FEDPANDEMIC: A CROSS-DEVICE FEDERATED LEARNING APPROACH TOWARDS ELEMENTARY PROGNOSIS OF DISEASES DURING A PANDEMIC

Anonymous authors
Paper under double-blind review

ABSTRACT

The amount of data, manpower and capital required to understand, evaluate and agree on a group of symptoms for the elementary prognosis of pandemic diseases is enormous. In this paper, we present FedPandemic, a novel noise implementation algorithm integrated with cross-device Federated learning for Elementary symptom prognosis during a pandemic, taking COVID-19 as a case study. Our results display consistency and enhance robustness in recovering the common symptoms displayed by the disease, paving a faster and cheaper path towards symptom retrieval while also preserving the privacy of patients' symptoms via Federated learning.

1 INTRODUCTION

Symptom prognosis and analysis are important tools of pandemic management, as medical conditions of the population could be gauged with these tools. However, appropriate symptoms and their exact effects were reported after mass collection and analysis during COVID-19 (Ghosh et al. (2020), Bennett & Carney (2011)). This not only consumed time but also required an immense amount of manual effort to anonymize the continuously-growing large corpus of client data. In this paper, we propose *FedPandemic*, a novel approach towards the elementary prognosis of diseases during a pandemic by cross-device Federated learning. We present a novel tool towards prominent symptom detection while retaining client privacy during an outbreak. This encourages collaborative efforts between the general public, smaller healthcare clinics/facilities, Non-Governmental Organizations (NGOs), hospitals and large network medical institutions. Federated learning (McMahan et al. (2016), Bonawitz et al. (2019)) enables one to send models to where the data resides, rather than sending the data to the cloud thereby respecting the privacy of the users. Federated learning empowers distributed learning by gaining generalized insights over the active client space on decentralized data over a large number of rounds. *FedPandemic* employs Word Embeddings as feature extractors for a binary classification model, which is trained using the Federated Averaging (FedAvg) Algorithm (McMahan et al. (2016)). The classifier is aimed to contribute towards preliminary medical examinations and prominent symptoms retrieval in the early stages of an outbreak. The model is developed in a mutable fashion to allow implementations of Secure Aggregation (Bonawitz et al. (2017)) or Differential Privacy (Wei et al. (2020)) for additional privacy use-cases.

FedPandemic is trained based on the statistics of symptoms as reported by Statista's collection of COVID-19 symptoms in Kenya (Faria (2021)), Germany (Koptyug (2021)), Italy (Stewart (2020)), United States (Elflein (2020)) and China (Thomala (2021)). The model employs and simulates different target clients with variable data sizes for learning. The implementation requires low computational prowess while still retaining high performance and client privacy making *FedPandemic* a potentially strong tool towards future symptom detection during an outbreak.

We summarize five major problems presented in current symptom prognosis tools:

- Time Consumption in centralized aggregation by a single institution.
- Data Security of clients participating in such statistics.
- Manual and Logistic Costs for data anonymization of COVID-19 data to protect client privacy.

- Logistic, Manual and Infrastructural costs for over heading such a project.
- Local Bias induced by smaller aggregators and analyzers.

Prominent symptom detection is an integral part of pandemic management and control. If these symptoms are detected and retrieved at the earliest, the process of elementary prognosis will be facilitated faster. This may allow different governments to prevent the spread of such diseases. However, current technologies, require a large network of people maintaining and analyzing this data, which is extremely costly. With *FedPandemic*, we hope to overcome this problem by using Federated learning to provide client privacy and low-cost maintenance based learning.

2 PROPOSED METHODOLOGY

Federated learning is employed in a Cross-Device system, as we wish to enable the general public as individual contributors. The Federated Averaging algorithm (McMahan et al. (2016)) is used for faster learning of the aggregated model. We use word-embeddings for feature extraction on local devices, this allows us to use State-Of-The-Art encoders and also computationally resourceful GloVe (Pennington et al. (2014)) and Word2Vec embeddings. Word Embeddings produce a vector of fixed length as extracted features. This output is then fed into a client model, which is trained for a limited number of epochs E and then the weights of the updated model are returned to a centralised server. In this paper, we run multiple simulations on different contributors using GloVe.

A common word encoder is decided for implementation and a lightweight classifier is designed keeping in mind the embedder selected. This allows us to develop a model, while at the same time keeping computational costs low. The selected embedder (here, GloVe) and model architecture are declared for training and aggregation. However, only training on client symptoms would make the classifier biased. Hence, we randomly sample symptoms from a given medical corpus, which are then learnt as negative samples by the models (refer Figure 1).

The proposed methodology allows us to keep a pseudo data balance, thereby making our models robust to bias and underfitting. We believe that we propose the first implementation for symptom aggregation on a large-scale application that entertains both client-privacy as well as distributed learning. The procedure allows us to overcome some important issues of symptom analysis:

- Manual aggregation of data from multiple healthcare centres is not required.
- Common Symptoms that would be easily identified by the public, such as, high temperatures, fevers, cough and cold, would also be treated with prominence, giving the general public a better chance of discerning the infected.
- Retains client privacy; evading efforts required for data anonymization.
- Word Embeddings also allow semantically similar symptoms to be treated with prominence. This may aid researchers to study additional symptoms that the affected might be exhibiting.

3 EXPERIMENTS

The experiments were conducted on a single system, running multiple instances of client models. The system consisted of 8GB RAM and a GeForce GTX 1650, 4GB GPU. All training was done using the PyTorch framework and the base algorithm used for Federated Learning was FedAvg (McMahan et al. (2016)). Our approach involved four simulations represented by different aggregation steps, which can be employed by local authorities. Our presentation takes statistical numbers from data; as published on Statista. We choose GloVe as our encoder in our experiments because its embeddings are light-weight and easy-to-use in a Federated learning environment when compared to embeddings from other State-of-the-art encoders like BERT (Devlin et al. (2018)) and ELMO (Peters et al. (2018)).

