On the Practical Utility of Diagnoses

Ingo Pill
Institute of Software Engineering and Artificial Intelligence, TU Graz, Austria

Johan de Kleer
c-infinity, Mountain View, CA, USA

Abstract

Diagnoses address a universal need for explanations that arises whenever things go wrong or when we observe some unexpected behavior. The ubiquitous need for such explanations has led to an abundance of algorithmic concepts and computation variants for diagnostic processes. Comparisons have mainly focused on highlighting the virtues of individual algorithms or the required computational efforts. In this paper, we focus on evaluating the *results* of a diagnosis algorithm *in the context of an inspection and repair process*. We completely ignore the algorithm's computation concept and whether the observations would justifiably support multiple diagnoses. In particular, we focus on inspecting and discussing the *practical utility* of diagnoses as a measure of the unnecessary efforts one has to spend when fully repairing a system. We start with assessing the utility of a single diagnosis and then progress to evaluating ambiguity groups, i.e., sets of diagnoses. We propose corresponding metrics that investigate worst case and average performance, contrast them to the utility metric used in the DX Competition 2010, and discuss the metrics' results for several scenarios.

2012 ACM Subject Classification Computing methodologies → Causal reasoning and diagnostics

Keywords and phrases Model-based Diagnosis, Diagnosis, Algorithms

Digital Object Identifier 10.4230/OASIcs.DX.2026.9

Acknowledgements We would like to thank the anonymous reviewers for their detailed reviews and very valuable comments.

1 Introduction

As we discussed in detail in [8], comparing and evaluating diagnosis algorithms is a much more complex task than one would expect at first glance. Natural evaluation metrics include correctness and computational efficiency. An intuitive but also essential dimension that has been under-represented in the literature is that of how well a diagnosis algorithm's results support a user in an inspection and repair process. In other words, how well they inform a user about what is actually wrong with the system so that they can effectively repair it.

In this paper, we focus exactly on this question. Instead of focusing on algorithmic traits and computational efficiency, our sole focus is on the practical effectiveness of the reported diagnoses. Intuitively, we are interested in how much a diagnosis would improve the effort we have to spend inspecting and repairing the system. Intuitively, in the worst case, we need to investigate all of the components of a system. In the best case, if we were lucky or well informed by a diagnosis algorithm, we would know *exactly* which components are faulty, so that we need to investigate and repair *only* those faulty components and can avoid working on others. In practice, we will end up somewhere between these two extreme cases, and we would like to be as close to the ideal case as possible to minimize the efforts and costs of the inspection and repair process.

The utility of a diagnosis as described in [3] was a first step in this direction and was used to evaluate the diagnosis algorithms for the DX Competition 2010. However, as we will show for several examples, this previously advocated metric is imprecise and makes estimates that do not measure what we really want. In this paper, we will thus delve deeply into the

question of how to capture *diagnostic utility* in the first place, and we will propose *precise* metrics for the *average* and *worst case utility* (The best case is obvious). Those metrics will cover scenarios where we want to evaluate a single diagnosis as well as sets of diagnoses. We will furthermore propose extended variants that allow us to capture inspection/repair costs that vary between individual system parts. This is relevant in practice, as in real-world applications the costs of the resulting repair process can be much more important than diagnostic precision. In particular, we might require high precision only for some components that are expensive and accept an over-approximation for cheaply replaceable components – which we can deduce also from where we place our sensors and make our measurements. Inspecting and repairing a motor block is certainly much more costly than replacing a spark plug, for example.

Our paper is organized as follows. In the next section, we introduce our notation and offer basic definitions such as the previously published metric to capture a diagnosis' utility. In Section 3, we briefly discuss diagnosis inspired system inspection and repair processes in order to provide the necessary background for evaluating some diagnoses' utility. Our presentation of corresponding utility metrics is divided into Sections 4 and 5 focusing on single- and multi-diagnosis scenarios. We evaluate the average and worst cases and use several examples to illustrate our arguments. In Section 6, we consider how to incorporate more complex considerations of inspection and repair costs into a utility metric. We conclude with a final discussion and an outlook on future work.

2 Notation and Basic Definitions

In this section, we introduce our notation and provide basic definitions. We also describe the basic utility metric that was proposed in [3] and used for the DX Competition 2010.

Please note that our definitions occasionally deviate from what the reader might be familiar with from the literature and papers like [2, 10, 9]. The purpose is to develop evaluation metrics that are agnostic to the algorithmic concept used to derive diagnoses. A simple example is that we do not require diagnoses to be subset-minimal. For Reiter's diagnostic theory [10] it made sense since every super-set of a diagnosis solves the diagnosis problem as well in the context of that theory and we want to avoid unnecessary computations. In our case, however, we do not require subset-minimality in a diagnosis nor do we make any assumption how a diagnosis was computed.

Without loss of generality (see Corollary 8), we assume that a system SD consists of components $x_i \in X$, and that we have some observations OBS about the behavior of SD.[1] Following the literature [2, 10], from SD, X and OBS we construct a diagnosis problem (SD, X, OBS). A diagnosis Δ for a diagnosis problem points us to a subset $X_C \subseteq X$ of faulty components that can explain OBS. A diagnosis assigns to each component $x_i \in X$ a mode $m \in M_i$, where M_i is the finite set of modes in which x_i can be.

The set M_i must contain the nominal mode m_{nom} and at least one fault mode. The latter can be the completely unconstrained general fault mode m_{gen} or a finite non-zero number of concrete fault modes m_k. The general fault mode implements a *Weak Fault Model (WFM)* [2, 10] that does not place restrictions on the faulty behavior of x_i. With concrete fault modes that implement a *Strong Fault Model (SFM)*, we can pose concrete constraints on the faulty behavior. Technically, there is no requirement to describe *exact* behavior (that

[1] As usual, we will also formally refer to a system model with SD and it will be clear from the context whether we refer to the system as an entity or to its model.

defines the component's output for every input combination) , but only some constraint. In practice, we tend to describe exact faulty behavior, though, like when we describe a stuck-at-0 fault for a logic gate. The increased precision over a WFM approach comes at the cost of a larger search space. That is, the search space grows significantly as formalized in Def. 4. Another side effect is that we limit our diagnostic search to the implemented fault modes at the algorithmic level: If we do not model a fault, we will not be able to find it.

We consider nominal mode assignments in a diagnosis Δ to be optional, so that only those tuples (x_i, m_j) in Δ that assign a non-nominal mode $m_j \in M_i \setminus \{m_{nom}\}$ to some component x_i need to be present. All components $x_i \in X$ for which Δ does not contain an explicit assignment are defaulted to the nominal mode.

▶ **Definition 1** (diagnosis problem, diagnosis, subset-minimal diagnosis)**.** *Given a system SD, a set OBS of observations about SD's behavior, a set of components X, and for each $x_i \in X$ a set of modes M_i that x_i can take and where M_i contains x_i's nominal mode m_{nom} and at least one fault mode. A* diagnosis Δ *for a* diagnosis problem (SD, X, OBS) *is defined as a set of mode assignments (x_j, m_k), such that the component x_j is assigned a mode $m \in M_j$ and such that there are no two assignments (x_j, m_k) and (x_l, m_m) that refer to the same component $(x_j = x_l)$. Any $x_j \in X$ for which Δ contains no explicit mode assignment $(x_j, *)$ is assigned m_{nom} by default. A diagnosis Δ is* subset-minimal *if and only if there is no diagnosis Δ' s.t. $\Delta' \subset \Delta$. In other words, there must be no other diagnosis Δ' whose set of explicit mode assignments is a subset of the explicit mode assignments defined by Δ.*

We would like to point out to the reader that, unlike Reiter's theory [10], our definition does not hint at a concept for computing diagnoses. We also do not assume guaranteed features about a diagnosis, such as that *OBS* is consistent with *SD* under the fault assumptions suggested by Δ (as suggested in the same theory). Similarly, we do not require that a diagnosis has a probability greater than zero, as we did in [8]. We simply consider the case where diagnoses are provided as the results of *some* diagnosis algorithm (DA).

▶ **Definition 2** (health state variables)**.** *Please note that for simplicity reasons, for a diagnosis problem (SD, X, OBS) as of Def. 1, we define a set of* health state variables *H that contains for each $x_i \in X$ a corresponding h_i that can take as a value any mode $m_j \in M_i$.*

▶ **Corollary 3.** *Without loss of generality, we can also define a diagnosis problem as (SD, H, OBS). Similarly, a diagnosis and a subset-minimal diagnosis can be defined according to Def. 1 but where we refer to H instead of X.*

▶ **Definition 4** (diagnosis space)**.** *The diagnosis space Θ of a diagnosis problem as of Def. 1 is defined by the product of the components' mode sets M_i*

$$\Theta = \prod_{x_i \in X} M_i$$

If and only if we implement a weak fault model (so that there is the nominal and one unconstrained fault mode m_{gen}) or if there is a single constrained fault mode, we will also write $\Theta = 2^X$. Similarly, we will also write $\Theta = M^X$ in case we have a single universal mode set M that is applicable to all *components (or health state variables).*

▶ **Definition 5** (set of injected or actual faults, ideal diagnosis result)**.** *Ideally, a diagnosis Δ derived by some algorithm is identical to the set $\Delta_Q \in \Theta$ of actual faults in SD. The set Δ_Q might relate to faults that were injected artificially or that occurred naturally.*

The ideal result is user-centric in that we do not refer to complete sets of justifiable diagnoses based on a theory. The goal is to identify the actual faults. Thus, inherent limitations such as limited observability that would have complete algorithms such as RC-Tree [9] or GDE [2] compute multiple justified diagnoses are not taken into account.

