Process Mining applied to Laboratory Workflow: a comparison between two sites

Alessandro Sulis, Mauro Del Rio, Francesca Frexia, Paolo Anedda, Giovanni Busonera,
Marco Cogoni, Gianluigi Zanetti
CRS4
Pula, CA, Italy
firstname.lastname@crs4.it

***Abstract*—In this work we present a Process Mining analysis performed on event logs coming from two fully automated Laboratory pre-analytical sites. Our purpose is to empirically discover the relevant workflow models and compare them with the theoretical (conceptual) workflow. The goal is to discover any unexpected behavior of the site workflow, and to refine and correct the theoretical model according to what the models themselves are evidencing, following a quantitative and objective approach.**

I. Introduction

Process Mining technique applications to the healthcare domain increased in the last decade, at first to discover clinical workflows from real data [1] [2], later to also evaluate the conformance to medical guidelines of actual behavior [3] [4]. Business Process Management and Business Intelligence have been used in healthcare processes descriptions [5] [6] [7] [8], but they often lead to models too influenced by stakeholders visions of the workflow, therefore unable to capture aspects outside their specific vision. Process Mining instead, starting from event logs recorded during process running, helps to create dynamic end-to-end models, describing the current behavior of heterogeneous workflows.

The preanalytical phase in laboratory medicine is a perfect example of how a medical procedure can highly vary according to the clinical site: best practices and protocols exist, but the combination of human resources, devices, test panels and management rules can result in extremely disparate configurations in different phlebotomy rooms. This peculiarity is among the reasons that preanalytical phase has partially been left behind in the process of automation which invested laboratories in the last twenty years; anyway, the high error rate occurrence of this phase induced many vendors to enrich automation support in the pre-analytical workflow, from test ordering to sample labeling [9] [10] [11].

This paper reports the results of a preliminary analysis of the first example of automation for the phlebotomy process combined with a traceability system, generating events at each relevant step of the workflow[1]. We analyzed traceability event logs coming from two different hospital sites, in order to discover the real workflows and to compare the discovered models to the theoretical representation: we will show how process mining techniques can be helpful for the settings and the fine tuning of an automated system in sites with different workflows, giving an objective and quantitative representation of reality.

The remainder of the paper is structured as follows. Section II provides a short description of the clinical context (laboratory medicine), followed by a overview of process mining techniques devoted to workflow discovery, including the framework we developed to perform the analysis(Section III). The results will be analyzed in Section IV and the last two sections will be dedicated to related works (Section V) and conclusions (Section VI).

II. Laboratory Workflow

The classification of a laboratory workflow is not universal and can vary according to different authors, but traditionally laboratory medicine workflow is divided into three main phases:

- *preanalytical phase*, which consists of test ordering, patient identification and sample collection, transportation and preparation;
- *analytical phase*, including all the steps to perform the analysis required by the physician;
- *postanalytical phase*, consisting in reporting and distribution of test results.

The activities to go from exam prescriptions to results (Total Testing Process TTP- model) [13] [14] [15] are not confined to the laboratory, so an additional distinction was made to divide what happens inside and outside the laboratory, introducing the definition of *pre-preanalytical* (from ordering to sample arrival in the laboratory) and *post-postanalytical* (reception and interpretation of laboratory reports) and limiting preanalytical to sample preparation before the analysis and postanalytical to validation and reporting. These definitions are functional to identify mistakes problems origin, since the importance of each of these steps is specified in the official definition by International Organization for Standardization (ISO) 22367 of *laboratory error*: ̈failure of planned action to be completed as intended, or use a wrong plan to achieve an aim, occurring at any part of the laboratory cycle, from ordering examinations to reporting results and appropriately interpreting and reacting to them¨ [16]. To evaluate and improve accuracy and quality in the testing process, externally quality assurance programs (EQA) and rules for internal quality control (IQC)

[1]Inpeco ProTube [11]

