Do We Really Sample Right in Model-Based Diagnosis?

Patrick Rodler[1] and **Fatima Elichanova**[2]

University of Klagenfurt

[1]patrick.rodler@aau.at
[2]fatimael@edu.aau.at

Abstract

Statistical samples, in order to be representative, have to be drawn from a population in a random and unbiased way. Nevertheless, it is common practice in the field of model-based diagnosis to make estimations from (biased) best-first samples. One example is the computation of a few most probable possible fault explanations for a defective system and the use of these to assess which aspect of the system, if measured, would bring the highest information gain.

In this work, we scrutinize whether these statistically not well-founded conventions, that both diagnosis researchers and practitioners have adhered to for decades, are indeed reasonable. To this end, we empirically analyze various sampling methods that generate fault explanations. We study the representativeness of the produced samples in terms of their estimations about fault explanations and how well they guide diagnostic decisions, and we investigate the impact of sample size, the optimal trade-off between sampling efficiency and effectivity, and how approximate sampling techniques compare to exact ones.

1 Introduction

Suppose we intend to predict the outcome of the next election and we conduct a poll where we ask only, say, university professors for whom they are going to vote. By this strategy, we will most likely not gain insight into the real sentiment in the population regarding the election. The problem is simply that professors are most probably not representative of all people. In model-based diagnosis, however, such kind of samples are often used as a basis for making decisions that rule the efficiency of the diagnostic process.

Model-based diagnosis [1; 2] deals with the detection, localization and repair of faults in observed systems such as programs, circuits, knowledge bases or physical devices of various kinds. One important prerequisite to achieve these goals is the generation of *diagnoses*, i.e., explanations for the faulty system behavior in terms of potentially faulty system components. A sample of diagnoses can be either

- *directly analyzed*, e.g., to manually discover or make estimations about the actual fault [3; 4], to aid proper algorithm choice [5; 6], or to support users in test case specifications or repair actions [7; 8; 9; 10; 11], or

- *used as an input or guidance to diagnostic algorithms*,

where we focus on the second bullet in this work.

An important class of diagnostic algorithms that are guided by a set of precomputed diagnoses are *sequential diagnosis (SD)* approaches [2; 12]. They use a sample of diagnoses to compute optimal system measurements that allow to efficiently and systematically rule out invalid diagnoses until a single or highly probable one remains. Since achieving (global) optimality of the sequence of system measurements is intractable in general [13], state-of-the-art SD techniques usually rely on local optimization [14] using one out of numerous *heuristics* [2; 15; 16; 17; 18] as optimality criteria. These heuristics can be seen as functions that, based on a given sample of diagnoses, map measurement candidates to one numeric score each, and finally select the one measurement with the best score. In most cases, these functions use two features of the sample:[1]

(F1) the *diagnoses' probabilities* (to estimate the probability of each measurement outcome), and

(F2) the *diagnoses' predictions of the measurement outcome* (to estimate diagnosis elimination rates).

Literature offers a wide range of different methods and ways to generate samples of diagnoses, among them ones that return a *specific sample* (which includes exactly a predefined subset of all diagnoses) and others that compute an *unspecific sample*, e.g., in a heuristic [20], stochastic [21] or simply undefined way [22][2] (where no guarantee can be given regarding diagnosis selection for the sample).

Many existing SD approaches draw on samples of the specific type in that they build upon *best-first* samples, such as maximum-probability or minimum-cardinality diagnoses [9; 16; 19; 23; 24; 25; 26]. While perhaps often being motivated by the desideratum that the most preferred/likely candidate(s) should be known at any stage of the diagnostic process, e.g., to allow for well-founded stopping criteria, the use of such non-random samples is highly questionable from the statistical viewpoint.

In this work we challenge the validity of the following statistical law in the domain of model-based diagnosis:

A randomly chosen unbiased sample from a population

---

[1]For this work, we make the assumption that measurements have uniform costs. If that is not the case, then the measurement cost is another factor that flows into the assessment of measurement candidates [15; 19].

[2]Note, the algorithm described in [22] can be modified for heuristic diagnosis computation, as will be explained in Sec. 3.2.

allows (on average) better conclusions and estimations about the whole population than any other sample.

The motivation behind this inquiry is to either understand why this fundamental principle does not apply to the particular domain of model-based diagnosis, or to rationalize the necessity of random sampling as part of diagnostic algorithms and to foster research in this direction.

The particular contributions are:

- We analyze a range of real-world diagnosis problems and gain insight into the quality of *three specific* (best-first, random and, as a baseline, worst-first) *and three unspecific* (approximate best-first, approximate random, approximate worst-first) *sample types*.
- We assess a sample type's quality based on
  - its *theoretical representativeness*, i.e., how well it allows to estimate the aspects (F1 and F2) that determine the heuristic score of measurements, and
  - its *practical representativeness*, i.e., its performance achieved in a diagnosis session wrt. time and number of measurements.
- We investigate the *impact of the*
  - *sample size*,
  - particular used *heuristic*, and
  - tackled *diagnosis problem*

  on the sample's representativeness.

This work is organized as follows: We provide a brief account of theoretical foundations in Sec. 2. Our evaluations (dataset, sample types, sampling techniques, evaluation criteria, research questions, experiment settings, and results) are discussed in Sec. 3. In Sec. 4, we address limitations of our research, and conclude with Sec. 5.

2 Preliminaries

We briefly characterize concepts from model-based diagnosis used in this work, based on the framework of [16; 9] which is more general [27] than Reiter's theory [1].[3]

Diagnosis Problem. We assume that the diagnosed system, consisting of a set of components $\{c_1, \ldots, c_k\}$, is described by a finite set of logical sentences $\mathcal{K} \cup \mathcal{B}$, where $\mathcal{K}$ (possibly faulty sentences) includes knowledge about the behavior of the system components, and $\mathcal{B}$ (correct background knowledge) comprises any additional available system knowledge and system observations. More precisely, there is a one-to-one relationship between sentences $ax_i \in \mathcal{K}$ and components c_i, where ax_i describes the nominal behavior of c_i (*weak fault model*). E.g., if c_i is an AND-gate in a circuit, then $ax_i := out(c_i) = and(in1(c_i), in2(c_i))$; $\mathcal{B}$ in this case might contain sentences stating, e.g., which components are connected by wires, or observed circuit outputs. The inclusion of a sentence ax_i in $\mathcal{K}$ corresponds to the (implicit) assumption that c_i is healthy. Evidence about the system behavior is captured by sets of positive (P) and negative (N) measurements [1; 2; 28]. Each measurement is a logical sentence; positive ones $p \in P$ must be true and negative ones $n \in N$ must not be true. The former can be, depending on the context, e.g., observations about the system, probes or required system properties. The latter model properties that must not hold for the system, e.g., if $\mathcal{K}$ is

---

[3]The main reason for using this more general framework is its ability to handle negative measurements (things that must *not* be true for the diagnosed system) which are helpful, e.g., for diagnosing knowledge bases [28; 16].

$$\mathcal{K} = \begin{array}{llll} \{ax_1 : A \to \neg B & ax_2 : A \to B & ax_3 : A \to \neg C \\ ax_4 : B \to C & ax_5 : A \to B \vee C & \} \end{array}$$

$\mathcal{B} = \emptyset$	$P = \emptyset$	$N = \{\neg A\}$

Table 1: Example DPI stated in propositional logic.

a biological knowledge base to be debugged, a negative test case might be "every bird can fly" (think of penguins). We call $\langle \mathcal{K}, \mathcal{B}, P, N \rangle$ a *diagnosis problem instance (DPI)*.