In this work, we present four variants of simulations:

- **Simulation I:** Large Medical Institutes (Baseline)

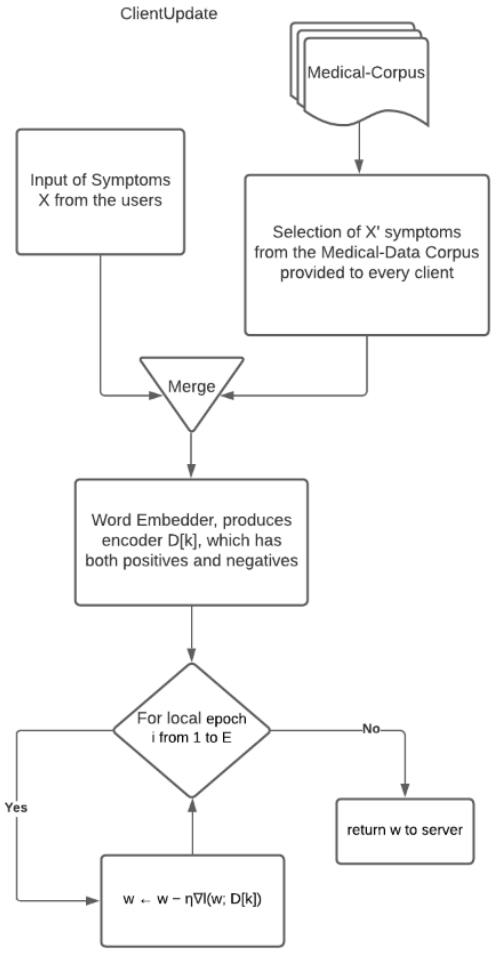

Figure 1: Proposed Methodology of *FedPandemic* for ClientUpdate in cross-device learning of prominent symptoms retrieval. The steps followed at the server are shown in Figure 3.

This simulation aims towards reproducing aggregation by large medical institutes. In this simulation, we distribute the entire corpus, into 20 institutions or clients and train a federated model. Each institution has been given an equal number of sample cases (60,000 samples). This simulation is definite as large medical institutes will already have enough data to ensure that they can select which symptoms are prominent.

- **Simulation II:** Medium Ranged Medical Institutes, like Hospitals, NGOs, etc.

  This simulation offers to cluster and pick symptoms from a larger collaborating group. However, even this group is large enough to accurately classify prominent symptoms. In this case, the data is not equally distributed and ranges between 10,000 to 20,000 samples.

- **Simulation III:** Small Ranged Medical Institutes, like clinics and health care centres

  This simulation is the most practical one, as these institutes may be able to actively collaborate for training such a model. Each client will have samples ranging from 500 to 2,000.

- **Simulation IV:** Individual/Family Contributions

  This simulation is the toughest to learn and provides the most realistic sample which could be implemented for the preliminary search of symptoms. Each client contains samples between 2 and 12.

Simulations	Size Range	No. of clients	Local epochs	Global epochs
I	(60000, 60000)	20	5	5
II	(10000, 20000)	80	5	5
III	(500, 2000)	900	5	5
IV	(2, 12)	100000	5	5

Table 1: Simulations characteristics

These simulations are pulled from the given distribution (refer Figure 4) and aim to replicate real-world usage of *FedPandemic*. We provide experimental results on a few prominent symptoms with different noise levels against the prediction output (refer Figure 2). We also display experiment results for different noise levels for target symptoms shown by greater than and lesser than 10% of the Survey Population (refer Figure A.2).

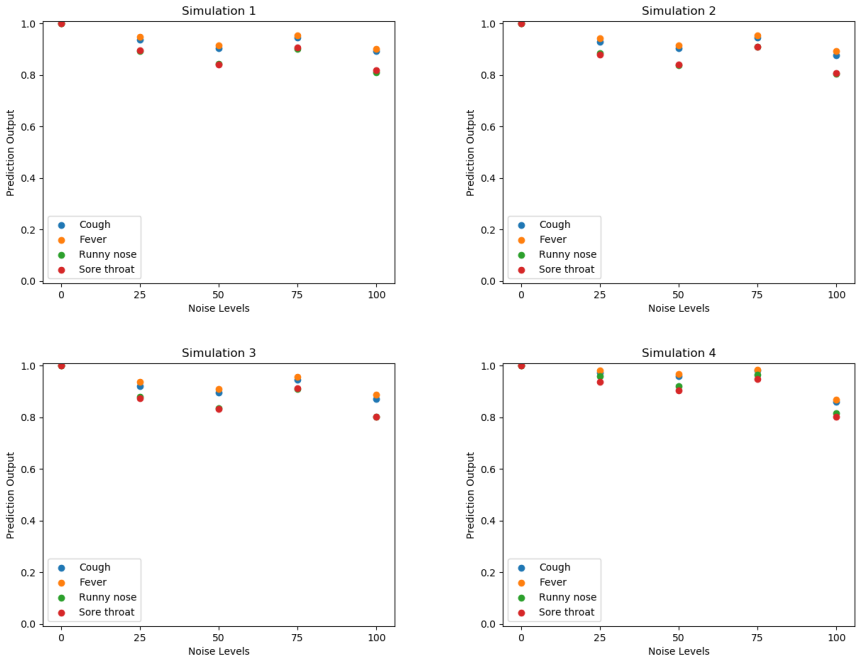

Figure 2: Prediction output versus Noise levels for four of the most common symptoms across all countries.

For the results, there are two important segments we take into account.

