Constructing a software platform by harnessing emerging data science tools for improved analytics and monitoring of female genital schistosomiasis and related health impacts

Background

Even amongst neglected tropical diseases (NTDs), a class of diseases that disproportionally affect the poorest nations in the world, female genital schistosomiasis (FGS) could be classed as one of the most neglected. FGS, a subset of schistosomiasis, is a chronically disabling gynaecological condition caused by the parasite *Schistosoma haematobium* [1]. The distribution and disease burden of FGS remains poorly understood, with few countries reporting nationwide statistics. However, the **best epidemiological estimates put the total number of cases at around 20-50 million people**, mainly within endemic areas of Sub-Saharan Africa [1]. Anywhere between 33-75% of girls and women living in *S. haematobium* endemic areas suffer from FGS [1].

The chronic sequelae align closely with those of sexually transmitted diseases and include pain and bleeding during intercourse, menstrual disorders, infertility, sexual dysfunction, and ectopic pregnancies [1]. These symptoms are caused by an inflammatory response to *S. haematobium* eggs and adult worms that can travel into the tissue of the cervix, uterus, ovaries, fallopian tubes, vulva, and vagina. FGS-associated coital bleeding has also been linked to an increase in the risk of HIV transmission in multiple studies with previous studies having suggested that controlling FGS is one approach to reducing HIV incidence in schistosome-endemic areas [1].

There is presently no way to distinguish between stages of FGS infection, and the disease progression dynamics are poorly understood. A deeper understanding of the temporal dynamics of FGS is key to developing strong control measures for such a ubiquitous parasite, where a single adult worm can live up to 30 years in the right conditions [2]. Knowledge of disease features across time is necessary to make conclusions on the impact of environmental and ecological drivers, mass drug administration efficacy and spatial distributions. From an environmental perspective, waterborne parasite transmission is very efficient when temperature and humidity are within ideal species limits [2]. Schistosomiasis transmission is known to be very focal; however, the FGS-specific spatial characteristics have yet to be accurately described. A recent study of schoolgirls in South Africa suggested there is some spatial association with certain waterbodies compared to others but noted several limitations to the spatial analysis [2].

Diagnosis of FGS involves one of the three following presentations on a colposcope image: single grains or clusters of 'sandy patches', 'sandy patches' as homogenous yellow areas, or rubbery papules [3]. However, accurate detection of these lesions is demanding, requiring expensive equipment and highly trained reviewers to read the images. Colposcopes, the cheapest models around USD$2000, require stable sources of electricity and sufficient infrastructure to produce the high-quality images needed for accurate diagnosis [1]. Rechargeable handheld colposcopes have recently been manufactured. However, due to the lower resolution images, they have a lower sensitivity and specificity for FGS [3].

Even with colposcopes, accurate diagnosis of FGS is difficult. The symptomology overlaps significantly with many common sexually transmitted diseases and cervicovaginal conditions [4]. Furthermore, even amongst expert reviewers there is significant disagreement in diagnosis, likely due to a lack of a diagnostic reference standard. A recent study of two experts discovered that there was a discordant diagnosis in 31.8% of cases [3]. This finding highlights the inadequate nature of human visual diagnosis using colposcopes.

In rural and remote areas especially, the combination of the lack of diagnostic infrastructure and inadequately trained staff has a significant follow-on effect on the accurate assessment of disease prevalence, distribution and, therefore, effective treatment and prevention strategies. There are also many gaps in understanding the relationship between disease progression, symptoms and severity. Furthermore, there are concerns that FGS is going under-detected in non-endemic regions or migratory communities where clinicians may be unfamiliar with the clinical signs of FGS. There is clearly a need to improve and support diagnosis, surveillance, and decision-making in communities where FGS is endemic. This project offers a simple and tangible solution in the form of a software platform that can read colposcope images and consider the risk profile of the patient to provide clinicians with an FGS diagnosis recommendation.

