Petsoulas, B. Cox, A. Honeyman, and J. Barlow. Facilitating stroke care planning through simulation modelling. *Health Informatics Journal*, 16(2):129–143, 2010.
- [25] P.C. Bell. Visual interactive modelling in operational research: successes and opportunities. *Journal of the Operational Research Society*, 36(11):975–982, 1985.
- [26] P.C. Bell and R.M. O’Keefe. Visual Interactive Simulation - History, recent developments, and major issues. *SIMULATION*, 49(3):109–116, 1987.
- [27] S. Ben Othman, H. Zgaya, S. Hammadi, A. Quilliot, A. Martinot, and J.-M. Renard. Agents endowed with uncertainty management behaviors to solve a multiskill healthcare task scheduling. *Journal of Biomedical Informatics*, 64:25–43, 2016.
- [28] I. Bendavid, Y.N. Marmor, and B. Shnits. Developing an optimal appointment scheduling for systems with rigid standby time under pre-determined quality of service. *Flexible Services and Manufacturing Journal*, 30(1-2):54–77, 2018.
- [29] M.W. Bending, P. Trueman, K.V. Lowson, H. Pilgrim, P. Tappenden, J. Chilcott, and J. Tappenden. Estimating the direct costs of bowel cancer services provided by the national health service in england. *International Journal of Technology Assessment in Health Care*, 26(4):362–369, 2010.
- [30] J.H. Bettencourt-Silva, G.S. Mannu, and B. de la Iglesia. Visualisation of integrated patient-centric data as pathways: Enhancing electronic medical records in clinical practice. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9605 LNCS:99–124, 2016.
- [31] A. Bijlani, A.E. Hebert, M. Davitian, H. May, M. Speers, R. Leung, N.E. Mohamed, H.S. Sacks, and A. Tewari. A multidimensional analysis of prostate surgery costs in the united states: Robotic-assisted versus retropubic radical prostatectomy. *Value in Health*, 19(4):391–403, 2016.
- [32] L. Bleser, R. de Depreitere, K. de Waele, K. Vanhaecht, and J.E.A. Vlayen. Defining pathways. *Journal of Nursing Management*, 14(7):553–563, 2006.
- [33] C. Born, M. Carbajal, P. Smith, M. Wallace, K. Abbott, S. Adyanthaya, E.A. Boyd, C. Keller, J. Liu, W. New, T. Rieger, B. Winemiller, and R. Woestemeyer. Contract optimization at texas children’s hospital. *Interfaces*, 34(1 SPEC. ISS.):51–58, 2004.
- [34] J. Bowers, G. Mould, and C. Marshall. Location of services and the impact on healthcare quality: Insights from a simulation of a musculoskeletal physiotherapy service. *Journal of the Operational Research Society*, 66(7):1212–1221, 2015.
- [35] J. Bowles, M.B. Caminati, and S. Cha. An integrated framework for verifying multiple care pathways. *Proceedings - 11th International Symposium on Theoretical Aspects of Software Engineering, TASE 2017*, pages 1–8, 2018.

- [36] B.D. Bradley, S.R.C. Howie, T.C.Y. Chan, and Y.-L. Cheng. Estimating oxygen needs for childhood pneumonia in developing country health systems: A new model for expecting the unexpected. *PLoS ONE*, 9(2): e89872, 2014.
- [37] S.C. Brailsford, T.B. Bolt, G. Bucci, T.M. Chaussalet, N.A. Connell, P.R. Harper, J.H. Klein, M. Pitt, and M. Taylor. Overcoming the barriers: A qualitative study of simulation adoption in the nhs. *Journal of the Operational Research Society*, 64(2):157–168, 2013.
- [38] S.C. Brailsford, V.A. Lattimer, P. Tarnaras, and J.C. Turnbull. Emergency and on-demand health care: Modelling a large complex system. *Journal of the Operational Research Society*, 55(1):34–42, 2004.
- [39] S. Brailsford, P. Harper, B. Patel and M. Pitt. An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, 3(3):130–140, 2009.
- [40] S. Brice, P. Harper, T. Crosby, D. Gartner, E. Arruda, T. England, E. Aspland and K. Foley. Factors influencing the delivery of cancer pathways: A summary of the literature. *Journal of Health Organization and Management*, 35(9):121–139, 2021.
- [41] B. Brown, P. Balatsoukas, R. Williams, M. Sperrin, and I. Buchan. Multi-method laboratory user evaluation of an actionable clinical performance information system: Implications for usability and patient safety. *Journal of Biomedical Informatics*, 77:62–80, 2018.
- [42] S. Bruzzi, P. Landa, E. Tànfani, and A. Testi. Conceptual modelling of the flow of frail elderly through acute-care hospitals: An evidence-based management approach. *Management Decision*, 56(10):2101–2124, 2018.
- [43] R.L. Burdett and E. Kozan. An integrated approach for scheduling health care activities in a hospital. *European Journal of Operational Research*, 264(2):756–773, 2018.
- [44] R.L. Burdett, E. Kozan, M. Sinnott, D. Cook, and Y.-C. Tian. A mixed integer linear programming approach to perform hospital capacity assessments. *Expert Systems with Applications*, 77:170–188, 2017.
- [45] Cancer Research UK. Your Cancer Type. Last Accessed: 17/02/2021. <https://www.cancerresearchuk.org/about-cancer/type>
- [46] Cancer Research UK. Types of Lung Cancer. Last Accessed: 17/02/2021. <https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/types>
- [47] Cancer Research UK. National optimal lung cancer pathway. Last Accessed: 17/02/2021. https://www.cancerresearchuk.org/sites/default/files/national_optimal_lung_pathway_aug_2017.pdf
- [48] B. Cardoen and E. Demeulemeester. Capacity of clinical pathways - a strategic multi-level evaluation tool. *Journal of Medical Systems*, 32(6):443–452, 2008.

- [49] F. Caron, J. Vanthienen, K. Vanhaecht, E. Van Limbergen, J. Deweerdt, and B. Baesens. A process mining-based investigation of adverse events in care processes. *Health Information Management Journal*, 43(1):16–25, 2014.
- [50] M.E. Celebi, H.A. Kingravi and P.A. Vela. A comparative study of efficient initialization methods for the K-Means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.
- [51] P. Chemweno, L. Brackenier, V. Thijs, L. Pintelon, A. Van Horenbeek, and D. Michiels. Optimising the complete care pathway for cerebrovascular accident patients. *Computers and Industrial Engineering*, 93:236–251, 2016.
- [52] P. Chemweno, V. Thijs, L. Pintelon, and A. Van Horenbeek. Discrete event simulation case study: Diagnostic path for stroke patients in a stroke unit. *Simulation Modelling Practice and Theory*, 48:45–57, 2014.
- [53] P. Chemweno, V. Thijs, L. Pintelon, A. Van Horenbeek, and J. Samyn. Simulating the transfer logistics for stroke patients between a stroke unit and the rehabilitation center of a large university hospital. *ILS 2014 - 5th International Conference on Information Systems, Logistics and Supply Chain*, 2014.
- [54] J. Chen, L. Sun, C. Guo, W. Wei, and Y. Xie. A data-driven framework of typical treatment process extraction and evaluation. *Journal of Biomedical Informatics*, 83:178–195, 2018.
- [55] J. Chen, W. Wei, C. Guo, L. Tang, and L. Sun. Textual analysis and visualization of research trends in data mining for electronic health records. *Health Policy and Technology*, 6(4):389–400, 2017.
- [56] Ciw: 2.1.0. The Ciw library developers. Last Accessed: 17/02/2021. <https://github.com/CiwPython/Ciw>
- [57] Ciw Documentation. Ciw: 2.0.1. The Ciw library developers. Last Accessed: 17/02/2021. <https://ciw.readthedocs.io/en/latest/index.html>
- [58] Ciw Documentation. Ciw: 2.0.1. Tutorial VII: Multiple Classes of Customer. The Ciw library developers. Last Accessed: 17/02/2021. https://ciw.readthedocs.io/en/latest/Tutorial-II/tutorial_vii.html
- [59] Ciw Documentation. Ciw: 2.0.1. How to Set Arrival & Service Distributions. The Ciw library developers. Last Accessed: 17/02/2021. https://ciw.readthedocs.io/en/latest/Guides/set_distributions.html
- [60] Ciw Documentation. Ciw: 2.0.1. How to Set a Seed. The Ciw library developers. Last Accessed: 17/02/2021. <https://ciw.readthedocs.io/en/latest/Guides/seed.html>
- [61] Ciw Documentation. Ciw: 2.0.1. Tutorial IV: Trials, Warm-up & Cool-down. The Ciw library developers. Last Accessed: 17/02/2021. https://ciw.readthedocs.io/en/latest/Tutorial-I/tutorial_iv.html
- [62] Ciw Issue 171, Fixed Number of Customers from Each Class. Raised by KAI10 on May 20 2020. Solved by geraintpalmer May 27 2020. Closed June 4 2020.

- Last Accessed: 17/02/2021. <https://github.com/CiwPython/Ciw/issues/171>
- [63] L. Claxton, R. Hodgson, M. Taylor, B. Malcolm, and R. Pulikottil Jacob. Simulation modelling in ophthalmology: Application to cost effectiveness of ranibizumab and aflibercept for the treatment of wet age-related macular degeneration in the united kingdom. *PharmacoEconomics*, 35(2):237–248, 2017.
- [64] T. Comans, M. Raymer, S. O’Leary, D. Smith, and P. Scuffham. Cost-effectiveness of a physiotherapist-led service for orthopaedic outpatients. *Journal of Health Services Research and Policy*, 19(4):216–223, 2014.
- [65] T.A. Comans, A.T. Chang, L. Standfield, D. Knowles, S. O’Leary, and M. Raymer. The development and practical application of a simulation model to inform musculoskeletal service delivery in an australian public health service. *Operations Research for Health Care*, 15:13–18, 2017.
- [66] D.A. Cook, K.J. Sorensen, J.A. Linderbaum, L.J. Pencille, and D.J. Rhodes. Information needs of generalists and specialists using online best-practice algorithms to answer clinical questions. *Journal of the American Medical Informatics Association*, 24(4):754–761, 2017.
- [67] K. Cooper, R. Davies, J. Raftery, and P. Roderick. Use of a coronary heart disease simulation model to evaluate the costs and effectiveness of drugs for the prevention of heart disease. *Journal of the Operational Research Society*, 59(9):1173–1181, 2008.
- [68] P-N. Tan, M. Steinbach and V. Kumar. Introduction to data mining. *Boston, MA : Pearson Addison Wesley*, Chapter 8, page 500, 2005.
- [69] J. Coughlan, J. Eatock, and T. Eldabi. Evaluating telemedicine: A focus on patient pathways. *International Journal of Technology Assessment in Health Care*, 22(1):136–142, 2006.
- [70] G.J. Crane, S.M. Kymes, J.E. Hiller, R. Casson, A. Martin, and J.D. Karnon. Accounting for costs, qalys, and capacity constraints: Using discrete-event simulation to evaluate alternative service delivery and organizational scenarios for hospital-based glaucoma services. *Medical Decision Making*, 33(8):986–997, 2013.
- [71] E.A. Crawford, P.J. Parikh, N. Kong, and C.V. Thakar. Analyzing discharge strategies during acute care: A discrete-event simulation study. *Medical Decision Making*, 34(2):231–241, 2014.
- [72] CRediT Authorship Contribution Statement. Elsevier. <https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement>. Last Accessed: 07/06/2021.
- [73] C.S.M. Currie and R.C.H. Cheng. A practical introduction to analysis of simulation output data. *2016 Winter Simulation Conference (WSC)*, 118–132, 2016.
- [74] A. Dagliati, L. Sacchi, A. Zambelli, V. Tibollo, L. Pavesi, J.H. Holmes, and R. Bellazzi. Temporal electronic phenotyping by mining careflows of breast cancer patients. *Journal of Biomedical Informatics*, 66:136–147, 2017.

- [75] N. Dalkey and O. Helmer. Experimental application of the DELPHI method to the use of experts. *Management Science*, 9(3):351–515, 1962.
- [76] F.J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.
- [77] datetime library. Python. <https://docs.python.org/3/library/datetime.html>. Last Accessed: 14/06/2021.
- [78] Y. Dauxais, T. Guyet, D. Gross-Amblard, and A. Happe. Discriminant chronicles mining: Application to care pathways analytics. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10259 LNAI:234–244, 2017.
- [79] R. Deja, W. Froelich, G. Deja, and A. Wakulicz-Deja. Hybrid approach to the generation of medical guidelines for insulin therapy for children. *Information Sciences*, 384:157–173, 2017.
- [80] K. de Luc, D. Kitchiner, A. Layton, E. Morris, Y. Murray and S. Overill. Developing care pathways: the handbook. *Radcliffe Medical Press*, page 179, 2001.
- [81] E. Demir, R. Lebcir, and S. Adeyemi. Modelling length of stay and patient flows: methodological case studies from the uk neonatal care services. *Journal of the Operational Research Society*, 65(4):532–545, 2014.
- [82] E. Demir and D. Southern. Enabling better management of patients: Discrete event simulation combined with the star approach. *Journal of the Operational Research Society*, 68(5):577–590, 2017.
- [83] P. Devapriya, C.T.B. Strömblad, M.D. Bailey, S. Frazier, J. Bulger, S.T. Kemberling, and K.E. Wood. Stratbam: A discrete-event simulation model to support strategic hospital bed capacity decisions. *Journal of Medical Systems*, 39(10):130, 2015.
- [84] F. Dexter, A. Macario, and E.U. Dexter. Computer simulation of changes in nursing productivity from early tracheal extubation of coronary artery bypass graft patients. *Journal of Clinical Anesthesia*, 10(7):593–598, 1998.
- [85] G. Du, Z. Jiang, X. Diao, and Y. Yao. Knowledge extraction algorithm for variances handling of cp using integrated hybrid genetic double multi-group cooperative pso and dpso. *Journal of Medical Systems*, 36(2):979–994, 2012.
- [86] G. Du, Z. Jiang, X. Diao, and Y. Yao. Intelligent ensemble t-s fuzzy neural networks with rcdpso_dm optimization for effective handling of complex clinical pathway variances. *Computers in Biology and Medicine*, 43(6):613–634, 2013.
- [87] G. Du, Z. Jiang, X. Diao, Y. Ye, and Y. Yao. Variances handling method of clinical pathways based on t-s fuzzy neural networks with novel hybrid learning algorithm. *Journal of Medical Systems*, 36(3):1283–1300, 2012.
- [88] G. Du, Z. Jiang, Y. Yao, and X. Diao. Clinical pathways scheduling using hybrid genetic algorithm. *Journal of Medical Systems*, 37(3):9945, 2013.

- [89] R. Dunbar, P. Naidoo, N. Beyers, and I. Langley. Operational modelling: The mechanisms influencing tb diagnostic yield in an xpertw mtb/rif-based algorithm. *International Journal of Tuberculosis and Lung Disease*, 21(4):381–388, 2017.
- [90] S.R. Earnshaw, A.P. Brogan, and C.L. McDade. Model-based cost-effectiveness analyses for prostate cancer chemoprevention: A review and summary of challenges. *Pharmacoeconomics*, 31(4):289–304, 2013.
- [91] J. Eatock, J. Lord, M. Trapero-Bertran, and A. Anagnostou. Discrete event simulation of whole care pathways to estimate cost-effectiveness in clinical guidelines. *Proceedings - Winter Simulation Conference*, pages 1447–1458, 2016.
- [92] M. Elbattah and O. Molloy. Towards improving modeling and simulation of clinical pathways: Lessons learned and future insights. *SIMULTECH 2015 - 5th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, Proceedings*, pages 508–514, 2015.
- [93] T.G. Erdogan and A. Tarhan. A goal-driven evaluation method based on process mining for healthcare processes. *Applied Sciences (Switzerland)*, 8(6):894, 2018.
- [94] T.G. Erdogan and A. Tarhan. Systematic mapping of process mining studies in healthcare. *IEEE Access*, 6:24543–25567, 2018.
- [95] A.V. Esensoy and M.W. Carter. High-fidelity whole-system patient flow modeling to assess health care transformation policies. *European Journal of Operational Research*, 266(1):221–237, 2018.
- [96] Fauman, M. Do physicians use practice guidelines? *Psychiatric Times*, 23(7):13, Gale Academic Onefile, 2006.
- [97] O. Fennelly, C. Blake, F. Desmeules, D. Stokes, and C. Cunningham. Patient-reported outcome measures in advanced musculoskeletal physiotherapy practice: a systematic review. *Musculoskeletal Care*, 16(1):188–208, 2018.
- [98] N. Fenton and M. Neil. Comparing risks of alternative medical diagnosis using bayesian arguments. *Journal of Biomedical Informatics*, 43(4):485–495, 2010.
- [99] R. Feyrer, U. Kunzmann, and M. Weyand. Computer-assisted process simulation: A suitable instrument for process optimization in hospitals [computerunterstützte prozesssimulation: Ein beitrag zur prozessoptimierung im op]. *Zentralblatt für Chirurgie*, 131(4):347–353, 2006.
- [100] R. Feyrer, U. Kunzmann, M. Weyand, and R. Cesnjevar. Process optimization by means of a computerized process simulation model in cardiac surgery. *Disease Management and Health Outcomes*, 14(2):91–97, 2006.
- [101] A.J. Fong, M. Smith, and A. Langerman. Efficiency improvement in the operating room. *Journal of Surgical Research*, 204(2):371–383, 2016.
- [102] A.A. Funkner, A.N. Yakovlev, and S.V. Kovalchuk. Data-driven modeling of clinical pathways using electronic health records. *Procedia Computer Science*, 121:835–842, 2017.

- [103] A.A. Funkner, A.N. Yakovlev, and S.V. Kovalchuk. Towards evolutionary discovery of typical clinical pathways in electronic health records. *Procedia Computer Science*, 119:234–244, 2017.
- [104] H. Furuhashi, K. Araki, T. Ogawa, and M. Ikeda. Effect on completion of clinical pathway for improving clinical indicator: Cases of hospital stay, mortality rate, and comprehensive-volume ratio. *Journal of Medical Systems*, 41(12):206, 2017.
- [105] P.H. Garthwaite, J.B. Chilcott, D.J. Jenkinson, and P. Tappenden. Use of expert knowledge in evaluating costs and benefits of alternative service provisions: A case study. *International Journal of Technology Assessment in Health Care*, 24(3):350–357, 2008.
- [106] D. Gartner, I.V. Arnolds, and S. Nickel. Improving hospital-wide patient scheduling decisions by clinical pathway mining. *Studies in Health Technology and Informatics*, 216:1066, 2015.
- [107] D. Gartner and R. Kolisch. Scheduling the hospital-wide flow of elective patients. *European Journal of Operational Research*, 233(3):689–699, 2014.
- [108] D. Gartner and R. Padman. Improving hospital-wide early resource allocation through machine learning. *Studies in Health Technology and Informatics*, 216:315–319, 2015.
- [109] J.A. George, R. Koka, T.J. Gan, E. Jelin, E.F. Boss, V. Strockbine, D. Hobson, E.C. Wick, and C.L. Wu. Review of the enhanced recovery pathway for children: perioperative anesthetic considerations [les programmes de récupération rapide pour les enfants: considérations anesthésiques périopératoires]. *Canadian Journal of Anesthesia*, 65(5):569–577, 2018.
- [110] M. Ghasemi and D. Amyot. Process mining in healthcare: A systematised literature review. *International Journal of Electronic Healthcare*, 9(1):60–88, 2016.
- [111] J. Gillespie, S. McClean, L. Garg, M. Barton, B. Scotney, and K. Fullerton. A multi-phase des modelling framework for patient-centred care. *Journal of the Operational Research Society*, 67(10):1239–1249, 2016.
- [112] A.S. Gordon, A.H. Marshall, and M. Zenga. Predicting elderly patient length of stay in hospital and community care using a series of conditional coxian phase-type distributions, further conditioned on a survival tree. *Health Care Management Science*, 21(2):269–280, 2018.
- [113] Graphviz. Python library. Last Accessed: 17/02/2021. <https://graphviz.readthedocs.io/en/stable/index.html>
- [114] M.M. Günel and M. Pidd. Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of Simulation*, 4(1):42–51, 2010.
- [115] S. Guo, K. Xu, R. Zhao, D. Gotz, H. Zha, and N. Cao. Eventthread: Visual summarization and stage analysis of event sequence data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):56–65, 2018.