- We have to make sure symptoms unrelated to COVID-19 are closer to 0, and those which have a higher chance of exhibition are closer to 1.
- At the same time, we have to make sure that only important symptoms are closer to 1, i.e. a limited number of important features, such as Cough, Fever and Runny Nose.

4 Conclusion and Future Work

In this paper, we showcase a novel approach using Federated learning towards Elementary Symptom Prognosis in order to preserve client privacy and improve faster response times during a pandemic. Our experiments include various noise levels and the accuracy levels drop consistently as the noise values are increased (refer Table 2). We also notice similar readings for 50% noise value and 75% noise values when experimented with most of the prominent symptoms (see Figure 2 and Table 2) which could probably be attributed to accumulating more relevant symptoms as noise levels are

increased. We hope to improve our method by making it more robust to malicious attacks and improve the training model by incorporating data from other countries as well.

REFERENCES

Belinda Bennett and Terry Carney. Review paper: Pandemic preparedness in asia: A role for law and ethics? *Asia Pacific Journal of Public Health*, 23(3):419–430, 2011. doi: 10.1177/1010539511408411. URL https://doi.org/10.1177/1010539511408411. PMID: 21551132.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pp. 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349468. doi: 10.1145/3133956.3133982. URL https://doi.org/10.1145/3133956.3133982.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konecný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. *CoRR*, abs/1902.01046, 2019. URL http://arxiv.org/abs/1902.01046.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

John Elflein. Percentage of people with covid-19 in the united states from january 22 to may 30, 2020 who had select symptoms. 2020. URL https://www.statista.com/statistics/1127573/covid-19-symptoms-us/.

Julia Faria. Knowledge of coronavirus (covid-19) symptoms among the kenyan population from may 30 to june 6, 2020. 2021. URL https://www.statista.com/statistics/1131792/knowledge-of-coronavirus-symptoms-in-kenya/.

Aritra Ghosh, Srijita Nundy, and Tapas K. Mallick. How india is dealing with COVID-19 pandemic. *Sensors International*, 1:100021, 2020. doi: 10.1016/j.sintl.2020.100021. URL https://doi.org/10.1016/j.sintl.2020.100021.

Evgenia Koptyug. Most frequent symptoms caused by the coronavirus (covid-19) in germany in 2021. 2021. URL https://www.statista.com/statistics/1105523/coronavirus-covid-19-symptoms-most-frequent-germany/.

H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016. URL http://arxiv.org/abs/1602.05629.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL http://arxiv.org/abs/1802.05365.

Conor Stewart. Most common symptoms in covid-19 deceased patients in italy 2020. 2020. URL https://www.statista.com/statistics/1110903/most-common-symptoms-in-covid-19-deceased-patients-in-italy/.

Lai Lin Thomala. Breakdown of 55,924 sample patients infected with novel coronavirus covid-19 in china as of february 22, 2020, by symptom. 2021. URL https://www.statista.com/statistics/1105492/china-common-symptoms-of-coronavirus-covid-19-patients/.

K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020. doi: 10.1109/TIFS.2020.2988575.

A Appendix

A.1 Noise Implementation

We present multiple simulations with different levels of noise simulated. In our experiments, we introduce four different noise levels:

1. **0 (Baseline)**: Perfect simulation, without any noise. The experiment can be thought of as a setting where only individuals infected with the novel coronavirus or the pandemic in consideration. There was no influence of any other symptoms the clients may have been facing during learning.

2. **25**: Close to ideal simulation. Here the noise levels are increased by 25%. That is, an individual may report symptoms other than that from the disease in consideration with a 25% probability. This setting is closer to reality, as the general public would not know whether the symptoms they feel are relevant to the pandemic or not.

3. **50**: Here noise levels have been set to 50%. That is, an individual may report symptoms other than that from the disease in consideration with a 50% probability. Another step closer to reality and probably the closest, as most citizens are still healthy or if suffering would recognize new symptoms easily. However, noise would still be generated due to the large sample space.

4. **75**: Noise levels are set to 75%. In this case the citizens have a higher chance of entering symptoms that are not related to the pandemic in question, however, due to the frequency presented in the total population, insignificant/unassociated symptoms would be lost.

5. **100**: There is almost a 100% chance that, if a person does not put a symptom related to the virus, they will put another arbitrary symptom. This noise level is entirely based on the frequency analysis shown during a pandemic. The only reason our federated model is expected to converge and learn in such a case, is the sheer frequency of pandemic victims. Therefore, we specifically target our project towards pandemics like novel coronavirus.

Algorithm 1 Noise Implementation

Input:
Survey_Population ← *People using the application*;
Prominent_COVID19_symptoms ← [cs_1, cs_2, ..., cs_n]; *the actual symptoms displayed as per the research conducted by Statista. Ex: [Fever, Cough, Headache]*
Symptom_Probability ← [ps_1, ps_2, ..., ps_n]; *probability of prominent symptoms that was displayed by the research conducted by Statista. Ex: [0.37, 0.2, 0.1]*
Medical_Corpus ← [s_1, s_2, ..., s_m, cs_1, cs_2, ..., cs_n]; *medical corpus present in GloVe which the public may identify. Ex: [Stomach Ache, Anemia, Red Eyes, ..., Fever, Cough, Headache, ..., ulcers, etc.]*
random.random() ← *random number sampled from the normal distribution N(0,1).*
Algorithm:
for $i \in$ Survey_Population **do**
  Symptoms_Displayed = [];
  for $s \in$ Prominent_COVID19_Symptoms **do**
    if random.random() $<$ Symptom_Probability[s] **then**
      Symptoms_Displayed.append(s);
    else if random.random() $<$ Noise_Level **then**
      Symptoms_Displayed.append(random_symptom(Medical_Corpus));
    end if
  end for
end for

