A Sensor-based Learning Public Health System

Robert Steele
Medical University of South Carolina
steelerj@musc.edu

Andrew Clarke
The University of Sydney
acla6821@uni.sydney.edu.au

Abstract

New smartphone technologies for the first time provide a platform for a new type of on-person, public health data collection and also a new type of informational public health intervention. In such interventions, it is the device via automatically collecting data relevant to the individual's health that triggers the receipt of an informational public health intervention relevant to that individual. This will enable far more targeted and personalized public health interventions than previously possible. However, furthermore, sensor-based public health data collection, combined with such informational public health interventions provides the underlying platform for a novel and powerful new form of learning public health system. In this paper we provide an architecture for such a sensor-based learning public health system, in particular one which maintains the anonymity of its individual participants, we describe its algorithm for iterative public health intervention improvement, and examine and provide an evaluation of its anonymity maintaining characteristics.

1. Introduction

The recent rapid growth in both the capabilities and uptake of mobile devices with sensors or smartphones capable of acting as health sensor platforms has the potential to advance public health data collection and intervention. Whilst the majority of research and commercial focus to-date has been on mobile devices and sensors as a tool for individual health data capture, monitoring and feedback, the full implications for public health have been less well explored.

In this paper we build upon an underlying platform that provides a smartphone-based system for anonymized population health data capture and intervention [1] to present a novel sensor-based Learning Public Health System (LPHS). The underlying platform from prior research provides a novel methodology whereby 1) public health data can be collected without the individual being identified or subject to re-identification based upon their data; and 2) enables targeted public health interventions to be distributed, performed and evaluated without the need for the identifying details of an individual to ever leave their mobile device. The novel contribution of this current paper is the description and evaluation of a learning public health system and its iterative algorithm for refining public health interventions that can be built upon this underlying platform.

The underlying platform from previous research does not need a fully trusted central server, which might prove impractical on population-scale applications [1]. Beyond de-identification the approach taken also resolves the risk of re-identification based on quasi-identifiers, in the form of information known about individuals that could potentially be used to match with and re-identify the submitted data. The conventional approach to address this type of risk is to use a trusted server or aggregation point to combine and obfuscate/alter data to the point where k-anonymity [2] is assured for a data set, such that any individual is indiscernible from k other records based on quasi-identifiers.

The proposed Learning Public Health System involves an iterative algorithm that is extensible to numerous types of health sensor data collection, public health application areas and types of public health intervention.

2. Related work

The use of participatory sensing is of increasing interest in a number of application areas including air quality and pollution sensing [3] through the use of external air quality sensors, urban area noise level data [4], urban traffic analysis through the use of vehicle mounted sensors [5] and vehicle fuel efficiency [6], amongst many other applications.

The rich capabilities of participatory sensing have garnered interest in its usage for a range of such applications. This has in turn spurred a number of different approaches to resolving or decreasing the implicit security and privacy concerns when involving individuals in sensing/data collection. The more conventional approach would use a trusted server, then

URI: http://hdl.handle.net/10125/41577
ISBN: 978-0-9981331-0-2

k-anonymity [2] or a variant, to anonymize the data before it is accessible for research/analysis. Of course this approach suffers from the need for a fully trusted server as well as issues of a single point of failure in terms of privacy breaches. Alternatively, other approaches have improved on this by removing some sensitive information before submission (removal of identifiers and communications anonymity) with a central point of trust [7] to provide an anonymous approach. While this is quite effective when the sensing is collecting data on something not specific to the individual, this alone is not well-suited to a model where quasi-identifiers are a key submission component (such as in the case of collection of public health data) as de-identification protection is still implemented at a central trusted point.

To resolve the issue of requiring a fully trusted server, alternative approaches include decentralized participatory sensing networks [8] using user interaction/awareness as part of the approach or keeping the data managed by the participant [9] and stringent user-definable access control mechanisms to manage sharing. While these approaches may be extensible to some aspects of Health Participatory Sensing Networks (HPSNs) [10], they typically have not incorporated the need and importance of health interventions, an important aspect in HPSNs and a capability that does not have a direct parallel in most participatory sensing systems. Additionally, the capabilities that are beneficial in other areas may make these approaches overly complex for individuals, limiting their feasibility for a large scale implementation. For this reason, amongst others, the approach of users consciously building a “web of trust” as per [8] and the personal data vault of [9] were not used in our approach to a sensor-based LPHS. The LPHS however draws on some aspects of the anonymizing capabilities provided by a HPSN [10] in building the middle layer of a LPHS (see Figure 1).

Whilst the concepts of a Learning Health System [11] as put forward by the US Institute of Medicine have been published, there has yet to be work published in relation to the technical mechanisms for implementing this for public health interventions via such technologies as smartphones, sensors and anonymizing networks as are put forward in this current paper. This represents a significant contribution of this current work.

3. A sensor-based learning public health system architecture

The overall system architecture (Figure 1) involves a LPHS server that communicates with mobile devices through an anonymizing HPSN to provide communications anonymity, and mobile devices that incorporate local processing and privacy thresholds to maintain data anonymity/privacy/de-identification.

