


# Towards a Traceable Data Model Accommodating Bounded Uncertainty for DST Based Computation of *BRCA1/2* Mutation Probability With Age


**Lorenz Gillner**

(University of Applied Sciences Wismar, Germany)

 <https://orcid.org/0009-0007-8244-5810>, [lorenz.gillner@hs-wismar.de](mailto:lorenz.gillner@hs-wismar.de)

**Ekaterina Auer**

(University of Applied Sciences Wismar, Germany)

 <https://orcid.org/0000-0003-4059-3982>, [ekaterina.auer@hs-wismar.de](mailto:ekaterina.auer@hs-wismar.de)

**Abstract:** In this paper, we describe the requirements for traceable open-source data retrieval in the context of computation of *BRCA1/2* mutation probabilities (mutations in two tumor-suppressor genes responsible for hereditary BREast or/and ovarian CAncer). We show how such data can be used to develop a Dempster-Shafer model for computing the probability of *BRCA1/2* mutations enhanced by taking into account the actual age of a patient or a family member in an appropriate way even if it is not known exactly. The model is compared with PENN II and BOADICEA (based on undisclosed data), two established platforms for this purpose accessible online, as well as with our own previous models. A proof-of-concept implementation shows that set-based techniques are able to provide better information about mutation probabilities, simultaneously highlighting the necessity for ground truth data of high quality.

**Keywords:** HBOC, *BRCA1/2*, mutation probability, data integration, data fusion, interval analysis, Dempster-Shafer evidence theory, provenance, AI based data mining

**Categories:** B.8, E.2, G.1.0, G.1.10, G.3, H.3, H.4.2, I.6.4, J.3

**DOI:** 10.3897/jucs.112797

## 1 Introduction

Pathogenic variants (also known as mutations) in cells are responsible for human diseases, most notably cancer. The majority of cancer cases are caused by the so-called somatic mutations, that is, changes in the DNA sequence of non-reproductive cells. However, if such changes occur in the reproductive cells (germline mutations), they can be inherited/passed on to the next generation, which accounts for a high cancer risk not only for specific persons but for their entire families. In particular, variants occurring in *BRCA1* and *BRCA2* genes are the most well-known mutations leading to hereditary breast cancer (BC). In comparison to approximately 13% risk of developing BC during a lifetime in the general female population, the risk of developing BC by 70 – 80 years of age is increased to 55% – 72% if there are hereditary mutations in *BRCA1* and to 45% – 69% if a hereditary *BRCA2* mutation is detected<sup>1</sup>. Additionally, the likelihood of contracting ovarian cancer (OC) becomes higher. The corresponding phenomenon is called the hereditary BC and OC (HBOC) syndrome<sup>2</sup>.

<sup>1</sup> <https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet>

<sup>2</sup> To avoid the implication that this syndrome affects only women, scientists also start to name it King syndrome.

Since there is a strong correlation between HBOC and *BRCA1/2* mutations, a possible strategy for identifying persons or families at high risk is to take a look at whether they carry a corresponding pathogenic variant. That is the reason why genetic testing and counseling for such kinds of mutations gain more and more importance nowadays. Although actual diagnostic genetic testing for DNA changes is getting cheaper, it still cannot be recommended for everyone. Therefore, mutation probability prediction software is used in genetic counseling (investigating if patients have a hereditary risk of a disease). Such software attempts to compute mutation probabilities for patients based on specific indicators (e.g., their family history, ethnicity or origin) without the actual process of testing. At the moment, there exist several established risk assessment (RA) tools and questionnaires for the purpose of predicting mutation or cancer risk, some of them accessible online (e.g., PENN II<sup>3</sup> or BOADICEA<sup>4</sup>). However, such tools are often based on undisclosed or untraceable data leading to questions about validity areas and credibility of results. For example, risks computed by PENN II for a *BRCA1/2* mutation are occasionally quite different from BOADICEA for the same person, without the possibility for non-expert patients to check or understand why or even to decide what prediction to trust (cf. Section 2.3). In [Auer and Luther 2021], we provided an overview of the state of the art, HBOC-related (meta)studies, tools and questionnaires and pointed out further possible problems with their use.

One of the first challenges for a research team trying to develop a new model for *BRCA1/2* based on traceable data is to obtain ground truth information. Medical professionals with access to large amounts of data from clinical trials publish statistical results, but the exact composition of the study population or models used for estimates often remain unclear. Genetic research organizations might disclose the algorithms used to generate a model, but its factual basis remains inaccessible to outsiders [Guerrini et al. 2017]. Although many publications in the area of medical science are open access (OA), the relevant source material is not. All this is justified by data protection legislation: A person's DNA constitutes sensitive medical information since it is a biometric trait. Therefore, it must be handled with utmost care to prevent invasion of individuals' privacy or identity theft.

This leads to a dilemma within genetic research since the correlation between certain types of cancer and the genetic profiles of patients is often exactly the subject of interest. Studying it requires access to confidential records, often obtained in collaboration with specialized institutions. A few projects dedicated to the collection of anonymized clinical data are publicly available over the internet. Two such repositories are the GDC<sup>5</sup> and ICGC<sup>6</sup> data portals. They allow users to explore data sets on various cancer cases from all over the world, with supplemental molecular samples including mutations. However, those projects severely restrict the access to data on germline variants by the requirements of a rigorous registration process. Simple germline variants (SGVs) differ from simple somatic mutations (SSMs) in that they appear in every DNA sample of a donor (and possibly their relatives), which can lead to the identification of the individual in question. However, for modeling the risk of carrying an inherited pathogenic *BRCA1/2* variant, an opportunity to use SSMs does not suffice.

Aside from the data access hurdle, a big difficulty in devising appropriate risk assessment tools in the context of HBOC is the uncertainty in the underlying data, its

<sup>3</sup> <https://pennmodel2.pmacs.upenn.edu/penn2/>

<sup>4</sup> <https://ccge.medschl.cam.ac.uk/boadicea/>

<sup>5</sup> <https://gdc.cancer.gov/>

<sup>6</sup> <https://dcc.icgc.org/>

major (interconnected) sources being data provenance and collection. A good record of data provenance facilitates reproducibility, validation and belief in the reliability of scientific results [Pasquier et al. 2017]. This point, especially important in the area of medical science, is still not implemented to a sufficient degree there. For example, although GDC contains data base items about the specific history of a family or patient connected to a certain variant (e.g., ethnicity, origin, age), they are often left empty (at least, in the publicly accessible somatic case). Frequently, the way of collecting the data is not documented sufficiently. For example, cohort composition or exact criteria for the choice of test persons as often as not remain unclear within a survey. Besides, there is inherent uncertainty present in the data since patients are often unsure about the specifics of their family history. Although it is difficult to remove this latter cause of uncertainty, the reduction in the former two can and should be addressed by astute researchers.

Uncertainty is present in practically every kind of real-life application, but it is especially high in the case of medical studies. In the field of uncertainty quantification, two main sources are discerned: aleatory (due to randomness) and epistemic (due to the lack of knowledge). State of the art tools such as PENN II or BOADICEA rely on crisp data in combination with the classical probability theory to take into account aleatory uncertainty, working with arithmetic means or medians in the presence of the epistemic one. In many cases, intervals represent the available, uncertain but bounded data better than crisp numbers. They can be propagated from inputs to outputs of a (static or dynamic) model using interval analysis (IA) [Moore et al. 2009]. Moreover, a way of combining the probabilistic and set-based reasoning is offered by the co-called imprecise probability [Bradley 2019], in particular, the Dempster-Shafer evidence theory (DST) [Shafer 1976].

The goals of this paper are, first, to formulate the requirements on the optimal HBOC database from the viewpoint of risk assessment with a focus on reliability including traceability and provenance. Second, we propose an interval DST model for computing *BRCA1/2* mutation probabilities based on data from OA publications as a proof of concept that taking into account epistemic uncertainty explicitly provides improved quality information for a patient. This model helps to incorporate uncertainty about the ages of the involved persons better. We compare the results with those from PENN II and BOADICEA as well as with those from our earlier model that did not differentiate to the same degree wrt. age. Here, the importance of careful extraction of ground truth data can be seen: we provide results of the same (old) model from [Auer and Luther 2021] based solely on data from [Frank et al. 2002] versus aggregated data extracted from several OA publications. The data in the developed database is traceable to the publication where they were provided.

The paper is structured as follows. First, we provide a short overview of the background methods and tools we rely on in Section 2. After that, we formulate the requirements on the (ideal) HBOC data base and provide a proof-of concept implementation filled by data from OA publications in Section 3. Next, we present our two-phase DST model improving that from [Auer and Luther 2022] by taking into account the uncertainty about age using intervals in Section 4. Conclusions and an outlook on our future work are in the last section.

## 2 Background

In this section, interval analysis and the Dempster-Shafer theory are described briefly. Additionally, we overview the main features of PENN II and BOADICEA, two models

based on conventional probability theory and available online for computing the *BRCAl/2* mutation probability and the risk of breast cancer, respectively.