Example 1 *(DPI)* Table 1 depicts a DPI stated in propositional logic. The "system" (the knowledge base itself in this case) comprises five "components" $c_1, \ldots, c_5$, and the "normal behavior" of c_i is given by the respective axiom $ax_i \in \mathcal{K}$. No background knowledge ($\mathcal{B} = \emptyset$) or positive measurements ($P = \emptyset$) are given from the start. But, there is one negative measurement (i.e., $N = \{\neg A\}$), which stipulates that $\neg A$ must *not* be an entailment of the correct system (knowledge base). Note, however, that $\mathcal{K}$ (i.e., the assumption that all "components" are normal) in this case does entail $\neg A$ (e.g., due to the axioms ax_1, ax_2) and therefore some axiom ("component") in $\mathcal{K}$ must be faulty. □

Diagnoses. Given that the system description along with the positive measurements (under the assumption $\mathcal{K}$ that all components are healthy) is inconsistent, i.e., $\mathcal{K} \cup \mathcal{B} \cup P \models \bot$, or some negative measurement is entailed, i.e., $\mathcal{K} \cup \mathcal{B} \cup P \models n$ for some $n \in N$, some assumption(s) about the healthiness of components, i.e., some sentences in $\mathcal{K}$, must be retracted. We call such a set of sentences $\mathcal{D} \subseteq \mathcal{K}$ a *diagnosis* for the DPI $\langle \mathcal{K}, \mathcal{B}, P, N \rangle$ iff $(\mathcal{K} \setminus \mathcal{D}) \cup \mathcal{B} \cup P \not\models x$ for all $x \in N \cup \{\bot\}$. We say that $\mathcal{D}$ is a *minimal diagnosis* for dpi iff there is no diagnosis $\mathcal{D}' \subset \mathcal{D}$ for dpi. The set of minimal diagnoses is representative of all diagnoses under the weak fault model [29], i.e., the set of all diagnoses is equal to the set of all supersets of minimal diagnoses. Therefore, diagnosis approaches usually restrict their focus to only minimal diagnoses. We furthermore denote by $\mathcal{D}^*$ the *actual diagnosis* which pinpoints the actually faulty axioms, i.e., all elements of $\mathcal{D}^*$ are in fact faulty and all elements of $\mathcal{K} \setminus \mathcal{D}^*$ are in fact correct.

Example 2 *(Diagnoses)* For our DPI in Table 1 we have four minimal diagnoses, given by $\mathcal{D}_1 := [ax_1, ax_3]$, $\mathcal{D}_2 := [ax_1, ax_4]$, $\mathcal{D}_3 := [ax_2, ax_3]$, and $\mathcal{D}_4 := [ax_2, ax_5]$. For instance, $\mathcal{D}_1$ is a minimal diagnosis as $(\mathcal{K} \setminus \mathcal{D}_1) \cup \mathcal{B} \cup P = \{ax_2, ax_4, ax_5\}$ is both consistent and does not entail the given negative measurement $\neg A$. □

Diagnosis Probability Model. In case useful meta information is available that allows to assess the likeliness of failure for system components, the probability of diagnoses (of being the actual diagnosis) can be derived. Specifically, given a function p that maps each sentence (system component) $ax \in \mathcal{K}$ to its failure probability $p(ax) \in (0, 1)$, the probability $p(\mathcal{D})$ of a diagnosis $\mathcal{D} \subseteq \mathcal{K}$ (under the common assumption of independent component failure) is computed as the probability that all sentences in $\mathcal{D}$ are faulty, and all others are correct, i.e., $p(\mathcal{D}) := \prod_{ax \in \mathcal{D}} p(ax) \prod_{ax \in \mathcal{K} \setminus \mathcal{D}} (1 - p(ax))$. Each time a new measurement is added to the DPI, the probabilities of diagnoses are updated using Bayes' Theorem [2].

Example 3 *(Diagnosis Probabilities)* Reconsider the DPI from Table 1 and let probabilities $\langle p(ax_1), \ldots, p(ax_5) \rangle = \langle .1, .05, .1, .05, .15 \rangle$. Then, the probabilities of all minimal diagnoses from Example 2 are $\langle p(\mathcal{D}_1), \ldots, p(\mathcal{D}_4) \rangle =$

$\langle.0077, .0036, .0036, .0058\rangle$. E.g., $p(\mathcal{D}_1)$ is calculated as $.1 * (1 - .05) * .1 * (1 - .05) * (1 - .15)$. The normalized diagnosis probabilities would then be $\langle.37, .175, .175, .28\rangle$. Note, this normalization makes sense if only a proper subset of all diagnoses is known. □

Measurement Points. We call a logical sentence a *measurement point (MP)* if it states one (true or false) aspect of the system under consideration. E.g., if the system is a digital circuit, the statement $out(c_i) = 1$, which states that the output of gate c_i is high, is an MP. In case of the system being, say, a biological knowledge base, $\forall X (bird(X) \rightarrow canFly(X))$ is an MP. Assuming an oracle orcl (e.g., an electrical engineer for a circuit, or a domain expert for a knowledge base) that is knowledgeable about the system, one can send to orcl MPs m and orcl will classify each m as either a positive or a negative measurement, i.e., $m \mapsto \text{orcl}(m)$ where $\text{orcl}(m) \in \{P, N\}$.

Measurements to Discriminate between Diagnoses. MPs come into play when there are multiple diagnoses for a DPI and the intention is to figure out the actual diagnosis. Hence, given a set of diagnoses $\mathbf{D}$ for a DPI between which we want to discriminate, the MPs m of particular interest are those for which each classification $\text{orcl}(m)$ is inconsistent with some diagnosis in $\mathbf{D}$ [2; 9]. We call such MPs *informative* (wrt. $\mathbf{D}$). In other words, each outcome of a measurement for some informative MP will invalidate some diagnosis.

Each MP m partitions any set of (minimal) diagnoses $\mathbf{D}$ into subsets $\mathbf{D}_m^+$, $\mathbf{D}_m^-$ and $\mathbf{D}_m^0$ as follows:

- Each $\mathcal{D} \in \mathbf{D}_m^+$ is consistent (only) with $\text{orcl}(m) = P$. (*diagnoses predicting the positive outcome*)
- Each $\mathcal{D} \in \mathbf{D}_m^-$ is consistent (only) with $\text{orcl}(m) = N$. (*diagnoses predicting the negative outcome*)
- Each $\mathcal{D} \in \mathbf{D}_m^0$ is consistent with both outcomes $\text{orcl}(m) \in \{P, N\}$. (*uncommitted diagnoses*)

Thus, an MP m is informative iff both $\mathbf{D}_m^+$ (diagnoses invalidated if $\text{orcl}(m) = N$) and $\mathbf{D}_m^-$ (diagnoses invalidated if $\text{orcl}(m) = P$) are non-empty sets.

