

The Application of Process Mining to Care Pathway Analysis in the NHS

Bushra Siddiqi, BSc, MSc

School of Public Health

Department of Primary Care and Public Health

Imperial College London

Doctor of Philosophy (PhD) 2017

Supervisors: Prof. Paul Aylin, Mr. Erik Mayer, Mr. Justin Vale

DEDICATION

To my dear parents...the reason I am whom I am, and the reason I am where I am. I owe them my education and life.

To my husband Najeeb, whose unconditional love and sacrifice knows no bounds, whose dream and aspirations I have fulfilled with this degree.

And last but not the least, to my lovely children Ahmad, Aisha, Abudi and Basil who patiently persevered while their mother sought this degree.

I am forever indebted to all of you.

STATEMENT OF ORIGINALITY

The research contained in this thesis is my own and has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due references are made.

COPYRIGHT DECLARATION

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives license. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work

ACKNOWLEDGEMENTS

I could not have completed this thesis and PhD without the exceptional guidance and mentoring of my supervisors Prof. Paul Aylin and Mr. Erik Mayer. I would like to express my deepest gratitude to them for their unparalleled supervision, patience, valuable critique and constant encouragement throughout my 4 years.

I would also like to thank Mr. Farzan Ramzan for being a constant source of guidance, encouragement and help throughout my journey. He has always gone above and beyond the call of duty to help me in every aspect of this PhD and has been a true brother I could always depend on.

Special thanks to the brilliant BIU team at St. Mary's: Garth Jones, Jashvin Kerai, Abdulrahim Mulla, Usman Maqbool, Deepak Rawat, Samir Mashdanny and Dhana Mani. They have been extremely generous in sharing their knowledge, resources and precious time with me. The bond we have developed in those 3 amazing years will last a lifetime and I am blessed to call them my extended family. I would like to offer a very special word of appreciation to Abdulrahim Mulla for going out of his way to help and support me at the toughest of times. His effort and sincerity has been truly valued.

I would also like to thank Ismail Omar and Elsie Mensah for helping me understand the ins and outs of Prostate Cancer referrals and treatments as well facilitating the case note audit. Their unconditional and tremendous help at all times has overwhelmed me.

Gratitude also goes out to Phawan Hurhangee for his precious consultation, advice and feedback regarding the Prostate Cancer Pathway.

This acknowledgement would be incomplete without extending my deep appreciation to the entire department of Primary Care and Public Health at Imperial College London for their warmth and love in welcoming me and making me feel right at home. Special thanks are extended to all members of the doctoral cohort who have been a pleasure to be around with. Their endless encouragement and appreciation meant a lot to me.

Last but not the least, I would like to whole heartedly thank all my family and friends in Riyadh, Bahrain and London (The Patels!) for their unparalleled support, encouragement and non-stop prayers that have made me reach the place I am at today. You guys are my backbone and I could not have made it without each and every one of you!

Table of Contents

CHAPTER 1: INTRODUCTION	20
1.1 MOTIVATION AND INSPIRATION	21
1.2 IMPORTANCE OF THE RESEARCH AND CONTRIBUTION TO KNOWLEDGE	22
1.3 RESEARCH QUESTIONS, AIMS AND OBJECTIVES.....	23
1.4 THESIS STRUCTURE.....	26
1.5 SUMMARY	28
CHAPTER 2: BACKGROUND	29
2.1 INTRODUCTION	30
2.2 PROCESS MINING	33
2.2.1 <i>History of Process Mining</i>	34
The Sixties.....	36
The Seventies and Eighties	36
The Nineties and Onwards	36
Current State of Process Mining.....	38
Domains That Use Process Mining	39
2.2.2 <i>Process Mining Requirements</i>	39
2.2.3 <i>Process Mining Challenges and Criteria</i>	41
Challenges When Extracting Event Logs.....	41
Quality Criteria of the Discovered Model.....	43
2.2.4 <i>Process Mining Techniques</i>	44
Process Discovery	44
Conformance Checking.....	46
Enhancement.....	46
2.2.5 <i>Process Mining Perspectives</i>	47
The Control Flow Perspective	47
The Organisational Perspective	47
The Case Perspective	48
The Time/Performance Perspective	48
2.2.6 <i>Process Mining Tool</i>	49
2.2.7 <i>Process mining in Healthcare</i>	49
Challenges in Healthcare	50
Healthcare Processes.....	55

2.3	CARE PATHWAYS.....	57
2.3.1	<i>What are Care Pathways?</i>	57
2.3.2	<i>The Need and Requirement of Care Pathways</i>	59
2.3.3	<i>The Strengths and Weaknesses of Care Pathways</i>	60
2.3.4	<i>Format and Design of Care Pathways</i>	63
2.4	PROSTATE CANCER.....	66
2.4.1	<i>Diagnosis and Referral of Prostate Cancer</i>	67
2.4.2	<i>Prostate Cancer Risk Management Programme</i>	69
2.4.3	<i>Economics of Prostate Cancer</i>	69
2.4.4	<i>Quality & Safety Indicators, Minimum Data Set & Auditing Requirements for Prostate Cancer</i> 70 Prostate Cancer Quality Indicators.....	71
	Minimum Dataset (MDS) for Prostate Cancer.....	73
	Auditing Requirements for Prostate Cancer.....	76
2.5	SUMMARY	79
CHAPTER 3: SYSTEMATIC REVIEW OF PROCESS MINING APPLICATIONS IN HEALTHCARE		80
3.1	INTRODUCTION.....	81
3.1.1	<i>Contributions of this chapter</i>	81
3.2	BACKGROUND	82
3.2.1	<i>Mining techniques for the control-flow perspective</i>	82
	Alpha algorithm.....	82
	Heuristics miner	83
	Fuzzy miner	83
	Genetic miner.....	84
	Pattern mining	84
	Declarative miner	84
3.2.2	<i>Mining techniques for the organisational perspective</i>	85
	Social network miner	85
	Organisational miner	85
	Role hierarchy miner	86
3.2.3	<i>Mining with the performance perspective</i>	86
	Performance analyser	86
	Dotted chart analysis	86
3.2.4	<i>Mining with the case perspective</i>	87
	Decision Point Analysis.....	87

3.2.5	<i>The PROM framework and DISCO software</i>	87
3.3	REVIEW QUESTIONS.....	89
3.4	AIMS AND OBJECTIVES	90
3.5	METHODS.....	91
3.5.1	<i>Overview of Methodology</i>	91
3.5.2	<i>Eligibility Criteria</i>	91
3.5.3	<i>Information Sources</i>	92
3.5.4	<i>Searching.....</i>	92
3.5.5	<i>Screening</i>	93
3.5.6	<i>Data extraction.....</i>	93
3.6	RESULTS.....	94
3.6.1	<i>Healthcare processes analysed.....</i>	96
	Anesthesia	97
	Arthritis.....	98
	Asthma.....	98
	Cancer.....	99
	Congestive heart failure	101
	Dentistry	102
	Diabetes.....	102
	Different diseases	102
	Emergency department.....	103
	Eye treatment.....	104
	General medicine and cardiology	105
	ICU	105
	Outpatient processes	105
	Pre-Surgery	106
	Stroke	107
	Treatment processes	107
	Urology	107
3.6.2	<i>Pre-processing techniques used in healthcare.....</i>	108
	Clustering.....	109
	Aggregation	110
	Filtering activities	111
	Adding attributes.....	112
	Deriving fields.....	112

Merge data (log integration)	112
Log-model alignment.....	112
Detecting and correcting typos, outliers, and missing values	113
3.6.3 <i>Process mining techniques used in healthcare</i>	113
3.7 DISCUSSION	121
3.7.1 <i>Healthcare processes analysed.....</i>	121
3.7.2 <i>Preprocessing techniques used.....</i>	122
3.7.3 <i>Mining and visualisation techniques used with respect to the different process mining perspectives</i>	123
3.8 PROMINENT PUBLICATIONS AFTER THE SYSTEMATIC REVIEW.....	124
3.9 STRENGTHS AND LIMITATIONS OF THE SYSTEMATIC REVIEW	125
3.10 CONCLUSIONS	126
3.11 SUMMARY	126
CHAPTER 4: PROCESS MODEL CONSTRUCTION.....	127
4.1 INTRODUCTION	128
4.1.1 <i>Contributions of this chapter.....</i>	129
4.2 BACKGROUND	129
4.2.1 <i>Submission of Secondary Care Data at the National Level.....</i>	130
4.2.2 <i>Flow of Secondary Care Data at the Local Level.....</i>	131
4.2.3 <i>Logistics of the Referral Workflow</i>	132
4.2.4 <i>Construction of Event Logs</i>	133
Goal of process mining	133
Scope of process mining	133
Focus of process mining	134
Extraction of event logs.....	134
4.2.5 <i>Data Cleansing and Preprocessing</i>	136
4.3 DOMAIN UNDERSTANDING	137
4.3.1 <i>Identifying Data Sources.....</i>	137
4.3.2 <i>Selected Data Sources</i>	138
Cerner Database	138
Somerset Database.....	139
Pathology Database.....	139
Radiology Database	139
Radiotherapy Database	139

Surgery Database.....	139
4.4 DATA PREPARATION	140
4.4.1 <i>Data Linkage</i>	140
Step 1: Filter Inpatient and Outpatient Appointment Tables from Cerner DB.....	142
Step 2: Filter Lab DB on PSA and Link	143
Step 3: Filter Lab DB on Biopsy and Link	144
Step 4: Filter Radiology DB and Link.....	145
Step 5: Filter MDT table from Somerset DB and Link.....	145
Step 6: Filter Surgery DB and Link	146
Step 7: Filter Chemotherapy table from Somerset DB and Link.....	146
4.4.2 <i>Output of data linkage</i>	147
4.4.3 <i>Data Extraction and Log Preparation</i>	147
4.4.4 <i>Output of data Extraction</i>	149
4.5 LOG PREPARATION	149
4.6 DISCUSSION	152
4.6.1 <i>Challenges</i>	153
4.6.2 <i>Limitations</i>	154
4.6.3 <i>Tips on improving data quality</i>	154
4.7 SUMMARY	155
CHAPTER 5: DATA VALIDATION	156
5.1 INTRODUCTION	157
5.1.1 <i>Contributions of this chapter</i>	159
5.2 METHODS.....	159
5.2.1 <i>Stage 1: Preparing for the audit</i>	159
5.2.2 <i>Stage 2: Selecting criteria</i>	160
5.2.3 <i>Stage 3: Measuring level of performance</i>	161
Collect data.....	161
Compare performance with criteria.....	166
5.3 RESULTS.....	166
5.3.1 <i>The Audit Data</i>	167
5.3.2 <i>The Audit Data Against DLE</i>	168
Metric 1: Total number of records found.....	168
Metric 2: Matching patients have same first date of appointment	169
Metric 3: Matching patients have a similar PSA value	170

Metric 4: Matching patients have the same appointment priority (TWW).....	170
5.4 DISCUSSION	171
5.5 SUMMARY	172
CHAPTER 6: DESCRIPTIVE RESULTS	173
6.1 INTRODUCTION	174
6.1.1 <i>Contributions of this chapter</i>	174
6.2 LOG INSPECTION RESULTS.....	174
6.2.1 <i>Five (5)-Year Cohort</i>	176
Patient Demographics	177
Patient Referral Phase	181
Patient Diagnostics Phase.....	182
Patient Treatment Phase.....	184
6.2.2 <i>Two (2)-Year Cohort</i>	185
Patient Demographics	188
Patient Referral Phase	191
Patient Diagnostics Phase.....	193
Patient Treatment Phase.....	194
6.3 LIMITATIONS	196
6.4 SUMMARY	196
CHAPTER 7: PROCESS MINING AND VISUALISATION OF THE PATHWAY	197
7.1 INTRODUCTION	198
7.1.1 <i>Contributions of this chapter</i>	199
7.2 BACKGROUND	199
7.2.1 <i>Pathway Analysis Using Process Mining Perspectives</i>	199
7.2.2 <i>The A-Priori Process Model</i>	200
7.2.3 <i>Current Cancer Pathway Mapping and Analysis Techniques</i>	204
7.3 METHODS.....	204
7.3.1 <i>Preprocessing and Initial Log Inspection</i>	204
7.3.2 <i>Clustering</i>	206
7.3.3 <i>Process Discovery and Visualisation</i>	206
7.4 RESULTS.....	209
7.4.1 <i>The LCA Guideline Pathway</i>	209
7.4.2 <i>The PM Cluster: 2-year Cohort (2013-2015)</i>	213

Control Flow Mining Results.....	213
TWO WEEK WAIT REFERRALS.....	215
URGENT REFERRALS	219
ROUTINE REFERRALS.....	222
Performance Mining Results	226
TRACE CLUSTER ANALYSIS.....	226
BOTTLENECK ANALYSIS	229
LCA GUIDELINE COMPLIANCE	239
7.5 DISCUSSION	244
7.6 SUMMARY	249
CHAPTER 8: EVALUATION OF PROCESS MINING VISUALISATIONS.....	250
8.1 INTRODUCTION	251
8.1.1 <i>Contributions of this chapter</i>	251
8.2 BACKGROUND	252
8.2.1 <i>Types of Evaluation Methods</i>	252
8.2.2 <i>How to do the Evaluation</i>	253
8.2.3 <i>Evaluating User Experience (UX)</i>	254
User Experience (UX) Methods	254
Informal Evaluation	254
Usability Test	255
Field Observation	255
Microsoft's Reaction Cards	255
8.3 METHODS.....	256
8.4 RESULTS.....	259
8.4.1 <i>Based on Words Chosen</i>	259
8.4.2 <i>Based on Professions</i>	262
8.4.3 <i>Based on Dimensions</i>	266
8.5 DISCUSSION	266
8.5.1 <i>Lessons Learnt from Evaluation</i>	267
8.5.2 <i>Limitations</i>	267
8.6 SUMMARY	268
CHAPTER 9: DISCUSSION AND CONCLUSION.....	269
9.1 DISCUSSION	270

9.2	SUMMARY OF RESULTS.....	271
9.2.1	<i>Systematic review</i>	273
9.2.2	<i>Process model construction</i>	274
9.2.3	<i>Data validation</i>	274
9.2.4	<i>Process mining</i>	275
9.2.5	<i>Evaluation</i>	276
9.3	STRENGTHS AND WEAKNESSES	277
9.4	ORIGINALITY OF WORK.....	279
9.5	RECOMMENDATIONS AND FUTURE WORK.....	279
	REFERENCES	282
	APPENDICES	293
	APPENDIX A	294
	<i>PRISMA Checklist</i>	294
	APPENDIX B	297
	<i>Data Extraction Spread sheet for Systematic Review</i>	297
	APPENDIX C	306
	<i>Sample ASP and TSQL Code Snippets</i>	306
	APPENDIX D	309
	<i>Prostate Pathway Full Metrics</i>	309
	APPENDIX E	310
	<i>Two-week Wait Cancer Referral Form</i>	310
	APPENDIX F	311
	<i>Microsoft's Product Reaction Cards (concise list)</i>	311

List of Figures and Tables

FIGURE 1: RESEARCH QUESTIONS OF MY STUDY	23
FIGURE 2: CPAM PHASES IN MY THESIS	32
FIGURE 3: CPAM ROAD MAP BY CARON ET AL.....	33
FIGURE 4: TRENDS IN INFORMATION SYSTEMS RELEVANT TO BPM [22]	35
FIGURE 5: THREE WAVES OF PROCESS EVOLUTION [23].....	35
FIGURE 6: BPM LIFECYCLE	37
FIGURE 7: RELATIONSHIP BETWEEN PROCESS MODEL AND EVENT LOG	41

FIGURE 8: BALANCING THE FOUR QUALITY DIMENSIONS [35].....	43
FIGURE 9: PETRI NET OF WORKFLOW LOG	46
FIGURE 10: TOTAL HEALTHCARE EXPENDITURE PER CAPITA UK 1997-2012	51
FIGURE 11: CANCER WAITING TIME STANDARDS.....	51
FIGURE 12: PERCENTAGE STILL WAITING/HAVING WAITED MORE THAN 18 WEEKS (MORE THAN SIX WEEKS FOR DIAGNOSTICS)	54
FIGURE 13: CARE PATHWAY DEVELOPMENT PROCESS (ADAPTED FROM: DE LUC [71] AND DAVIS ET AL [73])	65
FIGURE 14: MULTIDIMENSIONAL QUALITY OF CARE MEASURES BASED ON DONABEDIAN FRAMEWORK AND MAYER ET AL ADAPTATION	70
FIGURE 15: NICE GUIDELINE FOR PATIENTS SUSPECTED WITH PROSTATE CANCER [81]	72
FIGURE 16: PRISMA FLOW DIAGRAM FOR DATA EXTRACTION.....	95
FIGURE 17: STUDIES AND THEIR COUNTRY ORIGINS.....	96
FIGURE 18: HEALTHCARE PROCESSES ANALYSED	97
FIGURE 19: PREPROCESSING TECHNIQUES USED IN HEALTHCARE.....	108
FIGURE 20: PROCESS MINING TECHNIQUES FOUND IN THE REVIEW	113
FIGURE 21: PHASES 1 (EVENT LOG LINKAGE) AND 2 (EVENT LOG PREPROCESSING) OF THE CPAM ROADMAP	128
FIGURE 22: SUBMISSION OF CDS TO SUS DB AND FROM THERE TO HES	131
FIGURE 23: CDE WORKFLOW.....	132
FIGURE 24: OVERVIEW OF STEPS TO LINK THE SIX DATABASES	141
FIGURE 25: STEPS FOR CERNER DATA FILTERATION.....	142
FIGURE 26: STEPS FOR LAB FILTERATION ON PSA AND LINKAGE	143
FIGURE 27: STEPS FOR LAB FILTERATION ON BIOPSY AND LINKAGE	144
FIGURE 28: STEPS FOR RADIOLGY DATA FILTERATION AND LINKAGE	145
FIGURE 29: STEPS FOR MDT DATA FILTERATION AND LINKAGE	145
FIGURE 30: STEPS FOR SURGERY DATA FILTERATION AND LINKAGE.....	146
FIGURE 31: STEPS FOR CHEMOTHERAPY DATA FILTERATION AND LINKAGE	146
FIGURE 32: PHASE 3 (VALIDATION OF DATA EXTRACTION) OF THE CPAM ROADMAP.....	157
FIGURE 33: WELCOME SCREEN OF THE AUDIT TOOL	163
FIGURE 34: MAIN AUDIT FORMS SCREEN	163
FIGURE 35: TWW REFERRAL AUDIT FORM	164
FIGURE 36: CLINICAL LETTER AUDIT FORM	165
FIGURE 37: BAR CHART OF DAYS DIFFERENCE BETWEEN THE AUDIT AND DLE DATA	169
FIGURE 38: 5-YEAR COHORT CANCER FREQUENCY AND INTERVENTIONS FLOW CHART.....	175
FIGURE 39: ACTIVITIES/YEAR IN 5-YEAR COHORT	176
FIGURE 40: PATIENT DISTRIBUTION BASED ON POSTAL CODE IN 5-YEAR COHORT	177
FIGURE 41: AGE GROUP DISTRIBUTION OF PATIENTS IN 5-YEAR COHORT	178

FIGURE 42: MARITAL STATUS OF PATIENTS IN 5-YEAR COHORT	179
FIGURE 43: ETHNICITY IN 5-YEAR COHORT.....	180
FIGURE 44: RELIGION IN 5-YEAR COHORT.....	180
FIGURE 45: APPOINTMENT PRIORITIES IN 5-YEAR COHORT	181
FIGURE 46: APPOINTMENT TYPE IN 5-YEAR COHORT	181
FIGURE 47: REFERRING SOURCE IN 5-YEAR COHORT	182
FIGURE 48: PSAs PERFORMED/YEAR IN 5-YEAR COHORT.....	183
FIGURE 49: BIOPSIES PERFORMED/YEAR IN 5-YEAR COHORT	183
FIGURE 50: RADIOLOGICAL PROCEDURES PERFORMED/YEAR IN 5-YEAR COHORT.....	183
FIGURE 51: RADIOTHERAPIES PERFORMED/YEAR IN 5-YEAR COHORT	184
FIGURE 52: SURGERIES PERFORMED/YEAR IN 5-YEAR COHORT	184
FIGURE 53: CHEMOTHERAPIES/HORMONE THERAPIES PERFORMED/YEAR IN 5-YEAR COHORT	185
FIGURE 54: ACTIVITIES/YEAR IN 2-YEAR COHORT	186
FIGURE 55: 2-YEAR COHORT CANCER FREQUENCY AND INTERVENTIONS FLOW CHART.....	187
FIGURE 56: PATIENT DISTRIBUTION BASED ON POSTA CODE IN 2-YEAR COHORT	188
FIGURE 57: AGE GROUP DISTRIBUTION OF PATIENTS IN 2-YEAR COHORT.....	189
FIGURE 58: MARITAL STATUS OF PATIENTS IN 2-YEAR COHORT	190
FIGURE 59: ETHNICITY IN 2-YEAR COHORT.....	190
FIGURE 60: RELIGION IN 2-YEAR COHORT.....	191
FIGURE 61: APPOINTMENT PRIORITIES IN 2-YEAR COHORT	191
FIGURE 62: APPOINTMENT TYPE IN 2-YEAR COHORT	192
FIGURE 63: REFERRING SOURCE IN 2-YEAR COHORT	192
FIGURE 64: PSAs PERFORMED/YEAR IN 2-YEAR COHORT.....	193
FIGURE 65: BIOPSIES OERFORMED/YEAR IN 2-YEAR COHORT.....	193
FIGURE 66: RADIOLOGICAL PROCEDURES PERFORMED/YEAR IN 2-YEAR COHORT.....	194
FIGURE 67: RADIOTHERAPIES PERFORMED/YEAR IN 2-YEAR COHORT	195
FIGURE 68: SURGERIES PERFORMED/YEAR IN 2-YEAR COHORT	195
FIGURE 69: CHEMOTHERAPIES/HORMONE THERAPIES PERFORMED/YEAR IN 2-YEAR COHORT	196
FIGURE 70: PHASES 4 (EVENT LOG FILTRATION) AND 5 (PATHWAY ANALYSIS) OF THE CPAM ROADMAP	198
FIGURE 71: LOCALISED PROSTATE CANCER LCA + NICE GUIDELINES	203
FIGURE 72: LCA PROSTATE CANCER PATHWAY FLOWCHART WITH HIGHLIGHTED ACTIVITIES	211
FIGURE 73: THE LCA STANDARD PATHWAY PROCESS FLOW DIAGRAM (100% ACTIVITIES AND 100% PATHS) – MADE WITH PROM 6 INDUCTIVE MINER	212
FIGURE 74: PM 2-YEAR COHORT CANCER FREQUENCY AND INTERVENTIONS FLOW CHART.....	214
FIGURE 75: SCREENING OF PATIENTS TO REACH THE SURGICAL/RADIOTHERAPY INTERVENTIONS IN TWW REFERRAL	215

FIGURE 76: 2-YEAR COHORT TWW REFERRAL CLUSTER PROCESS FLOW DIAGRAM (100% ACTIVITIES AND 100% PATHS) SHOWING FREQUENCY OF PATIENTS AND SOJOURN TIMES	218
FIGURE 77: SCREENING OF PATIENTS TO REACH THE SURGICAL/RADIODIOTHERAPY INTERVENTIONS IN URGENT REFERRAL	219
FIGURE 78: 2-YEAR COHORT URGENT REFERRAL CLUSTER PROCESS FLOW DIAGRAM (100% ACTIVITIES AND 100% PATHS) – MADE WITH PROM 6 INDUCTIVE MINER	221
FIGURE 79: SCREENING OF PATIENTS TO REACH THE SURGICAL/RADIODIOTHERAPY INTERVENTIONS IN ROUTINE REFERRAL	222
FIGURE 80: 2-YEAR COHORT ROUTINE REFERRAL CLUSTER PROCESS FLOW DIAGRAM (100% ACTIVITIES AND 100% PATHS) – MADE WITH PROM 6 INDUCTIVE MINER	225
FIGURE 81: 10 CLUSTERS IN THE TWW REFERRAL	227
FIGURE 82: 6 CLUSTERS IN THE ROUTINE REFERRAL	228
FIGURE 83: TRACES ON THE TWW REFERRAL PATHWAY	230
FIGURE 84: BOTTLENECKS IN TRACES 1, 2, 3 OF TWW REFERRAL	232
FIGURE 85: BOTTLENECKS IN TRACE 4 OF TWW REFERRAL	232
FIGURE 86: BOTTLENECKS IN TRACES 5, 6 OF TWW REFERRAL	233
FIGURE 87: BOTTLENECKS IN TRACES 7, 8, 9, 10 OF TWW REFERRAL	233
FIGURE 88: BOTTLENECKS IN URGENT REFERRAL	235
FIGURE 89: TRACES ON THE ROUTINE REFERRAL PATHWAY.....	236
FIGURE 90: BOTTLENECKS IN TRACES 1, 2, 3 OF ROUTINE REFERRAL	238
FIGURE 91: BOTTLENECKS IN TRACE 4 OF ROUTINE REFERRAL	238
FIGURE 92: FIRST TWW APPOINTMENT METRIC COMPLIANCE.....	239
FIGURE 93: 62-DAY FIRST TREATMENT METRIC COMPLIANCE.....	240
FIGURE 94: FIRST TWW APPOINTMENT METRIC COMPLIANCE.....	243
FIGURE 95: 62-DAY FIRST TREATMENT METRIC COMPLIANCE.....	243
FIGURE 96: PHASE 6 (MEDICAL CONFIRMATION AND EVALUATION) OF THE CPAM ROADMAP	251
FIGURE 97: WORD CLOUDS FOR INDIVIDUAL PROTOTYPE DESIGNS [203].....	256
FIGURE 98: EXCEL SHEET TO PREPARE DATA TO GENERATE WORD CLOUDS IN WORDLE	258
FIGURE 99: WORDLE ADVANCED WEBSITE TO GENERATE WORD CLOUDS	258
FIGURE 100: FREQUENCY OF REACTION WORDS CHOSEN FOR THE EVALUATION	260
FIGURE 101: WORD CLOUD MADE OF THE EVALUATION RESULTS.....	262
FIGURE 102: PUBLIC HEALTH RESEARCHER CHOICES OF WORDS	263
FIGURE 103: PHYSICIAN CHOICES OF WORDS	263
FIGURE 104: CANCER QUALITY ASSURANCE MANAGER CHOICES OF WORDS	264
FIGURE 105: TOP 3 PROFESSIONS COMBINED WORDS CHOSEN	265

TABLE 1: SUMMARY OF AIMS AND OBJECTIVES	24
TABLE 2: EVENT LOG EXAMPLE WITH DUMMY DATA.....	39
TABLE 3: WORKFLOW LOG.....	45
TABLE 4: ACTIVITY AND PERFORMANCE OF THE TWO-MONTH WAIT STANDARD FOR DIFFERENT CANCER SITES 2013/14.....	52
TABLE 5: AGE CATEGORISED RAISED PSA LEVEL GUIDES (NHS).....	68
TABLE 6: SUBSET OF RAND QUALITY INDICATORS FOR EARLY STAGE PROSTATE CANCER [96]	71
TABLE 7: DANIELSON ET AL'S PROSTATE CANCER PRE-TREATMENT QUALITY INDICATORS [95]	72
TABLE 8: TABLE OF GLOBAL CORE DATA ELEMENTS FOR PROSTATE CANCER DIAGNOSIS AND REFERRAL BASED ON COSD DATA [102]	73
TABLE 9: AUDITING SUGGESTIONS FOR SUSPECTED PROSTATE CANCER BY NICE GUIDELINES [114]	77
TABLE 10: COMMON MINING APPROACHES TO PROCESS MINING TASKS BASED ON PERSPECTIVES [121]	89
TABLE 11: ELECTRONIC DATABASES SEARCHED	92
TABLE 12: NUMBER OF REFERENCES SEARCHED, KEYWORDS AND FILTERS APPLIED	94
TABLE 13:: COMPLETE LIST OF STUDIES WITH THEIR MINING TECHNIQUES USED WITH RESPECT TO MINING PERSPECTIVES	114
TABLE 14: PROSTATE CANCER RELATED DATABASE IN BIU	138
TABLE 15: RECORDS FILTERED FROM EACH DB AND TOTAL RECORDS LINKED	147
TABLE 16: EVENT LOG INFORMATION IN EACH TABLE	148
TABLE 17: DATA EXTRACTION RESULTS.....	149
TABLE 18: SNAPSHOT OF PATIENT-BY-PATIENT FLAT FILE	151
TABLE 19: PERFORMANCE METRICS FOR AUDIT	160
TABLE 20: TWW REFERRAL AUDIT FORM DATA	167
TABLE 21: CLINICAL LETTER AUDIT FORM DATA	168
TABLE 22: TOTAL NUMBER OF MATCHING RECORDS FOUND	168
TABLE 23: FIRST APPOINTMENT DATE MATCHES BETWEEN AUDIT AND DLE DATA	169
TABLE 24: DIFFERENCE BETWEEN THE PSA VALUES FOUND.....	170
TABLE 25: SUMMARY OF TWW AND NON-TWW APPOINTMENTS BETWEEN THE AUDIT AND DLE GROUPS	170
TABLE 26: LOG SUMMARY OF TWW CLUSTER	216
TABLE 27: START ACTIVITY FREQUENCY OF DIFFERENT PATHS IN THE TWW REFERRAL POST GP APPOINTMENT	216
TABLE 28: LOG SUMMARY OF URGENT CLUSTER.....	220
TABLE 29: LOG SUMMARY OF ROUTINE CLUSTER.....	223
TABLE 30: START ACTIVITY FREQUENCY OF DIFFERENT PATHS IN THE ROUTINE REFERRAL POST GP APPOINTMENT	223
TABLE 31: PERFORMANCE OF 10 TWW REFERRAL CLUSTERS.....	227
TABLE 32: PERFORMANCE OF 4 ROUTINE REFERRAL CLUSTERS.....	228
TABLE 33: DURATION COMPARISON BETWEEN ALL THE DIFFERENT REFERRALS	229

TABLE 34: BOTTLENECKS IN THE TWW REFERRAL	231
TABLE 35: MAXIMUM DELAYS IN HANDOVER IN TWW REFERRAL	234
TABLE 36: BOTTLENECKS IN THE URGENT REFERRAL.....	234
TABLE 37: MAXIMUM DELAYS IN HANDOVER IN ROUTINE REFERRAL	235
TABLE 38: BOTTLENECKS IN THE ROUTINE REFERRAL.....	237
TABLE 39: METRIC AND COMPLIANCE TABLE WITH CANCER WAITING TIMES	242
TABLE 40: EVALUATION METHODS CLASSIFIED ACCORDING TO TYPE AND METHODS	253
TABLE 41: TOP 4 POSITIVE WORDS CHOSEN WITH THEIR COMMENTS FROM PARTICIPANTS	261
TABLE 42: WORDS CHOSEN BASED ON THE 5 DIMENSIONS.....	266
TABLE 43: FULFILLMENT OF RESEARCH OBJECTIVES.....	271

ABSTRACT

Background

Prostate cancer is the most common cancer in men in the UK and the sixth-fastest increasing cancer in males. Within England survival rates are improving, however, these are comparatively poorer than other countries. Currently, information available on outcomes of care is scant and there is an urgent need for techniques to improve healthcare systems and processes.

Aims

To provide prostate cancer pathway analysis, by applying concepts of process mining and visualisation and comparing the performance metrics against the standard pathway laid out by national guidelines.

Methods

A systematic review was conducted to see how process mining has been used in healthcare. Appropriate datasets for prostate cancer were identified within Imperial College Healthcare NHS Trust London. A process model was constructed by linking and transforming cohort data from six distinct database sources. The cohort dataset was filtered to include patients who had a PSA from 2010-2015, and validated by comparing the medical patient records against a Case-note audit. Process mining techniques were applied to the data to analyse performance and conformance of the prostate cancer pathway metrics to national guideline metrics. These techniques were evaluated with stakeholders to ascertain its impact on user experience.

Results

Case note audit revealed 90% match against patients found in medical records. Application of process mining techniques showed massive heterogeneity as compared to the homogenous path laid out by national guidelines. This also gave insight into bottlenecks and deviations in the pathway. Evaluation with stakeholders showed that the visualisation and technology was well accepted, high quality and recommended to be used in healthcare decision making.

Conclusion

Process mining is a promising technique used to give insight into complex and flexible healthcare processes. It can map the patient journey at a local level and audit it against explicit standards of good clinical practice, which will enable us to intervene at the individual and system level to improve care.

CHAPTER 1: INTRODUCTION

This chapter provides an introduction to the thesis. Section 1.1 gives the motivation and inspiration behind choosing this research. Section 1.2 gives the importance of this research and the contributions made to knowledge. Section 1.3 presents the research questions, aims and objectives of this study. Section 1.4 provides a road map and structure of the various chapters in the thesis and what each chapter entails. Chapter 1 is concluded with Section 1.5 which provides a summary of the entire chapter and how the following chapter is linked.

1.1 MOTIVATION AND INSPIRATION

Not all organisations have a deep understanding of how their processes are executed in reality; where they fall short; or how can they be streamlined. Healthcare organisations are exposed to very flexible and constantly changing environments where the actual execution of processes often deviates from the supposed behaviour. Process mining techniques can deliver a deep insight into the many perspectives of processes to quickly give a snapshot of what is actually happening in the back end to have an eye opening moment about the process flow in the organisation. Process mining is a paradigm that comes from the business process management research field, which allows the discovery and graphical representation of human-understandable models that represent the real execution of a process.

The lack of transparency in the patient pathway is the biggest motivation for this research. Patients travel in their journey through a clinical pathway and every service/department they encounter is oblivious of the rest of the journey of the patient. The patient faces bottlenecks and delays, but no one knows exactly where these delays happen as no one can see the pathway in its entirety. Therefore, my objective was to bring the true pathway into the forefront so better and more informed decisions can be taken to improve the quality of healthcare.

1.2 IMPORTANCE OF THE RESEARCH AND CONTRIBUTION TO KNOWLEDGE

Importance

“Measurement is the first step that leads to control and eventually to improvement. If you can't measure something, you can't understand it. If you can't understand it you can't control it. If you can't control it, you can't improve it” - H. James Harrington

It is critical in order to be able to improve quality of care to understand a particular healthcare system. Caregivers, patients and administrators lack a general view of the movement of patients because it is a complex pathway. The complexity makes it difficult to map the patients' journey. The real value is the insight that is gained from having a detailed picture of how the patients move within a specific pathway. This gives the practitioners an understanding of where their patients get delayed, shows them a common recurring pattern amongst clusters of patients and aids them in sound decision making.

With this study, I am offering a technique that will allow a more transparent understanding of the patient's journey within a prescribed healthcare system. I have chosen prostate cancer care in the acute setting as an exemplar.

Contribution to knowledge

With the help of this study, I am able to:

- Use local database systems to generate a picture of patient journeys using prostate cancer as a model. I have determined what information is being captured at each step of the patient's pathway and have linked various systems required to perform analysis
- Perform a case note audit of extracted data against data found in the medical records of patients to highlight deficiencies and compare the outcomes
- Provide a systematic review of how process mining techniques and perspectives have been used in healthcare; where there needs to be more concentration and focus; which disciplines have used process mining; and which countries have championed it

- Show how I can use visualisation tools from process mining to improve understanding of patient journeys in a complex healthcare system. I am also able to identify and visualise bottlenecks, deviations and delays in the current pathway. I have compared actual journeys against the standard prescribed pathways

1.3 RESEARCH QUESTIONS, AIMS AND OBJECTIVES

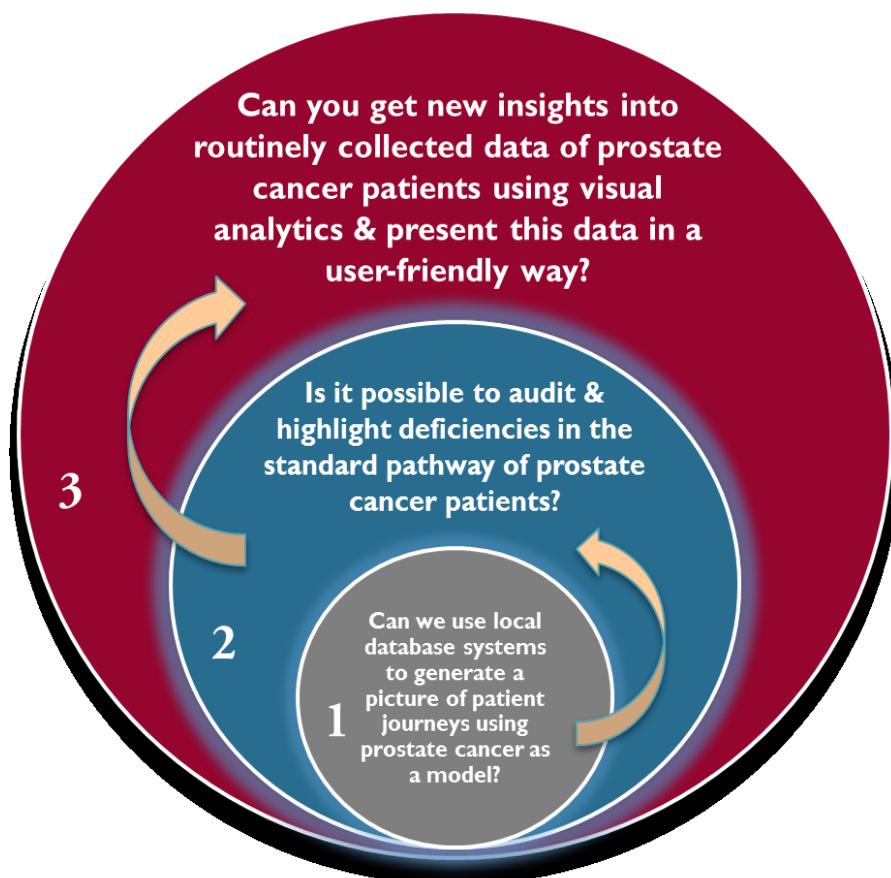


FIGURE 1: RESEARCH QUESTIONS OF MY STUDY

This work aims to analyse prostate cancer patient treatment pathways within the west and south of London in order to identify delays, bottlenecks and deviations from a known standard pathway (i.e. the pathway suggested by the London Cancer Alliance). These deviations will aid in capacity planning and eventually the development of a technique or model to highlight the nonconformities and gaps in care.

Figure 1 highlights my research questions as seen from different perspectives. At the highest level forms the question that gives the final output tool that shows the actual pathway followed by prostate cancer patients and highlights the deviations and bottlenecks in care. The lowest base level question is my starting point of analysing routinely collected data to formulate the actual pathway.

The following table (Table 1) summarises my research questions, aims and objectives:

TABLE 1: SUMMARY OF AIMS AND OBJECTIVES

Research Question/Aims	Objectives	Methods
<p>Question 1: Can we use local database systems to generate a picture of patient journeys using prostate cancer as a model?</p> <p>Aim: Analyse routinely collected data from current prostate cancer pathways</p>	<ol style="list-style-type: none"> 1. Do a literature review on the development of care pathways 2. Determine the clinical information systems being used in the urology clinic setting at St. Mary's Hospital in London 3. Determine what information is being captured at each step of the patient's pathway 4. Link various systems required to perform descriptive analysis 	<ul style="list-style-type: none"> • Studies selected will cover the development, usage, importance and types of care pathways specifically prostate cancer • Obtain NHS honorary contract • Arrange to meet IT staff/Urology staff for an overview of the systems in use • Get training for the necessary systems if required • Informally discuss with clinicians and/or medical staff on prostate cancer patient's journey in the hospital system • map the patient's journey at each step of treatment collecting information entered at each stage • create a separate record of the variables + computer systems • create a visual pathway of patient journey • Study the different databases, outline data structure and identify business rules • Clean data • Link using appropriate matching technique • Perform descriptive analysis • Report results of analysis • Discuss the analysis with the clinicians • Feed the results back to the clinicians

<p>Question 2: Is it possible to audit & highlight deficiencies in the standard pathway of Prostate Cancer patients?</p> <p>Aim: Perform a case note audit of extracted data against data found in the medical records of patients</p>	<p>5. Do a literature review on cancer quality indicators required minimum data set for reporting cancer, and audit requirements for prostate cancer</p> <p>6. Create an audit data collection tool</p> <p>7. Audit the hospital pathway against a known standard to find gaps in care</p>	<ul style="list-style-type: none"> Studies selected will cover protocols for cancer referral, diagnosis, and treatment, variables for auditing prostate cancer patient pathways Create a tool for collecting data for auditing using the variables and quality indicators obtained in the literature review of objective 5 Get ethical approval Seek the help of a clinical fellow to assist with an audit of the system using the tool created in objective 6. The audit will cover a retrospective date range of 10 months Measures to audit: Indications for referral Compare audit outcomes against case notes Report the results of the retrospective audit
<p>Question 3: Can you get new insights into routinely collected data of Prostate Cancer patients using visual analytics & present this data in a user-friendly way back to the physicians?</p> <p>Aim: Use a technique to identify bottlenecks and deviations in the current pathway and display this information back to the clinicians</p>	<p>8. Do a systematic review on how process mining has been used in healthcare</p> <p>9. Use the analytical and visualisation technique in a healthcare setting</p> <p>10. Evaluate the visualisation technique</p>	<ul style="list-style-type: none"> Studies selected will cover only the discovery of knowledge from process mining tools in healthcare Follow a methodical framework of data preparation and analysis Analyse the pathway through various process mining perspectives Do process discovery and comparison to the LCA guideline Highlight deficiencies, bottlenecks and gaps Research on different methods of evaluation Evaluate using a selected UX method Give end product to panel of experts in the field like Urology staff/IT staff/clinical staff to get feedback

1.4 THESIS STRUCTURE

This thesis is structured into 9 chapters with a reference list and appendices following the 9th chapter. It includes 6 empirical chapters, 2 introductory and background chapters and 1 discussion and concluding chapter.

Following is an account of what each chapter contains:

Chapter 1: Introduction	<p>This chapter gives an introduction to the research</p> <p>It talks about the motivation and inspiration that led to the research</p> <p>It presents the research questions, aims and objectives</p> <p>Finally, it presents the thesis structure</p>
Chapter 2: Background	<p>This chapter presents the background of the fundamental concepts used throughout the thesis</p> <p>It introduces the concept , history, techniques and challenges of process mining that is the primary technique used in the research to analyse and display the actual patient pathway</p> <p>It introduces the concept, requirement, strengths and format of care pathways to form the basis of studying the current care pathway used in the research</p> <p>Finally, it introduces the diagnosis, referral, risk management and economics of Prostate Cancer that forms the case study of my research</p>
Chapter 3: Systematic Review of Process Mining Applications in Healthcare	<p>This chapter presents a systematic review of process mining applications in healthcare</p> <p>The review provides an analysis of research evidence relating to the applicability of the discovery aspect of process mining in the healthcare domain and particularly in the study of care processes.</p> <p>Using process mining as a technique, the results display the various healthcare processes analysed, as well as the different pre-processing techniques and process mining methodologies used in healthcare</p> <p>The chapter forms a rigorous groundwork of where the gaps in literature are so that my reserach can contribute accordingly</p>

Chapter 4: Process Model Construction	<p>This chapter initiates phases 1 and 2 of the Clinical Pathway Analysis Method (CPAM) road map that I will be using throughout my research</p> <hr/> <p>It describes the methods I used in constructing the process model through various data integration, linkage, extraction and preparation steps</p> <hr/> <p>The outputs of this chapter (i.e. event log) form the input to the next phase in the CPAM: Validation of data linkage and extraction, in chapter 5</p> <hr/>
Chapter 5: Data Validation	<p>This chapter continues to phase 3 of the CPAM road map</p> <hr/> <p>It takes outputs from Chapter 4 (the patients' event log produced by data linkage and extraction algorithms) and validates them by performing a case note audit against the same patients found in the medical records</p> <hr/> <p>The results show the patient matches and targets met for the performance metrics when I compared the case note audit data against my date linkage and extraction (DLE) patients</p> <hr/> <p>Before proceeding to the analysis, the outputs of this chapter give me a validation that the patients I have extracted via my linkage algorithm conform to the patients that have actually been seen in the clinic within that time frame</p> <hr/>
Chapter 6: Descriptive Results	<p>This chapter presents a first inspection of the log and segregates the event log into two cohorts based on the availability of data</p> <hr/> <p>It provides basic statistics of the log such as the total number of events in the log, the average number of activities per case with their minimum and maximum values and the number of occurrences of specific activities</p> <hr/> <p>Outputs of this chapter give an overview of the statistics alongside the inspection of the log that later helps in filtering the log down</p> <hr/>

Chapter 7: Process Mining and Visualisation of the Pathway	<p>This chapter continues to phases 4 and 5 of the CPAM road map</p> <p>It begins by preprocessing and filtering the event log based on techniques learnt from the systematic review in Chapter 3</p> <p>It then goes on to the exploratory pathway analysis and visualisation using process mining techniques and perspectives learnt in Chapter 2</p> <p>Results are displayed and analysed based on control flow and performance mining aspects and are segregated into two clusters (following the cohorts made earlier in chapter 6)</p>
Chapter 8: Evaluation of Process Mining Techniques and Visualisation	<p>This chapter concludes the last phase 6 of the CPAM road map</p> <p>It describes the method I used to evaluate the visualisations produced from the process mining techniques I am using to analyse care pathways</p> <p>The evaluation was done using Microsoft Reaction Cards on 13 individuals from different professions</p> <p>Results of the evaluation were grouped into 5 different dimensions which aided me to analyse how different aspects of the visualisations were perceived</p>
Chapter 9: Discussion and Conclusion	<p>This chapter concludes the thesis</p> <p>It provides a summary of results from chapters 3, 4, 6, 7 and 8 of the thesis</p> <p>It presents the strengths and weaknesses of the research</p> <p>It shows the novelty of the work and gives recommendations for future work</p>

1.5 SUMMARY

This chapter gave a brief insight into the motivation and inspiration behind the work that went into this research. It also presented the importance of this study and the contributions it has made to knowledge and the research community. The chapter also provided the research questions, aims and objectives and a detailed road map of the chapters that follow this introduction. The next chapter, Chapter 2: Background, sets the stage for the three fundamental concepts that are presented in this thesis: Process Mining, Care Pathways and Prostate Cancer.

CHAPTER 2: BACKGROUND

This chapter introduces background concepts used throughout this thesis. Since the core of my research utilises the concept of process mining, section 2.2 starts with a detailed discussion on process mining techniques, perspectives and the various domains it is used in. Section 2.3 introduces the concept of care pathways which is relevant to understanding how current clinical care pathways are structured to better understand the flow of patients in my study. In Section 2.4, prostate cancer is discussed on a clinical, administrative and economic level and forms the basis of my case study for this thesis. Chapter 2 is concluded with Section 2.5 which provides a summary of the entire chapter and how the following chapter is linked.

2.1 INTRODUCTION

There are more than 200 types of cancers found in the UK each having different symptoms and modes of treatment. In 2010, more than 324, 500 people were diagnosed with cancer making it “the number one fear for the British public”[1]. According to the 2012 statistics, UK has the 22nd highest cancer-rate in the world with around 267 out of 100,000 people affected by this disease every year [2]. Breast, lung, bowel and prostate cancers together account for over half of all new cancers each year. Overall cancer incidence rates have increased by one quarter since 1975 and the increase is largely accredited to earlier diagnosis and aging of the population [3]. The EUROCARE-4 cancer survey, that monitors cancer survival in Europe, showed that the UK ranked 9th for male cancer mortality rates (where first equals lowest rates), and 22nd for female cancer mortality rates, compared with 27 other European countries [4]. It is evident from numerous studies in the past decade that cancer survival in the UK has been low compared to Europe and the United States [5-7]. In order to contend with the situation and suggest ways of improving patient care, it is important to find explanations as to why cancer survival is low in the UK. Several possible explanations exist that could substantiate poor cancer survival: Delays in approaching the care-provider and hence referrals to the secondary care provider [8]; wide variation in treatment pathways [9], suboptimal treatment [10] and socio-economic inequalities [11, 12]. Currently, in most UK hospitals and health care centres, there is scant information on quality or outcomes of care [13]. There is clearly a need for more sophisticated system measures that look at the whole patient journey so that failings and deviations in the delivery of healthcare, particularly in the handover of care between different services, could be identified. The multidisciplinary structured care plan used to map the entire

journey of a patient with a particular condition is known as an integrated care pathway. Mapping the patient journey at a local level and auditing it against explicit standards of good clinical practice will enable us to intervene both at the individual level and at the system level to improve care. As many hospital information systems contain rich amounts of semi-structured data about different processes in the hospital, it becomes difficult to use regular Business Process Management (BPM) or Workflow Management (WFM) techniques and it is more practical to use process mining techniques to give insights into the real execution of healthcare processes.

Before introducing the background concepts of processing mining and care pathways used in my work, I want to present a layout of the roadmap that I am following throughout the methodology and analysis of my work in this thesis.

My thesis is constructed along the concept of the Clinical Pathway Analysis Method (CPAM) proposed by Caron et al in [14]. In this methodology, process mining-based analytics is used to provide valuable insights into clinical pathways by looking at audit trails (patient traces) of previous pathway instances. Moreover, this methodology provides various kinds of performance-related information such as flow time of patients, the utilization of performers (or departments) and execution frequencies of events. It is also suited to assess guideline compliance and analyse adverse events in a healthcare setting.

Ideally, The CPAM follows a seven-phase approach as shown in Figure 3. However, due to the scope of my PhD and limitation of time, I have followed five phases from this approach in addition to one new phase that I consider important and relevant in validating my work.

The following diagram (Figure 2) presents the phases I am following in my study:

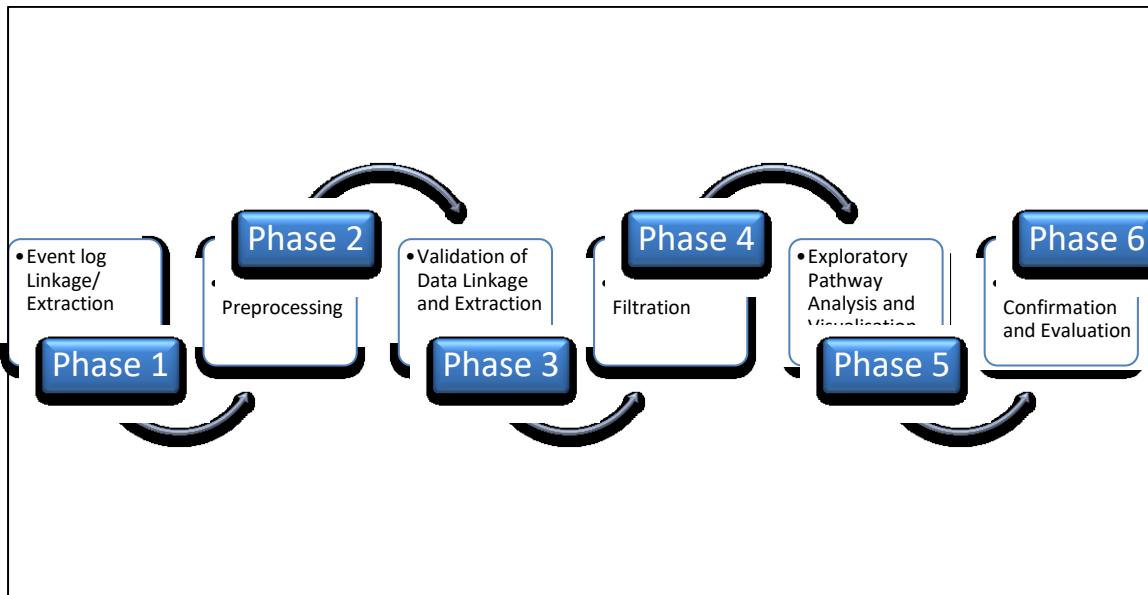


FIGURE 2: CPAM PHASES IN MY THESIS

- Phase 1: Project Definition and Event log Linkage/Extraction
- Phase 2: Event Log Preprocessing and Preparation
- Phase 3: Validation of Data Linkage and Extraction (new phase not in CPAM)
- Phase 4: Event Log Filtration and Perspective Selection
- Phase 5: Exploratory Pathway Analysis and Visualisation
- Phase 6: Medical Confirmation and Evaluation of Process Mining Techniques

The Following are the original phases of the CPAM Road Map by Caron et al:

Phase I	Project definition & event log extraction <ul style="list-style-type: none">• Define pathway scope definition• Identify event sources• Select event log attribute• Construct event log											
Phase II	Event log preprocessing <ul style="list-style-type: none">• Select log format and transform event log• Deal with log divergence• Event log specific operations, e.g. (re)grouping of activities, scrambling personal information, etc.											
Phase III	Perspective selection <ul style="list-style-type: none">• Identify interesting perspective (patient, treatment, diagnosis, department or drug)• Perform necessary log filtering operations• Identify and describe potential information losses, provide a solution to deal with the information losses											
Phase IV	Exploratory pathway analysis <table border="1" style="width: 100%; border-collapse: collapse;"><thead><tr><th style="text-align: left; padding: 2px;"><i>Functional analysis</i></th><th style="text-align: left; padding: 2px;"><i>Process analysis</i></th><th style="text-align: left; padding: 2px;"><i>Organisational analysis</i></th><th style="text-align: left; padding: 2px;"><i>Data analysis</i></th></tr></thead><tbody><tr><td style="padding: 2px;"><ul style="list-style-type: none">• Existence/absence of activities• Activity co-existence• Additional analyses</td><td style="padding: 2px;"><ul style="list-style-type: none">• Workflow discovery• Process variant analysis• Additional analyses</td><td style="padding: 2px;"><ul style="list-style-type: none">• Social network analysis (teams, hand-overs, interactions)• Task allocation• Additional analyses</td><td style="padding: 2px;"><ul style="list-style-type: none">• Data-driven conditions• Correlations data and pathway structure• Additional analyses</td></tr></tbody></table>				<i>Functional analysis</i>	<i>Process analysis</i>	<i>Organisational analysis</i>	<i>Data analysis</i>	<ul style="list-style-type: none">• Existence/absence of activities• Activity co-existence• Additional analyses	<ul style="list-style-type: none">• Workflow discovery• Process variant analysis• Additional analyses	<ul style="list-style-type: none">• Social network analysis (teams, hand-overs, interactions)• Task allocation• Additional analyses	<ul style="list-style-type: none">• Data-driven conditions• Correlations data and pathway structure• Additional analyses
<i>Functional analysis</i>	<i>Process analysis</i>	<i>Organisational analysis</i>	<i>Data analysis</i>									
<ul style="list-style-type: none">• Existence/absence of activities• Activity co-existence• Additional analyses	<ul style="list-style-type: none">• Workflow discovery• Process variant analysis• Additional analyses	<ul style="list-style-type: none">• Social network analysis (teams, hand-overs, interactions)• Task allocation• Additional analyses	<ul style="list-style-type: none">• Data-driven conditions• Correlations data and pathway structure• Additional analyses									
Phase V	Medical confirmation <ul style="list-style-type: none">• Review by medical expert(s) of results• Comparison with medical guidelines• Determine whether the results represent local conditions• Externalisation of knowledge											
Phase VI	Advanced pathway analysis <table border="1" style="width: 100%; border-collapse: collapse;"><thead><tr><th style="text-align: left; padding: 2px;"><i>Efficiency analysis</i></th><th colspan="3" style="text-align: left; padding: 2px;"><i>Quality and conformance analysis</i></th></tr></thead><tbody><tr><td style="padding: 2px;"><ul style="list-style-type: none">• Bottleneck analysis• Number & duration of diagnosis & treatment cycles• Performance analysis and comparison• Additional analyses</td><td colspan="3" style="padding: 2px;"><ul style="list-style-type: none">• Rule-based pathway analysis• Conformance & delta analysis• Analysis of adverse events• Root-cause analysis for variation• Additional analyses</td></tr></tbody></table>				<i>Efficiency analysis</i>	<i>Quality and conformance analysis</i>			<ul style="list-style-type: none">• Bottleneck analysis• Number & duration of diagnosis & treatment cycles• Performance analysis and comparison• Additional analyses	<ul style="list-style-type: none">• Rule-based pathway analysis• Conformance & delta analysis• Analysis of adverse events• Root-cause analysis for variation• Additional analyses		
<i>Efficiency analysis</i>	<i>Quality and conformance analysis</i>											
<ul style="list-style-type: none">• Bottleneck analysis• Number & duration of diagnosis & treatment cycles• Performance analysis and comparison• Additional analyses	<ul style="list-style-type: none">• Rule-based pathway analysis• Conformance & delta analysis• Analysis of adverse events• Root-cause analysis for variation• Additional analyses											
Phase VII	Improvement of pathway <ul style="list-style-type: none">• Adapt clinical pathway models according to new insights• Reinforce existing to-be models											

FIGURE 3: CPAM ROAD MAP BY CARON ET AL.

2.2 PROCESS MINING

Process mining aims at “extracting process knowledge from event logs” [15] and provides this knowledge by deriving process models from the “observed system behavior” [16] in order to improve and examine actual processes [17]. Process mining helps in giving an insight into the exact processes, as well as answering questions related to compliance and performance like where are there bottlenecks, how can they be removed, and why are there deviations from the path [18].

2.2.1 HISTORY OF PROCESS MINING

The concept of process mining, which emerged a little over a decade ago, is related to data mining, machine-learning and Business Intelligence (BI). These techniques also aim at knowledge discovery, performance measurement and prediction; however, process mining is more process-centric and focuses on discovery, conformance checking and other process analysis. Traditional data-mining approaches are not process-centric and they take as input a set of records and output decision-trees, trends and patterns and collection of clusters. Process mining, on the other hand, starts from events and looks at an end-to-end process model. It is bridging the gap between classical process model analysis and data oriented analysis like data mining and machine learning because it is focusing on processes but at the same time using the real data. Process mining has only become available recently, but it is “mature enough to be applied to (care) processes of any type and of any complexity” [19]

Process mining is based on the foundations of Business Process Management (BPM) and Business Process Analysis (BPA). Although both these concepts are usually used interchangeably, the notion of BPA consists of the aim to equip organisations with knowledge about how their processes function, to help detect disparities between set guidelines and actual practice [20], and the concept of BPM is a more holistic management approach that focuses at end-to-end process performance and customer quality. In the end, both concepts aim for optimising process performance [21].

To fully understand how process mining emerged and how it is placed within the BPM discipline, it would be interesting to see how BPM evolved historically over the years. I will present two different outlooks towards the evolution of BPM. The first, described by Van der Aalst from the standpoint of information systems development; and the other by Lusk et al from the viewpoint of businesses led by technology. In Figure 4, Van der Aalst has shown the ongoing trends for information systems that are relevant to BPM and how the trend shifted towards process orientation, redesign, and organic growth. The figure shows how today’s information systems are made up of a number of layers that contain various applications performing specific functions. The first central layer consists of the operating system (the software that controls all the hardware); the second layer consists of generic applications (like text editors, spreadsheet programs); the third layer consists of domain specific applications (like applications specific to a particular enterprise e.g. clinical decision support systems

in hospitals); and the fourth layer consists of tailor-made applications (applications special to specific organisations) [22]. The various layers we see today in information systems have evolved over decades.

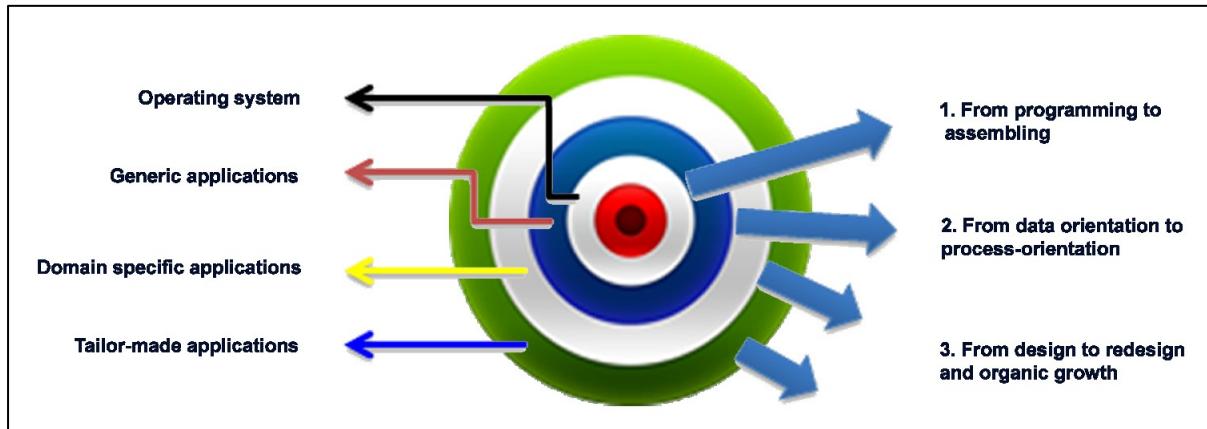


FIGURE 4: TRENDS IN INFORMATION SYSTEMS RELEVANT TO BPM [22]

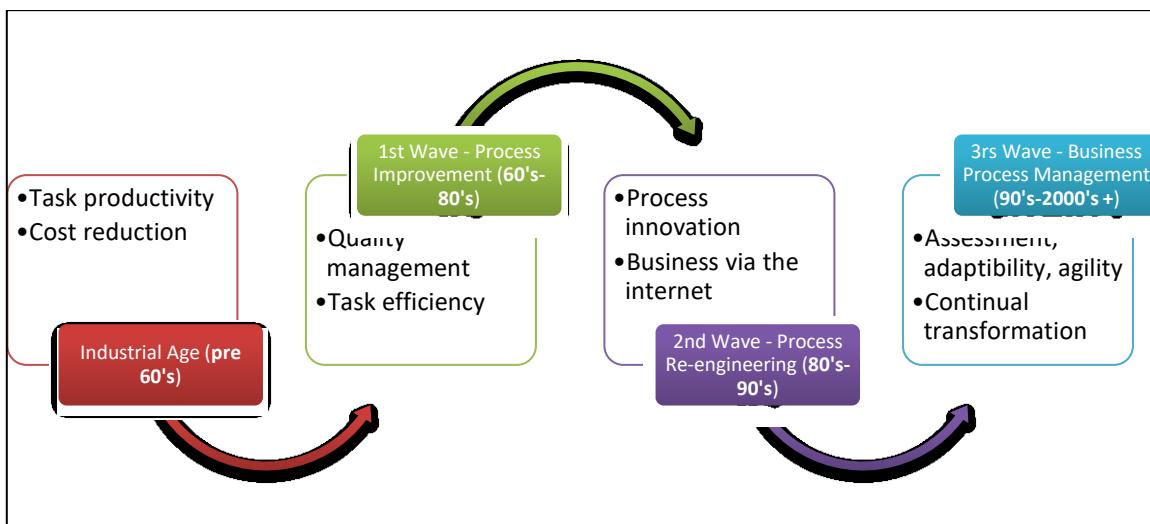


FIGURE 5: THREE WAVES OF PROCESS EVOLUTION [23]

Lusk et al [23] have also measured the growth of BPM as three waves of process evolution (Figure 5). They describe the evolution of BPM as a result of business tools, methodologies, technological

advancements as well as standards. The following sections briefly describe how BPM eventually changed its focus to being process-centric along the historical timeline as perceived by Van der Aalst [22] and Lusk et al [23].

THE SIXTIES

In the sixties, information systems were built on top of a limited operating system and the second and third layers did not exist. These systems, therefore, relied only on tailor-made applications. The trend in this era started to shift from programming to assembling complex software[22].

Lusk et al have described the latter part of the industrial age as an age where the standards and controls in organisations were more mechanical and deterministic and the focus was more on task productivity and cost reduction. As technology advanced in the 60's and 70's, this led to a rapid change in quality improvement initiatives as well as managing tasks efficiently due to increased competition. Technology thus became a process driver. This was called the 1st wave of process evolution – Process Improvement.

THE SEVENTIES AND EIGHTIES

The seventies and eighties were dominated by data-driven approaches. More and more focus was placed on retrieving data, storing and managing data. As a result data mining and data modeling became popular but process mining and process modeling were overlooked.

Similarly, Lusk et al have described that the growing use of computers in these decades led to increased data gathering and techniques for interpreting the gathered data results. Like Van der Aalst, Lusk et al have reported a noticeable inclination towards data-centric approaches and after a decade of statistical analysis and handling data, the focused moved to handling data in a meaningful way. This was called the 2nd wave of process evolution – Process Re-engineering.

THE NINETIES AND ONWARDS

The Nineties witnessed a shift from sensibly planned designs to redesign and organic growth. As information systems are rapidly changing in real-time, lesser systems are being built from scratch and existing modules are being molded to adapt to newer requirements. The BPM systems are

either located in the second layer (generic applications) or are integrated in the third layer (domain specific applications). Examples of BPM systems residing in the second layer are workflow management systems. BPM systems prevent work processes from being hard-coded into tailor-made applications and offer support for re-design and swift adaptation to new requirements.

Lusk et al describes this era as the “coming of age” of process-centric business. Technology shifted to being a process enabler from being just a process driver. As applications could be utilized regardless of the operating system under, “process management” could now be distinct and separated from “systems management” and “business management”. Processes can now be measured, managed, and integrated with technology. Enterprises can now be labeled as “Adaptive enterprises” as processes can be manipulated to adapt to the needs. This was the era where BPM began and flourished. This era was called the 3rd wave of process evolution since “the evolution of Business Process Management (BPM) as a customer-centric and process centric approach to improving business results entered its third wave” [23]

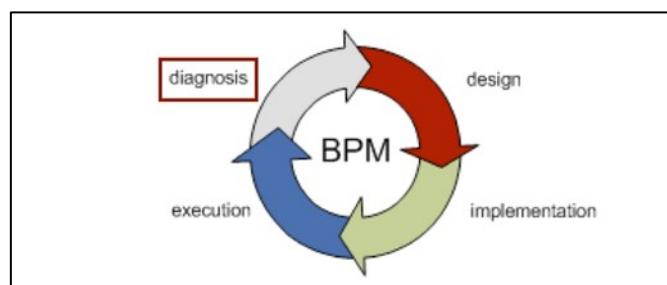


FIGURE 6: BPM LIFECYCLE

Figure 6 describes how BPM reiterates around the design, implementation, execution, diagnosis (analysis) and re-design of processes. Process mining fits into the diagnosis or analysis portion of this cycle. Traditional BPM approaches start by modeling the process, however, process mining starts by discovering the actual processes from data and understanding the processes that are already there [24].

“BPM is a well-designed, implemented, executed, integrated, monitored, and controlled management approach, which strives to continuously improve and analyze key operations in line with organizations’ strategies” [25]

BPM is said to be composed of the following six core elements [25]:

1. Strategic alignment – processes need to be designed, implemented, maintained, and assessed in line with the strategic priorities of the organization.
2. Governance –establishing accountability of the roles and responsibilities within all levels of the management process.
3. Methods – within BPM, the set of tools and techniques that are used to support and initiate the activities along the process life cycle.
4. Information technology (IT) – IT-based solutions are important elements in BPM
5. People – the human capital of an organization for the effective implementation of BPM.
6. Culture – the shared values of the people forming part of the organization

Based on a research indexed in PubMed that broadly lists articles using BPM in healthcare (until 2014), the trend shows a lower presence of BPM in the health care field. The apparent slower endorsement of BPM research in health care is a reflection of the disjointed health care systems with separate data sets for various settings, thereby preventing in-depth and system-wide process examinations [25]. The issues facing healthcare currently include rising costs, variations in quality, diversity in consumers, and concerns about value in return on investment. To address these issues, BPM not only helps to develop standardized processes within health care systems but also helps to minimize the variation in quality of health care delivery and errors. It also helps to select the right information management and technology resources to manage these processes. Furthermore, BPM can also help manage patient flow and information flow, which facilitate managing waiting times in health care delivery [25].

CURRENT STATE OF PROCESS MINING

Process mining research at TU/e (Eindhoven University of Technology) started in 1999 and later became the hub of process mining in the world. When it first started, there was little event data available and hence process mining was undeveloped. Today, with the availability of abundant event data, process mining has seen a tremendous momentum with implementations seen in various industries. The number of actively participating researchers in process mining from around the world

is increasing, with many who have made their names high in this field. More and more software vendors are adding process mining functionalities to their software.

DOMAINS THAT USE PROCESS MINING

Process mining has been applied to a variety of domains as it creates a relation between the actual behaviour with modeled behaviour. These domains include: software engineering [26], Ubiquitous Mobile Systems [27], healthcare [see section 1.2.7], banking [28], insurance [29], e-government, production, customer relationship management, supply chain optimization [30], and remote monitoring to name a few [19].

2.2.2 PROCESS MINING REQUIREMENTS

Process mining starts with event data. Every row in the table corresponds to an instance of an event and also contains different attributes pertinent to that event like the case ID, activity (event) name, timestamp and other relevant information [31]. A fragment of an example log containing dummy data and names is shown in Table 2. The table contains 12 events for 3 cases. Each line represents one event belonging to a particular case (events are grouped per case). Each event has an activity, a resource (optional) and a cost (optional). In some logs there can be more than one resource (performer of the activity), and in some logs this can be missing. The cost (or similar) attribute is a data attribute used to imply additional process knowledge. The event ID is used for distinguishing one event from another. The timestamp field shows the time the activity was completed (used to calculate performance properties). In some logs this field can be more detailed to include a start and end time. The minimal requirements from an event log are that any event can be related to a case and an activity, and that the events in a case are ordered. Therefore, the “case ID” and “Activity” fields are the minimum necessity for process mining [32].

TABLE 2: EVENT LOG EXAMPLE WITH DUMMY DATA

Case ID	Event ID	Properties			
		Timestamp	Activity	Resource	Cost
1	1234560	10/01/2015	Outpatient Appt.	John	130
	1234569	13/01/2015	Lab	Bill	100

	1234566	20/01/2015	MRI	Sandra	300
	1234565	20/02/2015	Surgery	Josh	700
2	9876541	13/02/2015	Outpatient Appt.	Peter	120
	9876544	23/02/2015	MRI	Sandra	300
	9876548	01/03/2015	Lab	Liz	120
	9876542	25/03/2015	Surgery	Nathan	850
3	2468002	15/03/2015	Outpatient Appt.	Peter	125
	2468009	22/03/2015	Lab	Bill	230
	2468000	02/04/2015	MRI	Mary	350
	2468005	01/05/2015	Surgery	Nathan	950

With advancements in data management technology, data warehouses (called Work Flow Management Systems) now house detailed data points belonging to each event that can easily be utilized for process mining. If these data points are not available, traditional data extraction techniques can be used to develop a process log database derived from a dataset of these event points in order to create a process model showing the entire activity [33], [34].

Figure 7 shows the relationship between a process model and an event log. There are three levels that can be seen: process model level, instance level, and the event level [19].

The process model level consists of the processes and activities. This level can be seen as a bird's eye view of all the activities.

The instance level, taking us one level deeper, consists of the cases and the activity instances. Finally the event level consists of the events and their attributes. The relationship between these different concepts is illustrated using cardinalities.

Example:

- Each process can have many activities but each activity can have only one process associated with it
- A process can have many cases but each case can refer to one process only
- A case can have many activity instances but an activity instance can refer to one case only
- Each activity instance can have many events

- Each event can have many event attributes

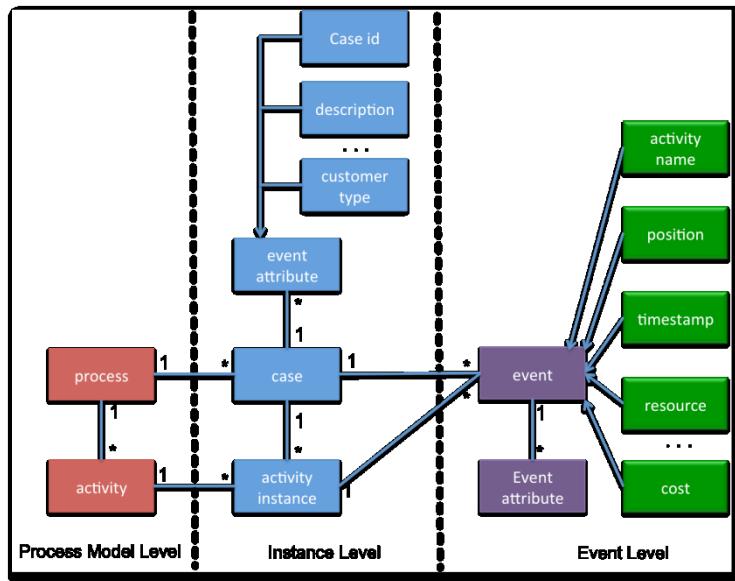


FIGURE 7: RELATIONSHIP BETWEEN PROCESS MODEL AND EVENT LOG

2.2.3 PROCESS MINING CHALLENGES AND CRITERIA

CHALLENGES WHEN EXTRACTING EVENT LOGS

Van Der Aalst has described the following common challenges faced when extracting event logs [32]:

Correlation

This is a simple requirement in the event log that needs all the cases to be related to each other. You should be able to identify events and their corresponding cases even if they are scattered across multiple tables and complex systems.

Timestamps

Events need to be ordered per case especially if the data is coming from different sources then a timestamp sort these events. This can become challenging if multiple local clocks are used and

recordings are delayed or partially recorded (like only date and no time). As a result the ordering of the events can become unreliable.

Snapshots

Event logs typically provide just a snapshot window of the longer running process. Some cases may have a lifetime longer or shorter than the event log window. It is best to solve this problem by removing the incomplete cases.

Scoping

Information systems have thousands of tables with business-relevant data. This makes it harder to decide which tables to include and which not. Domain knowledge is thus needed to scope the required data according to its availability and ability to answer the questions needed.

Granularity

Sometimes the events in an event log are at a different level of granularity than what the end users want. This can be solved by several approaches to pre-process the event log before using it.

Noise

Noise here refers to rare and infrequent behaviours (outliers or having “too much data”) usually caused by human or machine-related errors and does not refer to incorrect logging. Fortunately, there are many approaches now to filter out the noise e.g. heuristic mining, genetic mining and fuzzy mining.

Incompleteness

The notion of completeness is very important and is related to noise. It refers to having “too little data” possibly due to manual mishandling. Exceptions that are recorded just once should not be made part of the normal workflow in order to avoid any erroneous underlying dependencies.

QUALITY CRITERIA OF THE DISCOVERED MODEL

As the above challenges were related to the event log they do not say much about the quality of the resulting model. There are other dimensions that determine the quality of the process model as seen in Figure 8 [32]:

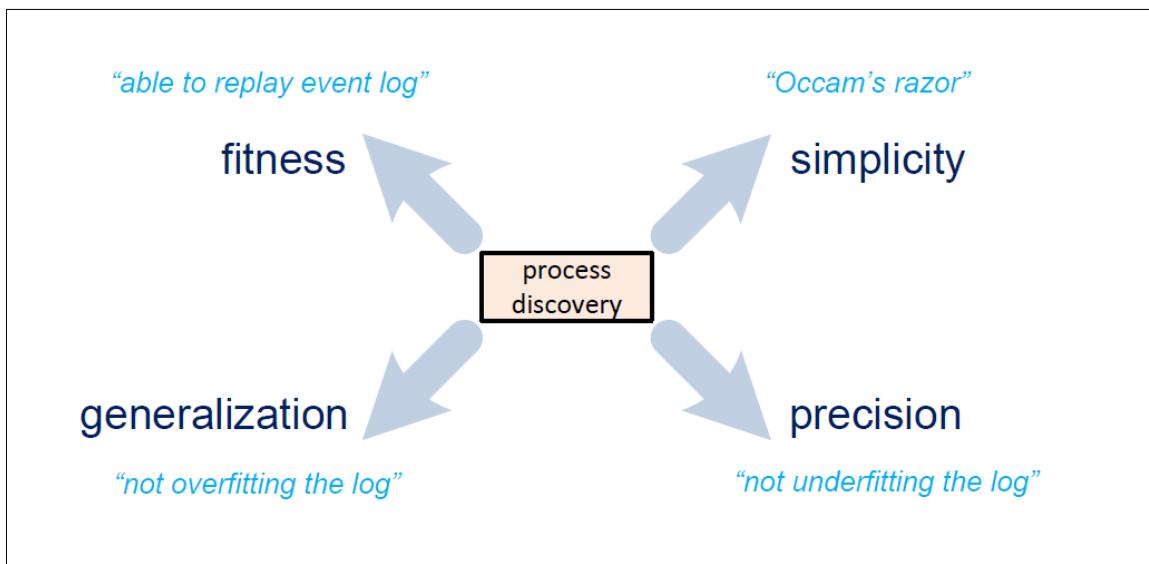


FIGURE 8: BALANCING THE FOUR QUALITY DIMENSIONS [35]

Fitness

A model with good fitness allows for the behaviour seen in the log. A model is said to have good fitness if all the traces can be replayed from beginning to end by the model.

Simplicity

In the context of process discovery this means the simplest model that can explain the behaviour in the log is the best model.

Precision

A model is precise if it doesn't allow for too much extra behaviour. If it is not precise it is "underfitting" which means it over generalizes and allows for behaviours not seen in the log.

Generalisation

A model should generalize and not restrict behaviour. A model that does not generalize is “over-fitting” meaning it allows only the exact behaviour recorded in the log.

2.2.4 PROCESS MINING TECHNIQUES

There are three types of process mining techniques: Process discovery, Conformance checking, and Enhancement [36], [37].

PROCESS DISCOVERY

Process discovery is one of the most challenging process-mining tasks. This technique takes an event log as input and creates process models automatically without a previously known standard model (a-priori model). Example of an algorithm which is used for process discovery is the α -algorithm that takes an event log, scans it, and produces a process model called Petri net or a Business Process Model and Notation (BPMN) model that shows the behaviour seen in the event log. An example of this process model would be to show the typical steps that are followed in a process. Although the α -algorithm is a very simple algorithm for this purpose, it is a good starting point to illustrate how process discovery can be implemented. Process discovery is done using the *control-flow perspective*, as discussed in section 1.2.5.

To demonstrate the principle of process discovery in more detail, I will take an example of the workflow log shown in Table 3 [38]. The log contains information about five cases. The log shows that:

- There are four cases: 1, 2, 3, and 4 for which the tasks A, B, C, and D are executed.
- For the fifth case only three tasks are executed: tasks A, E, and D.
- Each case starts with the execution of A and ends with the execution of D.
- If task B is executed, then also task C is executed.
- For some cases task C is executed before task B.

Based on the information provided in the workflow log (shown in Table 2) and by assuming that the cases are representative and cover a wide range of observed behaviour, we can deduce the process model shown in Figure 9, called a Petri net model. A Petri net model is an elegant and mathematically rigorous modeling tool for “systems that exhibit concurrency, synchronisation and randomness.” [39]. The following points can be concluded from the model in Figure 9:

- The Petri net starts with task A and finishes with task D.
- After executing A there is a choice between either executing B or C in parallel or just executing task E.
- To execute B and C in parallel two no observable tasks (AND-split and AND-join) have been added.

TABLE 3: WORKFLOW LOG

Case identifier	Task identifier
Case 1	Task A
Case 2	Task A
Case 3	Task A
Case 3	Task B
Case 1	Task B
Case 1	Task C
Case 2	Task C
Case 2	Task D
Case 5	Task A
Case 4	Task C
Case 2	Task B
Case 2	Task D
Case 5	Task A
Case 4	Task D
Case 1	Task C
Case 3	Task D
Case 3	Task B
Case 4	Task E
Case 5	Task D
Case 5	Task D
Case 4	Task D

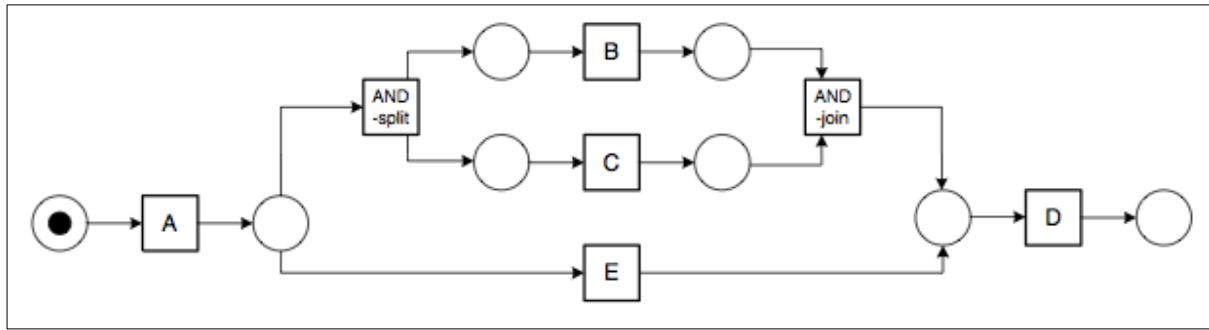


FIGURE 9: PETRI NET OF WORKFLOW LOG

CONFORMANCE CHECKING

In this technique, an existing process model (either constructed by hand or discovered) is compared with the event log of the same process to show how much a reality (depicted by the event log) deviates or conforms to a known process model. Conformance checking is used to detect deviations, analyse them and show their severity. It is used to show the alignment between the reality and a process model and repair models that are not well aligned.

