# *Portinari*: A Data Exploration Tool to Personalize Cervical Cancer Screening

Sagar Sen\*, Manoel Horta Ribeiro<sup>†</sup>, Raquel C. de Melo Minardi<sup>†</sup>, Wagner Meira Jr.<sup>†</sup> and Mari Nygård<sup>‡</sup>

\* Certus V&V Center and HPV Research Group,

Simula Research Laboratory and Cancer Registry of Norway, Oslo, Norway,

Email: sagar@simula.no

<sup>†</sup> Department of Computer Science,
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil,
Email: {manoelribeiro, meira, raquelcm}@dcc.ufmg.br
<sup>‡</sup> HPV Research Group,
Cancer Registry of Norway, Oslo, Norway,
Email: mari.nygard@kreftregisteret.no

Abstract-Socio-technical systems play an important role in public health screening programs to prevent cancer. Cervical cancer incidence has significantly decreased in countries that developed systems for organized screening engaging medical practitioners, laboratories and patients. The system automatically identifies individuals at risk of developing the disease and invites them for a screening exam or a follow-up exam conducted by medical professionals. A triage algorithm in the system aims to reduce unnecessary screening exams for individuals at low-risk while detecting and treating individuals at high-risk. Despite the general success of screening, the triage algorithm is a one-sizefits all approach that is not personalized to a patient. This can easily be observed in historical data from screening exams. Often patients rely on personal factors to determine that they are either at high risk or not at risk at all and take action at their own discretion. Can exploring patient trajectories help hypothesize personal factors leading to their decisions? We present Portinari, a data exploration tool to query and visualize future trajectories of patients who have undergone a specific sequence of screening exams. The web-based tool contains (a) a visual query interface (b) a backend graph database of events in patients' lives (c) trajectory visualization using sankey diagrams. We use Portinari to explore diverse trajectories of patients following the Norwegian triage algorithm. The trajectories demonstrated variable degrees of adherence to the triage algorithm and allowed epidemiologists to hypothesize about the possible causes.

*Index Terms*—cervical cancer screening, socio-technical system, data exploration, graph databases, knowledge discovery, Portinari, interactive visualization

## I. INTRODUCTION

Software systems in several countries have been deployed to facilitate the prevention of cancer in society. For instance, at the Cancer Registry of Norway, a software system guides eligible women in the age group between 25 and 69 years to attend cervical cancer screening exams. They take tests commonly known as the *Pap smear* at three year intervals or go through follow-up exams in case of a higher risk. Women are sent invitation and reminder letters to their postbox or via *digipost*<sup>1</sup>. Many women attend screening exams and follow up exams and this has reduced deaths due to cervical cancer

1https://www.digipost.no/

by up to 80% [1]. Outcomes of all screening exams are sent back to the cancer registry and registered in a relational database. This feedback loop of the software system with the society manifests itself into a *socio-technical system*. The reliability of the socio-technical system greatly depends not only on how bugfree the software system is but also on how well we are able to use it to mobilize and engage people (medical practitioners, nurses, epidemiologists, and patients in the general public) into taking action against cervical cancer while keeping health expenditure to a minimum.

The main challenge of a socio-technical system for cervical cancer screening is to simultaneously: (a) Minimize overscreening of patients due to its possible harmful effects [2] [3] (b) Increase screening attendance of under-screened patients to reduce cancer burden on the state or costs incurred from cancer treatment [4] [5]. A triage algorithm on the population aims to optimize screening by inviting a patient to attend a screening exam [6] [7]. The triage algorithm is implemented to automatically identify eligible patients in the population registry of a country and send invitations and reminders to them for screening. For instance, the triage algorithm in Norway<sup>2</sup>, shown in Figure 1, automatically invites women who turn 25 or immigrate into Norway at a later age to a cytology exam. The algorithm also invites women to a follow-up considering the outcome of the previous exam of a patient as detected in the anonymized cervical cancer screening database at the Cancer Registry of Norway. For instance, if a woman is tested with cytology ASC-US/LSIL she is immediately invited to take a test for the humanpappilomavirus (HPV). This is because most cases of cervical cancer are due to a persistent infection of a high-risk variant of the human papillomavirus (HPV) [8]. If the HPV test turns out to be positive, the patient is invited to a new cytology and HPV test in 6 to 12 months. This is because the human immune system can often eliminate the virus on its own. If the HPV infection persists the patient undergoes a colposcopy which can have harmful effects due to the biopsy.

<sup>2</sup>https://www.kreftregisteret.no/globalassets/gammelt/cervix/flytskjema.pdf

The screening guidelines have been very successful, reducing the incidence of cervical cancer by 70% [9]. The number of cervical cancer cases per year is around 300 instead of an estimated 1000 if the screening program did not exist.

Despite the general success of the screening program we observe in an anonymized cervical cancer screening database that people take or do not take tests based on their perceived risk of being diagnosed with cervical cancer. Low risk may be perceived due to many personal factors such as a healthy lifestyle and taking precautions against contamination. Fear and additional symptoms (which could be bleeding) not detected by screening exams can contribute to the perception of higher risk. Those who perceive their risk to be high may take too many tests which may have harmful effects. Those who perceive their risk to be low perhaps due to lifestyle factors may take very few or no screening exams at all. Therefore we ask, can exploring patient trajectories give us clues towards personalizing cervical cancer screening? We believe that software tools to explore peoples' trajectories can help improve the social objective in a socio-technical system. In our case, the social objective amounts to personalizing cervical cancer screening and consequently reducing number of cancer cases.

