Random vs. Best-First: Impact of Sampling Strategies on Decision Making in Model-Based Diagnosis[*]

Patrick Rodler

University of Klagenfurt, Austria
patrick.rodler@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 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 wrt. 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 an election and 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 wrt. the election. The problem is 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 (Reiter 1987) deals with the detection, localization and repair of faults in observed systems such as programs, circuits, knowledge bases or physical devices. 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 *(i) directly analyzed*, e.g., to manually discover or make estimations about the actual fault (Stern et al. 2013; Rodler et al. 2019), to aid proper algorithm choice (Slaney 2014), or to support users in test case specifications or repair actions (Kalyanpur 2006; Meilicke 2011; Schekotihin, Rodler, and Schmid 2018), or *(ii) used as an input or guidance to diagnostic algorithms*. We focus on *(ii)* in this work.

An important class of diagnostic algorithms that are guided by a set of precomputed diagnoses are *sequential diagnosis* approaches (de Kleer and Williams 1987; de Kleer and Raiman 1993). They use a sample of diagnoses to compute informative system measurements that allow to efficiently and systematically rule out spurious diagnoses until a single or highly probable one remains. Since achieving (global) optimality of the sequence of measurements is intractable in general (Pattipati and Alexandridis 1990), state-of-the-art sequential diagnosis methods usually rely on local optimization (de Kleer, Raiman, and Shirley 1992) using one out of numerous *heuristics* (Moret 1982; de Kleer and Williams 1987; Shchekotykhin et al. 2012; Rodler 2018) 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, i.e., the

(F1) diagnoses' probabilities, and

(F2) diagnoses' predictions of the measurement outcome,

which allow to estimate the probabilities and diagnosis elimination rates of the measurement outcomes.

Literature offers a range of techniques to generate samples of diagnoses, among them ones that return a *specific sample* (de Kleer and Williams 1987; Reiter 1987; Rodler 2020a) (which includes exactly a predefined subset of all diagnoses), and others that compute an *unspecific sample*, e.g., in a heuristic (Abreu and Van Gemund 2009), stochastic (Feldman, Provan, and Van Gemund 2008) or simply undefined way (Shchekotykhin et al. 2014) (where no guarantee can be given wrt. diagnosis selection for the sample).

Many existing sequential diagnosis approaches draw on

[*] An earlier version (Rodler and Elichanova 2020) of this work was accepted and presented at the Int'l Workshop on Principles of Diagnosis 2020 (DX'20). Please find an animated presentation, a poster, and a video presentation of the present work at: tinyurl.com/3shsn7p8, tinyurl.com/mjt9j4aw, and tinyurl.com/yahw5dcs .

samples of the specific type in that they build upon *best-first* samples, such as maximum-probability or minimum-cardinality diagnoses (de Kleer and Williams 1989; de Kleer 1991; de Kleer and Raiman 1995; Gonzalez-Sanchez et al. 2011; Shchekotykhin et al. 2012; Zamir, Stern, and Kalech 2014; Rodler and Herold 2018; Rodler 2022). While perhaps often being motivated by the desideratum to know the most preferred or likely candidate(s) 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 allows (on average) better conclusions and estimations about the whole population than any other sample.*

The particular contributions are: We

- analyze real-world diagnosis cases and gain insight into the quality of *three specific* (best-first, random and, as a baseline, worst-first) *and three unspecific* (approximate best-first / random / worst-first) *sample types*.
- assess a sample type's quality based on *(i)* its *theoretical representativeness*, i.e., how well it allows to estimate aspects *(F1)* and *(F2)* that determine the heuristic score of measurements, and *(ii)* its *practical representativeness*, i.e., its performance achieved in a diagnosis session wrt. time and number of measurements.
- investigate the *impact of the (i) sample size*, *(ii)* particular used *heuristic*, and *(iii)* tackled *diagnosis problem* on the sample's representativeness.[1]

The remainder of this work is organized as follows: Sec. 2 provides theoretical foundations and a problem motivation. The conducted evaluations (dataset, sample types, sampling techniques, evaluation criteria, research questions, experiment settings, and results) are discussed in Sec. 3. Research limitations are addressed in Sec. 4, before we conclude with Sec. 5. Finally, Appendix A explains one of the used sampling techniques in more detail.

2 Basics and Problem Motivation

We outline the basics of model-based diagnosis, based on the framework of (Rodler 2015) which is more general (Rodler and Schekotihin 2018) than Reiter's theory (Reiter 1987).[2]

Diagnosis Problem Assume a diagnosed system, consisting of a set of components $\{c_1, \dots, c_k\}$ and described by a finite set of logical sentences $K \cup B$, where K (possibly faulty sentences) includes knowledge about the behavior of the system components, and B (correct background knowledge) comprises any additional available system knowledge and system observations. More precisely, there is a one-to-one relationship between sentences $s_i \in K$ and components c_i, where s_i describes (only) the nominal behavior of c_i (*weak fault model* (Feldman, Provan, and Van Gemund 2009)). E.g., if c_i is an AND-gate in a circuit, then $s_i := out(c_i) = and(in1(c_i), in2(c_i))$; 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 s_i in 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 (de Kleer and Williams 1987; Reiter 1987; Felfernig et al. 2004). 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 K is a biological knowledge base to be debugged, a negative test case might be "every bird can fly" (think of penguins). We call $\langle K, B, P, N\rangle$ a *diagnosis problem instance (DPI)*.

