Healthcare Pathway Discovery, Conformance, and Enrichment

Christina Lin[a], Andreas W. Kempa-Liehr[a,*], Randall Britten[b,c], Delwyn Armstrong[d], Jonathan Wallace[d], Michael O'Sullivan[a]

[a]*Department of Engineering Science, The University of Auckland, 70 Symonds St, Auckland, New Zealand*
[b]*Auckland District Health Board, 2 Park Road, Auckland, New Zealand*
[c]*was at Orion Health, 181 Grafton Rd, Auckland, New Zealand*
[d]*Waitemata District Health Board, 124 Shakespeare Rd, Auckland, New Zealand*

Abstract

Background and purpose

Healthcare pathways define the execution sequence of clinical activities as patients move through a treatment process, and they are critical for maintaining quality of care. The aim of this study is to investigate the utilization of business process modelling (BPM) to design an adaptive healthcare pathway mining methodology, with particular emphasis on producing pathway models that are easy to interpret for clinicians without a sufficient background in process mining.

Method

This study utilizes the business process-mining software ProM to design a process mining pipeline for healthcare pathway discovery, conformance analysis, and enrichment using hospital records. The efficacy of the BPM approach is demonstrated via two case studies that apply the proposed process mining pipeline to discover appendicitis and cholecystitis pathways from hospital records. Machine learning methodologies based on probabilistic programming are utilised to explore pathway features that influence patient recovery time.

Results

The produced appendicitis and cholecystitis pathway models are easy for clinical interpretation and provide an unbiased overview of patient movements through the treatment process. Analysis of the discovered pathway model enables reasons for longer than usual treatment times to be explored and deviations from standard treatment pathways to be identified. A probabilistic regression model that estimates patient recovery time based on the information extracted by the process mining pipeline is developed and has the potential to be very useful for hospital scheduling purposes.

*Corresponding author
Email address: a.kempa-liehr@auckland.ac.nz (Andreas W. Kempa-Liehr)

Conclusion

This study establishes the application of the business process modelling tool ProM for the improvement of healthcare pathway mining methods. The proposed pipeline for healthcare pathway discovery has the potential to support the development of machine learning models to further relate healthcare pathways to performance indicators such as patient recovery time.

Keywords: Healthcare pathway; Process mining; Electronic health record; Probabilistic programming

1. Introduction

Healthcare pathways are critical for reducing clinical variability, affecting operational excellence, and maximizing health outcomes [1]. They define the execution sequence of clinical activities as patients move through a treatment process, a department, a hospital, or a wider health organization [2]. The accurate definition of healthcare pathways and patient conformance to those pathways is an issue of increasing relevance as precision medicine enables targeted approaches and diagnostic splitting. The proliferation of pathway branches is exponential, and pathways are increasingly non-linear.

Most healthcare pathways result from clinician-led practice rather than explicit pathway design via a consensus model and systems approach. In addition, healthcare pathways "shift" dynamically as steps in the pathway are altered or resources change along the pathway. If no explicit redesign of pathways is performed, then the providers of the pathways (and its associated resources) may be unaware of the change [3]. Pathway discovery (identification of pathways without a priori knowledge), conformance analysis (including gaps in care and clinical variability) and pathway enrichment (enrichment of a priori models with additional event data) are critical for healthcare services now, and into the future [4].

Past studies have shown that there is potential for informative healthcare pathways to be extracted from hospital health records [5], [6], but there is currently no consensus on a systematic healthcare pathway mining method that supports explicit design and conformance analysis of concise and comprehensible healthcare pathway models. The research described in this paper investigates the utilization of Business Process Modelling (BPM) as outlined by Becker et al. [7] to provide a scaffold for healthcare pathway discovery, conformance analysis and enrichment. The main objectives of applying BPM to healthcare data include:

1. Pathway discovery
    - Investigate the potential of ProM (a process-mining software package) to discover healthcare pathways from hospital records [8].
2. Conformance analysis
    - Apply BPM conformance analysis to discovered healthcare pathways.

- Improve detection of possible non-conformance and explain anomalies.

3. Data enrichment
   - Investigate correlation between healthcare pathway and performance indicators (e.g., patient length-of-stay, readmission rate).

2. Healthcare Pathway Mining Methodology

This chapter outlines the proposed process mining pipeline designed for mining healthcare pathways from hospital records using business process modelling tools. This study adopts the scientific computing practices recommended by Wilson et al. [9], [10] to ensure that all results are reproducible. The proposed process mining pipeline consists of three major phases that correspond to each of the three main objectives (i.e. pathway discovery, conformance, and enrichment). An overview of the process mining pipeline designed for this study is shown in Fig. 1. The following sections elaborate the three phases of the proposed pipeline and their respective objectives.

This study evaluates ProM (version 6.7) as the main process mining tool[1]. ProM is an open-source process mining software that is effective for construction of business models from input data files. The use of ProM in the healthcare sector has not been thoroughly explored in the past, but Van Der Aalst et al. have demonstrated the applicability of the software for industrial processes [11]. ProM is chosen for this study because it has an intuitive user interface and supports many process mining plug-ins [8]. The process mining and conformance analysis plug-ins supported by ProM are well documented. All these features of ProM make it easy for the process mining steps in the pipeline (see Fig. 1) to be repeated by users with no background in process mining.

2.1. Healthcare Pathway Discovery

Healthcare pathway discovery is the first phase of the proposed process mining pipeline. It consists of two steps: clinical data processing and pathway visualization, which are conducted iteratively until a concise model is produced. The aim is to use patient healthcare records stored in hospital information systems to design a concise pathway model that is easy for clinical interpretation. Therefore, clinical input is critical to the selection of appropriate processing methods.