▶ **Definition 6** (faulty component set, faulty component candidate set)**.** *For the set of injected faults Δ_Q, we define the* faulty component set $X_F \subseteq X$ *as the set of components x_i that Δ_Q assigns a non-nominal mode. Similarly, for a diagnosis Δ we define the* faulty component candidate set $X_C \subseteq X$ *as the set of components that Δ assigns a non-nominal mode.*

▶ Remark 7. Please note that with our definition of Δ_Q, we implicitly assume that the set of health state variables H in a diagnosis problem was modeled precisely enough, so that an algorithm *can* find the faults. In particular, we assume that $\Delta_Q \in \Theta$.

▶ Remark 8. We assume that a system consists of components. Natural components that have been used in the literature would be statements or functions in a program, gates or blocks in a digital system, physical components in a mechanical system, or any other type of entity or group of entities in *SD*. Technically, there is no restriction to any scheme for defining what a component is. We can literally have any part of *SD* as $x_i \in X$. A component could thus consist of two clauses in a conjunctive normal form description that belong to two individual physical components (like gates in a digital circuit). At the same time, we must take into account that, in general, there might be parts of *SD* (like a wire between two components) that do not belong to any component $x_i \in X$ so that they are not within the scope of any $h_i \in H$.
An important assumption that usually remains hidden in formalizations of a component-based system is that any specific part (like a sentence) of *SD* must not belong to more than one component $x_i \in X$. Not making this assumption would break some effects regarding parsimony/subset-minimality that are actively exploited in computation concepts [10].

In our basic definitions, we do not consider probabilities. This is in contrast to [8], where we assumed a diagnosis to report a probability for each $\Delta' \in \Theta$. As visible in the remainder of this paper, we will still discuss how to adopt probabilities if they are available.

In 2010, the organizers used the following metric to evaluate algorithms in the DX Competition. This metric has the property that $m_{utl\_DXC} = 1$ when $\Delta = \Delta_Q$ (perfect diagnosis), and decreases as false positives or false negatives increase. In Section 4, compared to our metrics, we can see that it suffers from problems and reports unexpected results for certain scenarios. Sometimes, the reported value is even worse than the worst case value that we can precisely compute with our own metrics.

$$m_{utl\_DXC}(\Delta, \Delta_Q, X) = 1 - \frac{|X_C \setminus X_F| * (|X_C| + 1)}{|X| * (|X_C \setminus X_F| + 1)} - \frac{|X_F \setminus X_C| * (|X \setminus X_C| + 1)}{|X| * (|X_F \setminus X_C| + 1)} \quad (1)$$

3 Diagnosis-Inspired System Inspection and Repair

When trying to assess the practical utility of diagnoses, we have to think about how we can exploit a diagnosis during inspection and repair [4]. In order to investigate this, let us first focus on a single diagnosis. This refers to cases where there is either only one diagnosis for a given diagnosis problem, where we isolated a *lead diagnosis* as a working hypothesis, where the diagnosis algorithm reported only one diagnosis, or where we compare multiple diagnoses.

In principle, a diagnosis $\Delta \in \Theta$ defines a set $X_C \subseteq X$ of components that we assume to be faulty (see Def. 6). Depending on the diagnosis space Θ, a diagnosis provides us for each $x_i \in X_C$ with either information that there is *something* wrong with x_i (m_{gen} for a WFM

computation), or detailed information about *what* is wrong with it for an SFM approach[2]. The latter simplifies inspection & repair, but has other side effects, as we discussed in the previous section. However, let us ignore this difference for now and work with X_C. That is, since we can compute it in either case. Please note that we will elaborate on this abstraction in Sec. 6 where we discuss a detailed consideration of costs and the impact of Θ.

Simply put, we should inspect & repair all $x_i \in X_C$. That is, since a diagnosis suggests faulty components and we, ideally, have $X_C = X_F$. Assuming that no fault probabilities are available, we would do so in arbitrary order. For the inspection and repair steps, we would make the implicit/hidden assumption that we have *perfect bug understanding* and similarly that we perform a *perfect repair* [8]. That is, we will always (and immediately) detect a fault when inspecting a component, and we can and will perfectly repair it. An inspection and repair procedure will thus always succeed if the right components are worked on.

In addition, we assume that we can recognize a fully repaired system. Once all faulty components $x_i \in X_F$ have been inspected and repaired, we can thus recognize that the system is fault-free.[3] A welcome effect of this assumption is that if a diagnosis Δ' is a proper super-set of the actual faults Δ_Q (s.t. $\Delta_Q \subset \Delta'$), we can stop the inspection and repair process once we inspected all x_i in X_F rather than having to inspect and repair all $x_i \in X_C$[4]. In this case, Δ suggests for some components that they are faulty while they are actually healthy. We coin these components $x_i \in X_C \setminus X_F$ *false negatives*.

▶ **Definition 9** (false negatives). *For a given diagnosis Δ as defined in Def. 1, the actual faults Δ_Q as of Def. 5, as well as X_C and X_F as of Def. 6, the set of* false negatives *can be derived as $X_{FN} = X_C \setminus X_F$.*

We coin the defective components that Δ fails to report *false positives*. If we suffer from such undetected defect components (s.t. $X_F \setminus X_C \neq \emptyset$), we must also investigate $x_i \in X \setminus X_C$.

▶ **Definition 10** (false positives). *For a given diagnosis Δ as of Def. 1, the actual faults Δ_Q as of Def. 5, as well as X_C and X_F as of Def. 6, the set of* false positives *can be derived as $X_{FP} = X_F \setminus X_C$.*

In many cases, a DA returns more than one diagnosis. For instance, when an algorithm is not precise enough, or when the diagnosis problem has multiple plausible solutions. The latter happens often in practice when the (limited) observations do not allow us to distinguish competing hypotheses. Indeed, when implementing a well-founded and complete analytical concept [10, 2, 9], an algorithm is forced to return multiple diagnoses in these ambiguous cases. Another source for such ambiguity is Θ. When using a weak fault model, it is well-known [10] that when we add a component to a diagnosis, the resulting Δ' can still explain the observed behavior. To illustrate the ambiguity, consider the example of two inverters in series [8]. Assume that we can observe the in- and output of the circuit and that the output is wrong. Since we have no observation about the connecting wire, there is no way to compute which inverter is faulty. Even both inverters could be faulty – in opposition to Occam's razor.

The main question now is how we would actually use a set Θ' of diagnoses. In some cases, each diagnosis might be assigned a probability, so that we rank them from the most probable diagnosis to the least likely one.[5] If we have no probabilities, we could, e.g., take

[2] Such as when a logic gate suffers from a stuck-at-0 fault.
[3] The conclusion is valid only for inspected scenarios and faults that *manifested* on the outputs in *OBS*.
[4] Please note that, as described in [8], this assumption implicitly requires that healthy and inspected/repaired components must not break (again) between inspections of the individual components.
[5] In arbitrary local order whenever there are multiple diagnoses with the same probability

into account the cardinality and/or the frequency with which the components in Δ appear in Θ' to order diagnoses. We could then walk through the diagnoses in this order, testing every new component along the way and discarding from Θ' those diagnoses that we can rule out whenever we learn whether some component is healthy or not.

There is also the option to not rank the diagnoses, but to combine them in a single set. Formally, we derive the union X'_C of the candidate spaces X_C of the individual diagnoses $\Delta_i \in \Theta'$. We would then inspect the components x_i in this *multi-diagnosis fault candidate set*. The order in which we inspect X'_C could be completely random, or, as suggested earlier, we could rank the components according to the frequency with which they appear in diagnoses.

▶ **Definition 11** (multi-diagnosis fault candidate set)**.** *Given a set $\Theta' \subseteq \Theta$ of diagnoses, and computing X_{C_i} for each $\Delta_i \in \Theta'$, the* multi-diagnosis fault candidate set *X'_C is defined as*

$$X'_C \subseteq X = \bigcup_{\Delta_i \in \Theta'} X_{C_i}$$

Obviously, there are an abundance of processes that we could adopt to take advantage of Θ' in an inspection and repair process. As we discussed in [8], all have their individual advantages and disadvantages. In practice, we would tailor the choice to the specific application domain and the structure of the diagnosis problem.

For our evaluation, we need to draw conclusions that generalize. We thus focus on X'_C and a random inspection order. This allows us to avoid any influence of the structure of a specific diagnosis problem and a specific repair process on the average utility. In particular, it is easy to see that the ordering of diagnoses or components[6] would depend on the structural aspects of the specific diagnosis problem which is (a) difficult to predict/estimate and (b) biased toward the sample. Inspecting/repairing the $x_i \in X'_C$ in random order allows us to circumnavigate those problems in a structured way and without introducing a new bias.

