How to kill them all: An exploratory study on the impact of code observability on mutation testing

Qianqian Zhu *, Andy Zaidman, Annibale Panichella

Software Engineering Research Group, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands

Abstract

Mutation testing is well-known for its efficacy in assessing test quality, and starting to be applied in the industry. However, what should a developer do when confronted with a low mutation score? Should the test suite be plainly reinforced to increase the mutation score, or should the production code be improved as well, to make the creation of better tests possible? In this paper, we aim to provide a new perspective to developers that enables them to understand and reason about the mutation score in the light of testability and observability. First, we investigate whether testability and observability metrics are correlated with the mutation score on six open-source Java projects. We observe a correlation between observability metrics and the mutation score, e.g., test directness, which measures the extent to which the production code is tested directly, seems to be an essential factor. Based on our insights from the correlation study, we propose a number of “mutation score anti-patterns”, enabling software engineers to refactor their existing code or add tests to improve the mutation score. In doing so, we observe that relatively simple refactoring operations enable an improvement or increase in the mutation score.

© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Mutation testing has been a very active research field since the 1970s as a technique to evaluate test suite quality in terms of the fault-revealing capability (Jia and Harman, 2011). Recent advances have made it possible for mutation testing to be used in industry (Petrovic et al., 2018). For example, PIT/PITest (Coles, 2019a) has been adopted by several companies, such as The Ladders and British Sky Broadcasting (Coles, 2019b). Furthermore, Google (Petrovic and Ivanovic, 2018) has integrated mutation testing with the code review process for around 6000 software engineers.

As mutation testing gains traction in the industry, a better understanding of the mutation score (one outcome of mutation testing) becomes essential. The existing works have mainly linked the mutation score with test quality (Inozemtseva and Holmes, 2014; Li et al., 2009; Yao et al., 2014; Just et al., 2017) (i.e., how good is the test suite at detecting faults in the software?) and mutant utility (Yao et al., 2014; Just et al., 2017) (i.e., how useful is the mutant?). However, in our previous study, we have observed that certain mutants could be killed only after refactoring the production code to increase the observability of state changes. In such cases, test deficiency is not the only reason for the survival of mutants. Still, some issues in the production code, such as code observability, result in difficulties to kill the mutants. Unlike previous works (e.g., Inozemtseva and Holmes, 2014; Li et al., 2009; Yao et al., 2014; Just et al., 2017), our goal is to bring a new perspective to developers that enable them to understand and reason about the mutation score in the light of testability and observability. Thereby, developers can make a choice when confronting low mutation scores: (1) adding new tests, (2) refactoring the production code to be able to write better tests, or (3) ignoring the surviving mutants.

To this aim, our study consists of two parts: firstly, we investigate the relationship between testability/observability and mutation testing in order to find the most correlated metrics; secondly, based on what we observe from the correlations, we define anti-patterns or indicators that software engineers can apply to their code to kill the surviving mutants. We start by investigating the relationship between testability/observability metrics and the mutation score inspired by the work of Bruntink and van Deursen (2006). Testability is defined as the “attributes of software that bear on the effort needed to validate the software product” (ISO, 1991; Bruntink and van Deursen, 2006). Given our context, an important part of testability is observability, which is a measure of how well internal states of a system can be inferred, usually through the values of its external outputs (Staatset al., 2011). Whalen et al. (2013) formally defined observability as follows:

An expression in a program is observable in a test case if the value of an expression is changed, leaving the rest of the program...
intact, and the output of the system is changed correspondingly. If there is no such value, the expression is not observable for that test. Compared to testability that covers various aspects of a project (e.g., inheritance and cohesion), observability specifically addresses the extent to which the value change of expression is observable in a test case.

Our first three research questions steer our investigation in the first part of our study:

**RQ1** What is the relation between testability metrics and the mutation score?

**RQ2** What is the relation between observability metrics and the mutation score?

**RQ3** What is the relation between the combination of testability and observability metrics and the mutation score?

After investigating the relationship between testability/observability and mutation testing, we still lack insight into how these relationships can help developers to take actions when facing survival mutants. That is why, based on the observations from RQ1–RQ3, we define anti-patterns or indicators that software engineers can apply to their code/tests to ensure that mutants can be killed. This leads us to the next research question:

**RQ4** To what extent does the removal of anti-patterns based on testability and observability help in improving the mutation score?

In terms of the methodology that we follow in our study, for RQ1–RQ3, we use statistical analysis on open-source Java projects to investigate the relationship between testability, observability, and the mutation score. For RQ4, we perform a case study with 16 code fragments to investigate whether the removal of anti-patterns increases the mutation score.

2. Background

In this section, we briefly introduce the basic concepts of and related works on mutation testing, testability metrics, and our proposed metrics for quantifying code observability.

2.1. Mutation testing

Mutation testing is defined by Jia and Harman (2011) as a fault-based testing technique that provides a testing criterion called the mutation adequacy score. This score can be used to measure the effectiveness of a test suite regarding its ability to detect faults (Jia and Harman, 2011). The principle of mutation testing is to introduce syntactic changes into the original program to generate faulty versions (called mutants) according to well-defined rules (mutation operators) (Motlow, 2002). The benefits of mutation testing have been extensively investigated and can be summarized (Zhu et al., 2018b) as (1) having better fault exposing capability compared to other test coverage criteria (Mathur and Wong, 1994; Frankl et al., 1997; Li et al., 2009; Inozemtseva and Holmes, 2014). Using structural coverage metrics alone might be misleading because, in many cases, statements might be covered, but their consequences might not be asserted (Inozemtseva and Holmes, 2014). Another factor is that well-developed open-source mutation testing tools (e.g., PIT/PiTest) have contributed to mutation testing being applied in the industrial environments (Petrovic et al., 2018; Petrovic and Ivanovic, 2018; Coles, 2019).

However, questions still exist about mutation testing, especially regarding the usefulness of a mutant (Just et al., 2017). The majority of the mutants generated by existing mutation operators are equivalent, trivial, and redundant (Kurtz et al., 2014; Just et al., 2017; Brown et al., 2017; Papadakis et al., 2018; Jimenez et al., 2018), which reduces the efficacy of the mutation score. If a class has a high mutation score while most mutants generated are trivial and redundant, the high mutation score does not promise high test effectiveness. A better understanding of mutation score and mutants is thus important.

To address this knowledge gap, numerous studies have investigated how useful mutants are. Example studies include mutant subsumption (Kurtz et al., 2014), stubborn mutants (Yao et al., 2014), and real-fault coupling (Just et al., 2014; Papadakis et al., 2018). These studies paid attention to the context and types of mutants as well as the impact of the test suite, while the impact of production code quality has rarely been investigated. We have seen how code quality can influence how hard it is to test (Bruntink and van Deursen, 2006) (called software testability Freedman, 1991), and since mutation testing can generally be considered as “testing the tests”, production code quality could also impact mutation testing, just like production code quality has been shown to be correlated with the presence of test smells (Spadini et al., 2018). Due to the lack of insights into how code quality affects the efforts needed for mutation testing, especially in how to engineer tests that kill all the mutants, we conduct this exploratory study. Our study can help researchers and practitioners deepen their understanding of the mutation score, which is generally related to test suite quality and mutant usefulness.