Aims

The goal of this project will be to develop an open-source, point-of-care software platform that combines convolutional neural networks (CNNs) trained on colposcope images and spatial and temporal risk profiles to assist in FGS diagnostic decision making. This aims of this project are as follows:

1. Train convolutional neural networks (CNN) on available high-quality colposcopy images to detect cases of FGS.
2. Investigate the applications of AI-powered image enhancement and pixel upscaling in colposcope images. Then, determine if using this AI improves the sensitivity and specificity of handheld alternatives to colposcopes (smartphones, digital cameras, handheld colposcopes),
3. Program a low-power, off-grid device to be used as an electronic diary to understand FGS's temporal dynamics and symptomology. This device will be used in an exploratory longitudinal FGS study of women already recruited within the *Schista!* Study in Zambia.
4. Investigate the spatial and temporal drivers of FGS then use Bayesian networks to develop risk profiles to strengthen FGS diagnosis recommendations.

Methodology

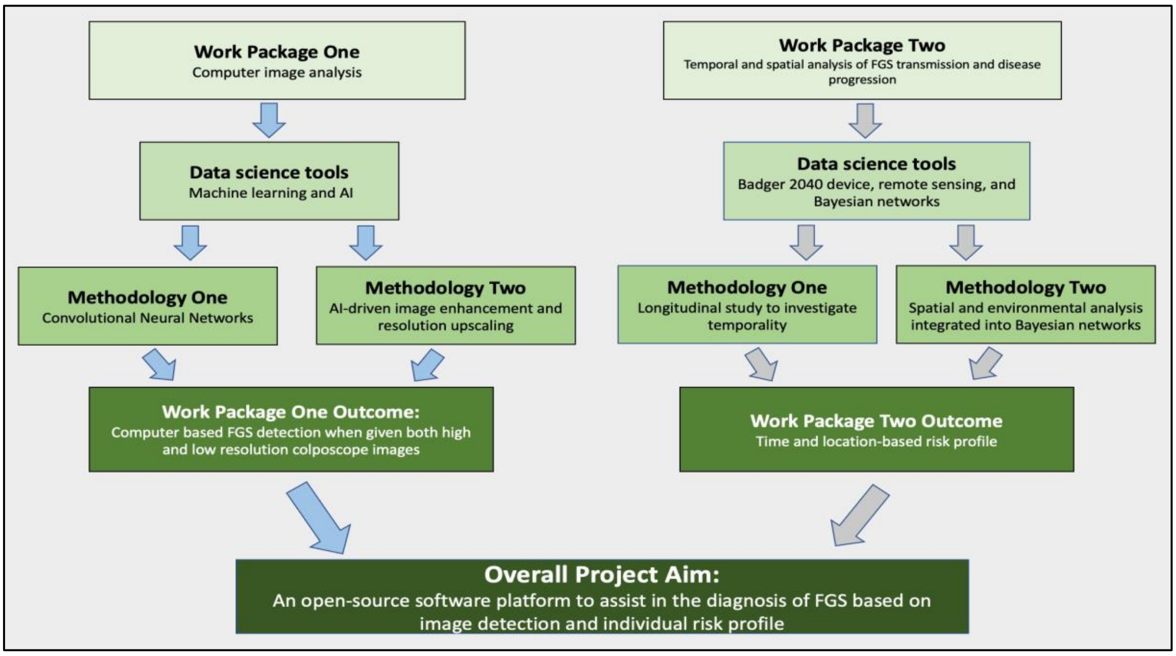

Figure 1: A conceptual framework of the multiple work packages needed to make the FGS diagnostic software platform a success.

Work package 1: Convolutional neural networks (CNN) are a cutting-edge deep learning method of object identification and image classification. In essence, they give computers the ability to 'see' images. Examples of open-source CNNs in action include coral reef monitoring (https://reefcloud.ai/), deforestation forecasting (https://github.com/PatBall1/DeepForestcast) and lung cancer screening with computer tomography images. For this project, CNNs will be trained using high-quality colposcopy images to support diagnostic decision-making using Tensorflow and Keras software. Holmen et al. have previously investigated the possibility of using computer image analysis with colposcope images and discovered that the characteristic cervical changes caused by FGS make it possible to train computers to detect the disease [4]. Holmen et al. used computer colour analysis, and the sensitivity was already 80.5%, with a specificity of 66.2%. The hope is that using CNNs will yield even higher results [4].

