Handling data imperfection-False data inputs in applications for Alzheimer's patients

Fatma Ghorbel, Fayçal Hamdi, Nassira Achich, Elisabeth Metais

Handling Data Imperfection - False Data Inputs in Applications for Alzheimer's Patients

Fatma GHORBEL [a,b], Fayçal HAMDI [a], Nassira ACHICH [a,b], Elisabeth METAIS [a]

[a] *CEDRIC Laboratory, Conservatoire National des Arts et Métiers (CNAM), Paris, France*

[b] *MIRACL Laboratory, University of Sfax, Sfax, Tunisia*

Abstract

Handling data imperfection is a crucial issue in many application domains. This is particularly true when handling imperfect data inputs in applications for Alzheimer's patients. In this paper we first propose a typology of imperfection for data entered by Alzheimer's patients or their caregivers in the context of these applications (mainly due to the memory discordance caused by the disease). This topology includes nine direct and three indirect imperfection types. The direct ones are deduced from the data inputs e.g. uncertainty and uselessness. The indirect imperfection types are deduced from the direct ones, e.g. the redundancy. We then propose an approach, called DBE_ALZ, that handles false data entry by estimating the believability of each data input. Based on the proposed typology, the falsity of these data is related to five imperfection types: uncertainty, confusion, typing error, wrong knowledge and inconsistency. DBE_ALZ includes a believability model that defines a set of dimensions and sub-dimensions allowing a qualitative estimation of the believability of a given data input. It is estimated based on its reasonableness and the reliability of its author. Compared to related work, the data input reasonableness is measured not only based on common-sense standard, but also based on a set of personalized assertions. The reliability of the patient is estimated based on the progression of the disease and the state of his memory at the moment of entry. However, the reliability of the caregiver is estimated based on his age and his knowledge about the data input's field. Based on the believability model, we estimate quantitatively the believability of the data input by defining a set of metrics associated to the proposed dimensions and sub-dimensions. The measurement methods rely on probability and fuzzy set theories to reason about uncertain and imprecise knowledge (Bayesian networks and Mamdani fuzzy inference systems). Three languages are supported: English, French and Arabic. Based on the generated believability degrees, a set of decisive actions are proposed to guarantee the quality of the data inputs e.g., inferring or not based on a given data. We illustrate the usefulness of our approach in the context of the Captain Memo memory prosthesis. Finally, we discuss the encouraging results derived from the evaluation step.

Keywords: Applications for Alzheimer's Patients, Imperfection Types, False Data Inputs, Believability.

1. Introduction

In the context of the VIVA[1] project ("*Vivre à Paris avec Alzheimer en 2030 grâce aux nouvelles technologies*"), we have proposed a memory prosthesis, called Captain Memo [1], to help Alzheimer's patients to palliate mnesic problems. This prosthesis supplies a set of services. Among these services, one is devoted to "remember things about people" i.e., it helps users to remember their convivial surroundings and relatives. In Captain Memo, personal data of the patient are structured semantically using an ontology called, PersonLink[2] [2]. This multicultural and multilingual OWL 2

---

[1] http://viva.cnam.fr/
[2] http://cedric.cnam.fr/isid/ontologies/files/PersonLink.html

ontology enables the storing, modeling and reasoning about interpersonal relationships (e.g., husband, aunt and half-brother). Three languages are supported: English, French and Arabic.

Captain Memo is proposed to be used by patients with earliest symptoms of Alzheimer. The main objective is to improve the autonomy of the patients by making them active in entering data. However, these particular users, living in uncertainty, may introduce imperfect data. For example, an Alzheimer's patient, who has only one son named Paul, could enter that he also has a daughter named Juliette. This information will then be used by the Captain Memo genealogic service that will generate a wrong genealogic tree that states that Juliette is the daughter of the patient and the sister of Paul. The wrong fraternity link is automatically inferred from the wrong input. Existing applications for this category of users, deals with this problem by only considering data inputs given by the caregivers and make the patient passive (they assume that the data inputs given by the caregivers are more accurate than the patient' ones). These inputs have to be assessed also as, for instance, the caregivers may not necessarily know all information related to the patient's private life. This paper addresses the issue of handling imperfect data inputs in applications for Alzheimer's patients. Data may be given by Alzheimer's patients or their caregivers. Our contributions consist of two parts:

- On the one hand, we focus on **classifying the different types of imperfection that may affect data entered by the Alzheimer's patients or their caregivers**. In the literature, several typologies of data imperfection have been proposed. Some are generic and others are specific to a given domain. However, to the best of our knowledge, there is no typology of imperfection of data inputs in the context of applications for Alzheimer's patients. We propose a typology of imperfection of these data. It offers direct and indirect imperfection types. The direct ones are deduced from the entered data e.g., uncertainty and uselessness. The indirect imperfection types are the ones which can be deduced from the direct ones. For instance, the redundancy can be generated from the uselessness.

- On the other hand, we focus on **handling false data inputs in applications for Alzheimer's patients** by proposing an approach called DBE_ALZ. Based on the proposed typology, the falsity of these data is related to five types of imperfection: uncertainty, confusion, typing error, wrong knowledge and inconsistency. To deal with false data inputs, a process estimating the believability of these inputs should be added. This process, which assigns to each data input a believability degree, will allow these applications to accept only data inputs having high believability degrees. In the literature, several approaches focus on data believability estimation. These approaches are only a source of inspiration for us. To the best of our knowledge, there is no approach to estimate the believability of data inputs in applications for Alzheimer's patients. For instance, the existing approaches do not take into consideration the patient's particular profile. The proposed DBE_ALZ approach includes a model that defines a set of dimensions and sub-dimensions allowing a qualitative estimation of the believability of a given data input. This estimation is based on the reasonableness of the data and the reliability of its author. Compared to related work, the data input reasonableness is measured not only in common-sense standard, but also on a set of personalized rules. The reliability of the patient is estimated based on the progression of the disease and the state of his memory at the moment of entry. The reliability of the caregiver is estimated based on his age and his knowledge of the data input's field. Furthermore, based on the proposed model, we estimate quantitatively the data input believability by defining a set of metrics associated with the proposed dimensions and sub-dimensions. This step is based on probability and fuzzy set theories to reason about uncertain and imprecise knowledge (Bayesian networks and Mamdani fuzzy inference systems). Based on the generated believability degrees, a set

of decisive actions are proposed to guarantee the quality of the data inputs e.g., inferring or not based on a given data.

The rest of the present paper is organized as follows. Section 2 is devoted to present some preliminaries and related work. Section 3 details the proposed typology of data inputs imperfection in applications for Alzheimer's patients. Section 4 details the DBE_ALZ approach for handling false data inputs in the context of these applications. Section 5 presents the DBE_ALZ approach-based prototype, proposes a semantic representation of the believability degrees in ontology, and illustrates the usefulness of our work within the context of the Captain Memo memory prosthesis. Section 6 presents and discusses some experimental results related to the evaluation study. Finally, Section 7 draws conclusions and future research directions.

2. Preliminaries and Related Work

The present work is closely related to the three following research areas: *(1)* data imperfection, *(2)* data quality *(3)* and data believability.

2.1 Data Imperfection

Available data in information systems are mostly imperfect [3, 4]. Data imperfection can have several forms and types. In the literature, several typologies have been proposed [5]. We distinguish two main categories: generic typologies and domain-dependent typologies.

One of the typologies associated to the first category is the one proposed by Niskanen [6]. He proposes a typology of the non-precision of the data. He defines four concepts which are uncertainty, imprecision, ambiguity and generality. Uncertainty is a concept associated with error definitions. Imprecision is a concept related to expressions having several meanings. Ambiguity appears when there are several points of view about the same subject. Generality is the multiple representation of reality according to the level of detail. Bouchon-Meunier, in her first work [3], distinguishes only two types of imperfection which are uncertainty and imprecision. Then, in her second work [7], she distinguishes a third factor of imperfection, which is the incompleteness. She defines uncertainty by the validity of the data. Imprecision is due to the vague or approximate nature of the used semantic. Incompleteness is related to the lack of data. Klir and Yuan [8] restrict their typology on only data uncertainty. The authors divide the uncertainty into two types which are fuzziness and ambiguity. Ambiguity refers to conflict and non-specificity. Smets [9] establishes a classification of data imperfection divided into three types which are imprecision, inconsistency and uncertainty. Imprecision is relative to the data content which may be vague. Inconsistency is related to conflicting data. Uncertainty is related to errors.

Several typologies are also proposed to identify data imperfection forms in a specific domain. Gershon [10] focuses on imperfection types related to information which might be provided to analysts or decision makers. This typology proposes six types which are incomplete information, inconsistency, information too complicated, uncertainty, corrupt information and the quality of the presentation. Fisher [11] proposes a typology of uncertainty related to geographic data. The author classifies these data into well or badly defined data. If the data are well defined, they are subject to uncertainty. In the other case, data imperfection is due to imprecision, ambiguity and/or incompleteness. Two types of ambiguity are recognized, namely disagreement and lack of specificity. Olteanu [12] proposes a typology based on the one proposed by Fisher [11] to classify the imperfection of a set of textual data describing ethnographic objects. Four imperfection types are

distinguished: imprecision, uncertainty, level of detail and incompleteness. Imprecision concerns the difficulty of expressing knowledge clearly and precisely. Uncertainty concerns a doubt about the validity of information. The level of detail is related to knowledge presented in several granularities. Incompleteness refers to the absence of data. Casta [13] establishes a typology of the imperfection of data related to the economic activity. It is divided into uncertainty, imprecision and error. Desjardin et al. [14] rely on the typology presented by Fisher [11] to propose a typology of imperfection adapted to the context of archeological data. They classify imperfection into uncertainty, imprecision, ambiguity and incompleteness. The uncertainty occurs when there is a doubt about the validity of knowledge. Imprecision is the difficulty in expressing the knowledge clearly. Ambiguity is the difficulty in agreeing. Incompleteness is related to missing or partial knowledge. Snoussi [15] proposes a typology of imperfection related to spatial data. The author distinguishes three types of imperfection: imprecision, inconsistency and uncertainty. Imprecision occurs when the true value is located in a defined subset of values. Inconsistency is related to conflict or inconsistency. Uncertainty is the partial knowledge about the true value of data. Sta [16] proposes several types of imperfect data during the data retrieval and data integration processes in smart cities. Four imperfection types are proposed: uncertainty, imprecision, vagueness and missing. Uncertainty reflects the lack of knowledge. Imprecision is related to non-specificity. Vagueness is related to ambiguous data. Missing information reflects the not found or incomplete data. Achich et al. [5] propose a typology of temporal data imperfection. This typology is divided into nine imperfection types of both numeric and natural language-based temporal data which are uncertainty, typing error, imprecision, missing, circumlocution, uselessness, inconsistency, incompleteness and redundancy.

We note that there is no universal typology of data imperfection. Some generic typologies correspond better to a reality than others. Most existing typologies are domain-dependent. To the best of our knowledge, there is no typology of imperfection related to data inputs in applications for Alzheimer's patients.

Most mentioned typologies share three common concepts which are imprecision, uncertainty and incompleteness. Also, there are no definitive definitions of terms used to qualify imperfect data [5]. Imperfection types are interdependent. Indeed, according to Bouchon-Meunier [7], incompleteness leads to uncertainty and imprecision can also be associated to incompleteness. Besides, according to Smets [9], imprecision always refers to incompleteness and imprecision may be the source of uncertainty. Finally, data can be subject to several types of imperfection at the same time. Indeed, according to [17], a data may be imprecise and certain, precise and uncertain or imprecise and uncertain.

2.2 Data Quality

Data inputs in the context of applications for Alzheimer's patients are subject to various types of imperfection. However, data imperfection impacts the quality of data [18]. In this section, we survey the topic of data quality.

Data quality is defined in several ways in the literature [19]. Its most cited definition is proposed by Juran [20]. It is defined as "fitting to use". According to Huang et al. [21], it is defined as "the degree to which a set of characteristics of data fulfills requirements". Other researchers define it in relation to a given domain. For example, in the geographic area, it is defined as "the precision and spatial accuracy of the data collected" [22].