- [116] B. Han, L. Jiang, and H. Cai. Abnormal process instances identification method in healthcare environment. *Proc. 10th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications, TrustCom 2011, 8th IEEE Int. Conf. on Embedded Software and Systems, ICESS 2011, 6th Int. Conf. on FCST 2011*, pages 1387–1392, 2011.
- [117] A. Happe and E. Drezen. A visual approach of care pathways from the french nationwide snds database – from population to individual records: the epeps toolbox. *Fundamental and Clinical Pharmacology*, 32(1):81–84, 2018.
- [118] L. He, S. Chalil Madathil, A. Oberoi, G. Servis, and M.T. Khasawneh. A systematic review of research design and modeling techniques in inpatient bed management. *Computers and Industrial Engineering*, 127:451–466, 2018.
- [119] Healthcare Improvement Scotland. Management of lung cancer, 2014. Last Accessed: 17/02/2021. <https://www.sign.ac.uk/media/1075/sign137.pdf>
- [120] K. Helbig, M. Römer, and T. Mellouli. A clinical pathway mining approach to enable scheduling of hospital relocations and treatment services. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9253:242–250, 2015.
- [121] A.R. Hevner, S.T. March, J. Park and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.
- [122] Z.M. Hira and D.F. Gillies. Identifying significant features in cancer methylation data using gene pathway segmentation. *Cancer Informatics*, 15:189–198, 2016.
- [123] K. Hoad, S. Robinson and R. Davies. Automating warm-up length estimation. *Journal of the Operational Research Society*, 61(9):1389–1403, 2010.
- [124] H. Huang, T. Jin, and J. Wang. Extracting clinical-event-packages from billing data for clinical pathway mining. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10219 LNCS:19–31, 2017.
- [125] Z. Huang, W. Dong, P. Bath, L. Ji, and H. Duan. On mining latent treatment patterns from electronic medical records. *Data Mining and Knowledge Discovery*, 29(4):914–949, 2015.
- [126] Z. Huang, Y. Bao, W. Dong, X. Lu, and H. Duan. Online treatment compliance checking for clinical pathways. *Journal of Medical Systems*, 38(10):123, 2014.
- [127] Z. Huang, W. Dong, H. Duan, and H. Li. Similarity measure between patient traces for clinical pathway analysis: Problem, method, and applications. *IEEE Journal of Biomedical and Health Informatics*, 18(1):4–14, 2014.
- [128] Z. Huang, W. Dong, L. Ji, and H. Duan. Predictive monitoring of clinical pathways. *Expert Systems with Applications*, 56:227–241, 2016.
- [129] Z. Huang, W. Dong, L. Ji, C. Gan, X. Lu, and H. Duan. Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of Biomedical Informatics*, 47:39–57, 2014.

- [130] Z. Huang, W. Dong, L. Ji, C. He, and H. Duan. Incorporating comorbidities into latent treatment pattern mining for clinical pathways. *Journal of Biomedical Informatics*, 59:227–239, 2016.
- [131] Z. Huang, W. Dong, L. Ji, L. Yin, and H. Duan. On local anomaly detection and analysis for clinical pathways. *Artificial Intelligence in Medicine*, 65(3):167–177, 2015.
- [132] Z. Huang, C. Gan, X. Lu, and H. Huan. Mining the changes of medical behaviors for clinical pathways. *Studies in Health Technology and Informatics*, 192(1-2):117–121, 2013.
- [133] Z. Huang, Z. Ge, W. Dong, K. He, and H. Duan. Probabilistic modeling personalized treatment pathways using electronic health records. *Journal of Biomedical Informatics*, 86:33–48, 2018.
- [134] Z. Huang, X. Lu, and H. Duan. On mining clinical pathway patterns from medical behaviors. *Artificial Intelligence in Medicine*, 56(1):35–50, 2012.
- [135] Z. Huang, X. Lu, and H. Duan. Latent treatment pattern discovery for clinical processes. *Journal of Medical Systems*, 37(2):9915, 2013.
- [136] Z. Huang, X. Lu, and H. Duan. Similarity measuring between patient traces for clinical pathway analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7885 LNAI:268–272, 2013.
- [137] Z. Huang, X. Lu, H. Duan, and W. Fan. Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, 46(1):111–127, 2013.
- [138] P. Hulshof, N. Kortbeek, R. Boucherie, E. Hans and P. Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems*, 1(2):129–175, 2012.
- [139] R.D. Hurrion. The design, use and required facilities of an interactive visual computer simulation language to explore production planning problems. *Ph.D. Thesis, University of London, England*, 1976.
- [140] R.D. Hurrion. Visual interactive simulation an aid to decision making. *Omega*, 6(5):419–426, 1978.
- [141] J.E. Hurwitz, J.A. Lee, K.K. Lopiano, S.A. McKinley, J. Keesling, and J.A. Tyndall. A flexible simulation platform to quantify and manage emergency department crowding. *BMC Medical Informatics and Decision Making*, 14(1):50, 2014.
- [142] Irish Cancer Society. Lung cancer action plan, 2019. Last Accessed: 17/02/2021. <https://www.cancer.ie/sites/default/files/2020-02/Irish%20Cancer%20Society%20Lung%20Action%20Plan%202019.pdf>
- [143] P. Jaccard. The distribution of the flora in the alpine zone. *The New Phytologist*, XI(2):37-50, 1912.
- [144] M.A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

- [145] M.A. Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14(5-7):491–498, 1995.
- [146] W.E. Winkler. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Bureau of the Census*, 1990.
- [147] P. Joranger, A. Nesbakken, G. Hoff, H. Sorbye, A. Oshaug, and E. Aas. Modeling and validating the cost and clinical pathway of colorectal cancer. *Medical Decision Making*, 35(2):255–265, 2015.
- [148] J. Karnon. Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. *Health Economics*, 12(10):837–848, 2003.
- [149] J. Karnon and T. Jones. A stochastic economic evaluation of letrozole versus tamoxifen as a first-line hormonal therapy: For advanced breast cancer in postmenopausal patients. *PharmacoEconomics*, 21(7):513–525, 2003.
- [150] T.M. Kashner, T.J. Carmody, T. Suppes, A.J. Rush, M.L. Crismon, A.L. Miller, M. Toprac, and M. Trivedi. Catching up on health outcomes: The texas medication algorithm project. *Health Services Research*, 38(1 I):311–331, 2003.
- [151] M. Keshtkaran, J. Hearne, B. Abbasi, and L. Churilov. Stroke care systems: Can simulation modeling catch up with the recent advances in stroke treatment? *Proceedings - Winter Simulation Conference*, pages 1379–1390, 2016.
- [152] KESS2. Knowledge Economy Skills Scholarships. Last Accessed: 17/02/2021. <http://kess2.ac.uk/>
- [153] L. Kinsman, T. Rotter, E. James, P. Snow and J. Willis. What is a clinical pathway? Development of a definition to inform the debate. *BMC Medicine*, 8(1):31, 2010.
- [154] L.M. Kolarczyk, H. Arora, M.W. Manning, D.A. Zvara, and R.S. Isaak. Defining value-based care in cardiac and vascular anesthesiology: The past, present, and future of perioperative cardiovascular care. *Journal of Cardiothoracic and Vascular Anesthesia*, 32(1):512–521, 2018.
- [155] R. Konrad, B. Tulu, and M. Lawley. Monitoring adherence to evidence-based practices: A method to utilize hl7 messages from hospital information systems. *Applied Clinical Informatics*, 4(1):126–143, 2013.
- [156] S.V. Kovalchuk, A.A. Funkner, O.G. Metsker, and A.N. Yakovlev. Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification. *Journal of Biomedical Informatics*, 82:128–142, 2018.
- [157] A.P. Kurniati, O. Johnson, D. Hogg, and G. Hall. Process mining in oncology: A literature review. *Proceedings of the 6th International Conference on Information Communication and Management (ICICM)*, pages 291–297, 2016.

- [158] M.M.H. Lahr, D.-J. Van Der Zee, G.-J. Luijckx, P.C.A.J. Vroomen, and E. Buskens. A simulation-based approach for improving utilization of thrombolysis in acute brain infarction. *Medical Care*, 51(12):1101–1105, 2013.
- [159] M.M.H. Lahr, D.-J. Van Der Zee, G.-J. Luijckx, P.C.A.J. Vroomen, and E. Buskens. Centralising and optimising decentralised stroke care systems: a simulation study on short-term costs and effects. *BMC Medical Research Methodology*, 17(1):1–12, 2017.
- [160] G.T. Lakshmanan, S. Rozsnyai, and F. Wang. Investigating clinical care pathways correlated with outcomes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8094 LNCS:323–338, 2013.
- [161] P. Landa, M. Sonnessa, E. Tànfani, and A. Testi. Multiobjective bed management considering emergency and elective patient flows. *International Transactions in Operational Research*, 25(1):91–110, 2018.
- [162] D.C. Lane and E. Husemann. System dynamics mapping of acute patient flows. *Journal of the Operational Research Society*, 59(2):213–224, 2008.
- [163] I. Langley, E. Adams, B. Doulla, and S.B. Squire. Operational modelling to guide implementation and scale-up of diagnostic tests within the health system: Exploring opportunities for parasitic disease diagnostics based on example application for tuberculosis. *Parasitology*, 141(14):1795–1802, 2014.
- [164] I. Langley, B. Doulla, H.-H. Lin, K. Millington, and B. Squire. Modelling the impacts of new diagnostic tools for tuberculosis in developing countries to enhance policy decisions. *Health Care Management Science*, 15(3):239–253, 2012.
- [165] I. Langley, H.-H. Lin, S. Egwaga, B. Doulla, C.-C. Ku, M. Murray, T. Cohen, and S.B. Squire. Assessment of the patient, health system, and population effects of xpert mtb/rif and alternative diagnostics for tuberculosis in tanzania: An integrated modelling approach. *The Lancet Global Health*, 2(10):e581–e591, 2014.
- [166] E. Lanzarone and A. Matta. The nurse-to-patient assignment problem in home care services. *International Series in Operations Research and Management Science*, 173:121–139, 2012.
- [167] E. Lanzarone, A. Matta, and E. Sahin. Operations management applied to home care services: The problem of assigning human resources to patients. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 42(6):1346–1363, 2012.
- [168] E. Lanzarone, A. Matta, and G. Scaccabarozzi. A patient stochastic model to support human resource planning in home care. *Production Planning and Control*, 21(1):3–25, 2010.
- [169] Averill M. Law. Simulation modeling and analysis. *McGraw-Hill Education*, Fifth Edition, 2014.
- [170] C. Legány, S. Juhász, and A. Babos. Cluster validity measurement techniques. *AIKED'06: Proceedings of the 5th WSEAS International Conference on Ar-*

- tificial Intelligence, Knowledge Engineering and Data Bases*, pages 388–393, 2006.
- [171] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8), 1966.
- [172] X. Li, H. Liu, J. Mei, Y. Yu, and G. Xie. Mining temporal and data constraints associated with outcomes for care pathways. *Studies in Health Technology and Informatics*, 216:711–715, 2015.
- [173] Y. Li, J.A. Schneider, and D.A. Bennett. Estimation of the mediation effect with a binary mediator. *Statistics in Medicine*, 26(18):3398–3414, 2007.
- [174] B. Lievrouw. Needleman-Wunsch Demo. Last Accessed: 17/02/2021. <https://blievrouw.github.io/needleman-wunsch/>
- [175] M.E. Lim, T. Nye, J.M. Bowen, J. Hurley, R. Goeree, and J.-E. Tarride. Mathematical modeling: The case of emergency department waiting times. *International Journal of Technology Assessment in Health Care*, 28(2):93–109, 2012.
- [176] F.-R. Lin, S.-C. Chou, S.-M. Pan, and Y.-M. Chen. Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics*, 62(1):11–25, 2001.
- [177] Y.C. Lin. Development of an ontology-based flexible clinical pathway system. *WSEAS Transactions on Information Science and Applications*, 6(12):1941–1952, 2009.
- [178] J. Lismont, A.-S. Janssens, I. Odnoletkova, S. vanden Broucke, F. Caron, and J. Vanthienen. A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways. *Computers in Biology and Medicine*, 77:125–134, 2016.
- [179] J. Liu, Z. Huang, X. Lu, and H. Duan. An ontology-based real-time monitoring approach to clinical pathway. *Proceedings - 2014 7th International Conference on BioMedical Engineering and Informatics (BMEI)*, pages 756–761, 2014.
- [180] R. Liu, R.V. Srinivasan, K. Zolfaghar, S.-C. Chin, S.B. Roy, A. Hasan, and D. Hazel. Pathway-finder: An interactive recommender system for supporting personalized care pathways. *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1219–1222, 2015.
- [181] Z. Liu, D. Rexachs, F. Epelde, and E. Luque. An agent-based model for quantitatively analyzing and predicting the complex behavior of emergency departments. *Journal of Computational Science*, 21:11–23, 2017.
- [182] J. Lord, S. Willis, J. Eatock, P. Tappenden, M. Trapero-Bertran, A. Miners, C. Crossan, M. Westby, A. Anagnostou, S. Taylor, I. Mavranouzouli, D. Wonderling, P. Alderson, and F. Ruiz. Economic modelling of diagnostic and treatment pathways in national institute for health and care excellence clinical guidelines: The modelling algorithm pathways in guidelines (mapguide) project. *Health Technology Assessment*, 17(58):1–150, 2013.

- [183] R.M. Luque-Baena, D. Urda, M. Gonzalo Claros, L. Franco, and J.M. Jerez. Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords. *Journal of Biomedical Informatics*, 49:32–44, 2014.
- [184] M. Mahdavi, T. Malmström, J. van de Klundert, S. Elkhuisen, and J. Visers. Generic operational models in health service operations management: A systematic review. *Socio-Economic Planning Sciences*, 47(4):271–280, 2013.
- [185] K. Maheshwari, J. Cywinski, P. Mathur, III Cummings, K.C. R. Avitsian, T. Crone, D. Liska, F.X. Campion, K. Ruetzler, and A. Kurz. Identify and monitor clinical variation using machine intelligence: a pilot in colorectal surgery. *Journal of Clinical Monitoring and Computing*, 33(4):725–731, 2018.
- [186] D. Maier. The complexity of some problems on subsequences and supersequences. *Journal of the ACM*, 25(2):322–336, 1978.
- [187] M. Maliapen and B.C. Dangerfield. A system dynamics-based simulation study for managing clinical governance and pathways in a hospital. *Journal of the Operational Research Society*, 61(2):255–264, 2010.
- [188] M.M. Malik, S. Abdallah, and M. Ala’raj. Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review. *Annals of Operations Research*, 270(1-2):287–312, 2018.
- [189] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and W. Van Der Aalst. Process mining techniques: An application to stroke care. *Studies in Health Technology and Informatics*, 136:573–578, 2008.
- [190] J. Marynissen and E. Demeulemeester. Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research*, 272(2):407–419, 2019.
- [191] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. Last Accessed: 17/02/2021. <https://matplotlib.org/>
- [192] Matplotlib Navigation Toolbar. Matplotlib: A 2D graphics environment. Last Accessed: 17/02/2021. https://matplotlib.org/3.1.1/users/navigation_toolbar.html
- [193] S. McClean, L. Garg, B. Meenan, and P. Millard. Using markov models to find interesting patient pathways. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, pages 713–718, 2007.
- [194] S. McClean, T. Young, D. Bustard, P. Millard, and M. Barton. Discovery of value streams for lean healthcare. *2008 4th International IEEE Conference Intelligent Systems (IS)*, 1:32–38, 2008.
- [195] Medline plus. Acute vs. Chronic Conditions. Last Accessed: 17/02/2021. <https://medlineplus.gov/ency/imagepages/18126.htm>
- [196] J. Meier, A. Dietz, A. Boehm, and T. Neumuth. Predicting treatment process steps from events. *Journal of Biomedical Informatics*, 53:308–319, 2015.

- [197] A. Meisami, J. Deglise-Hawkinson, M.E. Cowen, and M.P. van Oyen. Data-driven optimization methodology for admission control in critical care units. *Health Care Management Science*, pages 1–18, 2018.
- [198] R. Meskarian, M.L. Penn, T. Monks, M.A. Taylor, J. Klein, S.C. Brailsford, and P.R. Benson. Utilisation of health and social care services by the over 65s population. a system dynamics study. *Proceedings of the Operational Research Society Simulation Workshop 2016, SW 2016*, pages 218–227, 2016.
- [199] W. Michalowski, S. Wilk, A. Thijssen, and M. Li. Using a bayesian belief network model to categorize length of stay for radical prostatectomy patients: Using a bayesian belief network to categorize los. *Health Care Management Science*, 9(4):341–348, 2006.
- [200] M.A. Miranda, S. Salvatierra, I. Rodríguez, M.J. Álvarez and V.Rodríguez. Characterization of the flow of patients in a hospital from complex networks. *Health Care Management Science*, 23(1):66–79, 2020.
- [201] O. Mohammed and R. Benlamri. Developing a semantic web model for medical differential diagnosis recommendation. *Journal of Medical Systems*, 38(10):79, 2014.
- [202] T. Monks, C.S.M. Currie, B.S. Onggo, S. Robinson, M. Kunc and S.J.E. Taylor. Strengthening the reporting of empirical simulation studies: Introducing the STRESS guidelines. *Journal of Simulation*, 13(1):55–67, 2019.
- [203] T. Monks, K. Pearn, and M. Allen. Simulation of stroke care systems. *Proceedings - Winter Simulation Conference*, pages 1391–1402, 2016.
- [204] T. Monks, M. Pearson, M. Pitt, K. Stein, and M.A. James. Evaluating the impact of a simulation study in emergency stroke care. *Operations Research for Health Care*, 6:40–49, 2015.
- [205] T. Monks, M. Pitt, K. Stein, and M. James. Maximizing the population benefit from thrombolysis in acute ischemic stroke: A modeling study of in-hospital delays. *Stroke*, 43(10):2706–2711, 2012.
- [206] T. Monks, D. Worthington, M. Allen, M. Pitt, K. Stein, and M.A. James. A modelling tool for capacity planning in acute and community stroke services. *BMC Health Services Research*, 16(1):1–8, 2016.
- [207] A. Najjar, D. Reinharz, C. Girouard, and C. Gagné. A two-step approach for mining patient treatment pathways in administrative healthcare databases. *Artificial Intelligence in Medicine*, 87:34–48, 2018.
- [208] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.
- [209] The NHS Cancer Plan, A plan for investment A plan for reform. NHS, 2000. Last Accessed: 17/02/2021. https://www.thh.nhs.uk/documents/_Departments/Cancer/NHSCancerPlan.pdf

- [210] NHS Providers. The NHS Provider Sector. Last Accessed: 17/02/2021. <http://nhsproviders.org/topics/delivery-and-performance/the-nhs-provider-sector>
- [211] NHS Wales. National optimal pathway for lung cancer, 2019. Last Accessed: 17/02/2021. http://www.cancerservicesdirectory.wales.nhs.uk/sitesplus/documents/1112/NOP_Lung%20Cancer%20bFINAL.pdf
- [212] A. Noro, J.W. Poss, J.P. Hirdes, H. Finne-Soveri, G. Ljunggren, J. Björnsson, M. Schroll, and P.V. Jonsson. Method for assigning priority levels in acute care (maple-ac) predicts outcomes of acute hospital care of older persons - a cross-national validation. *BMC Medical Informatics and Decision Making*, 11(1):39, 2011.
- [213] Northern Ireland Cancer Network. Lung Pathway. Last Accessed: 17/02/2021. https://nican.hscni.net/wpfd_file/lung-pathway/
- [214] A. Novikov. PyClustering: Data Mining Library. *Journal of Open Source Software*, 4(36):1230, 2019.
- [215] B.S.S. Onggo, N.C. Proudlove, S.A. D’Ambrogio, A. Calabrese, S. Bisogno, and N. Levialedi Ghiron. A bpmn extension to support discrete-event simulation for healthcare applications: An explicit representation of queues, attributes and data-driven decision points. *Journal of the Operational Research Society*, 69(5):788–802, 2018.
- [216] Y.A. Ozcan, E. Tanfani, and A. Testi. A simulation-based modeling framework to deal with clinical pathways. *Proceedings - Winter Simulation Conference*, pages 1190–1201, 2011.
- [217] R. O’Keefe. Design science, the design of systems and operational research: Back to the future? *Journal of the Operational Research Society*, 65(5):673–684, 2014.
- [218] Y.A. Ozcan, E. Tanfani, and A. Testi. Improving the performance of surgery-based clinical pathways: a simulation-optimization approach. *Health Care Management Science*, 20(1):1–15, 2017.
- [219] G.I. Palmer, V.A. Knight, P.R. Harper and A.L. Hawa. Ciw: An open-source discrete event simulation library. *Journal of Simulation*, 13(1):68–82, 2019.
- [220] G. Palmer. Modelling deadlock in queueing systems. Thesis, Cardiff University, 2018.
- [221] A. Partington, M. Wynn, S. Suriadi, C. Ouyang, and J. Karnon. Process mining for clinical processes: A comparative analysis of four australian hospitals. *ACM Transactions on Management Information Systems*, 5(4):1–18, 2015.
- [222] PathSimR. Bristol, North Somerset and South Gloucestershire Clinical Commissioning Group (CCG). The Health Foundation, 2020. Last Accessed: 08/03/2021. <https://github.com/nhs-bnssg-analytics/PathSimR>
- [223] Developing a Versatile Tool for Modelling Pathway Capacity in NHS Organisations. Bristol, North Somerset and South Gloucestershire Clinical Commissioning Group (CCG). The Health Foundation, 2020. Last

