Requirements Engineering of a Herpes Simplex Virus Patient Registry: Alpha Phase

Svitlana Surodina
Skein LTD
Ching Lam
University of Oxford
Svetislav Grbich
Skein LTD
Madison Milne-Ives
University of Oxford
Michelle van Velthoven
University of Oxford
Edward Meinert (e.meinert14@imperial.ac.uk)
Imperial College London https://orcid.org/0000-0003-2484-3347

Software

Keywords: Data Collection, Herpes Simplex, Registries, Machine Learning, Risk Identification, Artificial intelligence, Medical diagnosis, Medical services

Posted Date: July 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-38387/v1

Requirements engineering of a Herpes Simplex Virus patient registry: Alpha phase

Svitlana Surodina[1], Ching Lam[2,3], Svetislav Grbich[1], Madison Milne-Ives[2], Michelle van Velthoven[2], and Edward Meinert[2,4]

[1]Skein LTD, Kemp House, 152-160 City Road, London, EC1V 2NX

[2]Digitally Enabled PrevenTative Health (DEPTH) Research Group, Department of Paediatrics, University of Oxford, Oxford, UK, OX3 9DU

[3]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK, OX3 7DQ

[4]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, UK, W6 8RP

Corresponding author: Edward Meinert (e-mail: edward.meinert@paediatrics.ox.ac.uk).

This work was supported in part from the European Institute of Innovation and Technology (EIT) Health (Grant 18654), which is supported by the EIT, a body of the European Commission. EM is supported by the Sir David Cooksey Fellowship at the University of Oxford.

ABSTRACT

Background. Collecting data from people with herpes simplex virus is challenging because of poor data quality, low user engagement, and concerns around stigma and anonymity. This project aimed to improve data collection for a real-world HSV registry by identifying predictors of HSV infection and selecting a limited number of relevant questions to ask new registry users in order to determine the HSV infection risk group.

Methods. The US National Health and Nutrition Examination Survey (NHANES, 2015-16) database has confirmed HSV1 and HSV2 status of American participants (14-49 years) as well as a wealth of demographic and health-related data. Two datasets – for HSV1 and HSV2 – were formed using this database, and an anonymous lifestyle-data based questionnaire with a Random Forest algorithm was devised using Python. The algorithm was optimised to reduce the number of questions and to identify risk groups for HSV. Data was split into subsets to train and test the model.

Results. The model selected a reduced number of questions from the NHANES questionnaire that predicted HSV infection risk with high accuracy scores of 0.91 and 0.96 and high recall scores of 0.88 and 0.98 for HSV1 and HSV2 datasets, respectively. The number of questions was reduced from 150 to an average of 40, depending on age and gender, that together provides high predictability of the infection

Conclusions. This machine-learning algorithm for risk identification of people infected with HSV can be used in a real-world evidence registry to collect relevant lifestyle data. A current limitation is the absence of real user data and integration with electronic medical records that would enable model learning and improvement. Future work will explore model adjustments, anonymisation options, explicit permissions and standardised data schema that meet GDPR, HIPAA and third-party interface connectivity requirements.

Keywords Data Collection, Herpes Simplex, Registries, Machine Learning, Risk Identification, Artificial intelligence, Medical diagnosis, Medical services

I. BACKGROUND

Patient data in medical registries is an important source of information for screening, treatment and research purposes. However, lack of high-quality data is a problem, especially in stigmatised diseases like Herpes Simplex Virus (HSV). Other challenges of data collection from people living with HSV also include low user engagement and concerns around stigma and anonymity. Various lifestyle predictors for HSV infection (e.g. sexual activity, number of partners) and recurrences (e.g. diet, exercise, sleep) can be used to select the most relevant questions for an HSV registry and improve data collection and analysis.

Machine learning tools and Artificial Intelligence (AI) techniques are widely used in medical research. Machine learning has previously been applied to medical diagnosis and patient data insights in oncology [1], for the diagnosis of heart, liver, diabetic, dengue, and hepatitis diseases [2], to predict suicidal behaviour using longitudinal electronic health record (EHR) data [3], and to classify whether patients have Alzheimer's disease [4]. However, current applications of machine learning are usually limited to pre-existing, structured datasets and do not address the problems of first-person data collection from patients and the resulting limitations of data completeness, quality, and validity.