(Estimated) Properties of Measurement Points. Since not all informative MPs are equally utile, the consideration of additional properties of MPs allows a more fine-grained preference rating of MPs. In fact, if $\mathbf{D}$ includes all diagnoses for the given DPI, the partition $\langle \mathbf{D}_m^+, \mathbf{D}_m^-, \mathbf{D}_m^0 \rangle$ allows to determine, for each measurement outcome $c \in \{P, N\}$, its *diagnosis elimination rate* $er(\text{orcl}(m) = c)$ as well as its *probability* $p(\text{orcl}(m) = c)$ [17]:

$$er_m^+ := er(\text{orcl}(m) = P) = \frac{|\mathbf{D}_m^-|}{|\mathbf{D}|}$$

$$er_m^- := er(\text{orcl}(m) = N) = \frac{|\mathbf{D}_m^+|}{|\mathbf{D}|}$$

$$p_m^+ := p(\text{orcl}(m) = P) = P_m^+ + \tfrac{1}{2} P_m^0$$

$$p_m^- := p(\text{orcl}(m) = N) = P_m^- + \tfrac{1}{2} P_m^0$$

where $P_m^X := \sum_{\mathcal{D} \in \mathbf{D}_m^X} p(\mathcal{D})$ for $X \in \{+, -, 0\}$.

In practice, the calculation of all diagnoses is often infeasible and diagnosis systems rely on a subset of the minimal diagnoses $\mathbf{D}$ to *estimate* these properties of MPs. In the following, we denote by $\hat{er}_{m,\mathbf{D}}^+$ and $\hat{er}_{m,\mathbf{D}}^-$ the *estimated elimination rate* for positive and negative measurement outcome for MP m computed based on $\mathbf{D}$. Similarly, we refer by $\hat{p}_{m,\mathbf{D}}^+$ and $\hat{p}_{m,\mathbf{D}}^-$ to the *estimated probability* of a positive and negative measurement outcome for m and $\mathbf{D}$. Importantly, these estimated values depend on *both* the MP m *and* the used sample $\mathbf{D}$ of diagnoses. Note that all four estimates attain values in $[0, 1]$ for any MP m, and in $(0, 1)$ if the MP m is informative. Moreover, $\hat{p}_{m,\mathbf{D}}^+ + \hat{p}_{m,\mathbf{D}}^- = 1$ and $\hat{er}_{m,\mathbf{D}}^+ + \hat{er}_{m,\mathbf{D}}^- \leq 1$ where the difference $1 - (\hat{er}_{m,\mathbf{D}}^+ + \hat{er}_{m,\mathbf{D}}^-)$ is the rate of uncommitted diagnoses, which are not affected by the measurement at m.

Example 4 *(Measurement Points and their Properties)* Assume again our DPI from Table 1 and let all minimal diagnoses be known, i.e., $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_4\}$ (cf. Example 2). Then, e.g., $m_1 := A \rightarrow C$ is an informative MP wrt. $\mathbf{D}$ since $\mathbf{D}_{m1}^+ = \{\mathcal{D}_1, \mathcal{D}_3\} \neq \emptyset$ and $\mathbf{D}_{m1}^- = \{\mathcal{D}_2, \mathcal{D}_4\} \neq \emptyset$. E.g., $\mathcal{D}_1 \in \mathbf{D}_{m1}^+$ holds because $(\mathcal{K} \setminus \mathcal{D}_1) \cup \mathcal{B} \cup P = \{ax_2, ax_4, ax_5\} \supset \{A \rightarrow B, B \rightarrow C\} \models m_1$ and thus m_1 can be no negative measurement under the assumption $\mathcal{D}_1$. In a similar way, we obtain that $\mathcal{D}_2 \in \mathbf{D}_{m1}^-$ due to $(\mathcal{K} \setminus \mathcal{D}_2) \cup \mathcal{B} \cup (P \cup \{m_1\}) = \{ax_2, ax_3, ax_5, m_1\} \supset \{A \rightarrow \neg C, A \rightarrow C\} \models \neg A$ where $\neg A$ is a negative measurement; hence, m_1 cannot be a positive measurement under the assumption $\mathcal{D}_2$. In contrast, e.g., $m_2 := B$ is a non-informative MP because $\mathbf{D}_{m2}^+ = \emptyset$.

Assuming the (normalized) probabilities from Example 3, we obtain probabilities $\hat{p}_{m1,\mathbf{D}}^+ = .545, \hat{p}_{m1,\mathbf{D}}^- = .455$ and elimination rates $\hat{er}_{m1,\mathbf{D}}^+ = .5, \hat{er}_{m1,\mathbf{D}}^- = .5$ for m_1. Note: *(1)* If we have at hand a different sample $\mathbf{D}$, the estimations for one and the same MP might vary substantially. E.g., suppose $\mathbf{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$; then $\hat{p}_{m1,\mathbf{D}}^+ = .758, \hat{p}_{m1,\mathbf{D}}^- = .242$ and $\hat{er}_{m1,\mathbf{D}}^+ \approx 0.33, \hat{er}_{m1,\mathbf{D}}^- \approx .67$. *(2)* If the sample gets smaller (wrt. subset-inclusion), then the number of informative MPs might shrink, and vice versa. E.g., m_1 becomes non-informative if, e.g., $\mathbf{D} = \{\mathcal{D}_1, \mathcal{D}_3\}$, and thus might be disregarded by diagnosis systems. Consequently, larger (smaller) samples will tend to provide a richer (sparser) selection of MP candidates. □

Evaluating Measurement Points Using Heuristics. To quantitatively assess the preferability of different MPs, state-of-the-art sequential diagnosis systems rely on heuristics that perform a one-step-lookahead analysis of MPs [14]. A *heuristic* h is a function that maps each MP m to a real-valued score $h(m)$ [30]. At this, $h(m)$ quantifies the *utility of the expected situation after knowing the outcome for MP* m. The MP with the best score according to the used heuristic is then chosen as a next query to the oracle.

Well-known heuristics incorporate exactly the two discussed features, i.e., the estimated elimination rates and estimated probabilities, into their computations [17]. So, different heuristics correspond to different functions of these estimates, e.g.:

- *information gain (ENT)* [2] uses solely the probabilities and prefers MPs where $P_m^0 = 0$ and $|\hat{p}_{m,\mathbf{D}}^+ - \hat{p}_{m,\mathbf{D}}^-|$ is minimal [30];
- *split-in-half (SPL)* [16] considers only the elimination rates and favors MPs with $\hat{er}_{m,\mathbf{D}}^+ + \hat{er}_{m,\mathbf{D}}^- = 1$ and minimal $|\hat{er}_{m,\mathbf{D}}^+ - \hat{er}_{m,\mathbf{D}}^-|$ [30];
- *risk optimization (RIO)* [31] takes into account both features by computing a dynamically re-weighted function of ENT and SPL;
- *most probable singleton (MPS)* [17; 30] also regards both features by giving preference to MPs that maximize the probability of a maximal elimination rate.

For details on these and other heuristics see [17] for a theoretical analysis and [18] for an empirical evaluation.