Healthcare pathways generally have much higher levels of complexity than standard business processes, and unprocessed clinical data contains too many clinical variations for a clean and concise pathway model to be mined [12, 13]. Each pathway variant is a unique activity sequence of a complete patient trace. The ProM plug-in `Explore Event Log` extracts pathway variants from patient

[1] `http://www.promtools.org`

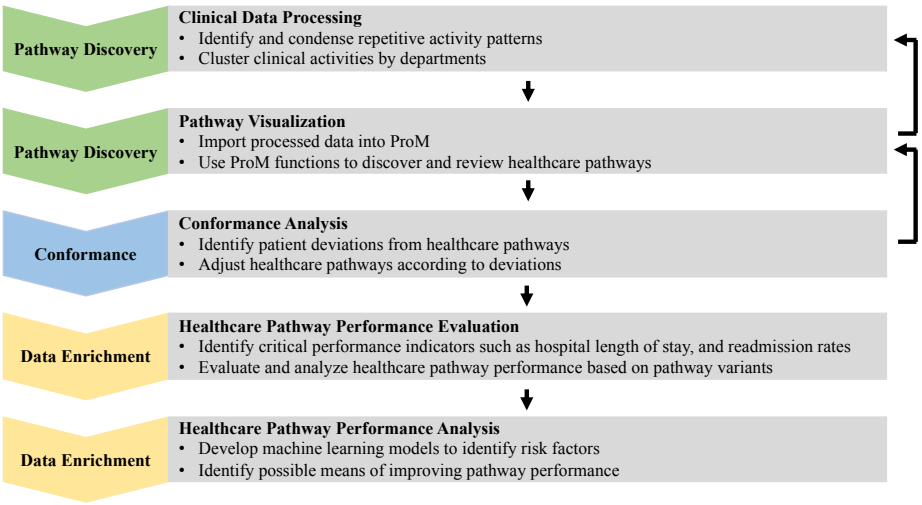

Figure 1: The process mining pipeline comprises three sections, which are subdivided into five steps: Clinical data processing, pathway visualization, conformance analysis, healthcare pathway conformance evaluation, and prescriptive analytics of healthcare pathways. The first three steps are connected in two iterative cycles.

traces, and the total number of pathway variants is an indicator of the level of clinical variation between patient traces.

In order to reduce healthcare pathway variations to a meaningful pathway model, the pathway variants visualized by the plug-in `Explore Event Log` are examined closely to determine the most suitable processing methods. There are three effective methods for reducing clinical variations without filtering patient traces:

Cluster clinical activities that are similar in nature so that the range of activities is reduced to a manageable size, e.g., 'Abdomen CT scan' and 'Pelvis CT scan' could be clustered into a single activity under 'CT scan'.

Merge consecutive clinical activities that are performed consecutively into a single activity, e.g., a patient receiving the same medication five times on the same day could be regarded as a single activity.

Condense repetitive activity patterns that repeat but exhibit variable cycle length. These patterns indicate an activity that must be performed periodically while the patient is waiting for a different activity to begin, e.g., lab tests to monitor a patient's condition, medication to prevent infection. These repetitive patterns could be condensed into a single, parallel activity.

Clinical input is highly recommended at this step particularly for complex or unfamiliar healthcare pathways.

2.2. Healthcare Pathway Conformance Analysis

Conformance analysis evaluates the degree to which a healthcare pathway model captures movement of patient traces by comparing the pathway model to clinical data. Accuracy of the discovered healthcare pathway model is validated if the majority of the patient traces conform to the model. Patient traces rarely all follow identical pathways, so the healthcare pathway model is not expected to capture all patient traces. The objective is to discover a healthcare pathway model that captures the fundamental structure of most patient traces.

ProM offers tools for conformance analysis of healthcare pathway models: Its plug-in `Inductive Visual Miner` compares patient traces from input clinical data to a healthcare pathway model and indicates patient traces which are deviating from the pathway model. For this purpose, the pathway model is visualized as a process tree, which is a hierarchical map comprised of decision nodes and tasks representing clinical activities [14]. Therefore, process trees enable the identification of pathway branches throughout the healthcare pathway model (c.f. Sec 3.1).

If valid patient traces deviate from a healthcare pathway model, adjustments are made to the model to improve patient conformance. A typical example might be the introduction of a new form of treatment, which has not been included into the model yet. Including these findings into the model leads to an iterative approach between pathway discovery and conformance analysis (Fig. 1). Conversely, conformance analysis can identify where invalid patient traces deviate from the model and investigate the reason for the discrepancy, e.g., clinicians following obsolete pathways or data errors.

2.3. Healthcare Pathway Data Enrichment

Data enrichment of healthcare pathways is the third phase of the discussed process mining pipeline (Fig. 1). It comprises two steps: Healthcare pathway performance evaluation and healthcare pathway performance analysis.

The main objectives of evaluating healthcare pathway performance are to understand the strengths and weaknesses of the current pathway design, and to identify potential methods of improvement. Possible indicators of healthcare pathway performance include waiting times of clinical activities, hospital length of stay, recovery time, and readmission rates [15]. Most of these indicators can be calculated or estimated using standard clinical timestamps. Postoperative Length of Stay (PLS), which is measured from end of surgery to discharge, can also be considered as patient recovery time. For surgical healthcare pathways, PLS is one of the critical indicators for evaluating healthcare pathway performance [16].

Analysing the performance of healthcare pathways with respect to pathway variants and other possible influencing factors like demographics or patient specific pathway characteristics, e.g. surgery duration (SD), is the final step of the process mining pipeline (Fig. 1). Due to the fact, that most pathway performance indicators do not follow normal distributions, while exhibiting significant stochastic volatility, neither classical hypotheses tests [17], nor point-predicting

machine learning models are appropriate for analysing healthcare pathways. Instead, probabilistic machine learning models [18] are used for extracting interpretable models from healtcare pathways (Sec. 3.2). For this purpose, feature engineering [19] from the patients's pathway traces (e.g. SD, pathway variant), demographics (e.g. age), as well as medical documentation like written diagnosis, time series, or images becomes important. In order to demonstrate this approach, the following case study discusses a probabilistic machine learning model for PLS, which takes into account pathway variants (Sec. 3.1), as well as demographics, and SD (Sec. 3.2).

3. Case Study: Appendicitis Healthcare Pathways

This section discusses the healthcare pathway discovery process for an appendicitis case study. For this purpose, two years' worth of data from 2015 to 2017 on 448 appendicitis patients have been analysed. A second case study on 52 cholecystitis patients is presented in Appendix A. The patient records for both case studies were collected from North Shore Hospital in Auckland, New Zealand. The data were de-identified and an ethics approval for this research was obtained.