2.2. Existing object-oriented metrics for testability

The notion of software testability dates back to 1991 when Freedman (1991) formally defined observability and controllability in the software domain. Voas (1992) proposed a dynamic technique coined propagation, infection, and execution (PIE) analysis for statistically estimating the program’s fault sensitivity. More recently, researchers have aimed to increase our understanding of testability by using statistical methods to predict testability based on various code metrics. Influential works include that of Bruntink and van Deursen (2006), in which they explored the relationship between nine object-oriented metrics and testability. To explore the relation between testability and mutation score (RQ1), we first need to collect several existing object-oriented metrics that have been proposed in the literature. In total, we collect 64 code quality metrics, including both class-level and method-level metrics that have been the most widely used. We select those 64 metrics because they measure various aspects of a project, including basic characteristics (e.g., NLOC and NOMT), inheritance (e.g., DIT), coupling (e.g., CBO and FIN), and cohesion (LCOM). A large number of those metrics, such as LCOM and HLTH, have been widely used to explore software testability (Bruntink and van Deursen, 2006; Gao and Shih, 2005) and fault prediction (Aripholm and Briand, 2006; Hall et al., 2011).

We present a brief summary of the 64 metrics in Table 1 (method-level) and Tables 2–3 (class-level). We computed these metrics using a static code analysis tool provided by JHawk (JHawk, 2019).
To explore the relation between observability and mutation score (RQ2), we first need a set of metrics to quantify code observability. According to Whalen et al. (2013)’s definition of observability (as mentioned in Section 1), we consider that code observability comprises two perspectives: that of production code and that of the test case. To better explain these two perspectives, let us consider the example in Listing 1 from project jfreechart-1.5.0 showing the method `setSectionPaint` and its corresponding test. This method sets the section paint associated with the specified key for the `PiePlot` object, and sends a `PlotChangeEvent` to all registered listeners. There is one mutant in Line 3 that removes the call to `org/jfree/chart/plot/PiePlot::fireChangeEvent`. This mutant is not killed by `testEquals`. Looking at the observability of this mutant from the production code perspective, we can see that the `setSectionPaint` method is void; thus, this mutant is hard to detect because there is no return value for the test case to assert. From the test case perspective, although `testEquals` invokes the method `setSectionPaint` in Line 14 and 17, no
Table 3

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PACK</td>
<td>Number of Packages imported</td>
<td>The number of packages imported by the class</td>
</tr>
<tr>
<td>RFC</td>
<td>Response for Class</td>
<td>The value of the Response For Class metric for the class. (One of the Chidamber and Kemerer metrics)</td>
</tr>
<tr>
<td>MPC</td>
<td>Message passing</td>
<td>The value of the Message passing metric for the class</td>
</tr>
<tr>
<td>CBO</td>
<td>Coupling between objects</td>
<td>The value of the Coupling Between Objects metric for the class. (One of the Chidamber and Kemerer metrics)</td>
</tr>
<tr>
<td>FIN</td>
<td>Fan In</td>
<td>The value of the Fan In (Afferent coupling (Ca)) metric for the class</td>
</tr>
<tr>
<td>FOUT</td>
<td>Fan Out</td>
<td>The value of the Fan Out (Efferent coupling (Ce)) metric for the class</td>
</tr>
<tr>
<td>R-R</td>
<td>Reuse Ratio</td>
<td>The value of the Reuse Ratio for the class</td>
</tr>
<tr>
<td>S-R</td>
<td>Specialisation Ratio</td>
<td>The value of the Specialisation Ratio metric for the class</td>
</tr>
<tr>
<td>NSUP</td>
<td>Number of Superclasses</td>
<td>The number of superclasses (excluding Object) in the hierarchy of the class</td>
</tr>
<tr>
<td>NSUB</td>
<td>Number of Subclasses</td>
<td>The number of subclasses below the class in the hierarchy. (NOC — one of the Chidamber and Kemerer metrics)</td>
</tr>
<tr>
<td>MI</td>
<td>Maintainability Index</td>
<td>The Maintainability Index for the class, including the adjustment for comments</td>
</tr>
<tr>
<td>MNC</td>
<td>Maintainability Index</td>
<td>The Maintainability Index for the class without any adjustment for comments</td>
</tr>
<tr>
<td>COH</td>
<td>Cohesion</td>
<td>The value of the Cohesion metric for the class</td>
</tr>
<tr>
<td>DIT</td>
<td>Depth of Inheritance Tree</td>
<td>The value of the Depth of Inheritance Tree metric for the class. (One of the Chidamber and Kemerer metrics)</td>
</tr>
<tr>
<td>LCOM2</td>
<td>Lack of Cohesion of Methods (variant 2)</td>
<td>The value of the Lack of Cohesion of Methods (2) metric for the class. This uses the LCOM2 calculation. (One of the Chidamber and Kemerer metrics)</td>
</tr>
<tr>
<td>CCOM</td>
<td>Number of Comments</td>
<td>The number of comments associated with the class</td>
</tr>
<tr>
<td>CML</td>
<td>Number of Comment Lines</td>
<td>The number of comment Lines associated with the class</td>
</tr>
<tr>
<td>CLOC</td>
<td>Lines of Code</td>
<td>The number of lines of code in the class and its methods</td>
</tr>
</tbody>
</table>

### Listing 1: Example of method `org.jfree.chart.plot.PiePlot.setSectionPaint` and its test

```java
public void setSectionPaint(Comparable key, Paint paint) {
    this.sectionPaintMap.put(key, paint);
    fireChangeEvent(); // mutant: remove this method
}

@Test
public void testEquals() {
    PiePlot plot1 = new PiePlot();
    PiePlot plot2 = new PiePlot();
    assertTrue(plot1.equals(plot2));
    assertTrue(plot2.equals(plot1));
    // sectionPaintMap
    plot1.setSectionPaint("A", new GradientPaint(1.0f, 2.0f,
                                                     Color.BLUE, 3.0f, 4.0f, Color.WHITE));
    assertFalse(plot1.equals(plot2));
    plot2.setSectionPaint("A", new GradientPaint(1.0f, 2.0f,
                                                    Color.BLUE, 3.0f, 4.0f, Color.WHITE));
    assertFalse(plot1.equals(plot2));
    ...
}
```

Proper assertion statements are used to examine the changes of `fireChangeEvent()` (which is used to send an event to listeners).

Starting with two angles of code observability, we come up with a set of code observability metrics. Since our study is a starting point to design metrics to measure the code observability, we start with the simple and practical metrics, which are easy for practitioners to understand and apply.

First of all, we consider the return type of the method. As discussed in Listing 1, it is hard to observe the changing states inside a void method because there is no return value for test cases to assert. Accordingly, we design two metrics, `is_void` and `non_void_percent` (shown in 1st and 2nd rows in Table 5). The metric `is_void` is to examine whether the return value of the method is void or not. The metric `non_void_percent` addresses the return type at class level which measures the percent of non-void methods in the class. Besides these two, a void method might change the field(s) of the class it belongs to. A workaround to test a void method is to invoke getters. So `getter_percentage` (shown in 3rd row in Table 5) is proposed to complement `is_void`.

