Development of Personalized Health Tracking Tool Using Machine Learning

Team DataStation

Mohammad Aljubran
Stanford University
aljubrmj@stanford.edu

Mohammed Alkhalifah
University of Kansas School of Medicine
alkhalifaa2@gmail.com

Moayd Alkhalifah
Johns Hopkins Aramco Healthcare
alkhalifah.moayd@gmail.com

Problem Statement

Disability-adjusted life year (DALY) is a quantitative measure of disease burden. The National Institute of Health (NIH) describes DALY as the number of years lost due to illness, disability, or premature death within a given population. It accounts for years lost due to both mortality and morbidity. One DALY can be thought of as a unit of measurement that equates to a lost year of 'healthy' life. The World Health Organization (WHO) adds that DALY can be used as means to quantify the gap between current health level and an ideal health condition.

More than 30% of the world DALY rates and over 50% of the global mortality can be explained by behavioral, environmental and occupational, and metabolic risks, which provides opportunities for guidance and prevention strategies [4]. However, many of these preventions are already implemented to tackle these critical risks. In addition, modifying many of these preventable risk factors traditionally mandates that patients visit healthcare centers many times. Yet, access to healthcare is variable amongst different communities and may not be sufficient to capture the ill population. Furthermore, controlling modifiable risk factors (e.g. high blood pressure, smoking, drug use, etc.) requires high levels of personal awareness and cooperation at the patient end to enable positive change.

Nature of Proposed Solution

In view of the prominence of modifiable risk factors, we are introducing a tool that can augment, track and provide individualized prevention recommendations guidelines. The proposed solution is a web/smartphone-based system that tracks individuals risk factors, health improvement or decline over time, and the individual's probability of developing a disease, disability, or premature death.

This tool will continue to provide continuous personalized risk factor modification recommendations based on the individual's overall health status at any given time. It will also provide an insight into the community's most common risk factors and diseases based on the

information obtained from the individuals' tool. In addition, these data will forecast the health status trajectory within a given community. Collectively, it will raise health awareness at the personal level, and also assist policymakers in designing risk factor modification strategies and developing healthcare systems.

Data Preprocessing

Tableau Prep Builder 2019.1 and Tableau Desktop 2019.1 are used to preprocess, visualize, and analyze the given datasets. Tableau offers an integrated platform to combine, shape, and clean data to capture actionable insights and implicit trends. It enables efficient and live visual data exploration and analytics to answer deeper questions of 'where', 'when', 'who', 'why', and 'how'.

Raw data can be incomplete, noisy, inconsistent, and plagued with errors. If analyzed directly without cleaning, misleading results and conclusions could be reached—'garbage in, garbage out'. Optimal preprocessing requires domain expert understanding of the dataset at hand. The health sector is the domain of interest, specifically health metrics and evaluation using DALY. As seen in **Fig. 1**, basic data preprocessing can generally be summarized in four steps: cleaning, integration, transformation, and reduction [7]. The following will highlight the different data preprocessing steps in a brief while illustrating the concepts and observations with examples from the datasets.

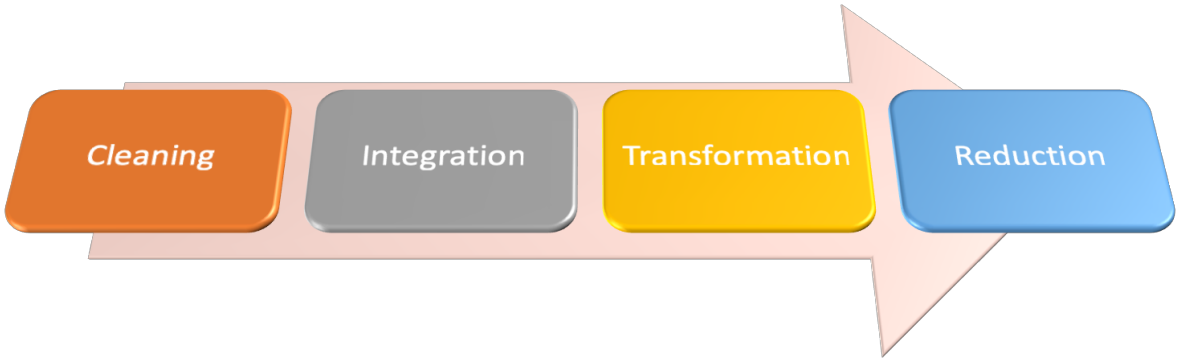

Fig. 1—Data Preprocessing Steps: Cleaning, integration, transformation, and reduction are the four basic data preprocessing steps that must be considered and applied carefully before visualizing and analyzing datasets.

Data Cleaning

This first step is centered around identifying missing data, adjusting for noise, and recognizing outliers. Both Primary and Secondary Datasets suffer from incomplete data where, for instance, the Primary Dataset highlights DALY mean and 95% confidence interval limits of all the given 22 cause categories only for specific age groups: '<20 years', '95 Plus', and 'Age-Standardized'. Meanwhile, the other age groups have DALY values only for two cause categories, namely 'Diarrheal Diseases' and 'Lower Respiratory Infections'. This complicates the comparison of DALY values between the different age groups for various cause categories within the Primary

Dataset. With respect to this limitation, analysis of the other attributes will be limited to the data provided for the 'Age-Standardized' group while the others will be filtered out. Similarly, the Secondary Dataset omits DALY values for specific cause categories across specific countries between 2010 and 2012.

Data Integration

Integrating data involves joining different datasets into one source for mining purposes. Note that the Secondary Dataset only highlights the 'Age-Standardized' group which is meant to enable comparison between different countries. Since the Saudi Arabia data of this age group within the Primary Dataset is already included in the Secondary Dataset, there is no need to join and unify these two datasets.

Data Transformation

This involves partially or fully reforming the dataset to enable efficient and appropriate forms of mining. DALY is provided in three different metrics: 'Number', 'Percent', and 'Rate'. Each of these implies a different meaning, and thus should be analyzed separately. This study and proposal analysis will filter out 'Number' and 'Percent' and only keep 'Rate'. The latter inherently enables comparison between different attributes and categories since it is measure per 100,000 population. Note that the DALY mean and 95% confidence interval limits involve negative values, which can be viewed as a health gain or period that someone lives beyond their life expectancy. However, to be more conservative in reporting DALY, these negative values are all set to zero [13].

Data Reduction

It is means of aggregation and compression that aims to reduce the data volume to accelerate analysis and limit computational power requirements. Reducing the Primary Dataset involved dropping a few attributes to facilitate analysis and generate a focused solution. These attributes are 'ID' (holds no technical value), 'Measure' (holds only one value, DALY), 'Location' (holds only one value, Saudi Arabia), and 'Rei_Type' (holds no value to the proposed solution).

Data Insights

Given the preprocessed data above, there are evident limitations in the data which introduces uncertainty. This can be resolved by mining the literature for complementary datasets in order to guarantee a more complete survey and increase the confidence level in the conclusions. In the meantime, we will illustrate a few means of data analytics and draw conclusions based on the preprocessed data at hand.