Noise Levels	Accuracy
0%	1.0
25%‘	0.75
50%	0.75
75%	0.6875
100%	0.5625

Table 2: Noise Levels vs Accuracy after four global epochs

A.2 Experimental Data

Country	USA	Country	Kenya
Total	3,73,883	Total	14,616
Fever, cough, or shortness of breath	69.8%	Fever	90.05%
Fever	43.15%	Dry cough	74%
Cough	50%	Difficulty in breathing	70%
Shortness of breath	28%	Sneezing	61.75%

Country	China	Country	Germany	Country	Italy
Total	55924	Total	7,47,900	Total	34142
Fever	87.9%	Cough	40%	Fever	75%
Dry cough	67.7%	Fever	28%	Dyspnoea	73%
Fatigue	38.1%	Runny nose	26%	Cough	38%
Sputum production	33.4%	Sore throat	21%	Diarrhea	6%

Table 3: These are the 4 most prominent symptoms as well as the total number of participants from every country included in our dataset.

For the work presented in this paper, we employ simulations for data collection and cross-device setup. As there is no similar objective or dataset, the paper has taken the liberty to implement these simulations based on real-world data recordings. We take the data collected by Statista for COVID-19 symptoms in different countries. The data presented is a statistical representation of the % of

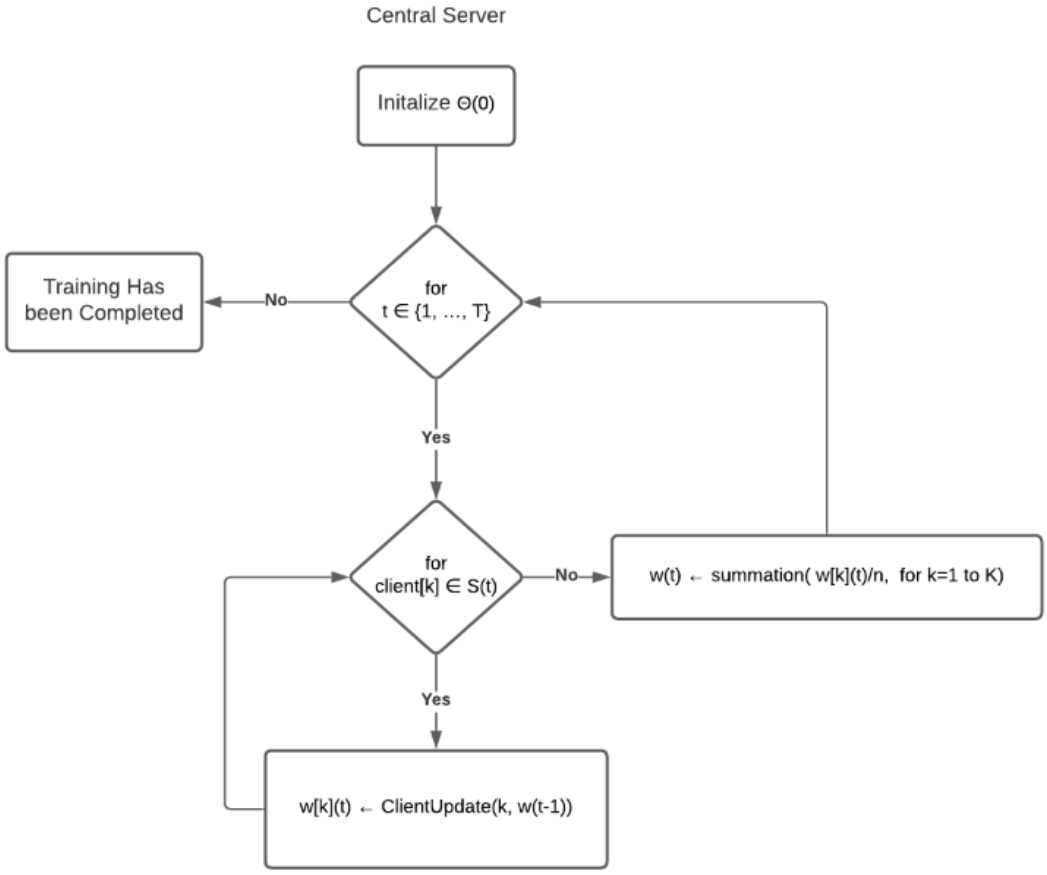

Figure 3: Function of the Central Server for running *FedPandemic*.

people exhibiting a certain symptom. The total number of samples taken from the entire corpus makes up 1,226,465 people's data. The countries whose data has been used for simulation are:

- Kenya
- Germany
- Italy
- United States
- China

We present the statistics of each of these countries in Table 3, and the aggregate statistics we pulled in Table 4. The data presented in Table 4 has been visualized in Figure 4.

Country	USA	China	Germany	Italy	Kenya
Total	373883	55924	747900	34142	14616
Fever	161330	49157	209412	25606	13161
Cough	186941	37860	299160	12973	10815
Shortness of breath	104687	10401	0	0	0
Myalgia	134784	0	0	0	0
Runny nose	22619	18678	194454	0	9025
Sore throat	74402	7773	157059	0	4194
Headache	127867	7605	0	0	8038
Nausea/Vomiting	42435	2796	0	0	0
Abdominal pain	28228	0	0	0	0
Diarrhea	71037	2069	0	2048	0
Loss of smell or taste	30845	0	157059	0	0
Fatigue	0	21307	0	0	0
Muscle Pain	0	8276	0	0	0
Chills	0	6375	0	0	0
Nasal Congestion	0	2684	0	24923	10231
Pneumonia	0	0	7479	0	0

Table 4: This table consists of absolute values of the people showing these symptoms according to country. We extract the base probability distribution on which the simulations are performed.

