




Proceeding Paper

Lean Six Sigma: Application of the Methodology in Data Processing for Cancer Registry [†]

Paolo Fidelbo ^{*}, Rosalia Ragusa , Carmelo Viscosi, Andrea Benedetto, Eleonora Salamone, Angelo Pisano and Margherita Ferrante 

Registro Tumori Integrato di Catania-Messina-Enna, Azienda Ospedaliero Universitaria Policlinico “G. Rodolico–San Marco”, 95123, Catania, Italy; ragusa@unict.it (R.R.); ettore.viscosi@gmail.com (C.V.); studiogeo.benedetto@gmail.com (A.B.); e.salamone@policlinico.unict.it (E.S.); a.pisano@policlinico.unict.it (A.P.); marfer@unict.it (M.F.)

^{*} Correspondence: paolo.fidelbo@gmail.com

[†] Presented at the International One Health Conference 2022, Catania, Italy, 27–28 September 2022.

Abstract: Since 2020, the Catania-Messina-Enna Cancer Registry (CR) has operated a transformational and incremental program while also applying a Lean Six Sigma methodology (LSS) to optimize the processes and reduce waste. Each project aimed to raise the performance of the CR while also providing the opportunity for human resources to express their talent. In this context, a machine learning project was developed to reduce the time spent on raw free-text histopathological reports that contain relevant information for cancer evaluation. The ability to extract meaningful information from histopathology reports is really important because reports provide crucial insights into the morphology and topography of cancer, enabling operators to validate oncology cases with the utmost diagnostic precision. However, the CR faced a significant challenge due to the extensive volume of written natural language reports, where only a small fraction contains pertinent information for cancer evaluation. In this paper, we describe how we applied the LSS method, the observed difficulties, and the benefits achieved by adopting a machine learning algorithm as a strategic solution in the Improve phase.

Keywords: project management; Lean Six Sigma; cancer registry



Citation: Fidelbo, P.; Ragusa, R.; Viscosi, C.; Benedetto, A.; Salamone, E.; Pisano, A.; Ferrante, M. Lean Six Sigma: Application of the Methodology in Data Processing for Cancer Registry. *Med. Sci. Forum* **2023**, *19*, 12. <https://doi.org/10.3390/msf2023019012>

Academic Editors: Chiara Copat, Antonio Cristaldi, Gabriela Fernandez, Melissa Jimenez, Gomez Tagle, Paolo Lauriola, Graziella Machado, Valerio Paolini, Prisco Piscitelli, Domenico Vito, Gea Oliveri Conti, Carla Albanese, Alfina Grasso, Carol Maione and Lorenzo Laquinteri

Published: 30 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Lean Six Sigma (LSS) is “a methodology that maximizes shareholder value by achieving the fastest rate of improvement in customer satisfaction, cost, quality, process speed and invested capital” [1]. It shows how Lean and Six Sigma methods complement and reinforce each other. Moreover, LSS has a commonality in the field of application. Lean focuses on reducing “chronic” waste and Six Sigma focuses on reducing the variation and thereby its “adverse effects” [2].

Transformational projects usually have to cope with different risks that could bring about the failure of the project if not adequately addressed. In order to reduce the probability of failure, the project manager applied the international standard of the Project Management Institute [3], pairing it with coaching techniques and business analysis with frequent sponsor validation.

Histopathology reports serve as one of the most important sources of information for the operators of the CR. A prominent challenge arises from the reliance on the text that implies the presence of heterogeneous expressions due to the natural language used by the authors of the documents and due to the low level of standardization within the reports. Moreover, the proportion of pertinent information for cancer evaluation constitutes merely a fraction of the overall annual histopathological reports. Operators usually need to read all of the documents just to identify those that help them to identify an oncology case and then validate it at the utmost level of precision.

Since 2020, the Catania-Messina-Enna Cancer Registry (CR) has operated a transformational program while also applying a Lean Six Sigma methodology (LSS) to optimize the processes and reduce waste.

The project management methodology was also tailored to the specific peculiarity of the CR collaborators, existing organizational process assets, and enterprise environmental factors.

In this paper, we describe how we applied the LSS method, the observed difficulties, and the benefits achieved by adopting a machine learning algorithm as a strategic solution in the Improve phase. In this paper, we will also look at the machine learning project [4] that aimed to better manage histopathology reports.

2. Methodology

2.1. LSS Methodology

The LSS methodology applies a cycle (called DMAIC) made of five distinct phases: Define, Measure, Analyze, Improve, and Control [5].