Secondly, we come up with the access control modifiers. Let us consider the example in Listing 2 from project commons-lang-LANG_3_7. The method `getMantissa` in class `NumberUtils` returns the mantissa of the given number. This method has only one mutant: the return value is replaced with "return if (getMantissa(str, str.length()) != null) null else throw new RuntimeException". This mutant should be easy to detect given an input of either a legal String object (the return value is not null) or a null string (throw an exception). This “trivial” mutant is not detected because the method `getMantissa` is private. The access control modifier `private` makes it impossible to test the method `getMantissa` directly, for this method is only visible to methods from class `NumberUtils`. To test this method, the test case must first invoke a method that calls method `getMantissa`. From this case, we observe that access control modifiers influence the visibility of the method, so as to play a significant role in code observability. Thereby, we take access control modifiers into account to quantify code observability, where we design `is_public` and `is_static` (shown in 4th and 5th rows in Table 5).

1 This mutant is generated by Return Values Mutator in PIT (Coles, 2019c). In Listing 2, `getMantissa(str, str.length())` returns a String object. When the return value of a method is an object, the mutator replaces non-null return values with null and throw a `java.lang.RuntimeException` if the un-mutated method would return null.
The third point we raise concerns fault masking. We have observed that mutants generated in certain locations are more likely to be masked (Copinath et al., 2017), i.e., the state change cannot propagate to the output of the method. The first observation is that mutants that reside in a nested class. The reasoning is similar to mutants that reside in nested sections of code, namely that a change in intermediate results does not propagate to a point where a test can pick it up. Thus, we come up with is_nested (in 6th row in Table 5). Another group of mutants is generated inside nested conditions and loops. These can be problematic because the results of the mutations cannot propagate to the output, and the tests have no way of checking the intermediate results within the method. Accordingly, we define nested_depth (shown in 7th row in Table 5) and a set of metrics to quantify the conditions and loops (shown in 8th through 13 rows in Table 5). The last observation is related to mutants that are inside a long method (the reason is similar to the mutants inside nested conditions and loops), thus, we design method_length (shown in 14th row in Table 5).

The next aspect we consider is test directness. Before we dig into test directness, we take Listing 3 as an instance. Listing 3 shows the class Triple from project commons-lang~LANG_3_7, which is an abstract implementation defining the basic functions of the object, and that consists of three elements. It refers to the elements as “left”, “middle” and “right”. The method hashCode returns the hash code of the object. Six mutants are generated for the method hashCode in class Triple. Table 4 summarises all the mutants from Listing 3. Of those six mutants, only Mutant 1 is killed, and the other mutants are not equivalent. Through further investigation of method hashCode and its test class, we found that although this method has 100% coverage by the test suite, there is no direct test for this method. A direct test would mean that the test method directly invoking the method (production code) (Athanasiou et al., 2014). The direct test is useful because it allows to control the input data directly and to assert the output of a method directly. This example shows that test directness can influence the outcome of mutation testing, which denotes the test case angle of code observability. Previous works such as Huo and Clause (2016) also addressed the significance of test directness in mutation testing. Therefore, we design two metrics, direct_test_no. and test_distance (shown in 15th and 16th row in Table 5), to quantify test directness. Those two metrics represent the test case perspective of code observability.

Last but not least, we take assertions into considerations. As discussed in Listing 1, we have observed that mutants without appropriate assertions in place (throwing exceptions is also under consideration) cannot be killed, as a prerequisite to killing a mutant is to have the tests fail in the mutated program. Schuler and Zeller (2013) and Zhang and Mesbah (2015) also drew a similar conclusion to ours. Accordingly, we come up with three metrics to quantify assertions in the method, assertion_no., assertion-McCabe_Ratio and assertion_density (shown in 17th-19th rows in Table 5). The assertion-McCabe_Ratio metric (Athanasiou et al., 2014) is originally proposed to measure test completeness by indicating the ratio between the number of the actual points of testing in the test code and the number of decision points in the production code (i.e., how many decision points are tested). For example, a method has a McCabe complexity of 4, then in the ideal case, we would expect 4 different assertions to test those linear independent paths (in this case this ration would be 1), but if the ratio is lower than 1, it could be an indication that either not all paths are tested, or that not all paths are tested in a direct way. The assertion_density metric (Kudrjavetset al., 2006) aims at measuring the ability of the test code to detect defects in the parts of the production code that it covers. We include those two metrics here as a way to measure the quality of assertions. These three metrics are proposed based on the test case perspective of code observability.

To sum up, Table 5 presents all the code observability metrics we propose, where we display the name, the definition of each metric, and the category.

### 3. Experimental setup

To examine our conjectures, we conduct an experiment using six open-source projects. We recall the research questions we have proposed in Section 1:

- **RQ1:** What is the relation between testability metrics and the mutation score?
- **RQ2:** What is the relation between observability metrics and the mutation score?
- **RQ3:** What is the relation between the combination of testability and observability metrics and the mutation score?
- **RQ4:** To what extent does removal of anti-patterns based on testability and observability help in improving the mutation score?

#### 3.1. Mutation testing

We adopt PIT (Version 1.4.0) (Coles, 2019a) to apply mutation testing in our experiments. The mutation operators we adopt are the default mutation operators provided by PIT (Coles, 2019c): Conditionals Boundary Mutator, Increments Mutator, Invert Negatives Mutator, Math Mutator, Negate Conditionals Mutator, Return Values Mutator, and Void Method Calls Mutator. We did not adopt the extended set of mutation operators provided PIT, as the operators in the default version are largely designed to be stable (i.e., not too easy to detect) and minimise the number of equivalent mutations that they generate (Coles, 2019c).

#### 3.2. Subject systems

We use six systems publicly available on GitHub in this experiment. Table 6 summarises the main characteristics of the selected projects, which include the lines of code (LOC), the number of tests (#Test), the total number of methods (#Total Methods), the number of selected methods used in our experiment (#Selected), the total number of mutants (#Total Mutants), and the killed mutants (#Killed). In our experiment, we remove the methods with no generated mutant by PIT, thus resulting in the number of
Table 4
Summary of mutants from Listing 3.

<table>
<thead>
<tr>
<th>ID</th>
<th>Line no.</th>
<th>Mutator</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>Negated conditional</td>
<td>Killed</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>Replaced return of integer sized value with (x == 0 ? 1 : 0)</td>
<td>Survived</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Replaced XOR with AND</td>
<td>Survived</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Negated conditional</td>
<td>Survived</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>Replaced XOR with AND</td>
<td>Survived</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>Negated conditional</td>
<td>Survived</td>
</tr>
</tbody>
</table>

Table 5
Summary of code observability metrics.

