Um Estudo Comparativo de Mecanismos de Privacidade Diferencial sobre um *Dataset* de Ocorrências do ZIKV no Brasil*

Daniel de Oliveira[1], Eduardo Rodrigues[2], Serafim Costa[2], Paulo Amora[2], Asley Caldas[2], Kary Ocaña[3], Marco Horta[4], Ana Maria de Filippis[4], Vânia Vidal[2], Javam Machado[2]

[1]Universidade Federal Fluminense (UFF)

danielcmo@ic.uff.br

[2]Universidade Federal do Ceará (UFC)

{eduardo.rodrigues, serafim.costa, paulo.amora, javam.machado}@lsbd.ufc.br

vvidal@lia.ufc.br

[3]Laboratório Nacional de Computação Científica (LNCC)

karyann@lncc.br

[4]Fundação Oswaldo Cruz (Fiocruz)

{marco.horta,abispo}@fiocruz.br

Resumo. *Nos últimos anos, o Governo Brasileiro tem organizado uma série de iniciativas para informatizar o SUS, com o objetivo de melhorar sua eficiência. Uma dessas iniciativas é o GAL (Gerenciador de Ambiente Laboratorial). O GAL tem como objetivo proporcionar a gerência das rotinas laboratoriais e o acompanhamento das etapas para realização dos exames. Adicionalmente, o GAL permite a extração de dados, que podem ser usados por gestores nas diversas esferas. Entretanto, essa exportação de dados pode não ser confiável, e levar a sérios riscos de violação de privacidade, uma vez que exibe dados pessoais de indivíduos. Simplesmente mascarar os elementos de identificação (nome, CPF, etc.) ou disponibilizar apenas resultados agregados pode não proporcionar proteção suficiente. Nesse cenário, técnicas mais elaboradas de privacidade de dados, como a Privacidade Diferencial (PD), se fazem necessárias. Este artigo apresenta um estudo que compara a aplicação de diferentes mecanismos de PD sobre os dados extraídos do GAL. Em especial, utilizamos como estudo de caso os dados de exames de casos suspeitos do Vírus da Zika (ZIKV) no Brasil.*

1. Introdução

Durante as últimas décadas, as comunidades da computação e da saúde têm despendido grandes esforços para prover soluções computacionais na área da saúde [Li et al. 2019]. Segundo [Shishvan et al. 2018], essas soluções variam desde o monitoramento de pacientes internados até a gerência de recursos na área da saúde. Esse tipo de iniciativa pode ser percebido também no Brasil [Silva et al. 2019]. Um dos exemplos é o *Sistema Gerenciador de Ambiente Laboratorial* (GAL)[1] do SUS. O GAL tem como objetivo informatizar a rede laboratorial de saúde pública brasileira, *i.e.,* ele registra informações de amostras de origem humana e animal que possam ter sido expostas à doenças, possibilitando que profissionais da saúde possam consultar,

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores agradecem também ao CNPq e FAPERJ por financiarem parcialmente a pesquisa.

[1]https://gal.nacional.sus.gov.br/

extrair e inferir conhecimento a partir dos dados para desempenhar vigilâncias epidemiológicas. Os dados exportados pelo GAL são fundamentais para identificar a circulação, distribuição e epidemiologia de diversos vírus no país.

Entretanto, essa exportação dos dados para análises e descoberta de padrões pode não ser totalmente confiável, e levar a sérios riscos de violação de privacidade, uma vez que pode exibir dados pessoais de pacientes. A área de saúde tem sido um alvo importante no que tange o acesso indevido à informação sensível [Dagher et al. 2018], uma vez que os registos de saúde contêm frequentemente informações privadas e sensíveis dos pacientes, *e.g.,* nomes, CPFs, endereço e particularmente o acometimento de doenças. Dessa forma, a anonimização dos dados na área de saúde se torna uma tarefa prioritária.