To resolve these challenges, a patient registry solution using machine learning was designed. Machine learning-based systems are particularly useful for generating new knowledge and insights without having a-priori hypotheses. The concept of 'data farming' or 'evidence farming' addresses the problems of collecting data directly from users [5]. Data can be 'organic' - grown and harvested if suitable environmental conditions are provided. Websites and online platforms where patients provide their own data often meet these conditions. These data can include medical information like disease diagnosis and laboratory results, medications the patients are taking, and subjective data from self-report questionnaires. One example of the data-farming paradigm is the online enterprise PatientsLikeMe.com.

Based on data generated in this way, machine learning models can be developed, in particular decision trees can be applied to analyse the flows of user-generated content. Decision trees are decision support models that are used to determine the strategy to reach a certain goal that is the most efficient and the most likely to be successful [6]. Application of decision tree analysis is widespread, but is mostly used in a basic way on highly-structured data, without applications in real-time data collection situations [7].

Machine learning methods such as Random Forest have great potential in the field of Real World Data (RWD)[8]. Large 'data lakes' have been created by aggregating information from hospital EHRs, including

unstructured and semi-structured data generated by questioning patients. This can be explored with machine learning methods to identify clinically meaningful patterns. These data lakes can be used to track patients longitudinally, providing a large body of data that wouldn't be available in the typical randomized controlled trial. By using RWD and tracking patients longitudinally, the necessity for certain traditionally conducted late-phase trials could be reduced or eliminated altogether. Drugs that have successfully passed phase I and II trials, and have evidence supporting their efficacy and safety, could be given to patients and their real-world experiences could be tracked [9]. Additionally, machine learning methods on RWD could be used to easily identify potentially eligible patients for clinical trials, depending on certain criteria.

CHALLENGES OF PATIENT REGISTRIES

A previous analysis of the challenges medical registries are facing, and user/patient and researcher use cases, identified several problems [10]:

- A common problem in digital data collection in general is that to collect sufficient data, the patient questionnaire tends to be long, which means there is often a high drop-out rate. However, a large amount of data points on demographic and lifestyle can potentially result in better insights and allow for more targeted clinical trial selection and recruitment. Therefore, a critical challenge for patient data collection is to reduce the amount of data entered by the users while maximising the usefulness, quality, and information content of the collected data.
- Any patient registry needs to be designed with the patient and their experience in mind (patient-centricity), meaning that the data collection process serves the expectations and needs of the patients.
- Selection bias during direct data collection can become more profound since there would be a subset of users who are more likely to complete an extensive questionnaire, including those who are more computer-literate, have more time, have more frequent and/or severe symptoms, and/or are keen to be informed of relevant clinical trials.
- Another common user concern is about privacy and control over data. This is especially true for sensitive and/or personally identifiable information.

This project aimed to develop an innovative machine learning method for patient data collection and predictive analytics to address problems of data availability and quality in medical registries. This approach can be applied to HSV patient registries and more broadly to user data collection for improved clinical and population health insights.

II. METHODS

A. OVERVIEW OF PROPOSED SOLUTION

This project designed an algorithm to optimize data collection questionnaires for an HSV patient registry. Integrating a decision tree-based technology into a patient registry can reduce the number of questions users have to answer while increasing the data content and reducing informational entropy for each user's record. A pre-existing dataset was used to train the model and the question sets were optimised to generate the most complete information to screen users who are prone to HSV and collect initial user data in an efficient manner. For non-diagnosed users, the model identifies risk groups based on the questions and received answers, and the answers provided by users who were clinically diagnosed with HSV can be used to further train and improve the model.

HSV screening was chosen as the goal for the prototype solution with an aim to provide an efficient way of engaging users who otherwise might not be willing to participate in research or log recurrence data. The solution schema is represented in Fig. 1.

Two types of patient users ('External Users') were considered when designing the system: new and returning, and an HSV researcher (Internal User). Since the barriers are most pronounced when an external user starts using the system for the first time, an HSV preliminary screening tool can serve as an incentive to answer the list of questions and get pre-diagnosed. From the homepage, new users are guided through an anonymous questionnaire that is optimised to reduce information entropy, minimise questions and their complexity whilst obtaining the maximum amount of information.

