Assessing Data Quality –
A Probability-based Metric for Semantic Consistency

Authors:

Heinrich, Bernd, Department of Management Information Systems, University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany, Bernd.Heinrich@ur.de
Klier, Mathias, Institute of Technology and Process Management, University of Ulm, Helmholtzstr. 22, 89081 Ulm, Germany, Mathias.Klier@uni-ulm.de
Schiller, Alexander, Department of Management Information Systems, University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany, Alexander.Schiller@ur.de
Wagner, Gerit, Department of Management Information Systems, University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany, Gerit.Wagner@ur.de

Citation: Bernd Heinrich, Mathias Klier, Alexander Schiller, Gerit Wagner, Assessing Data Quality –
A Probability-based Metric for Semantic Consistency, Decision Support Systems,Volume 110, June 2018, Pages 95-106,
ISSN 0167-9236, DOI 10.1016/j.dss.2018.03.011
(https://www.sciencedirect.com/science/article/pii/S0167923618300599)

Assessing Data Quality –
A Probability-based Metric for Semantic Consistency

Abstract:

We present a probability-based metric for semantic consistency using a set of uncertain rules. As opposed to existing metrics for semantic consistency, our metric allows to consider rules that are expected to be fulfilled with specific probabilities. The resulting metric values represent the probability that the assessed dataset is free of internal contradictions with regard to the uncertain rules and thus have a clear interpretation. The theoretical basis for determining the metric values are statistical tests and the concept of the p-value, allowing the interpretation of the metric value as a probability. We demonstrate the practical applicability and effectiveness of the metric in a real-world setting by analyzing a customer dataset of an insurance company. Here, the metric was applied to identify semantic consistency problems in the data and to support decision-making, for instance, when offering individual products to customers.

Keywords: data quality, data quality assessment, data quality metric, data consistency

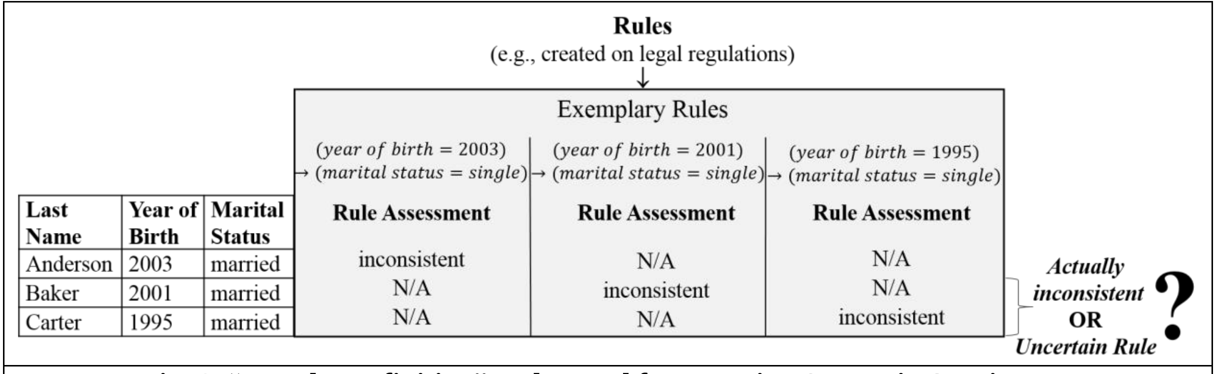
1 Introduction

Making use of large amounts of internal and external data becomes increasingly important for companies to gain competitive advantage and enable data-driven decisions in businesses (Ngai, Gunasekaran, Wamba, Akter, & Dubey, 2017). However, data quality problems still impede companies to generate the best value from data (Moges, van Vlasselaer, Lemahieu, & Baesens, 2016; Witchalls, 2014). Overall, poor data quality amounts to an average financial impact of $9.7 million per year and organization as reported by recent Gartner research (Moore, 2017). In particular, 63% of the respondents of a survey by Moges, Dejaeger, Lemahieu, and Baesens (2011, p. 639) indicated that “inconsistency (value and format) and diversity of data sources are main recurring challenges of data quality”.

Data quality can be defined as the “agreement between the data views presented by an information system and that same data in the real world” (Orr, 1998, p. 67). In this regard, data quality is a multidimensional construct comprising different dimensions such as accuracy, consistency, completeness, and currency (Batini & Scannapieco, 2016; Redman, 1996; Zak & Even, 2017). In the following, we focus on consistency, in particular semantic consistency, as one of the most important dimensions (Blake & Mangiameli, 2011; Shankaranarayanan, Iyer, & Stoddard, 2012; Wand & Wang, 1996). We define semantic consistency as the degree to which assessed data is free of internal contradictions (cf. also Batini & Scannapieco, 2016; Heinrich, Kaiser, & Klier, 2007; Redman, 1996).

Contradictions are usually determined based on a set of rules (Batini & Scannapieco, 2006; Heinrich et al., 2007; Mezzanzanica, Cesarini, Mercorio, & Boselli, 2012). Thereby, a rule represents a proposition consisting of two logical statements, where the first statement (antecedent) implies the second (consequent). For instance, in a database containing master data about customers in Western Europe, such a rule may be $year\ of\ birth = 2003 \rightarrow marital\ status = single$. Stored customer data regarding a married customer born in 2003 would contradict this rule, indicating a consistency problem.

Existing data quality metrics for semantic consistency are based on rules which are considered as "true by definition" (cf. Section 2). This means that the rules have to be true for all of the assessed data and any violation indicates inconsistent data. Examples for such rules are provided in Figure 1, which also shows some selected records of a customer database serving as a basis for our discussion:



Rules
(e.g., created on legal regulations)

Exemplary Rules

Last Name	Year of Birth	Marital Status	$(year\ of\ birth = 2003) \rightarrow (marital\ status = single)$ Rule Assessment	$(year\ of\ birth = 2001) \rightarrow (marital\ status = single)$ Rule Assessment	$(year\ of\ birth = 1995) \rightarrow (marital\ status = single)$ Rule Assessment
Anderson	2003	married	inconsistent	N/A	N/A
Baker	2001	married	N/A	inconsistent	N/A
Carter	1995	married	N/A	N/A	inconsistent

Fig. 1. "True by Definition" Rules used for assessing Semantic Consistency

Due to the fact that in Western Europe marriage is only legally allowed for people of age 16 and older, for an assessment in the year 2018, a value for *year of birth* of 2003 (antecedent of the first rule in Figure 1) implies the value *single* for *marital status* (consequent of the first rule), which is a typical example for a "true by definition" rule. In this case, the value *married* for *marital status* of the first record is assessed as inconsistent. However, for the assessment of semantic consistency it can be necessary to also consider rules such as $year\ of\ birth = 2001\ \rightarrow marital\ status = single$ (second rule in Figure 1). Here, one has to distinguish: On the one hand, violations of such a rule – assessed in 2018 – may indeed indicate an erroneous value which could have resulted from a random or systematic data error (cf. Alkharboush & Yuefeng Li, 2010; Fisher, Lauria, & Matheus, 2009). On the other hand, violations may stem from the fact that the rule is not "true by definition", but only fulfilled with a specific probability. For example, some customers may indeed have married at the age of 16. Therefore, a violation of this rule does not necessarily imply that such data is inconsistent and of low quality. This also holds for other years of birth (e.g., $year\ of\ birth = 1995$; third rule in Figure 1) or, in general, for other antecedents and consequents or applications where rules cannot be considered as "true by definition". Hence, we are confronted with rules with uncertain consequent, to which we refer as uncertain rules in the following. To the best of our knowledge, none of the existing approaches aiming to measure semantic consistency has considered such relevant uncertain rules yet.

Thus, to (1) consider uncertain rules in a well-founded way and (2) ensure a clear interpretation of the resulting metric values, we propose a data quality metric for semantic consistency based on probability theory. To address uncertain rules, the metric delivers an indication rather than a statement under certainty regarding the degree to which assessed data is free of internal contradictions. We argue that the well-founded methods of probability theory are adequate and valuable to deal with uncertain rules. More precisely, the theoretical basis for determining the metric values are statistical tests and the concept of the p-value, allowing the clear interpretation of the metric values as probabilities.

The remainder of the paper is structured as follows. In the next section, we discuss related work and the research gap. Then, we present a probability-based metric for semantic consistency and outline possible ways to instantiate this metric. In the fourth section, we illustrate the case of an insurance company to demonstrate the practical applicability and effectiveness of the metric. Finally, we briefly summarize the findings and conclude with a discussion of limitations and directions for further research.

2 Related Work

The data quality dimension consistency is seen "as a multi-faceted dimension" (Blake & Mangiameli, 2009, p. 3) which can be defined in terms of representational consistency, integrity, and semantic consistency (Blake & Mangiameli, 2009). Since these three aspects stem from different domains, they overlap in some cases. Representational consistency requires that data are "presented in the same format and are compatible with previous data" (Blake & Mangiameli, 2009; Wang & Strong, 1996). Integrity is often defined as entity, referential, domain, column, and user-defined integrity (Blake & Mangiameli, 2009; Lee, Pipino, Strong, & Wang, 2004). Entity integrity requires that data values considered as primary keys are unique and different from NULL. Referential integrity states that, given two relations, if an attribute is a primary key in one of them and is contained as a foreign key in the other one, the non-NULL data values from the second relation must be contained in the first one (Lee et al., 2004). Domain and column integrity require data values to be part of a predefined domain (e.g., $income \in \mathbb{R}^+$) and user-defined integrity requires the satisfaction of a set of general rules. Finally, semantic consistency refers to the absence of contradictions between different *data values* based on a rule set (Blake & Mangiameli, 2009; Heinrich et al., 2007; Lee, Pipino, Funk, & Wang, 2006; Liu & Chi, 2002; Mecella et al., 2002; Mezzanzanica et al., 2012; Redman, 1996; Scannapieco, Missier, & Batini, 2005). Generally, semantic consistency is equivalent to user-defined integrity.

In this paper, we focus on semantic consistency due to two reasons. First, assuring semantic consistency is crucial for decision support, as decision-making is typically based on data values. Second, both representational consistency and integrity have already been extensively studied in literature (Blake & Mangiameli, 2009, 2011). Semantic consistency, however, is a field of research which gains more and more importance in the course of growing data volumes and their thorough analysis.