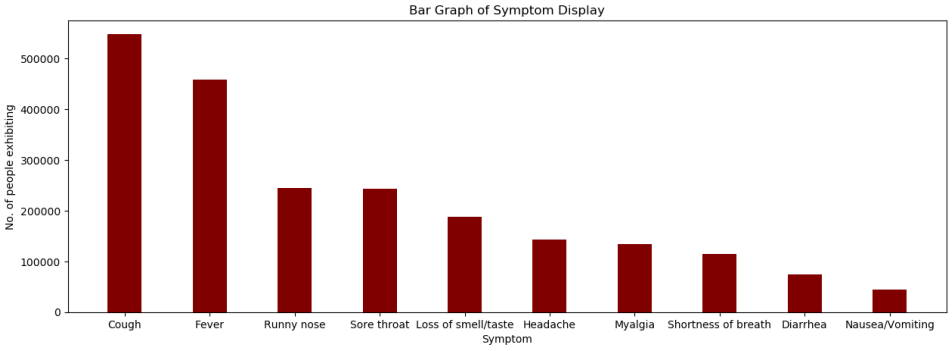

Figure 4: This figure is bar graph representation, of the total populations displaying said symptoms. In retrospect, these values represent the general display such symptoms as the total number of participates makes up 1.2 Billion people. The perturbations are generated from the same distribution in order to approach realistic sample spaces.

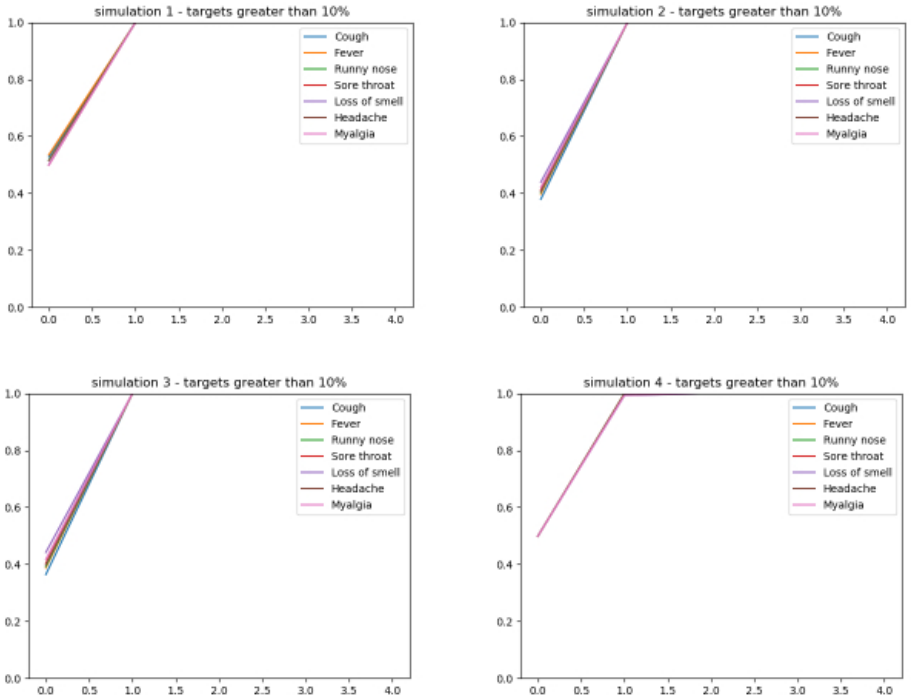

Figure 5: Simulations for 0% noise and target Symptoms exhibited by greater than 10% of the Survey population.

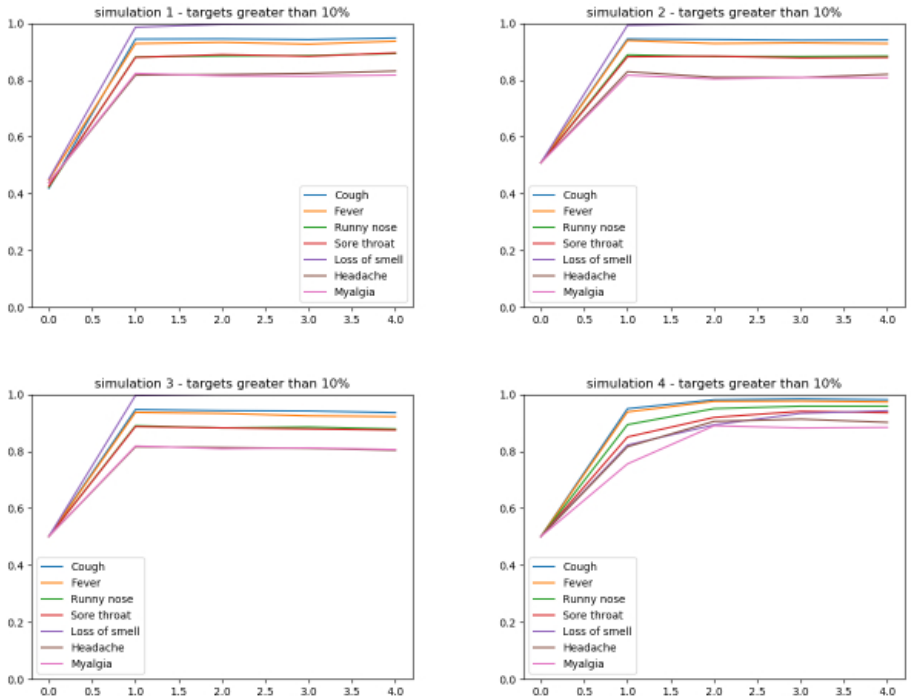

Figure 6: Simulations for 25% noise and target Symptoms exhibited by greater than 10% of the Survey population.

