Legal Knowledge and Information Systems
J. Savelka et al. (Eds.)

doi:10.3233/FAIA241244

A Case-Based-Reasoning Analysis of the COMPAS Dataset

Wijnand VAN WOERKOM [a,1], Davide GROSSI [b,c,d], Henry PRAKKEN [a]
and Bart VERHEIJ [b]

[a] *Department of Information and Computing Sciences,*
Utrecht University, The Netherlands
[b] *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence,*
University of Groningen, The Netherlands
[c] *Amsterdam Center for Law and Economics,*
University of Amsterdam, The Netherlands
[d] *Institute for Logic, Language and Computation,*
University of Amsterdam, The Netherlands

Abstract. In this paper we build on a formal model of reasoning with dimensions to analyze data from the COMPAS program—a widely used and studied tool for predicting recidivism. We extend the underlying theory of the model by introducing a notion of consistency and apply it to assess whether COMPAS follows this principle in its risk assessments and supervision level recommendations. Our analysis yields three key findings. First, the program's risk score assignments appear highly inconsistent, but we argue this is due to important input features missing from the dataset. Second, the program's recommended supervision levels do exhibit a high degree of consistency. Third, we uncover errors in the dataset related to the conversion of raw scores to decile scores. These findings cast doubts on previous studies conducted on the COMPAS dataset, and demonstrate the need for evaluation studies like ours.

Keywords. explainable artificial intelligence, case-based reasoning, COMPAS dataset, consistency analysis

1. Introduction

One of the concerns of the AI & law literature has been the development of formal models of *a fortiori* case-based reasoning. Examples of early contributions to this strand of research have been the development of the HYPO and CATO programs [1,2], the formal model of Prakken and Sartor in [3], and Horty's work on precedential constraint in [4,5]. More recently, [6] extended Horty's work in order to support the development of explainable AI methods, by adapting his model to handle cases with outcomes that are not binary but instead vary along dimensions. This paper continues the line of research from [6]. Specifically, we apply this extended model to analyze the COMPAS dataset [7].

The COMPAS program (Correctional Offender Management Profiling for Alternative Sanctions), developed by Northpointe (now Equivant) in 1998, is a risk and need

[1] Corresponding author; e-mail: w.k.vanwoerkom@uu.nl.

assessment tool designed to predict recidivism. It assigns *risk scores* based on various factors like a defendant's age and criminal history. The news outlet ProPublica published a dataset in [7] with information on COMPAS risk scores produced for defendants in Broward County, Florida, between 2013 and 2014. This dataset has colloquially become known as the COMPAS dataset. Based on an analysis of this data, ProPublica alleged that COMPAS racially discriminates in its decision-making. While the consensus in the literature is that ProPublica's analysis was flawed [8], the dataset they published remains in use, and the broader discussion on bias in recidivism risk prediction and the opacity of data-driven systems continues.

In [9], we applied Horty's theory of precedential constraint to analyze aspects of the COMPAS dataset. However, we could not use Horty's theory to measure the COMPAS program itself, as COMPAS produces outputs that are not binary but take values in dimensions. (Instead, the analysis focussed on the binary labels indicating whether the defendants in the dataset did or did not recidivate.) In [6], we adapted Horty's model to accommodate decisions taking values in dimensions, and remarked as a point of future work that this modification enabled an analysis of the COMPAS program itself.

Our aim in this work is to analyze the COMPAS program using the model we presented in [6]. We believe that developing explainable AI methods for decision-making systems, particularly in the legal domain, requires testing against real-world datasets, such as the COMPAS dataset. It is particularly suitable for analysis with this model due to the dataset's continued relevance, and because it represents a typical example of systems that require explainable AI methods: data-driven systems making decisions with ethical, social, or legal consequences.