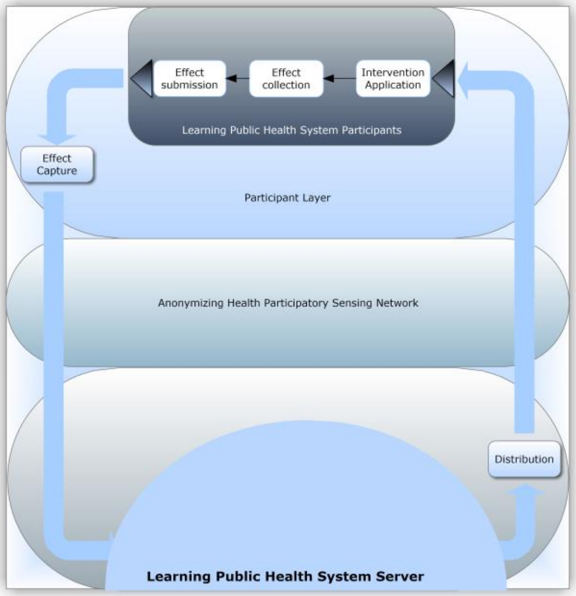
Effect submission
Effect collection
Intervention Application
Learning Public Health System Participants
Effect Capture
Participant Layer
Anonymizing Health Participatory Sensing Network
Distribution
Learning Public Health System Server

Figure 1. Learning public health system architecture

There are two primary data transmissions from and to the LPHS server respectively: (1) public health interventions are distributed from the LPHS server, and (2) intervention effect capture/ anonymized data collection submissions are sent to the LPHS server. The core functionality components of the LHPS server are (1) distribution of public health interventions; (2) aggregation of public health data; (3) analysis; and (4) support for public health intervention refinement.

The fundamental architecture can support different levels of public health intervention and public health data capture. The capabilities of the end user mobile devices as well as the level of participation in the public health interventions/ data collection tasks of the individual users of these devices will have implications for this also. We discuss these functional capabilities in the following subsections.

3.1 Smartphone-provided capabilities

When an individual utilizes just a smartphone without additional external sensors and the user is not required to take additional actions, this configuration has the advantage that it has the greatest level of existing deployment and ease of adoption – that is, smartphones without additional external sensors are the most common smartphone usage case.

3.2 Peripheral device sensors

An individual participating in the LPHS can also have the situation that they have additional external sensors connected to their smartphone. Increasingly additional health sensor data capture is available such as vital signs and blood constituent sensors [1]. An emerging area of application, but one still with substantive implementation challenges is the automated capture of dietary and nutritional intake information [12].

3.3 Intervention capabilities

The LPHS provides inputs to the individual while participating in the LPHS, to affect the health-related actions. Whilst an 'active' participatory sensing model for a typical sensing task might focus on achieving more complete data collection in terms of spatial/temporal range, LPHS-related active sensing would be more concerned with affecting a health-related action and hence have a component equating to a public health intervention. As such, the instigation to carry out 'active' sensing activities could essentially constitute a public health intervention input. Additionally, for LPHS purposes and key to the nature of a sensor-based LPHS, this allows for immediate and continuous feedback on the effectiveness of campaigns upon targeted groups.

4. Learning public health system algorithm

The LHPS algorithm (Figure 2) includes the following steps:

1. distribution of a public health intervention
2. effect capture
3. analysis of effect
4. identification of most and least successful effects
5. intervention modification and tailoring
6. return to step 1

The connection of participatory sensing with a learning public health system and this algorithm are novel contributions to the current literature, with the overall algorithm developed newly for the LPHS.

4.1 Distribution of a public health intervention

A public health intervention in the case of a sensor-based LPHS consists of an informational intervention: that is, some form of information - textual, multimedia or other - is distributed to the individual's device to affect positively a behavior that has a bearing on the individual's health. As of current technology, this device might most likely be the individual's smartphone, but this could be any device and the nature of the most suitable device for the receipt of an intervention will inevitably change over time even if only just due to technological advancement.

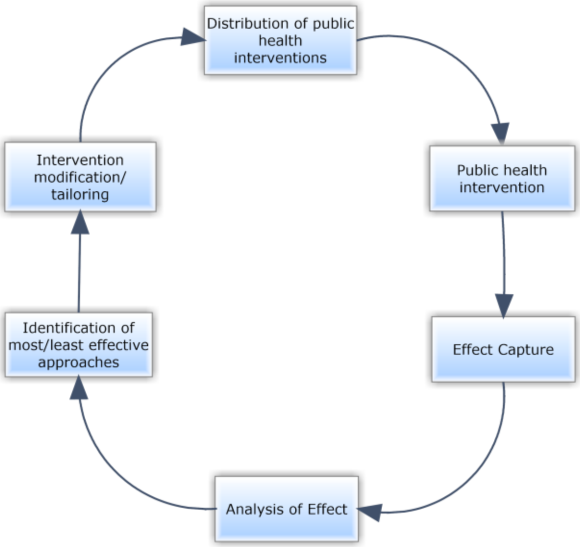
Distribution of public health interventions
Public health intervention
Effect Capture
Analysis of Effect
Identification of most/least effective approaches
Intervention modification/ tailoring

Figure 2. Learning public health system iterative algorithm

Firstly, a cohort to be the recipients of a public health intervention is identified. How this cohort may be refined is described in the following sub-sections, but an initial cohort is identified based upon some criterion/ criteria. For example, this may include the group who has high blood pressure or who are considered at risk of high blood pressure. The data necessary for this initial cohort identification can be determined via traditional clinical electronic records, or can be obtained anonymously via the underlying smartphone-based HPSN platform described in previous work [1, 10].