Example 5 *(Heuristics)* Reconsider our DPI from Table 1 and the MP m_1 from Example 4, and let all minimal diagnoses be known, i.e., $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_4\}$. Further, let $m_3 := A \wedge \neg B \rightarrow C$. Note that m_3 is informative (wrt. $\mathbf{D}$), $\mathbf{D}^+_{m_3} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$, $\mathbf{D}^-_{m_3} = \{\mathcal{D}_4\}$, and the estimations $\hat{p}^+_{m_3,\mathbf{D}} = .72, \hat{p}^-_{m_3,\mathbf{D}} = .28$ and $\hat{er}^+_{m_3,\mathbf{D}} = .25, \hat{er}^-_{m_3,\mathbf{D}} = .75$. Hence, given the two MP candidates $\{m_1, m_3\}$, the heuristic SPL would select m_1 (since a half of the known diagnoses are eliminated for each outcome). Similarly, ENT would prefer m_1 to m_3 (because for m_1 roughly a half of the probability mass is eliminated for each outcome).

However, assume that a used diagnosis sampling technique just outputs the sample $\mathbf{D} = \{\mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4\}$. In this case, we obtain the probability estimates $\hat{p}^+_{m1,\mathbf{D}} = .28, \hat{p}^-_{m1,\mathbf{D}} = .72$ and $\hat{p}^+_{m3,\mathbf{D}} = .55, \hat{p}^-_{m3,\mathbf{D}} = .45$, respectively. Hence, using ENT, the chosen MP would be m_3 (the worse MP, as shown above). If sampling would yield $\mathbf{D} = \{\mathcal{D}_2, \mathcal{D}_4\}$, then m_3 would not even be an informative MP (wrt. $\mathbf{D}$) on the one hand, and m_3 would be the (theoretically) optimal MP according to SPL on the other hand. This example shows the dramatic impact the used sampling technique can have on diagnostic decisions. □

Sequential Diagnosis (SD) aims at generating a sequence of informative MPs such that a single (highly probable) diagnosis remains for the given DPI, while minimizing the number of MPs needed (oracle inquiries are usually expensive). A generic SD process iterates through the following steps until (the Bayes-updated) $p(\mathcal{D})$ for some $\mathcal{D} \in \mathbf{D}$ exceeds a probability threshold σ:

S1 Generate a sample of minimal diagnoses $\mathbf{D}$ for the current DPI (initially, the given DPI).
S2 Choose a (heuristically optimal) informative MP m wrt. $\mathbf{D}$ (using a selection heuristic h).
S3 Ask the oracle orcl to classify m.
S4 Use the classification $\text{orcl}(m)$ to update the DPI, by adding m to the positive measurements if $\text{orcl}(m) = P$ and to the negative measurements if $\text{orcl}(m) = N$.

3 Evaluation

We conducted extensive experiments using a dataset of real-world diagnosis cases (Sec. 3.1) to study six different diagnosis sample types (Sec. 3.2 and 3.3) wrt. the accuracy of estimations and diagnostic efficiency (Sec. 3.4) in different scenarios in terms of sample size (number of diagnoses computed) and measurement selection heuristic used. The concrete research questions are explicated in Sec. 3.5 and the experiments are detailed in Sec. 3.6. Finally, in Sec. 3.7, we present and discuss the obtained results.

3.1 Dataset

In our experiments we drew upon the set of real-world diagnosis problems from the domain of knowledge-base debugging shown in Table 2. Note that every model-based diagnosis problem (according to Reiter's original characterization [1]) can be represented as a knowledge-base debugging problem [27], which is why considering knowledge-base debugging problems is without loss of generality. To obtain a representative dataset we chose it in a way it covers a variety of different problem sizes, theorem proving complexities, and diagnostic metrics (number of diagnoses, their sizes, number of conflicts, number of components). These metrics are depicted in the columns of Table 2. In order to

Table 2: Dataset used in experiments (sorted by 2nd column).

KB $\mathcal{K}$	$|\mathcal{K}|$	expressivity [1]	#D/min/max [2]	#C/min/max [2]
University (U) [3]	50	$\mathcal{SOIN}^{(D)}$	90/3/4	4/3/5
IT [4]	140	$\mathcal{SROIQ}$	1045/3/7	7/3/7
UNI [4]	142	$\mathcal{SROIQ}$	1296/5/6	6/3/10
MiniTambis (M) [3]	173	$\mathcal{ALCN}$	48/3/3	3/2/6
Transportation (T) [3]	1300	$\mathcal{ALCH}^{(D)}$	1782/6/9	9/2/6
Economy (E) [3]	1781	$\mathcal{ALCH}^{(D)}$	864/4/8	8/3/4
DBpedia (D) [5]	7228	$\mathcal{ALCHF}^{(D)}$	7/1/1	1/7/7
Cton (C) [6]	33203	$\mathcal{SHF}$	15/1/5	6/3/7

1): Description Logic expressivity [32]; the higher the expressivity, the higher is the complexity of consistency checking (conflict computation) for this logic.
2): #D/min/max denotes the number/the minimal size/the maximal size of minimal diagnoses for the DPI resulting from KB $\mathcal{K}$. Same notation for conflicts.
3): Sufficiently hard diagnosis problems from evaluations in [16], which were also used, e.g., in [18; 33; 34].   **4):** Diagnosis problems studied in [3; 35].
5): Faulty version of DBpedia ontology, see https://bit.ly/2ZO2qYZ.
6): Diagnosis problem used in scalability tests in [16]. The second scalability problem used in [16] was not included in the dataset since computation of all minimal diagnoses was infeasible (within hours of computation) for it.

implement the random sampling of diagnoses, another requirement to the dataset was that all the used problems allow the computation of *all* minimal diagnoses within tolerable time for our experiments (single digit number of minutes).

3.2 Sample Types

We examined the following types of diagnosis samples:

T1 best-first (bf)
T2 random (rd)
T3 worst-first (wf)
T4 approximate best-first (abf)
T5 approximate random (ard)
T6 approximate worst-first (awf)

By "best-first" / "worst-first", we mean the most / least probable minimal diagnoses. The types *T3* and *T6* serve as baselines. We refer to *T1*, *T2* and *T3* as *specific sample types* because we know the properties of the sample (exactly the k best or worst diagnoses, or k unbiased random ones) in advance by employing (expensive) sampling techniques that guarantee these properties. On the other hand, we call *T4*, *T5* and *T6 unspecific sample types* and adopt (usually less costly) heuristic techniques to provide them. In the following, we denote a sample of type Ti including k minimal diagnoses by $S_{Ti,k}$.

3.3 Sampling Techniques

The approaches we used for generating the samples for a given DPI $dpi = \langle \mathcal{K}, \mathcal{B}, P, N \rangle$ were:
T1: We used uniform-cost HS-Tree [9, Sec. 4.6] and stopped it after k diagnoses were computed. Due to the best-first property of the algorithm, these are provenly [9, Prop. 4.17] the k diagnoses with the highest probability among all minimal diagnoses.
T2, T3: We generated all[4] minimal diagnoses **allD** for dpi. For *T2*, we selected k random elements from this set by means of the Java (v1.8) pseudorandom number generator. For *T3*, we picked the k diagnoses with lowest probability.[5] The

---

[4]This is generally intractable [36]. So, this approach to random sampling is not viable in practice and just used for the purpose of our evaluation. As said in Sec. 3.1, we chose our dataset so that computation of **allD** was feasible within reasonable time.