3.1. Appendicitis Pathway Discovery and Conformance Analysis

The appendicitis pathway model, which has been generated by ProM's plugin `Explore Event Log`, is shown in Fig. 2 in the form of a pathway variant plot. The pathway variants are extracted without activities related to medication (i.e. preoperative and postoperative cefuroxime/metronidazole) because the clinicians confirmed that antibiotics are usually taken while the patient is waiting for surgery or discharge. The duration of these activities are therefore highly variable and result in a high number of unique pathway variants.

The pathway variant plot visualizes the 13 pathway variants of the appendicitis model sorted from the most common pathway (index 0) to the least common pathway (index 12). The top four variants account for approximately 88% of the patient traces (Tab. 1). All clinical activities are represented by a start event and a stop event. The activities are colour coded such that the same colour refers to the same clinical activity. The most common pathway variant (index 0) only consists of anesthesia and surgery, while the second most common variant (index 1) also includes preoperative X-ray. The pathway variant indices are used in Sec. 3.2 as one-hot encoded feature of the probabilistic machine learning model.

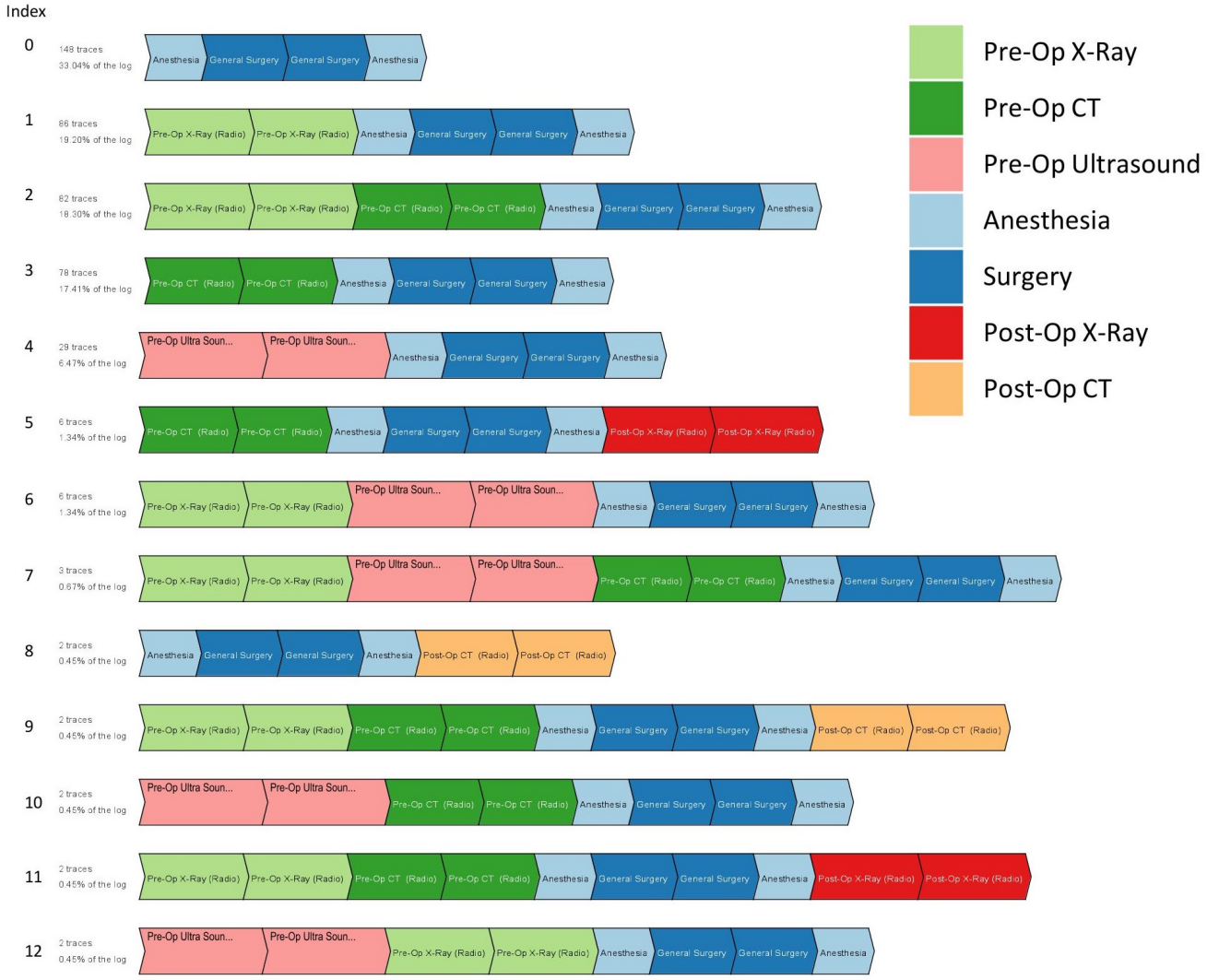

Figure 2: Appendicitis pathway variant plot auto-generated by ProM's plugin `Explore Event Log`. The plot visualizes the sequences of start and stop events for the different pathways. For the purpose of readability the legend was added and the statistics on the left are repeated in Tab. 1. The top four variants account for approximately 88% of the patient traces. They are modelled as one-hot encoded features V0, V1, V2, and V3 in Sec. 3.2, while pathway variants V4–V12 together represent the base model.

Table 1: Statistics of appendicitis patient traces shown in Fig. 2.

Variant	0	1	2	3	4	5	6
Patients	148	86	82	78	29	6	6
Percentage	33.04%	19.20%	18.30%	17.41%	6.47%	1.34%	1.34%
Variant	7	8	9	10	11	12	
Patients	3	2	2	2	2	2	
Percentage	0.67%	0.45%	0.45%	0.45%	0.45%	0.45%	

Table 2: Definitions of new pathway model notations.

Notation	Symbol	Definition
Orange Diamond		Decision Point, indicating exclusive choice
Orange Connector		Pathway Connection Point
Green Connector		Starting Point
Red Connector		Finishing Point

The first stage of the appendicitis pathway model visualized by `Inductive Visual Miner` is shown in Fig. 3. Unlike the pathway variants, this appendicitis pathway model incorporates activities representing antibiotics. The model indicates that 42 patients perform ultrasound and 183 patients perform X-ray upon admission.

