Identifying the Main Causes of Medical Data Incompleteness in the Smart Healthcare Era

Colin Wilcox, Soufiene Djahel and Vasileios Giagos
Department of Computing and Mathematics, Manchester Metropolitan University, UK

***Abstract*—Incomplete data due to discrepancies between medical data sources and their storage methods represents a serious concern as it may lead to the loss, or misrepresentation of important medical information. This concern is anticipated to grow in the era of smart healthcare as the volume, variety and speed at which medical data is collected will increase significantly. This paper aims to identify the main causes of data incompleteness in the medical domain, discuss some techniques currently used to build a complete medical picture and highlight how they may affect the consistency and accuracy of the collected data. It also outlines future research directions to efficiently handle data incompleteness and its consequences.**

***Index Terms*—Medical systems, Healthcare, Medical data incompleteness, Medical data accuracy.**

I. Introduction

The digital revolution and the associated increase in the volume of data in today's society has raised many questions regarding personal information accuracy, security and the ethical usage of an individual's personal information [1]. Although this problem is not unique to the medical industry, it is critical that personal medical information is kept secure and protected against unauthorized access, reducing the opportunities for manipulation and potential exploitation. In the future, it is foreseen that the information associated with an individual will not be restricted to just direct medical data but will also contain more granular information such as a person's movement and access history within medical facilities. Therefore, the quality and accuracy of such data is essential to preserving the security of access. Modern techniques involving deep learning and data analysis can be used to address the specific issues around data incompleteness while preserving an individual's ability to access permissible resources [2]. It is worth mentioning that in smart healthcare era, medical systems will collect data through various smart medical devices, which could also include fitness devices like the myriad of wearables available in the market and used to collect medical and non-medical data. The consumer storage technology used to store such data can lead to data incompleteness, as discussed in Section II.

There are many well-known benefits regarding the impact of electronic medical records on patient safety and the speed of patient diagnosis. However, there is evidence [3] that many medical records are poorly documented, which introduces challenges to determine their efficacy. There is some discrepancy between what a medical professional considers to be 'complete' and the more widely accepted definition. Any data transition to electronic storage of information has a direct financial implication, and ineffective or incomplete medical records lead to a waste of resources, in terms of time, effort, and money. Ineffective records can undermine patient safety and lead to potential medical errors. The study conducted in [4] indicated that both quantitative and qualitative (medication etc.) inaccuracies together with incompleteness (documentation etc.) may exist in medical records. Moreover, the study presented in [3] has shown that electronic medical record inaccuracies and incompleteness in plastic surgery cases may come in many forms. The researchers studied a sample of plastic surgery patients to determine the usefulness of their medical records and identified that although basic personal information was recorded correctly (i.e., name, date of birth etc.), there were issues with the more critical information, in particular medical and surgical histories. The most common errors focused around incomplete medication details, allergies, and intolerance to certain drugs as well as missing medical procedures. These omissions may have an impact when making current potential medical decisions (both in terms of delays and diagnosis) and could affect patient safety.

In the remainder of this paper, we will focus on highlighting the main sources of medical data incompleteness, categorizing incomplete data into three main categories, and providing an overview on the main approaches used to reduce the degree of such data incompleteness. We will also discuss some techniques currently used to build a complete medical picture, such as surveys, medical records and claims data, and outline their respective advantages and limitations with regard to data completeness and accuracy. Finally, we present potential future research directions to efficiently deal with data incompleteness and avoid or alleviate its consequences.

II. An Overview on Data Incompleteness

Due to the long-term nature of an individual's medical history, personal medical data has a tendency to be stored historically in a localized manner, across different potentially incompatible systems with each using a different method of encoding and information capture. This lack of consistency increases the risk of data being misrepresented or badly stored in the local data medium and can lead to errors or misrepresentation of data. A generalised approach to the aggregation of data from such incompatible sources can be shown in Figure 1. In this technique, a data adapter for each type of data source is used to convert the original data into a standardised format. As new data source types are included a new adapter must be added to the system. It is worth mentioning that several data sources might be of the same type and thus can use the same