**Approach:** To personalize cervical cancer screening, we exploit the screening outcomes of patients over a period of 20 years to understand if screening can be personalized. We achieve this in two steps:

- 1) Data from a socio-technical system for cervical cancer is available in the form of events in the life of patients. In cervical cancer screening an event corresponds to attendance to an exam such as a HPV test along with date and type of diagnosis. We transform these events from its flat form of transaction records into sequences of connected events for individual patients in a *graph database* implemented in Neo4J [10].
- 2) The main contribution of the paper is a web-based data exploration tool Portinari<sup>3</sup>, to explore and visualize individual<sup>4</sup> trajectories by querying the graph database. The interface of the tool at a glance is shown in Figure 2. The interface allows users (epidemiologist specializing in cervical cancer screening) to visually specify a query graph representing a sequence of exams and respective diagnoses taken by an individual. The user may also specify constraints such as the time between any two diagnosis and/or age ranges. Portinari automatically generates future trajectories of patients who underwent the input sequence of exams and diagnosis by matching similar patients in the graph database. Portinari visualizes the outcome as a sankey diagram [11] of a patient trajectory for a finite number of subsequent steps using the query subgraph as the origin. Sankey diagrams are a specific type of flow diagram, in which the width of the arrows is shown proportionally to the flow quantity. The

flow quantity in this paper typically represents number of individuals going from one exam to another.

We use Portinari to explore a graph database of screening exams from the Norwegian Cervical Cancer Screening program. The large anonymized database contains high quality data events from about 906713 distinct women in Norway, such as the date and results of various types of exams related to cervical cancer screening [12] [13]. Portinari was able to assess many of the recommendations of the Norwegian triage algorithm, and also find trajectories where the guidelines could use some further personalization. For instance, an interesting finding was that there were several patients who returned within 30 days of having normal cytology. Few of them got a more advanced test such as the histology done while some others were diagnosed with HPV positive. This was in contrast to the triage algorithm's recommendation to return in three years. This result clearly shows that the cytology exam is not perfect and there are other symptoms such as bleeding that are not taken into account to personalize recommendations. Tools for data exploration such as Portinari are a necessity to inspect and validate rapidly evolving socio-technical software systems, particularly in public health. Data is not static in these systems and is constantly updated by societal behavior. Policy makers should consider outliers and personalized scenarios to update their triage algorithms to enhance prevention.

The rest of the paper is organized as follows. In Section II, we present the graph database representation of the events from screening exams for a population. The data exploration system, *Portinari*, is described in Section III. We made explorations inspired by the Norwegian triage algorithm using *Portinari* and the results and its implications are presented in Section IV. Related work is presented in Section V. Finally, we present conclusions and future work in Section VI.

# II. GRAPH DATABASE OF SCREENING EVENTS

Events from societal behavior, such as people attending exams, are often stored by information systems as unrelated records of transactions. In cervical cancer screening, a screening exam and its diagnosis for a patient is such a record in a relational database. Creating temporal queries in relational databases is complex [14] [15]. Therefore, we advocate the representation of the temporal ordering using a graph database.

A patient's trajectory is in fact is these records put together in a chronological sequence. This is ideally modeled as a *directed property graph* where nodes represent an exam event along with properties such as diagnosis type and stage of disease and edges represent the *sequence between exams* with properties such as time between exams. A graph database representation of exam records allows querying and reasoning about patient trajectories as a whole. Below we describe both a tabular representation of the patients attending a screening exam and the transformation to a graph database representation.

**Tabular Representation:** The cervical cancer screening database in Norway contains high quality events related to cervical cancer screening exams, dates and results of screening

<sup>&</sup>lt;sup>3</sup>Available upon request

<sup>&</sup>lt;sup>4</sup>We use the terms *patient* and *individual* interchangeably in the article.



Fig. 1: Triage algorithm for cervical cancer screening in Norway.



Fig. 2: *Portinari's* overall interface: (a) A drag and drop canvas to create a query graph where nodes represent *events* and edges representing *order of events* (b) Constraints created on properties of nodes and edges in the query graph (c) Constraint creation form to choose property and specify a constraint (d) A query form to specify a property to observe in events (e.g. diagnosis) of trajectories, number of events to show in the trajectory (e.g. 2) for patients satisfying the query graph (e), (f) Upto to two interactive Sankey diagrams displaying trajectories for the chosen event property (e.g. diagnosis).

and diagnostic tests performed, as well as information about treatment of pre-cancers. It is estimated to be close to 100% complete as it is mandatory by law to report all screening tests, cancers, emigration and deaths [12] [13]. The database contains about records of 5 million exams taken by 0.9 million women from the year 1992 to 2014. Each record in the database is specified by a set of fields as presented in

Table I. We anonymize the original database before running all the experiments in this article. The anonymization was done by replacing all dates (birth date, diagnosis date, and censor date) to the 15th of the month, and perturbing every month by fuzzy factor of number between +/- 4. The resulting database had a very low re-identification risk as evaluated by the anonymization tool ARX [16] while preserving the general

TABLE I: Fields in the cervical cancer screening database.

ID	Internal numeric identifier.
birthdate	Month and year of birth of the woman.
diagnosisdate	Date of diagnosis.
type	Type of visit/exam.
diagnosis	Result given by the exam/visit.
stage	Stage for cancer diagnosis.
lab_nr	Code for the laboratory, which is a 5-digit number.
region	Norwegian health care region.
censordate	Date for emigration/death/cervical cancer diagnosis.



