EVALUATION OF FAIRNESS DEFINITIONS FOR HEALTH CARE SYSTEMS

BRUNA ROBERTA SEEWALD DA SILVA

EVALUATION OF FAIRNESS DEFINITIONS FOR HEALTH CARE SYSTEMS

Proposta de dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Flavio Vinicius Diniz de Figueiredo

Belo Horizonte

Junho de 2019

BRUNA ROBERTA SEEWALD DA SILVA

EVALUATION OF FAIRNESS DEFINITIONS FOR HEALTH CARE SYSTEMS

Dissertation proposal presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Flavio Vinicius Diniz de Figueiredo

Belo Horizonte

June 2019

List of Figures

A.1 Questions from study 2: [Saxena et al., 2018] 16

List of Tables

4.1 Timetable 11

B.1 Possible sensitive attributes available at MIMIC-III database 17

Contents

List of Figures **iv**

List of Tables **v**

1 Introduction **1**

2 Background and Related Work **4**

2.1 Machine Learning and Definitions 4

2.1.1 General view of supervised learning 4

2.1.2 Fairness definitions 5

2.2 Health care 7

2.2.1 Health and Fairness 7

2.2.2 MIMIC-III 7

3 Research Goals and Methodology **8**

4 Timetable **10**

Bibliography **12**

Attachment A Questions from the studies **15**

A.1 Study 1 15

A.2 Study 2 15

A.2.1 Examples of demographic questions 16

B MIMIC-III attributes **17**

Chapter 1

Introduction

With the increasing demand for algorithms and machine learning (ML) techniques as support tools for today's problems, there is also an increasing concern regarding the societal impact of such tools. In particular, the machine learning community is now focused on problems related to the fairness [Beutel et al., 2019, Grgic-Hlaca et al., 2016, Saxena et al., 2018] and accountability [Diakopoulos et al., 2017] of ML systems.

In these systems, fairness may be considered as the absence of algorithmic bias. That is, some definitions of fairness are concerned with features from individuals or groups and do not accurately represent the population (e.g., biases towards a race or gender). However, representativeness is not enough to capture fairness. Nowadays, data-driven decision making systems are being used in settings such as: employment [Miller, 2015], credit lending [Petrasic et al., 2017] and criminal justice [Barry-Jester et al., 2015]. The decisions made with the support of such systems impact people's lives and for this reason, finding suitable definitions of fairness in ML context is a active area of research.

Due to the relevance of this topic, multiple fairness definitions have been proposed in the computer science literature [Grgic-Hlaca et al., 2016, Saxena et al., 2018]. For instance, Narayanan [Narayanan, 2018] has recently presented a tutorial detailing 21 definitions of fairness. Some of such have mathematical and statistical notions, others, in contrast, stem from philosophical definitions. As a consequence, there is no clear agreement among researches over a particular definition of fairness [Gajane and Pechenizkiy, 2017]. The problem is even more complex when we consider that definitions are incompatible with each other [Kleinberg et al., 2016].

Even though automated algorithms are now responsible for some important decisions, they are not perfect. That is, errors can occur. Due to such errors, these algorithms should be accountable to the public. In this context, [Diakopoulos et al., 2017] states that *"accountability includes an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms"*. As an important example in this case, news articles about finance, weather or education are being produced by such algorithms. But there is a even more problematic situation, algorithms curation of contents are already responsible for influential news-dissemination. A famous case was the 2016 US elections, when Facebook was the main source of information about politics and the government. The news disseminated in the newsfeed affected users voting preferences, especially those undecided [Diakopoulos, 2016]. Even though these are not health applications, similar concerns naturally exists when algorithms mediate decisions form the health domain.

Up until now, most of previous work with regards to fairness and accountability have focused on datasets that explore criminal cases or credit loans. In such settings, fairness metrics related to representativeness are important in order to address social and historical inequities [Saxena et al., 2018]. However, when the focus changes to health care domain, representativeness by itself is harder to be justified. Consider a couple of examples. Initially, diseases like skin cancer will naturally have a higher incidence rate towards people with lighter skin. Another example is breast or prostate cancer, where some genders are naturally more affected than others. Given these characteristics, it is not yet clear how can researchers quantify fairness on health care. In this sense, when there is a causal relationship between a sensitive attribute (e.g., gender) and a disease, it is expected that lack of representativeness will not be an issue. However, when such causal notion is not present, then we have a representation problem.

It is important to understand this distinction, since different outcomes by protected class may not be unfair if they are motivated by biologic differences. Nonetheless, they would indeed be unfair if they result from suboptimal treatment decisions made in the past or the inability to follow-up treatment (e.g., distance, money, transportation) [Goodman et al., 2018].