## 2.1 Basic concepts of interval analysis

Interval analysis [Moore et al. 2009] is a well-known tool for result verification with applications in computer-assisted proofs, engineering, computer graphics, medical science and many others. With the help of IA, it is possible to prove formally, using a suitable fixed point theorem the assumptions of which can be checked on a computer reliably, that the result of a computer simulation is correct (given that the underlying code is correct). This takes into account such factors as rounding, conversion, discretization or truncation errors. The results are intervals with bounds expressed by floating point numbers which with certainty contain the exact solution to the formal model. Since the methods work with sets, they can be used for propagating bounded uncertainty, usually from the inputs to the outputs, in a deterministic way although there are also approaches to solve inverse propagation problems [Merheb et al. 2013, Desrochers and Jaulin 2017]. A common drawback of such rigor-preserving methods, caused by the dependency problem or the wrapping effect [Lohner 2001], is the possibility of too wide solution sets – an inherent problem of naive IA that more sophisticated techniques with result verification (e.g., affine or Taylor model based approaches) address [de Figueiredo and Stolfi 2004, Neumaier 2003, Makino and Berz 2004].

A real interval  $[\underline{x}, \bar{x}]$ , where  $\underline{x} \in \mathbb{R}$  is the lower,  $\bar{x} \in \mathbb{R}$  the upper bound, is defined as

$$[\underline{x}, \bar{x}] = \{x \in \mathbb{R} | \underline{x} \leq x \leq \bar{x}\} ,$$

usually for  $\underline{x} \leq \bar{x}$ . Crisp numbers  $x \in \mathbb{R}$  can be represented by point intervals with  $\underline{x} = \bar{x} = x$ . For an operation  $\circ = \{+, -, \cdot, /\}$  and two intervals  $[\underline{x}, \bar{x}]$ ,  $[\underline{y}, \bar{y}]$ , the corresponding interval operation can be defined as

$$\begin{aligned} [\underline{x}, \bar{x}] \circ [\underline{y}, \bar{y}] &:= \{x \circ y \mid \forall x \in [\underline{x}, \bar{x}], y \in [\underline{y}, \bar{y}]\} \\ &= [\min(\underline{x} \circ \underline{y}, \underline{x} \circ \bar{y}, \bar{x} \circ \underline{y}, \bar{x} \circ \bar{y}), \max(\underline{x} \circ \underline{y}, \underline{x} \circ \bar{y}, \bar{x} \circ \underline{y}, \bar{x} \circ \bar{y})] , \end{aligned}$$

that is, the result of an interval operation is also an interval. For normal interval division, it is assumed that  $0 \notin [\underline{y}, \bar{y}]$  although it is possible to allow divisor intervals to contain zero in extended interval arithmetics (see, e.g., [Kahan 1968]). The general formula can be simplified for a given operation  $\circ$  (e.g.,  $[\underline{x}, \bar{x}] - [\underline{y}, \bar{y}] = [\underline{x} - \bar{y}, \bar{x} - \underline{y}]$ ). An interval with floating point numbers as bounds can be obtained for any real interval in a verified way by using the concept of outward rounding. Based on interval arithmetic described above, which includes the possibility to evaluate functions over intervals, higher-level methods, for example, for solving systems of algebraic or differential equations, can be formulated to provide their error bounds (i.e., result verification) automatically.

## 2.2 Basic concepts of the Dempster-Shafer theory

The Dempster-Shafer theory [Ayyub and Klir 2006] facilitates synthesis between data and information and is increasingly used for uncertain data fusion, especially in the context of AI systems [Tang et al. 2023]. It combines evidence from different sources and provides a measure of confidence that a certain event occurs. A classical, additive,

discrete probability density function defines the probability (given by a crisp number) that a random variable  $X$  is equal to its certain realization  $x_i$  (again, a crisp number, i.e., real or point value). The finite DST allows us to assign a (crisp) probability to the event that a realization of  $X$  belongs to a given set (e.g., an interval  $[\underline{x}_i, \bar{x}_i]$ )<sup>7</sup>. The result is given in terms of the lower and upper limits (belief and plausibility) on the probability of a subset of the frame of discernment  $\Omega$ . A random DST variable can be characterized by its basic probability assignment (BPA)  $m$ . If  $A_1, \dots, A_n$  are the sets of interest where each  $A_i \in 2^\Omega$ , then  $m$  is defined by

$$\begin{aligned} m : 2^\Omega &\rightarrow [0, 1], & m(A_i) &= p_i, \quad i = 1 \dots n, \\ m(\emptyset) &= 0, & \sum_{i=1}^n m(A_i) &= 1. \end{aligned} \quad (1)$$

The mass of the impossible event  $\emptyset$  is equal to zero. Every  $A_i$  with  $m(A_i) \neq 0$  is called a focal element. The sum of masses of focal elements should be equal to one. If the sum is greater than one in BPAs provided by the experts, then a normalization can be carried out as  $\tilde{m}(A_i) := m(A_i) / \sum_{i=1}^n m(A_i)$ . If the sum is less than one, then the same normalization can be used or a new focal element  $A_{n+1} = \Omega$  can be introduced to accommodate the missing probability. The latter variant only makes sense for computing the lower limit  $Bel(Y)$  whereas the former variant could inflate the belief function too much.

The plausibility ('worst case') and belief ('best case') functions can be defined with the help of the BPAs for all  $i = 1 \dots n$  and any  $Y \subseteq \Omega$  as

$$Pl(Y) := \sum_{A_i \cap Y \neq \emptyset} m(A_i), \quad Bel(Y) := \sum_{A_i \subseteq Y} m(A_i). \quad (2)$$

These two functions represent a possibility to define an upper and a lower non-additive monotone measure [Ayyub and Klir 2006], respectively, on the true probability.

If there is evidence for the same issue from two or more sources, the BPAs have to be aggregated. In [Ferson et al. 2003], there is a good overview of the available aggregation methods, for example, Dempster's rule

$$m_{12}(A_i) = \frac{\sum_{\forall A_j \cap A_k = A_i} m_1(A_j)m_2(A_k)}{1 - \sum_{\forall A_j \cap A_k = \emptyset} m_1(A_j)m_2(A_k)} \quad (3)$$

with  $A_i \neq \emptyset$ ,  $m_{12}(\emptyset) = 0$  or mixing and averaging:

$$m_{1\dots n}(A_i) = \sum_{k=1}^n w_k \cdot m_k(A_i), \quad \sum_{k=1}^n w_k = 1. \quad (4)$$

Although Dempster's rule in (3) is a fair way to combine conflict-free evidence, it cannot always be applied in the context of automatic data extraction since conflicts cannot be

<sup>7</sup> Analogous considerations can be made for continuous random variables

dismissed a priori there. Recently, a possibility to circumvent this has been proposed in [Tang et al. 2023].

As described, for example, in [Auer et al. 2010], it is possible to work with interval BPAs instead of crisp ones. The meaning of such an interval BPA (IBPA) is then as follows: The probability that a realization of a random variable  $X$  belongs to a certain set is itself uncertain (but bounded). The computation of  $Pl(Y)$ ,  $Bel(Y)$  and any kind of aggregation can work in the same way as for crisp BPAs if interval arithmetic is used instead of floating point arithmetic. Since we cannot define the inverse wrt. addition in interval arithmetic [Moore et al. 2009], the condition  $\sum_{i=1}^n m(A_i) = 1$  for interval  $m$  cannot be fulfilled. It turns into the relaxed property  $1 \in \sum_{i=1}^n m(A_i)$ . The respective belief function signifying a lower limit on the strength of evidence is then itself an interval function having a lower and upper bound. Note that we do not try to countermeasure the relaxation of the summation property in this paper. It can be done in principle as shown, for example, in [Piegat and Dobryakova 2020] for linguistic probabilities (not for the DST) using interval arithmetic type 2. To use such kinds of IBPAs and compare the results to those presented in this paper is a topic of our future work.

### 2.3 Two existing web platforms for predicting *BRCA1/2* mutation probabilities

With over 8 million articles, PubMed<sup>®</sup> Central<sup>8</sup> is by far the most comprehensive public, English-language source for life science publications. More than half of the articles are available as OA. Nonetheless, or precisely because of this information flood, it is quite difficult for non-experts to find ground truth about the hereditary breast and ovarian cancer risk in their family.

In an early standard publication that had been used for prediction for many years, Tables from [Claus et al. 1994] estimate cumulative BC probability based on a survey considering mainly age-specific risk factors in combination with the family history and using a Bayesian model (with data on 4730 patients with confirmed BC matched against 4688 control subjects). Another relatively early paper [Frank et al. 2002] provides predictions for mutations in *BRCA1/2* correlated with such risk indicators as age of diagnosis, personal and family history, and ethnicity (also compiled in tables), for the cohort of over-all 10000 participants (of Ashkenazi-Jewish and non-Ashkenazi-Jewish ethnicity). Frank tables could be seen as corresponding to ground truth since they contain observed frequencies. However, they are old, occasionally contradictory and consider only relatively small cohorts. The models PENN II and BOADICEA offer easy, questionnaire-type online interfaces to their respective models computing (among others) *BRCA1/2* mutation probabilities according to the risk indicators they consider important in a more detailed way. However, the actual data their models are based on are undisclosed.