Fig. 3: Transformation of (a) tabular data for a patient with a given ID to (b) graph for the patient.

## characteristics of screening behaviour.

Graph Database Representation: Anonymized tabular data of exam records for each patient is transformed to a graph database as illustrated in Figure 3. Exam records are processed in chronological sequence to create nodes representing events such as exams with properties such as diagnosis type, stage, laboratory number. The edges, labeled next, contain the time elapsed between two events stored as a property. In this paper we specify time in number of days. We create additional edges between the nodes that are not consecutive but two or more events in the future as shown in Figure 3. This inclusion of edges between hoping over nodes in time facilitates creation and faster execution of temporal queries between two nonconsecutive events. For instance, a query that asks to return all patients who got cancer in 400 days from their first cytology positive simply will require matching a next edge with number of days < 400 days. This saves computation time by avoiding matching all events and edges.

Every patient's trajectory is created and stored as a graph in the database implemented in Neo4J. An example Cypher statement to create a patient's trajectory is shown in Listing 1. The statement is generated from the table for the individual shown in Figure 3(a). We first create nodes p1, p2, p3, and p4 that correspond to the exams taken by the patient, then the *1*-*hop* edges with the time between consecutive exams, then the *2*-*hop* edges and the *3*-*hop* edge. The database does not contain relationships between patients and hence the set of subgraphs of patients are unconnected. In the future, if relationships such as friendship or family are made available by approval from an ethical committee we can easily create edges in the schema-less graph database with this information. This is the main advantage of using a graph database. Similarly, genetic information from patient's blood serum in the Janus bank [17] of the Cancer Registry of Norway and lifestyle factors such as smoking, alcohol consumption obtained from a survey [18] can be introduced into the graph database to enrich it for more insight into societal processes.

CREATE
(p1:Event // Create first node
{PatientID:1, DiagnosisDate:"15.05.1992", BirthDate:"12.08.1960", ExamType:" cyt",
Diagnosis:13, MorphologyCode:76700, CDate:NULL, LaboratoryNbr:19, Region:2}),
(p2:Event // Create second node
{PatientID:1, DiagnosisDate:"15.09.1992", BirthDate:"12.08.1960", ExamType:" cyt",
Diagnosis:12, MorphologyCode:69000, CDate:NULL, LaboratoryNbr:19, Region:2}),
(p3:Event // Create third node
{PatientID:1, DiagnosisDate:"15.11.1994", BirthDate:"12.08.1960", ExamType:" cyt",
Diagnosis:20, MorphologyCode:76700, CDate:NULL, LaboratoryNbr:19, Region:2}),
(p4:Event // Create third node
{PatientID:1, DiagnosisDate:"15.03.1995", BirthDate:"12.08.1960", ExamType:" cyt",
Diagnosis:12, MorphologyCode:76700, CDate:NULL, LaboratoryNbr:19, Region:2}),
// 1-hop edges
$(p1) - [:Next1 {SinceLast:123}] - >(p2),$
$(p2) - [:Next1 {SinceLast:791}] -> (p3),$
$(p3) - [:Next1 {SinceLast:116}] - >(p4),$
// 2-hop edges
$(p1) - [:Next2 {SinceLast:914}] - >(p3), (p2) - [:Next2 {SinceLast:807}] - >(p4),$
// 3-hop edges
(p1) -[:Next3 {SinceLast:1030}] ->(p4)

Listing 1: An example Cypher query generated to create a patient trajectory in a graph database.

# **III. DATA EXPLORATION SYSTEM**

*Portinari* is a data exploration system with the aim of exploring and possibly validating how software systems impact social behavior. Socio-technical systems can comprise of software systems in a feedback loop with the society where data about peoples' behavior is gathered over time. Portinari is a tool that allows inspection, exploration, and visualization of peoples' trajectories in a socio-technical system. It is a generic tool that can be used to query and visualize trajectories of events in any socio-technical system. Data in socio-technical systems evolve over time and hence need to be constantly explored to gain insight into the effectiveness of an organized social process such as the triage algorithm in a cervical cancer screening program. Portinari was developed in light of such online monitoring of societal behavior. In this article, we use Portinari to explore trajectories of patients in the Norwegian cervical cancer screening program. In the following subsections, we present an overview of Portinari and describe the two principal steps in using the tool.

## A. System Overview

*Portinari* relies on a graph database of individuals' trajectories. For instance, we use the graph database of people attending to cervical cancer screening as described earlier in Section II. It expects the graph database to contain nodes representing events and directed edges representing the chronological order between between events for each individual. The graph database is a collection of directed graphs of people's trajectories. The graph database is queried by *Portinari*'s webbased interface as shown in Figure 2. The interface contains two parts:

Querying Interface: Portinari has a web-based querying interface (Figure 2(a-d)) that permits modeling a *query graph* representing a partial trajectory of an individual along with constraints. The length of the partial trajectory has an upper bound of the maximum number of exams a woman has ever taken (54 at present). Epidemiologists, explore finite sequences from the triage algorithm graph limited to two or three nodes representing exams. The query graph is transformed into query in the language Cypher [19] (a language similar to SQL but for graph databases) that is executed on the backend graph database. The result of the query is the set of all future trajectories of patients that matched the input partial trajectory. **Results Interface:** The result (Figure 2(e-f)) is visualized as an interactive sankey diagram [20]. Sankey diagrams are a specific type of flow diagram, in which the width of the arrows is shown proportionally to the flow quantity. We use sankey diagrams to show how many people flow from an event to other events. Portinari allows visualization of two sankey diagrams simultaneously, enabling the user to compare two scenarios.