Secondly, for a particular cohort, an appropriate informational public health intervention is designed/ chosen. This will be a manual process, whereby public health intervention designers will determine what is the appropriate content, frequency of communication, mode etc. to be used.

The LPHS server provides the central component of the system. It will initiate the distribution of the informational intervention and this will be sent anonymously to the identified cohort, to their end-user mobile devices, but through the intermediary of the anonymizing HPSN (see Figure 1).

The public health intervention distribution mechanism is in theory scalable to very large numbers of recipients, that is, scalable to the sub-population or national population level.

4.2 Effect capture

The nature of the sensor-based LPHS and the underlying smartphone-based anonymizing information system, is that data from the recipients of the public health intervention, about the effect of the intervention, can start to be collected almost immediately. Again this information is collected from the recipients via the underlying anonymizing platform so that no individuals are identified.

The duration to wait into the course of the intervention for data collection to occur so as to assess the intervention, is nevertheless a function of the intervention itself. It would be determined by the intervention designers as to what duration of time should pass before the collected health sensor data would be a meaningful indicator of any potential effect of the intervention.

The LPHS server captures the incoming sensing data, but once again via the intermediary anonymizing layer. The underlying HPSN [10] provides a mix network [13] or onion network [14], which provides for anonymity of the submitter as well as secure communication. Such approaches utilize a chain of proxy servers, which can provide anonymity for both parties, though in this case it is only required for the mobile device user. Though this creates additional implementation complexity the potential benefit to real privacy is significant, with the only remaining significant privacy threat being the content of the data submitted.

As our approach incorporates submissions of variable resolution (that is submissions for the effect of the same intervention can provide more or less detail), the LPHS server works to integrate this data and provide any data cleansing as necessary.

For the minimum resolution of data the aggregation is straightforward as the more detailed submissions are just summarized to the same level [15].

4.3 Analysis of effect

Once the data of the cohort which has experienced the public health intervention is received by the LPHS server, these data can be analyzed in relation to various characteristics.

At one level, Online Analytical Processing (OLAP)-like analysis can occur slicing and dicing this data according to particular demographic factors or demographic combinations. Of interest will be which sub-cohorts saw improvement (or otherwise) from the intervention, and to what extent. The measures of improvement would again be a factor determined from clinical expertise input and from the designers of the public health intervention.

The LPHS server can also capture and calculate other metrics of interest for public health analysis by health organizations, other than specifically relevant to the evaluation of a given public health intervention.

4.4 Identification of most and least successful effects

OLAP-like analysis of the intervention effectiveness data will allow the determination of which sub-cohorts had the best effect from the public health intervention and which had the worst or least effect. This OLAP analysis will break this down according to demographic and demographic combinations.

In addition to an OLAP-like analysis of the received intervention effectiveness, machine learning and predictive analytics techniques can also be utilized upon the received intervention effectiveness data. Machine learning techniques could include clustering to identify particular clusters which either responded well to the intervention or which did not. These clusters may be described in a more complex way than being just based upon specific demographics. In any approach that utilized machine learning approaches, the intervention itself can be described in terms of parameter/ model inputs such as type of messaging, duration, frequency etc.

The intervention effectiveness data can also be used for the purpose of the application of supervised machine learning algorithms and in particular predictive analytics approaches. The intervention effectiveness data in effect constitutes a labeled data set that can be used to train, test and create a predictive model. Utilizing these techniques, the created predictive model could be used to predict which individuals may be most receptive to the given public health intervention in future.

4.5 Intervention modification and tailoring

Once the results of the analysis of a given public health intervention are known these can be potentially utilized in a number of ways to refine the public health intervention.

Firstly, based upon the OLAP analysis, the sub-cohorts for which the public health intervention was least successful, can have a modified public health intervention designed and applied. This would be determined ultimately manually by the public health intervention designers taking into account a wide range of factors. For example, there could be perturbations made to the prior public health intervention, that the intervention designers may manually determine may be improvements.

Secondly, the machine learning techniques can also provide an automated tool to help determine which factors were important in the success of an intervention, both in terms of the nature of the intervention itself, but also the characteristics of the sub-cohorts which either responded well or otherwise. A developed predictive model can be used to make predictions as to which intervention may be successful for a given sub-cohort for future interventions.

A combination of human and computationally-derived insights can inform the choices of the public health intervention designers. Of particular interest would be making improvements to interventions sent to sub-cohorts for which improvements in health measures were not seen following the original intervention. Making improvements for interventions that were sent to sub-cohorts for which improvements in health measures were seen, also of course will be important.