ENHANCEMENT

This technique uses information acquired from the event log to project it onto the known process model for improvement and extension. The aim of this technique is to change and improve the a-priori model. There are two types of enhancements: *Repair*, which modifies the model to better reflect the reality. An example of this would be if the event log has activities that can happen in any order, then the process model is changed to reflect that; and *Extension*, that adds a new insight to the model by cross-correlating it with the event log. An example of this would be to show the bottlenecks, frequency and waiting times, quality metrics, etc.

Traditionally, process-mining techniques discussed are used in an offline setting using what is called “post mortem” data. This means that only completed cases are used. Process mining can also be performed in an online setting and is called *operational support*. Here, “pre mortem” data is considered as cases that are still running (and have a partial trace) can still be influenced and are responded to online. Three activities are identified for operational support:

1. Detect: This activity compares the partial trace with a normal model and detects any violations and provides an immediate response. It is also called conformance checking “on the fly”
2. Predict: This activity generates predictions about an event. E.g. the expected completion of time for a running activity can be predicted by comparing it to a similar case in the past
3. Recommend: This activity guides the user in selecting the next activity based on historic information

2.2.5 PROCESS MINING PERSPECTIVES

Related closely to the above-discussed process-mining techniques are the different process-mining perspectives. These perspectives provide a good classification of the different aspects of process mining analysis [32]:

THE CONTROL FLOW PERSPECTIVE

The control-flow perspective focuses mainly on the ordering of the activities. It helps in answering the question “how does the process occur?” The control-flow perspective of a process establishes the dependencies among its tasks. Which tasks precede which other ones? Are there concurrent tasks? Are there loops? In short, what is the process model that summarizes the flow followed by most/all cases in the log? This information is important because it gives you feedback about how cases are actually being executed in the organization. When mining this perspective, the goal is to show all the possible paths in an event log. The process models shown thus far show only the control-flow perspective. Examples of the modeling techniques to show this perspective are: Petri net, BPMN and UML to name a few.

THE ORGANISATIONAL PERSPECTIVE

The organisational perspective focuses on information about resources hidden in the log, i.e. which performers are involved in performing the activities and how they are related (e.g. people, systems, roles, departments). It answers questions like: “Who is performing the particular activities?” The goal here is to:

- Structure the organization by classifying people in terms of roles and organizational units
- Show relations between individual performers

In this perspective, we analyse the relation between the resources and the activities e.g. the mean number of times a resource performs an activity in a case.

Social Network Analysis is a method that presents data on interpersonal relationships in a graphical format and is used in the organisational perspective. A social network consists of nodes representing organisational entities and arcs representing relationships. Both the nodes and the arcs can have weights.

THE CASE PERSPECTIVE

The case perspective focuses on properties of cases. It helps in answering questions like: “what are the characteristics of a case that influence a particular decision?” Cases can be characterized by their path in the process; by the originators working on a case; or by the performance information. However, cases can also be characterised by the values of the corresponding data elements. This perspective is used to make decisions and shows which data is relevant and should be included in the model.

THE TIME/PERFORMANCE PERSPECTIVE

The time perspective focuses on the timing and frequency of events provided the events hold timestamps in the log. The granularity of the timestamps indicates what level of statistics can be collected. This perspective helps in answering the question like: “Where are the bottlenecks in my process?” It also helps in measuring service levels, monitor the utilization of resources and predict the remaining running time or delay times of running cases. This perspective can provide various kinds of performance-related information like:

- Visualisation of service waiting times
- Bottleneck detection and analysis
- Flow time analysis

- Utilisation and frequencies analysis

2.2.6 PROCESS MINING TOOL

The ProM (PROcess Mining) framework is the de facto tool for process mining aimed at covering the entire process mining spectrum. ProM is an open source extensible framework that supports a wide variety of process mining techniques in the form of plug-ins, i.e. people can add new process mining techniques by adding plug-ins without spending any efforts on the loading and filtering of event logs and the visualization of the resulting models. An example is the plug-in implementing the α -algorithm to automatically derive Petri nets from event logs. The ProM framework accepts event logs from various information systems as its input. In order to standardize the input supplied to ProM, a common format for the input called Mining Extensible Markup Language (MXML) format was developed and used. The format is XML based and is defined by an XML schema. When the ProM framework was being initially developed, it provided algorithms only for the discovery of process models. But now it offers other functionality like analysis of event logs, conversion of a model into another, exporting a model to a file, finding social network between different originators of a process, etc. ProM can be downloaded from www.processmining.org. [40]

2.2.7 PROCESS MINING IN HEALTHCARE

Healthcare has become home to vast competition that compels hospitals to remain in the race by achieving cost-effective quality of care. This standard of care, however, cannot entirely depend upon increasing the resources like physicians and beds. A great contributor to achieving this quality depends upon improving and streamlining the healthcare systems and processes (care plan), which a patient goes through [18], [20], [36], [41]. Improving the patient flow across an organisation decreases the likelihood of harming patients as well decreasing healthcare costs. Amongst the main methods researched for assessing patient flow across organisations include: analysing basic routinely collected data about service usage that involve retrospective analysis of data over a set period of time [10].

Many hospital information systems contain rich amount of data about different processes in the hospital. These processes are usually semi-structured with many exceptions and different stakeholders leading to complex decision-making. The execution of a semi-structured process is not completely implemented through a formal workflow model as information required for the model may only be partially recorded [42]. With these characteristics, it becomes impossible to use regular BPM or workflow management techniques. However, the abundant data collected in hospital information systems can be used to change care processes. This data is extracted using process-mining techniques [19]. As process mining gives insights to the real execution of healthcare processes, there is a growing uptake of this technique in the healthcare domain.

CHALLENGES IN HEALTHCARE

Healthcare is facing many different challenges. Care organisations are under enormous pressure to do “more for less” [19]. Some of the most prominent challenges that healthcare has to deal with is the need to improve productivity while reducing waiting times and costs. As treatment, diagnostics, and patient care become more specialized, costs are on the rise. Patients have greater expectations from the healthcare setting while the resources; funding and infrastructure remain the same. Clinicians and administrators perceive the best patient care is given with increased medical expertise. However, they fail to realise that an improvement in the systems and processes is a great contributor to increased quality of care [43].

To get a brief idea about the national health statistics, Figure 10 shows total healthcare expenditure per capita in UK between 1997 and 2012. Total expenditure on healthcare in the UK per capita has increased every year since 1997 however rates of growth have slowed since 2009. Average annual growth rates for total expenditure on healthcare per capita in the UK stood at 7.4% between 1997 and 2009. From 2009 to 2012 average annual growth has shrunk to 0.8%. In 2012, total healthcare expenditure per capita rose by 1.2%. In 2008, total healthcare expenditure per capita rose by 6.5% [44].

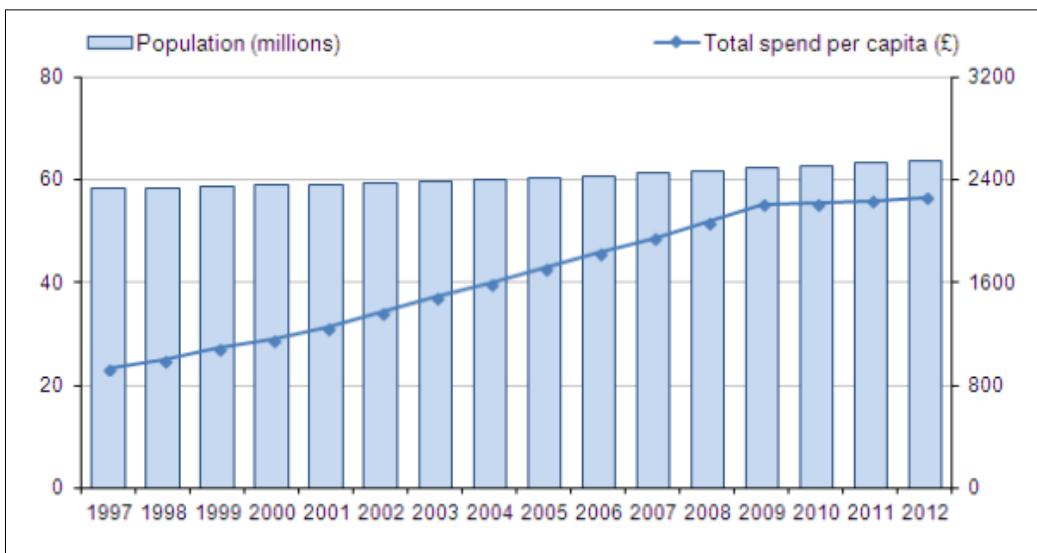


FIGURE 10: TOTAL HEALTHCARE EXPENDITURE PER CAPITA UK 1997-2012

(SOURCE: OFFICE FOR NATIONAL STATISTICS)

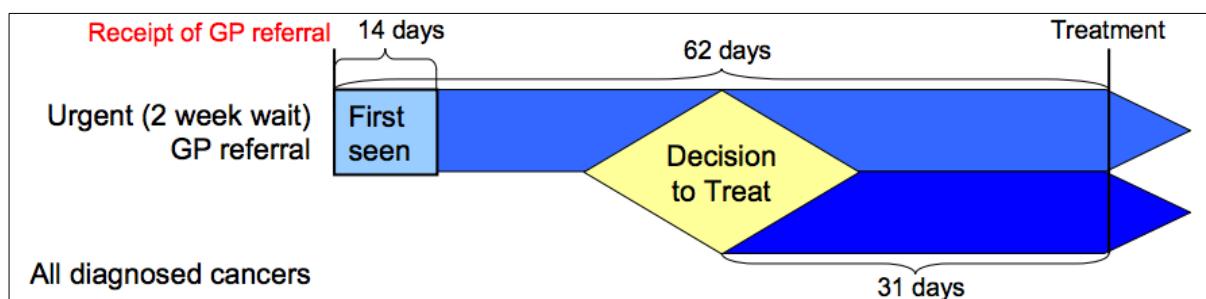


FIGURE 11: CANCER WAITING TIME STANDARDS

The NHS has set maximum waiting time standards for access to healthcare (Figure 11). There are a number of government pledges on waiting times, including [45]:

- A maximum one month (31 day) wait from the date a decision to treat (DTT) is made to the first definitive treatment for all cancers
- A maximum 31 day wait for subsequent treatment where the treatment is surgery
- A maximum 31 day wait for subsequent treatment where the treatment is a course of radiotherapy
- A maximum 31 day wait for subsequent treatment where the treatment is an anti-cancer drug regimen
- A maximum two month (62 day) wait from urgent referral for suspected cancer to the first definitive treatment for all cancers
- A maximum 62 day wait from referral from an NHS cancer screening service to the first definitive treatment for cancer
- A maximum 62 day wait for the first definitive treatment following a consultant's decision to upgrade the priority of the patient (all cancers)
- A maximum two week wait to see a specialist for all patients referred with suspected cancer symptoms

Moreover, Table 3 covers the 62-day wait for first treatment seen in 2013/14 following an urgent GP referral. It covers patients starting a first definitive treatment for a new primary cancer following an urgent GP referral for suspected cancer. The operational standard states that 85% of patients should be seen within 62 days of the referral date.

TABLE 4: ACTIVITY AND PERFORMANCE OF THE TWO-MONTH WAIT STANDARD FOR DIFFERENT CANCER SITES 2013/14

Cancer Report Groups	Total No. of patients seen	% Seen within 62 days
All Cancers	125,275	86.0
Breast Cancer	21,176	97.0
Lower Gastrointestinal Cancer	11,725	78.8
Lung Cancer	12,075	78.5
Other Cancer	30,120	80.9
Skin Cancer	20,767	96.7
Urological Malignancies	29,412	81.6

According to Table 3, the number of patients recorded under the 62-day standard increased by 6.2% from 2012/13 to 2013/14. No specific cancer site was responsible for this but rather a general increase across all cancers. The performance saw a large drop in the third and fourth quarters leading to the standard being failed in Q4 2013/14. This is the first breach of the operational standards since they were introduced. The performance for individual cancers showed that breast and skin cancers remained high and constant whereas urological and lung cancers had large decreases.

Based on The King's Fund Quarterly Monitoring Report (QMR) for 2015 [46], contractual penalties for missing referral-to-treatment waiting times performance standards were dropped in 2015 as part of a 'managed breach' policy to deal with patients still waiting to be seen and waiting over 18 weeks.

This 2015 breach is reflected in the latest figures shown in Figure 12 with waiting times for both non-admitted (outpatient) and admitted (inpatient) patients breaching in February 2015.

- The proportion of admitted patients waiting longer than 18 weeks rose to 13 per cent, the highest since this target was introduced.
- The proportion of non-admitted patients waiting more than 18 weeks rose to 5.3 per cent. This is the third breach of the non-admitted referral-to-treatment (RTT) target in the past four months.
- The number of patients still waiting to begin their treatment (both admitted and non-admitted) reduced to 6.9 per cent, which suggests the managed breach is having some positive impact.
- The proportion of patients waiting more than 6 weeks for a diagnostic test has now missed its target (1 per cent) for the past 15 months in a row.

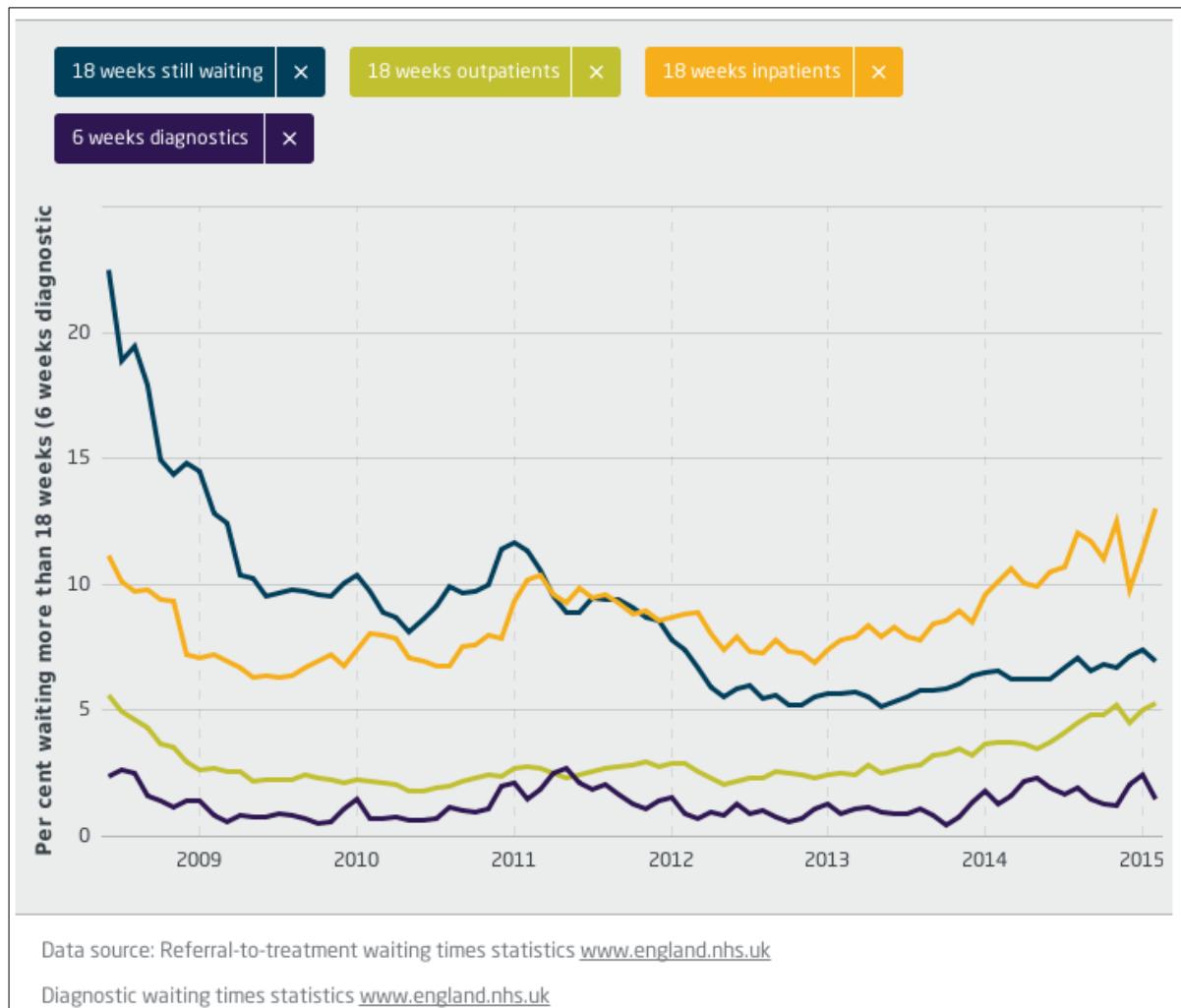


FIGURE 12: PERCENTAGE STILL WAITING/HAVING WAITED MORE THAN 18 WEEKS (MORE THAN SIX WEEKS FOR DIAGNOSTICS)

All the above statistics show the pressure on today's healthcare systems. Achievement of the national cancer waiting times (CWT) standards is considered by patients and the public to be an indicator of the quality of cancer diagnosis, treatment and care NHS organisations deliver. Delivering timely cancer pathways is crucial for the following reasons:

- Despite improving survival rates, cancer is the fourth leading cause of death in the UK
- Patients continue to present late to their GP with their symptoms, resulting in delayed referral
- There is variation in 2 week wait (2WW) referrals across the country suggesting that GPs are not always identifying suspicious symptoms

- Once a patient has been referred, they want to be told “It’s not cancer” as soon as possible or have their treatment planned in a timely manner
- Where the diagnosis is cancer, a speedy diagnostic pathway is critical for 62 day compliance

Despite consistent achievement of the cancer standards at a national level, it is recognized that many organisations either struggle to maintain compliant performance on a consistent basis or achieve below-standard performance.

One approach to improving these issues is to focus on the numerous, complex, time-consuming and non-trivial processes that occur in these organisations and try to improve and/or re-design them by analysing them. Traditionally, this analysis would be done by conducting interviews and surveys. However, this approach is very time consuming and provides a very subjective view of the processes since the stakeholders involved tend to have a very ideal scenario in mind and don’t consider the other situations. In order to give objective suggestions to improving these processes, the event log of these processes needs to be analysed using process mining techniques [19]. Event log data may come from a wide variety of sources [47]:

- A database system (e.g., patient data in a hospital)
- A comma-separated values (CSV) file or spreadsheet
- A transaction log (e.g., a trading system)
- A business suite/ERP system (SAP, Oracle, etc.)
- A message log (e.g., from IBM middleware)
- An open API providing data from websites or social media

HEALTHCARE PROCESSES

Healthcare is the “diagnosis, treatment and prevention of diseases in order to improve a person’s wellbeing” [19]. Healthcare processes are “highly dynamic, complex and increasingly multidisciplinary” [20] and many different kinds of healthcare processes exist with different characteristics.

It is important to point out that process mining only deals with care that is directly related to or provided to the patients. The focus is on operational processes that are concerned with the logistics

of work processes. The logistics deal with medical steps and the preparation required for these steps (E.g. making an appointment). Process mining of operational processes does not look at:

- process of individual physician decision for diagnosis and treatment
- medical interpretation of activities
- results of these activities (like blood test results)

Levels of care

Within the healthcare organisation, distinct levels of care exist corresponding to the particular needs of the patient:

Primary Care: Refers to the work of medical professionals that serve as the first point of consultation for a patient. In primary care patients with common health problems (like flu, hypertension) and patients requiring preventive measures (like vaccinations) are seen.

Secondary Care: Refers to the level of care immediately after the primary care and deals with patients requiring more specialized clinical expertise (e.g. patients with cancer). The physicians and other medical staff that see the patients in this level are generally not the first level of consultation for that patient. Secondary care is usually short-termed whereby a specialist provides expert opinion and intervention that primary care physicians are not equipped to provide.

Tertiary Care: Refers to the third level of care involving the management of rare and complex cases beyond the scope and capabilities of secondary and primary care. The patients referred to tertiary care are usually inpatients (e.g. Neonatal ICU). It is a highly specialized and technology driven level of care.

Sometimes a fourth level of care is also distinguished: this is *Emergency Care*. Emergency healthcare professionals provide this care to patients requiring urgent emergent care. Their role is to evaluate, manage and treat patients with unexpected illnesses and injuries. Their mission is also to provide care to patients who have no other means of care.

Health care industries are data rich as a result of embracing the notion of paperless system in a big scale [34]. Chapter 2 illustrates that process mining can be successfully applied in the healthcare domain.

2.3 CARE PATHWAYS

With the advent of equity in human rights in the 1960s and 1970s [48], healthcare organisations began advocating more towards patient-centred care where the patient's participation, autonomy, needs, and collaborative communication between the physician and patient were taken as a high precedence in care. The movement towards this new standard of care was seen primarily in the United States (US), United Kingdom (UK), Europe and Asia. Soon enough, it was adopted as a major theme in healthcare systems around the world [49].

With patient centred care comes the struggle to integrate the delivery and management of healthcare services. This is in line with the UK government's aim to place integrated care at the "heart of the programme for NHS reform" and its importance and urgency are evident when Goodwin et al mention that "Developing integrated care should assume the same priority over the next decade as reducing waiting times had during the last" [50]. For an effective patient centred care approach, it is imperative that health care professionals have good clinical skills and are skilled in communicating with the patient and/or family as well as other multidisciplinary professionals in the continuum of care. Failure to achieve this will result in adverse outcomes and will prove to be a barrier to achieving patient centred care. Hence, an approach that is structured, well planned, evidence-based, and considers the needs of individual patients is required to attain this new standard of care - and this requirement gave birth to Care Pathways (CPs).

Caution, however, should be practiced when taking the evidence-based value of care pathways. As more good quality research is made available to support the validity of care pathways, it is best to utilize them as a means of standardizing and reducing variations in practice, as the evidence-based validity of the outcomes of care pathways has been greatly challenged.

2.3.1 WHAT ARE CARE PATHWAYS?

The concept of care pathways originated in the aviation and construction industries in the 1950s. Known as Critical Pathways, they were additionally used in process management in the 1960s. Later, in the mid-1980s, they were adopted to be used systematically for the first time at the New England

Medical Centre in Boston (USA) for healthcare management using case management plans to serve a financial incentive in response to Diagnostic Related-Groups (DRGs) [51].

The term "Care Pathway" has been used extensively since its conception to mean a multidisciplinary structured care plan used to map the journey of a patient with a particular condition from the time the patient enters the custody of the care givers until the entire continuum of care.

The National Pathways Association (NPA), which supports the development of care pathways in the UK, defines care pathways as an integrated care pathway that "determines locally agreed, multidisciplinary practice based on guidelines and evidence where available, for a specific patient/user group. It forms all or part of the clinical record, documents the care given and facilitates the evaluation of outcomes for continuous quality improvement"[52].

Numerous definitions have emerged in the past one and half decades that highlight the necessity of care pathways to have a multidisciplinary step by step approach in providing quality care. Kinsman et al [53] have identified three major articles that described the characteristics of a clinical pathway: Campbell et al [54] have described care pathways as "structured multidisciplinary care plans which detail essential steps in the care of patients with a specific clinical problem". De Bleser et al [55] characterised the care pathways into 16 subcategories based on how detailed the definition of a clinical pathway is and how it encompasses nouns, characteristics, aims and outcomes in the definition. Vanhaecht et al [56] defines care pathways based on Donabedians' paradigm of structure, process and outcome; where structure represents the context in which care is delivered, process describes the set of actions and services which make up the care, and outcome denotes the result of the services on the patients.

Several synonyms are used interchangeably with care pathways: Integrated Care Pathways, Clinical Pathways, Multidisciplinary pathways of care, Pathways of Care, Care Maps, Collaborative Care Pathways [57], as well as Co-ordinated Care Pathways, Care Maps, or Anticipated Recovery Pathways [54].

Panella et al [58] have outlined that the most prominently used term in Medline is *Clinical Pathway*, whereas, *Critical Pathway* is still the internationally accepted term used in Medical Subheadings [55]. However, it was decided to use 'care pathway' as a generic term to avoid confusions resulting from juggling concepts of *clinical* and *hospital* varying in different European languages like Dutch, Italian,

French and German [58]. De Bleser et al [55], with their extensive review, came up with 84 definitions derived from the 16 sub categories of the care pathways. De Luc [52], found 17 different definitions. In the UK, the most widely accepted term is Integrated Care Pathway [51].

An interesting argument presented by Schrijvers et al [59] demonstrates how each of the different terminologies in use today differ in their meanings and assumptions and how the term 'Care Pathway' takes the lead. In their article, Schrijvers et al support the definition of Care Pathway provided by Vanhaecht et al [56] who state that "The aim of a care pathway is to enhance the quality of care across the continuum by improving risk-adjusted patient outcomes, promoting patient safety, increasing patient satisfaction, and optimizing the use of resources". They continue to argue that a Care Pathway is integrated and not fragmented by nature and should not be explicitly called *Integrated Care Pathway*, nor should it be called a *Clinical Pathway* as it is not only bound to a clinic but covers the continuum of care outside the clinic from outpatient to after discharge care.

2.3.2 THE NEED AND REQUIREMENT OF CARE PATHWAYS

The emergence of care pathways in the UK in the early 1990s is closely related to their development and use in the US. In the US, the need to develop a care plan surfaced as a result of rising health care costs levied by third party insurance companies. The development of pathways, therefore, were seen as "a way of managing high risk, high volume, and high cost patient populations" [60]. In the UK, however, cost was not the driving force behind the development of care pathways, but rather the requisite to improve quality of care by using evidence-based practice [55, 60, 61].

The rationale for developing care pathways in UK emerges from one of the six fundamental principles which the Department of Health (DoH) proposes in their white paper entitled "The New NHS". The paper proposes "to shift the focus onto quality of care so that excellence is guaranteed to all patients, and quality becomes the driving force for decision-making at every level of the service" [62].

Hence, the establishment of the National Institute for Clinical Excellence (NICE) took place to monitor healthcare so it can abide by the principles of good, quality, equitable health for all.

De Luc [52] identified the following reasons for developing care pathways:

- To deliver consistent high quality care
- To reduce variations and risks in practice
- To introduce integrated care across healthcare sectors, disciplines and agencies
- To make care more focused
- To serve as a tool for communication between clinicians and patients, users and carers
- To structure clinical documentation
- To form the basis of benchmarking
- To make planning for resources, training, education and costs more efficient and organised

In patient centred care, the patient's holistic needs are considered and the health services provided cater to the needs of the patient from physiological, psychological, social and financial aspects. Therefore, a well-developed care plan aims to reduce not only the cost, but also promotes a knowledgeable, inter-related clinical environment that promotes the virtues of patient autonomy, individualism, compassion and cooperative decision-making to foster satisfaction, content and improved outcomes.

2.3.3 THE STRENGTHS AND WEAKNESSES OF CARE PATHWAYS

Care pathways aim to improve the quality and continuity of care. They do this by providing continuous unified care and groundwork of communication between multidisciplinary teams that together aim to follow evidence-based guidelines and protocols to achieve enhanced care for the patient. As a result, the pathways offer benefits to the patients, clinicians and the administrators by providing evidence-based information at any given time.

True patient-focused care, as advocated by many healthcare institutions, can only be possible if the patient is truly involved in the continuum of care and all services of care are prepared, skilled and focused towards that concept.

With the introduction of care pathways in healthcare, the patient is elevated from a passive level to a more proactive level where decisions are made by the patient and for the patient.

Campbell et al [54] and Bower et al [63] have identified some benefits of care pathways as:

- Facilitation to the use of local evidence-based protocols in clinical practice
- Encouragement to multidisciplinary communication and care planning
- Promotion to more patient-focused care by promoting patient transparency and expected patient progress
- Reduction of paper work and saving of time
- Support evaluation of care practices through variance management

In the same paper, however, Campbell et al [54] have addressed their concerns regarding the studies included in their paper that were all describing benefits of care pathways. They concluded in the end a publication bias whereby all the studies favored care pathways and as such did not provide reliable evidence [54, 64].

De Luc [52] have outlined the benefits of care pathways from two different angles:

1. Benefits from the process of development of care pathways:

- Streamline the patient's journey
- Encourage multidisciplinary team development
- Promote consistency in practice
- Highlight bottlenecks and duplications
- Clarify roles and responsibilities in the team
- Ensure care is developed from the patient's perspective

2. Benefits from the use of care pathways:

- Provides an outline of anticipated course of treatment
- Integrates evidence-based guidelines
- Retrospective review is made easier by structured documentation
- Provides an aid to daily management of individual patients
- A dynamic tool that can be continuously reviewed and refined

Schrijvers et al [59] compares the benefits of care pathways with the theoretical benefits found in management theories like Critical Path Method, Lean Six Sigma, Business Process Redesign, and the Theory of Constraints. Based on these, the benefits include:

- Shortened production (care) times as a consequence of shortened waiting times between divisions of the same organisation
- The increase in coherence between departments involved in the production (care) process
- Reducing the risk of errors
- Cost reduction of the production (care) process by standardisation
- Clearly defined job roles increase employee job satisfaction

However, the disadvantages which Schrijvers et al [59] discuss include:

- Lack of creativity for employees as the entire process is standardized
- Lack of personal communication between patient and physician
- A reduction of choices for the patients
- Giving more time to patients may compromise quality of care
- Increase in costs of implementing the pathway
- With increased control of errors and defects, patients with poor physical conditions tend to get limited care
- A decreased job satisfaction due to decrease in work variation and passion for the professionals

As is evident from other published articles [50, 51, 65] the benefits of care pathways are multifold. Care pathways are robust tools that do not make the decision for the caregiver, but provide guidelines, evidence, and a route that support good decision-making. They encourage the formation of a multidisciplinary team that acts in unison to provide quality care; they promote the use of research evidence alongside clinical expertise to provide solid underpinning; and last but certainly not least, they put the patient in the forefront of all decisions, services and the provision of information to promote and respect patient-centred care.

Bearing in mind the numerous benefits of care pathways found in the literature, it is extremely important to point out that numerous studies have found several limitations, confounding factors and biases that invalidate the efficacy of the studied care pathways.

Every et al [66] have emphasized that the claim some studies make that care pathways are associated with cost savings is invalid as those studies have not implemented the pathways in a controlled and scientific manner and have not followed a careful study design.

Van Herck et al [67] have also mentioned that a careful methodology in evaluating the benefits of care pathways is extremely important and that many of the studies they searched emphasized on financial, clinical outcomes and process effects and did not cover team and service effects.

Similarly, Rotter et al [68] did a systematic review to critically appraise studies to find that the majority of studies had low quality study designs to evaluate the effectiveness of care pathways. Additionally, they also suggested confounding factors like the introduction of case-mix, hospital policy changes and quality initiatives as introduction of bias in these studies.

Other issues of concern in clinical pathways are not to consider all variations in pathways as something negative and unwarranted. This is due to the fact that patients should be treated as individuals different from one another and a pathway cannot be standardized for each one of them. In fact, as Every et al suggest, the critical pathway is more applicable to ideal patients with uncomplicated illnesses [66].

In another systematic review conducted by Kwan et al [69] regarding the use of clinical pathways for stroke, no evidence was found that care pathways provided additional benefit over standard medical care in relation to clinical outcomes like death or discharge

2.3.4 FORMAT AND DESIGN OF CARE PATHWAYS

The characteristics of care pathways differ from place to place depending on individual goals and aims of the clinical area developing them. However, they should all be "realistic and achievable, not idealistic" [70].

The development of care pathways requires time, commitment, and experience. Care pathways can be tailored to fit a number of different patients and disease conditions with a number of different time frames to suit the needs of the patients, thereby making them flexible enough to be applied to a variety of clinical settings.

A review of the literature revealed a number of similarities in the development of a format and design for care pathways. The defining characteristics of care pathway development include [54, 56, 71-74] (See Figure 13):

- Identification of the topic of interest
- Identification of multidisciplinary team members
- Identification of care pathway scope, aims and objectives, along with the expected outcomes and measurements after a process of consensus
- Attainment of information from literature searches, local guidelines and interviews
- A current process map of care
- Development of the care pathway
- Piloting the care pathway
- On-going review of the care pathway

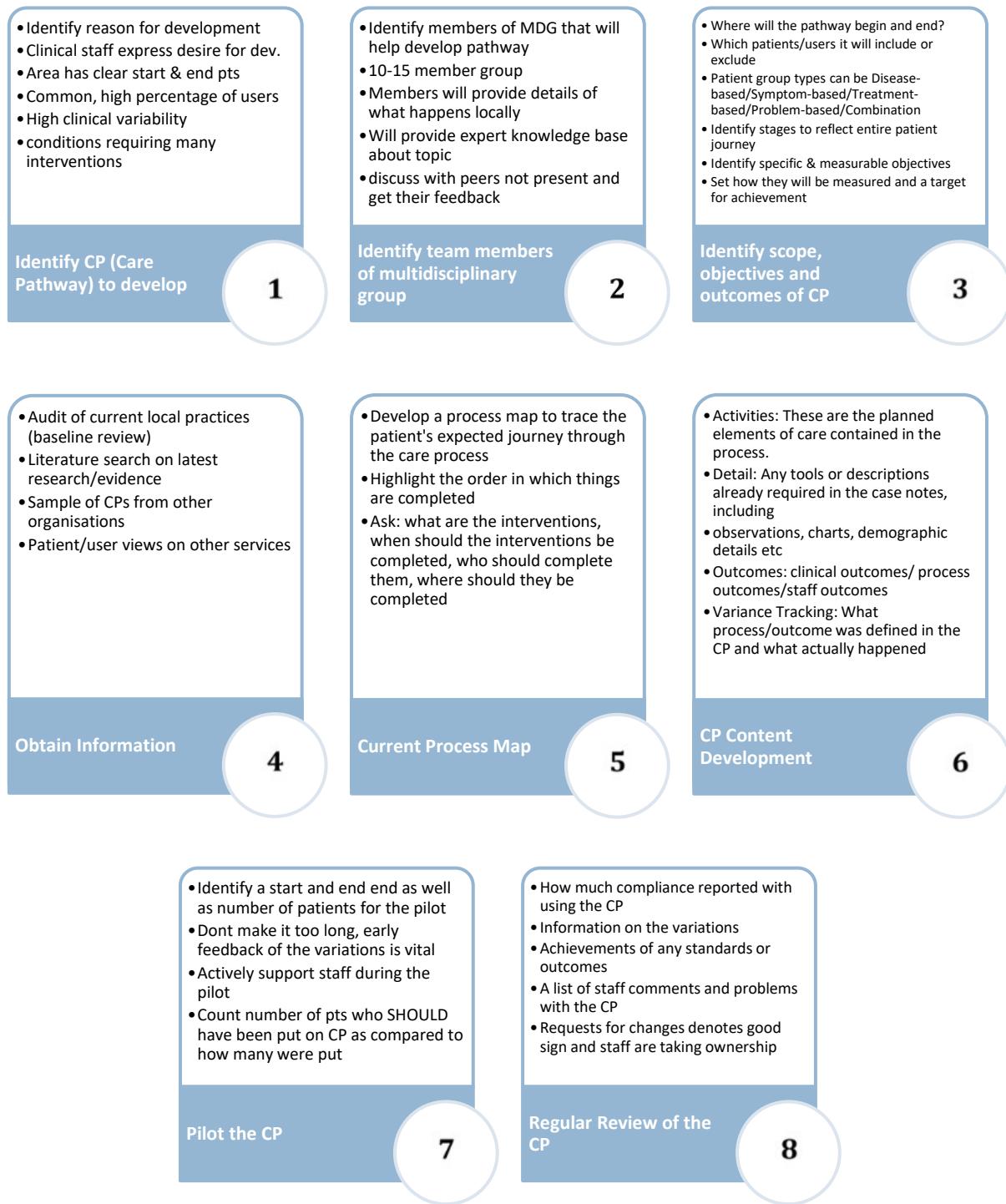


FIGURE 13: CARE PATHWAY DEVELOPMENT PROCESS (ADAPTED FROM: DE LUC [71] AND DAVIS ET AL [73])

2.4 PROSTATE CANCER

Prostate cancer is the most common cancer in men in the UK, with over 40,000 new cases diagnosed every year [75] and with more than 10,000 deaths every year [76]. Although the five-year survival rates of prostate cancer have been improving every year (~82% for 2009 [77]), it is still a major challenging public health issue. Incidence rates for prostate cancer are projected to rise by 12% in the UK between 2014 and 2035, to 233 cases per 100,000 males by 2035.

There is evidence that obvious trends in prostate-cancer mortality could be artefactual. The relationship between the observed national mortality trends and the uptake of PSA testing raises inconsistencies [78]. In the late 1980's, the rapid adoption of PSA screening led to age-adjusted prostate cancer incidence rates doubling in less than a decade. Slowly as PSA screening stabilized and prostate cancer cases in the population got Then, as PSA screening use stabilized and the pool of latent prostate cancer cases in the population was reduced, so did the incidence rates. Currently, both the incidence of early-stage prostate cancer and rates of PSA screening have declined and coincide with 2012 recommendations to omit PSA screening from routine primary care for men. However, more follow-up is required to see if the decline is associated with mortality or not [79].

Older men (above the age of 50) have a higher risk of getting prostate cancer and the risk increases with age. Men who have a family history of prostate cancer as well as men of black African and black Caribbean descent are more at risk [80].

When men suffer from prostate cancer, the usual symptoms are [81]:

- Blood in the urine (although more commonly it is associated with bladder or kidney cancer)
- Lower back pain
- Bone pain
- Weight loss, especially in the older men
- Erectile dysfunction (an inability to get or keep an erection firm enough for sexual activity)

2.4.1 DIAGNOSIS AND REFERRAL OF PROSTATE CANCER

Prostate cancer is conventionally diagnosed using several different measures: Digital Rectal Examination (DRE); the Prostate-Specific Antigen (PSA) blood test; Trans-Rectal Ultrasound (TRUS); TRUS-guided prostate biopsy; as well as imaging techniques like magnetic resonance imaging (MRI), computerised tomography (CT) scan, x-ray, and bone scan. Since each test separately is not sufficient to confirm diagnosis, a combination of these tests is performed to increase accuracy.

Digital Rectal Examination (DRE)

The DRE is the first examination patients having urinary symptoms, prostate problems, or at a risk of having prostate cancer undergo. The test should be performed after the patient's informed consent [82]. The DRE is considered to be the least invasive and simplest of all the examinations for assessing urological patients. For prostate cancer detection, it is considered a mandatory test. The DRE is an examination of the rectum. It assists in finding the extent of the spread of cancer as well as determining its clinical stage. Upon assessment of the prostate, cancer can either be ruled as malignant if the prostate appears to be abnormal and enlarged with nodules; or benign if it's enlarged but smooth and firm[83]. There are, however, some tumors that cannot be felt with a DRE, whereas others can only be detected with a DRE [84]. It is important, thus, to accompany the DRE with other tests to confirm prostate cancer as DRE is not a perfect test.

Prostate-Specific Antigen (PSA) blood test

PSA test is an important test for identifying patients with suspected prostate cancer, and monitoring cancer patients for a relapse. Prostate-specific antigen (PSA) is a protein produced by cells of the prostate gland. During a prostate abnormality, the PSA leaks out of the prostate and into the bloodstream thereby increasing the level of PSA in the blood. If high levels of PSA are found in the blood, it is more likely that the patient will have prostate cancer.

Although PSA test is the most important test for prostate cancer patients, it does come with its pros and cons and should only be administered after thorough counseling. A PSA test offers the benefits of early cancer detection even before the symptoms appear as well as an early stage detection when the cancer is curable. Moreover, it is used for the periodic monitoring of cancer patients for cancer recurrence. PSA testing, however, cannot be used as a precise indication to confirm cancer and further tests like DRE and biopsy are required for confirmation. A high level of PSA in the blood is not

sufficient to validate prostate cancer. Additional factors do exist that can contribute to the elevated levels of PSA in the blood. These factors may include: prostatitis (inflammation of the prostate) as well as benign prostatic hyperplasia (BPH) which is an enlargement of the prostate [83-86]. Other factors that contribute to an elevated PSA include: age, urine infection, vigorous exercise, ejaculation, biopsy, operations or catheter insertions on bladder or prostate gland, and medicines [87].

Some cancers have the characteristics of producing low PSA levels in the blood. In these cases, the tumor can grow large without even being detected. For this reason, a DRE exam is required alongside a PSA test to detect such cancers [84].

The PSA level naturally increases in men as their age increases. The following (Table 5) is a guide to raised PSA levels as provided by the NHS:

TABLE 5: AGE CATEGORISED RAISED PSA LEVEL GUIDES (NHS)

Age	PSA Measurement
50-59	>=3ng/ml
60-69	>=4ng/ml
70+	>=5ng/ml

Trans-Rectal Ultrasound (TRUS)

In trans-rectal ultrasound, sound waves are used to make an image of the prostate which then helps in examining the prostate and determining its accurate size. This test is not reliable enough to rule out prostate cancer and neither should it be used to screen asymptomatic men. Its main purpose is to assist in TRUS-guided biopsy [83],[88].

TRUS-guided prostate biopsy

When prostate cancer is suspected through the results of a DRE or raised PSA value, a confirmation needs to be done through a prostate biopsy where small cores of tissues are removed from the prostate and examined pathologically [89].

The aim of the biopsy is to histologically diagnose prostate cancer. The patient needs to be informed about the advantages and disadvantages of a biopsy before administration.

If the cancer has spread outside the prostate, a biopsy may not always be needed and other imaging techniques like scans can be adopted [90].

2.4.2 PROSTATE CANCER RISK MANAGEMENT PROGRAMME

In the UK there is no national screening programme for prostate cancer. Instead, individuals who wish to gain balanced information regarding prostate cancer PSA testing have an informed choice programme provided through the Prostate Cancer Risk Management Programme (PCRMP). General practitioners have been provided with information packs to counsel patients enquiring about PSA testing [91].

2.4.3 ECONOMICS OF PROSTATE CANCER

Cancer costs the UK economy £15.8bn a year. The total economic cost of prostate cancer is £800 million. Healthcare spending represents a cost of £90 per person in the UK population [92]. Several major studies have provided estimates of the cost of prostate cancer for different countries. These costs are used to explore economic burden of prostate cancer in the treatment and prevention of the disease. In [93], Roehrborn et al. have identified that a high proportion of the cancer costs are incurred in the first year after diagnosis. In 2006, this amounted to 106.7-179.0 million euros in the European countries where these data were available (UK, Germany, France, Italy, Spain and the Netherlands). The variations in costs from one country to another were due to differences in incidence and management practices. Per patient costs depend on cancer stage at diagnosis, survival and choice of treatment. Despite declining mortality rates, costs are expected to rise owing to increased diagnosis, diagnosis at an earlier stage and increased survival. Unless new strategies are devised to increase the efficiency of healthcare provision, the economic burden of prostate cancer will continue to rise.

2.4.4 QUALITY & SAFETY INDICATORS, MINIMUM DATA SET & AUDITING REQUIREMENTS FOR PROSTATE CANCER

Quality control initiatives have largely been seen previously in the business and manufacturing industries. In health care, the Donabedian framework introduced in 1966 has been the basis of most of the modern day research in quality health indicators. According to Miller, the Donabedian quality care appraisal “involves the accumulation of evidence demonstrating that the best available treatment plans were used in optimal fashion”[94].

Donabedian categorised the appraisal of quality into three domains: structure, process, and outcome [95].

Structure refers to factors that influence the context in which care is delivered like the infrastructure and resources (e.g. buildings, equipment, and staff). *Process* measures examine what is actually done for patients (e.g. therapies and procedures used). *Outcome* refers to the results of the processes and therapies after the medical care is given (e.g. health status and satisfaction of patients) [96, 97].

A more contemporary framework for surgery is proposed by Mayer et al [98] (Figure 14) who suggest that a weight should be placed on a number of different quality measures capable of adapting with time and have classified these measures into clinical pathway measures (that incorporate the three measures adapted from Donabedian in addition to a fourth measure of healthcare economics) and patient-reported measures (that include patient related treatment outcomes, health related quality of life, and patient satisfaction).



FIGURE 14: MULTIDIMENSIONAL QUALITY OF CARE MEASURES BASED ON DONABEDIAN FRAMEWORK AND MAYER ET AL ADAPTATION

PROSTATE CANCER QUALITY INDICATORS

Evaluation of the quality of care given to patients with prostate cancer has been very limited due to the absence of valid quality measures. One of the most important limitations of prostate cancer indicators available today are the lack of good evidence that links structures and processes to specific outcomes [96]. Although there have been several advances in the treatment of prostate cancer over the past decade, there have been several controversies regarding the early stage management and treatment of prostate cancer [99]. Due to this fact, several groups have developed quality indicators for prostate cancer. One of these groups is the RAND non-profit organization that has developed quality indicators for early stage prostate cancer. RAND has based its framework on the Donabedian model and has produced these indicators after thorough literature reviews, focus groups with patients as well as interviews with clinicians in the field. These measures have also been professionally validated by expert consensus and scientific evidence. The measurement framework consisted of 49 indicators and 14 covariates for assessment of quality. Table 6 shows an example of a subset of RAND indicators for early stage-prostate cancer [96]:

TABLE 6: SUBSET OF RAND QUALITY INDICATORS FOR EARLY STAGE PROSTATE CANCER [96]

Structure	Process	Outcome
Volume (number) of patients treated	Digital rectal examination, pretreatment clinical stage, total PSA, Gleason grade	Primary treatment failure indicated by three consecutive increasing PSA values after primary treatment by radiation therapy
Board certification of providers	Family history assessment	Patient satisfaction with treatment choice
Availability of radiation therapy services	Documented assessment of comorbidity	10-year overall survival

Danielson et al [95] have used a modified Delphi technique to develop the quality indicators for the process of care in localized prostate cancer Radiation Therapy. The Delphi technique uses a series of questionnaires given in rounds to reach a consensus. Table 7 shows pre-treatment indicators as developed by Danielson.

TABLE 7: DANIELSON ET AL'S PROSTATE CANCER PRE-TREATMENT QUALITY INDICATORS [95]

Prostate Cancer Pre-Treatment Quality Indicators					
Gleason Score	Needle biopsy comments on Gleason grade and score	Risk category	Extra-capsular extension	TNM Stage	CT pelvis for high risk patients
PSA Determination	Opportunity to see other specialist	Assessment of co-morbidities	Discussion of treatment options	Bone scan for high risk patients	
Digital rectal exam	Evaluation of sexual function	Evaluation of bowel function	Evaluation of urinary function	Discussion of complications	

Within the UK, the National Collaborating Centre for Cancer (NCCC) developed the National Institute of Clinical Excellence (NICE) clinical guidelines which provide recommendations about the treatment and care of people with specific diseases and conditions in the NHS in England and Wales [99]. The following NICE guideline (Figure 15) is developed for patients suspected of having prostate cancer [81]:

- Patients presenting with symptoms suggesting prostate cancer should have a digital rectal examination (DRE) and prostate-specific antigen (PSA) test after counseling. Symptoms will be related to the lower urinary tract and may be inflammatory or obstructive.
- Prostate cancer is also a possibility in male patients with any of the following unexplained symptoms:
 - erectile dysfunction
 - haematuria
 - lower back pain
 - bone pain
 - weight loss, especially in the elderly.
 These patients should also be offered a DRE and a PSA test.
- Urinary infection should be excluded before PSA testing, especially in men presenting with lower tract symptoms. The PSA test should be postponed for at least 1 month after treatment of a proven urinary infection.
- If a hard, irregular prostate typical of a prostate carcinoma is felt on rectal examination, then the patient should be referred urgently. The PSA should be measured and the result should accompany the referral. Patients do not need urgent referral if the prostate is simply enlarged and the PSA is in the age-specific reference range.
- In a male patient with or without lower urinary tract symptoms and in whom the prostate is normal on DRE but the age-specific PSA is raised or rising, an urgent referral should be made. In those patients whose clinical state is compromised by other comorbidities, a discussion with the patient or carers and/or a specialist in urological cancer may be more appropriate.
- Symptomatic patients with high PSA levels should be referred urgently. If there is doubt about whether to refer an asymptomatic male with a borderline level of PSA, the PSA test should be repeated after an interval of 1 to 3 months. If the second test indicates that the PSA level is rising, the patient should be referred urgently.

FIGURE 15: NICE GUIDELINE FOR PATIENTS SUSPECTED WITH PROSTATE CANCER [81]

MINIMUM DATASET (MDS) FOR PROSTATE CANCER

A minimum dataset for prostate cancer means a combination of a standard registration dataset with a clinical dataset and contains only the minimum amount of information required that allows for an accurate, timely and accessible health care data. The minimum dataset will be the source of defining the patient pathway for each prostate cancer episode; identifying patient subgroups; managing prostate cancer as well as capturing important outcomes [100, 101]. The following table (Table 8) shows the globally accepted core data elements for prostate cancer diagnosis and referral. The countries reviewed include: UK, US, Canada, Australia, Denmark and New Zealand. The table is based on the new national standard for reporting cancer in NHS England: Cancer Outcomes and Services Dataset (COSD) [102].

TABLE 8: TABLE OF GLOBAL CORE DATA ELEMENTS FOR PROSTATE CANCER DIAGNOSIS AND REFERRAL BASED ON COSD DATA [102]

Parameters	Variables Collected	Country	References
Demographic Data	NHS Number/PIN/National Health Index No.	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Local Patient Identifier	UK, Canada, Australia	[102] [103] [105]
	Patient Record Type	Canada, USA	[103] [108]
	DOB	UK, Canada, Denmark, New Zealand, USA	[102] [103] [104] [106, 107] [108]
	Age at Diagnosis	Denmark, New Zealand, USA	[104] [106, 107] [108]
	Family History of Prostate Cancer		
	Provider Code	UK, Australia	[102] [105]
	Reporting Province/Territory	Canada, Australia	[103] [105]
	Family Name	UK, Canada, Australia, New Zealand	[102] [103] [105] [106, 107]
	Type of current Surname	Canada	[103]
	First Given Name	UK, Canada, Australia, New Zealand	[102] [103] [105] [106,

			107]
	Family Name at birth	UK, Canada	[102] [103]
	Address at Diagnosis	UK, Canada, Australia, New Zealand, USA	[102] [103] [105] [106, 107] [108]
	Postal Code at Diagnosis	UK, Canada, New Zealand	[102] [103] [107]
	Gender	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Province/Territory/Country of Birth	Canada, Denmark, USA	[103] [104] [108]
	Patient's GP	UK	[102]
	GP Code	UK	[102]
	Ethnicity	UK, Australia, New Zealand, USA	[102] [105] [106, 107] [108]
	Marital Status	USA	[108]
Referral Data	Source of referral for outpatients	UK	[102]
	Referral start date	UK, New Zealand	[102] [107]
	Date first seen (by provider)	UK, New Zealand	[102] [106, 107]
	Consultant Code	UK	[102]
	Care professional main specialty code	UK	[102]
	Organisation site code (of provider first seen)	UK, Australia	[102] [105]
	Date first seen (by cancer specialist)	UK	[102]
	Organisation site code (of provider first cancer specialist)	UK	[102]
	Cancer referral patient status	UK, Australia	[102] [105]
	Cancer symptoms first noted date	UK	[102]
Diagnosis Data	Primary Diagnosis (ICD)	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Date of Diagnosis	UK, Canada, Denmark, Australia, New	[102] [103]

CHAPTER 2: BACKGROUND

		Zealand, USA	[104] [105] [106, 107] [108]
	Method of Diagnosis	Canada, USA, New Zealand	[103] [108] [107]
	Laterality	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Basis of Diagnosis	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Morphology/Histology	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Grade	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Stage	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Metastatic Site (Topography)	UK, Canada, Denmark, Australia, New Zealand, USA	[102] [103] [104] [105] [106, 107] [108]
	Behaviour	Canada, Denmark, New Zealand, USA	[103] [104] [106, 107] [108]
	Extent of Diagnosis	New Zealand, Australia	[106, 107] [105]
	Date of recurrence	UK, Australia, New Zealand	[102] [105] [107]
Death Data	Death Date	UK, Canada, Denmark, Australia, New	[102] [103]

		Zealand, USA	[104] [105] [106, 107] [108]
	Death Location Type	UK, Canada	[102] [103]
	Death Cause	UK, Canada, Australia, USA, New Zealand	[102] [103] [105] [107] [108]

AUDITING REQUIREMENTS FOR PROSTATE CANCER

"An audit is the review of clinical care, using objective measures, against explicit criteria for good clinical practice" [109]. The information obtained from audits helps to improve the clinical practice by ways of continuously monitoring the quality of care. The Department of Health has continuously endorsed clinical audits as an important and strategic way to improve, measure, and control the quality of care. Clinical audits have been previously used for physicians to monitor their performance. However, the focus is now shifting on utilising it as a mechanism to improve the quality of care patients receive as a whole.

The National Patient Safety Agency [110] has identified some key issues in cancer diagnosis in primary care. A wide variation in patient referral times was observed (e.g. patients with prostate cancer spent longer times in primary care). The three types of prominent delays identified include patient delays, provider delays and system delays. According to the report issued by the agency, provider delays usually, but not necessarily, refer to delays in primary care. The themes found in both their literature reviews and workshops suggest a misattribution of cancer symptoms to other conditions, a lack of error reporting, as well as a lack of adherence to guidelines with variances. System delays were a result of various communication issues, cancellations, lack of proper delivery of results, lack of communication between GP and hospital, administrative problems.

There are no specific criteria that currently apply to primary care in respect of cancer diagnosis. A number of audits regarding cancer referrals in secondary care have taken place in the past couple of years but have not seen ways of decreasing the referral delays between primary and secondary care [111-113]. The NICE referral guidelines for suspected cancer have come up with suggestions for audit as seen in Table 9 [114]:

TABLE 9: AUDITING SUGGESTIONS FOR SUSPECTED PROSTATE CANCER BY NICE GUIDELINES [114]

Criterion	Exception
1. Patients being referred with suspected cancer are offered a) information about the likely diagnosis, b) what to expect from the specialist service, and c) advice about seeking further help whilst awaiting the specialist consultation.	1 a) Patients who do not want information 1 b) nil 1 c) nil.
2. Patients presenting with classical features of the cancers are a) suspected of having cancer and b) initial investigation or referral is arranged at the first consultation. c) to be set locally; d) to be set locally.	2 a) nil 2 b) patients who refuse referral or investigation 2 c) none 2 d) none.
3. Patients referred for suspected cancer have had preliminary investigations undertaken in primary care as recommended in the guideline.	3) Patients who refuse investigations.

These suggestions, however, "could present operational challenges" [109] as they are all concerned with information about the likely diagnosis of patients with classical features of cancer.

In order to understand why an audit is necessary, it is essential to first understand the pathway a patient follows as well as the roles of the GP and the hospital specialist to fully understand where the gaps and variations in care are most likely to occur.

A patient typically follows the following pathway in primary care [110]:

1. Patient with symptoms or concern decides to seek medical assistance from primary care
2. Appointment made with GP or practice nurse
3. Evaluation of symptoms through the use of guidelines and shared decision making is done
4. Blood and imaging tests are ordered
5. Follow up is given with results
6. Referral is made to secondary care
7. Assessment is done in secondary care

The foremost problem a GP may face while diagnosing patients with prostate cancer would be inaccurate diagnosis due to non-specific symptoms and co-morbidities at presentation. As a GP investigates thousands of patients who do or do not have cancer, it is challenging to pick up an accurate diagnosis of cancer with unusual symptoms. For this reason, missed referrals and inappropriate prioritization of urgency may occur leading to delays in referral and treatment. It is thus important to bear in mind that robust methods of diagnosis as well as more accurate thresholds for referrals and investigations are required to attain improved cancer outcomes.

The following components of cancer diagnosis review and audit have been suggested by Baughan et al [115]:

- Patient diagnosis
- Date patient first noticed symptoms and Date patient first reported symptoms to primary care
- Date of decision to refer and Date referral sent
- Priority given to referral (e.g. emergency, urgent, routine)
- Use of any specific cancer referral pro forma
- Method of sending referral (e.g. electronic, secure fax, post)
- Date patient first seen by specialist
- Date patient was told the diagnosis and Date GP informed of diagnosis
- Reflective comments on patient pathway through primary care
- Audit of cancer diagnosis in primary care

The Healthcare Quality Improvement Partnership (HQIP) programme has put forth a proposal for the specification development for the national prostate cancer audit to be compared against the NICE guidelines for prostate cancer. They have proposed a minimum data set that includes: "ethnicity, socio-economic deprivation, characteristics of cancer, co-morbidities, as well as treatments that were offered and provided". The type of information to be collected in their prospective audit includes [116]:

- Characteristics of new patients, how prostate cancer was detected, and process of referral
- Diagnostic and staging process details
- Planning of initial treatment and Given treatments

- Patient experience and health outcomes (using patient-reported measures to be collected at 6 and 18 months after diagnosis)
- Overall and disease-free survival.

2.5 SUMMARY

This chapter introduced the background concepts relevant to the core theme of this research. It begins by giving a detailed background on process mining which is the main technique used in this study to highlight the actual pathway taken by patients. It then goes on to describe care pathways to better understand how clinical pathways are developed, followed and traced. The background chapter then ends by giving contextual information on prostate cancer that forms the basis of the case study used in this research. The next chapter, Chapter 3: Systematic Review of Process Mining Applications in Healthcare, provides an analysis of research evidence relating to the applicability of the *discovery* aspect of process mining in the healthcare domain and particularly in the study of medical care processes. Chapter 3 forms the essential groundwork necessary to understand where the gaps in literature are in terms of using process mining in healthcare.

CHAPTER 3: SYSTEMATIC REVIEW OF PROCESS MINING APPLICATIONS IN HEALTHCARE

This chapter provides a systematic literature review of ways in which the discovery aspect of process mining is applied to healthcare to analyse and visualize care processes. It is structured as follows: Section 3.2 starts with a detailed background on mining methodologies and visualisation techniques segregated by the process mining perspective. Section 3.3 talks about my primary and sub review questions. Section 3.4 puts forth my aims and objectives with this systematic review. Section 3.5 describes my methods in detail including the eligibility criteria information sources screening and data extraction techniques. Section 3.6 presents the results of my systematic review. Section 3.7 gives a detailed discussion of the results. Section 3.8 talks about new prominent papers that were published after my systematic review. Section 3.9 discusses the strengths and limitations of the systematic review. Section 3.10 presents a conclusion that is drawn from the findings of the review. Chapter 3 is concluded with section 3.11 that gives a summary of the entire chapter and how the following chapter is linked.

3.1 INTRODUCTION

As discussed in chapter 1, the idea of process mining is to discover, monitor and improve real processes by extracting knowledge from event logs. There are three types of process mining techniques: Process discovery, Conformance checking, and Enhancement. Closely related to these techniques are four process mining perspectives: Control-flow, Organisational, Case and Time (or Performance) perspectives.

This systematic literature review provides an analysis of research evidence relating to the applicability of the discovery aspect of process mining in the healthcare domain and particularly in the study of care processes. A PRISMA checklist was used to show where each item in the checklist was used in the systematic review (See Appendix A)

3.1.1 CONTRIBUTIONS OF THIS CHAPTER

In this chapter, I have conducted a systematic literature review and found that the discovery aspect of process mining has been widely applied in several care processes in the healthcare arena. I have

also shown which pre-processing techniques have been commonly used as well as the mining and visualisation techniques used with respect to different process mining perspectives.

3.2 BACKGROUND

Before I present results on the research found in the healthcare literature, the following is a brief background on the mining methodologies and visualisation techniques segregated by the process mining perspectives:

3.2.1 MINING TECHNIQUES FOR THE CONTROL-FLOW PERSPECTIVE

Control-flow discovery aims at the automatic extraction of a process model from an event log, i.e., the inference of a structural representation of the underlying process based on historic data.

Typically, events in these logs are only expected to (i) refer to an activity from the business process, (ii) refer to a case (i.e., process instance), and (iii) be totally ordered. Therefore, the event log can be considered as a set of event sequences. The control-flow perspective of a process establishes the dependencies among its tasks. Which tasks precede which other ones? Are there concurrent tasks? Are there loops? In short, what is the process model that summarizes the flow followed by most/all cases in the log? This information is important because it gives you feedback about how cases are actually being executed in the organization. As the main focus of process discovery is the control-flow perspective, a majority of the studies use this perspective to initially gain an understanding on the process behaviour in their data set. The rest of the perspectives are often enhanced over this fundamental perspective.

Alpha algorithm

The α -algorithm was one of the first process discovery algorithms that could adequately deal with concurrency (explicit causal dependencies and parallel tasks). Given a simple event log it extracts some footprint from the event log and uses this footprint to directly construct a process model (Petri net) that can replay the log [32]. The alpha algorithm uses a step-wise mathematical approach to analyze an event log to mine the process.

Pros: Can mine many processes, simple and robust

Cons: Invisible activities cannot be mined; the resulting Petri Net can hold two nodes that refer to the same task; algorithm cannot deal with non-free choice; and algorithm cannot handle noise (exceptional/infrequent behavior) in the event log.

Visualisation produced: Petri Nets, workflow models

Heuristics miner

The Heuristic Miner was the second process mining algorithm, closely following the alpha algorithm. It was developed by Dr. Ton Weijters, who used a heuristic approach to address many problems with the alpha algorithm, making this algorithm much more suitable in practice (<https://fluxicon.com/blog/2010/10/prom-tips-mining-algorithm/>). Heuristics Miner is a practical applicable mining algorithm that can deal with noise, and can be used to express the main behavior (i.e. not all details and exceptions) registered in an event log. It mines the control-flow perspective of a process model, and considers the order of the events within a case instead of the order of events among cases [117]. It is best to use it when you have real-life data with not too many different events, or when you need a Petri net model for further analysis in ProM.

Pros: Deals with noise; the resulting process model called a Heuristic net can be converted to other types of process models, such as a Petri net for further analysis in ProM

Cons:

Visualisation produced: Heuristic nets

Fuzzy miner

The Fuzzy miner is one of the more recent process discovery algorithms, and was developed by Fluxicon co-founder Christian W. Günther in 2007. It is the first algorithm to directly address the problems of large numbers of activities and highly unstructured behavior. The Fuzzy miner uses significance/correlation metrics to interactively simplify the process model at desired level of abstraction. It is also the first mining algorithm to use a mapping representation to process mining. It is best to use it when you have complex and unstructured log data, or when you want to simplify the model in an interactive manner

Pros: Deals with noise and leaves out unimportant activities or hides them in clusters; seamless process simplification and highlighting of frequent activities and paths; animates the event log on

top of the created model to get a feeling for the dynamic process behavior; faster than any other approach

Cons: The fuzzy models cannot be converted to other types of process modeling languages

Visualisation produced: Fuzzy models

Genetic miner

Genetic Process Mining mimics the process of evolution in biological systems. A genetic search is an example of a very global search strategy because the quality or fitness of a candidate model is calculated by comparing the process model with all traces in the event log. These algorithms start with an initial population of individuals. Every individual is assigned a fitness measure to indicate its quality. Populations evolve by selecting the fittest individuals and generating new individuals using genetic operators such as crossover (combining parts of two or more individuals) and mutation (random modification of an individual) [118]

Pros: robust to noise; duplicate tasks; hidden tasks; non-free choice constructs; and loops

Cons: resource intensive as finding the proper fitness function requires several hours to several days of simulation; Genetic algorithms do not scale well with complexity

Visualisation produced: Genetic models

Pattern mining

Pattern mining is an essential data mining task, with a goal of discovering knowledge in the form of repeated patterns.

Visualisation produced: Frequent pattern models

Declarative miner

The declarative workflow models have been introduced to deal with flexibility in processes. Its rationale is based on the idea of applying some constraints and then tying the execution of some activities to either the enabling, requiring or disabling of other activities. What is not explicitly prohibited by such constraints is implicitly considered legal. Declarative models for workflows are based on taxonomy of constraint templates. Constraints are thus instances of constraint templates, applied to specific activities. A collection of constraints constitute altogether a declarative workflow.

ConDec, which is now renamed to Declare, is the most used language for modeling declarative workflows in the community of Business Process Management [119].

Pros: allows for a higher level of abstraction than imperative process models

Cons: difficult to produce a declarative model for imperative specifications. Requires a different way of thinking

Visualisation produced: Declarative models

3.2.2 MINING TECHNIQUES FOR THE ORGANISATIONAL PERSPECTIVE

In this section, we answer questions regarding the social (or organizational) aspect of an institute/department. The questions are: How many people are involved in a specific case? What is the communication structure and dependencies among people? How many transfers happen from one role to another role? Who are important people in the communication flow? Who subcontracts work to whom? Who work on the same tasks?

Social network miner

The social network miner plugin in PROM reads a process log and generates social networks that can be used as a starting point for Social network analysis. The main idea of this technique is to monitor how individual process instances are routed between actors. We can apply several techniques to analyze the social networks, e.g., find interaction patterns, evaluate the role of an individual in an organization, etc.

Visualisation produced: Social network models, Fruchterman–Reingold layout

Organisational miner

The organizational mining technique enables a better understanding of the functional structure of the organization (based on models) and ultimately improves the underlying processes.

Visualisation produced: Organisational models

Role hierarchy miner

The process mining role hierarchy mining technique enables the investigation of role-related relationships

Visualisation produced: Role hierarchy graphs

3.2.3 MINING WITH THE PERFORMANCE PERSPECTIVE

In this perspective, process mining focuses on performance issues such as flow time, the utilization of performers or execution frequencies. More precisely, we want to enhance the process model with information about execution times and waiting times for the activities. The execution time is the time between the start and the completion of the activity. The waiting time is the time between the point at which the last activity that is a direct predecessor of this activity was completed and the moment at which the execution of the activity itself is started. Moreover, we also want to enhance the process model with probabilities for taking alternative paths, and with information about the case generation scheme.

Performance analyser

Allows an insight into the performance related information of cases. This plug-in focuses on analyzing time-related aspects of the process instances, the average/minimum/maximum throughput time of cases, which paths take too much time on average, how many cases follow these routings, what are the critical sub-paths for these routes, and what is the average service time for each task to name a few.

Visualisation produced: Performance models

Dotted chart analysis

The dotted chart is a chart similar to a Gantt chart. It shows a spread of events of an event log over time. The basic idea of the dotted chart is to plot dots according to the time.

Visualisation produced: Dotted chart

3.2.4 MINING WITH THE CASE PERSPECTIVE

In this perspective, we want to gain more insight into the cases of that process. More precisely, we want to discover data dependencies that influence the routing of a case. To analyze the choices in a process we first need to identify those parts of the model where the process splits into alternative branches, also called decision points. Based on data attributes associated to the events in the log we subsequently want to find rules for following one route or the other. In this perspective we answer questions regarding the execution patterns in the event log. A few questions are: What are the most frequent paths in the process? Are there any loop patterns in the process? What is the distribution of all cases over the different paths through the process? Can I select a subset of traces where particular paths were executed? Can I simplify the log by abstracting the most frequent paths? The case perspective is by default a perspective that most studies follow even if they explicitly don't mention it. This is because the case perspective is not confined to a particular set of algorithms and techniques but rather it is inferred when other techniques are used like clustering, trace alignment, performance, logical temporal language (LTL), dotted chart and decision point analysis to study the cases in a log.

Decision Point Analysis

Decision point analysis aims at deriving decision rules at alternative branching in process models. In a first step, the underlying process model is discovered. If the resulting process model contains decision points, the corresponding decision rules are analyzed using decision trees (data mining).

Pros: able to deal with process models containing both invisible and duplicate activities

Cons: if an alternative branch contains *only* invisible or duplicate activities, it cannot be detected as a decision class

Visualisation produced: Decision tree

3.2.5 THE PROM FRAMEWORK AND DISCO SOFTWARE

ProM is an extensible framework that supports a wide variety of process mining techniques in the form of plug-ins. Some of these plugins go beyond process mining (like doing process verification,

converting between different modeling notations etc.). ProM is platform independent as it is implemented in Java, and can be downloaded free of charge from the www.processmining.org website.

DISCO is a complete process mining software built by former leading academics with more than eight years of process mining experience. The Disco miner is based on the Fuzzy miner, but has been further developed in many ways. The Fuzzy Miner was the first mining algorithm to introduce the "map metaphor" to process mining, including advanced features like seamless process simplification and highlighting of frequent activities and paths. For Disco, they have used the scientifically proven approach of the Fuzzy Miner and combined it with extensive experience from their own practice and user testing. The result is a mining algorithm that, while providing reliable and trustworthy results for data sets of arbitrary complexity, can be operated and understood efficiently by domain experts with no prior experience in process mining [120].

Common approaches to process mining tasks with respect to the different perspectives

Because of the plug-in structure of the PROM framework, there are multiple algorithms that can be used to extract process information from a process log. An example is the following:

Mining algorithms for extracting the control flow of a process:

- Heuristics miner
- Fuzzy Miner
- Mining Organizational-Related Information about a process:
- Social Network Miner
- Organizational Miner
- Role Hierarchy Miner
- Staff Assignment Miner

Checking conformance of the mined process with a predefined model:

- Conformance Checker
- LTL Checker

Mining performance related information about a process:

- Basic Log Statistics plug-in
- Basic Performance Analysis
- Performance Analysis with Petri nets
- Dotted Chart Analysis

The following table (Table 10) shows the common process mining approaches based on tasks and perspectives:

TABLE 10: COMMON MINING APPROACHES TO PROCESS MINING TASKS BASED ON PERSPECTIVES [121]

Perspectives ▶	Control flow	Organisation	Case	Performance
Tasks ▼				
Discovery	<ul style="list-style-type: none"> • Alpha miner • Heuristic miner • fuzzy miner • genetic miner 	Social network miner	Data mining methods	Data mining methods
Conformance Checking	<ul style="list-style-type: none"> • Petri-net based conformance checking • Data mining based evaluation approaches • flexible conformance checking 	LTL checker	LTL checker	LTL checker
Enhancement		Organisational miner	Decision mining	Performance mining with petri net

3.3 REVIEW QUESTIONS

Healthcare is characterized by highly complex and extremely flexible patient care processes known as care flows. Process mining has already been successfully applied in the service industry. I reviewed the applicability of process mining to the healthcare domain by initially dividing my research question into the following 4 components (PICO):

- **P**atient/Person: who does this relate to?
Healthcare
- **I**ntervention (or cause, prognosis): what is the intervention or cause?
Discovery aspect of process mining
- **C**omparison (Is there something to compare the intervention to?)
Other techniques/graphs of giving insight to pathway information
- **O**utcome (What outcome are you interested in?)
An improved insight into the clinical pathway

The main question and sub-question for the review will provide an understanding into the topic addressed.

- *The main review question is:*
“In what ways has the discovery aspect of process mining been applied to healthcare to analyse and visualize care processes?”
- *With the sub question:*
“Which process mining perspective: control-flow perspective, organizational perspective or case perspective is most/least used in the process discovery and analysis of healthcare processes?”

3.4 AIMS AND OBJECTIVES

The aim is to ensure that a systematic and transparent evidence-base is established by following these objectives:

- Review a range of published literature in the use of the discovery aspect of process mining in healthcare
- Report on key process mining perspectives used in analysing the different care processes arising in the literature.
- Identify gaps in literature

- Provide an analysis and commentary

3.5 METHODS

3.5.1 OVERVIEW OF METHODOLOGY

The process used for this literature review is systematic and follows the following phases:

- **Searching**: the systematic identification of potentially relevant studies.
- **Screening**: the application of pre-determined inclusion and exclusion criteria derived from the review question to report titles, abstracts and full texts.
- **Data-extraction**: the in-depth examination of studies, meeting the pre-determined inclusion and exclusion criteria, to assess the quality of the study and extract evidence in support of the in-depth review.
- **Synthesis**: the development of a framework for data analysis and identification of key themes.
- **Reporting and dissemination**: presentation of the review findings.

3.5.2 ELIGIBILITY CRITERIA

The inclusion and exclusion criteria were derived from concepts inherent in both the main review question and the sub-question, and are as follows:

Studies ***included*** that:

- are written in English
- are conducted after 2001 and until March 2016
- draw on published and/or unpublished research
- focus on the use of the discovery aspect of process mining in the field of healthcare particularly to analyse a care process
- focus on one of the following perspectives in process mining: control-flow, organizational, case, or performance

Studies **excluded** that:

- were not written in English
- were conducted before 2001 and after March 2016
- were not based on empirical research
- did not focus on the use of the discovery aspect of process mining in healthcare
- did not focus on any of the perspectives of process mining e.g.: control-flow, organizational, case, or performance.

3.5.3 INFORMATION SOURCES

The following databases (in Table 11) were searched:

TABLE 11: ELECTRONIC DATABASES SEARCHED

Electronic databases
PubMed
IEEE Xplore Digital Library
Web of Science
Ovid (Global Health, HMIC, Medline, Maternity and Infant Care)
Google Scholar
ScienceDirect
Processmining.org list of research papers

3.5.4 SEARCHING

The first step in the searching process was to identify papers, research reports (PhD/MSc), white papers and articles that were broadly related to the use of process mining in healthcare. Using prior knowledge and experience, potentially relevant papers were filtered out using the above electronic database sources. Keywords were derived from the main and sub review question. The keywords were then combined strategically and were tested on a couple of databases as a preliminary literature search to see the number of successful results obtained from each effective combination. Those keywords were then used to further search other databases and refined accordingly to produce the best results. A log of search strings, number of results obtained, and filters used were kept for each database. In order to reduce the number of papers found through the search terms, a

filter was used that only included papers that were conducted after 2001, were written in English and predominantly addressed the use of process mining in healthcare.

3.5.5 SCREENING

The results of each search string were first assessed on screen to see if they met the inclusion and exclusion criteria of the study. The references of the papers that met the criteria were then imported into EndNote to capture the bibliographic information and start the initial screening process on the basis of title and abstract. The references found from other websites were entered manually.

3.5.6 DATA EXTRACTION

A PRISMA flow diagram¹ was used to record the number of studies extracted at each stage of the identification, screening, inclusion and exclusion process. Based on an initial search, the first set of papers were imported. Duplicates were removed using the functionality in EndNote and additionally by manual means when EndNote failed to identify the duplicates due to inconsistencies in the imported information. Screening based on title and abstract was then performed on the remaining papers. When the title was not informative or transparent enough, the abstract was read to help in the screening process. Papers that met one or more of the exclusion criteria were removed and kept in a separate “deleted” folder. A final list of papers was then obtained to be included in the full-text assessment. The full texts for each of the documents were obtained online using Imperial College London institution access privileges. Further studies were excluded after the full texts had been obtained and read.

A data extraction spreadsheet (see Appendix B), comprising of 10 categories to aid in the extraction of relevant information, was designed in MS Excel that supported the process of reporting the findings and analysing the studies. The 10 categories were the following:

1. Bibliographic information

¹ <http://prisma-statement.org/PRISMAStatement/FlowDiagram.aspx>

2. Does the study report as an outcome an impact on the control-flow perspective, organisational perspective, case perspective or time perspective
3. Research Question
4. Study Location (country, setting)
5. Time Frame
6. Healthcare process analysed
7. Data collection methods
8. Data analysis methods
9. Preprocessing techniques
10. Visualisation techniques

3.6 RESULTS

Table 12 summarises the searching and screening process from the various information sources.

TABLE 12: NUMBER OF REFERENCES SEARCHED, KEYWORDS AND FILTERS APPLIED

Electronic databases	Search terms	No. of papers found after applying limits: English language and years 2001 and up
PubMed	Process Mining healthcare	7
IEEE Xplore Digital Library	Process mining healthcare	338
Web of Science	“Process Mining” AND healthcare	34
Ovid (Global Health, HMIC, Medline, Maternity and Infant Care)	Process Mining AND healthcare	16
Google Scholar	“process mining”, “healthcare”, - “data mining”	584

ScienceDirect	"process mining" healthcare	105
Processmining.org list of research papers	Papers in healthcare	397

In the PRISMA flow diagram as seen in Figure 16, initially the total number of references imported was 1481. Screening based on title and abstract was then performed on the remaining 1392 papers.

A final list of 88 papers was then obtained to be included in the full-text assessment. A further 44 studies were excluded after the full texts had been obtained and read.