Many questions can be posed and answered using the data analytics we performed. Below are sample questions with the corresponding analysis and conclusions:

1. What is the most recent global DALY geographical distribution?

The Secondary Dataset provides statistics up to 2017 which will be used as the year of interest to compare the DALY rates of different countries. As seen in **Fig. 2**, A heatmap is generated for the

2017 global DALY rate, showing Japan as the best-performing while the Central African is the worst-performing. Meanwhile, Saudi Arabia is at the lower end of the DALY rate spectrum for that year.

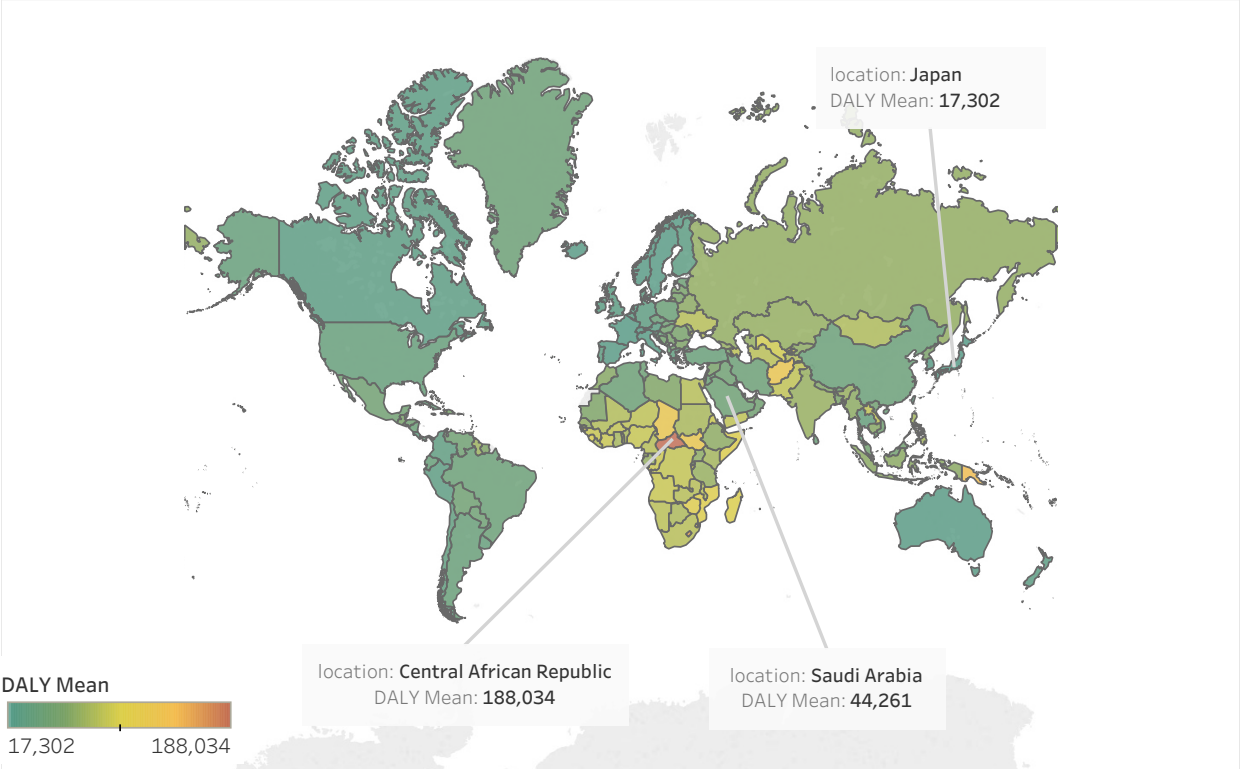

Fig. 2—2017 Global DALY Rate: Heatmap of the 2017 global DALY rate, showing Saudi Arabia in comparison to the top-performing country, Japan, and the bottom-performing country, Central African Republic.

2. What is the worldwide top DALY cause and risk as of today and where are they located?

The Secondary Dataset is used again where the summation of all DALY rates across the reported world countries are compared for different causes. As seen in **Fig. 3**, cardiovascular diseases, diabetes and kidney diseases, and respiratory infections and tuberculosis are the top three, respectively. Similarly, **Fig. 4** shows that the top 2017 risk factors are dietary risks, child and maternal malnutrition, and high systolic blood pressure.

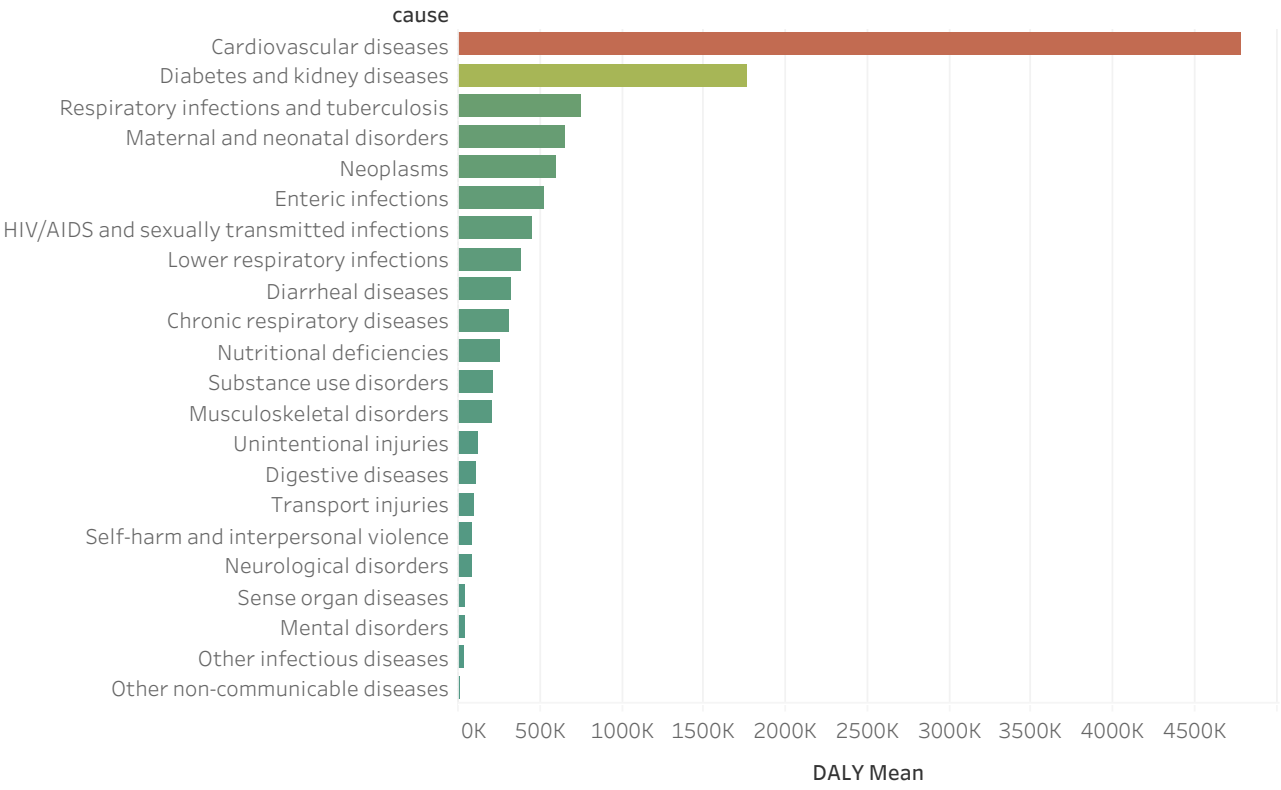

Fig. 3—2017 Global DALY Rates Per Cause: Bar chart that shows the global 2017 ranking of the different causes in terms of DALY rate per 100,000 population.