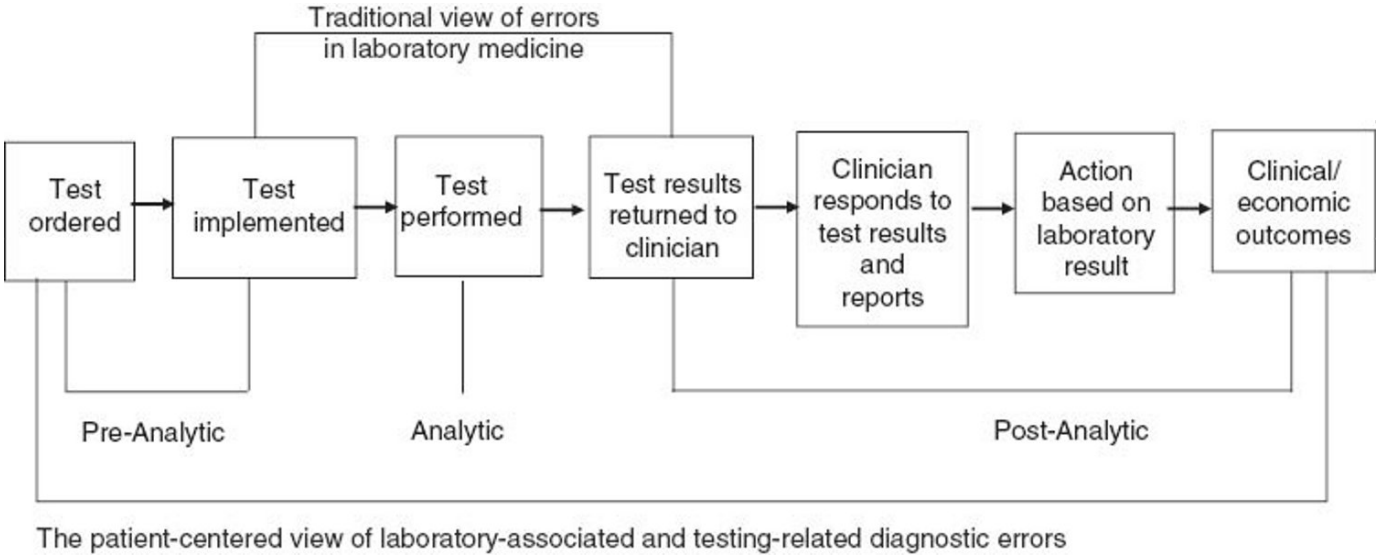

Fig. 1. Workflow elements model, based on analysis of workflow definitions [12].

have been developed [17] [18] and methodologies derived from the industrial sector have been introduced, like process automation, lean and six sigma strategies [19] [20].The first objective of the optimization in the past years was the analytical phase and the results are extremely good: this process, considering the six sigma evaluation, has a close to 5 sigma performance (0.002%), value which is the best among healthcare processes (1-2 sigma) and is comparable with non-healthcare industries [21]. Experimental studies [22] [23] shows that the analytical phase is less affected by errors, as can be seen in fig, and that pre-pre and preanalytical steps before analysis are the more crucial. The most frequent errors are related to order entry, patient/specimen misidentification, sample collection, inappropriate container, handling, storage and transportation, sorting and routing, pour-off, aliquoting and centrifugation [21].

Automation, lean and sigma strategies also applied to the preanalytical phase showed their efficacy [24] [25] [26] [27] and this approach was considered also from pre-preanalytical, introducing a series of systems like BCROBO [9] and EOS Lab.E.L® [10], with good results in clinical contexts [28]. Inpeco ProTube [11] system provides support to patient secure identification and correct sample preparation, recording a series of traceability events, which we consider as input for our analysis.

III. Process Mining for workflow discovering

As reported by literature, Process Mining has the aim *"to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in todays (information) systems"* [29]. It is a young discipline, placed in the middle between Business Intelligence and Business Process Management and useful to bridge the gap between them: classical data mining concepts are enriched with a process model-driven approach, with several use cases related to process model discovering from data, bottlenecks detection, performance evaluations.

In the previous sections, we highlighted the importance of the laboratory pre-analytical phase, and how it can be improved through the adoption of technologies able to automatize some critical steps, ensure the traceability of all produced and exchanged data, obtaining a fully-traced workflow of the pre-analytical phase itself. In this context, Process Mining techniques can help to perform a detailed analysis of the workflow, starting from its logs, which can be seen as the input of the analysis. Event logs, sequentially recorded, are grouped into activities, which are in turn related to a workflow (process) case: a case groups all the events and activities belonging to the same workflow instance.