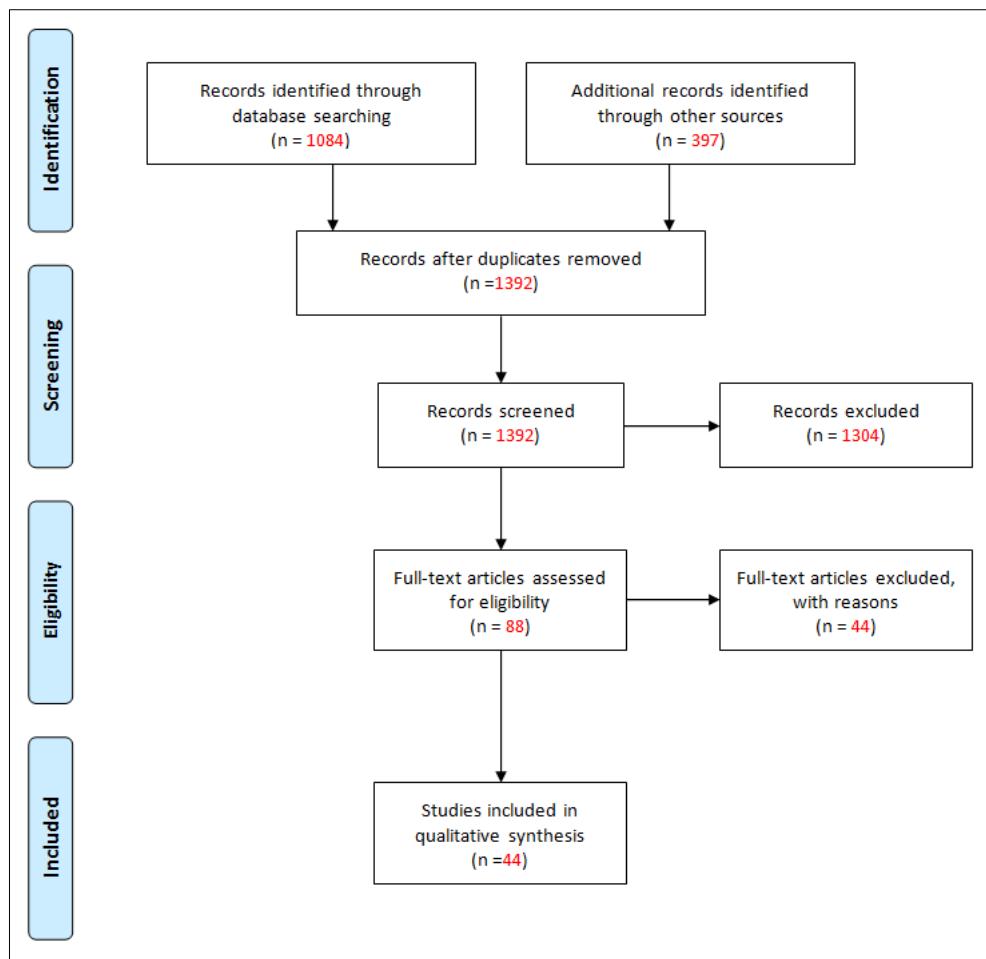


FIGURE 16: PRISMA FLOW DIAGRAM FOR DATA EXTRACTION

From the systematic review of the literature, using the key word search, 44 studies were identified for in-depth study. Countries of origin for all the studies are outlined in Figure 17:

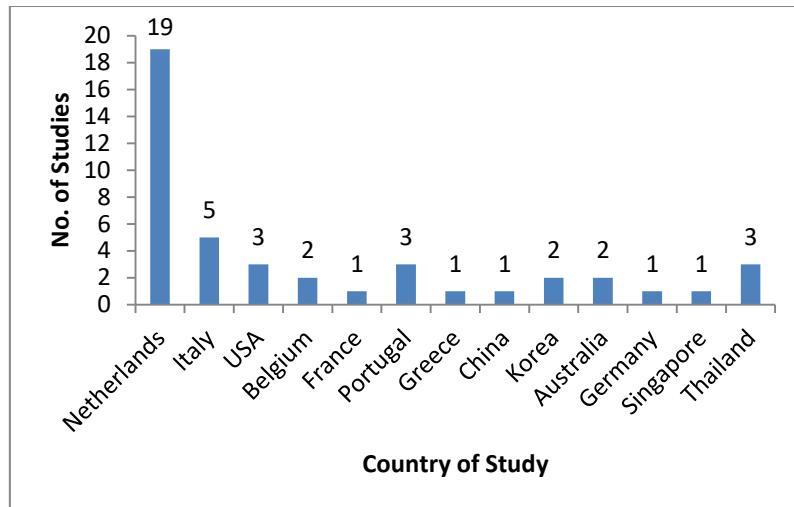


FIGURE 17: STUDIES AND THEIR COUNTRY ORIGINS

The data were synthesised according to four main categories that are related to the underlying concepts of the review question and sub-questions. These categories are:

- Healthcare processes analysed
- Pre-processing techniques used in healthcare
- Mining techniques used in healthcare
- Visualisation techniques used in healthcare

3.6.1 HEALTHCARE PROCESSES ANALYSED

Process mining has been used to diagnose problems in a broad range of clinical processes. Figure 18 shows the number of studies that analysed different healthcare processes:

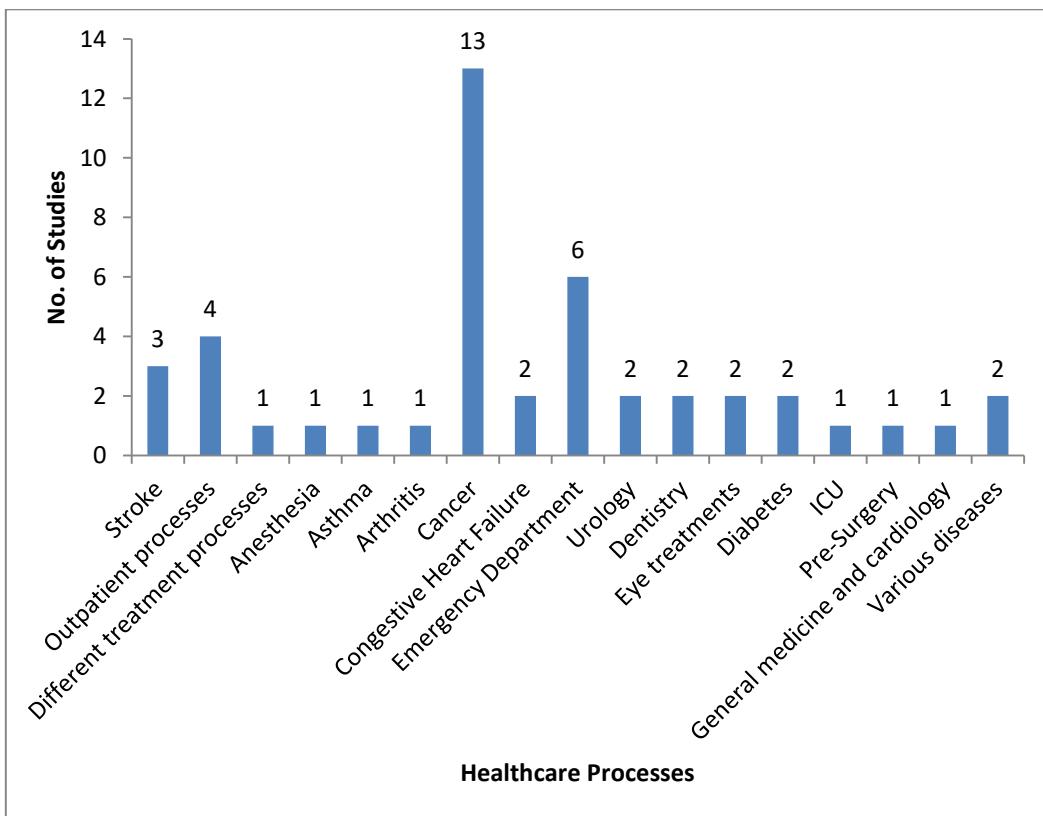


FIGURE 18: HEALTHCARE PROCESSES ANALYSED

It can be seen from Figure 18 that process mining has been used in 17 different care processes. The maximum number of studies has been established in the **cancer** domain followed by the **emergency department** and **outpatient processes** data sets. The following is a detailed description of each healthcare process analysed:

ANESTHESIA

One of the first research studies by Kaymak et al. [122] to study the diagnostic–therapeutic cycle (medical treatment process) was performed in **Netherlands** to assess the applicability of process mining methods to the discovery of clinical process models for the anesthesia procedure during Endoscopic Retrograde Cholangiopancreatography (ERCP). ERCP is a therapeutic and diagnostic

procedure used for diagnosing several conditions of the bile duct and pancreas. As an endoscope is used to look in the stomach and duodenum, the patient needs to be sedated and hence anesthesia is an important factor in this procedure. Since different patients require different levels of sedation based on their vital signs and other relevant observations, the performance of the anesthesia is of vital importance. All relevant information needed for this study was stored in the anesthesia system of the hospital and a variant range of cases were selected to find a process model using process mining techniques.

ARTHRITIS

A Master's Thesis research into the possibility of discovering the care flow of rheumatoid arthritis patients was performed in **Netherlands** by Zhou et al. [123]. In order to improve the quality and efficiency of the rheumatoid arthritis care path followed by the patients, rheumatologists have mapped this pathway and have come up with the most ideal path of a rheumatoid arthritis patient going through the hospital. This study attempts to research the applicability of process mining on acquiring objective process information in the healthcare domain as well as finding out the suitability of using the CRISP-DM data mining framework on a process mining project. Data required for the study was made available through two sources: care registration data (for management purposes) and the CS-EZIS information system which collects all data concerning the patients.

ASTHMA

A research study by Antonelli et al. [124] was conducted in **Italy** to analyse the healthcare network functioning based on an ontology framework. The raw data was collected from an Italian healthcare territorial agency and the diagnostic pathway data was collected and merged in a database through the data warehouse comprising of Hospital Discharge Records, Ambulatory Care Records, and Emergency Department Records. The purpose of this study was to give the medical managers a status of the services under their responsibility as well as suggesting improvements to the system inefficiencies through the use of process mining. The aim of the study was to focus on a specific pathology, like asthma, and analyze the mobility of patients across the different medical centers placed in that territory.

CANCER

There have been a total of 13 studies related to some form of cancer that utilizes the discovery aspect of process mining. Research in the field of ***gynecological oncology*** has taken the lead and a total of 8 papers were found all from ***Netherlands*** reporting on insights in these care flows. Other cancer areas include lung cancer, gastric cancer, breast cancer, colorectal cancer, bladder cancer, intestinal cancer and colon cancer Mans et al. [125] published one of the first studies in gynecological oncology to research the possibility to mine complex and flexible hospital processes giving insight into these processes. The purpose of the study was to identify care paths and collaborations between the different departments. Raw data was collected from the billing information system of the AMC hospital in Amsterdam. The process of gynecological oncology is supported together by several departments mainly gynecology, radiology and different labs.

In another study conducted by Ramos et al. [126] in gynecological oncology, the author proposed a method for healthcare process analysis that could enhance the process-related information needed by these organizations in order to yield process improvements. This master's thesis proposes a validation of an existing method developed by Mans et al.[127] on patients with mamma care and diabetes foot that was designed with the same purpose but in a different healthcare setting.

The PhD thesis of Mans [128] focused on the support of organisational healthcare processes, which capture organisational knowledge, by workflow technology. Workflow technology represents a class of software tools that enable advanced modeling and execution of processes. By applying this technology, a reduction of labor costs and efficient process execution is achieved. The study deals with the gynecological oncology data from the AMC hospital, Netherlands.

Bart [129], evaluate and refine the methodology by Rebuge et al. [20] to perform a quick scan on healthcare process-related data (gynecological oncology) with the purpose of presenting relevant, useful and objective process information for healthcare organizations.

Bose et al. [130] adopted a systematic approach to show that the gynecological oncology processes are in fact rather simple and often sequential and also the cases share a lot in common with very little deviations from the main path.

De Weerdt et al. [131] examine the possibility of intelligent analysis of clinical pathway data based on process mining techniques to deliver valuable insights into the actual carrying out of a gynecological oncology care process in the AMC hospital data in Netherlands. They use versatile data analysis methods in a drill up (general data insights) as well as in a drill down (therapeutic data insights) mode. They benefit from the technique by enhancement of control-flow patterns with other data dimensions.

Caron et al. [132] propose a clinical pathway analysis method for extracting valuable medical and organizational information on past diagnosis-treatment cycles that can be attributed to a specific clinical pathway, namely gynecological oncology. They do this by applying process mining techniques to three areas of healthcare settings: analyzing recurring patterns, pathway variants and exceptional/adverse events.

Mans et al. [19] have demonstrated the applicability of process mining using a real case of a gynecological oncology process in a Dutch hospital and have been able to show that process mining can be used to provide new insights that facilitate the improvement of existing care flows.

Poelmans et al. [133] used a unique combination of process discovery and data discovery techniques to discover process inefficiencies, exceptions and variations of an existing **breast cancer** care process in **Belgium**. Their work also helped the care process manager to gain insight into exactly what's happening on the working floor. The study has shown that neither process mining nor data discovery techniques alone are sufficient for discovering knowledge gaps in domains like healthcare and that a combination of both these techniques is needed.

Huang et al. [134] developed a new process mining approach that not only discovered which critical medical behaviors are performed and in which order, but also provided comprehensive knowledge about quantified temporal orders of medical behaviors in clinical pathways. The proposed approach was evaluated via data-sets extracted from Zhejiang Huzhou Central hospital of **China** with regard to six specific diseases, i.e., **bronchial lung cancer**, **gastric cancer**, cerebral hemorrhage, **breast cancer**, infarction, **and colon cancer**. They believe that in comparison to the traditional process mining techniques, their proposed approach can discover closed clinical pathway patterns that has not been done before.

Mans et al. [135] did a systematic search and came up with questions that are frequently posed by medical professionals in process mining projects. They found out which process mining data can be found in current Hospital Information Systems and does it allow for solving those questions. A data spectrum which classifies the typical event data found in such systems was also produced. They did a case study on **colorectal** patients in **Netherlands** to apply process mining techniques and answer the questions posed by the medical professionals.

In another study, Mans et al. [17] studied the possibilities of process mining within hospitals by presenting a healthcare reference model which lists the typical data that exists within a HIS, and based on this reference model, they presented interesting kinds of process mining analyses that can be performed on intestinal cancer patients. In their case study they selected the services that have been performed for the patients suffering from large **intestine cancer** until the surgical intervention.

De Leoni et al. [136] propose the implementation of a framework for the analysis of the execution of declarative processes on **bladder cancer** patients. They present a novel log preprocessing and conformance checking approach tailored towards declarative models using a log-based alignment techniques. They have shown how alignments provide a very powerful tool when relating observed behavior with modeled behavior.

CONGESTIVE HEART FAILURE

Helmering et al. [137] presented an approach to automate the method of documenting clinical workflows, compared workflows against one another to detect variances and standardize process execution, and optimize clinical workflows to identify workflow efficiencies and inefficiencies. The study, performed on **US** data at the Mercy Health System, showed a significant promise for process mining of clinical processes and rich data sets comprising process activities and related contextual data elements that can be extracted from the EHR and formatted for process mining

Lakshmanan et al. [138] have applied process mining in combination with frequent pattern mining to investigate clinical care pathways correlated with outcomes on traces of congestive heart failure patients, where the traces are first clustered to remove outliers. Their work on trace clustering, frequent pattern mining and overlay of frequent patterns on a mined model has been applied as new features in in a collaborative SaaS environment called BPI.

DENTISTRY

Mans et al. [139] propose a process-oriented methodology for evaluating the impact of IT on a dental process in **Netherlands** ahead of its implementation. In their method, process mining and discrete event simulation are used for obtaining detailed knowledge on a business process execution (through process mining), and building a model which accurately mimics the discovered process that can be used for exploring and evaluating various redesigns (through discrete event simulation).

In another study in dentistry in **Netherlands**, Mans et al. [140] have concentrated on the usefulness of process mining in the “single crown on implants” process and have shown process discovery results based on various perspectives and how the discovery of process behaviour is important even if it involves multiple parties. They have also validated their results by the people involved in the implant process.

DIABETES

Riemers [127], in his master’s thesis, designed a step by step method for the control and diagnosis phases of the BPM life cycle in a healthcare environment (diabetic foot and mammacare processes), in **Netherlands**, using process mining and/or visual analytics. They have concluded that a combination of the process mining and visual analytics approach resulted in visualization options that are more related to activities in the treatment process and the relations found between these activities. Hence, the basics of process mining are visualized by means of visual analytics

Dagliati et al. [141] show how temporal and process mining techniques can be employed together to extract comprehensible and clinically meaningful process models from Type 2 Diabetes patient event logs in an **Italian** hospital information system. The authors identify the best clinical and diagnostic route to follow by applying knowledge discovery methods in this context to improve care processes.

DIFFERENT DISEASES

Rattanavayakorn et al. [142] analyze and investigate the relationships between staff and resources in a hospital using process mining social network miner technique with respect to Working Together

metric. Their study, performed in **Thailand** on different disease processes, can show which doctors performed better in groups, devoting considerable amount of time and responsibility to deal with treatment process of patients' diseases in different sections and situations, and also which doctors were idle not working or interacting with others during the treatment processes.

Krutanard et al. [143] also worked on a dataset in **Thailand** for various diseases to discover a holistic model/graph representing the different role/positions (and structural functions) of the doctors in different levels of an estate governmental hospital in Bangkok. As a result of their work, the hospital administrators could better handle and improve the overall treatment process of the patients and physicians by accelerating the speed of the task performing, increasing the quality of the service work and eliminating the duplicate/redundant tasks during the healing process.

EMERGENCY DEPARTMENT

Alves et al. [144] present an approach that aims at finding communities inside a social network. The approach was implemented in the ProM framework and was demonstrated on the log files of an Emergency Department in **Portugal**. The algorithm was successful in discovering correct communities; demonstrated usefulness of the Modularity concept; derived information about social network and depicted business process based on extracted information from social network

Rebuge et al. [20] studied the radiology workflow process in an emergency department in **Portugal**. Their main goal was to devise a methodology based on process mining that leads to the identification of regular behavior, process variants, and exceptional medical cases. Moreover, the proposed methodology can provide insight into the flow of healthcare processes, their performance, and their adherence to institutional guidelines.

Ferreira et al. [145] use hierarchical clustering together with the concept of modularity to analyze social networks obtained from large event logs of an emergency department in **Portugal**. The clustering of users into communities allows the analysis and visualization of the social network at different levels of abstraction. Based on the analysis of the social network using the working together metric, they discovered which specialists work with other specialists as well as which specialties never work together with others.

Lewis et al. [146] aim to gain insight into patient journey data to identify problems that could cause access block. Pattern analysis combined with statistical data analysis were adapted to discover inpatient flow process patterns and their correlation with patient types, ward types, waiting time and Length of Stay (LOS) in an emergency department in **Australia**.

In Matthew's master's thesis, [147], a theoretical framework is produced to combine process mining with traditional process improvement techniques to create a more robust, data-centric framework for process improvement. Using the emergency department in **USA** as a case study, process mining has proved to be a feasible alternative for implementation with traditional process improvement methodologies. The results from process mining can facilitate process improvement tools and enhance the methods involved in PI projects.

Delias et al. [148] propose a methodology that follows a process mining approach to trace cluster customers' flows and produce effective summarizations. The proposed methodology can address the problem of complex healthcare models by delivering a small number of simpler process maps and hence provide practical usefulness as they enhance the understanding of an actual complex process and provide operational support on how it can be improved.

EYE TREATMENT

Xiaojin et al. [149] propose an approach based on process mining to identify clinical pathways for patients with multiple conditions through historical event logs of patients with similar conditions. The proposed methodology is implemented in a **Singapore** hospital with a case study of patients with eye problems and autoimmune disease. The pathway identified can be utilized as a basis to design patient centered clinical guidelines. The mined pathway illustrates sequence of activities for a patient among all healthcare services across boundaries, thereby eliminating duplicate services.

Overduin [150] has shown in his master's thesis how process mining techniques help in determining the link between the execution of a clinical treatment process and its effectiveness. This research is novel since it is the first process mining case study that focuses specifically on process effectiveness. It shows that by operationalizing process effectiveness based on well-chosen, relevant performance indicators, very useful insights can be gained on the link between the execution of a clinical process and its effectiveness. A case study on the cataract treatment process in **Netherlands** was conducted.

The cataract treatment process was chosen since this process is clinically straightforward, well academically researched, well documented, and the high patient volume makes statistical analysis of process mining results possible.

GENERAL MEDICINE AND CARDIOLOGY

Lewis et al. [151] gain insight into patient journeys from the point of admission until the patient is discharged from the hospital. The paper outlines how unstructured event data were processed to derive the event logs needed as an input for process mining in the absence of a Process Aware Information System (PAIS). Using the processed event data, process mining was then applied for an evidence-based process model discovery of patient journeys from start to end at Flinders Medical Centre (**FMC Australia**).

ICU

Boere's master's [152] analyses and improves an ICU weaning protocol process with the help of data analysis and process mining techniques applied on data from a clinical support system. ICU weaning protocol is used to discontinue mechanical ventilation on patients after cardiac surgery as soon as possible but under the right circumstances for the patient. This study uses process mining techniques to show that automatic generation of patient route improvements options for a medical care process is possible. The author claims that it is the first study to have addressed the application of process mining on a medical treatment process.

OUTPATIENT PROCESSES

Kim et al. [153] applied process mining techniques to discover outpatient care processes in a **Korean** hospital and confirmed that process mining techniques can be useful even in a healthcare environment where a variety of devices and systems are used. Unlike previous studies, this study tested if there is a recognizable difference in practices within hospitals by analyzing the machine-driven process and comparing it to the expert-driven process model. Moreover, frequent process patterns can be utilized for re-assignment of resources, such as hospital staff and departments. In addition, it may also be useful for relocating resources based on distance and department relationships.

Cho et al. [154] have proposed a methodology to analyse outpatient processes in a hospital in **Korea**. The methodology includes data integration, data exploration, data analysis, and discussion steps. They have developed an outpatient processes analysis framework consisting of the above steps. They derived the process model and compared it with the standard model in the hospital. Moreover, they analyzed the process patterns according to patient types and conducted performance analysis and made a simulation model using the analysis results.

Zhichao et al. [155] illustrate a process mining based methodology for healthcare processes management and improvement in an outpatient clinic in Chicago, Illinois, **USA**. This method is able to discover meaningful knowledge of the clinical care processes by mining event logs. The results suggest that this methodology is a useful and flexible tool for healthcare process performance improvement. Based on the results from process mining, a discrete event simulation model is developed to analyze the length of stay for patients and identify the impact of critical resources. Sensitive analysis for different operational scenarios and staffing is also carried out to provide recommendations for clinic management and improvement.

Micilo et al. [156] study the feasibility and advantages of using of Real Time Location System (RTLS) to get a complete and correct log file. The RTLS automatically records events according to patient locations in the service. The log file obtained can contain the pathway tracks followed by the patients, thereby enabling the use of Process Mining in order to make a diagnosis and propose improvements. The case study was done on outpatient processes in a hospital in **France**.

PRE-SURGERY

In Fei et al.'s study [157], a process mining application, with several mining plug-ins, is executed to perform a thoroughly analysis of the care processes. The main idea of this study is to use the discovered model as an objective start point to deploy systems that support the execution of care processes or as a feed-back mechanism to check if the prescribed clinical pathways for a specific set of patients can fit the executed ones. They did their case study on pre-operation activities with a pre-defined set of surgeries in a hospital in Belgium.

STROKE

Mans et al. [158] studied the possibility of applying process mining techniques to clinical data to gain a better understanding of different clinical pathways adopted by different hospitals and for different groups of patients. They observed a difference in treatment strategies between different hospitals, visualized pre-hospitalisation pathways and identified bottlenecks. Two data sets were used: One refers to the clinical course of ischemic stroke patients from their hospital admission to discharge (clinical data set), and the other one refers to the pre-hospital phase (pre-hospital behaviour data set). The case study was conducted and compared between four **Italian** hospitals.

Quaglini [159] proposes the use of process mining techniques to discover not only individuals' error, but also chains of responsibilities and manage the clinical risk. Both supervised and unsupervised process mining will be addressed. The former compares real processes with a known process model (e.g. a clinical practice guideline), while the latter mines processes from rough data, without imposing any model. The case study was done on stroke patients in **Italian** hospitals.

Montani et al. [160] developed a framework to analyse the quality of stroke management processes using process mining and case retrieval techniques, relying on a novel distance measure. This work showed that process mining and case retrieval techniques can be applied successfully to clinical data to gain a better understanding of different medical processes adopted by different hospitals (and for different groups of patients). The case study was done in **Italian** hospitals.

TREATMENT PROCESSES

Jaisook et al. [161] investigate the performance of a private hospital treatment processes in **Thailand** based on event logs. The "Time Performance" of the process instances of the collected event logs from different wards/sections of a hospital were emphasized in order to better visualize and study the behavior of patients referring to the following sections/wards (as well as the hospital's administrators/personnel attending to each case) during the entire treatment processes.

UROLOGY

Rovani et al. [162] report a case study in **Netherlands** that shows how process mining techniques can be used to mediate between event data reflecting the clinical reality and clinical guidelines

describing best-practices in medicine. Declarative models are used as they allow for more flexibility and are more suitable for describing healthcare processes that are highly unpredictable and unstable. The techniques have been applied to a case study on Cryptorchidism in the Urology department. The results demonstrate that the techniques are feasible and that the toolset based on ProM and Declare is indeed able to provide valuable insights related to process conformance.

3.6.2 PRE-PROCESSING TECHNIQUES USED IN HEALTHCARE

A various number of pre-processing techniques were used in the studies. These techniques are outlined in Figure 19:

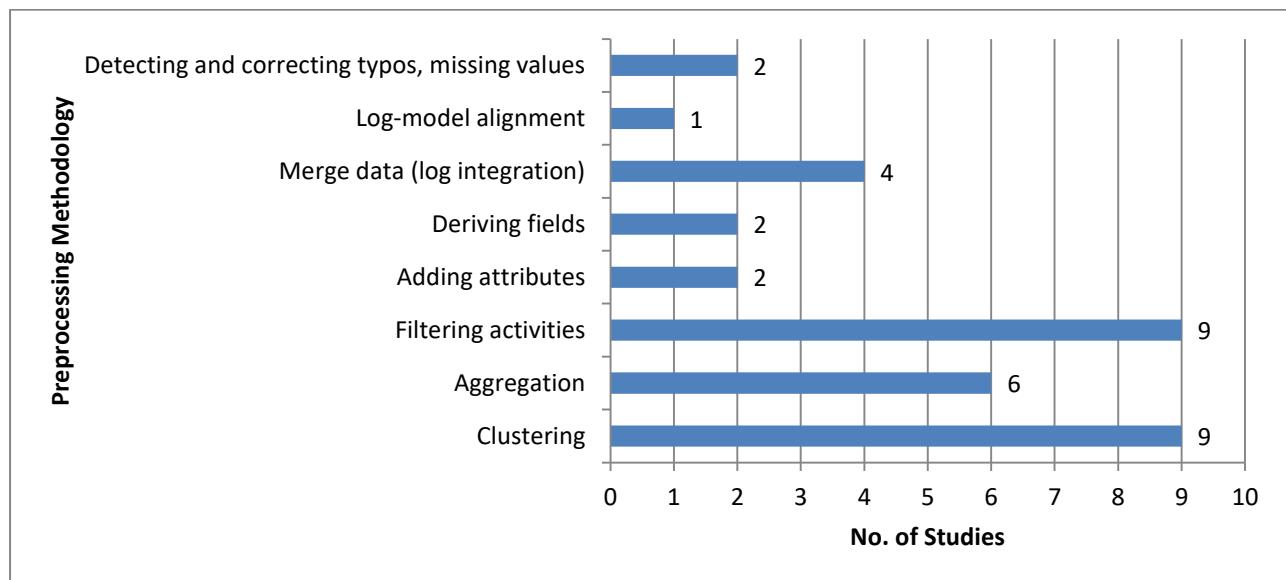


FIGURE 19: PREPROCESSING TECHNIQUES USED IN HEALTHCARE

The most prominent techniques of preprocessing found in the literature include 12 practices: Filtering; Aggregation; Clustering (Trace, Spectral, Sequence); Adding attributes; Deriving fields; Merging data (log integration); Log alignment; and detecting, correcting typos, outliers, and missing values.

Of the studies included in my review, 26 studies have used the crucial step of preprocessing on their data. The maximum number of studies used the clustering and filtering method (9) followed by aggregation (6)

The following is a discussion of the studies and their choice of preprocessing techniques:

CLUSTERING

Bose et al. [130] advocate the preprocessing of the log as an essential step in gaining meaningful insights. They propose a few means of segregating homogenous cases based on different perspectives. First they encode the event log into traces where each trace is the sequence of activities corresponding to a case, they then group similar traces into clusters and visualize these clusters by aligning the traces. In [138], however, Lakshmanan et al. use trace clustering to first eliminate outliers in the patient data, and then use the DBScan clustering algorithm to group process traces together based on an execution footprint. They uniquely represent the sequence of events within a trace as a string where the position of characters within that string represents the temporal occurrence of events within a trace. Lewis et al. [146] have also used clustering to find inliers and outliers.

To overcome the problem that clustering interferes with the amount of noise and ad-hoc behavior present in real-world logs and complicates the models of the generated clusters, Sequence clustering is used by Bart [129] and Rebuge et al. [20]. It is a technique mostly used in bioinformatics to discover the properties of sequences by grouping them into clusters and assigning each sequence to one of those clusters [163]. Unlike trace clustering that extracts features from traces, sequence clustering focuses on the sequential behavior of traces. Bart [129] has followed the following steps for sequence clustering: 1) running the sequence clustering algorithm, 2) building a diagram for cluster analysis, 3) understanding the regular behavior of the process, 4) understanding the process variants and infrequent behavior, 5) performing hierarchical sequence clustering (if needed), and 6) selecting the most interesting clusters for further analysis

In [148], Delias et al. have decided to use the spectral clustering technique as it demonstrated good results in the literature, and gives a recommendation about the number of clusters. They have developed a systematic framework where they first model the process, followed by assessing the degree of variability of the model (low or high) and if the variability was high, a similarity metric was calculated and traces were clustered into groups.

In [145], Ferriera and Alves used a hierarchical clustering technique along with the concept of modularity to analyze social networks. Hierarchical clustering aggregates nodes that are close together, measures the similarity between clusters to decide whether any pair of clusters should be merged together, and then repeats this merging iteratively. The hierarchical clustering approach provides a range of cluster configurations from having a cluster for each individual node to having a single cluster that contains all nodes.

Riemers [127] used manual and automatic clustering techniques on aggregated data.

AGGREGATION

Aggregation is the skipping of unnecessary low level activities and the merging of significant low level activities into singular high level ones.

Mans et al. [125], [19] and [128] simplified the log by: 1) finding an activity that is representative of low level activities, namely that activity is always executed. All other (low level) activities in the log are simply discarded, 2) low level activities in groups without a representative by (1) defining a representative, (2) mapping all activities from the group to this representative and (3) removing repetitions of events from the log. Ramos [126] renamed the activities and grouped events at the level of a visit to a certain department per day and continued increasing/assessing the aggregation with medical professionals until desired patterns were seen in the data. According to the clinicians, this methodology did not show a good level of detail from which it was possible to obtain interesting insights. Moreover, as compared to the clustering techniques, the level of aggregation adopted in this study, if taken solely as a preprocessing technique, would be very time consuming, lengthy, repetitive, and manual. Riemers [127] used aggregation performed on two levels, namely for activities and DBC-codes. Antonelli et al. [124] have applied aggregation to their log file after merging it. De Weerdt et al. [131] have abstracted the data by replacing the bursts of events belonging to the same organizational unit by the name of the organizational unit itself. In this way, a clinical pathway in terms of the unique activities performed by different organizational units is transformed into sequences of departments.

Lakshmanan et al. [138] have removed redundancies in event names by using publicly available medical vocabularies and developing a set of hierarchical category names to replace each individual event name in the medications, labs, and diagnoses event classes.

FILTERING ACTIVITIES

Filtering is the removal of unnecessary activities that offer no substantial benefit to the analysis.

In [126] Ramos has shown how using the MagnaView visual analytics tool, the users can create filters for unnecessary data. Moreover, for process mining filtering activities were done using the *Event Log filter* in the advanced filter options of the ProM tool.

Riemers [127] has performed two types of filtering: (1) manual (by presenting a list of activities to the specialists who in turn choose the most important one) and (2) automatic filtering via the ProM tool. Ferreira et al. [145] excluded the activities performed by other members of the staff, such as nurses and medical imaging personnel; and all process instances (cases) having a single doctor (i.e. doctors working alone) were excluded (filtered) as well.

Kaymak et al. [122] decided to disaggregate part of the data and look at more focused processes. For this purpose, the data set was filtered to consider only monitoring /controlling heart rate and blood pressure versus controlling the oxygen levels in the blood.

Helmering et al. [137] have only filtered out the congestive heart failure (CHF) patients undergoing a radiology procedure, whereas Lakshmanan et al. [138] have segregated positive and negative outcome CHF patients and filtered patients based on predefined criteria.

Zhou [123] has made the preprocessing stage as part of the CRISP-DM methodology and has five distinct steps: select data; clean data; construct data; integrate data; and format data. In the first step, he filters the data to be used in the project. In the second step he cleans the data by adding missing values. In the third step, he constructs the data by deriving attributes and generating records. In the fourth step, he integrates and merges the data from different tables into one table. In the last step, he formats the data to suit the modeling tool to be used. This is the only study using the maximum number of preprocessing steps as part of the data preparation stage.

Jaisook et al. [161] used MS Access database to select and choose those tables and fields of the data that were compatible and needed with the objectives of the study. [157] extracted pre-operation activities with a pre-defined set of surgeries from the medical information system. After this dataset was obtained, it is filtered according to a set of constraints (e.g., importance of the activity to the surgery, the appointment dates etc.).

ADDING ATTRIBUTES

When the event log will not have sufficient required attributes available to commence process mining, the preprocessing step of adding attributes needs to be undertaken. Ramos [126] performed the step of adding minutes to the timestamps, and Riemers [127] added time artificial time stamps.

DERIVING FIELDS

Deriving new fields from the available fields is not a trivial task. Lewis et al. [151] had to derive the date fields by concatenating the separate date and time fields available in the raw data files. Zhou [123] has also used this technique in another study (see filtering technique for more description).

MERGE DATA (LOG INTEGRATION)

This is the process of merging or integrating two different logs together into one based on some common unique identifier in both the logs.

Antonelli et al. [124] merge data of different nature to collect the patient movements inside the network. They do this by creating a controlled vocabulary in terms of set of ontologies to give data a meaning despite the different original data structure. Mans et al. [140] had to resort to manual linkage of two logs as the patient identifier was not the same in each. Lewis et al. [151] also transformed and merged two different sets to construct an event log. [123] has also used this technique in another study (see filtering technique for more description)

LOG-MODEL ALIGNMENT

The log-model alignment concept is based on the principle of creating an alignment of an event log and a process model. Each trace in the event log is related to a possible path in the process model. In [136], de Leoni et al. have adapted this approach to be used in a declarative mining setting where events in the log are mapped to executions of activities in the process model by using the A* algorithm to find the optimal alignment.

DETECTING AND CORRECTING TYPOS, OUTLIERS, AND MISSING VALUES

Although this is a standard technique used implicitly in all the studies, explicitly two papers have mentioned it: In Zhichao et al's paper [155] it was executed automatically in MATLAB. Zhou [123] also used this technique as the cleaning step in their data preparation stage of the CRISP-DM methodology.

3.6.3 PROCESS MINING TECHNIQUES USED IN HEALTHCARE

This section of the report examines evidence related to the review sub-question, which focuses on **“Which process mining perspective: control-flow perspective, organizational perspective or case perspective is most/least used in the process discovery and analysis of healthcare processes?”**

Figure 20 outlines the different mining techniques used in the literature. These techniques have been described in detail earlier in the background section 3.2. Table 13 gives a detailed list of the studies that used the different mining techniques segregated by the mining perspective.

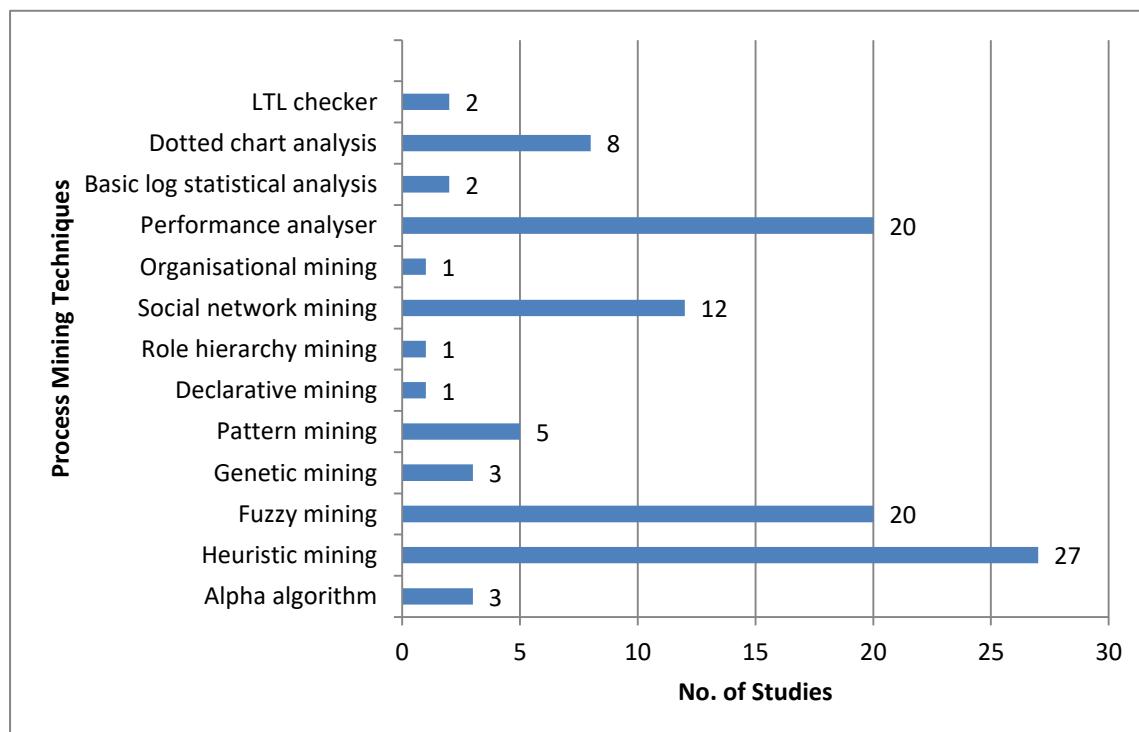


FIGURE 20: PROCESS MINING TECHNIQUES FOUND IN THE REVIEW

TABLE 13:: COMPLETE LIST OF STUDIES WITH THEIR MINING TECHNIQUES USED WITH RESPECT TO MINING PERSPECTIVES

Studies	Control flow perspective				Organisational Perspective			Performance Perspective		Case Perspective			
	Alpha	Heuristics	Fuzzy	Genetic	Pattern	Declare	Role hierarchy	Social Network	Organisational	Performance Analyser	Basic log Statistical Analysis	Dotted chart Analysis	LTL checker
[125]	X	X						X		X			
[158]		X								X			
[123]	X	X								X			
[126]	X	X	X					X		X			
[127]		X								X			
[159]		X											
[133]													
[144]										X			
[157]		X		X					X		X		
[128]	X	X	X						X		X		
[20]		X								X			
[130]				X									
[140]		X								X			

Studies	Control flow perspective					Organisational Perspective			Performance Perspective		Case Perspective		
	Alpha	Heuristics	Fuzzy	Genetic	Pattern	Declare	Role hierarchy	Social Network	Organisational	Performance Analyser	Dotted chart Analysis	Basic log Statistics	LTL checker
[134]													
[145]													
[122]													
[149]													
[129]													
[137]	X	X	X	X						X			
[146]							X						
[132]													
[139]											X		
[136]										X			
[152]		X	X	X							X		
[153]			X	X									
[150]												X	

Studies	Control flow perspective					Organisational Perspective			Performance Perspective		Case Perspective		
	Alpha	Heuristics	Fuzzy	Genetic	Pattern	Declare	Role hierarchy	Social Network	Organisational	Performance Analyser	Dotted chart Analysis	Basic log Statistical Analysis	LTL checker
[131]			X										
[160]			X						X				
[135]			X										
[17]			X						X		X		
[147]			X						X	X	X		
[138]			X				X						
[154]			X				X				X		
[151]			X								X		
[155]			X									X	
[141]					X								
[142]						X							
[124]									X				
[162]													

Studies	Control flow perspective				Organisational Perspective			Performance Perspective		Case Perspective		
	Alpha	Heuristics	Fuzzy	Genetic	Pattern	Declare	Role hierarchy	Social Network	Organisational	Performance Analyser	Dotted chart Analysis	LTL checker
[143]		X						X				
[19]			X								X	
[156]				X								
[148]					X							
[161]						X						

Visualisation techniques used in healthcare

1. Heuristic model
2. Fuzzy model
3. Social network model
4. Performance model
5. Organisational model
6. Alpha algorithm model
7. Genetic model
8. Petri net model
9. Statistical charts
10. Self-organizing maps (SOM)
11. Declare model
12. Concept lattices
13. Dotted chart
14. Trace model alignment graphs
15. Frequent pattern model
16. Comp model
17. Role hierarchy graph
18. Workflow patterns
19. Fruchterman–Reingold layout

There were 10 prominent papers that used process mining covering all the four perspectives of **control-flow, organisational, case and performance:**

- Mans et al. [125], Ramos [126], Fei et al. [157], Mans [128], Rebuge [20], Overduin [150] and Matthews [147] all used **fuzzy** miner and/or **heuristic** miner for the control flow; **social** network for the organisational perspective; and **performance** analysis for the performance perspective. Ramos and Fei, however, also used the **genetic** miner for comparison purposes, while Matthews used the **organisational** miner as well for the organisational perspective along with the **dotted** chart to show overall events and performance information of the log.
- Mans et al. [19] and Bart [129] also used the **fuzzy** miner and/or **heuristic** miner for the control flow, but they used the **dotted** chart instead of the performance analysis along with the **social** network miner for organisational information.
- Krutanard et al. [143] used **fuzzy** miner and **role hierarchy miner** along with **performance** analysis to discover a holistic model/graph representing the different role/positions (and structural functions) of the doctors in different levels.

Studies covering three perspectives like the **control-flow, performance and case** included 11 papers:

- Zhou [123] and Mans et al. [17] used **Heuristics** miner with **fuzzy** miner for control flow and combined it with **Performance** analysis and **dotted** chart analysis.
- Boere [152] and Cho et al. [154] used **heuristic** miner with **fuzzy** miner along with **performance** analysis only. [154] added an additional **Pattern** mining to discover the control flow.
- Lewis et al. [151], Zhichao et al. [155] and Jaisook et al. [161] all used **alpha/heuristic/fuzzy** mining with basic log **statistics** for performance information.
- Other prominent studies included Riemers [127] and Mans et al. [158] [**Heuristic** miner + **performance**]; Mans et al. [135] [**Heuristic + Dotted** chart]; and Mans et al.[139] who used **alpha** algorithm, **performance** analysis, **dotted** chart analysis and discrete event simulation.

The only study which combined the **control-flow, organisational and performance** perspectives was the study conducted by:

Mans et al. [140] who used **Heuristic** miner, **social network** and **performance** analysis to mine their data

Many studies also concentrated on just two perspectives like the **control flow and performance** (5 studies) and included papers like:

- Poelmans et al. [133] who use a unique combination of process discovery techniques and data discovery techniques. They use a **statistical** approach using hidden markov models to model the workflow. They believe that these probabilistic models offer a greater degree of flexibility and are a better option for healthcare, where traditional process mining techniques do not work well. For data discovery they use formal concept analysis (FCA). It is a data analysis technique that supports the user in analyzing the data and discovering unknown dependencies between data elements;
- Helmering et al. [137] who used **alpha** miner, **heuristics** miner, **genetic** miner, and **performance** sequence diagram
- Lewis et al. [146] who used Heuristic miner, pattern mining along with basic log statistical analysis to extract the performance information
- De Leoni et al. [136] have generated Different Declare models, through the **Declare Miner** plug-in in ProM with different numbers of activities (5, 10, 15 and 20) and nearly two constraints per activity. The types of constraints were precedence, responded existence and response
- Montani et al. [160] who used only the **heuristic** miner and **performance** analysis

Studies also had a combination of **Control flow and case perspectives** (10 studies) and included papers from:

- Bose et al. [130], De Weerdt et al. [131], Delias et al. [148], and Antonelli et al. [124] all used **fuzzy** miner
- Huang et al. [134] used a novel approach of mining closed clinical pathway **patterns** from clinical workflow logs that regularly record medical behaviors in patient-care journeys. The approach is based on sequence pattern mining algorithms which first mine clinical activity sequences at first, and then (2) mine chronicles on the sequences to generate closed clinical pathway patterns.
- Kaymak et al. [122] and Caron et al. [132] decided to use the **heuristic** mining approach based on literature reviews and professional advice. Kaymak et al. [122], however, realized that the models produced were too complex to understand. They concluded that for more successful applications of process mining in the health care, the methods need to incorporate medical knowledge into model search and preprocessing so that the search space is made smaller, enabling the algorithms to discover optimal models that describe the process.
- Kim et al. [153] used **heuristic, fuzzy** and **comp, pattern** mining
- Montani et al. [160] used the **fuzzy** miner
- Lakshmanan et al. [138] used frequent **pattern** mining and **heuristics** mining
- Dagliati et al. [141] used **heuristic** mining

Only one study using a combination of the two perspectives **Control-flow and organisational** was found by:

- Rattanavayakorn et al.[142] using the **fuzzy** miner with **social** network miner

Finally, studies involving only one perspective include:

Control-flow (4 studies) –

- Quaglini [159] and Xiaojin et al. [149] used the **heuristic** miner for control-flow discovery
- Rovani et al. [162] used the **declare** miner
- Micilo et al. [156] used RTLS to get a complete and correct log file and then used **fuzzy** mining

Organisational (2 studies) –

- Alves [144] and Ferreira et al. [145] used **social** network

3.7 DISCUSSION

I start off the discussion by examining the evidence that has emerged from the literature related to the main review question: “***In what ways has the discovery aspect of process mining been applied to healthcare to analyse and visualize care processes?***”

Four categories have been identified: Healthcare processes analysed, Pre-processing techniques used in healthcare, mining methodologies and visualisation techniques used in healthcare.

3.7.1 HEALTHCARE PROCESSES ANALYSED

Process mining has been used to diagnose problems in a broad range of clinical processes. Research into the different healthcare processes analysed through the discovery aspect of process mining has resulted in 17 different care processes. The maximum number of studies has been established in the ***cancer*** domain followed by the ***emergency department*** and ***outpatient processes*** data sets.

Due to the complexity of care and requirements of cancer, there is an urgent need to improve the cost and clinical effectiveness of cancer care pathways. A reason why the majority of the studies were oncology-based (cancer) is because Process mining offers the opportunity to develop deeper understanding of this complexity and help improve cancer care pathways and outcomes for cancer patients [164].

With increasing demand for medical services, emergency departments are facing problems such as overcrowding and dissatisfaction. Improving the processes of emergency departments has been the focal point of healthcare management [165] and thus the use of process mining helps in exploring these processes and improving services.

Outpatient care services are the gateways to providing the best, fastest and most cost-effective service to patients. It is a multi-step process approach that includes reception, consultation, treatment, test, and payment. Errors in the process become a key cause of discomfort for patients, as well as a factor that may interfere with their treatment. Because of the complexities in the nature of the patient’s condition and the individuality of services the a patient should receive, there is a definite need to discover and provide effective hospital processes to patients to reduce the time and cost and to provide patients with high-quality service [153]. Process mining provides one of the

objective solutions to uncover the journey of the patient in the system and hence studies resort to using this as a major alternative to traditional methods.

3.7.2 PREPROCESSING TECHNIQUES USED

Before any technique of process mining is applied on data, most of the available data from the information systems/workflow management systems need to go through a crucial and time consuming activity called **Preprocessing**. Preprocessing is the ‘the identification of appropriate data and then the process of pre-processing the data to derive the event log’ [151]. The data that is made available needs to be preprocessed to obtain meaningful models and to prepare the data to suit the data requirements of process mining. Preprocessing of the log is concerned with extracting the log, focusing only on the events that occur most frequently; focusing on a specific patient group; filtering unnecessary low level activities; simplifying or aggregating significant low level activities; and formatting the data to suit the requirements amongst many other ways. The techniques most commonly used for preprocessing, as per the literature, were filtering and clustering.

To address the problems associated with unstructured and diverse event logs, an approach is used where the “event log is clustered iteratively such that each of the resulting clusters corresponds to a coherent set of cases that can be adequately represented by a process model” [166]

To identify groups of traces that behaviour in a similar way, we apply the process of trace clustering. Trace clustering is similar to regular clustering except that the input is an event log which effectively is a collection of traces. Every trace in the event log is an object in the clustering task. By clustering an event log we can obtain more homogeneous groups of traces that can be analyzed independently from one another, which improves the quality of mining results for flexible environments. Since we want to cluster all the traces that are behaviorally similar, we use an algorithm to find the distance measure between traces [167]. Clustering is one of the most prominent preprocessing methodologies used in complex processes as it segregates the event log into more manageable and similar groups making it easier to analyse.

Filtering, as clearly seen in the literature, was also one of the most prominent methodologies used. Filtering is one of the basic ways of achieving an event log that is clean and free of unnecessary activities and events that contribute to the cluttering of the event log and bear no importance to the

data studied. Filtering can be done in a two ways: either manually (if the event log is not complex and large); otherwise using automatic techniques found in the process mining software.

3.7.3 MINING AND VISUALISATION TECHNIQUES USED WITH RESPECT TO THE DIFFERENT PROCESS MINING PERSPECTIVES

In the last few years, many different process mining approaches have been developed that deal with different perspectives of a given process. These approaches vary in the way they are used and the way they deal with the different challenges associated with event logs.

As discovered in the literature, for the control flow perspective, the maximum number of studies used the Heuristic mining approach (27 studies), followed by Fuzzy mining (20 studies) and then Pattern mining (5 studies). This is not surprising as the Heuristics Miner is a very practical mining algorithm that can deal with noise. The alpha algorithm that preceded the heuristics miner required the mined log to be complete and free of any noise. However, this was not practically possible. Hence, with the arrival of the heuristics miner, the sensitivity regarding noise and incompleteness was eliminated and the miner was readily accepted and used for real-life data [168]. Similarly, with the advent of the fuzzy miner, a large amount of confusing behavior which might have resulted with the heuristic miner was cleaned, and the models were mined faster and more reliably.

Only 40% (18 papers) used the organisational perspective as part of their research to study the event log. This was the only perspective to be least used and researched. The organisational perspective has a very important role as it evaluates the relations between people, teams and departments in the entire organisation [169]. By using this perspective, organisations can improve the flow of communication and provide the higher administration a clear view of how the work is handled and shared between departments. For the Organisational perspective, the maximum studies utilized the Social network miner (12 studies).

As for the performance perspective, 66% (29 papers) used it as part of their research to study the event log. This is one of the most promising perspectives as it gives the ability to “analyze the actual run-time behaviour of processes and obtain precise information about their performance in near real-time” [170]. The studies mostly used the performance analyser (20 studies), which is the most common way of performance mining, and the dotted chart analyser (8 studies).

Finally, most of the studies (31 papers) effectively used the case perspective, unless the study explicitly concentrated on the organisational aspects of the event log. This is because it is assumed when a study is doing the control flow or performance perspectives; they are already dealing with the case perspective since they are closely looking at the case-by-case scenario in their mining methodology. Hence, nearly all the studies incorporated this perspective even though they explicitly did not mention it.

The most prominent mining techniques for healthcare processes found in the literature are a combination of different techniques divided according to the process mining perspective. The most effective way to mine datasets was seen to be a holistic approach covering three or more perspectives to come up with a process model that can be used in effective analysis and communication. The most popular combination of perspectives was *control-flow with performance and case*. The least explored combination of two-perspective studies was the *control-flow along with the organisational perspective* and the perspective combination most commonly used was the *control-flow with case*. From amongst the studies concentrating only on one perspective in the study, the most common trend was to see the *control-flow* of the process model.

3.8 PROMINENT PUBLICATIONS AFTER THE SYSTEMATIC REVIEW

As my systematic review only included papers until March 2016, there were several noteworthy papers that were published immediately after this review that are worth mentioning in this section.

1. Process Mining in Healthcare: A literature review (Rojas et al.)

This review covers 74 papers all of which were analysed according to 11 main aspects: process and data types; frequently posed questions; process mining techniques, perspectives and tools; methodologies; implementation and analysis strategies; geographical analysis; and medical fields. The most commonly used categories and emerging topics have been identified, as well as future trends, such as enhancing Hospital Information Systems to become process-aware [171].

2. Process Mining for Healthcare Process Analytics (Erdoğan T et al.)

This review gives an overview of studies on process mining applications in the healthcare domain. It gives a generic classification scheme with respect to the attributes of research, contribution type, application context, process modelling type, modelling

notation/language, process mining algorithm type, and benefits. Moreover it gives a categorization of conformance verification studies with respect to the classification scheme. Finally it gives a description of the features of a tool, which uses the process mining technique for conformance verification, to carry out healthcare process analytics.[172]

3. Process Mining in Oncology: A Literature Review (Kurniati et al.)

This is the first systematic literature review that supports the use of process mining in oncology. It highlights the potential value of process mining for improving cancer care processes. It provides a useful overview of the current work undertaken in oncology using process mining. It helps researchers to choose process mining algorithms, techniques, tools, methodologies and approaches. Finally, it identifies research opportunities in this new field of study [164].

4. Towards Process Mining of EMR Data - Case Study for Sepsis Management (Vries et al.)

This paper provides insight into the steps required to perform process mining to EMR data in the domain of sepsis treatment. Process mining was used to follow selected events derived from the sepsis pathway purely from EMR data. Using process mining techniques, the authors analysed beyond the mere presence or absence of events and also addressed correct versus incorrect order with respect to a model that represented best practice. [173]

3.9 STRENGTHS AND LIMITATIONS OF THE SYSTEMATIC REVIEW

Originality/value – The contribution of this review is to provide a summary of the current trends in the use of process-mining in the healthcare area. A review of the work in this new and expanding area has been provided that identifies prominent papers that deal with the process discovery aspect of process mining from a multi-perspective point of view. This is the most comprehensive and up-to-date review covering 44 full text published papers describing the mining methodologies and visualisation techniques used in discovering process models in a healthcare setting.

Limitations: I have not reviewed the conformance and enhancement aspects of process mining in this literature review and have only focused on the discovery aspect.

3.10 CONCLUSIONS

It is evident from the literature that there is significant promise for process mining of complex and flexible clinical processes giving insight into these processes. In these flexible environments, the techniques that are most beneficial are the ones that can deal with large amounts of noise. Clearly, in the literature, the heuristics miner was the most popular to be used for discovering a process model. This is largely due to the fact that it deals better with noise than its predecessors. However, towards the mid 2013's, the use of the fuzzy miner became increasingly popular as the most outstanding advantage of this miner was its interactive and explorative nature and the ability to look at all aspects of the process at once. As one of the major drawbacks of a healthcare event log is that it contains many distinct, low level activities that produce a spaghetti-like process model, by including a preprocessing step, like the popular clustering and filtering techniques, before deriving the process model helps in reducing the amount of activities to an amount that is more manageable and relevant. This increases the process model's readability and comprehensibility. Additionally, to best benefit from the true potential of process mining, it is better to take on a multi-perspective approach to have a more holistic idea about the processes in a clinical system. With this review, I have concluded that the discovery aspect of process mining has been widely applied in the healthcare area, particularly in the Cancer and Emergency department care processes. In the future, more emphasis should be placed on the organisational aspect of process mining, as fewer studies were found contributing in this area. Since processes emerge because of human decision making, it would be interesting to see more research on the working behaviour of physicians and the precise input each physician contributes during the treatment process.

3.11 SUMMARY

This chapter conducted a systematic review to find published literature pertaining to the use of the discovery aspect of process mining in healthcare. The review gave an overview of the various algorithms, pre-processing and mining techniques used in healthcare. It also found gaps in literature relating to the different process mining perspectives that have been minimally used in healthcare. The next chapter, Chapter 4: Process Model Construction, initiates the first two phases of the CPAM roadmap and is the starting point of constructing an event log that will lead to process mining.

CHAPTER 4: PROCESS MODEL CONSTRUCTION

This chapter begins the first two phases of the CPAM roadmap (Figure 21). It talks about how I constructed the process model through various data integration, extraction and preparation steps. Section 4.2 starts with a background discussion on concepts like submission of secondary care data at the local and national levels; construction of event logs; record linkage techniques; and log preparation and pre-processing methodologies. Sections 4.3 – 4.5 describe the event log linkage and pre-processing phases and the exact methodology followed in identifying clinical information systems; linking various data sources; and constructing and pre-processing the event log to make it suitable for process mining and initial log inspection. Section 4.6 presents a discussion about issues and lessons learnt from this construction phase. Chapter 4 is concluded with section 4.7 that gives a summary of the entire chapter and how the following chapter is linked.

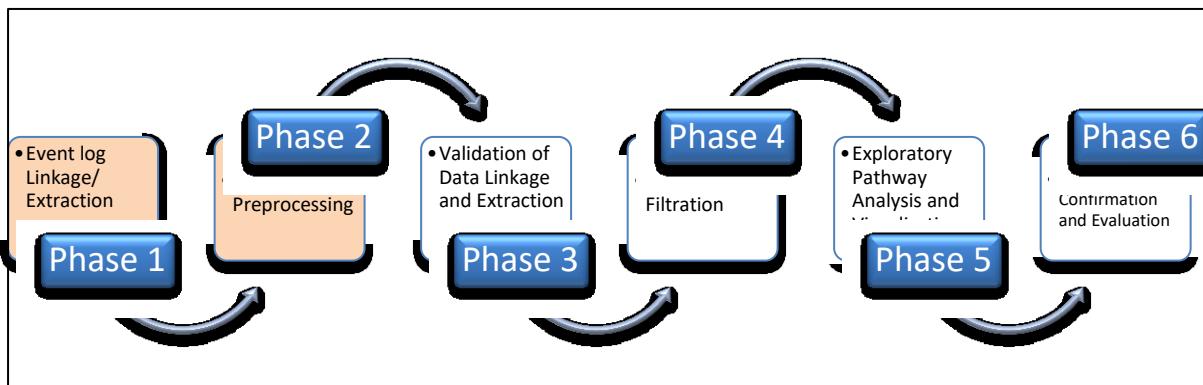


FIGURE 21: PHASES 1 (EVENT LOG LINKAGE) AND 2 (EVENT LOG PREPROCESSING) OF THE CPAM ROADMAP

4.1 INTRODUCTION

As we have seen earlier in the introduction chapter, Process mining heavily relies on event data. Although there are several powerful techniques of conducting process mining in various domains, a limiting factor is the preparation of event data [174]. In many occasions, the available data does not exist in an event log format and needs to be created from existing information systems. To do so, a lot of domain knowledge is required so that we select the right data [175]. Most information systems do not record events explicitly. Only process-aware information systems (e.g., BPM/WFM systems) record event data. To create an event log, we often need to gather and link data from

different data sources where events exist only implicitly and need to be extracted usually from conventional databases. Although the underlying relational databases are rich with data, there are no explicit references to events, cases, and activities. Instead, there are tables containing records and these tables are connected through key relationships. Tools such as XESame and ProM import provide some support, but the event logs still need to be constructed properly by querying the database and converting database records into events [174].

4.1.1 CONTRIBUTIONS OF THIS CHAPTER

In this chapter I have constructed the event log by linking 6 databases relevant to prostate cancer using linkage algorithms developed in TSQL and ASP script. After the data linkage, I then extracted and transposed the required data to be converted into a format that can be used in a process mining software. Following this, the data was cleaned and pre-processed to include only the data that was required for the analysis. The algorithms and underlying programming code necessary for the linkage, extraction and transposition was all developed from scratch and is scalable to any size and type of care process studied.

4.2 BACKGROUND

Before I discuss the methodology of how I constructed my event log and process models, it is necessary to understand how the current system works at both the local and national level to gain a better understanding of how to extract the data from the different data sources involved, and how event logs are constructed and pre-processed to make them suitable for process mining. The data for constructing the event log for this case study was taken from the Business Intelligence Unit (BIU) at the St. Mary's Hospital campus of the Imperial College Healthcare NHS Trust Hospitals. The reason this trust was chosen is because I needed an exemplar case study. Since I am already doing this research as part of the Imperial College London, I readily had access to this data (with my honorary contract).

The BIU is responsible for [176]:

- Populating Imperial enterprise data warehouse using ETL processes with both clinical and activity data for the whole Trust.

- Keeping quality of the data with enterprise data warehouse up to 99.9% accurate,
- Reporting any data quality error to respective team/supplier as per the protocol designed.
- Submitting both statutory and non-statutory reports both nationally and for day to day function of the Trust.
- Keeping all Trust and departmental performance dashboards up-to-date for efficiency and planning.
- Providing Finance Team all activity data within the Trust and providing SUS submission to commissioners.

4.2.1 SUBMISSION OF SECONDARY CARE DATA AT THE NATIONAL LEVEL

In the United Kingdom, within the National Health Service (NHS), primary care specialists work in the community and refer patients to secondary (hospitals) or tertiary care centres.

Clinical Commissioning Groups (CCGs) were created in 2012 and replaced Primary Care Trusts in April 2013. They are clinically-led legal NHS bodies responsible for the planning and commissioning of health care services for their local area [177]. At the primary care level, GPs submit their data to the CCGs and NHS England. At the secondary care level, hospital trusts submit their data to the CCGs as well.

All the activities the secondary care hospital does are submitted back to the CCGs in the form of a report called Commissioning Data Set (CDS) which contains all the data related to inpatient, outpatient, maternity, emergency etc. and based on this report the CCGs provide payment to the secondary care hospital.

The CDS is a national template that UK follows and is the primary mechanism for the national reporting of secondary care activity. As seen in Figure 22, CDS is securely submitted to the Secondary Uses Service (SUS) database in XML (Extended Markup Language) format. SUS is a data warehouse containing this patient-level information. Data in SUS can be patient identifiable, anonymised or pseudonymised as required for the user's needs. NHS providers and commissioners can use this data for secondary uses: purposes other than primary clinical care. SUS provides a range of services and functionality which you can use to analyse, report and present this data. SUS data are therefore held in a secure environment that maintains patient confidentiality to national standards.

CDS form the basis of the Hospital Episode Statistics (HES) data set. HES is a big enterprise data warehouse and gets consolidated reports from every trust in the form of SUS reports and other clinical submissions like Cancer data, Renal data, etc. HES data is therefore derived from SUS. [178]

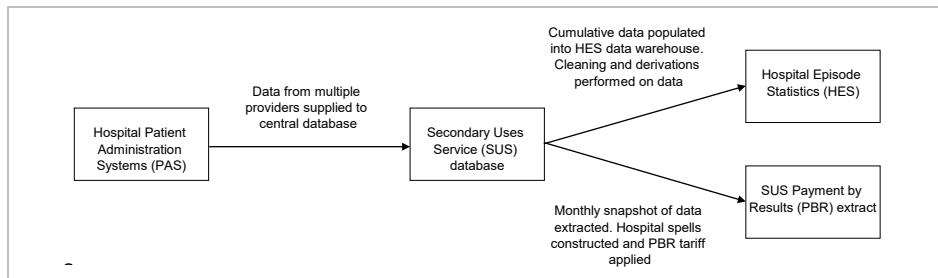


FIGURE 22: SUBMISSION OF CDS TO SUS DB AND FROM THERE TO HES

4.2.2 FLOW OF SECONDARY CARE DATA AT THE LOCAL LEVEL

As depicted in Figure 23, Cerner® solutions “enable physicians, nurses and other authorized users to share data and streamline processes across an entire organization. An online digital chart displays up-to-date patient information in real time, complete with decision-support tools for physicians and nurses” [179]. Cerner Millennium®, provided by Cerner, is the healthcare information system used at Imperial College Healthcare NHS Trust hospitals. It delivers high quality care to patients, safely and cost effectively. Cerner front-end stores its data in the Millennium database. Millennium database is an OLTP (Online Transaction Processing) system which forms the source of data for the data warehouse. PIEDW (PowerInsight Enterprise Data Warehouse) is the data warehouse for the Millennium database. From the PIEDW, each day 50+ CDE (Commissioning Data Extract) files are dispatched into an FTP location. From the FTP location they are put into dump tables. If there is still information remaining that is needed and not found in those CDE files, it is pulled out from PIEDW using the “Informatica” tool in the form of CDE or some other flat file. The CDE files are placed into a dump table and from the dump table staging tables are created based on patient MRN number and the activity (like an insert update or delete). An archive of the dump table and deleted records is also kept. From the CDE layer (which is the staging layer), the final reporting tables are created like the inpatient, outpatient etc. These are the fundamental data warehouse tables. From these tables the SUS reports are generated and sent off [178].

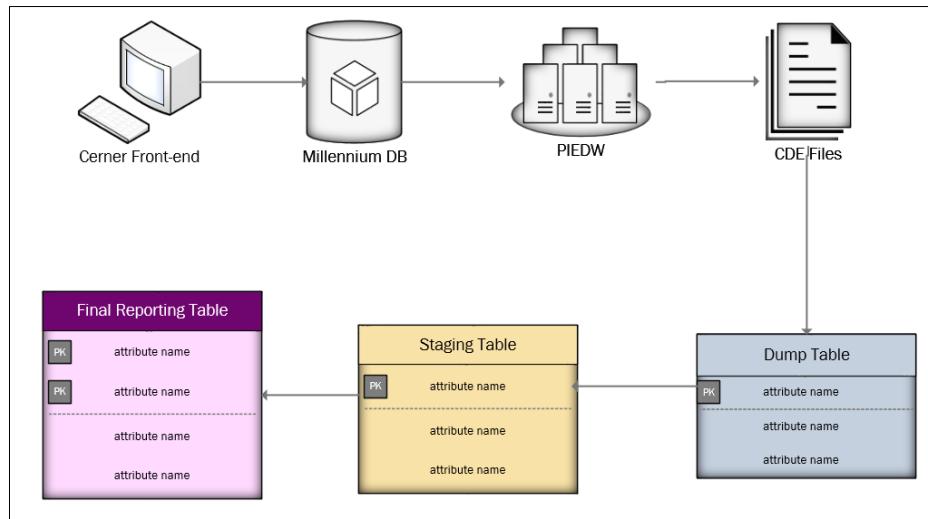


FIGURE 23: CDE WORKFLOW

4.2.3 LOGISTICS OF THE REFERRAL WORKFLOW

- Business comes from two sources: Inpatient and Outpatient. Inpatient is divided into two: Elective and non-elective. Elective means through a waiting list when the GP is sending. Non Elective means emergency. Elective patients are usually on an 18 week pathway.
- GP sends a referral via email, letter or fax to the hospital consultant's secretary. The consultant is either chosen by the GP or the patient himself through the CAB system (Choose and Book)
- Secretary receives this referral and books the patient based on the consultant's diary.
- Once patient is booked, the booking confirmation letter is sent back to the patient and a copy is sent to the GP.
- Before attendance, the patient is logged into the outpatient waiting list
- When the patient comes as a new patient, they are added to the patient record (outpatient table) by the Secretary and patient's status is then changed to *attended* or *did not attend* accordingly. [178]

4.2.4 CONSTRUCTION OF EVENT LOGS

Despite the availability of robust process-mining techniques and their successful applicability in various application domains, a limiting factor is the preparation of event data as data are usually not in a form that can be analyzed easily and need to be extracted and converted to event logs first. [174].

Before we start constructing the event log by using data from the information systems, we need to know where the information is stored and how to extract it and for that we need to understand the characteristics of the process mining project in detail. The construction of an event log depends on the desired view on the data. Even within the same process mining project there may be a desire to view the process from different angles. This can result in defining multiple conversions to extract different event logs from the same data source. Therefore, it is necessary to determine the following three steps before any event logs can be generated: Goal, Scope and focus of process mining [180]

GOAL OF PROCESS MINING

The goal of the process mining project should be clear before event log construction can start. In many cases it is required to know how the flow of the processes is executed as this knowledge is not available pre-hand. In such cases the goal would be to visualize the actual flow of processes as recorded in the event logs. Another goal would be to analyse certain performance or organisational aspects of the process. In some cases, however, there is no clear definition of the analysis that should be conducted. In these cases a more general empirical analysis of the event log on different aspects is done as a first step. The findings of such an analysis can allow the examination of further details of the process.

SCOPE OF PROCESS MINING

After the goal of the project is determined, the scope should be clarified as it determines what should or shouldn't be included in the event log extraction. The scope is partly determined by the goal of the project as it narrows down what part of the overall process should be investigated.

FOCUS OF PROCESS MINING

After both the goal and scope of the project are determined, the focus of the project allows us to concentrate on certain parts of the project where extra detail should be included and extracted. This can result in more fine-grained events and additional attributes to be included in the area that has more focus.

EXTRACTION OF EVENT LOGS

Once the above steps are executed and a clear plan is set in terms of goals and focus of the process mining project, we begin with the extraction of the event log. Event log extraction is a very important step in the process mining process and should not be overlooked. An event log generally consists of event traces which record the activities that were executed on a certain process instance (or case). Traces and events can contain several attributes that record information about the execution of tasks on cases. An example of such attributes would be information such as the name of the activity, when the task started or stopped and who executed it [180]. Although the underlying databases are laden with data, there are no explicit references to events, cases, and activities. The database tables contain records and the tables are connected to each other through relationship keys and thus the challenge is to convert these tables and records into event logs [174]. Information may be scattered over several different types of systems (e.g. ERPs, BPM systems, database systems). In such a diverse environment, the goal is to extract from all different sources to a common integrated environment to allow process mining [175]. If the data extracted is not taken from a data warehouse, an Extract, Transform, and Load (ETL) process needs to be performed manually before constructing the event log. ETL is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse. It involves the following tasks [181]:

- Extracting the data from source systems (e.g. ERP, workflow management systems), and converting it into an integrated data warehouse format which is ready for the transformation process.
- Transforming the data that involves applying business rules (e.g., calculating new measures and dimensions), cleaning (e.g., mapping NULL to 0 or "Male" to "M" and "Female" to "F" etc.), filtering (e.g., selecting only certain columns/data to load), joining together data from

multiple sources (e.g., lookup, merge), transposing rows and columns, applying any kind of simple or complex data validation (e.g., if the first 3 columns in a row are empty then reject the row from processing)

- Loading the data into a data warehouse or data repository

In situations when a database exists, several approaches are available to extract events [174, 175, 182]. The most common is the classical extraction in which events are manually obtained by linking various data sources and tables together. To do so, a lot of domain knowledge is required in order to select the right data [175]. If the application is small, usually human judgment would suffice about the case matching between two records. However, for applications with large amounts of data, this method becomes impractical and sophisticated linkage or matching techniques using the assistance of computers is required [183].

Record Linkage

Record linkage is the identification and collation of records from one or more data sources that are believed to be related to the same entity [184].

Linkage becomes easy when unique identifiers (e.g. NHS number), or some other element or group of elements that uniquely identify a given person or episode, are readily available. This approach is referred to as *deterministic matching*. Deterministic algorithms determine whether record pairs agree or disagree on a given set of identifiers. Match status can be assessed in a single step or in multiple steps. In a single-step strategy (called “exact” or “all-or-none” deterministic); records are compared all at once on the full set of identifiers to see for possible matching. A record pair is classified as a match if the two records agree, character for character, on all identifiers and the record pair is uniquely identified. A record pair is classified as a non-match if the two records are not uniquely identified or disagree on any of the identifiers. In a multiple-step strategy (called “approximate” or “iterative” deterministic); records are matched in a series of progressively less restrictive steps in which record pairs that do not meet a first round of match criteria are passed to a second round of match criteria for further comparison. If a record pair meets the criteria in any step, it is classified as a match. Otherwise, it is classified as a non-match. [183, 185, 186]

Record linkage, however, becomes more challenging when records don’t have this unique identification; when information is recorded in non-standard format; or when the files are large

scale. In this case, other variables like names, addresses, and/or dates of birth are then used to simulate human pattern recognition when deciding the matching process. This approach is called *probabilistic matching*.

Although human judgment is still considered to be the “gold standard” in matching records, a combination of deterministic and probabilistic computer methods, along with human judgment, will often be the best approach [183, 185].

4.2.5 DATA CLEANSING AND PREPROCESSING

Before process analysis techniques such as process mining can be applied on these event logs they need to be prepared and pre-processed in a way that allows for a seamless import into the process mining software of choice. A typical event log consists of the following [120]:

- Each event corresponds to an activity that was executed in the process
- Multiple events are linked together in a process instance or case
- Logically, each case forms a sequence of events—ordered by their timestamp.

In an event log, the following minimum elements need to be identified: Case ID, Activity, and Timestamp. The data that we have linked and constructed need to be prepared in such a way that these elements (and other additional elements of importance) are identified and arranged accordingly. This requires a number of preprocessing and cleansing steps to be undertaken e.g. simplify the data by skipping unnecessary low level activities and by merging the significant low level activities in singular high level ones [187], re-naming the events to a more understandable and friendly format, adding timestamps where required, etc.

Once the event logs are preprocessed, they need to be converted to a standardized format. For this purpose the Architecture of Information Systems research group at Eindhoven University of Technology specified the MXML event log format (used with PROM ver. 5). The MXML format has proven its use as a standardized way to store event logs for use in process mining. However, almost no information system in practice record their event logs directly in this format. During the use of the MXML standard several problems with its way of storing event related data have been discovered. To solve the problems encountered with MXML and to create a standard that could also be used to store event logs from many different information systems directly a new event log format

is under development by the IEEE Task Force Process Mining. This new event log format is named XES (used with PROM ver. 6 and stands for eXtensible Event Stream [180]).

As discussed in section 2.2.5, the DISCO software allows for the import and seamless integration of event logs in the standard format. DISCO has been designed to make the data import easy by automatically detecting timestamps, remembering configuration settings, and by loading data sets with unprecedented speed. DISCO is also fully compatible with the academic toolsets ProM 5 and ProM 6. By importing and exporting the event log standard formats MXML and XES, advanced users can seamlessly move back and forth between DISCO and ProM if they want to benefit from the cutting edge research technologies developed in academia [120].

4.3 DOMAIN UNDERSTANDING

As discussed in section 3.2.4, I began constructing my process model by first identifying the goal of my process mining project. In my case study, my goal was to find out where the bottlenecks and loop holes were in the flow of prostate cancer patients and how well does the actual pathway of these patients conforms to the standards laid out by the NICE clinical guidelines as well as the Prostate Cancer Risk Management Program. As my goal was covering a performance analysis aspect followed by a conformance check, I decided to conduct a more explorative analysis on my event log to gain an initial understanding of the pathway in terms of simple log statistics. Therefore, to narrow down to a scope of the project, I needed to find all data that covered the flow of prostate cancer patients in the secondary care from the time the patient enters the custody of the caregivers until the entire continuum of care. As a result, I had to focus on the referral phase: inpatient/outpatient appointments; the diagnostic phase: laboratory results, radiology exams; and the treatment phase: chemotherapy, radiotherapy, and surgery. Once this was decided, I was ready to identify the respective databases and start the linkage process to create my event log.

4.3.1 IDENTIFYING DATA SOURCES

My case study on prostate cancer involved a wide range of data from several departments at the Imperial College Healthcare NHS Trust. In order to accurately link this data across systems and departments, it was necessary to gather a list of relevant data sources containing routinely collected

data and clinical information systems that are used to register user activity of patients diagnosed with prostate cancer. To understand, analyse and link this data, I consulted with IT colleagues in the Business Intelligence Unit (BIU) as well as Urology physicians at St. Mary's Hospital, London and found out what data is available for prostate cancer. The databases I had access rights to were all application databases and hence were Online Transaction Processing Systems (OLTP) that recorded every transaction in the trust. As opposed to a data warehouse, that is an Online Analytical Processing System (OLAP), the trust databases did not facilitate querying and analysis usually found in warehouses. After a series of collaborations and meetings with the Urology clinic staff, as well as a detailed study of data dictionaries for each information system, I identified the databases relevant to prostate cancer that covered the following types of information systems: Inpatient and outpatient appointments, MDT meetings, Pathology, biopsy, radiology, chemotherapy, radiotherapy, and surgery.

4.3.2 SELECTED DATA SOURCES

The databases identified in Table 14 covered the administrative and clinical services provided to cancer patients.

TABLE 14: PROSTATE CANCER RELATED DATABASE IN BIU

Database Name	Type of Information contained
Cerner Database (previously ICHIS)	Outpatient, inpatient (New data and historic data before 2012), and demographic data
Somerset Database	Chemotherapy + Multidisciplinary team meeting (MDT) data
Pathology Database	Labs data
Radiology Database (ARIA)	Imaging data
Radiotherapy Database	Radiotherapy data
Surgery Database (Theatre Man)	Surgical data

CERNER DATABASE

The Cerner database, which is a hospital Patient Administration System (PAS) and is a successor to the legacy ICHIS system, contains patient demographic information as well as outpatient and inpatient appointment information for all patients along with their clinical coding data (OPCS codes for procedures and ICD-10 codes for diagnosis)

SOMERSET DATABASE

The Somerset database gets its data from the Somerset Cancer Register (SCR). SCR is a software application developed by the NHS, designed to collect relevant data throughout the patient's cancer journey. The data collection, in addition to supporting patient care, supports the National Clinical Audits, Surgeon Level Reporting and the Cancer Waiting Times [188]. The Somerset database contains information pertaining to cancer referrals, lab data, imaging data, chemotherapy data, MDT (multidisciplinary team meeting) data, brachytherapy data, etc. I was only interested in the MDT and chemotherapy tables from Somerset as the rest of the information I was going to link with the main respective source tables like pathology and radiology.

PATHOLOGY DATABASE

The biochemistry & histopathology lab system includes important blood reports and histopathological reports with respective tumor markers. This database was important to extract information regarding PSA tests as well as prostate biopsies that would aid in filtering prostate cancer patients.

RADIOLOGY DATABASE

The radiology database (previously known as ARIA) contains textual information (medical reports) relevant to radiological imaging information.

RADIOTHERAPY DATABASE

The radiotherapy system contained information relevant to radiotherapy treatments and results

SURGERY DATABASE

The surgery database gets its data from the TheatreMan theatre management system. TheatreMan is a modular, flexible and adaptable theatre and day surgery management system specifically designed with a comprehensive set of features to enable total management of the patient episode in a surgical environment [189].

4.4 DATA PREPARATION

To provide an accurate and holistic view of each patient, data needs to be collected and linked from different data sources. Since the data was residing in separate databases and not in a data warehouse that would contain cleaned, transformed data ready for process mining, I resorted to doing the ETL process manually. The first step was extracting and linking data from the databases identified in section 3.3.2.

4.4.1 DATA LINKAGE

I opted to use the Deterministic Record Linkage strategy as the same NHS number is shared among each database making it easier to make an integrated database. This technique picked a unique identifier (in my case the NHS number or PKEY) and selected the records that share the same value for that identifier as belonging to the same patient and considered the unmatched records as belonging to different patients. When exceptions occurred (like a missing NHS number or PKEY) then further record linkage rules (e.g. Local Hospital Number) were created to handle the exception and include the maximum records.

The data integration started off by considering the information system with the main inpatient and outpatient appointments (Cerner DB) as the central system on which the rest of the data from other information systems would be subsequently joined to.

Queries and scripts in Transact SQL (TSQL) were written and SQL views were created to amalgamate the records in a separate MS SQL server staging database by means of Active Server Pages (ASP) classic routines that would automatically import the required data (See Appendix C). The steps taken to integrate and link the data are depicted and explained in figures 20-27.