Example 1 *(DPI)* Assume a DPI stated in propositional logic with $K := \{s_1 : A \to \neg B, s_2 : A \to B, s_3 : A \to \neg C, s_4 : B \to C, s_5 : A \to B \vee C\}$. The "system" (here the knowledge base K itself) comprises five "components" $c_1, \dots, c_5$, and the "normal behavior" of c_i is given by the respective sentence $s_i \in K$. No background knowledge ($B = \emptyset$) or positive measurements ($P = \emptyset$) are given from the start. But, there is one negative measurement ($N = \{\neg A\}$), which stipulates that $\neg A$ must *not* be an entailment of the correct system. Note, however, that K (i.e., the assumption that all "components" are normal) in this case does entail $\neg A$ (e.g., due to the sentences s_1, s_2) and thus some sentence ("component") in K must be faulty. □

Diagnoses If the system description along with the positive measurements (under the assumption K that all components are healthy) is inconsistent, i.e., $K \cup B \cup P \models \bot$, or some negative measurement is entailed, i.e., $K \cup B \cup P \models n$ for some $n \in N$, some healthiness assumption(s) of components, i.e., some sentences in K, must be retracted. We call such a set of sentences $D \subseteq K$ a *diagnosis* for the DPI $\langle K, B, P, N\rangle$ iff $(K \setminus D) \cup B \cup P \not\models x$ for all $x \in N \cup \{\bot\}$. We say that D is a *minimal diagnosis* for *dpi* iff there is no diagnosis $D' \subset D$ for *dpi*. The set of minimal diagnoses is representative of all diagnoses under the weak fault model (de Kleer, Mackworth, and Reiter 1992), i.e., the set of all diagnoses is equal to the set of all supersets of minimal diagnoses. Thus, diagnosis approaches usually restrict their focus to only minimal diagnoses. We call a diagnosis D^* *the actual diagnosis* iff all elements of D^* are in fact faulty and all elements of $K \setminus D^*$ are in fact correct.

Example 2 *(Diagnoses)* For our DPI from Ex. 1 we have four minimal diagnoses, given by $D_1 := [s_1, s_3]$, $D_2 := [s_1, s_4]$, $D_3 := [s_2, s_3]$, $D_4 := [s_2, s_5]$. E.g., D_1 is a minimal diagnosis as $(K \setminus D_1) \cup B \cup P = \{s_2, s_4, s_5\}$ is consistent

[1] While our evaluations particularly aim at better understanding the effect of different sample types in the common model-based diagnosis setting where a precomputed sample of diagnoses serves as guidance for diagnostic decisions, the investigation of other techniques which forgo the sampling of diagnoses, e.g., by using probabilistic logical or graphical models, such as (Pearl 1988; Srinivas 1994; Mengshoel et al. 2010; Siddiqi and Huang 2011; Domingos et al. 2016), is beyond the scope of this work.

[2] This framework allows to represent things that must *not* be true for a diagnosed system, which is helpful, e.g., for diagnosing knowledge bases or ontologies (Shchekotykhin et al. 2012).

and does not entail the negative measurement $\neg A$. □

Diagnosis Probabilities If 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) $s \in K$ to its failure probability $p(s) \in (0,1)$, the probability $p(D)$ of a diagnosis $D \subseteq K$ (under the common assumption of independent component failure) is computed as $p(D) := \prod_{s \in D} p(s) \prod_{s \in K \setminus D}(1-p(s))$. Each time a new measurement is added to the DPI, probabilities of diagnoses are updated using Bayes' Theorem (de Kleer and Williams 1987).[3]

Example 3 *(Diagnosis Probabilities)* Recall the DPI from Ex. 1 and let $\langle p(s_1), \dots, p(s_5)\rangle = \langle .1, .05, .1, .05, .15\rangle$. Then, the probabilities of all diagnoses from Ex. 2 are $\langle p(D_1), \dots, p(D_4)\rangle = \langle .0077, .0036, .0036, .0058\rangle$. E.g., $p(D_1)$ 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 knowledge base, $\forall X(bird(X) \to canFly(X))$ is an MP. Assuming an oracle orcl (e.g., an 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., $m \mapsto \mathsf{orcl}(m)$ where $\mathsf{orcl}(m) \in \{P, N\}$.