To start our analysis, we expand the model of [6] with a notion of consistency, which is a straightforward adaptation of Horty's concept of consistency in [4,5]. Using this notion, we then measure the consistency of various aspects of the COMPAS dataset by means of a threefold analysis. First, we consider the degree to which the risk scores produced by COMPAS adhere to the a fortiori constraint principle of our model, based on the features gathered by ProPublica. We find that this is not the case, and conclude this is because the features gathered by ProPublica are only a subset of the features used by the COMPAS program. We exemplify why this can cause the impression that the program does not obey the a fortiori principle. Secondly, we check if the program adheres to the a fortiori principle in determining its recommended supervision level, as a function of its risk scores. We find that in this case the program does adhere to the a fortiori principle. This second analysis brings to light an error in ProPublica's dataset—namely, a flaw in the conversion of the so-called *raw* scores to *decile* scores. This error is the topic of our third analysis. We show that the values corresponding to this conversion in the dataset are inconsistent, according to both our proposed formalized notion of consistency and the description of the scores given by the COMPAS developers in [10]. We give a concrete example of this, and discuss some of its possible causes. We consider it a success of our model and of our approach that it has brought this issue to light, which seems to have gone unnoticed despite the COMPAS dataset facing much scrutiny over the years.

In Section 2 we describe the model of [6], including our new notion of consistency. In Section 3 we give a threefold analysis of the COMPAS data through the lens of our model. Finally, in Section 4, we summarize our findings and suggest directions for future work.

2. A Model of a Fortiori Reasoning with Dimensions

In this section we describe the formal model we will use for the analyses to come in Section 3. In Section 2.1, we recall the basic ideas and definitions of [6], and in Section 2.2 we add a notion of consistency, analogous to the definition used by Horty in [4,5].

2.1. Knowledge Representation: Dimension Hierarchies, Fact Situations, and Case Bases

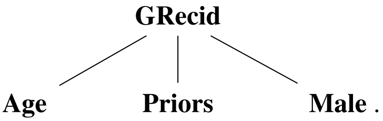
Definition 1. A *dimension* $(d, \preceq)$ is a partially ordered set. A *dimension hierarchy* (D, H) is a finite set of dimensions D with a relation H on D, such that the transitive closure of H is irreflexive. A dimension is called *base-level* if it is H-minimal, and *abstract* otherwise; the sets of base-level and abstract dimensions are respectively denoted by B and A.

As a notational convenience we will write $\mathsf{H}(d)$ for the pre-image of d under the hierarchical structure H; so more specifically, $\mathsf{H}(d) = \{e \in D \mid \mathsf{H}(e,d)\}$. In the context of a decision about a dimension d based on its immediate subordinates in $\mathsf{H}(d)$, we may refer to the dimension d as the *target* dimension, and those in $\mathsf{H}(d)$ as the *input* dimensions.

Definition 2. A *fact situation* X for a dimension hierarchy (D, H) is a choice function on a subset $\mathrm{dom}(X) \subseteq D$; i.e. $X : \mathrm{dom}(X) \to \bigcup D$ is a function satisfying $X(d) \in d$ for all $d \in \mathrm{dom}(X)$. We say X is *complete* if $\mathrm{dom}(X) = D$. The set of all fact situations is denoted by $\mathcal{F}$, or by $\mathcal{F}(D)$ if we want to stress with respect to which dimensions the situation is defined. A *case base* $\mathcal{C}$ is a finite subset $\mathcal{C} \subseteq \mathcal{F}$ of fact situations.

A dimension is a set of values d together with an order $\preceq$ indicating defeasible support for its overlying dimensions in the hierarchy. More precisely, given dimensions d, e in a dimension hierarchy (D, H) such that $\mathsf{H}(d,e)$, then if X and Y are fact situations such that $X(d) \preceq Y(d)$ we generally expect this to imply $X(e) \preceq Y(e)$.

Example 3. To illustrate Definitions 1 and 2, we give an example of a dimension hierarchy based on the recidivism risk domain of the COMPAS system:



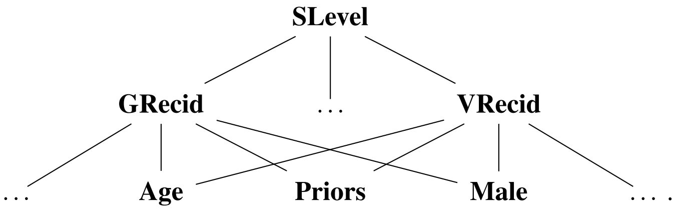
(2.1)

Above we have drawn a hierarchical structure between dimensions influencing a general recidivism risk score: age at assessment, number of prior offenses, and whether the individual is male or not. Formally we represent these as the followings sets and orders:

$$\mathbf{GRecid} = (\{1, 2, \ldots, 10\}, \leq), \qquad \mathbf{Age} = (\{18, 19, 20, \ldots\}, \geq),$$
$$\mathbf{Priors} = (\{0, 1, 2, \ldots\}, \leq), \qquad \mathbf{Male} = (\{0, 1\}, \leq).$$