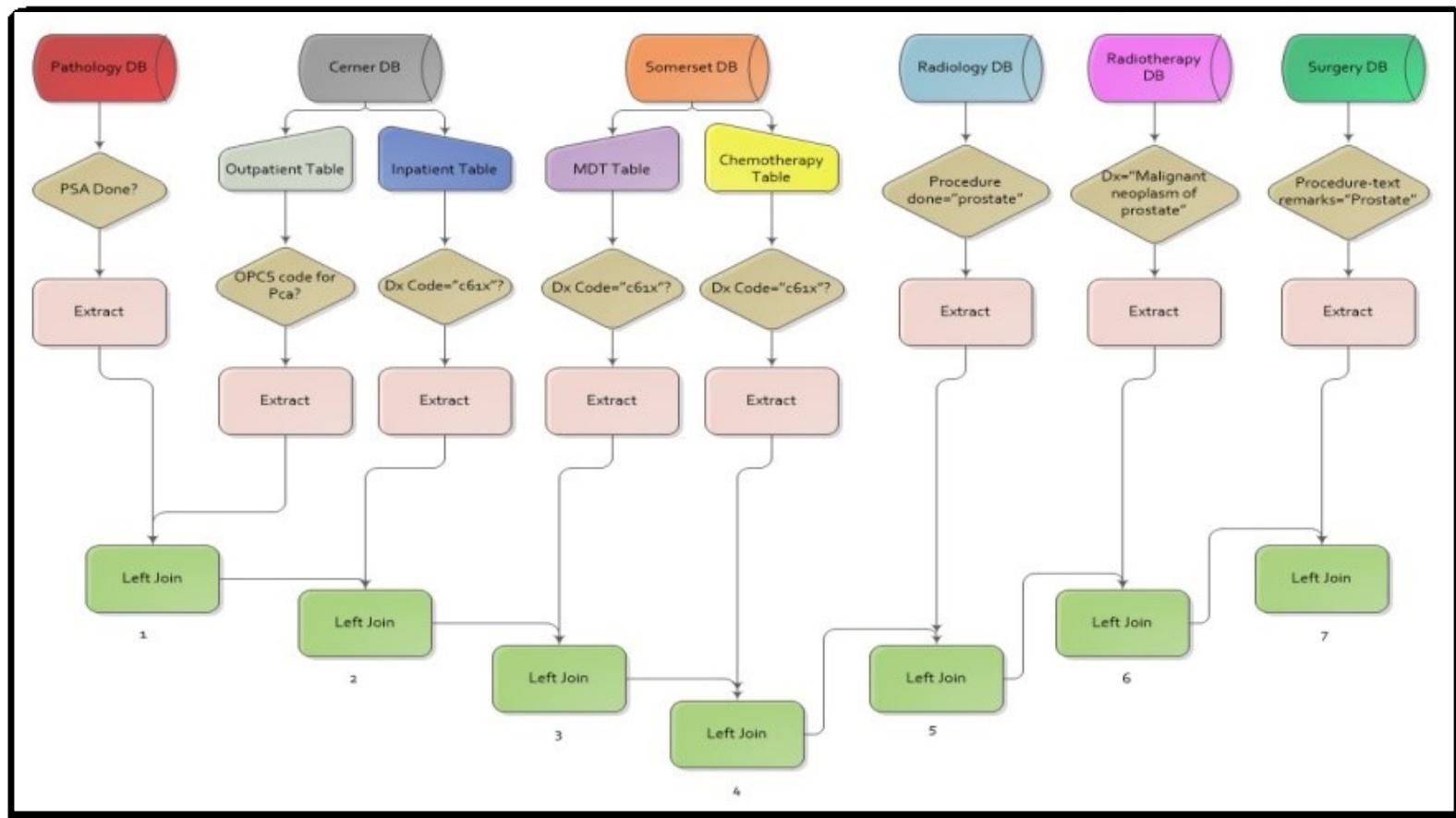


FIGURE 24: OVERVIEW OF STEPS TO LINK THE SIX DATABASES

STEP 1: FILTER INPATIENT AND OUTPATIENT APPOINTMENT TABLES FROM CERNER DB

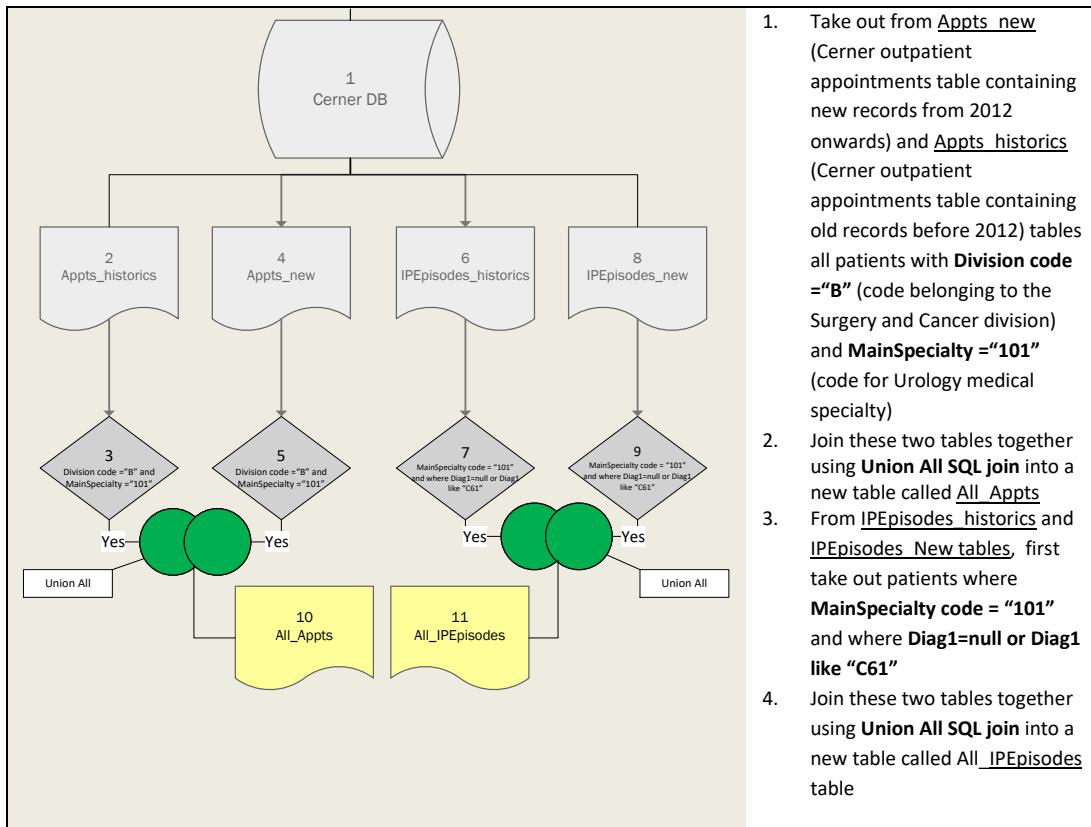


FIGURE 25: STEPS FOR CERNER DATA FILTERATION

1. Take out from Appts_new (Cerner outpatient appointments table containing new records from 2012 onwards) and Appts_histories (Cerner outpatient appointments table containing old records before 2012) tables all patients with **Division code** = “B” (code belonging to the Surgery and Cancer division) and **MainSpecialty** = “101” (code for Urology medical specialty)
2. Join these two tables together using **Union All SQL join** into a new table called All_Appts
3. From IPEpisodes_histories and IPEpisodes_New tables, first take out patients where **MainSpecialty code** = “101” and where **Diag1=null** or **Diag1 like “C61”**
4. Join these two tables together using **Union All SQL join** into a new table called All_IPEpisodes table

STEP 2: FILTER LAB DB ON PSA AND LINK

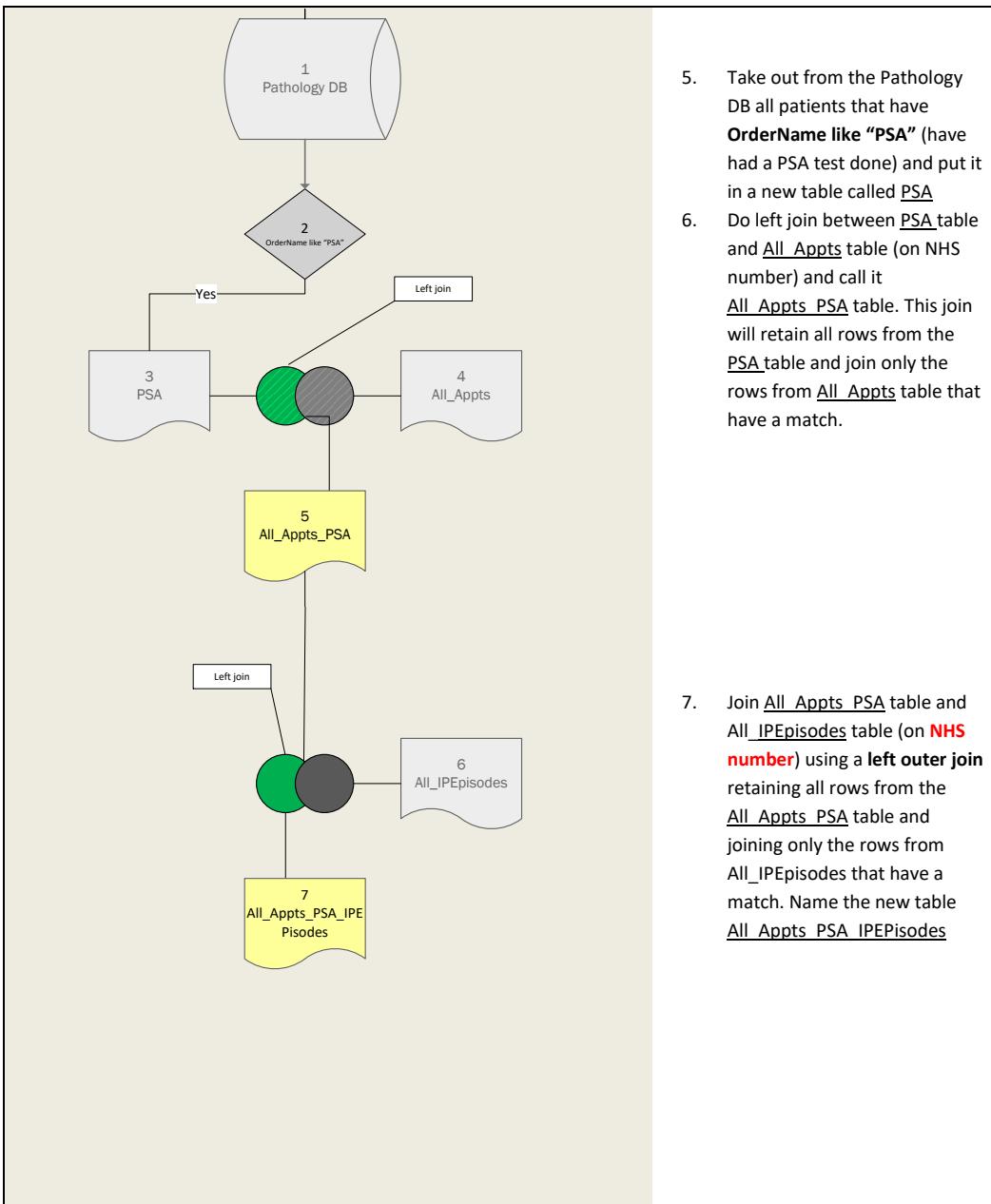


FIGURE 26: STEPS FOR LAB FILTERATION ON PSA AND LINKAGE

5. Take out from the Pathology DB all patients that have **OrderName like “PSA”** (have had a PSA test done) and put it in a new table called PSA
6. Do left join between PSA table and All_Appts table (on NHS number) and call it All_Appts_PSA table. This join will retain all rows from the PSA table and join only the rows from All_Appts table that have a match.
7. Join All_Appts_PSA table and All_IPEpisodes table (on **NHS number**) using a **left outer join** retaining all rows from the All_Appts_PSA table and joining only the rows from All_IPEpisodes that have a match. Name the new table All_Appts_PSA_IPEEpisodes

STEP 3: FILTER LAB DB ON BIOPSY AND LINK

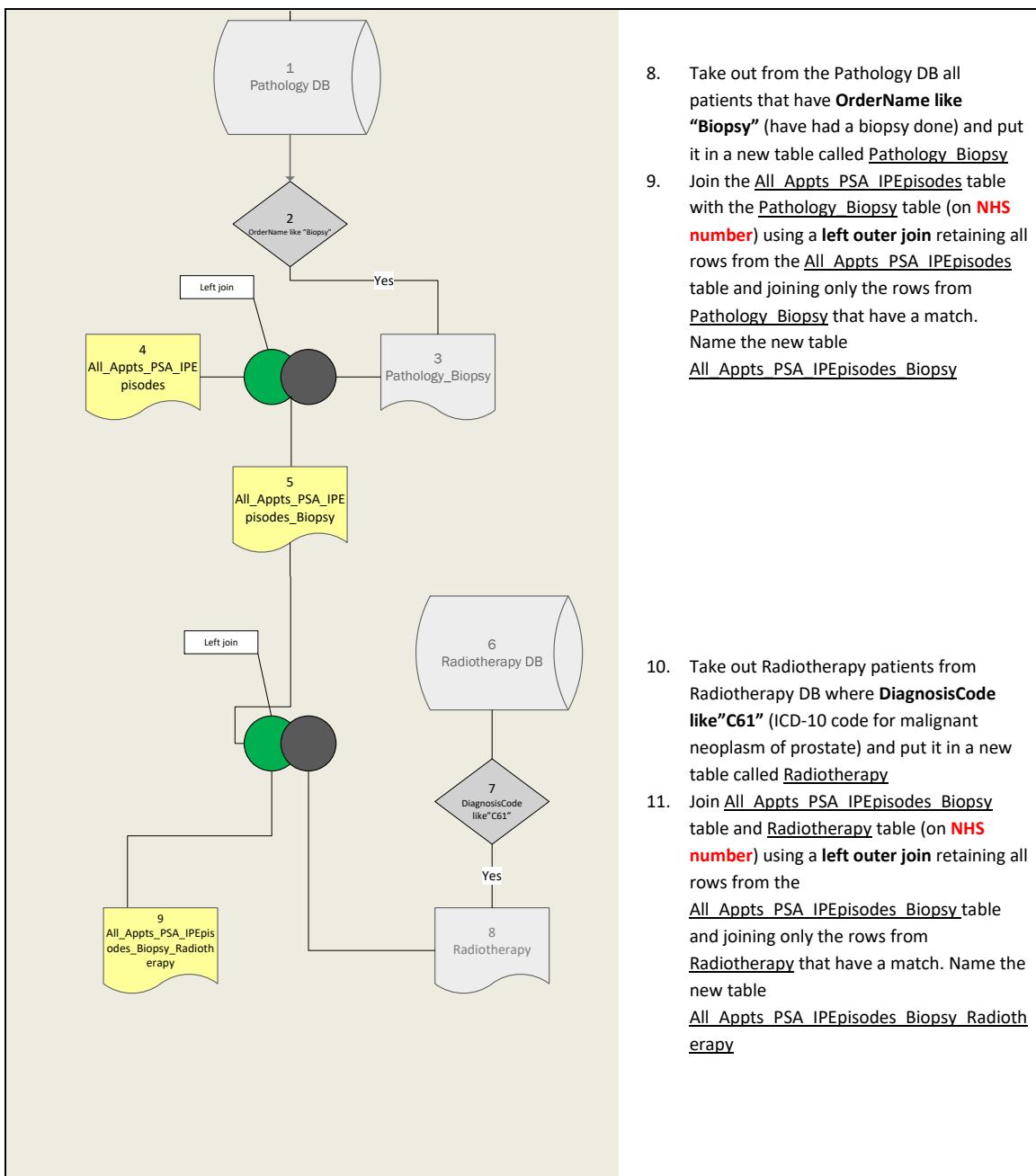


FIGURE 27: STEPS FOR LAB FILTERATION ON BIOPSY AND LINKAGE

8. Take out from the Pathology DB all patients that have **OrderName** like **"Biopsy"** (have had a biopsy done) and put it in a new table called Pathology_Biopsy
9. Join the All_Appts_PSA_IPEEpisodes table with the Pathology_Biopsy table (on **NHS number**) using a **left outer join** retaining all rows from the All_Appts_PSA_IPEEpisodes table and joining only the rows from Pathology_Biopsy that have a match. Name the new table All_Appts_PSA_IPEEpisodes_Biopsy
10. Take out Radiotherapy patients from Radiotherapy DB where **DiagnosisCode** like "**C61**" (ICD-10 code for malignant neoplasm of prostate) and put it in a new table called Radiotherapy
11. Join All_Appts_PSA_IPEEpisodes_Biopsy table and Radiotherapy table (on **NHS number**) using a **left outer join** retaining all rows from the All_Appts_PSA_IPEEpisodes_Biopsy table and joining only the rows from Radiotherapy that have a match. Name the new table All_Appts_PSA_IPEEpisodes_Biopsy_Radiotherapy

STEP 4: FILTER RADIOLOGY DB AND LINK

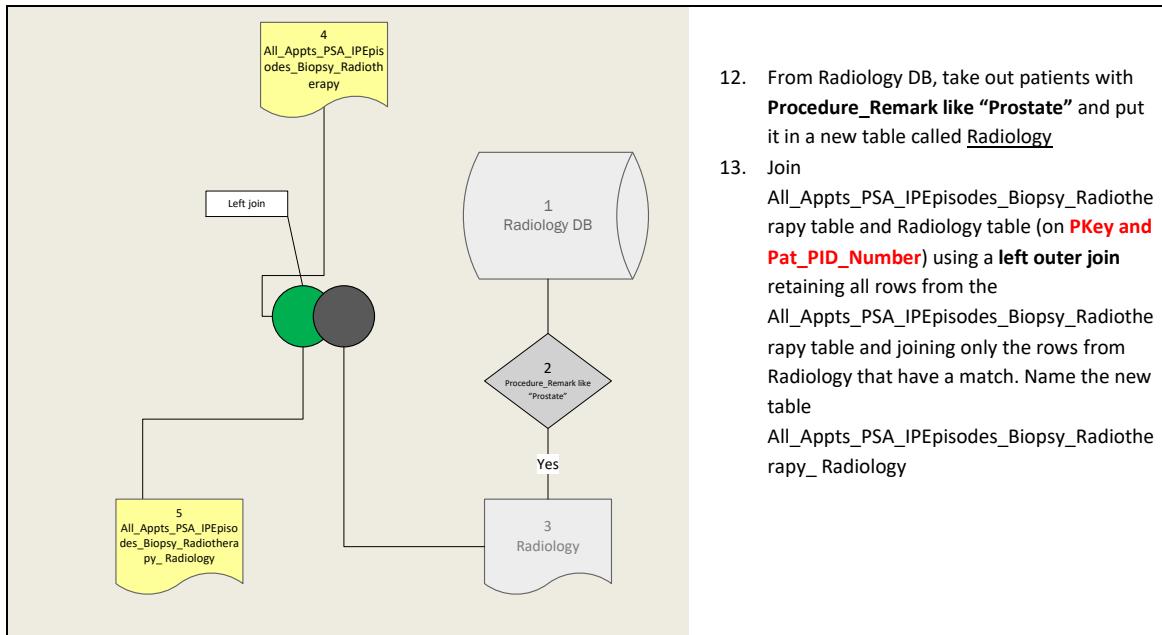


FIGURE 28: STEPS FOR RADIOLOGY DATA FILTERATION AND LINKAGE

STEP 5: FILTER MDT TABLE FROM SOMERSET DB AND LINK

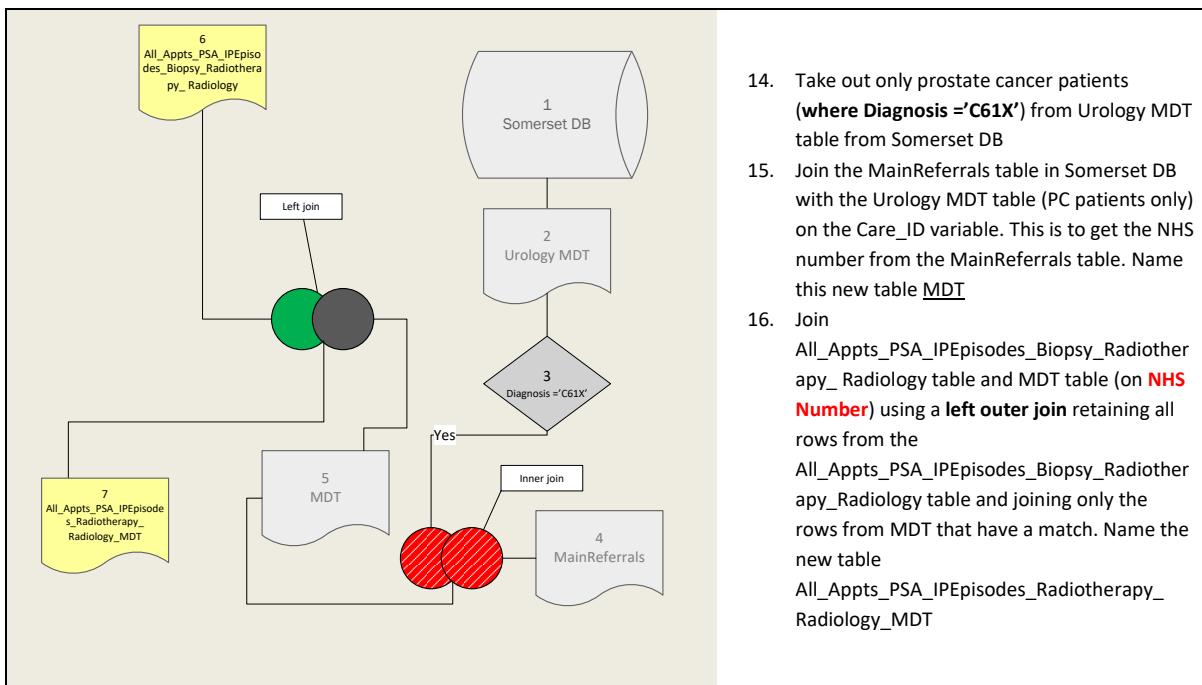


FIGURE 29: STEPS FOR MDT DATA FILTERATION AND LINKAGE

12. From Radiology DB, take out patients with **Procedure_Remark** like “**Prostate**” and put it in a new table called **Radiology**
13. Join **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy** table and **Radiology** table (on **PKey and Pat_PID_Number**) using a **left outer join** retaining all rows from the **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy** table and joining only the rows from **Radiology** that have a match. Name the new table **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology**

14. Take out only prostate cancer patients (**where Diagnosis = 'C61X'**) from Urology MDT table from Somerset DB
15. Join the **MainReferrals** table in Somerset DB with the Urology MDT table (PC patients only) on the **Care_ID** variable. This is to get the NHS number from the **MainReferrals** table. Name this new table **MDT**
16. Join **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology** table and **MDT** table (on **NHS Number**) using a **left outer join** retaining all rows from the **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy** table and joining only the rows from **MDT** that have a match. Name the new table **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology_MDT**

STEP 6: FILTER SURGERY DB AND LINK

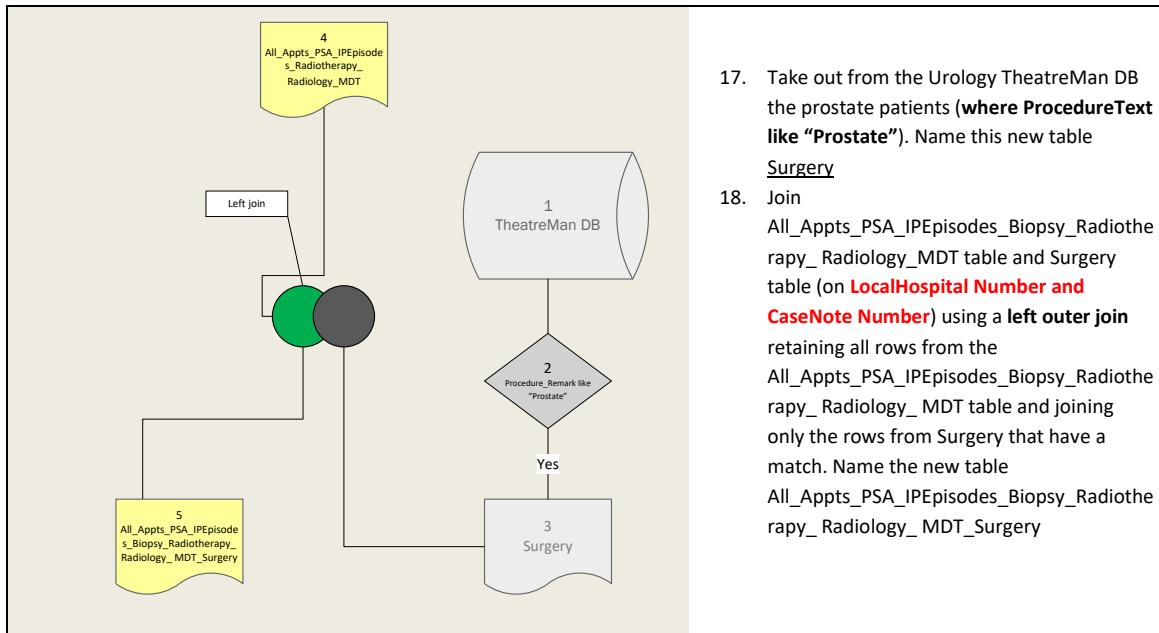


FIGURE 30: STEPS FOR SURGERY DATA FILTERATION AND LINKAGE

STEP 7: FILTER CHEMOTHERAPY TABLE FROM SOMERSET DB AND LINK

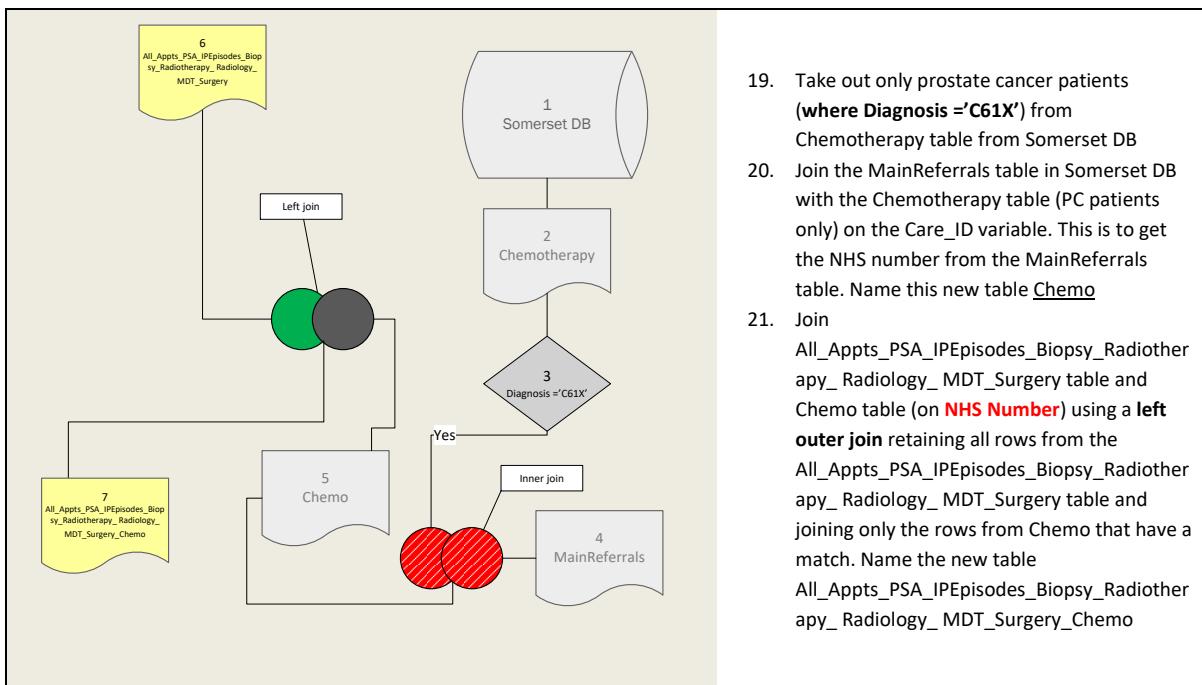


FIGURE 31: STEPS FOR CHEMOTHERAPY DATA FILTERATION AND LINKAGE

17. Take out from the Urology TheatreMan DB the prostate patients (**where ProcedureText like "Prostate"**). Name this new table **Surgery**
18. Join **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology_MDT** table and **Surgery** table (on **LocalHospital Number** and **CaseNote Number**) using a **left outer join** retaining all rows from the **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology_MDT** table and joining only the rows from **Surgery** that have a match. Name the new table **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology_MDT_Surgery**

19. Take out only prostate cancer patients (**where Diagnosis = 'C61X'**) from **Chemotherapy** table from Somerset DB
20. Join the **MainReferrals** table in Somerset DB with the **Chemotherapy** table (PC patients only) on the **Care_ID** variable. This is to get the NHS number from the **MainReferrals** table. Name this new table **Chemo**
21. Join **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology_MDT_Surgery** table and **Chemo** table (on **NHS Number**) using a **left outer join** retaining all rows from the **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology_MDT_Surgery** table and joining only the rows from **Chemo** that have a match. Name the new table **All_Appts_PSA_IPEpisodes_Biopsy_Radiotherapy_Radiology_MDT_Surgery_Chemo**

Once all the records were linked, they were then examined for accuracy and completeness (manually or using scripts), and a huge, de-normalized, linked staging table was created containing all the possible activities of the patients. This table served as the input for the next log preparation (Transformation and transposition) phase.

4.4.2 OUTPUT OF DATA LINKAGE

Based on my goal and scope, only the records containing information belonging to urology/prostate cancer starting from June 2009 until June 2015 were filtered in all the databases before linkage, instead of using the entire cohort of patients. Table 15 shows the number of records extracted from each database:

TABLE 15: RECORDS FILTERED FROM EACH DB AND TOTAL RECORDS LINKED

Table Name	Records filtered
Outpatient appointments	324,355
Inpatient appointments	4366
Lab DB on PSA	75,968
Lab DB on Biopsy	66,268
Radiology DB	135,282
MDT table	7029
Surgery DB	20,620
Chemotherapy table	6341
Total number of records after linking all the DBs	3,809,063

4.4.3 DATA EXTRACTION AND LOG PREPARATION

After the data linkage, the required data were extracted and transformed from the integrated database. The data extraction and transformation was done in Microsoft SQL Server 2012 using TSQL queries and codes written in ASP classic for creating the event log. In order to extract data, it is necessary to know how this data will be used in the analysis phase (and by what software packages) and thus the extraction needs to be done accordingly. For example, my data extraction needed to be utilised by a process mining software and needed to follow the form of event logs to apply the analysis techniques of process mining. Thus, not only vital event log information such as patient ID,

activity, resource ID, and timestamp was extracted, but also additional supplementary information such as the type of patient, type of diagnosis, etc. was also extracted. A table of the important event log variables belonging to each table is shown in Table 16. A number of scripts in ASP Classic environment (using VBScript) were written to automate the process of data extraction directly into an event log format so that no manual intervention was necessary (See Appendix C).

TABLE 16: EVENT LOG INFORMATION IN EACH TABLE

Table Name (Patient activity)	Case ID	Timestamp	Resource ID	Other important variables
Outpatient Appointments	NHSNumber	Date_of_Appointment	OPCS code	DateRaisedByGP, OriginalGPReferralDate, ReferralRequestReceivedDate, ClinicName1, ConsultantName, ReferringGP, ReferralConsultant, ReferringSourceNational, GPUrgentFlag, AttendedFlag, AttendedOrDidNotAttendNational, DateOutcomeRecorded, OutcomeCodeLocalDescription, ReasonForAppointmentDescription, AppointmentTypeLocal, AppointmentPriorityLocal, HospitalCodeDescription, SiteCode, MainSpecialityCodeLocal, DateOfBirth, AgeAtStartOfSpell, DateOfDeath, SexCodeLocal, EthnicCodeLocal, MaritalStatus, ReligionLocal, RegisteredGP, RTTStartDate, RTTEndDate
PSA	NHSNumber	Collection_date	Result	LabDept, OrderCode, OrderName, ordercomment, TestCode, TestName, Result, ResultUnits
Biopsy	NHSNumber	Collection_date_bio	Result_bio	LabDept, OrderCode, OrderName, ordercomment, TestCode, TestName, Result_bio, ResultUnits
Radiology	NHSNumber	Appointment_date	Rad_Procedure_code	-none-
MDT	NHSNumber	MDT_date	-none-	-none-
Radiotherapy	NHSNumber	Course_start	Procedure_code_rtherapy	PatientType, Diagnosis, Intent, CourseEnd, ActivityCode_Rtherapy, ProcedureComment, ProcedureDateTime, ActivityCategoryCode
Inpatient (Admissions)	NHSNumber	Procedure_date	Inp_Procedure	AdmissionDate, WhoAdmitted, WhereAdmitted, DateDecidedToAdmit, HospitalCode, IntendedManagementNational, SourceOfAdmissionNational, PointOfDelivery, DischargeMethodNational, DischargeDestinationNational, DischargeDate, WhereDischarged, WhoDischarged, DIAG1
Surgery	NHSNumber	Surgery_date	Surgery_Procedure_text	PrimarySurgeonCode, AdmissionType, OperationType, Theatre
Chemotherapy	NHSNumber	Chemo_date	Therapy_type	Drug_regimen

In the log preparation step, the aim is to prepare the available data in such a way that it can easily be used for the next step as well as for the actual process mining. The steps that are required to transform the data to a usable event log relies more on the intuition and process knowledge of the researcher than on a predefined method. Examples of activities are the renaming and aggregation of events. After creating the event log, the linked staging table that contained a dump of all the episodes of the patients, was separated on a patient-by patient basis. For this, a special script was written that would take out the unique patients from the staging table and find all their subsequent episodes and arrange them together (See Appendix C).

4.4.4 OUTPUT OF DATA EXTRACTION

The study has been executed on the care processes of 27,419 patients under prostate cancer belonging to the Imperial College NHS Trust group of hospitals. All diagnosis and treatment activities that were performed for these patients during a period of 5 years, from January 1, 2010 to December 31, 2014 were collected. A range window of 6 months before (from 1 June 2009) and 6 months after (1 June 2015) was taken in order to take into account patients still completing a previous path or completing a future path.

Table 17 shows the results of the data extraction.

TABLE 17: DATA EXTRACTION RESULTS

Table Name	Records filtered
Total no. of instances (events)	148,898
No. of distinct patients	27,419
No. of distinct activities (processes)	9

4.5 LOG PREPARATION

In order to make sense about the most frequent path or reason about throughput time of cases, I had to pre-process (or clean/filter) the logs. The filtering routines, for which a special script was written to automate the cleaning process, included:

- Giving names to all activities based on the database it was linked from (e.g. Outpatient Appointment, Labs, Imaging, etc.)
 - Deriving timestamp fields (Start date of process, end date of process), which is an essential aspect of process mining. Since the timestamp was not already available in a mining-friendly format, the dates were extracted from the dataset using scripts and were populated in timestamp variables. The end date of process, where not available, was kept as the start date of the process.
 - Selecting only relevant columns to load and skipping the low relevance columns that were not important for the objective
 - Translating coded values into user friendly values that would be later used in visualisation
- Some examples included:
- replacing gender codes “1”and “M” to “Male”
 - specialty code “101”, “370”, etc. to “Urology”, “Medical Oncology”, etc.
 - Ethnic code “A”, “B”, etc. to “British”, “Irish” etc.
 - Marital Status code “S”, “M”, etc. to “Single”, “Married”, etc.
 - OPCS codes “M701”, “M703” to “Aspiration of prostate”, “Rectal needle biopsy of prostate”, etc.
- Deriving calculated values like “Age”, Referral to treatment <63 days, Referral to MRI < 10 days (based on the London Cancer Alliance standards)
 - Transposing multiple columns into multiple rows based on the business requirement and activities performed
 - Generating a surrogate key (pseudo key) for each patient using VBScript that allowed patient anonymity to be preserved by not referring to the NHS number
 - Filling the missing values or null values in the timestamps with appropriate values based on discussions with the physicians (e.g. no radiotherapy end date for certain patients)
 - Change date values to day/month/year format
 - Filling out missing codes (e.g. Main Specialty code)

TABLE 18: SNAPSHOT OF PATIENT-BY-PATIENT FLAT FILE

Trial	DATE_PROC	DATE_PROC2	PROCESS	TimeOfAppointment	DateRaisedByGP	OriginalGPReferralDate	ReferralRequestReceivedDate	ClinicName1
1	2013-09-08	2013-09-08	OUTPATIENT APPOINTMENT	2013-09-08 10:00:00	2013-09-08		2013-09-08	STONE CLINIC 2
1	2013-09-09	2013-09-09	OUTPATIENT APPOINTMENT	2013-09-09 10:00:00			2013-09-09	BLADDER CYSTOSCOPIC
1	2013-09-14	2013-09-14	OUTPATIENT APPOINTMENT	2013-09-14 10:00:00			2013-09-14	GEN UROLOGY CLIN
1	2013-09-14	2013-09-14	LABS PSA				2013-09-14	
2	2013-09-15	2013-09-15	OUTPATIENT APPOINTMENT	2013-09-15 10:00:00			2013-09-15	RAPID ACCESS PRO
2	2013-09-15	2013-09-15	LABS PSA				2013-09-15	
3	2013-09-16	2013-09-16	IMAGING				2013-09-16	
3	2013-09-16	2013-09-16	IMAGING				2013-09-16	
3	2013-09-19	2013-09-19	LABS PSA				2013-09-19	

ConsultantName	ReferringGP	ReferringConsultant	ReferringSourceNational	GPUrgentFlag	AttendedFlag	AttendedOrDidNotAttendNational	DateOutcomeRecorded
			referral from a GP	Y	Y	Arrived late but seen	2013-09-08
			other - initiated by the CONSULTANT	N	Y	Arrived late but seen	
			other - initiated by the CONSULTANT	N	Y	Arrived late but seen	
						Arrived late but seen	
			referral from a GP	N	Y	Attended on time	
						Attended on time	
						Arrived late but seen	
						Arrived late but seen	
						Arrived late but seen	

OutcomeCodeLocalDescription	ReasonForAppointmentDescription	AppointmentTypeLocal	AppointmentPriorityLocal	HospitalCodeDescription	SiteCode	MainSpecialityCodeLocal
		F	U	Charing Cross Hospital	RYJ02	Urology
		R	R	Charing Cross Hospital	RYJ02	Urology
		R	R	Charing Cross Hospital	RYJ02	Urology
				Charing Cross Hospital	RYJ02	
		F	T	Charing Cross Hospital	RYJ02	Urology
				Charing Cross Hospital	RYJ02	
				Charing Cross Hospital	RYJ02	
				Charing Cross Hospital	RYJ02	
				Charing Cross Hospital	RYJ02	

Once the records underwent an extensive transformation procedure, the final output was exported to a Microsoft Excel spreadsheet format (Table 18). From MS Excel, the file was then directly imported into my chosen process mining tool DISCO. Table 16 shows a snapshot of the MS Excel file containing the patient by patient data.

4.6 DISCUSSION

The construction of my process model went through various data integration, extraction and preparation steps to achieve the required event log necessary for the process mining stage. The initial steps of understanding the prostate cancer domain included a thorough understanding of a generic cancer pathway and from that focusing on the prostate cancer pathway as a case study. To make sense out of the pathway, it was not sufficient to only meet with physicians and other colleagues handling oncology data, but I also made sure I met with interface and dashboard designers that are currently involved in designing a simplified cancer patient journey map to get an idea about the pathway from their point of view. Once my understanding was clear, I utilised my findings from the literature search on the prostate cancer minimum required dataset and identified the databases that can provide me with those variables.

The linkage process was a lengthy, reiterative process that involved linkage and testing to see if the cohort of patients was representative and covered all patients potentially having prostate cancer. I decided to use all patients having a PSA done in the years between 2010 and 2014. For purposes of including cases that began their appointment before or after this time range, I increased the window from June 2009 until June 2015 so that we can be sure that all patients are fully covered and complete.

Before filtration of the relevant records could take place, it was necessary for me to understand in every database what specific information could yield only patients belonging to prostate cancer. Once those filtration criteria were decided and tested, I went ahead and filtered only the prostate cancer patients from within all the databases. It was also important to decide at that point which kind of join to use between the different tables at each step in order to retain the necessary patient records.

While performing the case study some limitations and challenges of the application or the technology used were discovered. These are discussed in the following section.

4.6.1 CHALLENGES

- Acquiring an honorary contract at the Imperial College NHS Trust to give me access to the databases, took a minimum of six months that greatly delayed the data extraction phase
- The databases had no previous documentation, data dictionary, or entity-relationship diagrams available that could assist me in getting a general idea about the data contained in them. I had to understand what data is collected in each database and how it was organised.
- The BIU is home to a collection of different clinical databases residing under one server. The databases are all application databases and hence are Online Transaction Processing Systems (OLTP) that record every transaction in the trust. As opposed to a data warehouse, that is an Online Analytical Processing System (OLAP), the trust databases did not facilitate querying and analysis and it needed to be done manually or using specially written scripts.
- The reiterative process of linking the databases to find the most representative cohort took many months to perfect and even when I thought I achieved the most optimal cohort, I went back several times to link from scratch as there was always a better way of linking and getting a greater sample than the one I previously used. Looking back retrospectively, it would have been ideal if I perfected my cohort in the very beginning before proceeding to the next steps, as the whole reiterative process of re-extracting the data to find a representative cohort caused major delays.
- The ETL process involved millions of records and due to the limitation of resources in the computer systems, the queries took to an unexpected time to execute (sometimes involving hours and days), as well as countless system crashes, run-time errors, and script time-out errors. These challenges were resolved by modifying the queries to make them more optimized, increasing the amount of disk space, and segregating the data based on quarters in a year (over the 6 year span) to execute queries in batches of data.

4.6.2 LIMITATIONS

The current study mainly has limitations of the constructed event log from the available raw data in the BIU data repositories. The event log inherently suffered from the following limitations:

- A limited accuracy of the timestamps as some of the radiotherapy processes had no ending time stamps whereby a dummy timestamp was created with the same as the start date
- Data was noisy and incomplete. This was due to the fact that:
 - The constructed event log was not retrieved from a proper OLAP data warehouse
 - A multitude of blank values were found in the entire dataset including the variables used for analysis, leading to a skewed result
 - The hospitals mentioned in the raw dataset were mainly blanks and the remainders were hard coded as “Charing Cross” thereby disallowing any other hospital code to be populated in the database. This problem was taken care of by the database administrators later on when pointed out
 - There were no biopsy results and dates found in the pathology database. The BIU unit never utilised biopsy results and thus the only code I had to rely on were the OPCS code of “Rectal Needle Biopsy” that only constitutes 18.76% of the entire dataset. The unit later received a restricted view to the biopsy database, however, the information was not sufficient
 - Imaging and MRI data was incomplete for many years of the cohort
 - MDT information was not captured for the years preceding 2013. I tried retrieving this information from any historic archives but it was not recorded anywhere.

4.6.3 TIPS ON IMPROVING DATA QUALITY

Data quality is vital for good decision making. The following are tips to preserve this data quality [190]:

- Maintain a good data dictionary: A data dictionary is important for documenting data stored in the database. It is important to Identity all data elements, provide definitions of those data elements, and specify validation rules (expected value of each data element)

- Clean your data: As data quality is also concerned with usability in analytics it is more important to ensure data is clean by addressing and taking care of several issues like sentence case, telephone numbers, dates, abbreviations, spelling, etc.
- Originate data from one source: To avoid data conflict and data error, avoid replicating and re-distributing local copies of data and repeatedly duplicating the data from the same source. Generate the data once only.
- Be vigilant about missing data: Be careful about filling in missing values by statistically guessing it. This leads to errors in analysis
- Perform regular reviews of your data to uncover anomalies: If you want to really understand your data and ensure data quality, you have to dive into the data and review it to understand any irregularities

4.7 SUMMARY

This chapter began the first and second phases of the CPAM road map. It talked about how I selected my data sources and linked them using specially developed algorithms to extract data relevant to the prostate cancer pathway. I then pre-processed and prepared the data so the final output would be an event log ready to be used in a process mining software. The next chapter, Chapter 5: Data Validation, continues the second phase of the CPAM roadmap and provides a means of validating whether the patients extracted in this chapter using my algorithms conform to the patients actually seen in clinic. This validation is an important step before starting the analysis.

CHAPTER 5:

DATA

VALIDATION

This chapter continues on to the third phase of the CPAM roadmap (

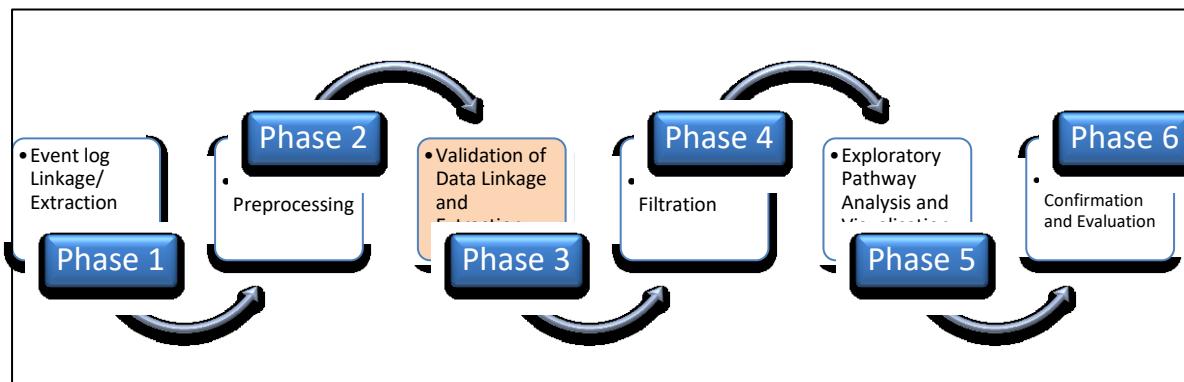


Figure 32). Section 5.2 describes the methods I used in conducting the case-note audit. It talks about the three stages of preparation, selection criteria and measuring performance. Section 5.3 presents the results of my audit including comparison metrics between my extracted data and the audit data. Section 5.4 presents a discussion on the results of the audit. Chapter 5 is concluded with section 5.5 that gives a summary of the entire chapter and how the following chapter is linked.

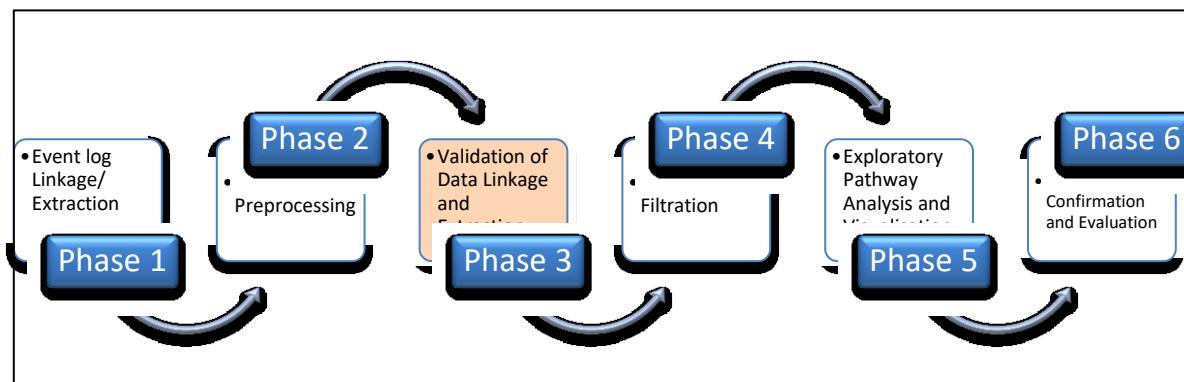


FIGURE 32: PHASE 3 (VALIDATION OF DATA EXTRACTION) OF THE CPAM ROADMAP

5.1 INTRODUCTION

Good record-keeping is a vital part of healthcare and is essential for the delivery of safe and effective patient care. It keeps a track of the patient's journey through the healthcare system and protects the accountability of the staff who delivers that care [191]. Moreover, it also ensures that high standards

and continuity of care is maintained all the time. Here are some tips for good record keeping [192, 193]:

- Write legibly so that everyone can read your writing (if electronic records are not available)
- Include the date and time in all your entries (if electronic records are not available)
- Avoid abbreviations to prevent ambiguity
- Do not alter an entry or disguise an addition to avoid tampering with medical records and initiating an investigation. Computerised record systems have an audit trail that will allow alterations to be discovered.

- Avoid unnecessary comments that are offensive, personal or unprofessional and could damage your credibility
- Beware of computer defaults and tick boxes that can lead to erroneous entries due to oversight
- Provide patient compliance to advice or treatment in order to decide on whether or not to provide advanced treatments
- It is best practice to record a patient's risk of developing disease by providing a risk assessment in the patient record
- Check dictated letters and notes. Electronically typed letters and notes have the advantage of legibility, but do have problems related to the quality of recording or misunderstandings of medical terminology. They should be checked and signed by the doctor who dictated them
- Check, evaluate and initial every report or letter before it is filed in the patient's records. For electronic records, any abnormal findings should be documented and action appropriately logged
- Be familiar with the Data Protection Act which allows patients access to view their records or to receive a copy, subject to exceptions.

One additional way to ensure the standard of a record is through a clinical audit.

An audit is “any summary of clinical performance of health care over a specified period of time aimed at providing information to health professionals to allow them to assess and adjust their performance” [194]. It can be used in several different ways depending on the aspects of performance being audited, the interests of the stakeholders and the availability of information. For instance, an audit can be used to check compliance to clinical guidelines. Moreover, an audit can be based on routinely collected data from patient databases or on data collected via patient or clinical surveys.

Before proceeding to the inspection of prostate cancer patient logs, it is important that I validate and test that the patients I have extracted via my linkage algorithm conform to the patients that have actually been seen in the clinic within that time frame as mentioned in the case note forms. The best technique to do that was using a case note audit. The purpose of a database audit is to prove that data have been correctly entered from the paper case report form into the computer

database. It determines the reliability, completeness and comprehensiveness of the data entered. The audit does not address issues and discrepancies written in the case note form, but rather the aim is to find out how much discrepancy is between the case note form and the data extracted or stored in the database.

5.1.1 CONTRIBUTIONS OF THIS CHAPTER

In this chapter I have developed an audit collection tool using ASP script and HTML that would allow the auditor to populate online two-week wait cancer referral and clinical notes forms with selective data retrieved from medical records. With the aid of this online population of a back-end relational database, a statistical comparison was done to validate the patients found through the linkage algorithms. The bespoke audit tool developed can be used to audit any cancer-related care process with a slight tweaking of the variables captured.

5.2 METHODS

In the methodology of conducting my audit, I followed a 3 stage approach for preparing for the audit, selecting the audit criteria and measuring the level of performance. I will describe each stage in detail below.

5.2.1 STAGE 1: PREPARING FOR THE AUDIT

This is the initial and first stage of the audit. In this stage, I identified the problem and also the resources that are available for my audit. After linking and extracting my data and before I proceeded into the final analysis of my event log, I did not know whether the number of patients that my linkage is producing are in fact the same number of patients with the same characteristics as seen in the clinic in that particular time frame. There was no way of cross-checking these details unless I performed a clinical audit to validate the accuracy of my extracted data.

After identifying the problem, I then had to find a way of conducting the audit. As I held an honorary contract with Imperial College NHS trust hospitals, I was not allowed to access the clinical notes of

the patients. For this reason, I had to appoint a clinical fellow in the field of Urology to conduct this audit on my behalf.

5.2.2 STAGE 2: SELECTING CRITERIA

After several meetings with the clinical fellow and the lead clinician, we decided on a sample size of the data that will be a representative subset of the whole database. A sample of the data is audited for the practical reason that a 100% audit would take too much time. We wanted to have explicit selection criteria so that we could ensure that our data was precise and only essential information was collected. Our selection criteria were to measure the number of patients seen retrospectively in the prostate cancer clinic in the 10 months between 03/01/2013 and 03/11/2013. The selection of patients would be random and each case note form will be reviewed to take out only the relevant and essential information as set out by the Prostate Cancer Risk Management Program (PCRMP) and laid out by the NICE guidelines. The minimum number of patients that needed to be extracted was set to 200 patients in order to have a representative sample for those 10 months (as that is roughly how many patients are seen in the clinic for that period).

The following are the comparison metrics that we wanted to test (Table 19):

TABLE 19: PERFORMANCE METRICS FOR AUDIT

	Metric
1	Same number of matching patients found
2	Matching patients have same first date of appointment
3	Matching patients have a similar PSA value
4	Matching patients have the same appointment priority (TWW)

5.2.3 STAGE 3: MEASURING LEVEL OF PERFORMANCE

COLLECT DATA

For the collection of data, since we did not have the resources and time to rely on the hospital's audit team who could help us with the audit, I created a web-based audit collection tool using Active Server Pages (ASP) classic for the back end server-side code and Hyper Text Markup Language (HTML) for the front-end web page design. I used MS SQL Server as the database management system to collect my data. For security of the information held, the tool was protected by a user ID and password combination and kept locally on the NHS trust computer. Moreover, for patient confidentiality reasons, we created pseudo-keys for all the NHS numbers and used them as unique identifiers instead of the NHS numbers. A log of the mapped NHS numbers and pseudo-keys was kept secure separately. These keys were kept securely in the clinical-fellow's hospital desktop. The tool can be used for retrospective or prospective data collection.

The purpose of this audit is to help validate information routinely collected from the databases against what's actually recorded in the notes and outlined in the PCRMP. After a series of meetings with prostate cancer physicians, an initial draft of the audit tool was made on paper, which elaborated what variables needed to be extracted to compose an audit from. Once these variables were finalized by the physicians, we developed two sets of forms containing only the important criteria set out by the national guidelines when receiving patients on a two-week wait referral. Each of these forms is discussed below.

Form 1: Two-week wait Referral Audit Form

In this form I want to capture the fields related to the cancer two-week wait referral of a standard proforma used by GPs (see Appendix E). Moreover, I combined the variables suggested in the PCRMP and NICE guidelines to come up with an online adaptation form. The new online form captures the following information from the paper-based two-week wait referral form:

- Demographic information like: audit no., age, date of decision to refer, etc.
- Referral Indications like: was PSA done with GP, level of referred PSA and date of PSA, etc.

Form 2: Clinical Letter Audit Form

In this form I want to capture the fields found in a standard clinical letter dictated by the physician. These are the fields that contain information summarizing the first clinic appointment for a patient referred via a TWW form. The online form captures the following information:

- Demographic information like: audit no., patient ethnicity, date of clinic, etc.
- Diagnostic information like: Was PSA repeated in clinic, repeat PSA value, was MRI recorded, was biopsy ordered, etc.
- Additional symptoms/confounders like: Previous biopsy, family history of prostate cancer, lower urinary tract symptoms, etc.

An online prototype of the first version of the audit tool was then created. Once the prototype was approved and tested for errors it was subsequently installed on the clinical fellow's computer to start a pilot run. After an initial pilot run for one week, based on the fellow's feedback the prototype was adjusted and errors were fixed. Finally, the application was ready to be launched and used.

The following are screenshots taken from the Prostate Cancer Audit Tool. Figure 33 is the main welcome screen to the tool. Once you log in with your credentials, the main Audit forms screen (Figure 34) appears. From this page you have the option to either start filling out the audit forms with your audit data, editing or viewing a previously entered patient, or then choosing other options from the menu above. Figure 35 and Figure 36 show the two-week wait and clinical letter audit forms.

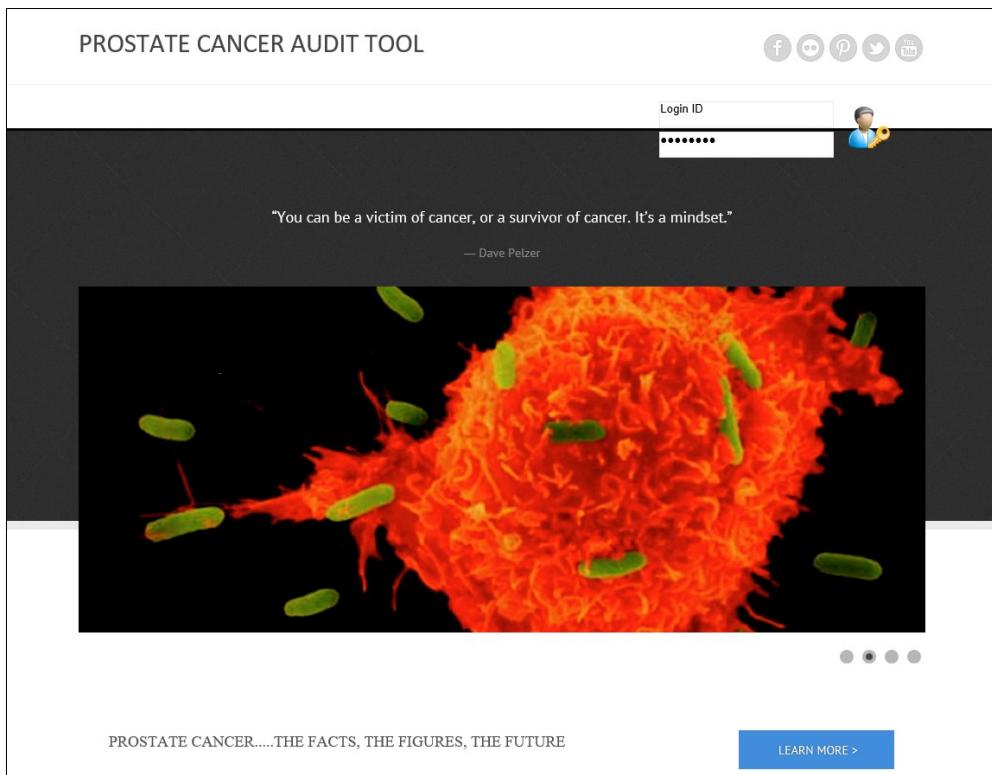


FIGURE 33: WELCOME SCREEN OF THE AUDIT TOOL

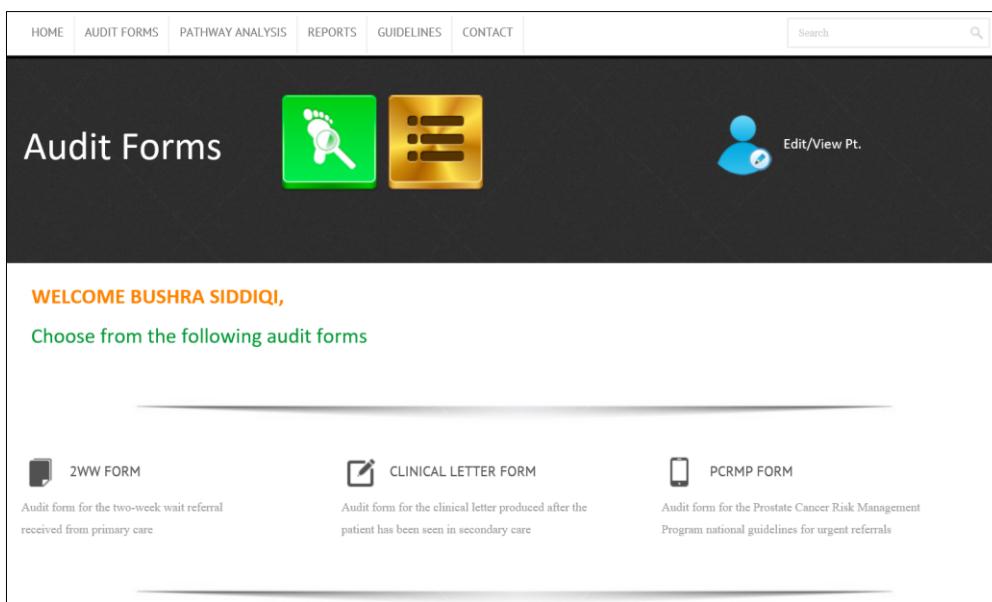


FIGURE 34: MAIN AUDIT FORMS SCREEN

Two Week Wait (2WW) Referral Audit Form

Patient Audit No.*	<input type="text"/>	Was 2WW form used?	<input checked="" type="radio"/> Yes	<input type="radio"/> No
Date of decision to refer*	<input type="text"/> 	Patient Age	- Select - 	
<hr/>				
Was patient informed of suspected prostate cancer?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded	
Was patient given 2WW leaflet?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded	
Was patient told they will be seen within 2 weeks?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded	
Did patient have a previous diagnosis of cancer?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded	
<hr/>				
Referral Indications				
<input type="checkbox"/> On DRE, hard and irregular prostate <input type="checkbox"/> Raised/rising age-specific PSA, with clinically malignant prostate or bone pain, or unexplained urological symptoms <input type="checkbox"/> Asymptomatic with age specific raised PSA in men (<75 years) with negative MSU				
Other Comments	<input type="text"/>			
Was PSA Done?*	<input type="radio"/> Yes	<input checked="" type="radio"/> No	PSA Value	<input type="text"/>
Date of PSA	<input type="text"/> 			
Was PSA repeated? <input type="radio"/> Yes <input checked="" type="radio"/> No				
<small>* Mandatory field</small>				
<input type="button" value="Submit"/> <input type="button" value="Reset"/>				

FIGURE 35: TWW REFERRAL AUDIT FORM

Patient Audit No.*	<input type="text"/>		
Date of clinic*	<input type="text"/>	Patient ethnicity	- Select - <input type="button" value="▼"/>
Patient Age	- Select - <input type="button" value="▼"/>		
<hr/>			
Did patient know implications of PSA/referral?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Was patient given 2ww information leaflet?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Was patient told they will be seen within 2 weeks?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Did patient have a previous diagnosis of cancer?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
<hr/>			
DRE findings	<input type="radio"/> Normal	<input type="radio"/> Abnormal (cancer)	<input type="radio"/> Abnormal (benign) <input type="radio"/> Not Recorded
Was PSA repeated from clinic?	<input type="radio"/> Yes	<input checked="" type="radio"/> No	Repeat PSA Value <input type="text"/>
Was MSU sent from clinic?	<input type="radio"/> Yes	<input checked="" type="radio"/> No	Result <input type="text"/>
Was MRI ordered?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Was prostate biopsy organised?	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
<hr/>			
Additional Symptoms/Confounders			
Lower Urinary Tract Symptoms (LUTS)	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Dysuria/UTI	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Haematuria	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Catheter in situ	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Bone Pain	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Patient on 5 Alpha reductase inhibitor	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Previous prostate biopsy	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Previous TURP	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded
Family history of prostate cancer	<input type="radio"/> Yes	<input type="radio"/> No	<input checked="" type="radio"/> Not Recorded

* Mandatory field

FIGURE 36: CLINICAL LETTER AUDIT FORM

It took the clinical fellow 5 months (due to various personal hindrances) to collect all the data from each individual case note form and populate the two online forms separately with the data extracted manually. The fellow first finished collection of the TWW referral form via the paper-based referral forms found in the patient's medical file, then the clinical letters were retrieved for the same patients (if found) and the second form was populated electronically. The fellow stopped until a minimum collection of 200 patients was achieved within the retrospective 10 month time span.

COMPARE PERFORMANCE WITH CRITERIA

This is the analysis stage. In this stage I compared the data collected and recorded in the electronic audit tool against the criteria and standards set out in stage 2. For the analysis of the data collected by the audit, I first joined the two database tables that separately contained the referral information and the clinical letter information of the patients. This new table now held all the information of the patients from the referral until the time they were first seen in the clinic (if applicable). I then did an inner join between this new table and my entire cohort to find the matching patients based on NHS number. Once the matches were found they were dropped into a table and then further analysed in Stata to see the comparisons. The matching patients were then compared against the metrics set in Table 19 to see if they reached the target or not. A series of different statistical tests were done to see the strength of the match. Moreover, for the patients that did not match in my cohort, I took out a detailed pathway of exactly where their footprints were in the database systems. To further validate my findings, I also got a data quality check done through a registrar on those unmatched patients to see if there was any missing information.

5.3 RESULTS

Process Mining Data source	Cerner (for outpatient appointments), Pathology Database (for PSA)
Time frame	03/1/2013 to 03/11/2013 (10 months)
Audit	An audit was undertaken by a clinical fellow in Nov 2014 focusing on prostate cancer patients. 241 patients were randomly selected and audited, however 3 patients did not have proper NHS numbers in the CRFs and were excluded. So N=238
Data Linkage and Extraction	For the mentioned time frame, the same patients from the audit were identified using data linkage and extraction (DLE) techniques
Aim	To compare results from Audit versus DLE, with a view to demonstrating the effectiveness/accuracy of my DLE techniques.
Comparators	Comparing the results from Audit vs. DLE techniques using: <ul style="list-style-type: none"> ○ Number of matched patients ○ Dates of appointment ○ PSA Results ○ TWW Appointment Priority
Inclusion criteria	Records with PSA value recorded within the clinical information system N=214 (n=27 records were excluded as 3 did not have a proper NHS number and all 27 did not have a PSA recorded in my DLE)

My audit has involved a retrospective case note review of 238 paper-based patient records. Only patients that had been referred onto the caseload from 03/01/2013 until 03/11/2013 with a valid NHS number were reviewed.

5.3.1 THE AUDIT DATA

The following tables (Table 20 and Table 21) give the descriptive results of the data extracted from the case note forms during the audit.

TABLE 20: TWW REFERRAL AUDIT FORM DATA

	Frequency	Percentage
No. of patients referred = 238		
Median age of patients	71 years	-
No. of standard TWW form referrals	191	79%
No. of patients with PSA done with GP	222	92%
No. of patients informed of suspected prostate cancer	106	44%
No. of patients given TWW leaflet	50	21%
No. of patients told they will be seen within 2 weeks	147	61%
No. of patients with previous diagnosis of cancer	18	7%

TABLE 21: CLINICAL LETTER AUDIT FORM DATA

	Frequency	Percentage
No. of patients seen in clinic = 233		
No. of PSAs ordered in clinic	233	100%
No. of MSUs sent from clinic	233	100%
No. of MRIs ordered in clinic	66	28%
No. of biopsies ordered in clinic	115	49%
No. of patients with (Lower Urinary Tract Symptoms) LUTs	146	63%
No. of patients with Dysuria/UTI	28	12%
No. of patients with Hematuria	20	9%
No. of patients on 5 Alpha Reductase Inhibitor	7	3%
No. of patients with bone pain	12	5%

5.3.2 THE AUDIT DATA AGAINST DLE

The following sections show the results of the matches and targets met for the performance metrics when I compared the case note audit data against my date linkage and extraction (DLE) patients.

METRIC 1: TOTAL NUMBER OF RECORDS FOUND

For my first metric, there were 214 records found out of 238. This was a 90% match (Table 22).

TABLE 22: TOTAL NUMBER OF MATCHING RECORDS FOUND

Total No. of Records found via DLE technique	214
Percentage of total found in Audit (N=238)	90%

METRIC 2: MATCHING PATIENTS HAVE SAME FIRST DATE OF APPOINTMENT

In my second metric, I want to measure whether the Audit and DLE first dates of appointment have an expected match of 75%. The following Table 23 summarises the matches and non matches in both the datasets.

TABLE 23: FIRST APPOINTMENT DATE MATCHES BETWEEN AUDIT AND DLE DATA

Match	N	(%)
Exact match	163	(76.2)
Non exact Match	51	(23.8)
Total	214	(100)

The following chart (Figure 37) displays the days difference against the percentage of patients where N=214. The maximum number of patients (76.2%) had the exact same appointment dates (difference=0)

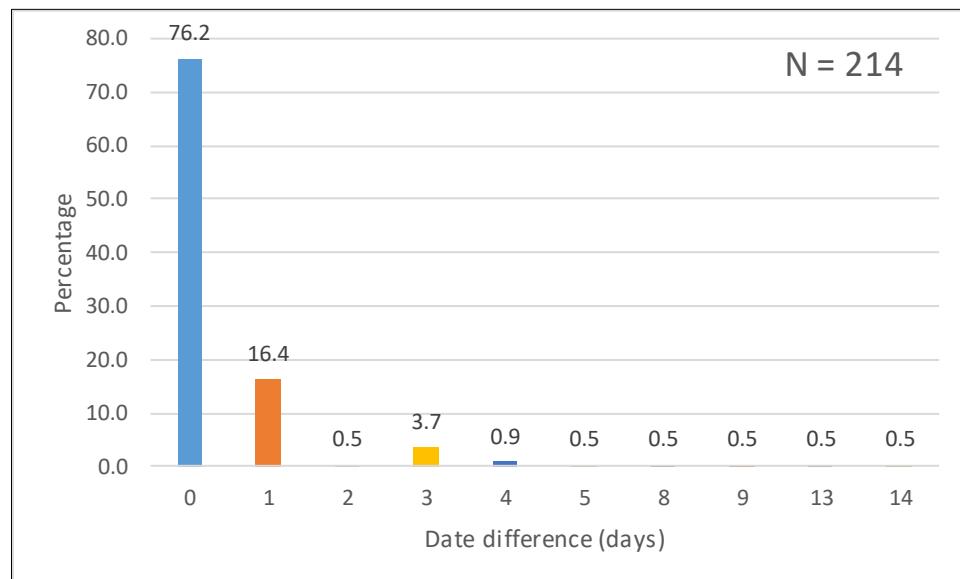


FIGURE 37: BAR CHART OF DAYS DIFFERENCE BETWEEN THE AUDIT AND DLE DATA

METRIC 3: MATCHING PATIENTS HAVE A SIMILAR PSA VALUE

To test this metric, I selected all values where the PSA is not null for both Audit and DLE datasets (N=189). I first measured the difference between the PSA values. The following table (Table 24) summarises the difference in numbers between the PSA values found in Audit and DLE.

TABLE 24: DIFFERENCE BETWEEN THE PSA VALUES FOUND

Difference	N	(%)
0	186	(98.41)
4	1	(0.53)
10	1	(0.53)
16	1	(0.53)
Total	189	(100)

METRIC 4: MATCHING PATIENTS HAVE THE SAME APPOINTMENT PRIORITY (TWW)

The following table (Table 25) summarises the Two-week wait (TWW) and non-TWW appointments found in the Audit and DLE groups

TABLE 25: SUMMARY OF TWW AND NON-TWW APPOINTMENTS BETWEEN THE AUDIT AND DLE GROUPS

Group	Non-TWW	Yes-TWW	Total	Percentage
Audit	39	175	214	81.7%
DLE	5	209	214	97.6%
Total	44	384	428	

5.4 DISCUSSION

Data validation has been an important step in verifying that the data I am extracting are fit for their intended use. The underlying reason to get valid data after performing an audit is that data present in the clinical databases will affect my process mining analysis, which in turn will affect treatment decisions and outcomes, which lastly and most importantly will affect patient healthcare.

The reason I created a web-based electronic Prostate Cancer Audit Tool was to make it easier in the future to add more patients, retrieve and edit patients in a centralised way. Moreover, with the original concept in mind, this audit tool was intended to be used for reporting purposes as well. However, due to the scope and timing constraints of my PhD, I could not continue making the reporting module.

As the data was kept centrally located on the hospital server, any changes or updates made by the auditor was immediately reflected in the back end. Once the data was collected in an SQL database from the back-end, it was easier to join it with my own cohort tables and start the comparison.

I got a 90% match between the records found using the DLE method against the Audit method; a 76.2% match between the first dates of appointment in both the groups; a 98.4% match between the PSA values in both the groups; and a higher number of 2 week waits than the Audit (97.6% for DLE vs. 81.7% for Audit). For the higher number of 2 week wait records, there are a number of possible explanations for this: It can be due to someone going back and re-entering the data, or it could also be that the TWW priority entered was wrong and was only used as a way of getting the diagnostic tests done faster.

For the 24 patients that did not match with the DLE group, the following is an explanation:

As my cohort is designed to capture patients who have ever had a PSA in their entire medical journey within my timeframe, for the above 24 patients I could not find a PSA in the pathology system I had access to and hence they were not part of my cohort. However, upon requesting a registrar for a quick data quality check on these 24 patients, I have deduced the following:

- There were 4 patients who had a PSA written in clinic and they actually had a PSA done in the labs however they did not show up in the pathology database that I used to extract my

- patients. From those 4 patients, 2 had a DNA, 1 was discharged for no cancer and 1 did not have any MDT meeting.
- There were 10 patients for whom a PSA was written in clinic, but they actually never got a PSA done in labs. From those 10 patients, 1 had no records found in data quality check, 4 patients had DNA, 1 patient had no MDT and was referred back to another hospital, 1 patient was sent back to come as routine appointment, 1 patient had no comments in MDT diagnosis, 1 patient had prostate cancer and was referred from Kent and 1 patient was diagnosed with no cancer.
- There were 8 patients who did not have a PSA written in clinic neither did they have any PSA done from the data quality check. From these 8 patients, 7 had a DNA and 1 was not fit for NHS treatment and was sent back.
- There were 2 patients who had no PSA written in clinic for them, but they actually did a PSA as per data quality check. From those 2 patients, 1 had DNA and the other private patient had cancer but his pathway was closed.

Overall, from the unmatched records, 2 had cancer but for both those patients my pathology database did not have a PSA registered and hence I could not pick them up with my algorithm.

5.5 SUMMARY

This chapter continued on to the third phase of the CPAM road map which is the data validation phase. In this chapter I discussed my methodology in preparing for a case not audit that was required to validate the patients I extracted through my data linkage technique. I discussed my selection criteria as well as how I developed a collection tool to gather case note data related to the two-week wait form and the clinical letter form. I then compared the statistics between the patients from the audit and patients from my data linkage and extraction technique to get an idea of how well my algorithm succeeded in linking and extracting the patients. The next chapter, Chapter 6: Descriptive Results, gives us a first look at the event log by providing basic statistics on the nature of the log that will aid in filtering it to only contain the relevant details for process mining.

CHAPTER 6: DESCRIPTIVE RESULTS

This chapter provides a first look at the basic statistics of the event log that would help in filtering the log and preparing it for process mining. Section 6.2 presents the log inspection (descriptive) results of the 5-year and 2-year cohorts. Section 6.3 presents the limitations of the descriptive results. Chapter 6 is concluded with section 6.4 that gives a summary of the entire chapter and how the following chapter is linked.

6.1 INTRODUCTION

After the log preparation has been completed from section 3.4.3, the next step of the methodology was to do a first inspection of the log. This meant that I needed to familiarize myself with the content of the log. This included gathering basic statistics of the log such as the total number of events in the log, the average number of activities per case with their minimum and maximum values and the number of occurrences of specific activities. The overview of the statistics alongside with the inspection of the log helps in filtering the log down. The demographic breakdown is important because it is important to know if there are any inequalities between the cohorts produced and if the analysis I am doing generalizable to other populations.

6.1.1 CONTRIBUTIONS OF THIS CHAPTER

In this chapter I have presented the initial basic statistics on the extracted event log. The Log inspection covers two cohorts divided by date range. The first cohort spans 5 years (from 2010-2015) and the second cohort spans two years (from 2013-2015). The results include graphs and charts related to the patient demographics, as well as the referral, diagnostics and treatment cycles of the patient's prostate cancer pathway.

6.2 LOG INSPECTION RESULTS

I have divided my descriptive results into two sections: the 5-year cohort and the 2-year cohort. As can be seen from Figure 38, my complete cohort started off by any patient having a PSA test done in

the 5 years from 01/01/2010 until 01/01/2015. The first initial look at the process model of the complete 5-year event log resulted in a heavily cluttered, unreadable process flow diagram. Hence, to reduce the spaghetti diagram, I performed filtering of the log based on years and the start of a GP appointment (TWW/Urgent or Routine) within the timeframe of the years chosen. The reason why I chose the start of the first GP appointment as a mandatory starting point for my event log is because of the innumerable number of patients that had a pathway being carried over into my date range from the previous years (outside my range) and were showing as cases without proper starting points. Moreover, there were many patients that came in as walk-in or routine patients, had their PSA done as part of a regular check-up and left. Therefore, in order to avoid including these kind of patients who probably never developed cancer, or who had their cancer pathway started at a previous point in time outside my range, I decided to put a filter to include only the cases that began their journey inside my time frame with a proper GP referral (either TWW, urgent or routine) and ended up having either a prostatectomy or radiotherapy to ensure that we are capturing only the patients who had prostate cancer.

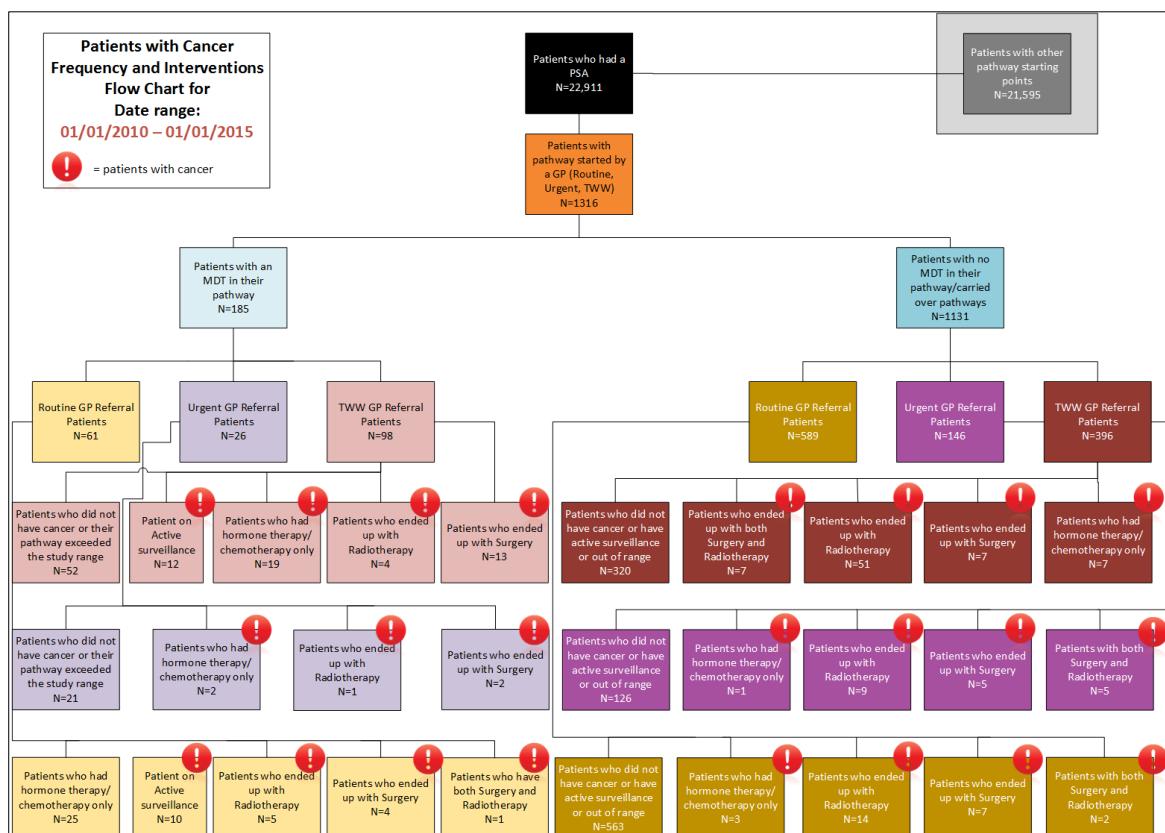


FIGURE 38: 5-YEAR COHORT CANCER FREQUENCY AND INTERVENTIONS FLOW CHART

6.2.1 FIVE (5)-YEAR COHORT

The following are descriptive results of 1316 patients undergoing a GP routine/TWW/urgent referral in the 5 years between 01/01/2010 to 01/01/2015.

Figure 39 lists the count of patients for each activity broken down by year. The view is filtered on activity. It can be seen that MDT information is missing from years 2010-2012, and imaging information is missing for years 2010, 2011, and 2013.

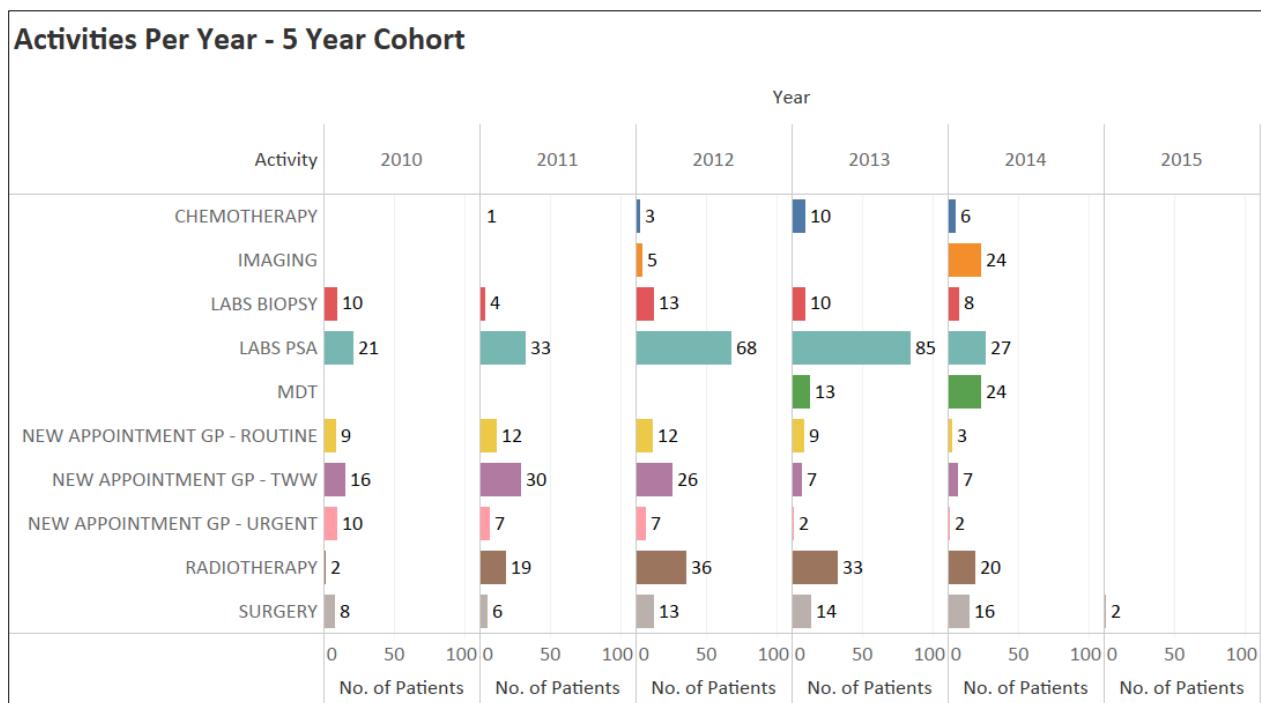


FIGURE 39: ACTIVITIES/YEAR IN 5-YEAR COHORT

PATIENT DEMOGRAPHICS

Starting off with the patient demographics, the total number of patients that were referred by a GP was **1316**. The average age of the referred patients in the 5-year cohort was **67 years**.

Figure 40 shows the distribution of patients based on postal code. The majority of the patients live in the W9 (brown dot) or Maida Vale area.