Dimensions allow estimating qualitatively the data quality [23]. According to Berti-Equille [24], quality dimensions can take on meaning depending on the application field's specificities. No general agreement exists on the exact meaning of each dimension [25]. Besides, it is common to find different

terms referring to the same dimension. There is no a universal classification of data quality dimensions [23, 25]. According to [26], the classification of quality dimensions depends on the project needs, the context of the study and the orientation which the analyst wants to give to his evaluation. Quality dimensions are defined within a quality model. Each dimension can have several sub-dimensions. The sub-dimensions do not provide quantitative measures and are therefore associated with one or more metrics allowing a quantitative estimation of quality. For each metric, one or more measurement methods are provided. In some works, the terms dimension and metric are not differentiated.

In the literature, several approaches aim to identify the different quality dimensions. We classify these approaches into two categories: approaches proposing generic quality models and others proposing specific quality models related to a given domain. *(1)* The work of Wang and Strong [27] is considered as one of the main references in the data quality research area. The authors define a generic data quality model. It proposes twenty dimensions organized into four categories. For instance, the intrinsic category includes dimensions which express the natural quality of the data. It includes the following dimensions: accuracy, believability, objectivity and reputation. Another generic data quality model is the one presented by Naumann and Roker [28]. The authors identify three approaches to identify quality dimensions: semantic oriented approach (based only on the meanings of dimensions), processing oriented approach, (classifying the dimensions according to their deployment in the different stages of data processing) and objective-oriented approach (a classification of the dimensions according to the defined objectives). They define 22 dimensions such as believability, interpretability, relevancy and reputation. Batini et al. [25] survey approaches defining data quality dimensions. According to them, four dimensions, which are accuracy, completeness, consistency and timeliness, are presented as the most commonly discussed in the literature. *(2)* In the literature, there are a considerable number of approaches proposing quality models related to a given domain. For instance, Akoka et al. [29] define four quality dimensions for data integration systems. In [30], the scope is limited to the social web. The authors identify five categories, which include 42 quality dimensions. In the context of medical informatics, Liaw et al. [31] analyze 245 papers focusing on data quality. According to them, the most used dimensions are completeness, accuracy, correctness, consistency and timeliness. Juddoo and George [32] introduce a quality model in the health Big Data domain. They define eight dimensions which are accuracy, usefulness, confidence, availability, validity, completeness, consistency and reliability.

The literature provides a wide range of methodologies for data quality estimation and improvement e.g., Total Data Quality Management (TDQM) [27], Datawarehouse Quality Methodology (DWQ) [33], Data Quality in Cooperative Information Systems (DaQuinCIS) [34], Methodology for the Quality Assessment of Financial Data (QAFD) [35], Canadian Institute for Health Information methodology (CIHI) [36], Comprehensive methodology for Data Quality management (CDQ) [37], Istat [38] and Data Quality Meta DataWarehouse (DQMDW) [39]. Batini et al. [25] provide a systematic and comparative description of such methodologies. Most of these methodologies are proposed in a specific context. We believe that data subject to different types of imperfection should not be estimated using a single quality score as provided based on existing methodologies. For instance, we cannot take into account the incompleteness and uncertainty imperfection types which affect the same data based on only one quality score. For this reason, we propose our own algorithm for data quality estimation and improvement. Compared to related work, it suggests that for each type of imperfection or a set of types of imperfection, it is necessary to estimate a quality score. Based on this latter, one or more corrective action(s) must be proposed.

```
1. Identify the different data imperfection types
2. For each imperfection type or a number of imperfection types
    2.1.  Identify the associated quality dimension(s)
    2.2.  For each quality dimension
            Identify the associated sub-dimension(s)
    2.3.  For each quality sub-dimension
            Identify the associated metrics
    2.4.  Identify the measurement method(s) based on the identified metrics
    2.5.  Estimate the data quality based on the identified measurement
    method(s)
    2.6.  Analyze and propose a set of corrective actions
```

Algorithm 1. Data quality estimation and improvement algorithm.

2.3 Data Believability

In this paper, we are limited to handle imperfection types related to false data inputs. The associated quality dimension is the believability which is reviewed in detail in this section.

Believability is considered as one of the most important quality dimensions when estimating data quality [27]. Wang and Strong define the data believability as "the extent to which data are accepted or regarded as true, real and credible" [27]. In [40] and [41], the authors propose the same definition. We have extended this definition by adding the relationship with the context of use. Thus, we define it as the extent in which the data seem, *in a specific context*; as true, real and verisimilar. Estimating data believability is based on the definition of a set of dimensions, sub-dimensions and metrics. We distinguish generic approaches to estimate data believability and specific ones related to a given domain.

Only few generic approaches have been proposed to estimate data believability e.g., [42] and [43]. Lee et al. [42] propose three dimensions to evaluate data believability: *(1)* the believability of the source which is defined as the data originate from a trustworthy source, *(2)* the data believability according to common-sense standard, and *(3)* the believability based on the temporality of data which is the extent to which a data value is credible based on proximity of transaction time to valid times and derived from data values with overlapping valid times. These dimensions remain quite general [43]. They allow only qualitative estimation. No formal metrics are presented. Prat and Madnick [43] propose an approach to estimate quantitatively the data believability based on provenance metadata i.e., the origin and subsequent processing history of data. This approach uses the believability dimensions proposed by [42]. It is three folds: *(1)* definition of metrics for estimating the believability of data sources, *(2)* definition of metrics for estimating the believability of data resulting from one process run, and *(3)* estimation of data believability as a whole based on the two mentioned metrics.

Others researchers believe that estimating data believability is context-dependent. Indeed, several approaches have been proposed, especially in the social media context. Shankaranarayanan et al. [44] represent a model to estimate the data believability in social media. They propose two main dimensions named "source credibility" and "domain expertise of data consumer". The first dimension is based on three sub-dimensions which are "identity", "expertise" and "reputation". In [45], the authors estimate the believability of characters in interactive narrative using the following dimensions: "behavior coherence", "change with experience", "awareness", "behavior understandability", "personality", "visual impact", "predictability" and "social and emotional expressiveness". Saikaew and Noyunsan [46] evaluate the believability of the Facebook posts by means of eight dimensions: "number of likes", "number of comments", "number of shared posts",

260   "number of contained links", "number of pictures", "number of hashtags", "number of videos" and "availability of geographical metadata". Nilsson and Alserud [47] propose a study to estimate the believability of information sources in social media based on the following dimensions: "identity", "reputation" and "domain expertise". Reuter et al. [48] explore also the believability of content in social media. In [19, 49], the authors share the view that information source is a critical factor which
265   affects the believability. To the best of our knowledge, there is no approach which aims to estimate the believability of data inputs in the context of applications for Alzheimer's patients.

3.   Our Typology of Data Inputs Imperfection in Applications for Alzheimer's Patients

Based on Algorithm 1, the first step in estimating and ameliorating the quality of data inputs in the context of applications for Alzheimer's patients consists of determining the different types of
270   imperfection that may affect these data. Data may be given by patients or their caregivers. In this section, we introduce a typology of data inputs imperfection given in the context of these applications. We are inspired from existing typologies and real examples collected from evaluations done in the context of some of our previous works. We distinguish direct and indirect data inputs imperfection types. The direct ones are deduced from the given data inputs e.g., uncertainty and confusion. The
275   indirect types are deduced from the direct ones. For instance, the inconsistency can be generated from the uncertainty, confusion, typing error and wrong knowledge. Figure 1 shows this typology.
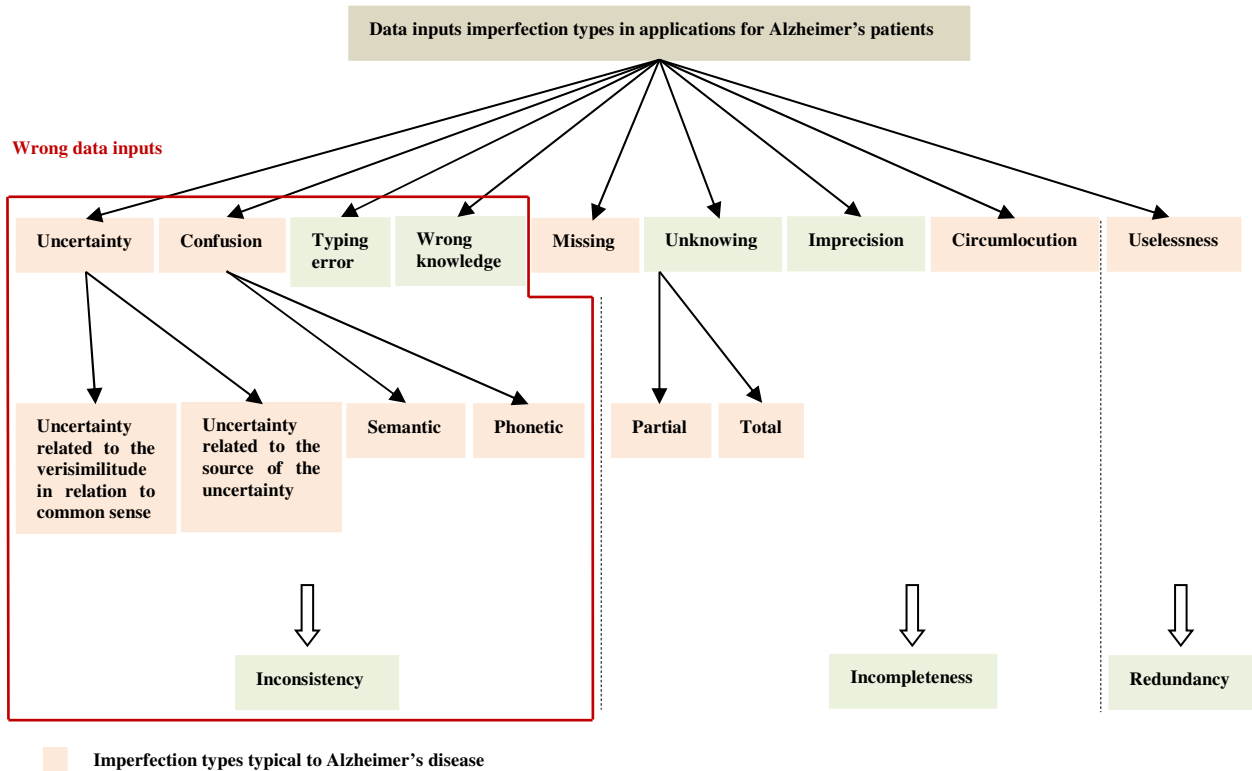

280   **Figure 1.** Data inputs imperfection types in applications for Alzheimer's patients.

3.1 Direct Data Inputs Imperfection Types

We distinguish nine direct imperfection types which may affect data inputs in applications for Alzheimer's patients: "uncertainty", "confusion", "typing error", "wrong knowledge", "missing", "unknowing", "imprecision", "circumlocution" and "uselessness". We detail them one by one and we illustrate them by some examples.

"Uncertainty": We use the same definition as the one proposed by Smets [4]. Each data input is either true or false. Uncertainty has two classifications. *(1)* The first classification is related to the verisimilitude of the data input in relation to common sense. Firstly, this classification includes uncertain data inputs which respect common-sense rules. These data can be either true or false. For instance, if the patient enters that his father is Olivier, a 95-year-old man with the same family name as the patient. This is verisimilar; but it is false. Secondly, this classification also includes uncertain data inputs which are not verisimilar compared to general common-sense rules. Subsequently, these data are false. For example, the patient enters "My father is Anna". This data input is false as Anna is a woman. *(2)* The second classification is related to the source of the uncertainty (uncertainty related to the reliability of the author and uncertainty of the author when entering the data input). The uncertainty related to the reliability of the author includes the data inputs given by Alzheimer's patients. These data are uncertain as Alzheimer's disease affects the mental faculties of the patients. The uncertainty related to the author when entering the data includes data inputs given by users (Alzheimer's patients or their caregivers) saying that they are not sure of their veracity. For example, the patient enters "I am not sure that her daughter's name is Maya" and "I think that my niece has two children".