Measurements to Discriminate among Diagnoses MPs are useful for identifying the actual diagnosis among multiple diagnoses for a DPI. Hence, given a set of diagnoses $\mathbf{D}$ for a DPI to discriminate between, the MPs m of particular interest are those for which each classification $\mathsf{orcl}(m)$ is inconsistent with some diagnosis in $\mathbf{D}$ (de Kleer and Williams 1987; Rodler 2015). 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 $\mathbf{D}^0_m$:

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

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

(Estimated) Properties of Measurement Points Since not all informative MPs are equally utile, the consideration of additional properties of MPs enables a more fine-grained preference rating of MPs. In fact, if $\mathbf{D}$ includes all diagnoses for the given DPI, the partition $\langle \mathbf{D}^+_m, \mathbf{D}^-_m, \mathbf{D}^0_m\rangle$ allows to determine, for each measurement outcome $c \in \{P, N\}$, its *diagnosis elimination rate* $er(\mathsf{orcl}(m) = c)$ as well as its *probability* $p(\mathsf{orcl}(m) = c)$ (Rodler 2018):

$$\begin{aligned}
er^+_m &:= er(\mathsf{orcl}(m) = P) = \tfrac{|\mathbf{D}^-_m|}{|\mathbf{D}|} \\
er^-_m &:= er(\mathsf{orcl}(m) = N) = \tfrac{|\mathbf{D}^+_m|}{|\mathbf{D}|} \\
p^+_m &:= p(\mathsf{orcl}(m) = P) = P^+_m + \tfrac{1}{2}P^0_m \\
p^-_m &:= p(\mathsf{orcl}(m) = N) = P^-_m + \tfrac{1}{2}P^0_m
\end{aligned}$$

where $P^X_m := \sum_{D \in \mathbf{D}^X_m} p(D)$ 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 & Properties)* Assume our DPI from Ex. 1 and let all minimal diagnoses be known, i.e., $\mathbf{D} = \{D_1, \dots, D_4\}$ (cf. Ex. 2). Then, e.g., $m_1 := A \to C$ is an informative MP wrt. $\mathbf{D}$ since $\mathbf{D}^+_{m1} = \{D_1, D_3\} \neq \emptyset$ and $\mathbf{D}^-_{m1} = \{D_2, D_4\} \neq \emptyset$. E.g., $D_1 \in \mathbf{D}^+_{m1}$ because $(K \setminus D_1) \cup B \cup P = \{s_2, s_4, s_5\} \supset \{A \to B, B \to C\} \models m_1$ and thus m_1 can be no negative measurement under the assumption D_1. Similarly, $D_2 \in \mathbf{D}^-_{m1}$ due to $(K \setminus D_2) \cup B \cup (P \cup \{m_1\}) = \{s_2, s_3, s_5, m_1\} \supset \{A \to \neg C, A \to C\} \models \neg A$ where $\neg A$ is a negative measurement; hence, m_1 can be no positive measurement under the assumption D_2. In contrast, e.g., $m_2 := B$ is a non-informative MP because $\mathbf{D}^+_{m2} = \emptyset$.

Assuming the (normalized) probabilities from Ex. 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 $\mathbf{D} = \{D_1, D_2, D_3\}$; 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)* Smaller (larger) samples will tend to provide a sparser (richer) selection of MP candidates. E.g., m_1 becomes non-informative if $\mathbf{D} = \{D_1, D_3\}$, and thus might be disregarded by diagnosis systems. □

[3] We adopt this way of modeling diagnostic uncertainty as it is widely used in model-based diagnosis, and because conformance with existing frameworks is important for the meaningfulness, significance and applicability of our results to the literature. Nevertheless, note that more flexible approaches to uncertainty in model-based diagnosis have been proposed, cf., e.g., (Lucas 2001).

Evaluating Measurement Points by Heuristics To quantitatively assess the preferability of different MPs, state-of-the-art diagnosis systems rely on heuristics that perform a one-step-lookahead analysis of MPs (de Kleer, Raiman, and Shirley 1992). A *heuristic* h is a function that maps each MP m to a real-valued score $h(m)$ (Rodler 2016), where $h(m)$ quantifies the *utility of the expected situation after knowing the outcome for MP* m. The MP with the best score as per 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 (Rodler 2018). So, different heuristics correspond to different functions of these estimates, e.g.: *(1) information gain (ENT)* (de Kleer and Williams 1987) uses solely the probabilities and prefers MPs where $P^0_m = 0$ and $|\hat{p}^+_{m,\mathbf{D}} - \hat{p}^-_{m,\mathbf{D}}|$ is minimal; *(2) split-in-half (SPL)* (Shchekotykhin et al. 2012) considers only the elimination rates and favors MPs with $\hat{er}^+_{m,\mathbf{D}} + \hat{er}^-_{m,\mathbf{D}} = 1$ and minimal $|\hat{er}^+_{m,\mathbf{D}} - \hat{er}^-_{m,\mathbf{D}}|$ (Rodler 2016); *(3) risk optimization (RIO)* (Rodler et al. 2013) takes into account both features by computing a dynamically re-weighted function of ENT and SPL; *(4) most probable singleton (MPS)* (Rodler 2016, 2018) also regards both features by giving preference to MPs that maximize the probability of a maximal elimination rate. For details on heuristics see (Rodler 2016, 2018) for theoretical analyses and (Shchekotykhin et al. 2012; Rodler et al. 2013; Rodler and Schmid 2018; Rodler and Eichholzer 2019) for empirical evaluations.