## B. Creating a Query Graph

The first step in using *Portinari* is to create a *query graph*. *Portinari*'s query mechanism has a drag-and-drop interface where the user builds the query graph representing *sequence of events* in a patient's life to be found in the graph database. The user may: (a) Add new nodes representing events (b) Create edges between nodes representing the sequence of events (c) Delete nodes. In the cervical cancer screening context, for instance, a node is an exam taken by a patient and the diagnosis she received. The edges represent a chronological ordering between two exams.

**Constraints:** The query interface allows to add constraints such as "*BIGGER THAN*", "*SMALLER THAN*", "*EQUALS TO*" and "*NOT EQUALS TO*" to properties in both nodes and edges. A specific node will be matched in the graph database if it satisfies *all the associated constraints* on the node. Constraints on edges allow for instance the specification of a time range between two events. Constraints are specified in the *constraints form* as shown earlier in Figure 2(c). The constraints on a node are automatically rendered in query graph right below nodes and edges. They are also shown on top of the *constraints form* (Figure 2(b)).

**Multiple Partial Trajectories:** The user may create more than one partial trajectory in the query graph. The query system considers partial trajectories such as in Figure 4(a) as patients that match both the trajectory n0 - e2 - n1 and n3 - e4 - n1. Here n0, n1, n3 are event nodes and e2, e4 are edges between events in the multiple path partial trajectory.

**Query form:** The query form is shown in Figure 2(d). In this final step the user specifies additional parameters to generate a sankey diagram using the *query graph* as input. The user may



Fig. 4: (*a*) A query graph in *Portinari* is visually specified where nodes represent events and edges represent order between events. Constraints on nodes and edges can be specified and appear below them (*b*) *Portinari* generates a sankey diagram in its results interface, displaying the trajectory of all individual that satisfy the graph query.

specify the following parameters: (a) Number of subsequent events the user wants to visualize considering all patients satisfying the query graph as origin. (b) The timeline of a chosen attribute the user wants to visualize in the trajectory rendered as the Sankey diagram. (c) The time range for next event considering the query graph as origin.

**Example of use:** Consider the scenario where a user wants to visualize the diagnosis of patients that had a specific pattern of exams and returned for an exam within 60 days. Lets assume that the cohort of interest is composed of patients who were diagnosed *Normal* or *Unsatisfactory* in the cytology exam (codes 11 and 10 respectively), and then, in less than a year, were diagnosed with HPV positive (code 0) in any region. An user employs the drag-and-drop interface to create a query graph with two possible paths and constraints as shown in Figure 4(*a*). The constraints on the variable diagnosis for the events are shown under the node. While the constraint on the time < 365 days is shown under the edges. The Cypher query

generated by *Portinari* from the visual interface is shown in Listing 2. The nodes n0, n1, n3 and edges e2, e4 in the query correspond to the query graph in Figure 4(a). The node nf1 and the edge ef1 are used as placeholders to obtain the first exam taken by after patient pattern specified in the query graph.

MATCH ( n0:Event { Diagnosis:11 })
MATCH ( n1: Event { Diagnosis:0 })
MATCH ( n3:Event { Diagnosis:10 })
<b>MATCH</b> $(n0) - [e_2] - (n_1)$
WHERE e2. SinceLast < 365
<b>MATCH</b> $(n1) - [e4] - (n3)$
WHERE e3. SinceLast < 365
MATCH $(n3)-[ef1:Next1]->(nf1)$
WHERE ef1. SinceLast < 60
RETURN COUNT(distinct n3), nfl.Diagnosis

Listing 2: An Example Cypher query generated to create a patient trajectory in a graph database.

Finally, the user specifies the number of steps in future and the attribute he/she wants to visualize in the *query form*. The user also can specify the time range for the immediate next event for all patients matching the query graph. As mentioned previously, the user wants the next diagnosis for patients that returned in more than 60 days. Therefore, the user can specify this constraint in the *query form* and submit the query to obtain the results. We present how the results are rendered in the following subsection.

# C. Generating Patient Trajectories as Sankey Diagrams

Portinari presents results as an interactive sankey diagram of the subsequent patient trajectories who match the query graph. The sankey diagram as shown in Figure 4(b) consists of vertical bars and flows between the bars. The leftmost vertical bar in the sankey diagram is called the Origin. It represents the total number of patients who match the query graph specified in the query interface. The subsequent bars in the sankey diagram represent the number of patients for different values of a chosen property of an event. In Figure 4(b), the different values correspond to the property diagnosis of the event screening exam. We are typically interested in the trajectory of diagnosis for patients. The sankey diagram is interactive. Hovering the mouse cursor over the nodes displays a tool-tip informing how many patients converge to that event and their percentage in the population. Flow width between bars in the sankey diagram is directly proportional to the number of patients that took this specific path. Hovering the cursor over the flows displays a tool-tip informing the number of patients that went from the event represented by the source of the flow to the event represented by the target of the flow. The tool-tip also displays what percentage of patients the flow represents in the source node. The width of the flow gives the user a perception of how likely people are to follow one path compared to another. For instance, the path to Cytology:Normal is far more likely than the one to Cancer in Figure 4(b). Portinari can render up to two sankey diagrams allowing the comparison of two different query graphs.