While the complex process tree notation of ProM's `Inductive Visual Miner` plugin is optimal for detailed analysis, it has been reformulated under new notations for easy clinical interpretation. The new model notations are summarized in Table 2, and the reformulated appendicitis pathway model is shown in Fig. 4. The section of the model labelled as 'Stage 1' in Fig. 4 corresponds to the process tree shown in Fig. 3. This is the final pathway model that has been compiled based on clinical input to account for valid patient deviations, and all patient traces conform to the updated pathway. The most common pathway variant (index 0) shown in Fig. 2 corresponds to the horizontal path from start to finish in Fig. 4.

3.2. Data Enrichment of Appendicitis Pathways

The following section investigates the question of whether the pathway variants of the appendicitis case study are relevant features or covariates for explaining the stochastic volatility of PLS (Fig. 5a). This is quite a challenging task, because the individual healing process is expected to depend on personal factors like age [21] as well as the individual severity of the appendicitis inflamation, which in general is unknown at this stage of the data analysis, but might be captured by proxies like surgery duration (SD).

For the purpose of this analysis, patients' traces with complete demographic information (415 samples) are selected and the patients' PLS, which is calculated as the time interval between end of surgery and discharge, is converted

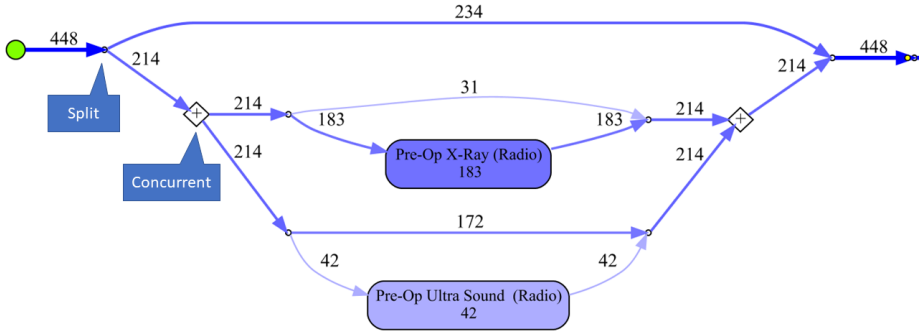

Figure 3: First stage of the appendicitis pathway model generated by ProM. The following stages have been omitted for the purpose of readability. The process indicates that 234 patients do not have any preoperative imaging diagnostics, while 214 patients enter the imaging diagnostics branch. Please refer to Leeman's manual on `Inductive Visual Miner` for details on the model notations [20].

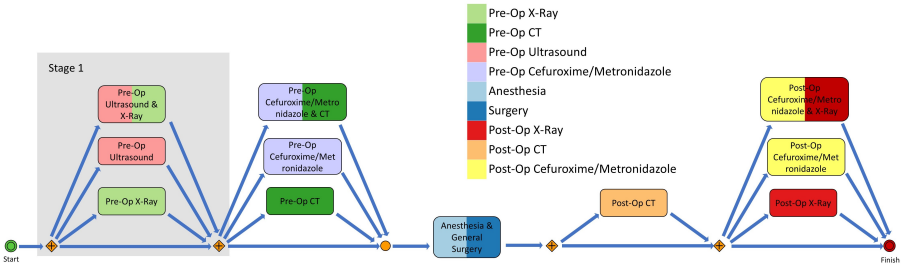

Figure 4: Appendicitis pathway model using nomenclature of Tab. 2. All preoperative and postoperative activities belong to radiology and pharmacy departments. Preoperative antibiotics are taken in the second stage of the treatment pathway.

into the number of days after surgery. Therefore, a PLS of one corresponds to a patient who has been discharged the day after surgery. Three of the $N = 415$ patient traces had a PLS of zero, because the surgery ended shortly after midnight and the patients were discharged on the same day. In order to simplify our analysis, the interpretation of PLS was broadend to the number of night rests after surgery, such that the PLS of these three patients could be casted to one.

Analysing the PLS of the 415 appendicitis patients reveals a histogram which closely resembles a Geometric distribution (Fig. 5a)

$$Pr(PLS = k) = \text{Geom}(k; p) = (1 - p)^{k-1} p^k, \tag{1}$$

parametrized by probability p of being discharged on the k^{th} day, respectively k nights of rest after surgery. However, estimating p as inverse mean of the

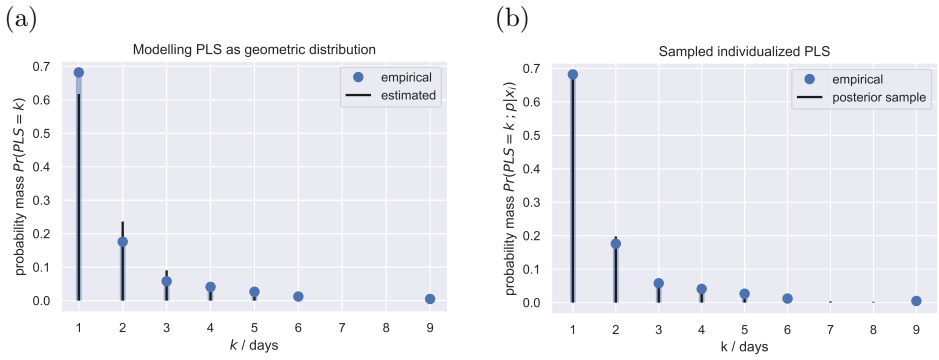

Figure 5: Modelling PLS as geometric distribution without patient specific model (a) and with individualised posterior (b) from model B (Eq (4)). Sampling from patient specific posterior distributions replicates the observed PLS quite well.

observed PLS to

$$\hat{p} = \frac{N}{\sum_{i=1}^{N} PLS_i} \approx 61.8\% \tag{2}$$

reveals that the average discharge probability $\hat{p}$ does not generalize well over the cohort (Fig. 5a), because the number of patients being discharged on the first day after surgery is underestimated and the number of patients being discharged on the second and third day after surgery is overestimated.