Example 5 *(Heuristics)* Reconsider our DPI from Ex. 1 and the MP m_1 from Ex. 4, and let $\mathbf{D} = \{D_1, \dots, D_4\}$. Further, let $m_3 := A \land \neg B \to C$. Note, m_3 is informative (wrt. $\mathbf{D}$), $\mathbf{D}^+_{m3} = \{D_1, D_2, D_3\}$, $\mathbf{D}^-_{m3} = \{D_4\}$, and the estimations $\hat{p}^+_{m3,\mathbf{D}} = .72, \hat{p}^-_{m3,\mathbf{D}} = .28$ and $\hat{er}^+_{m3,\mathbf{D}} = .25$, $\hat{er}^-_{m3,\mathbf{D}} = .75$. 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 a used sampling technique outputs the sample $\mathbf{D} = \{D_2, D_3, D_4\}$. In this case, we obtain the probability estimates $\hat{p}^+_{m1,\mathbf{D}} = .28, \hat{p}^-_{m1,\mathbf{D}} = .72$ as well as $\hat{p}^+_{m3,\mathbf{D}} = .55, \hat{p}^-_{m3,\mathbf{D}} = .45$. So, using ENT, the chosen MP would be m_3 (the worse MP, as shown above). If sampling would yield $\mathbf{D} = \{D_2, D_4\}$, then m_1 would not even be an informative MP (wrt. $\mathbf{D}$) on the one hand, and m_3 would be the (theoretically) optimal MP as per SPL on the other hand. This example shows the dramatic impact the used sampling technique can have on diagnostic decisions. □

Sequential Diagnosis 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 sequential diagnosis process iterates through the following steps until (the Bayes-updated) $p(D)$ for some $D \in \mathbf{D}$ exceeds a probability threshold σ:

(S1) Generate a sample of minimal diagnoses $\mathbf{D}$ for the current 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 $\mathsf{orcl}(m)$ to update the DPI, by adding m to the positive measurements if $\mathsf{orcl}(m) = P$, and to the negative measurements if $\mathsf{orcl}(m) = N$.

3 Evaluation

We conducted extensive experiments on a dataset of real-world diagnosis cases (Sec. 3.1) to study six different diagnosis sample types (Secs. 3.2 and 3.3) wrt. the accuracy of estimations and diagnostic efficiency (Sec. 3.4). The experiments (Sec. 3.6) target five specific research questions (Sec. 3.5) and their results are analyzed in detail (Sec. 3.7).

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 Tab. 1. Note, every model-based diagnosis problem (according to Reiter's theory (Reiter 1987)) can be represented as a knowledge-base debugging problem (Rodler and Schekotihin 2018), 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 and size of diagnoses, number of components). These metrics are depicted in the columns of Tab. 1. In order to 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), and

(T6) approximate worst-first (awf).

By "best-first" / "worst-first", we mean the most / least probable minimal diagnoses. 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 (tendentially more 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. Throughout, we denote a sample of type Ti including k minimal diagnoses by $S_{Ti,k}$.

3.3 Sampling Techniques

The methods we used to generate the samples for a given DPI $dpi = \langle K, B, P, N \rangle$ are:

***T1*:** We used uniform-cost HS-Tree (Rodler 2015, Sec. 4.6) and stopped it after k diagnoses were computed. Due to the

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

1): Logical expressivity (Baader et al. 2007); the higher it is, the higher is the complexity of consistency checking (diagnosis 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 K.

3): Hardest problems from evaluations in (Shchekotykhin et al. 2012), which were also used, e.g., in (Horridge, Parsia, and Sattler 2009; Ji et al. 2014).

4): Problems studied in (Rodler et al. 2019; Rodler 2020a).

5): Faulty version of DBpedia ontology, see bit.ly/2ZO2qYZ.

6): Problem used in scalability tests in (Shchekotykhin et al. 2012).

Table 1: Experiment dataset (sorted by 2nd column).

best-first property of the algorithm, it is guaranteed (Rodler 2015, Prop. 4.17) that these are the k diagnoses with the highest probability among all minimal diagnoses.

***T2*, *T3*:** We generated all[4] minimal diagnoses $\mathbf{allD}$ for dpi (this can be done by any sound and complete diagnosis computation method, e.g., HS-Tree (Reiter 1987)). 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.

***T4*, *T5*, *T6*:** We used Inv-HS-Tree (Shchekotykhin et al. 2014) to supply the samples. First, we added all $s \in K$ to a list L. For *T5*, we randomly shuffled L. For *T4* and *T6*, we sorted L in descending and ascending order of probability $p(s)$, respectively. Finally, we let plain Inv-HS-Tree operate on this list L to supply a sample of size k (see Appendix A for additional explanations).

3.4 Evaluating Samples

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

Theoretical Representativeness (T-Rep): A sample type Ti is the more representative, the better the *(i)* probability estimates $\langle \hat{p}^+_{m,\mathbf{D}}, \hat{p}^-_{m,\mathbf{D}} \rangle$ for MPs m match the respective actual values $\langle p^+_m, p^-_m \rangle$, *(ii)* 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 (P-Rep): A sampling technique Ti is the more representative, the lower the *(i)* number of measurements required, *(ii)* time needed for sampling (diagnosis computation) in a sequential diagnosis session until the actual diagnosis is isolated from spurious ones, where $\mathbf{D} = S_{Ti,k}$ in each sequential diagnosis iteration.