Underlining this importance, literature discusses several data quality problems and root causes which lead to inconsistencies with respect to data values (Kim, Choi, Hong, Kim, & Lee, 2003; Laranjeiro, Soydemir, & Bernardino, 2015; Oliveira, Rodrigues, & Henriques, 2005; Rahm & Do, 2000; Singh & Singh, 2010). These root causes are typically categorized in two ways. First, referring to the steps in the data management process (i.e., data entry/capturing, data transformation, data aggregation, data processing, etc.). And second, whether inconsistencies are caused by a single source or by multiple sources. Given this, a common and highly relevant root cause for inconsistencies are error-prone operative data entries via one single source (cf. Rahm & Do, 2000; Singh & Singh, 2010). This may be, for example, a call center employee, the person himself referred to in the considered record (e.g., a customer entering master data via a web application) or a damaged data capturing device (e.g., a malfunctioning sensor). In all these scenarios, inconsistencies regarding, for instance, two data values of a customer record may arise. In the case of a call center employee or the customer himself, it is possible that only one of the two data values is correctly entered or changed. The second data value, however, may be entered or changed erroneously (or not at all). For instance, the value for $year\ of\ birth$ may be correctly entered as 2003, the value for $marital\ status$, however, may be erroneously entered as $married$. Similarly, parts of a customer's address may be entered incorrectly, leading to an inconsistency. A second prevalent root cause concerns the steps data aggregation and integration in the data management process with respect to multiple sources (e.g., different databases; cf. Rahm & Do, 2000; Singh & Singh, 2010). Here, contradictory data values of, for instance, customer records may arise in scenarios in which the same customers are stored in multiple databases of departments and units of a company (e.g., after a merger). Contradictions may result from the integration of attributes or their values, for example when databases are integrated for a coordinated and comprehensive customer management. For instance, in one database, the $marital\ status$ of a customer may be stored as $single$, but in a second database, the value for $name\ of\ spouse$ of the same customer may not be equal to NULL, indicating that the customer is $married$. Faulty business rules used for data transformation and leading to contradicting data values (cf. Singh & Singh, 2010) constitute another important scenario and root cause among many others, stressing the relevance of semantic consistency.

In the following, for reasons of simplicity, we will use the term consistency instead of semantic consistency. To provide an overview of existing works on metrics for consistency, we concentrate on metrics that are (i) formally defined (e.g., by a closed-form mathematical function) and (ii) result in a numerical metric value representing the consistency of the data values to be assessed. In that sense, we do not consider approaches that aim to identify potentially (in)consistent data values without providing numerical metric values for (in)consistency (e.g., Bronselaer, Nielandt, Mol, & Tré, 2016; Fan, Geerts, Tang, & Yu, 2013; Mezzanzanica et al., 2012). Table 1 presents existing metrics for consistency satisfying (i) and (ii). They follow the idea that consistency of data values can be determined based on the number of fulfilled rules, with a higher number of fulfilled rules implying higher consistency.

We discuss these metrics with regard to (1) the way they assess consistency and (2) the interpretation of the resulting metric values. Related to (1) the first three rows of Table 1 with the light grey background contain metrics that assign weights to the fulfillment and violation of rules. The next two rows with the white background provide metrics assessing consistency purely as “true” or “false” regarding the fulfillment and violation of rules. The last two rows with the dark grey background contain metrics relying on conditional functional dependencies (CFDs; Bohannon, Fan, Geerts, Jia, & Kementsietsidis, 2007; Cong, Fan, Geerts, Jia, & Ma, 2007).

		Source	Metric
“True by definition” rules	Assign weights to the fulfillment/ violation of rules	Alpar and Winkelsträter (2014); Hipp et al. (2001); Hipp et al. (2007)	t: record; N: number of relevant rules for t; L: number of irrelevant rules for t; w_n^-, w_n^+, w_l^0: weights; $r_n(t) = \begin{cases} 0, \text{if } t \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$; $h_n(t) = \begin{cases} 0, \text{if rule } r_n \text{ is relevant for } t \\ 1 \text{ else} \end{cases}$; $cons(t) = \sum_{n=1}^{N}(w_n^- r_n(t) + w_n^+(1 - r_n(t)))(1 - h_n(t)) + \sum_{l=1}^{L} h_l(t)\, w_l^0$
		Hinrichs (2002)	g: data value; N: number of relevant rules for g; w_n: weights; $r_n(g) = \begin{cases} 0, \text{if } g \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$; $cons(g) = \frac{1}{\sum_{n=1}^{N} w_n r_n(g) + 1}$
		Kübart et al. (2005)	t: record; N: number of relevant rules for t; $w_n^- \geq 0$: weights; $r_n(t) = \begin{cases} 0, \text{if } t \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$; $incons(t) = \sum_{n=1}^{N} w_n^- r_n(t)$
	Assess only fulfillment/ violation of rules	Cordts (2008); Pipino et al. (2002)	g: data value; N: number of relevant rules for g; $r_n(g) = \begin{cases} 0, \text{if } g \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$; $cons(g) = 1 - \frac{\sum_{n=1}^{N} r_n(g)}{N}$
		Heinrich et al. (2007); Heinrich and Klier (2015a)	g: data value; N: number of relevant rules for g; $r_n(g) = \begin{cases} 0, \text{if } g \text{ fulfills rule } r_n \\ 1 \text{ else} \end{cases}$; $cons(g) = \prod_{n=1}^{N}(1 - r_n(g))$
	Use CFDs	Abboura et al. (2016)	a: attribute; N: number of relevant CFDs for a $cons(a) = \prod_{n=1}^{N} support(r_n) \cdot conf(r_n)$
		Wang et al. (2016)	D_B: database; S: number of tuples in D_B; $C_{min}(D_B)$: minimum set of tuples in D_B such that $D_B \backslash C_{min}(D_B)$ fulfills all CFDs $incons(D_B) = \frac{\lvert C_{min}(D_B) \rvert}{S}$

Table 1. Existing Metrics for Consistency

All metrics in the first three rows of Table 1 (Alpar & Winkelsträter, 2014; Hinrichs, 2002; Hipp et al., 2001; Hipp et al., 2007; Kübart et al., 2005) assign weights to the fulfillment and violation of rules. The considered rules correspond to association rules. For a given set of records, association rules are implications of the form $X \rightarrow Y$ that satisfy specified constraints regarding minimum support and minimum confidence (cf. Agrawal, Imieliński, & Swami, 1993; Srikant & Agrawal, 1996). Thereby, rule support $supp(X \rightarrow Y)$ is defined as the fraction of records that fulfill both antecedent X and consequent Y of the rule; rule confidence $conf(X \rightarrow Y)$ denotes the fraction of records fulfilling the antecedent X that also fulfill the consequent Y (cf. Agrawal et al., 1993). An example of an association rule is $year\ of\ birth = 1995\ \rightarrow marital\ status = single$. If 80% of the records in the database with $year\ of\ birth = 1995$ also fulfill $marital\ status = single$ and 5% of the records fulfill both $year\ of\ birth = 1995$ and $marital\ status = single$, it follows that the support of the rule is 5% while its confidence is 80%. To treat the violation of distinct association rules r_n as differently severe when assessing consistency, for example Alpar and Winkelsträter (2014) and Hipp et al. (2007) use the rule confidence $conf(\ r_n)$ to determine respective weights. In particular, the idea of these authors is to assign a weight of $conf(\ r_n)$ to the fulfillment of a rule and a weight of $-conf(\ r_n)$ to its violation. In order to determine consistency concerning several rules and a set of data values, the weights are calibrated and summed up.

While these approaches treat the violation of distinct rules as differently severe (i.e., depending on rule confidence), they assess the fulfillment of a rule to always be an indicator for high consistency by assigning a positive weight (i.e., $conf(\ r_n)$) to rule fulfillments and vice versa. As only association rules above a chosen minimum threshold for confidence based on the dataset to be assessed are determined, rules below this threshold remain unconsidered in these approaches. However, rules with a lower confidence are also highly relevant for assessing consistency, as they can be an important indicator for inconsistent data. For example, a rule (confidence) stating that 30% of 17-year-olds are stored as being married would certainly help to identify inconsistencies because a much smaller percentage of 17-year-olds is actually married in Western Europe. In addition, solely using the rule confidence based on the assessed data can lead to misleading results if a large part of the data to be assessed is erroneous: For instance, if 90% of all 17-year-olds are erroneously stored as being married in a database, a corresponding association rule and its rule confidence is determined (given a minimum rule confidence of e.g. 80%). On this basis, however, the 10% of 17-year-olds which are accurately stored as *not* being married would be considered as inconsistent. More generally, these approaches assess all rules with high confidence as “true by definition” and penalize violations against them as inconsistent.

To conclude, these metrics provide first, promising steps concerning the treatment of violations of distinct rules as differently severe. However, rules with low confidence are ignored and rules with high confidence are seen as "true by definition". Further, the resulting values of these metrics suffer from a lack of clear interpretation (cf. (2)). Indeed, it remains unclear what a particular metric value actually means, obstructing its use for decision support. This is due to the summation of the (calibrated) weights (representing the rule confidences as "measures of consistency"). To illustrate this, we again consider the example of a customer database. A customer record may fulfill some association rules (e.g., the values for *zip code* and *city*) and violate others (e.g., the values for *marital status* and *year of birth)*. The respective calibrated weights are summed up, but the result of the summation is a real number with no clear interpretation (e.g., in terms of a probability whether the considered record is consistent). Furthermore, the metric values are, in general, not interval-scaled and do not have a defined minimum and maximum. This may seriously hinder their usefulness for decision support: For example, in a second assessment of the customer data at a later point in time, the mined association rules and their confidence can differ from the first assessment. Then, a higher (or lower) metric value of the same, unchanged record in the second assessment does not necessarily represent higher (or lower) actual consistency. In fact, the consistency of the record may still be the same.

The metrics in the next two rows of Table 1 with the white background (Cordts, 2008; Heinrich et al., 2007; Heinrich & Klier, 2015a; Pipino et al., 2002) assess the consistency of data values only by "true" or "false" statements regarding the fulfillment and violation of the considered rules. On this basis, they provide a clear interpretation of the metric values in terms of the percentage of data values consistent with respect to the considered rules (cf. (2)). These approaches, however, treat all rules equally as "true by definition" and thus have similar limitations as the metrics discussed above.