Finally, we note that the veracity of the data inputs depends on the time factor. Indeed, a data input can be true for a given period of time and false for another period. For instance, interpersonal relationships change over time. Take the example of the data input provided by the patient "Pierre is my neighbor". It is false if Pierre moves to another house. Several other examples can be cited e.g., "I am going to visit my sister during this week" and "We are celebrating Stephen's birthday this Monday".

"Confusion": Data entered by Alzheimer's patients may be subject to confusion which may be semantic or phonetic. Semantic confusion is related to using a word which is semantically close to the wanted term. For example, the patient enters "pear" instead of "apple". Phonetic confusion (occurring at a more advanced stage) is related to using a word which is phonetically close to the desired term. For example, the patient enters "feet" instead of "meat".

"Typing error": Data entered by Alzheimer's patients or their caregivers may be subject to an unintentional error. For instance, the patient may enter a wrong data input due to a typo. He may type "1993" instead of "1939" or "19088" instead of "1908" for the year of birth. He may also write "François" instead of "Françoise".

"Wrong knowledge": In good faith, authors can make a mistake and believe that the data input is true when it is false. This type of imperfection is not related to Alzheimer's patients. Indeed, everyone can be wrong.

"Missing": Data entered by Alzheimer's patients may be subject to problems related to failing to remember. Indeed, memory impairments are the main symptom of Alzheimer's disease. Missing can be either total or partial. Total missing includes data inputs which the patients completely forget. For example, the patient enters "I forget when I visited my brother". In this example, the missing is total. He forgets totally temporal data. As a second example, he enters "It was in March but I forget the year". In this example, the missing is partial. He remembers the month. However, he forgets the year.

"Unknowing": Alzheimer's patient or a caregiver's knowledge fields do not obviously cover all the requested data inputs. For example, the patient's brother does not necessarily know all information

related to his brother's colleagues. Another example, the patient does not enter the dates of birth of his grandparents, his aunt and his doctor. We cannot bind the fact that he suffers from memory problems to the fact that he does not enter these data. This problem is not typical for Alzheimer's patients, since even young and healthy people do not necessarily know all these data.

"Imprecision": Data inputs given by the Alzheimer's patients or their caregivers may be imprecise. However, this imperfection type is accentuated with people suffering from Alzheimer's as they use more generic words. The increasing use of generic words leads to a vague discourse [50]. For instance, the patient enters "John was married to Béatrice from early 2000s to by 2016", two measures of imprecision are involved. On the one hand, the temporal data "early 2000s" is imprecise in the sense that it could mean, from 2000 to 2004; on the other hand, the temporal data "by 2016" is imprecise in the sense that it could mean from 2014 to 2016. Other examples can be cited: "John married to Maria just after he was graduated with a PhD" and "I moved to Nantes when I was young".

"Circumlocution": Data inputs given by Alzheimer's patients may be subject to a circumlocution. Indeed, the Alzheimer's patients suffer from a decline in their denominational capacities. For example, the patient enters "Our neighbor who lives on the third floor" instead of "Mrs. Ledoux". Another example, he enters "Maya is my son's daughter" in reference to "Maya is my granddaughter". As a last example, he enters "The first day of the week" instead of "Monday".

"Uselessness": This imperfection type means that the given data input is useless. It does not any added value and it is not managed by the application. For instance, a data input is useless if it can be inferred from the previously entered ones.

A data input can be subject to several imperfection types at the same time. For example, the Alzheimer's patient enters that "the name of my caregiver starts with an 'S'. It is maybe Sophie or Sylvie". This statement presents four types of imperfection which are uncertainty, missing (partial), confusion and imprecision.

3.2 Indirect Data Inputs Imperfection Types

We distinguish three indirect imperfection types which may affect data inputs in applications for Alzheimer's patients: inconsistency, incompleteness and redundancy. These imperfections are deduced from the direct imperfection types. They are considered as an obvious consequence.

"Inconsistency": To define this imperfection type, we are based on the definition proposed by [17]. According to the authors, inconsistency refers to "the existence of contradictory data on the same object". Inconsistency can be generated from four direct imperfection types: uncertainty, confusion, typing error and wrong knowledge.

"Incompleteness": To define the incompleteness imperfection type, we are based on the definition proposed by [7]. It is related to data lack. Incompleteness can be generated from four direct imperfection types: missing, unknowing, imprecision and circumlocution. For instance, the patient enters "My grandson will come next month"; "next month" is imprecise and so incomplete. It could be any day of the month.

"Redundancy": Redundancy is normal in speech, but not in a knowledge base, due to the inconsistency that may result in case of update and difficulty in data visualization. For instance, the patient enters "A long time ago, from many years, my grandparents dead". In this data input, the user indicates twice the temporal data input which is useless and redundant. Uselessness leads to redundancy.

3.3 Finer Level of Granularity

The proposed typology depends on the user's profile. A first example concerns the "spouse" relationship which may be defined between a man and a woman, a man and a man, a woman and a woman or a man and several women. Indeed, if the patient living in Europe enters that he is married to two women at the same time; this data input is obviously false. However, it may be true if the patient lives in a country allowing this type of relationship (for example: Gulf countries). Other examples related to some relationships which exist only in the same countries may be cited such as the "Godmother" and "surrogate mother" relationships. As a last example, if the patient enters "I will visit the museum next weekend"; "next weekend" is an imprecise data input. It could be "Saturday" or "Sunday" if the patient lives in Europe, for example. The weekend differs from the one in the Arab World which is "Friday" and "Saturday".

Based on Algorithm 1, we cannot estimate and ameliorate the quality of data inputs subject to several types of imperfection based on only one quality score. In this paper, we are limited to handle only five imperfection types related to false data inputs (uncertainty, confusion, typing error, wrong knowledge and inconsistency). The associated data quality dimension is the believability.

4. Our Approach for Handling False Data Inputs in Applications for Alzheimer's Patients

In this section, we propose an approach, called DBE_ALZ, to handle false data inputs in the context of applications for Alzheimer's patients. Data may be given by Alzheimer's patients or their caregivers. Based on the proposed typology, the falsity of these data is due to five imperfection types: uncertainty, confusion, typing error, wrong knowledge and inconsistency. To deal with this issue, a process assigning to each data input a believability degree should be added to these applications. Based on Algorithm 1, the proposed approach is two folds. *(1)* Firstly, we propose a believability model allowing a qualitative estimation of the believability by defining a set of dimensions and sub-dimensions. *(2)* Secondly, based on this model, we propose a set of metrics allowing a quantitative estimation of the believability. The measurement methods are based on probability and fuzzy set theories to reason about uncertain and imprecise knowledge (Bayesian networks and Mamdani fuzzy inference systems). Three languages are supported: English, French and Arabic. Based on the generated believability degrees, a set of corrective actions are proposed to ameliorate the quality of the data inputs e.g., considering only data having high believability degrees. A believability degree associated to a given data input may be updated based on new data inputs.

4.1 Qualitative Estimation of the Believability of Data Inputs in Applications for Alzheimer's Patients

Based on the established state of the art and the specificities of applications for Alzheimer's patients, we propose a model to estimate qualitatively the believability of data inputs in the context of these applications; as shown in Figure 2. This model proposes two main dimensions named "data reasonableness" and "author reliability". Two dimensions which inherit from the "author reliability" dimension are also proposed: "Alzheimer's patient reliability" and "caregiver reliability". The "data input reasonableness regarding common sense" and "data input reasonableness regarding personalized assertions" sub-dimensions estimate the "data input reasonableness" dimension. The two sub-dimensions "stage of Alzheimer's disease" and "state of the moment" are used to estimate the

"Alzheimer's patient reliability" dimension. The two sub-dimensions "age" and "knowledge field" aim to estimate the "caregiver reliability" dimension.
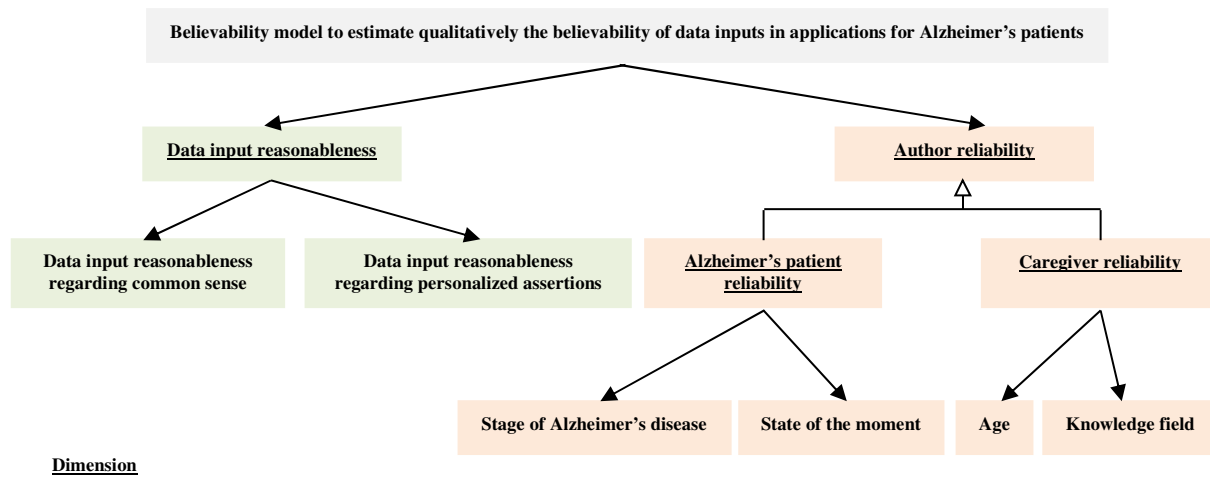

Figure 2. The proposed believability model to estimate qualitatively the believability of data inputs in applications for Alzheimer's patients.

4.1.1 Data Input Reasonableness

We define the "data input reasonableness" dimension as the verisimilitude of the data in relation to a set of verification assertions. Thanks to these assertions, we cannot assert the veracity of the data, but only its verisimilitude. However, we can assert its falsity. For example, the Alzheimer's patient enters that "Olivier is the father of Pierre". If the verification assertion concerning the age of the father and the age of the son (the age of the father must be greater than the age of the son) is not verified, we affirm the falsity of this data input. Otherwise, we cannot assert its veracity. In [42] and [43], the authors estimate the reasonableness of the data based on a set of verification assertions in relation to common sense. However, the data believability depends on the experience and preferences of the user [51]. We distinguish verification assertions related to common-sense and personalized ones. Thereafter, we propose the two following sub-dimensions to estimate the "data input reasonableness" dimension:

- "Data input reasonableness regarding common sense": This sub-dimension is defined as the verisimilitude of the data based on a set of verification assertions in relation to common sense. For example, the Alzheimer's patient enters "My wife gave birth at the age of ninety-seven". Based on the common sense-based verification assertion concerning the woman's fertility age, we affirm the falsity of this data input. As a second example, he enters "Paul and Pierre are twins. Paul was born in 2000 and Pierre was born in 2015". We affirm the falsity of this data input based on the common sense-based assertion related to the fact that the twins have the same birthday.

- "Data input reasonableness regarding personalized assertions": This sub-dimension is defined as the verisimilitude of the data based on a set of personalized verification assertions. These assertions depend on the user's private life and background. A personalized verification assertion is true to a given probability (the weight of the assertion) and under given condition(s). They depend on a number of parameters specified by an expert. In the context of the Captain Memo memory prosthesis, these assertions

depend on the culture and language. A first example concerns the first names of the father and the son. Indeed, in England, the first name of the son is very probable to be that of the father. However, most likely it is not the same in other cultures. Thereafter, the weight of the verification assertion (the father's first name is the same as that of the son) depends on the country/culture. A second example concerns the cousin relationship. Indeed, in the French language, there are, depending on the masculine or feminine gender, two terms qualify this relationship: « *cousin* » and « *cousine* ». In English, there is only one term which qualifies it: "cousin". So, we cannot associate an assertion of verification which concerns the masculine or feminine gender to verify a cousin relationship entered in English language.