PENN II [Lindor et al. 2010] is a mathematical model giving predictions about probabilities of *BRCA1* and *BRCA2* mutations based on logistic regression derived from 861 family histories of European and North American origin and taking into account Mendelian logic. The risk of a genetic defect is the same for the proband and the family if the proband is diagnosed with cancer. If not, the risk of the proband is reduced depending on the degree of relationship to the family member with a cancer diagnosis according to the principles of Mendelian genetics. This free tool offered by the University of Pennsylvania considers the following risk indicators: presence and ages of BC, presence of OC alone or with BC, bilaterality, diagnosis in both mother and daughter, male BC, presence of pancreatic and prostate cancers, Ashkenazi-Jewish or non-Ashkenazi-Jewish

<sup>8</sup> <https://www.ncbi.nlm.nih.gov/pmc/>

ethnicity as well as the degree of relation of the patient to the BC/OC case in the family. In Figure 1, screenshots of the PENN II questionnaire and result pages are shown for the web interface of the model.

**Part A. Select the side of the family being evaluated:**  Maternal  Paternal

**Part B. Please provide the following information:**

1. Presence of Ashkenazi (Eastern European) Jewish ancestry in the family?  no  yes
2. Number of women in the family diagnosed with both breast and ovarian cancer?  (0-100)
3. Number of women in the family diagnosed with ovarian, fallopian tube, or primary peritoneal cancer (in the absence of breast cancer)?  (0-100)
4. Number of breast cancer cases in the family diagnosed in women under the age of 50?  (0-100)
5. What is the age of the youngest breast cancer diagnosis in the family?  (18-130)
6. Presence of mother-daughter breast cancer diagnoses in the family?  no  yes
7. How many women with bilateral breast cancer in the family? (Note: Count women with cancer in both breasts, not two primaries in one breast.)  (0-100)
8. Number of men diagnosed with breast cancer in the family?  (0-100)
9. Presence of pancreatic cancer in the family?  no  yes
10. Number of men diagnosed with prostate cancer in the family?  (0-100)

**Part C. Closest relative with breast or ovarian Cancer:**

**Part D. Patient Information (Optional - For use on printable report only):**

1. Patient's first name
2. Patient's last name
3. Patient's age
4. Clinic location

[Risk Form](#) | [Admin](#)

**The Penn II Risk Model**

**Penn II *BRCA1/BRCA2* Mutation Prediction Result**

	Individual	Family
<b>Risk of <i>BRCA1</i> Mutation</b>	5.0%	5.0%
<b>Risk of <i>BRCA2</i> Mutation</b>	6.0%	6.0%

Figure 1: Screenshot of PENN II web interface from <https://penmodel2.pmacs.upenn.edu/penn2/>: questions (above) and computed probabilities (below) for the disease pattern from this subsection

BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) [Lee et al. 2019] is a part of the originally OA *CanRisk Tool* from the University of Cambridge which is in the process of being commercialized. Its goal, in contrast to PENN II, is to compute the risk of breast cancer based on a variety of genetic and non-genetic indicators. That is, BOADICEA is concerned, more broadly than PENN II, with general cancer risk and not only HBOC syndrome related *BRCA1/2*

mutation probabilities. Nonetheless, BOADICEA incorporates indicators based on family history and computes respective probabilities as a part of its analysis. Its questionnaire – still available for free at the moment after registration – contains questions about the following HBOC-related risk factors (cf. Figure 2): presence of BC, contralateral BC, OC and/or pancreatic cancer; family history including relation degrees and ages of cancer diagnosis; ancestry (again differentiating only between Ashkenazi-Jewish or non-Ashkenazi-Jewish). The questionnaire contains many more questions concerning other (non-genetic) factors. Note that although the ages can be specified as intervals, arithmetic means are used for further computations.

The screenshot displays the BOADICEA web interface, divided into three main sections: Medical History, Polygenic Risk Scores, and Family History.

**Medical History:**

- Questions about endometriosis and tubal ligation with radio button options (Yes, No, Unknown).
- Questions about oophorectomy and mastectomy with Yes/No buttons.
- A note: "Please note for women that have undergone BSO pre-menopausally, it can be assumed that their periods stopped at the age at BSO (in the Women's Health section). The current data on the effect of 'natural' and 'BSO-induced' menopause on breast cancer risk, suggest similar associations [Lancet Oncol. 2012(2)]."
- Questions about various cancers (Breast, Contralateral Breast, Ovarian, Pancreatic) with Yes/No buttons and an age input field for the first breast cancer question.

**Polygenic Risk Scores:**

- Question: "Has a SNP array / Polygenic Risk Score (PRS), ever been run?" with a No button.

**Family History:**

- Question about Ashkenazi Jewish ancestry with radio button options (Ashkenazi Jewish, Other / Unknown).
- A pedigree chart showing a family tree with a proband (orange circle) labeled "no. 100 breast ca. 35".
- Individual's Details panel on the right:
  - Name: imc
  - Sex: Male (selected), Female
  - Status: Alive (selected), Deceased
  - Age: 35
  - Year of birth: 1998
  - Proband: (checkbox checked), Exclude: (checkbox unchecked)
  - Diagnosis: Age, Pathology, Gene Tests
  - approx: (checkbox checked)
  - Breast: 30-40
  - Contralateral breast: -
  - Ovarian: -
  - Pancreatic: -

Figure 2: Screenshot of BOADICEA web interface (a user account needed) from <https://ccge.medschl.cam.ac.uk/boadicea/> showing only HBOC-related questions for the disease pattern considered in this subsection

At the moment, such general-purpose AI systems as ChatGPT<sup>9</sup> are not able to provide a differentiated answer about HBOC based on available publications. ChatGPT can be used to summarize and explain knowledge (with various challenges arising from this in scientific research and exchange), see [Schäfer 2023], therefore, it is in principle capable of finding out the appropriate information. For example, a broadly formulated question

<sup>9</sup> <https://chat.openai.com/>, May 3 version, accessed on May 8 2023



Are you capable of giving predictions on genetic mutation risk?

is answered by ChatGPT as “As an AI language model, I am capable of providing information and explanations about genetic mutations and their associated risks based on available scientific data”, with an additional caution to consult a gene specialist. Confronted with a concrete disease pattern, it answers mostly in general terms as

“The presence of a family history of breast cancer can indicate an increased risk of developing the disease”.

Only one (partial) pattern out of those we consider in Section 4 elicited actual probabilities as answers. The answer to the question

What’s the probability of carrying a pathogenic BRCA variant if a woman was diagnosed with breast cancer under the age of 40?

is

...“The exact probability of carrying a pathogenic BRCA1 or BRCA2 variant can vary depending on a variety of factors, including personal and family medical history, ethnicity, and other factors. However, some studies have estimated that the frequency of BRCA1 or BRCA2 mutations in individuals diagnosed with breast cancer before the age of 40 ranges from 8% to 17%”,

again with a caution that “Genetic counseling and testing can provide more personalized information about an individual’s specific risks and can inform decisions about management and prevention strategies”. ChatGPT does not provide information about the studies it mentions. As a comparison, PENN II computes the individual risk of *BRCA1* mutation as 5%, *BRCA2* as 6% (equal to the familial risk in this case), BOADICEA as 3.19% and 2.93%, respectively. As we can see, the probabilities suggested by all of the tools are different, posing the question of which one to trust.

Obviously, the ChatGPT answers are quite impressive for a general-purpose AI system. Nonetheless, they still indicate the need for specific data mining (AI based) strategies in OA publications devoted to reporting *BRCA1/2* frequencies from the available literature (cf. Subsection 3.3, 3.4). They should be combined with mathematical models taking into account both aleatory and epistemic uncertainty in data for computing probabilities to be used during genetic counseling with a person or a family (cf. Section 4).

### 3 Design of an Optimal HBOC Database

The field of *BRCA1/2* research is developing rapidly. Recent technological advancements such as next generation sequencing (NGS) have enabled fast variant discovery in samples donated by cancer patients. Since not all variants necessarily cause cancer, they first need to be assessed by experts to classify their pathogenicity [Plon et al. 2008]. There are numerous databases dedicated to the classification of (*BRCA1/2*) variants available online: BRCA Exchange<sup>10</sup>, ARUP *BRCA1/2*<sup>11</sup>, LOVD<sup>12</sup>, ClinVar<sup>13</sup>. These data can

<sup>10</sup> <https://brcaexchange.org/>

<sup>11</sup> <https://arup.utah.edu/database/BRCA/>

<sup>12</sup> <https://www.lovd.nl/>

<sup>13</sup> <https://www.ncbi.nlm.nih.gov/clinvar/>

help patients who already underwent genetic testing to understand their test results better. During genetic counseling, however, this information is not available yet. Therefore, risk assessment tools have to be able to estimate the mutation probability based only on knowledge about indicator factors in personal or family history.

While mutation risk can be estimated by relying on different theories (e.g., logistic regression, Bayesian networks, decision trees, cf. the overview in [Auer and Luther 2021]), the common feature of all strategies is the link between clinical data and variants' classification. All RA models usually need empirical data (ground truth). To obtain it, modeling is preceded by an extensive long-term survey in the best case. If this is not possible, the data from previous studies can be consolidated for the same purpose within a meta-study [Wang et al. 2021]. However, general standards about how to collect and report empirical data or record the meta-data about them are mostly missing, making definite conclusions about ground truth very difficult. As described in the introduction, the sparse availability of open-access cancer-related medical records for HBOC complicates the development of new RA models even further.