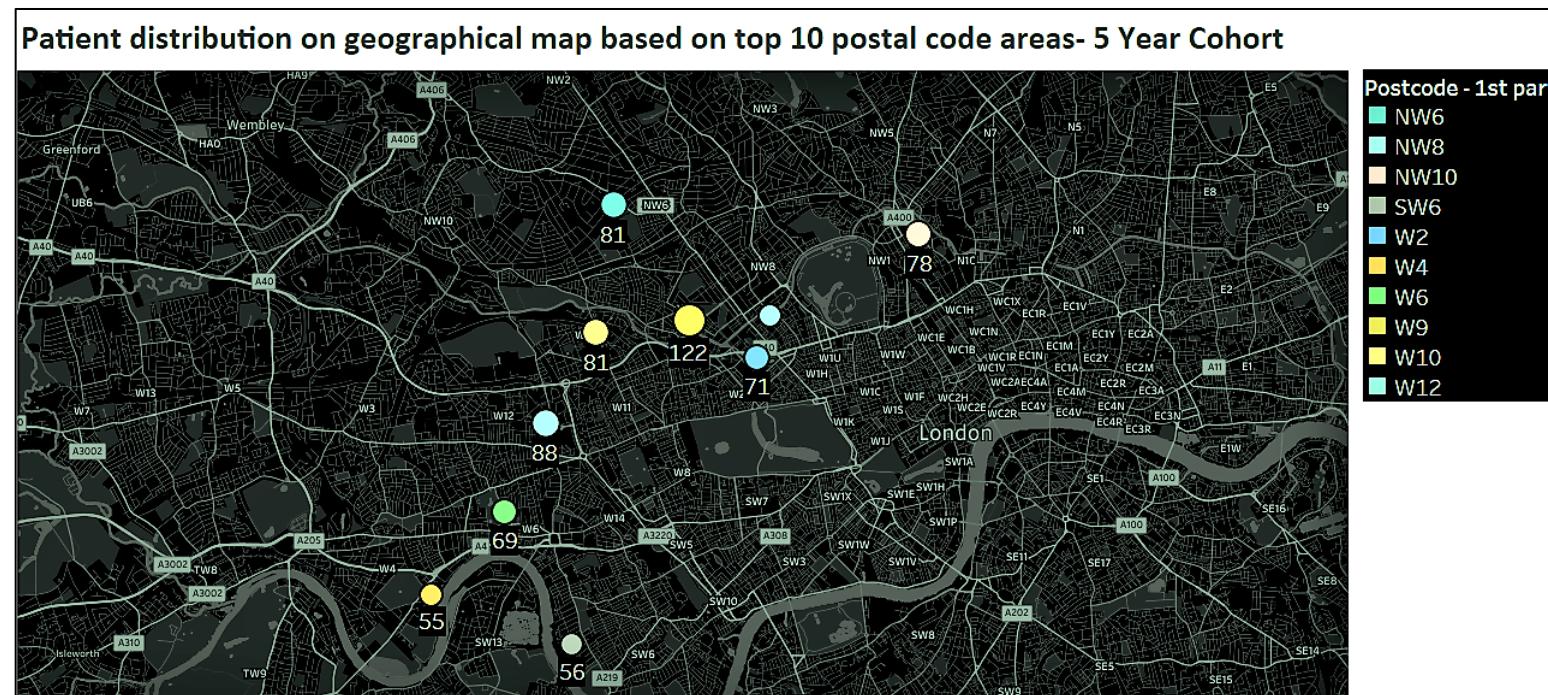


FIGURE 40: PATIENT DISTRIBUTION BASED ON POSTAL CODE IN 5-YEAR COHORT

Figure 41 shows the percentage of distinct patients for each age group. The view is filtered on Age Group, which keeps 50-59, 60-69 and 70-79. Percentages are based on the whole table. Patients belonging to the 70-79 group made up 38% of the patients.

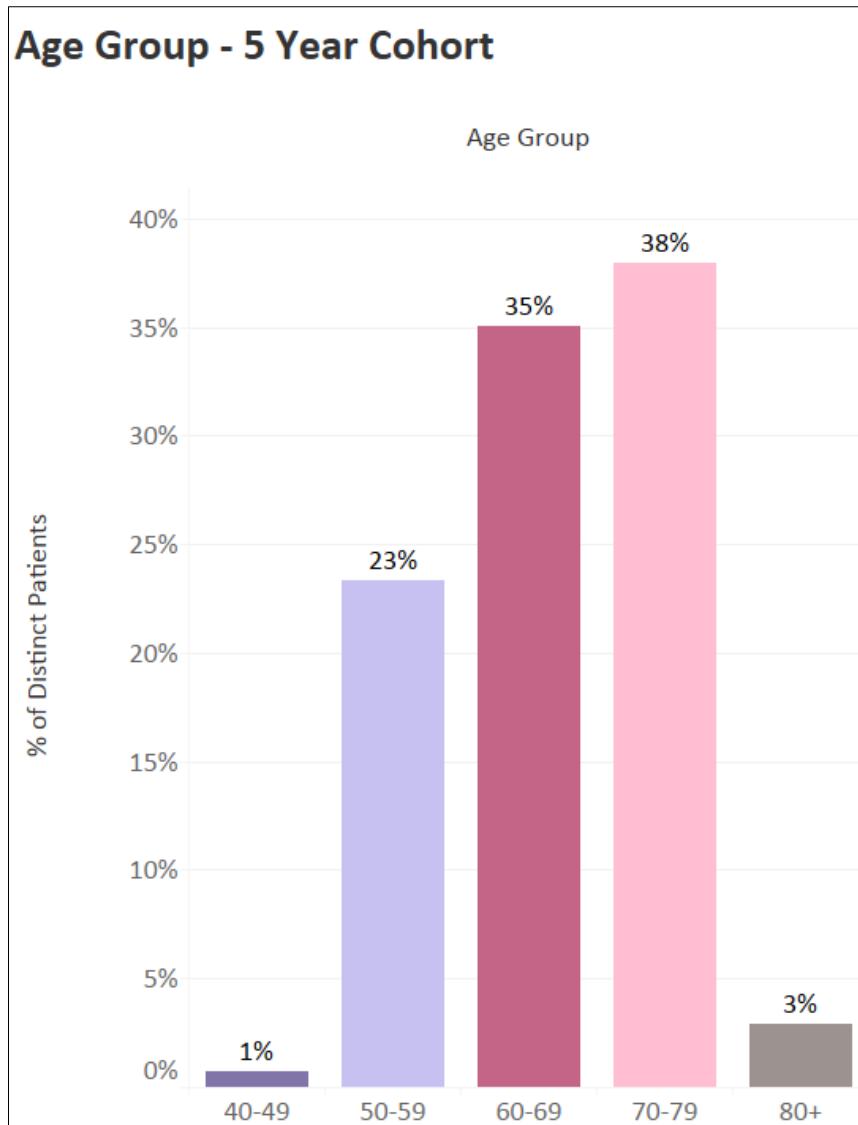


FIGURE 41: AGE GROUP DISTRIBUTION OF PATIENTS IN 5-YEAR COHORT

Figure 42 shows percentage of distinct patients for each marital status. The view is filtered on marital status, which keeps Married, Separated and Single. Percentages are based on the whole table. Married patients made up 78% of the cohort.

Figure 43 shows percentage of distinct patients for each ethnicity. The view is filtered on top 10 ethnicities, which has multiple members selected. Percentages are based on the whole table. 52% of the patients were of British ethnicity.

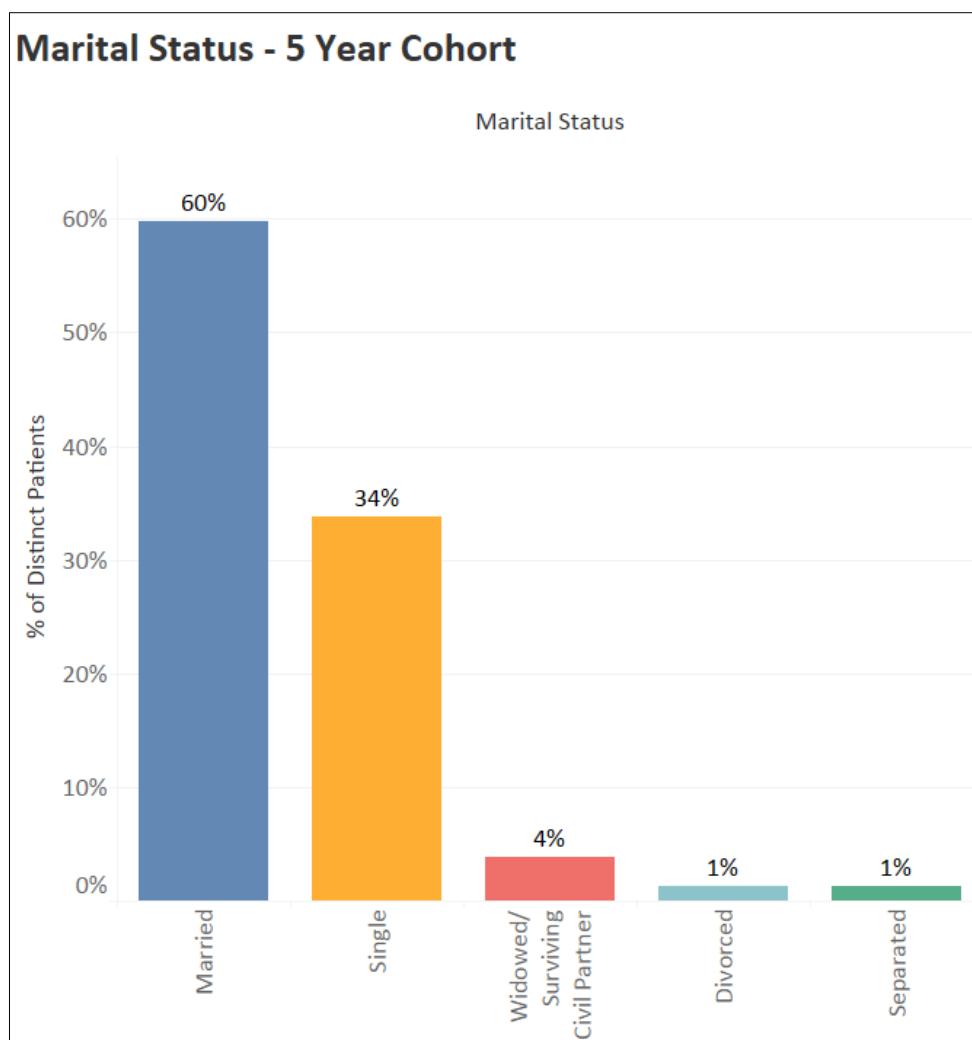


FIGURE 42: MARITAL STATUS OF PATIENTS IN 5-YEAR COHORT

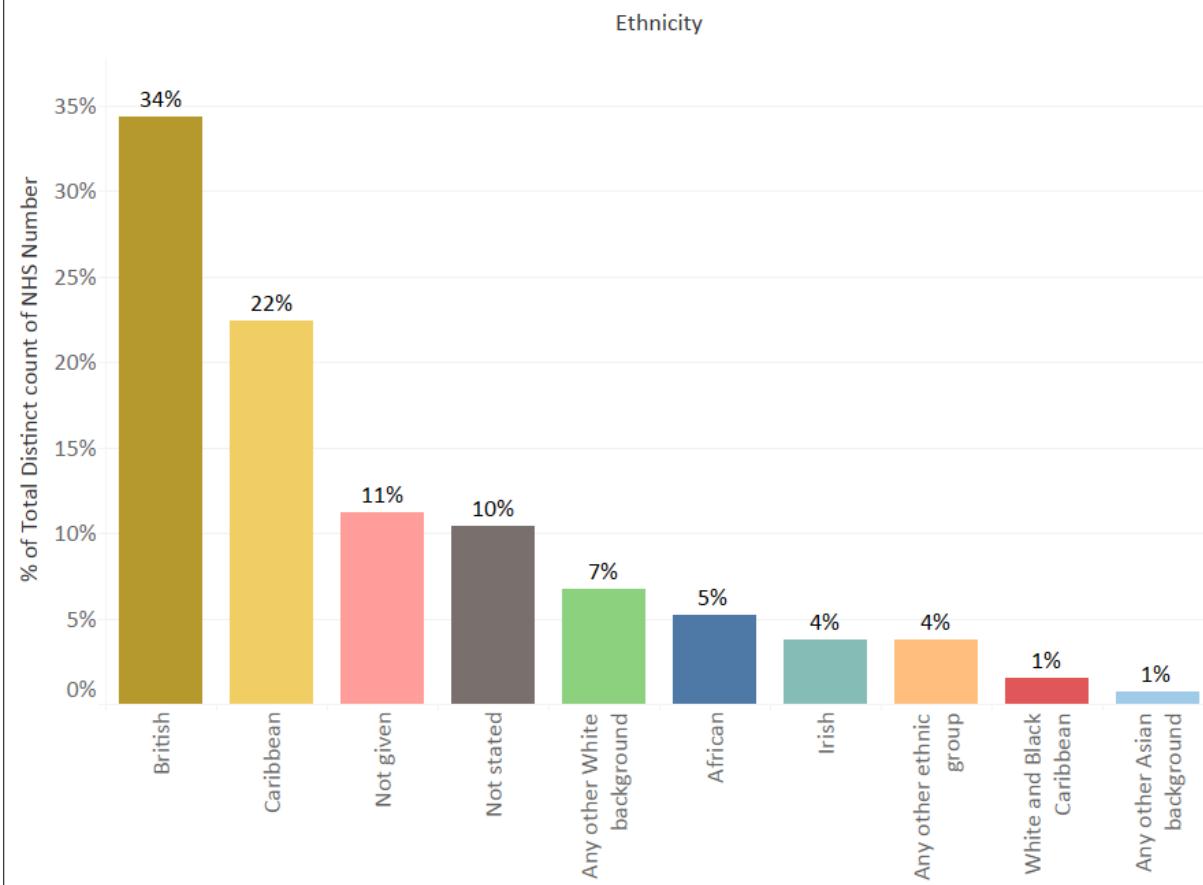
Top 10 Ethnicities - 5 Year Cohort


FIGURE 43: ETHNICITY IN 5-YEAR COHORT

Figure 44 shows percentage of distinct patients for each religion. The view is filtered on religion, which keeps Christian, Muslim and Declines to Disclose. Christians make up 89% of the cohort.

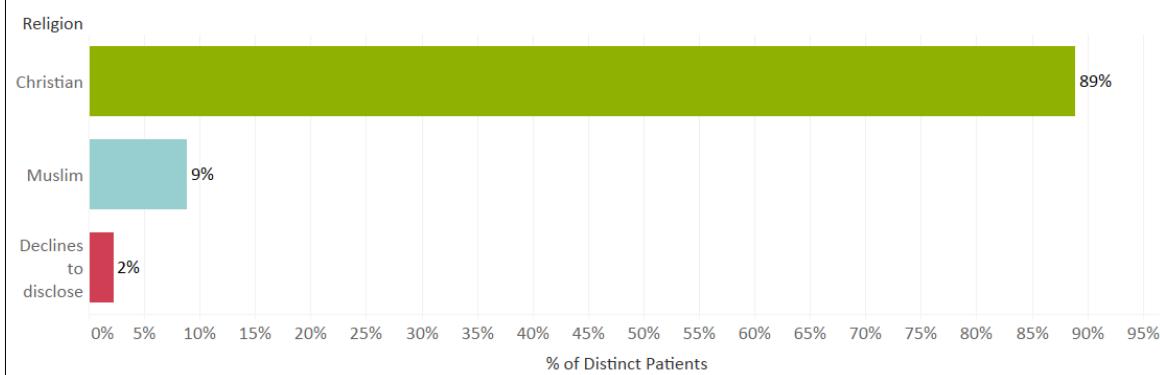
Religion - 5 Year Cohort


FIGURE 44: RELIGION IN 5-YEAR COHORT

PATIENT REFERRAL PHASE

Figure 45 shows the percentage of absolute count of patients for each appointment referral priority. The view is filtered on appointment referral priority, which has multiple members selected. Percentages are based on the whole table. 53% of the referrals were routine referrals followed by TWW that was 29%.

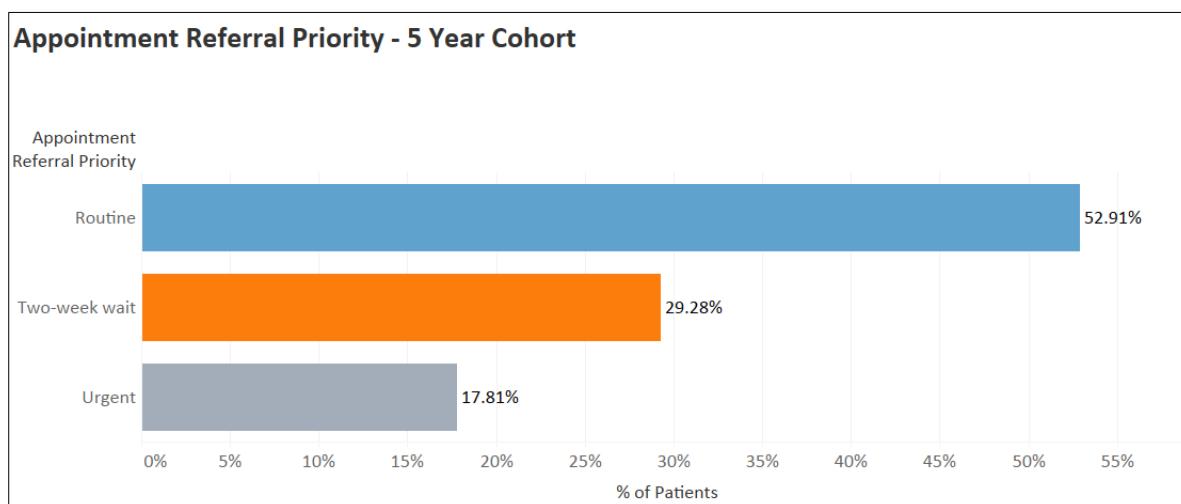


FIGURE 45: APPOINTMENT PRIORITIES IN 5-YEAR COHORT

Figure 46 shows percentage of absolute count of patients for each appointment type. The view is filtered on appointment type, which keeps Follow-up and New. The majority of patients (65%) had follow-up appointments in their pathway.

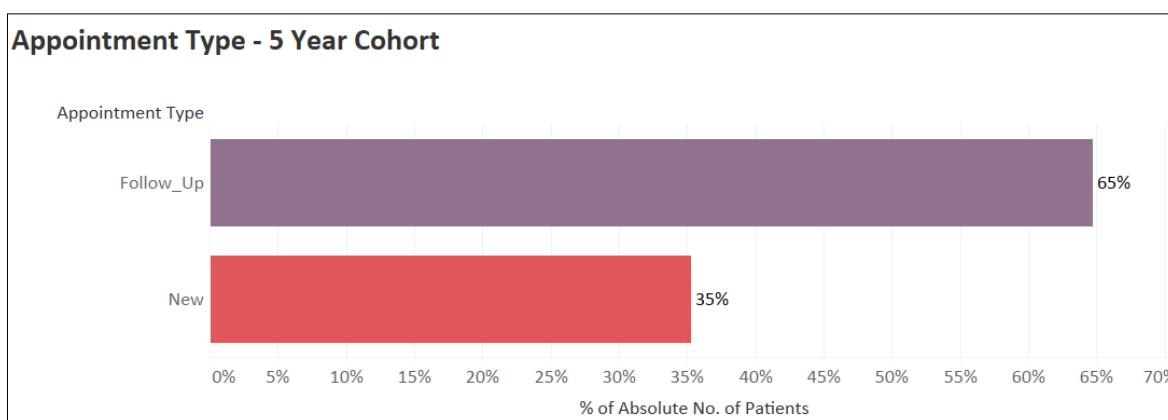


FIGURE 46: APPOINTMENT TYPE IN 5-YEAR COHORT

Figure 47 shows percentage of absolute count of patients for each referring source. The view is filtered on referring source, which has multiple members selected. The majority of the patients (56%) had a GP referral in their pathway

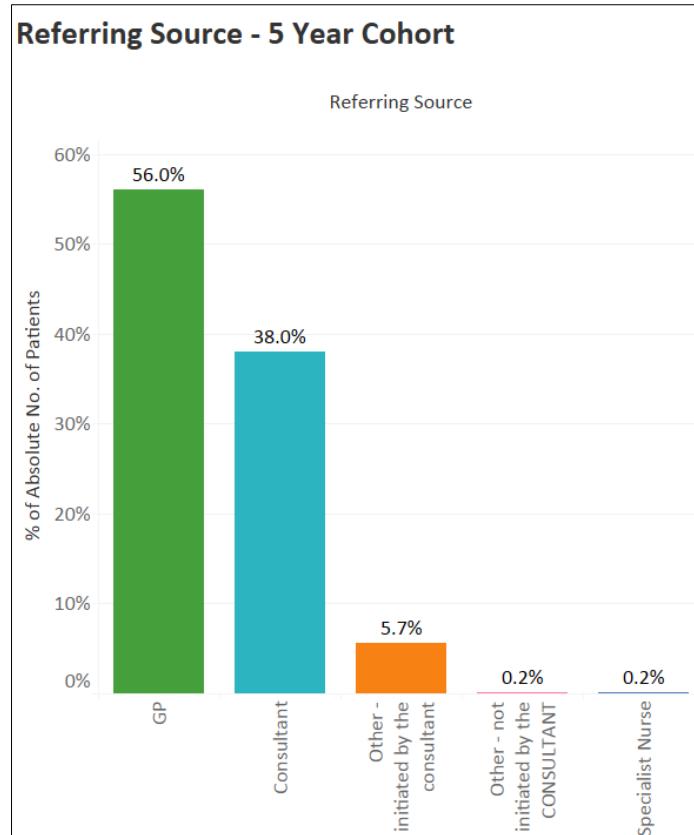


FIGURE 47: REFERRING SOURCE IN 5-YEAR COHORT

PATIENT DIAGNOSTICS PHASE

Figure 48 shows the trend of the distinct count of PSAs/year. Figure 49 shows the trend of distinct count of biopsies/year. It can be seen that the maximum PSAs were registered in 2013 and the maximum biopsies were registered in 2012 as well.

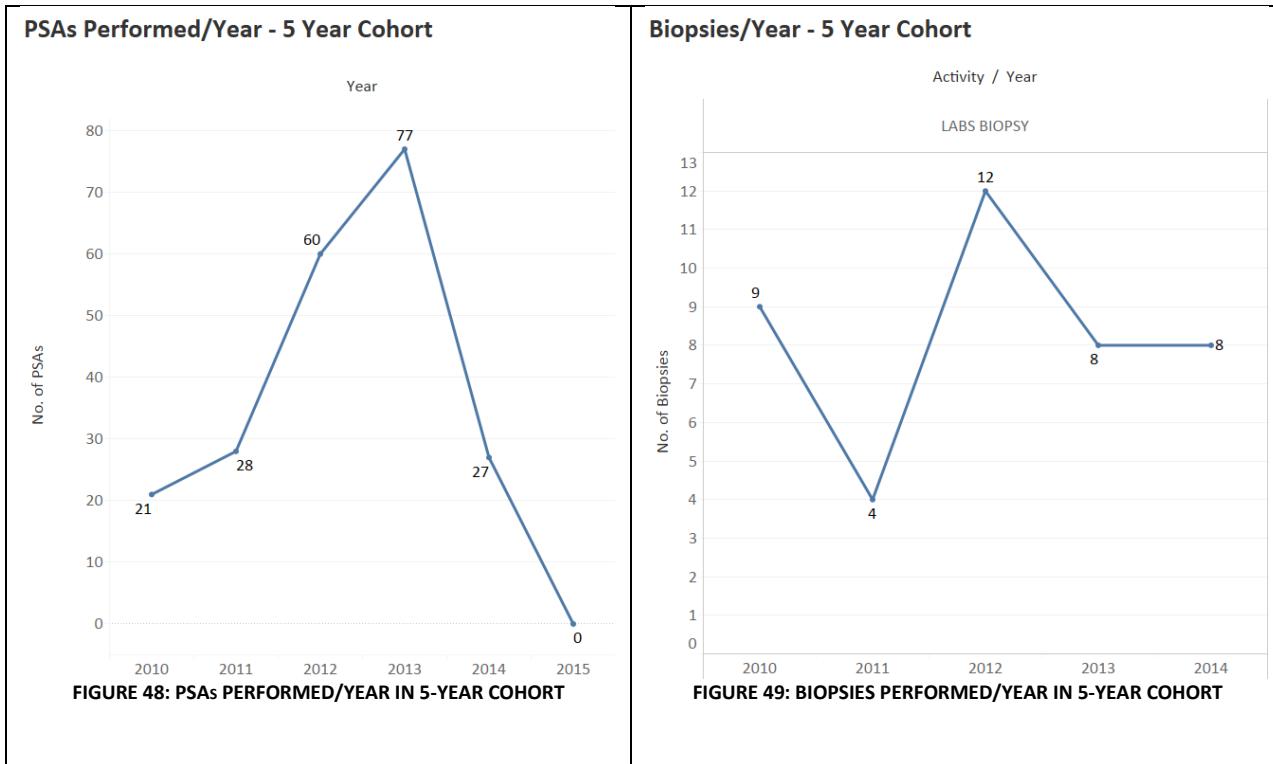


Figure 50 shows a distinct count of radiological procedures/year. The view is filtered on radiology procedure, which has multiple members selected. The majority of the patients had an MRI of the pelvis for prostate cancer in 2014. Note, the imaging information is missing for 2010, 2011 and 2013.

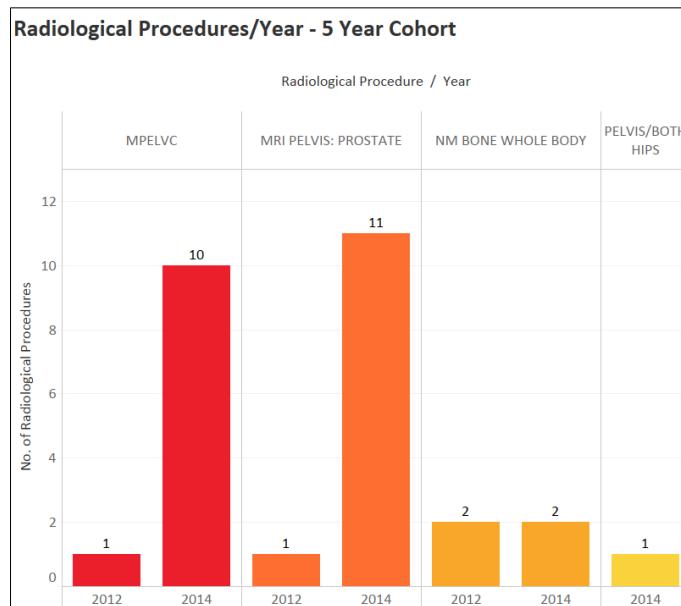


FIGURE 50: RADILOGICAL PROCEDURES PERFORMED/YEAR IN 5-YEAR COHORT

PATIENT TREATMENT PHASE

Figure 51 shows a distinct count of radiotherapy procedures (EBRT)/year. The view is filtered on radiotherapy procedure and keeps the EBRT procedure. It can be seen that the majority of radiotherapies were performed in 2012.

Figure 52 shows a distinct count of surgical procedures (Prostatectomy)/year. The view is filtered on surgery procedure and keeps the Prostatectomy procedure. It can be seen that the majority of the surgeries were performed in 2014.

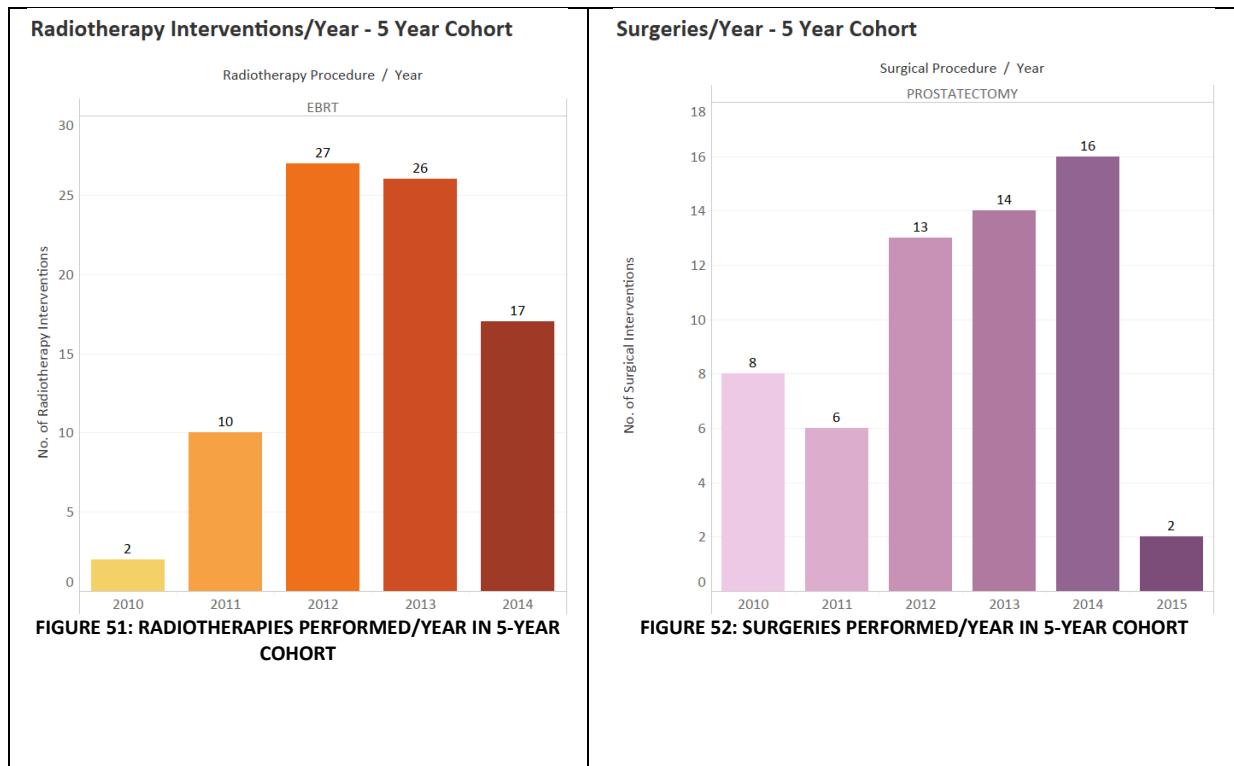


Figure 53 shows a distinct count of chemotherapies/hormone therapies/year. The view is filtered on Chemotherapy/hormone therapy drug, which has multiple members selected. The drug used most frequently was Bicalutamide.

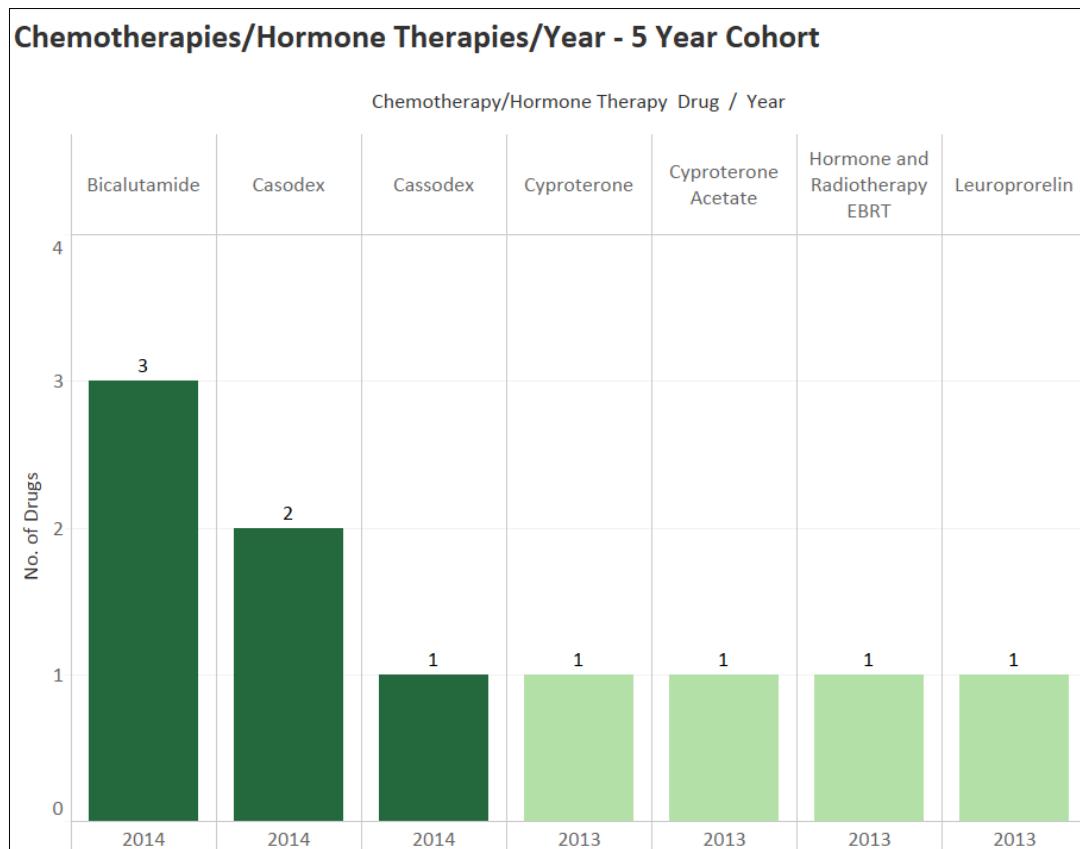


FIGURE 53: CHEMOTHERAPIES/HORMONE THERAPIES PERFORMED/YEAR IN 5-YEAR COHORT

6.2.2 TWO (2)-YEAR COHORT

As can be seen from Figure 39, big limitation in my study was the availability of MDT data before 2013. This was because there was an issue with MDT data quality and the commissioners were not allowed to submit. A large part of the biopsy data and Imaging data was also missing as the data warehouse had long-term issues in acquiring this information from the custodians of the database. As a result, in my 5-year cohort, I noticed a lot of missing information and unwarranted deviations from the norm specifically due to the lack of MDT data. Moreover, due to the missing MDT information, it was becoming difficult to differentiate whether the patient was on active surveillance or did their pathway get genuinely delayed. As every patient who is seen in the cancer clinic must go through an MDT, in order to overcome my problems in the 5-year cohort, I decided to only analyse patients in the two years from 2013-2015 in which I am guaranteed that every patient went through

an MDT discussion. The following are descriptive results of patients undergoing a GP routine/TWW/urgent referral in the 2 years between 01/01/2013 to 01/01/2015

Figure 54 lists the count of patients for each activity broken down by year. Colors represent the different activities. The view is filtered on activity, which excludes F/U APPOINTMENT, NEW APPOINTMENT CONS - ROUTINE and NEW APPOINTMENT CONS – URGENT as these activities were causing additional cluttering and were therefore removed from my process mining analysis. It can be seen that imaging information is still missing from year 2013.

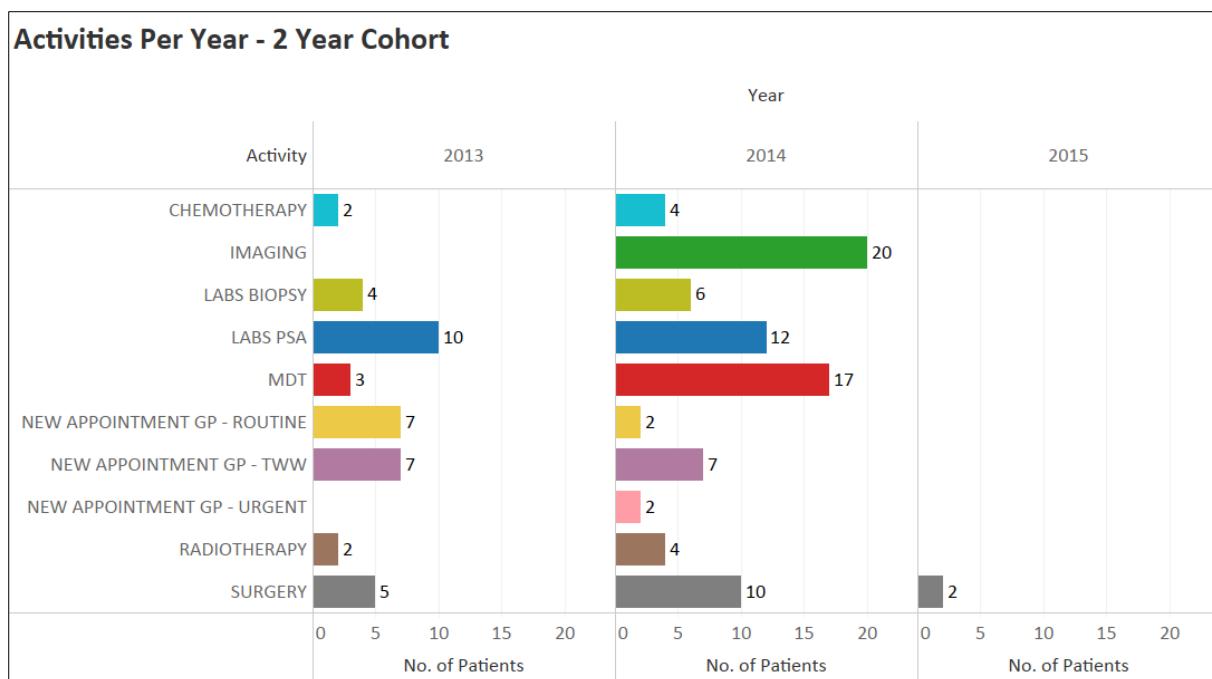


FIGURE 54: ACTIVITIES/YEAR IN 2-YEAR COHORT

As can be seen from Figure 55, my 2-year cohort started off by any patient having a PSA test done in the 2 years from 01/01/2013 until 01/01/2015. In order to avoid including patients whose timeline started with activities other than a GP appointment, I performed filtering of the log based on years and the start of a GP appointment (TWW/Urgent or Routine) within the timeframe of the years chosen.

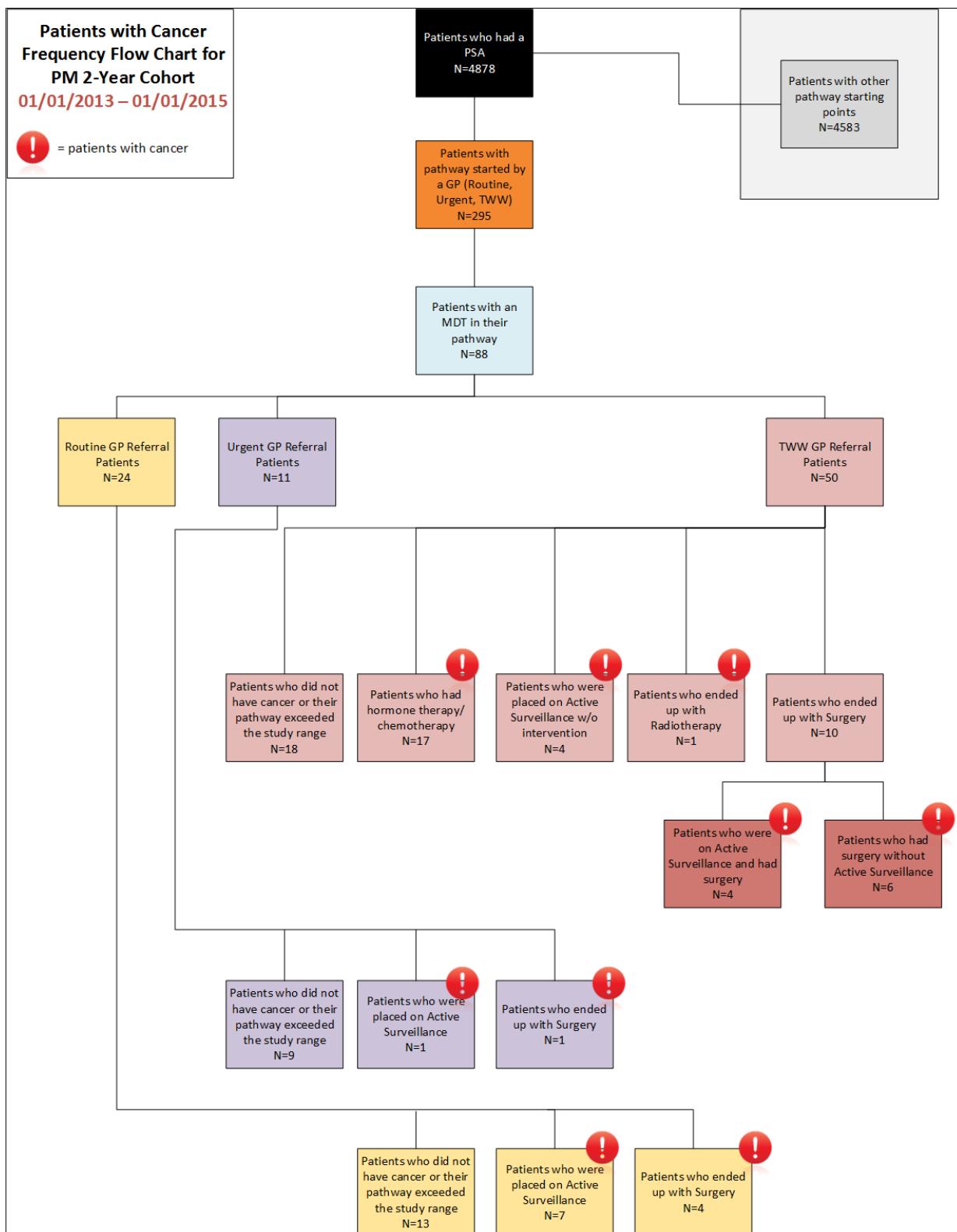


FIGURE 55: 2-YEAR COHORT CANCER FREQUENCY AND INTERVENTIONS FLOW CHART

PATIENT DEMOGRAPHICS

Starting off with the patient demographics, the total number of patients that were referred by a GP was **295**. The average age of the referred patients in the 2-year cohort was **65**.

Figure 57 shows the distribution of patients based on postal code. The majority of the patients live in the NW6 (orange dot) or Kilburn/West Hampstead area.

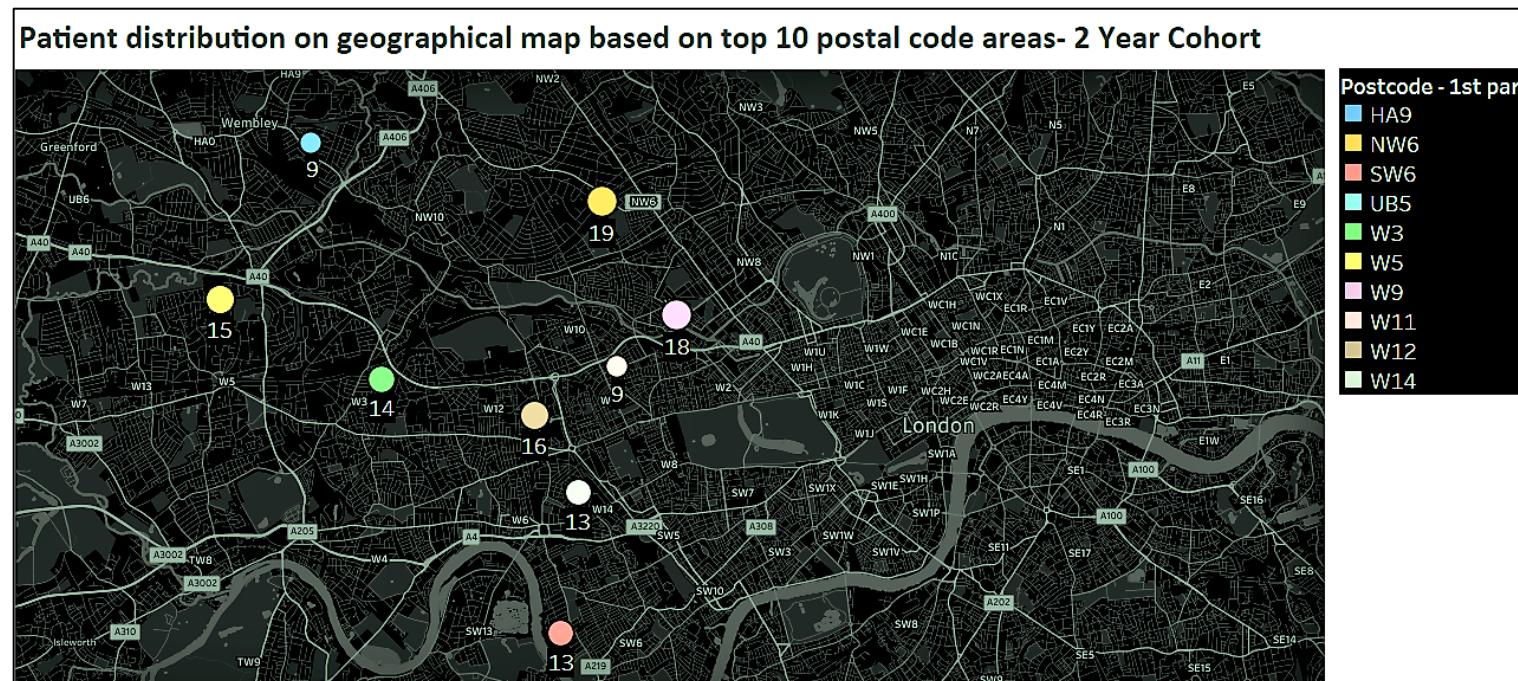


FIGURE 56: PATIENT DISTRIBUTION BASED ON POSTA CODE IN 2-YEAR COHORT

Figure 57 shows the percentage of distinct patients for each age group. The view is filtered on Age Group, which keeps 50-59, 60-69 and 70-79. Percentages are based on the whole table. Patients belonging to the 60-69 group made up 43% of the patients.

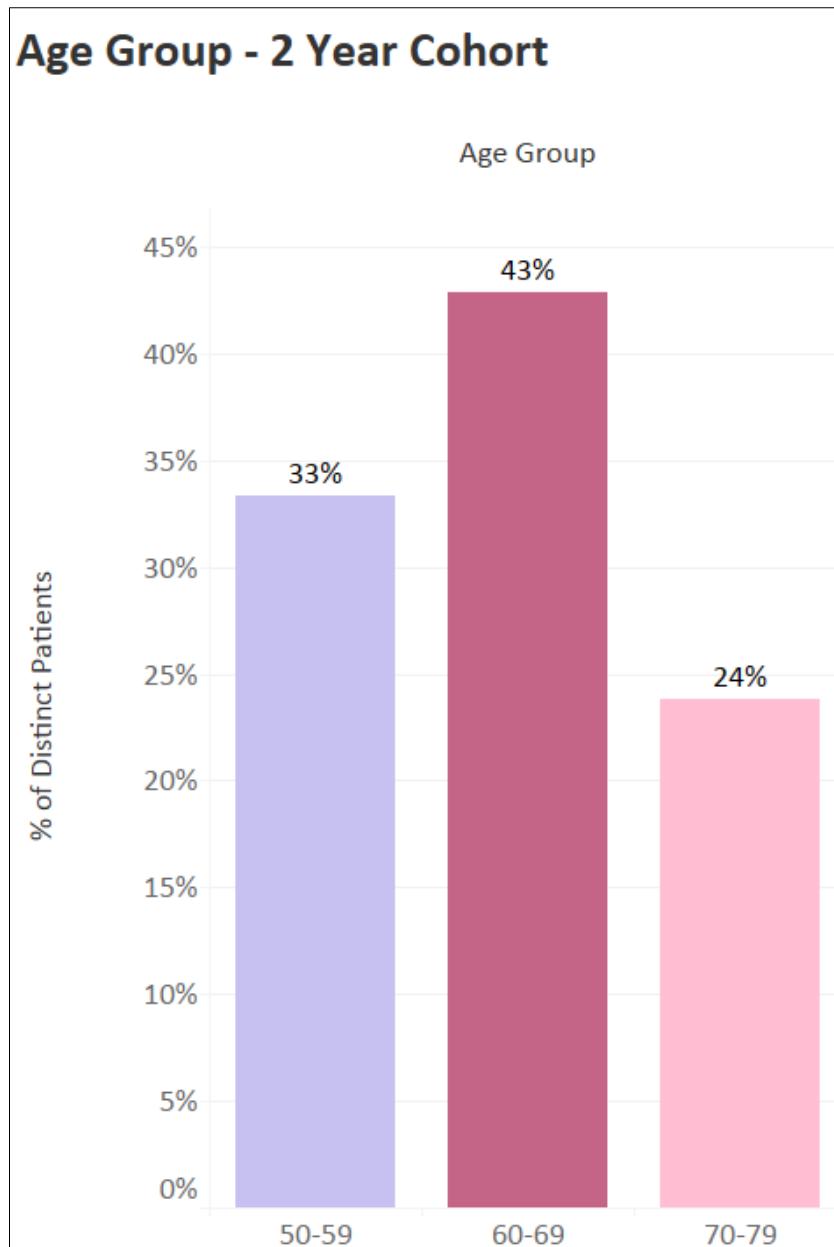


FIGURE 57: AGE GROUP DISTRIBUTION OF PATIENTS IN 2-YEAR COHORT

Figure 58 shows percentage of distinct patients for each marital status. The view is filtered on marital status, which keeps Married, Separated and Single. Percentages are based on the whole table. Married patients made up 78% of the cohort.

Figure 59 shows percentage of distinct patients for each ethnicity. The view is filtered on top 6 ethnicities, which has multiple members selected. Percentages are based on the whole table. 52% of the patients were of British ethnicity.

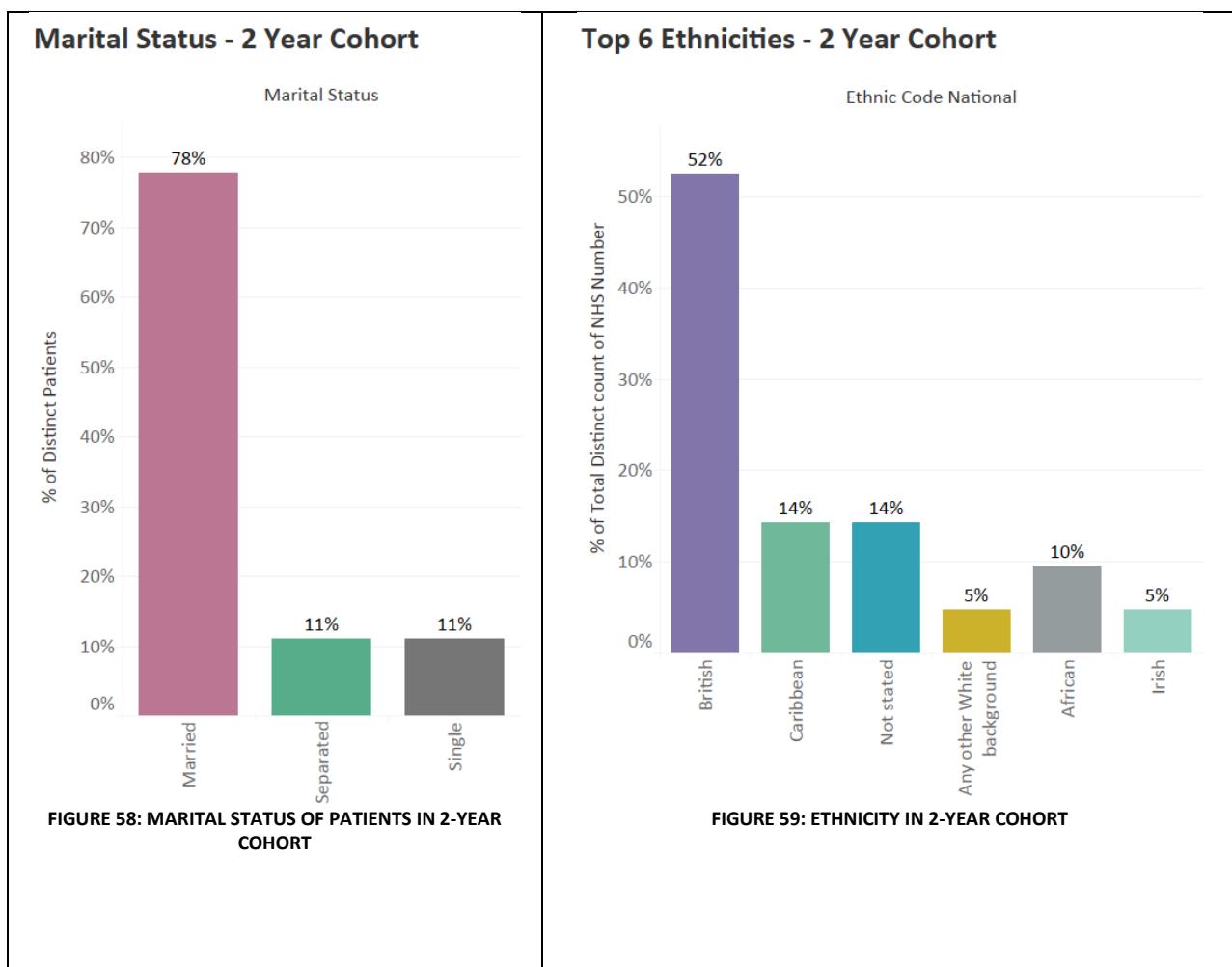


FIGURE 58: MARITAL STATUS OF PATIENTS IN 2-YEAR COHORT

FIGURE 59: ETHNICITY IN 2-YEAR COHORT

Figure 60 shows percentage of distinct patients for each religion. The view is filtered on religion, which keeps Christian and Other. Christians make up 86% of the cohort.

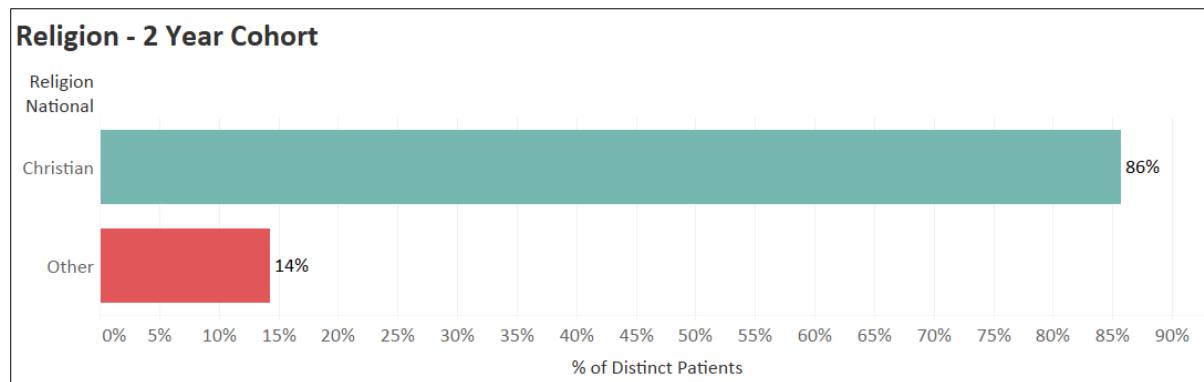


FIGURE 60: RELIGION IN 2-YEAR COHORT

PATIENT REFERRAL PHASE

Figure 61 shows the percentage of absolute count of patients for each appointment referral priority. The view is filtered on appointment referral priority, which has multiple members selected. Percentages are based on the whole table. 53% of the referrals were routine referrals followed by TWW that was 31%.

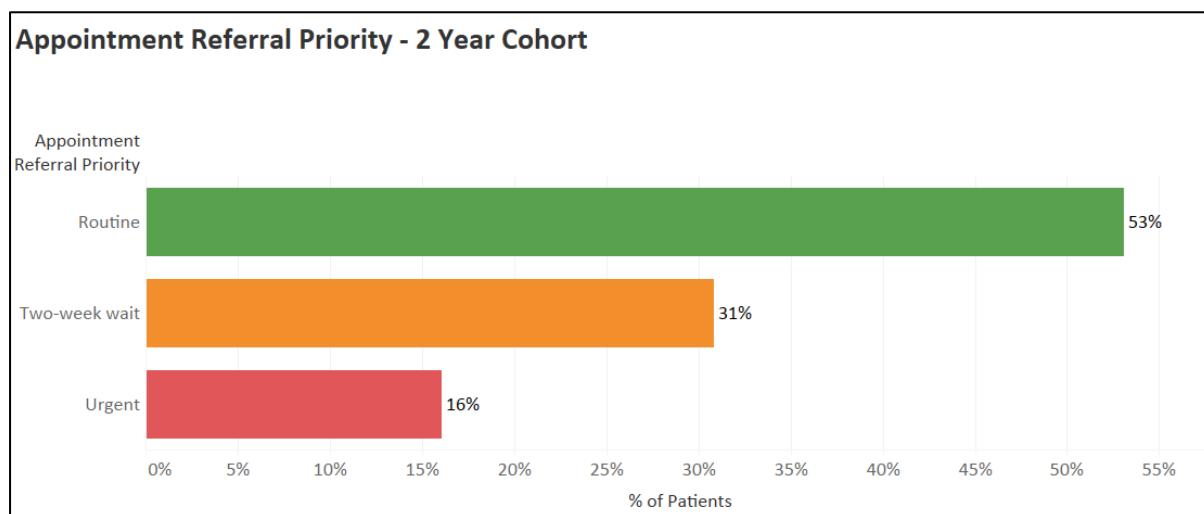


FIGURE 61: APPOINTMENT PRIORITIES IN 2-YEAR COHORT

Figure 62 shows percentage of absolute count of patients for each appointment type. The view is filtered on appointment type, which keeps Follow-up and New. The majority of patients (58%) had follow-up appointments in their pathway.

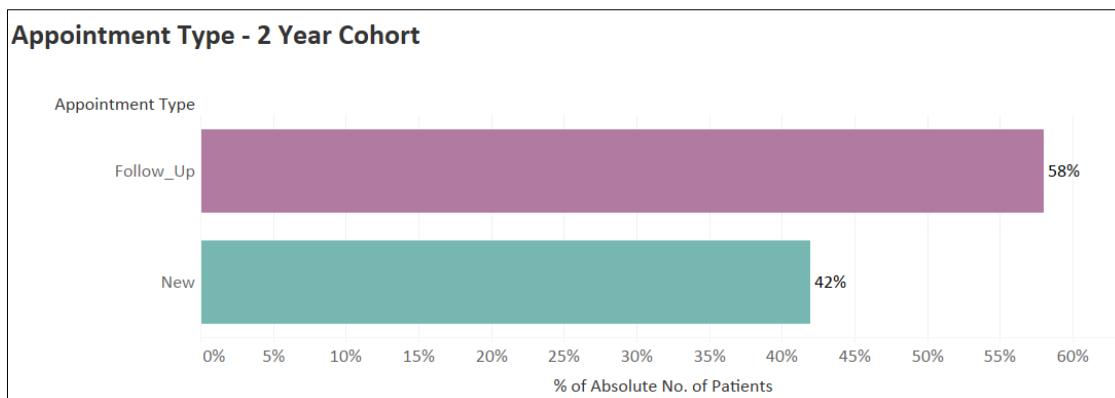


FIGURE 62: APPOINTMENT TYPE IN 2-YEAR COHORT

Figure 63 shows percentage of absolute count of patients for each referring source. The view is filtered on referring source, which has multiple members selected. The majority of the patients (53%) had a GP referral in their pathway

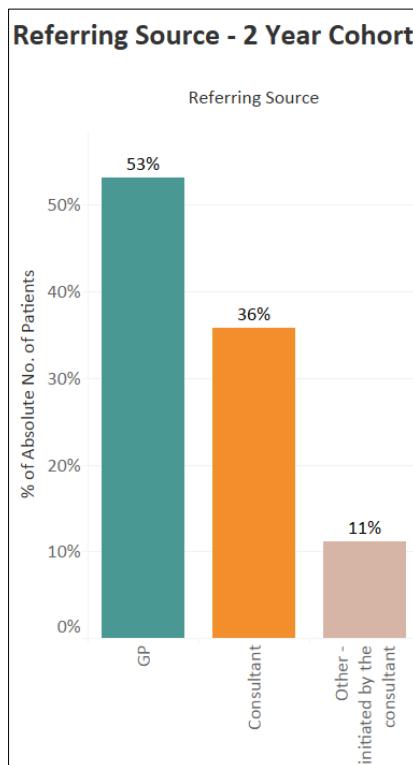


FIGURE 63: REFERRING SOURCE IN 2-YEAR COHORT

PATIENT DIAGNOSTICS PHASE

Figure 64 shows the trend of the distinct count of PSAs/year. Figure 65 shows the trend of distinct count of biopsies/year. It can be seen that the maximum PSAs were registered in 2014 and similarly the maximum biopsies were registered in 2014 as well.

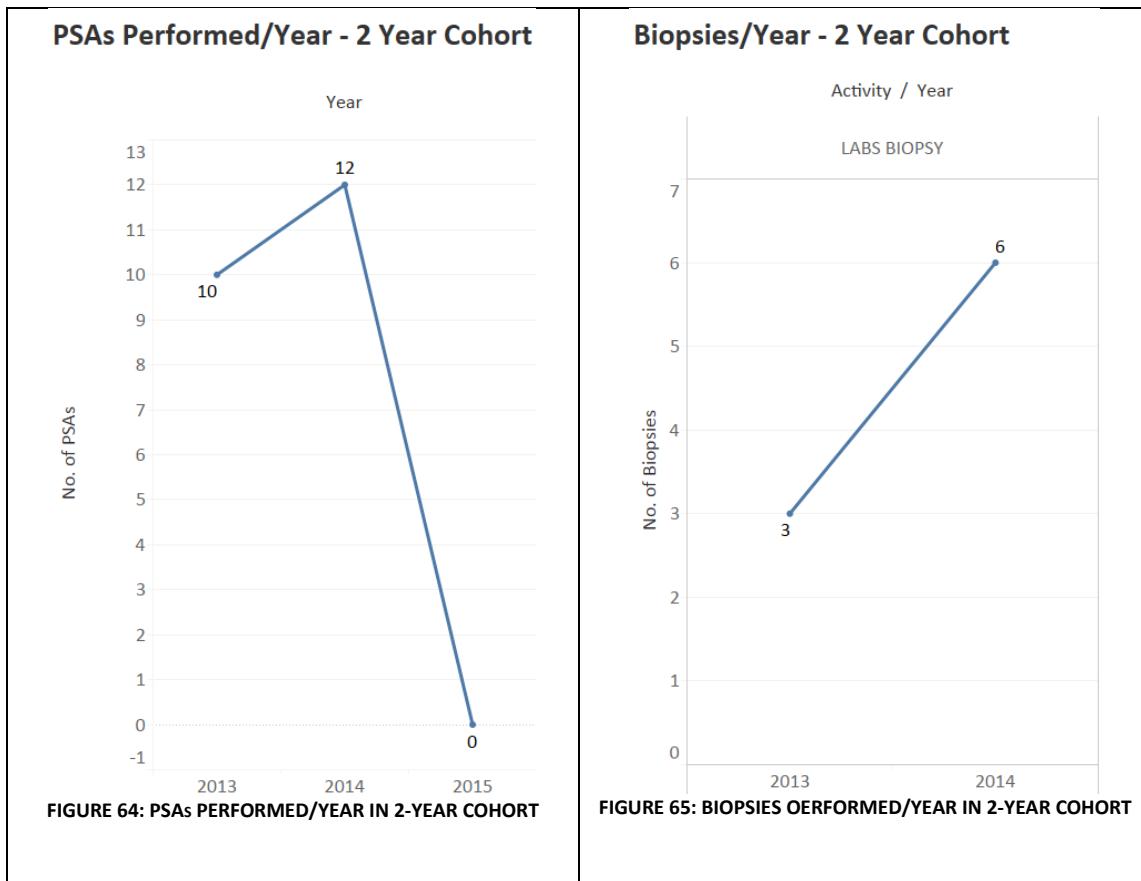


Figure 66 shows a distinct count of radiological procedures/year. The view is filtered on radiology procedure, which has multiple members selected. The majority of the patients had an MRI of the pelvis for prostate cancer in 2014. Note, the imaging information is missing for 2013.

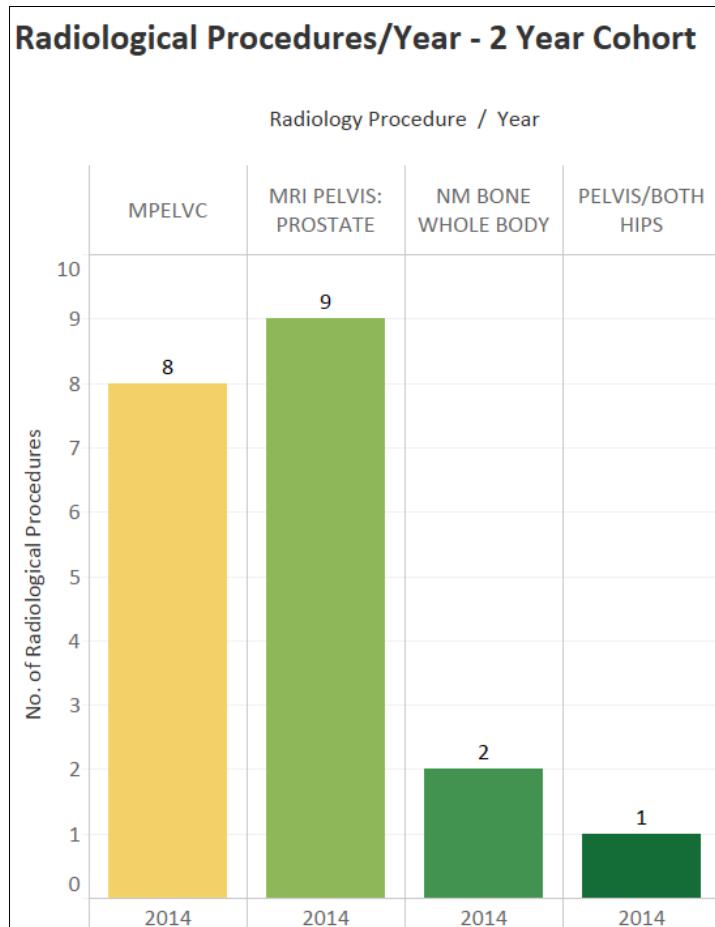


FIGURE 66: RADIOLOGICAL PROCEDURES PERFORMED/YEAR IN 2-YEAR COHORT

PATIENT TREATMENT PHASE

Figure 67 shows a distinct count of radiotherapy procedures (EBRT)/year. The view is filtered on radiotherapy procedure and keeps the EBRT procedure. It can be seen that the majority of radiotherapies were performed in 2014.

Figure 68 shows a distinct count of surgical procedures (Prostatectomy)/year. The view is filtered on surgery procedure and keeps the Prostatectomy procedure. It can be seen that the majority of the surgeries were performed in 2014 (10)

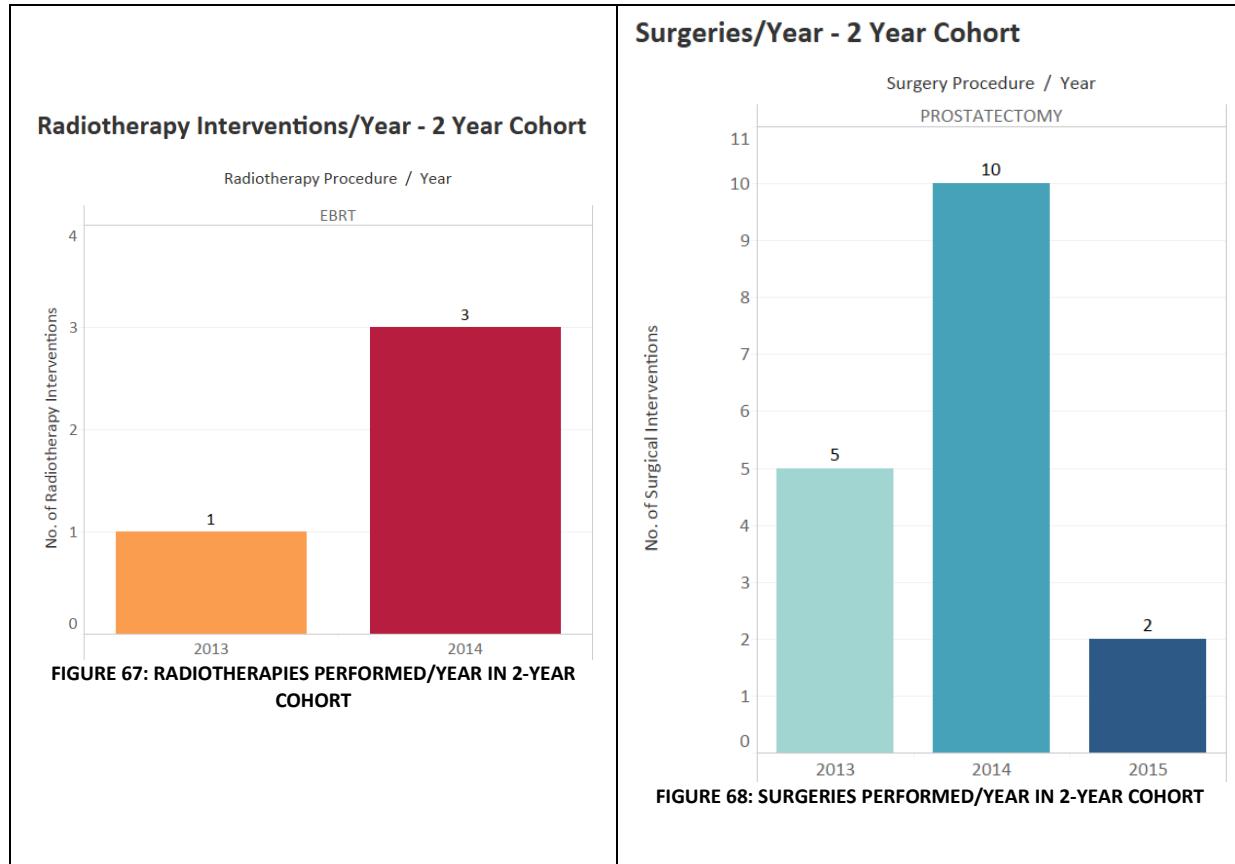


Figure 69 shows a distinct count of chemotherapies/hormone therapies/year. The view is filtered on Chemotherapy/hormone therapy drug, which has multiple members selected. The drug used most frequently was Casodex/Cassodex.

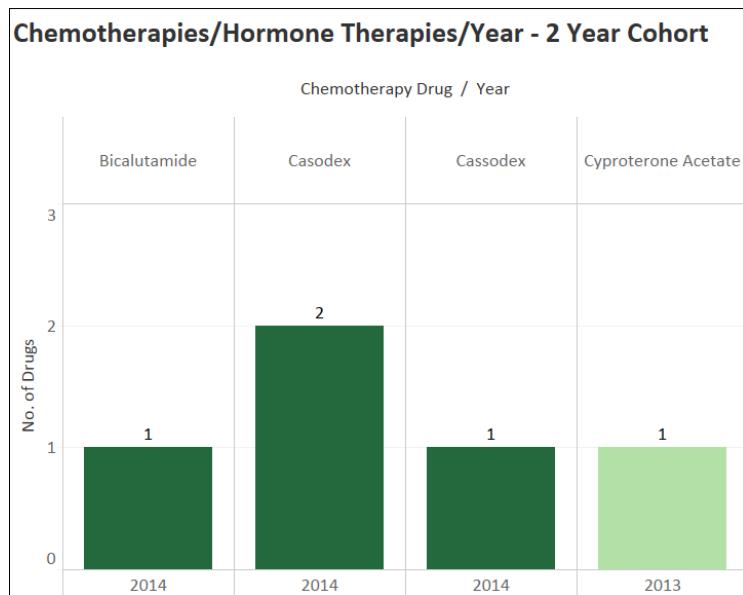


FIGURE 69: CHEMOTHERAPIES/HORMONE THERAPIES PERFORMED/YEAR IN 2-YEAR COHORT

6.3 LIMITATIONS

I have only looked at the simple demographics in this thesis and have not seen cross references of these demographics with the patient pathways. It would be interesting to see whether there were any differences between different ethnic groups and how they progressed in their pathways.

6.4 SUMMARY

This chapter gives us a first look at the event log by providing basic statistics related to the patient demographics, patient referral, diagnostics and treatment phases. It divided the log into two cohorts: a five year and a two year cohort. The five year cohort included all the patients with and without missing MDT and radiology information; whereas the two year cohort only included patients with complete information. This chapter was necessary to start the first initial filtration of the data into relevant cohorts that hold important information for the analysis. The next chapter, Chapter 7: Process Mining and Visualisation of the Pathway, continues with phases 4 and 5 of the CPAM roadmap and provides the main pre-processing, analytical and visual results after applying process mining techniques.

CHAPTER 7: PROCESS MINING AND VISUALISATION OF THE PATHWAY

*This chapter continues on to phases 4 and 5 of the CPAM roadmap (Figure 70). It begins by giving a background on: how a pathway is analysed using process mining techniques; the national standard *a-priori* process model (which in my case study is the London Cancer Alliance –West and South (LCA) Best Practice Prostate Pathway flowchart); and the current cancer pathway mapping techniques at Imperial College Healthcare NHS Trust. The chapter then describes my methods in detail and begins by pre-processing and filtering the event log (phase 4 of the CPAM roadmap) and goes on to the exploratory pathway analysis and visualisation (phase 5 of the CPAM roadmap). Results are displayed for the 2-year cohort (following the cohorts made earlier in chapter 6). For that cohort I have displayed the control flow mining and performance mining results for each patient referral type. The results also include a comparison between the LCA guideline metrics and my findings. The chapter is concluded with a discussion of my findings and results.*

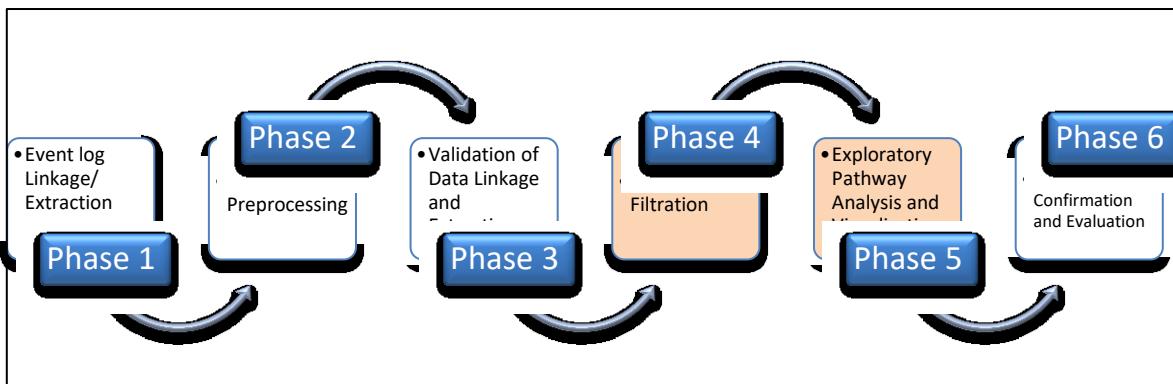


FIGURE 70: PHASES 4 (EVENT LOG FILTRATION) AND 5 (PATHWAY ANALYSIS) OF THE CPAM ROADMAP

7.1 INTRODUCTION

As discussed earlier in section 1.2.7, healthcare processes are usually semi-structured, complex and highly flexible leading to many exceptions and difficult decision-making. Research into the literature [132, 195] shows how the actual pathway followed by patients deviates from the standard clinical pathway and how analysing these pathways would improve the quality, resource optimisation and safety of patients. Many healthcare organisations are struggling to achieve performance compliance with the national standards and hence an objective analysis of the processes of these healthcare systems can contribute to improving the performance and quality of care.

7.1.1 CONTRIBUTIONS OF THIS CHAPTER

In this chapter, I have introduced a visualisation technique that shows what happens in reality in the prostate cancer pathway as opposed to what is theoretically supposed to happen as per the national LCA guidelines. It has helped in answering questions related to performance like where are the bottlenecks and deviations and how can they be removed. It has been novel because this method of giving insight into clinical pathways has not been used before in the NHS and particularly not for prostate cancer pathways anywhere in the world (See chapter 3: Systematic Review).

This exemplar is valuable in its own right as the visualisations highlighted gaps and delays in care that the clinicians did not know previously existed. For example, I have shown that patients going through a TWB referral had to wait an average of 38 days before they get their MRI done and 142 days before they get their biopsy done. What I have shown to the clinicians and managers in Imperial College Healthcare NHS Trust has been proven to be useful even though given the limitations on data quality. They have gone back to the data to see if these delays and bottlenecks are actually happening.

My algorithms, techniques and approach are generic enough to be applied to any discipline as long as data are available from the back end and can be transformed into the correct format.

7.2 BACKGROUND

7.2.1 PATHWAY ANALYSIS USING PROCESS MINING PERSPECTIVES

Not all analyses are considered relevant for research. To make sure that the analysis is not done ineffectively and is beneficial to the researcher, a list of interesting analyses questions needs to be generated. An example of such questions inspired by Boere are the following [152]:

- How long does a patient receive radiotherapy on average? (General performance)
- Does the process differ for young and old patients? (General performance)
- Are there any route deviations from the protocol? (Route conformance)
- What is the probability that the next patient undergoes the regular treatment pattern? (Data conformance)

Process mining in a healthcare setting, as suggested by [132] and [195] can be applied to the following three broad areas: Discovering and analysing recurring patterns in the pathway, Analysing pathway variants and analysing exceptional/adverse events.

During discovery of recurring patterns, the process discovery and visualisation techniques provide a complete insight on a model of the as-is patient pathway in one single visual. It gives you feedback about how cases are actually being executed in the organization and highlights recurring trends and patterns.

In the analysis of pathway variants (clusters), the mining algorithm segregates and characterises the pathway into several clusters based on a broad set of medical conditions and characteristics of the patients. Variants can be identified through different approaches like correlation analysis, filtration of traces on common characteristics, clustering pathway traces, and interesting factor combinations by experts (e.g. treatment and diagnosis code combinations).

The analysis of exceptional cases and adverse events can be accomplished by detecting inconsistencies between a standard a-priori model and the discovered process model (conformance checking) as well as looking at process variants.

7.2.2 THE A-PRIORI PROCESS MODEL

As my case study concentrates on prostate cancer, I have used the London Cancer Alliance –West and South (LCA) Best Practice Prostate Pathway flowchart as my guideline. The LCA guidelines are identified and mandated by the LCA Urology Pathway Group. They are not intended to be a comprehensive set of clinical guidelines but detail the necessary sequencing and timeliness of the various elements of the prostate cancer pathway to ensure it is delivered within the 62 day target (see section 1.2.7) [196]. Additionally, I have also reviewed the NICE guidelines for suspected prostate cancer referral [81] and incorporated them within the LCA guidelines to give additional detail and depth where necessary (see Figure 71).

It is important to point out here that the LCA guideline tells us in theory how patients with a prostate cancer diagnosis referral should move in their pathway to reach radiotherapy, prostatectomy or an active surveillance intervention. They have the assumption that the patients are one homogenous

group and they are trying to direct them through this guideline to reach a proper intervention. They are outlining the tasks the patients should follow along with their intended timelines to reach a decision on treatment. The percentage of patients that have achieved a completed a milestone in the required timeline is the performance metric with which these trusts are being evaluated on.

Figure 71 shows an adaptation of the process model for urgent two-week wait (TWW) referrals of patients with prostate cancer taken from the LCA pathway [196]. The following is an explanation of the pathway model:

- The process starts with an urgent GP TWW prostate cancer referral being received at day 0.
- By day 7, the patient is reviewed and their risk of having prostate cancer is assessed by evaluating lower urinary tract symptoms and sexual health, bloods and digital rectal examination, as well as subsequent diagnostic and staging investigations being requested according to clinical findings and protocol. If the patient is found to have no risk for prostate cancer, the patient is removed from the pathway. Otherwise, based on the stage of prostate cancer, the next steps are carried out. In my case study, I am following the localized prostate cancer stage and hence the LCA pathway in Figure 71 is adapted to show the staging requirements for localized prostate cancer.
- If watchful waiting is opted for, then the patient is placed on the watchful waiting pathway.
- If the patient is at a low-risk localized prostate cancer stage, then by day 7-14 a prostate biopsy should be done. If the patient is at a high-risk localized prostate cancer stage (determined by palpable DRE, PSA level >10 and high clinical suspicion), then by day 7-14 the patient is sent for an MRI pre-biopsy to guide biopsy strategy when uncertainty exists (possibly to avoid biopsy).
- Once the biopsy is done (if needed), by day 14-21 the biopsy results and further staging investigations are discussed in an MDT (Multidisciplinary Team) meeting as well as any hormones are started if appropriate. A meeting can also be set with the patient at this point to discuss the treatment options.
- By day 21-42 the staging and treatment options and the decision to treat (DTT) should be finalized in an MDT meeting.
- By day 62, the first definitive treatment (FDT) should be made.
- If the DTT was active surveillance (usually for low risk-localised prostate cancer), then the patient is taken out of the cancer pathway and yearly follow-ups are initiated.

- In the first year after the decision, within 3-4 months a PSA test should be taken; within 6-12 months a DRE should be done and within 12 months a re-biopsy should be taken.
 - In years 2-4 after the decision, within 3-6 months a PSA test should be taken and within 6-12 months a DRE should be done.
 - Finally, in year 5+, within 6 months a PSA test should be taken and within 12 months a DRE should be done.
-
- If the DTT was radical treatment (usually for intermediate risk localized prostate cancer), then treatments like brachytherapy, radiotherapy, or hormone therapy are initiated and a PSA test is checked from 6 weeks to 6 months after treatment
 - If the DTT was surgery (usually for high-risk localized prostate cancer), then a surgical treatment like prostatectomy is carried out

The LCA pathway focuses on the referral-to-treatment times in a prostate cancer pathway. LCA have also developed prostate pathway metrics based solely on the Cancer Waiting Times (CWT) and Cancer Outcomes Services Dataset (COSD) data items. I have evaluated these metrics (see full metrics in Appendix D) in my results section based on my prostate cancer dataset and model findings.