For the following probabilistic machine learning models, the individualized discharge probability p_i is modelled as a generalized linear model (GLM) using the inverse logistic function (logit) as a link function and the geometric distribution for generating the likelihood. For the following analysis, we are comparing two different models of discharge probability p_i as

$$\begin{aligned}\text{logit}(p_i) \sim SD_i + \log(age_i) + V_{0,i} + V_{1,i} + V_{2,i} + V_{3,i} + V_{0,i} \times \log(age_i) + \\ V_{1,i} \times \log(age_i) + V_{2,i} \times \log(age_i) + V_{3,i} \times \log(age_i),\end{aligned} \tag{3}$$

which we refer to as *model A*, and

$$\begin{aligned}\text{logit}(p_i) \sim SD_i + \log(age_i) + V_{0,i} + V_{1,i} + V_{2,i} + V_{3,i} + V_{0,i} \times \log(age_i) + \\ V_{1,i} \times \log(age_i),\end{aligned} \tag{4}$$

which we refer to as *model B*. These two models have been selected because they were the most credible based on the widely applicable information criterion (WAIC) and leave-one-out cross-validation (LOO) statistics [22]. The explanatory variables $V_{j,i}$ are categorical with

$$V_{j,i} = \begin{cases} 1: & V_{j,i} = j, \\ 0: & \text{otherwise.} \end{cases} \tag{5}$$

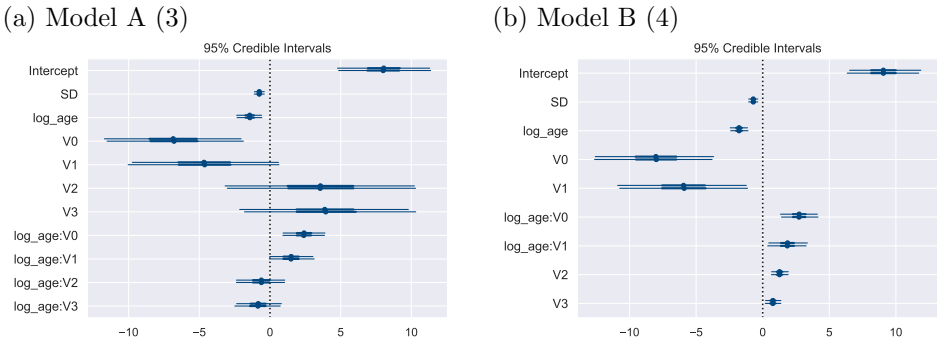

Figure 6: Credible intervals for models A (3) and B (4).

Therefore, the case $V_{0,i} = V_{1,i} = V_{2,i} = V_{3,i} = 0$ corresponds to variants V4–V12 of Fig. 2. Models A and B make the simplifying assumption that the individualised probability of discharge $Pr(PLS_i = k|x_i)$ can be estimated from

$$x_i = (SD_i, \log(age_i), V_{0,i}, V_{1,i}, V_{2,i}, V_{3,i}). \tag{6}$$

being available at the end of surgery, when future treatment steps and therefore the complete pathway variant for the respective i^{th} patient have been decided on:

$$\begin{aligned} Pr(PLS_i = k|x_i) &= \text{Geom}\left(p(x_i)\right) \\ &= \left(1 - p(x_i)\right)^{k-1} \left(p(x_i)\right)^k \end{aligned} \tag{7}$$

with $p(x_i)$ being modelled either by Eqn. (3) or (4).

The statistical model has been fitted using PyMC3 [23] version 0.24.2 using the No-U-Turn Sampler (NUTS) with 20,000 samples, 2,000 tuning steps, 2 chains, and an acceptance rate of 90%. The 95% credible intervals of the estimated model coefficients are shown in Fig. 6. The Gelman-Rubin convergence statistic Rhat is close to one for all coefficients (Tab. 3). However, for Model A the coefficients of the interaction terms $\log(age_i) \times V_2$ and $\log(age_i) \times V_3$ are close to zero with negative expectation values but a 25% and 15% probability of being larger than zero (Tab. 3). Therefore, we have decided to focus the following analysis of the fitted model on Eq. 4 for which all credible intervals exclude zero (Fig. 6b). However, this does not change the overall trend depicted in Fig. 7. It is important to note that both the coefficients of SD and $\log(age)$ are negative. This means that for the base model of variants V4–V12 as well as variants V2 and V3, the probability of discharge decreases with increasing SD and age. Compared to the base model, Variants V2 and V3 have a higher discharge propability because both coefficients are positive. The situation is more complicated for variants V0 and V1, because their coefficients are negative but the coefficients of their interaction terms with $\log(age)$ are positive, which basically counterbalances the effect of $\log(age)$.

Table 3: Coefficients of model B. Rhat is the potential scale reduction factor. A value of one indicates convergence.

	mean	hpd_2.5	hpd_97.5	Rhat
Intercept	9.09	6.48	11.85	0.999988
SD	-0.71	-1.059	-0.371	0.999987
log_age	-1.781	-2.46	-1.154	0.999988
V0	-8.04	-12.56	-3.70	0.999980
V1	-5.95	-10.82	-1.186	1.000013
log_age:V0	2.75	1.365	4.12	0.999980
log_age:V1	1.864	0.423	3.32	1.000013
V2	1.267	0.646	1.901	1.000062
V3	0.766	0.199	1.363	1.000046

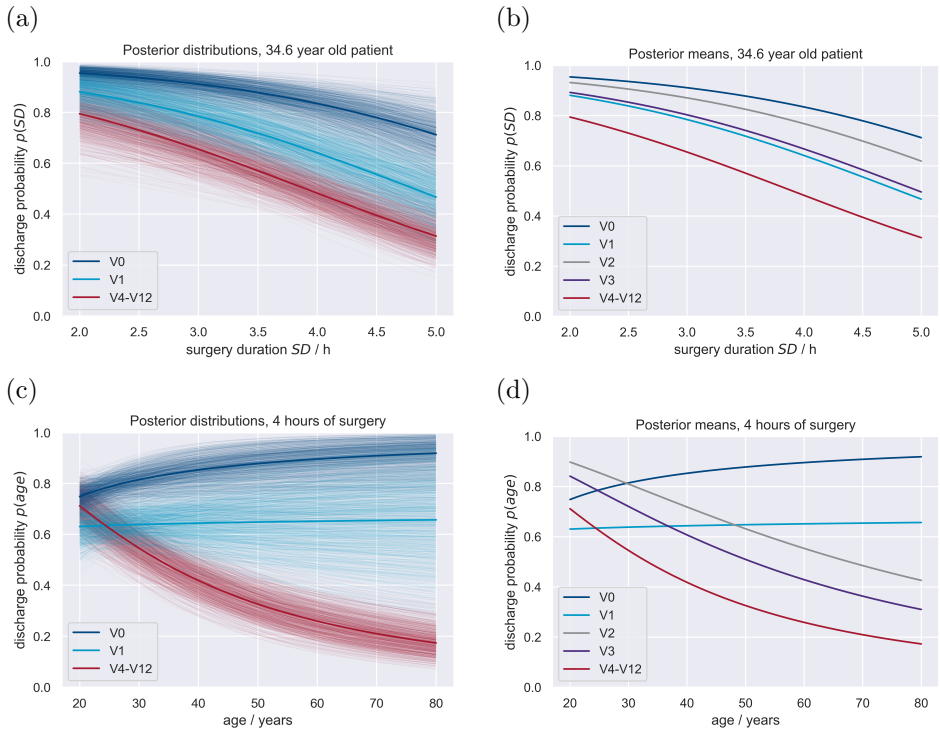

Figure 7: Posterior distributions and corresponding means for different pathway variants of model B (4). Trends with respect to SD and age are similar for model A (4).