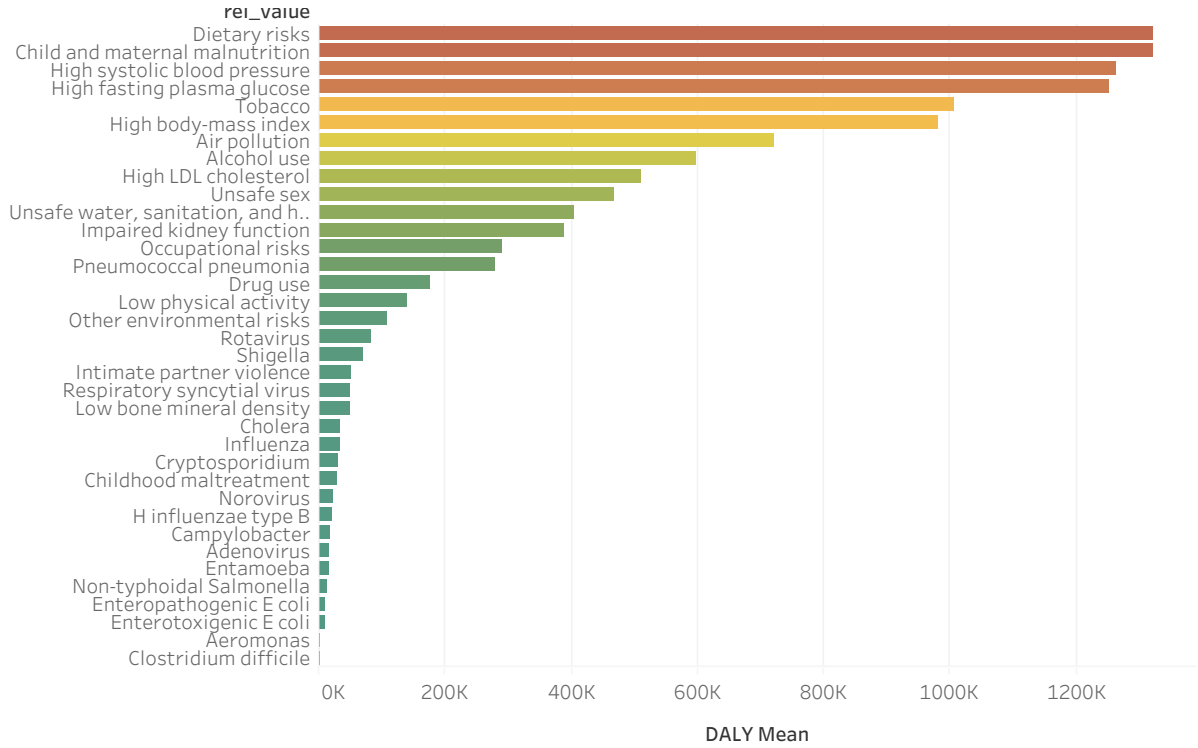

Fig. 4—2017 Global DALY Rates Per Risk: Bar chart that shows the global 2017 ranking of the different risk factors in terms of DALY rate per 100,000 population.

3. What is the most contributing cause to the Saudi Arabia DALY over time?

The Primary Dataset is used where the summation of all DALY rates across the reported cause categories are compared over time. Note that the data preprocessing step limited this comparison to the 'Age-Standardized' group as it retains all relevant causes which allows for an unbiased comparison of the different causes. As seen in **Fig. 5**, cardiovascular diseases, diabetes and kidney diseases, and maternal and neonatal disorders are the top three, respectively.