A anonimização de dados é uma técnica de preservação de privacidade que objetiva modificar valores dos atributos de um *dataset* com o objetivo de ocultar a identidade e/ou informações sensíveis de indivíduos. Todavia, essa não é uma tarefa simples, e apenas mascarar os elementos de identificação do indivíduo ou disponibilizar resultados agregados pode não proporcionar proteção suficiente. Por exemplo, consideremos um *dataset* contendo 100 exames de pacientes, dos quais apenas 2 são relativos à população indígena. Mesmo que os nomes, CPFs e endereços sejam mascarados ou omitidos, ainda pode ser possível identificar os indivíduos por outros atributos como a etnia. O desafio se encontra então em disponibilizar dados anonimizados para consulta que apresentem um nível de perturbação que não permita identificar os indivíduos nesse *dataset*, mas que ao mesmo tempo não torne os dados "inúteis". Assim, a escolha da técnica de anonimização é uma tarefa importante, uma vez que esta modificação pode acarretar em perda de informação, o que implica na diminuição da utilidade dos dados.

Diversas técnicas de privacidade e anonimização de dados já foram propostas na literatura [Warner 1965, Erlingsson et al. 2014, Dwork et al. 2006]. Uma das técnicas mais utilizadas é a Privacidade Diferencial [Dwork et al. 2006] (PD), que provê garantias de privacidade por meio do uso de um algoritmo aleatório, denominado mecanismo. Os mecanismos mais comuns existentes na literatura para garantir a PD são o Exponencial e o de Laplace. Além disso, outros mecanismos como a Resposta Randômica (*Randomized Response - RR*) [Warner 1965] podem ser aplicados. Cada um desses mecanismos possui vantagens e desvantagens, e pode ser mais adequado para um determinado *dataset*. Avaliar a utilidade de cada um deles para um determinado *dataset* se torna uma importante tarefa a ser desempenhada. O presente artigo tem como objetivo realizar um estudo comparativo de um conjunto de mecanismos de PD sobre um *dataset* extraído do GAL. Esse *dataset* contém os exames realizados com pacientes com casos suspeitos ou confirmados do Vírus da Zika (ZIKV) em diversos estados do Brasil (com nomes e CPFs sintéticos). Por meio de uma série de consultas analíticas, o estudo busca determinar o mecanismo mais adequado para ser aplicado para este tipo de *dataset*. Esse artigo se encontra organizado em 3 seções além da Introdução. Na Seção 2, o referencial teórico e os trabalhos relacionados são discutidos. A Seção 3 apresenta o estudo comparativo, e, finalmente a Seção 4 conclui o presente artigo.

2. Privacidade de Dados

A Privacidade Diferencial (PD), técnica de privacidade para anonimização de dados, é um modelo matemático que tem como objetivo permitir análises estatísticas sobre um conjunto de dados, sem comprometer a privacidade dos indivíduos. Ela assegura que qualquer resposta à uma determinada consulta, tem ocorrência igualmente possível e independe da presença, ou ausência, de um indivíduo no *dataset*. Dessa forma, a PD insere um mecanismo de aleatoriedade na adição do ruído à resposta de consultas realizadas sobre o *dataset*. Em especial, neste artigo iremos realizar uma análise sobre os mecanismos de Laplace, Exponencial e RR. A seguir,

discutimos com mais detalhes cada um desses mecanismos.