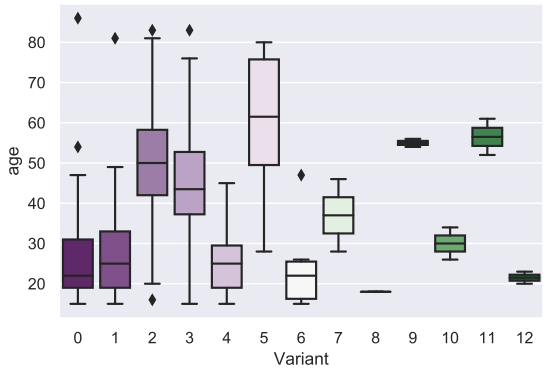

Figure 8: Boxplot of patients' age for pathway variants of the appendicitis model.

This is visualized in Fig. 7, which shows the posterior distributions of discharge probability p_i as function of SD (Fig. 7(a)) and age (Fig. 7(c)). While the discharge probability decreases with increasing SD for all pathway variants (Fig. 7(b)), the effect is different for age. The discharge probability p_i decreases for Variants V2, V3, and V4-V12 with increasing age of the patients, while its expectation value is constant for $V1$ and even increases by nearly 10% for V0. Note, that the corresponding graphics for model A (3), which are not shown here, resemble nearly the same correlations. From the clinical perspective, the difference in effects of age with respect to the pathway variants might be related to the fact that the most common pathway variants V0 and V1 are predominantly followed by younger patients (Fig. 8) and possibly older patients, for whom no complications are expected. However, the uncertainty of discharge probabilities for V1 is significantly larger compared to V0.

4. Conclusion

Healthcare pathways are critical for maintaining quality of care and improving health outcome for all patients, but there is no consensus on a healthcare pathway mining pipeline suitable for hospital implementation that supports pathway discovery from hospital health records. Business process modelling methods are used to design a process mining pipeline that produces concise and comprehensible healthcare pathway models from hospital records, and supports conformance analysis and enrichment of the discovered pathways. The proposed process mining pipeline successfully constructs concise pathway models for the appendicitis and cholecystitis case studies. The produced healthcare pathway models are easy for clinical interpretation and provide an unbiased overview of real patient traces through the treatment process. Probabilistic machine learning models for predicting postoperative length of stay, using information extracted by the process mining pipeline, is showing promising results. This

means that the proposed mining pipeline has the potential to support the development of machine learning models to further relate healthcare pathways to performance indicators. This study has established the use of business process modelling methods for the improvement of healthcare pathway mining methods, and there is value in investigating the capabilities of other business process modelling tools for healthcare pathway mining purposes.

Summary points
What was already known:

- Healthcare pathways are critical for reducing clinical variability, affecting operational excellence, and thereby maximizing health outcomes.
- Most healthcare pathways result from clinician-led practice rather than explicit pathway design via a consensus model and systems approach.
- There is currently no consensus on a systematic healthcare pathway mining method that supports explicit design and conformance analysis of concise and comprehensible healthcare pathway models.

What this study adds:

- The use of business process modelling methods improves the automatic mapping of healthcare pathways from clinical data.
- The application of business process modelling methods to healthcare pathway enables deviations from typical treatment pathways to be identified, all using standard clinical data timestamps.
- Probabilistic machine learning models enable the prediction of patient specific discharge probabilities, which lead to patient specific postoperative length of stay distributions.

Author Contributions

CL performed data analysis and processing using ProM, and created the new pathway visualisation method. AKL developed the probabilistic machine learning component of the work. MS and RB contributed during the conceptualisation of the project. MS, RB and AKL contributed high level steering and guidance during the pathways discovery phase of the work. DA assisted with data acquisition and understanding the context of the data. JW provided guidance on clinical subject matter. MS provided overall leadership of the project. All authors approved the final manuscript.

Acknowledgements

This research has been funded by the Precision Driven Health research partnership (project number 1209).

Declaration of Competing Interests

The authors declare that they have no competing interests for this study.

References

[1] F.-r. Lin, S.-c. Chou, S.-m. Pan, Y.-m. Chen, Mining time dependency patterns in clinical pathways, International Journal of Medical Informatics 62 (1) (2001) 11–25. doi:10.1016/S1386-5056(01)00126-5.

[2] Z. Huang, W. Dong, L. Ji, C. He, H. Duan, Incorporating comorbidities into latent treatment pattern mining for clinical pathways, Journal of Biomedical Informatics 59 (2016) 227–239. doi:10.1016/J.JBI.2015.12.012.

[3] Y. Zhang, R. Padman, L. Wasserman, N. Patel, P. Teredesai, Q. Xie, On Clinical Pathway Discovery from Electronic Health Record Data, IEEE Intelligent Systems 30 (1) (2015) 70–75. doi:10.1109/MIS.2015.14.

[4] K. Baker, E. Dunwoodie, R. G. Jones, A. Newsham, O. Johnson, C. P. Price, J. Wolstenholme, J. Leal, P. McGinley, C. Twelves, G. Hall, Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy, International Journal of Medical Informatics 103 (2017) 32–41. doi:10.1016/j.ijmedinf.2017.03.011.

[5] X. Xu, T. Jin, Z. Wei, J. Wang, Incorporating Topic Assignment Constraint and Topic Correlation Limitation into Clinical Goal Discovering for Clinical Pathway Mining, Journal of Healthcare Engineering (2017) 1–13doi:10.1155/2017/5208072.