Dealing with fairness and accountability is a complex matter. For instance, on one side, the accuracy of ML systems will depend on certain sensitive attributes. On the other, when these systems are used on decision making (for example, priority attention for elderly people), biased decisions reinforce social and historical inequalities which directly affect society.

Motivated by the above discussion, this dissertation will focus on the human side of fairness in health care systems. In particular we are concerned with scenarios where ML algorithms could assist health care professionals. Millions of inputs from doctors and patients are generated on a daily basis and could be applied for the training of

ML algorithms. However, historical data may have patterns that reinforce inequalities in access for treatments or medications. Therefore, using this data for training may perpetuate these disparities [Rajkomar et al., 2018].

To understand how humans perceive fairness, an exploratory data analysis of possible sensitive attributes in medical database will be conducted. Next, based on our findings from the first step, we shall employ surveys to understand how doctors, practitioners and patients of health care systems perceive issues of fairness and sensitive attributes. This second step will follow a similar methodology as [Grgic-Hlaca et al., 2016] and [Saxena et al., 2018]. Finally, this knowledge will be applied to enhance machine learning algorithms that take fairness into consideration.

The rest of this proposal is organized as follows: Chapter 2 presents a background about machine learning and related work; Chapter 3 addresses our research goals and methodology that will be applied to each goal; Chapter 4 contains the milestones from this dissertation and a timetable with them. Finally, Chapters 5 and 6 contain references and attachments, respectively.

Chapter 2

Background and Related Work

In this Chapter we shall discuss background and work related to this dissertation. In particular, we begin by presenting a general overview of Machine Learning in Section 2.1. Next, we shall detail previous efforts regarding fairness in the health domain in Section 2.2.

2.1 Machine Learning and Definitions

Machine Learning can be defined as a set of methods used to find patterns from datasets. Even though several approaches exist for ML, we shall here focus on the most commonly used statistical machine learning [Murphy, 2012]. More specifically, our discussion is focused on supervised learning. The problem in this type of learning is to correctly predict outputs based on new data or make decisions under uncertainty.

2.1.1 General view of supervised learning

As stated, this dissertation will solely focus on supervised learning. In supervised learning problems the goal is to learn how to map inputs $\mathbf{x}_i$ to outputs y_i. More formally, given a training-set $\mathcal{D}$ with N examples of input-output pairs:

$$\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^{N} \tag{2.1}$$

where $\mathbf{x}_i$ is a vector of features, for example, i. Each feature in the vector is indexed by j, and thus $\mathbf{x}_{ij} \in \mathbb{R}$ and $y_i \in \mathbb{R}$ value and each y_i was generated by an unknown function $y_i = f(x_i)$. Our goal is thus to discover a function h that approximates the true function f. This function h is called **hypothesis** [Russell and Norvig, 2016].

When we consider the whole set of candidate hypothesis, these may be viewed as functions conditioned on parameters $\mathbf{\Theta}$. The goal of supervised learning is thus to learn the parametric hypothesis, $\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$, that most accurately generalizes the dataset $\mathcal{D}$. Here, $\hat{y}_i$ is a prediction. Thus, to capture generality it is necessary to measure the accuracy of the model based on the chosen hypothesis (defined below). This is usually done with a test-set, which has distinct pairs from the training-set. If the outputs are a finite set of values, also called categorical features, this is then a classification problem. When the outputs are numbers, it is then a regression problem.

Examples of classification models that we may use include:

- Decision Trees
- Naive Bayes
- K Nearest Neighbors
- Linear Discriminant Analysis
- Support Vector Machines
- Random Forest

It is also important to define some terms used in these models. They will be necessary to understand some definitions presented in Section 2.1.2.1. They are:

- Outcomes: which are the categories to be predicted;
- Accuracy: which means the fraction of right predictions measured by the model. More formally:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2.2}$$

2.1.2 Fairness definitions

One recent focus of ML is on fairness. As discussed in Chapter 1, fairness may be considered the absence of algorithmic bias. However, since several systems are now data-driven, the decisions made by these systems may affect negatively the life of human beings, reinforcing social and historical inequalities in society. For this reason, finding suitable definitions of fairness is an important and active area of research.

Since 2011, the number of fairness definitions grew exponentially. To this date, more than 20 definitions exist and each one has details and differences that make them difficult to co-exist in the same situation. As we are trying to understand which sensitive features should be relevant or irrelevant to fairness in health care, we investigated several definitions of fairness. As a result, we decided to study the relevance of such characteristics reproducing the steps presented in [Grgic-Hlaca et al., 2016] and [Saxena et al., 2018].