O mecanismo de Laplace [Dwork et al. 2006] é a forma mais comum de se obter PD para uma dada consulta. O mecanismo adiciona ruído randômico obtido por meio da distribuição de Laplace à resposta original da consulta. A distribuição de Laplace é centralizada em 0 e com parâmetro escala b tem a distribuição $Lap(z|b) = \frac{1}{2b}exp(-\frac{|z|}{b})$. Assim, dada uma consulta $f : \mathbb{N}^{|x|} \rightarrow \mathbb{R}^k$ o mecanismo é expressado como $M_L(x, f, \epsilon) = f(x) + (Y_1...Y_k)$, onde $Y_i \sim Lap(\frac{\Delta f}{\epsilon})$ i.i.d. (variáveis aleatórias independentes e identicamente distribuídas). Portanto, a quantidade de ruído adicionado depende da sensibilidade global Δf e do parâmetro de privacidade ϵ. A sensibilidade é a métrica utilizada para mensurar o impacto de um indivíduo, ao ser removido do *dataset*. Quanto maior for o valor da sensibilidade, maior será a quantidade de ruído adicionada. O mecanismo Exponencial foi projetado para responder perguntas categóricas (não-numéricas), escolhendo a "melhor" resposta. De forma a ser capaz de comparar classes categóricas, é necessário uniformizar as entradas do mecanismo. Assim, a uniformização é realizada por meio de uma função de utilidade, que deve ser capaz de prover uma ordem para as classes. A função de utilidade mapeia um registro a um determinado valor para todos os registros e a sensibilidade Δf é calculada por $\Delta f = \max_{x,y:||x-y||_1 \leq 1} |u(x) - u(y)|$. De posse de Δf, o mecanismo é expresso como $M_E(x, u) = f(x)+(Y_1...Y_k)$, onde $Y_i \sim exp(\frac{\epsilon u(x)}{2\Delta f})$. A Resposta Randômica (RR) é uma técnica originalmente aplicada para obter respostas privadas em entrevistas. O objetivo é garantir que os entrevistados respondam questões sensíveis, *e.g.,* sexualidade ou crença religiosa, mantendo a confidencialidade das respostas [Warner 1965]. Para isto, é adicionado um processo de aleatoriedade na resposta do entrevistado, mascarando se a resposta é verdadeira ou não. Por exemplo, é perguntado ao usuário se ele é a favor da legalização das drogas. Como processo de aleatoriedade, o entrevistado joga uma moeda, em segredo, e responde "sim" se der cara, ou responde a verdade se der coroa. O RAPPOR (*Randomized Aggregatable Privacy-Preserving Ordinal Response*), proposto por [Erlingsson et al. 2014], é um mecanismo para coleta de dados estatísticos com fortes garantias de privacidade que utiliza a técnica de RR. Para qualquer valor coletado, o RAPPOR entrega uma forte garantia de privacidade para o indivíduo, que limita a informação privada divulgada, medida pelo limiar ϵ-privacidade diferencial.

2.1. Trabalhos Relacionados

Alguns trabalhos na literatura executam estudos comparativos de técnicas de privacidade de dados. [Nascimento et al. 2018] comparam técnicas de anonimização e perturbação de dados em um *dataset* da área de saúde. Em especial, os autores focam nas abordagens *k-anonimato* e *l-diversidade*. Diferentemente da PD, tanto o *k-anonimato* quanto a *l-diversidade* são modelos sintáticos, *i.e.,* se usadas essas técnicas, elas podem permitir um usuário malicioso com conhecimento prévio ser capaz de identificar os indivíduos no *dataset*. [Kifer and Machanavajjhala 2011] analisam o mecanismo de Laplace para verificar se o mesmo pode ser aplicado para dados de redes sociais. O objetivo dos autores é utilizar o teorema *no-free-lunch* para afirmar que é impossível fornecer privacidade e utilidade sem realizar suposições sobre os dados. Entretanto, os autores não comparam outros mecanismos em sua avaliação.

3. Estudo Comparativo

Nessa seção são apresentadas as características do *dataset* utilizado no estudo comparativo, as consultas utilizadas como base para o estudo e as discussões acerca dos resultados obtidos. Para realização do estudo, foi exportado um *dataset* do GAL contendo 104 atributos e 25.133 registros relativos à exames realizados por pacientes suspeitos e confirmados do ZIKV. A qualidade

dos dados foi avaliada de acordo com o percentual dos registros preenchidos para cada atributo. Foram utilizados os parâmetros do SINAN para qualidade (ruim ($< 70\%$), regular ($70\% - 89\%$) e excelente ($\geq 90\%$)). Foi utilizado nesse estudo um *dataset* com completude regular. De forma a comparar os diferentes tipos de mecanismos de PD, fizemos uso de um conjunto de consultas, definidas por especialistas, que pode ser submetido ao *dataset* do GAL. Para cada uma das consultas, aplicamos cada um dos mecanismos sobre a resposta destas consultas e calculamos a medida de erro relativo ($ERel$) entre resposta original e a anonimizada. $ERel$ pode ser definido como $ERel = \frac{|x-x'|}{x}$, onde x representa o valor original e x' o valor com ruído. Desta forma, planeja-se analisar, quanto à utilidade, como cada mecanismo se comporta para as consultas sobre o *dataset*. As consultas utilizadas foram: (i) "Qual o percentual de grávidas infectadas com o ZIKV no Brasil (Q1)?"; (ii) "Qual a quantidade de casos suspeitos por Estado (Q2)?"; (iii) "Qual a quantidade de casos suspeitos por mês no ano de 2016 de um determinado estado (Q3)?".