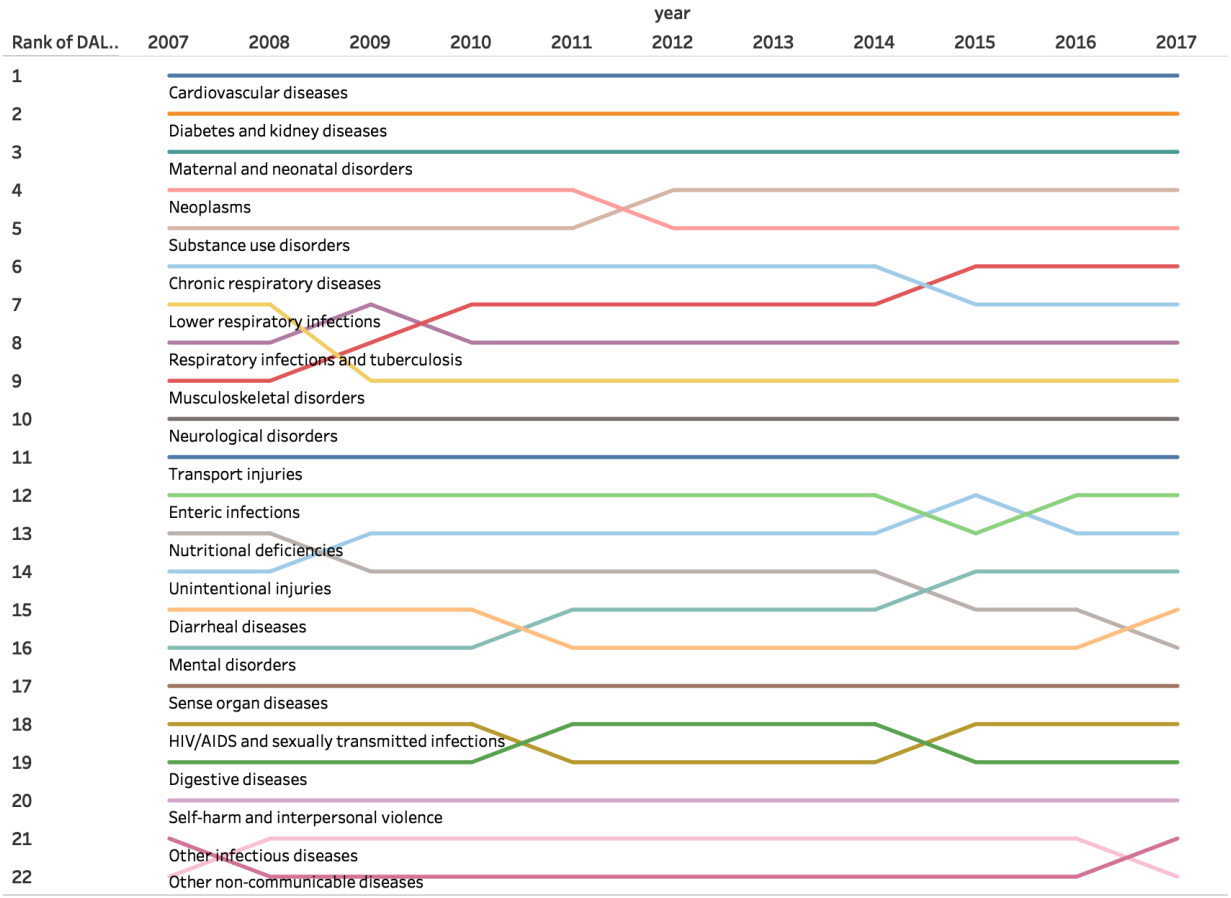

Fig. 5—Saudi Arabia DALY Rates Per Cause: Rank chart that shows the order of different causes in terms of DALY rate within Saudi Arabia over 2007 to 2017.

4. What is the Saudi Arabia DALY trend over time for these causes and risk factors?

The Primary Dataset is used again where the summation of all DALY rates across the reported cause and risk categories are compared over time. Note that the data preprocessing step limited this comparison to the 'Age-Standardized' group as it retains all relevant causes and risk factors which allows for an unbiased comparison of the different categories. **Figs 6 and 7** show these trends and results for causes and risk factors, respectively.

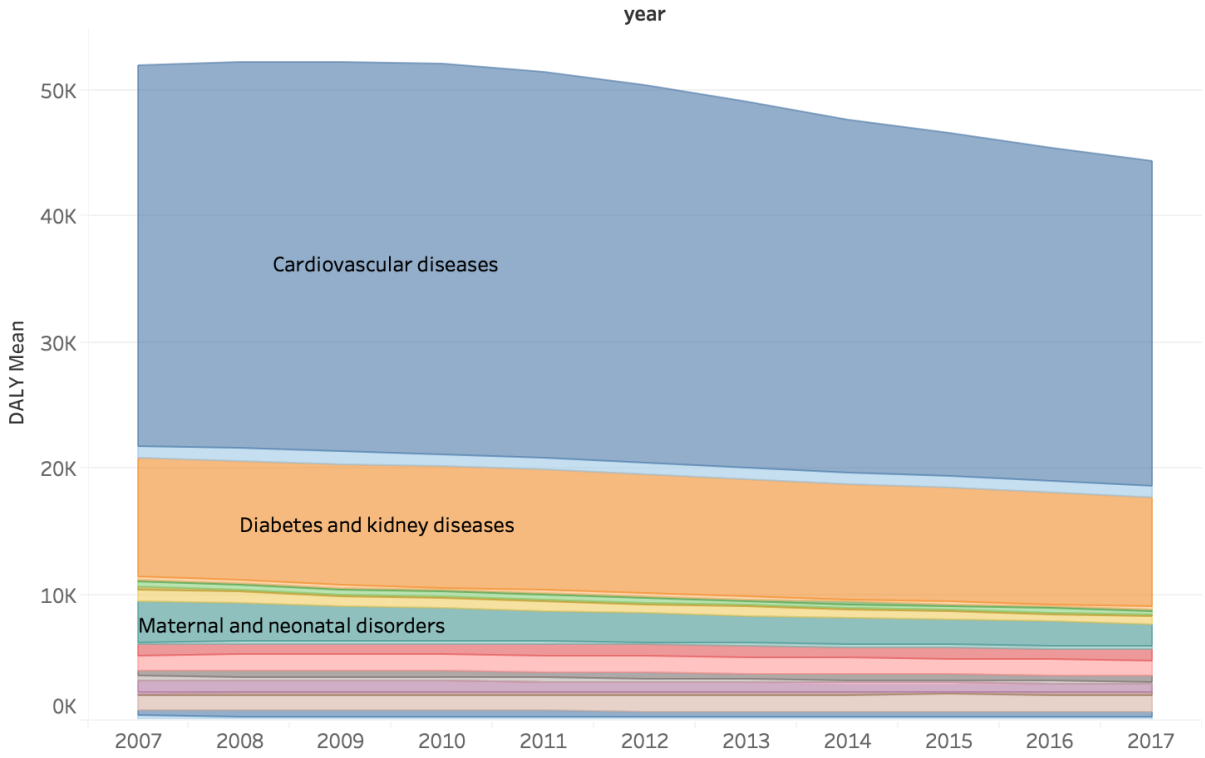

Fig. 6—Saudi Arabia DALY Per Cause Over Time: Area plot showing the contribution of the different causes to DALY rates in Saudi Arabia over time.

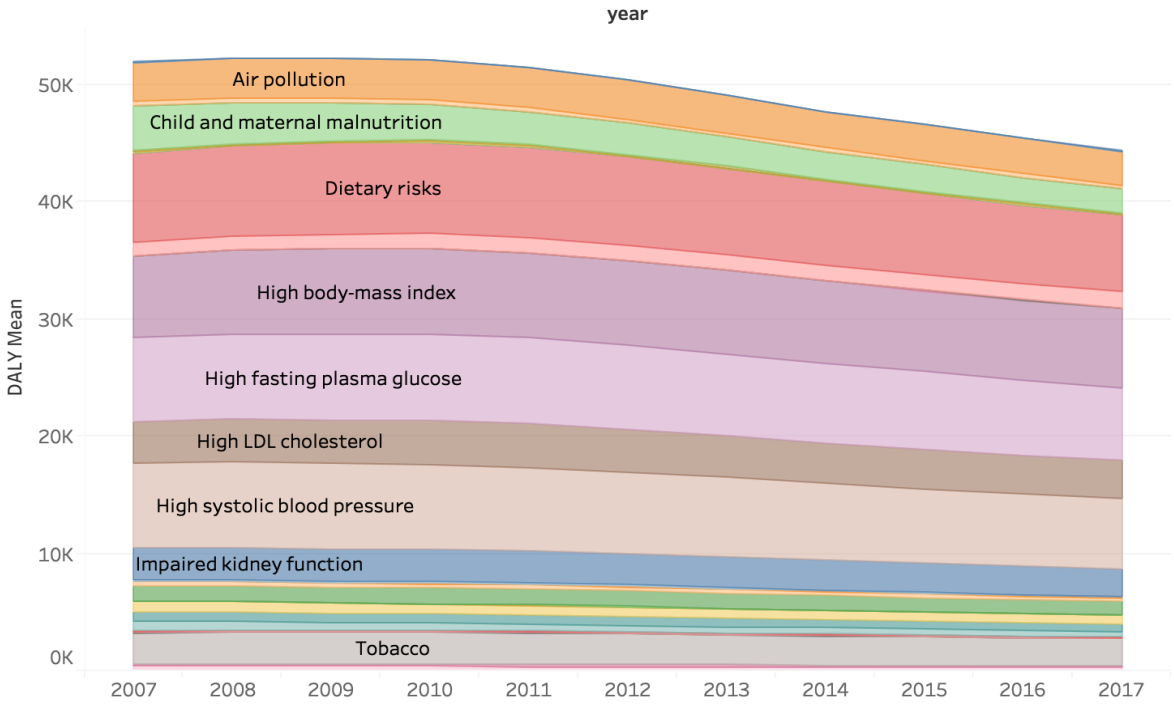

Fig. 7— Saudi Arabia DALY Per Risk Factor Over Time: Area plot showing the contribution of the different risk factors to DALY rates in Saudi Arabia over time.