The associated orders indicate the influence of these dimensions on recidivism risk, which are well established in the literature on recidivism [11]. For example, the **Age** dimension is ordered by the greater-than relation $\geq$, so that the link $\mathsf{H}(\mathbf{Age}, \mathbf{GRecid})$ indicates that the younger the defendant is the higher the recidivism risk score should be. For example, consider two individuals described by fact situations X and Y; then, if $X(\mathbf{Age}) \preceq Y(\mathbf{Age})$, meaning $X(\mathbf{Age}) \geq Y(\mathbf{Age})$, this should defeasibly imply that $X(\mathbf{GRecid}) \leq Y(\mathbf{GRecid})$.

Example 4. The hierarchy depicted in (2.1) contains just one abstract dimension, but an idea dating back to the work by Aleven on the CATO program in [2] is that abstract factors can themselves be subordinate to higher-level dimensions. An example of this in the COMPAS program is its *recommended supervision level* summary statistic—a score ranging from 1–4, intended to assist in determining the appropriate intensity of supervision. It is stated in [10] that "the violence and recidivism risk potential factors are [its] main drivers." Assuming both of these are based on the base-level dimensions of (2.1), possibly in addition to other dimensions, we obtain a hierarchy with the following general shape:



(2.2)

Formally we can represent the **VRecid** dimension—corresponding to a violent recidivism risk assessment—by $(\{1,2,\ldots,10\},\leq)$, and the **SLevel** dimension by $(\{1,2,3,4\},\leq)$.

2.2. *Case Base Constraint and Consistency*

The primary notion in this model—and in those like it [5,12,13,4]—is that of the constraint induced by precedent cases. As mentioned, the support relation between dimensions is defeasible in that values in other dimensions can have a mitigating effect. For instance, in Example 3 we could have fact situations X and Y such that both $X(\mathbf{Age}) \preceq Y(\mathbf{Age})$, suggesting $X(\mathbf{GRecid}) \preceq Y(\mathbf{GRecid})$; and $Y(\mathbf{Priors}) \preceq X(\mathbf{Priors})$, suggesting $Y(\mathbf{GRecid}) \preceq X(\mathbf{GRecid})$. It is up to the decision-maker to weigh these different values of the fact situation against each other when coming to a decision. The idea of constraint is that, if the relation holds for *all* underlying dimensions, then the implication should no longer be defeasible. In other words, if $X(e) \preceq Y(e)$ holds for all $e \in \mathsf{H}(d)$, this should imply $X(d) \preceq Y(d)$. In this scenario the value of X on d is upper bounded by that of Y. This works both ways—if $Y(e) \preceq X(e)$ for all $e \in \mathsf{H}(d)$, then $Y(d) \preceq X(d)$ should hold as well, meaning X is lower bounded by Y on d.

Definition 5. Given a case base $\mathcal{C}$ and a value $v \in d$, a fact situation X is *lower bounded* in d by $\mathcal{C}$ to v, denoted by $\mathcal{C} \vDash v \preceq X(d)$, if and only if either

(1) v is the least element of d, or
(2) $v \preceq X(d)$, or
(3) $d \in A$ and there is $Y \in \mathcal{C}$ satisfying $v \preceq Y(d)$ such that $\mathcal{C} \vDash Y(e) \preceq X(e)$ holds for all $e \in \mathsf{H}(d) \cap \mathrm{dom}(Y)$.

The *upper bound* in d by $\mathcal{C}$ to v, denoted $\mathcal{C} \vDash X(d) \preceq v$, is defined analogously.

Remark 6. Definition 5 differs slightly compared to that in [6, Definition 4]. Firstly, in the present work we do not assume that the links in the hierarchy have a polarity. We do this because the examples we discuss do not require link polarity. Furthermore, the constraint induced by a hierarchy with link polarity can always be reproduced by a hierarchy without

Table 1. Three example fact situations X, Y, and Z for the general recidivism risk domain depicted in (2.1).