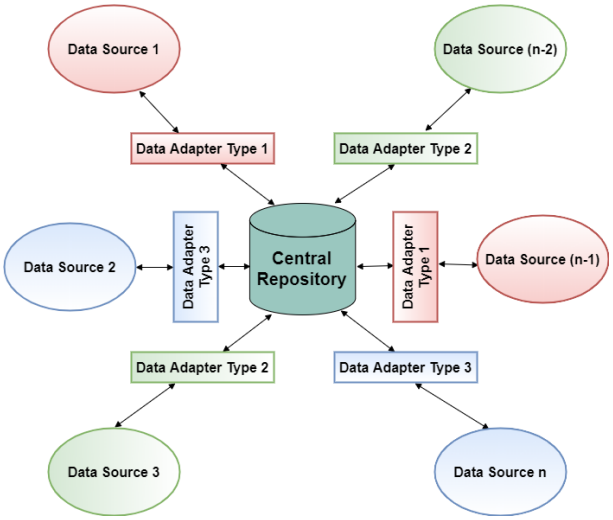

Fig. 1. Aggregating data from multiple sources

adapter, such as data sources 1 and (n-1) for example. Errors or misrepresentation of data are caused by several factors including human mistakes (typographical errors, bad transcription), erroneous (or inaccurate) measurements or faulty equipment resulting in missing or vague attribute values for certain records [5]. Regardless of the cause the result is inaccurate data with lower quality which significantly affects its potential use and effectiveness.

Incomplete data can be categorized based on whether it may be ignored or not, as shown in Figure 2. We can distinguish three categories of missing data [2]. (i) MCAR (Missing Completely at Random): the nature of missing data that has no identifiable pattern to its omission and so is considered to be randomly missing. (ii) MAR (Missing At Random): the distribution of missing values for an attribute depends on the observed data, but does not depend on other missing values. (iii) MNAR (Missing Not At Random): there is some sort of pattern or reason to the data being missing; the distribution of missing values for an attribute depends on other missing values. This third category has been included here for completeness purposes only and as such we shall focus on MCAR and MAR only in this paper.

There are two broad approaches when attempting to reduce the degrees of data incompleteness from a given data set. The first consists in systematically removing all attributes from data records which have missing values and then removing all records which have missing attributes. If we consider a domain of data to be a two-dimensional table of values, this would be analogous to removing all the columns (attributes) and then any rows (which had missing columns) [6]. While this approach ensures the creation of a complete data set (the largest complete subset) it discards partial information (e.g. censoring in epidemiology). Furthermore, the size and effectiveness of this approach lessens when the missing data is not clustered, that is it is more spread out, since this will mean more attributes and data rows will be removed, producing a small final subset of data. Such an approach is more effective when the percentage of records with missing data is relatively low or the same attributes are missing in a large number of records, so as to provide a complete data set without losing a large proportion of the information that has been recorded.

The second approach does not remove any information from the data set at all, but instead uses imputation techniques as a means of patching the holes using a variety of statistical methods. Imputation techniques can take two main forms depending on whether the data is anonymized [2] or not [7]. Such strategies may include data imputation according to predefined rules (e.g. use the modal or average attribute value) or a prediction from a model. The model training can be done on a subset of the available complete records but can also make use of external sources [2], [4]. This 'best guess' approach may be absolute or based on a statistical likelihood that a given field has a certain value (for example, a missing binary gender field has an 80% likelihood of being male). There are many different approaches that can be used such as [8], and [9]. The suitability of this approach depends on how varied the data set is. The larger the number of records which have a similar set of attributes to the incomplete data record under scrutiny, the more confident an observer will be in accepting any missing attributes being replaced using this data subset. Neither approach is perfect and can introduce assumptions that would inevitably skew the original meaning of the modified data.

There are many strategies and techniques that could be used as a basis for data imputation [4]. Traditional data agnostic techniques can be used with both MAR and MCAR data types. They are more simplistic, fast, and easy and can produce unbiased results [2]. However, such techniques may affect potential data correlations, reduce statistical effectiveness, and lead to the loss of the source of data. On the other hand, modern data-centric approaches can also be used with both data types (MAR and MCAR), provide unbiased estimates, preserve sample size and statistical relationships, and are available in a range of software packages using machine learning techniques. However, these techniques are not effective for sparsely populated data sets, can be mathematically intensive which may be difficult to program, and some of their algorithms may be computationally expensive [2].