Focusing on the Saudi Arabia visualizations, we note that many of modifiable risk factors with the highest DALY values (e.g. dietary risks, high body mass index, etc.) cannot be fully controlled by policy changes. Hence, alleviating these risks rather requires higher levels of awareness amongst the Saudi population, which emphasizes the need for personalized health solutions that can assist unhealthy citizens improve their health condition through positive practices.

Proposed solution

The proposed solution is a personalized online application that provides users with quantitative and qualitative insights on their health condition and where they fall across the societal spectrum. DALY is used as a health condition measure, where the application is operated by an algorithm that is connected to a live and continuously growing DALY dataset. This is a user-friendly solution where the individual is only requested to answer a few questions about their health condition. Then, they are prompted with about their personal health level and the recommended guidelines to improve on their condition.

A prototype of this solution is developed using the Primary Dataset. It incorporates a machine learning (ML) model that predicts DALY based on a given input set of attributes. Samuel (1959) describes ML as "a field of study that gives computers the ability to learn without being explicitly programmed" [11]. Mitchell (1998) gives a more modern and articulated definition of ML: a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measure by P, improves with experience E [9]. With the live nature of the proposed solution where the dataset is being dynamically updated, the selected model must allow for dynamic modelling. ML, and specifically the field of reinforcement learning, is well suited for this objective.

Using the MATLAB implementation[1], a fully connected artificial neural network (ANN) is trained to achieve the proposed predictive solution. ANN is a fully interconnected network of layers and neurons which feeds input examples forward to layers, and activates neurons with activation functions to map the corresponding output. Backpropagation is then applied to compute the gradient of the cost function, and thus update the ANN weights. These weights are initiated using the Nguyen-Widrow algorithm which ensures that the initially active region of the network neurons is evenly distributed to cover the input space [10]. Meanwhile, logarithmic sigmoid is used as the hidden layer neuron transfer function while a purely linear transfer function is used at the output layer neurons. Regularized mean-square error is used as the cost function in this application.

Note that in the presence of activation functions, the ANN cost function is usually nonconvex and requires a robust optimization solver. To locate the optimal ANN weights, the Levenberg-Marquardt algorithm is used as it is generally stable and fast [5, 8]. This technique

[1] Tableau preprocessed data and MATLAB codes with 'README' instructions can be found at our GitHub page: https://github.com/aljubrmj/Health-Datathon---Team-DataStation.git

alternates between the Gauss-Newton and gradient descent methods to approximate the Hessian matrix by using a performance/learning gain measure which is adjusted adaptively throughout the learning process to control the learning rate and ensure robust convergence. Meanwhile, the regularization parameter is optimized using the Bayesian regularization backpropagation algorithm which locates and trains on effective weights while effectively turning off the irrelevant counterparts [1, 3, 6]. The following steps briefly explain and illustrate the major application development steps:

- *Preprocess Dataset*: As seen above, it is critical to first preprocess and analyze the data before attempting to draw conclusions or fit predictive models. For the purposes of this solution prototype development illustration, the preprocessing steps conducted above are used with the exception that negative DALY values are rather maintained. Hence, this prototype uses the reported DALY rate mean values, and it is developed for the 'Age-Standardized' group for the reasons highlighted in the data preprocessing section.
- *Prepare Model Inputs and Outputs*: Model inputs involve four attributes which are 'Gender', 'Cause', 'Rei_Type', and 'Year' while the output variable is only 'DALY Mean' in units of year lost per 100,000 population. Note that 'Gender', 'Cause', and 'Rei_Type' are categorical variables, so they are encoded with numerical values to facilitate the construction of the proposed model. Input features are also standardized (normalized) to account for the difference in magnitude between the encoded categorical features and the 'Year' feature. This step limits oscillation through the optimization process that is conducted to solve for the optimal ANN model weights.
- *ANN Model Training*: ANN is used to construct the proposed predictive model. Since this is a predictive model, categories related to 2017 are used as testing data while categories related to the year 2016 are used as validation data. Meanwhile, all other categories for the years of 2007-2015 are used as training data. These models involve many hyperparameters which must be properly tuned to find the best fit to the data. Learning curve and bias-variance analysis is conducted using a Bayesian optimizer to avoid underfitting or overfitting the given data. The resultant optimal architecture is seen in **Fig. 8**.

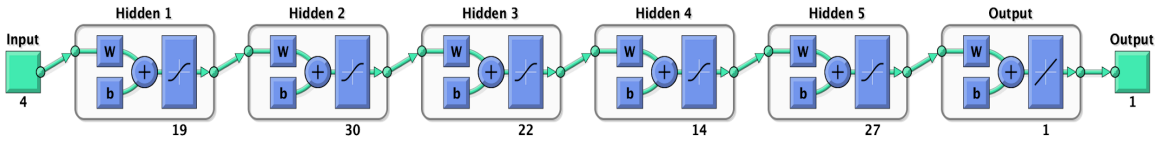

Fig. 8— Optimal ANN Architecture: Bayesian optimization yielded this optimal ANN architecture with 5 layers and 19, 30, 22, 14, and 27 neurons for each hidden layer, respectively.

To verify the performance of the model fit, a regression fit is constructed for all training, validation, and testing data, seen in **Fig. 9**. It is clear that the model performs well on all datasets with all points falling on the 45-degree line. To further verify the fit of the model, DALY prediction over time of different combinations of categories is plotted with the prediction shown up to 2020 with very sensible results, seen in **Fig. 10**.