It is important to note that sometimes we would specifically look for a biased metric. In a *batch repair* process as suggested by Stern and Kalech in [13] and [14], we, e.g., leverage synergies during repair. Those arise when the costs of repairing a group of components are lower than the sum of the individual repair costs. Their concept of repairing components in batches differs significantly from our step-wise repair assumption and would certainly require an individual metric. As a final note, we would like to specifically note that we assume that we have to work with the available diagnoses only, i.e., since we are interpret those diagnoses. Additional active testing for diagnosis/hypothesis discrimination [1] that is potentially supported by additional signal probing [5] (so that we arrive at solving a sequential diagnosis [11] problem), as well as approaches that integrate diagnosis and repair like [16] are thus not directly applicable and not in our scope.[7]

4 The Practical Utility of a Single Diagnosis

Consider the utility of a single diagnosis and let us first focus on a WFM such that $\Theta = 2^X$. Ideally, the DA would return the actual faults Δ_Q as the diagnosis result (see Def. 5). Intuitively, Δ_Q also represents the *minimal set of components that we need to inspect and repair* to arrive at a fully repaired system. Consequently, we cannot circumvent spending inspection and repair efforts for $x_i \in \Delta_Q$ and we in turn do not spend any *avoidable* efforts

[6] As noted before, we could inspect the $x_i \in X'_C$ according to the number of times they appear in $\Delta_j \in \Theta'$.
[7] The underlying concepts would allow us to compute educated orders for repairing x_i from X_C or X'_C.

if $\Delta = \Delta_Q$. Intuitively, our metric should thus return the maximum utility value for this scenario. In our work, we target a metric $m_{utl}(\Delta, \Delta_Q, X)$ that provides a value in the interval $[0, 1]$, so that we should have $m_{utl}(\Delta, \Delta_Q, X) = 1$ if and only if $\Delta = \Delta_Q$. The minimum value of 0 indicates that we spend the maximum of those efforts that we could theoretically avoid. Practically speaking, it corresponds to the worst case in which we have to investigate *all* components $x_i \in X$, and especially all $x_i \in X \setminus X_F$ since we cannot avoid the $x_i \in X_F$. Having to inspect all components seems a bit extreme, and one might thus be tempted to make the misled assumption that we would never see such a case in practice. But, as we will show later, we can suffer from a diagnosis with utility 0 even for small examples. For the range between the two extremes, we aim for a linear decrease in $m_{utl}(\Delta, \Delta_Q, X)$. This entails that the utility decreases by $\frac{1}{|X \setminus X_F|}$ for any additional (unnecessarily) investigated component since we can compute the precise upper bound on the number of avoidable components as X_F's complement $X_A = |X \setminus X_F|$.

▶ **Definition 12** (necessary/unavoidable and unnecessary/avoidable inspection space)**.** *The entire inspection space is defined via the set of components X (or alternatively H). The* necessary or unavoidable *inspection space $X_U \subseteq X$ is the set of faulty components X_F as of Def. 6. The* unnecessary or avoidable *inspection space X_A is defined as its complement $X_A = X \setminus X_U = X \setminus X_F$.*

▶ **Corollary 13.** *If the avoidable inspection space X_A is empty, then the worst and average utility of a diagnosis is* 1 *by default.*

Proof. (sketch) If we have $X_A = \emptyset$, then $X_F = X$ due to $X_A = X \setminus X_F$. In this case, we have to inspect and repair all $x_i \in X$, no matter which x_i a diagnosis suggests to be faulty. ◀

▶ **Assumption 14.** *Due to Corollary 13, we have that the utility of a diagnosis is* 1 *if $X_A = \emptyset$. For all the metric formulae that we provide, we thus assume $X_A \neq \emptyset$. The reason is twofold. First, we don't need any computation in this case. And second, we would otherwise need to define the special case of $X_A = \emptyset$ for most equations which we will explain for Eq. 2.*

Intuitively speaking, the envisaged metrics capture the fraction of the *avoidable inspection space* X_A that we can *actually* avoid when we take into account a diagnosis Δ.[8] So, let us formally define the metrics for the *worst case* and *average* utility of a diagnosis, starting with a consideration of the worst case. Recall from Sec. 3 how we would exploit some diagnosis Δ. We first investigate all $x_i \in X_C$. Only if this does not result in a working system (s.t. $\Delta_Q \not\subseteq \Delta$), would we investigate components in the complement of X_C. *False positives (see Def. 10)*, i.e., components $x_i \in X_F$ that were reported to work OK so that $x_i \notin X_C$, and *false negatives (see Def. 9)*, i.e., components x_i that are OK ($x_i \notin X_F$) but are reported as defective by Δ (s.t. $x_i \in X_C$) should have an impact on the metric. In fact, in their union they define the actual additional efforts that we need to spend on our inspection/repair process. For instance, if and only if there are no false positives s.t. $\Delta_Q \setminus \Delta = \emptyset$, we can define a diagnosis Δ's worst-case utility via the following equation:

$$m_{utl_{wc}}(\Delta, \Delta_Q, X) = 1 - \frac{|X_C \setminus X_F|}{|X_A|} \tag{2}$$

Taking into account Eq. 2, we can easily see the motivation for the Assumption 14. That is, $X_A = \emptyset$ would lead to a division by 0 so that we would need to define a specific case. Let us now interpret the formula and recall that we basically deduct from the maximum value

[8] By the above obvious conjecture, we cannot avoid looking at the $x_i \in \Delta_Q$ since they are faulty.

1 the fraction of the avoidable inspection space X_A that we actually have to inspect. In a worst case scenario, we would first inspect the false negatives, which means that we have to investigate *all* components $x_i \in X_C$. Since there are no false positives, we stop the inspection after the last $x_i \in X_C$ since we recognize that the system is fault free. If we also suffer from false positives, we also need to check components not in X_C since then $X_F \setminus X_C \neq \emptyset$. In a worst case scenario, we inspect the false positives last which leads us to the following formula:

$$m_{utl_{wc}}(\Delta, \Delta_Q, X) = 1 - \frac{|X_C \setminus X_F|}{|X_A|} - \frac{|X \setminus (X_F \cup X_C)|}{|X_A|} \tag{3}$$

The interested reader will have deduced that we actually have $m_{utl_{wc}} = 0$ then. Formally, this follows from the observation that $|X_C \setminus X_F| + |X \setminus (X_F \cup X_C)| = |X \setminus X_C| = |X_A|$. Nevertheless, we can observe that the two fractions each capture the individual fractions of the unnecessary efforts associated with false negatives and positives. The first captures the efforts associated with false negatives, and the second those connected to false positives.

Now that we have a clear picture about the worst case utility, let us switch our focus to investigating the *average case utility*. Like for the worst case evaluation, let us first assume that there are no false positives. The main difference versus the worst case scenario is that we do not inspect the false negatives first. Rather, we inspect $x_i \in X_C$ in a completely random order – as we discussed in Section 3. When X_C now contains y-many faulty components, and z-many false negatives, [9] the question is how many components from X_C would we inspect on average until all $x_i \in X_F$ have been covered.

Order statistics and in particular the scenario of a hypergeometric distribution[10] provide us with an answer. Consider a simple example where we draw balls from a bowl. Assume that there are n balls (and that we do not replace balls). k-many balls are red and the others green for $0 \leq k \leq n$. Using the following equation, we can estimate the average number of draws that we need to recover all the red balls. The intuition behind the formula is that we estimate the position of the last red ball in a random ordering.

$$E[draws/position] = \frac{k * (n+1)}{k+1} \tag{4}$$

Obviously, this estimate is equal to the number of draws that we need on average to retrieve all the red balls. Recall that, in principle, we are interested not in the total number of picks, but rather in the unnecessary ones that we could have avoided to make. In other words, the picks that produced a green ball instead of a red one. To derive this number, we simply subtract k from the estimate, since we need these k picks to recover all the red balls.

$$E[avoidable\ draws] = E[draws/position] - k = \frac{k * (n+1)}{k+1} - k = \frac{k * (n-k)}{k+1} \tag{5}$$

When transferring *Eq.* 5 to our diagnosis-oriented inspection/repair process, k is the number of faulty components s.t. $k = y = |X_F|$, n the number of components in X_C s.t. $n = y + z = |X_C|$, and $z = n - k = |X_C| - |X_F| = |X_C \setminus X_F|$.[11] For our metric, we also need to divide the number of avoidable inspections/repairs by X_A. This ensures that we subtract the *fraction* of X_A that we could not actually avoid. Please note that we use the subscript X_C to E to indicate that we take into account only false negatives at this time.

$$E[avoidable\ inspections]_{X_C} = \frac{|X_F| * |X_C \setminus X_F|}{(|X_F| + 1)}$$

[9] Since we assume there are no false positives, $y = |X_F|$ and we have $z = |X_C \setminus X_F|$, and $y + z = |X_C|$.
[10] `https://mathworld.wolfram.com/HypergeometricDistribution.html`
[11] Note that $X_F \subseteq X_C$ due to the absence of false positives.

$$m_{utl_{avg}}(\Delta, \Delta_Q, X) = 1 - \frac{E[avoidable\ inspections]_{X_C}}{|X_A|} = 1 - \frac{|X_F| * |X_C \setminus X_F|}{|X_A| * (|X_F| + 1)} \tag{6}$$

Now that we know how to handle false negatives, let us work towards characterizing the effects false positives have on the utility. An important observation that we need to make is that even false negatives must be handled differently *whenever* there are false positives. That is, whenever there are false positives, we cannot receive feedback that the system is repaired when investigating components $x_i \in X_C$. Consequently, we must inspect the entire X_C and cannot avoid the inspection of *any* false negative. With $\frac{|X_C \setminus X_F|}{|X_A|}$, the impact of false negatives on $m_{utl_{avg}}(\Delta, \Delta_Q, X)$ is thus the same as for the worst case utility (see Eq. 3)). Since there are false positives, we must also inspect at least some $x_i \in X \setminus X_C$. To calculate an estimate, how many $x_i \in X \setminus X_C$ we need to inspect and repair on average, we again adapt Eq. 4. This time, the set of components "to draw from" is of size $n = |X \setminus X_C|$, and $k = |X_F \setminus X_C|$ is the number of false positives that we need to find:

$$E[avoidable\ inspections]_{X \setminus X_C} = \frac{|X_F \setminus X_C| * (|X \setminus X_C| - |X_F \setminus X_C|)}{|X_F \setminus X_C| + 1}$$

$$m_{utl_{avg}}(\Delta, \Delta_Q, X) = 1 - \frac{|X_C \setminus X_F|}{|X_A|} - \frac{|X_F \setminus X_C| * (|X \setminus X_C| - |X_F \setminus X_C|)}{|X_A| * (|X_F \setminus X_C| + 1)} \tag{7}$$

So, let us evaluate our metrics in the context of a few examples, and let us compare the results to those for the 2010 DX Competition metric captured in Eq. 1 in Section 2.