One highly visible and current example of where incomplete or inaccurate medical data has a direct impact on individuals' health is the current COVID-19 crisis. In this case, there were insufficient testing kits stockpiled within medical facilities, and as a result, many potentially infected people could not be tested in a timely manner leading to an inaccurate measure of the rate of spread of infection. Such a direct current example emphasizes the need for medical data to be as up to date and accurate as possible at all times.

III. Medical Data Collection Techniques

There are many different sources of medical data each of which has their own unique advantages and drawbacks with regard to the accuracy and completeness. These include but are not limited to traditional approaches such as surveys,

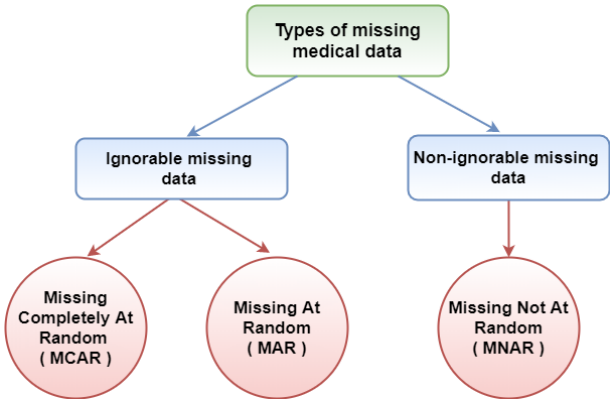

Fig. 2. Categorization of missing data

disease registries and medical claims data but would also include capturing physical attributes from smart devices such as wearable monitors, mobile phones and implants [10]. For the purpose of comparison and to establish any common trends we will restrict our analysis to surveys, claims data and medical records.

A. Questionnaires/Surveys

Depending on the way in which the survey is designed, many empirical data can be collected in a relatively short period of time, allowing large samples to be targeted quite readily. Like all types of data collection, questionnaires have their failings as well. People can introduce recall bias or ignore more complicated or personal questions, thereby introducing gaps in the sample data which is a reflection of the user sample. Secondly, the distribution mechanism may introduce an unintended bias in the data due to the exclusion of certain portions of the user domain. Such exclusions would produce significantly different results from those otherwise expected, for example, homeless people would likely not have access to an email distributed survey. Such biases would cause the results to be non-representative of the sampled population since it would be disproportionately skewed by particular traits or attributes that impacts on the results.

Another potential source of response bias, which can be prevalent in the medical domain, is the deliberate falsification of answers to sensitive medical questions where the subject may be unable or unwilling to provide certain key pieces of information for personal reasons. The distribution of missing data could be completely random, could be influenced by external factors or can even be influenced from the attribute that is missing. As an example, we will consider the subgroup of people with no fixed adobe responding to a question about medical conditions. It might be unlikely to have an email address (a socio-economic example of random missing information that is not related to a particular individual's medical condition). Furthermore, we can attribute a potential questionnaire non-response to a direct medical condition such as drug addiction (an example of "not at random" or response bias).

An important part of surveys design process involves creating clear and precise, and most importantly open questions to minimize the chance of misunderstanding and recording of incorrect information [11]. Questions should never lead the participant towards any particular answer. The expectation was that such open questions would lead to more standardization and better objectivity while reducing any directed questions that would lead to bias in the responses.

B. Medical Records

Personal medical records are a paper trail showing interactions between an individual and medical facilities over a period of time. These are personal to the individual but can be used to identify trends in patient characteristics as well as trends in health care access and quality. Electronic Health Records (EHRs) act as a computerized version of a traditional paper-based medical record. EHRs have been available for decades but the growth and increased use of computer technology in public space has meant that electronic recording of patient histories has only relatively recently been commonplace [12]. As such, there is a large legacy of paper based records which either need to be digitized (leading to the potential of miscoding/loss of data) or transition away from – leaving two disparate medical trails.

The electronic representation of medical records allows the information they contain to be easily shared and distributed between different authorized entities and allows a level of consistency that was not previously available. From the opposite standpoint however, such dissemination of data becomes problematic when the source of such data has missing, erroneous or inconsistent information. It is therefore important to make sure the source of all such bodies of data is as complete as possible. EHRs tend to be medically accurate, as the source of the information tends to come from medical sources; however, there can be a degree of misinterpretation and loss of medical context.