4.6 Repeat

The refined or modified public health interventions can subsequently be distributed. The same process of obtaining feedback via the sensor-based LPHS would also once again occur. In this way, an iterative learning public health system is enabled. Its intended effect, is that over time and via a number of iterations to incrementally improve upon public health interventions and thereby ultimately improve population health.

4.7 Case Study

It should be noted that the algorithm of the LPHS provides ‘learning’ in the sense that the iterative process is not one strictly limited to the application of ‘automated’ machine learning techniques, but the algorithm is such that steps 4 and 5 can utilize manual actions, automated machine learning/predictive models or a combination of both. Where machine learning/predictive models are used there are a plethora of such existing models available: support vector machines, k-means, decision trees, logistic regression, naïve bayes and ensemble approaches to name a few [16].

A simple case study where a largely automated machine learning technique such as support vector machines is applied might involve the following. An example intervention might be one that aims to lower the blood pressure of recipients with high blood pressure. The various characteristics of each individual would be known (albeit anonymously) via the LPHS, including such characteristics as demographic information and some physiological measures. These data would constitute the input features to the predictive model, in this case a support vector machine. The output variable for the support vector machine would be a categorical variable indicating the success/ level of success of the intervention on the individual: substantially lowered blood pressure, lowered blood pressure to a smaller degree, had no effect, increased blood pressure as some possible example success categories without quantifying the actual possible numerical ranges at this point. The values for this output variable i.e. the level of success of the intervention on a given individual, would be known from step 2 of the LPHS algorithm (Section 4.2 Effect Capture) and hence you would have a labeled data set on which to train and test the support vector machine predictive model. The support vector machine would then be trained on this labeled data set and would then provide a predictive model that for any new individual could now provide a prediction as to whether that particular individual would respond successfully (or in which category of success) to the intervention. This support vector machine predictive model could then be used in making the decisions as to which sub-set of the population to distribute a given intervention to.

For example, the decision might be made to only distribute the blood pressure public health intervention unchanged to those whom it is predicted their response will fall in the category of most successful response to the intervention.

5. Mechanism for intervention adjustment and targeting

A key capability of a learning public health system is the ability to redeploy the validated and tested targeted interventions to drive improved outcomes and participation.

Further the system needs to be able to provide not only the subjective evaluation of how participants who take part in interventions have been impacted but also that of control groups so as to provide a comparison. Due to the need for anonymity inherently part of our system, the control needs to be set at the mobile device level.

Additionally, to support the key capability of a public health system capable of learning and improving at a pace relevant to the modern world the system requires capabilities to modify the intervention strategy and approach dynamically, without losing the detail of the historical intervention pattern on the individual participant. This indicates the need for two types of intervention definition approaches, firstly a typical event type intervention that is deployed, utilized and then the effect captured. Secondly, an approach that is more similar to a continuous flow of or ‘stream’ of interventions, where the potential intervention streams are deployed, the stream of interventions are enacted over time and the effect capture is periodically submitted.

In the following subsections we detail the approach to allow local processing to define and maintain a combined intervention and control group of sufficient size and demographics to provide complementary information and maintain anonymity for the intervention participants. In addition, two main (and complementary) approaches, to achieve distribution of updateable public health interventions are covered: namely ‘hubs’ (see Section 5.3) and ‘streams’ (see Section 5.4) (Figure 3).

5.1 Establishment and monitoring of control subjects

The LPHS provides a level of anonymity, whereby the LPHS is not privy to the sensitive details of the individual or indeed interacts with the individual on a one-to-one basis. As such, it is not possible for the LPHS to explicitly define a control group. Within the capabilities of the LPHS, instead the decision to incorporate the user in the control or a specific public health intervention program will need to be decided at the individual’s device level. This process takes into account known demographic distributions to inform the decision and additionally, can take further logical decisions based on inclusion or exclusion rules defined within the intervention program.

The use of a decentralized control group decision making approach would of necessity require that the control group is larger than what would be strictly necessary to evaluate the performance of an intervention, we propose in this work that this can be kept to a level that would not overly impact the utility of the LPHS.

An additional challenge of the utilization of control groups within the LPHS is that if an individual is relegated to a control group, the motivation to participate and continue to collect data may be impacted.

To mitigate against this we suggest that the control group be structured so that while an individual may be in a control for a specific intervention program, they may be an active member of another mutually exclusive intervention program, as long as the goal and impacts do not overlap. Additionally, to retain users that would otherwise perhaps discontinue participation if they were allocated into a control is that local decision making could be made at the device to withdraw the user from the control group and mark the user’s data submissions with that additional metadata, indicating that the users data should not be considered during analytical reporting, while still allowing the individual to benefit from the LPHS.

5.2 Adjustment of intervention targeting and approach over time

A key guideline of the LPHS is to learn and then apply that gained knowledge to improve the operation of the LPHS. Therefore it is clear that the intervention programs, targets and approaches will evolve over time. This requires that the mechanics of replacing or updating the in-place intervention programs with consideration given to impact, flexibility and the retention of meaningfulness of previous results.