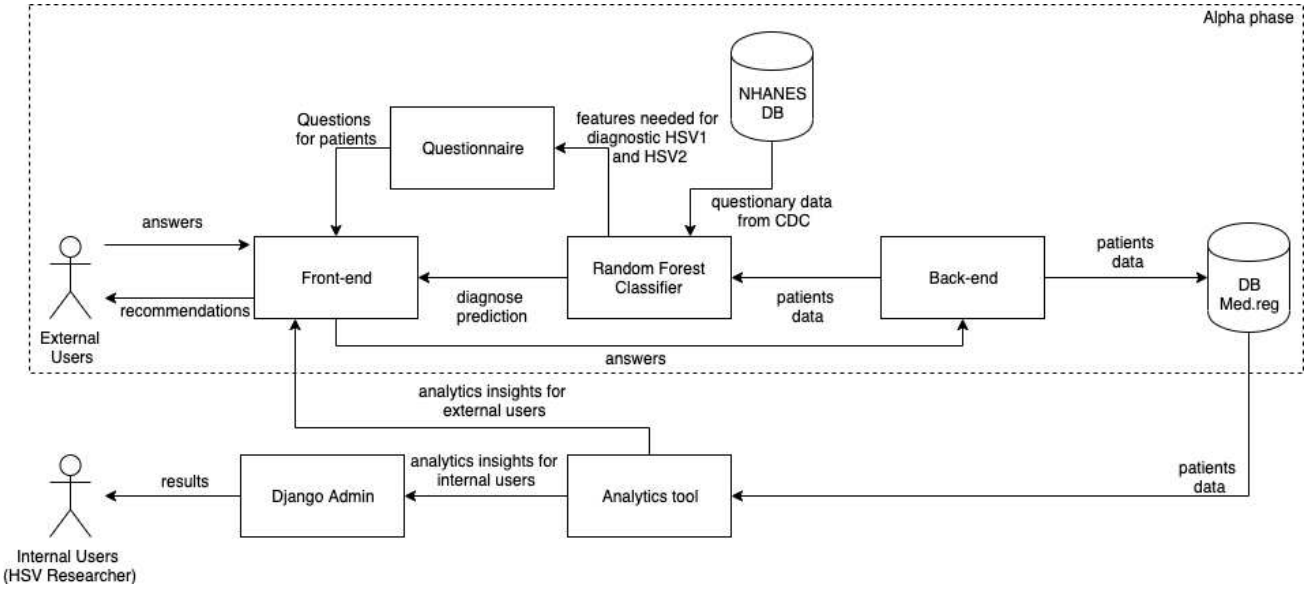

FIGURE 1. Schema of the technological solution.

B. SELECTION OF A DATABASE TO TRAIN AND TEST THE MODEL

To develop and train the model, an initial dataset that met the following requirements was needed:

- Open access, available without application or payment
- A clinically verified HSV diagnosis
- Large number of rows with valid HSV diagnostic data (over 1000)
- Cross-referenced interviews and physical examination data
- Extensive list of demographics, lifestyle, dietary, diagnostic data
- High density of data points, meaning that most of the data fields are populated
- Verifiable data quality and reliability

The initial search was conducted via Google and patient registry lists and included various sources dedicated to Human Herpes Simplex Virus and patient database catalogues. The data found on HSV patients consisted mostly of laboratory measurement data, which was not applicable for building a lifestyle-focussed questionnaire. The US National Health and Nutrition Examination Survey (NHANES) was the only one of the identified databases that met all of the above requirements. Its data is public domain and can be used freely without obtaining copyright permission [11].

C. DATABASE USED

NHANES, as a major program of the US National Centre for Health Statistics (NCHS), provided high density population data, gathered by high quality standards. It was created by a US government agency and detailed methodology and data provenance [12]. The dataset used was representative of the US population in 2015-2016 [13]. Moreover, it collected demographic information, enriched by detailed dietary, examination and laboratory data, all interconnectable via unique participant IDs.

The survey is unique in that it combines interviews and physical examinations. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

All NHANES participants visited a physician. Dietary interviews, body measurements, blood sampling and dental screening were included for all participants. Depending on the age of the participant, the rest of the examination included tests and procedures to assess the various aspects of health. HSV diagnosis was confirmed using diagnostic tests by physicians.

D. HSV1 AND HSV2 DATASETS

The demographics of the validation datasets HSV1 and HSV2 can be found in Fig. 2. Males and females 14-49 years old were represented in the dataset. HSV1 dataset contains data of 3,386 participants with 1,840 diagnosed with HSV1 and 1,546 confirmed negative. HSV2 dataset contains data of 2,813 participants with 478 positive and 2,335 negative for HSV2. The data on these participants, together with demographic information, formed the initial dataset.