- Accessed: 08/03/2021. <https://www.health.org.uk/improvement-projects/developing-a-versatile-tool-for-modelling-pathway-capacity-in-nhs-organisations>
- [224] PathSimR, Quick Start Guide. Bristol, North Somerset and South Gloucestershire Clinical Commissioning Group (CCG). The Health Foundation, 2020. Last Accessed: 08/03/2021. https://github.com/nhs-bnssg-analytics/PathSimR/blob/master/PathSimR_Shiny/documentation/quick_start_guide.pdf,
- [225] M.C. Peñaloza Ramos, P. Barton, S. Jowett, and A.J. Sutton. A systematic review of research guidelines in decision-analytic modeling. *Value in Health*, 18(4):512–529, 2015.
- [226] M. Peleg. Computer-interpretable clinical guidelines: A methodological review. *Journal of Biomedical Informatics*, 46(4):744–763, 2013.
- [227] J.M. Peña, J.A. Lozano and P. Larrañaga. An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recognition Letters*, 20(10):1027–1040, 1999.
- [228] A. Perer, F. Wang, and J. Hu. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*, 56:369–378, 2015.
- [229] S. Petrovic. A comparison between the silhouette index and the Davies-Bouldin index in labelling IDS clusters. *Proceedings of the 11th Nordic Workshop of Secure IT-Systems*, 2006.
- [230] H. Pilgrim, P. Tappenden, J. Chilcott, M. Bending, P. Trueman, A. Shorthouse, and J. Tappenden. The costs and benefits of bowel cancer service developments using discrete event simulation. *Journal of the Operational Research Society*, 60(10):1305–1314, 2009.
- [231] M. Pitt, T. Monks, S. Crowe, and C. Vasilakis. Systems modelling and simulation in health service design, delivery and decision making. *BMJ Quality and Safety*, 25(1):38–45, 2016.
- [232] J. Poirier, G.Y. Zou, and J. Koval. Confidence intervals for a difference between lognormal means in cluster randomization trials. *Statistical Methods in Medical Research*, 26(2):598–614, 2017.
- [233] C. Ponsard, R. De Landtsheer, Y. Guyot, F. Roucoux, and B. Lambeau. Decision making support in the scheduling of chemotherapy coping with quality of care, resources and ethical constraints. *ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems*, 1:460–470, 2017.
- [234] M. Prodel, V. Augusto, X. Xie, B. Jouaneton, and L. Lamarsalle. Discovery of patient pathways from a national hospital database using process mining and integer linear programming. *IEEE International Conference on Automation Science and Engineering*, pages 1409–1414, 2015.
- [235] ProM. Process Mining Group, Eindhoven University of Technology. Last Accessed: 17/02/2021. <https://www.promtools.org/doku.php>

- [236] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2020. Last Accessed: 08/03/2021. <https://www.r-project.org/>
- [237] S. Rachuba, A. Salmon, Z. Zhelev, and M. Pitt. Redesigning the diagnostic pathway for chest pain patients in emergency departments. *Health Care Management Science*, 21(2):177–191, 2018.
- [238] M. Ramos-Merino, L.M. Álvarez Sabucedo, J.M. Santos-Gago, and J. Sanz-Valero. A bpmn based notation for the representation of workflows in hospital protocols. *Journal of Medical Systems*, 42(10):181, 2018.
- [239] W. Rashwan, W. Abo-Hamad, and A. Arisha. A system dynamics view of the acute bed blockage problem in the irish healthcare system. *European Journal of Operational Research*, 247(1):276–293, 2015.
- [240] O. Rejeb, C. Pilet, S. Hamana, X. Xie, T. Durand, S. Aloui, A. Doly, P. Biron, L. Perrier, and V. Augusto. Performance and cost evaluation of health information systems using micro-costing and discrete-event simulation. *Health Care Management Science*, 21(2):204–223, 2018.
- [241] F. Rismanchian and Y.H. Lee. Process mining-based method of designing and optimizing the layouts of emergency departments in hospitals. *Health Environments Research and Design Journal*, 10(4):105–120, 2017.
- [242] S. Robinson. Simulation: the practice of model development and use. *Palgrave Macmillan*, 2014.
- [243] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro. Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, 61:224–236, 2016.
- [244] J. Roux, O. Grimaud and E. Leray. Use of state sequence analysis for care pathway analysis: The example of multiple sclerosis. *Statistical Methods in Medical Research*, 28(6):1651–1663, 2019.
- [245] D. Sarno and M.E. Nenni. Daily nurse requirements planning based on simulation of patient flows. *Flexible Services and Manufacturing Journal*, 28(3):526–549, 2016.
- [246] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [247] Scikit-learn documentation for Violin Plots. Scikit-Learn. Last Accessed: 17/02/2021. <https://scikit-learn.org/stable/modules/density.html>
- [248] Silhouette Score. Scikit-Learn. Last Accessed: 17/02/2021. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

- [249] V. Siless, S. Medina, G. Varoquaux, and B. Thirion. A comparison of metrics and algorithms for fiber clustering. *International Workshop on Pattern Recognition in NeuroImaging*, 190–193, 2013.
- [250] H.A. Simon. The sciences of the artificial. *MIT Press*, Cambridge, 3rd edition, 1996.
- [251] Sim.Pro.Flow. EmmaAspland, Github, 2021. Last Accessed: 17/02/2021. <https://github.com/EmmaAspland/Sim.Pro.Flow>
- [252] Sim.Pro.Flow Help File. EmmaAspland, Github, 2021. Last Accessed: 18/02/2021. https://github.com/EmmaAspland/Sim.Pro.Flow/blob/master/Sim.Pro.Flow/SimProFlow_Help.pdf
- [253] Sim.Pro.Flow Issues. EmmaAspland, Github, 2021. Last Accessed: 17/02/2021. <https://github.com/EmmaAspland/Sim.Pro.Flow/blob/master/Issues.txt>
- [254] Sim.Pro.Flow References. EmmaAspland, Github, 2021. Last Accessed: 17/02/2021. <https://github.com/EmmaAspland/Sim.Pro.Flow/blob/master/References.txt>
- [255] Sim.Pro.Flow Requirements. EmmaAspland, Github, 2021. Last Accessed: 17/02/2021. https://github.com/EmmaAspland/Sim.Pro.Flow/blob/master/requirements_versions.txt
- [256] SimPy. Discrete event simulation for Python. Team SimPy. Last Accessed: 17/02/2021. <https://simpy.readthedocs.io/en/latest/>
- [257] Simul8. Simul8 Corporation. Last Accessed: 17/02/2021. <https://www.simul8.com/>
- [258] C.-P. Shen, C. Jigjidsuren, S. Dorjgochoo, C.-H. Chen, W.-H. Chen, C.-K. Hsu, J.-M. Wu, C.-W. Hsueh, M.-S. Lai, C.-T. Tan, E. Altangerel, and F. Lai. A data-mining framework for transnational healthcare system. *Journal of Medical Systems*, 36(4):2565–2575, 2012.
- [259] N. Shukla, S. Lahiri, and D. Ceglarek. Pathway variation analysis (pva): Modelling and simulations. *Operations Research for Health Care*, 6:61–77, 2015.
- [260] N.F. Smedley, B.M. Ellingson, T.F. Cloughesy, and W. Hsu. Longitudinal patterns in clinical and imaging measurements predict residual survival in glioblastoma patients. *Scientific Reports*, 8(1):14429, 2018.
- [261] M. Snyder, and W. Zhou. Big data and health. *The Lancet Digital Health*, 1(6):e252–e254, 2019.
- [262] B.G. Sobolev, V. Sanchez, and C. Vasilakis. Systematic review of the use of computer simulation modeling of patient flow in surgical care. *Journal of Medical Systems*, 35(1):1–16, 2011.
- [263] E.M. Soffin and J.T. Yadeau. Enhanced recovery after surgery for primary hip and knee arthroplasty: A review of the evidence. *British Journal of Anaesthesia*, 117(suppl.3):iii62–iii72, 2016.

- [264] K.W. Soh, C. Walker, and M. O’Sullivan. A literature review on validated simulations of the surgical services. *Journal of Medical Systems*, 41(4):61, 2017.
- [265] L. Standfield, T. Comans, and P. Scuffham. Markov modeling and discrete event simulation in health care: A systematic comparison. *International Journal of Technology Assessment in Health Care*, 30(2):165–172, 2014.
- [266] Cancer Waiting Times. StatsWales. Last Accessed: 09/03/2021. <https://statswales.gov.wales/Catalogue/Health-and-Social-Care/NHS-Hospital-Waiting-Times/Cancer-Waiting-Times/Monthly/pre-February-2021>
- [267] A. Stefanini, D. Aloini, R. Dulmin, and V. Mininno. Linking diagnostic-related groups (drgs) to their processes by process mining. *HEALTHINF 2016 - 9th International Conference on Health Informatics, Proceedings; Part of 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, pages 438–443, 2016.
- [268] M. Stuit, H. Wortmann, N. Szirbik, and J. Roodenburg. Multi-view interaction modelling of human collaboration processes: A business process study of head and neck cancer care in a dutch academic hospital. *Journal of Biomedical Informatics*, 44(6):1039–1055, 2011.
- [269] H. Syed and A.K. Das. Identifying chemotherapy regimens in electronic health record data using interval-encoded sequence alignment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9105:143–147, 2015.
- [270] P. Tappenden, J. Chilcott, A. Brennan, H. Squires, and M. Stevenson. Whole disease modeling to inform resource allocation decisions in cancer: A methodological framework. *Value in Health*, 15(8):1127–1136, 2012.
- [271] S.J.E. Taylor, A. Anagnostou, T. Monks, C. Currie, B.S. Onggo, M. Kunc and S. Robinson. Applying the STRESS guidelines for reproducibility in modeling & simulation: Application to a disease modeling case study. *2018 Winter Simulation Conference (WSC)*, 739–748, 2018.
- [272] A. Teixeira and M.J.F. De Oliveira. Operations research on hospital admission systems: A first overview of the 2005-2014 decade. *Journal of Physics: Conference Series*, 616(1), 2015.
- [273] E. Tànfani and A. Testi. A simulation-based decision support tool to analyze clinical pathways in hospital. *International Series in Operations Research and Management Science*, 173:191–211, 2012.
- [274] Textdistance. Python library for comparing distance between two or more sequences by many algorithms. Last Accessed: 17/02/2021. <https://pypi.org/project/textdistance/>
- [275] A. Tolarczyk and K. Siwek. Sequential pattern recognition for medical records analysis. *Proceedings of 2016 17th International Conference Computational Problems of Electrical Engineering (CPEE)*, pages 1–3, 2016.

- [276] I.A. Trajano, J.B.F. Filho, F.R.C. Sousa, I. Litchfield and P. Weber. MedPath: A process-based modeling language for designing care pathways. *International Journal of Medical Informatics*, 146:104328, 2021.
- [277] S. Tsumoto, H. Iwata, S. Hirano, and Y. Tsumoto. Similarity-based behavior and process mining of medical practices. *Future Generation Computer Systems*, 33:21–31, 2014.
- [278] The unittest library. Last Accessed: 17/02/2021. <https://docs.python.org/3/library/unittest.html>
- [279] K. Uragaki, T. Hosaka, Y. Arahori, M. Kushima, T. Yamazaki, K. Araki, and H. Yokota. Sequential pattern mining on electronic medical records with handling time intervals and the efficacy of medicines. *Proceedings - IEEE Symposium on Computers and Communications*, pages 20–25, 2016.
- [280] E. Uzun Jacobson, S. Bayer, J. Barlow, M. Dennis, and M.J. MacLeod. The scope for improvement in hyper-acute stroke care in scotland. *Operations Research for Health Care*, 6:50–60, 2015.
- [281] J. van de Klundert, P. Gorissen, and S. Zeemering. Measuring clinical pathway adherence. *Journal of Biomedical Informatics*, 43(6):861–872, 2010.
- [282] D.-J. Van Der Zee, M.M.H. Lahr, G.-J. Luijckx, and E. Buskens. Simulation conceptual modeling for optimizing acute stroke care organization. *Proceedings - Winter Simulation Conference*, pages 1403–1414, 2016.
- [283] R. van Zelm, I. Janssen, K. Vanhaecht, A. de Buck van Overstraeten, M. Panella, W. Sermeus, and E. Coeckelberghs. Development of a model care pathway for adults undergoing colorectal cancer surgery: Evidence-based key interventions and indicators. *Journal of Evaluation in Clinical Practice*, 24(1):232–239, 2018.
- [284] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, and N. Litvak. Efficiency evaluation for pooling resources in health care. *OR Spectrum*, 34(2):371–390, 2012.
- [285] S.A. Vanderby, M.W. Carter, T. Noseworthy, and D.A. Marshall. Modelling the complete continuum of care using system dynamics: The case of osteoarthritis in alberta. *Journal of Simulation*, 9(2):156–169, 2015.
- [286] Velindre Cancer Centre. Last Accessed: 17/02/2021. <http://www.velindrecc.wales.nhs.uk/home>
- [287] V. Vogt, S.M. Scholz, and L. Sundmacher. Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data. *European Journal of Public Health*, 28(2):214–219, 2018.
- [288] Wales Cancer Network. Cancer Delivery Plan for Wales 2016-2020, 2016. Last Accessed: 17/02/2021. <http://www.walescanet.wales.nhs.uk/sitesplus/documents/1113/Cancer%20Delivery%20Plan%202016-2020.pdf>

- [289] Wales Cancer Network. National Optimal pathways. Last Accessed: 17/02/2021. <http://www.walescanet.wales.nhs.uk/national-optimal-pathways>
- [290] Wales Cancer Network. Single Cancer Pathway. Last Accessed: 17/02/2021. <http://www.walescanet.wales.nhs.uk/single-cancer-pathway>
- [291] Wales Cancer Network. Single Cancer Pathway Briefing Paper, 2017. Last Accessed: 17/02/2021. <http://www.walescanet.wales.nhs.uk/sitesplus/documents/1113/SCP%20Briefing%20Document1.pdf>
- [292] Wales Cancer Network. Single Suspected Cancer Pathway Definitions - pathway start date, 2018. Last Accessed: 17/02/2021. <http://www.walescanet.wales.nhs.uk/sitesplus/documents/1113/Point%20of%20suspicion%20%28SCP%29%20definitions%20-Final.pdf>
- [293] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018.
- [294] F. Walter, A. Webster, S. Scott, and J. Emery. The andersen model of total patient delay: A systematic review of its application in cancer diagnosis. *Journal of Health Services Research and Policy*, 17(2):110–118, 2012.
- [295] T. Wang, X. Tian, M. Yu, X. Qi, and L. Yang. Stage division and pattern discovery of complex patient care processes. *Journal of Systems Science and Complexity*, 30(5):1136–1159, 2017.
- [296] W. Wang, S. Nelson, and J.M. Albert. Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Statistics in Medicine*, 32(24):4211–4228, 2013.
- [297] Lung Cancer in Wales. Lung Cancer Survival and Survival by Stage. Welsh Cancer Intelligence and Surveillance Unit, Health Intelligence Division, Public Health Wales. 2015. Last Accessed: 17/02/2021. <https://phw.nhs.wales/services-and-teams/welsh-cancer-intelligence-and-surveillance-unit-wcisu/awareness-posters-and-information/lung-cancer/lung-cancer-docs/lung-cancer-in-wales-lung-cancer-survival-and-survival-by-stage/>
- [298] Written Statement: Single Cancer Pathway public reporting. Vaughan Gething AM, Cabinet Secretary for Health and Social Services. Welsh Government, 2018. Last Accessed: 17/02/2021. <https://gov.wales/written-statement-single-cancer-pathway-public-reporting>
- [299] Bar Chart. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Bar_chart
- [300] Cosine Similarity. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Cosine_similarity

-
- [301] Damerau-Levenshtein Distance. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Damerau%E2%80%93Levenshtein_distance
- [302] Directed Graph. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Directed_graph
- [303] L.A. Wolsey and G.L. Nemhauser. Integer and combinatorial optimization. *John Wiley & Sons*, 55, 1999.
- [304] Exponential Distribution. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Exponential_distribution
- [305] Floor and Ceiling Functions. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Floor_and_ceiling_functions
- [306] Histogram. Wikipedia. Last Accessed: 17/02/2021. <https://en.wikipedia.org/wiki/Histogram>
- [307] Jaccard Index. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Jaccard_index
- [308] Jaro-Winkler Distance. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance
- [309] k-Medoids. Wikipedia. Last Accessed: 17/02/2021. <https://en.wikipedia.org/wiki/K-medoids>
- [310] Levenshtein Distance. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Levenshtein_distance
- [311] Line Chart. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Line_chart
- [312] Longest Common Subsequence Problem. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Longest_common_subsequence_problem
- [313] Needleman-Wunsch Algorithm. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm
- [314] Silhouette (Clustering). Wikipedia. Last Accessed: 17/02/2021. [https://en.wikipedia.org/wiki/Silhouette_\(clustering\)](https://en.wikipedia.org/wiki/Silhouette_(clustering))
- [315] Violin Plot. Wikipedia. Last Accessed: 17/02/2021. https://en.wikipedia.org/wiki/Violin_plot
- [316] R.M. Wood and B.J. Murch. Modelling capacity along a patient pathway with delays to transfer and discharge. *Journal of the Operational Research Society*, 71(10):1530–1544, 2019.
- [317] World Health Organisation. Cancer. Last Accessed: 17/02/2021. <https://www.who.int/news-room/fact-sheets/detail/cancer>
- [318] wxPython. The GUI Toolkit for Python, The wxPython Team. Last Accessed: 17/02/2021. <https://www.wxpython.org/pages/overview/>

- [319] W. Xu, Y. Zhu, and Y. Geng. Development of an open metadata schema for clinical pathway (opencp) in china. *Methods of Information in Medicine*, 57(4):159–167, 2018.
- [320] X. Xu, T. Jin, and J. Wang. Summarizing patient daily activities for clinical pathway mining. *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6, 2016.
- [321] X. Xu, T. Jin, Z. Wei, C. Lv, and J. Wang. Tcpcm: Topic-based clinical pathway mining. *Proceedings - 2016 IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 292–301, 2016.
- [322] X. Xu, T. Jin, Z. Wei, and J. Wang. Incorporating domain knowledge into clinical goal discovering for clinical pathway mining. *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 261–264, 2017.
- [323] X. Xu, T. Jin, Z. Wei, and J. Wang. Incorporating topic assignment constraint and topic correlation limitation into clinical goal discovering for clinical pathway mining. *Journal of Healthcare Engineering*, 2017, 2017.
- [324] H. Yan, P. Van Gorp, U. Kaymak, X. Lu, L. Ji, C.C. Chiau, H.H.M. Korsten, and H. Duan. Aligning event logs to task-time matrix clinical pathways in bpmn for variance analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(2):311–317, 2018.
- [325] W.-S. Yang and S.-Y. Hwang. A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, 31(1):56–58, 2006.
- [326] X. Yang, R. Han, Y. Guo, J. Bradley, B. Cox, R. Dickinson, and R. Kitney. Modelling and performance analysis of clinical pathways using the stochastic process algebra PEPA. *BMC Bioinformatics*, 13(14):S4, 2012.
- [327] L. Yin, W. Dong, Z. Huang, L. Ji, X. Lv, and H. Duan. On detecting the changes of medical behaviors in clinical pathways. *Chinese Journal of Biomedical Engineering*, 34(3):272–280, 2015.
- [328] L. Yin, Z. Huang, W. Dong, C. He, and H. Duan. Utilizing electronic medical records to discover changing trends of medical behaviors over time. *Methods of Information in Medicine*, 56(S 01):e49–e66, 2017.
- [329] S. Yoo, M. Cho, S. Kim, E. Kim, S.M. Park, K. Kim, H. Hwang, and M. Song. Conformance analysis of clinical pathway using electronic health record data. *Healthcare Informatics Research*, 21(3):161–166, 2015.
- [330] K. Zander, M. Etheredge and K. Bower. Nursing case management: Blueprints for transformation. *New England Medical Center Hospital*, page 1128, 1987.
- [331] X. Zhang. Application of discrete event simulation in health care: A systematic review. *BMC Health Services Research*, 18(1):687, 2018.
- [332] Y. Zhang and R. Padman. Data-driven clinical and cost pathways for chronic care delivery. *American Journal of Managed Care*, 22(12):816–820, 2016.

-
- [333] Y. Zhang and R. Padman. An interactive platform to visualize data-driven clinical pathways for the management of multiple chronic conditions. *Studies in Health Technology and Informatics*, 245:672–676, 2017.
- [334] Y. Zhang, R. Padman and J. Levin. Paving the cowpath: data-driven design of pediatric order sets. *Journal of the American Medical Informatics Association*, 21(e2):e304e311, 2014.
- [335] Y. Zhang, R. Padman, and N. Patel. Paving the cowpath: Learning and visualizing clinical pathways from electronic health record data. *Journal of Biomedical Informatics*, 58:186–197, 2015.
- [336] M.E. Zonderland, R.J. Boucherie, and A. Al Hanbali. Appointments in care pathways: the geox/d/1 queue with slot reservations. *Queueing Systems*, 79(1):37–51, 2014.
- [337] A. Zwerling, R.G. White, A. Vassall, T. Cohen, D.W. Dowdy, and R.M.G.J. Houben. Modeling of novel diagnostic strategies for active tuberculosis - a systematic review: Current practices and recommendations. *PLoS ONE*, 9(10): e110558, 2014.