Over the lifetime of a patient, the individuals' medical record can be shared and stored on many different locations. While this promotes real-time access to data, it also exposes these isolated data repositories to the risk of becoming inconsistent. The process of synchronizing these isolated data becomes progressively more difficult as the number of end points increases. There are two possible solutions to this problem; the first is to have a single managed source for EHR, providing a single point of trust but also a single point of failure. The second is to manage the interconnections between all data locations and manage the real-time synchronization between them.

Consider the situation where we have n separately managed sources of a person's medical data. In a real world scenario, the data managed locally at each of these data sources could be changed and updated independently with a worst case scenario being all of these sources of information containing similar, but different, versions of a person's medical data. Since these sources are independent there is no need to ensure that the data is consistent with data from any of the other sources. The

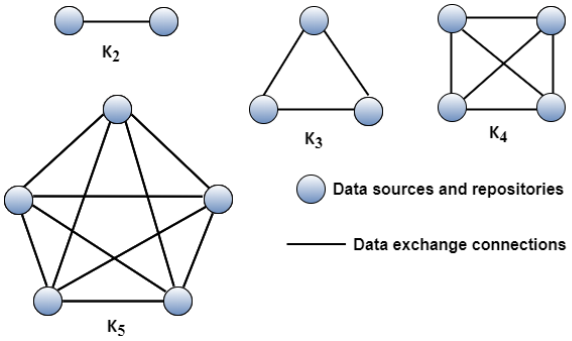

Fig. 3. Series of non-trivial n-node complete graphs

only requirement is that changes are self-consistent in their own isolated environments. With the centralized repository model, shown in Figure 1, there will be a need to manage each of n bidirectional connections between the set of endpoints and the central repository. Changes to data at individual endpoints must be synchronized with the central repository and notifications of the change should be sent to each connected nodes to ensure all endpoints are aware of the changes. The obvious issue here would be how potential conflicts of data are resolved when changes to the same piece of medical data occurs at different nodes within the same time frame. Which one is considered to be correct in this situation?

The second approach does away with a need for a central repository in favour of changes being propagated to all storage locations. From a topological perspective this results in a fully connected (complete), bidirectional graph with n nodes (local data sources). The connections are bidirectional as it is feasible that a node can push a change across the network. The number of bidirectional connections in this scenario can be shown to be $\frac{n(n-1)}{2}$. This is a quadratic function indicating that as the number of data sources, n, increases, the number of connections that need to be managed increases non-linearly. The family of graphs, K_n, shown in Figure 3, denotes the set of complete graphs containing n nodes and the graph plotted in Figure 4 shows how quickly the number of connections increases making this model less practical to manage and maintain.

1) Proof by Induction: Let S_k be the number of bidirectional connections in a fully connected network, K_k, containing k nodes. As explained above this can be written as:

$$S_k = \frac{k(k-1)}{2} \tag{1}$$

Adding a single new node to such a connected network will mean connecting this node to each of the already existing k nodes, a total of k new bidirectional connections. From this description the new fully connected network with $(k+1)$ nodes has a total number of bidirectional connections C, defined as follows.

$$C = S_k + k$$

According to Equation. 1 we can expand this as described below.

$$\begin{aligned} C &= \frac{k(k-1)}{2} + k \\ &= \frac{k}{2}((k-1)+2) \\ &= \frac{k}{2}(k+1) = S_{k+1} \end{aligned} \tag{2}$$

We can prove that this equation holds for all sizes of a fully connected network, using mathematical induction. Considering the smallest (non-trivial) fully connected network, K_2, with two nodes we have, according to Equation. 1, the following. $S_2 = \frac{2}{2}(2-1) = 1$, which is obviously correct as shown in Figure 3.

Taking the formula for S_k as the induction basis and S_{k+1} as the induction step we can therefore conclude that for any number of nodes, $n \in \mathbb{N}$ $(n \geq 2)$, in a fully connected network K_n, there will be a total of $\frac{n(n-1)}{2}$ bidirectional connections.

C. Claims Data

Medical claims data (administrative data) is similar to electronic medical data but the scope is much wider. This provides a much larger data pool for analysis purposes. The data is culled directly from physical notes of health care providers and tends to be recorded in the presence of the patient. This approach leaves certain facts open to interpretation by the transcriber, which in turn can lead to conflicting data for a patient.