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Definition</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>is_void</td>
<td>Whether the return value of the method is void or not</td>
<td>Return type</td>
</tr>
<tr>
<td>2</td>
<td>non_void_percent (class-level)</td>
<td>The percent of non-void methods in the class</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>getter_percentage</td>
<td>The percentage of getter methods in the class^</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>is_public</td>
<td>Whether the method is public or not</td>
<td>Access control modifiers</td>
</tr>
<tr>
<td>5</td>
<td>is_static</td>
<td>Whether the method is static or not</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>is_nested (class-level)</td>
<td>Whether the method is located in a nested class or not</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>nested_depth</td>
<td>The maximum number of nested depth (MDN from Section 2.2)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>(cond)</td>
<td>The number of conditions (if, if-else and switch) in the method</td>
<td>Fault masking</td>
</tr>
<tr>
<td>9</td>
<td>(cond(cond))</td>
<td>The number of nested conditions (e.g., if(if())) in the method</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>(cond/loop)</td>
<td>The number of nested condition-loops (e.g.,if(for())) in the method</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>(loop)</td>
<td>The number of loops (for, while and do-while) in the method (LOOP from Section 2.2)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>(loop/cond)</td>
<td>The number of nested loop-conditions (e.g.,for(if())) in the method</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>(loop/loop)</td>
<td>The number of nested loop-conditions (e.g.,for(for())) in the method</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>method_length</td>
<td>The number of lines of code in the method (NLOC from Section 2.2)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>direct_test_no.</td>
<td>The number of test methods directly invoking the method under test (production code)^5</td>
<td>Test directness</td>
</tr>
<tr>
<td>16</td>
<td>test_distance</td>
<td>The shortest method call sequence required to invoke the method (production code) by test methods^6</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>assertion_no.</td>
<td>The number of assertions in direct tests</td>
<td>Assertion</td>
</tr>
<tr>
<td>18</td>
<td>assertion-McCabe_Ratio</td>
<td>The ratio between the total number of assertions in direct tests and the McCabe Cyclomatic complexity</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>assertion_density</td>
<td>The ratio between the total number of assertions in direct tests and the lines of code in direct tests</td>
<td></td>
</tr>
</tbody>
</table>

^A getter method must follow three patterns (Zhang and Mesbah, 2015): (1) must be public; (2) has no arguments and its return type must be something other than void; (3) have naming conventions: the name of a getter method begins with “get” followed by an uppercase letter.

^If the method is not directly tested, then its direct_test_no. is 0.

^If the method is directly tested, then its test_distance is 0. The maximum test_distance is set Integer.MAX_VALUE in Java which means there is no method call sequence that can reach the method from test methods.

Table 6
Subject systems.

<table>
<thead>
<tr>
<th>PID</th>
<th>Project</th>
<th>LOC</th>
<th>#Tests</th>
<th>#Methods</th>
<th>#Mutants</th>
<th>#Killed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>#Total</td>
<td>#Selected</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Bukkit-1.7.9-R0.2</td>
<td>32373</td>
<td>432</td>
<td>7 325</td>
<td>2 385</td>
<td>7 325</td>
</tr>
<tr>
<td>2</td>
<td>commons-lang-LANG_3_7</td>
<td>77 224</td>
<td>4 068</td>
<td>13 052</td>
<td>2 740</td>
<td>13 052</td>
</tr>
<tr>
<td>3</td>
<td>commons-math-MATH_3_6_1</td>
<td>208 959</td>
<td>6 523</td>
<td>48 524</td>
<td>6 663</td>
<td>48 524</td>
</tr>
<tr>
<td>4</td>
<td>java-apns-apns-0.2.3</td>
<td>3 418</td>
<td>91</td>
<td>429</td>
<td>150</td>
<td>429</td>
</tr>
<tr>
<td>5</td>
<td>jfreechart-1.5.0</td>
<td>134 117</td>
<td>2 175</td>
<td>34 488</td>
<td>7 133</td>
<td>34 488</td>
</tr>
<tr>
<td>6</td>
<td>pysonar2-2.1</td>
<td>10 926</td>
<td>269</td>
<td>3 070</td>
<td>719</td>
<td>3 074</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>467 017</td>
<td>13 558</td>
<td>106 888</td>
<td>19 790</td>
<td>106 892</td>
</tr>
</tbody>
</table>

The results of the mutants that are killable for all of the subjects are shown in Columns 7–8 of Table 6. Fig. 1a shows the distribution of mutation scores among selected methods. The majority of the mutation scores are either 0 or 1. Together with Fig. 1b, we can see that the massive number of 0s and 1s are due to the low number of mutants per method. Most methods show less than 10 mutants, which is mainly due to most methods being short methods (NOS < 2 as shown in Fig. 2). Writing short methods is a preferred strategy in practice, for a long method is a well-known code smell (Beck et al., 1999). Besides, PIT adopts several optimisation mechanisms (Coles, 2019d) to reduce the number of mutants. Thus, the number of mutants (#Total Mutants) shown in Table 6 is fewer than the actual number of generated mutants. The large number of methods with low mutant number is an unavoidable bias in our experiment.

3.3. Tool implementation

To evaluate the code observability metrics that we have proposed, we implemented a prototype tool (coined Mutation Observer) to capture all the necessary information from both the program under test and the mutation testing process. This tool is openly available on GitHub (Zhu, 2019).
Our tool extracts information from three parts of the system under test (in Java): source code, bytecode, and tests. Firstly, Antlr (2019) parses the source code to obtain the basic code features, e.g., is public, is static, and (cond). Secondly, we adopt Apache Commons BCEL (Apache, 2019) to parse the bytecode. Then, java-callgraph (java-callgraph, 2019) generates the pairs of method calls between the source code and tests, which we later use to calculate direct test no. and other test call-related metrics. The last part is related to the mutation testing process, for which we adopt PIT (Version 1.4.0) (Coles, 2019a) to obtain the killable mutant results. An overview of the architecture of Mutation Observer can be seen in Fig. 3.

3.4. Design of experiment

3.4.1. RQ1–RQ3

Our investigation of the relationships between testability/observability metrics and the mutation score (RQ1–RQ3) is two-fold: in the first part, we adopt Spearman’s rank-order correlation to measure the pairwise correlations statistically between each metric (both testability and observability metrics) and the mutation score; in the second part, we turn the correlation problem into a binary classification problem (where we adopt Random Forest as the classification algorithm) to investigate how those metrics interact with one another.

**Pairwise correlations.** To answer RQ1, RQ2, and RQ3, we first adopt Spearman’s rank-order correlation to statistically measure the correlation between each metric (both testability and observability metrics) and the mutation score of the corresponding methods or classes. Spearman’s correlation test checks whether there exists a monotonic relationship (linear or not) between two data samples. The resulting coefficient \( \rho \) takes values in the interval \([-1; +1]\); the higher the correlation in either direction (positive or negative), the stronger the monotonic relationship between the two data samples under analysis. The strength of the correlation can be established by classifying into “negligible” (\(|\rho| < 0.1\)), “small” (\(0.1 \leq |\rho| < 0.3\)), “medium” (\(0.3 \leq |\rho| < 0.5\)), and “large” (\(|\rho| \geq 0.5\)) (Hinkle et al., 1988). Positive \( \rho \) values indicate that one distribution increases when the other increases as well; negative \( \rho \) values indicate that one distribution decreases when the other increases. To measure the statistical significance of Spearman’s correlation test, we look at \( p \)-values that measure the probability of an observed (or more extreme) result assuming that the null hypothesis is true. Any test size larger than the \( p \)-value leads to rejection, whereas using a test size smaller than the \( p \)-value fails to reject the null hypothesis (Hung et al., 1997). Here we consider the test size of 5% as the cutoff for statistical significance.