4.1.2 Author Reliability

In the literature, existing approaches consider that the reliability of the source is a main dimension to estimate the data believability (some authors also use the terms "reputation", "trustworthiness" or "credibility" instead of "reliability"). In the same way, we estimate the believability of a given data input based on the "author reliability" dimension. It is defined as the reliability accorded to the person entering the data. In the context of applications for Alzheimer's patients, we identify two main authors: The Alzheimer's patients and their caregivers (e.g., son, grandson, nurse, wife and friend). Therefore, we propose two dimensions which inherit from this dimension: "Alzheimer's patient reliability" and "caregiver reliability".

4.1.2.1 Alzheimer's Patient Reliability

The "Alzheimer's patient reliability" dimension is defined as the reliability accorded to the Alzheimer's patient entering the data at the moment of entry. To estimate qualitatively this dimension, we identify the following two sub-dimensions:

- "Stage of Alzheimer's disease": The reliability of an Alzheimer's patient depends on the disease's progression. It is obvious that early-stage Alzheimer's patients are more reliable than those suffering from this disease in later stage. Indeed, this disease is characterized by a gradual decrease in mental and memory faculties. At first, it is mainly the short-term memory which is affected (forgetting the recently learned information). For example, the patient does not remember all the dates of birth of his grandchildren and he may remember the dates of birth of his wife, children and parents. The mistakes made by a first-stage Alzheimer's patient affects more recent data compared to the old ones. The disease affects the long-term memory only in late stages (e.g., history and language) [52]. Thereafter, as this disease progresses, the reliability of the Alzheimer's patient declines.

- "State of the moment": The reliability of an Alzheimer's patient depends on the state of his mental and memory faculties at the moment of data entry. Indeed, this disease is punctuated by periods of lucidity (moments of clarity) and moments of bewilderment. Sometimes, patients have stunning moments of total lucidity. For most Alzheimer's patients, periods of total lucidity alternate with moments of straying, first occasional, then increasingly frequent.

4.1.2.2 Caregiver Reliability

The "caregiver reliability" dimension is defined as the reliability accorded to the patient's caregiver that enters the data at the moment of entry. To estimate qualitatively this dimension, we identify the following two sub-dimensions:

- "Age": The reliability of the patient's caregivers depends on their ages. As a first example, we cannot fully rely on the four-year-old grandson to enter data related to his grandmother. As a second example, we cannot also fully rely on the ninety-five-year-old spouse of the patient to remember all the necessary data related to her husband. She suffers very probably from memory impairment's related to the normal aging process. However, we can rely relatively on the thirty-year-old son of the patient to enter the required data related to his father.

- "Knowledge field": The patient's caregivers do not necessarily know all information related to the patient's private life. For example, we can rely on the patient's wife to enter data concerning the ascendants and descendants of his husband. However, we do not rely on her to enter data about his childhood friends. Another example, the friend of the patient knows some data related to their friends in common. However, he does not necessarily know, for example, all his family members and colleagues. The reliability of the patient's caregiver depends on his knowledge fields.

4.2 Quantitative Estimation of the Believability of Data Inputs in Applications for Alzheimer's Patients

In this section, we detail our work to estimate quantitatively the believability of data inputs in the context of applications for Alzheimer's patients; as shown in Figure 3. A degree of believability C (C $\in$ [0, 1], 0 and 1 represent, respectively, completely unbelievable and completely believable) is generated for each data input. It is based on the dimensions and sub-dimensions proposed by the believability model presented in the previous section. Three languages are supported: English, French and Arabic. We use probability and fuzzy set theories. Three main modules are proposed: "data input reasonableness estimation", "author reliability estimation" and "data input believability estimation".

4.2.1 Data Input Reasonableness Estimation

This module estimates the reasonableness of a given data input. It estimates quantitatively the "data input reasonableness" qualitative believability dimension. It generates a score R (R $\in$ [0, 1], 0 and 1 represent, respectively, completely unreasonable and completely reasonable). It is based on the probability theory to deal with uncertainty. Precisely, we use the Bayesian Network model. It is composed of three main sub-modules: "data input pattern matching", "verification rule pattern generation" and "verification rule fulfillment".
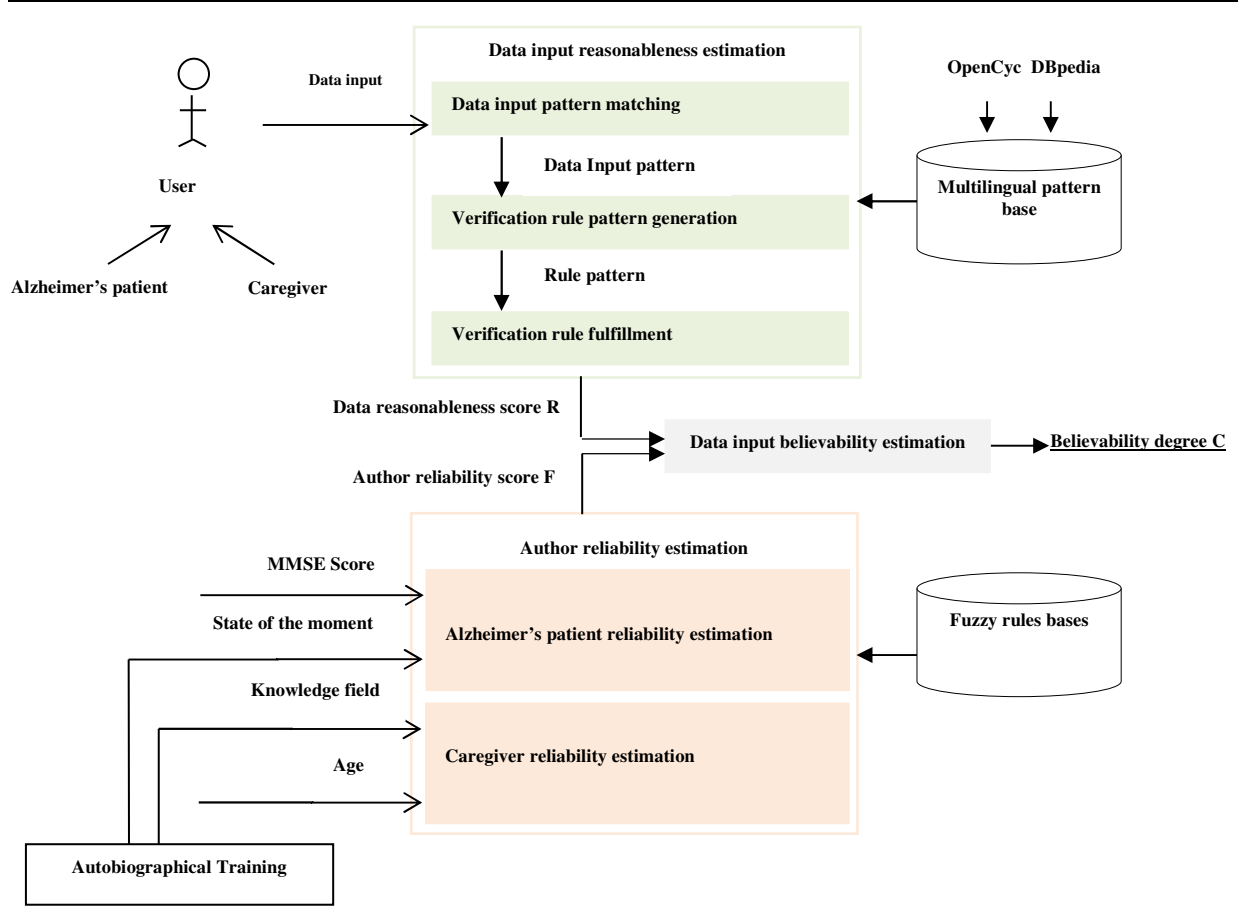
Figure 3. Quantitative estimation of the believability of data inputs in applications for Alzheimer's patients.

4.2.1.1 Data Input Pattern Matching

This sub-module takes as input the given data input I_K. It returns the associated pattern P_{IK} thanks to a multilingual semantic similarity-based matching process with the pre-established data input patterns base $P_I = \{P_{I1}, P_{I2}\ldots P_{IN}\}$. For instance, if we have the following data inputs: "Philippe is the father of Pierre", "Philippe is the papa of Pierre", "Philippe daddy Pierre" or "Philippe dad Pierre". The corresponding pattern is "Person X father Person Y". As a second example, if the patient enters "Maya is my cousin". The associated pattern is "Person X cousin Patient".

To determine the data input pattern, we estimate the semantic similarity scores between the data input and all proposed data input patterns. The associated pattern is the one having the highest score. We use our previous approach [53] which allows estimating the semantic similarity between two sentences. It supports three languages: English, French and Arabic. Compared to related work, it takes into account the semantic arguments notably the semantic class and thematic role. This approach is divided into three main steps named "preprocessing", "similarity score attribution" and "supervised learning". The preprocessing step allows determining the data input's language, tokenization, lemmatization and removing punctuation signs and stop words. We enrich this step by using the NOOJ platform [54] to determine the data input's named entities (persons, organizations and locations). The second step aims to estimate the similarity rating between the data input and the given pattern. Three similarity levels are measured: lexical, semantic and syntactico-semantic. The lexical similarity is computed using the lexical units which compose the data input and the given pattern to

extract the words which are lexically the same. The Jaccard coefficient is used [55]. The semantic similarity is computed using the semantic vector, Jaccard coefficient and Cosine similarity metric [56]. This measurement is reinforced by means of WordNet database [57] to extract the synonyms of each word. The estimation of the syntactico-semantic similarity consists of extracting the characteristics of the semantic arguments of the data input and the given pattern from the VerbNet database for the French and English languages [58] and LMF Arabic dictionary for the Arabic language [59]. The third step aims to use the automatic learning to define the appropriate coefficients for the measures described in the second step.

4.2.1.2 Verification Rule Pattern Generation

This sub-module takes as input the data input pattern P_{IK} and returns the associated verification rule pattern P_{RK}. $R_P = \{R_{P1}, R_{P2}\ldots R_{PN}\}$ is a verification rule patterns base. For each data input pattern I_{PK}, a rule pattern P_{RK} is associated. A verification rule pattern is defined as the following:

$$IF\ P_{IK}\ THEN\ A_{K/1}\ldots A_{K/N}$$

The resulting part consists of a conjunction and/or disjunction of one or more verification assertions $A_{K/I}$ which ought to be fulfilled to confirm that I_K is reasonable. Based on the proposed believability model, the data input reasonableness is estimated based on common sense-based verification assertions and personalized ones. *(1)* Common sense-based assertions estimate the data input reasonableness regarding common-sense standards. To define these assertions, we use the OpenCyc[3] ontology (version 4.0) which proposes pieces of knowledge composing human common-sense (for instance, a person has only one biological mother and one biological father). The OpenCyc knowledge base contains about 239,000 concepts and 2,093,000 facts. It is connected to Wikidata[4]. *(2)* Personalized assertions estimate the data input reasonableness based on user's background and private life. We associate a weight $W_{K/I}$ to each personalized assertion $A_{K/I}$ to estimate its validity according to the associated parameters ($W_{K/I} \in [0, 1]$, 0 and 1 represent, respectively, completely invalid and completely valid). These weights are determined by interrogating, via a set of SPARQL queries, Linked Open Data datasets e.g., DBpedia[5] and Freebase[6].

An assertion $A_{K/I}$ is defined based on one of the following two patterns:

$$P_{A1} = \{(V_1)_{C1},\ OP,\ (V_2)_{C2}\ ;\ W_{AI}\}$$

$$P_{A2} = \{(V_1)\ _{C1},\ OP,\ C\ ;\ W_{AI}\}$$

$(V_I)_{CI}$ represents a variable already saved in the knowledge base of the patient. Its believability degree is CI which was determined based on the proposed approach. C is a constant value. OP is an operator such as $=, \neq, <, >, \leq, \geq$ and $\in$. W_{AI} is the corresponding weight of the assertion (It is obvious that a verification assertion relating to the common-sense has a weight equal to 1).