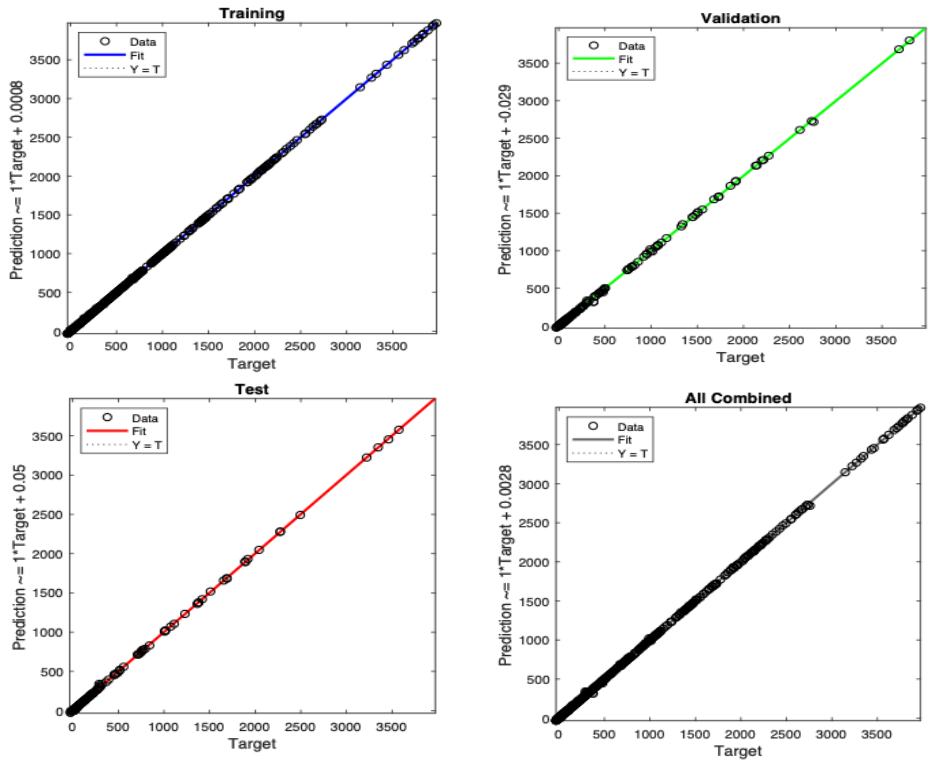

Fig. 9— ANN Regression Plot: Training, validation, and testing regression plots all reflect the good fit and generalization of this ANN model where all data point predictions output fall at the 45-degree line when plotted against the actual target output.

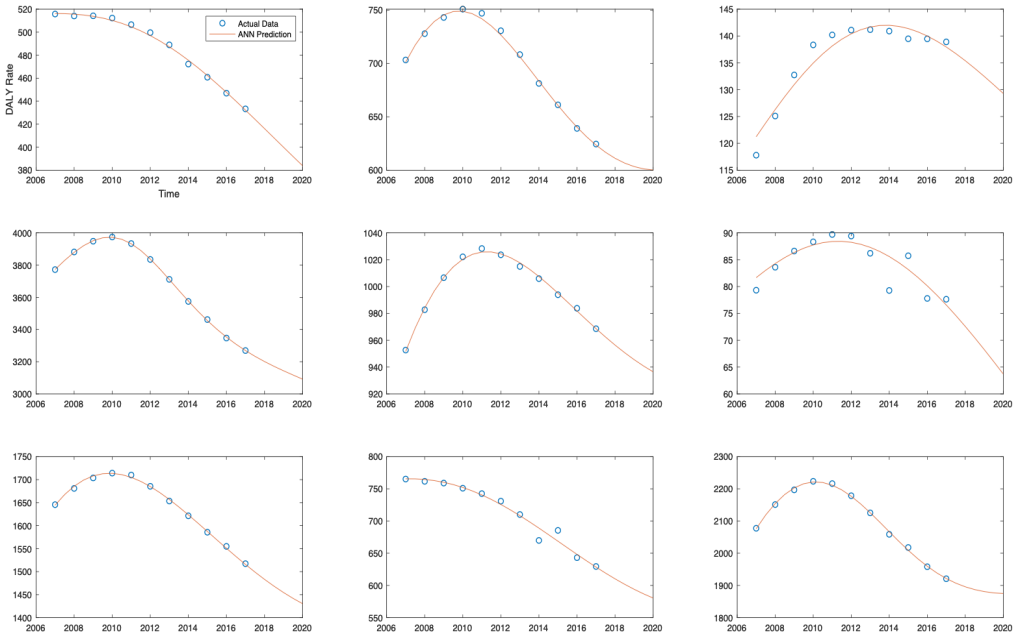

Fig. 10— ANN Model Forecast Evaluation: 9 different plots are generated for different combinations of categories to visualize the ANN model capabilities in forecasting DALY rates all the way to 2020. The orange solid line fits represent the ANN output while the blue dots represent the actual data for the selected set of attributes.

- *DALY Spectrum*: The ANN model is then used to predict DALY rate mean values for the year of 2018 with respect to all attribute combinations available in the datasets. These predictions are then used to fit a probability distribution function (PDF) for that given year, seen in **Fig. 11**. Based on the 0th, 50th, and 75th quantiles, the PDF is then segmented into three regions: safe, warning, and danger. The prototype model is also capable of repeating the same process for the years of 2019 and 2020.

Fig. 11— 2018 DALY PDF: 2018 DALY distribution is plotted and divided into three regions: safe (green), warning (yellow), and danger (red).

- *User Health Condition*: The user interaction with the app is briefly simulated to illustrate the extent of the proposed solution. **Fig. 12** shows how the user is prompted with questions on their personal condition in a stepwise manner. Based on the shown selection, the ANN model outputs its prediction and reflects it on the PDF spectrum of the DALY rate for that specific year, seen in **Fig. 13**. This example uses the following input categories: 'Gender' = male, 'Cause' = cardiovascular diseases, 'Rei_Type' = low physical activity, and 'Year' = 2018. The resultant DALY is 611 per 100,000 population where the patient falls in the danger zone.

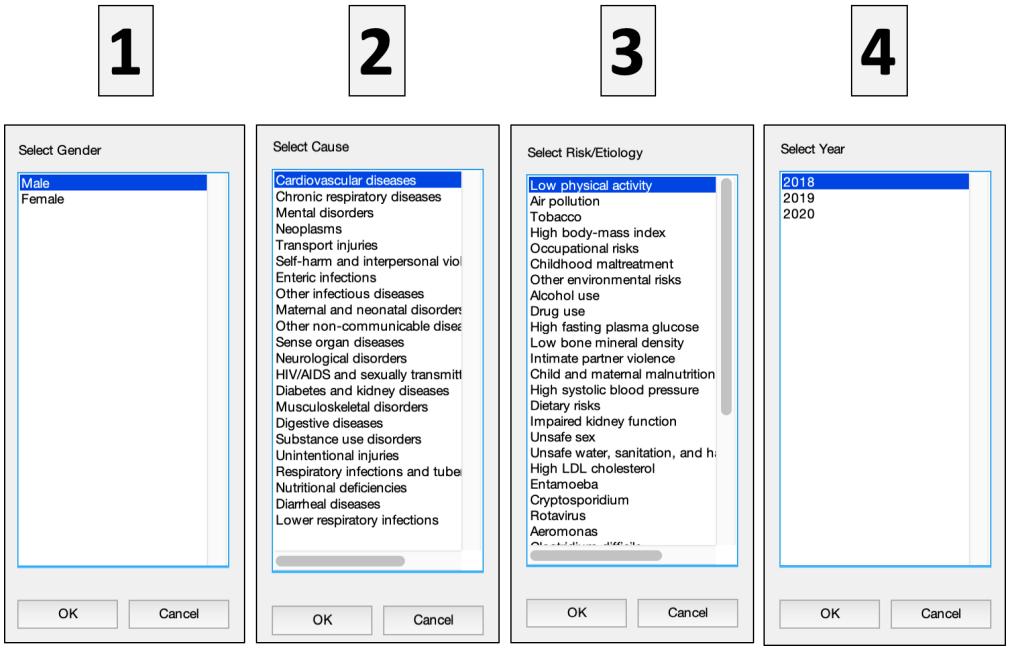

Fig. 12— User Input Interface: Prototype illustration of the proposed input questions which are then used by the ANN algorithm to generate a prediction on the patient's condition.