LBXHE1 - Herpes Simplex Virus Type 1

Variable Name: LBXHE1

SAS Label: Herpes Simplex Virus Type 1

English Text: Herpes Simplex Virus Type 1

Target: Both males and females 14 YEARS - 49 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Positive	1840	1840	
2	Negative	1546	3386	
3	Indeterminate	7	3393	
.	Missing	317	3710	

LBXHE2 - Herpes Simplex Virus Type 2

Variable Name: LBXHE2

SAS Label: Herpes Simplex Virus Type 2

English Text: Herpes Simplex Virus Type 2

Target: Both males and females 18 YEARS - 49 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Positive	478	478	
2	Negative	2335	2813	
3	Indeterminate	2	2815	
.	Missing	895	3710	

FIGURE 2. **Demographics of validation dataset [10]**

A complete dataset with questionnaires and results for the period 2015-2016 is available on the National Center for Health Statistics website [12]. The overall initial list of questions is listed in Additional file 1 and comprises over 600 questions. For the model, a smaller number of questions – around 150 – were selected (Appendix 2). These were chosen because many questions didn't have enough cross-referenced answer entries for data analysis.

E. TRAINING AND TESTING SUBSETS

Using the data science method train_test_split from the sklearn Python library, the confirmed negative or positive cases reported in NHANES were divided into two sub-datasets for training and validation of the model with a ratio of 0.8 to 0.2. The training dataset was used to train the model and the validation dataset was used for accuracy scoring. A threshold of 0.01 was experimentally defined to keep the list of questions as short as possible whilst maintaining good accuracy of the model). After checking feature importance values and determining that max_depth=9 gives the threshold of feature importance less than 0.01, questions below the threshold were considered less relevant and were excluded.

F. TOOLS AND TECHNOLOGY STACK FOR MODEL

A CART (Classification and Regression Trees) Random Forest (RF) model was used to generate the main questionnaire. XGboost approaches were also reviewed, but RF performed better than XGBoost and with less complexities of implementation in production. Due to high transparency and interpretability of CART models, a sequence of decision trees bagged into Random Forest ensemble were chosen. The average decision tree plot, together with feature importance, was used to explore the full list of questions and define the shortest chain of interdependencies leading to HSV screening with the highest probability of accuracy.

The Random Forest ensemble was built from a sequence of decision trees using a bagging method [14]. Bagged decision tree ensembles are used to define entropy and information gain from previously selected features or discriminants [15]. Binary splitting on features with maximal informational gain leads to fewer nodes in the trees (i.e. fewer relevant questions for diagnosing HSV).

The model was designed to process the data in the following way:

1. the initial dataset was divided into two subsets based on HSV type (1 or 2);
2. these datasets were split by the data science method train_test_split from the sklearn Python library to form training and validation subsets;
3. the HSV1 and HSV2 training sets were processed by Random Forest Classification estimators;
4. accuracy on the training datasets was optimized by tuning the max_depth parameter (controls the total depth of the tree, i.e. number of binary splitting levels);
5. accuracy was checked with validation subsets;
6. values of max_depth gave the threshold of sufficient feature importance, and all questions below that level were excluded;

7. the final questions became the exhaustive list of features for the trained Random Forest Classifier and used for the screening tool;
8. given real-life data (questionnaire responses and clinical diagnosis verification), the model can improve its precision.

G. ILLUSTRATION OF THE RANDOM FOREST TREE

The decision tree (here bagged into Random Forest ensemble) does sequences of binary splitting (splitting the sets of questions into two subgroups that produce the greatest distinction between positive and negative HSV diagnoses) until the resulting number of splittings is sufficient to explain the general tendencies of the dataset (until the model has learned hidden patterns in data). Splittings are performed on the most informative feature, i.e. the data feature having the highest information gain. In this particular case, a depth of 9 hierarchical levels of splitting was enough for the model to learn the connections between the data features and HSV1/HSV2 diagnosis.

The decision trees below show the final iteration of the Random Forest training process. The best results were achieved after 9 levels of branching, and further learning (splitting) would bring no meaningful improvements.

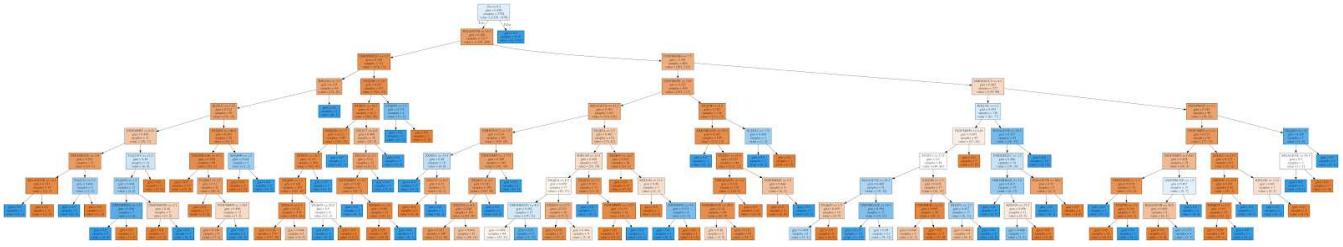

FIGURE 3. **Random Forest training process final decision tree for HSV1 testing subset.**