Taking the example of the paternity relationship related to patients living in Canada, mentioned in the last subsection, the corresponding verification rule pattern is the following:

IF ("Person X father Person Y") THEN ({Age (X) C1 > Age (Y) C2; WA1= 1} ∧ {Gender(X) C3 =

---

[3] http://www.cyc.com/opencyc/
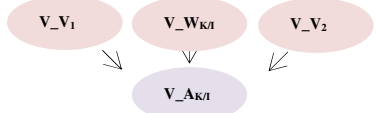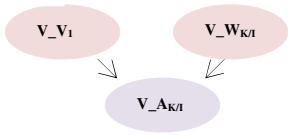[4] www.wikidata.org
[5] http://dbpedia.org/
[6] http://freebase.com/

"Man"; WA2 = 1} ⋀ {Last name(X) C4 = Last name(Y) C5; WA3= 0.99} ⋀ {Nationality(X) C6 = Nationality(Y) C7; WA4= 0.99} ⋀ {First name(X) C8 <> First name(Y) C9 WA5 = 0.99})

4.2.1.3 Verification Rule Fulfillment

This sub-module takes as input the verification rule pattern P_{RK} and returns the associated data input reasonableness score R. The verification rule pattern presents two sources of uncertainty. First, each personalized assertion is uncertain. Its certainty is equal to its associated weight. Second, each assertion refers to at least one variable V_I which represents a data stored in the knowledge base of the patient. Its certainty is equal to the believability degree C_I. We use Bayesian networks to reason about uncertain facts.

For each assertion $A_{K/I}$, a Bayesian network pattern $BN_{K/I}$ is associated to determine a score representing the extent to which the assertion is fulfilled. It is determined based on the pattern of the assertion, as shown in Table 1.

Table 1. Bayesian network patterns associated to assertion patterns.

Assertion pattern	Associated Bayesian network pattern
$P_{A1} = \{(V_1)_{C1}, OP, (V_2)_{C2} ; W_{AI}\}$	
$P_{A2} = \{(V_1)_{C1}, OP, C ; W_{AI}\}$	

V_V_I, $V\_W_{K/I}$ and $V\_A_{K/I}$ are three probabilistic variables. They represent, respectively, a data stored in the knowledge base of the patient, the assertion's weight and the probability of fulfilling the assertion.

For each verification rule pattern P_{RK}, we associate a Bayesian network pattern BN_K. It is formed by the N Bayesian networks $BN_{K/I}$ and a probabilistic variable R. This variable depends on the probabilistic variables $V\_A_{K/I}$, as shown in Figure 4.
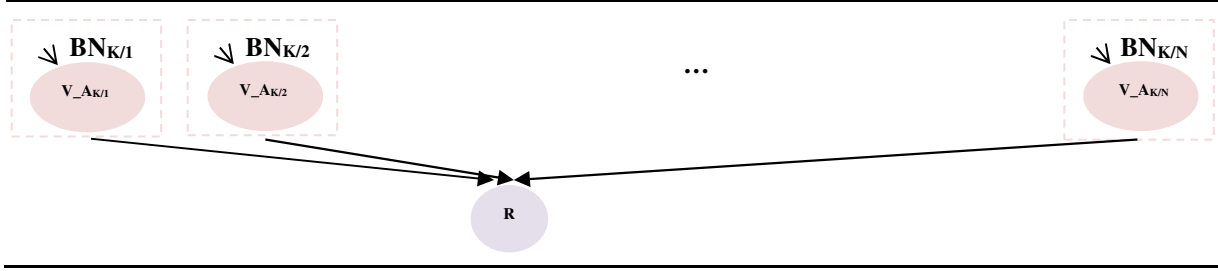
Figure 4. Bayesian network pattern associated to verification rule pattern.

Finally, we instantiate the Bayesian network pattern BN_K based on the data input I_K and the data
saved in the knowledge base of the patient. Taking the mentioned example of the paternity
relationship, the corresponding Bayesian Network pattern is shown in Figure 5.
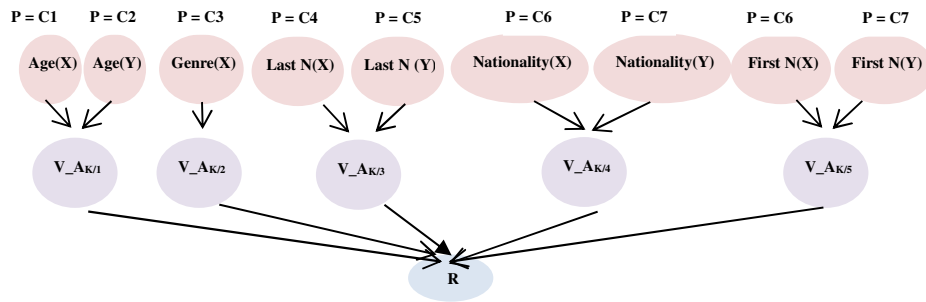

Figure 5. An example of a Bayesian network pattern.

4.2.2 Author Reliability Estimation

This module estimates the reliability of the data input's author (an Alzheimer's patient or a caregiver)
at the moment of entry. It estimates quantitatively the "author reliability" qualitative believability
dimension. It generates a score F ($F \in [0, 1]$, 0 and 1 represent, respectively, completely unreliable and
completely reliable). It is composed of two sub-modules: "Alzheimer's patient reliability estimation"
and "caregiver reliability estimation". The first sub-module is activated if the data input is given by
the patient and it aims to estimate quantitatively the quality dimension "Alzheimer's patient
reliability". The second one is activated if the data input is given by a person from the patient's
surrounding and it aims to estimate quantitatively the quality dimension "caregiver reliability". These
two sub-modules are based on the fuzzy set theory to deal with imprecision. Precisely, we use the
Mamdani fuzzy inference system. We use the gravity center defuzzification method. For the rest of
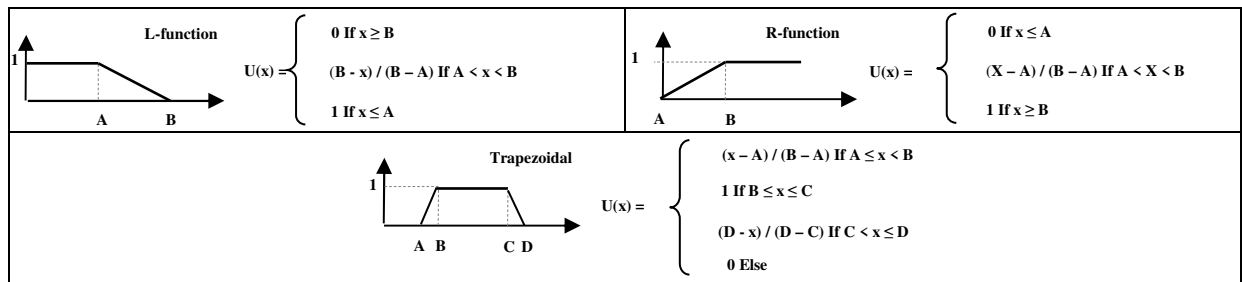this paper, we use the membership functions defined in [60] and shown in Figure 6.

Figure 6. L-function, R-function and Trapezoidal membership functions [60].

4.2.2.1 Alzheimer's Patient Reliability Estimation

Based on the believability model, estimating the Alzheimer's patient reliability is based on two sub-dimensions: "stage of Alzheimer" and "state of the moment". We associate to each sub-dimension a metric in order to estimate it quantitatively.

To determine the stage of Alzheimer, we associate a metric called "MMSE score". It is based on the Mini Mental State Examination (MMSE). It contains items assessing the orientation (time and place), memory recall, attention and calculation, object naming, verbal registration, language, and visuospatial/constructional performance. It is a 30-points questionnaire, which includes 12 questions. It is used extensively in clinical and research settings to measure cognitive impairments. Scores range from 0 to 30, with higher scores indicating better performance. According to [61], MMSE better than 20 means "mild stage", MMSE between 10 and 19 means "moderate stage", MMSE below than 10 means "severe stage" and very low MMSE means "terminal stage".

The memory faculties of an Alzheimer's patient depend on the actual moment. Sometimes, patients have stunning moments of total lucidity. To estimate the momentary memory capabilities, we propose a metric called "state of the moment score". To determine it, we use our earlier work [62]. It is a "Question and Answer" training, named "Autobiographical Training", which aims to refresh the patient's memory. It does not use general knowledge facts or false examples, but it uses the patient's private life as a knowledge source input. The generated questions are based on information that the patient introduced before. This training supports three languages: English, French and Arabic. Before entering data, the patient is asked to answer a set of questions generated by this training. A percentage representing the correct answers ("Successful Score") is calculated. Based on the last one, we can judge the memory state of the patient. A higher score indicates better momentary memory capabilities.

The metrics associated to the "stage of Alzheimer" and "state of the moment" dimensions are imprecise. To deal with imprecision, we implement this module as a Mamdani fuzzy inference system. It takes as input two fuzzy variables related to the two mentioned metrics, named "Alzheimer_Stage" and "Momentary_State_Memory", and returns the patient reliability score F based on the output fuzzy variable "Patient_Reliability". All the values presented in the following are validated by a neurologist doctor. The variable "Alzheimer_Stage" related to the "MMSE score" metric has four linguistic labels {Terminal_Stage, Severe_Stage, Moderate_Stage and Mild_Stage}. "Terminal_Stage" has the L-function membership function which has as parameters A = 2 and B = 5. "Severe_Stage" has the Trapezoidal membership function which has as parameters A = 2, B = 5, C = 8 and D = 12. "Moderate_Stage" has the Trapezoidal membership function which has as parameters A = 8, B = 12, C = 18 and D = 22. "Mild_Stage" has R-function membership function which has as parameters A = 18 and B = 22. The variable "Momentary_State_Memory" is related to the "state of the moment score" metric. It represents the "Successful Score" returned by the "Question and Answer" training. It has the following linguistic labels: {Confused, Average and Well-Remembered}. "Confused" has the L-function membership function which has as parameters A = 20 and B = 40. "Average" has the Trapezoidal membership function which has as parameters A = 20, B = 40, C = 60 and D = 80. "Well-Remembered" has the R-function membership function which has as parameters A = 60 and B = 80. The output fuzzy variable "Patient_Reliability" represents the reliability of the patient. It has the following linguistic labels: {Not_Reliable, Reliable and Very_Reliable}. "Not_Reliable" has the L-function membership function which has as parameters A = 0,2 and B = 0,4. "Reliable" has the Trapezoidal membership function which has as parameters A = 0,2, B = 0,4, C = 0,6 and D = 0,8. "Very_Reliable" has the R-function membership function which has as parameters A = 0,6 and B = 0,8. The generated score is based on a pre-established fuzzy rules base.

Example: The patient's MMSE score is 27. The score representing the percentage of correct answers generated by "Autobiographical Training" is 86%. The score representing the patient's reliability is 0,81.

4.2.2.2 Caregiver Reliability Estimation

Based on the proposed believability model, estimating the patient's caregiver reliability is based on two sub-dimensions: "knowledge field" and "age". We associate to each sub-dimension a metric in order to estimate it quantitatively.

The patient's caregivers do not necessarily know all information related to the patient's private life. To estimate the "knowledge field" dimension, we associated a metric named "knowledge field score". To estimate it, we also use our earlier work "Autobiographical Training". The caregiver is asked to give responses of a set of questions related to the data input's field. After responding all questions, a percentage of the correct answers is calculated ("Successful Score"). Based on this score, we estimate the extent to which the caregiver is reliable in this specific field. A higher score indicates a better reliability.

The reliability of the caregivers depends on their ages. We propose a metric called "age score" to estimate quantitatively the "age" dimension. We consider that all caregivers under the age of six and over ninety are unreliable. Users between the ages of thirty and sixty are very reliable. These ages are proposed by a neurologist doctor.