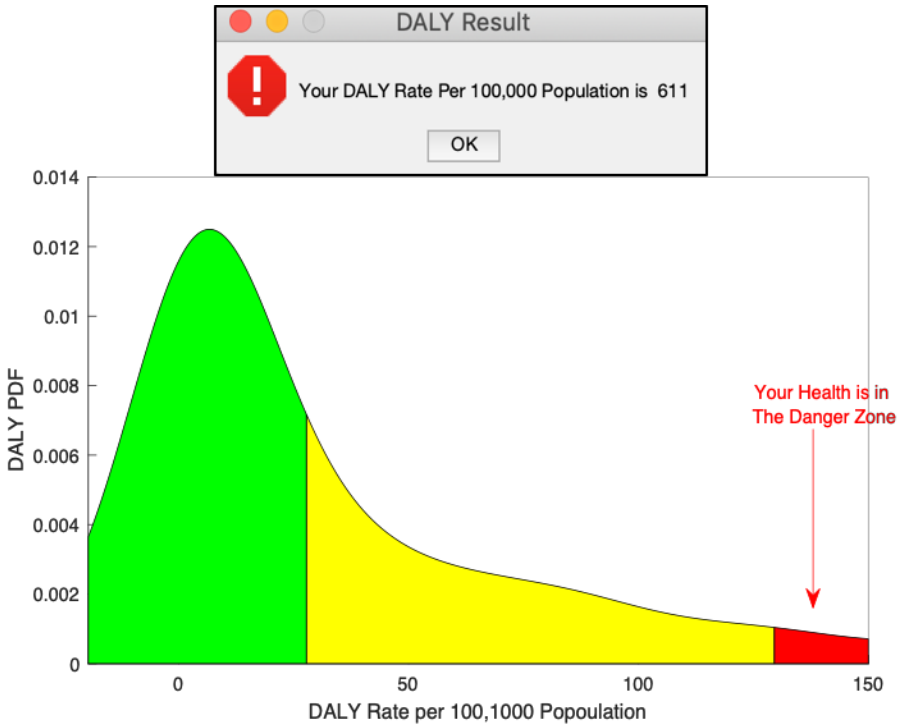

Fig. 13— Prototype Model Output: An example output for using the proposed ANN prototype for a patient with the following categories: 'Gender' = male, 'Cause' = cardiovascular diseases, 'Rei_Type' = low physical activity, and 'Year' = 2018. The given example shows DALY of 611 per 100,000 population where the patient falls in the danger zone. Note that the ANN model runs on the enumerated versions of the categorical variables.

The application then prompts the user with guidelines based on their DALY estimate and health condition. These guidelines are based on formal recommendations and guidelines provided by the Ministry of Health. Furthermore, the application is capable of providing statistical predictions of DALY for cause and risk category combinations that were not linked in the original data. This additional feature is powerful as it provides insights to academic researchers who are studying links and relationships between different cause and risk factors when surveys are not available. The tool is also useful for policymaking as it forecasts disease burden in the Kingdom. Since the application involves user data input, this solution is an ever-improving and learning tool that only becomes more accurate and robust over time.

Impact

This tool outputs continuous individualized risk factor modification recommendations based on the individual's overall health status at any given time. It also provides an insight into the community's most common risk factors and diseases based on surveys and data obtained from individuals, which can be used to forecast the societal health status trajectory. Collectively, this will help policymakers in designing risk factor modification strategies and developing improved healthcare systems.

Using this tool is expected to have collective as well as personal impact within the Saudi health sector. Plus, continuously tracking individuals' health status and future risks will provide significant premorbid prevention leading to a drop in DALY. For example, stroke is the second leading cause of death and the number one leading cause of disability globally. 90% of the stroke risk is modifiable and 74% of the risk is behavioral [2]. With this tool's capabilities to continuously track and provide tailored recommendations and guidelines to patients to modify these modifiable risk factors, the impact is indeed significant.

Solution Uniqueness

Sparked by Saudi Vision 2030, the Kingdom is experiencing modern innovative changes in many sectors including healthcare and artificial intelligence. The Saudi society is technologically responsive and adaptive which paves the path for innovative healthcare tools. 2018 statistics reveal that the number of internet users in Saudi Arabia is close to 30 million people [12]. The internet penetration in the country has now reached 91%, seen in **Fig. 14**. This implies that the proposed solution will be fairly accessible and attractive to users, unlike traditional solutions.

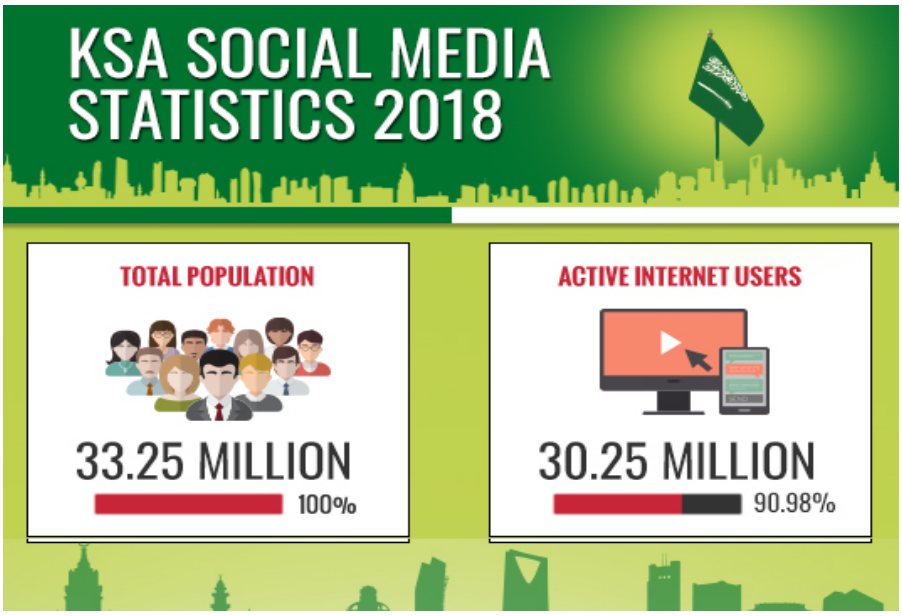

Fig. 14— Internet Use Statistics in Saudi Arabia: Statistics show that the internet use in the Kingdom has climbed to unprecedented levels, reaching up to 90% of the Saudi total population [12].