3.5 Research Questions

We investigate the following research questions:

R1 Which type of sample is best in terms of T-Rep?
R2 Which type of sample is best in terms of P-Rep?
R3 Are the results wrt. *R1* and *R2* consistent over different *(a)* sample sizes, *(b)* measurement selection heuristics, and *(c)* diagnosis problem instances?
R4 Does larger sample size imply better representativeness?
R5 Does better T-Rep translate to better P-Rep?

3.6 Experiments

We ran two experiments, EXP1 and EXP2, to study our research questions. Common to both of them are the following settings:

- We defined one DPI $dpi_K := \langle K, \emptyset, \emptyset, \emptyset \rangle$ for each K in Tab. 1. That is, we assumed each sentence (component) in K to be possibly faulty and left the background knowledge and the measurements void to begin with. To each $s \in K$, we randomly assigned a fault probability $p(s) \in (0, 1)$ in a way that syntactically equally (more) complex sentences have an equal (higher) probability (cf. (Shchekotykhin et al. 2012)). E.g., in our DPI from Ex. 1, elements of $\{s_1, s_3\}$ (one implication, one negation) and $\{s_2, s_4\}$ (one implication), respectively, would each be allocated the same probability, and the former two would have a higher probability than the latter (cf. Ex. 3).
- We precomputed all minimal diagnoses $\mathbf{allD}$ for each DPI dpi_K.
- We used all sample types Ti for $i \in \{1, \dots, 6\}$ (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: (T-Rep) For each dpi_K, for each k, and for each Ti, we computed a sample $\mathbf{D} = S_{Ti,k}$. We used *(i)* $\mathbf{D}$ 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 *(ii)* $\mathbf{allD}$ 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: (P-Rep) For each dpi_K, for each k, for each Ti, and for each of the four heuristics $h \in \{$ENT,SPL,RIO,MPS$\}$ (cf. Sec. 2), we executed 10 diagnosis sessions (loop *S1*–*S4*, Sec. 2) while in each session *(i)* searching for a different randomly selected target diagnosis $D^* \in \mathbf{allD}$ for dpi_K, *(ii)* starting from the initial problem dpi_K, *(iii)* with stop criterion $\sigma = 1$ (loop until a single minimal diagnosis remains, i.e., all others have been ruled out). At this, in each loop iteration, at step *S1*, a sample $\mathbf{D} = S_{Ti,k}$ is drawn for the current DPI, at step *S2*, an optimal informative MP wrt. h is selected, and at step *S3*, an automated oracle classifies

[4] This is generally intractable (Bylander et al. 1991). So, this approach to random sampling is not viable in practice and just used for evaluation purposes. As said in Sec. 3.1, we chose our dataset so that computation of $\mathbf{allD}$ was feasible.

each MP in a way the target diagnosis D^* is not ruled out. For our analyses, we recorded (sampling) times and number of measurements (i.e., loop iterations) throughout a session.

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 for the factors sample type (6 levels), diagnosis problem (8), sample size (5), and heuristic (4).[5]

Presentation *Tab. 2* 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 table column) means that Ti was better than Tj in more factor combinations of the respective scenario than vice versa. Equally ranked sample types are written in parentheses. 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. T-Rep in Sec. 3.4), and a lower avg. number of measurements (M) and, respectively, a lower avg. sample computation time (T) in diagnosis sessions (cf. P-Rep in Sec. 3.4).[6] The idea behind this representation is to give the user of a diagnosis system guidance how to set parameters (diagnosis computation algorithm, number of computed diagnoses, measurement selection heuristic) in order to have the highest chance of achieving best estimations (EXP1) / efficiency (EXP2).

Tab. 3 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: *R1*:[7] *(Elimination rate, criterion E, Tab. 2)* We see that rd is the sample type of choice, as one would expect. In numbers,

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
$h =$ ENT	bf	abf	ard	awf	rd	wf	M
$h =$ SPL	bf	ard	abf	rd	awf	wf	M
$h =$ RIO	rd	ard	awf	bf	abf	wf	M
$h =$ 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
$h =$ ENT	awf	abf	bf	ard	rd	wf	T
$h =$ SPL	$(abf$	awf	$bf)$	ard	rd	wf	T
$h =$ RIO	awf	$(abf$	ard	$bf)$	rd	wf	T
$h =$ MPS	bf	awf	ard	abf	rd	wf	T

Table 2: T-Rep and P-Rep: Rankings of sample types for various scenarios (EXP1 & EXP2).

	best sample type			
scenario	$t = 1$	$t = 10$	$t = 1$ (adj)	$t = 10$ (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

Table 3: Best sample types wrt. overall sequential diagnosis time for various scenarios (EXP2) under the assumption that the time for each measurement equals t minutes. Columns with the predicate "adj" show the results when assuming efficient algorithms for the sample types rd and wf such that they can be computed as fast as the sample type bf.

[5] Please find all further information on the experiments and the results (including the raw data, additional data analyses, pointers to the used code, and information on the computing infrastructure) at http://isbi.aau.at/ontodebug/evaluation.

[6] Tab. 2 does *not* inform 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). And, a higher ranked strategy is not necessarily *always* better than a lower ranked one.

[7] Remarks wrt. *R1*: *(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 co-

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 likely 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. That is, Inv-HS-Tree with a random sorting of its input (cf. Sec. 3.3) does not allow to simulate a random selection of diagnoses.