It is common practice for medical providers to use a standardized system for billing, using predefined codes to identify which health care services are provided [13]. This standardization allows relatively easy cross comparison on medical services that have been provided across different medical providers. A patient's privacy needs to be respected and so in order for such data to be used as part of an analysis process it must be anonymized so that individuals cannot be identified from the data. The ability to access such a large amount of information about the health of a population helps researchers to gain valuable statistical information across different patient demographics. Such demographic based analysis can help employers to identify targeted wellness initiatives based on their employees profile to reduce future medical claims as well as to highlight any trends or hidden risk factors in particular subsets of the workforce or working location.

D. Analysis and discussion

Each of these data collection methods has advantages and limitations. Surveys are popular among researchers as they provide a high representation of a population's capability as people tend to answer surveys and, therefore, the data being collected tends to be more representative of the population. This can be contrasted with other methods of data capture which may provide a less representative data set. The relative cost is quite low, the main cost being the production of the surveys as well as many convenient ways to gather the

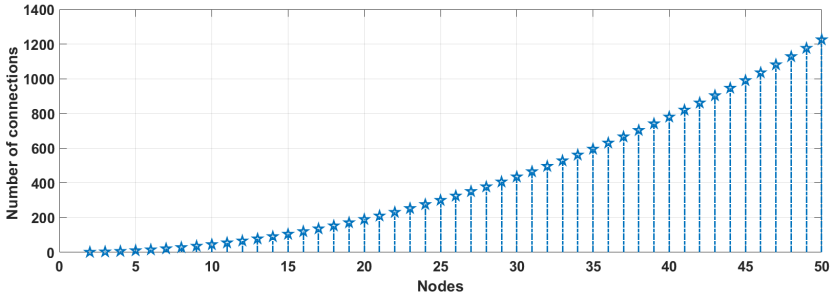

Fig. 4. Connections growth as nodes increase

results. Surveys offer little or no observer objectivity in that all participants get the same questions in the same manner and, therefore, any inherent bias in the delivery is removed. The two main issues with surveys are their lack of flexibility in design and their unsuitability for any conversational type interactions.

Medical claims data has a number of advantages as it provides a broader range of information relating to the patient and contains a full patient history showing all historic medical events and interactions. Medical records, however, would record a much narrower set of data restricted to the portion relating to actual care only. Such detailed record-keeping allows medical practitioners to assess whether or not a patient is taking a medication as directed or not. This evaluation cannot be done using other methods that may not record medical activity with this level of granularity. The downside of claims data is that it can be limited in the richness of the data provided which is limited to procedural and diagnostic events. Medical records, however, may provide a more data-rich set of information covering other medical attributes such as habits and other problems that claims data would not record. Electronic medical records on the other hand would record an individual's habits, vital signs, and history results from surveys, etc.

IV. Future Research Directions

Incomplete data sets provide a significant hindrance to real world applications of data analysis such as pattern discovery and data classification. This is magnified as the data sets get larger and more encompassing. Better ways of identifying incomplete data sets are required in order to remove the uncertainty of data imputation. Alternatively, a standardized method of recording medical data needs to be agreed and adopted in the wider industry to eventually make such partial data sets less and less of a relevance over time.

The consideration that seems to be neglected in the current research is the nature and distribution of the missing data itself. Pertinent information about the data itself may be interpreted based on the type and distribution of the gaps across the data domain. One way to remove the inconsistency in recording personal data is to use new technology to automatically capture more autonomic personal information without the need for human interaction, thereby removing the likelihood of encoding errors or misrepresentation of the facts. The advent of personal wearable devices, such as smart watches, allows certain personal metrics (e.g., heartbeat, temperature etc.) to be compared directly against an individual's personal medical record along with an accurate time-stamp of its recording [14]. Such measures are a ready made way to build a historic record over time of such information that is both unambiguous and complete.