5.3 Hub intervention approach

The ‘hubs’ referred to are the bundles of intervention-related data being distributed from the LPHS server composed of: 1) the intervention information itself; and 2) additional logic required to enable intervention updates. This approach involves a complex intervention logic and content pre-determined by the LPHS server that will have some ability to refer to additional or modified interventions through a pull based approach. The intervention is replaced/updated with new content/logic periodically to apply new learned approaches to public health interventions. Replacement is based on timed-expiry/refresh cycles.

This is in many ways the more straightforward approach, though perhaps not the most practical in a LPHS. Due to the nature of a continually, learning health system the reality is that a full in-place replacement may need to provide logic to transition an individual from the current stage they are in an intervention to an equivalent in the updated program. Or in the case where targeting has changed, the individual may need to be moved out of the intervention program. The necessity to maintain the older logic and provide continuity until an expected end point is reached, creates a compounding level of complexity for the hub approach.

However, the hub approach allows for a single focal point for the participant with new interventions being detected based on the data of the individual and retrieved in a pull based manner and applied. It also means there will be far less duplication than would be likely in a stream approach, whereby each individual sequence needs to contain the logic necessary to guide an individual’s path through the intervention program, and it is likely that there will be large amounts of similar content or logic in the sequences.

5.4 Intervention stream approach

Rather than a single intervention control logic and content block, logic is distributed throughout a stream of linked simpler interventions. As such the starting point for a stream based approach is that there will be a continuous and on-going intervention campaign, rather than a once off intervention. The stream concept allows for the public health intervention designers to target and analyze interventions as an on-going interaction or 'stream' rather than a structured and finite interaction pattern. The participants receive at intervals a sequence of interconnected health interventions (called nodes in the sequence), where intervention logic and stream decision making logic is stored and processed. This allows the public health intervention designers to plan as many or as few node sequences ahead as required and to replace individual node sequences only if needed when an improved approach is required. This also allows for a very concrete way to split separate competing approaches/controls without concern for accidental/unplanned interaction.

This approach allows the LPHS to target specific nodes in the sequence and replace or remove them without impacting the rest of the logic – indeed individuals that have already passed through those nodes won't be affected either, even in cases where they are still taking part in an intervention program.

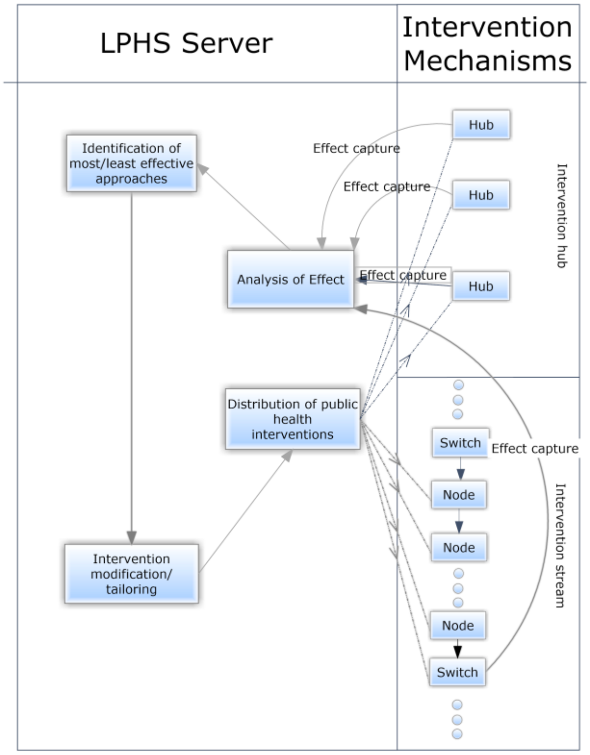

Figure 3. Learning public health sytem and intervention update mechanisms

5.4.1 Intervention switches

Intervention stream approaches will mostly be composed of lightweight intervention nodes that contain a single targeted intervention. However, at set intervals it'll be necessary to perform more detailed analysis and recalibration of the participant to a modified stream – this can occur at intervention switches. A switch is where a large number of streams come together at a specific point along the sequence. Similar to nodes, these switches contain logic for stream decision making though at a much more comprehensive level. Additionally, rather than performing interventions – these steps in the sequence execute the effect capture portion of the LPHS (see Figure 3).

6. Privacy threshold approach

The sensor-based LPHS by applying granular and modular restrictions upon data collection controlled by the user, reduces real privacy risks though high levels of user control of contribution and restrictions on data potentially usable for re-identification. Additionally, the use of a local processing approach to data submission and health interventions policies allows the on-device adaptation to achieve a data submission which matches the data request as closely as possible without breaching variable user defined privacy conditions.