Machine learning is only as good as the data used for training and there is a risk that there may not be enough high-quality images to adequately train the CNN. In an attempt to increase the volume of training images, AI-powered image enhancement and pixel upscaling software, such as Topaz Labs Gigapixel AI or Let's Enhance.io, will be assessed for their accuracy in upscaling colposcope images. If the enhancement is found to be true to life, then these images will be fed into the CNN to increase the sensitivity and specificity of the software platform. This upscaling software may also assist with accurately processing lower resolution images so that handheld devices (smartphones, digital cameras, handheld colposcopes) can be used.

Work package 2: To support the work done by CNNs, data will be collected on additional variables to increase model sensitivity further. These variables will be temporal, spatial, and environmental and will be assessed for their association with FGS using various regression techniques and geographic information systems (Eg. ArcGIS). To decode the complex web of disease transmission, the explanatory variables will be integrated into a Bayesian network to produce a risk model, which will be integrated into the software platform for more accurate diagnostic classifications.

As there are significant gaps in the data available on spatial, temporal and environmental drivers of FGS, data will be collected as part of this project using the following:

a) Badger 2040 devices: These are currently in a prototype stage, and investigation is needed as to whether they can be used in public health settings. The Badger 2040 device is the size of a credit card, cheap (approximately £20.00 per device), ultralow power consuming and python programmable. These e-ink devices will be programmed to be used as electronic diaries and given to women as part of a longitudinal study, nested within the *Schista! Study* in Zambia, to analyse temporal dynamics and symptom progression of FGS. Data protection is paramount in this sensitive subject matter and the benefit of the Badger device is that entered data is converted into a text file and then a QR code, so data cannot be read without the correct electronic data management system.
b) Spatial and environmental analysis: open-source data warehouses will provide data to perform correlation analyses and explanatory variables will be used to build Bayesian networks. Using Bayesian networks means that each patient can be assigned an ecological risk profile for FGS transmission.

Project Outcomes and their Public Health Impact

Schistosomiasis is second only to malaria in its public health impact, and FGS is a *particularly* neglected form of the disease [1]. Safe, accurate, and low-cost diagnostic solutions for FGS are not readily available, and there are still gaps in understanding FGS symptoms and transmission dynamics. The work completed in this project will help improve the current FGS landscape by providing a point-of-care diagnostic software platform to be used by clinicians, including those not speciality trained to review colposcope images. Early and accurate detection of FGS not only means earlier treatment of the disease in an individual but also has positive ramifications for controlling community transmission in endemic areas. In doing this, significant progress is made towards two sub-targets of the UN's third sustainable development goal by eliminating epidemics of NTDs and AIDS by 2030 and ensuring universal access to sexual and reproductive health.

While this project aims to develop a point-of-care software platform, each stage of the program will have significant independent value. For example, once it is confirmed that a CNN can be trained to detect FGS, it could easily be trained to rule out other cervical conditions simultaneously. Additionally, investigating whether AI image enhancement software can accurately upscale colposcopy image resolution would mean that handheld colposcopes could be used more confidently. If AI enhancement is good enough, it could also mean that other handheld devices, such as modified smartphones or digital cameras, could be used instead of expensive colposcope devices. The results of this project will also foster a deeper understanding of FGS disease dynamics by recording the temporality of symptoms using the Badger 2040 electronic diaries. Furthermore, once these diaries are programmed, they can be used for a multitude of applications and data recording for various other diseases.

As the climate changes, so does the distribution of infectious diseases. The resultant Bayesian network risk models will help predict FGS distribution, which is known to follow a focal, high-intensity pattern. While FGS is almost entirely contained within Sub-Saharan Africa, it can also be found in The Middle East and recently has been found as far north as Corsica, France [1]. This spread in distribution will have significant implications for diagnosis and control in areas where FGS is presently not endemic. To make matters worse, climate and conflict-driven migration mean that women with undiagnosed or chronic FGS are migrating to areas with 1) naïve populations and 2) clinicians unaware of the signs and symptoms of FGS [1]. By having a point-of-care software platform available, which will hopefully be integrated into general cervical screening programs, clinicians will not need an in-depth understanding of the disease to come to an accurate FGS diagnosis. In addition, the ability to input different locations, which will carry different ecological risk profiles, is also attractive in climate change and high population mobility.