A Figura 1(a) apresenta os resultados obtidos para a consulta Q1, que retorna apenas um valor numérico percentual. As barras azuis apresentam o valor da consulta antes da aplicação do mecanismo e as barras laranja os valores anonimizados. Pode-se perceber que apesar de os três mecanismos oferecerem resultados equivalentes em termos de $ERel$, o mecanismo que apresentou o menor $ERel$ foi o RR com $0,63\%$, seguido pelo mecanismo de Laplace com $2,25\%$, e, finalmente, do mecanismo exponencial com $10,79\%$. As Figuras 1(b), 1(e) e 1(h) apresentam os resultados das consultas Q2 para os mecanismos Exponencial, RR e Laplace, respectivamente. Nas barras azuis, são apresentados os resultados originais das consultas, enquanto que nas barras laranja, o resultado anonimizado pelo respectivo mecanismo. Podemos perceber que para todos os mecanismos a anonimização foi bem sucedida para os estados com as maiores quantidades de ocorrências. Entretanto, o $ERel$ não pode ser negligenciado no caso dos estados com poucas ocorrências no período. Dessa forma, a consulta Q3 foi avaliada para cada um dos mecanismos para os estados do Rio de Janeiro (RJ - maior quantidade de casos) e da Bahia (BA - menor quantidade de casos).

No caso da BA e do RJ para a consulta Q3 e uso do mecanismo exponencial (Figuras 1(c) e 1(d), respectivamente), observamos que o $ERel$ é bastante diferente para cada um dos estados. No caso do RJ, dada a grande quantidade de ocorrências, temos que $ERel < 2\%$. Entretanto, no caso da Bahia, o $ERel$ não pode ser negligenciado, onde $ERel_{BA} = 26,6\%$ em média, apesar de ter sido aplicada uma técnica que distribui o ruído ao longo dos meses considerados (o cenário de janeiro é o mais complicado, pois só foram considerados 4 casos). Dessa forma, a anonimização para o RJ foi capaz de manter a utilidade dos dados, enquanto que para a BA, pode-se perceber um maior $ERel$. O mesmo comportamento é encontrado para o RR (Figuras 1(f) e 1(g)), onde $ERel_{BA} = 59\%$, e para o mecanismo de Laplace (Figuras 2(a) e 2(b)), onde $ERel_{BA} = 61,1\%$. Apesar de os três mecanismos apresentarem $ERel$ não negligenciáveis para o estado da BA, existe uma diferença de qualidade entre os mesmos. Tomemos como exemplo o mês de junho de 2016. Nesse mês, temos 8 casos confirmados, porém os mecanismos exponencial, RR e Laplace retornam os valores 10, 4, 25, respectivamente, mostrando que mesmo com a utilidade reduzida, o mecanismo exponencial ainda foi o que apresentou o melhor resultado para essa consulta.

Em especial, no caso do mecanismo RR, o $ERel$ aumenta à medida que a quantidade de categorias também aumenta, como nas consultas Q2 e Q3. Isso acontece pois a quantidade de categorias tem um efeito inverso na probabilidade de se obter respostas verdadeiras. Logo, o mecanismo RR pode ser indicado para consultas com poucas categorias, como a Q1. Para contornar esta restrição do RR, é possível aplicar o mecanismo apenas sobre a resposta da consulta agregada (e não antes).