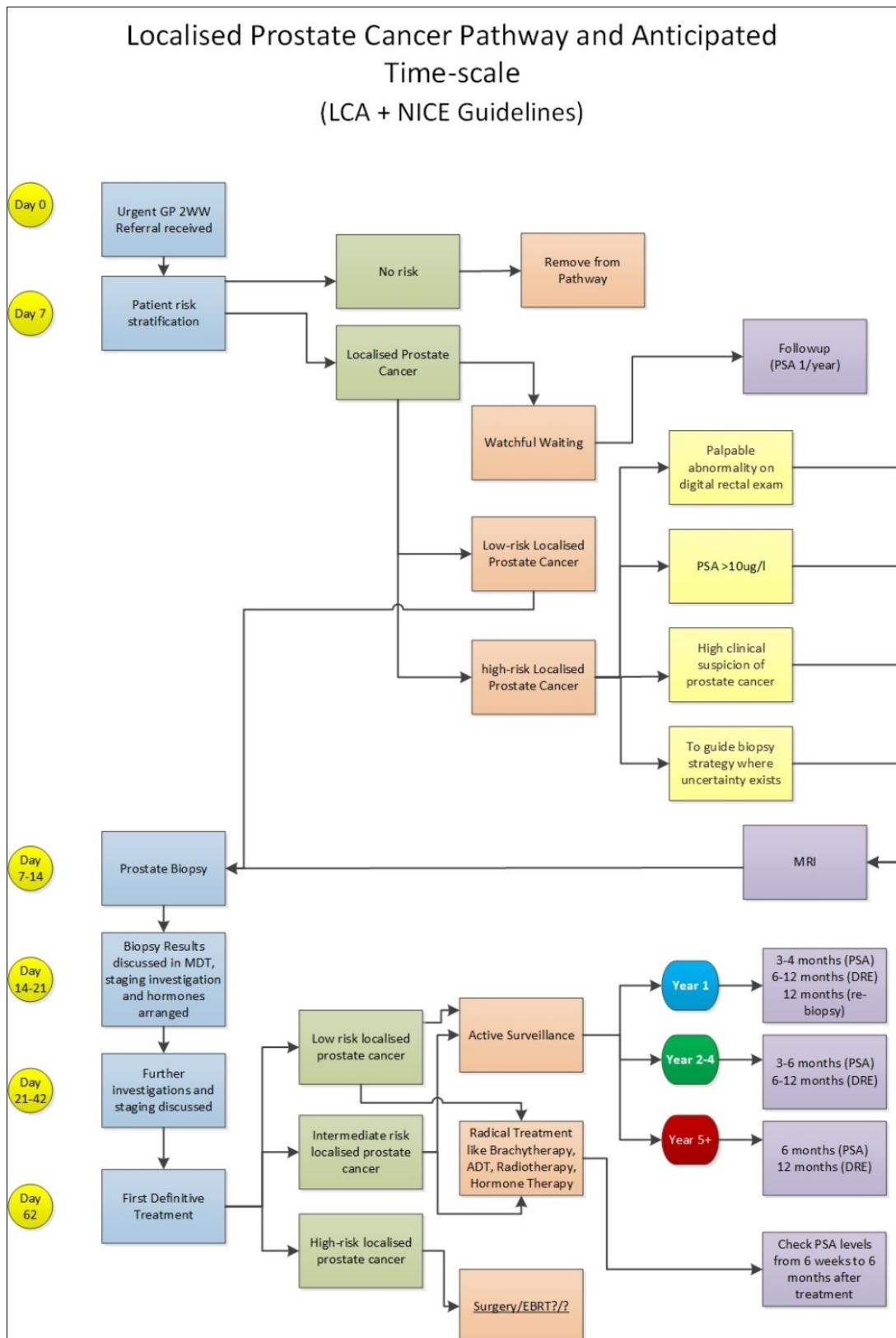


FIGURE 71: LOCALISED PROSTATE CANCER LCA + NICE GUIDELINES

7.2.3 CURRENT CANCER PATHWAY MAPPING AND ANALYSIS TECHNIQUES

Currently in the Cancer Quality Assurance unit at Imperial College Healthcare NHS Trust, the cancer pathways are mapped using MS Visio software for drawing diagrams of the expected pathway based on physician input and LCA/NICE guidelines. Several multidisciplinary people are involved in its construction including: the booking office (that is where the GP reference comes in), a lead clinician, someone from the performance team (MDT administration side) and a nursing representative. The consultant will explain to the team step-by-step what the pathway is based on the LCA guidelines and their clinical knowledge and a performance team member draws that pathway on MS Visio. It takes an average of three meetings to construct this pathway. The people on board need to know exactly the procedures and sometimes it is also good to have the business manager on board so he can tell you which activity comes where. The whole pathway is based on the 62-day pathway. They use the LCA pathway as a template and have adapted it for the Imperial College hospitals. The pathway that they create with MS Visio is the final output that the physician sees in order to know what the cancer pathway is. These pathways are then followed by reports and charts stating the frequency and reasons of patients who breached the milestones within the pathway. Data for the report is taken from the Open Exeter database (that populates its fields from the Somerset Cancer Database). Open Exeter contains clean and validated data supported by cancer waiting times and reasons for breach.

7.3 METHODS

7.3.1 PREPROCESSING AND INITIAL LOG INSPECTION

I evaluated several commercially available and academic process mining software that would be useful in the analysis of all the processes and pathways in my dataset. The selection of my preferred software was done after going through a checklist for selecting a process mining tool [197]. The first few criteria I looked at before evaluating from a process mining perspective was to see simplicity, ease of use, and availability. Moving further, I then evaluated the software based on the software's support for process mining functionalities and here's where I used the checklist to answer the following questions:

1. Which of the process mining activities are supported?
2. Which perspectives are supported?
3. What modeling notations are supported?
4. What import and export formats are supported?

Once the software was chosen (in my case DISCO – by Fluxicon and PROM 6.5.1), I requested the software company to provide us with an educational license (where necessary) so I could use the software free-of-cost for research purposes.

I then imported my extracted and prepared event log (see section 3.5) into DISCO to start process discovery of my processes. Regular contact with the clinicians increased my knowledge to better interpret the result of the analysis and process discovery. After an initial run of the process mining software and an early inspection of the log, it was evident that the complex process model was unreadable and spaghetti-like. This was not surprising as processes in the healthcare domain have a lot of variants based on different patients and diseases. To cut down on low-level events that did not lend any importance to the path and contributed in making the process more complicated, I performed the following preprocessing steps:

- If there were several outpatient events on the same day I took only the first event
- If the procedure in Surgery was prostate biopsy, I changed the name of the process from SURGERY to BIOPSY
- Biopsies were registered as a separate activity per biopsy core. I deleted the rest of the repetitive entries and just kept one
- All various kinds of prostatectomies (e.g. Laparoscopic, Robotic, Radical, etc.) were coupled into a single process called PROSTATECTOMY to avoid being overcrowded and all other forms of surgeries were excluded
- All Radiology activities that were relevant to prostate cancer (like MRI) were kept
- As the admission activity was always occurring on the same instance prior to the surgery, to minimize on the clutter of activities in the model, I omitted the admission
- As my scope only included Electron Beam Radiotherapy patients, I excluded all other patients having any other type of radiotherapy

7.3.2 CLUSTERING

Examining the full activity log and process model did not result in comprehensible visuals. Therefore, a filtering approach was needed to obtain clear and concise models that contain more homogenous behaviour. In order to simplify the process flow and concentrate on areas that were relevant and complete, I performed clustering of the dataset based on a range of years where I found complete data. My process mining (PM) cluster took into account the last **2 years** of my cohort (from 01/01/2013 until 01/01/2015). As described previously in Chapter 5, my reasons for choosing this cohort size were because of issues with MDT submission before the year 2013 when the commissioners were not allowed to submit their data. Hence, if the entire 5-year cohort was taken, the data would be incomplete and assumptions would be wrong. We have therefore chosen the final cohort to be from 2013 to 2015 to demonstrate the principles of process mining against the LCA guidelines standard.

Within the PM cluster I further segmented my findings based on the priority of referrals, namely, **Two-Week Wait (TWW)**, **Routine**, and **Urgent**. I also limited my findings to concentrating on the patients who ended up having a first definitive treatment of either **Radiotherapy** or **Surgery** (or both).

To create the PM cluster, I used the filtering facility in the DISCO Fluxicon software that helped me filter down to the exact years I needed. Furthermore, I then put a filter on the Appointment Priority and Referral Source variables to retrieve all the patients that were referred by a GP and were on a TWW or Routine or Urgent priority. I extracted the pathways of these three categories of patients until the actual surgical/radiotherapy event so I can go back and do reverse engineering from that point backwards to see how the patients reached that stage (how many variants).

7.3.3 PROCESS DISCOVERY AND VISUALISATION

Following preprocessing and clustering, my event log was ready to be discovered and analysed through the different perspectives. Part of my methods was to look at different packages that visualise the movement of the patients between processes in a user-friendly and intuitive format. The evaluated packages included: Geographical Information System (GIS) packages, simulation packages, animation packages, and diagramming packages. However, the process mining software

and the accompanying visualisations produced provided factual and graphical representations of the actual processes followed by the patients by relying mainly on the routinely collected data in the backend databases. The PM cluster was exported from DISCO in the XES format suitable to be directly read by PROM 6.5.1. Once imported and read in PROM, the PM cluster was ready to be analysed, visualized and compared with the LCA guidelines.

Before starting process discovery on the PM cluster, I prepared a process flow pathway for the patients following the timelines and order of events suggested in the guidelines by LCA. To do this, I populated a sample spread sheet table in MS Excel with dummy entries of 40 patients following all the different possible paths laid out by the LCA guidelines. After the physicians' approval, I continued to construct the flow diagrams of the guideline pathway in PROM and this was then used as a benchmark comparison for the process flow diagrams I constructed from my PM cluster.

The PM cluster was analysed using two process mining perspectives: Control-flow and Performance. For the control flow mining perspective of the PM cluster, I initiated process discovery by learning how a process performs in real-life and how it deviates from the intended behaviour. For this, I had to perform process exploration by using a combination of different algorithms. Process exploration iteratively utilises repeated parameter selection and tuning to perform process discovery whilst continuously evaluating the resulting process map. Before commencing with process exploration, I needed to decide what discovery technique I intended to use. Two of the techniques that I have concentrated on in my approach include a Directly-Follows (DF) based approach and an Inductive Mining (IM) approach. DF based tools (Like DISCO), although easy to use and have many features like log animation and extensive filtering, do not usually have executable semantics and hence deviations cannot be easily analysed. Moreover, they do not support parallelism meaning the state of the system unrealistically depends on the last executed process step and a clutter like diagram of loops is produced. With IM tools (like the Inductive Visual Miner (IvM) plugin in PROM), they allow for semantics, allow for fitness and support parallelism without getting lost in loops [198].

I first used the DISCO software (DF approach) to perform process discovery and examine the process model produced on the PM cluster. Within DISCO, the next-generation-fuzzy miner is used by default to mine the processes and produce fuzzy models. In PROM, I utilised the Inductive Visual Miner (IvM) plugin (IM approach) to generate the process models for the PM cluster. As both techniques had their pros and cons, I used both these techniques to formulate my results. I used the

DF approach used in DISCO to find the mean and median waiting times (sojourn times) that were harder to interpret in the IM approach, whereas, I used the IM approach used in PROM to draw my process flow diagrams that showed a more functional and practical visualisation. PROM supports life cycle data and has the ability to distinguish between subtle control flow aspects like concurrent and interleaved execution [199].

I first generated a high-level spaghetti-like overview of the flow of patients to get an initial idea of the complexity of the given event log of the PM cluster. The complexity was then reduced by applying activity and life cycle filters to the event log in Inductive Miner. Finally, a high-level diagram was created, showing the ‘happy flow’ as Leemans et al call it [198], through the process, i.e. the most common path taken by the patients.

Following the control flow mining perspective, I then analysed the PM cluster through a performance perspective to show the differences between the three referral priorities (TWW/Urgent/Routine) in terms of delays, and adherence to LCA guidelines. I divided this perspective into three categories: Trace cluster analysis, Bottleneck analysis and LCA guideline compliance.

Trace-cluster analysis explores the number of groups (clusters) of patients (traces) following a common track. In process mining terms, these are called trace variants. There are many plugins in PROM that provide this feature and amongst them is the Trace Variant Analysis plugin. This technique aims at organizing the traces in an event log in such a way that both common and exceptional behavior can be easily distinguished. It does this by (1) grouping similar traces in clusters and (2) visualising these clusters. The representations I got helped me visualize the cycle times of each cluster. To do trace-cluster analysis, I took out the top 10 trace clusters in each referral priority (based on the median duration it takes to complete each cluster). I then analysed the percentage of cases (traces) in each cluster and the minimum/maximum duration of each cluster to see which referral had consistency and what are the total time differences amongst all the three referrals.

Bottleneck analysis explores the delays or sojourn times (waiting times) between two activities in the pathway. This is the time it takes one activity to start after handover from the previous activity in the pathway. For that I segregated first all the possible traces (paths) each patient can take in all the different referrals. I then took out the waiting times of each activity (shown in days under the

activity names) to detect where the bottlenecks in the pathway were based on the timelines anticipated by the LCA guidelines.

Finally for the LCA compliance, I compared some of the cancer waiting times metrics stated in the LCA guidelines (See Appendix D) against the waiting times discovered in all the three referral pathways of my PM cohort to see how much they conform.

7.4 RESULTS

In this section, I will present my results that I have achieved using process mining as a tool to analyse and study the performance and conformance of the prostate cancer pathway to the national LCA guidelines. I have divided my results according to the process mining perspectives that I have applied on my derived model: Control flow mining and Performance mining. I will begin by mapping the prostate cancer referral pathway standard flow chart laid out by the LCA guidelines onto a process flow diagram.

7.4.1 THE LCA GUIDELINE PATHWAY

In Figure 72 of the LCA flow chart, I have highlighted the activities I am tracking in my PM cohort of patients and comparing them with this standard guideline. Figure 73 is the retrieved care process model of all the activities performed by patients undergoing the standard pathway laid out by LCA (highlighted in gold in Figure 72). These activities are entirely based and traceable on the LCA flow chart. The model was constructed by manually entering test patients in an event log and the process flow diagram was constructed using the Inductive Visual Miner in Prom 6.5.1. The model shows 100% of the activities and 100% of the different paths in the pathway. As I am only following the pathways of the patients that undergo radiotherapy or surgery as a final treatment, those are the only two treatment activities that you will see on the right hand side of the diagram.

Beginning at the left side of the model, when the patient enters the pathway through a first GP TWW appointment (indicated with a green activity bubble), the patient is directed to any one of the three diagnostic tests: Biopsy, PSA or MRI Imaging based on what the clinician decides. If the patient is directed by the clinician to go do a biopsy first (indicated by the orange coloured track), after the biopsy is done, the patients go through an MDT. Once the decision is made in an MDT meeting, the

patient either goes to further diagnostic tests (based on what was recommended in the MDT meeting) like MRI imaging and/or PSA test, or then goes straight to the first definitive treatment like Radiotherapy or Surgery.

If the patient is directed by the clinician to go do a PSA test first (indicated by the green coloured track), after the PSA is done, the patients either go directly to an MRI imaging or then straight to a Biopsy (as seen by the conditional branching in the diagram). After the biopsy, the patients follow on to an MDT meeting. After the MRI imaging, the patients can go through a Biopsy or then go straight to an MDT meeting. Once the decision is made in an MDT meeting, the patient then goes to further diagnostic tests (based on what was recommended in the MDT meeting) like MRI imaging, or then goes straight to the first definitive treatment like Radiotherapy or Surgery.

If the patient is directed by the clinician to go do an MRI Imaging first (indicated by the blue coloured track), after the MRI Imaging is done, the patients either go directly to an MDT, or then go through a biopsy and then an MDT. Once the decision is made in an MDT meeting, the patient then goes to further diagnostic tests (based on what was recommended in the MDT meeting) like a PSA test, or then goes straight to the first definitive treatment like Radiotherapy or Surgery.

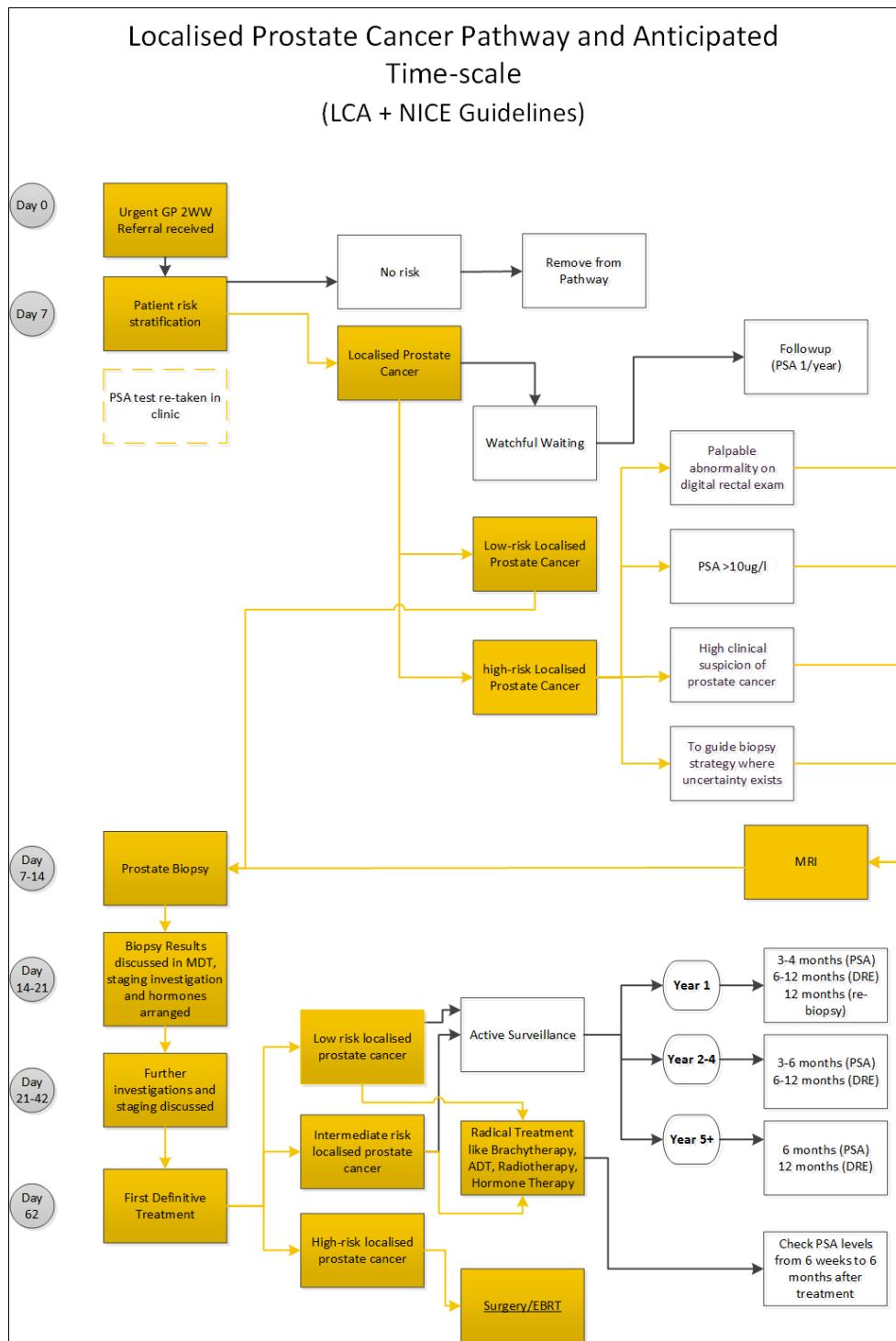


FIGURE 72: LCA PROSTATE CANCER PATHWAY FLOWCHART WITH HIGHLIGHTED ACTIVITIES

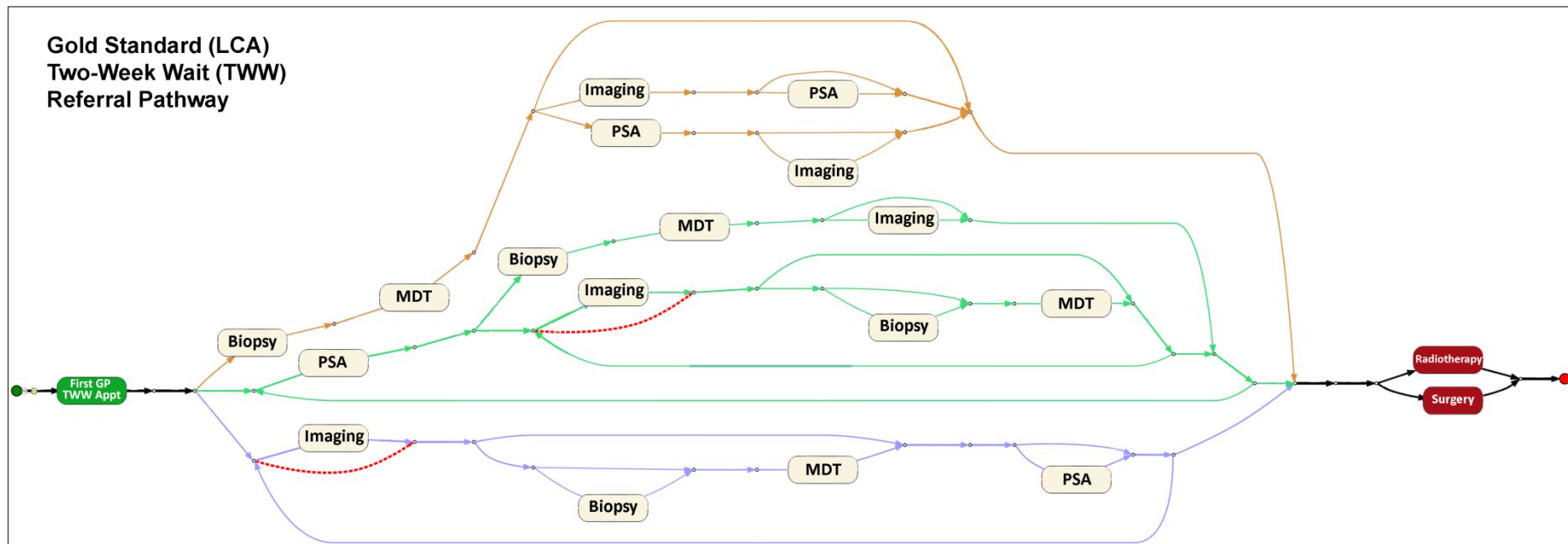


FIGURE 73: THE LCA STANDARD PATHWAY PROCESS FLOW DIAGRAM (100% ACTIVITIES AND 100% PATHS) – MADE WITH PROM 6 INDUCTIVE MINER

7.4.2 THE PM CLUSTER: 2-YEAR COHORT (2013-2015)

This is the main process mining (PM) cluster within my cohort. It covers all the referral pathways (TWW/Urgent/Routine) that have started within the 2-year time span from 01/01/2013 to 01/01/2015. In Figure 74, the entire hierarchy and breakdown of the frequency of patients with or without cancer following all the interventions is shown in this flow chart. The flow chart starts with our main criteria that every patient who has cancer would have undergone a PSA at least once in their journey. This is the starting point of our pathway and includes 4878 patients. Further on, the flowchart is divided into patients based on their first starting activity within that time frame. Thus, we have patients that start with various different starting points other than a proper GP referral, and this includes 4583 patients (shown as a grey block). These can include patients that are carried over from the previous years and have an on-going pathway. The other starting point is patients coming through a proper new GP (TWW, Urgent or Routine) referral that has its first starting point within that time frame. This set consists of 295 patients. This is the category of patients I am interested in and following in my analysis. Patients referred via a GP are then divided into patients who have a registered MDT meeting in their pathway (88 patients) followed by patients stratified according to different GP referrals: TWW (50), Routine (27) and Urgent (11). Within each of these referrals I am then seeing the frequency of patients that had a Surgery (10), Radiotherapy (1), Chemotherapy (17), are on Active surveillance (4) or do not have cancer (18). In my analysis, however, I am only following the Surgery (10) and Radiotherapy (1) patients.

CONTROL FLOW MINING RESULTS

I have divided the results in the control flow mining perspective into three categories based on the priority of the referrals: TWW, Urgent and Routine. I will begin with the TWW referral results.

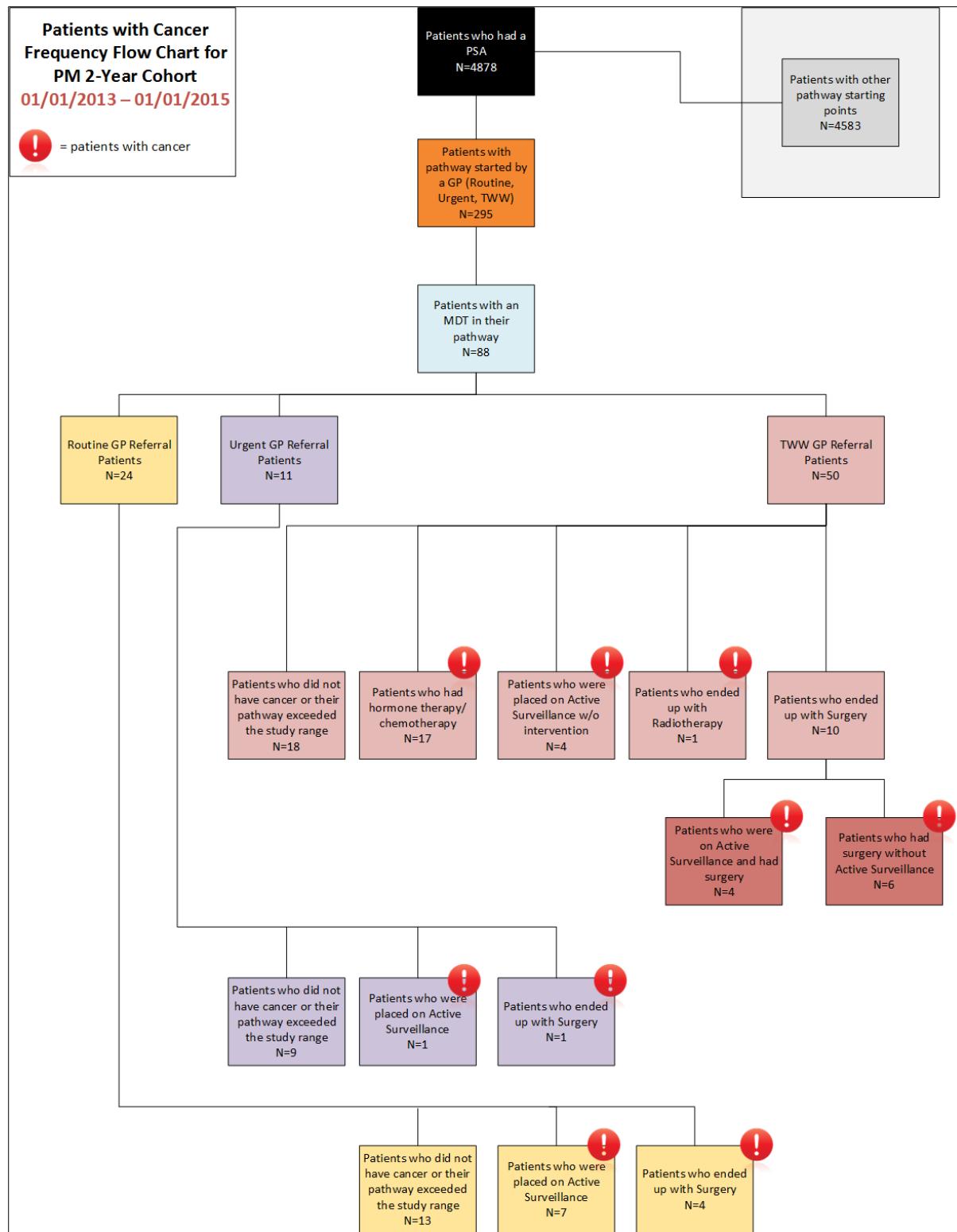


FIGURE 74: PM 2-YEAR COHORT CANCER FREQUENCY AND INTERVENTIONS FLOW CHART

TWO WEEK WAIT REFERRALS

This is the first segmentation of my PM cluster. It includes all patients that have entered the system through a GP two-week wait outpatient appointment (GP TWW OPA) referral and ended up having either radiotherapy (EBRT) or surgery (Prostatectomy) between the years 01/01/2013 and 01/01/2015. The following diagram (Figure 75) shows the exact screening process of how I selected the TWW referral patients who ended up with radiotherapy or surgery interventions

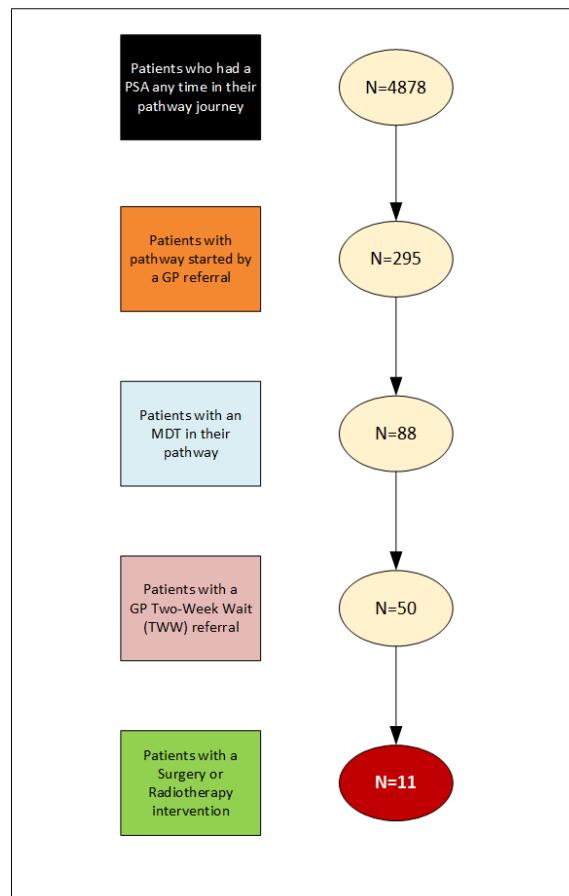


FIGURE 75: SCREENING OF PATIENTS TO REACH THE SURGICAL/RADIOThERAPY INTERVENTIONS IN TWW REFERRAL

Table 26 shows a summary of the TWW referral event log. It can be seen from this table that there are 11 cases (process instances) going through the TWW referral pathway (see Figure 75) with a total of 131 combined events amongst those cases. There are 8 different activity classes in this pathway. The activity with the maximum number of relative occurrences is the **Imaging** activity with a frequency of 24%. The activity with the least number of relative occurrences is the **Radiotherapy** activity with a frequency of 2%.

Table 27 shows us the relative frequency of the start activities in the different paths taken after the TWW referral outpatient appointment (OPA). There were four observed paths post the first appointment in a TWW referral: paths taken via the biopsy, Imaging, PSA and MDT routes. Patients who took an MDT route after their first appointment accounted for the maximum percentage of patients (36%). This was followed by patients who re-took a PSA test in clinic after their first appointment (27%) and others who went straight for an MRI (27%). Patients going for a biopsy after first appointment accounted for the least percentage of patients (9%).

TABLE 26: LOG SUMMARY OF TWW CLUSTER

Log Summary		
Total No. of Process Instances	11	
Total No. of Absolute Events	131	
Total No. of Activities	8	
Total No. Start Events	1	
Total No. End Events	2	
Activity Frequency		
	Absolute	Relative
Imaging	32	24%
MDT	28	21%
Surgery	22	17%
PSA	16	12%
Biopsy	12	9%
GP TWW OPA	11	8%
Chemotherapy/Hormone	8	6%
Radiotherapy	2	2%

TABLE 27: START ACTIVITY FREQUENCY OF DIFFERENT PATHS IN THE TWW REFERRAL POST GP APPOINTMENT

Start Activity	Relative Frequency of patients	Percentage
Biopsy	1	9%
Imaging	3	27%
MDT	4	36%
PSA	3	27%

Figure 76 depicts the high-level retrieved care process model of the overall activities performed by patients undergoing a TWW referral pathway from a GP and ending on a radiotherapy or surgery first definitive treatment. The model was constructed using the Inductive Visual Miner in Prom 6.5.1. For the sake of clarity and readability, the model uses 100% of the activities and 100% of the different paths in the pathway.

Beginning at the left side of the model, the total number of patients entering this cluster via a GP TWW OPA is 11 as depicted in Table 26. As the IvM supports parallelism, and since every patient can follow several different pathways, the model (with low level activities removed) shows patients taking 4 different routes from the start of their patient journey until radiotherapy or surgery treatments.

Patients going for a biopsy as the first diagnostic activity after the first TWW outpatient appointment go on to have an MDT meeting and then move on to have their first definitive treatment of Surgery or Radiotherapy.

Patients going for an MRI imaging as the first diagnostic activity after the first TWW outpatient appointment go on to have an MDT meeting. At this point, some patients get started on chemo/hormone therapy before their surgery/radiotherapy while others go straight for surgery or radiotherapy.

Patients going for an MDT as the first diagnostic activity after the first TWW outpatient appointment go on to have either a chemo/hormone therapy before their surgery/radiotherapy while others go straight for surgery or radiotherapy.

Patients going for a PSA test as the first diagnostic activity after the first TWW outpatient appointment go on directly for an MDT meeting and then treatments, or go through imaging and then an MDT meeting and then treatments.

Also indicated in the diagram are activities with long median sojourn times (number of days written under activity name). These are the activities that have to wait a long time before starting. The diagnostic test that takes the longest to start is the Biopsy test (median wait 151 days) as well as the PSA test (41 days).

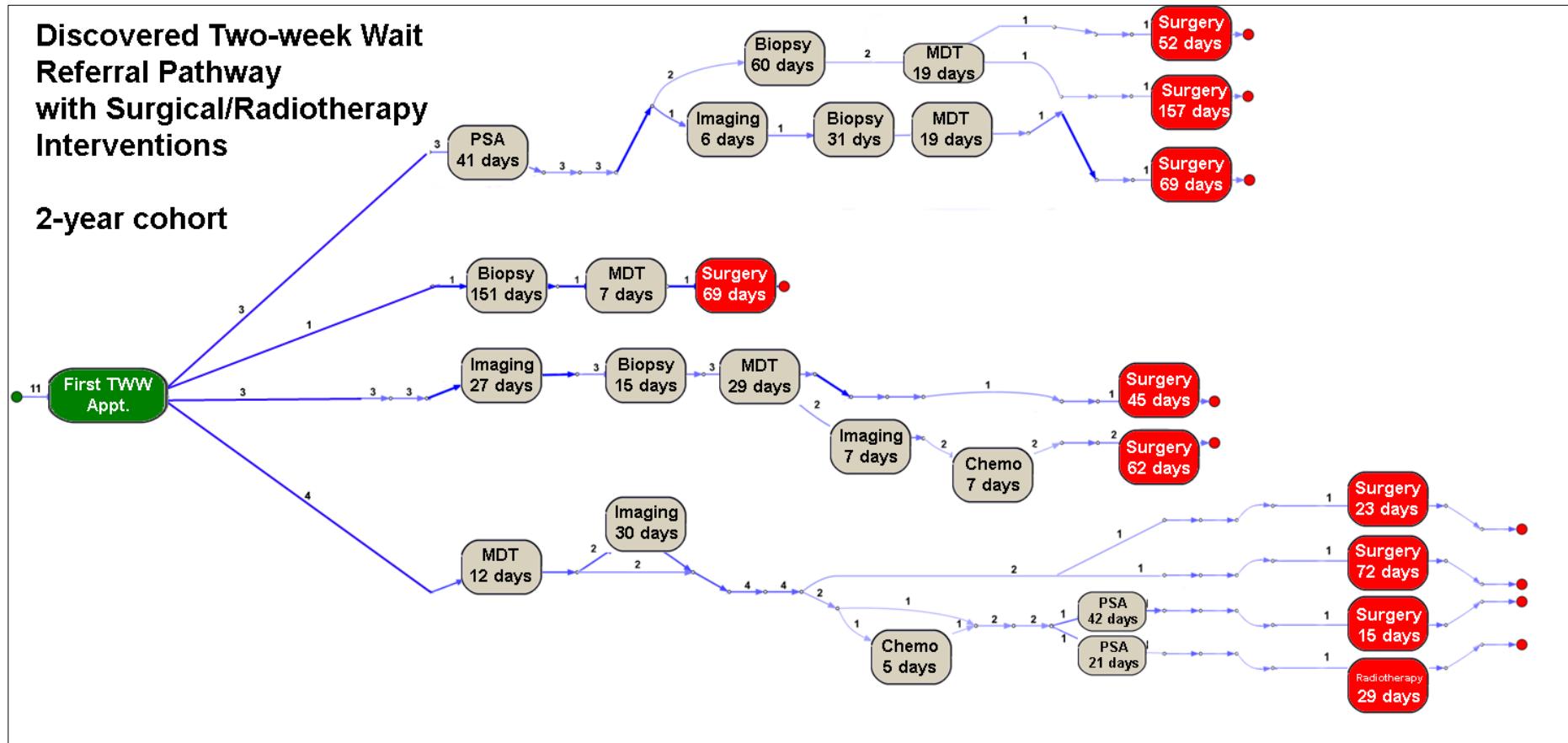


FIGURE 76: 2-YEAR COHORT TWW REFERRAL CLUSTER PROCESS FLOW DIAGRAM (100% ACTIVITIES AND 100% PATHS) SHOWING FREQUENCY OF PATIENTS AND SOJOURN TIMES

URGENT REFERRALS

This is the second segmentation of my PM cluster. It includes all patients that have entered the system through a GP urgent outpatient appointment (GP Urgent OPA) referral and ended up having either radiotherapy (EBRT) or surgery (Prostatectomy) between the years 01/01/2013 and 01/01/2015. The following diagram (Figure 77) shows the exact screening process of how I selected the urgent referral patients who ended up with radiotherapy or surgery interventions.

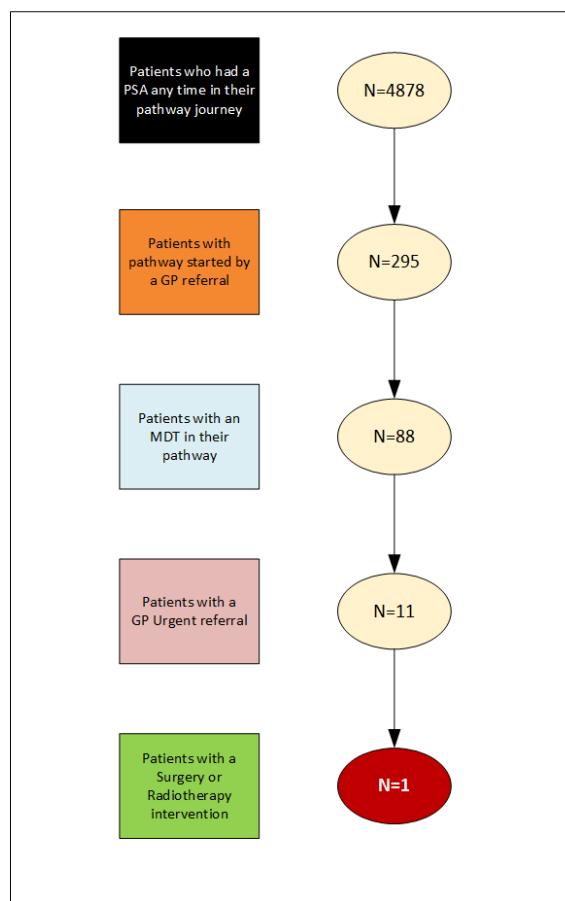


FIGURE 77: SCREENING OF PATIENTS TO REACH THE SURGICAL/RADIOOTHERAPY INTERVENTIONS IN URGENT REFERRAL

Table 28 shows a summary of the urgent referral event log. It can be seen from this table that there is 1 case (process instance) going through the urgent referral pathway with a total of 3 combined events for that case. There are 3 different activity classes in this pathway. All the activities happen only once in this pathway. The patient goes to an MDT meeting after the first GP Urgent appointment, followed by a surgical intervention.

TABLE 28: LOG SUMMARY OF URGENT CLUSTER

Log Summary		
Total No. of Process Instances	1	
Total No. of Absolute Events	3	
Total No. of Activities	3	
Total No. Start Events	1	
Total No. End Events	2	
Activity Frequency		
	Absolute	Relative
GP Urgent OPA	1	33%
MDT	1	33%
Surgery	1	33%

Figure 78 depicts the high-level retrieved care process model of the overall activities performed by the patient undergoing an urgent referral pathway from a GP and ending on a radiotherapy or surgery first definitive treatment. The model was constructed using the Inductive Visual Miner in Prom 6.5.1. For the sake of clarity and readability, the model uses 100% of the activities and 100% of the different paths in the pathway.

Beginning at the left side of the model, the total number of patients entering this cluster via a GP Urgent OPA is 1 patient as depicted in Table 28. This patient goes on to have an MDT meeting followed by a surgical prostatectomy.

As seen from Figure 78, the MDT meeting takes a median of 27 days to start (sojourn time) while the surgery takes 48 days to commence.

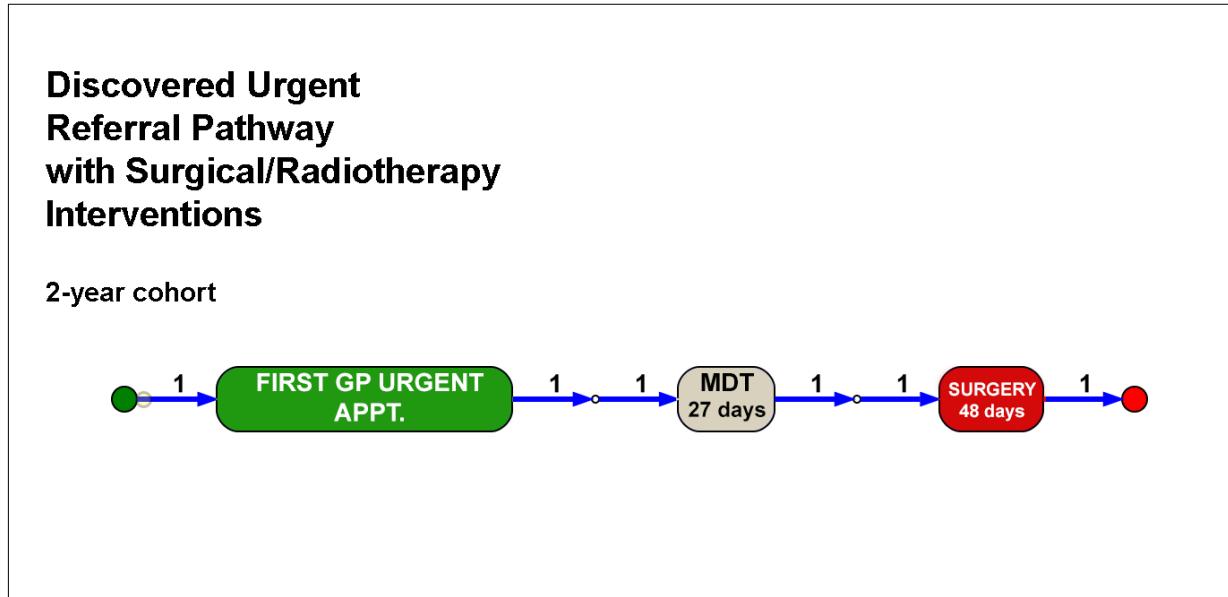


FIGURE 78: 2-YEAR COHORT URGENT REFERRAL CLUSTER PROCESS FLOW DIAGRAM (100% ACTIVITIES AND 100% PATHS) – MADE WITH PROM 6 INDUCTIVE MINER

ROUTINE REFERRALS

This is the third segmentation of my PM cluster. It includes all patients that have entered the system through a GP routine outpatient appointment (GP Routine OPA) referral and ended up having either radiotherapy (EBRT) or surgery (Prostatectomy) between the years 01/01/2013 and 01/01/2015. The following diagram (Figure 79) shows the exact screening process of how I selected the routine referral patients who ended up with radiotherapy or surgery interventions.

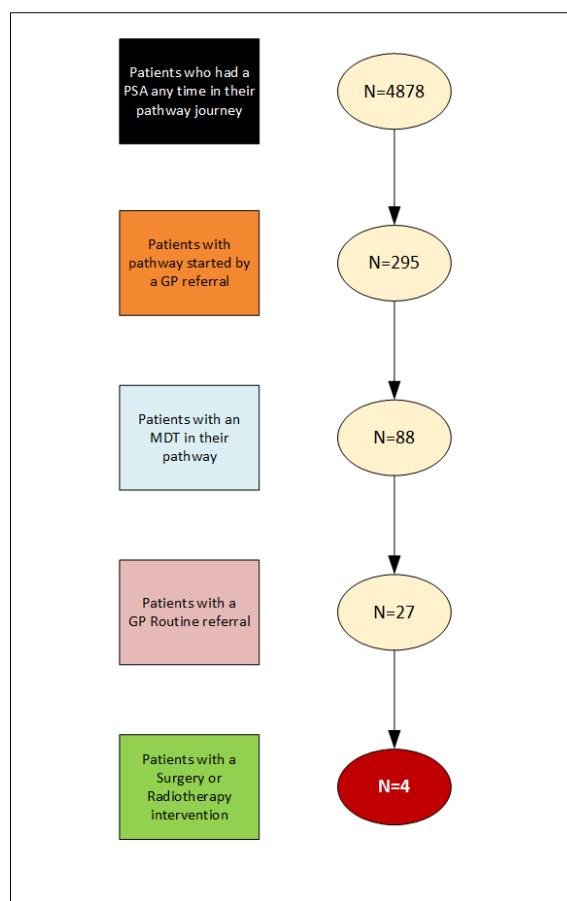


FIGURE 79: SCREENING OF PATIENTS TO REACH THE SURGICAL/RADIOThERAPY INTERVENTIONS IN ROUTINE REFERRAL

Table 29 shows a summary of the routine referral event log. It can be seen from this table that there are 4 cases (process instances) going through the routine referral pathway with a total of 31 combined events amongst those cases. There are 8 different activity classes in this pathway. The activity with the maximum number of relative occurrences is the **PSA** activity with a frequency of 23%. The activities with the least number of relative occurrences are the **Biopsy** and **Chemotherapy** activities with a frequency of 3% each.

Table 30 shows us the relative frequency of the start activities in the different paths taken after the TWG referral outpatient appointment (OPA). There were two observed paths post first appointment in a routine referral: paths taken via PSA, and paths taken via Imaging. Patients who re-took a PSA route after their first appointment accounted for the maximum percentage of patients (75%). This was followed by patients who went straight for an MRI imaging (25%).

TABLE 29: LOG SUMMARY OF ROUTINE CLUSTER

Log Summary		
Total No. of Process Instances	4	
Total No. of Absolute Events	31	
Total No. of Activities	8	
Total No. Start Events	1	
Total No. End Events	1	
Activity Frequency		
	Absolute	Relative
PSA	7	23%
MDT	4	13%
Surgery	4	13%
Imaging	4	13%
GP Routine OPA	4	13%
Radiotherapy	6	19%
Chemotherapy/Hormone	1	3%
Biopsy	1	3%

TABLE 30: START ACTIVITY FREQUENCY OF DIFFERENT PATHS IN THE ROUTINE REFERRAL POST GP APPOINTMENT

Start Activity	Relative Frequency of patients	Percentage
PSA	3	75%
Imaging	1	25%

Figure 80 depicts the high-level retrieved care process model of the overall activities performed by patients undergoing a routine referral pathway from a GP and ending on a radiotherapy or surgery first definitive treatment. The model was constructed using the Inductive Visual Miner in Prom 6.5.1. For the sake of clarity and readability, the model uses 100% of the activities and 100% of the different paths in the pathway.

Beginning at the left side of the model, the total number of patients entering this cluster via a GP Routine OPA is 4 as depicted in Table 29. As the IvM supports parallelism, and since every patient can follow several different pathways, the model (with low level activities removed) shows patients taking 3 different routes from the start of their patient journey until radiotherapy or surgery treatments.

Patients going for a PSA test as the first diagnostic activity after the first routine outpatient appointment go on to have either a biopsy or move straight to an MDT meeting. If patients go through a biopsy they eventually go through an MDT after that. After the MDT meetings, the patients then either go to chemotherapy (or hormone treatment) and then MRI imaging, or else go straight to their first definitive treatment of Surgery or Radiotherapy.

Patients going for an MRI imaging as the first diagnostic activity after the first routine outpatient appointment go on to have an MDT meeting followed by a re-take of PSA test and then eventually surgery or radiotherapy.

As seen from Figure 80, the diagnostic test that takes the longest to start is the PSA test (median wait 84 days). Within the activities following an MRI imaging, the PSA test again has the longest sojourn (waiting) time of 65 days.

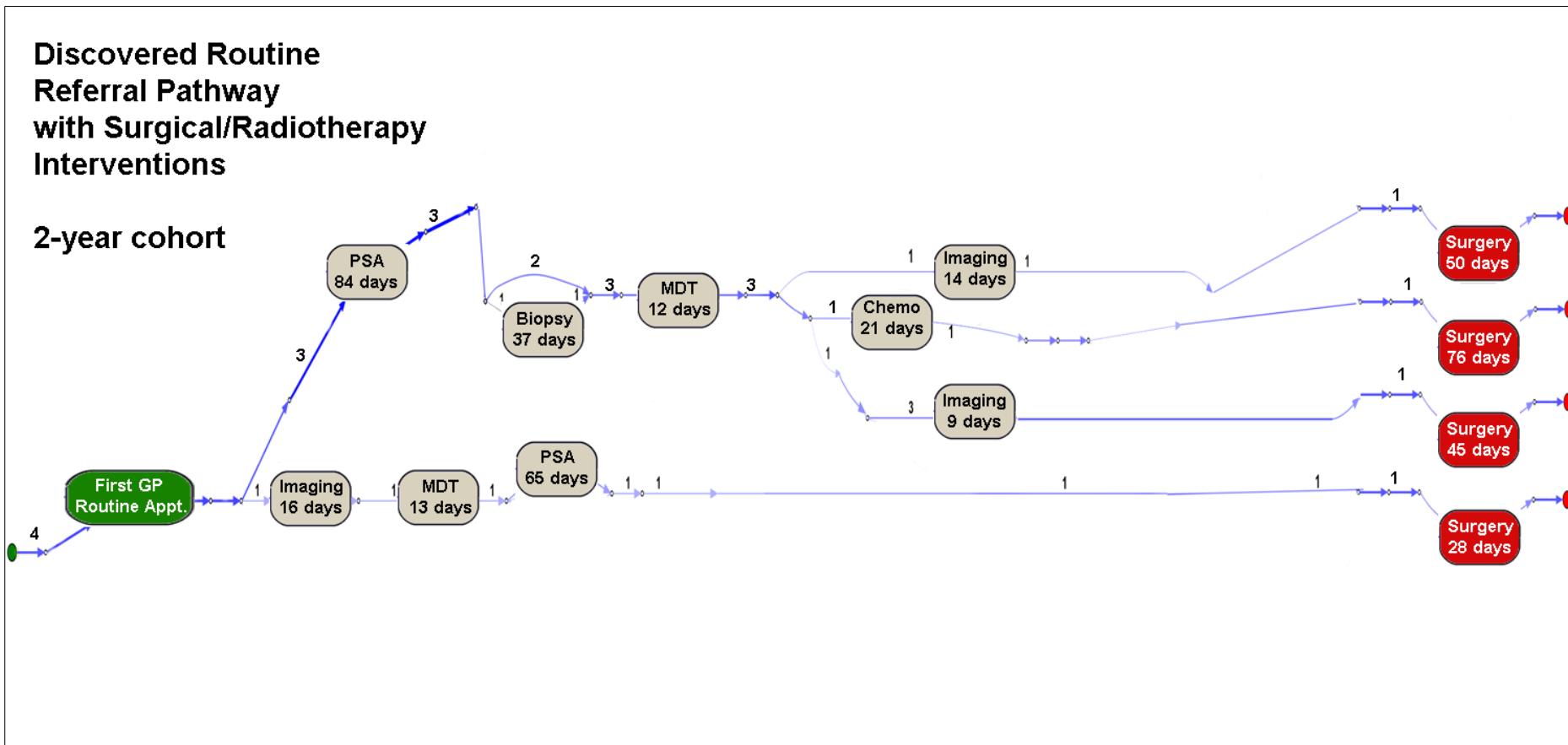


FIGURE 80: 2-YEAR COHORT ROUTINE REFERRAL CLUSTER PROCESS FLOW DIAGRAM (100% ACTIVITIES AND 100% PATHS) – MADE WITH PROM 6 INDUCTIVE MINER

PERFORMANCE MINING RESULTS

To present the results in the performance mining perspective, I am categorizing them into three categories: Trace cluster analysis, bottleneck analysis and LCA guideline compliance. All these categories will include results and comparisons from the three referral priorities: TWW, Urgent and Routine

TRACE CLUSTER ANALYSIS

Trace cluster analysis looks at all the various cluster groups within a pathway to see how many variants there are in the pathway and what are the cycle times (durations) of each cluster.

TWO WEEK WAIT REFERRALS

Within the TWW referral, after applying the Trace Variant plugin in PROM, I got 10 clusters from 11 cases, as seen in Figure 81. This means that nearly all the cases are following a unique pathway and there is extreme variance from one patient to another. Table 31 shows us that the cluster with the shortest sojourn time was cluster 7. It involves 4 activities: **GP/New/TWW Appt, MDT, Imaging** (repeated 3 times) and **Surgery**. It has a mean sojourn time of 44 days. The cluster with the longest sojourn time was cluster 10. It involves 5 activities: the **GP/New/TWW Appt, MDT, PSA test, Biopsy** and **Surgery**. It has a mean sojourn time of 229 days.

URGENT REFERRALS

Within the urgent referral, there was only one patient that was following 3 activities: **GP/New/TWW Appt, MDT and Surgery**. It had a mean sojourn time of 74 days. As there was only one patient in this referral, I am not showing the trace cluster or performance table for this referral.

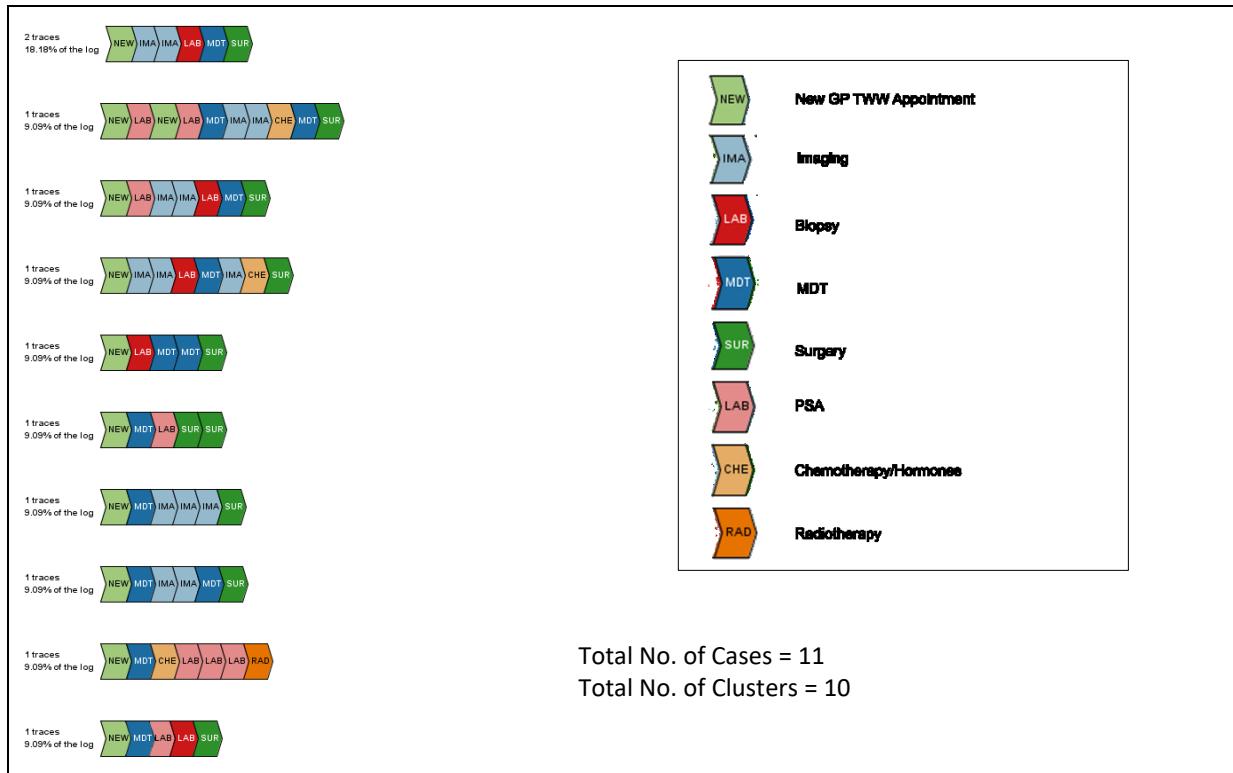


FIGURE 81: 10 CLUSTERS IN THE TWW REFERRAL

TABLE 31: PERFORMANCE OF 10 TWW REFERRAL CLUSTERS

Cluster (N=10)	No. of cases (N=11)	% of cases	Min Duration (days)	Max Duration (days)	Median Sojourn Time (days)	Mean Sojourn Time (days)
1	2	18%	76	147	112	112
2	1	9%	181	181	181	181
3	1	9%	217	217	217	217
4	1	9%	138	138	138	138
5	1	9%	228	228	228	228
6	1	9%	159	159	159	159
7	1	9%	44	44	44	44
8	1	9%	156	156	156	156
9	1	9%	160	160	160	160
10	1	9%	229	229	229	229

ROUTINE REFERRALS

Within the routine referral, after applying the Trace Variant plugin in PROM, I got 4 clusters from 4 cases, as seen in Figure 82. This means that all the cases are following a unique pathway and there is extreme variance from one patient to another. Table 32 shows us that the cluster with the shortest sojourn time was cluster 4. It involves 6 activities: **GP/New/TWW Appt, Imaging, MDT, PSA test (twice) and Surgery**. It has a mean sojourn time of 142 days. The cluster with the longest sojourn time was cluster 1. It involves 7 activities: **GP/New/TWW Appt, PSA test (twice), Biopsy, MDT, Chemotherapy/Hormones and Surgery**. It has a mean sojourn time of 453 days.

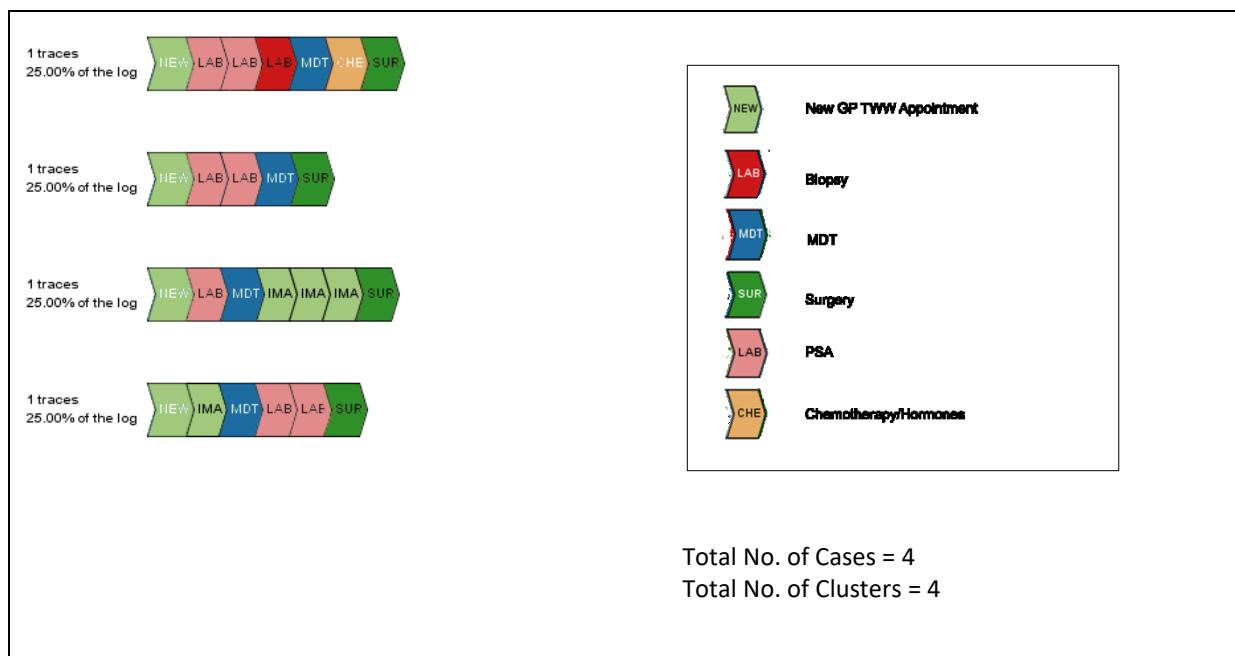


FIGURE 82: 6 CLUSTERS IN THE ROUTINE REFERRAL

TABLE 32: PERFORMANCE OF 4 ROUTINE REFERRAL CLUSTERS

Cluster (N=4)	No. of cases (N=4)	% of cases	Min Duration (days)	Max Duration (days)	Median Sojourn Time (days)	Mean Sojourn Time (days)
1	1		453	453	453	453
2	1		273	273	273	273
3	1		166	166	166	166
4	1		142	142	142	142

COMPARISON

TABLE 33: DURATION COMPARISON BETWEEN ALL THE DIFFERENT REFERRALS

Referral	Shortest Cluster Duration (days)	Longest Cluster Duration (days)	Mean Duration (days)	Mean wait until MRI (days)	Mean wait until Biopsy (days)
TWW	44	229	146	38	142
Urgent	74	74	74	N/A	N/A
Routine	142	453	259	92	121

In the comparison of the duration of the different referrals, we can see from Table 33 that TWW had a mean duration of 146 days whereas the routine referral had the longest duration of 259 days from referral until intervention. Patients going through a TWW referral had to wait an average of 38 days before they get their MRI done and 142 days before they get their biopsy done. In routine referrals, patients would wait an average of 92 days to get their MRI done and 121 days to get their biopsy done. Since urgent has only one patient I am not taking it into account.

BOTTLENECK ANALYSIS

In order to make use of the process flow diagrams from a performance perspective, I utilised the waiting times shown in days under the activity names to detect where the bottlenecks in the pathway were based on the timelines anticipated by the LCA guidelines. For each referral, I took out the different traces or paths the patients can possibly follow. Then I compared to see which traces in that referral were the shortest, longest, had the most bottlenecks and where were the delays in the handover from.

TWO WEEK WAIT REFERRALS

Within the TWW referral, as shown in Figure 83, I followed 10 traces. Traces 1, 2, 3 (Figure 84) belonged to the pathway taken by patients who followed a PSA immediately after their first TWW appointment. Trace 4 (Figure 85) was the path followed by patients going for a biopsy immediately after their first TWW appointment. Traces 5 and 6 (Figure 86) were the path followed by patients going for an MRI imaging immediately after their first TWW appointment. Traces 7, 8, 9, 10 (Figure 87) were the paths followed by patients going for an MDT immediately after their first appointment.

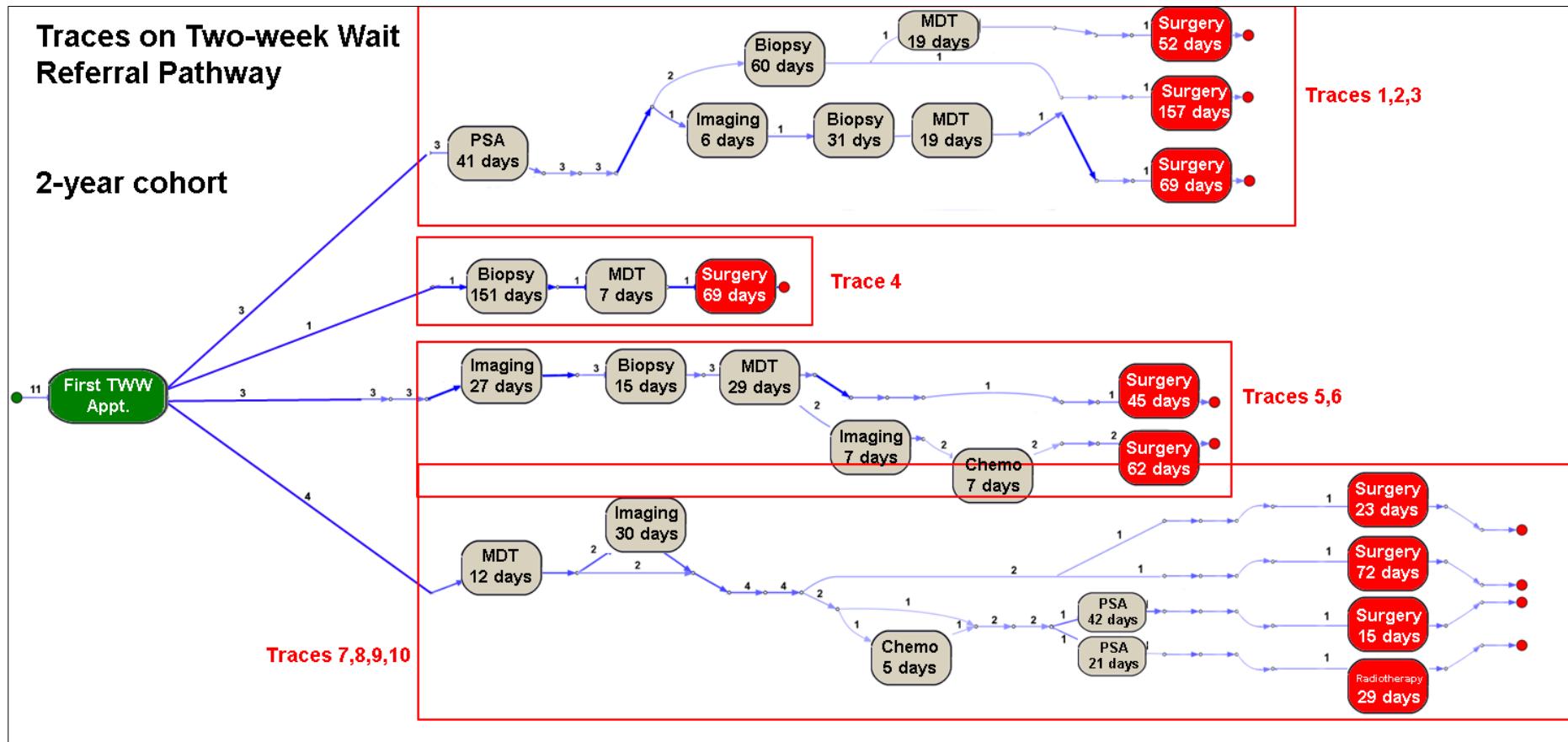


FIGURE 83: TRACES ON THE TWW REFERRAL PATHWAY

TABLE 34: BOTTLENECKS IN THE TWW REFERRAL

Trace #	Path	Total Duration (days)	No. of Bottlenecks in Path	Percentage of path that has bottlenecks	Delays in handover from	Ref.
1	GP-TWW-Appt.→PSA→Biopsy→MDT→Surgery	172	4	100%	<ul style="list-style-type: none"> • GP-TWW-Appt.→PSA • PSA→Biopsy • Biopsy→MDT • MDT→Surgery 	Figure 84
2	GP-TWW-Appt.→PSA→Biopsy→Surgery	248	3	100%	<ul style="list-style-type: none"> • GP-TWW-Appt.→PSA • PSA→Biopsy • Biopsy→Surgery 	Figure 84
3	GP-TWW-Appt.→PSA→Imaging→Biopsy→MDT→Surgery	166	4	80%	<ul style="list-style-type: none"> • GP-TWW-Appt.→PSA • Imaging→Biopsy • Biopsy→MDT • MDT→Surgery 	Figure 84
4	GP-TWW-Appt.→Biopsy→MDT→Surgery	227	2	50%	<ul style="list-style-type: none"> • GP-TWW-Appt.→Biopsy • MDT→Surgery 	Figure 85
5	GP-TWW-Appt.→MDT→Imaging→Surgery	65	2	67%	<ul style="list-style-type: none"> • MDT→Imaging • Imaging→Surgery 	Figure 86
6	GP-TWW-Appt.→MDT→Surgery	90	1	50%	<ul style="list-style-type: none"> • MDT→Surgery 	Figure 86
7	GP-TWW-Appt.→MDT→Chemo→PSA→Radiotherapy	67	1	25%	<ul style="list-style-type: none"> • PSA→Radiotherapy 	Figure 86
8	GP-TWW-Appt.→MDT→PSA→Surgery	69	1	33%	<ul style="list-style-type: none"> • MDT→PSA 	Figure 86
9	GP-TWW-Appt.→Imaging→Biopsy→MDT→Surgery	116	4	100%	<ul style="list-style-type: none"> • GP-TWW-Appt.→Imaging • Imaging→Biopsy • Biopsy→MDT • MDT→Surgery 	Figure 87
10	GP-TWW-Appt.→ Imaging→ Biopsy→MDT→Imaging→Chemo→Surgery	147	4	67%	<ul style="list-style-type: none"> • GP-TWW-Appt.→Imaging • Imaging→Biopsy • Biopsy→MDT • Chemo→Surgery 	Figure 87

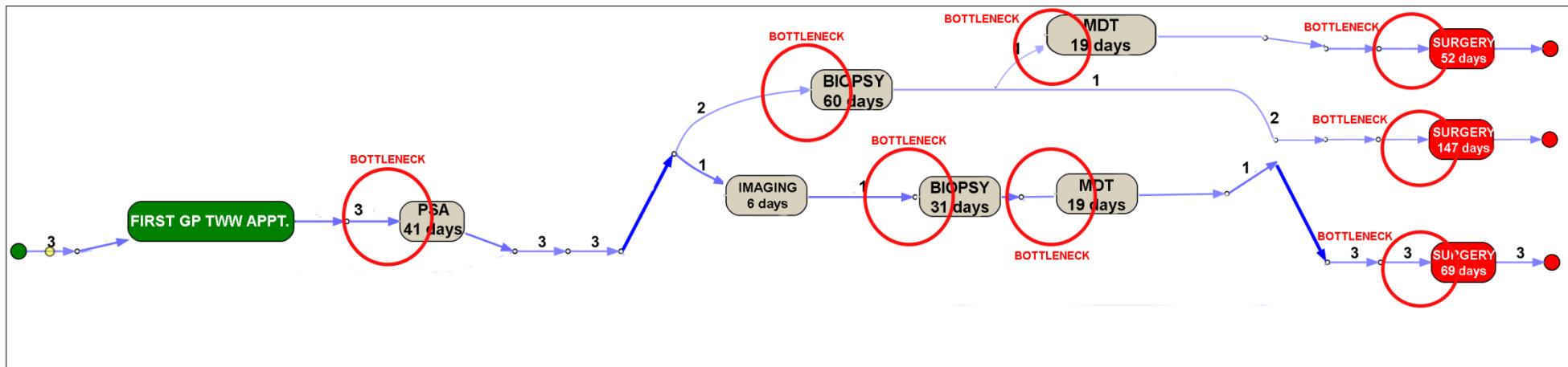


FIGURE 84: BOTTLENECKS IN TRACES 1, 2, 3 OF TWW REFERRAL

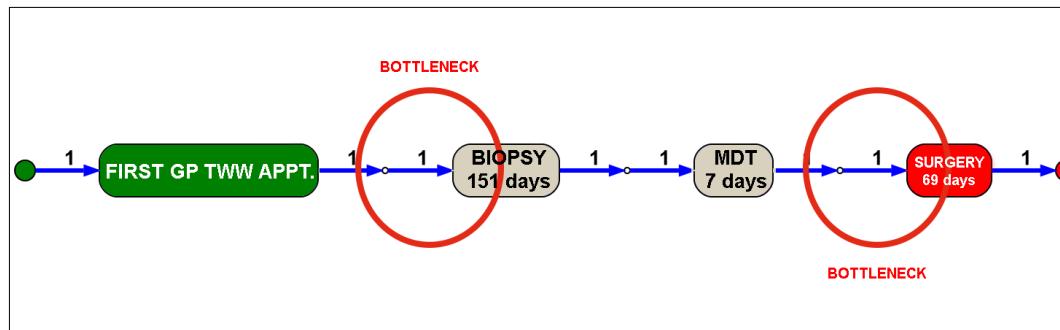


FIGURE 85: BOTTLENECKS IN TRACE 4 OF TWW REFERRAL

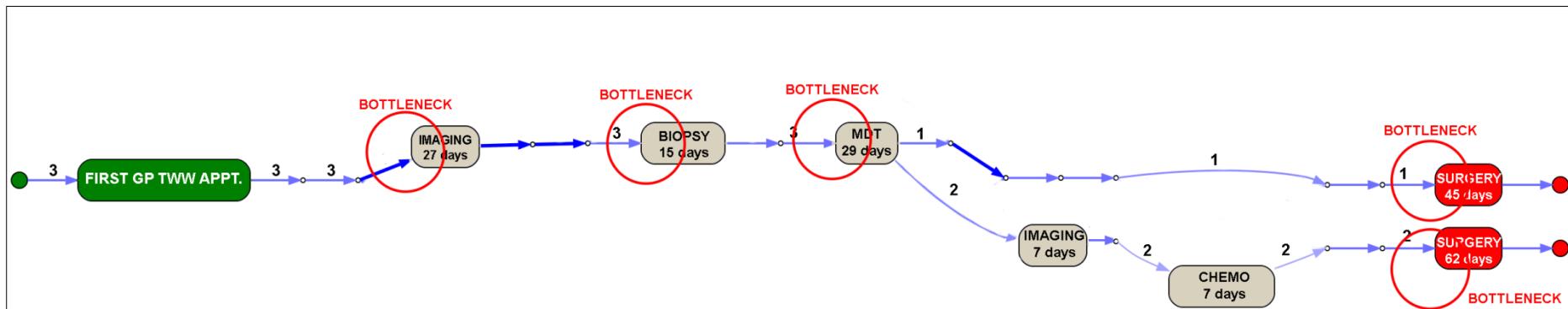


FIGURE 86: BOTTLENECKS IN TRACES 5, 6 OF TWW REFERRAL

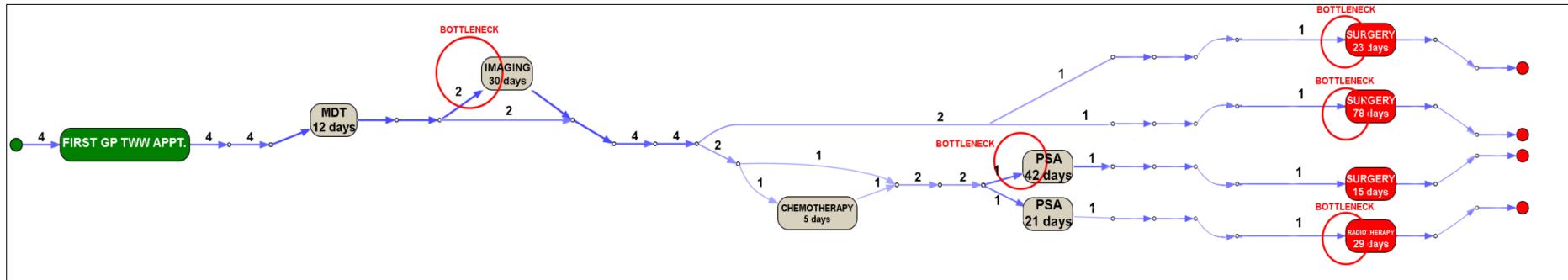


FIGURE 87: BOTTLENECKS IN TRACES 7, 8, 9, 10 OF TWW REFERRAL

Table 34 shows a breakdown of the traces and bottlenecks in the TWW referral pathway. The table highlights the 10 traces with their durations, bottlenecks and delays. The trace with the longest duration is trace 2. The traces with the maximum blockage in their paths are traces 1, 2, and 9. The maximum delays in handover are shown in Table 35:

TABLE 35: MAXIMUM DELAYS IN HANDOVER IN TWW REFERRAL

Delay maximum from which activity in pathway	No. of times	Delay mostly in handover from	Percentage of times the delay in handover takes place
MDT	7	MDT→Surgery	71%
GP-TWW-Appt.	6	G-TWW-Appt→PSA	50%
Biopsy	5	Biopsy→MDT	80%

In the above table, MDT has had the maximum number of delays (7 times) in handover mostly while handing over to Surgery.

URGENT REFERRALS

Within urgent referrals there was only one patient. This patient had one trace that was 75 days long and had a 100% blockage in its path. The delay in handover occurred amongst all the activities in the path.

TABLE 36: BOTTLENECKS IN THE URGENT REFERRAL

Trace #	Path	Total Duration (days)	No. of Bottlenecks in Path	Percentage of path that has bottlenecks	Delays in handover from	Ref.
1	GP-Urgent-Appt.→MDT→Surgery	75	2	100%	<ul style="list-style-type: none"> • GP-Urgent-Appt.→MDT • MDT→Surgery 	Figure 88

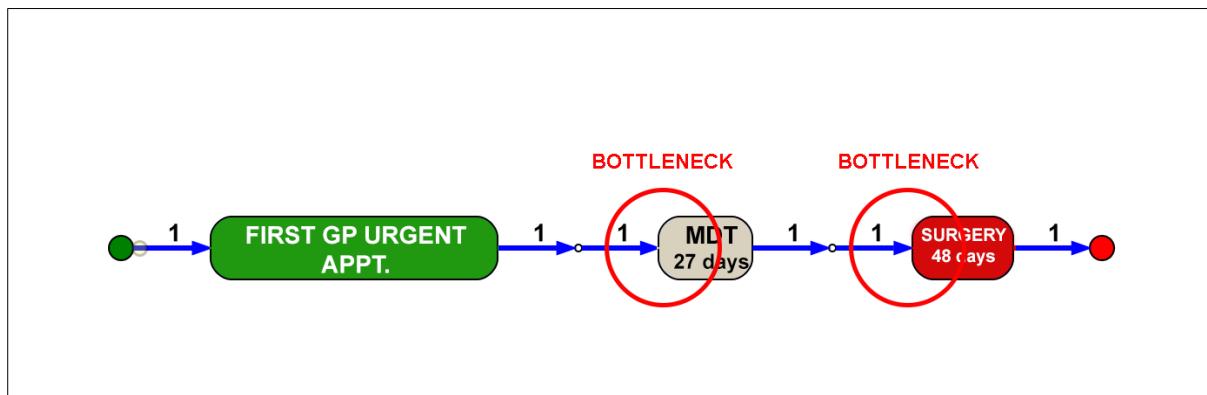


FIGURE 88: BOTTLENECKS IN URGENT REFERRAL

ROUTINE REFERRALS

Within the routine referral, as shown in Figure 89, I followed 4 traces. Traces 1, 2, 3 (Figure 90) belonged to the pathway taken by patients who followed a PSA immediately after their first routine appointment. Trace 4 (Figure 91) was the path followed by patients going for an MRI imaging immediately after their first routine appointment.

Table 38 shows a breakdown of the traces and bottlenecks in the routine referral pathway. The table highlights the 4 traces with their durations, bottlenecks and delays. The trace with the longest duration is trace 1. The traces with the maximum blockage in their paths are traces 1 and 2. The maximum delays in handover are shown in Table 37:

TABLE 37: MAXIMUM DELAYS IN HANDOVER IN ROUTINE REFERRAL

Delay maximum from which activity in pathway	No. of times	Delay mostly in handover from	Percentage of times the delay takes place
GP-routine-Appt.	4	G-routine-Appt → PSA	100%
PSA	3	PSA → MDT	67%

In the above table, GP-routine-Appt. has had the maximum number of delays (4 times) in handover mostly while handing over to PSA.