1. **Example 1:** $|X| = 10, \Delta_Q = \Delta = \{x_1, x_5, x_6, x_8\} : |X_C| = 4, |X_F| = 4, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 0, |X_A| = |X \setminus X_F| = 6$. In this scenario, we have neither false negatives nor false positives. Intuitively, the utility should be 1 since we make no unnecessary inspections and repairs.
   $m_{utl\_DXC} = 1 - \frac{0*5}{10*1} - \frac{0*7}{10*1} = 1$
   $m_{utl_{wc}} = 1 - \frac{0}{6} = 1$, $m_{utl_{avg}} = 1 - \frac{4*0}{6*5} = 1$
2. **Example 2:** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Delta = \{x_1, x_5, x_6, x_8, x_9\} : |X_C| = 5, |X_F| = 4, |X_C \setminus X_F| = 1, |X_F \setminus X_C| = 0, |X_A| = |X \setminus X_F| = 6$. In this scenario, there is one false negative but no false positive. Intuitively, the utility should thus be lower than 1 since we (potentially) need to investigate one additional component. For the worst case estimate, we can be more precise in that it should be lower by $1/X_A = 1/6$. The average utility should be higher than that. For this example, we can easily observe that the metric used for the DX competition did not catch the average case very well. In particular, the reported value is even *lower* than our precise worst case utility.
   $m_{utl\_DXC} = 1 - \frac{1*6}{10*2} - \frac{0*6}{10*1} = 0.7$
   $m_{utl_{wc}} = 1 - \frac{1}{6} = 0.8\dot{3}$, $m_{utl_{avg}} = 1 - \frac{4*1}{6*5} = 0.8\dot{6}$
3. **Example 3:** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Delta = \{x_1, x_2, x_5, x_6, x_8, x_9\} : |X_C| = 6, |X_F| = 4, |X_C \setminus X_F| = 2, |X_F \setminus X_C| = 0, |X_A| = |X \setminus X_F| = 6$. In this scenario, there are two false negatives but no false positives. Compared to Example 2, the values for all metrics should be lower since we have two false negatives instead of one. We can again observe that the DXC metric reports a lower average utility than our precise worst case utility estimate.
   $m_{utl\_DXC} = 1 - \frac{2*7}{10*3} - \frac{0*5}{10*1} = 0.5\dot{3}$
   $m_{utl_{wc}} = 1 - \frac{2}{6} = 0.\dot{6}$, $m_{utl_{avg}} = 1 - \frac{4*2}{6*5} = 0.7\dot{3}$

4. **Example 4:** $|X| = 10, \Delta_Q = \{x_1, x_5\}, \Delta = \{x_1, x_5, x_6, x_8\} : |X_C| = 4, |X_F| = 2, |X_C \setminus X_F| = 2, |X_F \setminus X_C| = 0, |X_A| = |X \setminus X_F| = 8$. Here we have two false negatives but no false positives. We thus need to investigate two additional components – as in Example 3. The score is still supposed to be different, since the amount of injected faults is different. That is, since this has an impact on the fraction of the components that we could avoid investigating. The DXC average utility metric suffers from the same issue as for the previous two examples and reports a utility that is below the worst case utility.

   $m_{utl\_DXC} = 1 - \frac{2*5}{10*3} - \frac{0*7}{10*1} = 0.\dot{6}$
   $m_{utl_{wc}} = 1 - \frac{2}{8} = 0.75$, $m_{utl_{avg}} = 1 - \frac{2*2}{8*3} = 0.8\dot{3}$

5. **Example 5:** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Delta = \{x_1, x_5, x_8, x_9\} : |X_C| = 4, |X_F| = 4, |X_C \setminus X_F| = 1, |X_F \setminus X_C| = 1, |X_A| = |X \setminus X_F| = 6$. In this scenario, there is one false negative (x_9) and one false positive (x_6). The worst case utility should be 0 since there is a false positive. For the computation of $m_{utl_{wc}}$, note that $|X_F \cup X_C| = 5$. In contrast to the previous three examples, the DXC average utility metric reports a value that is quite close to our average utility $m_{utl_{avg}}$.

   $m_{utl\_DXC} = 1 - \frac{1*5}{10*2} - \frac{1*7}{10*2} = 0.4$
   $m_{utl_{wc}} = 1 - \frac{1}{6} - \frac{5}{6} = 0$, $m_{utl_{avg}} = 1 - \frac{1}{6} - \frac{1*(6-1)}{6*2} = 0.41\dot{6}$

6. **Example 6:** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Delta = \{x_1, x_5, x_8\} : |X_C| = 3, |X_F| = 4, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 1, |X_A| = |X \setminus X_F| = 6$. In this scenario, there is no false negative but one false positive. Due to the false positive, the worst case utility should be 0. For the computation of $m_{utl_{wc}}$ note that we have $X_C \cup X_F = X_F$ for this case. Intuitively, the average utility should be 0.5, i.e., since there is no false negative, and we will find the single false positive in the middle of inspecting $X \setminus X_C$ on average. Our average case metric fulfills this intuition, while the DXC one reports a larger value.

   $m_{utl\_DXC} = 1 - \frac{0*4}{10*1} - \frac{1*8}{10*2} = 0.6$
   $m_{utl_{wc}} = 1 - \frac{0}{6} - \frac{6}{6} = 0$, $m_{utl_{avg}} = 1 - \frac{0}{6} - \frac{1*(7-1)}{6*2} = 0.5$

7. **Example 7:** $|X| = 10, \Delta_Q = \{x_1, x_4, x_5, x_6, x_7, x_8, x_9, x_A\}, \Delta = \{\{x_1, x_2, x_3, x_A\} : |X_C| = 4, |X_F| = 8, |X_C \setminus X_F| = 2, |X_F \setminus X_C| = 6, |X_A| = |X \setminus X_F| = 2$. This scenario is a quite interesting corner case. We have 8 faulty components, but only two (x_1, x_A) are part of the reported diagnosis Δ. Δ contains 2 false negatives (x_2, x_3) which are actually the only two components that are not faulty. Consequently, we have $X_F \cup X_C = X$ for our worst case evaluation. Despite the complicated scenario, the worst case utility should be 0 since we have false positives. In this specific scenario, also the avg. utility should be 0 since we indeed have to investigate *all* components. Our metric matches these expectations, while the DXC metric is slightly off.

   $m_{utl\_DXC} = 1 - \frac{2*5}{10*3} - \frac{6*7}{10*7} = 0.0\dot{6}$
   $m_{utl_{wc}} = 1 - \frac{2}{2} - \frac{0}{2} = 0$, $m_{utl_{avg}} = 1 - \frac{2}{2} - \frac{4*(6-6)}{2*7} = 0$

As we can see in the results, our metrics report values that match the intuitions and expectations we have about a metric that captures the utility of a diagnosis. In some cases also the DXC metric is close, but in others it is clearly off. Sometimes it reports a value that is even below that of the worst case utility, a situation that must not arise for a metric that is supposed to describe the average case.

There are a number of hidden assumptions that our formulae have been based on, such as that there is a health state variable for each component[12]. In practice, this might not be the case, so more general equations can be defined by considering H (the set of health state variables) instead of the component set$X$. For example, Eq. 7 could be reformulated to

$$m_{utl_{avg}}(\Delta, \Delta_Q, H) = 1 - \frac{|H_C \setminus H_F|}{|H_A|} - \frac{|H_F \setminus H_C| * (|H \setminus H_C| - |H_F \setminus H_C|)}{|H_A| * (|H_F \setminus H_C| + 1)} \tag{8}$$

We furthermore assumed that the cost of inspecting any component is unity. In reality, the costs for inspecting and repairing components differ between individual components. Consequently, we might want to consider a cost function, which we discuss in Section 6.

A most prominent assumption was that we focused first on a weak fault model with the diagnosis space $\Theta = 2^X$. The imminent question now is whether the move to a strong fault model with a larger Θ would require us to derive new formulae, and our answer is *yes and no*. That is, the answer depends on how we define the costs of inspecting a component.
If we approximate the costs with unity no matter the fault model, and if we assume perfect bug understanding, we can indeed use all formulae proposed so far also for the SFM case. Only if we do not assume unity, we need to use different formulae. Since this assumes a cost function, let us continue the discussion in Section 6.

5 Considering Ambiguity Groups

In many if not most practical cases, a DA will return more than one diagnosis – as we analyzed in Sec. 3. Like we argued in the same section, there are multiple options to exploit the set Θ' of reported diagnoses. An important and inherent challenge is that the structure of Θ' depends on the structure of the system, the observations, and Δ_Q. For illustration purposes, assume that we pick a certain element for inspection. Depending on whether the component is actually faulty, a certain set of diagnoses in Θ' will become invalid, so that a working set of hypotheses or X'_C would need to be re-computed. Depending on the metrics used to select a component, we might need to update also the ranking of the individual diagnoses and components that we follow in the inspection and repair process. For a precise average utility metric, we would need to estimate the effects of this process in a meaningful way which is a quite complex challenge. For this and other reasons raised in Sec. 3, we chose to focus on the multi-diagnosis fault candidate set X'_C as of Def. 11.

Like for single diagnoses, we initially assume $\Theta = 2^X$ and start our discussion with the worst case utility. It is easy to see that we can easily transfer the worst case utility equations for the single diagnosis case to cover multiple diagnoses by considering X'_C instead of X_C. If we assume that there is no false positive (see Def. 10), i.e., $X_F \setminus X'_C = \emptyset$, we arrive at:

$$m_{utl_{wc}}(\Theta', \Delta_Q, X) = 1 - \frac{|X'_C \setminus X_F|}{|X_A|} \tag{9}$$

Like for a single diagnosis, we basically deduct from the maximum utility value the fraction of the avoidable inspection space X_A that we actually have to inspect. If there are false positives, we need to also look at components in X'_C's complement since $X_F \setminus X'_C \neq \emptyset$. In the worst case, we inspect the false positives last, so that we arrive at:

$$m_{utl_{wc}}(\Theta', \Delta_Q, X) = 1 - \frac{|X'_C \setminus X_F|}{|X_A|} - \frac{|X \setminus (X_F \cup X'_C)|}{|X_A|} \tag{10}$$

[12] In fact, we argued primarily with X rather than H.