Finally, the metrics provided in the last two rows of Table 1 with the dark grey background (Abboura et al., 2016; Wang et al., 2016) assess consistency by using CFDs. A CFD is a pair $(X \rightarrow Y, T_i)$ consisting of a functional dependency $X \rightarrow Y$ (an implication of sets of attributes) and a certain tableau T_i (with $i \in \{1, 2, \dots, N\}$) which specifies values for the attributes in X and Y (cf. Bohannon et al., 2007 for details). To give an example, stating that records with $year\ of\ birth = 1995$ also fulfill $marital\ status = single$ can be represented by the following CFD: $(year\ of\ birth \rightarrow$

$marital\,status, T_1)$, with T_1 containing a row which includes 1995 as value for $year\ of\ birth$ and $single$ as value for $marital\,status$. A probabilistic CFD is a pair consisting of a CFD and its confidence, where support and confidence of a probabilistic CFD are defined analogously to association rules (Golab, Karloff, Korn, Srivastava, & Yu, 2008). Abboura et al. (2016) define the consistency of an attribute to be the product of the support of a (probabilistic) CFD multiplied by its confidence. The product is taken over all CFDs relevant for the considered attribute. Thus, analogous to the approaches in the first three lines of Table 1, the approach assesses the considered CFDs as "true by definition" and penalizes violations against them as inconsistent. This results in similar problems as outlined above. Additionally, the metric values do not provide a clear interpretation (cf. (2)). Wang et al. (2016) propose to determine a minimum subset of tuples in a database which – if corrected – would lead to the database fulfilling all CFDs. Then, the inconsistency of the database is measured by the ratio of the size of this minimum subset in relation to the size of the whole database.

Overall, existing metrics interpret their rules used for assessing consistency as "true by definition" resulting in several limitations. In particular, they do not deal with uncertain rules. Moreover, metrics which treat the violation of distinct rules as differently severe do not ensure a clear interpretation of the metric values. In the next section we address this research gap.

3 Probability-based Metric for Consistency

In this section, we present our metric for semantic consistency. First, we outline the general setting and the basic idea. Then, we describe methodological foundations which serve as a basis when defining the metric in the following subsection. Finally, we outline possible ways to instantiate the metric.

3.1 General Setting and Basic Idea

We consider the common relational database model and a database D_B to be assessed. A relation consists of a set of attributes $\{a_1, a_2, \dots, a_m\}$ and a set of records $T = \{t_1, t_2, \dots, t_n\}$. The data value of record t_j regarding attribute a_i is denoted by $\phi(t_j, a_i)$. In line with existing literature (cf. Section 2), we use a rule set R to assess consistency. Rules are propositions of the form $r: A \rightarrow C$, where A (antecedent) and C (consequent) are logical statements addressing either single attributes in D_B or relations between them.

As opposed to existing approaches, we do not treat rules as “true by definition”. Rather, we aim to consider uncertain rules that are expected to be fulfilled with specific probabilities.

This allows to determine metric values which represent the probability that the assessed dataset is free of internal contradictions with regard to these uncertain rules. More precisely, for a data value $\phi(t_j, a_i)$ in D_B and an uncertain rule r, we interpret consistency as the probability that $\phi(t_j, a_i)$ is free of contradictions with regard to r. A metric that results in a probability guarantees that the metric takes values in $[0;\ 1]$ and the metric values have a clear interpretation.

The following running example from our application context (cf. Section 4) illustrates the idea of our metric: An insurer strives to conduct a product campaign targeting only married customers younger than 20 years. If the data stored in the customer database is of low quality, wrong decisions and economic losses may result. For instance, if a customer younger than 20 years is erroneously stored as *married* in the database, contacting him with a product offer will generate costs and may lead to lower customer satisfaction. In case the insurer aims to assess the consistency of its customer database before conducting the campaign, existing metrics for consistency would consider the rule $r_1\colon year\ of\ birth > 1998 \rightarrow marital\ status = single$. This rule is selected because it is fulfilled by most people that are younger than 20 years (e.g., 95%), which goes along with a high rule confidence. However, such metrics would assess data regarding a married customer who is younger than 20 years as inconsistent. Thus, the determined metric values could not provide any support within the campaign.

Our metric, in contrast, additionally considers the rule $r_2\colon year\ of\ birth > 1998 \rightarrow marital\ status = married$ and the probabilities with which r_1 and r_2 are expected to be fulfilled (e.g., based on census data). In particular, our approach evaluates the actual fulfillment of r_1 and r_2 in the customer database in comparison to the expected distribution of rule fulfillment. For example, the number of married people that are younger than 20 years is generally low, meaning that r_2 is expected to be fulfilled only with a low frequency (e.g., 4.1%). Thus, if r_2 is fulfilled similarly infrequently in the customer database (e.g., 4.2%), the corresponding data of married customers is assessed to have a high probability of being consistent.

This interpretation of metric values as probabilities is viable because statistical tests and the concept of the p-value form the methodological foundation for determining the metric values (cf. Section 3.2). Moreover, by assessing consistency as a probability, the metric values for each customer can be integrated in decision support, for instance, into the calculation of expected values. Such a calculation may reveal that targeting a married customer younger than 20 years within the campaign is only beneficial if the consistency of the data of this customer – represented by a probability – is greater than 0.8. Thus, applying the rule r_2, the metric can be used to determine whether this threshold is met (note that this threshold is totally different from rule confidence, as confidence of r_2 is only 4.2%).

3.2 Methodological Foundations

3.2.1 Uncertain Rules

A rule $r: A \rightarrow C$ consists of logical statements A and C, with A and C describing single attributes or relations between different attributes in D_B. The simplest form of a logical statement S is defined as (Chiang & Miller, 2008; Fan et al., 2013):

$$\begin{array}{l} < attribute >< operator >< attribute > \\ \text{or} \\ < attribute >< operator >< constant > \end{array} \tag{1}$$

Here, $< attribute >$ is one of the attributes a_i and $< operator >$ is a binary operator such as $=,\ \geq,\ >,\ \neq$ or $substring\_of$. Simple logical statements can be linked by conjunction (AND, ∧), disjunction (OR, ∨) or negation (NOT, ¬) to form more complex logical statements. For instance, in the running example, we may have a rule of the following form:

$$r_3: year\ of\ birth > 1998\ \land gender = female \rightarrow marital\ status = single \tag{2}$$

Here, $year\ of\ birth > 1998$, $gender = female$, $marital\ status = single$, and $year\ of\ birth > 1998\ \land gender = female$ are logical statements. To determine whether a logical statement S is true or false for a record t of D_B, it can be applied to t by replacing each attribute a_i contained in S by $\phi(t, a_i)$. In other words, the corresponding data values of the record are inserted. We further define the set of records in D_B rendering S true as $fulfilling\ records(D_B, S) := \{t \in T |\ S(t) \text{ is } true\}$.

As an example, we can apply the antecedent $year\ of\ birth > 1998 \land gender = female$ and the consequent $marital\ status = married$ of the rule r_3 to a record t of the database D_B with $\phi(t, year\ of\ birth) = 2000$, $\phi(t, gender) = female$ and $\phi(t, marital\ status) = married$. As

$2000 > 1998$, $female = female$ and $married = married$, it follows $A(t)\ true$ and $C(t)\ true$. Thus, $t \in fulfilling\ records(D_B, A)$ and $t \in fulfilling\ records(D_B, C)$.

We call a rule $r: A \rightarrow C$ *relevant* for a record $t \in T$ if $t \in fulfilling\ records(D_B, A)$. If r is relevant for t, we say that t *fulfills* r, if $t \in fulfilling\ records(D_B\ , A \wedge C)$, and that t *violates* r otherwise. As mentioned above, we consider uncertain rules and not just rules which are "true by definition". To be more precise, an *uncertain rule* in our context is defined as:

$$(r: A \rightarrow C, p(r)) \tag{3}$$

An uncertain rule $(r: A \rightarrow C, p(r))$ has two components. It comprises a rule r containing the logical statements A (antecedent) and C (consequent) as well as a number $p(r) \in [0;\ 1]$ representing the probability with which r is expected to be fulfilled. The probability $p(r)$ allows to specify the uncertainty of the rule r. In contrast to existing approaches, this allows to consider rules that are unlikely to be fulfilled as well as almost certain rules or rules which are "true by definition" (i.e., the special case $p(r) = 1$) for the assessment of consistency. It is different from the confidence of an association rule as it is not based on the relative frequency of rule fulfillment in the dataset to be assessed. Moreover, the probability $p(r)$ is not used for selecting rules (e.g., with a high probability of being fulfilled), but rather for assessing consistency (the determination of uncertain rules will be outlined in Section 3.4.1).

3.2.2 Using Uncertain Rules for the Assessment of Consistency

Let $r: A \rightarrow C$ be a rule in the rule set R and let D_B be the dataset to be assessed. The rule r is expected to be fulfilled with probability $p(r)$. Hence, if the records in $fulfilling\ records(D_B, A)$ are consistent with regard to r, the application of r to such a record t can be seen as a Bernoulli trial with success probability $p(r)$, where success is defined as $t \in fulfilling\ records(D_B, A \wedge C)$. This is, because applying r to a record in $fulfilling\ records(D_B, A)$ has only two possible outcomes: The rule can either be fulfilled (with probability $p(r)$) or violated (with probability $1 - p(r)$). Thus, the Bernoulli trial can be represented by a random variable $r(t)$ resulting in $r(t) \sim Bern(p(r))$:

$$r(t) := \begin{cases} 1 & \text{if } t \in fulfilling\ records(D_B, A \wedge C) \\ 0 & \text{if } t \notin fulfilling\ records(D_B, A \wedge C), t \in fulfilling\ records(D_B, A) \end{cases} \tag{4}$$

Similarly, r can then be applied to all records t in D_B with $t \in fulfilling\ records(D_B, A)$ and the results can be summed up by the random variable $X(r) := \sum_{t \in fulfilling\ records(D_B, A)} r(t)$. As a sum of independent Bernoulli-distributed random variables, $X(r)$ follows a binomial distribution with parameters $|fulfilling\ records(D_B, A)|$ and $p(r)$: $X(r) \sim B(|fulfilling\ records(D_B, A)|, p(r))$. An illustration for such a distribution with parameters 100 and 0.5 is presented in Figure 2.