The mutation score\(^2\) is calculated by Eq. (1) (method-level).

\[
\text{mutation score}(A) = \frac{\# \text{ killed mutants in method } A}{\# \text{ total mutants in method } A} 
\]  

We adopt Matlab (MATLAB, 2019) to calculate the Spearman’s rank-order correlation coefficient between each metric and the mutation score. In particular, we used the statistical analysis

\(^2\) In the original equation for mutation score, the divisor is the number of non-equivalent. In our study, our main focus is the relation between testability/observability metrics and mutation score, rather than mutation score itself. In our previous literature review (Zhu et al., 2018b), we have found that treating all mutants as non-equivalent is a common method when the mutation score is used as a relative comparison. Therefore, we do not manually analyse the equivalent mutants, and treated all mutants as non-equivalent.
(corr function with the option of “Spearman” in Matlab’s default package).  

**Interactions.** Except for the pairwise correlations between metrics and mutation score, we are also interested in how those metrics interact with one another. First, we try regression models to predict mutation scores based on the metrics. However, all the regression models incur extremely high cross-validation errors, i.e., Root Relative Squared Errors (RRSEs) are > 70% (e.g., RRSE of linear regression is 76.62%). Therefore, we turn the correlation problem into a classification problem for better performance. For simplicity, we use 0.5 as the cutoff between HIGH and LOW mutation core because 0.5 is widely used as a cutoff in classification problems whose independent variable ranges in [0,1] (e.g., defect prediction (Zhang et al., 2016; Tosun and Bener, 2009)). We consider all the metrics to predicate whether the method belongs to classes with HIGH or LOW mutation score. One thing to notice here is that building a perfect prediction model is not our primary goal. Our interest is to see which metrics and/or combinations of the metrics contribute to the LOW mutation score by building the prediction models. Therefore, deciding different threshold values is outside the scope of this paper.

For prediction, we adopt Random Forest (Breiman, 2001) as the classification algorithm, where we use WEKA (Frank et al., 2016) to build the prediction model. Random Forest is an ensemble method based on a collection of decision tree classifiers, where the individual decision trees are generated using a random selection of attributes at each node to determine the split (Han et al., 2011). Besides, Random Forest is more accurate than one decision tree, and it is not affected by the overfitting problem (Han et al., 2011).

As our investigation includes testability and observability metrics, for each project, we compare three types of classification models: (1) a model based on merely existing testability metrics. (2) a model based on merely code observability metrics, and (3) a model based on the combination of existing and our observability metrics (overlapping metrics, e.g., method_length to NLOC, are only considered once). In particular, we include the model based on the combination of the two aspects for further comparison: to see whether the combination of the two aspects can work better than each aspect itself. To examine the effectiveness of Random Forest in our dataset, we also consider ZeroR, which classifies all the instances to the majority and ignores all predictors, as the baseline. It might be that our data is not balanced, as in that one project has over 90% methods with a HIGH mutation score. This could entail that the classification model achieving 90% accuracy is not necessarily an effective model. In this situation, ZeroR could also achieve over 90% accuracy in that scenario. Our Random Forest model must thus perform better than ZeroR; otherwise, the Random Forest model is not suitable for our dataset.

In total, we consider four classification models: (1) ZeroR (i.e., the constant classifier), (2) Random Forest based on existing metrics, (3) Random Forest based on code observability metrics, and (4) Random Forest based on the combination of existing metrics and code observability metrics. To build Random Forest, WEKA (Frank et al., 2016) adopts bagging in tandem with random attribute selection. We use WEKA’s default parameters to train the Random Forest model, i.e., “-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1”. To evaluate the performance of the classifier model (e.g., precision and recall), we use K-fold cross-validation with K = 10 (Kohavi et al., 1995).

In terms of feature importance, we apply scikit-learn (Pedregosa et al., 2011) to conduct the analysis. To determine the feature importance, scikit-learn (Pedregosa et al., 2011) implements “Gini Importance” or “Mean Decrease Impurity” (Breiman, 2001). The importance of each feature is computed by the probability of reaching that node (which is approximated by the proportion of samples reaching that node) averaged over total tree ensembles (Breiman, 2017). We use the method of feature_importances_ in sklearn.ensemble.RandomForestRegressor (scikit-learn, 2019) package to analyse the feature importance.

![Fig. 3. Overview of Mutation Observer architecture.](image-url)

---


---

3.4.2. **RQ4**

To answer RQ4, we first need to establish the anti-patterns (or smells) based on these metrics. An example of an anti-pattern rule generated from the metrics is `method_length > 20 and test_distance > 2`. In this case, it is highly likely that the method has a low mutation score. To obtain the anti-pattern rules, we adopt J48 to build a decision tree (Quinlan, 1993; Frank et al., 2016). We consider J48 because of its advantage in interpretation over Random Forest. After building the decision tree, we rank all leaves (or paths) according to instances falling into each leaf and accuracy. We select the leaves with the highest instances and accuracy ≥ 0.8 for further manual analysis, to understand to what extent refactoring of the anti-patterns can help improve the mutation score.
3.5. Evaluation metrics

For RQ1, RQ2, and RQ3, to ease the comparisons of the four classification models, we consider four metrics widely used in classification problems: precision, recall, AUC, and the mean absolute error.

In our case, we cannot decide which class is positive or not, or in other words, we cannot say HIGH mutation score is what we expect. We use a prediction model to investigate the interactions between those metrics or how they interact with each other. So we adopt weighted precision and recall, which also take the number of instances in each class into consideration.

**Weighted precision.** The precision is the fraction of true positive instances in the instances that are predicted to be positive: TP/(TP+FP). The higher the precision, the fewer false positives. The weighted precision is computed as follows, where \( p_{c_1} \) and \( p_{c_2} \) are the precisions for class 1 and class 2, and \( |c_1| \) and \( |c_2| \) are the number of instances in class 1 and class 2, respectively:

\[
\text{weighted precision} = \frac{p_{c_1} \times |c_1| + p_{c_2} \times |c_2|}{|c_1| + |c_2|}
\]

(2)

**Weighted recall.** The recall is the fraction of true positive instances in the instances that are actual positives: TP/(TP+FN). The higher the recall, the fewer false-negative errors there are. The weighted recall is computed as follows, where \( r_{c_1} \) and \( r_{c_2} \) are the recalls for class 1 and class 2, and \( |c_1| \) and \( |c_2| \) are the number of instances in class 1 and class 2:

\[
\text{weighted recall} = \frac{r_{c_1} \times |c_1| + r_{c_2} \times |c_2|}{|c_1| + |c_2|}
\]

(3)

**AUC.** The area under ROC curve, which measures the overall discrimination ability of a classifier. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test.

**Mean absolute error.** The mean of overall differences between the predicted values and actual values.

4. RQ1–RQ3 testability versus observability versus combination

We opt to discuss the three research questions, RQ1, RQ2, and RQ3, together, because it gives us the opportunity to compare testability, observability, and their combination in detail.