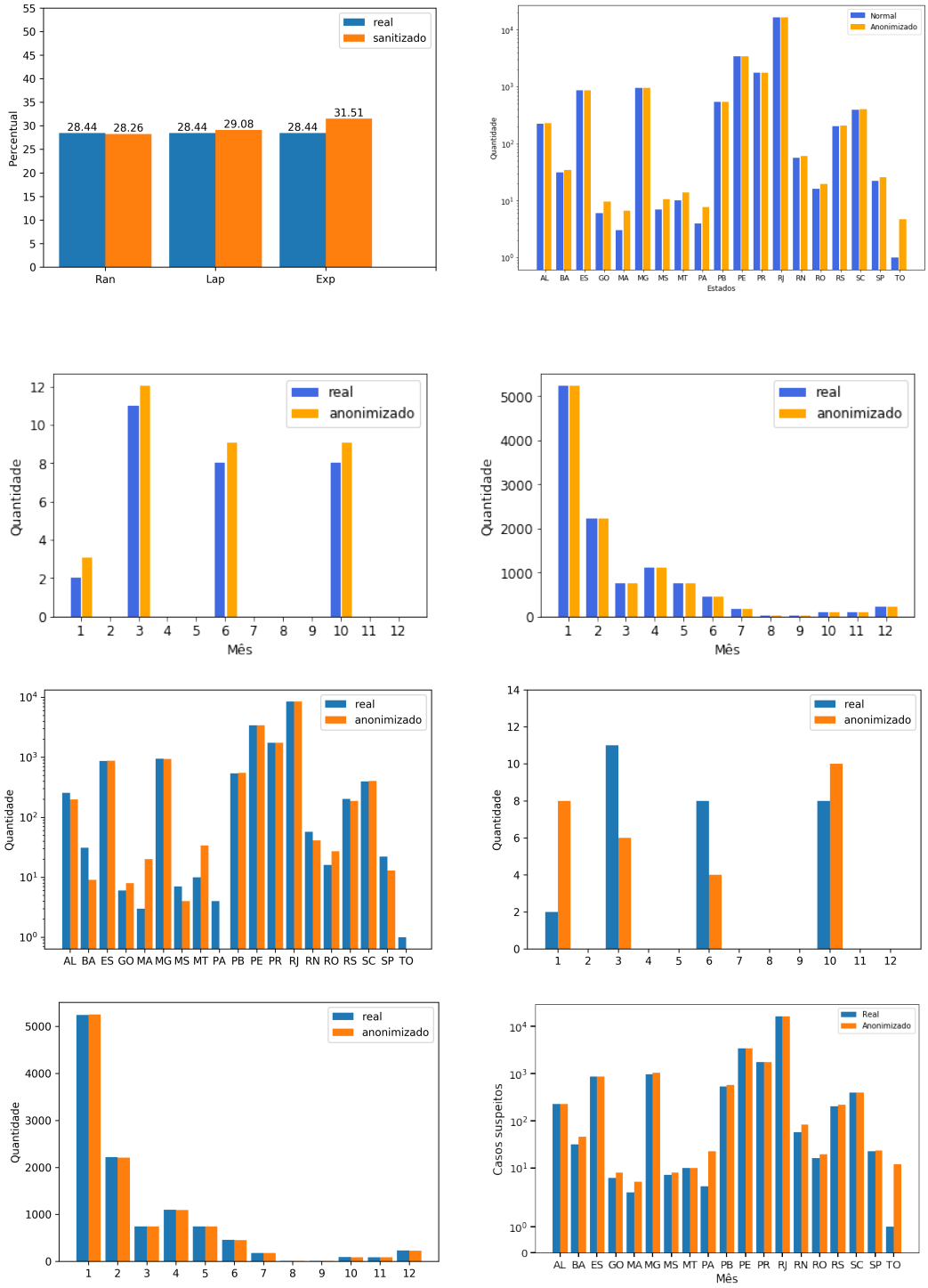

Figura 1. Repostas das consultas originais e anonimizadas (a) Q1 (b) Q2 - Mecanismo Exponencial (c) Q3/BA - Mecanismo Exponencial (d) Q3/RJ - Mecanismo Exponencial (e) Q2 - Randomized Response (f) Q3/BA - Randomized Response (g) Q3/RJ - Randomized Response (h) Q2 - Mecanismo Laplace