In this section, we first introduce our understanding of the concepts of data integration and data fusion (which are often used synonymously in literature but are actually not exactly the same) along with data provenance in Subsection 3.1. After that we propose a database design for storing ground truth in the context of the HBOC syndrome which can be understood as a first step towards a standard in Subsection 3.2. Note that the existing standards such as FHIR<sup>14</sup> or OMOP<sup>15</sup> cannot be used as they are, cf. [Bönisch et al. 2022]. The focus of the design proposed here is on flexibility: the model developers have to be able to perform the task of data fusion easily depending on the criteria they need the ground truth data on. Because *BRCA1/2* genes are the most researched ones in connection with breast and ovarian cancer, we consider only them, although the proposed database scheme can be extended to take into account further genes, for example, *CHEK2* or *PALB2*. In Subsection 3.3, we suggest an approach to extract information from OA publications into a database. This helps to avoid issues of classified information nature still present in the suggestion from 3.2. It is also a good way for making the data accessible without violating the patients' privacy. A proof of concept implementation illustrates the applicability of the data structure within the context of a DST based model for predicting *BRCA1/2* mutation probabilities.

### 3.1 Extracting data: Integration versus fusion; provenance

If several data sources are to be combined, the most important sub-processes to consider besides the data cleaning are

- consolidating various data schemata and
- combining the data objects.

If two or more SQL databases are considered as sources, then these sub-processes correspond roughly to the operations JOIN and MERGE, respectively. If, however, the data originate from different sources, the operations needed to be carried out are usually more complex, corresponding to *data integration* and *data fusion*. These terms are sometimes used interchangeably in the literature. However, they are separate, although

<sup>14</sup> <https://fhir.org/>

<sup>15</sup> <https://www.ohdsi.org/data-standardization/>

interconnected, processes, with data integration needing to be carried out prior to data fusion.

Data integration is extension of the structure and content of one data source by that of another [Li Lee and Ling 1995]. The basis of this operation is the so-called schema matching that checks the attributes of data sources for the three cases *identity*, *similarity*, *newness* and generates a global schema based on the findings using a top-down (e.g., by data harmonization) or a bottom-up approach depending on whether the number of data sources to combine is known beforehand [Arfaoui and Akaichi 2015].

The task of data integration is to build up a large body of information from many sources and to reduce it to a common data schema. By contrast, various data about one object have to be aggregated into a single data item during data fusion, for example, to combine multiple opinions on a particular issue into one statement. The data set to be combined is not allowed to contain duplicates, which might make further pre-processing necessary [Naumann et al. 2006, Dong et al. 2015]. If fusion is to be applied to non-numeric data, then the strategies of interaction (users decide which data is retained), selection (only the data from pre-selected sources is retained) and voting (data with the most frequent occurrence is retained) can be used. When working with numerical data, data synthesis using, for example, Bayesian networks, DST or fuzzy logic is to be preferred.

Not infrequently, the operations mentioned above are performed without giving much thought to the reliability of the data in the global database obtained in such a way. One part of making data reliable and helping to interpret data better is a record of their *provenance*. Similar approaches are being actively developed for intelligent log management of distributed applications (cf. [Harutyunyan et al. 2019]) and logging of data in general [Moreau et al. 2008] since many years.

Provenance is defined as “information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness”<sup>16</sup>. Moreover, data provenance contributes to reproducibility of scientific results through “both systematic and formal records of the relationships among data sources, processes, datasets, publications and researchers” [Pasquier et al. 2017]. Provenance means that scientists require for their research not only the data themselves but also the meta-data about the data. A possible meta-data representation is through an acyclic directed graph with its vertices reflecting the involved persons/organizations, data items and data transformations and its edges corresponding to the interactions between them. This can be the basis for exploiting the provenance involving the necessary stages of meta-data capture, storage and analysis/visualization. Although such standards as PROV [Missier 2017] exist, they are still not used widely, especially while recording or reporting medical data. Standardizing medical science wrt. provenance guidelines is a challenging task, to tackle which is nonetheless important and necessary as pointed out and partially attempted, for example, in [Martínez-García and Hernandez-Lemus 2022] (without using the word ‘provenance’, exactly) and [Gierend et al. 2023].

### 3.2 Requirements and possible design of a traceable HBOC database

Over the past decades, researchers have agreed on certain factors that act as indicators for pathogenic *BRCAl/2* variants. Aside from a patient’s biological sex, their nationality and ethnicity can be decisive. As suggested in previous studies [Hall et al. 2009, Ashton-Prolla

<sup>16</sup> <https://www.w3.org/TR/prov-overview/>

and Vargas 2014], certain ethnic groups – most prominently people of Ashkenazi-Jewish ancestry – are at a higher risk of carrying *BRCA1/2* mutations, mostly due to endogamy and geographically contained gene pools (so-called “founder mutations”). Since it is possible that a person is a mutation carrier but has not developed cancer yet (while other family members have), the ideal HBOC database should be able to store relationships between individuals. Moreover, the age of cancer patients at the time of diagnosis, the primary tumor location and subtype, as well as all variants occurring in *BRCA1/2* genes must be recorded. Registering every variant is necessary because its classification as pathogenic or non-pathogenic can change over time. It should also be indicated if a somatic or germline mutation occurred.

There are existing standards for recording and exchanging medical information such as FHIR or OMOP. However, the common goal of these standards is to keep a reliable record, for one actual patient, of the development of their disease history with time, including details on observations, conditions, prescriptions in order to provide best care. These are not the models on which databases like GDC or BRCA Exchange, relevant for the research in this paper, are based<sup>17</sup>. Such details are not necessary for obtaining information on mutation probabilities and would additionally disclose information openly violating data privacy legislation which is what we try to avoid as far as possible in our suggestion for a traceable database. Moreover, it has been demonstrated recently in [Bönisch et al. 2022] that “none of the data formats include all metadata, which is required to successfully operate the MeDIC<sup>18</sup> for the purpose of reliable data management”. Hence, we formulate the requirements specific to the HBOC research which can be understood as a first step towards a standard in this area. Obviously, there is a need for automated compression and anonymization of FHIR or OMOP patient data into any data format suitable for research on HBOC.

In [Auer and Luther 2022], we came to the conclusion based on the extensive analysis of the available publications that, in order to be able to perform data integration or fusion, it is necessary to choose studies with clearly described cohorts of large sizes which classify the included patients and their family members wrt.

- the risk factor single/multiple breast cancer (also, in the same person), bilateral, male breast cancer and ovarian cancer with the record about the respective first age of onset;
- standardized selection criteria (disease patterns), the a priori risk class;
- origin/ethnicity of the patient (e.g., the youngest family member with BC), family history including the degree of relationship and the first occurrence of the disease;
- detected tumor subtypes (e.g., triple negative breast cancer); and, finally,
- the quality and the trustworthiness of the data (e.g., use of public databases with a disclosed search strategy), ideally, its provenance or lineage.

However, the list of the important risk factors or selection criteria might change with the new developments in medical research. A future standard should be flexible enough to incorporate any such changes.

We consider a relational database design consisting of eight tables to be suitable for the task at hand: four tables for the entities *cancer*, *person*, *variant* and *project*, and four relational tables for the connections between them. The corresponding entity-relationship diagram is shown in Figure 3. This architecture allows multiple cases of

<sup>17</sup> See <https://gdc.cancer.gov/developers/gdc-data-model>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6324924/>

<sup>18</sup> Medical Data Integration Center

cancer in the same person, while a set of variants can be assigned to each patient, describing only the necessary genetic properties in the context of *BRCA1/2* genes.

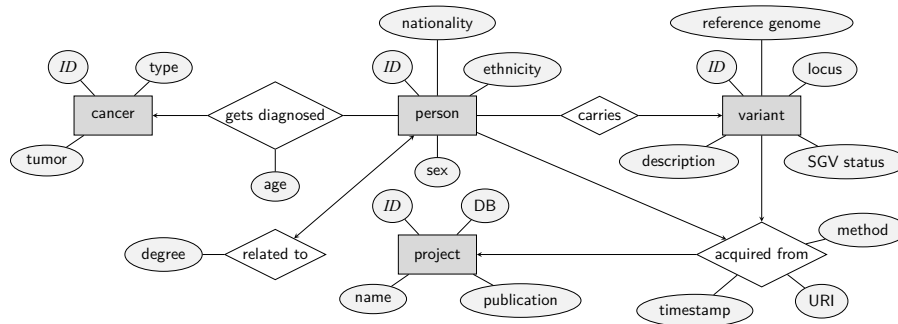


Figure 3: Entity-relationship diagram

By combining oncogenic databases with variant databases using data integration strategies, the standard HBOC database could be constructed without the need for conducting new large-scale, long-term surveys while retaining a good degree of data provenance because all data fusion instances possibly carried out for obtaining concrete values for risk factors can be traced to the corresponding database items as well as aggregation strategy be made clear or exchanged. As shown in Figure 3, each personal record should be associated with a project, to trace back its original study cohort. In well-established oncogenic databases like GDC and ICGC it is common practice to disclose research projects' goals and methodologies to ensure a record's credibility. Similarly, it must be transparent how a variant was discovered and classified. Patient records and variants both originate from third-party sources, so it is essential to document time and method of acquisition, as well as a unique resource identifier (URI) for online databases. The umbrella term *project* is used here for the research project that published a data record in a public database and possibly summarized its findings in an accompanying publication referred to by its DOI.