The core concept of local processing (on the user mobile device) of health data for the LPHS requires that individual components of a data submission have an associated quasi-identifier score (QIS). Additionally, as the components are made more generalized such as for example a submission including the city of submission rather than specific postcode, the QIS would be lower to reflect the increased generality. The approach also takes into account the case where multiple quasi-identifiers are submitted together as such a group of quasi-identifiers will have a combined QIS value that is assessed against privacy thresholds. The four core data components and their QISs used in determining the combined QIS (θ_{LTDM}) are Measures (M_{QIS}), Location (L_{QIS}), Temporal (T_{QIS}) and Demographic (D_{QIS}). For details on how the QIS is calculated, see [1].

6.1 Public health interventions and feedback

Although other participatory sensing applications do not have a public health intervention component, parallels can be drawn between some interventions and participatory sensing that involves tasking. The use of targeted or personalized tasks/interventions would usually involve the LPHS knowing enough detail about the individual to provide this capability. However, to provide a higher level of privacy, targeting/personalization can be performed on the local device based on the much more specific detail available there. Additionally, the use of an anonymizing HPSN restricts the risk of the LPHS being aware of which individual mobile devices have received particular interventions.

In a hub approach to public health interventions the collection of intervention effect data is similar to other data submissions where the type of intervention and the metrics of success can be considered the 'measure' and the other details, the additional data components. The same approach can be taken in regard to privacy thresholds to ensure that whilst a very specific intervention can be issued, it is not reported as the specific intervention type, if to do so would violate a privacy threshold.

Stream interventions allow the public health intervention designers to take a more active role in the reporting of public health interactions. Based on the design of the streams the public health intervention designers can structure the intervention programs so that data collection steps are part of specific points in the sequence. Further, if the reporting points are major switching points, where a large number of individuals will traverse, key metrics such as the time interval between switching points, the path entered by and the path by which the individual exited the switching point will be able to be collected, without disclosing the finer detail of the individual sequence and detailed path the individual took which will be much more granular and hence raise a re-identification risk.

A potential way to design the switching steps, the points in the intervention interaction where the effect capture is conducted and detailed intervention approach decisions are processed, is to allow individuals with more relaxed privacy thresholds to contribute additional data at multiple stages of aggregation before the switching point. That is, the various public health intervention pathways begin to aggregate at some sequence points before the switching point, providing finer-grained data.

7. Privacy evaluation

To demonstrate the operation of the LPHS in terms of anonymity maintenance we evaluate an example data submission for the Greater Sydney Metropolitan area based on real data distributions. This fits the purpose of a typical public health intervention effect capture as well as typically such initiatives are targeted to a large area. Additionally, our LPHS approach aims to provide high levels of privacy for participants at a significant scale of intervention effect capture. As such, we consider the use of known population data and an analysis of the likely *k* values of intervention effect capture at varying levels of detail will provide a straightforward approach to compare the effective privacy in terms of the risk of re-identification.

This area of Greater Sydney has a population of 4,391,674 as of last census. Using the Australian Bureau of Statistics census population statistics [17] we generated a random data set based on the relative size of the demographics, specifically looking at gender, age range and ancestry based on parents' country of origin. Additionally, to create plausible intervention effect capture we then generated intervention application and effect data. Additionally, while the consideration and inclusion of control groupings is part of the capability of this approach, it doesn't have a measurable impact on the privacy considerations. This is because a proportion of the entire participant group that is large enough to provide an adequate control would be larger than the k values we are concerned with. Additionally, as there are no public health interventions performed against the control they can't be further identified by the type of intervention applied.

Assessing the LPHS anonymity maintenance characteristics we generated the data set out to a specific number of participant's intervention effect capture numbers: 50,000, 100,000, 200,000, 400,000 and 800,000. We then tallied the number of effect captures for *k*, under the thresholds of 50> k >= 20, 20> k >= 5, and k < 5. Having a small *k* value for a specific demographic is undesirable, as it can allow for potential re-identification or inference based attacks to be used against the data set. As such we can consider these k groups to represent low risk, moderate risk and high risk.

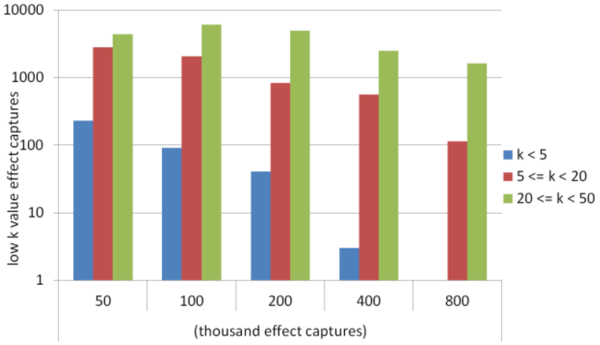

Figure 4. Effect capture *k* value without local processing

As can be seen in Figure 4, there were high numbers of effect captures with high or moderate risk at 50,000 submissions. Additionally, due to the large variation between the k value groups the chart is on a logarithmic scale. Over 3000 effect captures have a *k* value lower than 50 and 231 have a *k* value lower than 5. In practice this would be problematic in ensuring anonymity and privacy of data submissions. As, for example, if additional knowledge that an individual participates in the LPHS is available, it may be enough to perform re-identification of some individuals. As the number of effect captures are increased to 800,000 these risks diminish but there is still a reasonable potential chance of re-identification even at significant data collection levels of 400,000.