Different types of Process Mining [30] can be used to analyze a workflow, according to how events are used:

- *discovery* aims to infer a process model from event logs, without a-priori information;
- *conformance* compares an existing model (inferred or theoretical) with actual event logs, checking the conformance between reality and the model itself;
- *enhancement* improve, extending or repairing, the a-priori model using logs to obtain a better conformance between model and reality or adding new aspects and point of views.

Determining if workflow event data are following the theoretical model or not is among the most important questions Process Mining can answer to: indeed, it often happens that, even if the event data have been produced starting from a conceptual schema, the reality is different and not all workflow cases follow the theoretical model as expected. The so called Discovery type of Process Mining provides techniques to mine a workflow model from event logs, obtaining in this way a representation of the real process, which can be compared

to the theoretical one. Both theoretical and mined workflow models can be then analyzed using the Conformance Process Mining type, which tells us if the reality conforms to the model and viceversa.

Several algorithms and approaches can be followed to start a Process Mining analysis; in this section we will briefly describe the one of interest for the scope of our work, the Heuristic Miner. The goal of this algorithm is to mine a model of the workflow(process) by analyzing its event data in two steps: the first one looks for activities dependencies, the second one determines how an activity depends from the previous and the following. Looking the two phases from the point of view of the mined diagram model, the first one determines the arcs, while the second one determines the split or join gateways, which can be parallel or exclusive (AND/OR gateways). Some thresholds can also be applied in order to filter infrequent arcs or gateways. As for the algorithms, same multiplicity of notations is present in literature for workflow model representations. All workflow models shown in the following of this paper will be represented according the Business Process Models and Notation (BPMN), a well-known standard for business process modeling, based flowchart representation a simple to read, easily understandable by all business stakeholders. The main components of a BPMN diagrams are the activities, the arcs and the parallel and exclusive gateways, represented respectively with a "+" and a "X" sign.

A complete list of tools for Process Mining can be found in [31]. The main open source framework for Process Mining is ProM [31], representing the state-of-the-art, with about three hundreds algorithms/tools: the GUI is user-friendly, but it is not so easy to use it as a back-end for an external application. Another under development open source tool is PMLAB [32] [33], which provides a Python-based scripting environment for process mining, while the more used commercial solution is Disco [34].

These tools cover a large spectrum of applications, but our research interests (dealing with healthcare informatics and computational support to a wide range of bio-medical projects and connected to manage and analysis of large datasets with distributed computing technologies) required a flexible and modular solution, scalable and based on indipendent modules: for these purposes we developed pyMine, a open source Python library for Process Mining, soon available as an open-source project.

The analysis and the results presented in this paper have been obtained using pyMine modules for heuristic mining and conformance checking.

IV. Results

The study analyzes the laboratory pre-preanalytical workflow, through the use of process mining applied to event logs coming from two different laboratories, site A and site B (details about event number are in the following table). Given the theoretical model, our first objective was the discovery of the models deriving from the two sites' events and the comparison with the expected behavior, to quantify the "coverage" in two actual clinical settings of real workflows trajectories respect expected path.

Event logs for the two sites			
Site A		**Site B**	
Total Events	Total Cases	Total Events	Total Cases
221316	28308	299761	42462

TABLE I
NUMBER OF EVENTS

The process has been modeled according to system requirements before the installation, identifying these macro activities (the BPMN diagram of the theoretical model is shown in Figure 2):

- **IDENTIFICATION**: query and retrieve of patients information;
- **SEARCH ORDERS**: query and retrieve of patient orders;
- **TRANSCODING ERROR**: atomic activity indicating that an error occurred while computing the required tubes for the orders retrieved;
- **VERIFY ORDERS**: this activity is performed if some orders have to be filtered (according to the site configuration) or have some peculiarities (i.e. timed repetitions)
- **LABELING**: production of the labels;
- **LABELING SET OPTIONS**: configurations for the labeling;
- **RELABELING**: sample relabeling;
- **ALT TUBE**: choice of different tube type;
- **CHECKOUT**: confirmation that all tubes or part of them are filled and ready for transport;
- **ABORT**: interruption of the process caused by the operator

Each event belongs to an activity and is structured as follows:

- *case ID*: unique identifier of the case to which the event is referred;
- *timestamp*: reference timestamp of the event;
- *operator*: numerical identifier for the person who performed the action related to the event;
- *activity*: definition of the action the event belongs to;
- *lifecycle*: indication of the lifecycle attribute of the activity (i.e. START, END);
- *resources*: list of unique identifiers associated to the devices involved in the action related to the event.