Appendices

Appendix A

Preliminary Investigation

This appendix contains the results of the preliminary investigation from subsection 1.3.1 which asked experts to annotate the converted NOLCP (Figure A.1) with how they view the pathway in practice.

The experts consulted were from different care levels that interact with the pathway at various stages. For secondary care, Helen Howison, a specialist nurse at St Woolos hospital (Aneurin Bevan University Health Board) and for secondary/tertiary care, Dr Mick Button, a clinical oncologist and deputy clinical director at Velindre Cancer Centre.

Anything in red refers to a difference from the NOLCP. Figure A.2 displays the original annotations from Helen Howison at St Woolos which were interpreted to form the pathway in Figure A.3, and Figure A.4 displays the interpretation the pathway discussed with Dr Mick Button.

Due to the conclusions of the preliminary investigation from subsection 1.3.1, the pathways displayed in Figure A.2, A.3 and A.4 are incomplete and should not be taken as a true reflection of the pathway.

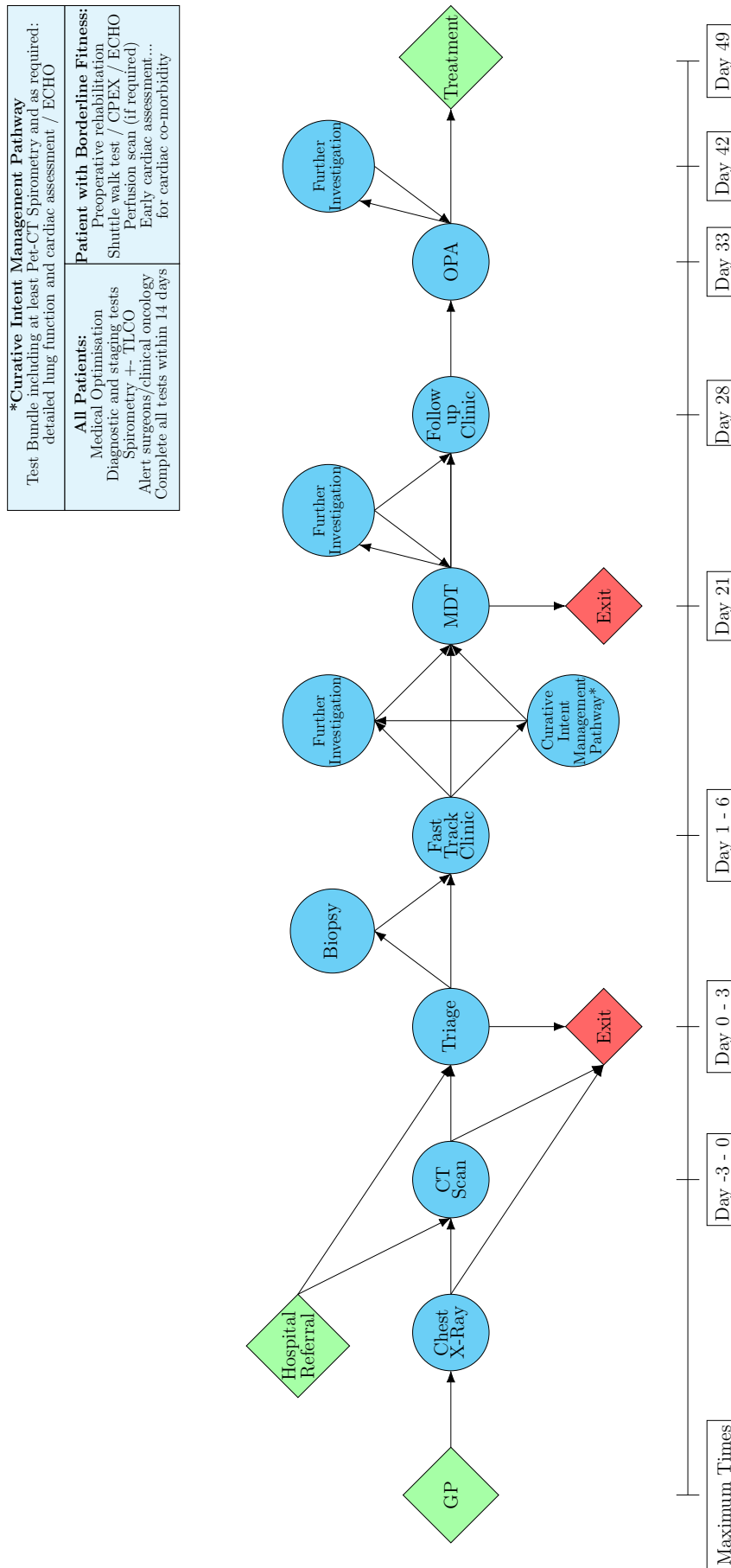


Figure A.1: Simplified National Optimal Lung Cancer Pathway.

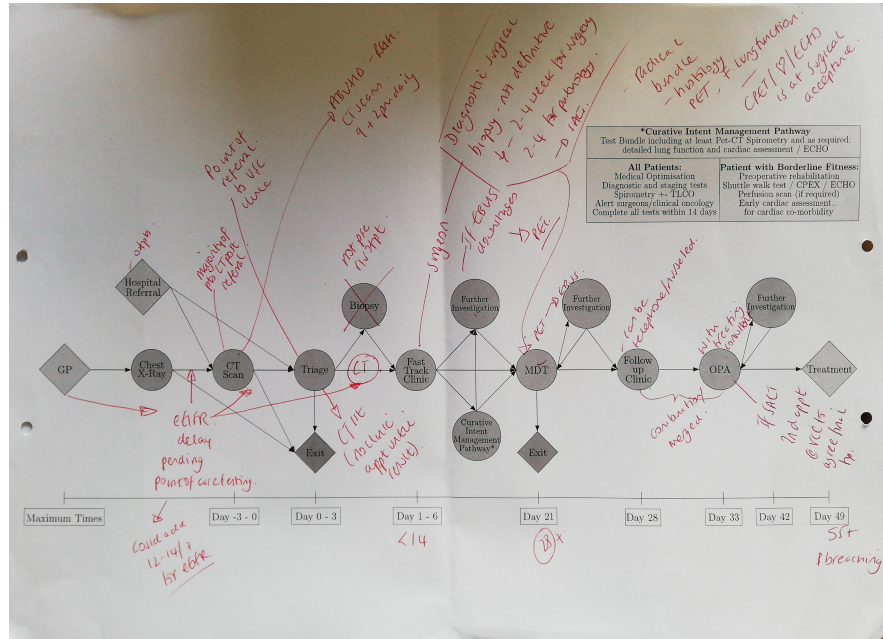


Figure A.2: Pathway Mapping Exercise: Original St Woolos.

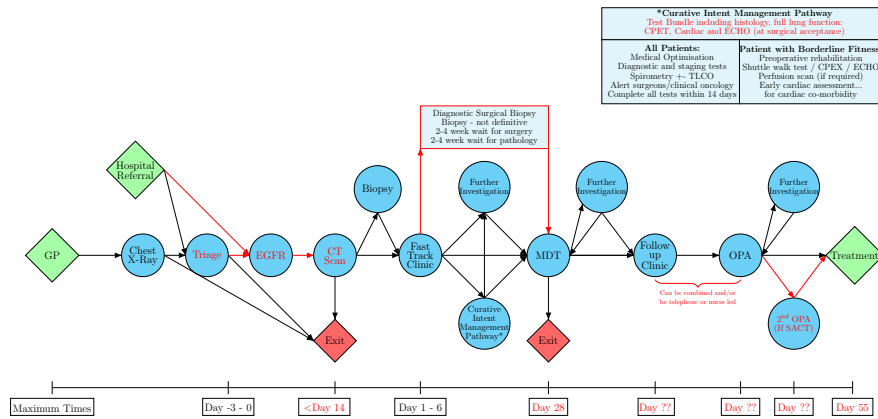


Figure A.3: Pathway Mapping Exercise: St Woolos Interpretation.

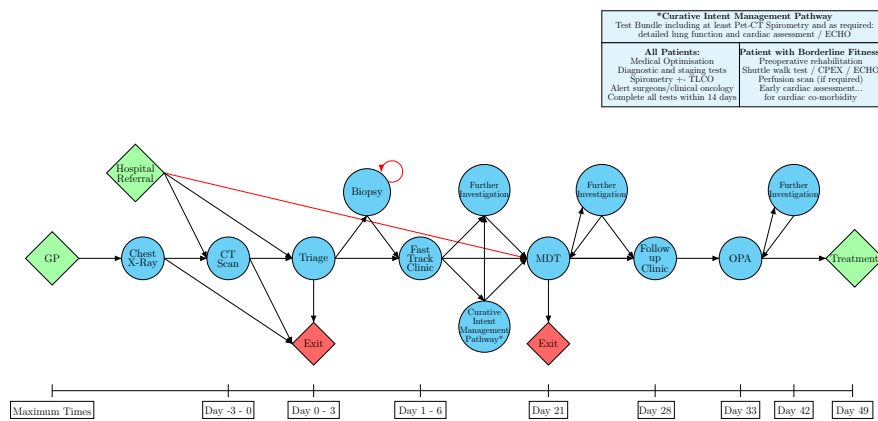


Figure A.4: Pathway Mapping Exercise: Velindre Interpretation.

Appendix B

Literature Review Tables

Table B.1: Literature Review Table For: Papers of Notable Contribution.

Reference	Type	Summary	Year Published
[7]	Case Study	Business Process Management technology applied to clinical pathways. Pediatric kidney transplantation case study.	2017
[37]	Case Study	Factors influencing successful adoption of general simulation tools within healthcare organisations.	2013
[41]	I.T.Artifact	Develops a Electronic audit and feedback. (e-A&F) system.	2018
[69]	Evaluation	Telemedicine research.	2006
[154]	Review	Overview of the past, present and future of preoperative cardiovascular care.	2018
[182]	Guidelines and Case Study	Test the feasibility of building full National Institute for Health and Care Excellence (NICE) guidelines models for cost-effectiveness analysis.	2013
[231]	Survey and Case Study	Delivery and design across healthcare planning. A case study in emergency stroke care is presented as an exemplar.	2015
[263]	Narative Review	Constructing enhanced recovery after surgery (ER-AHS) pathways for hip and knee arthroplasty.	2016
[270]	Framework	Developing health economic models of whole systems of disease and treatment pathways for resource allocation decisions.	2012
[293]	I.T Artifact	Synthea - an open-source software package that simulates the lifespan of synthetic patients.	2018
[329]	Conformance analysis	Conformance rates of actual usage of clinical pathway using Electronic Health Record log data.	2015

Table B.2: Literature Review Table For: Summary of Previous Literature Reviews.

Ref	Type	Summary	Dates included	No. Search Engines	No. Papers Included	Major Search Terms
[2]	Literature Review	Model-based cost-utility studies of depression.	2002-2010	4	14	Depression, models, computer simulation, cost utility, QALY
[55]	Literature Survey	Data mining for electronic health records (EHRs).	2000-2016	3	2516	Data mining, machine learning, artificial intelligence, mining, AND electronic health record (EHR)
[90]	Literature Review	Mathematical decision models that evaluated 5 α -reductable inhabitants and prostate cancer chemoprevention.	until 2013	2	7	Cost-effectiveness, cost-utility, decision-analytic, economic model, AND prostate cancer
[92]	Systematic Review	Clinical pathways.	NA	4	22	Clinical pathways, Critical pathways, Care maps, Integrated care pathways, Care pathways
[94]	Systematic Mapping	Process mining in healthcare.	2005-2017	10	172	"Process mining" in healthcare and clinical pathways
[97]	Systematic Review	Patient-reported outcome measures (PROMs) being utilized by Advanced practice physiotherapists (APPs).	until 2018	5	38	Physio Therapy, Advanced Practice, Patient-reported outcome measures
[101]	Systematic Review	Intraoperative efficacy improvement in surgery.	until 2015	1	38	Operating room, surgery, surgical AND Efficiency, Lean, Six Sigma
[109]	Literature Review	Compare adult and pediatric Enhanced Recovery After Surgery (ERAS) pathways.	1940-2018	1	NA/83	Enhanced Recovery, Fast Track, surgery AND child
[110]	Literature Review	Process mining in healthcare	until 2016	8	2371	Process mining, healthcare, clinic, hospital, care, health
[118]	Systematic Review	Inpatient bed management.	2013-2017	3	92	Bed management, bed assignment, bed planning, bed allocation
[151]	Literature Review	Simulation modelling in stroke care systems.	until 2014	1	30	Stroke, simulation, simulation model
[157]	Literature Review	Process mining in oncology	until 2016	5	37	Process mining, data mining, pathway analysis, AND patient flow, AND oncology
[175]	Literature Review	Mathematical modelling for evaluating waiting times in a hospital emergency department.	2000-2010	5	29	NA
[184]	Literature Review	Application of generic operational models in health services	until 2013	2	116	Operational model AND health, care, Clinical pathway, Simulation, Markov
[188]	Systematic Review	Data mining and predictive analytics in healthcare operations and supply chain management.	until 2015	2	22	Big data, data mining, process mining AND optimizations AND, healthcare
[190]	Literature Review	Multi-appointment scheduling in hospitals	1995-2017	2	56	Multi-appointment, integrated, holistic AND healthcare, AND scheduling
[226]	Literature Review	Computer-interpretable guidelines.	2001-2013	1	21	Electronic clinical guidelines, clinical pathway, clinical pathways, care pathway
[225]	Systematic Review	Good practice guidelines and contemporary developments.	1990-2014	7	33	NA
[243]	Literature Review	Process mining in healthcare	until 2016	3	74	Process mining, workflow mining, healthcare
[262]	Systematic Review	Simulation of changes in the delivery of surgical care.	1957-2007	8	34	Clinical path, patient flow, Markov, system dynamics, discrete event, agent based, statechart
[264]	Literature Review	Validated simulation models on hospital-wide surgical services.	2000-2016	4	22	Simulation, AND clinical pathway, care pathway, critical pathway, patient pathway, care map
[265]	Systematic Review	Comparing Markov modelling and discrete event simulation for cost-effectiveness analysis of healthcare technologies	1947-2012	3	22	Discrete event simulation, Markov, microsimulation, Monte-Carlo, economic
[272]	Literature Review	Operations Research applied to Hospital Administration Systems	2005-2014	6	152	Hospital, Admission, Systems
[283]	Literature Review	Evidence based model pathway for surgical patients with colorectal cancer	2006-2014	3	15	Clinical pathway, AND colorectal, cancer, enhanced recovery program
[294]	Systematic Review	Application of the Anderson's Model of Total Patient Delay to assess cancer diagnosis	1979-2009	4	10	NA
[331]	Systematic Review	Discrete event simulation applied to health-related topics and decision making.	until 2017	2	211	Discrete event simulation, AND health service, Patient, healthcare
[337]	Systematic Review	Mathematical modelling the cost-effectiveness of diagnostic strategies for active TB.	2000-2013	1	36	NA

Table B.3: Literature Review Table For: Frequency of Publications in JCR Category.

JCR Category	
MI	[3, 19, 21, 22, 24, 27, 29, 48, 54, 66, 70, 71, 74, 83, 85, 87, 88, 98, 104, 105, 129–131, 133–135, 137, 141, 147, 156, 173, 176, 183, 196, 201, 207, 212, 228, 232, 238, 258, 268, 281, 296, 319, 328, 335]
OR/MS	[10, 11, 20, 28, 33, 34, 38, 43, 44, 67, 81, 82, 95, 107, 111, 128, 161, 162, 187, 215, 230, 239, 245, 284, 325, 336]
HPS	[17, 18, 31, 49, 63, 64, 112, 148–150, 158, 164, 197, 199, 218, 237, 240, 332]
IE	[51, 168]
AN	[84, 185]
No ISSN	[23, 35, 53, 91, 102, 103, 116, 179, 193, 194, 198, 233, 275, 320–322]

Table B.4: Literature Review Table For: Frequency of Papers Applying Collection Method.

Obtained	
Data Driven	[6, 10, 11, 17, 19, 21, 48, 49, 54, 74, 78, 79, 85, 89, 93, 98, 102–104, 106–108, 112, 115, 117, 120, 122, 124, 125, 127–137, 150, 155, 156, 160, 168, 172, 176, 178, 179, 183, 185, 187, 193, 194, 196, 199, 201, 205–207, 212, 218, 228, 232, 234, 241, 258, 260, 267–269, 275, 277, 279, 285, 287, 295, 319–328, 332, 333, 335]
Collaboration	[3, 18, 22, 27, 38, 66, 82, 105, 141, 162, 198, 216, 237]
Both	[20, 30, 52, 64, 87, 147, 149, 181, 189, 204, 221, 240, 259, 280]
Other	[29, 31, 33, 35, 65, 91, 111, 116, 164, 180, 245]

Table B.5: Literature Review Table For: Frequency of Papers in Each Condition Area.

Condition	Focus	Applied
Acute	[24, 36, 51–53, 71, 81, 91, 98, 158, 159, 162, 189, 203–206, 212, 237, 280, 282]	[6, 23, 42, 54, 66, 95, 102, 103, 111, 125, 127–131, 135, 156, 176, 194, 221, 232, 234, 245, 259, 296, 322–324, 326, 327]
Chronic	[1, 17, 19, 29, 34, 64, 65, 67, 70, 74, 79, 89, 105, 122, 147, 149, 150, 163–165, 173, 230, 233, 260, 269, 285, 287, 332]	[21, 35, 63, 78, 82, 85–87, 115, 116, 120, 132–134, 136, 137, 148, 160, 172, 178, 180, 183, 196, 201, 207, 228, 240, 258, 267, 268, 295, 333, 335]
Surgical	[31, 84, 100, 199, 218]	[30, 48, 88, 93, 155, 177, 185, 216, 279]

Table B.6: Literature Review Table For: Frequency of Papers in Each Care Level.

Care Level	
Primary	[5, 19, 24, 29, 34, 36, 38, 64, 67, 158, 159, 162–165, 189, 203, 216, 221, 258, 268, 282, 285, 296]
Secondary	[3, 6, 11, 19–24, 27–31, 33, 38, 42–44, 48, 49, 51–53, 65–67, 70, 71, 81–83, 91, 93, 95, 99, 102–108, 111, 112, 115, 117, 124, 125, 127–131, 133, 135, 137, 141, 156, 158, 159, 161, 162, 176, 179–181, 185, 187, 189, 193, 194, 196–198, 203–205, 207, 215, 216, 218, 221, 230, 232, 234, 237–241, 245, 258–260, 268, 273, 275, 277, 279–282, 284, 285, 287, 319–328, 333, 335, 336]
Tertiary	[1, 18, 19, 21, 24, 29–31, 48, 74, 84, 88, 100, 105, 116, 120, 132, 134, 136, 137, 148, 149, 155, 160, 172, 177, 178, 185, 196, 203, 206, 216, 221, 230, 233, 240, 267–269, 279, 285, 295, 332]
Disease	[5, 10, 17, 21, 35, 36, 54, 63, 67, 78, 79, 85–87, 89, 98, 122, 147–150, 166–168, 173, 180, 183, 193, 199, 201, 212, 228, 260, 296, 335]
Home Care	[10, 24, 38, 51, 166–168, 198, 203]

Table B.7: Literature Review Table For: Frequency of Multiple Care Levels.

Multiple Care Levels	
Two	[5, 10, 30, 31, 36, 48, 51, 105, 137, 148, 149, 158, 159, 162, 166–168, 180, 185, 189, 193, 196, 198, 230, 240, 258, 260, 279, 282, 296, 335]
Three	[19, 21, 29, 38, 67, 216, 221, 268, 285]
Four	[24, 203]

Table B.8: Literature Review Table For: Frequency of Papers by Scope.

Scope	
Hospital	[11, 31, 34, 38, 42–44, 53, 65, 82, 95, 99, 106–108, 111, 120, 156, 159, 162–165, 187, 198, 203–206, 216, 230, 232, 239, 240, 258, 267, 268, 277, 284, 324, 325, 332]
Department	[20, 22, 23, 27, 28, 71, 81, 83, 88, 93, 100, 141, 161, 177, 181, 197, 215, 218, 221, 233, 237, 238, 241, 245, 259, 273, 279, 281, 336]
Clinical	[1, 3, 6, 18, 19, 21, 24, 29, 30, 33, 48, 49, 51, 52, 64, 66, 67, 70, 74, 84, 91, 102–105, 112, 115–117, 124, 125, 127–137, 148, 155, 158, 160, 172, 176, 178–180, 185, 189, 194, 196, 207, 234, 269, 275, 280, 282, 285, 287, 295, 319–323, 326–328, 333, 335]
Disease	[5, 10, 17, 21, 35, 36, 54, 63, 67, 78, 79, 85–87, 89, 98, 122, 147–150, 166–168, 173, 180, 183, 193, 199, 201, 212, 228, 260, 296, 335]

Table B.9: Literature Review Table For: Frequency of Papers Applying Method Type.