2.1.2.1 Process fairness

In the paper presented by [Grgic-Hlaca et al., 2016], the authors introduce the notion of **process fairness**. In this definition, fairness is measured considering the features that are fair to use when making a decision. This process relies only on the user's judgments. In this context, [Grgic-Hlaca et al., 2016] argues that *"an individual opinion about a feature may change after they learn how using the feature might affect the decision outcomes. For instance, a user who initially considered a feature unfair for use in predicting recidivism risk might change their mind and deem the feature fair to use after learning that using the feature significantly improves the accuracy of prediction. Similarly, learning that using a feature might increase or decrease disparity in decision outcomes for different demographic groups (e.g., whites vs. blacks or men vs. women) might make a user change their opinion on the fairness of using that feature in decision making".*

Considering this situation, three measures of process fairness were presented:

- feature-apriori fairness: a feature s is considered fair to use without a *priori* knowledge of how its usage affects outcomes.
- feature-accuracy fairness: a feature s is considered fair to use *if its usage increases the accuracy of the classifier.*
- feature-disparity fairness: a feature s is considered fair to use *even* if its usage increases a measure of disparity (e.g., disparate treatment or disparate impact) of the classifier.

Disparate treatment and disparate impact, as cited in feature-disparity fairness, refer to discrimination due to a sensitive attribute and are reflected in the different decisions (harmful to one group) made by the classifier when these attributes are considered. The first one is intentional and the second is not [Zafar et al., 2017].

2.2 Health care

The adoption of ML in health care promises significant improvement in services [Char et al., 2018]. Simple routine tasks are already automated and algorithms will soon incorporate more complex decisions. Currently, the main concern still is to provide advanced tools that can help doctors to predict adverse outcomes in patients and diagnose more accurately symptoms and diseases.

2.2.1 Health and Fairness

Similar to the breakthroughs that happened in artificial intelligence in the last few years (e.g., DeepBlue, DeepMind, Siri, self-driving cars) a digital transformation in terms of health care is also happening. However, there are different kinds of concerns when considering the health domain.

For instance, there is a legal definition, called *protected groups*, which refers to populations that suffered from biased experiences in the past and yet remain vulnerable to harm *by incorrect predictions or withholding of resources* [Rajkomar et al., 2018]. For this reason, [Rajkomar et al., 2018], [Goodman et al., 2018] and [Char et al., 2018] reinforce the idea of ensuring fairness into the models that will be used in health care. The most important achievement in this case is to benefit all patients.

2.2.2 MIMIC-III

We now discuss MIMIC-III ('Medical Information Mart for Intensive Care'), our selected database with medical data. MIMIC is a critical care database with deidentified clinical data from patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts - US. This database was created as a effort to provide freely accessible information to the scientific community and others, due to the lack of data available for research. It was developed by the MIT Lab for Computational Physiology comprising anonymous health data associated with approximately 40,000 critical care patients. It includes demographics, vital signs, laboratory tests, medications, and more [Johnson et al., 2016].

Chapter 3

Research Goals and Methodology

The general objective for this master's dissertation is narrowed down into three specific goals that are driven by three research questions (RQs):

RQ1: A exploratory data study of medical attributes considered sensitive.

Current applications are forbidden to use sensitive attributes due to laws (e.g, Law number 13.709/18 - Lei Geral de Proteção de Dados Pessoais), under which these attributes are considered discriminatory. Yet, as we discussed earlier in this proposal, recent studies demonstrate that they can not be ignored since they could preserve medical errors or overpass bias on predictions, which can put at risk patients lives [Mullainathan and Obermeyer, 2017, Rajkomar et al., 2018].

For this exploratory phase, we will analyze the MIMIC-III database [Johnson et al., 2016] which was presented in Section 2.2.2. Examples of sensitive attributes from this database are presented in Attachment B. There is also the possibility to gain access to a database with eletrocardiograms (ECGs) from the Telehealth Network of Minas Gerais, Brazil [Alkmim et al., 2012].

Telehealth Network provides services and develop studies that support health care professionals in providing tele-assistance, also performing tele-electrocardiography and teleconsultations. Nowadays, it connects 780 municipalities in the state of Minas Gerais. With this database will be possible to obtain attributes such as age, education level, marital status and income, as well as features linked to the ECGs exams. In particular, with this dissertation, we intend to recognize correlations of sensitive attributes with specific diseases.

RQ2: Understand how doctors and users from the health care system comprehend fairness and sensitive attributes.