[5]Note, the naive approach to generating the least probable diagnoses using a best-first diagnosis computation mechanism (such as

generation of **allD** can be done by any sound and complete diagnosis computation, e.g., HS-Tree [1].

T4, T5, T6: We used Inv-HS-Tree [22] to supply the samples. First, we added all $ax \in \mathcal{K}$ to a list L. For *T5*, we randomly shuffled L. For *T4* and *T6*, we sorted L in descending and ascending order of probability $p(ax)$, respectively. Finally, we let plain Inv-HS-Tree operate on this list L to supply a sample of size k.

Inv-HS-Tree uses k calls to a diagnosis computation method called Inverse QuickXPlain (Inv-QX) [37; 38; 39]. Each call of Inv-QX returns one well-defined minimal diagnosis $\mathcal{D}_L$ for dpi based on the strict total order of elements imposed by the sorting of the list L. Specifically, $\mathcal{D}_L$ is the minimal diagnosis with highest rank wrt. the antilexicographic order $>_{\text{antilex}}$ defined on sublists of $L = [l_1, \ldots, l_{|\mathcal{K}|}]$ [38]. At this, for sublists X, Y of L, we have $X >_{\text{antilex}} Y$ *(X has higher rank wrt. $>_{\text{antilex}}$ than Y)* iff there is some k such that $X \cap \{l_{k+1}, \ldots, l_{|\mathcal{K}|}\} = Y \cap \{l_{k+1}, \ldots, l_{|\mathcal{K}|}\}$ (both sublists are equal wrt. their lowest ranked elements in L) and $l_k \in Y \setminus X$ (the first element that differs between the sublists is in Y). E.g., if L includes the letters $a, b, \ldots, z$ in alphabetic order, then $X >_{\text{antilex}} Y$ for $X = [b, n, r, v]$ and $Y = [a, p, r, v]$ because both lists share $[r, v]$ and, after deleting these two letters from both X and Y, the now last element (p) of Y is ranked lower in L than the one (n) of X (see also [40, Sec. 3.2.5]).

That is, in the approximate best-first case *(T4)*, the computed diagnosis $\mathcal{D}_L = [d_1, \ldots, d_{n-1}, d_n]$ has the property that there is no other minimal diagnosis $\mathcal{D}' = [d_1', \ldots, d_r']$ where d_r' has higher probability than d_n, and among all minimal diagnoses that share the last element d_n, there is no other minimal diagnosis whose second-last element has a higher probability than d_{n-1}, and so forth. If we replace "higher probability" with "lower probability", we obtain a description of the diagnosis $\mathcal{D}_L$ returned in the approximate worst-first case *(T6)*. In the approximate random case *(T5)*, we reshuffle L before each call of Inv-QX, thereby trying to simulate a random selection. Note, Inv-HS-Tree guarantees that each Inv-QX call generates a *new* diagnosis by systematically "blocking" different elements in L which must not occur in the next diagnosis [22].

3.4 Evaluating Samples

We evaluate sample types based on what we call their theoretical and practical representativeness:

Theoretical Representativeness: A sample type Ti is the more representative, the better the

- probability estimates $\langle \hat{p}_{m,\mathbf{D}}^+, \hat{p}_{m,\mathbf{D}}^- \rangle$ for MPs m match the respective actual values $\langle p_m^+, p_m^- \rangle$,
- elimination rate estimates $\langle \hat{er}_{m,\mathbf{D}}^+, \hat{er}_{m,\mathbf{D}}^- \rangle$ for MPs m match the respective actual values $\langle er_m^+, er_m^- \rangle$

for samples $\mathbf{D} = S_{Ti,k}$.

Practical Representativeness: A sampling technique Ti is the more representative, the lower the

- number of measurements required,
- time required for sampling (diagnoses computation)

throughout a sequential diagnosis session until the actual diagnosis is isolated from spurious ones, where $\mathbf{D} = S_{Ti,k}$ in each sequential diagnosis iteration.

---

a uniform cost version of Reiter's HS-Tree [9]) and simply taking the reciprocals $p'(ax) := 1 - p(ax)$ instead of the probabilities $p(ax)$ for $ax \in \mathcal{K}$ (provably) does not work in general.

3.5 Research Questions

The goal of our evaluation is to shed light on the following research questions:

RQ1 Which type of sample is best in terms of theoretical representativeness?

RQ2 Which type of sample is best in terms of practical representativeness?

RQ3 Are the results wrt. *RQ1* and *RQ2* consistent over different *(a)* sample sizes, *(b)* measurement selection heuristics, and *(c)* diagnosis problem instances.

RQ4 Does larger sample size (more computed diagnoses) imply better representativeness?

RQ5 Does a better theoretical representativeness translate to a better practical representativeness?

3.6 Experiments

We conducted two experiments, EXP1 and EXP2, to investigate our research questions. Common to both experiments are the following settings: • We defined one DPI $dpi_{\mathcal{K}} := \langle \mathcal{K}, \emptyset, \emptyset, \emptyset \rangle$ for each $\mathcal{K}$ in Table 2. That is, we assumed each axiom (component) in $\mathcal{K}$ to be possibly faulty and left the background knowledge and the measurements void to begin with. To each $ax \in \mathcal{K}$, we randomly assigned a fault probability $p(ax) \in (0, 1)$ in a way that syntactically equally (more) complex axioms have an equal (higher) probability (cf. [9; 16]). E.g., in our DPI in Table 1, elements of $\{ax_1, ax_3\}$ (one implication, one negation) and $\{ax_2, ax_4\}$ (one implication), respectively, would each be allocated the same probability, and the former two would have a higher probability than the latter (cf. Example 3). • We precomputed all minimal diagnoses **allD** for each DPI $dpi_{\mathcal{K}}$. • We used all sample types Ti for $i \in \{1, \ldots, 6\}$ (cf. Sec. 3.2). • We used sample sizes (numbers of generated minimal diagnoses) $k \in \{2, 6, 10, 20, 50\}$.

The specific settings for each experiment were:

EXP1: (theoretical representativeness) For each $dpi_{\mathcal{K}}$, for each k, and for each Ti, we computed a sample $\mathbf{D} = S_{Ti,k}$. We used

- $\mathbf{D}$ to compute probability and elimination rate estimates $\langle \hat{p}_{m,\mathbf{D}}^+, \hat{p}_{m,\mathbf{D}}^- \rangle$ and $\langle \hat{er}_{m,\mathbf{D}}^+, \hat{er}_{m,\mathbf{D}}^- \rangle$, and
- **allD** to compute $\langle p_m^+, p_m^- \rangle$ and $\langle er_m^+, er_m^- \rangle$

for 50 (if so many, otherwise for all) randomly selected informative MPs wrt. $\mathbf{D}$. For each such MP, we thus had four estimates and four corresponding actual values, that we could compare against one another.