Method	
Data Mining or Machine Learning	[6, 10, 11, 19, 30, 35, 42, 49, 54, 66, 74, 78, 79, 85, 93, 98, 102–104, 106, 108, 115–117, 120, 122, 124, 125, 127–132, 134–137, 150, 155, 156, 160, 172, 176, 178–180, 185, 189, 193, 194, 199, 201, 207, 212, 221, 228, 234, 238, 241, 258, 260, 267–269, 275, 277, 279, 287, 295, 319–325, 327, 328, 332, 333, 335]
Simulation	[1, 3, 10, 17, 18, 20, 23, 24, 27–29, 31, 34, 36, 38, 48, 51–53, 63, 65, 67, 70, 71, 81–84, 91, 95, 99, 100, 105, 111, 141, 148, 156, 158, 159, 161–165, 173, 181, 187, 198, 203–206, 215, 216, 218, 230, 232, 237, 239, 240, 245, 259, 273, 280, 282, 284, 285, 296, 326]
Optimisation and Heuristics Stochastic Modelling	[5, 11, 20–22, 28, 33, 43, 44, 85–89, 106–108, 112, 133, 137, 149, 166, 167, 177, 183, 196, 233, 234, 241, 269, 281, 335]
	[10, 19, 64, 147, 168, 196, 197, 215, 284, 325, 335, 336]

Table B.10: Literature Review Table For: Frequency of Papers Applying Multiple Methods.

Multiple Methods	
Two	[11, 19, 20, 28, 85, 106, 108, 137, 156, 196, 215, 234, 241, 269, 284, 325]
Three	[10, 335]

Table B.11: Literature Review Table For: Graph of the Interaction Between Mapping, Modelling and Improving the Pathway.

Investigating Type	
Mapping (Ma)	[1, 5, 6, 10, 33, 42, 49, 54, 66, 74, 78, 85, 87, 88, 102, 103, 115, 117, 124, 125, 128, 129, 132–137, 150, 160, 172, 178, 179, 183, 189, 193, 194, 201, 207, 221, 228, 233, 234, 238, 258, 260, 267–269, 275, 277, 279, 287, 295, 319–325, 328, 332, 333, 335]
Modelling (Mo)	[23, 29–31, 34, 35, 43, 44, 53, 63, 67, 81, 86, 89, 95, 99, 108, 111, 116, 122, 127, 148, 149, 163, 165–167, 173, 181, 203–205, 212, 215, 232, 245, 259, 281, 282, 284, 285, 296, 336]
Improving (I)	[21, 71]
Ma & Mo	[3, 17, 19, 27, 28, 64, 91, 93, 98, 104–106, 112, 120, 128, 131, 147, 155, 156, 162, 168, 176, 177, 180, 185, 187, 196, 199, 230, 327]
Mo & I	[11, 24, 36, 48, 51, 65, 70, 83, 84, 100, 107, 158, 159, 161, 164, 197, 206, 239, 273]
All Types	[18, 20, 22, 38, 52, 79, 82, 141, 198, 216, 218, 237, 240, 241, 280, 326]

Table B.12: Literature Review Table For: Frequency of Papers Considering Outcome Measure.

Outcome	
Pathway Mapping	[1, 5, 6, 21, 30, 49, 54, 66, 74, 85–87, 89, 93, 102, 103, 115–117, 122, 124, 125, 127–137, 155, 160, 162, 163, 172, 178–180, 189, 193, 194, 196, 207, 221, 228, 234, 238, 259, 260, 267–269, 275, 277, 281, 282, 287, 295, 319–324, 328, 333, 335]
Time	[3, 11, 18, 20, 23, 24, 28, 34, 42–44, 48, 51–53, 65, 71, 81, 83, 88, 93, 95, 99, 100, 104, 106, 107, 112, 120, 132, 141, 156, 158, 166, 181, 185, 193, 197–199, 203–205, 215, 232, 233, 237, 239–241, 245, 273, 279, 284, 326, 327, 336]
Resource	[3, 10, 18, 22, 23, 27, 36, 38, 42, 44, 48, 52, 53, 71, 83, 95, 107, 108, 141, 161, 166–168, 198, 203, 206, 215, 216, 218, 232, 233, 237, 258, 273, 285, 326]
Cost	[17, 19, 22, 23, 29, 31, 33, 64, 67, 70, 84, 91, 105, 111, 147, 159, 165, 166, 177, 185, 187, 193, 203, 230, 240, 241, 332]
Patient Progression	[35, 42, 63, 64, 70, 78, 79, 82, 91, 111, 122, 147–150, 155, 164, 173, 176, 183, 194, 201, 203, 212, 280, 296]
Legal	[98, 325]

Table B.13: Literature Review Table For: Frequency of Considering Multiple Outcomes.

Multiple Outcomes	
Two	[3, 18, 22, 44, 48, 52, 53, 64, 70, 71, 83, 91, 93, 95, 107, 111, 122, 132, 141, 147, 155, 185, 194, 198, 215, 232, 233, 237, 240, 241, 273, 326]
Three	[23, 42, 166, 193]
Four	[203]

Table B.14: Literature Review Table For: Frequency of Papers Considering Decision Level.

Decision Level	
Strategic	[3, 11, 17, 22, 24, 29, 33, 36, 38, 42, 44, 48, 63, 65, 67, 70, 71, 79, 83, 91, 95, 99, 104, 105, 112, 147–149, 159, 163, 164, 189, 230, 239–241, 258, 273, 284, 285]
Tactical	[10, 18, 20, 22, 43, 51–53, 100, 106, 108, 111, 122, 156, 161, 165, 197, 198, 203, 205, 206, 233, 237]
Operational	[23, 27, 28, 44, 84, 88, 107, 128, 158, 161, 166–168, 187, 199, 204, 212, 218, 232, 245, 280, 336]
No Decision	[1, 5, 6, 19, 21, 30, 31, 34, 35, 49, 54, 64, 66, 74, 78, 81, 82, 85–87, 89, 93, 98, 102, 103, 115–117, 120, 124, 125, 127, 129–137, 141, 150, 155, 160, 162, 172, 173, 176–181, 183, 185, 193, 194, 196, 201, 207, 215, 216, 221, 228, 234, 238, 259, 260, 267–269, 275, 277, 279, 281, 282, 287, 295, 296, 319–328, 332, 333, 335]

Appendix C

Simulation

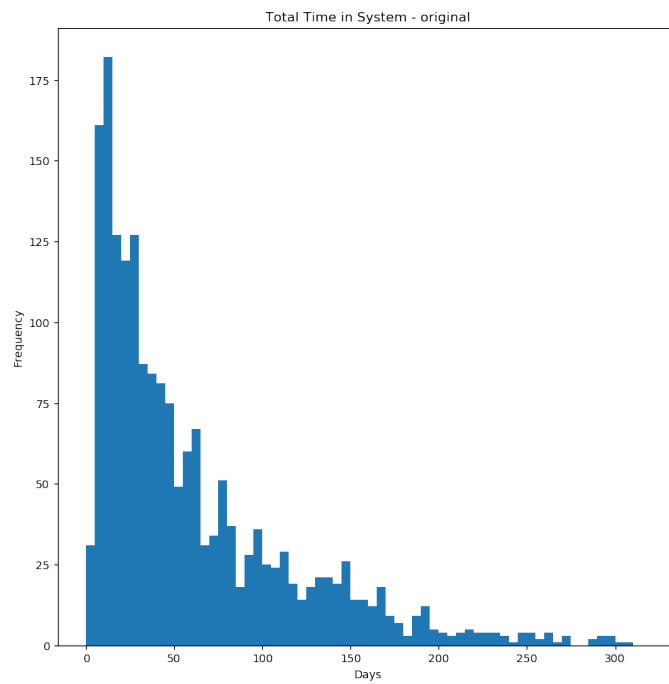


Figure C.1: Histogram of the Original Data Total Time in System.

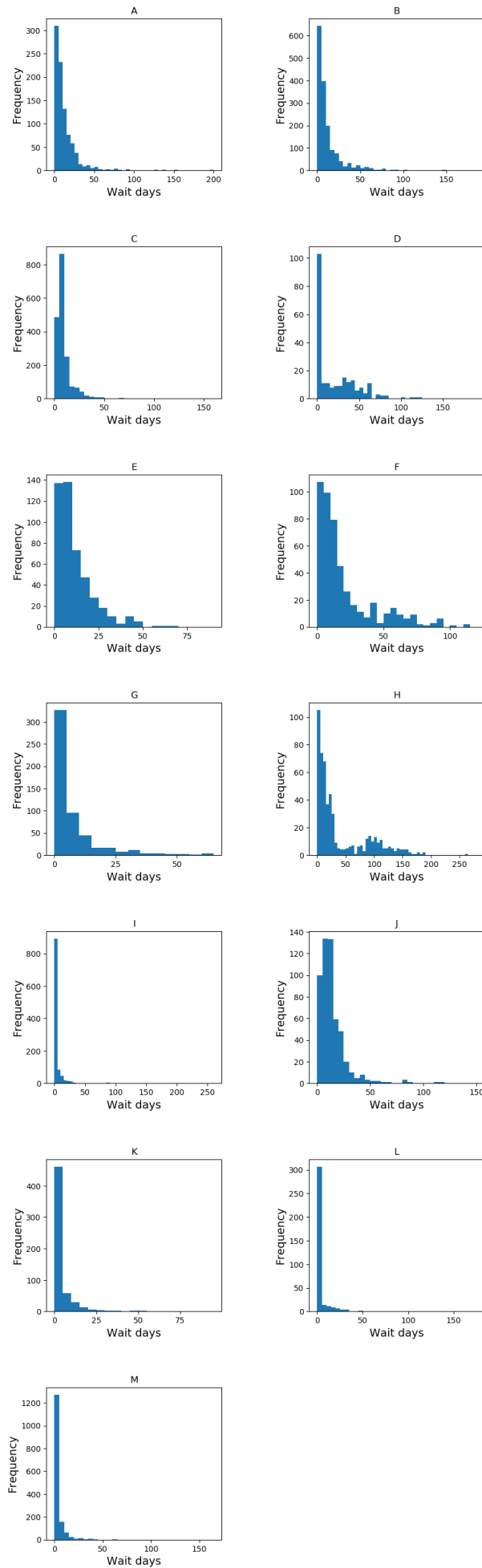


Figure C.2: Activity Waiting Time Histograms for Original Data.

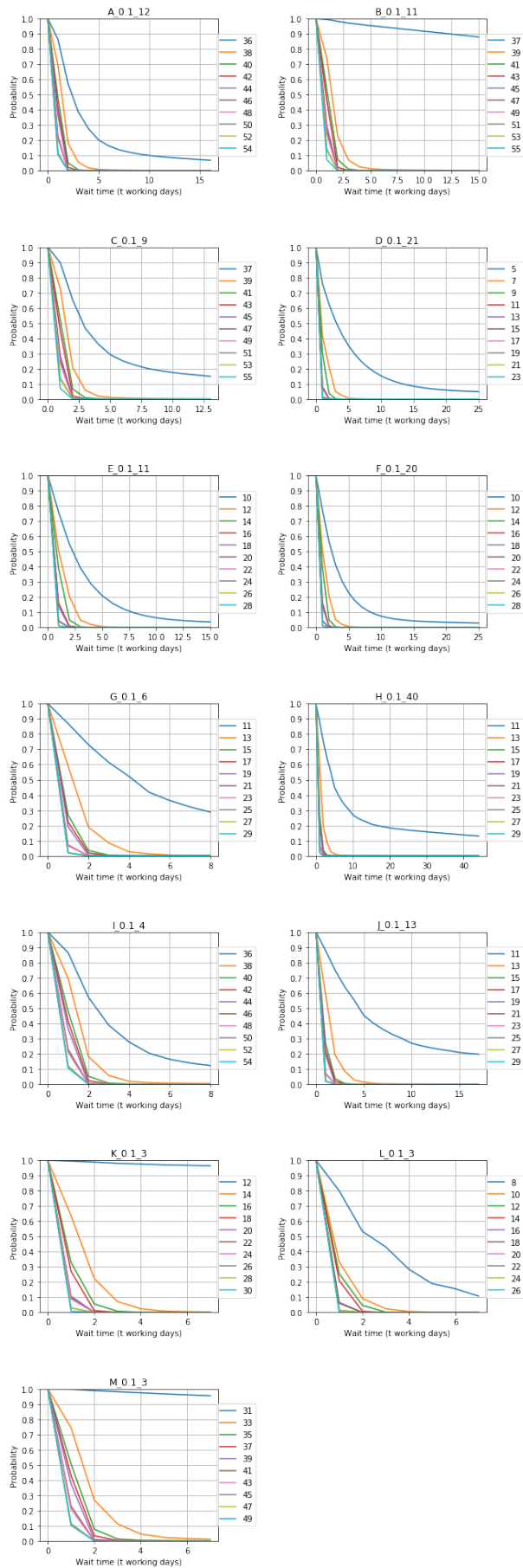


Figure C.3: Line Plot for Capacity Scenario 4.

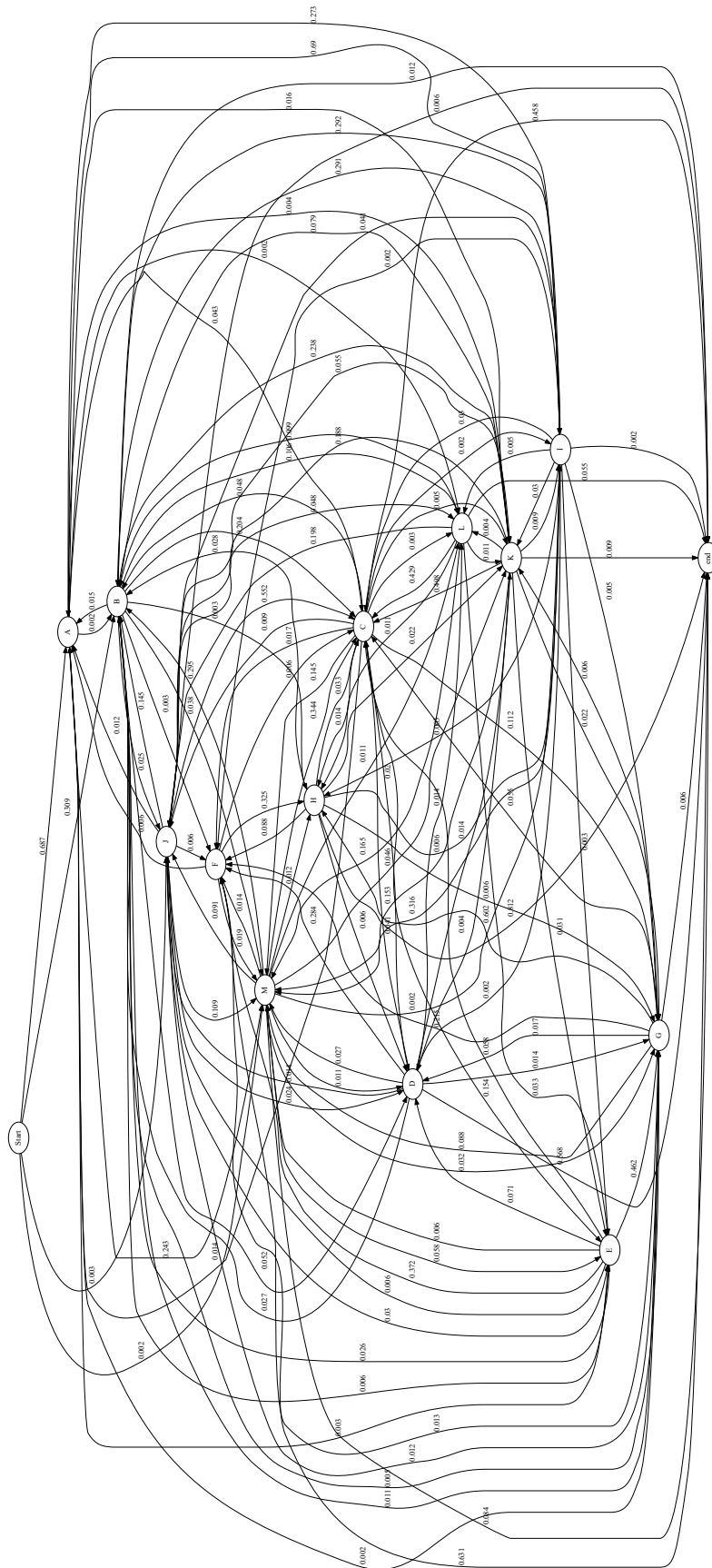


Figure C.4: Network Graph for Cluster Transitions Cluster 1.

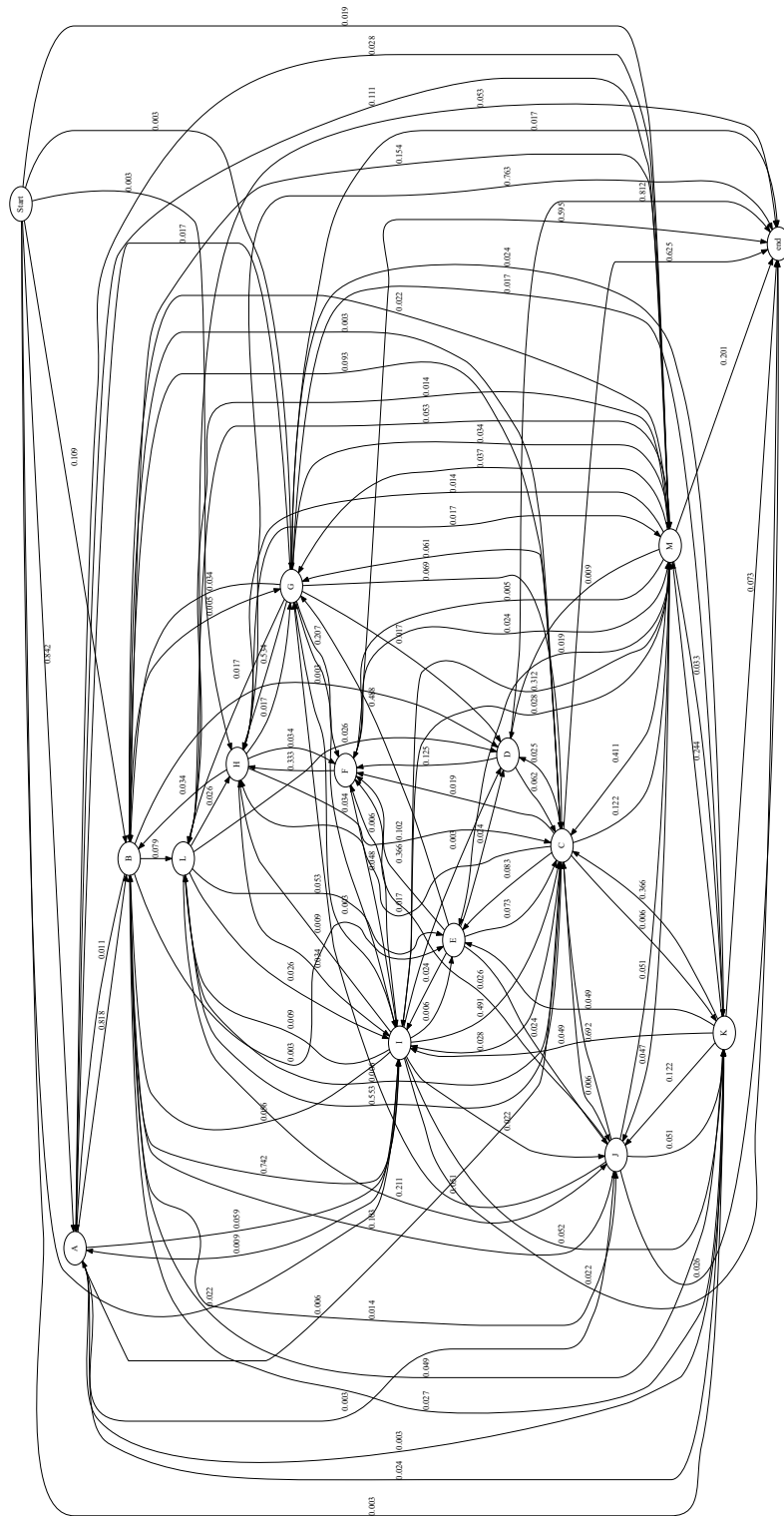


Figure C.5: Network Graph for *Cluster Transitions* Cluster 2.