Please note that Cor. 13 transfers also to X'_C and Assumption 14 is thus valid also when considering ambiguity groups. The average utility situation is more complex. Since being utterly precise is not an option due to the issues discussed in Sec. 3, we investigated estimations and present two variants. For $m_{utl_{avg\_est}}(\Theta', \Delta_Q, X)$, we compute X'_C and transfer the formulae we proposed for the single diagnosis case. If there are no false positives ($X'_C \setminus X_F = \emptyset$), we thus have:

$$m_{utl_{avg\_est}}(\Theta', \Delta_Q, X) = 1 - \frac{|X_F| * |X'_C \setminus X_F|}{|X_A| * (|X_F| + 1)} \tag{11}$$

If there are also false positives $X'_C \setminus X_F \neq \emptyset$, we would use

$$m_{utl_{avg\_est}}(\Theta', \Delta_Q, X) = 1 - \frac{|X'_C \setminus X_F|}{|X_A|} - \frac{|X_F \setminus X'_C| * (|X \setminus X'_C| - |X_F \setminus X'_C|)}{|X_A| * (|X_F \setminus X'_C| + 1)} \tag{12}$$

For both equations, we compute X'_C and randomly process components from X'_C and thus connect to order statistics. A welcome side effect of these formulae is that the obtained values compare nicely to those for the single diagnosis case. That is, when $|\Theta'| = 1$ and $X'_C = X_C$ the results are indeed the same as if we had used the single diagnosis formula.

We investigated also an alternative $m_{utl_{avg\_est'}}(\Theta', \Delta_Q, \rho, X)$ that does not focus on computing X'_C. To compute it, we, in principle, average over the average utility of all $\Delta_i \in \Theta'$. If we have probabilities for the individual diagnoses, we can also incorporate weights. To that end, we consider a function $\rho(\Delta)$ that gives us the probability of some $\Delta \in \Theta'$. With $1 = \sum_{\Delta_i \in \Theta'} \rho(\Delta_i)$ and $\rho(\Delta_i) = \frac{1}{|\Theta'|}$ in case the DA does not report probabilities, we have

$$m_{utl_{avg\_est'}}(\Theta', \Delta_Q, \rho, X) = \sum_{\Delta_i \in \Theta'} \rho(\Delta_i) * m_{utl_{avg}}(\Delta_i, \Delta_Q, X) \tag{13}$$

Let us now test the metrics for some examples that we derived from those used in Sec. 4.

- **Example 1':** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Theta' = \{\{x_1, x_5, x_6\}, \{x_5, x_8, x_1\}, \{x_6, x_8\}\}$: $|X_F| = 4, |X'_C| = 4, |X'_C \setminus X_F| = 0, |X_F \setminus X'_C| = 0, |X_A| = |X \setminus X_F| = 6$. In this variant of Example 1, the DA returns 3 diagnoses (Θ'), and none of these diagnoses represents Δ_Q. However, their union X'_C is equal to Δ_Q, so we have neither false negatives nor false positives. Intuitively, the worst case and average utility for our X'_C-oriented approach should therefore be 1 since the DA identified *all* faulty components and *only* faulty components. The fact that no diagnosis is equal to Δ_Q, is actually irrelevant to the repair and inspection process. In particular, even if we did not follow the process described in Section 3 but consider the diagnoses in Θ' in any imaginable way and order, would we inspect and repair faulty components only.
  $m_{utl_{wc}} = 1 - \frac{0}{6} = 1$, $m_{utl_{avg\_est}} = 1 - \frac{4*0}{6*5} = 1$
  For $m_{utl_{avg\_est_2}}$, we would compute the average over 3 bad single diagnosis utilities where Δ_i refers to the i-th diagnosis in Θ. For Δ_1, $|X_C| = 3, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 1$. For Δ_2, we have $|X_C| = 3, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 1$, and for Δ_3, $|X_C| = 2, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 2$. So we do not have false negatives, but false positives for each Δ_i.
  $m_{utl_{avg}}(\Delta_1) = 1 - \frac{0}{6} - \frac{1*(7-1)}{6*2} = 0.5$,
  $m_{utl_{avg}}(\Delta_2) = 1 - \frac{0}{6} - \frac{1*(7-1)}{6*2} = 0.5$,
  $m_{utl_{avg}}(\Delta_3) = 1 - \frac{0}{6} - \frac{2*(8-2)}{6*3} = 0.\dot{3}$
  $m_{utl_{avg\_est'}}(\Theta') = \frac{1}{3} * (0.5 + 0.5 + 0.\dot{3}) = 0.\dot{4}$
  From this single example, we can easily see that $m_{utl_{avg\_est'}}$ has serious problems in assessing the utility of Θ'. In fact, it is a good indicator of how good individual diagnoses are on average, but it does not tell us the utility of Θ'.

- **Example 2':** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Theta' = \{\{x_1, x_5, x_6\}, \{x_5, x_8, x_1, x_9\}, \{x_6, x_8\}\} : |X'_C| = 5, |X_F| = 4, |X'_C \setminus X_F| = 1, |X_F \setminus X'_C| = 0, |X_A| = |X \setminus X_F| = 6$. In this scenario, there is one false negative but no false positive for X'_C. Intuitively, the utility should thus be slightly lower than 1 since we (potentially) need to investigate one additional component. For the worst case estimate, we can be more precise in that it should be lower by $1/X_A = 1/6$.
  $m_{utl_{wc}} = 1 - \frac{1}{6} = 0.8\dot{3}$, $m_{utl_{avg\_est}} = 1 - \frac{4*1}{6*5} = 0.8\dot{6}$
  Now, let us look at how $m_{utl_{avg\_est'}}$ works for this example. For Δ_1, $|X_C| = 3, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 1$. For Δ_2, we have $|X_C| = 4, |X_C \setminus X_F| = 1, |X_F \setminus X_C| = 1$, and for Δ_3, $|X_C| = 2, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 2$. So we have false positives for each Δ_i, and for Δ_2 there is also one false negative.
  $m_{utl_{avg}}(\Delta_1) = 1 - \frac{0}{6} - \frac{1*(7-1)}{6*2} = 0.5$
  $m_{utl_{avg}}(\Delta_2) = 1 - \frac{1}{6} - \frac{1*(6-1)}{6*2} = 0.41\dot{6}$
  $m_{utl_{avg}}(\Delta_3) = 1 - \frac{0}{6} - \frac{2*(8-2)}{6*3} = 0.\dot{3}$
  $m_{utl_{avg\_est'}}(\Theta') = \frac{1}{3} * (0.5 + 0.41\dot{6} + 0,\dot{3}) = 0.41\dot{6}$
  Compared to the previous example, the false positive for Δ_2 further decreases $m_{utl_{avg\_est'}}$ – as expected and as should be. We again observe that this metric is more indicative of the distance between Δ_i and Δ_Q in terms of efforts, rather than describing the utility of Θ'. That is, for this example, there is only one false negative for which we have to waste our efforts when following our inspection and repair approach for Θ'. Since $m_{utl_{avg\_est'}}$ does not implement the desired intuitions, we skip it for the remaining examples.
- **Example 3':** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Theta' = \{\{x_1, x_5, x_6\}, \{x_5, x_8, x_1, x_9\}, \{x_5, x_6, x_8, x_2\}\} : |X'_C| = 6, |X_F| = 4, |X'_C \setminus X_F| = 2, |X_F \setminus X'_C| = 0, |X_A| = |X \setminus X_F| = 6$. This is a multi-diagnosis variant of Example 3 with the same values reported by the respective metrics. That is, X'_C is equal to X_C derived for Example 3.
  $m_{utl_{wc}} = 1 - \frac{2}{6} = 0.6\dot{6}$, $m_{utl_{avg\_est}} = 1 - \frac{4*2}{6*5} = 0.7\dot{3}$
- **Example 4':** $|X| = 10, \Delta_Q = \{x_1, x_5\}, \Theta' = \{\{x_1, x_5, x_6\}, \{x_5, x_8, x_1\}, \{x_5, x_6, x_8\}\} : |X'_C| = 4, |X_F| = 2, |X'_C \setminus X_F| = 2, |X_F \setminus X'_C| = 0, |X_A| = |X \setminus X_F| = 8$. In this scenario, X'_C contains two false negatives, but no false positives. Like for Example 3', we need to investigate two additional components. The score is still supposed to be different than for Example 3', since the amount of injected faults differs between the examples.
  $m_{utl_{wc}} = 1 - \frac{2}{8} = 0.75$, $m_{utl_{avg\_est}} = 1 - \frac{2*2}{8*3} = 0.8\dot{3}$
- **Example 5':** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Theta' = \{\{x_1, x_5, x_9\}, \{x_5, x_8\}, \{x_1, x_8\}\} : |X'_C| = 4, |X_F| = 4, |X'_C \setminus X_F| = 1, |X_F \setminus X'_C| = 1, |X_A| = |X \setminus X_F| = 6$. In this scenario, there are one false negative (x_9) and one false positive (x_6) in X'_C. The worst case utility should be 0 since there is a false positive. For the computation of $m_{utl_{wc}}$, note that $|X_F \cup X'_C| = 5$.
  $m_{utl_{wc}} = 1 - \frac{1}{6} - \frac{5}{6} = 0$, $m_{utl_{avg\_est}} = 1 - \frac{1}{6} - \frac{1*(6-1)}{6*2} = 0.41\dot{6}$
- **Example 6':** $|X| = 10, \Delta_Q = \{x_1, x_5, x_6, x_8\}, \Theta' = \{\{x_1, x_5\}, \{x_5, x_8\}, \{x_1, x_8\}\} : |X'_C| = 3, |X_F| = 4, |X'_C \setminus X_F| = 0, |X_F \setminus X'_C| = 1, |X_A| = |X \setminus X_F| = 6$. In this scenario, there is no false negative but one false positive. Due to the false positive, the worst case utility should be 0. For the computation of $m_{utl_{wc}}$ note that we have $X_C \cup X_F = X_F$ for this case. Intuitively, the average utility should be 0.5 as there is no false negative, and, on average, we will find the remaining false positive (x_6) in the middle of inspecting $X \setminus X_C$. $m_{utl_{wc}} = 1 - \frac{0}{6} - \frac{6}{6} = 0$, $m_{utl_{avg\_est}} = 1 - \frac{0}{6} - \frac{1*(7-1)}{6*2} = 0.5$
- **Example 7':** $|X| = 10, \Delta_Q = \{x_1, x_4, x_5, x_6, x_7, x_8, x_9, x_A\}, \Theta' = \{\{x_A, x_2\}, \{x_3\}, \{x_1, x_A\}\} : |X'_C| = 4, |X_F| = 8, |X'_C \setminus X_F| = 2, |X_F \setminus X'_C| = 6, |X_A| = |X \setminus X_F| = 2$. This scenario is a quite interesting corner case. We have 8 faulty components, but only two