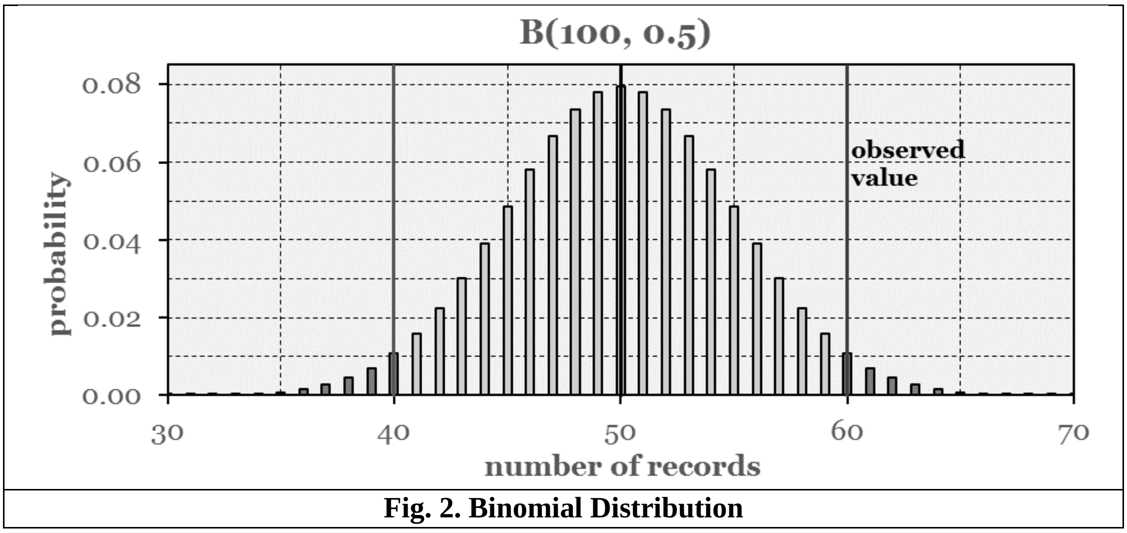


Fig. 2. Binomial Distribution

If the records in $fulfilling\ records(D_B, A)$ are consistent with regard to r and $p(r)$, it follows that $|fulfilling\ records(D_B, A \wedge C)|$ is distributed as the successes of $X(r)$. Thus, to determine the consistency of the records in $fulfilling\ records(D_B, A)$, the actual value of $|fulfilling\ records(D_B, A \wedge C)|$ is contrasted with the distribution of $X(r)$. In Figure 2, we observe $|fulfilling\ records(D_B, A \wedge C)| = 60$ and expected value $E[X(r)] = 50$, resulting in an indication of inconsistency.

Based on this idea, we develop a probability-based metric for consistency founded on the well-known concept of the (two-sided) p-value in hypothesis testing. Let $p'(r)$ be the relative frequency with which the rule r is fulfilled by a relevant record in the dataset D_B. If the relevant records are consistent with regard to r, then $p'(r)$ should correspond to $p(r)$ (e.g., 0.5 in Figure 2). Thus, in statistical terms, measuring consistency implies testing the null hypothesis $H_0: p'(r) = p(r)$ against the alternative hypothesis $H_1: p'(r) \neq p(r)$ for the binomially distributed random variable $X(r) \sim B(|fulfilling\ records(D_B, A)|, p(r))$. A two-sided alternative hypothesis is used because both too many and too few fulfillments of r indicate inconsistency: The more $|fulfilling\ records(D_B, A \wedge C)|$ deviates from $E[X(r)]$, the more the consistency of D_B decreases in regard to r.

This intuitive understanding is formalized by the two-sided p-value. It represents the probability that a value occurs under the null hypothesis which is equal to or more extreme than the observed value. For example, in Figure 2, $E[X(r)] = 50$ and observed value $|fulfilling\ records(D_B, A \wedge C)| = 60$. Since the distribution is symmetric, values $\geq\ 60$ and values $\leq\ 40$ are equal to or more extreme than the observed value. Following this, the two-sided p-value is calculated by summing up the probabilities $p(X(r) \geq 60)$ and $p(X(r) \leq 40)$, represented by the dark grey bars.

In our case, the observed value is $|fulfilling\ records(D_B, A \wedge C)|$ and the expected value is $E[X(r)]$. Thus, the p-value represents the probability that, under the null hypothesis, the random variable $X(r)$ yields a value equal to or more extreme than $|fulfilling\ records(D_B, A \wedge C)|$. Hence, it represents the probability that the assessed records in D_B are free of contradictions with regard to the rule r. The two-sided p-value of the random variable $X(r){\sim}B(|fulfilling\ records(D_B, A)|, p(r))$ with respect to the observed value $|fulfilling\ records(D_B, A \wedge C)|$ is denoted as follows:

$$p\text{-}value(X(r){\sim}B\big(|fulfilling\ records(D_B, A)|, p(r)\big), |fulfilling\ records(D_B, A \wedge C)|\big) \tag{5}$$

Note that we are aware of the discussion regarding the p-value (cf., e.g., Goodman, 2008) and since this is not the main focus of our paper, we follow the above standard interpretation. The outlined methodological foundations allow for a formal definition of our metric in the next subsection and ensure a clear interpretation of the metric values.

3.3 Definition of the Metric for Consistency

Let D_B be a database, $t_j \in T$ be a record in D_B, a_i be an attribute in D_B, and $r{:}\, A \rightarrow C$ with $p(r) \in [0;\ 1]$ be an uncertain rule such that a_i is part of r and $t_j \in fulfilling\ records(D_B, A \wedge C)$. We define the consistency of the data value $\phi(t_j, a_i)$ with regard to r as:

$$\begin{aligned} &Q_{Cons}\big(\phi(t_j, a_i), r{:}\, A \rightarrow C\big) := \\ &\quad p\text{-}value\big(X(r){\sim}B\big(|fulfilling\ records(D_B, A)|, p(r)\big), |fulfilling\ records(D_B, A \wedge C)|\big) \end{aligned} \tag{6}$$

This definition ensures that only attributes which are part of the antecedent or consequent and records which fulfill the rule are considered. The metric value $Q_{Cons}\big(\phi(t_j, a_i), r{:}\, A \rightarrow C\big)$ represents the probability that, if the relevant records are consistent with regard to r, the random variable $X(r)$ yields a value which is equal to or more extreme than $|fulfilling\ records(D_B, A \wedge C)|$.

The metric in Definition (6) measures consistency with regard to a single rule. If multiple rules can be used to assess the consistency of a specific data value, these rules can be aggregated for the assessment. This can be achieved by using conjunctions (AND, ∧). For example, let $r_1: A_1 \rightarrow C_1$ and $r_2: A_2 \rightarrow C_2$ be two rules available for the assessment of the data value $\phi(t_j, a_i)$. Then, it holds that $t_j \in fulfilling\ records(A_1 \wedge C_1 \wedge A_2 \wedge C_2)$ while a_i is part of A_1 or C_1 and part of A_2 or C_2, respectively. Thus, instead of the single rules r_1 and r_2, the aggregated rule $r_3: A_1 \wedge A_2 \rightarrow C_1 \wedge C_2$ can be considered and used to assess consistency in a well-founded manner by means of Definition (6). Analogously, an iterative aggregation can be applied if more than two rules are available.

Definition (6) allows the identification of data values which are likely to be inconsistent due to both random and systematic data errors. On the one hand, random data errors may lead to erroneous data values, thus contradicting a rule in the rule set R. On the other hand, systematic data errors may occur which usually bias the data values “in one direction” and thus cause $|fulfilling\ records(D_B, A \wedge C)|$ to differ considerably from $E[X(r)]$ for a rule $r: A \rightarrow C$ in R. Thus, for both random and systematic data errors, the considered p-value is low. As a result, both types of errors lead to low metric values indicating inconsistency of the corresponding data values with regard to R.

The metric in Definition (6) assesses consistency on the level of data values. On this basis, aggregated metric definitions for records, attributes, relations, and the whole database D_B can be determined. To do so, the weighted arithmetic mean of the metric values of the corresponding data values can be used similarly to, for example, Heinrich and Klier (2011). This allows the assessment of consistency on different data view levels and to support decisions relying on, for instance, the consistency of D_B as a whole.

3.4 Metric Instantiation

In this subsection, we describe how to instantiate our metric. In particular, we describe how uncertain rules can be obtained and how the metric values can be calculated.

3.4.1 Obtaining Uncertain Rules

For the application of the metric, it is crucial to determine an appropriate set of uncertain rules R and the corresponding values $p(r)$ for each uncertain rule $r \in R$. Generally, there are different possibilities to determine this rule set. We briefly describe the following three ways: (i) Analyzing a reference dataset, (ii) Conducting a study, and (iii) Surveying experts.

Ad (i): A promising option is to use a quality assured reference dataset D_R (in case such exists). This reference dataset D_R needs to be representative for the data of interest in D_B to allow the determination of meaningful uncertain rules. Such a reference dataset may, for example, be reliable historical data owned by the organization itself. With more and more external data being provided by recent open data initiatives, reliable publicly available data from public or scientific institutions (e.g., census data, government data, data from federal statistical offices and institutes) can be analyzed as well. The German Federal Statistical Office, for instance, offers detailed data about the population of Germany and thus for many attributes of typical master data (e.g., of customers). Further examples are traffic data as well as healthcare databases providing detailed (anonymized) data about diseases and patients. From such a reference dataset D_R, it is possible to determine uncertain rules for the assessment of D_B directly and with a high degree of automation. In the following, we exemplarily discuss three possible ways for determining uncertain rules based on a reference dataset.

First, an association rule mining algorithm (Agrawal et al., 1993; Kotsiantis & Kanellopoulos, 2006) can be applied to D_R. The resulting association rules can subsequently be used as input for the metric. Applying an association rule mining algorithm in this context differs from existing works using association rules for the assessment of consistency (e.g., Alpar & Winkelsträter, 2014). In our context the rules and their confidence are not determined based on the dataset to be assessed itself, but on a reference dataset, which prevents possibly misleading results in case part of the dataset to be assessed is erroneous. Moreover, using an association rule mining algorithm in our context means that uncertain rules with a rule confidence below a chosen threshold for minimum rule confidence are not excluded. Such rules with low confidence are beneficial for assessing consistency with the metric presented in this paper and, thus, should also be mined. This can be achieved using common association rule mining algorithms (e.g., the Apriori algorithm; Agrawal & Srikant, 1994).

Still, it is possible that for a specific data value, no association rule can be used to assess consistency because the data value is not part of an antecedent or consequent in any rule. Thus, we suggest further ways to determine or enhance a set of uncertain rules based on a reference dataset.