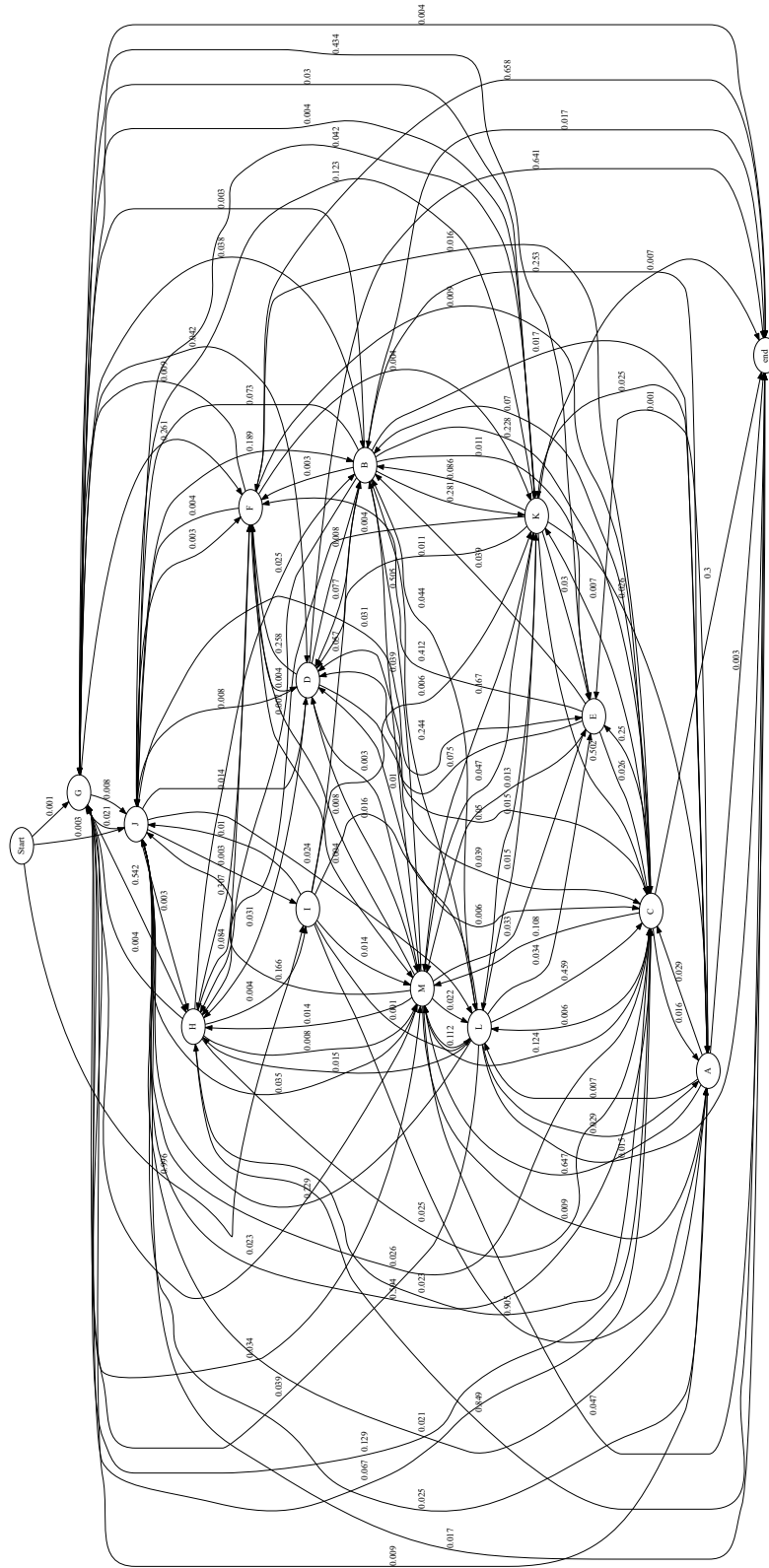


Figure C.6: Network Graph for Cluster Transitions Cluster 4.

Strengthening the Reporting of Empirical Simulation Studies (STRESS)

Discrete-event simulation guidelines STRESS-DES

Section/Subsection	Item	Recommendation	Addressed
1. Objectives			
Purpose of the model	1.1	Explain the background and objectives for the model.	The background, motivation, aim and objective have clearly been defined.
Model Outputs	1.2	Define all quantitative performance measures that are reported, using equations where necessary. Specify how and when they are calculated during the model run along with how any measures of error such as confidence intervals are calculated.	There are three main tables reported: top level results, activity frequency and activity waiting time. The results included in these tables are detailed further. The results are collected at the end of each run of the simulation and then 95% confidence intervals are calculated.
Experimentation Aims	1.3	If the model has been used for experimentation, state the objectives that it was used to investigate. <ul style="list-style-type: none"> a.) Scenario based analysis – Provide a name and description for each scenario, providing a rationale for the choice of scenarios and ensure that item 2.3 (below) is completed. b.) Design of experiments – Provide details of the overall design of the experiments with reference to performance measures and their parameters (provide further details in <i>data</i> below). c.) Simulation Optimisation – (if appropriate) Provide full details of what is to be optimised, the parameters that were included and the 	The four models each consider a different definition of the clinical pathway. The objective is to explore the various definitions of the clinical pathway.

algorithm(s) that was be used. Where possible provide a citation of the algorithm(s).

2. Logic

Base model overview diagram	2.1	Describe the base model using appropriate diagrams and description. This could include one or more process flow, activity cycle or equivalent diagrams sufficient to describe the model to readers. Avoid complicated diagrams in the main text. The goal is to describe the breadth and depth of the model with respect to the system being studied.	Visualisations for each model are provided in the sections preceding the reporting.	
Base model logic	2.2	Give details of the base model logic. Give additional model logic details sufficient to communicate to the reader how the model works.	The nature of the investigation – looking into more granular level of data at each step addresses this.	
Scenario logic	2.3	Give details of the logical difference between the base case model and scenarios (if any). This could be incorporated as text or where differences are substantial could be incorporated in the same manner as 2.2.	The nature of the investigation – looking into more granular level of data at each step addresses this.	
Algorithms	2.4	Provide further detail on any algorithms in the model that (for example) mimic complex or manual processes in the real world (i.e. scheduling of arrivals/appointments/operations/maintenance, operation of a conveyor system, machine breakdowns, etc.). Sufficient detail should be included (or referred to in other published work) for the algorithms to be reproducible. Pseudo-code may be used to describe an algorithm.	Detail is provided on how to assign arrivals and capacity.	
Components	2.5	2.5.1 Entities	Give details of all entities within the simulation including a description of their role in the model and a description of all their attributes.	Entities are referred to as individuals and introduced.
		2.5.2 Activities	Describe the activities that entities engage in within the model. Provide details of entity routing into and out of the activity.	The routing into and out of activities is described in detail through the pathway definitions.

2.5.3 Resources	List all the resources included within the model and which activities make use of them.	Resources are represented as capacity levels for activities i.e. the number of individuals that can be served each day.
2.5.4 Queues	Give details of the assumed queuing discipline used in the model (e.g. First in First Out, Last in First Out, prioritisation, etc.). Where one or more queues have a different discipline from the rest, provide a list of queues, indicating the queuing discipline used for each. If renegeing, balking or jockeying occur, etc., provide details of the rules. Detail any delays or capacity constraints on the queues.	From the queue the individuals are served on a First In First Out (FIFO) basis as default. There are no extra restrictions on the queues.
2.5.5 Entry/Exit Points	Give details of the model boundaries i.e. all arrival and exit points of entities. Detail the arrival mechanism (e.g. 'thinning' to mimic a non-homogenous Poisson process or balking)	The arrival and exit points are described in the prior sections. No arrival mechanism is applied.

3. Data

Data sources	3.1 List and detail all data sources. Sources may include: <ul style="list-style-type: none"> • Interviews with stakeholders, • Samples of routinely collected data, • Prospectively collected samples for the purpose of the simulation study, • Public domain data published in either academic or organisational literature. Provide, where possible, the link and DOI to the data or reference to published literature. <p>All data source descriptions should include details of the sample size, sample date ranges and use within the study.</p>	The data is introduced in the previous chapter, but main features such as sample size and date ranges are recapped.
--------------	---	---

Pre-processing	3.2	Provide details of any data manipulation that has taken place before its use in the simulation, e.g. interpolation to account for missing data or the removal of outliers.	Described in the previous chapter – data manipulation only occurred in the process of reducing the dataset to remove outliers.
Input parameters	3.3	List all input variables in the model. Provide a description of their use and include parameter values. For stochastic inputs provide details of any continuous, discrete or empirical distributions used along with all associated parameters. Give details of all time dependent parameters and correlation. Clearly state: <ul style="list-style-type: none"> • Base case data • Data use in experimentation, where different from the base case. • Where optimisation or design of experiments has been used, state the range of values that parameters can take. Where theoretical distributions are used, state how these were selected and prioritised above other candidate distributions.	The input parameters were discussed in detail throughout the chapter and previous chapter. These are recapped.
Assumptions	3.4	Where data or knowledge of the real system is unavailable what assumptions are included in the model? This might include parameter values, distributions or routing logic within the model.	The model build only includes information gathered from the data, therefore no assumptions are made to accommodate for unavailable data. However, as the model build is automated, assumptions are applied in areas such as capacity pattern.
4. Experimentation			
Initialisation	4.1	Report if the system modelled is terminating or non-terminating. State if a warm-up period has been used, its length and the analysis method used to select it. For terminating systems state the stopping condition.	A warm up period has been used and discussed.

		State what if any initial model conditions have been included, e.g., pre-loaded queues and activities. Report whether initialisation of these variables is deterministic or stochastic.	
Run length	4.2	Detail the run length of the simulation model and time units.	The model runs until 1865 individuals have exited the system. 1 is defined as one day.
Estimation approach	4.3	State the method used to account for the stochasticity: For example, two common methods are multiple replications or batch means. Where multiple replications have been used, state the number of replications and for batch means, indicate the batch length and whether the batch means procedure is standard, spaced or overlapping. For both procedures provide a justification for the methods used and the number of replications/size of batches.	
5. Implementation			
Software or programming language	5.1	State the operating system and version and build number. State the name, version and build number of commercial or open source DES software that the model is implemented in. State the name and version of general-purpose programming languages used (e.g. Python 3.5). Where frameworks and libraries have been used provide all details including version numbers.	Windows 10. The model was not executed in Sim.Pro.Flow, however the code for prototype v2 was used. Python 3.6.3 All versions of the libraries are as defined in the Sim.Pro.Flow requirements file.
Random sampling	5.2	State the algorithm used to generate random samples in the software/programming language used e.g. Mersenne Twister. If common random numbers are used, state how seeds (or random number streams) are distributed among sampling processes.	Ciw's seed function was used, where the seed was representative of the run number, starting at 0. This has been described.

Model execution	5.3	<p>State the event processing mechanism used e.g. three phase, event, activity, process interaction.</p> <p><i>Note that in some commercial software the event processing mechanism may not be published. In these cases authors should adhere to item 5.1 software recommendations.</i></p> <p>State all priority rules included if entities/activities compete for resources.</p> <p>If the model is parallel, distributed and/or use grid or cloud computing, etc., state and preferably reference the technology used. For parallel and distributed simulations the time management algorithms used. If the HLA is used then state the version of the standard, which run-time infrastructure (and version), and any supporting documents (FOMs, etc.)</p>	<p>Ciw uses the three phase approach, which is discussed in the introduction of ciw.</p> <p>The model was not run parallel.</p>
System Specification	5.4	<p>State the model run time and specification of hardware used. This is particularly important for large scale models that require substantial computing power. For parallel, distributed and/or use grid or cloud computing, etc. state the details of all systems used in the implementation (processors, network, etc.)</p>	<p>The time to complete one run was recorded, but the overall time to execute all the models was not.</p>
6. Code Access			
Computer Model Sharing Statement	6.1	<p>Describe how someone could obtain the model described in the paper, the simulation software and any other associated software (or hardware) needed to reproduce the results. Provide, where possible, the link and DOIs to these.</p>	<p>Although Sim.Pro.Flow itself was not used, the model could be recreated in SimPro.Flow with ease due to the automated build feature.</p> <p>The user would require the exact data, and need to implement the warm start.</p> <p>The data will not be available publicly.</p>

Appendix D

Sim.Pro.Flow

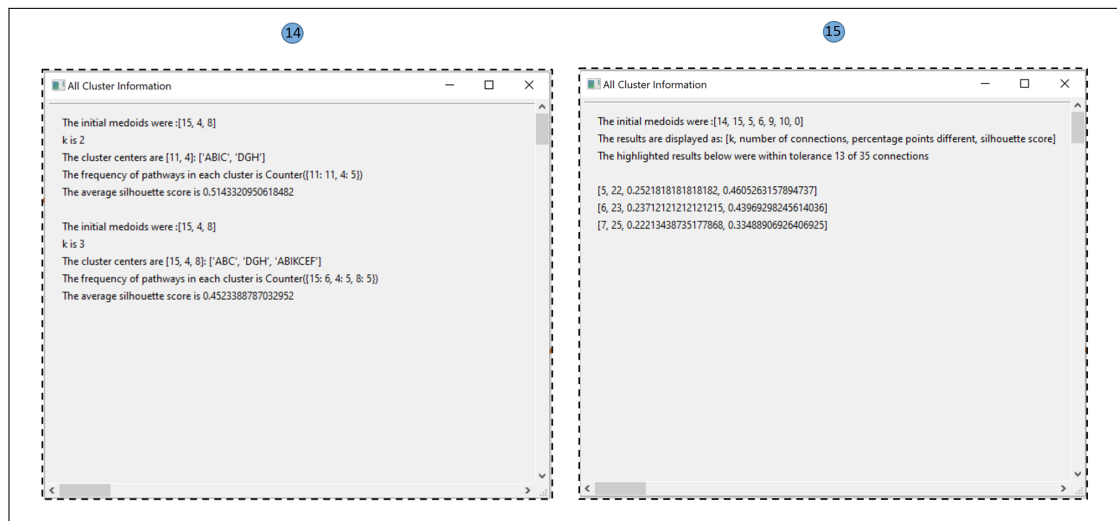


Figure D.1: Sim.Pro.Flow - Clustering Panel Results Pop Out Window.

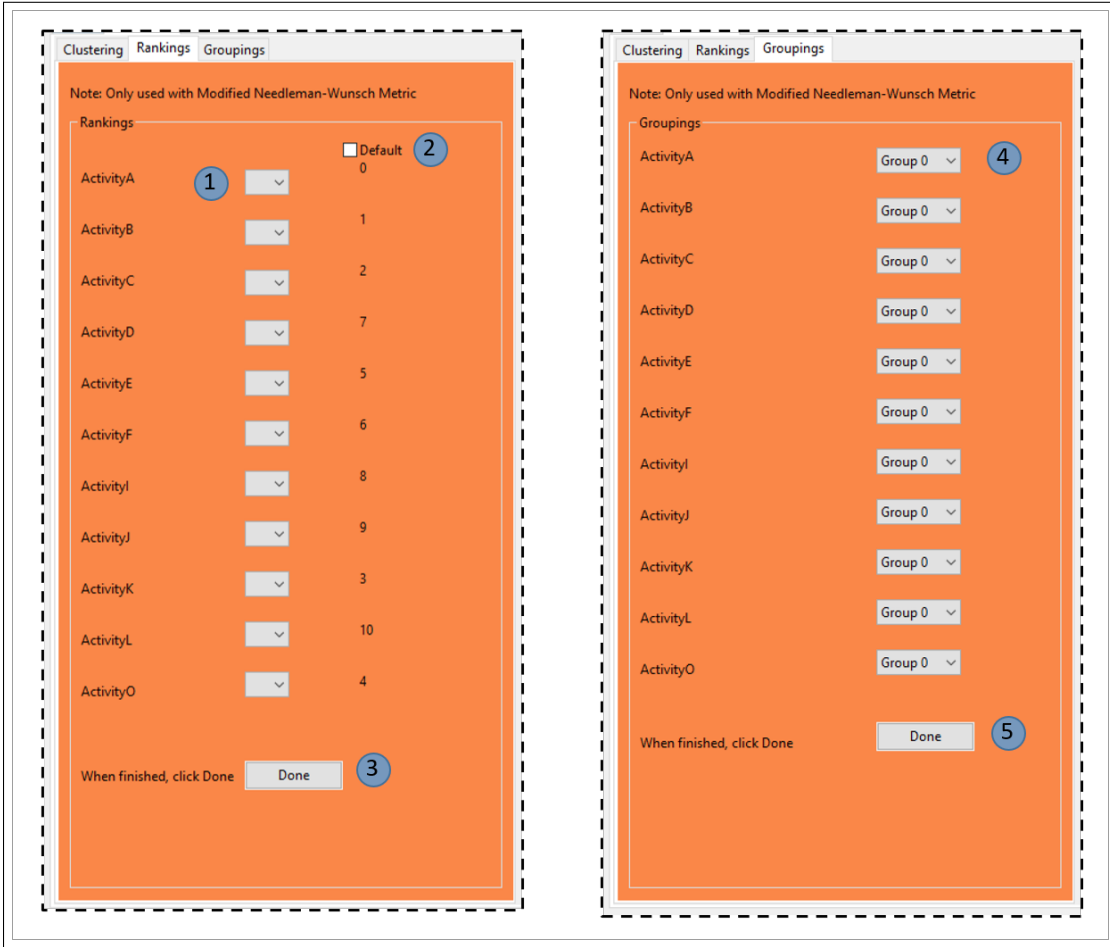


Figure D.2: Sim.Pro.Flow - Clustering Panel Subtabs for Inputting Rankings and Groupings.

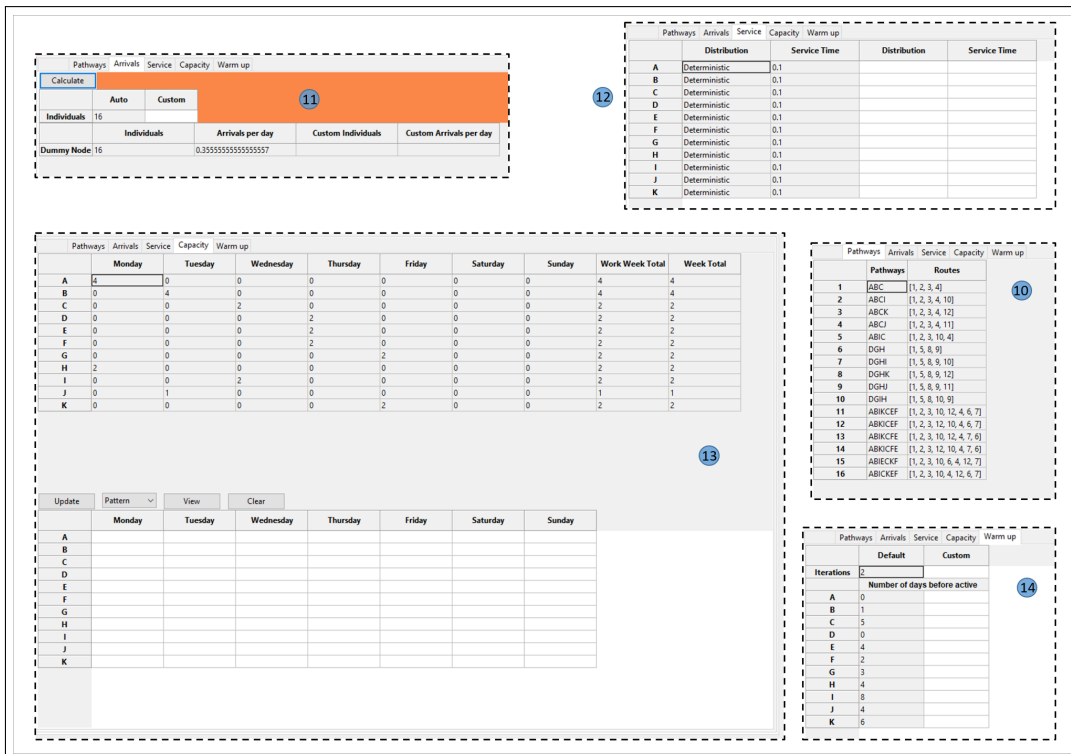


Figure D.3: Sim.Pro.Flow - Simulation Panel Input Parameters.

	Percentage Util 100%	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
A	0.0	34.38	nan	nan	nan	nan		
B	0.0	nan	34.38	nan	nan	nan		
C	50.0	nan	nan	68.75	nan	nan		
D	25.0	nan	nan	nan	31.25	nan		
E	12.5	nan	nan	nan	37.5	nan		
F	25.0	nan	nan	nan	37.5	nan		
G	28.57	nan	nan	nan	nan	35.71		
H	12.5	31.25	nan	nan	nan	nan		
I	37.5	nan	nan	62.5	nan	nan		
J	25.0	nan	25.0	nan	nan	nan		
K	42.86	nan	nan	nan	nan	57.14		

Figure D.4: Sim.Pro.Flow - Simulation Panel Utilisation Pop Out Window.

	Target (%)	Target Time (days)	No. of Increments	Increment Amount	Plot Max (days)	Run
A	90	1	10	2	6	Yes
B	90	1	10	2	6	Yes
C	90	4	10	2	9	Yes
D	90	1	10	2	6	Yes
E	90	3	10	2	8	Yes
F	90	1	10	2	6	Yes
G	90	2	10	2	7	Yes
H	90	3	10	2	8	Yes
I	90	7	10	2	12	Yes
J	90	3	10	2	8	Yes
K	90	5	10	2	10	Yes

Figure D.5: Sim.Pro.Flow - Capacity Panel for Capacity Investigation.