# IV. EXPLORATIONS

We hope that exploring the graph database of patient trajectories will give us insight into personalizing the cervical cancer screening program in Norway. We use *Portinari* to address the following broad questions:

**Q1:** What happens to patients who follow the triage algorithm? **Q2:** What happens to patients who *do not follow* guidelines in the triage algorithm?

**Q3:** Can we compare different trajectories taken by patients in the triage algorithm?

Question **Q1** is essential to understand whether guidelines in a public health program are effective. In the triage algorithm shown earlier in Figure 1, the most commonly followed guideline is simple (shown in bold) for women to understand and follow. Women are asked to perform a cytology exam when they enter Norway and are above 25 years of age or when they turn 25. If the result of the test is normal then they are asked to come back every three years for regular exams. Therefore, we model the query graph for such a scenario in Figure 5(a), where we are interested in the trajectories taken by patients who have two consecutive normal cytology exams. We use Portinari to find out what happens to these women if they come back within three and four years after two normal cytologies. In Figure 5, we see that about 88.2% of the women were still diagnosed with cytology normal while some had more tests done. We also observe 39 women (about 0.01%) who had a cancer (squamous cell carcinoma). These outliers despite following the triage algorithm, were still diagnosed with cancer. How could we have better screened these 39 women? Should they have undergone other tests in addition to the cytology exam? These are questions that exploration using Portinari can raise.

Question **Q2** concerns women who do not adhere to the triage algorithm. In Figure 5(c), we explore trajectories of women who return only 30 days after two consecutive normal trajectories. There were about 4337 women who returned only in 30 days, some of who got different tests such as HPV testing and histology. Some were diagnosed with HPV+ve while the histology revealed a *polyp* in others. A polyp is an abnormal growth of tissue projecting from a mucous membrane. The most likely hypothesis is that the women or their doctors perceived a higher risk probably based on additional evidence such as bleeding that is not revealed by a cytology exam. These deviations from what the triage algorithm expects strengthens the case for personalization and improvements in testing and evidence collection.

Finally, for question **Q3**, we compare two future trajectories for one query graph. In Figure 6, we construct a query graph illustrating the scenario where a cytology normal exam is followed by a HPV exam that turns out to be positive. We compare trajectories for women who return within one year as recommended by the triage algorithm in Figure 6(b) and women who return between one and two years or who are delayed beyond 2 years in Figure 6(c). We observe that majority of women that had HPV+ve test return within a year. This illustrates that they perceive their risk to be high when they have a positive test for the humanpapillomavirus. HPV sounds like HIV and does it somehow increase the gravity of the disease? Both trajectories show that few women have a



Fig. 5: (a) Query graph for women with two consecutive normal cytologies (b) Trajectory of women who come back in 3 to 4 years after two normal cytology exams (c) Trajectory of women with two consecutive normal cytologies but return in just 30 days.

persistent HPV infection. Which means that those who were diagnosed with HPV+ve first were not diagnosed a second time with HPV+ve. Many have a normal cytology diagnosis that follows the HPV+ve test. These trajectories explain the natural history of the HPV virus in the human body. Even the high risk variants of HPV are eliminated by the immune system in most cases. Comparison of two similar trajectories allows users to evaluate the trade-offs between coming on time as recommended by the triage algorithm or being delayed with respect to the screening algorithm. For policy makers and epidemiologists the comparison tool allows augmentation of the triage algorithm with personalized suggestions if necessary. For instance, people with weaker immune systems may need to be screened more frequently and asked to adjust their lifestyle compared to those who are in good health over a threshold number of screening exams.

# V. RELATED WORK

*Portinari* is a generic exploration tool for socio-technical systems which we use to facilitate prevention of cancer based on interactive visualization of the societal process of cervical cancer screening. It derives inspiration from prior work on interactive tools for visualizing electronic health records (EHR) and querying temporal data in the medical context. In the following section we review prior work on visualization of

EHRs, querying of temporal data and data-driven approaches in the medical context and explain how *Portinari* is positioned with respect to what has been developed.

#### A. Interactively Visualizing EHRs

There has been extensive research on the development of interactive visualization tools to explore and query EHRs [21] [22] [23] [24] [25]. They aim to aid in medical decision making, and to translate the costs of having an EHR into actual benefits for physicians and patients [26]. Some representative systems of this approach are:

*LifeLines* [21], which was developed in the late 1990s and allowed the visualization of a *single patient record*. The tool visualizes the timeline of cases, placements, assignments, and reviews for a patient. *Portinari* in contrast aggregates patients who are similar and shows trajectories over time or events for a given variable such as diagnosis. The width of the flow in a trajectory allows rapid comparison of two trajectories at the societal level instead of an individual. Aggregation also means that individual identity will not be revealed when it comes to communicating paths taken by patients.

*LifeFlow* [22], which uses a visualization inspired in an Icicle Tree to simultaneously display multiple event sequences. This is again hierarchical visualization of event sequences in the life of a unique patient. The complex visualization in the tool



Fig. 6: (a) Query graph for women who have a cytology normal and they take a HPV test and turn positive in the next exam (b) Trajectory of women who come back in one year after the query graph (c) Trajectory of women who come back after two years.

does not allow exploration of trajectories of multiple patients in one visualization. *LifeFlow* is useful when a patient has several exams and visits to a doctor and the health records need to be organized. It however, unlike *Portinari*, does not visualize societal behaviour.