The metrics associated to the "knowledge field" and "age" dimensions are imprecise. To deal with imprecision, we implement this module as a Mamdani fuzzy inference system. It takes as input two fuzzy variables related to the mentioned metrics, named "Knowledge_Field" and "Reliability_Age", and returns the caregiver reliability score F based on the output fuzzy variable "Caregiver_Reliability". All the values presented in the following are validated by a neurologist doctor. The variable "Knowledge_Field" related to the "knowledge field score" metric has the linguistic labels {Weak_Knowledge_Field, Good_Knowledge_Field and Very_Good_Knowledge_Field}. "Weak_Knowledge_Field" has the L-function membership function which has as parameters A = 20 and B = 40. "Good_Knowledge_Field" has the Trapezoidal membership function which has as parameters A = 20, B = 40, C = 60 and D = 80. "Very_Good_Knowledge_Field" has the R-function membership function which has as parameters A = 80 and B = 100. The variable "Reliability_Age" related to the metric representing the "Age" dimension has the linguistic labels {Age_Reliable and Age_Not_Reliable}. "Age_Reliable" has a personalized membership function (1 IF age ∈ [0, 6] or age >= 90; - (1/24) * age + (30/24) IF age ∈]6, 30] and (1/30) * age - 2 IF age ∈ [60, 90[). "Age_Not_Reliable" has the Trapezoidal membership function which has as parameters A = 6, B = 30, C = 60 and D = 90. The output fuzzy variable "Caregiver_Reliability" represents the same linguistic labels as the output fuzzy variable "Patient_Reliability". The generated score is based on a pre-established fuzzy rules base.

Example: The age of the caregiver is 32 years old. The score representing the percentage of the correct answers to the questions generated by Autobiographical Training is 95%. The reliability of this caregiver is 0,96.

4.2.3 Data Input Believability Estimation

This module takes as inputs the outputs of the two other modules. It returns the believability degree C of a given data input. The author reliability score F and the data reasonableness score R are imprecise.

We implement this module as a Mamdani fuzzy inference system that takes as input two fuzzy variables related to these scores and a fuzzy output variable named "Believability_Degree". The variable "Reasonableness_Score" related to the data input reasonableness score has the following linguistic labels: {Not_Reasonable, Reasonable and Very_Reasonable}. "Not_Reasonable" has the L-function membership function which has as parameters A = 0,2 and B = 0,4. "Reasonable" has the Trapezoidal membership function which has as parameters A = 0,2, B = 0,4, C = 0,6 and D = 0,8. "Very_Reasonable" has the R-function membership function which has as parameters A = 0,8 and B = 1. The second fuzzy input variable corresponds to the fuzzy output variable of the inference system associated to the "author reliability estimation" module. The variable "Believability_Degree" related to the data input believability has the following linguistic labels: {Not_Believable, Believable and Very_Believable}. "Not_Believable" has the L-function membership function which has as parameters A = 0,2 and B = 0,4. "Believable" has the Trapezoidal membership function which has as parameters A = 0,2, B = 0,4, C = 0,6 and D = 0,8. "Very_Believable" has the R-function membership function which has as parameters A = 0,8 and B = 1. A fuzzy rules base is pre-established.

5.   Validation

A Java-based prototype is implemented based on the DBE_ALZ approach. It uses jFuzzyLogic [63] (for implementing industry standards related to fuzzy logic), JavaBayes [64] (for implementing Bayesian networks), JENA[7] (for manipulating ontologies) and SPARQL-DL[8] API (for querying ontologies). Then, we propose a semantic representation which may be added to a given ontology to represent the believability degrees associated to a given data and their progression in time. Finally, we extend the PersonLink ontology by this semantic representation and we integrate the DBE_ALZ-prototype in the Captain Memo memory prosthesis to handle false data inputs.

5.1 Representing the Believability Degrees and their Progression in Ontology

In this section, we propose a semantic representation which can be added to a given ontology to represent the believability degrees associated to a given data and their progression in time.

We introduce a class named "Believability" to represent a believability degree associated to a given data input at a given date. It has two datatype properties. The first one, named "Has_Believability_Degree", represents the believability degree estimated based on the DBE_ALZ-based prototype. The second one, named "Has_Time", represents the associated date. The believability degree associated to a given data input may be updated based on new data inputs. Thus, for each update of the believability degree, a new instance of the class "Believability" is created.

To represent a stored data ("Value") structured using a datatype property named "Data_Property" related to an instance named "Instance_Value" and its associated believability degrees, we propose the semantic representation presented in Figure 7. We propose a class named "Data_Type". This class has a datatype property named "Has_Value" representing the saved data "Value". The original property "Data_Property" connects the "Instance_Value" instance (domain) with an instance of the "Data_Type" class (range). We propose an object property, named "Has_Believability". It connects the instance of the "Data_Type" class (domain) with an instance of the "Believability" class (range).

---

[7] https://jena.apache.org/
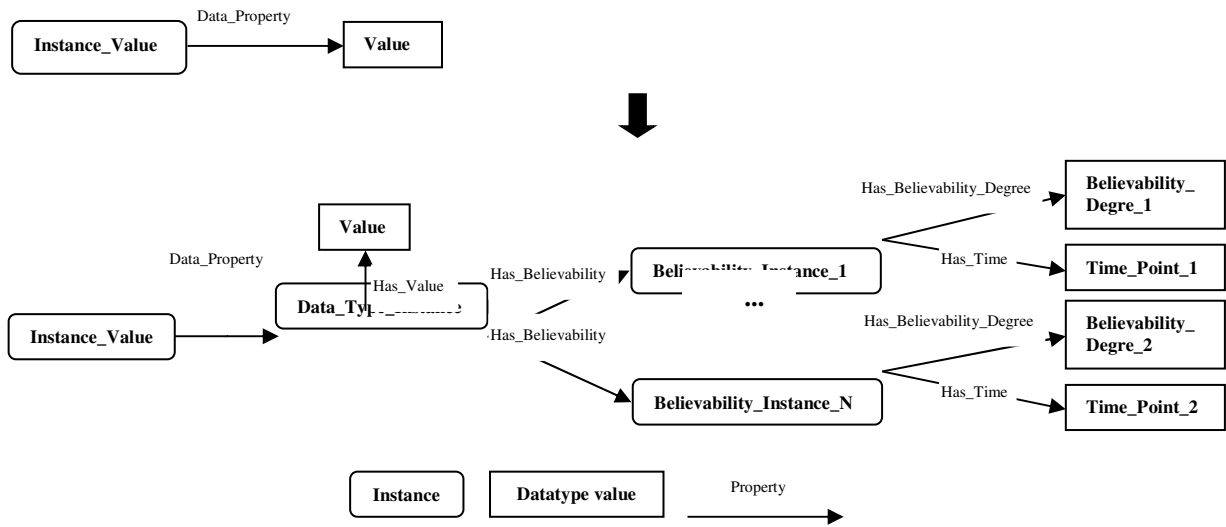[8] http://www.derivo.de/en/resources/sparql-dl-api.html

Figure 7. Representing believability degrees associated to a datatype property value.

To represent an object property named "Object_Property" connecting two instances named "Instance_Link_1" (domain) and "Instance_Link_2" (range) and its associated believability degrees, we propose the semantic representation presented in Figure 8. We introduce a class named "Believability_Link". For each object property, we associate two instances of this class named "Believability_Link_1" and "Believability_Link_2". The "Object_Property" links these two instances. We propose an object property, named "Is_Linked", to connect "Instance_Link_1" with "Believability_Link_1" and "Instance_Link_2" with "Believability_Link_2". The two instances of the "Believability_Link" class are related to an instance of the "Believability" class using the object property "Has_Believability".
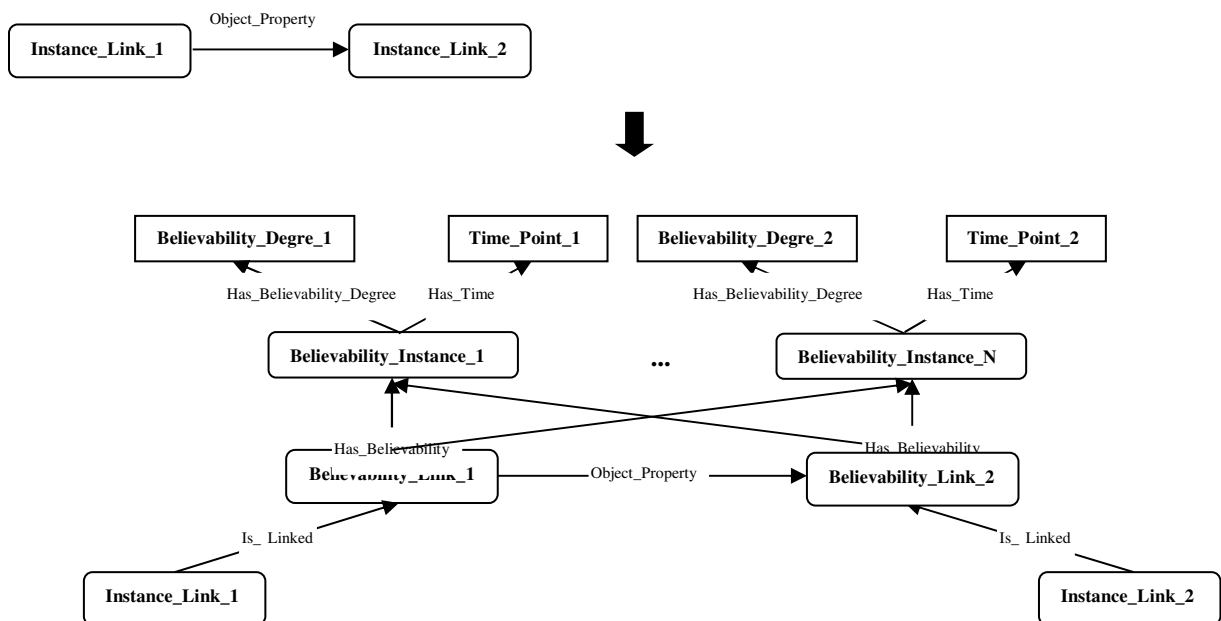

Figure 8. Representing believability degrees associated to an object property.

5.2 Application to the Captain Memo Memory Prosthesis

We integrate the DBE_ALZ-based prototype in the prototype of the Captain Memo memory prosthesis to handle false data inputs in the context of the PersonLink ontology.

We extend the PersonLink ontology based on the semantic representation detailed in section 5.1 to represent the believability degrees associated to each stored data. Let's take the following example: "In June 05th 2019, the patient enters "Maria is 5 years old". The associated believability degree is 0,46. In June 29th 2019, he enters "Maria is the wife of Sam". The associated believability degree is 0,91. Based on this data input, a verification assertion related to the marriage's age is activated. The believability degree associated to the age of Maria is updated to become 0,26. In December 02th 2019, the patient enters "Maria is 36 years old". The associated believability degree is 0,85". In PersonLink, the person's age is represented as a datatype property named "Has_Age" associated to the "Person" class. The spouse relationship is represented as an object property named "Wife_Of" connecting two instances of the "Person" class. Figure 9 shows the semantic representation of these data associated to their believability degrees in PersonLink.
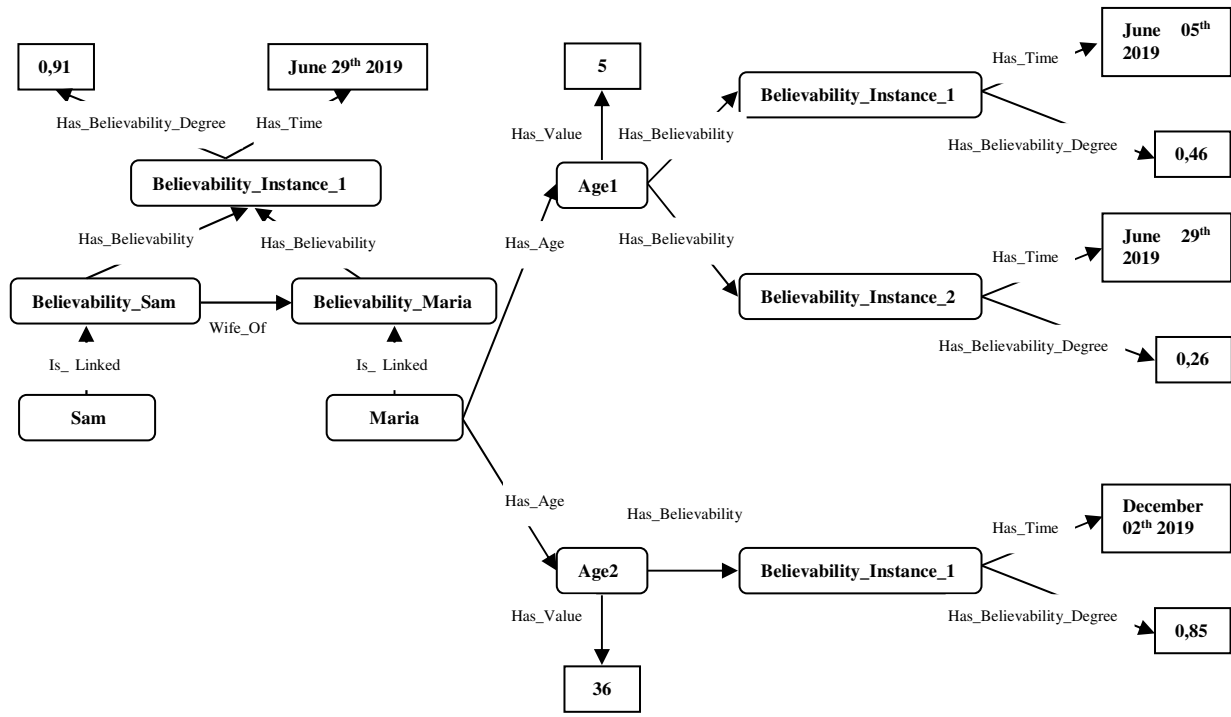

Figure 9. Representing believability degrees and their progression in time in PersonLink.