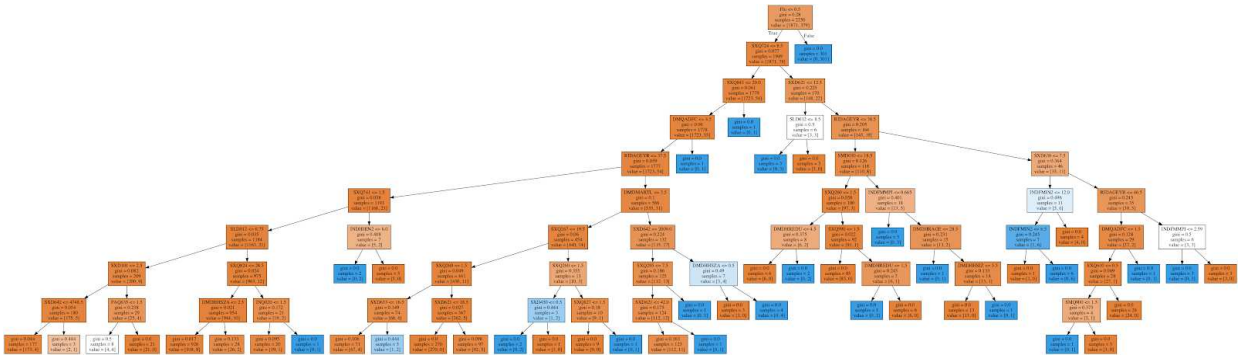

FIGURE 4. **Random Forest training process final decision tree for HSV2 testing subset.**

H. MODEL EVALUATION METRICS

The reduction in the number of questions needed to achieve high accuracy is an important success metric of the model. It can show the feasibility of using the model to reduce information entropy and encourage participants to

complete the questionnaire. To validate the performance of the model, two key metrics were used: accuracy and recall score. Accuracy is the overall precision of the model in identifying HSV positive patients from the questionnaire, whereas recall is a measure of the model's capability to identify true positives.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{True\ Positive}{Total\ Actual\ Positive}$$

FIGURE 5. **Recall formula.** [16]

Recall score is an important metric because it is preferable to identify a non-infected as being in the risk group, than vice versa. The lowest possible score is 0 (0%), the highest is 1 (100% probability of true prediction).

III. RESULTS

A. STAGE 1

The initial results of the Random Forest model computations are outlined in Table 1. As a result of the first stage of model development and testing, the number of questions was reduced from 150 to 62. The model selected a set of 62 questions that form shorter sequences for each user, based on their age and gender. On average, a user would be asked 40 questions, with a minimum of 21. The final list of questions can be found in Additional file 3.

B. STAGE 2

The lower accuracy score on the HSV1 dataset than the HSV2 dataset in Stage 1 meant that latent features inherited from NHANES survey were not strong enough as a predictor for HSV1 (Table 1). For example, the dataset had limited symptom-related interview questions. This led us to more literature research on HSV1 and HSV2 symptoms to introduce and test additional features to the model (additional questions). A sample feature was added into the model on the assumption that a significant proportion of people infected with HSV1/HSV2 virus types (up to 80% for various gender and virus types combinations in various sources [15]) may experience more general symptoms like fever, muscle aches and nausea. Therefore, a new question was added to the questionnaire, both for HSV1 and HSV2 types (*"Is your general feeling of discomfort or illness followed by one or more symptoms: fever, nausea, headaches, muscle pain, swollen lymph nodes or malaise?"*), and engineered the additional feature for the dataset, with a positive label in the 80% cases of the infected population.An additional question about symptoms with high presence in HSV-infected people was introduced, and resulted in improved the scores of the Random Forest model to train and test data predictions (Table 1).

TABLE I

STAGE 1 AND 2 ACCURACY AND RECALL SCORES

	Stage 1		Stage 2	
	Accuracy	*Recall*	*Accuracy*	*Recall*
HSV 1	0.61	0.83	0.91	0.88
HSV 2	0.83	0.90	0.96	0.98

Once the tool is in operation and is collecting real world data – a significant number of participants answer the questionnaire and their results are confirmed clinically - the model will gradually verify whether flu-like symptoms are in fact a strong predictor of HSV. The model will re-adjust the scoring method to exclude it as an important factor if this question turns out not to add much information.