The Define phase. The Define phase aims to state the problem, reach a consensus about it, define the project management approach, approve the project charter, and set the measurable variables that represent the process performance. A deep analysis of the actual process is performed and a flowchart is defined to better understand the process. Flowcharts are developed in an iterative approach until the right level of detail is gained. By the analysis of the flowchart, it is also possible to identify “non-value added activities” (NVAA), which are actions that do not increase the worth of what is delivered by the process.

The Measure phase. During the Measure phase, a plan to collect relevant data is designed and applied to create a meaningful picture of the actual state of the process. The raw data are then used to create a baseline of the performance of the process so that at the end of the project, it is possible to measure the increment (or decrement in the worst scenario) of the performance of the process.

The Analyze phase. The Analyze phase aims to identify the root causes of the problem and to define the action to solve the problems, which involves a deep use of statistics. This is the phase that applies the Six Sigma methodologies and provides insight into the process.

The Improve phase. The Improve phase aims to implement the solution suggested by the data collected and analyzed during the previous phases. One of the processes that the CR aimed to improve was related to the delivery of thousands of histopathology reports. In the next paragraphs, a project about machine learning is described as a solution that was implemented in one of the DMAIC cycles performed by the CR.

The Control phase. The Control phase measures the variables identified during the Define phase in order to numerically evaluate if a real improvement had been reached within the Improve phase. The Control phase was implemented in two distinct periods—a few weeks and two months—to separate the transition from the fully operational new process. During this phase, the process owner and other people involved were trained and sustained during the application of the designed changes in order to guarantee the adoption of the new process in the long run.

The LSS methodology focuses on reducing eight types of waste: overproduction, inventory, defects, motion, over-processing, waiting, transportation, and talent. All of them were addressed during the define phase. Each project was managed by a project team made of the domain expert (usually a biologist or a medical doctor), the process owner (the professional accountable for the process), two developers, and the Project Manager/LSS manager.

2.2. The Machine Learning Project

One of the processes that the CR aimed to improve was related to the delivery of thousands of histopathology reports. A project about machine learning was a solution implemented in one of the DMAIC cycles performed by the CR. Histopathology reports are important for CR as the operators evaluate them to define the morphology and topography

of cancer. Using a machine learning algorithm, the CR built a classification model that evaluates each free-text report and returns a score that indicates the probability that it could contain relevant oncologic information. The CR managed a project to identify relevant information in each report by using a classification model that assigned a score to all the reports. Then, operators were able to read the first report that could contain relevant oncologic information as detected by the algorithm.

The project implemented a supervised machine learning approach [6] that detects among new free-text diagnoses only the reports that contain useful information. This is achieved by testing the performance of the most common classification algorithms and performing an automatic classification of pathology free-text reports into a class of relevant or not-relevant oncologic information.

More than one million records were available to train the algorithm, so it was possible to split this large amount of data to reduce the time spent by the training and also gain a test dataset and do not conduct the k-fold cross-validation [7].

At the end of the training, the project team was able to evaluate several learning algorithms and identify the one with the highest performance.

3. Results

The transformational program was defined after a series of coaching interviews with all the internal stakeholders of the CR. Then, a prioritization matrix was defined for evaluating several parameters such as strategy coherence, time cycle of the process, concurrent projects, commitment, and so on. An overall score was assigned to each project to define a priority scale among the projects. Each project was approved by developing the project charter and then a stakeholders analysis helped the project team to mitigate risks and also to orient future choices.

The project involving machine learning was made possible thanks to the results of the other projects undertaken through the innovation program.

The primary deliverable of this project was the ability to process 1000 new free-text reports in 4 s. This result helped the CR to increase its performance: the CR serves the whole area of Eastern Sicily, where the estimated population, on 01-01-2021, amounted to 1,846,682, according to ISTAT. Related to the same area, the CR medical doctors manually identified 260,113 pathology reports as relevant in order to better qualify an oncologic case; additionally, 790,839 pathology reports were also defined as unuseful. The length of the reports varied from 22 to 5080 characters.

To better focus on the impact of the program, the list below cites the primary deliverables of the projects of the program:

- Introduction of Agile methodologies (SCRUM) into the Information Technology office;
- Refactoring the software, gaining operational speed, and reducing the database size by 20%;
- Introduction of versioning and a tracking platform for the software releases;
- Team-building sessions;
- Mapping of the overall process of the CR along with NVAA activities (see Figure 1);
- Training of the project team members while managing the project;
- Changes in the organizational structure, introducing the Project Manager role, and reassigning daily operations to different roles;
- Reducing the waste of talent and applying competencies in machine learning that were not identified among the employees of the CR;
- Increasing performance, from 6% to 100%, of the process designed to upload raster documents to the repository in 3 days (before the project, the waiting time was up to 50 days for single documents); Figure 2 shows how the variable Upload Time changed after the Improve phase.