Combining the two types of databases considered in this scenario can be expressed as a two-way extract-transform-load (ETL) process shown in Figure 4. The ETL process is standard in data integration/warehousing tasks [Denney et al. 2016] and consists of the following three steps:

**First**, relevant subsets are queried from all data sources. Although containing common domain-specific data, different databases often use distinct technology stacks, which may require a query be translated. Such a 'translation' restructures a single request to meet the grammatical rules of various other query languages. In our context, this step applies to both oncogenic and variant databases and is needed to extract all data on both *BRCA1/2* -related cancers and variants, respectively.

**Second**, the transformation step ensures that all extracted records follow the same global schema. This can be achieved by either schema matching (finding similar data fields and rearranging them in a common structure) or schema mapping (fitting similar data fields into a predefined data model). In our context, schema mapping is preferred,

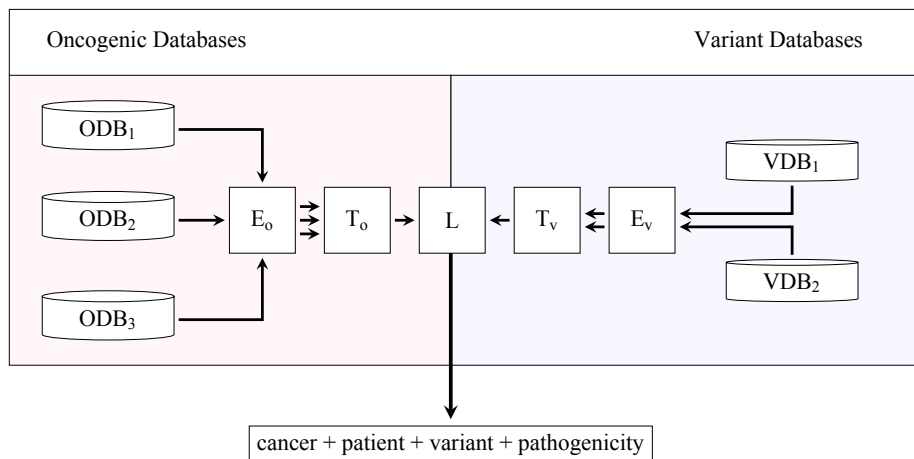


Figure 4: Two-way ETL scheme

since the required tables and fields are already known (cf. Figure 3). Up to this point, oncogenic and variant databases are handled separately and produce two homogeneous integrated datasets.

**Finally**, oncogenic and variant data need to be joined, based on variant names. This part is probably rather complicated because of synonymous nomenclature used for a single variant across multiple sources; possible duplicates would have to be eliminated. Note that the database BRCA Exchange actually stores synonyms for each registered variant and might be used as a thesaurus in this case. Once oncogenic patient records, variants and their pathogenicity have been linked, the merged dataset can be loaded thus providing absolute mutation frequencies as the basis for a new risk assessment model.

Although this design stores only a minimum amount of personal data, the question of privacy still arises because of the short DNA excerpts linked to patients, which may again require that access be restricted. In the next subsection we suggest an alternative database possibility to circumvent any privacy concerns.

### 3.3 Construction of an alternative database from scientific literature

The main issue with open access genetic data is the possibility of personal identification, even though data sources might be used only to calculate relative frequencies of mutation carriers among a group of patients with similar characteristics. The mentioned frequencies themselves, however, may in fact be published, because they do not contain any sensitive data anymore. To make use of this fact, we propose a second database design that does not depend on the first-hand genetic data at all. Instead, we utilize findings from OA publications by scientists with access to genetic samples linked to medical information and build a database upon findings from multiple sources.

Note that the alternative suggested here is a proposal of how to publish classified data without privacy violation concerns, achieved through aggregation. Although the degree of trustworthiness diminishes, this would give those developers of new RA models who

are not linked to big scale medical institutions at least some semblance of traceable data. The proposal from Subsection 3.2 is to be preferred if researchers are granted access to sensitive data. The approach from this subsection is an acceptable compromise if not. It provides data in a standardized form and documents meta-data to a certain degree by recording all source publications and subsequent aggregation strategies.

While developing the database, we focused primarily on data in the form of tables from medical publications. Obviously, graphical representations such as histograms contain valuable information, too, but are much more difficult to translate into exact data, both manually and automatically. Usually, the prevalence of *BRCA1/2* mutations presented in such research papers is stratified by attributes similar to those described at the beginning of Subsection 3.2 and is given as absolute frequencies. Because various publications might focus on different types of cancer, patient's characteristics or age intervals, the database used for storing the extracted information must support heterogeneous data structures. Therefore, we recommend the use of a NoSQL database. The data extracted from one publication may be structured in the following way:

```
[
  {
    case: {
      cancer: 'breast',
      type: 'bilateral',
      age: { lo: 18, hi: 29 }
    },
    total: 26236,
    genes: [
      { brca1: 4996 },
      { brca2: 2519 }
    ]
  }
]
```

The flexible nature of NoSQL databases makes using traditional database design principles to describe their entire structure somewhat difficult. Since records in a NoSQL database should be considered as linked/related objects of similar appearance rather than rows in a strictly constrained table, object-oriented modeling paradigms can be applied. The modeling of NoSQL databases is still a subject of ongoing research, with no established standard as yet. In Figure 5, we use an extension of the UML class diagram standard [Sparks 2011] to represent the database structure because of its widespread familiarity and readability. Its use of symbols has been modified to express an object's structure in the following way:

- atomic property; attribute in the classical sense
- = multiple instances of the same atomic property
- + composite property; the attribute itself is an object
- \* enumeration of composite properties
- ~ reference to another object
- ... anything.

Note that object nesting, a common operation in object-oriented databases, can also be represented as a classical relational model. However, such kind of a representation can be misinterpreted as a system of linked objects, which is not the case in our database. Objects of nested attributes can still have relations, while nested attributes are usually entirely private to the parent objects' scope.

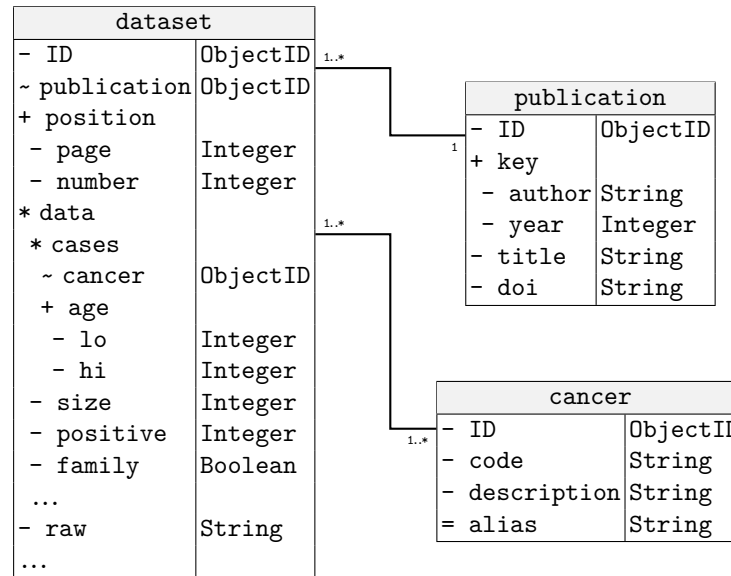


Figure 5: Publication database schema

The proposed database architecture in Figure 5 links tabular data (datasets) to its corresponding publication object, which in turn references the original source material by its DOI to keep the data traceable. Every object of type dataset is a single table with a unique position (page being the relative page number in the linked publication; number refers to the absolute table count per page). Contextually relevant table cells are described as data objects, covering the positive rate of a certain type of cancer (with optional aliases to compensate for synonyms) for a cohort of sample size in the age bracket lo to hi, and recording whether or not this sample concerns the family history. The original raw data should be present for validation.

### 3.4 Proof of concept literature database for computing *BRCAl/2* mutation probabilities

We tested the database design proposed in Subsection 3.3 wrt. its practicability by implementing it in the NoSQL database system MongoDB. Its extensive set of aggregation functions allowed us to perform most computations inside the database itself, for example, for construction of age intervals for certain risk factors or for retrieval of the relative frequency of pathogenic variants in every calculated interval. The data fusion of multiple sources for the same risk factors can be used to produce a general approximation of the mutation probability depending on different criteria.