*EventFlow* [23], which implemented a simple query language that allows users to specify intervals between two or more events and visualize amount of flow between them. *Portinari* goes one step further and visualizes alternative trajectories starting with an event sequence as an input.

*OutFlow* [24], which aggregates multiple event sequences into pathways, and allows users to explore external factors that influence specific event transitions in a pathway. *OutFlow* also allows clustering of events of high risk together. *OutFlow* and *Portinari* are very similar in terms of user experience and ease of learning. However, one of the fundamental differences between both is the query system allowing the specification of constraints on both temporal and non-temporal attributes giving higher control over the visualization of trajectories. Moreover, *Portinari* is open source and available upon request unlike IBM's Outflow which is proprietary.

*DecisionFlow* [25], which allows the analysis of highdimensional temporal event sequence data. It is a commercial tool that takes *OutFlow* one step further to make it a fullfledged tool for ad-hoc statistics and epidemiology. *Portinari* aims to be open-source alternative that integrates several statistical analysis features of *DecisionFlow* along with data mining algorithms to find explanations for patterns seen in patient trajectories.

Rind et al. present a comprehensive survey where some of

the above interactive visualization systems are compared and evaluated [26].

# B. Querying Temporal Data

Research has also been done to develop tools and methods to allow domain experts to perform queries in temporal data, which is often too technical to be done in traditional query languages such as SQL [14]. A strong motivation in this context is to allow domain experts to create cohorts, groups of individuals with common features. A representative system built for this purpose is COQUITO [14], a visual interface that assists the users to define cohorts with temporal constraints, giving information on the filtered population as the restrictions are added. Systems that aim to make temporal querying userfriendly include: (a) PatternFinder [27], where users can formulate queries on patient event histories with connected boxes. (b) QueryMarvel [28], where a comic strip metaphor is used in order to make the querying system more easy and fun to use. (c) DataPlay [29], which allows for a interactive trial-and-error query specification (d) DataMeadow [30], which provides the constructions of interactive queries through starplots. Textual temporal query languages such as T-SQL [31] and TQuel [15] are also relevant in our context.

As pointed by Krause *et al.* [14], some of aforementioned systems allow users begin with a overview of the health data, and then filter towards a desired pattern. We may refer to this as a *pattern recognition*) based approach where a user has a bird's eye view of the data and then he/she drills down in to patterns of interest.

*Portinari*'s, leverages the paradigm of *pattern specification* instead of *pattern recognition*, by allowing specification of queries of temporal sequence of events along with personal attributes as constraints instead of filtering to a cohort from a overview of the complex dataset. Textual temporal query languages along with tools such as *QueryMarvel*, *DataPlay*, *DataMeadow* also fall in the category of pattern specification approach.

# C. Data-Driven Methods in Medicine

Our work can also be set into a larger context of applying new data-driven computational methods and tools in medicine [32], particularly to public health. Data-driven approaches are used to analyze data, create predictive models, help clinicians take decisions and provide more personalized diagnosis and treatment to patients [32] [33] [34] [35]. Within this paradigm, unlike in traditional medical research, the data is not purposely collected to test specific questions, but obtained from legacy databases or data warehouses and used for secondary analysis [32].

Some representative work that fits into this scenario include the creation of disease trajectories for better understanding the correlation between diseases and their progression [36], and the system proposed by Zamora *et al.* to discover and analyse co-occurrence of diagnostics, intervantions and prescriptions [37]. *PARAMO* [38], is a platform developed by Ng *et al.* to simplify the process of building predictive models from *EHRs*.

*Portinari* can be set in the broader context of incorporating data-driven methods into exploration and verification of peoples' behaviour in socio-technical systems. Portinari tries to use data that has already been collected in order to extract useful knowledge to improve public health systems. However, it differs from much of the mentioned work as it relies on a user to query the system, and does not yet create a predictive model. *Portinari*'s idea is to allow epidemiologists and health care planners with a easy to use tool to test hypotheses and find patterns or generate new hypothesis for further investigation.

## VI. CONCLUSION

Data in a socio-technical systems constantly evolves with how society follows new trends and changes its behavior. The effectiveness of such a system relies on how well it is able to achieve its societal goal. We believe that software verification and validation must go beyond building bug-free software to evaluating whether its societal goals are met. For instance, in cervical cancer screening the societal goal is to screen people before they get cervical cancer. This article presents Portinari, a tool to explore and eventually verify and validate socio-technical systems developed for public health. Portinari leverages software technology such as graph databases and web-based interactive visualization to give users insight into a societal process in real-time as data evolves. We use it to evaluate an important part of the Norwegian triage algorithm for cervical cancer screening: cytology exams to be taken at three year intervals. We also use it to compare two different paths taken by patients who are recommend the HPV test.

Applications of *Portinari* in the cervical cancer screening program include: (a) Gaining insight into how effective a screening exam such as the Pap smear is in the early detection of cancer. Several, biotechnology companies propose different laboratory testing techniques. We can use Portinari to evaluate the long-term effectiveness of one test over another. This has several implications in terms of public spending in health. (b) Portinari can also be used by an individual in the general public to evaluate different options in waiting time before the next test is taken in order to minimize risk. However, this would require Portinari to be adapted to different user groups ranging from patient in the general public to experienced epidemiologists. Tools like Portinari aim to create a feedback loop between information systems run by government and the society it influences. Interactive visualizations using tools such as Portinari can spawn self-adaptivity in the sociotechnical system for cervical cancer screening if made publicly available.