are part of X'_C. X'_C contains 2 false negatives that are actually the only two components (x_2, x_3) that are not faulty. Consequently, we have $X_F \cup X'_C = X$ for our worst case evaluation. Despite the complicated scenario, the worst case utility should be 0 since we have false positives. In this specific scenario, also the avg. utility should be 0 since we indeed have to investigate all components due to the specific structure of the diagnosis and the diagnosis problem.
$m_{utl_{wc}} = 1 - \frac{2}{2} - \frac{0}{2} = 0$, $m_{utl_{avg\_est}} = 1 - \frac{2}{2} - \frac{6*(6-6)}{2*7} = 0$

While $m_{utl_{avg\_est'}}$ might seem intuitive at first glance, we can easily observe that its results are counterintuitive. The particular problem is that we might average over potentially bad single diagnosis utilities which causes us to miss that the information contained in Θ' in its entirety can lead to a *perfect* inspection and repair process. We still chose to report it in order to highlight this observation. $m_{utl_{avg\_est'}}$ on the other hand, leverages the synergies between the individual $\Delta_i \in \Theta'$ via its computation of X'_C. As we saw, even for non-ideal Δ_is such that none of them is equal to Δ_Q, we can achieve a perfect inspection and repair process with a utility score of 1. That is, as long as we have no false negatives[13] in any $\Delta_i \in \Theta'$, nor false positives[14] in X'_C, the utility of Θ' can and *should* indeed be 1.

▶ **Proposition 15.** *Given a diagnosis problem (SD, X, OBS) with diagnosis space $\Theta = 2^X$, the injected faults Δ_Q, and a multi-diagnosis solution $\Theta' \subseteq \Theta$ provided by some diagnosis algorithm. The worst case and average case utility for Θ' can have an ideal value of 1 even if Θ' does not contain Δ_Q as a diagnosis.*

Proof. The proof follows directly from the empirical evidence for Example 1', its discussion, and the definition of our inspection and repair process in Section 3. ◀

Like for the single diagnosis scenario, there is the question of whether an SFM diagnosis space Θ would require us to use different formulae to describe the utility of Θ'. And again we must answer this question taking the repair costs into account. However, the situation is slightly different since the individual $\Delta_i \in \Theta'$ might suggest different fault modes for the same component. So, if we approximate the costs of inspecting and repairing a component in X'_C with unity, no matter the fault modes suggested by the individual Δ_i, and if we assume perfect bug understanding, we can indeed use all formulae proposed so far also for the SFM case. Only if we do not assume unity, do we need to use different formulae and replace X'_C with a more complex data structure. Due to the inherent connection to cost functions, let us continue the discussion in the next section.

6 Taking Individual Costs Into Account

A prominent assumption made for the previously proposed metric is that inspection and repair costs should be equal for all components. This is certainly not the case in practice, as we can observe when comparing the inspection and repair costs for a motor block with those for a spark plug. Consequently, for real-world applications, we might be interested in metrics that take into account variations in those costs.

To define the corresponding variants of our metric, assume that there is a cost function $c(x_i) : X \to \mathbb{R}^+$ that reports for each component x_i (or health state variable if preferred) a positive real value indicating the costs of investigating and repairing x_i.

[13] Healthy components that are contained in one $\Delta_i \in \Theta'$ (see Def. 9).
[14] Faulty components that are part of no $\Delta_i \in \Theta'$ so that they are not part of X'_C (see Def. 10).

▶ **Definition 16.** *Given a set of components X, a cost function $c(x \in X)$ is defined as function $c() : X \to \mathbb{R}^+$ that maps each component $x_i \in X$ to a positive real value $r \in \mathbb{R}^+$ that describes the costs of inspecting* and *repairing x_i.*

Please note that splitting the costs for inspecting and repairing a component is indeed possible. Sometimes one would even be able to provide more exact results as we mentioned in Sec. 3. A simple example is when we have only two components and inspect the first one. Depending on the result, we will indeed need to repair only one of the components. This split adds some additional complexity to the formulae and argumentation line[15], though, so that we chose to consider unified costs in this manuscript. For considering costs instead of the *number* of false positives and false negatives, we make the following three updates:

1. We do not consider the cardinality of a diagnosis or the number of components in certain sets but the sum of the associated costs.
2. The *unavoidable costs* of the *unavoidable efforts* are given by the sum of the costs of inspecting and repairing all $x_i \in X_F$.
3. The *avoidable costs* of the *avoidable efforts* are given by the sum of the costs of inspecting and repairing all $x_i \in X \setminus X_F$

For a simpler notation, let us define a cost function $C(X')$ that computes for a given component subset $X' \subseteq X$ the costs for inspecting and repairing all $x_i \in X'$:

▶ **Definition 17.** *For a given subset X' of X, and a cost function $c()$ following Def. 16, let us define a cost function $C(X') = \sum_{x_i \in X'} c(x_i)$.*

Replacing the size $|X'|$ of a component subset with costs $C(X')$ allows us to generate variants of our previous equations that work with costs. In some of the earlier equations, we see a $+1$ for some terms. It had the meaning of one additional component. When considering costs, we thus replace this $+1$ with the adding average costs over all elements. The motivation is simple. In the original formula, we add 1 to the number of components in a specific subset, and now we add the average costs for one component to the sum of the costs.[16] Implementing this translation concept, we end up with the equations listed in the next subsection.

In Subsection 6.2, we will continue our discussion of whether using a strong fault model with multiple fault modes would require us to update our metrics.

6.1 The WFM Case

Based on the proposed transfer concept, we can derive the following metrics. Let us start with the single diagnosis metrics from Section 4. From Equations 2,3 (no false positives/with false positives) we can derive the worst case utility as (remember that $X_A = X \setminus X_F$):

$$mC_{utl_{wc}}(\Delta, \Delta_Q, X, C()) = 1 - \frac{C(X_C \setminus X_F)}{C(X_A)} \tag{14}$$

$$mC_{utl_{wc}}(\Delta, \Delta_Q, X, C()) = 1 - \frac{C(X_C \setminus X_F)}{C(X_A)} - \frac{C(X \setminus (X_F \cup X_C))}{C(X_A)} \tag{15}$$

[15] In order to be precise, we would then need to estimate costs when we know by deduction (and not inspection) that a component is actually wrong.

[16] In order to take the exact term into account more precisely, we could also use, e.g., the average over the complement of the subset considered in the term.

For the average utility, we get from Equations 6,7 (no false positives/with false positives) the following formulae, where $c_{avg} = \frac{1}{|X|} \sum_{x_i \in X} c(x_i)$ represents the average costs of inspecting and repairing some component.

$$mC_{utl_{avg}}(\Delta, \Delta_Q, X, C()) = 1 - \frac{C(X_F) * C(X_C \setminus X_F)}{C(X_A) * (C(X_F) + c_{avg})} \tag{16}$$

$$mC_{utl_{avg}}(\Delta, \Delta_Q, X, C()) = 1 - \frac{C(X_C \setminus X_F)}{C(X_A)} - \frac{C(X_F \setminus X_C) * (C(X \setminus X_C) - C(X_F \setminus X_C))}{C(X_A) * (C(X_F \setminus X_C) + c_{avg})} \tag{17}$$

Note that we can refer also to H instead of X like in Eq. 7. We only need to overload $C()$ to describe the costs for the SD parts in the scope of H rather than for some component:

$$mC_{utl_{avg}}(\Delta, \Delta_Q, H, C()) = 1 - \frac{C(H_C \setminus H_F)}{C(H_A)} - \frac{C(H_F \setminus H_C) * (C(H \setminus H_C) - C(H_F \setminus H_C))}{C(H_A) * (C(H_F \setminus H_C) + c_{avg})} \tag{18}$$

Proceeding to the multi-diagnosis utility metrics from Section 5, the formulae for the worst case utility can be deduced from Eqs. 9,10 (no false positives/with false positives):

$$mC_{utl_{wc}}(\Theta', \Delta_Q, X, C()) = 1 - \frac{C(X'_C \setminus X_F)}{C(X_A)} \tag{19}$$

$$mC_{utl_{wc}}(\Theta', \Delta_Q, X, C()) = 1 - \frac{C(X'_C \setminus X_F)}{C(X_A)} - \frac{C(X \setminus (X_F \cup X'_C))}{C(X_A)} \tag{20}$$

The cost-oriented variants for Eqs. 11, 12 (no false positives/with false positives) that encode our first average utility estimation variant are defined as follows:

$$mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C()) = 1 - \frac{C(X_F) * C(X'_C \setminus X_F)}{C(X_A) * (C(X_F) + c_{avg})} \tag{21}$$

$$mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C()) = 1 - \frac{C(X'_C \setminus X_F)}{C(X_A)} - \frac{C(X_F \setminus X'_C) * (C(X \setminus X'_C) - C(X_F \setminus X'_C))}{C(X_A) * (C(X_F \setminus X'_C) + c_{avg})} \tag{22}$$

Since the experimental evaluation in Sec. 5 showed that $m_{utl_{avg\_est'}}$ does not explain the utility of some Θ' very well, we do not provide a cost-related version. Now, let us revisit Examples 1-3,6-7 in order to illustrate the cost-based metrics. Note that cost functions are described as lists such that the costs are ordered according to the component ID.