As a second way, we propose the use of so-called *column rules*, which can also be determined in an automated manner. Using column rules to assess the consistency of D_B means that dependencies between different attributes are not considered. These rules consist of a tautological antecedent $\top$ (i.e., the logical statement A is always true) and $a_l = \phi(t_m, a_l)$ as a consequent for all records t_m in D_R and attributes a_l of D_R. This results in the rule set of the form $R_c = \{r{:}\, \top \rightarrow\ a_l = \phi(t_m, a_l)\}$, where the probability of a rule represents the relative frequency of occurrence of $\phi(t_m, a_l)$ in D_R. For example, for a record t in D_R with $\phi(t, year\ of\ birth) = 1997$ and $\phi(t, marital\ status) = single,$ $r_1{:}\, \top \rightarrow year\ of\ birth = 1997$ and $r_2{:}\, \top \rightarrow marital\ status = single$ would be added to R_c.

Third, so-called *row rules* can also be used. Row rules are very strict with regard to their fulfillment, as all of the data values of a record need to match. These rules with tautological antecedent $A = \top$ and $\bigwedge_{a_l}(a_l = \phi(t_m, a_l))$ as consequent for all t_m in D_R can be generated in an automated manner as well. This leads to the rule set of the form $R_r = \{r{:}\, \top \rightarrow \bigwedge_{a_l}(a_l = \phi(t_m, a_l))\ \}$, where the probability of a rule represents the relative frequency of occurrence of t_m in D_R. To give an example, for a record t in D_R with $\phi(t, year\ of\ birth) = 1997$ and $\phi(t, marital\ status) = single$ (and no other attributes in D_R), the rule $r_3{:}\, \top \rightarrow year\ of\ birth = 1997 \wedge marital\ status = single$ would be added to R_r.

These three ways for obtaining uncertain rules based on a reference dataset D_R were presented because of their general applicability. A large variety of further uncertain rules can be determined, for example by considering fixed attributes in the antecedent or by using different operators. Depending on D_B and the specific application, any of these possibilities (or a combination of them) can be favorable as the dependencies between attributes may vary. For instance, in a context where dependencies of attributes do not have to be analyzed at all, using column rules is promising. Another example is provided in Section 4, where uncertain rules based on a reference dataset from the German Federal Statistical Office are determined. In any of these ways, the relative frequency with which r is fulfilled in D_R can be calculated and used as $p(r)$. Thereby, based on D_R both rules and corresponding probabilities of fulfillment can be determined with a high degree of automation. This allows a use of multiple rule sets to focus on different aspects of the data to be assessed or to analyze the specific reasons for inconsistencies in the data (cf. Section 4).

When using a reference dataset D_R for determining rules, the number $|fulfilling\, records(D_R, A \wedge C)|$ of records in D_R fulfilling a rule needs to be sufficiently large to ensure reliable metric values with respect to this rule. To be more precise, the statistical significance of $p(r)$ needs to be assured. If an association rule mining algorithm is used, a suitable minimum support can be fixed to exclude rules based on a non-significant proportion of records. In any case, a statistical test can be applied in order to determine the minimal number of records required such that a rule has a significant explanatory power (cf. Section 4). Moreover, to provide a statistically reliable basis and to circumvent the aforementioned issue, rules can be aggregated (e.g., by using a disjunction). In this way, robust estimations of $p(r)$ can be obtained, allowing the determination of reliable metric values.

Ad (ii): If neither internal nor external reference data is available, conducting a study is a further possibility. For example, if a customer database is to be assessed, a random sample of the customers can be drawn and surveyed. The survey results can be used to determine appropriate uncertain rules by analyzing the customers’ statements. Moreover, the corresponding values of $p(r)$ for each rule r can be obtained by analyzing how many of the surveyed customers fulfill the rule. Thus, the input parameters for the metric are provided. As a result of the survey, one obtains quality assured data of the surveyed customers and can also assess the consistency of the data of customers not part of the survey.

Ad (iii): Another possibility is to use an expert-based approach (similar to Mezzanzanica et al., 2012; Baker & Olaleye, 2013; Meyer & Booker, 2001). Here, the idea is to survey qualified individuals. For rules in a customer database of an insurer taking into the account the attributes *number of insurance relationships*, *insurance group* and *fee paid*, insurance experts could be surveyed. Another example concerns very rare events such as insurance exclusions without reimbursement, for which not enough (reference) data is available. The experts can assess which rules are suitable to describe the expected structure of the considered data values and can specify the respective values of $p(r)$ for each rule.

3.4.2 Calculating the Metric Values

Based on a set of uncertain rules R with values $p(r)$ for each $r \in R$, the metric values can be calculated in an automated manner. The values $|fulfilling\ records(D_B, A)|$ and $|fulfilling\ records(D_B, A \wedge C)|$ can be determined efficiently via simple database queries. In addition, based on the value of $p(r)$, the corresponding binomial distribution can be instantiated. Then, the (two-sided) p-value with regard to $|fulfilling\ records(D_B, A \wedge C)|$ can be calculated in order to obtain the metric values.

In the literature, several different approaches to calculate the two-sided p-value have been proposed (Dunne, Pawitan, & Doody, 1996). These include doubling the one-sided p-value and clipping to one, summing up the probabilities less than or equal to the probability of the observed result, and more elaborate ways. In practical applications, for non-symmetric distributions, the approaches to calculate the two-sided p-value may lead to slightly different results. However, the larger the sample size (in our case $|fulfilling\ records(D_B, A)|$), the smaller the differences between the results of the different approaches are. This is due to the fact that for $p(r) \in (0;\ 1)$, the binomial distribution converges to the (symmetric) normal distribution (de Moivre-Laplace theorem).

4 Evaluation

In this section we evaluate (E1) the practical applicability as well as (E2) the effectiveness (Prat, Comyn-Wattiau, & Akoka, 2015) of our metric for consistency in a real-world setting. First, we discuss the reasons for selecting the case of a German insurer and describe the assessed customer dataset. Then, we show how the metric could be instantiated for this case. Subsequently, we present and discuss the results of the application. Finally, we compare the results with those of existing metrics for consistency.

4.1 Case Selection and Dataset

The relevance of managing customer data at a high data quality level is well acknowledged (cf. e.g., Even, Shankaranarayanan, & Berger, 2010; Heinrich & Klier, 2015b). The metric was applied in cooperation with one of the major providers of life insurances in Germany. High data quality of customer master data is critical for the insurer and plays a particularly important role in the context of customer management. However, the staff of the insurer suspected data quality issues due to negative customer feedback (e.g., in the context of product campaigns). Customers claimed to have a marital status different from the focused target group of campaigns. Thus, they either were not interested in the product offerings or were not even eligible to participate. To analyze these issues, we aimed to assess the consistency of the customers' marital status depending on their age.

This setting seemed particularly suitable for showing the applicability and effectiveness of our metric for the following reasons: First, the marital status of a customer is a crucial attribute for the insurer, because insurance tariffs and payouts often vary depending on marital status. Indeed, for example a customer whose marital status is erroneously stored as *widowed* may receive unwarranted life insurance payouts. Additionally, the marital status also significantly influences product offerings, as customers with different marital statuses tend to have varying insurance needs. In fact, as mentioned above, customers may even only be eligible for a particular insurance if they have a specific marital status. Second, interpretable metric values are of particular importance in this setting, for instance to facilitate the aforementioned product offerings. Third, using traditional rules which are "true by definition" is not promising here as except for children, who are always *single*, no marital status is definite or impossible for customers. For example, a 60-year-old customer may be *single, married, divorced, widowed,* etc., each with specific probability.

To conduct the analyses described above, the insurer provided us with a subset of its customer database. The analyzed dataset contains five attributes storing data about customers of the insurer born from 1922 onwards and represents the state of the customer data from 2016. The subset consists of 2,427 records which had a value for both the attribute *marital status* and the attribute *date of birth*. Each record represents a specific customer of the insurer. The *marital status* of the customers was stored as a numerical value representing the different statuses *single*, *married*, *divorced*, *widowed*, *cohabiting*, *separated* and *civil partnership*. As the marital statuses *cohabiting* and *separated* are not recognized by German law (Coordination Unit for IT Standards, 2014), we matched these statuses to the respective official statuses *single* and *married*. The *date of birth* was stored in a standard date format. On this basis, customers' age could easily be calculated and stored as an additional attribute *age*. Moreover, an attribute *gender* was available both in the customer dataset as well as in the data used for the instantiation of the metric (cf. following subsection). As gender may have a significant impact on marital status as well, we also included this attribute in our analysis. Each of the 2,427 records contained a value for *gender*, classifying the respective customer as either *male* or *female*.

4.2 Instantiation of the Metric for Consistency

In Section 3.4.1, we described possibilities to obtain a set of uncertain rules for the instantiation of our metric. In our setting, we were able to use publicly available data from the German Federal Statistical Office as a reference dataset and thus chose option (i). The German Federal Statistical Office provides aggregated data regarding the number of inhabitants of Germany having a specific marital status. We used the most recent data available, which is based on census data from 2011 and was published in 2014 (German Federal Statistical Office, 2014). The data is broken down by age (in years) as well as gender and includes all Germans regardless of their date of birth, containing in particular the data of the insurer's customers. Overall, the data from the German Federal Statistical Office seems to be an appropriate reference dataset for our setting and could be used to determine meaningful uncertain rules and the probabilities $p(r)$ for each rule r.