Traces on Routine Referral Pathway

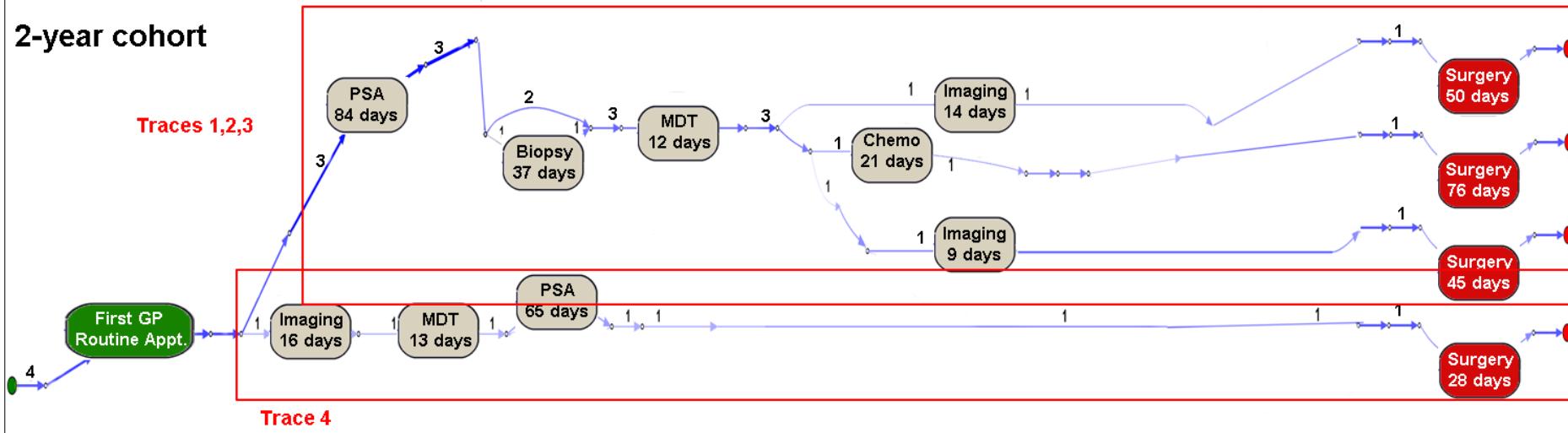


FIGURE 89: TRACES ON THE ROUTINE REFERRAL PATHWAY

TABLE 38: BOTTLENECKS IN THE ROUTINE REFERRAL

Trace #	Path	Total Duration (days)	No. of Bottlenecks in Path	Percentage of path that has bottlenecks	Delays in handover from	Ref.
1	GP- Routine -Appt. \rightarrow PSA \rightarrow Biopsy \rightarrow MDT \rightarrow Chemotherapy \rightarrow Surgery	230	5	100%	<ul style="list-style-type: none"> • GP-Routine -Appt.\rightarrowPSA • PSA\rightarrowBiopsy • Biopsy\rightarrowMDT • MDT\rightarrowChemotherapy • Chemotherapy\rightarrowSurgery 	Figure 90
2	GP- Routine -Appt. \rightarrow PSA \rightarrow MDT \rightarrow Surgery	146	3	100%	<ul style="list-style-type: none"> • GP-Routine -Appt.\rightarrowPSA • PSA\rightarrowMDT • MDT\rightarrowSurgery 	Figure 90
3	GP- Routine -Appt. \rightarrow PSA \rightarrow MDT \rightarrow Imaging \rightarrow Surgery	150	3	75%	<ul style="list-style-type: none"> • GP-Routine -Appt.\rightarrowPSA • PSA\rightarrowMDT • Imaging\rightarrowSurgery 	Figure 90
4	GP- Routine -Appt. \rightarrow Imaging \rightarrow MDT \rightarrow PSA \rightarrow Surgery	122	2	50%	<ul style="list-style-type: none"> • GP-Routine - Appt.\rightarrowImaging • MDT\rightarrowPSA 	Figure 91

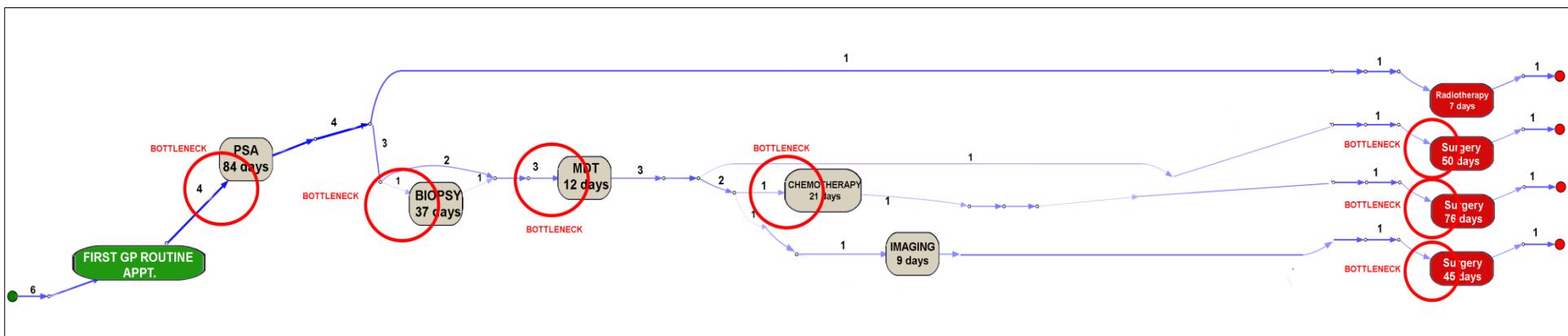


FIGURE 90: BOTTLENECKS IN TRACES 1, 2, 3 OF ROUTINE REFERRAL

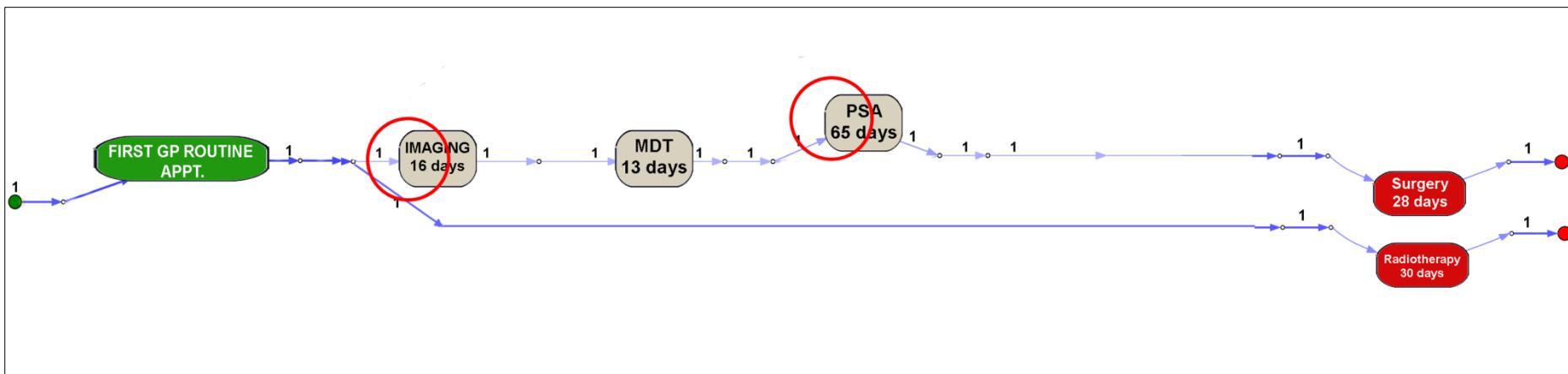


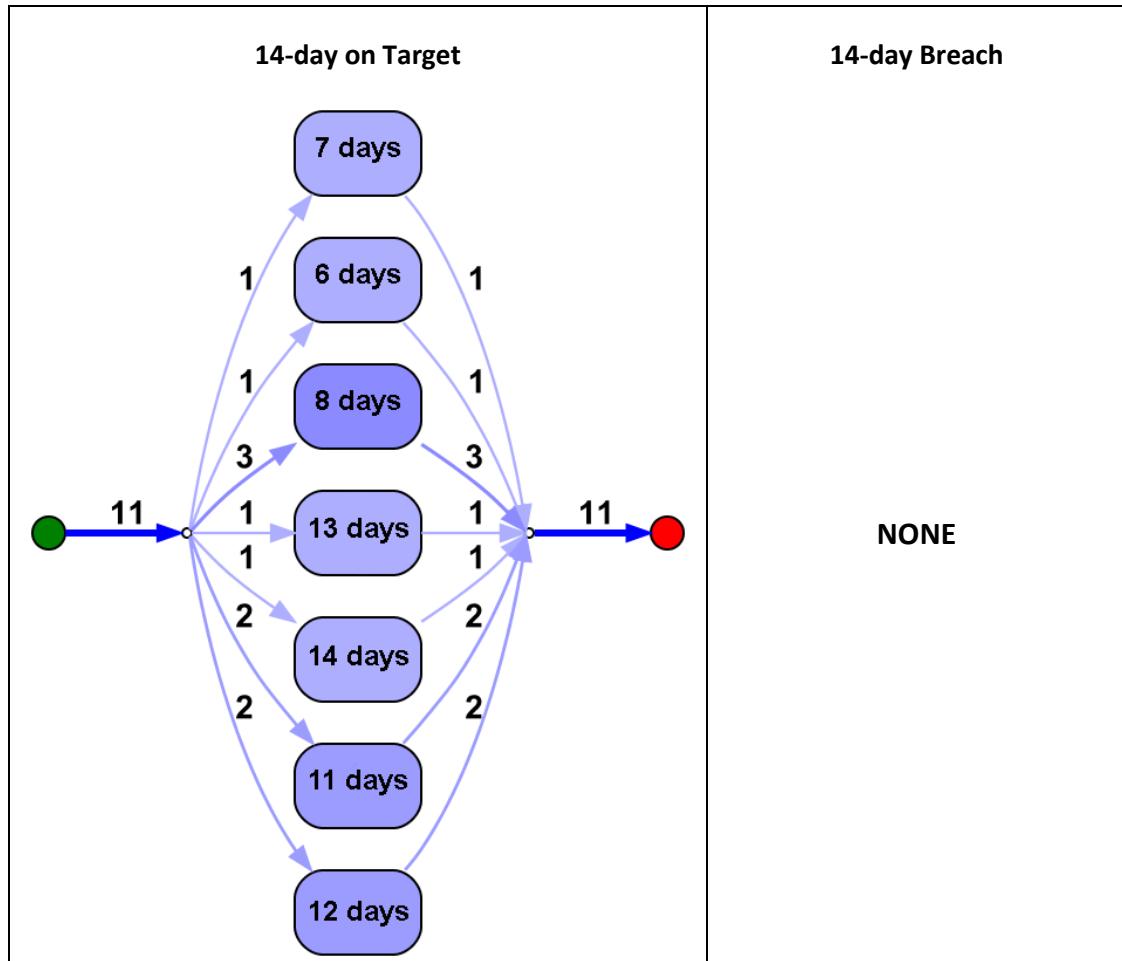
FIGURE 91: BOTTLENECKS IN TRACE 4 OF ROUTINE REFERRAL

LCA GUIDELINE COMPLIANCE

In this section I will present the results of compliance with the LCA and Cancer Waiting Times (CWT) standards. The LCA recognises the need to utilise existing data sources when monitoring compliance against best practice pathways. Therefore, they have developed metrics that are based solely on the CWT [196]. I will present the process flow diagrams of 2 metrics: First TWW Appointment and 62-day First Treatment.

TWO WEEK WAIT REFERRALS

FIRST TWW APPOINTMENT



62 DAY FIRST TREATMENT

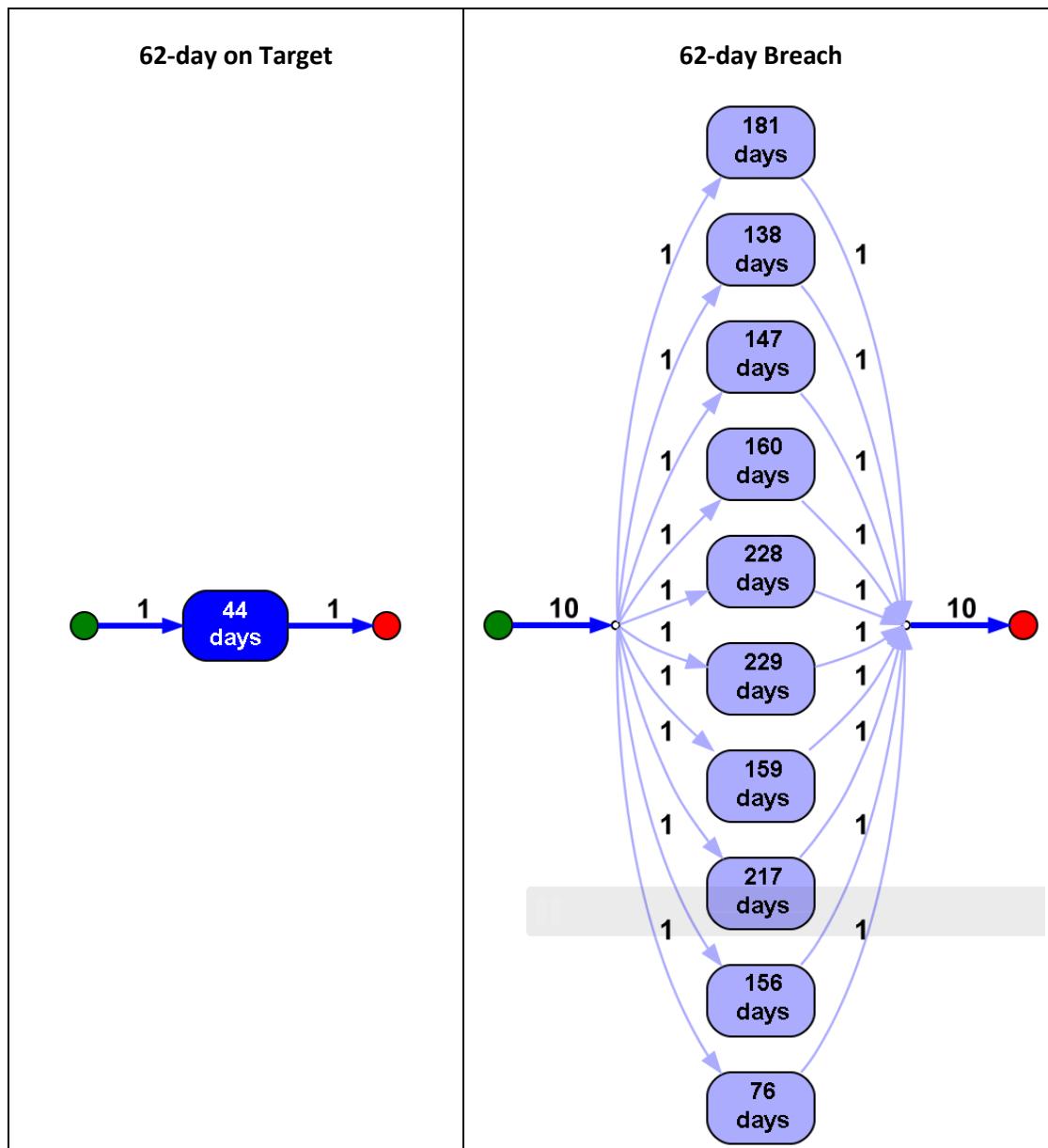


FIGURE 93: 62-DAY FIRST TREATMENT METRIC COMPLIANCE

Starting off with the First TWW Appointment metric, Figure 92 shows the compliance of my TWW cohort to the 14 day standard laid out by the CWT. This standard recommends a 14-day delay from the date of referral received to the date the first TWW appointment is made. From the total number of patients that had a TWW referral (N=11), patients that conformed to the TWW delay are all 11 patients. On your left hand is the flow diagram with the patients who have conformed to the standards. It can be seen by the diagram that the majority of the patients had an 8-day delay (3 patients). From the right hand flow diagram we can see the patients that breached this metric (0 patients).

For the 62-day First Treatment metric, Figure 93 shows us the number of patients whose date from referral received to first treatment date is <63 days. In my case I am assuming the first treatment date is either a radiotherapy or surgery date. The left side figure shows the patients that conformed to the metric (1 patient) and the right side figure shows the patients that breached this metric (10 patients).

The following Table 39 shows us the metrics and compliance percentages to the LCA and CWT standards. From this table we can see that for metric 1 my cohort complied 100% where the required target was 93%. For metric 2 my cohort complied by 9% where the required target was 85%. For metric 3 that requires the biopsy to be done within 14 days of referral, my cohort complied by 0%. For metric 4 that requires the MRI to be done within 10 days of referral, my cohort complied by 0%. Lastly, for metric 5 that requires an MRI to be done before biopsy, my cohort complied by 100%.

TABLE 39: METRIC AND COMPLIANCE TABLE WITH CANCER WAITING TIMES

No.	Metric	What are we measuring?	Data Item (s)	Cases on Target (N=1184)	Percentage on target	Target to reach
1	First 2ww appointment for prostate cancer patients	Date from referral to first appointment is to be < 14 days	2ww appointment date – 2ww referral date	11/11	100%	93%
2	62 day first treatment	Date from referral to first treatment <63 days	First treatment date – 2ww referral date	1/11	9%	85%
3	Biopsy	Date from referral to biopsy < 14 days	Sample collection date – 2ww referral date	0/7	0%	Not yet set by operational standards
4	MRI	Date from referral to MRI < 10 days	Procedure date (if imaging modality = MRI scan) – 2ww referral date	0/8	0%	Not yet set by operational standards
5	Pre Biopsy MRI	Date of MRI to be before date of biopsy	Sample collection date – Procedure date (if imaging modality = MRI scan)	4/4 MRI before Bx	100%	Not yet set by operational standards

URGENT REFERRALS

Since there was only one patient in the urgent referral, I will not do the LCA compliance of this patient as there are not enough patients and activities to perform a proper compliance on.

ROUTINE REFERRALS

FIRST TWW APPOINTMENT

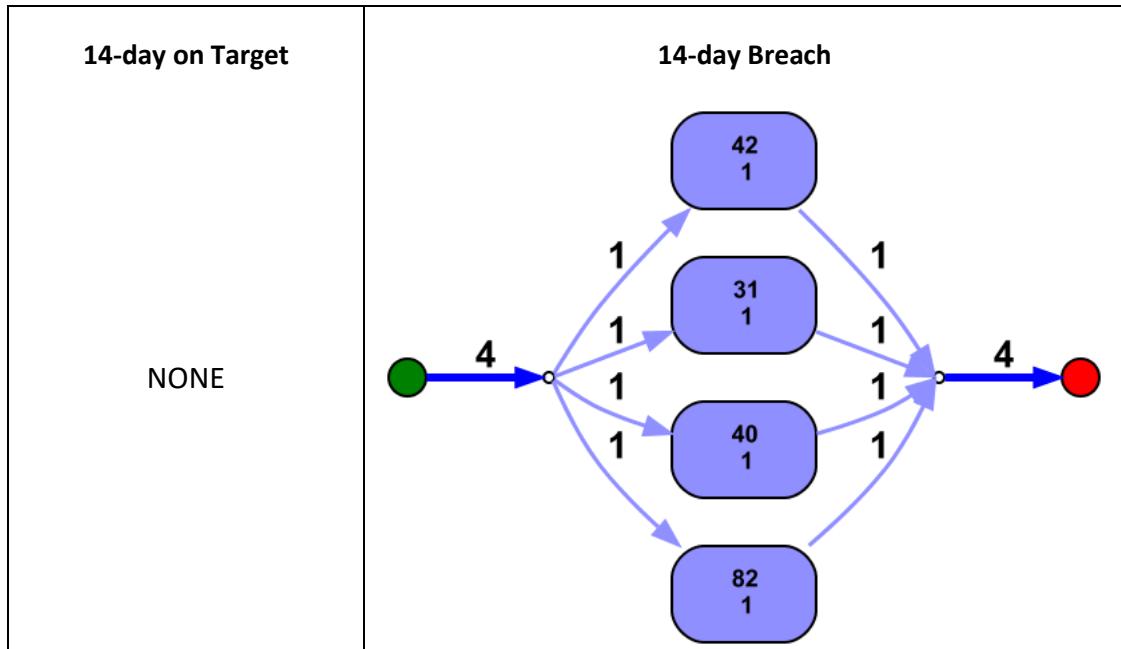


FIGURE 94: FIRST TWW APPOINTMENT METRIC COMPLIANCE

62 DAY FIRST TREATMENT

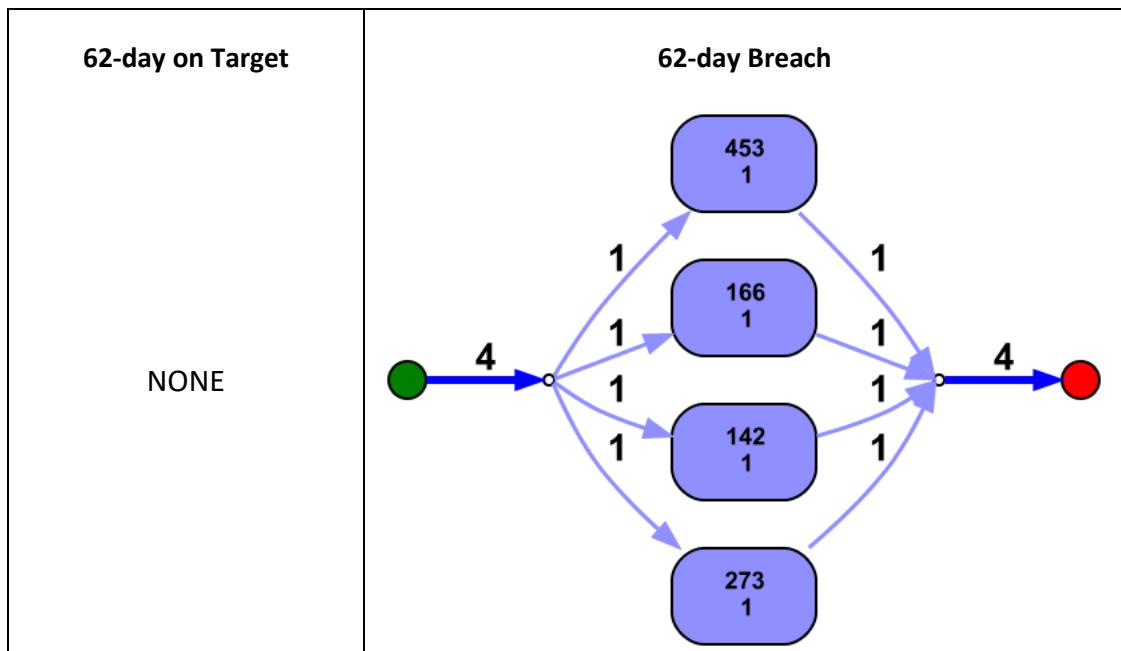


FIGURE 95: 62-DAY FIRST TREATMENT METRIC COMPLIANCE

Starting off with the First TWW Appointment metric, Figure 94 shows the compliance of my Routine cohort to the 14 day standard laid out by the CWT. This standard recommends a 14-day delay from the date of referral received to the date the first appointment is made. From the total number of patients that had a routine referral (N=4), patients that conformed to the TWW delay are none of the patients. On your left hand is the flow diagram with the patients who have conformed to the standards (0 patients). From the right hand flow diagram we can see the patients that breached this metric (all 4 patients).

For the 62-day First Treatment metric, Figure 95 shows us the number of patients whose date from referral received to first treatment date is <63 days. In my case I am assuming the first treatment date is either a radiotherapy or surgery date. The left side figure shows the patients that conformed to the metric (1 patient) and the right side figure shows the patients that breached this metric (10 patients).

Since all the TWW metrics were defied, I will not show the metric table for this referral.

7.5 DISCUSSION

In this chapter, I have focused on obtaining in an explorative way, insights into the prostate cancer healthcare process. I have given a transparent way of visualising this pathway with the use of process mining techniques that has aided in the discovery of bottlenecks and deviations in the pathway as well as highlighting delays that lead to performance issues.

To discover a suitable process model it is assumed that the event log contains a representative sample of behavior. If we randomly take 400 cases from a set of say 1000 cases, we would like to discover more or less a similar model every time. In every real-life log there are traces that appear only once in the log and may disappear when considering a sample of less cases. Moreover, there are dozens of cases and activities that are rare or infrequent. Such rare behaviors add to the “noise” of the event log and cannot be representative of the typical behaviour of the process. With that said, it is also important to point out that noise is not irrelevant. In fact, in a healthcare setting where we expect substantial variation from one patient pathway to another, we need to identify these exceptions and deviations to better direct our resources and decision-making.

As previously explained in Chapter 5, the first initial look at the process model of the complete 5-year event log resulted in a heavily cluttered, unreadable process flow diagram. Hence, to reduce the spaghetti diagram, I performed filtering of the log based on years and the start of a GP appointment (TWW/Urgent or Routine) within the timeframe of the years chosen. The reason why I chose the start of the first GP appointment as a mandatory starting point for my event log is because of the innumerable number of patients that had a pathway being carried over into my date range from the previous years (outside my range) and were showing as cases without proper starting points. Moreover, there were many patients that came in as walk-in or routine patients, had their PSA done as part of a regular check-up and left. Therefore, in order to avoid including these kind of patients who probably never developed cancer, or who had their cancer pathway started at a previous point in time outside my range, I decided to put a filter to include only the cases that began their journey inside my time frame with a proper GP referral (either TWW, urgent or routine) and ended up having either a prostatectomy or radiotherapy to ensure that we are capturing only the patients who had prostate cancer. With this filtering, I used a reverse engineering technique to go back and trace the exact path these patients took to reach their intervention.

I also explored the third definitive treatment choice: active surveillance by forward-tracking the patients from the time the treatment choice was made until either the patient received a surgical or radiotherapy treatment or then continued on active surveillance. However, I encountered many limitations in this cluster as it was getting increasing difficult to differentiate the starting point of the definitive active surveillance choice as compared to a normal delayed routine or TWW pathway of the patient. Moreover, the event log contained too few follow-up events related to the surgery and radiotherapy treatments after an active surveillance and made it difficult to discover some of the underlying control-flow structures. Hence, I concentrated on the Prostatectomy and Radiotherapy interventions only.

As mentioned previously, the lack of MDT data for the years 2010-2012 prompted me to only collect the results for the 2-year cohort from 2013-2015 despite the fact that I had very little patients that matched the inclusion criteria. MDT meeting is extremely vital in deciding whether the patient has cancer or not and the planned intervention. If MDT information is not found for a patient it is difficult to highlight the exact sequence of processes. Without MDT information, it becomes increasingly difficult to assess if the patient's pathway is delayed due to a bottleneck in the pathway or is the patient just on active surveillance. Hence, I decided to only look at the years with properly

recorded MDT information so that I can filter out the patients on active surveillance (by looking at their MDT notes) and in that way I was only left with the surgical and radiotherapy intervention patients to properly analyse. Having decided on a 2-year cluster (from 2013-2015) and further segregated my results based on the priority of referral, I divided my results into two sections based on the process mining perspective being analysed: Control flow and Performance.

Starting with the control-flow mining results of the TWW segmentation, it can be seen from Table 27 and Figure 76 that the majority of the patients (36%) started off directly with an MDT appointment after their first GP TWW appointment. Practically speaking, this is not possible as at least one diagnostic test is required (other than a GP PSA) to be included in the MDT discussion. This brings me to point out another deficiency we had with respect to the Biopsy data. We had a very difficult time acquiring biopsy data from the custodians of the histopathology database and it took us more than 3 years to come to an agreement with them to provide us with a restricted view of that data. Therefore, a large percentage of the pathway will have biopsy information missing in several places. One of the places that likely seems like it is missing biopsy information is the unexpected start of the pathway with an MDT meeting without a prior biopsy/MRI. Another fact that supports my inference is if we see the number of patients going to a direct biopsy after the first GP TWW appointment we can see that it's only 1 patient, which seems highly unlikely. Therefore the most possible assumption would be that the patients going on a direct MDT track are actually going through a biopsy first, but it is not available in my data. With that assumption in mind, if we can merge the two sets of patients separately going on biopsy and MDT tracks together as one and say that all of them go through a biopsy first then MDT, then in that way we can say that based on my data the majority of the patients then go through a biopsy as the first diagnostic test more than an MRI imaging. Similar assumptions can be deduced in the Urgent and Routine referrals. With the control flow mining results it can be seen that the majority of the path conforms to the activities indicated in the LCA process flow diagram, however performance –wise they have been very far off.

Moving on to the performance mining results, I have compared all the referral priorities equally although I am aware that the cancer performance metrics are only applicable to TWW referrals. The reason for this is that it is interesting to compare the delays, bottlenecks and pathway durations of patients who are being treated for prostate cancer via a TWW referral as opposed to patients being treated for cancer via a routine or urgent referral. Additional, it is also possible that patients are wrongly referred via the urgent or routine referrals when they should have been referred via TWW.

Looking at the TWW referral, it can be seen in the variant analysis that there are 10 distinct patient clusters from 11 patients. This shows that almost all the patients in the entire TWW segmentation are going through their own unique path. The shortest cluster in the TWW referral had a total duration of 44 days but the longest cluster had a total duration of 229 days indicating a severe blockage in its path.

The routine referral had 4 clusters out of 4 patients. It can also be seen with the trace duration of the routine referral that all the 4 traces took a very long time to complete (the shortest being 142 days). This shows that patients going through a routine referral, who end up having cancer, take a very long time until their intervention is started.

When comparing the median duration of days between a GP referral and the intervention to start, the TWW referral took 145 days whereas the routine referral took a much longer time of 259 days again proving that patients referred via a routine referral will take longer to reach intervention.

Patients going through a TWW referral had to wait an average of 38 days before they get their MRI done and 142 days before they get their biopsy done. Although both these timelines seem comparatively long, however, one of the biggest known problems in prostate cancer is its capacity in getting an MRI done [200]. Moreover, another delay in the pathway is caused by the turnaround time for a biopsy to be done and its report to be sent out. Hence, the results of such long wait times between these two activities. In routine referrals, this wait time is even more extended as patients would wait an average of 92 days to get their MRI done and 121 days to get their biopsy done.

A bottleneck analysis is very important and useful to interpret as it can clearly tell the stakeholders where the problems and delays are taking place in the processes of a cancer pathway and lets them make an informed decision about the measures that can be taken to deal with those issues. For the bottleneck analysis of the TWW referral there were severe bottlenecks in 26 out of 37 activity handovers. That means 70% of the pathway was blocked, hence the severe delays. The maximum bottlenecks were occurring in handover from MDT to surgery suggesting either a delay in decision making or then delays in follow-up appointments and patient thinking time leading to the intervention. Similarly, in the urgent referral 100 % of the pathway was blocked and in the routine referral 81% of the pathway was blocked (13 out 20 activities).

The results of the LCA and CWT compliance have shown that in the first metric of the TWW referral, all the 11 patients were conforming to the LCA metric of a 14-day delay to the first appointment. The majority of patients getting a TWW appointment were within 8 days. With this we conformed to 100% of the standards (where the required target was only 93%) as seen in Table 39.

For the second metric showing the 62-day first treatment my cohort was only able to comply by 9% (Table 39) as compared to the required 85%.

The remaining metrics shown in Table 39 show us the biopsy and MRI date standards and whether an MRI was done pre-biopsy as recommended by the national standards. None of the patients in my cohort were complying with these metrics. It is also important not to forget here that the patients I am getting are not representative enough of the entire cohort as I am missing on so much important MDT and Biopsy data.

In conclusion, I would like to add that the existing way of creating pathways and measuring performance and adherence to pathways is laid out clearly by LCA. These guidelines assume that there is a homogenous group of patients and they try to funnel these patients through the guideline flowchart that they have made as a standard. By using Process mining I am trying to show that in reality not everyone is adhering to that. I'm giving everyone the ability to have visibility and transparency to see what is happening in the pathway exactly and thus I'm showing the massive heterogeneity. The main message in this chapter is that you cannot benchmark performance against percentage compliance. If patients are not going according to the LCA guidelines because it is not appropriate for them then we are just disadvantaging and penalizing those trusts that are not getting the compliance, when in reality their patients are getting the correct care.

The LCA guideline is just a guideline - it guides the carers what to do. My research is not trying to undermine the guidelines. I am only showing a new method and technique of visualising the heterogeneity in the pathway.

7.6 SUMMARY

The chapter describes the process mining technique I have used to analyse the prostate cancer pathway and visually present it. I have described my initial pre-processing steps required to filter the event log. I have then shown the actual patient journey in the prostate cancer pathway for my 2-year cohort. I have analysed the cohort based on control flow mining and performance mining methods and have compared the derived waiting times using process mining against the known standard waiting times as portrayed in the LCA guidelines. The next chapter, Chapter 8: Evaluation of Process Mining Visualisations, continues with the last phase of the CPAM roadmap: phases 6, and provides an evaluation of the visualisations produced in this chapter through the Microsoft's Reaction Cards evaluation technique that was used on clinical, IT and administrative staff.

CHAPTER 8: EVALUATION OF PROCESS MINING VISUALISATIONS

This chapter continues on to the last phase of the CPAM roadmap: phases 6 (Figure 96). In this chapter an evaluation technique is presented that evaluates the visual aspect of the process mining technique used in this study to map the patient pathway. Section 8.2 starts with a background on types of evaluation methods; how to do an evaluation; and methods of evaluating user experience. Section 8.3 describes the methods I used in conducting the evaluation using Microsoft's Reaction Cards. Section 8.4 presents the results of my evaluation. Section 8.5 provides a discussion of my results as well as lessons learnt and limitations of the evaluation. This chapter is concluded with section 8.6 that gives a summary of the entire chapter.

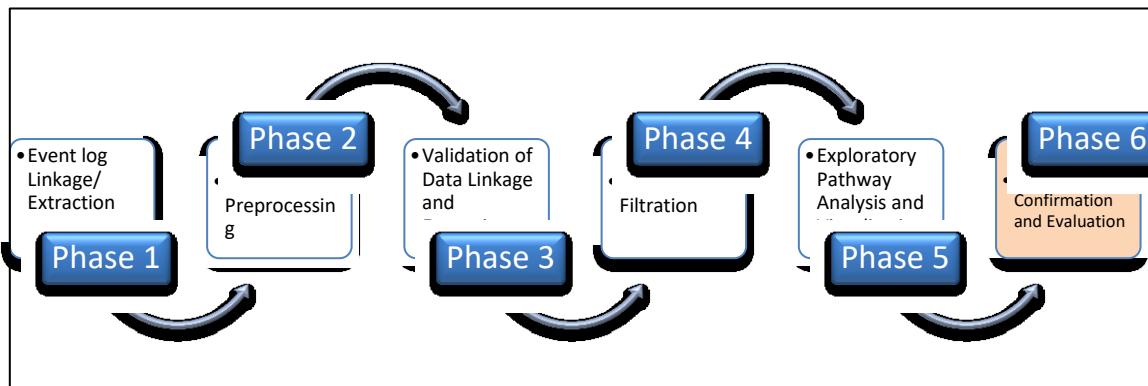


FIGURE 96: PHASE 6 (MEDICAL CONFIRMATION AND EVALUATION) OF THE CPAM ROADMAP

8.1 INTRODUCTION

Information visualization is the art of representing data in a way that it is easy to understand and manipulate and can aid us in making sense out of the information so we can use it in the most efficient way. As information visualization matures in its techniques and modalities, there is an increased demand and necessity in evaluating those techniques in the most useful way.

8.1.1 CONTRIBUTIONS OF THIS CHAPTER

In this chapter I have provided an evaluation of the visualisations produced by the process mining technique using a method called Microsoft's Reaction Cards. The evaluation has captured the users'

comments and description of how they feel about the visualisation. Furthermore, a quantitative analysis on the user's choice of words was also given to show a more tangible outcome concerning the user experience.

8.2 BACKGROUND

It is not sufficient to simply evaluate visualisations loosely but in fact a clear distinction should be made as to what type of evaluation is done and needed at what stage. A few questions that can help in finding the appropriate visualisation can include: Is the evaluation performed by an evaluation specialist or test users? Is the purpose of the evaluation to provide design feedback, or to demonstrate the usability of a particular product, or to objectively compare two or more interfaces? [201]. It is therefore essential that we make a clear distinction in the beginning as to what type of evaluation we intend to perform and what are the goals of our evaluation.

8.2.1 TYPES OF EVALUATION METHODS

Evaluation methods are divided into two types based on who performs the evaluation and what is the goal or purpose of the evaluation [201].

Evaluation methods based on *who performs* the evaluation:

1. Inspection methods: specialist evaluators inspect an interface and use their experience and judgment to assess it.
2. Testing methods: test users use one or more interfaces and observations or measurements are made.

Evaluation methods can also be classified according to *their purpose* (Table 40):

1. Exploratory: Exploratory evaluation provides feedback of how an interface is used and what it is used for.
2. Predictive: Predictive evaluation predicts user performance based on the interface design.
3. Formative: Formative evaluation provides design feedback in the form of a list of problems and its associated solutions.
4. Summative: Summative evaluation provides an overall assessment of a single interface or a comparison of multiple interfaces, and provides feedback in a quantitative form which is statistically analysed.

TABLE 40: EVALUATION METHODS CLASSIFIED ACCORDING TO TYPE AND METHODS

<i>Method</i>	<i>Type</i>	<i>Purpose</i>	<i>Description</i>
Observational Study	Testing	Exploratory	A longer term study following a small sample of users as they use an interface for their own tasks. Observations and anecdotal evidence are collected and assessed.
Action Analysis	Inspection	Predictive	An evaluator produces an estimate of the time an expert user will take to complete a given task, by breaking the task down into ever smaller steps and then summing up the atomic action times.
Heuristic Evaluation	Inspection	Formative	A small team of evaluators inspects an interface using a small checklist of general principles and produces an aggregate list of potential problems.
Guideline Checking	Inspection	Formative	An evaluator checks an interface against a detailed list of specific guidelines and produces a list of deviations from the guidelines.
Cognitive Walkthrough	Inspection	Formative	A small team walks through a typical task in the mind set of a novice user and produces a success or failure story at each step along the correct path.
Thinking Aloud	Testing	Formative	Representative test users are asked to think out loud while performing a set of typical tasks. The insight gained into why problems arise is used to produce a list of recommendations.
Guideline Scoring	Inspection	Summative	An evaluator scores an interface against a detailed list of specific guidelines and produces a total score representing the degree to which an interface follows the guidelines.
Questionnaires	Testing	Summative	After using one or more interfaces for some typical tasks, test users are asked to rate the interface(s) on a series of scales.
Formal Experiment	Testing	Summative	A larger sample of users performs a set of tasks on one or more interfaces. Objective measurement data is collected and statistically analysed.

The purpose of summative evaluation is to obtain a seal of approval (demonstrates superiority over other techniques); the purpose of formative evaluation is to improve a design (leads to better and more usable systems); the purpose of explorative evaluation is to find out, to provide knowledge (demonstrates utility or fitness for purpose); and the purpose of predictive evaluation is to predict problems users will encounter without actually testing the system.

8.2.2 HOW TO DO THE EVALUATION

Before embarking on an evaluation, it is necessary to think and decide on a few key points to make it easier to conduct the evaluation. These key points are:

- 1- Purpose of evaluation – It is important to know and be clear from the beginning what you are hoping to gain out the evaluation. Your aim should be to understand if, when and under what circumstance a visualisation or design technique works or is useful
- 2- Evaluation measures – What you are measuring should be relevant and useful to your goals and should not be merely measured without a purpose

- 3- Success and Failure – If your evaluation is a success or failure, what new knowledge will you have gained that you do not already know? If the answer is “nothing”, then you do not need to do the evaluation
- 4- Quantitative and qualitative measures – It is best to combine both quantitative and qualitative measures in order to show that certain behaviour occurs and to also know why it occurs.

8.2.3 EVALUATING USER EXPERIENCE (UX)

Evaluating User Experience (UX) refers to the act of using methods and tools to find out how a person perceives a system or visualisation either in a short or long time span. The goal is to understand whether or not the visualisation has accomplished what the participants desire to see. Therefore, UX focuses on having a good understanding of the needs and abilities of the users.

Some questions that can be considered useful when preparing for a UX evaluation can be:

- 1) What features are seen as useful?
- 2) What features are missing?
- 3) How can I change the features to make it better?
- 4) Is the tool understandable and can it be learned?

USER EXPERIENCE (UX) METHODS

UX evaluations can take varied forms. The following methods are a collection of just a few [202]:

Informal Evaluation

An informal evaluation is performed by giving a demo of the visualisation to a group of people who are experts in the domain. The users are allowed to play with the system while the evaluators take on a very informal approach to assess the evaluation. The method does not generally have a predefined structure or list of tasks to perform. It is the simplest and most common kind of evaluation. These types of evaluations can assess intuitiveness, functionality, design flaws, user preferences, and ideas for improvements and enhancements.

Usability Test

A usability test is carried out by observing how participants perform a set of predefined tasks. During the usability test, the evaluators take note of interesting remarks, behaviours and feedback from the participants as well as any problems that occur during the interaction. This method involves a careful preparation of tasks to be measured and feedback material like questionnaires and interviews. Its goal is to improve the overall final design by identifying major flaws and deficiencies in existing prototypes.

Field Observation

The main goal of field observations is to understand how users interact with the tool in a real setting where they are freely allowed to use and play around with the system in order to derive valuable information and patterns that can be used to improve the system. Sometimes, this kind of study can be followed by questionnaires or interviews to better understand the nature of the observed patterns.

Microsoft's Reaction Cards

Developed by Microsoft experts and first reported in 2002, the reaction cards are a collection of words that can be used at the end of a user research session, to capture a user's more emotional response to a prototype or visualisation. It was driven by the limitations of standard feedback mechanisms such as Likert scales. These reaction cards provide "a way for users to tell the story of their experience, choosing the words that have meaning to them as triggers to express their feelings – negative or positive – about their experience" [203].

The reaction cards can be used both as a tool to capture the user's comments and description of how they feel about a particular visualisation, as well as more quantitative analysis to give more tangible results concerning UX.

At the end of a research session or demo of a visualisation, participants are typically asked to select from the list of reaction cards a maximum of 5 adjectives (see Appendix F) that best reflect their experience with the system. After making a first selection, they are also requested to comment on their five choices, thus providing a better insight into user's experience. In its original design, there was a list of 118 cards to be used in the method; however, several studies have reported using only a limited selection of those adjectives. Results of reaction cards can be analysed and used in a variety

of ways. Classical graphs enable us to observe more precise distances between selected adjectives within a visualisation. However, the results are most commonly presented in the form of a word cloud (Figure 97).

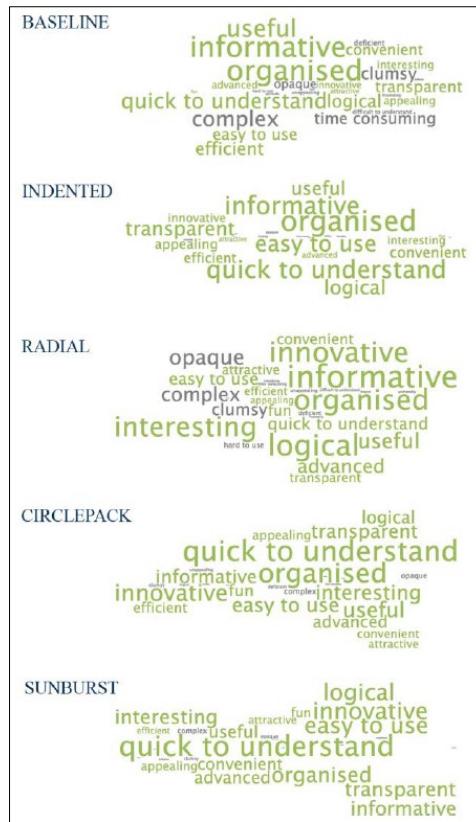


FIGURE 97: WORD CLOUDS FOR INDIVIDUAL PROTOTYPE DESIGNS [203]

8.3 METHODS

In order to conduct the evaluation with the maximum efficiency and least amount of time, I have chosen to use the Microsoft Reaction Cards method to do my UX evaluation. I invited two groups of people on two separate days to attend a presentation in which I demonstrated how process mining techniques can be used to visualise pathways of care. I showed them the TWW process flow diagrams from Chapter 6 as well as animated versions of the flow diagrams to present to them the full capabilities of process mining. I further showed them how process mining can be used to analyse performance issues from bottlenecks to variant analysis and clustering.

The people were grouped according to their professions as it was getting difficult to gather them all together in one day. I therefore had the MDT coordinators/ Cancer Quality Assurance administrators on one day; and the clinical researchers and physicians on the other day.

Using the version of Microsoft's Reaction Cards that held the concise number of 64 adjectives, I asked the participants at the end of my presentation to choose 5 cards that best described the process mining visualization they saw. Since I was meeting people face-to-face, I printed out the concise version of the list of words for people to choose from (e.g. by ticking or circling). Furthermore, I requested them to leave a brief description next to each word they chose so I could capture their exact feelings and the user's experience with the new technology. In the end, everyone wrote down their profession so it would be easier for me to segregate the results based on profession.

Once I got the results, I grouped the words into 5 different dimensions which aided me to analyse different aspects of the visualisations:

- Perceived ease of use (e.g. clear, effortless, friendly, intuitive, etc.),
- Perceived usefulness (e.g. helpful, relevant, valuable, meaningful, etc.)
- Perceived efficiency (e.g. effective, responsive, time-saving, fast, etc.)
- Appeal (e.g. appealing, attractive, desirable, impressive, novel, etc.)
- Engagement (e.g. engaging, exciting, entertaining, motivating, inspiring, etc.)

To present the results, I decided to use two techniques: a classical graph chart of the words and their frequencies used, and a word cloud to show the word(s) that were picked the maximum number of times by the users. To show a word cloud, I used a website called Wordle that generated these word clouds from text that I provided. The clouds give greater prominence to words that appear more frequently in the source text. Wordle gives the ability to tweak your clouds with different fonts, layouts, and color schemes [204].

In order to make a word cloud, I utilised an MS Excel sheet program specifically written to prepare user input data to be used in a wordle cloud. The Excel sheet had macros pre-built to help analyse the data. The spread sheet was divided into three columns (Figure 98). The first column contained all the words from the Microsoft Reaction Cards, the second column was left blank to enter the number of users who selected the words, and the last column was a derived column of hexadecimal numbers used specifically in wordle to generate a word cloud. I entered the number of users who

chose a particular word next to the respective word and the macro outputted the wordle string (in hexadecimal numbers) adjacent to it in the column to the right. If a word count was left to zero, the wordle string would assign the hexadecimal number FFFFFF to it.



Words	Number of users who chose this word	Wordle string
Accessible	0	Accessible:0:FFFFFF
Advanced	0	Advanced:0:FFFFFF
Ambiguous	0	Ambiguous:0:FFFFFF
Annoying	0	Annoying:0:FFFFFF
Appealing	0	Appealing:0:FFFFFF
Approachable	0	Approachable:0:FFFFFF
Attractive	0	Attractive:0:FFFFFF
Awkward	0	Awkward:0:FFFFFF
Boring	0	Boring:0:FFFFFF
Bright	0	Bright:0:FFFFFF
Business-like	0	Business-like:0:FFFFFF
Busy	0	Busy:0:FFFFFF
Clean	0	Clean:0:FFFFFF
Clear	0	Clear:0:FFFFFF
Cluttered	0	Cluttered:0:FFFFFF
Compelling	0	Compelling:0:FFFFFF

A	E	Number of users who chose this word	Wordle string
1		Words	
2	Accessible	7	Accessible:7:6A6A6A
3	Advanced	0	Advanced:0:FFFFFF
4	Ambiguous	1	Ambiguous:1:E9E9E9
5	Annoying	2	Annoying:2:D4D4D4
6	Appealing	8	Appealing:8:555555
7	Approachable	8	Approachable:8:555555
8	Attractive	5	Attractive:5:949494

FIGURE 98: EXCEL SHEET TO PREPARE DATA TO GENERATE WORD CLOUDS IN WORDLE

Once this was done, I then copied the entire hexadecimal codes of all the words and pasted them in the specified area designated in the Wordle website (Figure 99)



The screenshot shows the Wordle Advanced interface. At the top, there's a navigation bar with links for Home, Create, Credits, Forum, FAQ, and Advanced. Below that, the main title is "Wordle Advanced!". It has two input sections:

- Paste weighted words or phrases here:** This section includes an "Example" field containing "fruitbats:133", "llamas on parade:85.43", "zombies?:420", and "donuts:50". Below it is a text input field with a scroll bar and a "Go" button.
- OR** (separated by a horizontal line)
 - Paste weighted words with hex colors here:** This section includes an "Example" field containing "fruitbats:133:4411AA", "llamas in space:85.43:00FF40", "zombies!:420:6280AA", and "donuts:50:000000". Below it is a text input field with a scroll bar and a "Go" button.
 - Background color, in hex:** A text input field containing "FFFFFF".

FIGURE 99: WORDLE ADVANCED WEBSITE TO GENERATE WORD CLOUDS

8.4 RESULTS

I collected data from 13 end users who had never seen the product reaction cards. I have divided the results based on the words chosen per person and profession per words chosen. There were 23 out of 64 words that were chosen most frequently by the 13 participants. The 13 participants together covered 6 professions: Physician, Public Health Researcher, Cancer Quality Assurance Manager, Medical Writer, Administrative Manager, and Medical Microbiologist.

8.4.1 BASED ON WORDS CHOSEN

As can be seen from Figure 100, there were 23 words that were chosen most frequently. The majority of the participants (7) found the visualisation as an “Innovative” technology, followed by “Organized” and then “Efficient” and “Creative”. Words chosen by the least number of participants (1) included: “Relevant”, “Predictable”, “Essential”, “Dull”, “Confusing” and “Clean”. All the other 41 remaining words were not picked.

Table 41 shows the top 4 positive words chosen with their comments from the participants. Those who chose “Innovative” thought that the visualisation was “completely original”, “The research was something that had not been addressed before”, it was a “new approach”, and the “existing technology was used quite creatively to showcase (visually) and issue more transparently to allow stakeholders to intervene. Those who chose the word “Organised” thought that the visualisation was “well structured”, “maps were well organized” and “systematic”. Those who chose the word “Efficient” thought the visualisation was “Easy to be expanded to new settings and easy to collect and analyse on-going data”, “takes into account various factors in the process and highlights the areas in need of resources”. And finally those who chose the word “Creative” thought that the visualisation was a “new approach” and the way they were presented was “creative”.

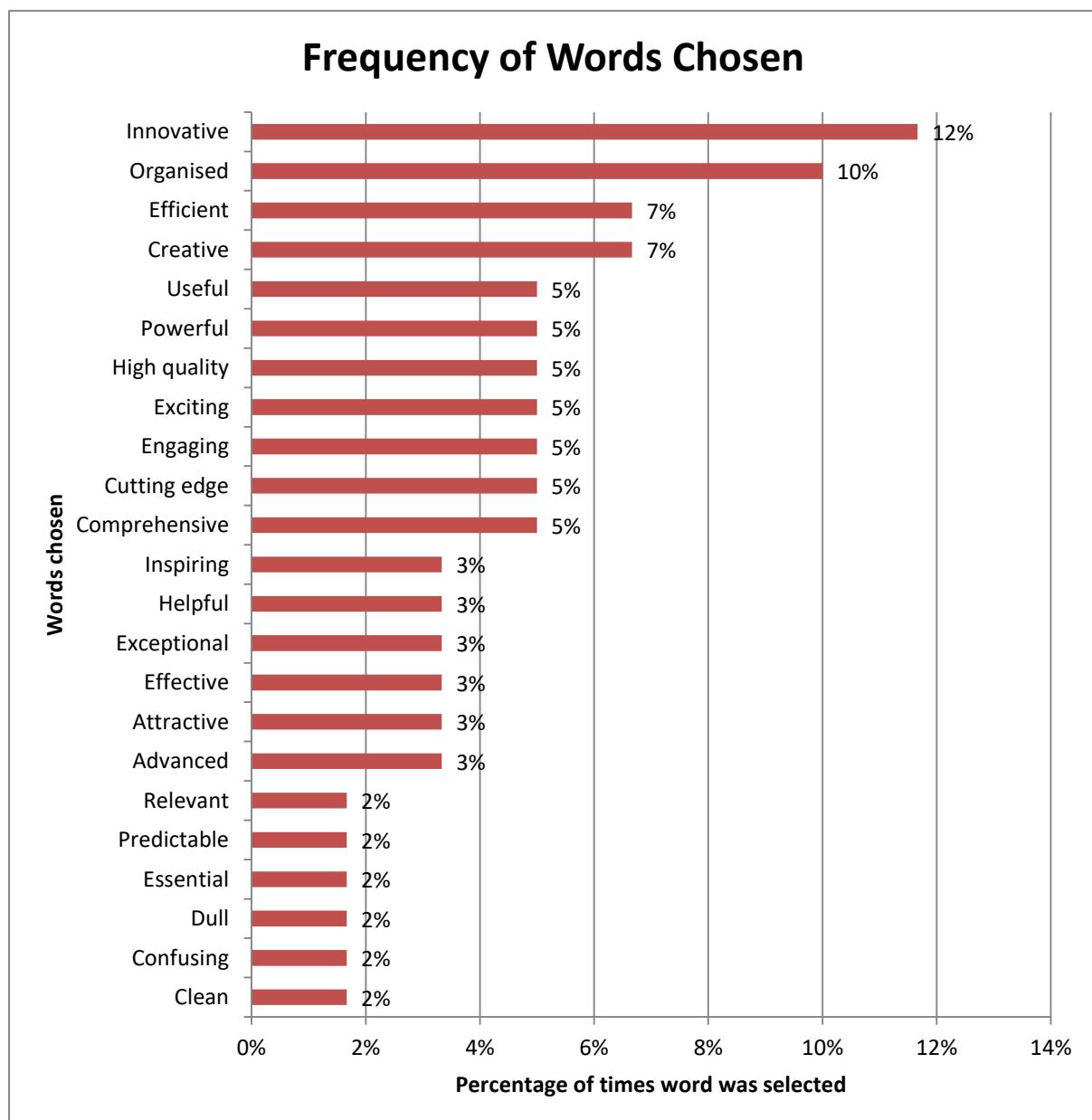


FIGURE 100: FREQUENCY OF REACTION WORDS CHOSEN FOR THE EVALUATION

TABLE 41: TOP 4 POSITIVE WORDS CHOSEN WITH THEIR COMMENTS FROM PARTICIPANTS

Top 4 Positive Words Chosen	Comments
Innovative	<ul style="list-style-type: none"> • “Brand new, completely original work” • “Nice process maps” • “Fascinating creation of coding to demonstrate your data” • “The research is something that has not been addressed before, There is a need for this technology to be used to help clinics be more productive” • “Used an existing technology quite creatively to showcase (visually) an issue more transparently to allow stakeholders to intervene” • “Methodology can be generalised for a multitude of clinical disease pathways” • “A new approach”
Organised	<ul style="list-style-type: none"> • “Well structured” • “Good structure” • “Excellent, well organised” • “Maps are well organised” • “Systematic”
Efficient	<ul style="list-style-type: none"> • “Efficient but needs more business modelling” • “Easy to be expanded to new settings and easy to collect and analyse on-going data” • “Takes into account various factors in the process and highlights the areas in need of resources”
Creative	<ul style="list-style-type: none"> • “The way the illustrations were presented was creative” • “New approach”

Figure 101 gives a picture of the word cloud that was created using Wordle. The word cloud clearly shows in larger text the words that were most frequently picked to express the participant's feelings towards the visualisation. It can be seen from the word cloud in Figure 101 that “Innovative” and “Organised” were the two most frequently picked words. Also according to the size of the words, the words: “Dull”, “Confusing”, “Relevant”, “Essential”, “Predictable” and “Clean” were the least frequently picked.



FIGURE 101: WORD CLOUD MADE OF THE EVALUATION RESULTS

8.4.2 BASED ON PROFESSIONS

I have also divided my results based on the professions of the participants. The following diagrams (Figure 102- Figure 104) show the top 3 professions and their choices of words to describe their feelings towards the visualisations. For the Public Health Researchers, the majority thought the visualisation was “Innovative”, “Efficient”, “Powerful”, “Useful” and “Comprehensive”. The majority of the physicians thought that the visualisation was “Attractive”, “Engaging”, and “Exciting”. Finally the majority of the Cancer Quality Awareness Managers thought the visualisation was “Organised”.

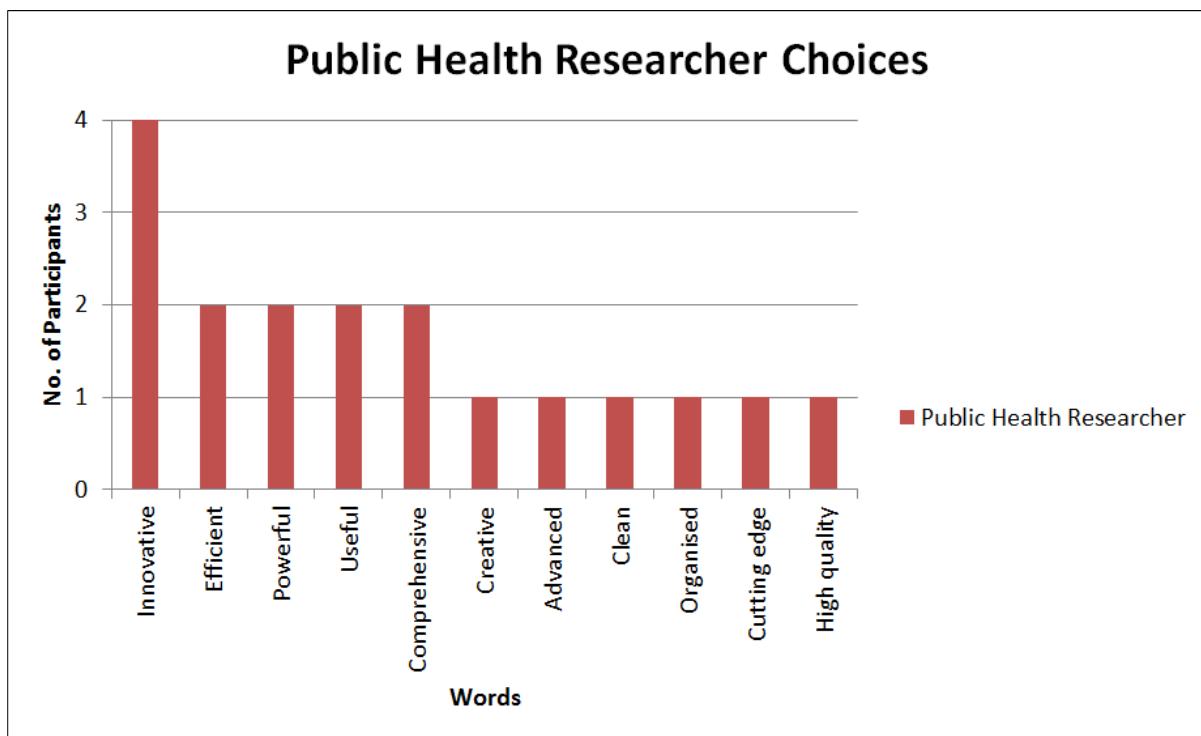


FIGURE 102: PUBLIC HEALTH RESEARCHER CHOICES OF WORDS

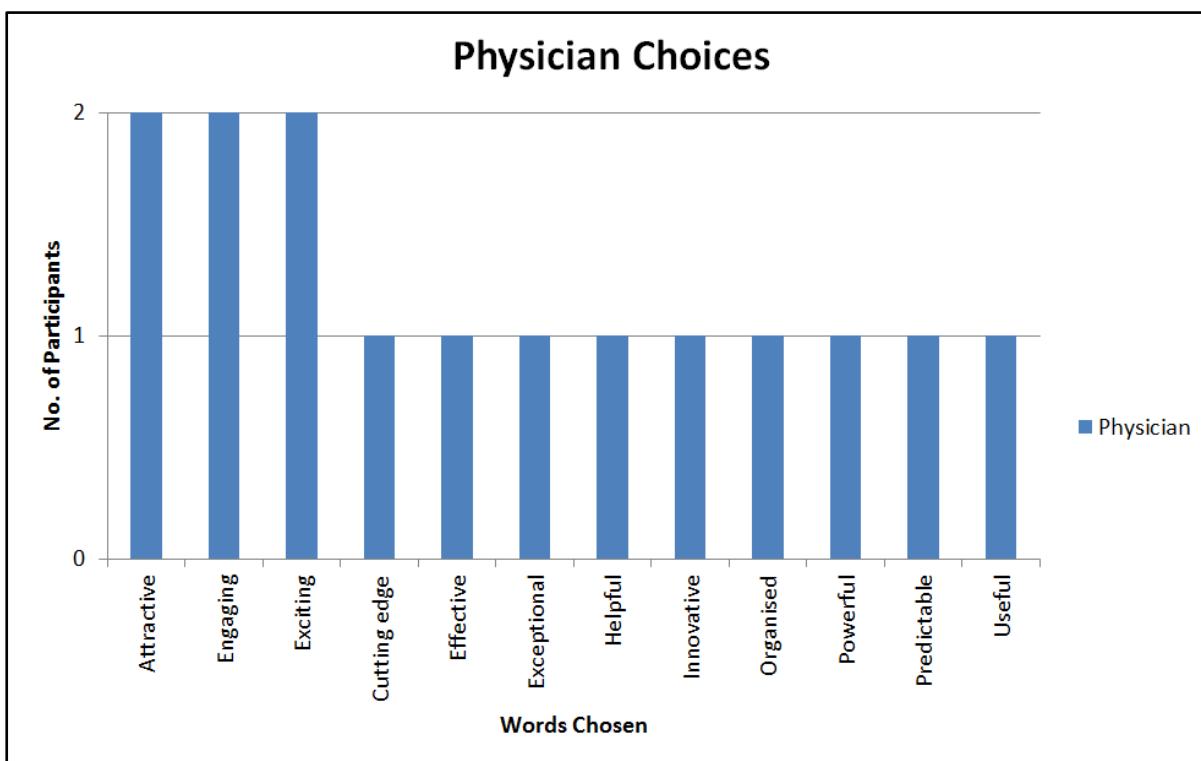


FIGURE 103: PHYSICIAN CHOICES OF WORDS

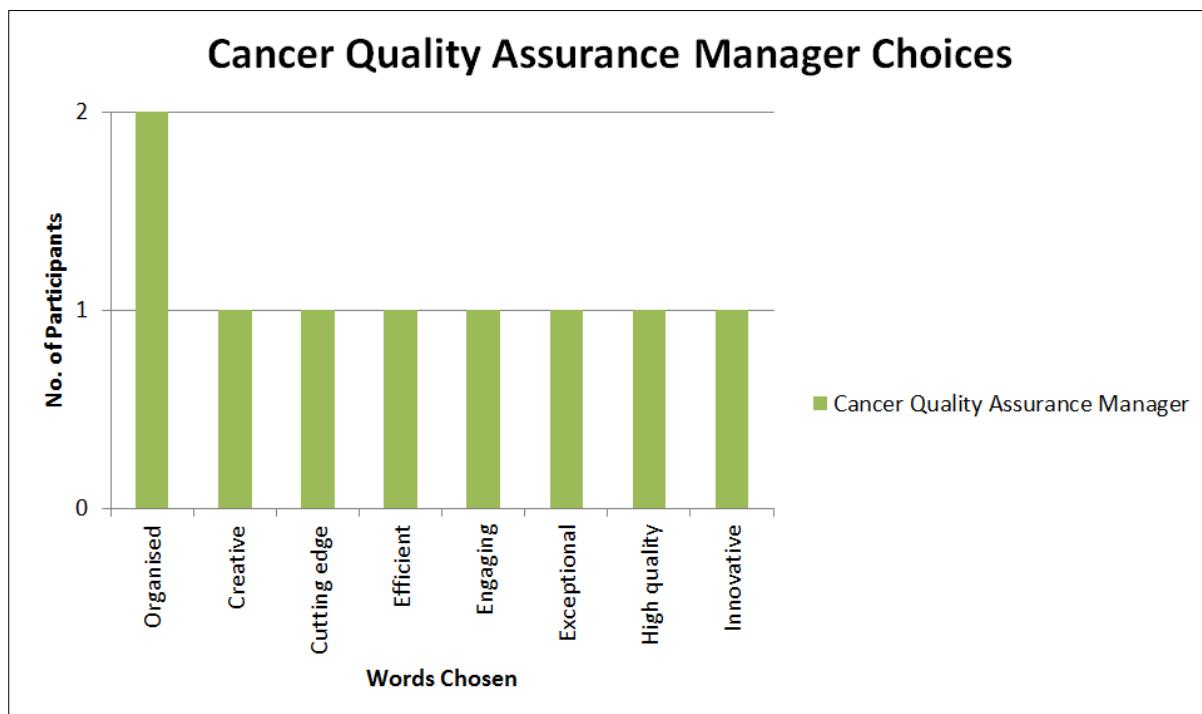


FIGURE 104: CANCER QUALITY ASSURANCE MANAGER CHOICES OF WORDS

Figure 105 shows us the combined choices of the top 3 professions. It can be seen that all the 3 professions: Public health researcher, Physician and Cancer quality assurance managers chose the words “Innovative” and “Cutting Edge” to describe the visualisation. The physicians and public health researchers both thought that the visualisation was “Useful” and “Powerful”. Whereas the Public health researchers and cancer quality assurance managers thought the visualisation was “High Quality”, “Efficient” and “Creative”. Finally, the physicians and cancer quality assurance managers both thought the visualisation was “Organised”, “Exceptional”, and “Engaging”. None of the top 3 professions gave any negative comments.

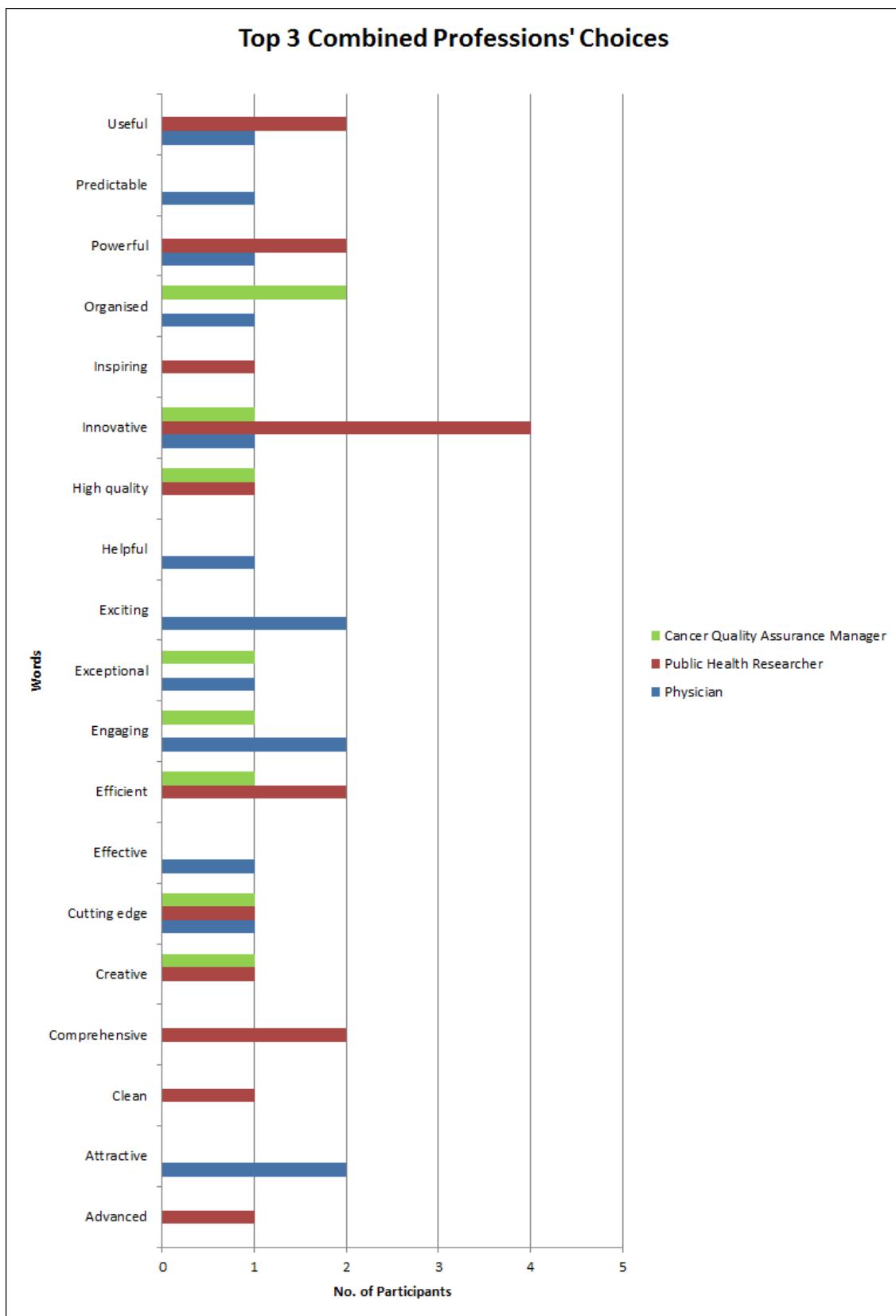


FIGURE 105: TOP 3 PROFESSIONS COMBINED WORDS CHOSEN

8.4.3 BASED ON DIMENSIONS

The following results in Table 42 show a breakdown of words based on the 5 dimensions of various aspects of the visualisation. It can be seen that for the perceived ease of use, the participants thought it was useful, clean yet confusing and dull. For the perceived usefulness of the visualisation, the participants thought it was relevant, helpful, predictable and essential. For the perceived efficiency of the visualisation, the participants thought it was effective, efficient, powerful, high quality, advanced and organised. For the overall appeal of the visualisation, the participants thought it was attractive, creative, cutting edge, exceptional, innovative and inspiring. Finally, for the level of engagement, the participants thought it was engaging, exciting and comprehensive.

TABLE 42: WORDS CHOSEN BASED ON THE 5 DIMENSIONS

Dimension	Words chosen
Perceived ease of use	useful, clean, confusing, dull
Perceived usefulness	relevant, helpful, predictable, essential
Perceived efficiency	effective, powerful, high quality, advanced, efficient, organised
Appeal	Attractive, creative, cutting edge, exceptional, innovative, inspiring
Engagement	Engaging, exciting, comprehensive

8.5 DISCUSSION

In the field of information visualization, positive user experience is extremely important if we wish to see users adopt and engage with the novel information visualization tools. It was necessary to portray to the users the importance of using this technology to visualise the patient's journey in the prostate cancer pathway. The main advantage of using Microsoft's Reaction Cards evaluation method was that it did not rely on a questionnaire or rating scales and users did not have to generate words themselves. From the usability perspective, it was quick to administer and useful data was obtained in a short period of time, including details about what users liked and disliked.

8.5.1 LESSONS LEARNT FROM EVALUATION

It was interesting to see how my segregation of two groups at the start of my evaluation journey gave rise to two categories of people with very distinctive thinking based on their backgrounds. The MDT and cancer quality group had critical feedback regarding the integrity and completeness of the data being portrayed. Even though I made it very clear to them that the evaluation is not on the quality of data, but rather the visualisation used to project data, nevertheless, their background and skills in data quality was overshadowing their ability to give the visualisation an important look. They did, however, praise the techniques used and gave positive feedback.

On the other hand, the clinical researchers and physicians group had the greatest uptake and most profound feedback amongst all. Both these professions thought the visualisations were innovative in their technology, were efficient, organised, powerful and cutting edge. All these strong positive words allude to the fact that the visualisation was well accepted and they would like to see it being used to guide stakeholders in their healthcare decision making. There was no doubt that the technology was thought as high quality and engaging.

The only negative words chosen by a person from the Management side was “Dull” and “Confusing” and the reasons behind “Dull” were not explicitly stated whereas it was “Confusing” for them because of “too many lines and circles”. Another comment that verbally flew by in the meeting room was that the process map resembled the London underground subway map. This remark does show that indeed the circles, lines and connections can get a bit confusing and overwhelming for the participants to follow if not clearly explained.

8.5.2 LIMITATIONS

The results from this evaluation technique cannot be generalized. By using this evaluation technique I am not making broad statements and assumptions about all potential users of the technology. The demonstrations were biased towards gathering information that I can use to judge the quality of the user experience for the participants who were in my evaluations and to suggest design changes through feedback and discussion. I am relying more on qualitative rather than quantitative data by using this technique. I am not looking for statistical significance in the results of this evaluation, only counts of adjectives (words) to indicate frequency and trends. Moreover, clearly there are other important aspects that need future work yet could not be completed due to the timescale of the

project. These include attempts to involve more evaluators as well as to look at elements that influence factors such as adherence to processes of care, general feasibility, expertise required, cost, etc. The evaluation is only looking at one aspect. In terms of feasibility, it has taken me four years working continuously to gain access to the data, linking the data, and cleaning the data. At the moment the trust information systems are not prepared in a way that would allow straightforward manipulation or scaling of these techniques. However, we are not far.

8.6 SUMMARY

This chapter demonstrates an evaluation technique using Microsoft's Reaction Cards to evaluate the visualisations and process models produced by the process mining techniques. It has shown the methods of administering this evaluation and displays the results based on key words chosen by the candidates, as well as results based on the professions of the candidates and a breakdown of words based on the 5 dimensions of various aspects of the visualisation, namely: Perceived ease of use, Perceived usefulness, Perceived efficiency, Appeal and Engagement. The next chapter, Chapter 9: Discussion and Conclusion, ends the thesis with a detailed discussion of the concepts, ideas, contributions and techniques used in this study as well as highlighting the strengths and weaknesses and recommendations for future work.

CHAPTER 9: DISCUSSION AND CONCLUSION

This chapter concludes my thesis. Section 9.1 gives a general discussion. Section 9.2 summarises my results. Section 9.3 draws the strengths and weaknesses of my approach. Section 9.4 outlines my work's potential, significance and originality. Section 9.5 concludes the thesis with recommendations and potential directions for future work.

9.1 DISCUSSION

A care pathway is a multidisciplinary structured care plan used to map the journey of a patient with a particular condition from the time the patient enters the custody of the care givers until the entire continuum of care. It aims to improve the quality and continuity of care by following evidence-based guidelines and protocols to achieve enhanced care for the patient. However, one of the biggest limitations of care pathways is its inability to capture any variations and exceptions made in the healthcare processes. Gaps exist between the clinical guidelines (that form the basis of the care pathways) and clinical practice since what is supposed to happen as per the guidelines and what actually happens in reality may not be the same. To address these gaps is a challenge for healthcare organisations and there is an urgent need for methods and tools to improve and streamline the healthcare systems and processes and present the results in a user friendly way back to the physicians to take appropriate actions. This is where process mining comes into place.

“Big data”, “data mining”, “business intelligence”, “process mining” - we are regularly bombarded with these terms more frequently now in our daily lives than ever before. We have witnessed in the last few decades an explosion in the volumes of data to which we are subjected to. What all these terms have in common is that they process large volumes of data that simply cannot be evaluated by hand anymore.

Processes are a very integral and structural part of any organisation. However, not all organisations have a deep understanding of how these processes are executed in reality; where they fall short; and how can they be streamlined. Process mining techniques can deliver a deep insight into the many perspectives of processes.

Healthcare organisations are exposed to very flexible and constantly changing environments where the actual execution of processes often deviates from the supposed behaviour. Because of a large

variation in the execution of processes in healthcare, they usually have a much less clear understanding of their processes. The variation in process execution is not a drawback but rather a trigger that allows healthcare processes to adjust to unique treatment options adapted to suit the patient and their wellbeing. It is therefore imperative that healthcare organisations not be strictly pre-defined and mandated to follow stringent guidelines, but a fair amount of flexibility should be given and fewer constraints should be placed to include a large set of possible behaviours as part of the main stream process.

9.2 SUMMARY OF RESULTS

The goal of the work presented in my thesis has been to show how process mining can be applied to the NHS (specifically Imperial College NHS Trust Hospitals) to show the heterogeneity in the prostate cancer pathway. The focus of my project is not on the quality of data extracted (of which there are several limitations) but rather on the process mining technique I am using and the ability I am giving to provide visibility and transparency to see what is happening in reality in the pathway as opposed to what is supposed to happen as per the guidelines.

The research objectives I put forth in the beginning of this thesis have been fulfilled in the following chapters and sections (Table 43):

TABLE 43: FULFILLMENT OF RESEARCH OBJECTIVES

Research Question/Aims	Objectives	Methods
<p>Question 1: Can we use local database systems to generate a picture of patient journeys using prostate cancer as a model?</p> <p>Aim: Analyse routinely collected data from current prostate cancer pathways</p>	<ol style="list-style-type: none">1. Literature review on the development of care pathways2. Determine the clinical information systems being used in the urology clinic setting at St. Mary's Hospital in London3. Determine what information is being captured at each step of the patient's pathway	<ul style="list-style-type: none">• Chapter 2 / Section 2.3• Studies selected covered the development, usage, importance and types of care pathways• Chapter 4 / Section 4.3• Information systems included: Outpatient/Inpatient, Pathology, Radiology, Surgery, Chemotherapy, MDT, Radiotherapy• Chapter 4 / Section 4.3• The information captured in each system was extracted via data dictionaries and entity-relationship diagrams of databases

	<p>4. Link various systems required to perform descriptive analysis</p> <ul style="list-style-type: none"> • Chapter 4 / Section 4.4 • Deterministic Record Linkage technique using the same NHS number was used to link the databases • Queries and scripts in TSQL were written to amalgamate the records in a separate staging database by means of ASP-classic routines • A descriptive analysis on the integrated linked database was done in Chapter 6 to show patient demographics, referral and treatment information
<p>Question 2: Is it possible to audit & highlight deficiencies in the standard pathway of Prostate Cancer patients?</p> <p>Aim: Perform a case note audit of extracted data against data found in the medical records of patients</p>	<p>5. Literature review on cancer quality indicators required minimum data set for reporting cancer, and audit requirements for prostate cancer</p> <ul style="list-style-type: none"> • Chapter 2 / Section 2.4 • Studies selected covered protocols for cancer referral, diagnosis, treatment, and variables for auditing prostate cancer patient pathways nationally and internationally <p>6. Create an audit data collection tool</p> <ul style="list-style-type: none"> • Chapter 5 / Section 5.2.3 • A data collection tool was created using HTML and ASP-classic for collecting data for auditing the extracted data against the actual case note data <p>7. Audit the hospital pathway against a known standard to find gaps in care</p> <ul style="list-style-type: none"> • Chapter 5 • Ethical approval was not required as a Urology clinical fellow was used for the collection process • The audit covered a retrospective date range of 10 months and collected approx. 200 patients • Outcomes were compared against case notes and showed a substantial match
<p>Question 3: Can you get new insights into routinely collected data of Prostate Cancer patients using visual analytics & present this data in a user-friendly way back to the physicians?</p> <p>Aim: Use a technique to identify bottlenecks and deviations in the current pathway and display this</p>	<p>8. Systematic review on how process mining has been used in healthcare</p> <ul style="list-style-type: none"> • Chapter 3 • 44 Studies were selected from 1392 and showed how the discovery aspect of process mining was applied to various healthcare processes • Review showed which mining methodologies and pre-processing techniques were widely used • Suggestions were made on which perspective needs more focus and what approach to process mining is better <p>9. Use the analytical and visualisation technique in a healthcare setting</p> <ul style="list-style-type: none"> • Chapter 7 • The CPAM roadmap was used to prepare and analyse the data using process mining

information back to the clinicians	10. Evaluate the visualisation technique	<ul style="list-style-type: none"> • The pathway was analysed through process discovery and various process mining perspectives • Deficiencies, bottlenecks and gaps in the pathway were highlighted and comparisons made to the LCA guideline • Chapter 8 • Evaluation of the visualisations were done using Microsoft's Reaction Cards technique • The evaluators included IT staff, clinical staff and administrative staff • The feedback on the visualisations was well received
------------------------------------	--	---

In the following subsections, I will start off by summarising my findings related to the systematic review, followed by how I constructed my process model and event log; my data validation results; my main process mining and discovery results and my evaluation results.