IV. DISCUSSION

The possibility of using optimisation algorithms to minimise the number of user-generated data points needed to accurately assess risk of HSV1 and HSV2 infection was successfully tested. The ultimate aim is to increase the quality and quantity of data collected and improve the probability of users disclosing their personally identifiable information and volunteering for clinical trials. The system was prototyped on the publicly available data of a small population of US citizens published in the NHANES database [10]. The next step is to integrate it into an independent backend module and connect it with the question-outputting and answer-collecting front-end to further improve the model and test its self-improvement capabilities on real world data.

The second stage illustrates how the same procedure could be repeated when more user data are added into the model (such as other symptoms with their distribution) and how a better accuracy score of the model achieved for the model if certain questions are strongly linked to HSV. . Similarly, other assumptions of HSV can be tested on real users by including the relevant variables into the questionnaire and model. Once real users start using the screening tool and the tool results are verified by a clinician through diagnostic tests, the model will self-learn and verify such assumptions

The model could also be improved by integrating more user data from electronic health records, to generate more insights on what questions can be more predictive of risk group.

A. FUTURE DIRECTIONS

Some of the additional functionalities that could be considered for the future research and system improvement are:

- Multi-class classification, where HSV1 and HSV2 data would be treated simultaneously. This way the machine learning assembling would help researchers find the additional patterns in HSV patients' habits in the generated response database.
- Add descriptive user segmentation to the model. By defining the most recurring patient behaviour and their profile type, the probability of gathering more relevant data could be improved.

Anonymisation options, explicit permissions, and standardised data schema that addresses GDPR, HIPAA and third-party interface (such as FHIR) will help create a platform based on the HSV screening tool. This could be used by researchers to complement clinical research, ease patient recruitment for clinical trials, and engage users with further data sharing and obtaining health insights.

After the screening tool results are received, the user could decide whether or not to be registered and added to the database. If they decide not to be added, their responses are saved anonymously in the database and can be used as an additional source of insights into populations that would not provide any data that required registration (addressing selection bias).

After registration, users could get access to dashboards that help them track their health and give personalised insights, news and advice. Consent processes would allow users to agree to notifications, give consent for trial involvement and use of data, and get in touch with a clinician. The user responses will be recorded in the database and the model will self-improve as each HSV screening tool result is confirmed by a valid diagnostic method.

Researchers would need to register and get verified by the system admin and login to their account before accessing pseudo-anonymised data. They would have the capability of filtering user data, sending invitations to trial and research groups, and creating further questions to identify trial eligibility. The various user flows identified for the platform are listed in Additional file 4.

B. SUGGESTED SUCCESS METRICS FOR FUTURE WORK

According to Gov.uk Service Design guidelines, the beta stage of the project will introduce and track key quantitative metrics of system performance. This tracking should include the following key metrics:

- Conversion rate: patient visits the registry to the start of the questionnaire
- Questionnaire start to completion: drop-off rate
- Completion to sign up to share personal data
- Number of users who completed the screening questionnaire

These metrics can be tracked by integrated database analytics and Google Analytics, which will also be important for accumulating user behaviour data for future analytics and development.

C. CONCLUSION

This project successfully developed, trained, and tested a model to predict risk of HSV1 and HSV2 infection based on an optimized set of demographic, lifestyle, and symptom questions. This machine learning determination of the questions with the best predictive value means that fewer questions need to be asked of patients who are completing patient registry surveys. One limitation of the project is that it used pre-collected survey data, and the model has not yet been trained and tested on real user data in the context of a patient registry or online questionnaire. Future research will address this area to improve the model and important anonymity, consent, interoperability, and data security concerns will be examined.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The datasets generated and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS], included in this published article [and its supplementary information files].

Competing interests: The authors declare that they have no competing interests. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Funding: This work was supported in part from the European Institute of Innovation and Technology (EIT) Health (Grant 18654), which is supported by the EIT, a body of the European Commission. EM is supported by the Sir David Cooksey Fellowship at the University of Oxford.

Authors' contributions: C.L. and E.M. conceived the study topic. S.S. and S.G. conducted the research and prototyped the model. S.S. prepared the first draft of the paper with revisions from M.v.V., E.M., C.L., M.M.-I. and S.S. All authors read and approved the final manuscript.

Acknowledgements: Not applicable

Additional files

Additional file 1. Questionnaire Data.doc *Initial set of questions extracted from NHANES*

Additional file 2. Questions for the model.doc *Questions selected to develop and train the model*

Additional file 3. Resulting questions.doc *The questioning algorithm and the question list selected by the model*

Additional file 4. User flows.pdf *The schema of user journey and interactions with the system by several user persona groups.*

REFERENCES

1. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI: **Machine learning applications in cancer prognosis and prediction**. *Comput. Struct. Biotechnol. J.* 2015, **13**:8–17.