- **Example 1/1':** $|X| = 10, \Delta_Q = X_F = \{x_1, x_5, x_6, x_8\}, |X_F| = 4, |X_A| = |X \setminus X_F| = 6$, two cost functions $C_1() = [1,1,1,1,1,1,1,1,1,1]$, $C_2() = [1,2,3,4,5,1,2,3,4,5]$ s.t $c_{avg_1} = 1$ and $c_{avg_2} = 3$.

  $\Delta = X_C = \Delta_Q : |X_C| = 4, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 0$
  $mC_{utl_{wc}}(\Delta, \Delta_Q, X, C_1()) = 1 - \frac{0}{1+1+1+1+1+1} = 1$ (via Eq. 14)
  $mC_{utl_{wc}}(\Delta, \Delta_Q, X, C_2()) = 1 - \frac{0}{2+3+4+2+4+5} = 1$ (via Eq. 14)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_1()) = 1 - \frac{(1+1+1+1)*0}{(1+1+1+1+1+1)*((1+1+1+1)+1)} = 1$ (via Eq. 16)

$mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_2()) = 1 - \frac{(1+5+1+3)*0}{(2+3+4+2+4+5)*((1+5+1+3)+3)} = 1$ (via Eq. 16)

$\Theta' = \{\{x_1, x_5, x_6\}, \{x_5, x_8, x_1\}, \{x_6, x_8\}\} : X'_C = \Delta_Q, |X'_C| = 4, |X'_C \setminus X_F| = 0, |X_F \setminus X'_C| = 0.$
$mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{0}{1+1+1+1+1+1} = 1$ (via Eq. 19)
$mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{0}{2+3+4+2+4+5} = 1$ (via Eq. 19)
$mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{(1+1+1+1)*0}{(1+1+1+1+1+1)*((1+1+1+1)+1)} = 1$ (via Eq. 21)
$mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{(1+5+1+3)*0}{(2+3+4+2+4+5)*((1+5+1+3)+3)} = 1$ (via Eq. 21)
Intuitively, since we have no false positives nor false negatives, the utility should be 1 regardless of the utilized cost function, whether we consider the worst or average case, and whether we have one or more diagnoses.

- **Example 2/2':** $|X| = 10, \Delta_Q = X_F = \{x_1, x_5, x_6, x_8\}, |X_F| = 4, |X_A| = |X \setminus X_F| = 6$, $C_1() = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$, $C_2() = [1, 2, 3, 4, 5, 1, 2, 3, 4, 5]$ s.t $c_{avg_1} = 1$ and $c_{avg_2} = 3$.

  $\Delta = \{x_1, x_5, x_6, x_8, x_9\} : |X_C| = 5, |X_C \setminus X_F| = 1, |X_F \setminus X_C| = 0$
  $mC_{utl_{wc}}(\Delta, \Delta_Q, X, C_1()) = 1 - \frac{1}{6} = 0.8\dot{3}$ (via Eq. 14)
  $mC_{utl_{wc}}(\Delta, \Delta_Q, X, C_2()) = 1 - \frac{4}{2+3+4+2+4+5} = 0.8$ (via Eq. 14)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_1()) = 1 - \frac{(1+1+1+1)*1}{(1+1+1+1+1+1)*((1+1+1+1)+1)} = 0.8\dot{6}$ (via Eq. 16)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_2()) = 1 - \frac{(1+5+1+3)*4}{(2+3+4+2+4+5)*((1+5+1+3)+3)} = 0.846$ (via Eq. 16)

  $\Theta' = \{\{x_1, x_5, x_6\}, \{x_5, x_8, x_1, x_9\}, \{x_6, x_8\}\} : X'_C = \{x_1, x_5, x_6, x_8, x_9\}, |X'_C| = 5,$
  $|X'_C \setminus X_F| = 1, |X_F \setminus X'_C| = 0.$
  $mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{1}{6} = 0.8\dot{3}$ (via Eq. 19)
  $mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{4}{2+3+4+2+4+5} = 0.8$ (via Eq. 19)
  $mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{(1+1+1+1)*1}{(1+1+1+1+1+1)*((1+1+1+1)+1)} = 0.8\dot{6}$ (via Eq. 21)
  $mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{(1+5+1+3)*4}{(2+3+4+2+4+5)*((1+5+1+3)+3)} = 0.846$ (via Eq. 21) As we can observe, if we assume unity for the costs we get the same results as when not using a cost-related metric. It is easy to see that this observation is true also if the cost function has a uniform distribution with a higher cost value like 2. Between C1 and C2, the expected penalty for C2 results from the fact that the false negative x_9 has higher (4) than average ($c_{avg_2} = 3$) costs.

- **Example 3/3':** $|X| = 10, \Delta_Q = X_F = \{x_1, x_5, x_6, x_8\}, |X_F| = 4, |X_A| = |X \setminus X_F| = 6$, $C_1() = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$, $C_2() = [1, 2, 3, 4, 5, 1, 2, 3, 4, 5]$ s.t $c_{avg_1} = 1$ and $c_{avg_2} = 3$.

  $\Delta = X_C = \{x_1, x_5, x_6, x_8, x_3, x_9\} : |X_C| = 6, |X_C \setminus X_F| = 2, |X_F \setminus X_C| = 0$
  $mC_{utl_{wc}}(\Delta, \Delta_Q, X, C_1()) = 1 - \frac{2}{6} = 0.\dot{6}$ (via Eq. 14)
  $mC_{utl_{wc}}(\Delta, \Delta_Q, X, C_2()) = 1 - \frac{3+4}{2+3+4+2+4+5} = 0.65$ (via Eq. 14)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_1()) = 1 - \frac{(1+1+1+1)*2}{(1+1+1+1+1+1)*((1+1+1+1)+1)} = 0.7\dot{3}$ (via Eq. 16)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_2()) = 1 - \frac{(1+5+1+3)*(3+4)}{(2+3+4+2+4+5)*((1+5+1+3)+3)} \approx 0.731$ (via Eq. 16)

  $\Theta' = \{\{x_1, x_5, x_6\}, \{x_5, x_8, x_1, x_9\}, \{x_5, x_6, x_8, x_3\}\} : X'_C = \{x_1, x_5, x_6, x_8, x_3, x_9\},$
  $|X'_C| = 6, |X'_C \setminus X_F| = 2, |X_F \setminus X'_C| = 0.$
  $mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{2}{6} = 0.\dot{6}$ (via Eq. 19)
  $mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{7}{20} = 0.65$ (via Eq. 19)
  $mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{4*2}{6*(4+1)} = 0.7\dot{3}$ (via Eq. 21)
  $mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{10*7}{20*(10+3)} \approx 0.731$ (via Eq. 21)

- **Example 6/6':** $|X| = 10, \Delta_Q = X_F = \{x_1, x_5, x_6, x_8\}, |X_F| = 4, |X_A| = |X \setminus X_F| = 6$, $C_1() = [1,1,1,1,1,1,1,1,1,1]$, $C_2() = [1,2,3,4,5,1,2,3,4,5]$ s.t $c_{avg_1} = 1$ and $c_{avg_2} = 3$.

  $\Delta = X_C = \{x_1, x_5, x_8\} : |X_C| = 3, |X_C \setminus X_F| = 0, |X_F \setminus X_C| = 1$
  $mC_{utl_{wc}}(\Delta', \Delta_Q, X, C_1()) = 1 - \frac{0}{6} - \frac{6}{6} = 0$ (via Eq. 15)
  $mC_{utl_{wc}}(\Delta', \Delta_Q, X, C_2()) = 1 - \frac{0}{2+3+4+2+4+5} - \frac{20}{20} = 0$ (via Eq. 15)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_1()) = 1 - \frac{0}{6} - \frac{1*(7-1)}{6*(1+1)} = 0.5$ (via Eq 17)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_2()) = 1 - \frac{0}{2+3+4+2+4+5} - \frac{1*(21-1)}{20*(1+3)} = 0.75$ (via Eq 17)
  $\Theta' = \{\{x_1, x_5\}, \{x_5, x_8\}, \{x_1, x_8\}\} : X'_C = \{x_1, x_5, x_8\}, |X'_C| = 3, |X_F| = 4, |X'_C \setminus X_F| = 0, |X_F \setminus X'_C| = 1, |X_A| = |X \setminus X_F| = 6$
  $mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{0}{6} - \frac{6}{6} = 0$ (via Eq. 20)
  $mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{0}{2+3+4+2+4+5} - \frac{20}{20} = 0$ (via Eq. 20)
  $mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{0}{6} - \frac{1*(7-1)}{6*(1+1)}$ (via Eq. 22)
  $mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{0}{20} - \frac{1*(21-1)}{20*(1+3)} = 0.75$ (via Eq. 22)

  Compared to the previous examples, we have to switch formulae since we have a false positive. We remind the reader that we have $X \setminus (X_F \cup X'_C) = X \setminus (X_F \cup X_C) = X_A$ for this example. Due to the false positive, the worst case metrics report a utility of 0, no matter the cost function.