As it was our aim to examine consistency of the marital status of customers depending on their age and gender, both attributes *age* and *gender* were part of the antecedent of the rules while the attribute *marital status* was contained in the consequent. To determine a rule set, we proceeded as follows: First, for each marital status *m*, each gender g and each possible value of age $a \in \mathbb{N}$, we specified rules of the following form:

$$r_{m,g}^{a}: (age = a) \wedge (gender = g) \rightarrow marital\ status = m \tag{7}$$

Second, we calculated the probabilities $p(r_{m,g}^{a})$ based on the data from the German Federal Statistical Office. Third, starting at an age of 0 years, we systematically aggregated these rules to rules of the form:

$$(age \geq a_1) \wedge (age < a_2) \wedge (gender = g) \rightarrow marital\ status = m \tag{8}$$

Here, $a_1, a_2 \in \mathbb{N}$ (with $a_1 < a_2$) specify an age group. The aggregation of the rules $r_{m,g}^{a}$ was performed to increase the number of records each rule was relevant for. However, age groups also have to be homogeneous and thus, the differences in probabilities of rule fulfillment within an age group were required to not exceed a specific threshold. More precisely, for a given value of a_1, the value a_2 was determined to be the maximum of all values $j \in \mathbb{N}$ for which $\left|p\left(r_{m,g}^{j}\right) - p\left(r_{m,g}^{k}\right)\right| \leq 0.1$ held for all

$a_1 \leq k \leq j$. In this way the following rule $\tilde{r}$ for single men between 42 and 49 was obtained:

$$\tilde{r}: (age \geq 42) \wedge (age < 50) \wedge (gender = male) \rightarrow marital\ status = single \quad (9)$$

Afterwards, for each rule $r \in R$ the probabilities $p(r)$ were calculated based on the data from the German Federal Statistical Office. For example, as approximately 26.4% of men between 42 and 49 are single according to the German Federal Statistical Office, this resulted in $p(\tilde{r})$=0.264. Moreover, a statistical test to the significance level of 0.05 was applied to ensure that each rule is based on a statistically significant number of relevant records in both the reference dataset and the customer dataset. Rules not fulfilling the test were excluded from further analysis to guarantee reliable metric results. This way, 37 different rules and corresponding probabilities were determined.

Each customer record of the insurer belonged to one of the age groups and had the value *male* or *female* for the attribute *gender* and the value *single*, *married*, *divorced*, *widowed* or *civil partnership* for the attribute *marital status* as represented by our rule set. Accordingly, a metric value could be determined for the value of the attribute *marital status* of each of these records. For instance, to assess the consistency of the marital status *single* of a 46-year-old male customer t, $\tilde{r}$ was used. The calculation of the metric value by means of Definition (6) yielded a consistency of 0.888:

$$Q_{Cons}(\phi(t, single), \tilde{r}) = p\text{-}value(X(\tilde{r}) \sim B(57{,}0.264), 14) = 0.888 \quad (10)$$

To calculate the two-sided p-value, we doubled the one-sided and clipped to one (Dunne et al., 1996).

4.3 Application of the Metric for Consistency and Results

Having instantiated the metric, we applied the metric to the 2,427 customer records by means of a Java implementation. The results for the marital status *widowed* seemed particularly interesting and alarming. Indeed, in contrast to the other marital statuses, analyses for this marital status revealed that the metric values were very low across all customer records. In fact, for the 1,160 records with a marital status of *widowed*, the metric value was always below 0.001 (cf. Table 2).

Gender	Age Group	Relative Frequency of Rule Fulfillment (Insurer Dataset)	Probability of Corresponding Rule (Statistical Office)	Value of the Metric for Consistency
male	0-74	0.139	0.012	0.000
	75-81	0.713	0.132	0.000

	>=82	0.744	0.313	0.000
female	0-60	0.096	0.016	0.000
	61-68	0.435	0.143	0.000
	69-73	0.676	0.248	0.000
	74-77	0.898	0.359	0.000
	78-80	0.950	0.483	0.000
	81-84	0.921	0.610	0.000
	>=85	0.918	0.754	0.000

Table 2. Results of the Metric for Consistency per Age Group and Gender

This means that for each record, the difference between actual rule fulfillment and expected rule fulfillment was so large that it is very unlikely to have occurred by chance. To be more precise, this probability was less than 0.001 for each record. Thus, with the results being based on a large number of records, consistency of *widowed* was assessed as very low with high statistical significance. This led to the conclusion that a previously undetected systematic bias had to be present in the customer data. In general, various different reasons could have led to this bias (e.g., a systematic data error such as a large number of young customers erroneously captured and stored as *widowed*). The bias was likely to cause serious problems for the insurer (e.g., due to negative effects on insurance tariffs and product offerings). We thus decided to investigate this issue further by analyzing each age group and gender based on the respective metric values, focusing on all rules with *marital status* = *widowed* in the consequent.

Table 2 illustrates the results of this analysis for all age groups. The first two columns display which customers were taken into account (rule antecedent). The third column shows the relative frequency of fulfillment of the respective rule (i.e., the proportion of customers in this age group and of this gender which had the marital status *widowed*). The penultimate column specifies the probability of the respective rule based on the data of the German Federal Statistical Office which was determined during the instantiation of the metric. Finally, the last column shows the corresponding metric value for consistency. Obviously, for a marital status of *widowed*, the bias in the data was so strong that the metric value was below 0.001 in each case. For example, the dataset included 107 female customers of age between 61 and 68 with marital status *widowed*, which results in a relative frequency of rule fulfillment of 0.435. The corresponding rule was:

$$(age \geq 61) \wedge (age < 69) \wedge (gender = female) \rightarrow marital\ status = widowed \tag{11}$$

The probability of this rule, however, was determined to be just 0.143 based on the data of the German Federal Statistical Office (i.e., 14.3% of female customers within that age group were expected to be *widowed*). Measuring consistency as the probability that the assessed data is free of internal contradictions with regard to this rule results in a metric value of 0.000 (rounded). This means that the actual rule fulfillment was so different from the expected rule fulfillment that it is very likely that a systematic bias was present in the customer data.

The results in Table 2 indicate that the relative frequency of rule fulfillment was considerably higher than the probability of the corresponding rule in each row. This means that a much larger number of customers than to be expected was considered as widowed by the insurer. A systematic bias of this magnitude in the insurer's customer data could result in severe economic losses for the insurer. Thus, we aimed to find the reason(s) for this potential data quality issue.

We suspected that a data capturing problem or a data integration problem might have occurred during some time in the past, resulting in many customers being erroneously stored as *widowed*. To analyze this presumption, we took the additional attribute *month of acquisition* of the dataset into account. It represents the month in which a person first became customer of the insurer by a standard date format. Of the 2,427 records, 931 records had a *month of acquisition* in the recent years 2013-2016, while 786 records exhibited a *month of acquisition* further in the past (until November 1951) and 710 records had a missing value for this attribute. We chose 2013 as threshold because the insurer data was structured differently from this year on. We created a new rule set including *month of acquisition* in the antecedent. This rule set was determined analogously to the procedure above (with slightly different age groups due to *month of acquisition*). For example, the rule for a widowed female customer in age group 55-68 acquired by the insurer in 2013-2016 was then given by:

$$\begin{gathered}(age \geq 55) \wedge (age < 69) \wedge (gender = female) \wedge (month\ of\ acquisition \in [2013, 2016]) \\ \rightarrow marital\ status = widowed\end{gathered} \tag{12}$$

The probabilities of the rules were again determined based on the German Federal Statistical Office data regarding the respective age, gender and marital status (e.g., 0.108 for the rule in (12)). The results from applying our metric using this new rule set are illustrated in Table 3. Here, we focus on the age group per gender with the highest number of widowed customers. The first two columns again specify which customers were taken into account. The probability of the rules for this age group and gender based on data from the German Federal Statistical Office is given in the third column. The fourth to sixth columns show the relative frequencies of rule fulfillment and the corresponding metric values for a missing *month of acquisition*, a *month of acquisition* before 2013 and a *month of acquisition* in 2013-2016.

Age Group	**Gender**	**Probability of Corresponding Rules**	**Relative Frequency of Rule Fulfillment (Insurer Dataset)/ Value of the Metric if Value of *month of acquisition is...***		
			...missing	**...before 2013**	**...in 2013-2016**
63-80	male	0.078	0.760 / 0.000	0.494 / 0.000	0.067 / 0.792
55-68	female	0.108	0.794 / 0.000	0.458 / 0.000	0.105 / 0.978

Table 3. Results of the Metric for Consistency considering the Month of Acquisition

This more detailed analysis shows that metric values are equal to 0.000 in the case of a missing *month of acquisition* or a *month of acquisition* before 2013, caused by very large relative frequencies of the data value *widowed* compared to the low probabilities of the corresponding rules. In contrast, for a *month of acquisition* in 2013-2016, relative frequencies and probabilities are much closer (0.067 and 0.078 resp. 0.105 and 0.108), resulting in higher metric values (0.792 resp. 0.978). We concluded that mainly records with a missing *month of acquisition* or a *month of acquisition* before 2013 were problematic and caused the consistency problems.

We discussed our findings with a board member of the insurer. He confirmed that an organizational restructuring had taken place in 2013. It included a revamp of the data capturing process, giving a reason why the customer data from 2013 onwards showed significantly higher consistency. However, it was not known that a data quality problem concerning the marital status *widowed* had existed beforehand. This problem was neither recognized nor solved during the restructuring process and thus still persisted in the customer data. Subsequently conducted internal evaluations of the insurer revealed that the *marital status* of customers had not been captured rigorously in the past and thus its values for customers with a value of *month of acquisition* before 2013 were not trustworthy. This clarified the too large relative frequency of *widowed* in the case of a *month of acquisition* before 2013 (and a missing *month of acquisition*, indicating an even more erroneous record). Further, the values of our metric for consistency allowed to quantify the too large relative frequency of *widowed* and to decide whether the deviation was significant. Thus, due to the clear interpretation, the metric values could then be used to decide which data values of *marital status* to consider as trustworthy in the future. Later on, the board member stated that initiatives to check the marital status of customers acquired before 2013 were started in order to rectify erroneous records (and, e.g., prevent unwarranted life insurance payouts). To do so, employees of the insurer began to analyze old paper-based documents containing customer data. Moreover, the insurer aimed to improve its data quality by contacting customers whose marital status was (highly) probably erroneous as identified by our metric. These initiatives facilitate an improved customer management, for instance regarding the design of future customer campaigns. In particular, a high data quality of the marital status of customers supports to conduct successful campaigns focusing on a specific target group of customers.

In addition, we analyzed the efforts for the instantiation and application of the proposed metric (in the sense of required time) as well as the corresponding benefits in this real-world setting. With respect to efforts, time was required to (i) find and prepare the census data of the German Federal Statistical Office, (ii) calculate the probabilities $p(r)$ based on the census data and conclude the rule set, (iii) assess the consistency by means of the metric and (iv) interpret and discuss the results. To conduct these four steps in our application setting, the following amount of time was necessary: With respect to (i), the data from the German Federal Statistical Office could be easily found online via a quick research. Due to their clear structure, preprocessing this data was not difficult after an initial familiarization. All in all, step (i) could be completed in one person-day. In another person-day, the rule set including the probabilities $p(r)$ was obtained and discussed. Indeed, the rule set and the probability for each rule could be determined in an automated manner. Based on this rule set, the assessment of the consistency of the dataset (iii) could be performed in less than one second using a Java implementation, which was realized in three further person-days. Of course, this effort is necessary only once and the implementation can be reused in further assessments, even in different application contexts. Finally, the results were interpreted and discussed (iv) both internally and in cooperation with the insurer in the course of two more person-days. Thus, the four steps (i) to (iv) to instantiate and apply the metric including the discussion of the findings resulted in overall efforts of about seven person-days.