[6] H. Iwata, S. Hirano, S. Tsumoto, Mining clinical pathway based on clustering and feature selection, in: Brain and Health Informatics. BHI 2013, Vol. 8211 of Lecture Notes in Computer Science, Springer, Cham, 2013, pp. 237–245. doi:10.1007/978-3-319-02753-1_24.

[7] J. Becker, M. Rosemann, C. von Uthmann, Guidelines of business process modeling, in: W. van der Aalst, J. Desel, A. Oberweis (Eds.), Business Process Management, Vol. 1806 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2000, pp. 30–49. doi:10.1007/3-540-45594-9_3.

[8] B. F. Van Dongen, A. K. A. De Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters, W. M. P. Van Der Aalst, The ProM framework: A new era in process mining tool support, in: G. Ciardo, P. Darondeau (Eds.), International conference on application and theory of petri nets, Vol. 3536 of Lecture Notes in Computer Science, Springer, 2005, pp. 444–454. doi:10.1007/11494744_25.

[9] G. Wilson, D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, P. Wilson, Best Practices for Scientific Computing, PLoS Biology 12 (1) (2014) e1001745. doi:10.1371/journal.pbio.1001745.

[10] G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, T. K. Teal, Good enough practices in scientific computing, PLOS Computational Biology 13 (6) (2017) e1005510. doi:10.1371/journal.pcbi.1005510.

[11] W. M. P. Van Der Aalst, H. A. Reijers, A. J. M. M. Weijters, B. F. Van Dongen, A. K. Alves De Medeiros, M. Song, H. M. W. Verbeek, Business process mining: An industrial application, Information Systems (2007) 713–732doi:10.1016/j.is.2006.05.003.

[12] Z. Huang, X. Lu, H. Duan, W. Fan, Summarizing clinical pathways from event logs, Journal of Biomedical Informatics 46 (1) (2013) 111–127. doi:10.1016/J.JBI.2012.10.001.

[13] G. M. Veiga, D. R. Ferreira, Understanding Spaghetti Models with Sequence Clustering for ProM, in: International Conference on Business Process Management, Springer, Berlin, Heidelberg, 2010, pp. 92–103. doi:10.1007/978-3-642-12186-9_10.

[14] S. Leemans, D. Fahland, W. Aalst, van der, Process and deviation exploration with inductive visual miner, in: L. Limonad, B. Weber (Eds.), BPM Demo Sessions 2014 (co-located with BPM 2014, Eindhoven, The Netherlands, September 20, 2014), CEUR Workshop Proceedings, CEUR-WS.org, 2014, pp. 46–50.

[15] T. Rotter, J. Kugler, R. Koch, H. Gothe, S. Twork, J. M. van Oostrum, E. W. Steyerberg, A systematic review and meta-analysis of the effects of clinical pathways on length of stay, hospital costs and patient outcomes, BMC Health Services Research 2008 8 (265) (2008) 1–15. doi:10.1186/1472-6963-8-265.

[16] S. D. Pearson, S. F. Kleefield, J. R. Soukop, E. Cook, T. H. Lee, Critical pathways intervention to reduce length of hospital stay, The American Journal of Medicinedoi:10.1016/S0002-9343(00)00705-1.

[17] S. Goodman, A dirty dozen: Twelve p-value misconceptions, Seminars in Hematology 45 (3). doi:10.1053/j.seminhematol.2008.04.003.

[18] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature 521 (7553) (2015) 452. doi:10.1038/nature14541.

[19] G. Dong, H. Liu (Eds.), Feature engineering for machine learning and data analytics, Taylor & Francis, Boca Raton, FL, 2018.

[20] S. J. Leemans, Inductive visual miner, http://www.leemans.ch/publications/ivm.pdf (2017).

[21] C. Polanczyk, E. Marcantonio, L. Goldman, L. Rohde, J. Orav, C. M Mangione, T. Lee, Impact of age on perioperative complications and length of stay in patients undergoing noncardiac surgery, Annals of internal medicine 134 (2001) 637–43. doi:10.7326/0003-4819-134-8-200104170-00008.

[22] A. Vehtari, A. Gelman, Gabry, Practical bayesian model evaluation using leave-one-out cross-validation and waic, Statistics and Computing 27 (5) (2017) 1413–1432. doi:10.1007/s11222-016-9696-4.

[23] J. Salvatier, T. V. Wiecki, C. Fonnesbeck, Probabilistic programming in Python using PyMC3, PeerJ Computer Science 2 (e55) (2016) 1–24. doi:10.7717/peerj-cs.55.

Appendix A. Cholecystitis Case Study

Appendix A.1. Cholecystitis Pathway Variants

This section shows the cholecystitis pathway variant plot auto-generated by ProM. Cholecystitis pathway variants are analyzed without activities from the 'antibiotics' sub-process and the 'monitoring labs' sub-process (see Fig. A.10 for the activities from the two sub-processes). Clinicians confirmed that these sub-processes are standard monitoring and maintenance systems while the patient is waiting for further diagnosis. Only analyzing activities from the primary cholecystitis pathway significantly reduces the level of clinical variation between patient traces.

The cholecystitis pathway model consists of 10 pathway variants. The 10 pathway variants of the cholecystitis pathway model are shown in Fig. A.9, and the number of patient traces that follow each pathway variant are listed in Table A.4. Pathway variants from Fig. A.9 are sorted from the most common pathway (index 0) to the least common pathway (index 9). The most common pathway variant (index 0) consists of anesthesia, surgery, and surgical pathology lab. The second pathway variant (index 1) includes surgery without anesthesia because of faulty clinical data.

Appendix A.2. Cholecystitis Pathway Model

The cholecystitis pathway model visualized by `'Inductive Visual Miner'` incorporates activities from the 'antibiotics' sub-process and the 'monitoring labs' sub-process. A breakdown of the cholecystitis pathway model into one primary pathway and two concurrent sub-processes is shown in Fig. A.10, and the model notations are summarized in Table 2. The first pathway model in Fig. A.10 is the primary pathway, followed by the 'antibiotics' sub-process and the 'monitoring labs' sub-process. Patient traces can execute any combination of the two sub-processes concurrently with the primary pathway. The eight patient traces that follow the second pathway variant (index 1) do not conform to this pathway model because of faulty clinical data. Based on this model, pre-operation haematology and chemistry labs tend to span the entire pre-operation process, while pre-operation antibiotics are taken closer to surgery.