	Age	**Priors**	**Male**	**GRecid**
X	30	1	0	5
Y	20	4	1	8
Z	25	2	0	?

link polarity. This is done by adding a copy of each dimension, corresponding to its inverse order: for any $d \in D$ with order $\preceq$, add a dimension d' with order $\succeq$, and replace negative links from d by positive links from d', so that the resulting hierarchy contains only positive links. A second difference stems from the addition of the points in Definition 5 regarding least and greatest elements; the absence of these points in [6] was just an oversight.

Having defined constraint we can define a notion of inconsistency: the degree to which the constraint induced by a case base contradicts itself. This definition is new with respect to [6], and is a straightforward adaptation of the definition used by Horty in [4,5].

Definition 7. Let $\mathcal{C}$ be a case base for a dimension hierarchy (D, H), $d \in D$ a dimension, and X a fact situation; X is *d-inconsistent* with respect to $\mathcal{C}$ if there are values $v \prec w \in d$ such that both $\mathcal{C} \vDash X(d) \preceq v$ and $\mathcal{C} \vDash w \preceq X(d)$; otherwise X is *d-consistent*.

Example 8. We illustrate this notion of consistency through an example related to the dimension hierarchy depicted in (2.1), taken from [6, Table 1], listed here in Table 1. Using Definition 5 we can calculate that $\{X, Y\} \vDash 5 \preceq Z(\mathbf{GRecid})$, and $\{X, Y\} \vDash Z(\mathbf{GRecid}) \preceq 8$. This means that Z should get a recidivism risk score between 5–8, and a value outside that range would qualify as inconsistent according to Definition 7. The reader is referred to [6, Section 5.3] for an example utilizing the recursive clause of Definition 5.

3. Data Analyses

In this section we analyze ProPublica's COMPAS dataset through the lens of our model of a fortiori constraint. We do so by interpreting the features of the data as dimensions, the rows of the dataset as cases, and the dataset itself as a case base. The statistic we are primarily interested in is case base consistency (Definition 7): the relative frequency of cases in the case base that are consistent with respect to it (along some dimension d).

Our findings are threefold. Firstly, we show that the risk scores outputted by the COMPAS program do not seem to adhere to the a fortiori principle with respect to the information gathered by ProPublica. We conclude that this is because this data is only a subset of the data used by COMPAS to arrive at its assessments, as is also pointed out in [14]. Secondly, we find that the recommended supervision level score—which the program outputs as a final summarizing recommendation to the user of the program—does satisfy our definition of a fortiori constraint to a high degree. Thirdly, we show that the raw scores outputted by COMPAS were inconsistently converted to decile scores.

We begin by giving more details on the COMPAS dataset and its features in Section 3.1, and then proceed to describe our three analyses in Sections 3.2, 3.3, and 3.4, respectively.[2]

[2] A complete overview of our results, as well as the code and data used to generate them, can be found at https://github.com/wijnanduu/AF-CBR-COMPAS.

3.1. Dimensions in the COMPAS Dataset

The COMPAS program outputs three primary risk scores: risk of failure to appear, risk of general recidivism, and risk of violent recidivism. Data on these scores produced by COMPAS in Broward County, Florida, between 2013 and 2014, were acquired by ProPublica through a public records request [15]. Other than the risk scores, the COMPAS program also profiles defendants on aspects such as criminal history, drug use, peer and family situations, work/education, and so forth. These profiles drive the risk predictions, need assessment, and treatment plans that COMPAS recommends [10, Section 5.1]. Unfortunately, these data were not provided to ProPublica in response to their public records request. To compensate for this, ProPublica separately built a profile of the criminal history of each person of which they received a COMPAS score, matching records based on name and date of birth. In total the dataset contains information on the personal details, criminal history, and associated COMPAS scores of some 11,000 people. An overview of the features of the dataset that we use in our analyses is given in Table 2. The listed orders were manually determined by us, but are in agreement with the literature on the subject, and seem to be in agreement with the way COMPAS uses them [11,10].

The COMPAS program outputs scores in what its developers call a *raw* format. This is a continuous scale, of which the unit does not appear to have any particular meaning, other than that higher values indicate a higher risk. For example, the program may output a recidivism risk score of 0.11, and what this quantity exactly represents is unclear; for instance, it is not a percentage, as the score can also be negative. The COMPAS program outputs these scores based on a regression model that was fit to historical data [16].