Processing the two datasets with heuristic mining techniques, we obtained the mined models for the two sites A and B shown in Figure 3: in the diagrams the arcs in red indicate all mined paths not allowed by the theoretical model. In order to evaluate how the mined models reflect the event data used to infer them, the *Fitness Indicator* [35] has been calculated, as the ratio of the cases accepted by the total number of the cases. Site A model obtained a fitness of 0.9994, while site B model fitness was 0.9950: these values indicate that the two models correctly represent real-life event logs.

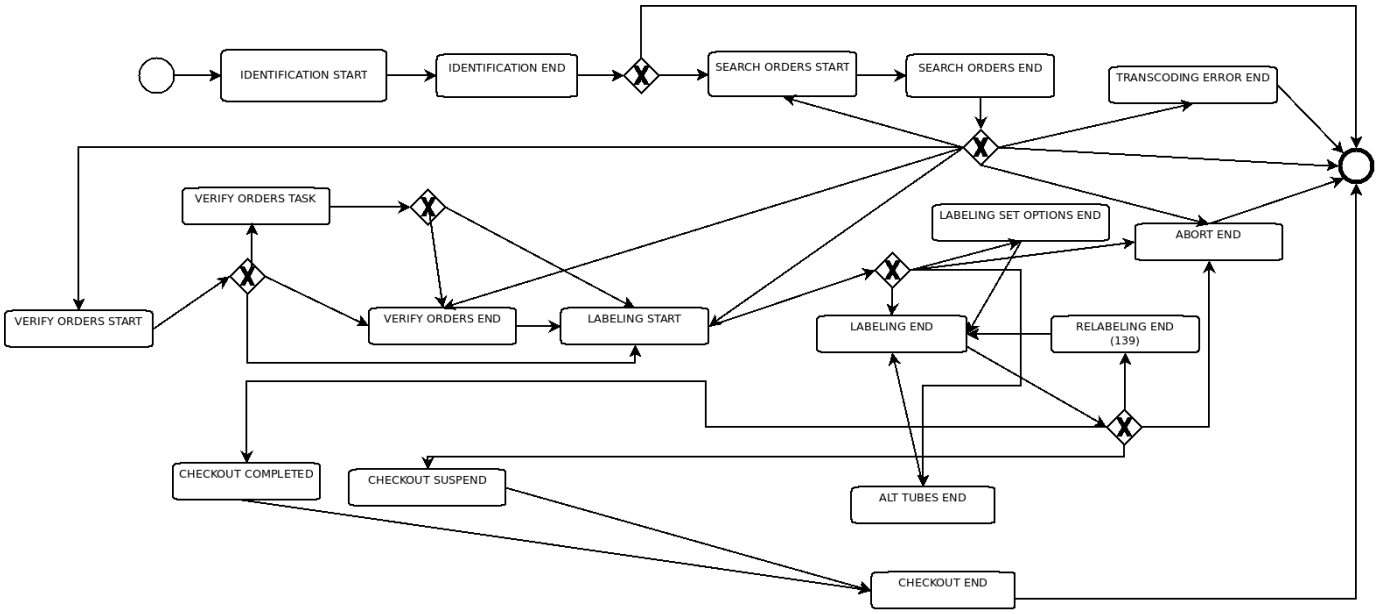

Fig. 2. Theoretical BPMN model of the Laboratory pre-preanalytical workflow.

At this preliminary level of study we can only comment the difference of topology between the two sites A and B: site A has a lower variability in terms of pathways than site B; furthermore, even if the most frequent path is the same in the two sites, in case of site A the percentage of cases that follow this path in relation to the total number of cases is higher than in site B. Another different behavior between A and B comes to light when comparing their topology to the theoretical model: site A model is very similar to the theoretical one, while this is not true for site B. Indeed, site A and theoretical model differ just in two paths, covered by a few cases (only 320 on 28308), while site B presents about 15000 cases following unexpected paths. This is confirmed by computing the Fitness Indicator of the two datasets on the theoretical model: site A presents a value of 0.9565 and site B a value of 0.3990.
Site B results must be further investigated. Possible reasons of such a topology discrepancy between B and the theoretical model could be:

- there is excessive noise in the data set;
- there are behaviors of the devices or the operators that are considered impossible by the theoretical model;
- the theoretical model is missing parts of the real workflow.