The need for greater diagnostic accuracy drives innovation, and the rapid expansion of technology and digital health in Sub-Saharan Africa makes this project an innovative, cost-effective, and viable option [5]. A brief look through the *WHO compendium of innovative health technologies for low-resource settings 2022* indicates that several successful rapid evaluation digital technologies are already being used in low-resource settings. For example, the WHO-backed 'Radify' radiological image analyser has successfully used a similar methodology to this project to create a device that can read chest x-rays. There are still significant barriers to accessing digital health solutions in rural and remote parts of SSA. For example, limited access to electricity and limited content in local languages. However, by creating a software tool that is used by clinicians, either on a mobile device or desktop, that is efficient with power use, there is the flexibility and room to make this platform a great success.

Candidate Suitability

In my pursuit of a PhD position, I am keen to undertake a methodology-heavy project where I can broaden and enhance my data science skills. I have a strong foundational knowledge that will support me during this project. For example, I use R, Python and other data science tools regularly in my current projects and have a keen interest in infectious diseases, particularly neglected tropical diseases (NTDs). My master's dissertation focussed on NTDs and developing location-based and ecological risk profiles, so I know the process very well. Outside of my university education, I have also successfully completed many online courses related to data analysis in public health. My ability to learn independently to meet the needs of a project makes me confident that I am ready to meet the technical challenges of a PhD.

My current work as [deleted] at [deleted] is focused on overcoming NTD surveillance and diagnostic challenges in Africa using data science, and the overlap with the content of this PhD is significant. Through this work, I have developed an understanding of how to work creatively and innovatively within the financial and logistic constraints of NTD research. Furthermore, I have become hyper-aware of the importance of developing open-source tools and data in infectious diseases, not least because a large amount of my current research includes using open-source, remotely sensed data.

I am continually inspired by the work being done in NTD research in the UK and overseas, and I know that collaboration with in-country teams is of the utmost importance. My work with [deleted] in [deleted], supporting local institutes in setting up [deleted], has provided me with valuable experience liaising with key in-country stakeholders and constantly adapting to the needs of the local population. Finally, I have huge career ambitions and aim to [delete]. My ultimate goal is to use novel technologies to support the eradication of NTDs through improved diagnostics and surveillance.

References

1. Bustinduy AL, Randriansolo B, Sturt AS, Kayuni SA, Leutscher PDC, Webster BL, et al. Chapter One - An update on female and male genital schistosomiasis and a call to integrate efforts to escalate diagnosis, treatment and awareness in endemic and non-endemic settings: The time is now,. Adv Parasitol [Internet]. 2022 [cited 2023 7 Jan]; 115:[1-44 pp.].
2. Livingston M, Pillay P, Zulu SG, Sandvik L, Kvalsvig JD, Gagai S, et al. Mapping Schistosoma haematobium for Novel Interventions against Female Genital Schistosomiasis and Associated HIV Risk in KwaZulu-Natal, South Africa. Am J Trop Med Hyg. 2021;104(6):2055-64.
3. Sturt A, Bristowe H, Webb E, Hansingo I, Phiri C, Mudenda M, et al. Visual diagnosis of female genital schistosomiasis in Zambian women from hand-held colposcopy: agreement of expert image review [version 1; peer review: awaiting peer review]. Wellcome Open Res [Internet]. 2023 [cited 2023 12 Jan]; 8(14).
4. Holmen SD, Kleppa E, Lillebø K, Pillay P, Lieshout Lv, Taylor M, et al. The First Step Toward Diagnosing Female Genital Schistosomiasis by Computer Image Analysis. Am J Trop Med Hyg [Internet]. 2015 [cited 2023 7 Jan]; 93(1):[80-6 pp.].
5. Ibeneme S, Karamagi H, Muneene D, Goswami K, Chisaka N, Okeibunor J. Strengthening Health Systems Using Innovative Digital Health Technologies in Africa. Front Digit Health [Internet]. 2022 [cited 2023 Jan 7]; 4(854339):[1-8 pp.].

TOTAL WORD COUNT = 2545
(Excludes headers)