2. Fatima M, Pasha M: **Survey of Machine Learning Algorithms for Disease Diagnostic**. *JILSA* 2017, **09**:1–16.

3. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, Nock MK, Smoller JW, Reis BY: **Predicting Suicidal Behavior From Longitudinal Electronic Health Records**. *Am. J. Psychiatry* 2017, **174**:154–162.

4. Ray S, Britschgi M, Herbert C, Takeda-Uchimura Y, Boxer A, Blennow K, Friedman LF, Galasko DR, Jutel M, Karydas A, Kaye JA, Leszek J, Miller BL, Minthon L, Quinn JF, Rabinovici GD, Robinson WH, Sabbagh MN, So YT, Sparks DL, Tabaton M, Tinklenberg J, Yesavage JA, Tibshirani R, Wyss-Coray T: **Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins**. *Nat. Med.* 2007, **13**:1359–1362.

5. Oquendo MA, Baca-Garcia E, Artés-Rodríguez A, Perez-Cruz F, Galfalvy HC, Blasco-Fontecilla H, Madigan D, Duan N: **Machine learning and data mining: strategies for hypothesis generation**. *Mol. Psychiatry* 2012, **17**:956–959.

6. Lee J, Jung Y, Shin S, Yoon T: **Analysis of HSV-1 and HSV-2 that cause herpes simplex with Apriori algorithm, decision tree, and support vector machine**. In *2017 19th International Conference on Advanced Communication Technology (ICACT)*. 2017:679–684.

7. Rau C-S, Wu S-C, Chien P-C, Kuo P-J, Chen Y-C, Hsieh H-Y, Hsieh C-H: **Prediction of Mortality in Patients with Isolated Traumatic Subarachnoid Hemorrhage Using a Decision Tree Classifier: A Retrospective Analysis Based on a Trauma Registry System**. *Int. J. Environ. Res. Public Health* 2017, **14**.

8. Panesar A: *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*. Apress; 2019.

9. Shah P, Kendall F, Khozin S, Goosen R, Hu J, Laramie J, Ringel M, Schork N: **Artificial intelligence and machine learning in clinical development: a translational perspective**. *NPJ Digit Med* 2019, **2**:69.

10. Surodina S, Lam C, de Cock C, van Velthoven M, Milne-Ives M, Meinert E: **Engineering Requirements of a Herpes Simplex Virus Patient Registry: Discovery Phase of a Real-World Evidence Platform to Advance Pharmacogenomics and Personalized Medicine**. *Biomedicines* 2019, **7**:100.

11. **Use of Agency Materials** [https://www.cdc.gov/other/agencymaterials.html].

12. **NHANES Questionnaires, Datasets, and Related Documentation** [https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2015].

13. **NHANES Response Rates and Population Totals** [https://wwwn.cdc.gov/nchs/nhanes/responserates.aspx].

14. **Ensemble methods** [https://scikit-learn.org/stable/modules/ensemble.html].

15. **Entropy: How Decision Trees Make Decisions** [https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8].