Analyzing the number of site B cases associated with the unexpected paths, some of these trajectories could fall in the first two categories, while the high incidence of those related to the VERIFY ORDERS activities suggests to us that probably these paths are effectively ”legal” and they should be added to the theoretical model.

V. Related Works

The myriad of perspectives connected to workflow definition and analysis is very well depicted in [36], which is a systematic literature review of workflow research articles published between 1995 and 2008, but also presents specific questions about methods for workflows analysis and connected evaluation metrics and about workflow definition and basic components. Starting from a first selection of 6221 articles, in the fields of engineering, basic sciences, healthcare and social sciences, the authors observed the absence of a unique definition of workflow (always varying according the different studies interests) and outlined a conceptual framework of workflow-related terms, depicted in Figure 4, with two levels, pervasive and specific, respectively presenting three (context, temporal and aggregate factors) and five (actors, artifacts, actions, characteristics and outcomes) components.

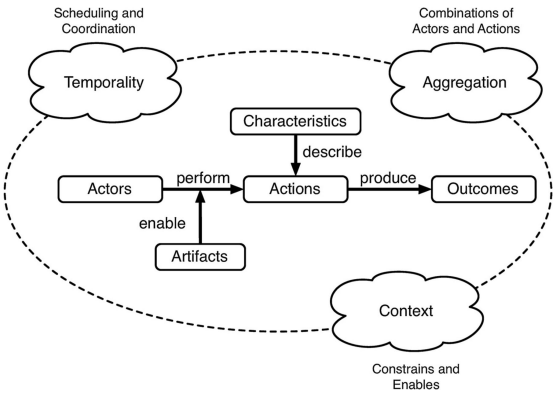

Fig. 4. Workflow elements model, based on analysis of workflow definitions [36].