In order to aid with the interpretation of the raw scores, they are converted to *decile scores*. This is done by ranking the raw scores of a normative group (i.e., a representative subset of the population) in ascending order, and then dividing them into ten equal-sized groups. For example, if a defendant's decile recidivism risk score is 7, then this means their raw score is higher than that of the lowest scoring 60% of the normative group, and lower than that of the highest scoring 30%. A raw risk score of 0.11 is much harder to interpret than its corresponding decile score of 7—hence the use of this conversion. ProPublica only use the decile scores for their analysis in [7].

The decile scores are then further simplified into what we will call *textual* scores, which is a low/medium/high indication. The conversion from decile scores to textual scores is simple: 1–4 is low, 5–7 is medium, and 8–10 is high [10, Table 2.1].

3.2. First Analysis: COMPAS Risk Scales

For our first analysis we compute the consistency percentages of the three primary COMPAS risk scores (failure to appear, general recidivism, and violent recidivism) based on the input dimensions listed in Table 2: age at assessment, age at first arrest, sex, number of priors, and partnership status. For each of the target dimensions we analyze the three different versions of the score: the raw, decile, and textual versions. To see the effects of our choice of dimension orders we also computed the scores for three different methods of determining the dimension orders: our manual choice listed in Table 2, a linear regression analysis, and a Pearson correlation analysis; see [18, Section 3.1.1] for an explanation of the automated methods. This results in a total of $3^3 = 27$ percentages. Each of the datasets contains the same 11,671 rows (but uses a different subset of the columns). An overview of the results can be found in Table 3.

Table 2. An overview of the dimensions present in the COMPAS dataset. An ascending order means that higher values indicate more support for overlying dimensions; and vice versa for a descending order.

Dimension	**Description**	**Order**
Age at assessment	Age of the defendant at the time of the COMPAS assessment.	Descending
Age at first arrest	Age of the defendant when they were first arrested. We manually compute this value from the `casearrest` table in the `compas.db` file provided by ProPublica. This feature was not used in [7].	Descending
Sex	A binary 0/1 value indicating whether the defendant is male.	Ascending
Number of priors	Number of offenses committed prior to the one that led to the COMPAS assessment, split by juvenile and adult priors in the dataset.	Ascending
Partnership status	A binary 0/1 value indicating whether the person was in a partnership at the time of the assessment. This feature was not used in [7].	Descending
Failure to Appear	Prediction of risk of failure to appear, "based largely on prior history [and] current charges for failure to appear, prior recidivism on community placement, general criminal involvement, and unstable residential ties and transience." [17]	Ascending
General recidivism	Score for predicting new offenses subsequent to the assessment, based on "the Criminal Involvement Scale, drug problems sub-scale, age at assessment, age at first adjudication, number of prior arrests, arrest rate, and the Vocational Educational Scale." [14]	Ascending
Violent recidivism	Score for predicting new violent offenses subsequent to the assessment, based on "age at assessment, age at first adjudication, the History of Violence Scale, the History of Noncompliance Scale, and the Vocational Educational Scale." [14]	Ascending
Supervision level	A score on a scale of 1–4 indicating the recommended intensity of the defendant's supervision, meaning low, medium, medium with override consideration, and high, respectively. [10, Section 5.1]	Ascending

Table 3. Consistency scores for the three main compas risk scores: failure to appear, general recidivism, and violent recidivism. These are computed based on the input dimensions that were gathered by ProPublica in [7].

Target	**Manual**			**Linear regression**			**Pearson correlation**		
	Raw	**Decile**	**Text**	**Raw**	**Decile**	**Text**	**Raw**	**Decile**	**Text**
FTA	0%	1%	7%	1%	1%	14%	0%	1%	7%
GRecid	0%	1%	3%	0%	2%	4%	0%	1%	3%
VRecid	0%	2%	7%	0%	0%	3%	0%	2%	7%

The main takeaway of the results in Table 3 is that the consistency percentages are very low. This is somewhat contrary to expectations, because the developers of COMPAS have given indication that the risk scales are largely based on regression models, which should theoretically operate according to an a fortiori principle. For example, it is stated in [10] that the **VRecid** score is computed as the following weighted sum:

$$\begin{aligned}\mathbf{VRecid} = {} & (-w_1 \cdot \text{age}) + (-w_2 \cdot \text{age at first arrest}) + (w_3 \cdot \text{history of violence}) \\ & + (w_4 \cdot \text{vocation education}) + (w_5 \cdot \text{history of noncompliance}).\end{aligned} \tag{3.1}$$

Table 4. Consistency scores for the recommended supervision level COMPAS output, based on the failure to appear, general recidivism, and violent recidivism risk scores. The consistency percentage is given for each of the three representations of these input dimensions: the raw scores, the decile scores, and the textual representation.