Similar to [Auer and Luther 2021], the general mutation risk was modeled as a combined Dempster-Shafer structure of the proband’s personal history and their family history. Partitioned into those two categories, each data source was weighted based on additional quality criteria. A source was considered more important if its sample size was sufficiently large with clearly divided age groups and a transparent classification strategy. For each case of cancer in a specific age group, its mass was set to the weighted average (cf. Eq. (4)) across all relevant sources, for the personal ( $m_p$ ) and familial ( $m_f$ ) mutation risk respectively. Familial and personal mass distributions were combined into a single Dempster-Shafer structure ( $m_{pf}$ ), using Dempster’s rule (cf. Eq. (3)). We used the R language to combine the aggregated mass assignments, and the `mongolite` package as an interface to our database. In Table 1, the obtained BPAs are displayed. A screenshot of the web application is shown in Figure 6.

Case	$m_p$	$m_f$	$m_{pf}$
BC <sub>&lt;40</sub>	0.020	0.043	0.006
BC <sub>&lt;50</sub>	0.075	0.030	0.015
BC <sub>≥50</sub>	0.034	0.108	0.024
BC <sub>bilateral</sub>	0.030	0.088	0.017
BC <sub>male</sub>	0.037	0.138	0.033
OC	0.151	0.104	0.103
BC <sub>&lt;50</sub> , OC	0.058	0.054	0.020
BC <sub>≥50</sub> , OC	0.119	0.245	0.191
$\Omega$	0.476	0.190	0.591

Table 1: Aggregated BPAs from publications

Due to limited data availability, only few scenarios were covered by this model (cf. Table 3). Additionally, poor source data quality led to lower weights in some cases, which in turn reduced the mutation risk significantly in comparison with the existing models. For example, the risk of a person whose father was diagnosed with breast cancer before the age of 50 would be calculated as  $Bel_{m_{pf}}(\{BC_{<50}, BC_{male}\}) = m_{pf}(BC_{<50}) + m_{pf}(BC_{male}) = 4.8\%$ , as compared to 10 – 18% predicted by PENN II. That is, better data are certainly needed but the frequencies can definitely be obtained as proposed in Subsection 3.3.

Our limited test data allowed a differentiation per case only between “below the age of 50” and “above the age of 50” at best. Since the age of onset [Buys et al. 2017] strongly influences the likelihood of being a carrier of a pathogenic *BRCA1/2* variant, we decided that a more granular approach to handling the overall risk as a function in dependence of a patient’s age was needed. A model for this is proposed in the following section.

In this first implementation, data acquisition for testing was performed manually. We found that there was still a need for more sophisticated AI-based text mining tools if the process is to be automated. While the problem of table extraction from text (PDF files in our case) has been solved<sup>19</sup>, automatic interpretation of extracted data in a certain context is a challenge yet to be overcome and a topic for future research.

<sup>19</sup> cf. <https://cran.r-project.org/web/packages/PDE/vignettes/PDE.html>

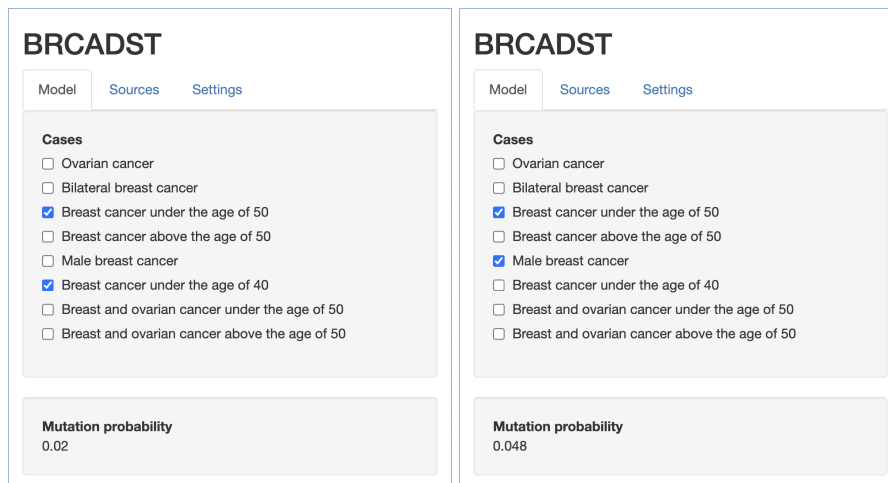


Figure 6: Screenshots of the web app based on the data structure from Subsection 3.3 for the disease pattern considered in Subsection 2.3 (left) and a further pattern from this subsection

#### 4 Modeling the *BRCA1/2* Mutation Probabilities as a Function of Age

As mentioned above, it is necessary to differentiate better in dependence on age in the developed DST model to obtain more realistic results. Since we were not able to arrive at ground truth for HBOC risk factors from the databases available on the internet due to privacy restrictions (cf. Subsection 3.2), we use the manually collected information from OA publications [Frank et al. 2002, Buys et al. 2017], supplemented by predictions from PENN II where necessary, directly. In this section, we present a proof of concept DST approach to assess the probability of *BRCA1/2* mutation based on finer age models. We consider it to be merely a proof of concept since the data we rely on are selective and possibly do not reflect ground truth yet. Nonetheless, we are able to achieve good correspondence with the state of the art systems, which are based on (partially) undisclosed data (cf. Subsection 2.3). Note that better data can be directly incorporated into the proposed model.

The age at the first cancer diagnosis is one of the most important indicators for the presence of a *BRCA1/2* mutation [Buys et al. 2017]. As suggested in [Auer and Luther 2022], the mutation probability can be modeled by combining age-based cumulative percentage curves for different risk factors. In this paper, we extend this idea by utilizing the Dempster-Shafer theory of evidence as the basis for combining multiple risk factors in dependence on the patient's age and the age of his/her relatives with a history of *BRCA1/2* -related cancer.

We consider the risk factors BC, male breast cancer (mBC), bilateral breast cancer (bBC), OC, breast and ovarian cancer (BCOC) in the history, BC and OC cancer in the same person (BCOCsp), and ethnicity (E) for ages between 60 and 20 in steps of 5 years. At the moment, we only differentiate between two ethnicities: general and Ashkenazi-Jewish (AJ). The masses for BPAs at each given age are shown in Table 2. Note that the

curves for risk indicators BCOC and BCOCsp are obtained by considering the sum of values for BC and OC and the additive factor sp, respectively. For all ages in the case of Ashkenazi-Jewish ethnicity, additive factors of 0.05 are used for BC and bBC; of 0.03 for OC; and of 0.01 for sp. For obtaining the cumulative probability curves shown in Figure 7 (general on the left, AJ on the right), we interpolate linearly between the values for the ages (as is usually done in statistics). In the same way as in the implementation from Subsection 3.4, we assume that the focal element  $\Omega$  containing all the considered risk factors is assigned the remainder of the probability since we compute only the lower bound on the risk (which, however, can itself be an interval, cf. Subsection 2.2). All other subsets of  $\Omega$  aside from those in Table 2 (and  $\Omega$ ) are supposed to have zero masses.

Age	BC	bBC	mBC	OC	sp
60	0.04	0.02	0.09	0.03	0.03
55	0.015	0.02	0.09	0.05	0.02
50	0.015	0.02	0.09	0.07	0.02
45	0.02	0.02	0.09	0.09	0.02
40	0.02	0.02	0.09	0.11	0.02
35	0.02	0.02	0.09	0.13	0.02
30	0.03	0.02	0.09	0.15	0.02
25	0.04	0.02	0.09	0.17	0.02
20	0.04	0.02	0.09	0.19	0.02

Table 2: BPAs for the chosen risk factors wrt. age

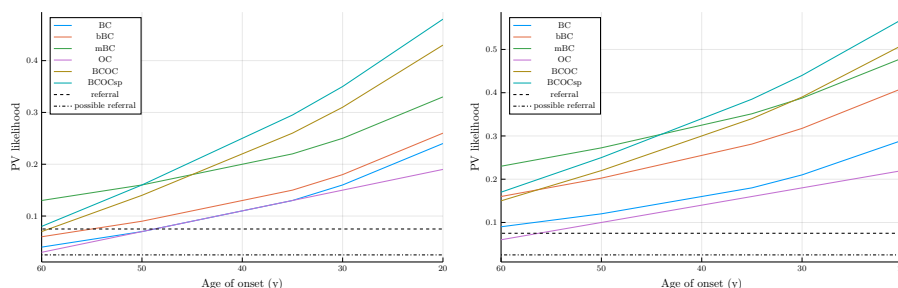


Figure 7: Cumulative risk curves for the mutation probabilities: general (left) and AJ (right) case

In Figure 7, there are two lines showing the threshold values for considered and recommended gene test referral (at 2.5% and 7.5%, respectively) that are given by a relatively recent meta-study [Pujol et al. 2021]. As the cumulative curves show, practically all individual mutation probabilities (e.g., as reported in [Buys et al. 2017]) are higher than those thresholds. In our opinion, it is necessary to reconsider those values, at least, if such models as PENN II are to be viewed as trustworthy.