# A. Future Work

During the collaborative project with the Cancer Registry of Norway we identified several areas in which software engineering research can benefit public health programs such as cervical cancer screening. We will consider the following ideas as part of future work:

Application of software V&V techniques: The triage algorithm on the population is a starting point to apply software V&V techniques to public health. The algorithm can either be represented in a declarative form (constraint satisfaction problem) or can be executed algorithmically to generate the so called *ideal* patient trajectories for screening. These trajectories can be searched for in real data to validate if a trajectory recommended by the triage algorithm indeed is effective. For instance, if someone who went to screening for three consecutive invitations got cancer despite a normal cytology exam then the triage algorithm is not optimal for at least one person. Similarly, one may also verify to what extent a patient follows a triage algorithm. The sequence of events in the life of a patient can be separated into predicates that can be verified against constraints in the triage algorithm using a formal method such as Alloy [39] or constraint logic programming [40].

**Data mining in socio-technical software systems:** The vast amount of data in public health often contains common patterns that serve as explanation for observed societal behavior. For instance, what is common in patients who regularly take the cytology exam? Is it their age, is it their low-risk lifestyle of not smoking, or is it correlated with an event such as the death of some important due to cancer? Data mining tools can help extract such explanations by, for instance, finding the maximum common sequence in the temporal event sequences of thousands of people. These explanations can also be seen as a hypothesis for further statistical investigation.

**User experience studies:** *Portinari* can be targeted to epidemiologists, health policy makers, and even the general public. User experience studies will give insight into improving the interface for specific user groups.

**Conceptual modeling of data:** Querying big data from public health can be very computationally expensive if the information we seek in the data is hard to match. In this paper, we explore the use of graph databases to transform a simple set of records to a sequence of events. Graph databases make querying very user-friendly, however, their performance to return results is still limited by the computational complexity of subgraph matching, which is NP-hard. Therefore, alternative models of data in socio-technical systems is an active research area. We intend to explore string representations and string matching as a promising alternative for improving performance.

### ACKNOWLEDGMENT

We thank the Norwegian Research Council for funding our work through the Certus-SFI scheme. We also thank the Cancer Registry of Norway and the Brazilian institutions CAPES, CNPq and FAPEMIG for supporting the work.

#### REFERENCES

 M. Arbyn, A. Anttila, J. Jordan, G. Ronco, U. Schenck, N. Segnan, H. Wiener, A. Herbert, and L. Von Karsa, "European guidelines for quality assurance in cervical cancer screening. —summary document," *Annals of Oncology*, vol. 21, no. 3, pp. 448–458, 2010.

- [2] T. Bernie, I. Les, G. Paul, K. Jan, W. David, and S. Chris, "A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, hemoccult," *Bmj*, vol. 317, no. 7158, pp. 559–565, 1998.
- [3] M. A. Nobbenhuis, J. M. Walboomers, T. J. Helmerhorst, L. Rozendaal, A. J. Remmink, E. K. Risse, H. C. van der Linden, F. J. Voorhorst, P. Kenemans, and C. J. Meijer, "Relation of human papilloma virus status to cervical lesions and consequences for cervical-cancer screening: a prospective study," *The Lancet*, vol. 354, no. 9172, pp. 20–25, 1999.
- [4] A. L. Frazier, G. A. Colditz, C. S. Fuchs, and K. M. Kuntz, "Costeffectiveness of screening for colorectal cancer in the general population," *Jama*, vol. 284, no. 15, pp. 1954–1961, 2000.
- [5] S. J. Goldie, L. Kuhn, L. Denny, A. Pollack, and T. C. Wright, "Policy analysis of cervical cancer screening strategies in low-resource settings: clinical benefits and cost-effectiveness," *Jama*, vol. 285, no. 24, pp. 3107–3115, 2001.
- [6] R. A. Smith, D. Manassaram-Baptiste, D. Brooks, M. Doroshenk, S. Fedewa, D. Saslow, O. W. Brawley, and R. Wender, "Cancer screening in the united states, 2015: A review of current american cancer society guidelines and current issues in cancer screening," *CA: a cancer journal for clinicians*, vol. 65, no. 1, pp. 30–54, 2015.
- [7] A. Anttila, G. Ronco, G. Clifford, F. Bray, M. Hakama, M. Arbyn, and E. Weiderpass, "Cervical cancer screening programmes and policies in 18 european countries," *British journal of cancer*, vol. 91, no. 5, pp. 935–941, 2004.
- [8] M. Schiffman, P. E. Castle, J. Jeronimo, A. C. Rodriguez, and S. Wacholder, "Human papillomavirus and cervical cancer," *The Lancet*, vol. 370, no. 9590, pp. 890–907, 2007.
- [9] S. Lönnberg, B. T. Hansen, T. Haldorsen, S. Campbell, K. Schee, and M. Nygård, "Cervical cancer prevented by screening: long-term incidence trends by morphology in norway," *International journal of cancer*, vol. 137, no. 7, pp. 1758–1764, 2015.
- [10] N. Developers, "Neo4j," Graph NoSQL Database [online], 2012.
- [11] W. Martin, "Visualizing risk: Health, gender and the ageing body," *Critical Social Policy*, vol. 32, no. 1, pp. 51–68, 2012.
- [12] I. K. Larsen, M. Småstuen, T. B. Johannesen, F. Langmark, D. M. Parkin, F. Bray, and B. Møller, "Data quality at the cancer registry of norway: an overview of comparability, completeness, validity and timeliness," *European journal of cancer*, vol. 45, no. 7, pp. 1218–1231, 2009.
- [13] E. F. Bilet, H. Langseth, S. Ø. Thoresen, and F. Bray, "Completeness of invasive cervical cancer at the cancer registry of norway," *Acta Oncologica*, vol. 48, no. 7, pp. 1070–1073, 2009.
- [14] J. Krause, A. Perer, and H. Stavropoulos, "Supporting iterative cohort construction with visual temporal queries," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 91–100, 2016.
- [15] R. Snodgrass, "The temporal query language tquel," ACM Transactions on Database Systems (TODS), vol. 12, no. 2, pp. 247–298, 1987.
- [16] F. Prasser, F. Kohlmayer, R. Lautenschläger, and K. A. Kuhn, "Arx-a comprehensive tool for anonymizing biomedical data," in *AMIA Annual Symposium Proceedings*, vol. 2014. American Medical Informatics Association, 2014, p. 984.
- [17] E. Jellum, A. Andersen, P. Lund-Larsen, L. Theodorsen, and H. Orjasaeter, "Experiences of the janus serum bank in norway." *Environmental health perspectives*, vol. 103, no. Suppl 3, p. 85, 1995.
- [18] B. T. Hansen, S. S. Hukkelberg, T. Haldorsen, T. Eriksen, G. B. Skare, and M. Nygård, "Factors associated with non-attendance, opportunistic attendance and reminded attendance to cervical screening in an organized screening program: a cross-sectional study of 12,058 norwegian women," *BMC Public Health*, vol. 11, no. 1, p. 1, 2011.
- [19] N. Team, "Cypher query language," 2013.
- [20] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive sankey diagrams," in *IEEE Symposium on Information Visualization*, 2005. INFO-VIS 2005. IEEE, 2005, pp. 233–240.
- [21] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman, "Lifelines: visualizing personal histories," in *Proceedings of the SIGCHI*