(Probability, criterion P, Tab. 2) Here, bf proved to be the predominantly superior technique in all depicted scenarios, whereas rd was, surprisingly, only the second best method. Closer analyses of the data revealed that the explanation for this is that often few of the most probable diagnoses already accounted 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, 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.

R2: *(Number of measurements, criterion M, Tab. 2)* 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 $k = 50$, in a diagnosis session using bf, 58 measurements had to be conducted to identify the actual fault vs. only 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 Tab. 2), tends to be no less effective than a real random strategy. Also, observe that wf was in fact the least favorable option in quasi all scenarios.

(Time for diagnosis session, criterion T, Tab. 2) Due to the 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). Hence, there appears to be a *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).

(Overall diagnosis time, criteria T & M combined, Tab. 3) As the outcome for *R2* is not clear-cut when viewing M and T separately, we investigate their combined effect, i.e., the overall (avg.) time for diagnosis sessions for the different sample types. In brief, the conclusions are:

(i) For small sample size (below 10), go with bf.
(ii) For sample size 10, use abf.
(iii) For sample size 20, take awf if the expected time for conducting measurements is low, and take bf else.
(iv) For large sample size (50), use rd if there is an efficient method for it, otherwise use bf.
(v) For ENT or SPL, adopt bf.
(vi) For RIO, if measurement time is short, use rd if there is an efficient method for it, else use bf; if measuring takes longer, use ard.
(vii) For MPS, use ard.

R3: For T-Rep, 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 Tab. 2). There is more variation when comparing results for different diagnosis problems. Nevertheless, results are fairly stable wrt. 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 P-Rep, we see more of a fluctuation over different sample sizes and heuristics, as discussed for *R2* above (cf. variation over entries of each column for M and T in Tab. 2). Examining 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.

R4: Our data indicates a clear trend that increasing sample

efficients were statistically significant ($\alpha = 0.05$), except for a few $k = 6$ scenarios and some scattered $k = 10$ cases.

size leads to better T-Rep (cf. discussion of *R1*). However, it also suggests that there is no general significant positive effect of larger sample size on P-Rep. 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 is in line with similar findings, albeit for lower sample sizes or other types of diagnosis problems, reported by (de Kleer and Raiman 1995; Rodler and Schmid 2018).

R5: From our data, we cannot generally conclude that a better T-Rep implies a better P-Rep (see discussion on *R1*, *R2* and *R3*). E.g., observe the performance of ard in the top (criteria E,P) vs. bottom (criteria M,T) part of Tab. 2. The likely cause of this is that all common measurement selection heuristics are based on a one-step-lookahead (de Kleer, Raiman, and Shirley 1992), where the approximate character of this analysis might lower the benefit of good estimations.

4 Research Limitations

Our evaluations do not come without limitations. In brief, there are the following threats to validity:[8] 1. For feasibility reasons, we *(i)* could not use *all* diagnoses, but only all minimal ones, to determine the real values (EXP1), and *(ii)* had to rely on problem instances that allow the generation of all minimal diagnoses in reasonable time. 2. We focused on binary-outcome measurements which are common in some but not all diagnosis sub-domains, e.g., in knowledge base debugging (Schekotihin, Rodler, and Schmid 2018), circuit diagnosis (de Kleer and Williams 1987), or matrix-based methods (Shakeri et al. 2000).$*$ 3. We did not evaluate *bf* sample types including minimum-cardinality diagnoses, but concentrated on most probable ones.$*$ 4. To keep the size of our dataset manageable, we *(i)* omitted less commonly used existing measurement selection heuristics, and *(ii)* included only a subset of all diagnosis computation methods (sample types) in literature in our analyses.$*$

5 Conclusions

This work addresses an important and fundamental question for model-based diagnosis, i.e., whether the fully biased diagnosis samples primarily used in the field provide a reasonable basis for decision making in spite of not being in line with statistical practices.

The bottom line of our investigations is: Somewhat surprisingly, the best-first samples including the most probable diagnoses commonly used in the field proved to be the best choice in a large fraction of the investigated cases. Yet, we also find that, for certain configurations of a diagnosis system, best-first samples imply drastic overheads compared to other sample types. Random samples, though enabling highly reliable estimations, often led to a worse diagnostic efficiency than biased ones. We discuss reasons for this phenomenon and make recommendations which configurations wrt. diagnosis computation algorithm, measurement selection heuristic and sample size users of diagnosis systems should adopt for best diagnostic performance. E.g., best-first samples are favorable for small sample sizes or when the information gain or split-in-half heuristics are used, whereas random ones are best for larger sample sizes or when adopting the risk optimization or most probable singleton techniques for measurement selection.

Further, our results suggest a time-information trade-off in diagnosis sampling, i.e., more efficient sampling tends to imply less effective measurements. Finally, we find that an approximate, and often efficient, sampling technique based on the Inv-HS-Tree algorithm (Shchekotykhin et al. 2014) in many cases provides a good balance between sampling efficiency and diagnostic effectivity. We believe that these findings, especially the given recommendations which sampling technique to use in particular diagnostic scenarios (cf. Sec. 3.7), can be of high value for both diagnosis researchers and practitioners.