Our model works as follows: For each examined pattern, one BPA is usually constructed for the individual ( $m_p$ ) and a further one for the person in her family history ( $m_h$ ) based on the cumulative curves in dependence on the ages given by a disease pattern and the risk indicator ( $m_i$  for  $i \in \{BC, bBC, mBC, OC, sp\}$ ). In cases where the age of onset for a family member is not known exactly, we use intervals for the probabilities (IBPAs). Because of the monotonic properties of the cumulative risk curves, we can assume the interval interpretation of the mass function  $m_i(\cdot)$  for a risk factor  $i$  to be  $[m_i(\bar{a}), m_i(\underline{a})]$ ,  $[\underline{a}, \bar{a}]$  being the age interval. If there are more than one individual mentioned in the family history, it is possible to consider additional BPAs for each mentioned person. At the moment, all (I)BPAs have the same importance and are merged using Dempster's rule of combination (cf. Eq. (3)). However, we plan to implement the possibility to apply Mendelian logic for aggregating since there is no difficulty to realize it in our model in principle (e.g., using the mixing and averaging rule, cf. Eq. (4)). In the final step, the belief function is calculated as given by Eq. (2), on the right, for all applicable risk factors employing the combined mass assignment  $m_{pf}$ .

We apply our model ("new") to the example disease patterns described below and compare them with the results from [Auer and Luther 2022] ("old"), with the app described in Subsection 3.4 ("app") where possible and with those from PENN II, BOADICEA and Frank tables where available (cf. Table 3). The shown intervals are rounded outwards to three decimal positions; crisp numbers are rounded to three decimal positions using rounding to the next number where necessary.

#### Non-Ashkenazi-Jewish

**Pattern 1:** Father with BC at 42, second degree relative with BC > 50

**Pattern 2:** Patient with BC < 40 and mother with BC < 50

**Pattern 3:** Patient with OC < 40 and her aunt with BC < 50

**Pattern 4:** Patient aged 22 with BC and OC; mother with bBC > 50

#### Ashkenazi-Jewish

**Pattern 5:** Patient (30-40 years of age) with BC, aunt with BC > 51

**Pattern 6:** Patient with BC at 45 years of age, sister with OC < 50

**Pattern 7:** Patient with OC < 50 and OC and BC > 50 in the family history

**Pattern 8:** Patient with OC and BC at 35 years of age, aunt with bBC over 50

The results for the new model show a good agreement with those from [Frank et al. 2002] and PENN II, which is not surprising: Although the models are different, we still rely on and incorporate data from those two sources into our model as ground truth. The results from BOADICEA show a good agreement for Patterns 4 and 5 only (they are also quite different from PENN II). Note that BOADICEA is designed to assess the general risk of contracting breast/ovarian cancer from a variety of genetic and non-genetic indicators, among others due to hereditary factors. That means it is not easy to reflect the considered disease patterns in exactly the same way as described above since a lot more questions about the patient need to be answered. Besides, both PENN II and BOADICEA are very sensitive wrt. the age of onset for the cancer cases (making the need of interval representations even more important). That is, the values for PENN II and BOADICEA

Pa.	New	Old	PENN II	BOADICEA	Frank	App
1	[0.251, 0.294]	0.257	0.27	0.059		0.072
2	[0.116, 0.317]	0.198	0.15 – 0.35	0.073	0.297	0.044
3	[0.232, 0.584]	0.358	0.12 – 0.4	0.036		0.137
4	[0.446, 0.549]	0.527	0.54	0.491		0.184
5	[0.149, 0.275]	0.319	0.24 – 0.35	0.298	0.318	
6	[0.296, 0.406]	0.313	0.31	0.662	0.415	
7	[0.225, 0.410]	0.417	0.41	0.727	0.412	
8	[0.459, 0.603]	[0.518, 0.656]	0.53	0.937		

Table 3: *BRCA1/2* mutation probabilities for the example disease patterns

from Table 3 are provided as a background reference and reflect the results that might be obtained by a non-expert patient filling out the respective questionnaires; they might change if a gene expert operates the interfaces. The results from the app (last column) show poor agreement systematically underestimating the risk for reasons explained in Subsection 3.4, although they are based on the same model as in [Auer and Luther 2022] (Column “old” with a medium to good agreement), demonstrating the importance of good quality data on ground truth.

The advantage of the new model proposed in this paper is that it now also supplies the lower (upper) bound on the belief function and not just an average or median value. Besides, since it can be made to work with the database from Subsection 3.3, it can incorporate traceable data. In this way, decisions about the computed probabilities can be explained comprehensively.

## 5 Conclusions and Future Work

In this paper, we made a first step towards a standard for collecting and storing HBOC-related data with a focus on its flexibility and reliability through provenance as well as methods with result verification. In Subsections 3.2 and 3.3, we suggested two database designs for data on HBOC, one containing classified information and one avoiding the restrictions due to privacy concerns through aggregation, respectively. The data extraction approach from 3.2 was designed but not yet implemented for HBOC, exactly because of privacy legislation restrictions. In Subsection 3.4, we implemented the design from 3.3 to incorporate data from several OA publications. We showed that the proposed database structure was well suited for obtaining aggregated frequencies for risk indicators needed, for example, in the context of a DST model for computing *BRCA1/2* mutation probabilities first introduced in [Auer and Luther 2021]. Lessons learned from working with this model were that a more graded approach to modeling risk indicators in dependence on the age of first cancer occurrence was necessary, resulting in the two-stage DST based technique presented in Section 4. It takes into account epistemic uncertainty by considering intervals if ages of persons from the family history are not known exactly and employs IA with result verification for all involved (DST-related) operations.

Although this implementation shows a good agreement with established tools while providing more information on *BRCA1/2* mutation probabilities through working with interval BPAs, there is still room for improvement. For example, this approach does not take into account the degree of relationship, that is, individual and family history are treated as equally relevant at the moment. Future improvements might include possible redistribution using further combination rules (aside from Dempster’s rule) to account for inheritance probability. Another interesting topic for future research is to apply interval

arithmetic type 2 in the context of this DST approach and compare with the ‘normal’ IA version wrt. both complexity and prediction accuracy.

Yet another point is that the issue of finding ground truth about HBOC syndrome from OA sources automatically is still largely unresolved due to multiple reasons starting with privacy issues and ending with the lack of related standards. As a manageable topic for future research, we will study the possibility of employing customized AI for data extraction about HBOC from OA papers.

## References

- [Arfaoui and Akaichi 2015] Arfaoui, N., Akaichi, J.: “Automating Schema Integration Technique Case Study: Generating Data Warehouse Schema from Data Mart Schemas”; Kozielski, S., Mrozek, D., Kasprowski, P., Malysiak-Mrozek, B., Kostrzewa, D. (eds): Proceedings of 11<sup>th</sup> International Conference: Beyond Databases, Architectures and Structures, Communications in Computer and Information Science, 521, Springer, 2015, 200–209.
- [Ashton-Prolla and Vargas 2014] Ashton-Prolla, P., Vargas, F.: “Prevalence and impact of founder mutations in hereditary breast cancer in Latin America”; *Genet. Mol. Biol.*, 37(1), 2014, 234–240.
- [Auer and Luther 2021] Auer, E., Luther, W.: “Uncertainty Handling in Genetic Risk Assessment and Counseling”; *JUCS - Journal of Universal Computer Science*, 27(12), 2021, 1347–1370.
- [Auer and Luther 2022] Auer, E., Luther, W.: “Dempster-Shafer Theory Based Uncertainty Models for Assessing Hereditary, BRCA1/2-Related Cancer Risk”; The 8<sup>th</sup> International Symposium on Reliability Engineering and Risk Management, 2022, 755–762.
- [Auer et al. 2010] Auer, E., Luther, W., Rebner, G., Limbourg, Ph.: “A Verified MATLAB Toolbox for the Dempster-Shafer Theory”; *Proc. of the Workshop on the Theory of Belief Functions*, 2010.
- [Ayyub and Klir 2006] Ayyub, B., Klir, G.: *Uncertainty Modeling and Analysis in Engineering and the Sciences*; 2006.
- [Bönisch et al. 2022] Bönisch, C., Kesztyüs, D., Kesztyüs, T.: “Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata”; *Scientific Data*, 2022, 659.
- [Bradley 2019] Bradley, S.: “Imprecise Probabilities”; Zalta, E. N. (ed): *The Stanford Encyclopedia of Philosophy*, Spring 2019 edn, Metaphysics Research Lab, Stanford University.
- [Buys et al. 2017] Buys, S. S., Sandbach, J. F., Gammon, A., Patel, G., Kidd, J., Brown, K. L., Sharma, L., Saam, J., Lancaster, J., Daly, M. B.: “A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes”; *Cancer*, 123(10), 2017, 1721–1730.
- [Claus et al. 1994] Claus, E. B., Risch, N., Thompson, W. D.: “Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction”; *Cancer*, 73(3), 1994, 643–651.
- [de Figueiredo and Stolfi 2004] de Figueiredo, L. H., Stolfi, J.: “Affine Arithmetic: Concepts and Applications”; *Numerical Algorithms*, 34(1–4), 2004, 147–158.
- [Denney et al. 2016] Denney, M., Long, D., Armistead, M., Anderson, J., Conway, B.: “Validating the Extract, Transform, Load Process Used to Populate a Large Clinical Research Database”; *International Journal of Medical Informatics*, 94(07), 2016.
- [Desrochers and Jaulin 2017] Desrochers, B., Jaulin, L.: “Thick Set Inversion”; *Artificial Intelligence*, 249, 2017, 1–18.
- [Dong et al. 2015] Dong, X. L., Berti-Équille, L., Srivastava, D.: “Data Fusion: Resolving Conflicts from Multiple Sources”; *CoRR*, 2015.

