

# EHR Data Pre-processing Facilitating Process Mining: an Application to Chronic Kidney Disease

Vojtech Huser, MD, PhD<sup>1,2</sup>, Justin Starren, MD, PhD<sup>1</sup>

<sup>1</sup>Marshfield Clinic, Marshfield, WI; <sup>2</sup>Morgridge Institute for Research, Madison, WI

## Abstract

*Process mining can be a useful visualization approach for analyzing EHR data. We present our experiments with data preparation techniques to improve process mining results on a large cohort of chronic kidney disease patients.*

## Introduction

Healthcare processes are notoriously complex and difficult to model. Process mining (PM) is one approach for analyzing large volumes of event data which can convert them into visual process definitions. Previous reports<sup>1</sup> on use of PM with healthcare data were limited in terms of origin of logs (simulated data), size of datasets (hundreds of patients) and scope (data about a single medical episode originating from a single medical department). We address some of these limitations with our analysis of lifetime data of a large cohort of chronic kidney disease patients and specifically focus on necessary pre-processing of raw EHR data. This poster will also provide general introduction to PM and demonstrate its application to studying progression of chronic kidney disease.

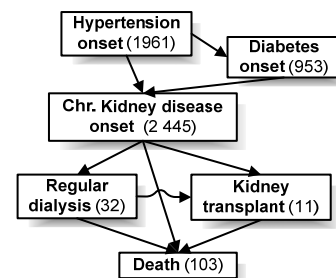
## Method

Lifetime EHR data of chronic kidney disease patients were extracted from data warehouse of Marshfield Clinic. A complete set of coded EHR data (such as visit data, laboratory results, diagnoses and procedures) was processed with our previously developed analytical package called RetroGuide<sup>2</sup> with the goal to test a range of healthcare-specific data abstraction and data filtering transformations which are not available directly in PM tools. This filtered dataset was then converted into a standard MXML process log format and loaded into an open-source PM tool called ProM<sup>2</sup> where a number of data mining, analytical and conversion algorithms were applied. The resulting graphical process models or other visualizations were evaluated for displaying clinically meaningful content and their ability to support side-by side comparison of different sub-populations of CKD patients (e.g., different locations, insurance coverage, or co-morbidities).

## Preliminary results

We have extracted 2.1 GB (15 million events) of coded EHR data on a cohort of 2445 CKD patients.

Within RetroGuide, we created a set of CKD-specific data pre-processing abstractions. For example, identification of disease onset dates or key treatment procedures, and determination of disease stages or level of control using coded laboratory, clinical, or aggregate data (e.g., eGFR, BP, HbA1c, specialty specific visit frequencies, or dialysis frequency). The following ProM mining algorithms were used: heuristic miner (example shown in Figure 1), alpha algorithm, association rules mining and grouping log conversion. Additional details about our methodology and results can be found on our project website<sup>2</sup>.



**Figure 1.** Simplified example of a mined process showing a subset of high-level disease milestones and the number of satisfying process instances.

## Discussion

Existing process mining tools and techniques do not produce clinically meaningful visualizations when applied to raw EHR data. We tested several data pre-processing transformations and demonstrate their usefulness for analysis of EHR data. PM can be a valid alternative to authoring process definitions within a workflow editor. This work on PM is part of our larger project focusing on use of a Workflow Management System within an EHR system<sup>2</sup>.

## References

1. Mans R, Schonenberg H, Leonardi G, Panzarasa S, Cavallini A, Quaglini S, van der Aalst W Process mining techniques: an application to stroke care, *Stud Health Technol Inform* 2008;136: 573-8.
2. Healthcare Workflow <http://healthcareworkflow.wordpress.com> [accessed: Mar 4, 2009]