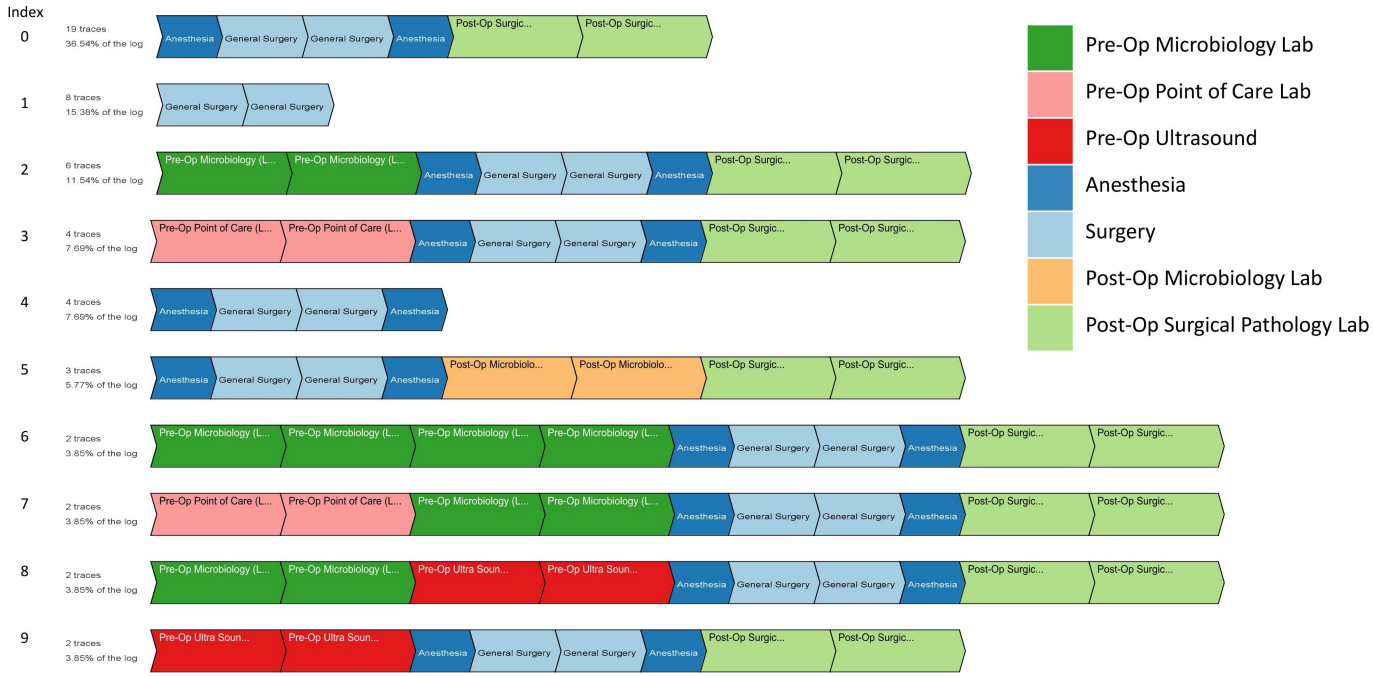

Figure A.9: Cholecystitis pathway variant plot auto-generated by ProM's plugin `Explore Event Log`. For the purpose of readability the legend was added and the statistics on the left are repeated in Tab. A.4. The top three variants account for approximately 63% of the patient traces.

Table A.4: Statistics of cholecystitis patient traces shown in Fig. A.9.

Variant	0	1	2	3	4
Patients	19	8	6	4	4
Percentage	36.54%	15.38%	11.54%	7.69%	7.69%
Variant	5	6	7	8	9
Patients	3	2	2	2	2
Percentage	5.77%	3.85%	3.85%	3.85%	3.85%

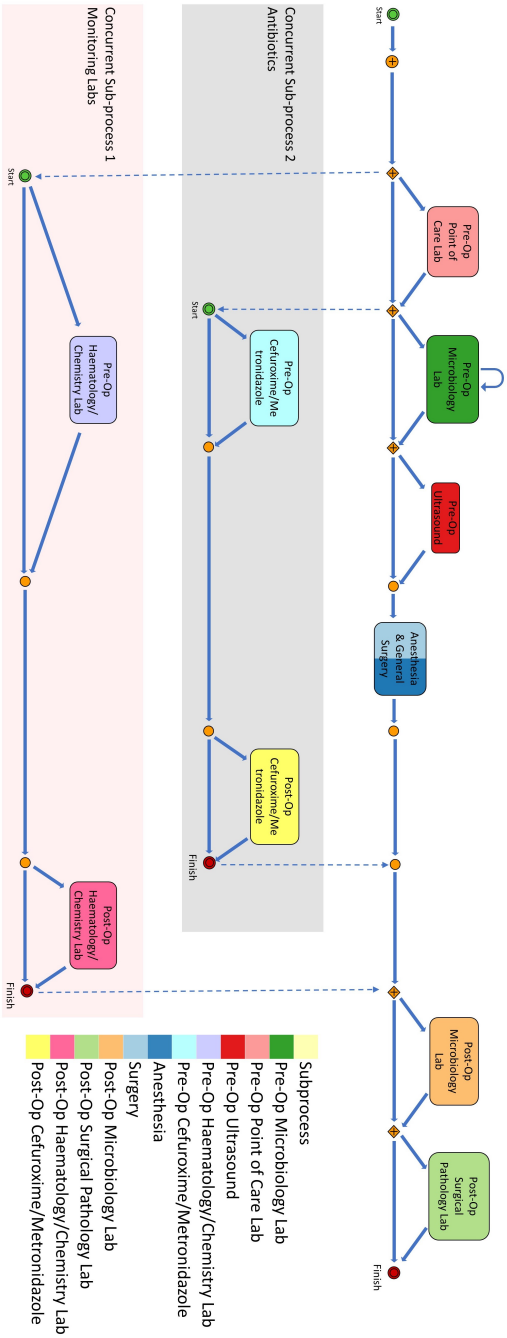

Figure A.10: Cholecystitis pathway model using nomenclature of Tab. 2. The model is broken down into one primary pathway and two sub-processes.