According to the Precision Medicine Initiative, precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." This approach is described in genetic medicine and it is analogous to personalized medicine which we propose to implement where the collective DALY dataset is used to formulate personalized solutions to individual patients. We believe that our tool and approach are unique in utilizing the currently available datasets in formulating personalized solutions and also offering forecast capabilities to predict individualized risk.

Accurately predicting future risk allows patients, healthcare providers, and policymakers to properly use policies and take preventative measures against main risks that adding to the highest number of DALY. This further allows doctors and researchers to make better decisions on treatments and prevention strategies for a particular disease and demographic setting.

Individuals who are interested in maintaining personal health and to maintain good of life will not hesitate to seek this tool to obtain individualized risk and personalized recommendations for modifiable environmental, occupational, behavioral, physiological, and metabolic risk factors. We provide a unique insight into the personalized health future by utilizing data-driven prediction, analysis, and interpretation. Once this tool grows/matures, users will be able to input additional personal health data, track health behaviors using personal devices, electronic medical records, social media accounts, etc. Meanwhile, current healthcare apps offer only tracking capabilities with no ability to interpret, analyze, correlate or predict personal health data.

Implementation and Feasibility

Preventions and primary care guidelines are readily available to reduce the burden of essential risks. Developing an app/website that allows individuals to input personal health-related data and utilize the proposed ML model to generate personalized evidence-based preventive measures and

guidance. In addition, it will predict a selective future risk. We showed proof of concept of our analytic protocol by utilizing the Primary Dataset to generate personalized risk predictor which is used in initiating a personalized prevention program. Consequently, this will assist in mitigating environmental, occupational, behavioral, physiological, and modifiable metabolic risk factors. Implementing the concept will require developing a user-friendly interface introduced by an app/website.

Offering monitoring tools, incentives through a feedback loop can encourage users to keep tracking their health and provide additional data. This can be offered as a free app/ subscription for individuals. Users will be prompted to consent/agree to share their de-identified personal health data in exchange for the different features they acquire and ultimately towards creating a healthy community. It is understood that many of the components that shape the wellbeing of individuals and communities have their origins outside the conventional healthcare system. Social and behavioral biometrics can be tracked and input into the system to provide feedback and personal coaching, seen in **Fig. 15**. This can be achieved using various personalized devices, e.g. wearable electronics.

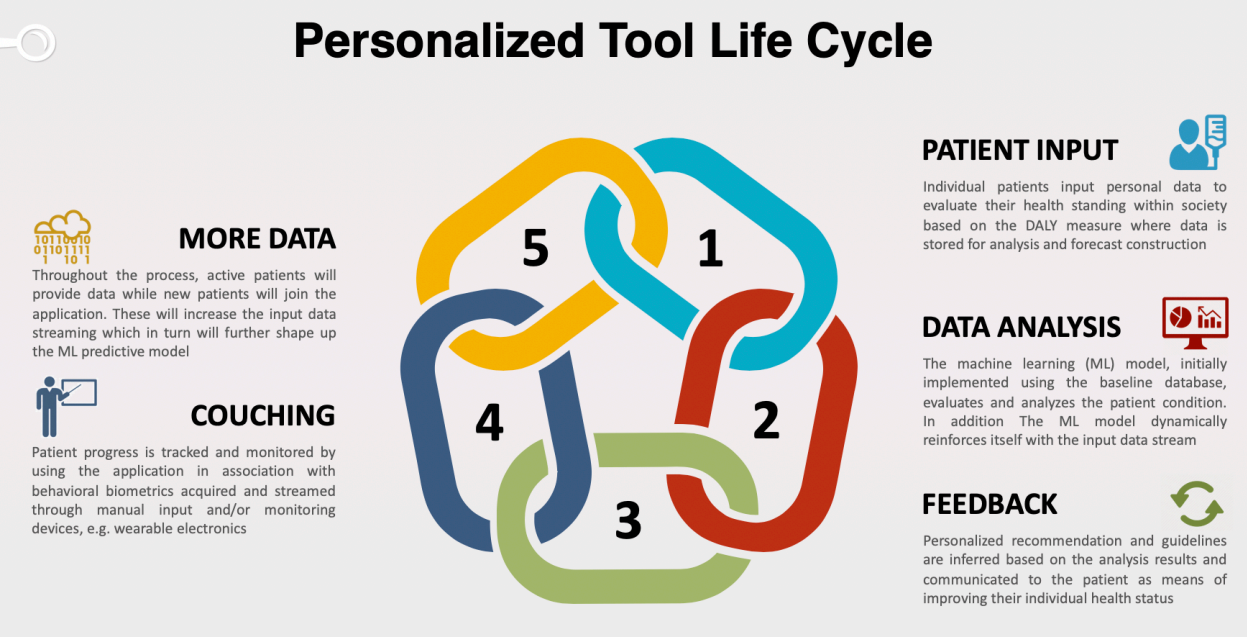

Fig. 15— Personalized Tool Life Cycle: This cyclic diagram shows the five major steps that the proposed solution goes through over time as it evaluates, tracks, and coaches patients while further shaping up the ML model through reinforcement learning.

References

[1] Burden, F., & Winkler, D. (2008). Bayesian regularization of neural networks. In Artificial neural networks (pp. 23-42). Humana Press.

[2] Feigin, V. L., Roth, G. A., Naghavi, M., Parmar, P., Krishnamurthi, R., Chugh, S., ... & Estep, K. (2016). Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet Neurology, 15(9), 913-924.

[3] Foresee, F. D., & Hagan, M. T. (1997, June). Gauss-Newton approximation to Bayesian learning. In Proceedings of International Conference on Neural Networks (ICNN'97) (Vol. 3, pp. 1930-1935). IEEE.

[4] Forouzanfar, M. H., Afshin, A., Alexander, L. T., Anderson, H. R., Bhutta, Z. A., Biryukov, S., ... & Cohen, A. J. (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1659-1724.

[5] Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. Quarterly of applied mathematics, 2(2), 164-168.

[6] MacKay, D.J. (1992). Bayesian Interpolation. Neural Computation, Vol. 4, No. 3, 1992, pp. 415 to 447).

[7] Malik, J. S., Goyal, P., & Sharma, A. K. (2010). A comprehensive approach towards data preprocessing techniques & association rules. In Proceedings of the 4th National Conference.

[8] Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. Journal of the society for Industrial and Applied Mathematics, 11(2), 431-441.

[9] Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (1990). Machine learning. Annual review of computer science, 4(1), 417-433.

[10] Nguyen, D., & Widrow, B. (1990, June). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In 1990 IJCNN International Joint Conference on Neural Networks (pp. 21-26). IEEE.

[11] Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development 44(1), 210–229.

[12] Saudi Arabia Social Media Statistics 2018 - Official GMI Blog. (2018). Retrieved from https://www.globalmediainsight.com/blog/saudi-arabia-social-media-statistics/

[13] Struijk, E. A., May, A. M., Beulens, J. W., de Wit, G. A., Boer, J. M., Onland-Moret, N. C., ... & Peeters, P. H. (2013). Development of methodology for disability-adjusted life years (DALYs) calculation based on real-life data. PLoS One, 8(9), e74294.