	Name	Mean Time in System	Median Time in System	Target (days, %)	No. Unique Pathways	Occurs Once	Occurs > Once	Total Transitions	Mean Transitions	Largest Transition	Day Last Arrival	Overall Period
1	original	15.31	13.0	[30, 81.25]	16	16	0	0	0.0	0	15	45

	Activity	original
1	A	11
2	B	11
3	C	11
4	D	5
5	E	6
6	F	6
7	G	5
8	H	5
9	I	10
10	J	2
11	K	8

	pathway	original	count	original
1	A	0.0		
2	B	1.0		
3	C	4.36		
4	D	0.0		
5	E	3.83		
6	F	1.33		
7	G	2.4		
8	H	3.4		
9	I	7.8		
10	J	3.5		
11	K	5.12		

Figure D.6: Sim.Pro.Flow - Simulation Panel Containing Simulation Results.

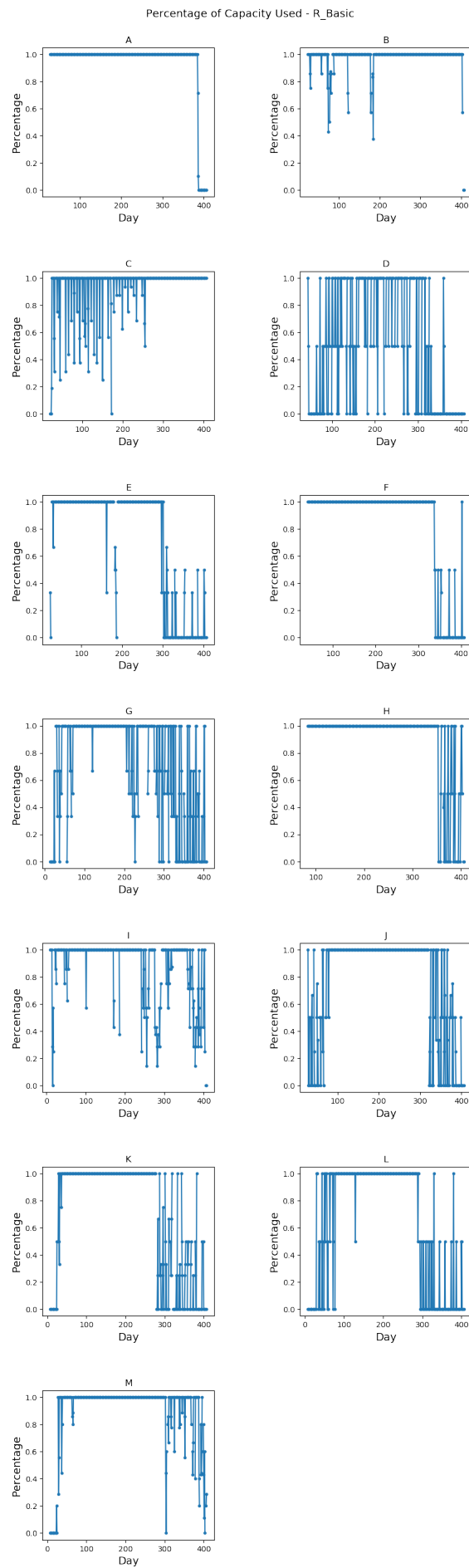
Table D.1: Output Files for Sim.Pro.Flow.

Note: Main location is the selected save location folder. >> indicates path, " indicates fixed name, [] indicates description of name that will change.

Action Button	Name	Description
Data Tab		
Select Save Location	Folder 'Network_diagrams'	Empty folder for Draw to output.
	Folder 'Plots	General folder to contain place specific folders for plots.
	'Plots' >> 'Capacity' Folder	Empty folder to contain plots from the Capacity tab.
	'Plots' >> 'Process_Violin_Plots' Folder	Empty folder to contain the process violin plots from the clustering tab.
	'Plots' >> 'Simulation' Folder	Empty folder to contain all plots from the simulation tab.
	'Plots' >> 'Trials' Folder	Empty folder to contain plots from running trials in the simulation tab.
	'Plots' >> 'Summary' Folder	Empty folder to contain the plots for the summary sheet.
	'Clustering_Transition_Matrix.xlsx'	Empty excel file to add sheets for the clustering transition matrix from the clustering tab.
	'Process_Centroids.xlsx'	Empty excel file to add sheets for the process based clustering centroids from the clustering tab.
	'Raw_Sim_Results.xlsx'	Empty excel file to add sheets for the raw simulation results from the simulation tab.
'Simulation_Difference_Matrix.xlsx'	Empty excel file to add sheets for the simulation difference matrix from the simulation tab.	
'Cluster_Centroids.xlsx'	Empty excel file to add sheets for the classic clustering centroids from the clustering tab.	
Select Columns/ Format	'SimProFlow_[DataName].xlsx'	Additional information added to a copy of the original data file, as shown in data types.
	'Plots' >> 'Simulation' >> 'Activity_Waits_original.png'	Plot of subplots containing histogram of the waiting time for each activity in the original data selection.
	'Plots' >> 'Simulation' >> 'TotalTime_original.png'	Histogram of the total time in system for the original data.
Create Summary Sheet	'Summary_Sheet.docx'	Word document produced containing summary information about the original data. Some of the information included is the number of individuals, number of pathways, mean, median and quartile time in system, as well as the following four plots.
	'Plots' >> 'Summary' >> 'Activity_Frequency.png'	Horizontal bar chart displaying the number of times (frequency) that each activity was performed.
	'Plots' >> 'Summary' >> 'Boxplot_Activity_Wait_Time.png'	Boxplot of the activity waiting times.
	'Plots' >> 'Summary' >> 'Heatmap_All_Pathways.png'	Heatmap displaying a representation of all pathways from the original data.
	'Plots' >> 'Summary' >> 'Histogram_Total_Time_in_System.png'	Histogram of the total time in system. <i>*This may be a different scale to 'TotalTime_original'.</i>

Action Button	Name	Description
Clustering Tab		
Create Matrix	Update 'Clustering_Transition_Matrix.xlsx'	Sheet added with distance metric code as sheet name. If Modified Needleman-Wunsch selected sheet name is MNW-[mgsns] where m, g, s and ns are the penalty values selected. Contains ixi distance matrix for pathways in same order as dataframe.
Save Centroids Checked	Classic Cluster Used: Update 'Cluster_Centroids.xlsx'	On first save create sheet Set_Medoids recording the set number and the initial centroids. This will subsequently be updated. Sheet added called [Metric]_Set_[SetNumber].df containing all the pathways in the dataframe and the index of the corresponding medoids of the cluster it belongs to.
	Process Cluster Used: Update 'Process_Centroids.xlsx'	On first save create sheet Set_Medoids recording the set number and the initial centroids. This will subsequently be updated. Sheet added called [Metric]_Set_[SetNumber] will contain the medoids for k (based on results type) and the number of pathways assigned to each medoids. Sheet added called [Metric]_Set_[SetNumber].df containing all the pathways in the dataframe and the index of the corresponding medoids of the cluster it belongs to.
Process Cluster	'Plots' >> 'Process_Violin_Plots' >> [Metric_Set_SetNumber_Type_k].png	Creates the violin plot for all values in [2, max k] to aid in decision of k values to use.
Simulation Tab		
Run Simulation	Update 'Raw_Sim_Results.xlsx'	Adds a sheet called [SimName] containing the raw simulation results. Each row is an individual with columns for id, waiting time at each activity, pathway, total time in system and customer class. Add a sheet called [SimName]-Util containing the Utilisation Table.
	Update 'Simulation_Difference_Matrix.xlsx'	Initially adds sheet called original. For each simulation adds a sheet called [SimName] containing the transition matrix excluding the start transitions and including the end transitions.
	'Plots' >> 'Simulation' >> 'Activity_Waits_'[SimName].png	Plot containing subplots of histograms for the waiting time for each activity.
	'Plots' >> 'Simulation' >> 'TotalTime_'[SimName].png	Histogram of the total time in system.
	'Plots' >> 'Simulation' >> 'Utilisation_Percent_'[SimName].png	Plot containing subplots of a line chart showing the percentage of capacity used each day for each activity.
	'Plots' >> 'Simulation' >> 'Utilisation_Queue_'[SimName].png	Plot containing subplots of a line chart showing the number of individuals remaining in the queue at the end of each day.

Action Button	Name	Description
<i>*Note in all draw diagrams a light grey line with no label represents a value of 1, whether that be one connection or a probability of 1.*</i>		
Auto Setup - Full Transitions - Draw	'Network_diagrams' >> 'Network_'[SimName].png	Directed graph of the full transitions network. Each node is an activity where an edge between two nodes represents the transition probability. The start and end nodes are included for visual aid. The start node itself is not the arrival node for the simulation.
Auto Setup - Cluster Transitions - Draw	'Network_diagrams' >> 'Network_'[SimName]'.Class'[_No.].png	Directed graph for each class for the cluster transitions. There will be a file created for each class/cluster where [_No.] is the class number. Each node is an activity where an edge between two nodes represents the transition probability. The start and end nodes are included for visual aid. The start node itself is not the arrival node for the simulation.
Auto Setup - Process Based - Draw	'Network_diagrams' >> 'Network_'[SimName_k_].png	Directed graph for the process medoids. Each node is an activity where an edge between two nodes represents the transition probability from within the set of medoids. The start and end nodes are included for visual aid. The start node itself is not the arrival node for the simulation.
	'Network_diagrams' >> 'Network_'[SimName_k_]'.adjust'[_Perc].png	Directed graph for the process medoids. Each node is an activity where an edge between two nodes represents the transition probability for the original transitions after adjusting for the adjust percentage ([_Perc]). The start and end nodes are included for visual aid. The start node itself is not the arrival node for the simulation.
	'Network_diagrams' >> 'Network_'[SimName_k_]'.pathways'.png	Visual representation of the raw medoids.
	'Network_diagrams' >> 'Network_'[SimName_k_]'.linked'.png	Visual representation of the raw medoids where activities are grouped by position. The boxes define the position of the activity (reading left to right). The boxes contain nodes representing all activities that occurred at that position in the set of medoids. The connections between activities in the groups represent the number of times that connection occurred in the set of medoids.
Simulation - Capacity Tab		
Calculate	'Plots' >> 'Capacity' >> 'Cal_Cap_'[Number].png	Plot containing subplots for each activity displaying the percentage seen within x axis days for the various values of weekly capacity investigated.
File Menu		
Export	'Raw_Variables.py'	Python file containing a dictionary for each of the arrivals, service, capacity and service options used for each simulation. The dictionary keys are the [SimName].
Save	'Results_Tables.xlsx'	The four results tables each saved as a sheet. This will be overwritten on each save.

Figure D.7: Capacity Utilisation Percentage Plot for *Raw Pathways* - seed 0.

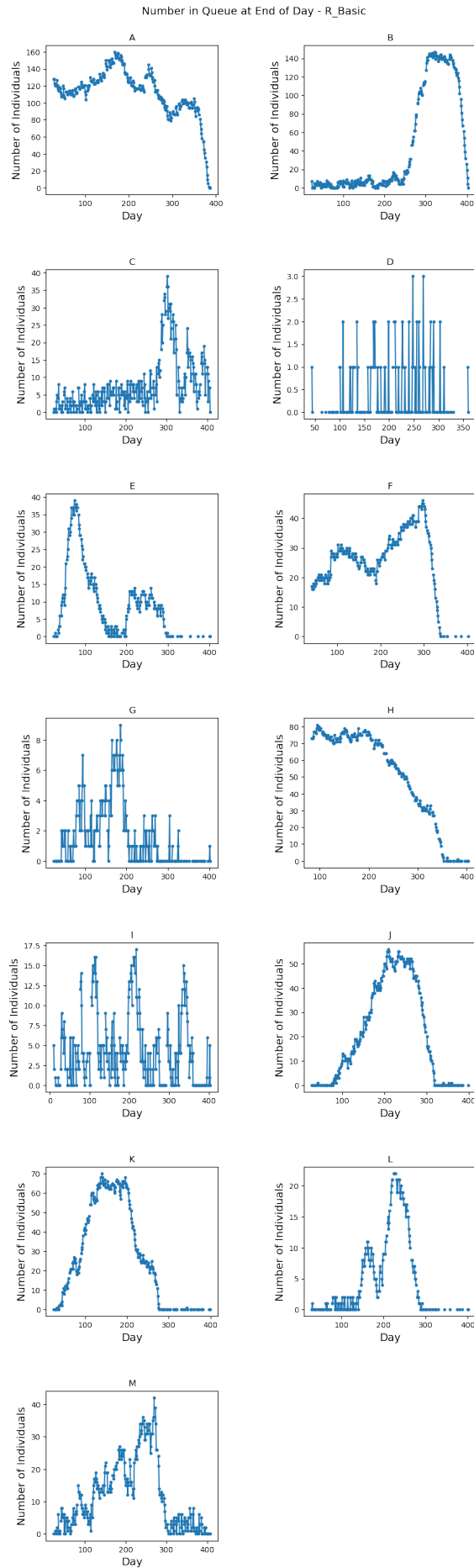


Figure D.8: Capacity Utilisation Queue Plot for *Raw Pathways* - seed 0.

* Layout adjusted and typographic mistakes corrected for thesis.

Summary Sheet

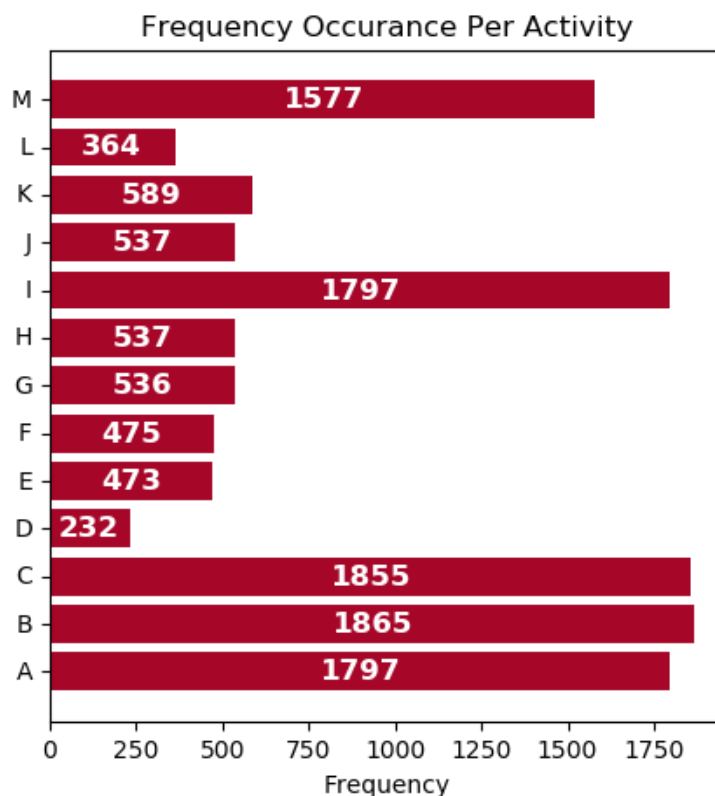
Data Summary

The data includes 1865 patients, 13 activities and 783 different pathways, where 576 pathways only occurred once. For reference, Table 1 shows a key of activities and their codes.

Code	Activity
A	DATE_FIRST_SEEN
B	DATE_OF_DIAGNOSIS
C	MDT_DISCUSSION_DATE
D	PROCEDURE_DATE
E	DECISION_TO_TREAT_DATE_CHEMO
F	CHEMOTHERAPY_START_DATE
G	DECISION_TO_TREAT_DATE_TELE
H	TELETHERAPY_START_DATE
I	DATE_OF_CT_SCAN
J	DATE_OF_PET_OR_PET_CT_SCAN
K	DATE_OF_BRONCHOSCOPY
L	DATE_OF_CT_GUIDED_BIOPSY
M	DATE_SPEC_NURSE_SEEN

Activity Summary

The frequency of occurrence for each activity can be seen in Figure 1.

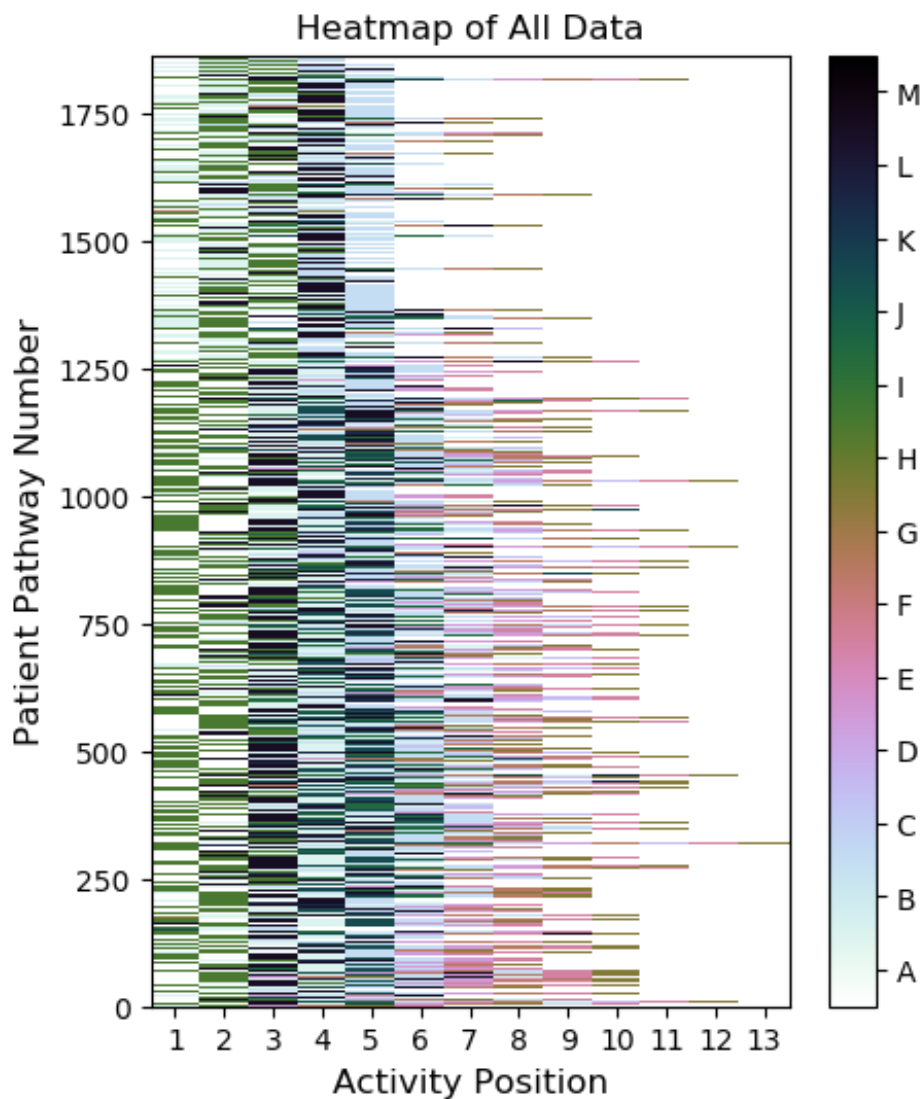


Pathway Summary

The 10 most popular pathways are:

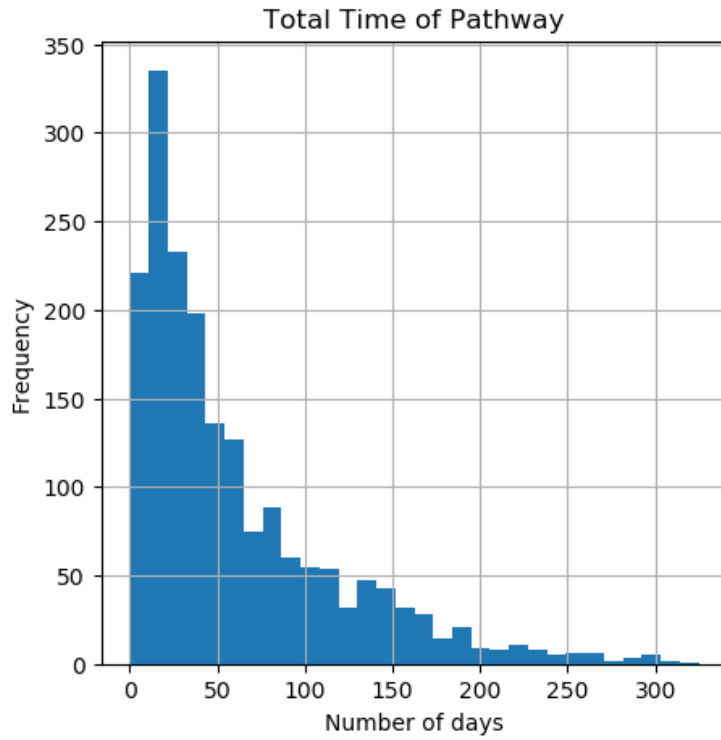
Pathway	Frequency
BIAMC	105
ABIC	74
ABIMC	63
IAMBC	45
ABICM	34
AIBC	33
AIMBC	31
IAMBKC	26
AMIBC	24
IAMBLC	22

The heatmap in Figure 2 is a visual representation of the all the pathways included in the data.