A Sampling by Means of Inv-HS-Tree

To supply a sample of size k, Inv-HS-Tree uses k calls to a diagnosis computation method called Inverse QuickXplain (Inv-QX) (Felfernig, Schubert, and Zehentner 2011; Shchekotykhin et al. 2014), based on the well-known QuickXplain algorithm (Junker 2004; Rodler 2020b). Each call of Inv-QX returns one well-defined minimal diagnosis D_L for the $dpi = \langle K, B, P, N \rangle$ based on the strict total order of elements imposed by the sorting of the list L (which includes all elements of K, cf. Sec. 3.3). Specifically, D_L is the minimal diagnosis with highest rank wrt. the antilexicographic order $>_{\mathsf{antilex}}$ defined on sublists of $L = [l_1, \dots, l_{|K|}]$ (Junker 2004). At this, for sublists X, Y of L, we have $X >_{\mathsf{antilex}} Y$ *(X has higher rank wrt. $>_{\mathsf{antilex}}$ than Y)* iff there is some k such that $X \cap \{l_{k+1}, \dots, l_{|K|}\} = Y \cap \{l_{k+1}, \dots, l_{|K|}\}$ (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 $a, b, \dots, z$ in alphabetic order, then $X >_{\mathsf{antilex}} Y$ 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.

So, in the approximate best-first case (*T4*), the computed diagnosis $D_L = [d_1, \dots, d_{n-1}, d_n]$ has the property that there is no other minimal diagnosis $D' = [d'_1, \dots, d'_r]$ 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 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 (Shchekotykhin et al. 2014).

[8] Those bullet points marked by a “$*$” we plan to address in terms of additional experiments as part of future work.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF), contract P-32445-N38. I thank Fatima Elichanova for her help in implementing the experiments, and Wolfgang Schmid for coordinating the execution of the experiments on our servers. Last but not least, I am grateful to Dietmar Jannach for his general support and valuable advice.

References

Abreu, R.; and Van Gemund, A. J. 2009. A low-cost approximate minimal hitting set algorithm and its application to model-based diagnosis. In *Symposium on Abstraction, Reformulation, and Approximation (SARA)*.

Baader, F.; Calvanese, D.; McGuinness, D. L.; Nardi, D.; and Patel-Schneider, P. F., eds. 2007. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

Bylander, T.; Allemang, D.; Tanner, M.; and Josephson, J. 1991. The computational complexity of abduction. *Artificial Intelligence*, 49: 25–60.

de Kleer, J. 1991. Focusing on Probable Diagnoses. In *AAAI Conference on Artificial Intelligence (AAAI)*.

de Kleer, J.; Mackworth, A. K.; and Reiter, R. 1992. Characterizing diagnoses and systems. *Artificial Intelligence*, 56.

de Kleer, J.; and Raiman, O. 1993. How to diagnose well with very little information. In *Int'l Workshop on Principles of Diagnosis (DX)*.

de Kleer, J.; and Raiman, O. 1995. Trading off the costs of inference vs. probing in diagnosis. In *Int'l Joint Conference on Artificial Intelligence (IJCAI)*.

de Kleer, J.; Raiman, O.; and Shirley, M. 1992. One step lookahead is pretty good. In *Readings in model-based diagnosis*.

de Kleer, J.; and Williams, B. C. 1987. Diagnosing multiple faults. *Artificial Intelligence*, 32(1): 97–130.

de Kleer, J.; and Williams, B. C. 1989. Diagnosis with behavioral modes. In *Int'l Joint Conference on Artificial Intelligence (IJCAI)*.

Domingos, P.; Lowd, D.; Kok, S.; Nath, A.; Poon, H.; Richardson, M.; and Singla, P. 2016. Unifying logical and statistical AI. In *ACM/IEEE Symposium on Logic in Computer Science (LICS)*.

Feldman, A.; Provan, G. M.; and Van Gemund, A. J. 2008. Computing Minimal Diagnoses by Greedy Stochastic Search. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Feldman, A.; Provan, G. M.; and Van Gemund, A. J. 2009. Solving Strong-Fault Diagnostic Models by Model Relaxation. In *Int'l Joint Conference on Artificial Intelligence (IJCAI)*.

Felfernig, A.; Friedrich, G.; Jannach, D.; and Stumptner, M. 2004. Consistency-based diagnosis of configuration knowledge bases. *Artificial Intelligence*, 152(2): 213–234.

Felfernig, A.; Schubert, M.; and Zehentner, C. 2011. An efficient diagnosis algorithm for inconsistent constraint sets. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 26(1): 53–62.

Gonzalez-Sanchez, A.; Abreu, R.; Gross, H.-G.; and Van Gemund, A. J. 2011. Spectrum-Based Sequential Diagnosis. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Horridge, M.; Parsia, B.; and Sattler, U. 2009. Lemmas for Justifications in OWL. *Description Logics*, 477.

Ji, Q.; Gao, Z.; Huang, Z.; and Zhu, M. 2014. Measuring effectiveness of ontology debugging systems. *Knowledge-Based Systems*, 71: 169–186.

Junker, U. 2004. QuickXplain: Preferred Explanations and Relaxations for Over-Constrained Problems. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Kalyanpur, A. 2006. *Debugging and Repair of OWL Ontologies*. Ph.D. thesis, University of Maryland, College Park.