4.1. Spearman’s rank order correlation

4.1.1. Testability

**Findings.** Table 7 presents the overall results of Spearman’s rank-order correlation analysis for existing code metrics. The columns of “rho” represent the pairwise coefficient between each code metric and the mutation score. The \( p\)-values columns denote the strength of evidence for testing the hypothesis of no correlation against the alternative hypothesis of a non-zero correlation using Spearman’s rank-order. Here we used 0.05 as the cutoff for significance. From Table 7, we can see that except for NSD, NLOC, MOD, EXOR, INST(class), NSUB(class), CDH(class) and S-R(class) (which, for convenience, we highlighted by underlining the value), the correlation results for the metrics are all statistically significant.

Overall, the pairwise correlation between each source code metric and the mutation score is not strong (\(|\text{rho}| < 0.27\)). We speculate the reason behind the weak correlations to be the collinearity of these code metrics. More specifically, Spearman’s rank-order correlation analysis only evaluates the correlation between individual code metric and mutation score. Some code metrics could interact with one another. For example, a long method does not necessarily have a low mutation score. Alternatively, another example: if there are more than four loops in a long method, then the method is very likely to have a low mutation score. That is also an example of collinearity, i.e., the number of loops and the method length are highly correlated.

From Table 7, we can see that the highest rho\(^4\) is ~0.2634 for both NSUP(class) standing for Number of Superclasses, and DIT(class), or Depth of Inheritance Tree. Followed by R-R(class), for Reuse Ratio, and HIER(class), for Hierarchy method calls. At first glance, the top 4 metrics are all class-level metrics. However, we cannot infer that class-level metrics are more impactful on the mutation score than method-level ones. In particular, it can be related to the fact that we have considered more class-level metrics than method-level ones in the experiment.

Additionally, we expected that the metrics related to McCabe’s Cyclomatic Complexity, i.e., COMP, TCC, AVCC and MAXCC would show stronger correlation to the mutation score. In fact, McCabe’s Cyclomatic Complexity has been widely considered a powerful measure to quantify the complexity of a software program, and it is used to provide a lower bound to the number of tests that should be written (Woodward et al., 1979; Gill and Kemerer, 1991; Fenton and Ohlsson, 2000). Based on our results without further investigation, we could only speculate that McCabe’s Cyclomatic Complexity might not directly influence the mutation score.

**Summary.** We found that the pair-wise correlations between the 64 existing source code metrics and the mutation score to be not so strong (\(|\text{rho}| < 0.27\)). The top 4 metrics with the strongest correlation coefficients are NSUP(class), DIT(class), R-R(class) and HIER(class).

4.1.2. Observability

**Findings.** Table 8 shows the overall results of Spearman’s rank-order correlation analysis for code observability metrics. From Table 8, we can see that except for method_length and (cond(loop)), whose \( p\)-value is greater than 0.05, the results of the other observability metrics are statistically significant. The overall correlation between code observability metrics and mutation score is still not strong (<0.5), but significantly better than existing code metrics (<0.27). The top five metrics are test_distance, direct_test_no., assertion-density, assertion-McCabe, and assertion_no. The metrics related to test directness, i.e., test_distance (~0.4923) and direct_test_no.(0.4177) are ranked first in terms of rho among all metrics that we consider (including existing code metrics in Section 2.2). This observation corresponds to our hypothesis in Section 2.3 that the methods with no direct tests are more challenging to kill mutants. In terms of rho values, the assertion related metrics are ranked after test directness related metrics; this confirms both our conjectures in Section 2.3 and what has been reported in the related literature (Schuler and Zeller, 2013; Zhang and Mesbah, 2015) that the quality of assertions can influence the outcome of mutation testing.

**Summary.** The correlations between code observability metrics and mutation score are not very strong (<0.5); however, they are significantly better than the correlations for existing code metrics. Test directness (test_distance and direct_test_no.) takes the first place of NSUP(class) in |rho| among all metrics (including existing ones in Section 2.2), followed by assertion-based metrics (assertion-density, assertion-McCabe and assertion_no).

\(^4\) In terms of absolute value.
### Table 7

Spearman results of existing code metrics for testability.

<table>
<thead>
<tr>
<th>Metric</th>
<th>rho</th>
<th>p-value</th>
<th>Metric</th>
<th>rho</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMP</td>
<td>0.0398</td>
<td>2.16E–08</td>
<td>NOC</td>
<td>0.1908</td>
<td>1.254E–161</td>
</tr>
<tr>
<td>NOCL</td>
<td>0.1047</td>
<td>3.27E–49</td>
<td>NOA</td>
<td>0.0423</td>
<td>2.723E–09</td>
</tr>
<tr>
<td>NOS</td>
<td>–0.0139</td>
<td>0.05024</td>
<td>CAST</td>
<td>–0.0182</td>
<td>0.01302</td>
</tr>
<tr>
<td>HLTH</td>
<td>0.0518</td>
<td>2.927E–13</td>
<td>HDIF</td>
<td>0.1334</td>
<td>2.691E–79</td>
</tr>
<tr>
<td>HVOC</td>
<td>0.0485</td>
<td>8.831E–12</td>
<td>NEXP</td>
<td>0.0288</td>
<td>5.135E–05</td>
</tr>
<tr>
<td>HEFF</td>
<td>0.0856</td>
<td>1.595E–33</td>
<td>LOMC(class)</td>
<td>0.0564</td>
<td>2.125E–15</td>
</tr>
<tr>
<td>HBBG</td>
<td>0.0518</td>
<td>3.163E–13</td>
<td>LOCAB(class)</td>
<td>0.0405</td>
<td>1.206E–08</td>
</tr>
<tr>
<td>CREF</td>
<td>0.0193</td>
<td>0.00653</td>
<td>AVCC(class)</td>
<td>0.0793</td>
<td>5.416E–29</td>
</tr>
<tr>
<td>XMET</td>
<td>0.0465</td>
<td>5.743E–11</td>
<td>NOS(class)</td>
<td>0.0824</td>
<td>3.826E–31</td>
</tr>
<tr>
<td>LMET</td>
<td>–0.0221</td>
<td>0.00191</td>
<td>HBUC(class)</td>
<td>0.0981</td>
<td>1.564E–43</td>
</tr>
<tr>
<td>NLOC</td>
<td>–0.0004</td>
<td>0.95</td>
<td>HEFF(class)</td>
<td>0.0982</td>
<td>1.213E–43</td>
</tr>
<tr>
<td>VDEC</td>
<td>0.0281</td>
<td>7.702E–05</td>
<td>UNVC(class)</td>
<td>0.0929</td>
<td>3.708E–39</td>
</tr>
<tr>
<td>TDDN</td>
<td>0.0408</td>
<td>6.34E–09</td>
<td>INST(class)</td>
<td>0.0045</td>
<td>0.523E</td>
</tr>
<tr>
<td>NAND</td>
<td>0.0357</td>
<td>5.191E–07</td>
<td>PACK(class)</td>
<td>–0.1029</td>
<td>0.956E–48</td>
</tr>
<tr>
<td>LOOP</td>
<td>0.0685</td>
<td>5.116E–22</td>
<td>RFC(class)</td>
<td>0.095</td>
<td>6.38E–41</td>
</tr>
<tr>
<td>MOD</td>
<td>0.0103</td>
<td>0.1482</td>
<td>CBO(class)</td>
<td>–0.0157</td>
<td>0.0274</td>
</tr>
<tr>
<td>NOPR</td>
<td>0.0067</td>
<td>3.807E–21</td>
<td>M(class)</td>
<td>0.0482</td>
<td>1.144E–11</td>
</tr>
<tr>
<td>EXCT</td>
<td>0.1125</td>
<td>8.723E–57</td>
<td>CCML(class)</td>
<td>0.1559</td>
<td>6.998E–108</td>
</tr>
<tr>
<td>MDN</td>
<td>0.053</td>
<td>8.3E–14</td>
<td>NLOCClass)</td>
<td>0.0756</td>
<td>1.692E–26</td>
</tr>
<tr>
<td>EXCR</td>
<td>–0.0067</td>
<td>0.3473</td>
<td>RVF(class)</td>
<td>–0.033</td>
<td>3.498E–06</td>
</tr>
<tr>
<td>HVOL</td>
<td>0.0512</td>
<td>5.719E–13</td>
<td>LOM2(class)</td>
<td>–0.0486</td>
<td>7.691E–12</td>
</tr>
<tr>
<td>VREF</td>
<td>0.0446</td>
<td>3.4E–10</td>
<td>MAXCC(class)</td>
<td>–0.0178</td>
<td>0.01245</td>
</tr>
</tbody>
</table>