Rather than using personal technological devices such as smart watches, it is also possible to adapt the environment with new technology. Many places, such as airports, are starting to use a combination of low powered Bluetooth beacon technology and Wi-Fi hot spots to create a network overlay on top of physical environments that allows the monitoring and tracking of people's movements within the space [15]. These ideas could be used to predict a person's future movements for the purpose of detecting illegal access to secure areas or detecting medical episodes affecting its mobility. Such technologies could work together to reinforce each other's readings to verify the accuracy of the information they are providing. This idea could easily be adapted to any public space, including medical facilities. However, it should be considered that such hybrid/composite solutions tend to be very environmental specific as, for example, Bluetooth beacons have a limited range in which they are useful and whose signal is adversely effected by reflective and glass surfaces, such as the walls commonly found in airports and hospitals. Similarly, many hospitals may not be fully covered by an adequate Wi-Fi system which could lead to drop outs in less well covered areas.

V. Conclusion

Lessons learned from this study suggest that current statistical techniques for 'filling in the gaps' in medical information data sets create further inconsistencies. Many approaches are concerned with just removing the gaps in the data rather than giving any consideration to the nature and distribution of the omissions which may in itself be significant. By analysis, patterns or commonalities in the locations and distribution of missing information may itself shed light on inadequacies in the data sources and the ways in which the original data was encoded. Advanced technology certainly has a part to play in any future solutions to gathering, validating and recording medical data. The caveat, however, is that external constraints

need to be considered as well as technological choices in order to develop the most effective solutions.

REFERENCES

[1] Z. Li, V. Sharma, and S. P. Mohanty. Preserving data privacy via federated learning: Challenges and solutions. *IEEE Consumer Electronics Magazine*, 9(3):8–16, 2020.

[2] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page. A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access*, 6:63279–63291, 2018.

[3] C. J. Hong, M. N. Kaur, F. Farrokhyar, and A. Thoma. Accuracy and completeness of electronic medical records obtained from referring physicians in a hamilton, ontario, plastic surgery practice: A prospective feasibility study. *Plast Surg (Oakv)*, 23(1):48–50, 2015.

[4] J. Luengo, S. García, and F. Herrera. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowl Inf Syst)*, 32:77–108, 2012.

[5] M. A. Bochicchio, L. Vaira, E. Cicinelli, and A. Vimercati. Dealing with incompleteness in multidimensional analysis of health records: An experience on fetal growth. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 1032–1038, 2015.

[6] E. Weitschek, G. Felici, and P. Bertolazzi. Clinical data mining: Problems, pitfalls and solutions. In *2013 24th International Workshop on Database and Expert Systems Applications*, pages 90–94, 2013.

[7] L. Rocher, J. M. Hendrickx, and Y. de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun*, 10, 2019.

[8] G. K. Vishwakarma, A. Bhattacharjee, J. Jose, and R. Kumar V. Cancer patients missing pain score information:- application with imputation techniques. *Epidemiology biostatistcs and public health*, 13(4), 2016.

[9] L. Kotze. Pns222 imputation techniques for missing covariates when modeling disease progression. *Value in Health*, 22:S323, 2019. ISPOR 2019: Rapid. Disruptive. Innovative: A New Era in HEOR.

[10] S. J. Olshansky, B. A. Carnes, Y. C. Yang, N. Miller, J. Anderson, H. Beltrán-Sánchez, and K. Ricanek. The future of smart health. *Computer*, 49(11):14–21, 2016.

[11] N. Gafni, A. Moshinsky, and J. Kapitulnik. A standardized open-ended questionnaire as a substitute for a personal interview in dental admissions. *Journal of Dental Education*, 67(3):348–353, 2003.

[12] N. Menachemi and T. H. Collum. Benefits and drawbacks of electronic health record systems. *Risk Management in Healthcare Policy*, 4:47–55, 2011.

[13] K. C. Wang, J. B. Patel, B. Vyas, M. Toland, B. Collins, D. J. Vreeman, S. Abhyankar, E. L. Siegel, D. L. Rubin, and C. P. Langlotz. Use of radiology procedure codes in health care: The need for standardization and structure. *RadioGraphics*, 37(4):1099–1110, 2017.

[14] H. Zhu, C. K. Wu, C. H. KOO, Y. T. Tsang, Y. Liu, H. R. Chi, and K. Tsang. Smart healthcare in the era of internet-of-things. *IEEE Consumer Electronics Magazine*, 8(5):26–30, 2019.

[15] D. Surian, V. Kim, R. Menon, A. G. Dunn, V. Sintchenko, and E. Coiera. Tracking a moving user in indoor environments using bluetooth low energy beacons. *Journal of Biomedical Informatics*, 98:103288, 2019.