Lucas, P. 2001. Bayesian model-based diagnosis. *International Journal of Approximate Reasoning*, 27(2): 99–119.

Meilicke, C. 2011. *Alignment incoherence in ontology matching*. Ph.D. thesis, Universität Mannheim.

Mengshoel, O.; Chavira, M.; Cascio, K.; Poll, S.; Darwiche, A.; and Uckun, S. 2010. Probabilistic model-based diagnosis: An electrical power system case study. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(5): 874–885.

Moret, B. M. 1982. Decision trees and diagrams. *ACM Computing Surveys*, 14(4): 593–623.

Pattipati, K. R.; and Alexandridis, M. G. 1990. Application of heuristic search and information theory to sequential fault diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(4): 872–887.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

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

Rodler, P. 2015. *Interactive Debugging of Knowledge Bases*. Ph.D. thesis, Alpen-Adria Universität Klagenfurt. http://arxiv.org/pdf/1605.05950v1.pdf.

Rodler, P. 2016. Towards Better Response Times and Higher-Quality Queries in Interactive Knowledge Base Debugging. Technical report, Alpen-Adria Universität Klagenfurt. http://arxiv.org/pdf/1609.02584v2.pdf.

Rodler, P. 2018. On Active Learning Strategies for Sequential Diagnosis. In *Int'l Workshop on Principles of Diagnosis (DX)*.

Rodler, P. 2020. Reuse, Reduce and Recycle: Optimizing Reiter's HS-Tree for Sequential Diagnosis. In *European Conference on Artificial Intelligence (ECAI)*.

Rodler, P. 2022. A formal proof and simple explanation of the QuickXplain algorithm. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10149-w

Rodler, P. 2022. Memory-limited model-based diagnosis. *Artificial Intelligence*, 305: 103681.

Rodler, P.; and Eichholzer, M. 2019. On the Usefulness of Different Expert Question Types for Fault Localization in

Ontologies. In *Int'l Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE)*.

Rodler, P.; and Elichanova, F. 2020. Do we really sample right in model-based diagnosis? In *Int'l Workshop on Principles of Diagnosis (DX)*. https://arxiv.org/pdf/2009.12178.

Rodler, P.; and Herold, M. 2018. StaticHS: A Variant of Reiter's Hitting Set Tree for Efficient Sequential Diagnosis. In *Annual Symposium on Combinatorial Search (SoCS)*.

Rodler, P.; Jannach, D.; Schekotihin, K.; and Fleiss, P. 2019. Are query-based ontology debuggers really helping knowledge engineers? *Knowledge-Based Systems*, 179: 92–107.

Rodler, P.; and Schekotihin, K. 2018. Reducing Model-Based Diagnosis to Knowledge Base Debugging. In *Int'l Workshop on Principles of Diagnosis (DX)*.

Rodler, P.; and Schmid, W. 2018. On the Impact and Proper Use of Heuristics in Test-Driven Ontology Debugging. In *Rules and Reasoning - Int'l Joint Conference (RuleML+RR)*.

Rodler, P.; Shchekotykhin, K.; Fleiss, P.; and Friedrich, G. 2013. RIO: Minimizing User Interaction in Ontology Debugging. In *Web Reasoning and Rule Systems (RR)*.

Schekotihin, K.; Rodler, P.; and Schmid, W. 2018. OntoDebug: Interactive Ontology Debugging Plug-in for Protégé. In *Foundations of Information and Knowledge Systems - Int'l Symposium (FoIKS)*.

Siddiqi, S.; and Huang, J. 2011. Sequential diagnosis by abstraction. *Journal of Artificial Intelligence Research*, 41: 329–365.

Shakeri, M.; Raghavan, V.; Pattipati, K. R.; and Patterson-Hine, A. 2000. Sequential testing algorithms for multiple fault diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(1): 1–14.

Shchekotykhin, K.; Friedrich, G.; Fleiss, P.; and Rodler, P. 2012. Interactive Ontology Debugging: Two Query Strategies for Efficient Fault Localization. *Web Semantics: Science, Services and Agents on the World Wide Web*, 12-13: 88–103.

Shchekotykhin, K.; Friedrich, G.; Rodler, P.; and Fleiss, P. 2014. Sequential diagnosis of high cardinality faults in knowledge-bases by direct diagnosis generation. In *European Conference on Artificial Intelligence (ECAI)*.

Slaney, J. K. 2014. Set-theoretic duality: A fundamental feature of combinatorial optimisation. In *European Conference on Artificial Intelligence (ECAI)*.

Srinivas, S. 1994. A probabilistic approach to hierarchical model-based diagnosis. In *Uncertainty in Artificial Intelligence (UAI)*.

Stern, R.; Kalech, M.; Feldman, A.; Rogov, S.; and Zamir, T. 2013. Finding all diagnoses is redundant. In *Int'l Workshop on Principles of Diagnosis (DX)*.

Zamir, T.; Stern, R. T.; and Kalech, M. 2014. Using Model-Based Diagnosis to Improve Software Testing. In *AAAI Conference on Artificial Intelligence (AAAI)*.