Include machine learning in health care can significantly improve services and provide higher assertive treatment for patients. However, it is necessary caution when inserting human bias on data that will be used for training. The goal here is to make a quantitative study with medical professionals and users from the health care system to understand their perception about fairness and sensitive attributes.

As a first analysis, all attributes contained in the base will be evaluated. Next, we will select a subset containing only sensitive attributes and those considered essential to the prediction of diseases. A survey will be elaborated with questions to evaluate fairness and whether the chosen subset has fair or unfair attributes to predict diseases. The survey will follow the model proposed by [Grgic-Hlaca et al., 2016] and [Saxena et al., 2018]. Examples of questions from surveys are presented in Attachment A.

RQ3: Evaluate different algorithms according to these sensitive attributes.

From the knowledge acquired with the survey about sensitive attributes in RQ1 and the quantitative study made on RQ2, the third research question will evaluate different machine learning algorithms applying the model proposed in [Grgic-Hlaca et al., 2016] and also analyze the trade-off between accuracy and fairness.

The quantitative study will compare notions of fairness for every attribute in all questions answered. Based on the fair ones, different classification models (e.g., Naive Bayes, Decision Tree) will be trained and the results evaluated, according to the accuracy of diagnosis and the selected fairness definition. As in previous works [Grgic-Hlaca et al., 2016], the results will be evaluated to comprehend if is possible to achieve and establish relevant process fairness, both for doctors and users. In this way, we intend to bring a human perception to the matter of machine learning usage as an auxiliary tool for medical diagnoses.

Chapter 4

Timetable

The milestones of this proposal are organized by month and indicated in Table 4.1. The activities will start in July 2019. The following activities are described as:

- MIMIC-III database analysis: As discussed in Chapter 3, this activity will be part of RQ1, which is a exploratory stage of this dissertation.
- Survey elaboration: As presented in Chapter 3, this activity will be part of RQ2 and is essential to understand which features are considered fair or unfair to use by health care professionals and patients.
- Submission to COEP-UFMG: As we are elaborating a survey that will be answered by human beings and will include personal information about them, it is necessary to submit this project to the UFMG Research Ethics Committee.
- Interview with health care professionals: With the approval of COEP-UFMG, the survey will be answered by health care professionals and patients from Hospital das Clínicas - UFMG.
- Quantitative analysis from surveys results and ML classifiers: Finally, as discussed in Chapter 3, this evaluation will be part of RQ3. Based on fair attributes, different models will be trained and we will analyze the results according to the accuracy of diagnosis and the selected fairness definition.

Table 4.1: Timetable in months

Milestones	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Literature review.	X	X	X	X	X	X	X	X	X	X	X	
MIMIC-III database analysis.	X	X	X									
Survey elaboration.				X	X							
Submission to COEP-UFMG.		X	X	X								
Interview with health care professionals.						X	X					
Quantitative analysis from surveys results and ML classifiers.							X	X	X			
Dissertation writing.					X	X	X	X	X	X	X	X
Journal publication submission.									X	X	X	X
Defence of master's dissertation.												X

Bibliography

[Alkmim et al., 2012] Alkmim, M. B., Figueira, R. M., Marcolino, M. S., Cardoso, C. S., Abreu, M. P. d., Cunha, L. R., Cunha, D. F. d., Antunes, A. P., Resende, A. G. d. A., Resende, E. S., et al. (2012). Improving patient access to specialized health care: the telehealth network of minas gerais, brazil. *Bulletin of the World Health Organization*, 90:373--378.

[Barry-Jester et al., 2015] Barry-Jester, A. M., Casselman, B., and Goldstein, D. (2015). The new science of sentencing. *The Marshall Project*, 4:2015.

[Beutel et al., 2019] Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., and Chi, E. H. (2019). Putting fairness principles into practice: Challenges, metrics, and improvements. *arXiv preprint arXiv:1901.04562*.

[Char et al., 2018] Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981.

[Diakopoulos, 2016] Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56--62.

[Diakopoulos et al., 2017] Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., et al. (2017). Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML*.

[Gajane and Pechenizkiy, 2017] Gajane, P. and Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

[Goodman et al., 2018] Goodman, S. N., Goel, S., and Cullen, M. R. (2018). Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*.

[Grgic-Hlaca et al., 2016] Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2.

[Johnson et al., 2016] Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

[Kleinberg et al., 2016] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

[Miller, 2015] Miller, C. C. (2015). Can an algorithm hire better than a human. *The New York Times*, 25.

[Mullainathan and Obermeyer, 2017] Mullainathan, S. and Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, 107(5):476--80.