9.2.1 SYSTEMATIC REVIEW

In the systematic review to show the ways in which the discovery aspect of process mining been applied to healthcare to analyse and visualize care processes, it is evident that there is significant promise for process mining of complex and flexible clinical processes particularly in the cancer and emergency department care processes. In these complex environments, the techniques that are most beneficial are the ones that can deal with large amounts of noise. Clearly, in the literature, the heuristics miner followed by the fuzzy miner was most popularly used for discovering a process model. As complex processes produce spaghetti-like process flow diagrams; by including a preprocessing step, like the popular clustering and filtering techniques, helps in reducing the amount of activities to an amount that is more manageable and relevant. Additionally, to best benefit from the true potential of process mining, it is better to take on a multi-perspective approach to have a more holistic idea about the processes in a clinical system.

9.2.2 PROCESS MODEL CONSTRUCTION

I constructed my process model by following various data linkage, extraction and preparation steps to achieve the required event log necessary for the process mining stage. My case study on prostate cancer involved a wide range of data from several departments at the Imperial College Healthcare NHS Trust. After receiving an honorary contract with the NHS, the databases I had access rights to were all application databases that did not facilitate querying and analysis usually found in warehouses. I identified the databases relevant to prostate cancer that covered the following types of information systems: Inpatient and outpatient appointments, MDT meetings, Pathology, Biopsy, radiology, chemotherapy, radiotherapy, and surgery.

For the linkage process, I opted to use the Deterministic Record Linkage strategy as the same NHS number is shared among each database making it easier to make an integrated database. I decided to use all patients having a PSA done in the years between 2010 and 2014 and considered the information system with the main inpatient and outpatient appointments (Cerner DB) as the central system on which the rest of the data from other information systems would be subsequently joined to.

Following the data linkage, I extracted and transformed the required data from the integrated database. The data extraction and transformation was done in Microsoft SQL Server 2012 using TSQL queries and codes written in ASP classic for creating the event log.

Finally, in the log preparation step, I prepared the available data by renaming and aggregating events and arranging them together on a patient-by-patient basis that could easily be used for the actual process mining stage.

9.2.3 DATA VALIDATION

Before proceeding to process discovery, it was essential that I validate and verify that the data I am extracting are fit for their intended use. This was done using a clinical database audit. For the audit, I requested a clinical fellow to review the case note forms of prostate cancer referral patients for 10 months from 03/01/2013 to 03/11/2013. The information extracted from the case note forms was populated into a web-based electronic Prostate Cancer Audit Tool that I created for this purpose.

Once the clinical fellow finished collecting the case note form patient data and populating an SQL database from the back-end, I joined it with my own cohort tables and started the comparison.

I got a 90% match between the records found using my data linkage and extraction method against the Audit method. I got a 76.2% match between the first dates of appointment in both the groups. I got a 98.4% match between the PSA values in both the groups. And finally I got a higher number of 2 week waits than the Audit (97.6% for my cohort vs. 81.7% for Audit).

For the 24 patients that did not match with my cohort, overall, only 2 of the 27 patients had cancer but for both those patients my pathology database did not have a PSA registered and hence I could not pick them up with my algorithm. Those 2 patients were privately referred.

9.2.4 PROCESS MINING

After the descriptive results in which I showed frequency charts of the demographic, referral, diagnostic and treatment phases of the pathway; a first initial look at the process model of the complete 5-year event log resulted in a heavily cluttered, unreadable process flow diagram. To reduce the spaghetti diagram, I performed filtering of the log based on years and included patients that started off with a GP appointment (TWW/Urgent or Routine) within the timeframe and ended up having either a prostatectomy or radiotherapy to ensure that we are capturing only the patients who had prostate cancer. With this filtering, I used a reverse engineering technique to go back and trace the exact path these patients took to reach their intervention.

As MDT meeting is extremely vital in deciding whether the patient has cancer or not and the planned intervention, the lack of MDT data for the years 2010-2012 prompted me to only collect the results for the 2-year cohort from 2013-2015, despite the fact that I had very little patients that matched the inclusion criteria. I further segregated my results based on the priority of referral and I divided my results into the Control flow mining and Performance mining perspectives.

If we see the control-flow mining results of the TWW segmentation, the pathway starts with four distinct points after the first GP TWW appointment: with biopsy, PSA, MRI imaging and MDT. The majority of the patients (36%) started off directly with an MDT appointment after their first GP TWW appointment. Practically speaking, this may or may not have been possible depending on many

significant issues like my deficiency in acquiring biopsy data. Nevertheless, the process model showed the heterogeneity in the pathway as compared to the homogenous LCA guidelines.

Moving on to the performance mining results, I have compared all the referral priorities against the cancer performance metrics given by the LCA guidelines. Within the TWW referral, I found 10 distinct patient clusters from 11 patients. This showed that almost all the patients in the entire TWW segmentation are going through their own unique path to reach the intervention. The shortest cluster in the TWW referral had a total duration of 44 days but the longest cluster had a total duration of 229 days indicating a severe blockage in its path.

The routine referral had 4 clusters out of 4 patients with the shortest duration to reach intervention being 142 days. This shows that patients going through a routine referral, who end up having cancer, take a very long time until their intervention is started.

Patients going through a TWW referral had to wait an average of 38 days before they get their MRI done and 142 days before they get their biopsy done. In routine referrals, this wait time is even more extended as patients would wait an average of 92 days to get their MRI done and 121 days to get their biopsy done.

When I did the bottleneck analysis of the TWW referral, I found severe bottlenecks in 26 out of 37 activity handovers. That means 70% of the pathway was blocked, hence the severe delays. The maximum bottlenecks were occurring in handover from MDT to surgery suggesting either a delay in decision making or then delays in follow-up appointments and patient thinking time leading to the intervention. Similarly, in the urgent referral 100 % of the pathway was blocked and in the routine referral 81% of the pathway was blocked (13 out 20 activities).

9.2.5 EVALUATION

To evaluate the process mining techniques and methods I am using to analyse care pathways, I conducted an evaluation using Microsoft Reaction Cards on 13 individuals from different professions (MDT coordinators/ Cancer Quality Assurance administrators, clinical researchers and physicians). I invited two groups of people on two separate days to attend a presentation in which I demonstrated how process mining techniques can be used to visualise pathways of care. Using the version of Microsoft's Reaction Cards that held the concise number of 64 adjectives, I asked the participants at

the end of my presentation to choose 5 cards that best described the process mining visualization they saw. Furthermore, I requested them to leave a brief description next to each word they chose so I could capture their exact feelings and the user's experience with the new technology.

Once I got the results, I grouped the words into 5 different dimensions which aided me to analyse different aspects of the visualisations.

It was interesting to see how my segregation of two groups at the start of my evaluation journey gave rise to two categories of people with very distinctive thinking based on their backgrounds. On the one hand, the MDT and cancer quality group had critical feedback regarding the integrity and completeness of the data being portrayed; on the other hand, the clinical researchers and physicians group had the greatest uptake and most profound feedback amongst all. Both these professions thought the visualisations were innovative in their technology, were efficient, organised, powerful and cutting edge. All these strong positive words allude to the fact that the visualisation was well accepted and they would like to see it being used to guide stakeholders in their healthcare decision making. There was no doubt that the technology was thought as high quality and engaging.

The only negative words chosen by a person from the Management side was “Dull” and “Confusing” and the reasons behind “Dull” were not explicitly stated whereas it was “Confusing” for them because of “too many lines and circles”.

9.3 STRENGTHS AND WEAKNESSES

This section highlights the most significant strengths and weaknesses of using process mining techniques to visualise pathways in a healthcare setting.

The benefits of process mining vary depending on whether it is used for increasing process efficiency, reducing errors, finding delays and bottlenecks, etc. The most important strength of process mining in healthcare is its transparency to visualise the processes exactly as they are being executed. Many organisations have been struggling for months or years to understand the exact cause of their inefficiencies. By using process mining, provided data are available in a format that can readily be used, it quickly gives these organisations a deep understanding of where they are going wrong in their processes, how much waste is produced in each activity, how to improve the processes and prioritise scarce resources. It gives an accurate, quantitative picture of what the

organization has been doing. The value of process mining is not in the fact that it can produce process flow diagrams automatically, but the real value is in the insight that is gained after looking at that detailed process flow diagram. The process flow diagram may not be perfect; it may lack intuitiveness and may be difficult to comprehend if the processes are complicated like in healthcare. Nevertheless, sometimes all you need is a quick snapshot of what is actually happening in the back end to have an eye opening moment about the process flow in your organisation [205].

Process mining is not just a theoretical approach that is trying to find its niche in the industry. It is a rigorous, fast and applicable approach that transforms businesses and goes far beyond traditional approaches available so far. It is an approach that is shaping the future.

While process mining in healthcare is suitable for all practical applications, it is also important to point out a number of limitations and challenges that may hinder its usage.

If the process is extremely unstructured (as in my case) where no two traces followed the same path, it becomes increasingly difficult to keep tuning the measurement options and simplifying the process through filtration, only to result in flow diagrams that are still cluttered.

Moreover, using the plugins provided by PROM to visualise the pathway resulted in process flow diagrams that were inherently difficult to visually see (due to the extreme heterogeneity in the pathway); and thus a majority of the time was spent in re-drawing and re-editing the output of PROM (in photo editing software) to make it more readable and printable on A4 size paper.

Another major weakness in the visualisations was its intuitiveness. The automatic generation of the visual process flow diagrams in PROM resulted in pathways drawn that did not necessarily provide an immediate positive insight into the pathway taken by the patients. It was after a lot of streamlining/manually moving the connection lines in other software that made the display easy to understand and follow. The flow diagrams did not spread out equally on the page to provide a clear view of the entire journey.

A lot of the complexity of healthcare processes comes from the heterogeneity of the patients that are treated. This is due to the fact that all patients are unique and complex in their own medical conditions and even though they are getting treated for the same illness, they will take their unique pathway thus making it quite challenging to perform process mining due to a lack of similarity [206].

It is important that the event log, prior to the initiation of process mining, is prepared in a very careful and robust way otherwise the process mining results would not be representative of the actual flow of patients. Creating the event log, if no prior event log information is available in the information systems, is a very time consuming and tedious job. The procedure became even more difficult and time-consuming for me as the data found in the NHS was not in an event log format and required extensive cleaning, linkage and transformation to make it useable for process mining. Therefore, the availability and quality of data is the key driving force in carrying out sound process mining techniques.

A limitation in my work is that the study was carried out in one NHS Trust. The findings may not be generalizable to other trusts, particularly if they use different IT systems.

9.4 ORIGINALITY OF WORK

The work presented in this thesis uses the concept of process mining to visualise the paths taken by prostate cancer patients in the NHS. Process mining techniques have not been used to analyse processes in the NHS before; similarly the techniques have also not been used to analyse prostate cancer pathways before. Therefore, my PhD presents a novel approach for the NHS and the prostate cancer domain. Moreover, I have written bespoke scripts and routines to link, extract and transform raw data from the NHS data warehouse. These scripts serve as a template to make the process scalable and repeatable for future projects in different healthcare processes analysed.

9.5 RECOMMENDATIONS AND FUTURE WORK

The work presented in this thesis is a first step towards showing a transparent view of the processes that shape up the prostate cancer pathway. Future work in this domain will have to concentrate on getting better quality, complete and validated data from the NHS to properly visualise and analyse the exact impact the processes are having on quality of care.

Additionally, more emphasis should be placed on the organisational aspect of process mining, as fewer studies were found contributing in this area. Since processes emerge because of human

decision making, it would also be interesting to see more research on the working behaviour of physicians and the precise input each physician contributes during the treatment process.

Process mining holds a great potential for the flexible environment of the NHS, especially because it has not been used in this realm so far. There is a great need for sophisticated system measures and analysis techniques that process mining can easily provide in a very stunning way.

The current use of BPM in healthcare systems can be simplified to improve organisational processes. This can be done in a couple of ways [207]:

- Create an integration interface between systems: Avoid manual exchange of information between two systems that need to “speak” to each other. This creates errors and typos and ambiguity. Instead create a proper integration interface that will allow seamless information exchange between two different processes
- Automate the processes: Automating a business process means making it executable. This means that people that take part on it will have tools that will guarantee its correct execution, productivity, and transparency [208]
- Standardize recurring and identical processes: This method aims to unify the procedures in organizations that use different practices to do the same process. It focuses on searching for standardization constantly when designing a new process. Business process standardisation includes: Setting the standard, reporting the standard, establishing adherence to the standard and encouraging the continuous improvement of the standard [209]
- Define business rules: The primary purpose of defining business rules is to facilitate decision making. They should be simple and ought to go through the flow of the process without delay or uncertainty

Further development of the proposed approach includes integration with predictive models to simulate the behavioral and clinical evolution of cases as well as integration with quality of life metrics to enable detailed analysis of patients’ experience.

In addition to this, I would also like to incorporate operational support features on pre-mortem data (data on-going and online with cases still completing). By doing that, we will be able to detect violations, predict completion times of a running activity and recommend the next activity based on historical information.

A novel way of automating processes is by using Robotic Process Automation (RPA). RPA is an approach for automating manual tasks in business. The focus of RPA is to carry out these tasks automatically on the existing software front-end. With traditional automated workflow systems, the focus is on automating structured data by data integration and scripting techniques. With RPA, the focus is on unstructured data and how an end-user can be replaced with software robots. RPA streamlines repetitive high-volume processes and automates manual work [210]. Process mining can be fused with RPA to provide a general view of which processes need to be automated, analyse the effect of using RPA, and monitor the use of RPA.

The following papers are in preparation for publication:

- Process discovery from prostate cancer event logs
- Linking disparate data sources to discover process models in prostate cancer pathways of care
- Multi-perspective Discovery of Processes in the Healthcare Domain – a Systematic Review

I hope my work marks only the beginning of more extensive research into the usage and application of process mining in the NHS.

REFERENCES

1. UK, C.R. *Cancer Stats: Cancer Statistics for the UK*. [cited 2013 3rd April]; Available from: <http://www.cancerresearchuk.org/cancer-info/cancerstats/keyfacts/>
2. Fund, W.C.R. *World Cancer Statistics: Overall 2013* [cited 2013 11th April]; Available from: http://www.wcrf-uk.org/research/cancer_statistics/world_cancer_statistics_overall.php
3. Health, N.I.O., *Cancer Fact Sheet*. 2010.
4. *Delayed Diagnosis of Cancer*. 2010, NHS - National Patient Safety Agency.
5. Walters S, M.C., Butler J, Brierley JD, Rachet B, Coleman MP *Comparability of stage data in cancer registries in six countries: Lessons from the International Cancer Benchmarking Partnership*. Int J Cancer, 2013. **132**(3): p. 676–685.
6. Coleman MP, F.D., Bryant H, Butler J, Rachet B, Maringe C, Nur U, Tracey E, Coory M, Hatcher J, McGahan CE, Turner D, Marrett L, Gjerstorff ML, Johannessen TB, Adolfsson J, Lambe M, Lawrence G, Meechan D, Morris EJ, Middleton R, Steward J, Richards MA, *Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data*. Lancet 2012. **377**: p. 127-138.
7. Berrino F, D.A.R., Sant M, Rosso S, Bielska-Lasota M, Coebergh JW, Santaquilani M, *Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995-99: results of the EUROCARE-4 study*. Lancet Oncol 2007. **8**(9): p. 773–783.
8. Allgar VL, N.R., *Delays in the diagnosis of six cancers: analysis of data from the National Survey of NHS Patients: Cancer*. Brit J Cancer, 2005. **92**: p. 1959-70.
9. Shabaruddin FH, E.R., Valle JW, Newman WG, Payne K. , *Understanding chemotherapy treatment pathways of advanced colorectal cancer patients to inform an economic evaluation in the United Kingdom*. Br J Cancer, 2007. **103**(3): p. 315-23.
10. Dejardin O, R.B., Morris E, Bouvier V, Jooste V, Haynes R, Coombes EG, Forman D, Jones AP, Bouvier AM, Launoy G, *Management of colorectal cancer explains differences in 1-year relative survival between France and England for patients diagnosed 1997-2004*. Br J Cancer, 2013. **108**(4): p. 775-83.
11. E, B.D.M., *Cervical screening and health inequality in England in the 1990s*. Journal of Epidemiology and Community Health 2003. **57**: p. 417–423.
12. Adams J, W.M., Forman D, *Are there socioeconomic gradients in stage and grade of breast cancer at diagnosis? Cross sectional analysis of UK cancer registry data*. BMJ, 2004. **329**(7458): p. 142.
13. Raleigh, V.S., *Getting the Measure of Quality*. 2010, The King's Fund: London.
14. Caron, F., et al., *A process mining-based investigation of adverse events in care processes*. Health Information Management Journal, 2014. **43**(1): p. 16-25.
15. M. Rovani, F.M.M., M. de Leoni, W.M.P. van der Aalst, R.S. Mans, A. Pepino, *Declarative process mining in healthcare*, in *BPM reports* ; 1411. 2014. p. 29p.
16. Lang, M., et al., *Process mining for clinical workflows: challenges and current limitations*. Stud Health Technol Inform, 2008. **136**: p. 229-34.
17. Mans, R., W. Aalst, and R. Vanwersch, *Process mining in healthcare opportunities beyond the ordinary*. 2013: narcis.nl.
18. Poelmans, J., et al., *Combining Business Process and Data Discovery Techniques for Analyzing and Improving Integrated Care Pathways*, in *Advances in Data Mining. Applications and Theoretical Aspects*, P. Perner, Editor. 2010, Springer Berlin Heidelberg. p. 505-517.
19. Mans, R., et al., *Process Mining in Healthcare*. Case study. 2015: wwwis.win.tue.nl.
20. Rebufé, A. and D.R. Ferreira, *Business process analysis in healthcare environments: A methodology based on process mining*. Information Systems, 2012. **37**(2): p. 99-116.

21. Matthijssen, P., *Business Process Management*, in *Business Process Management vs. Business Analysis?* 2013.
22. Aalst, W., *Business process management: a personal view*. Business Process Management Journal, Bradford, 2004. **10**(2): p. 135-139.
23. Lusk, S., S. Paley, and A. Spanyi, *The evolution of business process management as a professional discipline*. Evolution of BPM as a Professional Discipline, 2005.
24. Rozinat, A., *How is Process Mining Different From ...*, in *Flux Capacitor*.
25. Buttigieg, S.C., *Business process management in health care: current challenges and future prospects*. Information technology, 2016. **2**: p. 5.
26. Cook, J.E. and A.L. Wolf, *Discovering models of software processes from event-based data*. ACM Transactions on Software Engineering and Methodology (TOSEM), 1998. **7**(3): p. 215-249.
27. Weijters, A.T., *Process mining for ubiquitous mobile systems: an overview and a concrete algorithm*. 2004.
28. Bautista, A.D., L. Wangikar, and S.M.K. Akbar, *Process Mining Driven Optimization of a Consumer Loan Approvals Process*. BPI Challenge, 2012.
29. Rozinat, A. and W.M. van der Aalst. *Conformance testing: Measuring the fit and appropriateness of event logs and process models*. in *International Conference on Business Process Management*. 2005. Springer.
30. Maruster, L., et al., *Discovering distributed processes in supply chains*, in *Collaborative Systems for Production Management*. 2003, Springer. p. 219-230.
31. Rozinat, D.A., *Disco User's Guide*. 2013, fluxicon.
32. Der Aalst, V. and W.P. Mining, *Discovery, Conformance and Enhancement of Business Processes*. 2011, Springer: Berlin, Germany Heidelberg, Germany.
33. Rozinat, D.A. *How Process Mining Compares to BI*. 2014 [cited 2014 29th December]; Available from: <http://fluxicon.com/blog/2011/01/how-pm-compares-to-bi/>.
34. Perimal-Lewis, L., et al., *Gaining insight from patient journey data using a process-oriented analysis approach*, in *Proceedings of the Fifth Australasian Workshop on Health Informatics and Knowledge Management - Volume 129*. 2012, Australian Computer Society, Inc.: Melbourne, Australia. p. 59-66.
35. Van Der Aalst, W.M. and B.F. Van Dongen, *Discovering petri nets from event logs*, in *Transactions on Petri Nets and Other Models of Concurrency VII*. 2013, Springer. p. 372-422.
36. Song, M., et al., *A comparative study of dimensionality reduction techniques to enhance trace clustering performances*. Expert Systems with Applications, 2013. **40**(9): p. 3722-3737.
37. van der Aalst, W.P., *Distributed Process Discovery and Conformance Checking*, in *Fundamental Approaches to Software Engineering*, J. de Lara and A. Zisman, Editors. 2012, Springer Berlin Heidelberg. p. 1-25.
38. Van der Aalst, W.M., et al., *Workflow mining: a survey of issues and approaches*. Data & knowledge engineering, 2003. **47**(2): p. 237-267.
39. Chapman, N. *Petri Net Models - ISE-2 Surprise 97 Project*. 1997.
40. Gupta, S., *Workflow and process mining in healthcare*. Master's Thesis, Technische Universiteit Eindhoven, 2007.
41. Kirchner, K., et al., *Embedding Conformance Checking in a Process Intelligence System in Hospital Environments*, in *Process Support and Knowledge Representation in Health Care*, R. Lenz, et al., Editors. 2013, Springer Berlin Heidelberg. p. 126-139.

42. Wombacher, A., M. Iacob, and M. Haitsma. *Towards a performance estimate in semi-structured processes*. in *Service-Oriented Computing and Applications (SOCA), 2011 IEEE International Conference on*. 2011. IEEE.
43. Ho, E.T.L., *Improving waiting time and operational clinic flow in a tertiary diabetes center*. BMJ quality improvement reports, 2014. **2**(2): p. u201918. w1006.
44. Statistics, O.f.N. *Expenditure on Healthcare in the UK: 2012 2015* [cited 2016 27 June]; Available from: <http://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/articles/expenditureonhealthcareintheuk/2015-03-26>.
45. Parkin, E., *NHS maximum waiting times standards and patient choice policies*. 2016.
46. Fund, T.K.s. *Quarterly Monitoring Report*. 2014 [cited 2016 27 June]; Available from: <http://qmr.kingsfund.org.uk/2015/15/data>.
47. Aalst, W.V.D. *Event logs - What kind of data does process mining require?* 2016; Available from: <http://www.processmining.org/logs/start>.
48. Aita, V., et al., *Patient-centered care and communication in primary care practice: what is involved?* Patient Education and Counseling, 2005. **58**(3): p. 296-304.
49. Pelzang, R., *Time to learn: understanding patient-centred care*. British journal of nursing, 2010. **19**(14): p. 912-917.
50. Goodwin, N.S., Judith; Davies, Alisha; Perry, Claire; Rosen, Rebecca; Dixon, Anna; Dixon, Jennifer; Ham, Chris, *Integrated care for patients and populations: Improving outcomes by working together*. 2012.
51. Vanhaecht, K.P., Massimiliano; van Zelm, Ruben; Sermeus, Walter *An overview on the history and concept of care pathways as complex interventions*. International Journal of Care Pathways, 2010. **14**(3): p. 117-123.
52. De Luc, K., *Developing Care Pathways – the Handbook*. 2001, Oxford: Radcliffe Medical Press Ltd.
53. Kinsman, L., et al., *What is a clinical pathway? Development of a definition to inform the debate*. BMC Med, 2010. **8**: p. 31.
54. Campbell, H., et al., *Integrated care pathways*. BMJ, 1998. **316**(7125): p. 133-7.
55. De Bleser, L., et al., *Defining pathways*. J Nurs Manag, 2006. **14**(7): p. 553-63.
56. Vanhaecht K, D.W.K., Sermeus W, *The Impact of Clinical Pathways on the Organisation of Care Processes*. 2007: Leuven: ACCO.
57. Clinical, O. *Clinical Pathways*. [cited 2013 29/05/2013]; Available from: <http://www.openclinical.org/clinicalpathways.html>.
58. Panella, M. and K. Vanhaecht, *Is there still need for confusion about pathways?* International Journal of Care Pathways, 2010. **14**(1): p. 1-3.
59. Schrijvers, G., A. van Hoorn, and N. Huiskes, *The care pathway: concepts and theories: an introduction*. Int J Integr Care, 2012. **12**(Spec Ed Integrated Care Pathways): p. e192.
60. Currie, V.L.H., G. , *The use of care pathways as tools to support the implementation of evidence-based practice*. JOURNAL OF INTERPROFESSIONAL CARE, 2000. **14**(4).
61. De Luc, K., *Care pathways: an evaluation of their effectiveness*. J Adv Nurs, 2000. **32**(2): p. 485-96.
62. Health, D.o., *The New NHS Modern and Dependable: A National Framework for Assessing Performance*, D.o. Health, Editor. 1998.
63. Bower, K.A., *Clinical pathways: 12 lessons learned over 25 years of experience*. Int J Care Pathways, 2009. **13**: p. 78-81.

64. El Baz, N., et al., *Are the outcomes of clinical pathways evidence-based? A critical appraisal of clinical pathway evaluation research*. J Eval Clin Pract, 2007. **13**(6): p. 920-9.
65. Hasnain-Wynia, R., *Is evidence-based medicine patient-centered and is patient-centered care evidence-based?* Health Serv Res, 2006. **41**(1): p. 1-8.
66. Every, N.R., et al., *Critical pathways : a review. Committee on Acute Cardiac Care, Council on Clinical Cardiology, American Heart Association*. Circulation, 2000. **101**(4): p. 461-5.
67. Van Herck, P., K. Vanhaecht, and W. Sermeus, *Effects of Clinical Pathways: Do They Work?* Journal of Integrated Pathways, 2004. **8**(3): p. 95-105.
68. Rotter, T., et al., *The quality of the evidence base for clinical pathway effectiveness: room for improvement in the design of evaluation trials*. BMC Med Res Methodol, 2012. **12**: p. 80.
69. Kwan, J. and P. Sandercock, *In-Hospital Care Pathways for, Stroke: An Updated Systematic Review*. Stroke, 2005. **36**(6): p. 1348-1349.
70. Hammond, R., *Integrated Care Pathways*. 2002, THE CHARTERED SOCIETY OF PHYSIOTHERAPY: LONDON.
71. De Luc, K., *Developing Care Pathways - the Tool Kit*. 2001, Oxford: Radcliffe Medical Press Ltd.
72. Layton, A., F. Moss, and G. Morgan, *Mapping out the patient's journey: experiences of developing pathways of care*. Qual Health Care, 1998. **7 Suppl**: p. S30-6.
73. Davis, N., *The Integrated Care Pathways Guide to Good Practice*. 2005, Llanharan, Wales: National Leadership and Innovation Agency for Healthcare.
74. Hunter, B. and J. Segrott, *Re-mapping client journeys and professional identities: a review of the literature on clinical pathways*. Int J Nurs Stud, 2008. **45**(4): p. 608-25.
75. NHS. *NHS Choices: Prostate Cancer*. 2013 [cited 2013 21st October]; Available from: <http://www.nhs.uk/Conditions/Cancer-of-the-prostate/Pages/Introduction.aspx>.
76. UK, C.R. *Prostate cancer statistics*. 2013 [cited 2013 21st October]; Available from: <http://www.cancerresearchuk.org/cancer-info/cancerstats/types/prostate/?script=true>.
77. UK, C.R. *Prostate cancer survival statistics*. 2009 [cited 2013 Nov 5]; Available from: <http://www.cancerresearchuk.org/cancer-info/cancerstats/types/prostate/survival/>.
78. Oliver, S.E., M.T. May, and D. Gunnell, *International trends in prostate-cancer mortality in the "PSA era"*. International journal of cancer, 2001. **92**(6): p. 893-898.
79. Etzioni, R. and R. Gulati, *Recent trends in psa testing and prostate cancer incidence: A look at context*. JAMA Oncology, 2016. **2**(7): p. 955-956.
80. UK, P.C. *Who is at Risk?* 2013 [cited 2013 21st October]; Available from: <http://prostatecanceruk.org/information/who-is-at-risk>.
81. NICE, *Referral guidelines for suspected cancer: Clinical Guideline 27*. 2005.
82. UK, P.C. *Visiting your GP*. 2013; Available from: <http://prostatecanceruk.org/information/prostate-problems/psa-test>.
83. Burford, D.C.K., Michael; Austoker, Joan *Prostate Cancer Risk Management Programme - information for primary care - PSA testing in asymptomatic men*. 2009, Cancer Research UK.
84. *Localised prostate cancer - A guide for men and their families*. 2010, Cancer Council Australia.
85. Institute, N.C. *Prostate-Specific Antigen (PSA) Test*. 2013 [cited 2013 September 5]; Available from: <http://www.cancer.gov/cancertopics/factsheet/detection/PSA>.
86. Philip, J., et al., *Is a digital rectal examination necessary in the diagnosis and clinical staging of early prostate cancer?* BJU International, 2005. **95**(7): p. 969-971.
87. UK, P.C. *PSA Test*. 2013 [cited 2013 24 Sept]; Available from: <http://prostatecanceruk.org/information/prostate-problems/psa-test>.

88. Knowledge, H.M.S.H.H.P.P. *What is transrectal ultrasonography (TRUS)?* 2013 [cited 2013 21st October]; Available from: <http://www.harvardprostateknowledge.org/what-is-transrectal-ultrasonography-trus>.
89. Association, A.U. *Prostate Cancer.* 2013 [cited 2013 24 Sept]; Available from: <http://www.auanet.org/education/guidelines/prostate-cancer.cfm>.
90. UK, P.C. *Testing for Prostate Cancer.* 2013 [cited 2013 21st October]; Available from: <http://prostatecanceruk.org/information/prostate-cancer/testing-for-prostate-cancer>.
91. *Prostate Cancer Risk Management Programme.* 2013 [cited 2013 Nov 5]; Available from: <http://www.cancerscreening.nhs.uk/prostate/index.html>.
92. Oxford, U.o. *Cancer costs the UK economy £15.8bn a year.* 2012 [cited 2017 19 February]; Available from: <http://www.ox.ac.uk/news/2012-11-07-cancer-costs-uk-economy-£158bn-year>.
93. Roehrborn, C.G. and L.K. Black, *The economic burden of prostate cancer.* BJU international, 2011. **108**(6): p. 806-813.
94. Miller, D.C., J.E. Montie, and J.T. Wei, *Measuring the quality of care for localized prostate cancer.* J Urol, 2005. **174**(2): p. 425-31.
95. Danielson, B., et al., *Development of indicators of the quality of radiotherapy for localized prostate cancer.* Radiother Oncol, 2011. **99**(1): p. 29-36.
96. Miller, D.C. and C.S. Saigal, *Quality of care indicators for prostate cancer: progress toward consensus.* Urol Oncol, 2009. **27**(4): p. 427-34.
97. Kunkel, S., U. Rosenqvist, and R. Westerling, *The structure of quality systems is important to the process and outcome, an empirical study of 386 hospital departments in Sweden.* BMC Health Serv Res, 2007. **7**: p. 104.
98. Mayer, E.K., et al., *Appraising the quality of care in surgery.* World J Surg, 2009. **33**(8): p. 1584-93.
99. Cancer, N.C.C.f., *Prostate Cancer Diagnosis and Treatment: Clinical Guideline 58.* 2008, NICE.
100. Joe McDevitt, H.C., *A PROPOSED CORE NATIONAL CANCER DATASET.* 2010, NATIONAL CANCER REGISTRY IRELAND.
101. Ghaneie, M., et al., *Designing a Minimum Data Set for Breast Cancer: a Starting Point for Breast Cancer Registration in Iran.* Iran J Public Health, 2013. **42**(Supple1): p. 66-73.
102. *CANCER OUTCOMES AND SERVICES DATASET (COSD).* 2013, NCIN.
103. Canada, G.o.C.S. *Definitions, Data Sources and Methods.* 2013; Available from: [http://www23.statcan.gc.ca/imdb-pIX.pl?Function=showDirectPDF&fl=http://www23.statcan.gc.ca/imdb-bmdi/document/3207_D2_T9_V1-eng.pdf&flng=eng&a=1](http://www23.statcan.gc.ca/imdb/pIX.pl?Function=showDirectPDF&fl=http://www23.statcan.gc.ca/imdb-bmdi/document/3207_D2_T9_V1-eng.pdf&flng=eng&a=1).
104. Gjerstorff, M.L., *The Danish Cancer Registry.* Scand J Public Health, 2011. **39**(7 Suppl): p. 42-5.
105. Registry, M.O. *Cancer (clinical) DSS.* 2011 [cited 2013 29th October]; Available from: <http://meteor.aihw.gov.au/content/index.phtml/itemId/394731>.
106. Hauora, M.o.H.M. *New Zealand Cancer Registry - table of available data.* 2013 [cited 2013 21st October]; Available from: <http://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections/new-zealand-cancer-registry-nzcr/new-zealand-cancer-registry-table-available-data>.
107. HISO, *National Cancer Core Data Definitions Interim Standard 2012,* Ministry of Health, New Zealand.
108. SEER, *SEER PROGRAM CODING AND STAGING MANUAL 2013.* 2013, SEER.
109. Greg Rubin, S.M., Kathy Elliott, *National Audit of Cancer Diagnosis in Primary Care.* 2011, RCGP, NHS, NCIN.

110. Agency, N.P.S., *Delayed diagnosis of cancer: Thematic Review*. 2010, NHS.
111. Veronique Poirier, S.C., Julia Verne, *Breast cancer "Two week wait" referral audit*. 2003.
112. K Ruth, J.M., S Keohane, J Verne, D de Berker, V Poirier, *Two-week wait referrals for malignant melanoma: A clinical audit carried out across four UK Cancer Networks*. UK Association of Cancer Registries conference, 2006.
113. Philippa King, M.H., Veronique Poirier, Anton Kruger, *Audit of high grade B non-Hodgkin's lymphoma (NHL) in the South West Region 2001 - clinical data on 337 cases*. British Society for Haematology meeting, 2005.
114. NICE. *Referral guidelines for suspected cancer - Appendix D: Technical detail on the criteria for audit*. 2005 [cited 2013 4th October]; Available from: <http://publications.nice.org.uk/referral-guidelines-for-suspected-cancer-cg27/appendix-d-technical-detail-on-the-criteria-for-audit>.
115. Baughan, P., B. O'Neill, and E. Fletcher, *Auditing the diagnosis of cancer in primary care: the experience in Scotland*. Br J Cancer, 2009. **101 Suppl 2**: p. S87-91.
116. *Specification Development for the National Prostate Cancer Audit*. HQIP.
117. Weijters, A., W.M. van Der Aalst, and A.A. De Medeiros, *Process mining with the heuristics miner-algorithm*. Technische Universiteit Eindhoven, Tech. Rep. WP, 2006. **166**: p. 1-34.
118. de Medeiros, A.K.A., A.J. Weijters, and W.M. van der Aalst, *Genetic process mining: an experimental evaluation*. Data Mining and Knowledge Discovery, 2007. **14**(2): p. 245-304.
119. Di Ciccio12, C. and M. Mecella, *Studies on the discovery of declarative control flows from error-prone data*. Data-Driven Process Discovery and Analysis SIMPDA 2013, 2013: p. 31.
120. Rozinat, A., *Disco Tour*. 2013.
121. Rozinat, A., *Process mining: conformance and extension*. 2010, Technische Universiteit Eindhoven.
122. Kaymak, U., et al. *On process mining in health care*. in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*. 2012.
123. Zhou, J., *Process mining: acquiring objective process information for healthcare process management with the CRISP-DM framework*. Master's thesis. Eindhoven University of Technology, 2009.
124. Antonelli, D. and G. Bruno, *Application of Process Mining and Semantic Structuring Towards a Lean Healthcare Network*, in *Risks and Resilience of Collaborative Networks*, L.M. CamarinhaMatos, F. Benaben, and W. Picard, Editors. 2015. p. 497-508.
125. Mans, R.S., et al., *Application of Process Mining in Healthcare - A Case Study in a Dutch Hospital*, in *Biomedical Engineering Systems and Technologies*, A. Fred, J. Filipe, and H. Gamboa, Editors. 2008. p. 425-438.
126. Ramos, L., *Healthcare Process Analysis: Validation and Improvements of a Data-based Method using Process Mining and Visual Analytics*. 2009: Master's thesis, Eindhoven
127. Riemers, P., *Process Improvement in Healthcare: a Data-Based Method Using a Combination of Process Mining and Visual Analytics*. Master's thesis. Eindhoven University of Technology, 2009.
128. Mans, R., *Workflow support for the healthcare domain*. 2011, Technische Universiteit Eindhoven.
129. Bart, D.V., *Process Mining in Healthcare Systems: An Evaluation and Refinement of a Methodology*. 2012.
130. Bose, R. and v.d.W. Aalst, *Analysis of patient treatment procedures: The BPI Challenge case study*. 2011.

131. De Weerdt, J., et al., *Getting a grasp on clinical pathway data: An approach based on process mining*, in *Emerging Trends in Knowledge Discovery and Data Mining*. 2012, Springer. p. 22-35.
132. Caron, F., J. Vanthienen, and B. Baesens, *Healthcare Analytics: Examining the Diagnosis-treatment Cycle*. Procedia Technology. 2013: Elsevier.
133. Poelmans, J., et al., *Combining Business Process and Data Discovery Techniques for Analyzing and Improving Integrated Care Pathways*, in *Advances in Data Mining. Applications and Theoretical Aspects: 10th Industrial Conference, ICDM 2010, Berlin, Germany, July 12-14, 2010. Proceedings*, P. Perner, Editor. 2010, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 505-517.
134. Huang, Z., X. Lu, and H. Duan, *On mining clinical pathway patterns from medical behaviors*. Artificial intelligence in medicine, 2012. **56**(1): p. 35-50.
135. Mans, R., et al., *Process mining in healthcare: Data challenges when answering frequently posed questions*. Process Support and ..., 2013.
136. de Leoni, M., F.M. Maggi, and W.M. van der Aalst, *An alignment-based framework to check the conformance of declarative process models and to preprocess event-log data*. Information Systems, 2015. **47**: p. 258-277.
137. Paul Helmering, P.H., Vidya Iyer, Anil Kabra, Jeff Van Slette, *Process Mining of Clinical Workflows for Quality and Process Improvement*. 2012.
138. Lakshmanan, G.T., S. Rozsnyai, and F. Wang, *Investigating clinical care pathways correlated with outcomes*, in *Business process management*. 2013, Springer. p. 323-338.
139. Mans, R., et al., *A Process-oriented Methodology for Evaluating the Impact of IT: a Proposal and an Application in Healthcare*. Information Systems, 2013.
140. Mans, R., et al. *Mining processes in dentistry*. in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. 2012. ACM.
141. Dagliati, A., et al., *Temporal Data Mining and Process Mining Techniques to Identify Cardiovascular Risk-Associated Clinical Pathways in Type 2 Diabetes Patients*. 2014 Ieee-Embs International Conference on Biomedical and Health Informatics (Bhi), 2014: p. 240-243.
142. Rattanavayakorn, P. and ... *Analysis of the social network miner (working together) of physicians*. ICT and Knowledge ..., 2015.
143. Krutanard, C., P. Porouhan, and ... *Discovering organizational process models of resources in a hospital using Role Hierarchy Miner*. ICT and Knowledge ..., 2015.
144. Alves, C.d.C., *Social Network Analysis for Business Process Discovery*. The Technical University of Lisbon. 2010: fenix.tecnico.ulisboa.pt.
145. Ferreira, D.R. and C. Alves. *Discovering user communities in large event logs*. in *Business Process Management Workshops*. 2011. Springer.
146. Perimal-Lewis, L., et al. *Gaining insight from patient journey data using a process-oriented analysis approach*. in *Proceedings of the Fifth Australasian Workshop on Health Informatics and Knowledge Management-Volume 129*. 2012. Australian Computer Society, Inc.
147. Matthews, L., *Process mining to facilitate process improvement in a healthcare environment: An emergency department case study*. 2013: gradworks.umi.com.
148. Delias, P., et al., *Supporting healthcare management decisions via robust clustering of event logs*. Knowledge-Based Systems, 2015. **84**: p. 203-213.
149. Xiaojin, Z. and C. Songlin. *Pathway identification via process mining for patients with multiple conditions*. in *Industrial Engineering and Engineering Management (IEEM), 2012 IEEE International Conference on*. 2012.

150. Overduin, M.T., *Exploration of the link between the execution of a clinical process and its effectiveness using process mining techniques* 2013.
151. Perimal-Lewis, L., D. De Vries, and C.H. Thompson. *Health intelligence: Discovering the process model using process mining by constructing Start-to-End patient journeys*. in *Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management-Volume 153*. 2014. Australian Computer Society, Inc.
152. Boere, J.-J., *An analysis and redesign of the ICU weaning process using data analysis and process mining*. 2013, Maastricht University Medical Centre.
153. Kim, E., et al., *Discovery of outpatient care process of a tertiary university hospital using process mining*. Healthcare 2013: synapse.koreamed.org.
154. Cho, M., M. Song, and S. Yoo, *A Systematic Methodology for Outpatient Process Analysis Based on Process Mining*. Asia Pacific Business Process Management, 2014.
155. Zhichao, Z., W. Yong, and L. Lin. *Process mining based modeling and analysis of workflows in clinical care - A case study in a chicago outpatient clinic*. in *Networking, Sensing and Control (ICNSC), 2014 IEEE 11th International Conference on*. 2014.
156. Micilo, R., et al. *RTLS-based Process Mining: Towards an automatic process diagnosis in healthcare*. in *Automation Science and Engineering (CASE), 2015 IEEE International Conference on*. 2015.
157. Fei, H. and N. Meskens. *Discovering patient care process models from event logs*. in *8th International conference of modeling and simulation, MOSIM*. 2008. Citeseer.
158. Mans, R., et al., *Process mining techniques: an application to stroke care*. Stud Health Technol Inform, 2008. **136**: p. 573-8.
159. Quaglini, S., *Process mining in healthcare: a contribution to change the culture of blame*. Business Process Management Workshops, 2008.
160. Montani, S., et al., *Mining and retrieving medical processes to assess the quality of care*, in *Case-Based Reasoning Research and Development*. 2013, Springer. p. 233-240.
161. Jaisook, P. and W. Premchaiswadi. *Time performance analysis of medical treatment processes by using disco*. in *ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015), 2015 13th International Conference on*. 2015.
162. Rovani, M., et al., *Declarative process mining in healthcare*. Expert Systems with Applications, 2015. **42**(23): p. 9236-9251.
163. Ferreira, D., et al., *Approaching process mining with sequence clustering: Experiments and findings*, in *Business Process Management*. 2007, Springer. p. 360-374.
164. Kurniati, A.P., et al. *Process mining in oncology: A literature review*. in *Information Communication and Management (ICICM), International Conference on*. 2016. IEEE.
165. Salimifard, K., S.Y. Hosseini, and M.S. Moradi. *Improving Emergency Department Processes Using Coloured Petri Nets*. in *PNSE+ ModPE*. 2013. Citeseer.
166. Medeiros, A.K.A., et al., *Process Mining Based on Clustering: A Quest for Precision*, in *Business Process Management Workshops: BPM 2007 International Workshops, BPI, BPD, CBP, ProHealth, RefMod, semantics4ws, Brisbane, Australia, September 24, 2007, Revised Selected Papers*, A. Hofstede, B. Benatallah, and H.-Y. Paik, Editors. 2008, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 17-29.
167. van Oirschot, Y., et al., *Using Trace Clustering for Configurable Process Discovery Explained by Event Log Data*. 2014, Master's thesis.
168. Saravanan, M. and R. Rama Sree, *Evaluation of process models using heuristic miner and disjunctive workflow schema ALGORITHM for dyeing Process*. Int J Inform Technol Conver Serv (IJITCS), 2011. **1**(3): p. 47-68.

169. Cross, R., et al., *Knowing what we know:: Supporting knowledge creation and sharing in social networks*. Organizational dynamics, 2001. **30**(2): p. 100-120.
170. Ferreira, D.R. *ERCIM News - Performance Analysis of Healthcare Processes through Process Mining*. 2012 [cited 2017 23 February]; Available from: <http://ercim-news.ercim.eu/en89/special/performance-analysis-of-healthcare-processes-through-process-mining>.
171. Rojas, E., et al., *Process mining in healthcare: A literature review*. Journal of biomedical informatics, 2016. **61**: p. 224-236.
172. Erdogan, T. and A. Tarhan. *Process Mining for Healthcare Process Analytics*. in *Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2016 Joint Conference of the International Workshop on*. 2016. IEEE.
173. de Vries, G.-J., et al., *Towards Process Mining of EMR Data*. BIOSTEC 2017, 2017: p. 585.
174. van der Aalst, W.M., *Extracting event data from databases to unleash process mining*, in *BPM-Driving innovation in a digital world*. 2015, Springer. p. 105-128.
175. de Murillas, E.G.L., et al., *Connecting Databases with Process Mining: A Meta Model and Toolset*.
176. Rawat, D., *Responsibilities of the Business Intelligence Unit*, B. Siddiqi, Editor. 2017.
177. NHS Clinical Commissioners. [cited 2016 30 May]; Available from: <http://www.nhscc.org/ccgs/>.
178. Rawat, D., *Submission of Secondary Care Data*, B. Siddiqi, Editor. 2015.
179. Cerner - Hospitals and HEalth Systems. 2016 [cited 2016 11 June]; Available from: http://www.cerner.com/solutions/hospitals_and_health_systems/.
180. Buijs, J., *Mapping data sources to xes in a generic way*. Maters Thesis, 2010.
181. *ETL Process - Datawarehouse4U*. 2009 [cited 2016 10 June]; Available from: <http://datawarehouse4u.info/ETL-process.html>.
182. González-López de Murillas, E., W. van der Aalst, and H. Reijers, *Process mining on databases: Unearthing historical data from redo logs*. 2015.
183. Clark, D., *Practical introduction to record linkage for injury research*. Injury Prevention, 2004. **10**(3): p. 186-191.
184. Gu, L., et al., *Record linkage: Current practice and future directions*. CSIRO Mathematical and Information Sciences Technical Report, 2003. **3**: p. 83.
185. Herzog, T.N., F.J. Scheuren, and W.E. Winkler, *Data quality and record linkage techniques*. 2007: Springer Science & Business Media.
186. Dusetzina, S.B., et al., *Linking data for health services research: a framework and instructional guide*. 2014.
187. *Inspecting and cleaning an event log - PROM*. [cited 2016 26 June]; Available from: <http://www.promtools.org/doku.php?id=tutorial:preprocessing>.
188. Taunton and Somerset - About SCR. [cited 2016 22 June]; Available from: <http://www.somersetscr.nhs.uk/about.html>.
189. Trisoft - TheatreMan. [cited 2016 22 June]; Available from: <http://www.theatreman.co.uk/index.php/products/theatreman>.
190. Gallo, M., *Nine Tips to Improve Data Quality and Improve Your Decision Making*, in *The Light Touch*. 2013.
191. Bird, A. *A guide to case notes and record-keeping* Nursing in Practie, 2012.
192. MDU, *Good Record Keeping - Staying Patient Focused*. 2013.
193. Birkin, R., *Top ten tips for good record keeping*. 2017.

194. Signe Agnes Flottorp, G.J., Bernhard Gibis, Martin McKee, *Using audit and feedback to health professionals to improve the quality and safety of health care*. 2010, European Observatory on Health Systems and Policies.
195. Caron, F., et al., *Monitoring care processes in the gynecologic oncology department*. Computers in biology and medicine, 2014. **44**: p. 88-96.
196. Alliance, L.C., *LCA Best Practice Prostate Pathway*. 2013.
197. Coursera. *Lecture 55 - 6.4: Process Mining Software*. [cited 2017 13 Sept]; Available from: <https://www.coursera.org/learn/process-mining/lecture/uJS05/6-4-process-mining-software>.
198. Leemans, S.J., D. Fahland, and W.M. van der Aalst. *Exploring processes and deviations*. in *International Conference on Business Process Management*. 2014. Springer.
199. Leemans, S.J., D. Fahland, and W.M. van der Aalst. *Using life cycle information in process discovery*. in *International Conference on Business Process Management*. 2015. Springer.
200. Hurhangee, P., *Prostate Cancer Referral Pathways*, B. Siddiqi, Editor. 2017.
201. Andrews, K. *Evaluation comes in many guises*. in *AVI Workshop on Beyond time and errors (BELIV) Position Paper*. 2008.
202. Lam, H., et al., *Seven guiding scenarios for information visualization evaluation*. 2011.
203. Merčun, T. *Evaluation of information visualization techniques: Analysing user experience with reaction cards*. in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. 2014. ACM.
204. Feinberg, J. *Wordle*. 2014; Available from: <http://www.wordle.net/>.
205. Swenson, K., *What is the value of process mining?*, in *BPM*. 2012.
206. Rozinat, A., *4 Challenges for Process Mining in Healthcare*, in *Flux Capacitor*. 2011.
207. Veyrat, P., *5 awesome ideas for business process simplification*, in *Heflo*. 2016.
208. Peterson, *How to automate a business process*. 2017.
209. Veyrat, P., *Business Process Standardization: All you need to know.*, in *Heflo*. 2016.
210. Komulainen, R.M.O., *4 Steps to Robotic Process Automation Success with Process Mining*, in *QPR*. 2017.

APPENDICES

APPENDIX A

PRISMA CHECKLIST

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	80
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	81
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	82-89
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	89
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	N/A*
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	91
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	92
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	92

Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	93
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	93
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	N/A*
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	N/A*
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A*
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A*

Page 1 of 2

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	N/A*
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A*
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	94-95
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Appendix B
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A*

Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	96-120
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A*
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A*
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A*
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	121-124
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	125
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	126
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	N/A

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed.1000097

For more information, visit: www.prisma-statement.org.

Page 2 of 2

* Meta-analysis not done

APPENDIX B

DATA EXTRACTION SPREAD SHEET FOR SYSTEMATIC REVIEW

Study No.	Study Name	Citation	Date searched	Study year	Does the study report as an outcome an impact on the control-flow perspective, organisational perspective, case perspective or time perspective	Research Question	Study Location (country, setting)	Time Frame	Healthcare process analysed	Data collection methods	Data analysis methods	Results	Conclusions	Visualisation techniques	Evaluation means	Preprocessing	Notes
1	Application of Process Mining in Healthcare - A Case Study in a Dutch Hospital	13-Mar	2008	control-flow, organisational, performance, case	Can process mining be used to obtain insights into care flows	AMC hospital Amsterdam, Netherlands	2005-2006	Gynaecological oncology	billing system of AMC hospital	process discovery, social network mining, performance analysis	average sojourn times between events, throughput time for each process, relationship between medical departments, diagnostic results, overall process events	it is possible to mine complex and flexible hospital processes giving insight into these processes	Self Organising Map (SOM), heuristics model, fuzzy model, social network, dotted chart, meter chart	They compared their work with a flowchart. Flowchart construction took more time whereas process mining was semi-automatic	aggregation		
2	Process Mining Techniques an Application to Stroke Care	13-Mar	2008	control-flow, performance, case	Can process mining techniques be applied successfully to clinical data to gain a better understanding of different clinical pathways adopted by different hospitals and for different groups of patients	4 Italian hospitals		Ischemic stroke	MS Access stroke clinical data set, patient interviews	heuristics miner	Observed a difference in treatment strategies between different hospitals. Visualized the pre-hospitalisation pathways and identified bottlenecks.	process mining techniques can be applied successfully to clinical data to gain a better understanding of different clinical pathways adopted by different hospitals and for different groups of patients	heuristics model, petri net,				
3	Acquiring Objective Process Information for Healthcare Process Management with the CRISP-DM Framework		2009	performance, control flow, case	research the applicability of process mining on acquiring objective process information in the healthcare domain			rheumatoid arthritis		heuristics miner and the fuzzy miner, linear temporal logic (LTL) checker and the conformance checker, performance analysis miner		petri net, heuristic model, fuzzy model, dotted chart	filtering, Detecting and correcting typos, outliers, and missing values				

4	Healthcare Process Analysis: validation and improvements of a data-based method using process mining and visual analytics	13-Mar	2009	control flow, case, organisational, performance	can we provide healthcare organisations with process information that will improve their clinical pathways	AMC hospital Amsterdam, Netherlands	2005-2008	gynaecological oncology	billing system of AMC hospital	Heuristic miner, fuzzy miner, genetic miner, social network mining, performance analysis	average sojourn times between events, throughput time for each process, relationship between medical departments, diagnostic results, overall process events	A method was proposed for healthcare process analysis that could enhance the process-related information needed by these organizations in order to produce process improvements.	all activities visualisation by MagnaView, sequential pattern MagnaView, heuristic miner, performance self organising map (SOM), bar chart, fuzzy model, social network, genetic model, petri net model,	Checking the reliability and relevance of the method in satisfying the information needs for healthcare process improvements of the different stakeholders at the AMC process	Aggregation, adding attributes, filtering	
5	Process improvement in Healthcare A data-based method using a combination of process mining and visual analytics	13-Mar	2009	control flow, performance, case	can we design a method for the control and diagnosis phases of the BPM life cycle in a healthcare environment	Dutch hospital, Netherlands	2006-2008	mammacare, diabetes foot	From data warehouse to MySQL DB	process discovery, pattern analysis, performance analysis	waiting time till surgery, number and type of amputations, number of repeating consultations, Number of nursing days	The combination of the process mining and visual analytics approach resulted in visualization options that are more related to activities in the treatment process and the relations found between these activities than that are related to the monitoring of treatment processes.	all activities visualisation by MagnaView, sequential pattern MagnaView, heuristic miner, self organising map (SOM)	medical specialists and managers, three patients	Aggregation and adding attributes, clustering, filtering	
6	Process Mining in Healthcare: A Contribution to Change the Culture of Blame	13-Mar	2009	control flow	can process mining techniques be used as a mean to discover not only individuals' error, but also chains of responsibilities and manage the clinical risk	Mondino Hospital in Pavia and 4 Italian hospitals		stroke	from care flow management system	process discovery	heuristics model	two hospitals adopt different strategies	heuristics miner			
7	Combining Business Process and Data Discovery Techniques for Analyzing and Improving Integrated Care Pathways	13-Mar	2010	performance	can process mining be combined with data discovery techniques to analyse integrated care pathways	Belgium hospital	Jan - Jun 2008	breast cancer	process discovery techniques (hidden markov models) and data discovery techniques (formal concept analysis)	process inefficiencies, exceptions and variations	a combination of process discovery techniques and data discovery techniques were used to gain a deeper understanding of an existing breast cancer care process and the actual activities performed on the working floor	concept lattices, process models				
8	Social Network Analysis for Business Process Discovery		2010	Organisational perspective	In this work we present an approach that aims at finding communities inside a social network.	Portugal		emergency dept.	Organizational Miner Cluster, Cluster Social Network		Algorithm successful in discovering correct communities, demonstrated usefulness of Modularity concept; derived conclusions, information about social network and depicted business process based on extracted information from social network	social network graph,				
9	DISCOVERING PATIENT CARE PROCESS MODELS FROM EVENT LOGS		2010	control flow, organisational, case		Belgium hospital		pre-surgery	Heuristic and Genetic mining algorithms		performance sequence diagram, heuristic model, genetic model		filtering			
10	Workflow support for the healthcare domain	13-Mar	2011	control-flow, resource, performance perspectives	What are the main paths that are followed by patients? What are the most important collaborations with other depts.? Are there any performance bottlenecks in the process?	AMC hospital Amsterdam, Netherlands	2005-2008	gynaecological oncology cervical cancer surgery patients	billing system of AMC hospital	Process discovery, social network mining, performance analysis	average sojourn times between events, throughput time for each process,	it is possible to mine complex and flexible hospital processes giving insight into these processes	dotted chart, heuristics model, ERC model, petri net model, fuzzy model, social network, meter chart, bar chart		Aggregation	

11	Business process analysis in healthcare environments: A methodology based on process mining	2011	control-flow, organisational, performance, case	Main goal of our work is to devise a methodology based on process mining in order to support BPA in healthcare.	Hospital of São Sebastião, Portugal	Jan-July 2009	Radiology workflow in Emergency department	Medtriz database in HIS	process discovery, performance analysis	clustering plays a keyrole in identifying regular behavior, process variants, and infrequent behavior as well. This is done by means of a cluster diagram and a minimum spanning tree, which provide	we have presented a methodology based on process mining for the analysis of healthcare processes	heuristic model, petri net, cluster diagram, spanning tree, performance analyzer, social network	filtering	Organisational and performance perspectives less explored. Also, previous authors do not describe nor formalize a methodology for BPA in healthcare based on process mining. Rather, they tend to focus on a specific technique or a specific perspective of the process.
12	Analysis of Patient Treatment Procedures The BPI Challenge Case Study	29-Mar	2011	control-flow, case	can we develop a systematic approach for the analysis of hospital event log using control-flow and case	AMC hospital Amsterdam, Netherlands	gynecological oncology	raw event log	fuzzy miner and trace alignment	Through enhanced fuzzy mining and trace alignment	Adopting the systematic approach presented in this paper, we realized/Showed that the	Hasse diagram, histogram, activity burst diagrams fuzzy	none used	trace clustering, aggregation
13	Mining Processes in Dentistry	2012	control-flow, organisational, performance	process mining techniques are applied in order to demonstrate that, based on automatically stored data, detailed process knowledge can be obtained on dental processes, e.g. it can be discovered how dental processes are actually executed.	Dental practice Netherlands	2008-2010	Single crown implant Dentistry	dental lab DB	performance analysis, social network mining, process discovery (heuristic mining)	process mining is a technology that is of value to discover actual process behavior, even when it involves multiple parties across organizational boundaries, as is clearly the case in the dental domain	heuristics model, petri net, social network	with process owners	Log integration and event similarity change	shows validation results too. See!
14	On mining clinical pathway patterns from medical behaviors	2012	control-flow, case	can a process mining approach be used to find a set of clinical pathway patterns given a specific clinical workflow log and minimum support threshold	Zhejiang Huizhou Central hospital of China	2007-2009	bronchial lung cancer, gastric cancer, cerebral hemorrhage, breast cancer, infarction, and colon cancer	clinical pathway pattern mining	sequential order of activities, information about the time span between different pairs of activities	proposed approach provides the ability to discover clinical pathway patterns that cover the most frequent medical behaviors which are regularly encountered in clinical practice	workflow patterns	evaluated by clinical experts and hospital managers		
15	Discovering User Communities in Large Event Logs	2012	Organisational perspective	How to use hierarchical clustering together with the concept of modularity to analyze social networks obtained from large event logs	Portugal	emergency dept.	social network mining,			social network graph,		filtering, clustering		

16	On process mining in health care	2012	control flow, case	we argue that existing process mining methods fail to identify good process models, even for well-defined clinical processes.	Netherlands	anesthesia during Endoscopic Retrograde Cholangiopancreatography	process discovery (heuristic miner)	Although the algorithm could recognize some key activities, the evaluation with the medical expert has revealed that there were sequences in the model that did not make sense from a medical perspective and that the algorithm must have determined a sub-optimal process model	current algorithms of process mining are of limited use for analyzing health care processes. Even for well-defined processes of medium complexity	heuristic models by health professionals	Filtering
17	Pathway identification via process mining for patients with multiple conditions	2012	control flow, performance	can we develop an approach based on process mining to identify clinical pathways for patients with multiple conditions through historical event logs of patients with similar conditions	general hospital Singapore	2011	eye problem and autoimmune disease	EMR	A heuristics miner algorithm is developed and tested with a case study involving patients with multiple conditions	It is found that process mining can be applied in the healthcare context to identify pathways with correct ordering of healthcare services and process constructs. These pathways provide a reference process model for designing healthcare services that are catered to individual patients' needs.	heuristics model
18	Process Mining in Healthcare Systems: An Evaluation and Refinement of a Methodology	2012	control flow, performance, organisational, case	unveil the possible problems and issues that are encountered during the application of the original methodology and to come up with quick problem solutions via few adjustments in the methodology.	Isala, AMC, Netherlands	Isala Urology and AMC gynecological oncology	performance analysis, social network mining, process discovery (heuristic mining)	By including the preprocessing step combined with the questions concerning the selection of certain activities, it is now possible to reduce the amount of activities, but still keeping the relevant and interesting activities. This increases the process models readability and comprehensibility, because one of the major problems with regard to process mining in healthcare systems is that the log contains many distinct activities; especially with many being rather low level activities, which results in spaghetti-like models.	markov chain cluster, heuristic models, fuzzy models, dotted charts, social network	results presented to experts	sequence clustering

19	Process Mining of Clinical Workflows for Quality and Process Improvement	2012	control flow, performance	can automated process discovery and process mining be important to improving clinical processes	MERCY health system, USA	2011	Congestive heart failure	MERCY EHR event logs	process discovery (alpha miner, heuristics miner, genetic miner, performance analyser)	there is significant promise for process mining of clinical processes and Rich data sets comprising process activities and related contextual data elements can be extracted from the EHR and formatted for process mining	heuristic model, petri net.	filtering	
20	Gaining Insight from Patient Journey Data using a Process-Oriented Analysis Approach	2012	control-flow, performance, case	This study aims to gain insight into patient journey data to identify problems that could cause access block.	Flinders Medical Centre (FMC), Australia	Emergency dept.	Patient Journey Database	statistical analysis, pattern analysis, Performance Sequence Analysis	investigate the correlation of patterns and bottlenecks,	The process undertaken has proven to be a viable approach in analysing the inpatient journey.	pattern diagram, sequence diagram, workflow diagram	clustering	
21	Healthcare Analytics Examining the Diagnosis-Treatment Cycle	2013	control flow, case	propose a clinical pathway analysis method for extracting valuable medical and organizational information on past diagnosis-treatment cycles that can be attributed to a specific clinical pathway	European academic hospital	2005-2008	gynecological oncology	heuristic mining, statistical analysis,	presents a novel approach that is based on process mining techniques to acquire insight in the real sequence of healthcare activities performed on specific patients.				
22	process-oriented methodology for evaluating the impact of IT A proposal and an application in healthcare	2013	control flow, organisational performance, case	we have proposed a process-oriented methodology for investigating the impacts of IT. The methodology is based on both process mining and discrete event simulation.	Netherlands	digital dentistry	dental lab DB	discrete event simulation, process mining	The total throughput time, the total time spent by people in the lab, and the total time spent by a dentist.	By using process mining, an objective view is obtained how processes are really executed. Furthermore, the actual simulation phase in which impacts of digital technologies are evaluated can be started much quicker compared to the traditional approach, where simulation models are created manually. Also, the methodology allows for obtaining non-trivial quantifiable process insights that by following another methodology perhaps would not have been obtained.			
23	An alignment-based framework to check the conformance of declarative process models and to preprocess event-log data	2013	control flow, performance	propose the implementation of a framework for the analysis of the execution of declarative processes	Dutch hospital, Netherlands	bladder cancer		process discovery and conformance checking on declare models	a novel log preprocessing and conformance checking approach tailored towards declarative models has been presented and evaluated	execution time prediction and analysis through bar chart and transition diagrams, social network, trace model alignments,	log model alignment		
24	An analysis and redesign of the ICU weaning process using data analysis and process mining	2013	control flow, performance, case	Analyze and improve a medical treatment process with the help of data analysis and process mining techniques applied on data from a clinical support system	Maastricht University Medical Centre (MUMC), Netherlands	ICU Weaning protocol	patient data management system (PDMS) and Critical Care and Anesthesia (ICCA) system	regression analysis, heuristics miner, fuzzy miner	examining and improving the patient routes in a medical treatment process using process mining.	This study enhances process mining research with a study that uses process mining techniques to show that (automatic) generation of patient route improvements options for a medical care process is possible.	heuristic model, fuzzy model,	medical stakeholders	see interesting paper about exact way you should use too

25	Discovery of Outpatient Care Process of a Tertiary University Hospital Using Process Mining	2013	control flow, case	potential of a process mining technique to determine an outpatient care process that can be utilized for further improvements	Seoul National University Bundang Hospital, Korea	2012-2013	outpatient processes	heuristic mining and fuzzy mining	process mining techniques can be applied in the healthcare area, and through detailed and customized analysis in the future, it can be expected to be used to improve actual outpatient care processes.
26	Exploration of the link between the execution of a clinical process and its effectiveness using process mining techniques	2013	control flow, performance, organisational, case	Can process mining techniques help in determining the link between the execution of a clinical treatment process and its effectiveness	academical hospital of Maastricht, Netherlands	cataract treatment	SAP DB	fuzzy miner	This research is novel since it is the first process mining case study that focuses specifically on process effectiveness. It shows that by operationalizing process effectiveness based on well-chosen performance indicators that are relevant (to patients), very useful insights can be gained on the link between the execution of a clinical process and its effectiveness
27	Getting a Grasp on Clinical Pathway Data An Approach Based on Process Mining	2013	control flow, case, organisational	deriving useful insights from clinical pathway data by making use of process mining techniques	AMC hospital, Netherlands	gynecology oncology			major benefit of the technique is the enhancement of pure control flow patterns with other data dimensions
28	Mining and Retrieving Medical Processes to Assess the Quality of Care	2013	control flow, performance	process mining and case retrieval techniques, relying on a novel distance measure, to stroke management processes. Specifically, the goal of the framework is the one of analyzing the quality of stroke management processes	Italy	stroke	heuristic miner, case retrieval techniques	networked graph, fuzzy models	This work showed that process mining and case retrieval techniques can be applied successfully to clinical data to gain a better understanding of different medical processes adopted by different hospitals (and for different groups of patients).
29	Process Mining in Healthcare Data Challenges when answering frequently posed questions	2013	control flow, performance, organisational, case	which process mining data can be found in current Hospital Information Systems (HIS). Does it allow for solving frequently posed questions.	Netherlands	2009-2012	colorectal cancer	dotted chart, petri net,	Based on the questions posed by the medical professionals, data may be required from different data sources. This requires that links between the four systems of the spectrum are clear.

30	Process Mining in Healthcare Opportunities Beyond the Ordinary	2013	control flow, case, performance	What are the possibilities of process mining within hospitals?	umcm, Netherlands	2008-2012	intestinal cancer	heuristic miner, fuzzy miner, performance	we presented a healthcare reference model which exhaustively lists typical data that exists within a HIS and that can be used for process mining	dotted chart, performance model, heuristic model, fuzzy model, petri net
31	Process mining to facilitate process improvement in a healthcare environment An emergency department case study	2013	control flow, performance, organisational, case		USA	Emergency dept.	heuristic miner, fuzzy miner, organisational miner	process mining has proved to be a feasible and effective implementation with traditional process improvement methodologies. The results from process mining can facilitate process improvement tools and enhance the methods involved in PI projects	dotted chart, social network, heuristic model, fuzzy model, organisational model, values stream map	CLOSEST TO YOUR LIT REVIEW!!
32	Investigating Clinical Care Pathways Correlated with Outcomes	2013	control flow, case		USA	Congestive heart failure	process mining, frequent pattern mining	Trace clustering, frequent pattern mining and overlay of frequent patterns on a minimum model are implemented as new features in BPM as a result of our work	frequent pattern model, heuristic model	physician filtering and trace clustering
33	A Systematic Methodology for Outpatient Process Analysis Based on Process Mining	2014	control flow, case, performance	we suggest a method to analyze outpatient processes based on process mining	Korea	2012	outpatient processes	heuristic miner, fuzzy miner, pattern analysis, comp mining	heuristic model, fuzzy model, comp model, simulation,	very good for PhD see delta analysis
34	Health intelligence: Discovering the process model using process mining by constructing Start-to-End patient journeys	2014	control flow, case, performance	gain insight into patient journeys from the point of admission to the Emergency Department (ED) until the patient is discharged from the hospital	Australia	general medicine and cardiology	heuristic model, performance	This paper outlines how unstructured event data were processed to derive the event logs needed as an input for process mining in the absence of PAIS. Using the processed event data, process mining was then applied for an evidence-based process model discovery of patient journeys from start to end at Flinders Medical Centre (FMC).	heuristic model, petri net	deriving fields
35	Process Mining Based Modeling and Analysis of Workflows in Clinical Care - A Case Study in a Chicago Outpatient Clinic	2014	control flow, performance, case	an outpatient clinic in Chicago, Illinois, USA, is used as a case study to illustrate a process mining based method for healthcare processes management and improvement. This method is able to discover meaningful knowledge	USA	outpatient processes	alpha algorithm, fuzzy miner, simulation	The results suggest that this methodology is a useful and flexible tool for healthcare process performance improvement.	alpha algorithm model, fuzzy model, simulation model	detecting and correcting typos, outliers, and missing values

36	Temporal Data Mining and Process Mining Techniques to Identify Cardiovascular Risk-Associated Clinical Pathways in Type 2 Diabetes Patients	2014	control flow, performance, case	how temporal and process mining techniques can be employed together to extract comprehensible and clinically meaningful process models from the event logs extracted from healthcare information systems.	Italy	Type 2 Diabetes	heuristic miner	This work tackles the major challenges we faced managing complex clinical and administrative temporal data through a range of methods derived from temporal and process data mining research in order to derive meaningful healthcare pathways.	heuristic model	merging
37	Analysis of the Social Network Miner (Working Together) of Physicians	13-Mar	2015	Organisational perspective	analyze and investigate the relationships between staff and resources in a hospital using process mining social network miner technique with respect to working together metric	Thailand estate hospital	2014	"different disease processes"	MS Access and excel extractions	social network mining
38	Application of Process Mining and Semantic Structuring Towards a Lean Healthcare Network	2015	organisational perspective, case, control flow		Italy	asthma	fuzzy miner, semantic structuring	contribute in giving medical managers an accurate and deep understanding of the healthcare network functioning. There are several contributions: the first contribution is the definition of an ontology of the healthcare network: general concepts and relationships. Then, starting from the data organized in the model, the process mining analysis is used to extract information for the network evaluation.	fuzzy model, ontology model	merge data, aggregate
39	Declarative process mining in healthcare	2015	control flow, performance, case	how to use process mining techniques based on declarative models to analyze medical treatment processes	Isla, Netherlands	Cryptorchidism, urology	discovery, conformance	declare model	filtering	see important things

40	Discovering organizational process models of resources in a hospital using Role Hierarchy Miner	2015	Organisational perspective, case	Thailand estate hospital	various diseases	role hierarchy miner, fuzzy miner	we could discover a holistic model/graph representing the different role/positions (and structural functions) of the doctors in different levels of an estate governmental hospital in Bangkok, Thailand.	role hierarchy graph, fuzzy model	good to see steps to DISCO
41	PROCESS MINING IN HEALTHCARE A Case Study	2015	control flow, organisational performance	AMC, Netherlands	2005-2006	gynecological oncology	heuristic mining, clustering, dotted chart, social mining	The results show that process mining can be used to provide new insights that facilitate the improvement of existing care flows	heuristic model, social network, dotted chart
42	RTLS-based Process Mining Towards an automatic process diagnosis in healthcare	2015	control flow	France		outpatient processes	fuzzy miner	fuzzy model	
43	Supporting healthcare management decisions via robust clustering of event logs	2015	control flow, case	Greece		Emergency dept.	fuzzy miner, clustering	fuzzy model, spectral clustering	spectral clustering
44	Time performance analysis of medical treatment processes by using disco	2015	control flow, performance, case	Thailand		treatment processes	fuzzy mining	fuzzy model	filtering, Change attribute format

APPENDIX C

SAMPLE ASP AND TSQL CODE SNIPPETS

1. I used the following generic TSQL codes in performing the functions described in the linkage steps:

- **Union All** statement

```
SELECT * FROM TABLE1 UNION ALL  
SELECT * FROM TABLE2
```

- **Inner Join** statement

```
SELECT * FROM TABLE1 INNER JOIN  
SELECT * FROM TABLE2 ON TABLE1.NHSNUMBER = TABLE2.NHSNUMBER
```

- **Left Join** statement

```
SELECT * FROM TABLE1 LEFT OUTER JOIN  
SELECT * FROM TABLE2 ON TABLE1.NHSNUMBER = TABLE2.NHSNUMBER
```

- Filtering prostate cancer patients from a DB

```
SELECT * FROM TABLE1 WHERE (DIAGNOSISCODE LIKE '%C61%')
```

2. The following is a snippet of the pseudo code that allows the automatic extraction and transformation to event log format (note only a few tables are shown as an example, the code is similar for the rest of the tables with their respective variables):

```
SELECT ALL RECORDS WITH UNIQUE PKEY FROM STAGING_TABLE AND PUT IN A NEW RECORDSET CALLED RSNHSNUM ORDERED BY  
PKEY  
FOR EACH ROW OF THE RECORDSET DO  
    'GET SERVICE AND PROCEDURE INFO FOR EACH NHSNUMBER ONE BY ONE USING THE ABOVE SEL STAGING_TABLE ECTED  
    TABLE  
    'OUTPATIENT APPOINTMENTS TABLE VARIABLES  
    SELECT ALL RECORDS WITH UNIQUE DATEOFAPPOINTMENT, PKEY, NHSNUMBER, DATERAISEDBYGP,  
    ORIGINALGPREFERRALDATE, REFERRALREQUESTRECEIVEDDATE, CLINICNAME1, CONSULTANTNAME, REFERRINGGP,  
    REFERRALCONSULTANT, REFERRINGSOURCENATIONAL, GPURGENTFLAG, ATTENDEDFLAG,  
    ATTENDEDORDIDNOTATTENDNATIONAL, DATEOUTCOMERECORDED, OUTCOMECODELOCALDESCRIPTION,  
    REASONFORAPPOINTMENTDESCRIPTION, APPOINTMENTTYPELOCAL, APPOINTMENTPRIORITYLOCAL,  
    HOSPITALCODEDESCRIPTION, SITECODE, MAINSPECIALITYCODELOCAL, SURNAME, FORENAME, DATEOFBIRTH,  
    AGEATSTARTOFSPELL, DATEOFDEATH, SEXCODELOCAL, ETHNICCODELOCAL, MARITALSTATUS, RELIGIONLOCAL, ADDRESS1,  
    POSTCODE, REGISTEREDGP, RTTSTARTDATE, RTTENDDATE  
    FROM STAGING_TABLE  
    WHERE PKEY= CURRENT_ROW(PKEY) AND DATEOFAPPOINTMENT IS NOT NULL  
    PUT RECORDS IN A NEW RECORDSET CALLED RSAPPOINTMENTS ORDERED BY DATEOFAPPOINTMENT  
  
    'PSA TABLE VARIABLES
```

```

SELECT ALL RECORDS WITH UNIQUE COLLECTDT, NHSNUMBER, LABDEPT, ORDERCODE, ORDERNAME, ORDERCOMMENT,
TESTCODE, TESTNAME, RESULT, RESULTUNITS
FROM STAGING_TABLE
WHERE PKEY= CURRENT_ROW(PKEY) AND COLLECTDT IS NOT NULL
PUT RECORDS IN A NEW RECORDSET CALLED RSPSA ORDERED BY COLLECTDT
•
•
•
'RADIOLOGY TABLE VARIABLES
SELECT ALL RECORDS WITH UNIQUE APPOINTMENT_DATE, NHSNUMBER, PROCEDURE_CODE
FROM STAGING_TABLE
WHERE PKEY= CURRENT_ROW("PKEY") AND APPOINTMENT_DATE IS NOT NULL
PUT RECORDS IN A NEW RECORDSET CALLED RSRADIOLOGY ORDERED BY APPOINTMENT_DATE

'INSERT EACH RECORDSET TRANPOSED INTO A NEW FLAT FILE TABLE
FOR EACH ROW OF THE RSAPPOINTMENTS RECORDSET DO
    INSERT INTO NEW_FLAT_TABLE ALL THE RECORDS OF THE RSAPPOINTMENT RECORDSET
    'ADD ADDITIONAL VARIABLES CALLED: "DATE_PROC2" AND "PROCESS_NAME" WHICH WILL HOLD THE END DATE
    OF THE PROCESS AND THE NAME OF THE PROCESS RESPECTIVELY
        COPY APPOINTMENT_DATE VALUE TO DATE_PROC2 AND INSERT INTO NEW_FLAT_TABLE
        PROCESS_NAME="OUTPATIENT APPOINTMENT"
        INSERT PROCESS_NAME INTO NEW_FLAT_TABLE
        GO TO NEXT ROW
    ENDFOR
    •
    •
    •
    'DO THE SAME FOR ALL THE ABOVE RECORDSETS
GO TO NEXT ROW
ENDFOR

```

3. The following is the pseudo code for the script that made the patient-by patient episode arrangement possible:

```

SELECT ALL RECORDS WITH UNIQUE TRIAL, NHSNUMBER FROM STAGING_TABLE ORDER BY NHSNUMBER
SET TRIAL1=1
SET LAST_NUM=CURRENT_ROW("NHSNUMBER")
FOR EACH ROW IN THE RECORDSET DO
    CURRENT_NUM= CURRENT_ROW("NHSNUMBER")
    IF (CURRENT_NUM IS NOT EQUAL TO LAST_NUM) THEN
        INCREASE TRIAL1 BY 1

```

```
END IF  
UPDATE STAGING_TABLE SET TRIAL=TRIAL1 WHERE NHSNUMBER= CURRENT_ROW ("NHSNUMBER")  
LAST_NUM=CURRENT_NUM  
GO TO NEXT ROW  
ENDFOR
```

APPENDIX D

PROSTATE PATHWAY FULL METRICS

Metric No.	Metric	What are we measuring?	Data Item (s)	Source	Availability	Target
LCAPP1	First 2ww appointment for prostate cancer patients	Date from referral to first appointment is to be < 8 days	2ww appointment date – 2ww referral date	Cancer Waiting Times	Now	93%
LCAPP2	62 day first treatment	Date from referral to first treatment <63 days	First treatment date – 2ww referral date	Cancer Waiting Times	Now	85%
LCAPP3	Decision to treat	Date from referral to decision to treat < 31 days	Decision to treat date – 2ww referral date	Cancer Waiting Times	Now	Not yet set by the PG
LCAPP4	First 62 day treatment modality	% of patients receiving active monitoring as their first treatment	First treatment type	Cancer Waiting Times	Now	LCA comparison for outliers
LCAPP5	Biopsy	Date from referral to biopsy < 20 days	Sample collection date – 2ww referral date	COSD Core Data Item	2014	Not yet set by the PG
LCAPP6	MRI	Date from referral to MRI < 10 days	Procedure date (if imaging modality = MRI scan) – 2ww referral date	COSD Core Data Item	2014	Not yet set by the PG
LCAPP7	Pre Biopsy MRI	Date of MRI to be before date of biopsy	Sample collection date – Procedure date (if imaging modality = MRI scan)	COSD Core Data Item	2014	Not yet set by the PG
LCAPP8	% complete for all COSD items	To assess the validity of the data received as the COSD dataset is likely to be incomplete	Sample collection date; Procedure date	COSD Core Data Item	2014	Not yet set by the PG

APPENDIX E

TWO-WEEK WAIT CANCER REFERRAL FORM

URGENT SUSPECTED UROLOGY CANCER REFERRAL FORM		
PLEASE ENSURE THAT THIS FORM IS ATTACHED TO YOUR CHOOSE AND BOOK REFERRAL		
Hospital to which patient is being referred:		
Patient details		GP Details
NHS number:		Dr:
Surname:	Address:	
First Name:		
Age / D.O.B:	Tel:	
Address:		Email:
Postcode:		Date of decision to refer:
Tel day:	Tel eve:	Signature:
Have you informed the patient that you suspect a urology cancer? Y / N Have you given the patient the 2WW information leaflet Y / N Have you told the patient they will be seen within 2 weeks? Y / N Has the patient had a previous diagnosis of cancer? Y / N (Specify if known)		
Has the patient previously visited this hospital? Y / N Hospital number (if known):		First language: Interpreter required? Y / N
Prostate (please tick as appropriate) Either <input type="checkbox"/> hard, irregular prostate on digital rectal examination (DRE) Or <input type="checkbox"/> raised / rising age specific PSA with clinically malignant prostate or bone pain, or unexplained urological symptoms <input type="checkbox"/> asymptomatic with age specific raised PSA in men (<75 years) with negative MSU PSA value: ng/ml date of PSA test: If borderline, repeat in 1-3 months otherwise refer		

APPENDIX F

MICROSOFT'S PRODUCT REACTION CARDS (CONCISE LIST)

Entertaining	Patronizing	Irrelevant	Predictable	Organized
Innovative	Impersonal	Poor quality	Effective	Inviting
Convenient	Trustworthy	Professional	Stressful	Confusing
Cutting edge	Annoying	Familiar	Straight Forward	Efficient
Essential	Flexible	Powerful	Dated	Exciting
Attractive	Approachable	Simplistic	Difficult	Clean
High quality	Complex	Engaging	Dull	Desirable
Unrefined	Comfortable	Time-consuming	Unpredictable	Intimidating
Inconsistent	Satisfying	Fast	Exceptional	Useful
Easy to use	Comprehensive	Inspiring	Overwhelming	Unattractive
Consistent	Advanced	Busy	Undesirable	Friendly
Relevant	Personal	Rigid	Helpful	Reliable
Unconventional	Creative	Collaborative	Ineffective	