EXP2: (practical representativeness) For each $dpi_{\mathcal{K}}$, for each k, for each Ti, and for each of the four heuristics $h \in \{\text{ENT}, \text{SPL}, \text{RIO}, \text{MPS}\}$ (cf. Sec. 2), we executed 10 sequential diagnosis sessions (loop *S1–S4*, Sec. 2) while in each session

- searching for a different randomly selected target diagnosis $\mathcal{D}^* \in \textbf{allD}$ for $dpi_{\mathcal{K}}$,
- starting from the initial problem $dpi_{\mathcal{K}}$
- with stop criterion $\sigma = 1$ (loop until a single minimal diagnosis remains, i.e., all others have been ruled out),

where in each iteration through the loop at step

- *S1*, a sample $\mathbf{D} = S_{Ti,k}$ is drawn for the current DPI,
- *S2*, an informative MP that is optimal for h is selected,
- *S3*, an automated oracle classifies each MP in a way the predefined target diagnosis $\mathcal{D}^*$ is not ruled out.

For our analyses, we recorded (sampling) times and number of measurements (i.e., loop iterations) throughout a session.

Table 3: Theoretical and practical representativeness: Rankings of sample types for various scenarios (EXP1 and EXP2).

scenario	(best)		ranking			(worst)	criterion
all	rd	wf	bf	awf	(abf	ard)	E
$k = 6$	bf	wf	(rd	awf)	abf	ard	E
$k = 10$	rd	wf	bf	awf	abf	ard	E
$k = 20$	rd	wf	bf	awf	(abf	ard)	E
$k = 50$	rd	wf	awf	bf	ard	abf	E
all	bf	rd	awf	abf	ard	wf	P
$k = 6$	bf	(abf	rd	ard)	awf	wf	P
$k = 10$	bf	rd	(abf	awf)	(ard	wf)	P
$k = 20$	bf	rd	awf	(abf	ard	wf)	P
$k = 50$	bf	rd	awf	ard	wf	abf	P
all	bf	ard	abf	rd	awf	wf	M
$k = 2$	bf	abf	ard	awf	rd	wf	M
$k = 6$	bf	rd	ard	abf	awf	wf	M
$k = 10$	rd	abf	ard	bf	awf	wf	M
$k = 20$	ard	abf	awf	rd	bf	wf	M
$k = 50$	ard	rd	awf	bf	abf	wf	M
$h = \text{ENT}$	bf	abf	ard	awf	rd	wf	M
$h = \text{SPL}$	bf	ard	abf	rd	awf	wf	M
$h = \text{RIO}$	rd	ard	awf	bf	abf	wf	M
$h = \text{MPS}$	ard	abf	rd	awf	wf	bf	M
all	awf	bf	(abf	ard)	rd	wf	T
$k = 2$	abf	(ard	awf)	bf	rd	wf	T
$k = 6$	awf	(abf	ard)	bf	(rd	wf)	T
$k = 10$	awf	abf	(ard	bf)	rd	wf	T
$k = 20$	bf	ard	awf	abf	rd	wf	T
$k = 50$	bf	wf	(awf	rd)	ard	abf	T
$h = \text{ENT}$	awf	abf	bf	ard	rd	wf	T
$h = \text{SPL}$	(abf	awf	bf)	ard	rd	wf	T
$h = \text{RIO}$	awf	(abf	ard	bf)	rd	wf	T
$h = \text{MPS}$	bf	awf	ard	abf	rd	wf	T

3.7 Results

From our experiments, we obtained two large datasets, with $6*8*5 = 240$ (EXP1) and $6*8*5*4 = 960$ (EXP2) factor combinations, respectively, for the factors sample type (6 levels), diagnosis problem (8), sample size (5), and heuristic (4). Due to the high information content of our data and the paper length restrictions, we can only provide a very condensed presentation of the results, given in Tables 3 and 4.
Presentation. *Table 3* shows rankings of the sample types over different subsets of all factor combinations (referred to as *scenarios*; left column of the table). E.g., scenario "all" means all 240 (EXP1) / 960 (EXP2) cases aggregated, whereas "$k = 20$" denotes exactly the $240 : 5 = 48$ (EXP1) / $960 : 5 = 192$ (EXP2) cases where the sample size was set to 20. Results from EXP1 are depicted in the top part of the table (first ten rows); results from EXP2 in the bottom part. A sample type Ti being ranked prior to type Tj (middle column of table) means that Ti was better than Tj in more of the factor combinations of the respective scenario than vice versa. At this, the meaning of "better" (*criterion* for comparison; rightmost table column) is

- a higher Pearson correlation coefficient between estimated and real values for elimination rate (E) and, respectively, probability (P) estimations (cf. *theoretical representativeness* in Sec. 3.4), and
- a lower average number of measurements (M) and, respectively, a lower average sample computation time (T) in sequential sessions (cf. *practical representativeness* in Sec. 3.4).

Note, the table does *not* give information about *how much better (worse)* one Ti was than another, but only that it was a preferred choice to the other *in more (less) cases* (of a scenario). Moreover, the rankings do *not* mean that a higher ranked strategy was *always* better than a lower ranked one.

The idea behind this representation is to give the user of a diagnosis system a guidance how to set parameters (diagnosis computation algorithm, number of computed diagnoses, heuristic used for measurement selection) in order to have the highest chance of achieving best estimations (EXP1) / efficiency (EXP2).

Table 4 lists the best (ranked) sample types wrt. overall time per diagnosis session in EXP2 (cumulated system computation time plus cumulated time for all measurements) for different scenarios and assumptions (1min, 10min) of measurement conduction times. The two rightmost columns ("adj") show hypothetical results under the assumption that sample types *T2* (rd) and *T3* (wf)—which we naively simulated by means of brute force diagnosis computation in our experiments (cf. Sec. 3.3)—were as efficiently computable as sample type *T1* (bf). This allows to assess the added value of, e.g., *efficient* random diagnosis sampling techniques.
Discussion. We address each research question in turn:

RQ1:[6] *(Elimination rate, criterion E, Table 3)* We see that rd is the sample type of choice, as one would expect. In numbers, the median correlation coefficients over all cases per scenario for (best,worst) sample type for $k \in \{6, 10, 20, 50\}$ were $\{(0.76, 0.5), (0.83, 0.52), (0.95, 0.7), (0.98, 0.85)\}$, which reveals that estimations were altogether pretty good for all sampling techniques. However, for $k \geq 20$, coefficients for rd manifested a significantly lower variance than in case of all other techniques, i.e., all coefficients for rd concentrated in the interval [0.9,1], whereas lowest coefficients for all other techniques lay between less than 0.5 and 0.7. Moreover, it stands out that wf allowed almost as accurate estimations as rd. A possible explanation for these favorable results of wf is that there is usually a large number of minimal diagnoses with a very small probability, which is why the "sub-population" from which the wf diagnoses are "selected" tends to be larger (and thus more representative) than for other sample types, except for rd (where diagnoses are drawn at random from the *full* population). Finally, it is interesting that approximate methods (awf, ard, abf) produced less representative samples than exact ones. And, although rd comes out on top for E, its approximate counterpart ard shows the worst results.