These steps could be seen as part of a typical data quality assessment and improvement process (cf., e.g., Wang, 1998 and Figure 3). Here, in a preceding step, the data quality problems at hand have to be recognized and analyzed before the metric for semantic consistency is applied. For the insurer, this resulted in focusing on the consistency of the customers' marital status depending on their age. Similarly, in a succeeding step, initiatives to fix identified inconsistencies can be performed. Both complementing steps are related to the particular application context and, for instance, depend on the extent of identified semantic inconsistencies. In the case of the insurer, initiatives were conducted to improve the quality of the customer data to support future campaigns. In this regard, it is important to note that the efforts of the steps (i) to (iv) are reduced if an instantiated metric is reused in future consistency assessments. For example, data of new customers can be assessed using the same rules, probabilities (i.e., $p(r)$) and (tool) implementation. Only after some time (e.g., several years), an update of the underlying census data may become necessary to reflect demographical changes and to thus ensure valid results. However, even in this case, the four steps (i) to (iv) remain the same and the existing implementation can be used, resulting in smaller efforts compared to an initial conduction.

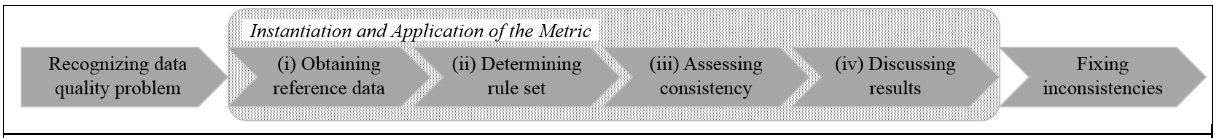

Instantiation and Application of the Metric
Recognizing data quality problem
(i) Obtaining reference data
(ii) Determining rule set
(iii) Assessing consistency
(iv) Discussing results
Fixing inconsistencies

Fig. 3. Core Steps to Instantiate and Apply the proposed Metric

Compared to the efforts for performing the steps (i) to (iv), which can be determined in a straightforward manner, the benefits of both (re)using the metric results (i.e., the resulting probabilities) and (re)using the improved data values are not easy to assess. From a methodological and decision-oriented perspective, both benefits can be estimated by comparing the effects resulting from decisions with respectively without considering the metric results and the improved data quality (for a detailed discussion cf. Heinrich & Hristova, 2016; Heinrich, Hristova, Klier, Schiller, & Szubartowicz, 2018). Not having or considering the metric results means that customers who actually have a marital status different from the focused target group are selected for the campaigns. Thus, products are offered to those customers wrongly. This may result in claims, which can be counted, assessed and attributed to a campaign as they arrive, allowing the quantification of their amount and severity. Preventing these claims by taking into account the metric results manifests a first benefit. However, such claims put forward to the insurer will just occur in a small number of cases and constitute only the "tip of the iceberg", as many customers would be annoyed by the campaign conducted based on low data quality, but not complain at all. The prevention of this decreasing customer satisfaction as a second (soft) benefit is difficult to measure. Moreover, using data values with improved data quality based on applying the consistency metric can lead to further improved decisions. More precisely, customers with corrected marital status can then be addressed in campaigns for which they would otherwise have been disregarded. Product sales being caused by these additionally considered customers constitute a third benefit resulting from fixing identified inconsistencies (thus representing a succeeding effect of applying the metric). In addition, both metric results and improved data quality cannot only be used in a single campaign, but also in future campaigns and customer interactions resulting in further benefits dependent on the particular application context (for a general decision-oriented framework comprising efforts and benefits of data quality assessment, we refer to Heinrich et al. 2018). Overall, in the case of the insurer, the efficiency can be supported; however, without any doubt efficiency has to be examined individually for each application context.

4.4 Comparison of the Results with existing Metrics for Consistency

In order to further evaluate our approach, we also instantiated and applied existing metrics for consistency (Alpar & Winkelsträter, 2014; Cordts, 2008; Heinrich et al., 2007; Heinrich & Klier, 2015a; Hinrichs, 2002; Hipp et al., 2001; Hipp et al., 2007; Kübart et al., 2005; Pipino et al., 2002) for the case of the German insurer and compared the results. Thereby, we used the same dataset and again focused on the attributes *age*, *gender* and *marital status*. To instantiate the existing metrics, we determined association rules with *marital status* in the consequent. The values for minimum support and minimum confidence were chosen in accordance with the respective works. In particular, each existing metric was instantiated using three different settings for minimum support and minimum confidence, leading to rule sets of different sizes: In Setting 1 (minimum support: 0.01, minimum confidence: 0.80), 26 association rules were determined. Setting 2 (minimum support: 0.00025, minimum confidence: 0.85) led to 111 rules and Setting 3 (minimum support: 0.0001, minimum confidence: 0.75) to 153 rules. Further, not all existing metrics provide values within the interval $[0; 1]$. Thus, to be able to compare the results, we transformed all metric values to this interval. This was done so that for each approach, the value 0 resp. 1 represent the minimal resp. maximal determined consistency.

For each approach and setting, we analyzed the minimum, average and maximum metric values over all records with marital status *widowed*. Regarding the existing approaches, the consistency of the attribute value *widowed* of the attribute *marital status* in the dataset is actually assessed to be rather high or even very high. Indeed, all approaches except the ones by Alpar and Winkelsträter (2014) and Hipp et al. (2007) assess the dataset as perfectly consistent or almost perfectly consistent (average metric value of at least 0.991). Even the metric values determined by the approaches of Alpar and Winkelsträter (2014) and Hipp et al. (2007) do not indicate a (critical) consistency problem as the average metric values are still at least 0.689 and thus rather high. Hence, existing approaches do not identify the severe consistency problem existing in the data and acknowledged by the insurer. In contrast, this problem is clearly indicated by the very low metric values (0.000 each as minimum, average and maximum metric value) determined by our approach using uncertain rules. The evaluation results are presented in Table 4 (higher metric values are represented by cells with darker background).

	Setting	**Minimum...**	**Average...**	**Maximum...**
		...metric value of records with marital status *widowed*		
Our proposed Metric	N/A	0.000	0.000	0.000
Alpar and Winkelsträter (2014); Hipp et al. (2007)	1	0.492	0.716	1.000
	2	0.483	0.689	1.000
	3	0.269	0.796	1.000
Hipp et al. (2001); Kübart et al. (2005)	1	1.000	1.000	1.000
	2	1.000	1.000	1.000
	3	0.556	0.996	1.000
Hinrichs (2002)	1	1.000	1.000	1.000
	2	1.000	1.000	1.000
	3	0.304	0.993	1.000
Cordts (2008); Pipino et al. (2002)	1	1.000	1.000	1.000
	2	1.000	1.000	1.000
	3	0.000	0.991	1.000
Heinrich et al. (2007); Heinrich and Klier (2015a)	1	1.000	1.000	1.000
	2	1.000	1.000	1.000
	3	0.000	0.991	1.000

Table 4. Comparison of the Results with existing Metrics for Consistency

To sum up, regarding (E1), the evaluation in a real-world setting demonstrated the practical applicability of our metric for consistency. Publicly available data could be used to determine a rule set with probabilities for each rule and instantiate the metric. Thereafter, the metric could be applied to identify consistency problems in the considered dataset. With respect to (E2), the evaluation also substantiated the effectiveness of our metric. Applying the metric multiple times (for increasingly detailed analyses) led to the identification of specific consistency problems in a real-world customer dataset, which, in comparison, was not supported when using existing metrics for consistency.

5 Conclusion, Limitations and Future Work

In this paper, we present a probability-based metric for the data quality dimension semantic consistency using uncertain rules. Existing approaches for measuring semantic consistency only consider rules that are “true by definition”, which means, the fulfillment of such a rule is always used as an indicator for high consistency. This impedes the consideration of rules that are expected to be *not* fulfilled for a higher number of data values. For example, a rule which is expected to be fulfilled only rarely, but is actually fulfilled very often in the assessed dataset, is an important indicator for inconsistent data. In addition, “true by definition” rules based on the assessed data can lead to misleading results if, for instance, a large part of the data is erroneous due to a systematic data error. Then the smaller part of accurately stored data values would be considered as inconsistent. Consequently, many consistency problems cannot be detected and assessed. We thus consider uncertain rules in the assessment of consistency by taking into account the probability with which a rule is expected to be fulfilled. This allows to determine a metric value which represents the probability that the dataset to be assessed is free of internal contradictions with regard to uncertain rules. The theoretical foundation for determining the metric values are statistical tests and the concept of the p-value. In particular, the fulfillment of a rule is modeled as a Bernoulli-distributed random variable. On this foundation, our metric is defined as the two-sided p-value of a binomial distribution. Thus, the metric values can be interpreted as the probability that the data values to be assessed do not contradict the considered rule set. This clear interpretation is relevant to support decision-making based on the metric values. We provide a formal metric definition and present different possibilities for the instantiation of the metric, in particular for determining a rule set. Further, we evaluate the practical applicability and effectiveness of our metric in a real-world setting by analyzing a customer dataset of an insurance company. Here, our metric could be applied to identify consistency problems in the data, which was not supported when using existing metrics for consistency.

There are also some limitations that may constitute the starting point for future research. To begin with, we evaluated our metric by analyzing a single customer dataset. Future research could, first of all, cover the application of the metric to additional datasets containing master data. Further, an application of the metric to different contexts such as, for example, sensor data is promising as well and has already yielded interesting results in an initial analysis we conducted. Moreover, for our application to the customer dataset, we determined a rule set based on reference data from the German Federal Statistical Office. Other ways to instantiate the metric are also feasible, but may require additional considerations (e.g., how to conduct a cost-efficient survey to determine the rule set). Future research should thus evaluate the application of other types of rules such as association rules, rules obtained by a survey and rules derived by experts. Moreover, the dataset we assessed contained about 2,400 records and is thus not very large. It would be interesting to apply the metric to a larger dataset and compare the results. Another possible path for future research is to develop elaborate aggregation procedures which take the statistical properties of the metric into account. For instance, an aggregation could be defined based on the sum of random variables following a Bernoulli distribution and thus also be interpreted as p-value. Finally, our metric is defined for structured data. However, in general, it can be extended to semi- and unstructured data by applying text mining methods such as inverted term frequency.