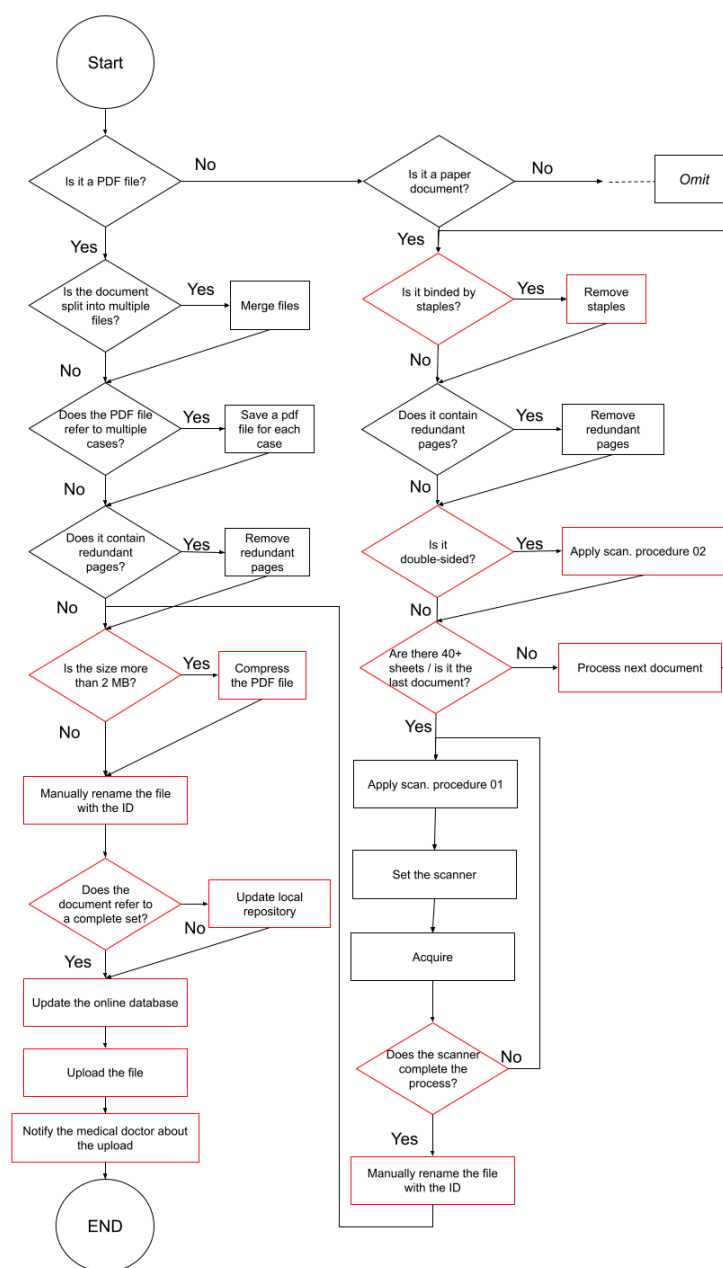


Figure 1. Example workflow from a project of the CR—red boxes represent NVAA.

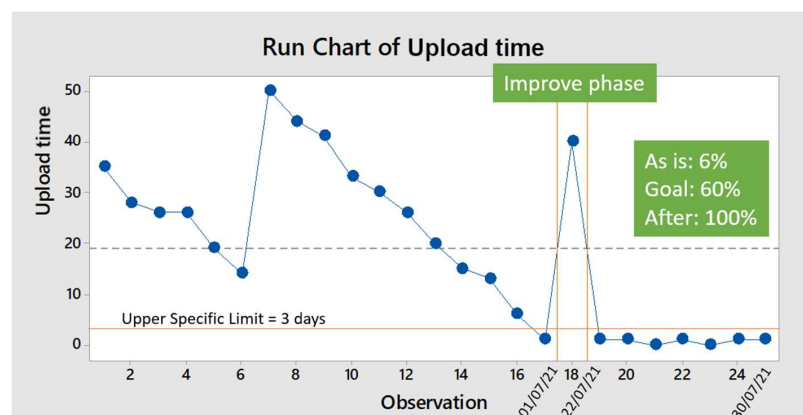


Figure 2. Run chart of the Upload Time variable from a project of the CR.

4. Discussion

The histopathological reports are written in natural language, and only a small part of the total annual histopathological reports contains relevant information for cancer evaluation. Filtering the right information is a key problem of the CR.