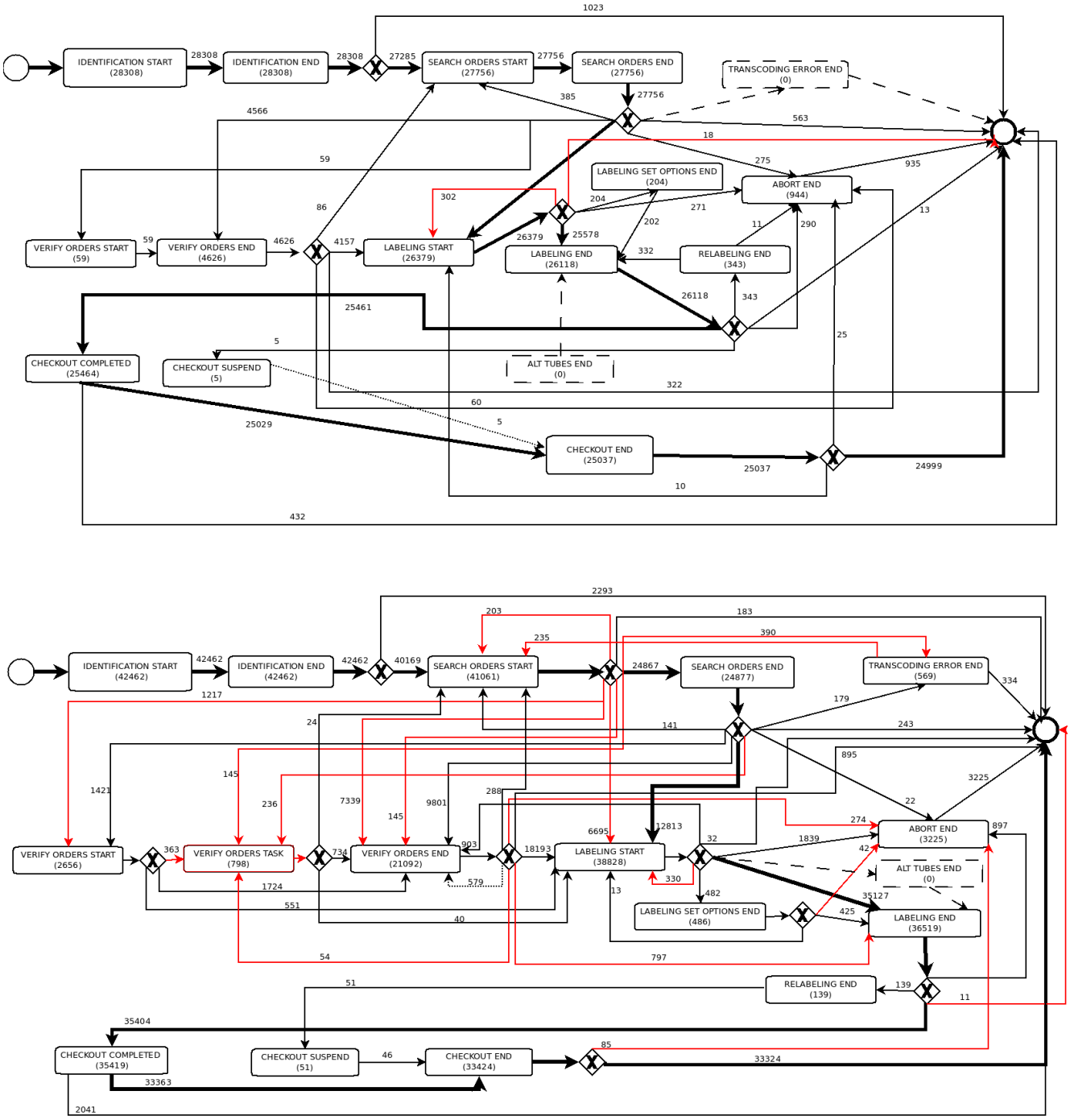

Fig. 3. BPMN mined models of the pre-preanalytical workflow for site A (above) and site B (below). Red arcs indicate all mined paths which are not allowed in the theoretical model. Frequencies for both nodes and arcs are shown; the most frequent paths are drawn with thicker lines.

Other important factors emerging from the article are the prevalence of qualitative approaches in the examined publications and the difficulties of representing the dynamic behavior of healthcare reality through a flow-chart or a model. Process Mining can offer significative solutions to these issues, as explained in the IEEE Task Force on Process Mining Manifesto [29] and in the main reference book in the field [30]. A complete review about literature related to the application of process mining in healthcare can be find online in [37]: the 59 papers are divided according to two types of process

mining, discovery (51 papers) and conformance (8 papers). To create a common framework for healthcare applications, [38] defines a healthcare reference model, organizing the classes of possible data for modeling of processes, the relationships potentially connecting them and some criteria to evaluate and improve event logs quality.

VI. Conclusions and Future Works

In this paper we presented an example of application of Process Mining techniques to automatically extract workflow models from events log coming from two real pre-preanalytical laboratory automation and traceability system sites: two models were discovered analyzing data from two real clinical sites, A and B. The results highlight the added value of the process mining approach, in terms of both discovering of undesired behaviors in the real workflow and providing guidance in the refinement of the theoretical model. Process Mining tools help to build quantitative and objective models of reality. The progressive automation of healthcare processes will make available an increasing amount of data and the extensive application of process mining strategy will help to o extract value from the latter.

Our work is continuing along two complementary directions. On one side, we are refining the process mining tools, with the algorithmic refiniments and extension converging in the pyMine framework. On the other, we are working with clinicians to extend the boundary of the analysis, considering wider processes with human and device actors, as for example the overall Laboratory testing process, with the aim to add some context information in order to guide the modeling with questions about the real system behavior (performance, bottlenecks, evolutions, and improvements).

Acknowledgments

This work has been partially supported by the Region of Sardinia, project PIA/2012-7633-974 TRACE (TRAceable Clinical Envirnoments).

References

[1] R. S. Mans, M. H. Schonenberg, M. Song, and P. J. M. Bakker, "Application of process mining in healthcare a case study in a dutch hospital," 2008.

[2] L. Maruster and R. J. Jorna, "From data to knowledge: a method for modeling hospital logistic processes," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 2, pp. 248–255, 2005. [Online]. Available: http://dx.doi.org/10.1109/TITB.2005.847194

[3] J. van de Klundert, P. Gorissen, and S. Zeemering, "Measuring clinical pathway adherence," *Journal of Biomedical Informatics*, vol. 43, no. 6, pp. 861 – 872, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046410001127

[4] R. Dunkl, K. A. Fröschl, W. Grossmann, and S. Rinderle-Ma, "Assessing medical treatment compliance based on formal process modeling," in *Information Quality in eHealth*, ser. Information Quality in eHealth (USAB 2011). -: Springer Lecture Notes, November 2011. [Online]. Available: http://eprints.cs.univie.ac.at/3079/

[5] R. Müller and A. Rogge-Solti, "Bpmn for healthcare processes," in *Proceedings of the 3rd Central-European Workshop on Services and their Composition, ZEUS*. CEUR Workshop Proceedings, volume 705, 2011, pp. 65–72.