4. Conclusão

A comunidade da saúde precisa divulgar dados para a sociedade. Ao mesmo tempo, a privacidade dos pacientes terá que ser assegurada. Entretanto, mesmo quando utilizamos técnicas elaboradas como a Privacidade Diferencial, pode não ser simples definir qual o melhor mecanismo a ser usado para o *dataset*. O presente artigo comparou três mecanismos de Privacidade

Diferencial aplicados sobre um *dataset* de ocorrências do ZIKV no Brasil. A partir dos resultados obtidos, foi possível inferir qual dos mecanismos mantém um certo nível de utilidade dos dados anonimizados. Pode-se observar que o mecanismo de Respostas Randômicas funciona melhor para consultas que retornam poucas categorias, enquanto que o mecanismo Exponencial apresentou melhores resultados para consultas com muitas categorias. Como trabalhos futuros, pretendemos executar experimentos em *datasets* maiores de ZIKV, além de incorporar os mecanismos avaliados na camada de consulta de um *Data Lake* que vem sendo desenvolvido em conjunto com a Fundação Oswaldo Cruz.

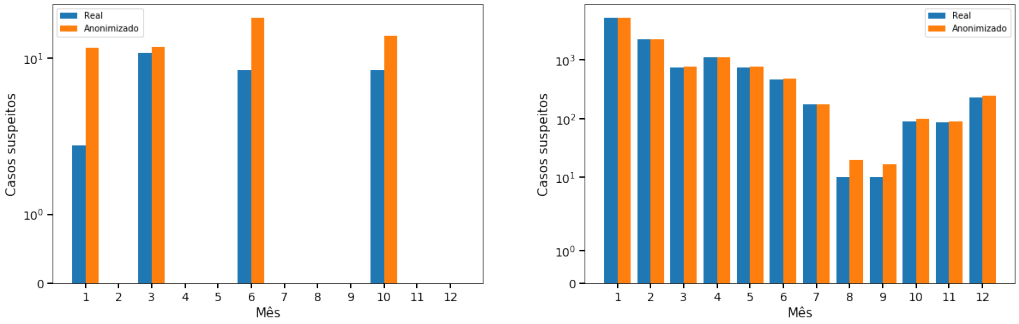

Figura 2. Repostas das consultas originais e anonimizadas (a) Q3/BA - Mecanismo Laplace (b) Q3/RJ - Mecanismo Laplace

Referências

[Dagher et al. 2018] Dagher, G. G., Mohler, J., Milojkovic, M., and Marella, P. B. (2018). Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. *Sustainable Cities and Society*, 39:283 – 297.

[Dwork et al. 2006] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T., editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Erlingsson et al. 2014] Erlingsson, Ú., Korolova, A., and Pihur, V. (2014). RAPPOR: randomized aggregatable privacy-preserving ordinal response. *CoRR*, abs/1407.6981.

[Kifer and Machanavajjhala 2011] Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *Proc. of the 2011 SIGMOD*, SIGMOD '11, pages 193–204, New York, NY, USA. ACM.

[Li et al. 2019] Li, S., Bamidis, P. D., Konstantinidis, S. T., Traver, V., Car, J., and Zary, N. (2019). Setting priorities for EU healthcare workforce IT skills competence improvement. *Health Inf. Journal*, 25(1).

[Nascimento et al. 2018] Nascimento, F., Vale, K. O., and Gorgônio, F. L. (2018). Um estudo comparativo entre algoritmos de proteção da privacidade aplicado à bases de dados na área de saúde. In *XXXIII SBBD, Rio de Janeiro, RJ, Brazil, August 25-26, 2018.*, pages 301–306.

[Shishvan et al. 2018] Shishvan, O. R., Zois, D., and Soyata, T. (2018). Machine intelligence in healthcare and medical cyber physical systems: A survey. *IEEE Access*, 6:46419–46494.

[Silva et al. 2019] Silva, A. B., Guedes, A., Síndico, S., Vieira, E., and de Andrade Filha, I. (2019). Registro eletrônico de saúde em hospital de alta complexidade: um relato sobre o processo de implementação na perspectiva da telessaúde. *Ciência & Saúde Coletiva*, 24:1133–1142.

[Warner 1965] Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.