CR medical doctors retrieve a huge amount of pathology reports from Medical Pathology Services and they read all the documents to identify the relevant ones and then capture the information in order to analyze a tumor case. The reports help to identify new neoplastic cases, morphological and topological characteristics, the year of first diagnoses, and determine the histopathological confirmation percentage, which is an important indicator of quality [8,9]. Removing the wasted time spent by the operators in reading non-relevant documents is a primary goal of the operation of the CR, so the extraction of the information from free text represents an important improvement and it is also challenging due to the implicit nature of free text. [10].

The LSS project was able to elicit the needs of the operators of the CR and identify several solutions aligned with the strategy of the CR. Mapping the needs of the operators helped the Project Manager to identify a solution that increased the engagement of the operators and also of the ML specialist that deployed the first ML algorithm for the CR, with a confidence interval of 95%. It is important to also note that there were already many NLP systems developed based on English text that, due to the peculiarity of the language and the nature of free text, do not provide relevant results with text written in Italian [11,12]. The algorithm developed by the CR is able to process 1,000 reports in 4 seconds and then operators can be notified of the relevant information in order to correctly code the right case. This helps to reduce the cost of coding activities, not only because the time spent by a human to identify information is obviously more than the time spent by the algorithm but also because it is also more accurate than extracting information by using other automatic approaches, such as keyword-based [13,14] and rule-based classification that identify a pattern of words [15].

Furthermore, machine learning, along with related disciplines such as deep learning and big data, is rapidly expanding in healthcare and represents a field with high potential for its application [16–19].

The project management and the LSS methodology was applied as a driver of the improvement of the performance of the CR: knowledge and competencies were already possessed by the team; the project manager aimed to foster the team to perform together and express their potential while also managing the engagement of all the stakeholders involved in the project.

5. Conclusions

The described machine learning project helped the CR shorten the time spent by its operators and this was only one of the projects undertaken by the CR. In order to obtain the relevant impacts from a set of projects, they were organized as an incremental series of projects, where the newest one leveraged from the results—or simply acquired information from failure—of the predecessor projects.

Implementing a specialized office to continuously improve several processes, as described in this paper, with a professional Project Management Professional (PMP®) certification holder and Lean Six Sigma Black Belt with competencies in coaching, has been demonstrated as a good approach to improve the performance of the CR.

The LSS methodology also appears to fit well in the innovation program of the CR, and the health industry could be the right place to find great economies of scale.

Change management aims to manage the transition to a new state of organization that can be maintained. Pilot projects and proof of concepts are only the first steps toward changing economic institutions such as the organization the CR operates. Further studies could be focused on how to leverage the experience described in this paper to create a new way of working inside the institution and to make this kind of project and organizational approach the new operational standard.

The bottom-up approach is not effective enough to achieve a structural change such as the one described and it is hard to find awareness about the importance of the culture of project management inside the institution. Moreover, professionals with deep project management skills are perceived as “external” from the institution, limiting the opportunities to implement changes. For example, it is simple to state that for non-medical professionals involved in the program described above there are undefined career paths, underpaid profiles, and the absence of institutional roles that match the competencies deployed by the project manager [3] or LSS professional [2]. On the other hand, if an organization has been able to successfully apply the deep knowledge of its professionals, remove talent waste, and develop a machine learning project, such as ours involving histopathology reports [4], this kind of environment would also incentivize professionals to leave such organizations.

Author Contributions: Conceptualization, P.F.; methodology, P.F.; software, C.V.; validation, R.R. and M.F.; investigation, P.F.; data curation, A.B., E.S. and A.P.; writing—original draft preparation, P.F.; writing—review and editing, P.F. and R.R.; supervision, M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. The study did not require ethical approval. The present analysis was considered a secondary analysis of data already collected for the purposes of the RTI and, as such, it was covered by the same consent statements and privacy provisions of the RTI that were already obtained previously; therefore, a further request to the Ethics Committee was not deemed necessary under present national laws.

Informed Consent Statement: Patient consent was waived as the present analysis was a secondary analysis of data that was covered by the same consent statements and privacy provisions of the RTI that had previously been obtained.