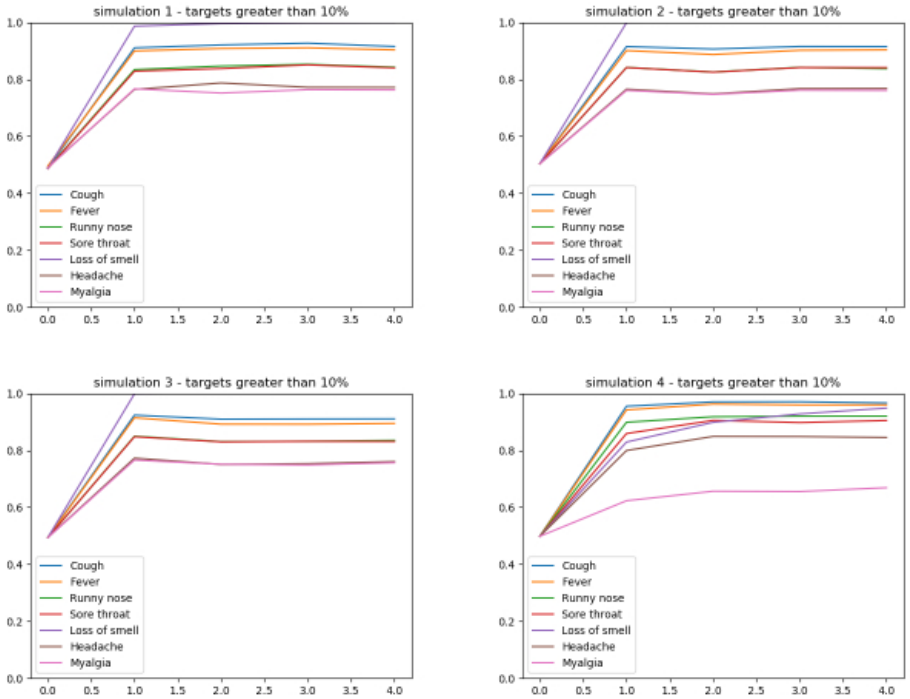

Figure 7: Simulations for 50% noise and target Symptoms exhibited by greater than 10% of the Survey population.

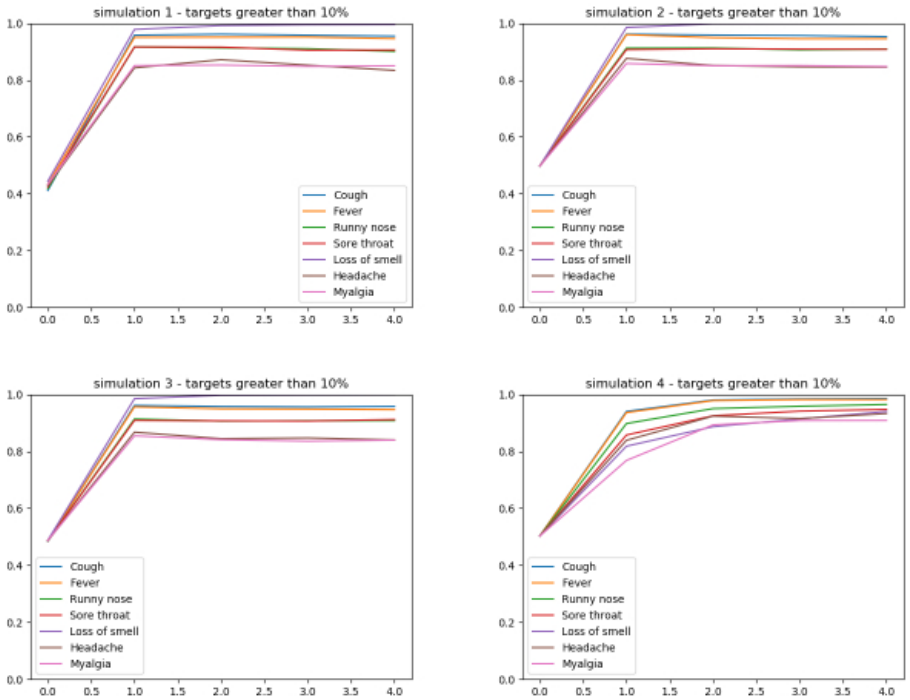

Figure 8: Simulations for 75% noise and target Symptoms exhibited by greater than 10% of the Survey population.

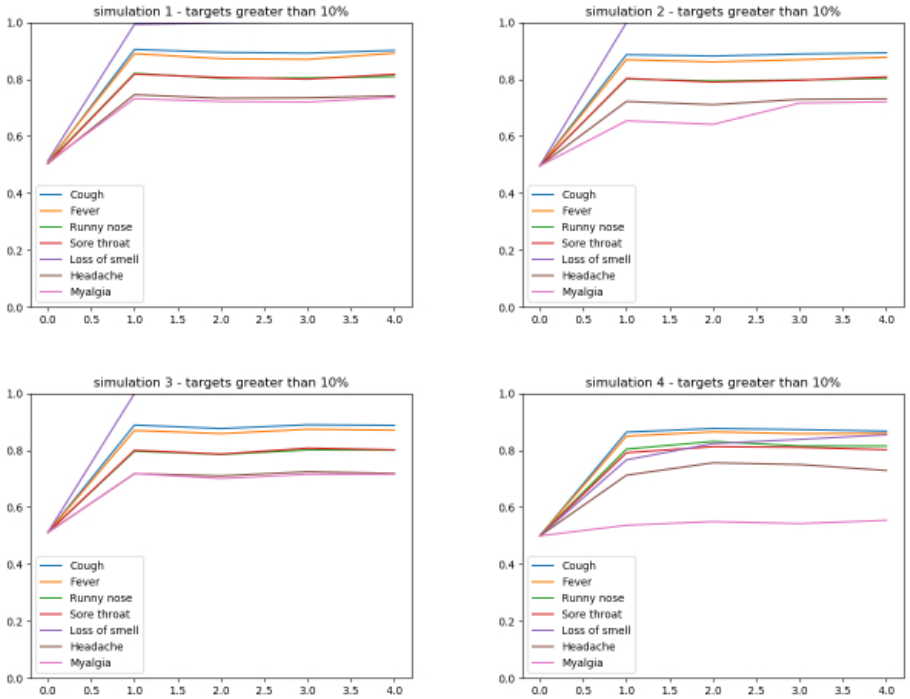

Figure 9: Simulations for 100% noise and target Symptoms exhibited by greater than 10% of the Survey population.