Target	**Manual**			**Linear regression**			**Pearson correlation**		
	Raw	**Decile**	**Text**	**Raw**	**Decile**	**Text**	**Raw**	**Decile**	**Text**
SLevel	84%	100%	72%	82%	22%	0%	84%	100%	72%

The last three summands of Eq. (3.1) correspond to scales that COMPAS uses but which are not present in ProPublica's dataset. This leads us to conclude that the most likely cause of the low consistency percentages of our analysis is due to missing dimensions in the data. Consider, for example, the fact situations X and Z in Table 1, and suppose that $Z(\mathbf{GRecid}) = 4$. Then Z would be **GRecid**-inconsistent according to Definition 7. However, the recidivism risk scores assignment might in reality be based on an additional drug-problem dimension **Drugs**—a 1–10 score ordered by $\leq$—which is missing from our hierarchy. Suppose, furthermore, that $X(\mathbf{Drugs}) = 8$ and $Z(\mathbf{Drugs}) = 4$; then Z appears inconsistent in the version of the hierarchy that does not include the drug-problem dimension, but is in fact consistent with respect to the hierarchy that does include it. Indeed, the COMPAS general recidivism risk score does rely on such a dimension according to [14], and this dimension is not present in ProPublica's dataset, so it is highly likely that our analysis is suffering from the effect illustrated by this example.

3.3. Second Analysis: Recommended Supervision Level

For our second analysis, we look at the consistency percentage of the recommended supervision level (**SLevel**) scores in the dataset. This score is an overall recommendation that the COMPAS program outputs, based on its various needs and risk scales. As mentioned in Example 4, this recommendation is primarily based on the general recidivism and violent recidivism risk scores, and so—unlike in the first analysis—we can now be (more or less) certain that we are using the same dimensions as the ones used by the COMPAS program. Each of the datasets we examined contained approximately 12,000 rows. The dimension orders are again determined either manually (as listed in Table 2), or automatically (by either the linear regression or Pearson correlation method).

The results of our analysis can be found in Table 4. We see that the scores are significantly higher than before—some 84% of decisions were consistent according to the manually determined dimension orders, when we use the raw risk scores as inputs. In fact, the consistency percentage goes up to 100% when we use the decile scores instead. Upon reflection, this is a very strange result: the conversion of raw scores to decile scores should be monotonic, and so should not be able to decrease the number of precedents that constrain a given fact situation. Therefore, we would expect the consistency percentage to remain the same or decrease, but never to increase, as we switch from raw to decile scores.

An examination of this change reveals that the cause of this heightened consistency percentage is an error in ProPublica's dataset regarding the conversion from raw scores to decile scores. To see this, consider the two fact situations X and Y listed in Table 5, taken from the dataset. If we look at the raw scores, we have the constraint $\{Y\} \vDash X(\mathbf{SLevel}) \preceq 3$ because X assigns lower values to all three of the raw COMPAS risk scores. This means

Table 5. Two examples of fact situations X,Y in the COMPAS dataset that are **SLevel**-inconsistent with respect to their raw scores, but **SLevel**-consistent with respect to their decile scores, due to a faulty conversion. The input dimensions are as in Table 4: the risk scores of failure to appear, general recidivism, and violent recidivism.

	FTA		**GRecid**		**VRecid**		**SLevel**
	Raw	**Decile**	**Raw**	**Decile**	**Raw**	**Decile**	
Y	21	3	0.14	7	−0.95	9	3
X	19	3	0.11	8	−1.21	8	4

Table 6. The first row gives the consistency of the conversion in the dataset compared to the specification of the conversion in [10, Table 2.3]; the second row gives the consistency percentage as a case base in our model.