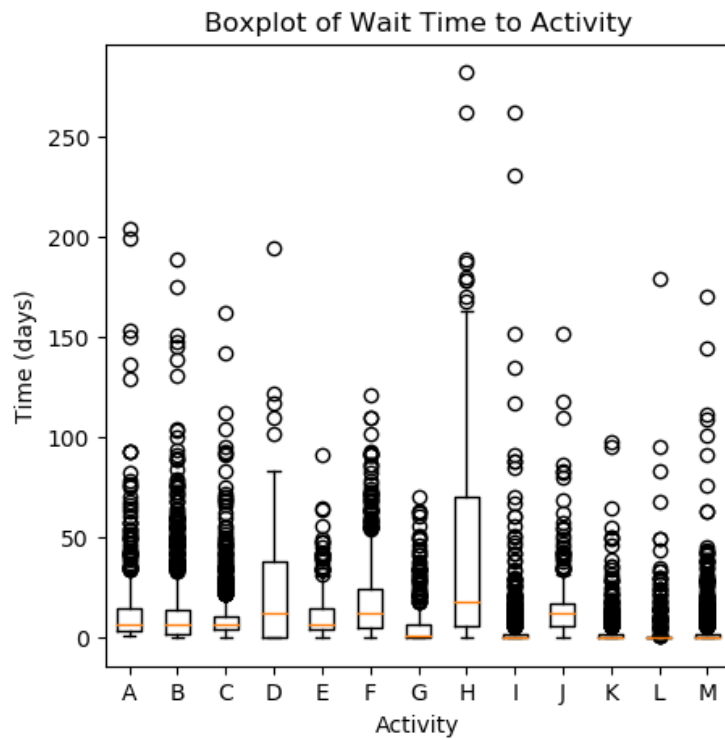


Time Summary

The mean, median, 25 percentile and 75 percentile total time was 60.65, 41.0, 17.0 and 84.0 days respectively. The histogram in Figure 3 displays the overall total times.



The time to each activity from the one that preceded it is displayed in the boxplot in Figure 4



Appendix E

Case Study

Table E.1: Top Level Results for Individual Adjustment Investigation.

Name	Mean TiS	Media TiS	Target	Overall Period
Original Standard	60 65.08, (63.34, 66.82)	41 56.72, (54.4, 59.04)	64.4 53.12, (50.34, 55.9)	362 407.66, (407.46, 407.86)
D - (2, 2, 1, 2, 2)				
D1 (1, 1, 1, 1, 1)	68.88, (66.63, 71.13)	58.72, (55.79, 61.65)	48.48, (44.4, 52.57)	408.09, (407.82, 408.36)
D2 (2, 2, 1, 1, 1)	65.48, (63.48, 67.48)	57.36, (54.42, 60.3)	51.4, (47.54, 55.26)	407.74, (407.51, 407.97)
D3 (2, 1, 1, 1, 1)	66.52, (64.29, 68.75)	57.6, (54.3, 60.9)	50.53, (46.55, 54.51)	407.89, (407.62, 408.15)
G - (2, 3, 2, 3, 3)				
G1 (2, 2, 2, 2, 2)	66.64, (64.91, 68.37)	58.64, (56.06, 61.22)	50.73, (46.99, 54.47)	408.45, (408.27, 408.64)
G2 (3, 3, 2, 2, 2)	65.48, (63.19, 67.77)	57.32, (54.43, 60.21)	50.87, (46.67, 55.08)	407.94, (407.73, 408.16)
G3 (3, 2, 2, 2, 2)	63.2, (61.26, 65.14)	54.56, (52.03, 57.09)	55.55, (52.88, 58.23)	407.94, (407.82, 408.06)
I - (7, 7, 7, 7, 8)				
I1 (7, 7, 7, 7, 7)	65.12, (63.29, 66.95)	56.88, (54.4, 59.36)	53.22, (49.96, 56.47)	408.05, (407.84, 408.26)
I2 (7, 7, 7, 7, 6)	66.16, (63.81, 68.51)	58.6, (55.3, 61.9)	49.55, (45.19, 53.91)	408.88, (408.42, 409.35)
I3 (7, 7, 7, 6, 6)	65.52, (63.12, 67.92)	58.0, (54.8, 61.2)	50.78, (46.63, 54.93)	410.19, (409.39, 411.0)
I4 (6, 7, 7, 7, 7)	65.76, (63.6, 67.92)	57.72, (54.78, 60.66)	50.87, (46.72, 55.02)	408.9, (408.46, 409.33)
K - (2, 4, 4, 3, 2)				
K1 (3, 4, 4, 3, 3)	64.16, (62.23, 66.09)	55.08, (52.27, 57.89)	53.99, (51.17, 56.8)	407.42, (407.18, 407.65)
K2 (4, 4, 4, 4, 3)	63.88, (61.7, 66.06)	55.96, (52.71, 59.21)	52.48, (48.43, 56.52)	407.68, (407.37, 408.0)
K3 (4, 4, 4, 3, 3)	63.44, (61.24, 65.64)	54.72, (51.58, 57.86)	55.09, (52.35, 57.84)	407.61, (407.4, 407.82)
C - (5, 7, 9, 16, 1)				
C1 (5, 7, 9, 9, 1)	77.76, (75.89, 79.63)	82.44, (80.05, 84.83)	24.88, (23.24, 26.51)	451.02, (450.94, 451.09)
C2 (5, 7, 9, 12, 1)	63.0, (61.14, 64.86)	54.44, (51.6, 57.28)	54.28, (50.73, 57.82)	407.65, (407.38, 407.91)
C3 (5, 7, 9, 10, 1)	73.6, (71.52, 75.68)	75.8, (72.87, 78.73)	29.85, (26.87, 32.83)	442.86, (442.71, 443.0)
C4 (5, 7, 9, 11, 1)	71.6, (70.06, 73.14)	71.56, (69.35, 73.77)	32.63, (30.58, 34.68)	434.9, (434.74, 435.05)
A - (6, 7, 10, 6, 6)				
A1 (7, 7, 10, 7, 7)	63.2, (61.7, 64.7)	54.56, (52.44, 56.68)	55.28, (53.74, 56.82)	402.63, (402.13, 403.13)
A2 (7, 8, 10, 7, 7)	65.08, (63.5, 66.66)	57.48, (55.25, 59.71)	52.95, (51.11, 54.78)	402.37, (402.02, 402.73)
A3 (7, 7, 10, 7, 6)	64.0, (61.88, 66.12)	55.2, (52.33, 58.07)	54.1, (51.16, 57.03)	404.3, (403.51, 405.09)
A4 (7, 7, 10, 6, 6)	63.24, (61.35, 65.13)	53.76, (51.22, 56.3)	55.34, (53.17, 57.51)	406.4, (406.15, 406.64)

Table E.2: Activity Waiting Time Results for Individual Adjustment Investigation for D, G, and I.

Activity	Original	D1 (1, 1, 1, 1, 1)	D2 (2, 2, 1, 1, 1)	D3 (2, 1, 1, 1, 1)
A	12.52	23.56, (21.49, 25.62)	23.11, (21.23, 24.99)	23.24, (21.2, 25.27)
B	11.86	9.69, (9.31, 10.07)	9.79, (9.5, 10.07)	9.76, (9.43, 10.09)
C	9.40	2.07, (1.98, 2.15)	1.95, (1.87, 2.04)	2.01, (1.93, 2.08)
D	21.91	35.62, (35.35, 35.9)	4.19, (3.93, 4.44)	14.75, (14.36, 15.13)
E	11.19	7.56, (7.45, 7.68)	7.63, (7.51, 7.74)	7.57, (7.45, 7.7)
F	20.20	15.15, (14.97, 15.33)	17.77, (17.6, 17.95)	16.78, (16.61, 16.95)
G	6.41	1.44, (1.37, 1.51)	1.45, (1.38, 1.52)	1.45, (1.38, 1.53)
H	40.21	33.72, (33.57, 33.87)	33.92, (33.78, 34.06)	33.86, (33.7, 34.01)
I	4.05	1.51, (1.38, 1.65)	1.58, (1.43, 1.73)	1.57, (1.39, 1.75)
J	13.66	13.97, (13.82, 14.13)	13.98, (13.81, 14.15)	13.97, (13.81, 14.13)
K	3.58	16.57, (16.38, 16.76)	16.53, (16.33, 16.72)	16.55, (16.32, 16.77)
L	3.85	5.34, (5.21, 5.47)	5.32, (5.17, 5.47)	5.34, (5.2, 5.49)
M	3.20	2.92, (2.84, 3.0)	3.0, (2.9, 3.09)	2.97, (2.86, 3.08)

Activity	G1 (2, 2, 2, 2, 2)	G2 (3, 3, 2, 2, 2)	G3 (3, 2, 2, 2, 2)
A	24.19, (22.6, 25.79)	23.56, (21.45, 25.67)	21.97, (20.17, 23.77)
B	9.82, (9.54, 10.1)	9.8, (9.5, 10.11)	9.58, (9.36, 9.8)
C	2.29, (2.2, 2.39)	1.97, (1.89, 2.05)	2.09, (2.03, 2.15)
D	0.72, (0.68, 0.77)	0.72, (0.69, 0.76)	0.79, (0.75, 0.84)
E	7.95, (7.75, 8.14)	7.64, (7.5, 7.77)	7.79, (7.62, 7.96)
F	9.7, (9.46, 9.93)	16.9, (16.67, 17.12)	13.96, (13.73, 14.18)
G	27.77, (27.49, 28.04)	4.77, (4.6, 4.93)	13.96, (13.74, 14.19)
H	15.87, (15.73, 16.0)	30.94, (30.75, 31.13)	22.73, (22.52, 22.94)
I	1.67, (1.54, 1.79)	1.69, (1.57, 1.82)	1.47, (1.37, 1.58)
J	14.02, (13.89, 14.15)	14.06, (13.88, 14.25)	14.04, (13.87, 14.21)
K	16.31, (16.06, 16.55)	16.47, (16.25, 16.69)	16.53, (16.32, 16.75)
L	5.3, (5.14, 5.47)	5.36, (5.2, 5.51)	5.32, (5.21, 5.44)
M	2.77, (2.68, 2.87)	2.99, (2.9, 3.08)	2.96, (2.89, 3.02)

Activity	I1 (7, 7, 7, 7, 7)	I2 (7, 7, 7, 7, 6)	I3 (7, 7, 7, 6, 6)	I4 (6, 7, 7, 7, 7)
A	22.36, (20.77, 23.95)	21.6, (19.75, 23.44)	18.79, (16.99, 20.58)	21.29, (19.63, 22.94)
B	9.81, (9.54, 10.08)	10.25, (9.88, 10.62)	10.5, (10.11, 10.89)	10.07, (9.72, 10.43)
C	1.94, (1.87, 2.01)	2.08, (1.98, 2.18)	2.08, (2.0, 2.16)	2.03, (1.93, 2.14)
D	0.69, (0.65, 0.72)	0.7, (0.66, 0.74)	0.76, (0.69, 0.82)	0.69, (0.65, 0.73)
E	7.57, (7.38, 7.75)	7.4, (7.24, 7.57)	7.3, (7.11, 7.48)	7.49, (7.35, 7.64)
F	17.95, (17.7, 18.2)	17.61, (17.37, 17.84)	17.23, (16.99, 17.47)	17.5, (17.26, 17.74)
G	1.37, (1.29, 1.45)	1.52, (1.44, 1.6)	1.6, (1.51, 1.68)	1.57, (1.48, 1.66)
H	33.79, (33.59, 33.98)	33.32, (33.06, 33.57)	32.95, (32.73, 33.16)	33.28, (33.08, 33.48)
I	2.54, (2.26, 2.81)	5.06, (4.43, 5.7)	8.05, (7.39, 8.72)	5.06, (4.28, 5.83)
J	14.01, (13.81, 14.21)	13.94, (13.76, 14.13)	13.66, (13.43, 13.88)	13.82, (13.56, 14.08)
K	16.22, (15.93, 16.52)	15.32, (15.04, 15.6)	14.52, (14.22, 14.82)	15.27, (14.94, 15.59)
L	5.39, (5.26, 5.52)	4.95, (4.71, 5.18)	4.71, (4.54, 4.89)	5.22, (4.95, 5.49)
M	2.92, (2.81, 3.04)	2.33, (2.18, 2.49)	1.78, (1.68, 1.88)	2.46, (2.24, 2.68)

Table E.3: Activity Waiting Time Results for Individual Adjustment Investigation for K, C and A.

Activity	K1 (3, 4, 4, 3, 3)		K2 (4, 4, 4, 4, 3)		K3 (4, 4, 4, 3, 3)			
A	12.52	23.22, (21.46, 24.99)	23.02, (21.01, 25.03)	22.62, (20.64, 24.6)				
B	11.86	9.55, (9.24, 9.85)	9.71, (9.38, 10.03)	9.57, (9.27, 9.86)				
C	9.40	1.71, (1.66, 1.77)	1.73, (1.67, 1.78)	1.71, (1.64, 1.77)				
D	21.91	0.96, (0.89, 1.02)	0.96, (0.91, 1.01)	0.99, (0.92, 1.06)				
E	11.19	9.13, (8.97, 9.29)	10.94, (10.74, 11.13)	10.32, (10.14, 10.5)				
F	20.20	19.68, (19.46, 19.91)	19.38, (19.18, 19.59)	19.5, (19.28, 19.71)				
G	6.41	2.26, (2.2, 2.32)	2.82, (2.72, 2.92)	2.67, (2.57, 2.77)				
H	40.21	35.63, (35.49, 35.76)	36.3, (36.17, 36.44)	36.05, (35.87, 36.23)				
I	4.05	1.45, (1.34, 1.55)	1.61, (1.49, 1.72)	1.53, (1.39, 1.66)				
J	13.66	16.02, (15.8, 16.25)	15.63, (15.44, 15.82)	16.15, (15.92, 16.38)				
K	3.58	5.89, (5.74, 6.03)	1.41, (1.34, 1.47)	2.78, (2.66, 2.89)				
L	3.85	5.7, (5.58, 5.83)	6.51, (6.33, 6.7)	6.28, (6.08, 6.49)				
M	3.20	3.85, (3.74, 3.96)	4.06, (3.97, 4.15)	3.97, (3.88, 4.06)				
Activity	C1 (5, 7, 9, 9, 1)		C2 (5, 7, 9, 12, 1)		C3 (5, 7, 9, 10, 1)		C4 (5, 7, 9, 11, 1)	
A	23.99, (22.14, 25.83)	22.39, (20.66, 24.13)	22.78, (20.79, 24.77)	23.09, (21.68, 24.51)				
B	9.81, (9.53, 10.08)	9.57, (9.29, 9.86)	9.53, (9.21, 9.84)	9.66, (9.41, 9.91)				
C	16.38, (16.18, 16.58)	1.59, (1.55, 1.63)	12.92, (12.73, 13.1)	9.79, (9.66, 9.92)				
D	0.61, (0.57, 0.65)	1.11, (1.04, 1.17)	0.63, (0.59, 0.67)	0.65, (0.61, 0.68)				
E	6.03, (5.94, 6.12)	11.56, (11.3, 11.81)	6.26, (6.15, 6.37)	6.59, (6.5, 6.68)				
F	15.67, (15.48, 15.87)	19.26, (19.02, 19.51)	16.78, (16.58, 16.97)	17.32, (17.11, 17.53)				
G	1.13, (1.08, 1.18)	3.03, (2.93, 3.13)	1.2, (1.12, 1.27)	1.29, (1.21, 1.36)				
H	30.56, (30.43, 30.69)	36.55, (36.42, 36.67)	31.42, (31.29, 31.56)	32.19, (32.01, 32.37)				
I	1.57, (1.46, 1.69)	1.67, (1.56, 1.79)	1.51, (1.39, 1.63)	1.53, (1.38, 1.68)				
J	14.0, (13.82, 14.17)	15.24, (15.05, 15.43)	13.98, (13.8, 14.17)	13.95, (13.76, 14.15)				
K	16.61, (16.37, 16.86)	0.44, (0.42, 0.46)	16.63, (16.45, 16.82)	16.55, (16.36, 16.74)				
L	5.27, (5.12, 5.41)	6.87, (6.63, 7.1)	5.31, (5.18, 5.43)	5.22, (5.09, 5.36)				
M	2.47, (2.4, 2.53)	4.18, (4.07, 4.29)	2.63, (2.58, 2.68)	2.85, (2.76, 2.94)				
Activity	A1 (7, 7, 10, 7, 7)		A2 (7, 8, 10, 7, 7)		A3 (7, 7, 10, 7, 6)		A4 (7, 7, 10, 6, 6)	
A	9.83, (9.06, 10.6)	8.26, (7.5, 9.03)	13.38, (11.82, 14.94)	16.9, (15.25, 18.54)				
B	11.37, (10.76, 11.97)	12.04, (11.43, 12.66)	11.23, (10.71, 11.75)	10.5, (10.13, 10.87)				
C	3.15, (3.07, 3.22)	3.35, (3.26, 3.45)	3.06, (2.98, 3.14)	2.71, (2.63, 2.8)				
D	0.83, (0.79, 0.87)	0.77, (0.72, 0.82)	0.78, (0.73, 0.83)	0.76, (0.7, 0.81)				
E	7.03, (6.86, 7.2)	6.84, (6.68, 7.01)	7.13, (7.02, 7.25)	7.41, (7.25, 7.57)				
F	17.4, (17.2, 17.61)	17.18, (16.96, 17.39)	17.83, (17.61, 18.04)	18.15, (17.98, 18.31)				
G	1.47, (1.37, 1.57)	1.32, (1.24, 1.4)	1.43, (1.35, 1.51)	1.46, (1.39, 1.54)				
H	33.14, (32.93, 33.36)	32.9, (32.7, 33.11)	33.51, (33.31, 33.71)	33.87, (33.75, 33.98)				
I	2.45, (2.11, 2.79)	3.51, (3.0, 4.02)	2.18, (1.89, 2.47)	1.72, (1.52, 1.92)				
J	15.55, (15.37, 15.72)	15.63, (15.39, 15.88)	14.91, (14.66, 15.16)	14.59, (14.41, 14.77)				
K	19.79, (19.62, 19.97)	20.32, (20.13, 20.51)	18.95, (18.71, 19.2)	17.92, (17.77, 18.08)				
L	6.68, (6.55, 6.81)	6.9, (6.77, 7.03)	6.3, (6.11, 6.48)	5.87, (5.74, 6.0)				
M	10.18, (9.86, 10.5)	12.03, (11.74, 12.32)	7.87, (7.68, 8.05)	5.18, (5.09, 5.28)				

Table E.4: Activity Frequency Results for Medoids Capacity Investigation.

	Medoids	Smoothed	Adjusted	Additional
A	1802.08, (1799.18, 1804.98)	1807.16, (1804.33, 1809.99)	1803.68, (1801.08, 1806.28)	1805.32, (1802.24, 1808.4)
B	1865.0, (1865.0, 1865.0)	1865.0, (1865.0, 1865.0)	1865.0, (1865.0, 1865.0)	1865.0, (1865.0, 1865.0)
C	1865.0, (1865.0, 1865.0)	1865.0, (1865.0, 1865.0)	1865.0, (1865.0, 1865.0)	1865.0, (1865.0, 1865.0)
D	64.56, (61.5, 67.62)	62.12, (59.24, 65.0)	61.8, (58.43, 65.17)	61.56, (59.24, 63.88)
E	194.16, (189.32, 199.0)	200.72, (196.93, 204.51)	197.28, (192.66, 201.9)	194.52, (189.98, 199.06)
F	194.16, (189.32, 199.0)	200.72, (196.93, 204.51)	197.28, (192.66, 201.9)	194.52, (189.98, 199.06)
G	222.8, (217.55, 228.05)	219.84, (214.67, 225.01)	217.16, (211.32, 223.0)	221.0, (215.63, 226.37)
H	222.8, (217.55, 228.05)	219.84, (214.67, 225.01)	217.16, (211.32, 223.0)	221.0, (215.63, 226.37)
I	1802.08, (1799.18, 1804.98)	1807.16, (1804.33, 1809.99)	1803.68, (1801.08, 1806.28)	1805.32, (1802.24, 1808.4)
J	64.56, (61.5, 67.62)	62.12, (59.24, 65.0)	61.8, (58.43, 65.17)	61.56, (59.24, 63.88)
K	169.76, (165.11, 174.41)	173.68, (169.05, 178.31)	174.88, (170.15, 179.61)	175.36, (171.15, 179.57)
L	72.6, (69.46, 75.74)	71.16, (68.01, 74.31)	73.64, (69.3, 77.98)	71.96, (69.16, 74.76)
M	1496.4, (1488.59, 1504.21)	1500.56, (1495.22, 1505.9)	1495.32, (1489.56, 1501.08)	1496.4, (1490.99, 1501.81)