### Table 8

Spearman results of code observability metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>rho</th>
<th>p-value</th>
<th>Metric</th>
<th>rho</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>is_public</td>
<td>–0.0639</td>
<td>2.35E–19</td>
<td>(cond(cond))</td>
<td>–0.0415</td>
<td>5.4E–09</td>
</tr>
<tr>
<td>is_static</td>
<td>0.1137</td>
<td>6.29E–58</td>
<td>(cond(loop)</td>
<td>0.0073</td>
<td>0.302</td>
</tr>
<tr>
<td>is_void</td>
<td>–0.1427</td>
<td>1.42E–90</td>
<td>(loop)</td>
<td>0.0685</td>
<td>5.12E–22</td>
</tr>
<tr>
<td>is_nested</td>
<td>0.0466</td>
<td>5.38E–11</td>
<td>(loop)</td>
<td>0.0216</td>
<td>0.00024</td>
</tr>
<tr>
<td>method_length</td>
<td>–0.0004</td>
<td>0.95</td>
<td>(loop)</td>
<td>0.0428</td>
<td>1.65E–09</td>
</tr>
<tr>
<td>nested_depth</td>
<td>0.053</td>
<td>8.3E–14</td>
<td>non Void_percent</td>
<td>0.2424</td>
<td>1.24E–26</td>
</tr>
<tr>
<td>direct_test_no</td>
<td>0.4177</td>
<td>0</td>
<td>getter Percent</td>
<td>–0.153</td>
<td>6.23E–104</td>
</tr>
<tr>
<td>test_distance</td>
<td>–0.4921</td>
<td>0</td>
<td>assertion-McCabe</td>
<td>0.0051</td>
<td>0</td>
</tr>
<tr>
<td>(cond)</td>
<td>0.023</td>
<td>0.000124</td>
<td>assertion-density</td>
<td>0.4096</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 9

Random forest results of code observability metrics vs. Existing metrics.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–</td>
<td>0.856</td>
<td>0.497</td>
<td>0.2465</td>
<td>0.927</td>
<td>0.93</td>
<td>0.961</td>
<td>0.1014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>–</td>
<td>0.913</td>
<td>0.498</td>
<td>0.1595</td>
<td>0.947</td>
<td>0.951</td>
<td>0.932</td>
<td>0.0775</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>–</td>
<td>0.815</td>
<td>0.499</td>
<td>0.3015</td>
<td>0.848</td>
<td>0.861</td>
<td>0.836</td>
<td>0.2039</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>–</td>
<td>0.507</td>
<td>0.468</td>
<td>0.5801</td>
<td>0.667</td>
<td>0.667</td>
<td>0.733</td>
<td>0.3821</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>–</td>
<td>0.62</td>
<td>0.5</td>
<td>0.4712</td>
<td>0.842</td>
<td>0.843</td>
<td>0.908</td>
<td>0.2347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>0.726</td>
<td>0.493</td>
<td>0.3982</td>
<td>0.73</td>
<td>0.743</td>
<td>0.804</td>
<td>0.2948</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>–</td>
<td>0.569</td>
<td>0.5</td>
<td>0.4905</td>
<td>0.862</td>
<td>0.862</td>
<td>0.928</td>
<td>0.2133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dir.</td>
<td>–</td>
<td>0.853</td>
<td>0.499</td>
<td>0.2513</td>
<td>0.945</td>
<td>0.946</td>
<td>0.949</td>
<td>0.0915</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non.</td>
<td>–</td>
<td>0.593</td>
<td>0.5</td>
<td>0.4829</td>
<td>0.853</td>
<td>0.853</td>
<td>0.923</td>
<td>0.2329</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.2. Random forest

**Classification effectiveness.** As discussed in Section 3.4, we compare the four models in terms of both our code observability metrics and the existing metrics, namely:

1. **ZeroR:** model using ZeroR approach
2. **existing:** Random Forest model based on existing code metrics
3. **code observability:** Random Forest model based on code observability metrics
4. **combined:** Random Forest model based on the combination of existing metrics and code observability metrics

Table 9 shows the results of the comparison of the four models. To make clear which model performs better than the others, we highlighted the values of the model achieving the best performance among the four in \textbf{bold}, that of second best in underline. For precision, recall, and AUC, the model with the best performance is the one with the highest value, whereas for the mean absolute error, the best scoring model exhibits the lowest value. For the ZeroR model, because this model classifies all the instances to the majority (i.e., one class), the precision of the minority is not valid due to 0/0. Thus, in Table 9, we mark the precisions by "\--".

From Table 9, we can see that the Random Forest models are better than the baseline ZeroR which only relies on the majority. This is the \textbf{prerequisite} for further comparison. Combined achieves the best performance (in 5 out of 6 projects) compared to the existing code metrics and code observability metrics in terms of AUC; this observation is as expected since combined considered both the existing and our metrics during training, which provides the classification model with more information.
The only exception is `java-apns-apns-0.2.3` (pid = 4). We conjecture that the number of instances (selected methods) in this project might be too small (only 150 methods) to develop a sound prediction model. In second place comes the model based on code observability metrics, edging out the model based on existing metrics.

For the overall dataset (the 7th row marked with “all” in Table 9), combined takes the first place in all evaluation metrics. In second place comes the code observability, slightly better than existing. Another interesting angle investigate further is the test directness. If we only consider the methods that are directly tested (the second to last row in Table 9), it is easy to understand that when the dataset only considers methods that are directly tested (or not), the test directness features in our model become irrelevant. However, we can see that the difference between existing metrics and ours are quite tiny (<3.4%).