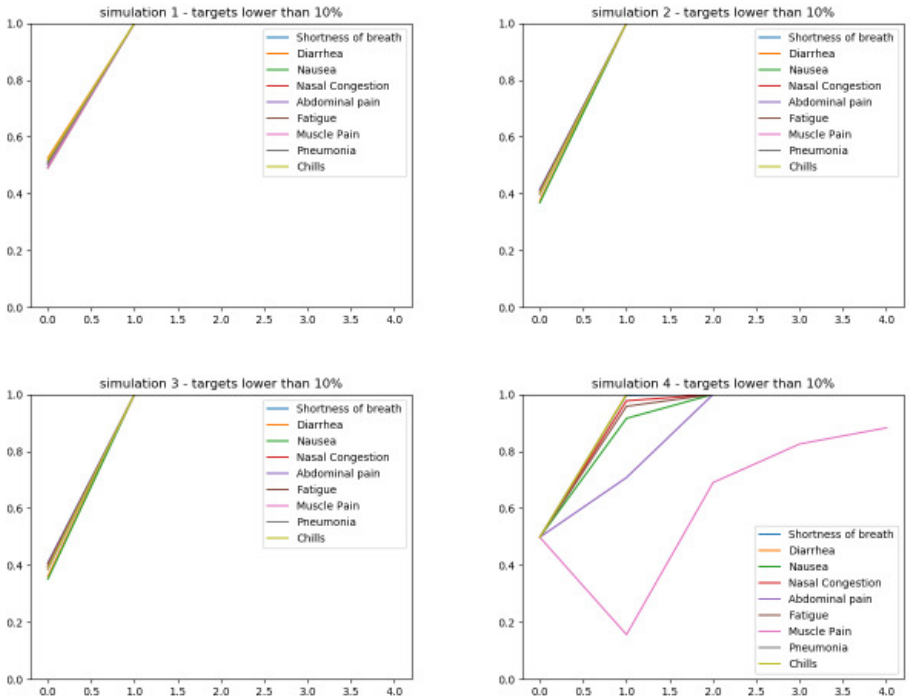

Figure 10: Simulations for 0% noise and target Symptoms exhibited by lesser than 10% of the Survey population.

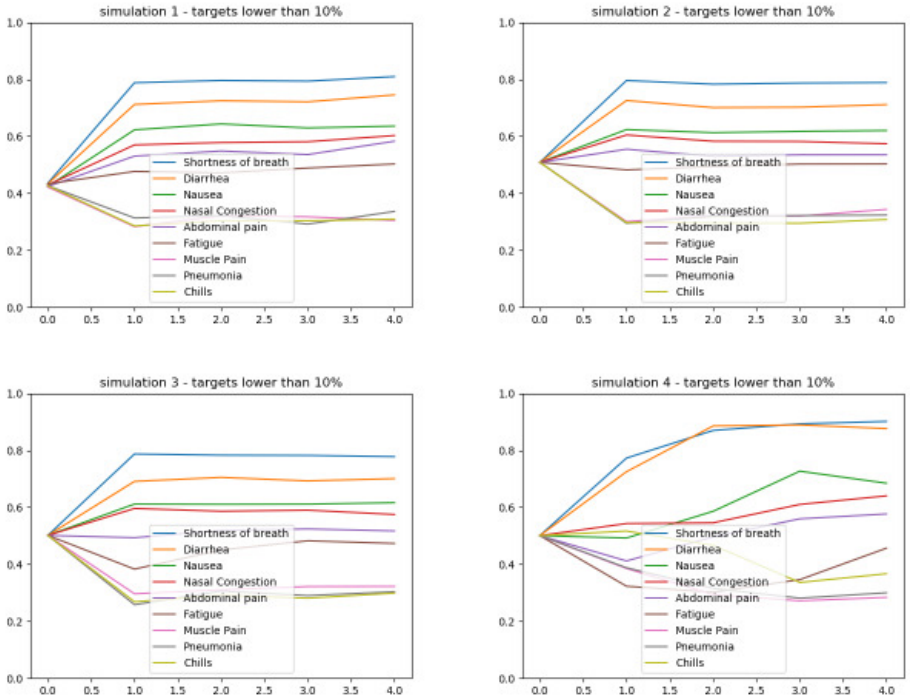

Figure 11: Simulations for 25% noise and target Symptoms exhibited by lesser than 10% of the Survey population.