To improve this result we implemented our demographic formula which is part of the local processing approach and set a reasonably conservative threshold value for D_{QIS}. The other QIS scores M_{QIS}, T_{QIS}, and L_{QIS} were not significant values of the θ_{LTDM} and were not adapted. As ancestry was the optional value in this effect capture it was adjusted. If a D_{QIS} value for an individual was over the threshold based on known population demographics ancestry details were excluded from the effect capture.

As demonstrated in Figure 5 this resulted in a dramatic decrease in the number of effect captures that had low *k* values, with less than a tenth of the unadjusted submission approach. Again, due to the large variation between the k value groups the chart is on a logarithmic scale. This differentiation increased as the number of effect captures increased with the adjusted submission approach reaching a safe level much sooner at ~400,000 and comprising as low as .5% of the effect captures being below the k< 5 threshold at even the 50,000 submission level (Figure 6).

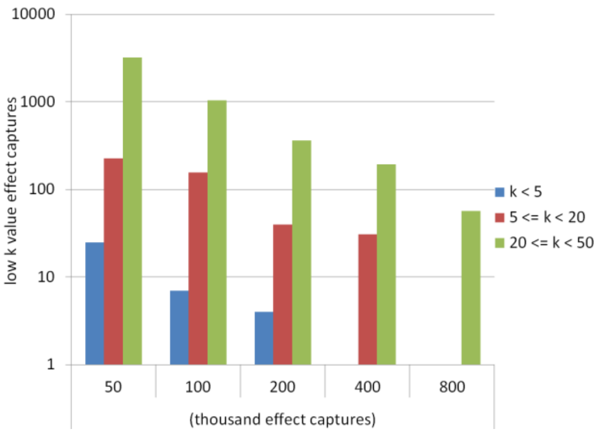

Figure 5. Effect capture *k* value with local processing

The threshold at the local device level could of course be adjusted either higher or lower based on the expected submission numbers. However, it performed quite respectably at the initial level with a significantly lower level of risk at the 50,000 submission level and close to no statistical risk at the 400,000 level which represents 9.1% of the area total population.

The limitation of this local processing approach as compared to a trusted server approach that performs *k*-anonymity, is that the number of other submissions cannot be known with certainty by the local device. As such, the privacy threshold is set at a conservative value to preserve privacy. However this means that when there are high levels of submissions more records are obfuscated/adjusted than was required.

In summary, for the example data set the LPHS performed favorably compared to the defined public health requirements and privacy limitations.

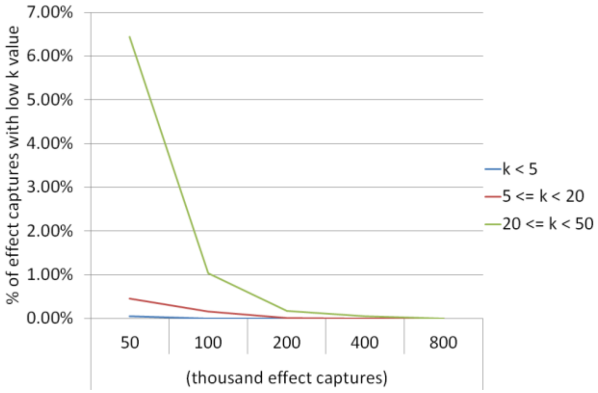

Figure. 6. Low k value effect captures as percentage of total effect capture

8. Discussion and future work

Our approach focused on alleviating privacy issues that would be inherent in developing such LPHSs. As such, the system would be quite resilient to extension via new sensors or sensor systems [18] as they would present just an additional data measure, where the key privacy restrictions are demographic, temporal and spatial-based. However, the extension of sensor capabilities potentially may reach the point where sensor systems are diagnostic in nature which would result in the measure itself being of a sensitive nature, in a similar manner to portions of a private electronic health record. These considerations can also be resolved within the bounds of the existing approach.

However, privacy and public perceptions of such LPHSs need to be further explored. As such, future work could include studies of perceived privacy of participatory sensing applications specific to the health domain. A useful extension of this approach would be to consider incentivization, adoption and health organization acceptance of such approaches.

In addition such LPHSs as described blur the lines between public health intervention and “ubiquitous computing”-based telehealth techniques [19]. Whereas the telehealth approach could use sensors in the care of an individual patient, the LPHS paradigm could involve providing targeted public health intervention benefits to a group of individuals.

Whilst the LPHS limits the amount of detail of the data flowing to the central server in the interests of maintaining anonymity, this does not preclude the maintenance of far more detailed health-related data on the individual's local device or portable personal health record [20]. More complex analysis of this data can also be carried out locally to benefit the healthcare of that individual [21], without transmitting this more complete data to the LPHS server.

9. Conclusion

This paper presents a sensor-based Learning Public Health System. This includes an iterative algorithm that can improve upon public health interventions over time via gathering feedback on the performance of previously distributed interventions. The paper also addresses the mechanisms for how iteratively updated interventions can be distributed. Finally, the LPHS has an emphasis upon maintaining the privacy of the individual participants in the LPHS who are the recipients of interventions. As such the anonymity preserving characteristics of the system are evaluated and results presented.