Based on the believability degrees generated by the DBE_ALZ-based prototype, a set of corrective actions are proposed to guarantee the quality of the services offered by Captain Memo. *(1)* Only data having a believability degree greater than 0,8 are taken into account. *(2)* Our approach is useful in case of contradictory data inputs. We rely only on the data having the highest believability degree. For instance, in Figure 10, only the data inputs related to Maria having believability degrees greater than 0,8 are shown. Besides, only the data input concerning the Maria's age which has the highest believability degree is shown (36 years old).
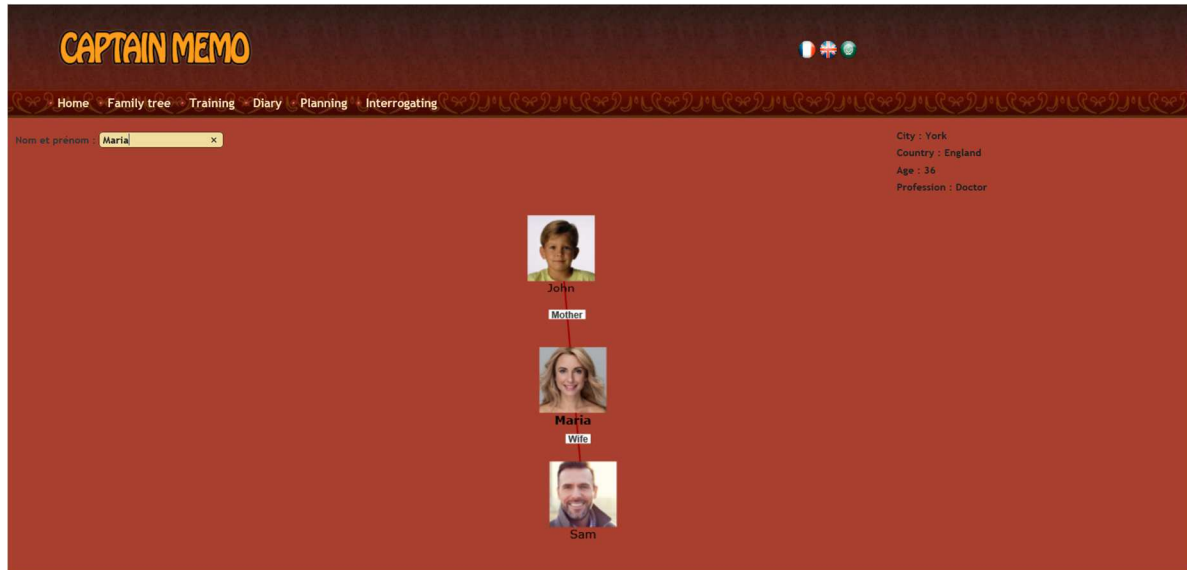
Figure 10. Taking into account only of data inputs having high believability degrees in Captain Memo.

6. Evaluation

The evaluation study was done in the context of the Captain Memo memory prosthesis. A total of 24 Alzheimer's patients P= {P_1 … P_{24}} and their associated caregivers C= {C_1 … C_{24}} were recruited to participate in this study. All caregivers are first-degree relatives e.g., son or wife. {P_1 … P_{15}} were early-stage Alzheimer's patients (MMSE score better than 20). The others were moderate stage Alzheimer's patients (MMSE score between 10 and 19). Their MMSE scores were ranged from 15 to 29 at the baseline. They were aged between 63 and 72 years old (median = 67 years). Most Alzheimer's patients were living in a nursing home in Tunisia. We asked each patient's legal sponsor for the consent letter. We excluded participants with overt behavioral disturbances, sever aphasia and sever auditory and/or visual loss.

This study was performed from July 2018 for about four months (14 weeks). The evaluation consisted of two test sessions for each patient per week. The test session duration depends on the cognitive performance of the Alzheimer's patient. At mean, it was about one hour. Each patient was asked to enter about 20 data inputs relative to his private life and background. The knowledge bases associated to the participants are structured using the PersonLink ontology.

Three scenarios are proposed:

- "Without DBE_ALZ" scenario: We do not integrate the DBE_ALZ-based prototype in the prototype of Captain Memo. All data entered by the Alzheimer's patient P_i are saved in $KB_{i/s1}$ (knowledge base corresponding to the data entered by the patient P_i based on the first scenario). $KB_{i/s1}$ are generated after 14 weeks of using Captain Memo. Each caregiver C_i is asked to identify only the true data inputs given by the patient. The last ones formed the gold standard knowledge base related to the first scenario $KB_{i/GS1}$.
- "DBE_ALZ @ 2 weeks" scenario: We integrate the DBE_ALZ-based prototype in Captain Memo. All data entered by the Alzheimer's patient P_i and having a believability degree higher than 0,8 are saved in $KB_{i/s2}$. $KB_{i/s2}$ are generated after 2 weeks of using Captain Memo (about 80 data inputs are saved in $KB_{i/s2}$). Each caregiver C_i is asked to identify only the true data inputs given by the patient. The last ones formed the gold standard knowledge base $KB_{i/GS2}$.

- "DBE_ALZ @ 14 weeks" scenario: We integrate the DBE_ALZ-based prototype in Captain Memo. All data inputs given by the Alzheimer's patient P_i and having a believability degree higher than 0,8 are saved in the knowledge base $KB_{i/s3}$. $KB_{i/s3}$ are generated after 14 weeks of using Captain Memo (about 560 data inputs are saved in $KB_{i/s3}$). Each caregiver C_i is asked to identify only the true data inputs given by the patient. The last ones formed the gold standard knowledge base $KB_{i/GS3}$.

We compare the generated $KB_{i/s1}$, $KB_{i/s2}$ and $KB_{i/s3}$ knowledge bases against the gold standard ones. We use the Precision evaluation metric. $P_{i@1}$ ($|KB_{i/s1} \cap KB_{i/GS1}| / |KB_{i/s1}|$), $P_{i@2}$ ($|KB_{i/s2} \cap KB_{i/GS2}| / |KB_{i/s2}|$) and $P_{i@3}$ ($|KB_{i/s3} \cap KB_{i/GS3}| / |KB_{i/s3}|$) represent, respectively, the Precision associated to the patient P_i according to the first, second and third scenarios. Table 2 shows the results.

Table 2. Evaluation's results.

		Precision according to the first scenario ($P_{i@1}$)	Precision according to the second scenario ($P_{i@2}$)	Precision according to the third scenario ($P_{i@3}$)
Early stage Alzheimer's patients	P_1	0,821	0,845	0,918
	P_2	0,723	0,750	0,884
	P_3	0,635	0,797	0,894
	P_4	0,751	0,814	0,958
	P_5	0,807	0,842	0,907
	P_6	0,841	0,878	0,945
	P_7	0,853	0,895	0,929
	P_8	0,810	0,837	0,917
	P_9	0,696	0,720	0,917
	P_{10}	0,694	0,814	0,923
	P_{11}	0,914	0,939	0,949
	P_{12}	0,896	0,912	0,933
	P_{13}	0,848	0,891	0,953
	P_{14}	0,817	0,845	0,917
	P_{15}	0,846	0,862	0,946
Mean (only early-stage Alzheimer's patients)		**0,796**	**0,842**	**0,926**
Moderate stage Alzheimer's patients	P_{16}	0,65	0,795	0,958
	P_{17}	0,621	0,752	0,881
	P_{18}	0,675	0,786	0,940
	P_{19}	0,560	0,658	0,788
	P_{20}	0,489	0,745	0,843
	P_{21}	0,532	0,658	0,908
	P_{22}	0,507	0,689	0,887
	P_{23}	0,382	0,604	0,701
	P_{24}	0,676	0,745	0,895
Mean (only moderate stage Alzheimer's patients)		**0,565**	**0,714**	**0,866**
Mean (all patients)		**0,710**	**0,794**	**0,903**

The overall means of the precision associated to the "DBE_ALZ @ 2 weeks" and "DBE_ALZ @ 14 weeks" scenarios (0,794 and 0,903) are better than the overall mean of the precision associated to the "Without DBE_ALZ" scenario (0,710). These results prove the efficiency of the proposed approach.

The overall mean of the precision associated to the "DBE_ALZ @ 14 weeks" scenario (0,903) is better than the overall mean of the precision associated to the "DBE_ALZ @ 2 weeks" scenario (0,794). This value is ameliorated as the data input reasonableness scores were improved. Indeed, the knowledge bases of the patients store more data from one navigation session to another. As a result, more fuzzy rules are activated to determine these scores.

The overall mean of the precision associated to the "Without DBE_ALZ" scenario for early-stage Alzheimer's patients (0,796) is better than the overall mean of the precision associated to the "Without DBE_ALZ" scenario for moderate stage Alzheimer's patients (0,565). Based on these results, we confirm that the reliability of Alzheimer's patients in first stage is better than those who survive into the final sub-stages of the disease process. This confirms our choice to estimate the reliability of Alzheimer's patients based on the progression of this disease.

7. Conclusion

This paper addresses the issue of imperfect data inputs in applications for Alzheimer's patients. Data may be given by Alzheimer's patients or their caregivers. Our first contribution consists of proposing a typology of imperfection of data inputs in the context of these applications. In the literature, several typologies of data imperfection have been proposed. Some are generic and others are specific to a given domain. However, to the best of our knowledge, there is no typology of imperfection of data inputs in the context of applications for Alzheimer's patients. We propose a typology of imperfection of these data which offers nine direct and three indirect imperfection types. The direct ones are identified from the given data inputs. The indirect imperfection types are deduced from the direct ones. In this paper, we are limited to handle only false data inputs which are related to five imperfection types: uncertainty, confusion, typing error, wrong knowledge and inconsistency. Our second contribution consists of proposing an approach, called DBE_ALZ, that handles false data entry by estimating the believability of each data input. The state-of-the-art shows that most existing believability models are domain-dependent. To the best of our knowledge, there is no approach which proposes dimensions or metrics to estimate the believability of data inputs in applications for Alzheimer's patients. The DBE_ALZ approach is two folds. Firstly, we propose a model which defines a set of dimensions and sub-dimensions allowing a qualitative estimation of the believability of the data inputs. The first dimension represents the data input reasonableness. Compared to related work, it is measured not only based on common-sense standard, but also based on a set of personalized rules. The second dimension aims to estimate the reliability of the Alzheimer's patients or their caregivers. To estimate the reliability of the patients, we are based on two sub-dimensions: "stage of Alzheimer" and "state of the moment". To estimate the reliability of the caregivers, we are based on two sub-dimensions: "age" and "knowledge field". Secondly, based on the proposed model, we estimate quantitatively the data input believability by defining a set of metrics associated to the proposed dimensions. We use Bayesian networks and Mamdani fuzzy inference systems to deal with uncertainty and imprecision. Three languages are supported: English, French and Arabic. Based on the generated believability degrees, a set of corrective actions are proposed to guarantee the quality of the data inputs e.g., considering only the data input having the highest believability degree in case of contradictory inputs. We implement a prototype based on the DBE_ALZ approach. We propose a semantic representation to associate the generated believability degrees to data inputs structured using an ontology. Our work is applied to the Captain Memo memory prosthesis. Finally, an evaluation of the proposed work is carried out with 24 Alzheimer's patients and their caregivers. The results are promising.

