Challenges for Automation of Public Health Data Analysis

Ravi Shankar

Grenoble, France
rsps1001@gmail.com

Abstract

Advancements in Machine Learning and Data Science are not adequately reflected in how public health data is handled today. There is a visible gap between the advances in computing and medical sciences. In this position paper, we present an example of data science applied to the automation of a repetitive process within a cervical cancer screening program. We discuss the challenges for automating public health data and share our insights to elevate artificial intelligence (AI) in public healthcare.

Keywords

Public health, Data analysis, Cancer research, Automation

1. Introduction

More than 80% of the cervical cancer cases and deaths in a year occur in low medium income countries (LMICs) where prevention and cervical screening resources are limited [1][2]. Recent research studies have used machine learning models to support the initial phase of screening for detection of cancerous lesions using colposcopic images or cervicography[3][4]. These techniques require tech-savvy healthcare workers who are very scarce per capita in these countries.

We aim to build a user-friendly automation that would allow medical experts to diagnose cancerous tissues of the cervix in a short period of time while reducing costs and technical experience required. This idea will work by combining heath and AI researchers' expertise and experiences.

The main problem we aim to address is diagnosing biopsied women within a cervical cancer program. Our motivation is driven by the importance and time consumption of pathology process (i.e., pathologists reading histological slides). In the pathology process, women testing positive on screening tests are referred to specialised examination (colposcopy) to collect biopsy samples from the cervix and then haematoxylin and eosin (H&E) histological slides are prepared to be reviewed by pathologists using a microscope. This is an eye-dependent process, therefore inter- and intra-variability is present, and external revision is often needed as part of quality assurance (QA). The full process including QA may not be affordable, particularly in LMICs. Hence, AI contributes to eliminate such variability while saving time and resources.

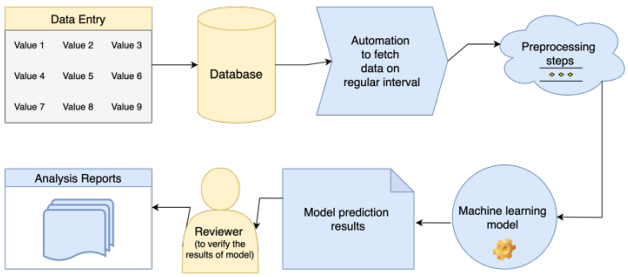

Figure 1: Project pipeline illustrating the automation of public health data analysis involving human reviewers who validate the Machine Learning model's prediction results.

Figure 1 illustrates the example of a proposed pipeline in which we aim to automate the steps from fetching of biopsy-based cervical data within a cervical cancer screening program. We then pre-process the fetched data, followed by training our machine learning (ML) model to make two or three prediction sets (ensuring QA) for human reviewers to validate, and finally generate the reports of the analysis. The current

Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland EMAIL: rsps1001@gmail.com (A. 1)

process (highlighted in yellow in Figure 1) excludes these steps for automation (highlighted in blue in Figure 1).

While the system works with the current process, the automation steps are currently done manually and repeatedly by a group of pathologists and statisticians. As the ML model does not exist in the current process, the analysis reports are produced after 2-3 stages of reviews involving multiple meetings to concur on the results. Including our proposed steps for automation in the current process will lower the burden of the experts and improve the timeframe up to 1/20 in comparison to the current process.

While our proposed project pipeline (Figure 1) forecasts optimal benefits for cervical cancer screening, in laying the groundwork, we were faced with critical challenges encompassing the realms of – technical, ethical, legal, and (most importantly) end user facing challenges. In this Workshop on "Healthy Interfaces (HEALTHI) 2022," we look forward to discussing our research on automation of public health data analysis. We hope to share our current challenges, methods, and future plans for AI powered healthcare.

2. Challenges for Automation of Public Health Data Analysis

In this section we generalise the problems we faced when implementing our project (Figure 1) to discuss the challenges for automation of public health data analysis:

1. **Trained Data Entry**: The first challenge is the considerable effort needed to change the conventional data entry practices. To automate, it is essential to construct database constraints, design helpful interfaces, and train non-tech savvy workers to log: complete, error free, and rightly formatted data.
2. **Patient Privacy**: Anonymising the data is important to preserving the privacy of personal health records of patients who sign up for the study. If possible, it should be mindfully made visible at the level of the interface to both the patients and their clinicians.
3. **Data Pre-processing**: Data pre-processing is the cleaning and preparation of data for the model and analysis tasks. This is a time-consuming underestimated challenge, if done improperly, it potentially hinders the performance and accuracy of the model and delays the overall study.
4. **Handling Large Datasets**: As public health studies are ongoing processes which include participants on a rolling basis, they can result in large datasets during overall period of the study (which might span several years). It is crucial to prepare for handling the data in batches for faster training of the model.
5. **Cross Validation by Experts**: Validation of results is a necessity with respect to training ML models. In healthcare-related data, cross validation by experts is much more important to prevent fatal diagnosis errors and to check for any potential biases in the model.
6. **Human Control**: It is important to have adequate human control so that the confidence of the predicted results is higher. Enabling human control via the automation process's interface allows to spot any discrepancies and malfunctioning.
7. **Transparency**: The interface should be made simple and transparent for both non-medical and other non-tech savvy stakeholders involved. The entire automation process should be comprehensible to all stakeholders involved for the project to succeed.
8. **Legal Efforts and Approval:** Last but the most important challenge is to succeed in the legal efforts and approvals required for the automation projects. Developing proof of concepts with publicly available datasets is one of the ways to prepare for the challenge of gaining legal approvals and other grants

3. Conclusion

AI powered public healthcare will foster a health structure in the future where the AI process drives the speed and accuracy of the diagnosis, treatment, and recovery. People will get the right diagnosis at the right time such that their treatment and recovery chances improve, thus improving chances of a good life. Furthermore, the cost efficiency brought by AI techniques will enable smart healthcare to be adapted to different healthcare structures in different countries, specifically in the low-income countries, so that healthcare becomes accessible and affordable there. This is a possibility only when AI researchers combine their expertise and experiences with health researchers. With this position paper we aim to contribute by informing both medical professionals and computer scientists of the challenges for automation of public health data analysis.

4. References

[1] Almonte, Maribel, Raúl Murillo, Gloria Inés Sánchez, Paula González, Annabelle Ferrera, M A Picconi, Carolina Wiesner, Aurelio Cruz-Valdéz, Eduardo Lazcano-Ponce, Jose Jeronimo, Catterina Ferreccio, Elena Kasamatsu, Laura Patricia Mendoza, Guillermo Rodríguez, Alejandro Calderón, Gino Venegas, Verónica Villagra, Silvio Alejandro Tatti, Laura Fleider, Carolina Terán, Armando Baena, María de la Luz Hernández, Mary-Luz Rol, Eric Lucas, Sylvaine Barbier, Arianis Tatiana Ramírez, Silvina Arrossi, Maria I. Rodriguez, E Díaz González, Marcela Celis, Sandra Martínez, Yuly Salgado, Marina Ortega, Andrea Verónica Beracochea, Natalia Pérez, Margarita M Rodríguez de la Peña, Maria de Sales Ramon, Pilar Hernández-Nevarez, Margarita Arboleda-Naranjo, Yessy Cabrera, Brenda Utrera Salgado, Laura García, Marco Antonio Retana, María Celeste Colucci, Javier A. Arias-Stella, Yenny Bellido-Fuentes, María Liz Bobadilla, Gladys Olmedo, Ivone Brito-García, Armando Méndez-Herrera, Lucía Cardinal, Betsy Flores, J F Márquez Peñaranda, Josefina Martínez-Better, Ana María Soilán, Jacqueline Figueroa, Benedicta Caserta, Carlos P. Sosa, Adrian A. Moreno, Juan Mural, Franco Doimi, Diana Giménez, Hernando Gutiérrez Rodríguez, Oscar Lora, Silvana Luciani, Nathalie Jeanne Nicole Broutet, Teresa M. Darragh and Rolando Herrero. “Multicentric study of cervical cancer screening with human papillomavirus testing and assessment of triage methods in Latin America: the ESTAMPA screening study protocol.” BMJ Open 2020 May 24;10(5): e035796. doi: 10.1136/bmjopen-2019-035796. PMID: 32448795; PMCID: PMC7252979.

[2] Bray F, Jemal A, Grey N, Ferlay J, Forman D. Global cancer transitions according to the Human Development Index (2008-2030): a population-based study. Lancet Oncol. 2012;13(8):790–801.

[3] Cho, BJ., Choi, Y.J., Lee, MJ. et al. Classification of cervical neoplasms on colposcopic photography using deep learning. Sci Rep 10, 13652 (2020). https://doi.org/10.1038/s41598-020-70490-4

[4] Liming Hu, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S Jaiswal, Brian Befano, L Rodney Long, Rolando Herrero, Mark H Einstein, Robert D Burk, Maria Demarco, Julia C Gage, Ana Cecilia Rodriguez, Nicolas Wentzensen, Mark Schiffman, An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening, JNCI: Journal of the National Cancer Institute, Volume 111, Issue 9, September 2019, Pages 923–932, https://doi.org/10.1093/jnci/djy225