References

[1] Clarke, A., & Steele, R., "Smartphone-based Public Health Information Systems: Anonymity, Privacy and Intervention". Journal of the Association for Information Science and Technology, 66(12), 2015, pp. 2596-2608.

[2] Kalnis, P., and Ghinita, G., "Spatial K-Anonymity", in (Liu, L., and Özsu, M.T., 'eds.'): Encyclopedia of Database Systems, Springer US, 2009, pp. 2714-2714.

[3] Predic, B., Zhixian, Y., Eberle, J., Stojanovic, D., and Aberer, K., "Exposuresense: Integrating Daily Activities with Air Quality Using Mobile Participatory Sensing", Pervasive Computing and Communications Workshops, 2013 IEEE International Conference on, 2013, pp. 303-305.

[4] Wisniewski, M., Demartini, G., Malatras, A., and Cudré-Mauroux, P., "Noizcrowd: A Crowd-Based Data Gathering and Management System for Noise Level Data", in (Daniel, F., Papadopoulos, G., and Thiran, P., 'eds.'): Mobile Web and Information Systems, Springer Berlin, 2013, pp. 172-186.

[5] Ganti, R., Mohomed, I., Raghavendra, R., and Ranganathan, A., "Analysis of Data from a Taxi Cab Participatory Sensor Network", in (Puiatti, A., and Gu, T., 'eds.'): Mobile and Ubiquitous Systems: Computing, Networking, and Services, Springer, 2012, pp. 197-208.

[6] Ganti, R.K., Pham, N., Ahmadi, H., Nangia, S., and Abdelzaher, T.F., "Greengps: A Participatory Sensing Fuel-Efficient Maps Application", Proceedings of the 8th international conference on Mobile systems, applications, and services, 2010, pp. 151-164.

[7] Cornelius, C., Kapadia, A., Kotz, D., Peebles, D., Shin, M., and Triandopoulos, N., "Anonysense: Privacy-Aware People-Centric Sensing", 6th international conf. on Mobile systems, applications, and services, 2008, pp. 211-224.

[8] Christin, D., "Impenetrable Obscurity Vs. Informed Decisions: Privacy Solutions for Participatory Sensing", Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on, 2010, pp. 847-848.

[9] Mun, M., Hao, S., Mishra, N., Shilton, K., Burke, J., Estrin, D., Hansen, M., and Govindan, R., "Personal Data Vaults: A Locus of Control for Personal Data Streams", Proceedings of the 6th International Conference on emerging Networking EXperiments and Technologies, 2010, pp. 1-12.

[10] Clarke, A., & Steele, R. "Health Participatory Sensing Networks". Mobile Information Systems, 10(3), 2014, pp. 229-242.

[11] Friedman, C. P., Wong, A. K., & Blumenthal, D., "Achieving a Nationwide Learning Health System". Science Translational Medicine, 2010, 2(57).

[12] Steele, R., "An Overview of the State of the Art of Automated Capture of Dietary Intake Information". Critical Reviews in Food Science and Nutrition, 55(13), 2015, pp. 1929-1938.

[13] Sampigethaya, K., and Poovendran, R., "A Survey on Mix Networks and Their Secure Applications", Proceedings of the IEEE, 94(12), 2006, pp. 2142-2181.

[14] Mauw, S., Verschuren, J.H.S., and Vink, E.P., "A Formalization of Anonymity and Onion Routing", in (Samarati, P., Ryan, P., Gollmann, D., and Molva, R., 'eds.'): Computer Security – Esorics 2004, Springer Berlin Heidelberg, 2004, pp. 109-124.

[15] Clarke, A., and Steele, R., "Summarized Data to Achieve Population-Wide Anonymized Wellness Measures", Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, 2012, pp. 2158-2161.

[16] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. "Top 10 algorithms in data mining." Knowledge and Information Systems, 2008, 14(1), 1-37.

[17] ABS, "Census Community Profiles Sydney"2011, http://www.censusdata.abs.gov.au/census_services/getproduct/census/2011/communityprofile/1GSYD, accessed 28/03/2013

[18] Swan, M. "Sensor mania! The Internet of Things, wearable computing, objective metrics, and the Quantified Self 2.0." Journal of Sensor and Actuator Networks 1.3, 2012, pp. 217-253

[19] Steele, R., & Lo, A. "Telehealth and Ubiquitous Computing for Bandwidth-constrained Rural and Remote Areas". Personal and Ubiquitous Computing, 2013, 17(3), 533-543.

[20] Steele, R., & Min, K. "HealthPass: Fine-grained Access Control to Portable Personal Health Records". 24th IEEE International Conference on Advanced Information Networking and Applications, 2010, pp. 1012-1019.

[21] Steele, R., & Lo, A. "Future Personal Health Records as a Foundation for Computational Health". In Computational Science and Its Applications–ICCSA 2009, 2009 (pp. 719-733). Springer Berlin Heidelberg.