(Probability, criterion P, Table 3) Here, bf proved to be the predominantly superior technique in all depicted scenarios. Although the fact that rd was only the second best method might be surprising at first sight, the likely explanation for this is that often few of the most probable diagnoses already account for a major part of the overall probability mass, which is why they are more reliable for estimations of P than a random sample. For the same reason, it comes as no revelation that wf samples turned out to be the least preferable means to estimate P. The medians of the correlation coefficients over all cases per scenario for (best,worst) sample type for $k \in \{6, 10, 20, 50\}$ were $\{(0.93, 0.6), (0.87, 0.64), (0.98, 0.74), (0.99, 0.86)\}$. Thus, again, all sampling methods enabled pretty decent estimations, even for small sample sizes of only six diagnoses. Similarly as for E, the va-

---

[6]Remarks wrt. *RQ1*: *(1)* We had to leave out the $k = 2$ scenarios as there were too few informative MPs which made these scenarios not reliably analyzable. *(2)* Values and rankings for other types of correlation coefficients (i.e., Spearman and Kendall) were very similar to the presented (Pearson) results. *(3)* Most correlation coefficients were statistically significant ($\alpha = 0.05$), except for a few $k = 6$ scenarios and some scattered $k = 10$ cases.

riance of correlation coefficients was significantly lower for the best sample type, bf, than for all others. However, the spread of correlation coefficients for the single sample types was noticeably larger for P than for E for $k \leq 20$, suggesting that only bf facilitates very reliable estimations of P for small to medium sample sizes.

RQ2: *(Number of measurements, criterion M, Table 3)* We find that bf was the best strategy if all data is considered; and it was the most suitable choice for heuristics ENT and SPL and for small sample sizes $\{2, 6\}$. On the other hand, it was the worst choice for the MPS heuristic where it led to substantial overheads (of up to $>100\,\%$) compared to other sample types, especially for large sample sizes. E.g., for the diagnosis problem U and $k = 50$, a diagnosis session using bf involved 58 measurements vs. 25 measurements if rd was used instead. What is somewhat surprising is that bf decidedly outperformed rd in the SPL scenarios, although the SPL function does not use any probabilities (where bf leads to better estimations), but solely the elimination rate (where rd produces better estimates). Further analyses are needed to better understand this phenomenon. Overall, rd compares favorably only against awf and wf, but its performance depends largely on the used heuristic. For RIO it is even the sample type of choice, and for MPS it clearly overcomes bf. For all four heuristics, one of the approximate methods was the second best method, among which ard led to good performance most consistently. In comparison with rd, ard was only (slightly) outweighed for RIO, but prevails for the other three heuristics. When considering large samples (20 or 50 diagnoses), ard even turned out to be the overall winner. This indicates that the QuickXPlain-based approximate random algorithm, in spite of its rather poor estimations (cf. E and P in Table 3), tends to be no less effective than a real random strategy. Finally, observe that wf was in fact the least favorable option in quasi all scenarios.

(Time for diagnosis session, criterion T, Table 3) Due to the naive brute force approach we used in our experiments to generate samples of type rd and wf, it comes as no surprise that these two methods perform most poorly in terms of T. When drawing our attention to the best strategies, we find that, in all but one ($h = $ SPL) of the shown scenarios, it is a different sample type that exhibited lowest time (T) than the one that manifested the lowest number of measurements (M). This prompts the conjecture of some time-information trade-off in diagnosis sampling—or: whenever the sampling process is most efficient (on avg.), the measurements arising from the sample are not most effective (on avg.). In particular, we recognize that, if an exact method is best for T (M), then an approximate method is best for M (T). And, unlike for M, ard tends to be worse than awf and abf in case of T.

(Diagnosis session time, criteria T, M combined, Table 4) Since the outcome for *RQ2* is not at all clear-cut when viewing M and T separately, we investigate their combined effect, i.e., the overall (avg.) length of diagnosis sessions for the different sample types. In brief, the conclusions are: • For small sample size below 10, go with bf. • For sample size 10, use abf. • For sample size 20, take awf if the expected time for conducting measurements is low, and take bf else. • Unless there is an efficient method for rd, use bf for large sample size (50), otherwise use rd. • For ENT or SPL, adopt bf. • For RIO, if measurement time is short, use bf if there is no efficient algorithm for rd, otherwise use rd; if measuring takes longer, use ard. • For MPS, use ard.

Table 4: Best sample types wrt. overall sequential diagnosis time for various scenarios (EXP2).

	best sample type (given that time for each measurement $= t$)			
scenario	$t = 1$min	$t = 10$min	$t = 1$min (adj)	$t = 10$min (adj)
all data	bf	bf	bf	bf
$k = 2$	bf	bf	bf	bf
$k = 6$	bf	bf	bf	bf
$k = 10$	abf	abf	abf	(rd, abf)
$k = 20$	awf	bf	awf	bf
$k = 50$	bf	bf	rd	rd
$h = $ ENT	bf	bf	bf	bf
$h = $ SPL	bf	bf	bf	bf
$h = $ RIO	bf	ard	rd	(ard, bf)
$h = $ MPS	ard	ard	ard	ard

RQ3: For theoretical representativeness, we observe pretty consistent (ranking) results over all sample sizes (cf. often equal entries in each column for each of the E and P criteria in Table 3). There is more variation when comparing results for different diagnosis problems. Nevertheless, results are fairly stable concerning the winning strategy: rd is in all cases the best (75 %) or second best (25 %) sample type for E, and bf is in all but one case the best (63 %) or second best (25 %) sample type for P. For practical representativeness, we see more of a fluctuation over different sample sizes and heuristics, as discussed for *RQ2* above (cf. variation over entries of each column for M and T in Table 3). Examining (ranking) results over different diagnosis problems reveals a similar picture, where however the rankings for T are decidedly more stable than those for M, meaning that relative sampling times are less affected by the particular problem instance than the informativeness of the samples.

RQ4: Our data indicates a clear trend that increasing sample size leads to better theoretical representativeness (cf. discussion of *RQ1*). However, it also suggests that there is no general significant positive effect of larger sample size on practical representativeness. While this is obvious for sampling time (T), i.e., generating more diagnoses cannot take less time, it is less so for the number of measurements (M). In fact, we even measured increases wrt. M in some cases (e.g., for MPS) as a result to drawing larger samples. This corroborates similar findings in this regard, albeit for lower sample sizes and other types of diagnosis problems, reported by [18; 25].

RQ5: From our data, we cannot generally conclude that a better theoretical implies a better practical representativeness (see discussion on *RQ1*, *RQ2* and *RQ3*). E.g., observe the performance of ard in the top vs. bottom part of Table 3. We surmise the cause of that to lie in the fact that *(1)* heuristics are based on a lookahead of one step only (where the approximate character of this analysis might counteract the benefit of good estimations), and *(2)* the added (information) value of additional diagnoses taken into a sample, regardless of how selected, decreases with the sample size (cf. the *law of diminishing marginal utility* [41] in economics).