Measure	**FTA**	**GRecid**	**VRecid**
Cut-points in [10, Table 2.3]	2%	6%	28%
A fortiori constraint	47%	23%	100%

that X is **SLevel**-inconsistent because $X(\mathbf{SLevel}) = 4 \succ 3$. However, when we look at the decile scores rather than the raw scores we get a different picture, because according to those we have $X(\mathbf{GRecid}) = 8 \not\preceq 7 = Y(\mathbf{GRecid})$; so Y no longer constrains X, and X has become **SLevel**-consistent. Of course, this should not happen: when a general recidivism risk score of 0.11 is higher than 70–80% of the normative group, then 0.14 should also be higher than 70–80% of the normative group.

3.4. Third Analysis: Raw-to-Decile Score Conversion

The fact that the consistency of the dataset in Table 4 (using the manual orders) jumps from 84% to 100% means that there are quite a few examples such as the one in Table 5. Further quantifying the extent of this problem is the goal of our third analysis.

The cut-off points that were supposedly used to convert the raw scores to decile scores are listed in [10, Table 2.3]. We have used these cut-points to manually convert the raw scores in the COMPAS dataset to decile points and compared the relative frequency of the resulting scores that match the decile scores given in the dataset; the results of this comparison are listed in the first row of Table 6. These clearly show that the cut-points given in [10, Table 2.3] were not the ones used to convert the raw scores to decile scores, or that the raw scores listed do not match the ones used to arrive at the decile scores.

By definition, the conversion of raw scores to decile scores should be monotonic: if one person's raw score is higher than another's, then their decile score should also be higher than the other's. We can measure the degree to which this holds simply by applying our a fortiori model, by considering the raw version of the dimensions to be influencing its decile version—the higher the raw score, the higher the decile score—and performing a case base consistency analysis. We have performed this analysis and listed its results in the second row of Table 6. We can see that in particular the general recidivism score was not converted in a monotonic way, while the violent recidivism score was perfectly converted—just not according to the decile cut-points listed in [10, Table 2.3]. We can deduce what cut points were used simply by looking at the minimum and maximum raw scores that have led to any given decile score; the resulting points are listed in Table 7.

Without looking deeper into this we can only speculate about the cause of these discrepancies. It may be that an error was made in the conversion from raw scores to decile

Table 7. The decile cut-points used to convert the raw violent recidivism risk scores in the COMPAS dataset to decile scores. Note that these values do not match those listed in [10, Table 2.3].

	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
VRecid	−1.39	−0.92	−0.60	−0.39	−0.19	−0.01	0.19	0.39	0.67	2.36

scores, or that it was done correctly but on the basis of multiple scales which changed throughout the years. It could also be that COMPAS uses different norm groups for the conversion, such as different groups for men and women, or different groups for older or younger individuals. (Though, this seems unlikely; the individuals in Table 5 are both male and of a somewhat similar age—23 and 31, respectively.) It could also be that the conversion was done correctly, but that the COMPAS scores were incorrectly entered into the dataset, or accidentally swapped between people. All we can say is that the violent scores seem to be correctly converted, but on the basis of cut-points that do not match the practitioners guide [10], and that there is no single set of cut-points that can explain the conversion from raw to decile scores found in this dataset.

4. Conclusion

In this work, we used the model of [6] to study the COMPAS dataset through the lens of a fortiori case-based reasoning. We drew three main conclusions on the basis of this analysis. Firstly, we found that the program does not seem to adhere to the a fortiori principle when computing its risk scores, but we argue that this is due to important input features missing from the data. Secondly, we saw that with respect to its recommended supervision level assessment, the program does adhere to the a fortiori principle. Thirdly, we found an error in the COMPAS dataset with respect to the conversion of raw-to-decile score conversions.

We consider it a success of our model that it has brought this conversion issue to light because to our knowledge it has been overlooked thus far in the extensive literature surrounding the COMPAS dataset. For example, Equivant's defense in [14] against the accusations by ProPublica does not mention it, and neither does the related critical work of [19]. Many studies performed based on the COMPAS dataset, such as [20], focus on the decile scores and often do not even mention the raw scores, simply stating that "COMPAS scores, ranging from 1 to 10, classify the risk of recidivism as low-risk (1 to 4), medium-risk (5 to 7), or high-risk (8 to 10)."

An evident source of future work is to further investigate the cause of this flaw. As mentioned in Section 3.4, many possible causes are imaginable: the conversion of raw scores to decile scores was not correctly executed, or it was correctly executed but using multiple different scales, or it was correctly executed but different raw scores were used than the ones listed, and so forth.