conference on Human factors in computing systems. ACM, 1996, pp. 221–227.

- [22] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman, "Lifeflow: visualizing an overview of event sequences," in *Proceedings of the SIGCHI conference* on human factors in computing systems. ACM, 2011, pp. 1747–1756.
- [23] M. Monroe, K. Wongsuphasawat, C. Plaisant, B. Shneiderman, J. Millstein, and S. Gold, "Exploring point and interval event patterns: Display methods and interactive visual query," *University of Maryland Technical Report*, 2012.
- [24] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2659–2668, 2012.
- [25] D. Gotz and H. Stavropoulos, "Decisionflow: Visual analytics for high-dimensional temporal event sequence data," *IEEE transactions on* visualization and computer graphics, vol. 20, no. 12, pp. 1783–1792, 2014.
- [26] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman, "Interactive information visualization to explore and query electronic health records," *Foundations and Trends in Human-Computer Interaction*, vol. 5, no. 3, pp. 207–298, 2011.
- [27] J. A. Fails, A. Karlson, L. Shahamat, and B. Shneiderman, "A visual interface for multivariate temporal data: Finding patterns of events across multiple histories," in 2006 IEEE Symposium On Visual Analytics Science And Technology. IEEE, 2006, pp. 167–174.
- [28] J. Jin and P. Szekely, "Querymarvel: A visual query language for temporal patterns using comic strips," in 2009 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). IEEE, 2009, pp. 207–214.
- [29] A. Abouzied, J. Hellerstein, and A. Silberschatz, "Dataplay: interactive tweaking and example-driven correction of graphical database queries," in *Proceedings of the 25th annual ACM symposium on User interface software and technology.* ACM, 2012, pp. 207–218.
- [30] N. Elmqvist, J. Stasko, and P. Tsigas, "Datameadow: a visual canvas for analysis of large-scale multivariate data," *Information visualization*, vol. 7, no. 1, pp. 18–33, 2008.
- [31] M. Coles, Pro T-SQL 2008 programmer's guide. Apress, 2008.
- [32] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *International journal of medical informatics*, vol. 77, no. 2, pp. 81–97, 2008.
- [33] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.
- [34] N. Ramakrishnan, D. Hanauer, and B. Keller, "Mining electronic health records," *Computer*, vol. 43, no. 10, pp. 77–81, 2010.
- [35] H. M. Krumholz, "Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system," *Health Affairs*, vol. 33, no. 7, pp. 1163–1170, 2014.
- [36] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen, and S. Brunak, "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients," *Nature communications*, vol. 5, 2014.
- [37] M. Zamora, M. Baradad, E. Amado, S. Cordomí, E. Limón, J. Ribera, M. Arias, and R. Gavaldà, "Characterizing chronic disease and polymedication prescription patterns from electronic health records," in *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on.* IEEE, 2015, pp. 1–9.
- [38] K. Ng, A. Ghoting, S. R. Steinhubl, W. F. Stewart, B. Malin, and J. Sun, "Paramo: A parallel predictive modeling platform for healthcare analytic research using electronic health records," *Journal of biomedical informatics*, vol. 48, pp. 160–170, 2014.
- [39] D. Jackson, "Software abstractions: Logic," Language, and Analysis. MIT Press, vol. 2012, 2006.
- [40] J. Jaffar and J.-L. Lassez, "Constraint logic programming," in Proceedings of the 14th ACM SIGACT-SIGPLAN symposium on Principles of programming languages. ACM, 1987, pp. 111–119.