[6] M. G. Rojo, E. Rolón, L. Calahorra, F. García, R. P. Sánchez, F. Ruiz, N. Ballester, M. Armenteros, T. Rodríguez, R. M. Espartero *et al.*, "Implementation of the business process modelling notation (bpmn) in the modelling of anatomic pathology processes," *Diagnostic pathology*, vol. 3, no. Suppl 1, p. S22, 2008.

[7] P. Brooks, O. El-Gayar, and S. Sarnikar, "Towards a business intelligence maturity model for healthcare," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, 2013, pp. 3807–3816.

[8] M. Spruit, R. Vroon, and R. Batenburg, "Towards healthcare business intelligence in long-term care: An explorative case study in the netherlands," *Computers in Human Behavior*, vol. 30, pp. 698–707, 2014.

[9] Techno Medica BC-ROBO. [Online]. Available: http://www.technomedica.co.jp/English/888.html

[10] EOS Lab.E.L®. [Online]. Available: https://www.cab.de/en/marking/label-printer/eos-series/?ref=overlay_langl

[11] INPECO ProTube. [Online]. Available: http://www.inpeco.com/EN/Protube/Pages/ProTube.aspx

[12] M. Plebani, "Laboratory-associated and diagnostic errors: a neglected link," *Diagnosis*, vol. 1, no. 1, pp. 89–94, 2014.

[13] S. R. Gambino, "Met and unmet needs of the automated clinical laboratory," *Analytical Chemistry*, vol. 43, no. 1, pp. 20A–20A, 1971.

[14] G. D. Lundberg, "Acting on significant laboratory results," *Jama*, vol. 245, no. 17, pp. 1762–1763, 1981.

[15] M. Plebani, M. Laposata, and G. D. Lundberg, "The brain-to-brain loop concept for laboratory testing 40 years after its introduction," *American journal of clinical pathology*, vol. 136, no. 6, pp. 829–833, 2011.

[16] International Organization for Standardisation. ISO/TS 22367: Medical laboratories - reduction of error through risk management and continual improvement. Geneva: International Organization for Standardisation, 2008. [Online]. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=40918

[17] L. Sciacovelli, L. Zardo, S. Secchiero, and M. Plebani, "Quality specifications in eqa schemes: from theory to practice," *Clinica chimica acta*, vol. 346, no. 1, pp. 87–97, 2004.

[18] L. Sciacovelli, L. Zardo, S. Secchiero, M. Zaninotto, and M. Plebani, "Interpretative comments and reference ranges in eqa programs as a tool for improving laboratory appropriateness and effectiveness," *Clinica chimica acta*, vol. 333, no. 2, pp. 209–219, 2003.

[19] N. Riebling and L. Tria, "Six sigma project reduces analytical errors in an automated lab." *MLO: medical laboratory observer*, vol. 37, no. 6, pp. 20–22, 2005.

[20] K. E. Blick, "Providing critical laboratory results on time, every time to help reduce emergency department length of stay how our laboratory achieved a six sigma level of performance," *American journal of clinical pathology*, vol. 140, no. 2, pp. 193–202, 2013.

[21] R. Hawkins, "Managing the pre-and post-analytical phases of the total testing process," *Annals of laboratory medicine*, vol. 32, no. 1, pp. 5–16, 2012.

[22] M. Plebani, "Errors in clinical laboratories or errors in laboratory medicine?" *Clinical Chemical Laboratory Medicine*, vol. 44, no. 6, pp. 750–759, 2006.

[23] P. Carraro and M. Plebani, "Errors in a stat laboratory: types and frequencies 10 years later," *Clinical chemistry*, vol. 53, no. 7, pp. 1338–1342, 2007.