With regard to our model, we would like to further test it by applying it to other datasets. An example that was used throughout [6] was that of bail data, and a bail dataset was made available in [21]. It would be interesting to see the results of a similar analysis performed in this work on the bail dataset of [21]. Secondly, a more extensive explainable AI method could be developed on the basis of our a fortiori model. Ideally this would include an argumentative structure that could underlie the generated explanations, in the style of [22].

Acknowledgements

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022.

References

[1] Rissland EL, Ashley KD. A Case-Based System for Trade Secrets Law. In: Proceedings of the First International Conference on Artificial Intelligence and Law. ICAIL '87. ACM Press; 1987. p. 60-6.

[2] Aleven V, Ashley KD. Evaluating a Learning Environment for Case-Based Argumentation Skills. In: Proceedings of the Sixth International Conference on Artificial Intelligence and Law. ICAIL '97. ACM; 1997. p. 170-9.

[3] Prakken H, Sartor G. Modelling Reasoning with Precedents in a Formal Dialogue Game. Artificial Intelligence and Law. 1998 Jun;6(2):231-87.

[4] Horty J. Rules and Reasons in the Theory of Precedent. Legal Theory. 2011 Mar;17(1):1-33.

[5] Horty J. Reasoning with Dimensions and Magnitudes. Artificial Intelligence and Law. 2019 Sep;27(3):309-45.

[6] van Woerkom W, Grossi D, Prakken H, Verheij B. Hierarchical *a Fortiori* Reasoning with Dimensions. In: Sileno G, Spanakis J, van Dijck G, editors. Legal Knowledge and Information Systems. JURIX 2023: The Thirty-sixth Annual Conference. IOS Press; 2023. p. 43-52.

[7] Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias [Newspaper article]. ProPublica. 2016.

[8] Lagioia F, Rovatti R, Sartor G. Algorithmic Fairness through Group Parities? The Case of COMPAS-SAPMOC. AI & Society. 2023 Apr;38(2):459-78.

[9] van Woerkom W, Grossi D, Prakken H, Verheij B. Landmarks in Case-Based Reasoning: From Theory to Data. In: HHAI2022: Augmenting Human Intellect. vol. 354 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2022. p. 212-24.

[10] Equivant. Practitioner's Guide to COMPAS Core; 2019.

[11] Yukhnenko D, Blackwood N, Fazel S. Risk Factors for Recidivism in Individuals Receiving Community Sentences: A Systematic Review and Meta-Analysis. CNS spectrums. 2020 Apr;25(2):252-63.

[12] van Woerkom W, Grossi D, Prakken H, Verheij B. Hierarchical Precedential Constraint. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. ICAIL '23. ACM Press; 2023. p. 333-42.

[13] Canavotto I, Horty J. Reasoning with Hierarchies of Open-Textured Predicates. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. ICAIL '23. ACM Press; 2023. p. 52-61.

[14] Dieterich W, Mendoza C, Brennan T. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe Inc. Research Department; 2016.

[15] Larson J, Mattu S, Kirchner L, Angwin J. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica; 2016.

[16] Brennan T, Dieterich W, Ehret B. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. Criminal Justice and Behavior. 2009 Jan;36(1):21-40.

[17] Northpointe. Measurement and Treatment Implications of COMPAS Core Scales [Manual]. Northpointe Institute for Public Management, Inc. A; 2009.

[18] van Woerkom W, Grossi D, Prakken H, Verheij B. *A Fortiori* Case-Based Reasoning: From Theory to Data. Journal of Artificial Intelligence Research. 2024 Oct;81:401-41.

[19] Flores AW, Bechtel K, Lowenkamp CT. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.". The Administrative Office of the US Courts. 2016;80(2):38-46.

[20] Dressel J, Farid H. The Accuracy, Fairness, and Limits of Predicting Recidivism. Science Advances. 2018 Jan;4(1):eaao5580.

[21] William Dieterich, Christina Mendoza, Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe Inc. Research Department; 2016.

[22] Prakken H, Ratsma R. A Top-Level Model of Case-Based Argumentation for Explanation: Formalisation and Experiments. Argument & Computation. 2022 Jan;13(2):159-94.