Data Availability Statement: The data that support the findings of our study are available on request from the Integrated Cancer Registry (RTI CT-ME-EN). Further information is available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Michael, L.G. *Lean Six Sigma Combining Six Sigma Quality with Lean Production Speed*; Mc Graw Hill: New York, NY, USA, 2022.
2. *ISO 18404:2015; Quantitative Methods in Process Improvement—Six Sigma—Competencies for Key Personnel and Their Organizations in Relation to Six Sigma and Lean Implementation*. International Organization for Standardization: Geneva, Switzerland, 2015.
3. Project Management Institute. *A Guide to the Project Management Body of Knowledge PMBOK Guide*, 16th ed.; Project Management Institute: Newtown Square, PA, USA, 2017.
4. Viscosi, C.; Fidelbo, P.; Benedetto, A.; Varvarà, M.; Ferrante, M. Selection of diagnosis with oncologic relevance information from histopathology free text reports: A machine learning approach. *Int. J. Med. Inform.* **2022**, *160*, 104714. [[CrossRef](#)] [[PubMed](#)]
5. *ISO 13053-2; Quantitative Methods in Process Improvement—Six Sigma—Part 2: Tools and Techniques*. International Organization for Standardization: Geneva, Switzerland.
6. Japkowicz, N. Supervised Versus Unsupervised Binary-Learning by Feedforward Neural Networks. *Mach. Learn.* **2001**, *42*, 97–122. [[CrossRef](#)]
7. Anguita, D.; Ghelardoni, L.; Ghio, A.; Oneto, L.; Ridella, S. The ‘K’ in K-fold Cross Validation. In Proceedings of the ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning 2012, Bruges, Belgium, 25–27 April 2012; Available online: <https://www.esann.org/sites/default/files/proceedings/legacy/es2012-62.pdf> (accessed on 1 March 2022).
8. Hammami, L.; Paglialonga, A.; Pruneri, G.; Torresani, M.; Sant, M.; Bono, C.; Baili, P. Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing techniques: A rule-based approach. *J. Biomed. Inform.* **2021**, *116*, 103712. [[CrossRef](#)] [[PubMed](#)]
9. Fabacher, T.; Godet, J.; Klein, D.; Velten, M.; Jegu, J. Machine learning application for incident prostate adenocarcinomas automatic registration in a French regional cancer registry. *Int. J. Med. Inform.* **2020**, *139*, 104139. [[CrossRef](#)] [[PubMed](#)]
10. Burger, G.; Abu-Hanna, A.; de Keizer, N.; Cornet, R. Natural language processing in pathology: A scoping review. *J. Clin. Pathol.* **2016**, *69*, 949–955. [[CrossRef](#)] [[PubMed](#)]

11. Friedman, C.; Shagina, L.; Lussier, Y.; Hripcsak, G. Automated encoding of clinical documents based on natural language processing. *J. Am. Med. Inform. Assoc.* **2004**, *11*, 392–402. [[CrossRef](#)] [[PubMed](#)]
12. Savova, G.K.; Masanz, J.J.; Ogren, P.V.; Zheng, J.; Sohn, S.; Kipper-Schuler, K.C.; Chute, C.G. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 507–513. [[CrossRef](#)] [[PubMed](#)]
13. Spasić, I.; Livsey, J.; Keane, J.A.; Nenadić, G. Text mining of cancer-related information: Review of current status and future directions. *Int. J. Med. Inform.* **2014**, *83*, 605–623. [[CrossRef](#)] [[PubMed](#)]
14. Jouhet, V.; Defosse, G.; Burgun, A.; Le Beux, P.; Levillain, P.; Ingrand, P.; Claveau, V. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods Inf. Med.* **2012**, *51*, 242–251.15. [[PubMed](#)]
15. Osborne, J.D.; Wyatt, M.; OWestfall, A.; Willig, J.; Bethard, S.; Gordon, G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 1077–1084. [[CrossRef](#)] [[PubMed](#)]
16. Bongers, M.E.; Thio, Q.C.; Karhade, A.V.; Stor, M.L.; Raskin, K.A.; Calderon, S.A.L.; Schwab, J.H. Does the SORG algorithm predict 5-year survival in patients with chondrosarcoma? An external validation. *Clin. Orthop. Relat. Res.* **2019**, *477*, 2296. [[CrossRef](#)] [[PubMed](#)]
17. Yamaguchi, T.; Inoue, K.; Tsunoda, H.; Uematsu, T.; Shinohara, N.; Mukai, H. A deep learning-based automated diagnostic system for classifying mammographic lesions. *Medicine* **2020**, *99*, e20977. [[CrossRef](#)] [[PubMed](#)]
18. Pastorino, R.; De Vito, C.; Migliara, G.; Glocker, K.; Binenbaum, I.; Ricciardi, W.; Boccia, S. Benefits and challenges of Big Data in healthcare: An overview of the European initiatives. *Eur. J. Public Health* **2019**, *29* (Suppl. S3), 23–27. [[CrossRef](#)] [[PubMed](#)]
19. Davenport, T.; Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc. J.* **2019**, *6*, 94–98. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.