- [Ferson et al. 2003] Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D. S., Sentz, K.: Constructing Probability Boxes and Dempster-Shafer Structures; Washington, D.C: United States. Dept. of Energy, 2003.
- [Frank et al. 2002] Frank, T. S., Deffenbaugh, A. M., Reid, J. E., Hulick, M., Ward, B. E., Lingenfelter, B., Gumpfer, K. L., Scholl, T., Tavtigian, S. V., Pruss, D. R., Critchfield, G. C.: “Clinical Characteristics of Individuals With Germline Mutations in BRCA1 and BRCA2: Analysis of 10,000 Individuals”; *J. Clin. Oncol.*, 20(6), 2002, 1480–1490.
- [Gierend et al. 2023] Gierend, K., Wodke, J. A.H., Genehr, S., Gött, R., Henkel, R., Krüger, F., Mandalka, M., Michaelis, L., Scheuerlein, A., Schröder, M., Zeleke, A., Waltemath, D.: “TAPP: Defining standard provenance information for clinical research data and workflows – Obstacles and opportunities”; *Companion Proceedings of the ACM Web Conference 2023*, ACM, 2023.
- [Guerrini et al. 2017] Guerrini, C. J., McGuire, A. L., Majumder, M. A.: “Myriad take two: Can genomic databases remain secret?”; *Science*, 356(6338), 2017, 586–587.
- [Hall et al. 2009] Hall, M. J., Reid, J. E., Burbidge, L. A., Pruss, D., Deffenbaugh, A. M., Frye, C., Wenstrup, R. J., Ward, B. E., Scholl, Th. A., Noll, W. W.: “BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer”; *Cancer*, 115(10), 2009, 2222–2233.
- [Harutyunyan et al. 2019] Harutyunyan, A. N., V. Poghosyan, A., M. Grigoryan, N., A. Hovhannisyan, N., Kushmerick, N.: “On Machine Learning Approaches for Automated Log Management”; *JUCS - Journal of Universal Computer Science*, 25(8), 2019, 925–945.
- [Kahan 1968] Kahan, W.: A more complete interval arithmetic. Lecture notes for a summer course, University of Toronto, 1968, Canada.
- [Lee et al. 2019] Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., de Villiers, C. B., Izquierdo, A., Simard, J., Schmidt, M. K., Walter, F. M., Chatterjee, N., Garcia-Closas, M., Tischkowitz, M., Pharoah, P., Easton, D. F., Antoniou, A. C.: “BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors”; *Genetics in Medicine*, 21(8), 2019, 1708–1718.
- [Li Lee and Ling 1995] Li Lee, M., Ling, T. W.: “Resolving structural conflicts in the integration of Entity-Relationship schemas”; Papazoglou, Michael P. (ed), *OOER '95: Object-Oriented and Entity-Relationship Modeling*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, 424–433.
- [Lindor et al. 2010] Lindor, N. M., Johnson, K. J., Harvey, H., Shane Pankratz, V., Domchek, S. M., Hunt, K., Wilson, M., Cathie Smith, M., Couch, F.: “Predicting BRCA1 and BRCA2 gene mutation carriers: comparison of PENN II model to previous study”; *Familial Cancer*, 9(4), 2010, 495–502.
- [Lohner 2001] Lohner, R.: “On the Ubiquity of the Wrapping Effect in the Computation of Error Bounds”. Kulisch, U., Lohner, R., Facius, A. (eds), *Perspectives on Enclosure Methods*. Springer-Verlag, 2001, 201–218.
- [Makino and Berz 2004] Makino, K., Berz, M.: Suppression of the Wrapping Effect by Taylor Model-Based Validated Integrators; MSUHEP 40910. Department of Physics, Michigan State University, East Lansing, MI 48824, 2004.
- [Martínez-García and Hernandez-Lemus 2022] Martínez-García, M., Hernandez-Lemus, E.: “Data Integration Challenges for Machine Learning in Precision Medicine”; *Frontiers in Medicine*, 8(1), 2022.
- [Merheb et al. 2013] Merheb, R., Mora, L., Palomo del Barrio, E.: “Parameter estimation in an uncertain and noisy environment via set inversion”; *IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES)*, 2013, 125-132.
- [Missier 2017] Missier, P.: “Provenance Standards”; Liu, L., Özsu, M. T. (eds), *Encyclopedia of Database Systems*, Springer New York, 2017, 1–8.

- [Moore et al. 2009] Moore, R. E., Kearfott, R. B., Cloud, M. J.: Introduction to Interval Analysis; Philadelphia: Society for Industrial and Applied Mathematics, 2009.
- [Moreau et al. 2008] Moreau, L., Ludäscher, B., Altintas, I., Barga, R. S., Bowers, S., Callahan, S., Chin, Jr., G., Clifford, B., Cohen, S., Cohen-Boulakia, S., Davidson, S., Deelman, E., Digiampietri, L., Foster, I., Freire, J., Frew, J., Futrelle, J., Gibson, T., Gil, Y., Goble, C., Golbeck, J., Groth, P., Holland, D. A., Jiang, S., Kim, J., Koop, D., Krenek, A., McPhillips, T., Mehta, G., Miles, S., Metzger, D., Munroe, S., Myers, J., Plale, B., Podhorszki, N., Ratnakar, V., Santos, E., Scheidegger, C., Schuchardt, K., Seltzer, M., Simmhan, Y. L., Silva, C., Slaughter, P., Stephan, E., Stevens, R., Turi, D., Vo, H., Wilde, M., Zhao, J., Zhao, Y.: “Special Issue: The First Provenance Challenge”; *Concurr. Comput. : Pract. Exper.*, 20(5), 2008, 409–418.
- [Naumann et al. 2006] Naumann, F., Bilke, A., Bleiholder, J., Weis, M.: “Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies”; *IEEE Data Eng. Bull.*, 29(2), 2006 21–31.
- [Neumaier 2003] Neumaier, A.: “Taylor Forms – Use and Limits”; *Reliable Computing*, 9(1), 2003, 43–79.
- [Pasquier et al. 2017] Pasquier, T., Lau, M. K., Trisovic, A., Boose, E. R., Couturier, B., Crosas, M., Ellison, A. M., Gibson, V., Jones, C. R., Seltzer, M.: “If these data could talk”; *Scientific Data*, 4(1), 2017, 170114.
- [Piegat and Dobryakova 2020] Piegat, A., Dobryakova, L.: “A Decomposition Approach to Type 2 Interval Arithmetic”; *International Journal of Applied Mathematics and Computer Science*, 30(03), 2020, 185–201.
- [Plon et al. 2008] Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., Hogervorst, F. B. L., Hoogerbrugge, N., Spurdle, A. B., Tavtigian, S. V., et al.: “Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results”; *Human mutation*, 29(11), 2008, 1282–1291.
- [Pujol et al. 2021] Pujol, P., Barberis, M., Beer, P., Friedman, E., Piulats, J. M., Capoluongo, E.D., Garcia Foncillas, J., Ray-Coquard, I., Penault-Llorca, F., Foulkes, W. D., Turnbull, C., Hanson, H., Narod, S., Arun, B. K., Aapro, M. S., Mandel, J.-L., Normanno, N., Lambrechts, D., Vergote, I., Anahory, M., Baertschi, B., Baudry, K., Bignon, Y.-J., Bollet, M., Corsini, C., Cussenot, O., De la Motte Rouge, T., M.Dubois de Labarre, Duchamp, F., Duriez, C., Fizazi, K., Galibert, V., Gladieff, L., Gligorov, J., Hammel, P., Imbert-Bouteille, M., Jacot, W., Kogut-Kubiak, T., Lamy, P.-J., Nambot, S., Neuzillet, Y., Olschwang, S., Rebillard, X., Rey, J.-M., Rideau, C., Spano, J.-P., Thomas, F., Treilleux, I., Vandromme, M., Vendrell, J., Vintraud, M., Zarca, D., Hughes, K. S., Alés Martínez, J. E.: “Clinical practice guidelines for *BRCA1* and *BRCA2* genetic testing”; *European Journal of Cancer*, 146, 2021, 30–47.
- [Schäfer 2023] Schäfer, M. S.: “The Notorious GPT: science communication in the age of artificial intelligence”; *Journal of Science Communication*, 22(02), 2023, Y02.
- [Shafer 1976] Shafer, G.: *A Mathematical Theory of Evidence*; Princeton: Princeton University Press, 1976.
- [Sparks 2011] Sparks, G.: *Methods and Tools*. Software Development Magazine - Project Management, Programming, Software Testing, 2011.
- [Tang et al. 2023] Tang, Y., Zhang, X., Zhou, Y., Huang, Y., Zhou, D.: “A new correlation belief function in Dempster-Shafer evidence theory and its application in classification”; *Scientific Reports*, 13(1), 2023, 2045–2322.
- [Wang et al. 2021] Wang, X.-M., Zhang, X.-R., Li, Z.-H., Zhong, W.-F., Yang, P., Mao, C.: “A brief introduction of meta-analyses in clinical practice and research”; *The Journal of Gene Medicine*, 23(5), 2021, e3312.