- **Example 7/7':** $|X| = 10, \Delta_Q = X_F = \{x_1, x_4, x_5, x_6, x_7, x_8, x_9, x_A\}, |X_F| = 8, |X_A| = |X \setminus X_F| = 2, C_1() = [2,2,2,2,2,2,2,2,2,2]$, $C_2() = [1,2,3,4,5,1,2,3,4,5]$ s.t $c_{avg_1} = 2$ and $c_{avg_2} = 3$.

  $\Delta = \{x_1, x_2, x_3, x_A\} : |X_C| = 4, |X_C \setminus X_F| = 2, |X_F \setminus X_C| = 6$
  $mC_{utl_{wc}}(\Delta', \Delta_Q, X, C_1()) = 1 - \frac{2+2}{2+2} - \frac{0}{2+2} = 0$ (via Eq. 15)
  $mC_{utl_{wc}}(\Delta', \Delta_Q, X, C_2()) = 1 - \frac{2+3}{2+3} - \frac{0}{2+3} = 0$ (via Eq. 15)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_1()) = 1 - \frac{4}{4} - \frac{(2+2+2+2+2+2)*(12-12)}{4*(12+2)} = 0$ (via Eq 17)
  $mC_{utl_{avg}}(\Delta, \Delta_Q, X, C_2()) = 1 - \frac{5}{5} - \frac{(4+5+1+2+3+4)*((4+5+1+2+3+4)-(4+5+1+2+3+4))}{(2+3)*((4+5+1+2+3+4)+3)} = 0$
  $\Theta' = \{\{x_A, x_2\}, \{x_3\}, \{x_1, x_A\}\} : X'_C = \{x_1, x_2, x_3, x_A\}, |X'_C| = 4, |X'_C \setminus X_F| = 2, |X_F \setminus X'_C| = 6$
  $mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{4}{4} - \frac{0}{4} = 0$ (via Eq. 20)
  $mC_{utl_{wc}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{5}{5} - \frac{0}{5} = 0$ (via Eq. 20)
  $mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_1()) = 1 - \frac{4}{4} - \frac{4*0}{4*14}$ (via Eq. 22)
  $mC_{utl_{avg\_est}}(\Theta', \Delta_Q, X, C_2()) = 1 - \frac{5}{5} - \frac{19*0}{5*22} = 0$ (via Eq. 22)

  As desired, the utility is 0 for this example, no matter the cost function and whether we consider the worst or average case.

6.2 Do Strong Fault Models Change the Game?

A question that we have been continuously revisiting in our discussion is that of whether the use of strong fault models should change the way we define our metrics. In the context of cost functions, our answer depends on the cost function's complexity. In particular, we can observe that, in contrast to a WFM situation, Θ' might contain different faulty mode assignments for one and the same component. So, $\Delta_1 \in \Theta'$ could suggest fault mode m_5 for component x_2, while $\Delta_4 \in \Theta'$ suggests m_6 instead. To that end, we could derive an intuitive alternative to X'_C in the form of a set X^*_C that contains all the individual mode assignments from Θ'. However, this would make sense only if we have cost estimates that depend on the suggested and actual fault modes. Enabling the corresponding precise and structured

estimates would require us to differentiate between inspection and repair costs.[17] Defining and analyzing such a complex cost function is a task that we have to leave for future work. As a coarser estimate, we suggest to keep using cost functions as they were defined for the WFM case. Our argument is simply that our assumptions of perfect fault-understanding and repair mean that we will be able to (a) understand the fault and (b) repair it with the estimated costs. Consequently, we then need to look at each component just once and can use X'_C (and our formulae) instead of X^*_C also for the SFM case.

7 Discussion and Conclusions

Motivated by the fact that the practical assessment of diagnoses has been an under-represented research area, we delved into the question of how to measure the utility of one or more diagnoses. As we outlined in Section 3, we considered the context of a general inspection and repair process without any interleaved active testing or diagnosis steps. We established the required notation and definitions as well as developed precise metrics for assessing the worst case and average impact of a diagnosis algorithm's results on the process. In order to support complex real-world scenarios, we developed metric variants that allow us to take into account individual inspection/repair cost-functions.

We illustrated the advantages of our metrics for a set of examples and contrasted their performance with that of a metric proposed for the 2010 DX Competition [3]. As we showed, the DX Competition metric sometimes reports values that are even below our worst case evaluations. In our evaluation, we found that our metrics are much more precise. The are also intuitive and adhere to expected traits.

We discussed several directions for future work. Of specific interest are complex cost functions that split inspection and repair costs, and thus could offer advantages in the context of SFM scenarios. In order to support a general assessment, we defined metrics that generalize well enough, but complex repair processes (like batch repair) and elaborate concepts for ranking components/diagnoses could profit from precisely tailored metrics that consider the specifics of these processes. Such an investigation will need to be accompanied by experiments that investigate whether these dedicated variants show real benefits in practice. Only the isolation of such benefits could justify the additional effort needed to, e.g., define and maintain complex cost functions.

An important aspect of future experiments will be to compare the diagnostic information returned by complex analytic diagnosis approaches such as GDE [2] or RC-Tree [9] with those of computationally less complex ones, such as spectrum-based fault localization [6] and its recent extensions [12]. Data-driven solutions [15, 7] will be of specific interest in these experiments. Assessing the results with our utility metric will allow us to depict a better picture of the trade-off between computational complexity in the diagnostic process and the resulting efficiency of the inspection and repair process.

References

1 R. Alur, C. Courcoubetis, and M. Yannakakis. Distinguishing tests for nondeterministic and probabilistic machines. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '95, pages 363–372, 1995. doi:10.1145/225058.225161.

[17] As a bonus, this would allow us to effectively consider special scenarios like when we can deduce that the sole remaining component is faulty.

2 J. de Kleer and B. C. Williams. Reasoning about Multiple Faults. In *5th National Conf. on Artificial Intelligence Volume 1: Science*, pages 132–139, 1986.

3 A. Feldman, T. Kurtoglu, S. Narasimhan, S. Poll, D. Garcia, J. de Kleer, L. Kuhn, and A. van Gemund. Empirical evaluation of diagnostic algorithm performance using a generic framework. *International Journal of Prognostics and Health Management*, pages 1–28, 2010.

4 G. Friedrich, G. Gottlob, and W. Nejdl. Formalizing the repair process. In *10th European Conference on Artificial Intelligence*, pages 709–713, 1992.

5 A. Hou. A theory of measurement in diagnosis from first principles. *Artificial Intelligence*, 65(2):281–328, February 1994. doi:10.1016/0004-3702(94)90019-1.

6 J. Jones and M. Harrold. Empirical evaluation of the tarantula automatic fault-localization technique. In *20th IEEE/ACM International Conference on Automated Software Engineering, ASE 2005*, pages 273–282, November 2005. doi:10.1145/1101908.1101949.

7 A. Lundgren and D. Jung. Data-driven fault diagnosis analysis and open-set classification of time-series data. *Control Engineering Practice*, 121:105006, 2022.

8 I. Pill and J. de Kleer. Assessing Diagnosis Algorithms: Of Sampling, Baselines, Metrics and Oracles. In *36th Int. Conf. on Principles of Diagnosis and Resilient Systems (DX 2025)*, volume 136 of *Open Access Series in Informatics (OASIcs)*, pages 5:1–5:19, 2025. doi:10.4230/OASIcs.DX.2025.5.

9 I. Pill and T. Quaritsch. RC-Tree: A variant avoiding all the redundancy in Reiter's minimal hitting set algorithm. In *IEEE Int. Symp. on Software Reliability Engineering Workshops (ISSREW)*, pages 78–84, 2015.

10 R. Reiter. A Theory of Diagnosis from First Principles. *Artificial Intelligence*, 32(1):57–95, 1987. doi:10.1016/0004-3702(87)90062-2.

11 P. Rodler. Sequential model-based diagnosis by systematic search. *Artificial Intelligence*, 323:103988, 2023. doi:10.1016/j.artint.2023.103988.

12 J. Schleich and F. Wotawa. Combining Dynamic Slicing and Spectrum-Based Fault Localization - A First Experimental Evaluation. In *36th Int. Conf. on Principles of Diagnosis and Resilient Systems (DX)*, volume 136 of *Open Access Series in Informatics (OASIcs)*, pages 3:1–3:18, 2025. doi:10.4230/OASIcs.DX.2025.3.

13 R. Stern and M. Kalech. Repair planning with batch repair. In *25th Int. Workshop on Principles of Diagnosis*, 2014. URL: http://dx-2014.ist.tugraz.at/papers/DX14_Wed_AM_S1_paper4.pdf.

14 R. Stern, M. Kalech, and H. Shinitzky. Implementing troubleshooting with batch repair. *CEUR Workshop Proceedings*, 1507:113–118, 2015. 26th Int. Workshop on Principles of Diagnosis. URL: https://ceur-ws.org/Vol-1507/dx15paper15.pdf.

15 H. S. Steude, A. Diedrich, I. Pill, L. Moddemann, D. Vranješ, and O. Niggemann. Data-driven diagnosis for large cyber-physical-systems with minimal prior information. In *2026 IEEE Conference on Artificial Intelligence (CAI)*, pages 325–332, 2026.

16 S. Thiébaux, M. Cordier, O. Jehl, and J. Krivine. Supply restoration in power distribution systems: a case study in integrating model-based diagnosis and repair planning. In *12th Int. Conf. on Uncertainty in Artificial Intelligence*, pages 525–532, 1996.