[Murphy, 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[Narayanan, 2018] Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*.

[Petrasic et al., 2017] Petrasic, K., Saul, B., Greig, J., Bornfreund, M., and Lamberth, K. (2017). Algorithms and bias: What lenders need to know. *White & Case*.

[Rajkomar et al., 2018] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., and Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866--872.

[Russell and Norvig, 2016] Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[Saxena et al., 2018] Saxena, N., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D., and Liu, Y. (2018). How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. *arXiv preprint arXiv:1811.03654*.

[Zafar et al., 2017] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171--1180. International World Wide Web Conferences Steering Committee.

Attachment A

Questions from the studies

A.1 Study 1

The first study, presented in [Grgic-Hlaca et al., 2016] had the following questions. They were part of the survey presented to users when examining the ProPublica COMPAS dataset:

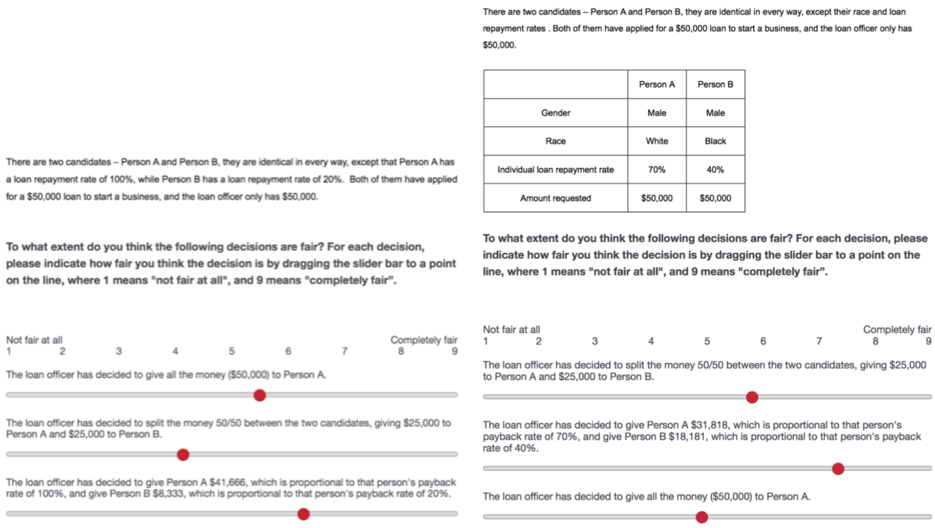
Q. 1: Do you believe it is fair or unfair to use information about this feature when estimating the offender's risk of recidivism?

Q. 2: Do you believe it is fair or unfair to use information about this feature when estimating the offender's risk of recidivism, if it makes the estimation more accurate?

Q. 3: Do you believe it is fair or unfair to use information about this feature when estimating the offender's risk of recidivism, if it makes black people more likely to be assessed as having a higher risk of recidivism than white people?

A.2 Study 2

The second one was presented in [Saxena et al., 2018]. It is composed of two questions and the users had to selected the level of fairness in each situation. There was a second survey with demographic questions, but answering was optional.



(a) Question 1 (b) Question 2

Figure A.1: Questions from study 2: [Saxena et al., 2018]

A.2.1 Examples of demographic questions

- 1. Do you identify as:
  - Male
  - Female
  - Other (please specify):
- 2. Do you identify as:
  - Spanish, Hispanic, or Latino
  - White
  - Black or African-American
  - American-Indian or Alaskan Native
  - Asian
  - Asian-American
  - Native Hawaiian or other Pacific Islander
  - Other (please specify):
- 3. What is your age ?

Attachment B

MIMIC-III attributes

This chapter describes some sensitive attributes available in MIMIC-III database. The INSURANCE, LANGUAGE, RELIGION, MARITAL STATUS, ETHNICITY columns occur in the ADMISSIONS table, LANGUAGE column occur in the PATIENTS table. They are described as follows:

- INSURANCE is the type of medical insurance of the patient.
- LANGUAGE is the primary language of the patient.
- RELIGION is the stated religion of the patient.
- MARITAL STATUS is the marital status of the patient.
- ETHNICITY is the stated ethnicity of the patient.
- GENDER is the genotypical sex of the patient.

Sensitive attributes	
INSURANCE	VARCHAR(255)
LANGUAGE	VARCHAR(10)
RELIGION	VARCHAR(50)
MARITAL STATUS	VARCHAR(50)
ETHNICITY	VARCHAR(200)
GENDER	VARCHAR(5)

Table B.1: Possible sensitive attributes available at MIMIC-III database