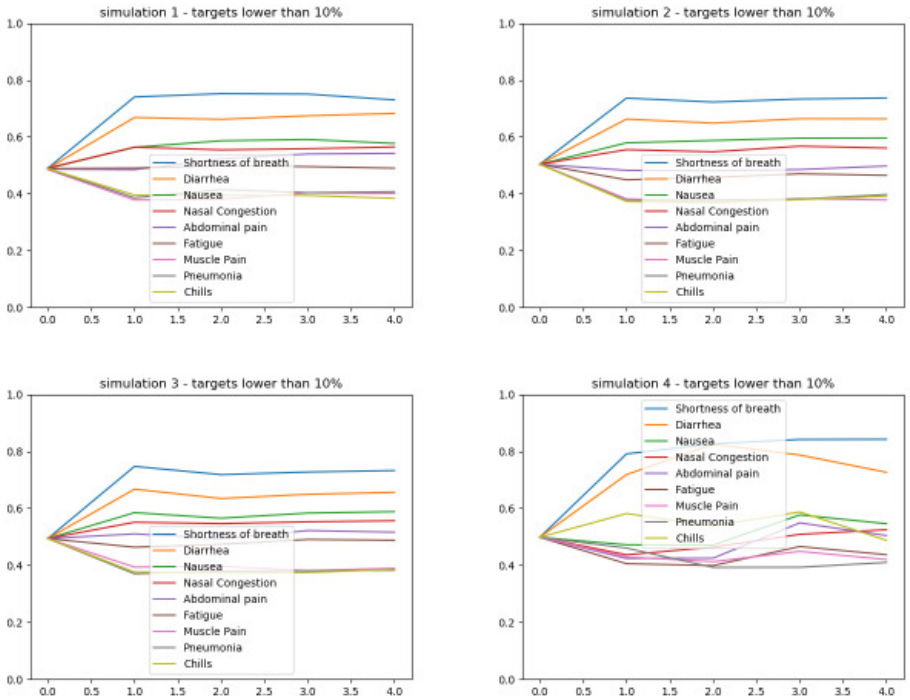

Figure 12: Simulations for 50% noise and target Symptoms exhibited by lesser than 10% of the Survey population.

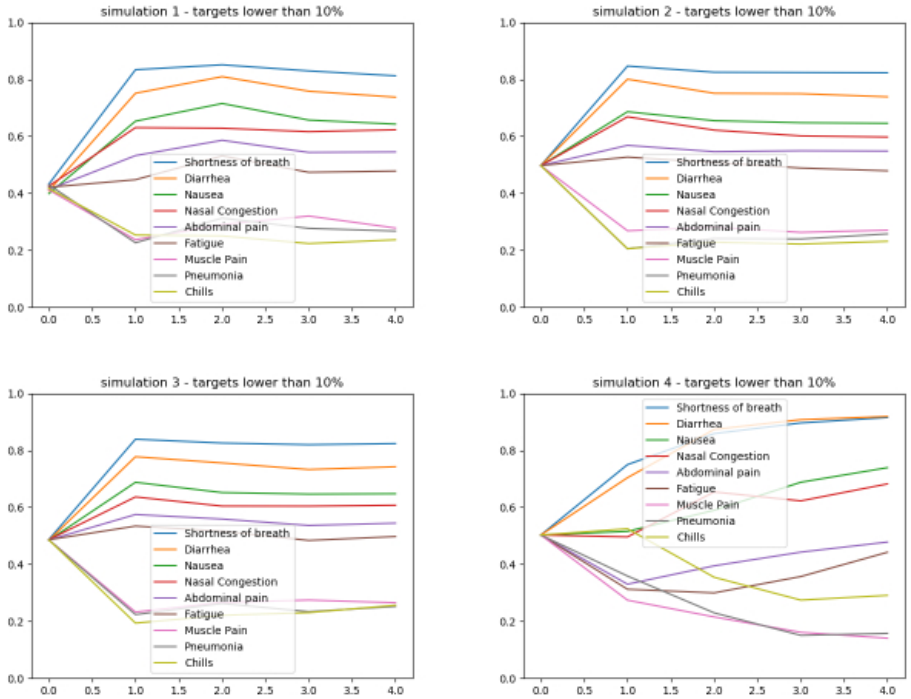

Figure 13: Simulations for 75% noise and target Symptoms exhibited by lesser than 10% of the Survey population.

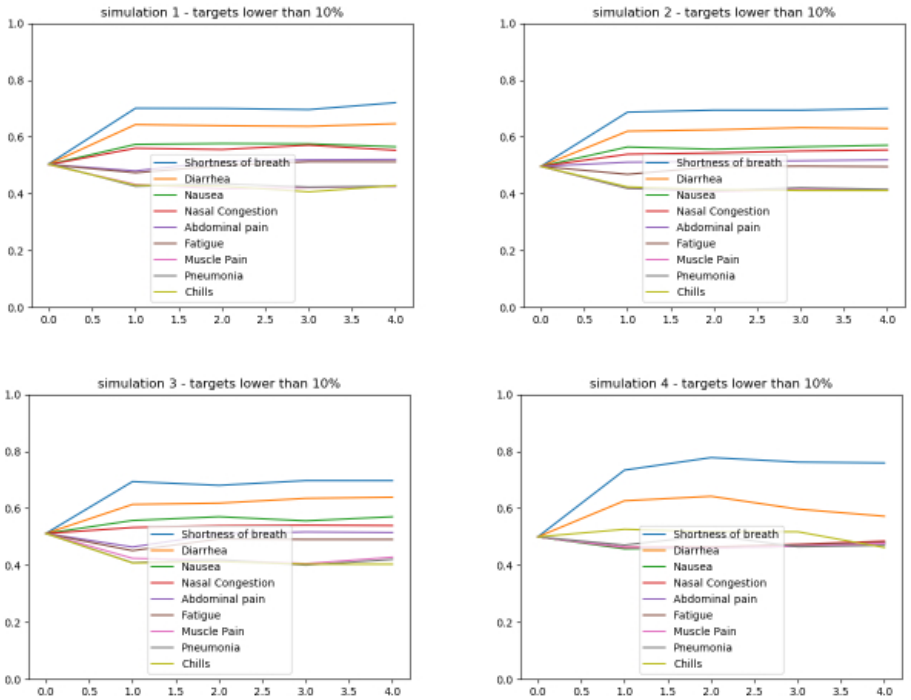

Figure 14: Simulations for 100% noise and target Symptoms exhibited by lesser than 10% of the Survey population.