The DBE_ALZ approach for handling false data inputs is mainly proposed in the context of applications for Alzheimer's patients. However, on the one hand, it can be used in applications for the elderly or patients suffering from progressive mental illnesses. On the other hand, this approach can be used in completely different contexts. For instance, it can be used in police interrogations to estimate the believability of the suspect's responses to questions asked by the interrogator. In this case, we use the DBE_ALZ approach stepwise to estimate the reasonableness of the responses.

However, we ought only to propose other sub-dimensions and associated metrics to estimate the reliability of the suspect.

The generated degrees of believability may help the neurologist doctor to evaluate the mental and memory capabilities of the patient in time. The idea is to elaborate a statistical curve representing the progression of the believability degrees related to data inputs given by the patient in time. Indeed, it is difficult for the doctor to assess the progression of the disease during the limiting time of the consultation; especially as the mental symptoms related to this disease may vary from a moment to another. This curve, which takes data over several days, is therefore useful.

Future works mainly concern two axes. Firstly, we plan to explore person-related big data resources. Secondly we plan to propose a new approach based on machine learning capabilities to evaluate quantitatively the credibility of each data input given by the user.

References

[1]   E. Métais, F. Ghorbel, N. Herradi, F. Hamdi, N. Lammari, D. Nakache, N. Ellouze, F. Gargouri, A. Soukane, Memory Prosthesis, Non-pharmacological Therapies in Dementia 3(2), 2015

[2]   N. Herradi, F. Hamdi, E. Métais, F. Ghorbel, A. Soukane, PersonLink: An Ontology Representing Family Relationships for the Captain Memo Memory Prosthesis, ER 2015 Workshops, 2015

[3]   B. Bouchon-Meunier, *La logique floue:«Que sais-je?»*, Presses universitaires de France, 1993

[4]   P. Smets, Imperfect Information: Imprecision-Uncertainty en Uncertainty Management in Information Systems: from Needs to Solutions, 1999

[5]   N. Achich, F. Ghorbel, F. Hamdi, E. Metais, F. Gargouri, A Typology of Temporal Data Imperfection, International Conference on Knowledge Engineering and Ontology Development, 2019

[6]   V. Niskanen, Introduction to imprecise reasoning. Uncertainty, decision making and knowledge engineering, 1989

[7]   B. Bouchon-Meunier, *La logique floue et ses applications*, 1995

[8]   G. J. Klir, B. Yuan, Fuzzy sets and fuzzy logic: theory and applications, PHI New Delhi, pp. 443-455, 1995

[9]   P. Smets, Imperfect information: Imprecision and uncertainty, Uncertainty management in information systems, pp. 225-254, 1997

[10] N. Gershon, Visualization of an imperfect world, IEEE Computer Graphics and Applications, 18(4), pp. 43–45, 1998

[11] P. Fisher, A. Comber, R.A. (Richard) Wadsworth , *Nature de l'incertitude pour les données spatiales*, 2005

[12] A. M.Olteanu, S. Mustière, A. Ruas, E. Desveaux, *L'apport de données spatiales pour une base de données ethnographique*, SAGEO, 2006.

[13] J. F. Casta, *Incertitude et comptabilité*, 2009

[14] E. Desjardin, O. Nocent, C. De Runz, Prise en compte de l'imperfection des connaissances depuis la saisie des données jusqu' à la restitution 3d, pp. 385–396, 2012

[15] M. Snoussi, P. A. Davoine, Methodological proposals to handle imperfect spatial and temporal information in the context of natural hazard studies. Revue Internationale de Géomatique, 23(3-4), pp. 495-517, 2013

[16] H. B. Sta, Quality and the efficiency of data in "Smart-Cities", Future Generation Computer Systems, 74, pp. 409-416, 2017

[17] E. Gavignet, E. Leclercq, N. Cullot, M. Savonnet, *Raisonner en logique modale sur l'incertitude liée aux données-Application en archéologie*. Revue Internationale de Géomatique, 26(4), pp. 467-490, 2016

[18] C. De Runz, *Imperfection, temps et espace: modélisation, analyse et visualisation dans un SIG archéologique*, Doctoral dissertation, 2008

[19] L. Cai, Y. Zhu, The challenges of data quality and data quality assessment in the big data era, Data Science Journal, 14, 2015

[20] J. M. Juran, Quality control handbook, 1962

[21] H.Huang, B. Stvilia, C. Jörgensen, H. W. Bass, Prioritization of data quality dimensions and skills requirements in genome annotation work, Journal of the American Society for Information Science and Technology, 63(1), pp. 195-207, 2012

[22] A. Rodríguez-Pose, Economic geographers and the limelight: institutions and policy in the World Development Report 2009, Economic Geography, 86(4), pp. 361-370, 2010

[23] C. Aubin, *Indicateurs de performance & Qualité des données: Vers une démarche industrielle dans un grand Hôpital Français*, Gestion et Ingénierie des Systèmes Hospitaliers-GISEH, 2012

[24] L. Berti-Equille, *Un état de l'art sur la qualité des données*, *Ingénierie des systèmes d'information*, 9(5-6), pp. 117-143, 2004

[25] C. Batini, C. Francalanci, C. Cappiello, A. Maurino, Methodologies for data quality assessment and improvement, ACM computing surveys, 41, pp. 1 - 52, 2009

[26] S. Ben Hassine, *Évaluation et requêtage de données multisources: une approche guidée par la préférence et la qualité des données: application aux campagnes marketing B2B dans les bases de données de prospection*, Doctoral dissertation, 2014

[27] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of management information systems, 12(4), pp. 5-33, 1996

[28] F. Naumann, C. Rolker, Assessment methods for Information Quality criteria, International Conference on Information Quality, pp. 148-162, 2000

[29] J. Akoka, L. Berti-Équille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, M. Quafafou, *Evaluation de la qualité des systèmes multisources : Une approche par les patterns*, Qualité des Données et des Connaissances, 2008

[30] M. Schaal, B. Smyth, R. M. Mueller, R. MacLean, Information quality dimensions for the social web, International Conference on Management of Emergent Digital EcoSystems, pp. 53-58, 2012

[31] S. T. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, A. Talaei-Khoei, Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature, International journal of medical informatics, 82(1), pp. 10-24, 2013

[32] S. Juddoo, C. George, Discovering the most important data quality dimensions in health big data using latent semantic analysis, 2018

[33] M. A. Jeusfeld, C. Quix, M. Jarke, Design and analysis of quality information for data warehouses. International Conference on Conceptual Modeling, pp. 349-362, 1998

[34] M. Scannapieco, A.Virgillito, C. Marchetti, M. Mecella, R. Baldoni, The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems, Information systems, *29*(7), pp. 551-582, 2004

[35] F. D. Amicis, C. Batini, A methodology for data quality assessment on financial data, Studies in Communication Sciences, 4(2), pp. 115-137, 2004

[36] J. Long, C. Seko, A Cyclic-Hierarchical Method for Database Data-Quality Evaluation and Improvement, Advances in Management Information Systems-Information Quality, 2005

[37] C. Batini, M. Scannapieco, Data quality, Data-centric Systems and Applications, 2006

[38] P. D. Falorsi, M. Scannapieco, Principi Guida per la Qualità dei Dati Toponomastici nella Pubblica Amministrazione (in Italian), ISTAT, 2006

[39] A. Maydanchik, Data quality assessment, Technics publications, 2007

[40] L. Pipino, Y. Lee, R. Wang, Data quality assessment, Commun. ACM 45, pp. 211–218, 2002

[41] T. Hong, Contributing factors to the use of health-related websites, Health Commun 11(2), pp. 149–165, 2006

[42] Y.W. Lee, L.L. Pipino, J.F. Fund, R.Y. Wang, Journey to Data Quality, The MIT Press, Cambridge, 2006

[43] N. Prat, S. Madnick, Measuring data believability: a provenance approach, Hawaii International Conference on System Sciences, p. 393, Los Alamitos, CA, USA, 2008

[44] G. Shankaranarayanan, B. Iyer, D. Stoddard, Quality of social media data and implications of social media for data quality, International Conference on Information Quality (ICIQ 2012), pp. 311–325, 2012

[45] P. Gomes, A. Paiva, C. Martinho, A. Jhala, Metrics for character believability in interactive narrative, International conference on interactive digital storytelling, pp. 223-228, 2013

[46] K. R. Saikaew, C. Noyunsan, Features for Measuring Credibility on Facebook Information. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9(1), pp. 174–177, 2015

[47] M. Nilsson, F. Alserud, Who can an organization believe in social media? Exploring the process of believability assessment, 2017

[48] C. Reuter, M. A. Kaufhold, R. Steinfort, Rumors, Fake News and Social Bots in Conflicts and Emergencies: Towards a Model for Believability in Social Media, ISCRAM, 2017

[49] M. J. Metzger, A. J. Flanagin, Credibility and trust of information in online environments: The use of cognitive heuristics. Journal of Pragmatics, 59, pp. 210-220, 2013

[50] M. Barkat-Defradas, M. Sophie, L. R. Duarte, D. Brouillet, *Les troubles du langage dans la maladie d'Alzheimer*, 2008

[51] S. Watts, G. Shankaranarayanan, A. Even, Data quality assessment in context: A cognitive perspective, Decision Support Systems, 48(1), pp. 202-211, 2009

[52] M. A. Farage, K. W. Miller, F. Ajayi, D. Hutchins, Design principles to accommodate older adults, Global journal of health science, 4(2), 2, 2012

[53] W. Wali, F. Ghorbel, B. Gragouri, F. Hamdi, E. Metais, A Multilingual Semantic Similarity-Based Approach for Question-Answering Systems, International Conference on Knowledge Science, Engineering and Management, pp. 604-614, 2019

[54] M.Silberztein, T. Váradi, M. Tadić, Open source multi-platform NooJ for NLP, COLING 2012: Demonstration Papers, pp. 401-408, 2012

[55] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, Using of Jaccard coefficient for keywords similarity, International Multiconference of Engineers and Computer Scientists, 1(6), pp. 380-384, 2013

[56] G. Salton, Automatic information organization and retrieval, McGraw Hill Text, 1968

[57] G. A. Miller, WordNet: a lexical database for English, Communications of the ACM, 38(11), pp. 39-41, 1995

[58] K. Kipper-Schuler, VerbNet: A broad-coverage, comprehensive verb lexicon (Ph. D. thesis), University of Pennsylvania, Philadelphia, PA, 2006

[59] A. Khemakhem, B. Gargouri, A. B. Hamadou, G. Francopoulo, ISO standard modeling of a large Arabic dictionary, Natural Language Engineering, 22(6), pp. 849-879, 2016

[60] L. A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning—II, Information sciences, 8(4), pp. 301-357, 1975

[61] M. F. Folstein, S. E. Folstein, P. R. McHugh, "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician, Journal of psychiatric research, 12(3), pp. 189-198, 1975

[62] F. Ghorbel, N. Ellouze, E. Métais, F. Hamdi, F. Gargouri, MEMO_Calendring: A smart reminder for Alzheimer's disease patients, International Conference on Smart, Monitored and Controlled Cities, pp. 41-47, 2017

[63] P. Cingolani, J. Alcala-Fdez, jFuzzyLogic: a robust and flexible Fuzzy-Logic inference system language implementation, International Conference on Fuzzy Systems, pp. 1-8, 2012

[64] F. Cozman, The JavaBayes system, ISBA Bulletin. 7(4), pp. 16–21, 2001