[24] M. Cankovic, R. C. Varney, L. Whiteley, R. Brown, R. D'Angelo, D. Chitale, and R. J. Zarbo, "The henry ford production system: Lean process redesign improves service in the molecular diagnostic laboratory: a paper from the 2008 william beaumont hospital symposium on molecular pathology," *The Journal of Molecular Diagnostics*, vol. 11, no. 5, pp. 390–399, 2009.

[25] T. J. Persoon, S. Zaleski, and J. Frerichs, "Improving preanalytic processes using the principles of lean production (toyota production system)," *American journal of clinical pathology*, vol. 125, no. 1, pp. 16–25, 2006.

[26] G. Lima-Oliveira, G. Lippi, G. L. Salvagno, E. Danese, M. Montagnana, G. Brocco, M. Voi, G. Picheth, and G. C. Guidi, "Does laboratory automation for the preanalytical phase improve data quality?" *Journal of laboratory automation*, p. 2211068213488892, 2013.

[27] A. W. Jekelis, "Increased instrument intelligencecan it reduce laboratory error?" *Journal Information*, vol. 39, no. 3, 2005.

[28] G. Da Rin, "Pre-analytical workstations: a tool for reducing laboratory errors," *Clinica Chimica Acta*, vol. 404, no. 1, pp. 68–74, 2009.

[29] W. M. P. v. d. Aalst, A. Adriansyah, A. K. A. d. Medeiros, F. Arcieri, T. Baier, T. Blickle, R. P. J. C. Bose, P. v. d. Brand, R. Brandtjen, J. C.

A. M. Buijs, A. Burattin, J. Carmona, M. Castellanos, J. Claes, J. Cook, N. Costantini, F. Curbera, E. Damiani, M. de Leoni, P. Delias, B. F. v. Dongen, M. Dumas, S. Dustdar, D. Fahland, D. R. Ferreira, W. Gaaloul, F. v. Geffen, S. Goel, C. W. Gnther, A. Guzzo, P. Harmon, A. H. M. t. Hofstede, J. Hoogland, J. Espen Ingvaldsen, K. Kato, R. Kuhn, A. Kumar, M. La Rosa, F. Maggi, D. Malerba, R. S. Mans, A. Manuel, M. McCreesh, P. Mello, J. Mendling, M. Montali, H. Motahari Nezhad, M. zur Muehlen, J. Munoz-Gama, L. Pontieri, J. Ribeiro, A. Rozinat, H. Seguel Prez, R. Seguel Prez, M. Seplveda, J. Sinur, P. Soffer, M. S. Song, A. Sperduti, G. Stilo, C. Stoel, K. Swenson, M. Talamo, W. Tan, C. Turner, J. Vanthienen, G. Varvaressos, H. M. W. Verbeek, M. Verdonk, R. Vigo, J. Wang, B. Weber, M. Weidlich, A. J. M. M. Weijters, L. Wen, M. Westergaard, and M. T. Wynn, "Process mining manifesto," in *BPM 2011 Workshops, Part I*, vol. 99. Springer-Verlag, 2012, pp. 169–194. [Online]. Available: http://www.win.tue.nl/ieeetfpm/doku.php?id=shared:process_mining_manifesto

[30] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[31] ProM Tools. [Online]. Available: http://www.promtools.org

[32] PMLAB. [Online]. Available: http://www.cs.upc.edu/~jcarmona/PMLAB

[33] J. Carmona Vargas, M. Solé *et al.*, "Pmlab: An scripting environment for process mining," 2014.

[34] Discover your processes. [Online]. Available: https://fluxicon.com/disco/

[35] A. Rozinat and W. M. van der Aalst, "Conformance testing: Measuring the fit and appropriateness of event logs and process models," in *Business Process Management Workshops*. Springer, 2006, pp. 163–176.

[36] K. M. Unertl, L. L. Novak, K. B. Johnson, and N. M. Lorenzi, "Traversing the many paths of workflow research: developing a conceptual framework of workflow terminology through a systematic literature review," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 265–273, 2010.

[37] Process Mining in Healthcare Papers. [Online]. Available: http://www.healthcare-analytics-process-mining.org/

[38] R. Mans, W. M. P. van der Aalst, and R. J. B. Vanwersch, *Process Mining in Healthcare - Evaluating and Exploiting Operational Healthcare Processes*, ser. Springer Briefs in Business Process Management. Springer, 2015. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16071-9