16. Powers DM: **Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation (Tech. Rep.)**. *Adelaide, Australia* 2007.

17. Kimberlin DW, Rouse DJ: **Clinical practice. Genital herpes**. *N. Engl. J. Med.* 2004, **350**:1970–1977.

Figures

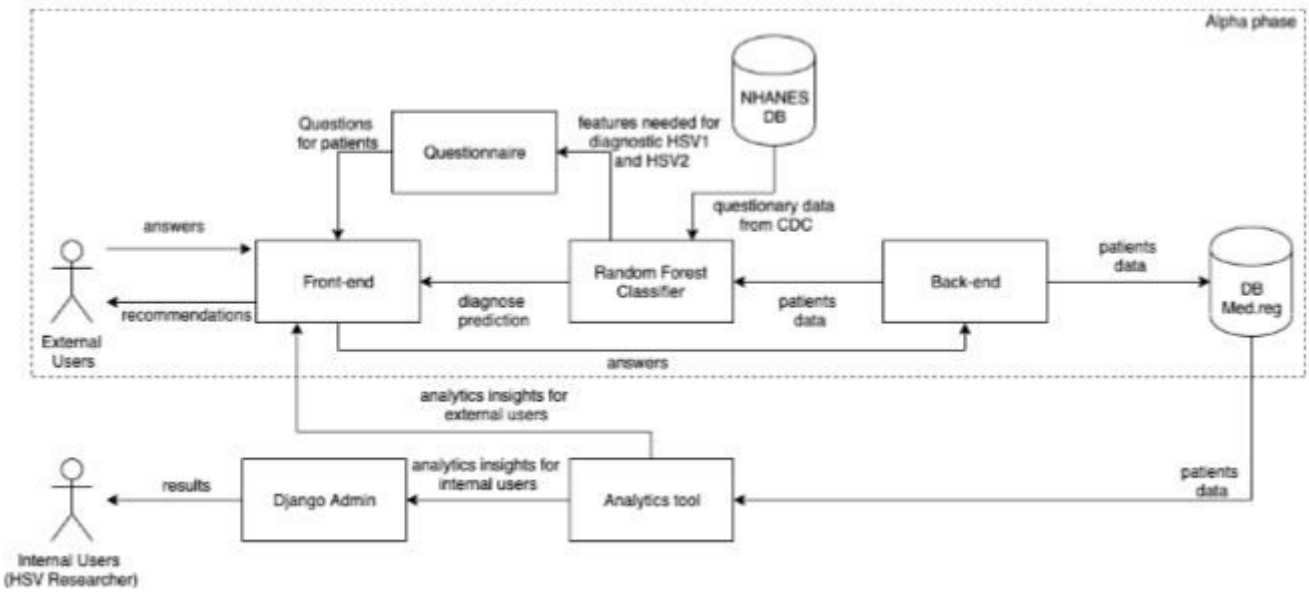

Figure 1

Schema of the technological solution.

LBXHE1 - Herpes Simplex Virus Type 1

Variable Name: LBXHE1
SAS Label: Herpes Simplex Virus Type 1
English Text: Herpes Simplex Virus Type 1
Target: Both males and females 14 YEARS - 49 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Positive	1840	1840	
2	Negative	1546	3386	
3	Indeterminate	7	3393	
.	Missing	317	3710	

LBXHE2 - Herpes Simplex Virus Type 2

Variable Name: LBXHE2
SAS Label: Herpes Simplex Virus Type 2
English Text: Herpes Simplex Virus Type 2
Target: Both males and females 18 YEARS - 49 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Positive	478	478	
2	Negative	2335	2813	
3	Indeterminate	2	2815	
.	Missing	895	3710	

Figure 2

Demographics of validation dataset [10]

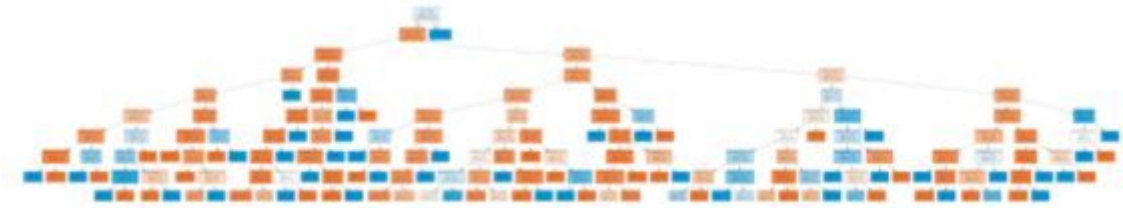

Figure 3

Random Forest training process final decision tree for HSV1 testing subset.

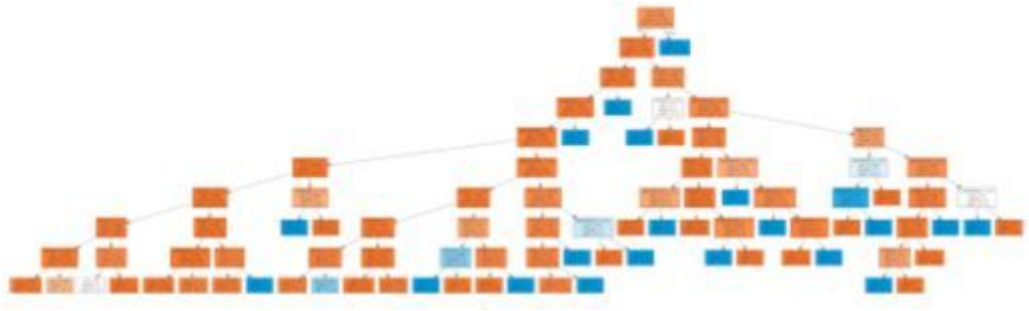

Figure 4

Random Forest training process final decision tree for HSV2 testing subset.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{True\ Positive}{Total\ Actual\ Positive}$$

Figure 5

Recall formula. [16]

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile2.Questionsforthemodel.docx
- Additionalfile1.Questionnairedata.docx