4 Research Limitations

Our evaluations do not come without limitations. In brief, there are the following threats to validity:[7] • For feasibility reasons, we *(1)* did not use *all* diagnoses to determine the real values (EXP1), but just all minimal ones, and *(2)* used

---

[7] Those bullet points marked by a "∗" we plan to address in terms of additional experiments as part of future work.

only such problem instances that reasonably allow the generation of all minimal diagnoses. • We focused on binary-outcome measurements which are common in some, but not all diagnosis sub-domains, e.g., in ontology and KB debugging [9; 42], circuit diagnosis [2], or matrix-based methods [43].∗ • We did not evaluate *bf*/*wf* sample types including minimum/maximum-cardinality minimal diagnoses, but concentrated on most/least probable ones.∗ • To keep the size of our dataset manageable, we *(1)* omitted less commonly used existing heuristics, and *(2)* included only a subset of available diagnosis computation methods into our analyses.∗ • In EXP2, differences between sample types underlying the rankings were in many cases not statistically significant ($\alpha = 0.05$) due to the relatively low number of 10 diagnosis sessions we ran for each factor combination (note, EXP2 already takes several weeks of computation time with these settings). So, conclusions from our data must be treated with caution for the time being.∗

5 Conclusions

We tested six diagnosis sampling techniques wrt. the quality of their estimations used by measurement selection heuristics, and their achieved performance in terms of diagnostic efficiency. Whereas random sampling, in line with statistical theory, leads to highly reliable estimations, this benefit is only conditionally reflected by the performance exhibited by random samples in diagnosis sessions, e.g., if the sample size is large or one specific heuristic is adopted. It turned out that, inspite of their missing statistical foundation, fully biased best-first samples including the most probable diagnoses were the best method in most investigated scenarios.

References

[1] R Reiter. A Theory of Diagnosis from First Principles. *AIJ*, 32(1):57–95, 1987.

[2] J de Kleer, B Williams. Diagnosing multiple faults. *AIJ*, 32(1):97–130, 1987.

[3] P Rodler, D Jannach, K Schekotihin, P Fleiss. Are query-based ontology debuggers really helping knowledge engineers? *KBS*, 179:92–107, 2019.

[4] R Stern, M Kalech, A Feldman, S Rogov, T Zamir. Finding all diagnoses is redundant. In *DX'13*.

[5] D Jannach, T Schmitz, K Schekotihin. Parallelized hitting set computation for model-based diagnosis. In *AAAI'15*.

[6] J Slaney. Set-theoretic duality: A fundamental feature of combinatorial optimisation. In *ECAI'14*.

[7] A Kalyanpur. *Debugging and Repair of OWL Ontologies*. PhD thesis, Univ. of Maryland, 2006.

[8] K Schekotihin, P Rodler, W Schmid. Ontodebug: Interactive ontology debugging plug-in for Protégé. In *FoIKS'18*.

[9] P Rodler. *Interactive Debugging of Knowledge Bases*. PhD thesis, Univ. Klagenfurt. *CoRR*, abs/1605.05950, 2015.

[10] C Meilicke. *Alignment incoherence in ontology matching*. PhD thesis, Univ. Mannheim, 2011.

[11] K Schekotihin, P Rodler, W Schmid, M Horridge, T Tudorache. Test-driven ontology development in Protégé. In *ICBO'18*.

[12] J de Kleer, O Raiman. How to diagnose well with very little information. In *DX'93*.

[13] K Pattipati, M Alexandridis. Application of heuristic search and information theory to sequential fault diagnosis. *IEEE T. Syst. Man Cyb.*, 20(4):872–887, 1990.

[14] J de Kleer, O Raiman, and M Shirley. One step lookahead is pretty good. In *Readings in Mod.-Based Diag.*, 1992.

[15] B Moret. Decision trees and diagrams. *CSUR*, 14(4):593–623, 1982.

[16] K Shchekotykhin, G Friedrich, P Fleiss, P Rodler. Interactive Ontology Debugging: Two Query Strategies for Efficient Fault Localization. *JWS*, 12-13:88–103, 2012.

[17] P Rodler. On active learning strategies for sequential diagnosis. In *DX'17*.

[18] P Rodler, W Schmid. On the impact and proper use of heuristics in test-driven ontology debugging. In *RuleML'18*.

[19] A Gonzalez-Sanchez, R Abreu, H Gross, and A van Gemund. Spectrum-based sequential diagnosis. In *AAAI'11*.

[20] R Abreu, A van Gemund. A low-cost approximate minimal hitting set algorithm and its application to model-based diagnosis. In *Symp. on Abstract. Reform. Approx.*, 2009.

[21] A Feldman, G Provan, A van Gemund. Computing minimal diagnoses by greedy stochastic search. In *AAAI'08*.

[22] K Shchekotykhin, G Friedrich, P Rodler, P Fleiss. Sequential diagnosis of high cardinality faults in knowledge-bases by direct diagnosis generation. In *ECAI'14*.

[23] J de Kleer. Focusing on probable diagnoses. In *AAAI'91*.

[24] J de Kleer, B Williams. Diagnosis with behavioral modes. In *IJCAI'89*.

[25] J de Kleer, O Raiman. Trading off the costs of inference vs. probing in diagnosis. In *IJCAI'95*.

[26] T Zamir, R Stern, M Kalech. Using model-based diagnosis to improve software testing. In *AAAI'14*.

[27] P Rodler, K Schekotihin. Reducing model-based diagnosis to knowledge base debugging. In *DX'17*.

[28] A Felfernig, G Friedrich, D Jannach, M Stumptner. Consistency-based diagnosis of configuration knowledge bases. *AIJ*, 152(2):213–234, 2004.

[29] J de Kleer, A Mackworth, R Reiter. Characterizing diagnoses and systems. *AIJ*, 56, 1992.

[30] P Rodler. Towards better response times and higher-quality queries in interactive knowledge base debugging. Technical report, Univ. Klagenfurt. *CoRR*, abs/1609.02584v2, 2016.

[31] P Rodler, K Shchekotykhin, P Fleiss, G Friedrich. RIO: Minimizing User Interaction in Ontology Debugging. In *RR'13*.

[32] F Baader, D Calvanese, D McGuinness, D Nardi, P Patel-Schneider. *The Description Logic Handbook*. CUP, 2007.

[33] M Horridge, B Parsia, U Sattler. Lemmas for justifications in owl. *DL'09*.

[34] Q Ji, Z Gao, Z Huang, M Zhu. Measuring effectiveness of ontology debugging systems. *KBS*, 71:169–186, 2014.

[35] P Rodler. Reuse, Reduce and Recycle: Optimizing Reiter's HS-tree for sequential diagnosis. In *ECAI'20*.

[36] T Bylander, D Allemang, M Tanner, J Josephson. The computational complexity of abduction. *AIJ*, 49:25–60, 1991.

[37] A Felfernig, M Schubert, C Zehentner. An efficient diagnosis algorithm for inconsistent constraint sets. *AI for Engin. Design Analysis Manufact.*, 26(1):53–62, 2011.

[38] U Junker. QuickXPlain: Preferred Explanations and Relaxations for Over-Constrained Problems. In *AAAI'04*.

[39] P Rodler. Understanding the QuickXPlain algorithm: Simple explanation and formal proof. *CoRR*, abs/2001.01835, 2020.

[40] P Rodler, W Schmid, K Schekotihin. A generally applicable, highly scalable measurement computation and optimization approach to sequential model-based diagnosis. *CoRR*, abs/1711.05508, 2017.

[41] R Easterlin. Diminishing marginal utility of income? Caveat emptor. *Soc. Indic. Res.*, 70(3):243–255, 2005.

[42] P Rodler, M Eichholzer. On the Usefulness of Different Expert Question Types for Fault Localization in Ontologies. In *IEA/AIE'19*.

[43] M Shakeri, V Raghavan, K Pattipati, A Patterson-Hine. Sequential testing algorithms for multiple fault diagnosis. *IEEE T. Syst. Man Cyb.*, 30(1):1–14, 2000.