6 References

Abboura, A., Sahri, S., Baba-Hamed, L., Ouziri, M., & Benbernou, S. (2016). Quality-based online data reconciliation. *ACM Transactions on Internet Technology (TOIT), 16*(1:3).

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, pp. 207–216).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)* (Vol. 1215, pp. 487–499).

Alkharboush, N., & Yuefeng Li. (2010). A decision rule method for assessing the completeness and consistency of a data warehouse. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).* https://doi.org/10.1109/WI-IAT.2010.245

Alpar, P., & Winkelsträter, S. (2014). Assessment of data quality in accounting data with association rules. *Expert Systems with Applications, 41*(5), 2259–2268.

Baker, E., & Olaleye, O. (2013). Combining experts: Decomposition and aggregation order. *Risk Analysis, 33*(6), 1116–1127.

Batini, C., & Scannapieco, M. (2006). *Data-Centric Systems and Applications: Concepts, Methodologies and Techniques*: Springer.

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality*: Springer.

Blake, R. H., & Mangiameli, P. (2009). Evaluating the semantic and representational consistency of interconnected structured and unstructured data. In *Proceedings of the Americas Conference on Information Systems (AMCIS).*

Blake, R. H., & Mangiameli, P. (2011). The effects and interactions of data quality and problem complexity on classification. *Journal of Data and Information Quality (JDIQ), 2*(2), 8:1–28.

Bohannon, P., Fan, W., Geerts, F., Jia, X., & Kementsietsidis, A. (2007). Conditional functional dependencies for data cleaning. In *IEEE 23rd International Conference on Data Engineering (ICDE)* (pp. 746–755).

Bronselaer, A., Nielandt, J., Mol, R. de, & Tré, G. de (2016). Ordinal assessment of data consistency based on regular expressions. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 317–328).

Chiang, F., & Miller, R. J. (2008). Discovering data quality rules. *Proceedings of the VLDB Endowment*, *1*(1), 1166–1177.

Cong, G., Fan, W., Geerts, F., Jia, X., & Ma, S. (2007). Improving data quality: Consistency and accuracy. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB)* (pp. 315–326).

Coordination Unit for IT Standards. (2014). *Data set for the reporting system. Uniform Federal and State Part (DSMeld)*. Datensatz für das Meldewesen. Einheitlicher Bundes-/Länderteil (DSMeld). (In German).

Cordts, S. (2008). Implementation of a data quality service for the evolutional data quality improvement in relational database management systems (in German) (Dissertation), University of Hamburg.

Dunne, A., Pawitan, Y., & Doody, L. (1996). Two-sided p-values from discrete asymmetric distributions based on uniformly most powerful unbiased tests. *The Statistician*, *45*, 397–405.

Even, A., Shankaranarayanan, G., & Berger, P. D. (2010). Evaluating a model for cost-effective data quality management in a real-world CRM setting. *Decision Support Systems*, *50*, 152–163.

Fan, W., Geerts, F., Tang, N., & Yu, W. (2013). Inferring data currency and consistency for conflict resolution. In *IEEE 29th International Conference on Data Engineering (ICDE)* (pp. 470–481).

Fisher, C. W., Lauria, E. J. M., & Matheus, C. C. (2009). An accuracy metric: Percentages, randomness, and probabilities. *Journal of Data and Information Quality (JDIQ)*, *1*(3), 1–21. https://doi.org/10.1145/1659225.1659229

German Federal Statistical Office. (2014). Current population. Retrieved from https://www.destatis.de/EN/FactsFigures/SocietyState/Population/CurrentPopulation/CurrentPopulation.html

Golab, L., Karloff, H., Korn, F., Srivastava, D., & Yu, B. (2008). On generating near-optimal tableaux for conditional functional dependencies. *Proceedings of the VLDB Endowment*, *1*(1), 376–390.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology* (Vol. 45, pp. 135–140).

Heinrich, B., & Hristova, D. (2016). A quantitative approach for modelling the influence of currency of information on decision-making under uncertainty. *Journal of Decision Systems*, *25*(1), 16–41. https://doi.org/10.1080/12460125.2015.1080494

Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2018). Requirements for Data Quality Metrics. *Journal of Data and Information Quality (JDIQ)*, *9*(2), 12.

Heinrich, B., Kaiser, M., & Klier, M. (2007). Metrics for measuring data quality - foundations for an economic oriented management of data quality. In *Proceedings of the 2nd International Conference on Software and Data Technologies (ICSOFT).* INSTICC/Polytechnic Institut of Setúbal Barcelona, Spain.

Heinrich, B., & Klier, M. (2011). Assessing data currency—a probabilistic approach. *Journal of Information Science*, *37*(1), 86–100.

Heinrich, B., & Klier, M. (2015a). Data quality metrics for an economically oriented quality management (in German). In *Daten-und Informationsqualität* (pp. 49–67). Springer.

Heinrich, B., & Klier, M. (2015b). Metric-based data quality assessment—Developing and evaluating a probability-based currency metric. *Decision Support Systems*, *72*, 82–96.

Hinrichs, H. (2002). Data quality management in data warehouse systems (in German) (Dissertation). Universität Oldenburg.

Hipp, J., Güntzer, U., & Grimmer, U. (2001). Data quality mining - making a virtue of necessity. In *DMKD*.

Hipp, J., Müller, M., Hohendorff, J., & Naumann, F. (2007). Rule-Based Measurement Of Data Quality In Nominal Data. In *12th International Conference on Information Quality (ICIQ)* (pp. 364–378).

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery, 7*(1), 81–99.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 71–82.

Kübart, J., Grimmer, U., & Hipp, J. (2005). Rule-based outlier search for data quality analysis (in German). *Datenbank-Spektrum, 14*, 22–28.

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A survey on data quality: Classifying poor data. In *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)* (pp. 179–188).

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006). *Journey to Data Quality*: The MIT Press.

Lee, Y. W., Pipino, L., Strong, D. M., & Wang, R. Y. (2004). Process-embedded data integrity. *Journal of Database Management (JDM), 15*(1), 87–103.

Liu, L., & Chi, L. N. (2002). Evolutional data quality: a theory-specific view. In *Proceedings of the Seventh International Conference on Information Quality (ICIQ)* (pp. 292–304).

Mecella, M., Scannapieco, M., Virgillito, A., Baldoni, R., Catarci, T., & Batini, C. (2002). Managing Data Quality in Cooperative Information Systems. In R. Meersman & Z. Tari (Eds.), *Lecture Notes in Computer Science. On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE* (Vol. 2519, pp. 486–502). Springer. https://doi.org/10.1007/3-540-36124-3_28

Meyer, M. A., & Booker, J. M. (2001). *Eliciting and analyzing expert judgment: a practical guide* (Vol. 7): SIAM.

Mezzanzanica, M., Cesarini, M., Mercorio, F., & Boselli, R. (2012). Towards the Use of Model Checking for Performing Data Consistency Evaluation and Cleansing. In Laure Berti-Equille, Isabelle Comyn-Wattiau, & Monica Scannapieco (Eds.), *Proceedings of the 17th International Conference on Information Quality (ICIQ)* (pp. 163–177). MIT.

Moges, H.-T., Dejaeger, K., Lemahieu, W., & Baesens, B. (2011). Data quality for credit risk management: new insights and challenges. In *Proceedings of the 16th International Conference on Information Quality (ICIQ)* (pp. 632–646). Adelaide, Australia.

Moges, H.-T., van Vlasselaer, V., Lemahieu, W., & Baesens, B. (2016). Determining the use of data quality metadata (DQM) for decision making purposes and its impact on decision outcomes—An exploratory study. *Decision Support Systems, 83*, 32–46.

Moore, S. (2017). How to Create a Business Case for Data Quality Improvement. Retrieved from http://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/

Ngai, E. W. T., Gunasekaran, A., Wamba, S. F., Akter, S., & Dubey, R. (2017). Big data analytics in electronic markets. *Electronic Markets*, *27*(3), 243–245. https://doi.org/10.1007/s12525-017-0261-6

Oliveira, P., Rodrigues, F., & Henriques, P. (2005). A Formal Definition of Data Quality Problems. In *Proceedings of the Tenth International Conference on Information Quality (ICIQ).*

Orr, K. (1998). Data quality and systems theory. *Communications of the ACM, 41*, 66–71.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211–218.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems*, *32*(3), 229–267. https://doi.org/10.1080/07421222.2015.1099390

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Bulletin on Data Engineering, 23*(4), 3–13.

Redman, T. C. (1996). *Data Quality for the Information Age*: Artech House.

Scannapieco, M., Missier, P., & Batini, C. (2005). Data Quality at a Glance. *Datenbank-Spektrum*, *14*, 6–14.

Shankaranarayanan, G., Iyer, B., & Stoddard, D. (2012). Quality of Social Media Data and Implications of Social Media for Data Quality. In Laure Berti-Equille, Isabelle Comyn-Wattiau, & Monica Scannapieco (Eds.), *Proceedings of the 17th International Conference on Information Quality (ICIQ)* (pp. 311–325). MIT.

Singh, R., & Singh, K. (2010). A descriptive classification of causes of data quality problems in data warehousing. *International Journal of Computer Science Issues*, *7*(3), 41–50.

Srikant, R., & Agrawal, R. (1996). Mining quantitative association rules in large relational tables. In *ACM SIGMOD Record* (Vol. 25, pp. 1–12).

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95.

Wang, H., Li, J., & Gao, H. (2016). Data Inconsistency Evaluation for Cyberphysical System. *International Journal of Distributed Sensor Networks*, *12*(8).

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58–65.

Witchalls, C. (2014). Gut & gigabytes: Capitalising on the art & science in decision making. Retrieved from http://www.economistinsights.com/business-strategy/analysis/gut-gigabytes

Zak, Y., & Even, A. (2017). Development and evaluation of a continuous-time Markov chain model for detecting and handling data currency declines. *Decision Support Systems*, *103*, 82–93.