**Feature importance analysis.** Tables 10 and 11 show the top 15 features per project (and overall) in descending order. We can see that for five out of the six projects (including the overall dataset), test_distance ranks first. This again supports our previous findings that test directness plays a significant role in mutation testing. The remaining features in the top 15 vary per project; this is not surprising, as the task and context of these projects vary greatly. For example, Apache Commons Lang (Column “2” in Table 10) is a utility library that provides a host of helper methods for the `java.lang` API. Therefore, most methods in Apache Commons Lang are public and static; thus, `is_public` and `is_static` are not among the top 15 features for Apache Commons Lang. A totally different context is provided by the JFreeChart project (Column “5” in Table 10). JFreeChart is a Java chart library, whose class encapsulation and inheritance hierarchy are well-designed, so `is_public` appears among the top 15 features.

Looking at the overall dataset (Column “all” in Table 11), there are eight metrics from our proposed code observability metrics among the top 15 features. The importance of test_distance is much higher than the other features (>4.83X). In second place comes `PACK(class)`, or the number of packages imported. This observation is easy to understand since `PACK(class)` denotes the complexity of dependency, and dependency could influence the difficulty of testing, especially when making use of mocking objects. Thereby, dependency affects the mutation score. Clearly, more investigations are required to draw further conclusions. The third place in the feature importance analysis is taken by NOCL, which stands for the Number of Comments. This observation is quite interesting since NOCL is related to how hard it is to understand the code (code readability). This implies that code readability might have an impact on mutation testing.

As for the methods with direct tests (Column “dir.” in Table 11), `is_void` takes the first position, which indicates that it is more difficult to achieve a high mutation score for void methods. Considering the methods without direct tests (Column “non-dir.” in Table 11), test_distance again ranks first.

Another observation stems from the comparison of the performance of assertion related metrics in the feature importance analysis and the Spearman rank order correlation results (in Section 4.1). For Spearman’s rank order correlation, we can see that

---

**Table 10**

Feature importance of classification model (1).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>test_distance</td>
<td>0.35</td>
<td>test_distance</td>
<td>0.15</td>
<td>test_distance</td>
<td>0.13</td>
<td>test_distance</td>
<td>0.48</td>
</tr>
<tr>
<td>NLOC(class)</td>
<td>0.15</td>
<td>HIER(class)</td>
<td>0.12</td>
<td>NOCL</td>
<td>0.05</td>
<td>method_distance</td>
<td>0.03</td>
</tr>
<tr>
<td>NOCL</td>
<td>0.05</td>
<td>NOCL</td>
<td>0.03</td>
<td>COMP</td>
<td>0.03</td>
<td>EXCT</td>
<td>0.04</td>
</tr>
<tr>
<td>CREF</td>
<td>0.03</td>
<td>NLOC</td>
<td>0.05</td>
<td>MI(class)</td>
<td>0.03</td>
<td>NOCL</td>
<td>0.03</td>
</tr>
<tr>
<td>MINC(class)</td>
<td>0.03</td>
<td>NOCL</td>
<td>0.04</td>
<td>is_static</td>
<td>0.02</td>
<td>CAST</td>
<td>0.03</td>
</tr>
<tr>
<td>non_void_percent</td>
<td>0.02</td>
<td>MI(class)</td>
<td>0.04</td>
<td>non_void_percent</td>
<td>0.02</td>
<td>HDF</td>
<td>0.03</td>
</tr>
<tr>
<td>HDIF</td>
<td>0.02</td>
<td>assertion-density</td>
<td>0.03</td>
<td>HVOC</td>
<td>0.02</td>
<td>(Cond)</td>
<td>0.02</td>
</tr>
<tr>
<td>NOS(class)</td>
<td>0.02</td>
<td>CREF</td>
<td>0.03</td>
<td>HEFF</td>
<td>0.02</td>
<td>VREF</td>
<td>0.02</td>
</tr>
<tr>
<td>PACK(class)</td>
<td>0.02</td>
<td>HDF</td>
<td>0.03</td>
<td>CREF</td>
<td>0.02</td>
<td>is_void</td>
<td>0.01</td>
</tr>
<tr>
<td>TCC(class)</td>
<td>0.02</td>
<td>PACK(class)</td>
<td>0.03</td>
<td>VREF</td>
<td>0.02</td>
<td>direct_test_no</td>
<td>0.01</td>
</tr>
<tr>
<td>LMC(class)</td>
<td>0.02</td>
<td>method_length</td>
<td>0.02</td>
<td>NEXP</td>
<td>0.02</td>
<td>assertion_no</td>
<td>0.01</td>
</tr>
<tr>
<td>HLTH</td>
<td>0.01</td>
<td>HVOC</td>
<td>0.02</td>
<td>HEFF(class)</td>
<td>0.02</td>
<td>non_void_percent</td>
<td>0.01</td>
</tr>
<tr>
<td>HVOC</td>
<td>0.01</td>
<td>HEFF</td>
<td>0.02</td>
<td>PACK(class)</td>
<td>0.02</td>
<td>assertion-density</td>
<td>0.01</td>
</tr>
<tr>
<td>HEFF</td>
<td>0.01</td>
<td>LMET</td>
<td>0.02</td>
<td>CBD(class)</td>
<td>0.02</td>
<td>HLTH</td>
<td>0.01</td>
</tr>
<tr>
<td>XMET</td>
<td>0.01</td>
<td>NOA</td>
<td>0.02</td>
<td>CCM(class)</td>
<td>0.02</td>
<td>HVOC</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Table 11**

Feature importances of classification model (2).

<table>
<thead>
<tr>
<th>Metric</th>
<th>Metric Imp.</th>
<th>Metric</th>
<th>Metric Imp.</th>
<th>Metric</th>
<th>Metric Imp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBO(class)</td>
<td>0.09</td>
<td>test_distance</td>
<td>0.29</td>
<td>is_void</td>
<td>0.22</td>
</tr>
<tr>
<td>HDIF</td>
<td>0.07</td>
<td>PACK(class)</td>
<td>0.06</td>
<td>PACK(class)</td>
<td>0.13</td>
</tr>
<tr>
<td>NQI(class)</td>
<td>0.06</td>
<td>NOCL</td>
<td>0.05</td>
<td>HDIF</td>
<td>0.05</td>
</tr>
<tr>
<td>test_distance</td>
<td>0.04</td>
<td>is_void</td>
<td>0.03</td>
<td>NOS</td>
<td>0.04</td>
</tr>
<tr>
<td>non_void_percent</td>
<td>0.03</td>
<td>EXCT</td>
<td>0.03</td>
<td>assertion-density</td>
<td>0.03</td>
</tr>
<tr>
<td>HVOC</td>
<td>0.03</td>
<td>non_void_percent</td>
<td>0.02</td>
<td>NEXP</td>
<td>0.03</td>
</tr>
<tr>
<td>HEFF</td>
<td>0.03</td>
<td>CREF</td>
<td>0.02</td>
<td>direct_test_no</td>
<td>0.02</td>
</tr>
<tr>
<td>CREF</td>
<td>0.03</td>
<td>HDF</td>
<td>0.02</td>
<td>assertion_no</td>
<td>0.02</td>
</tr>
<tr>
<td>XMET</td>
<td>0.03</td>
<td>MI(class)</td>
<td>0.02</td>
<td>assertion-McCabe</td>
<td>0.02</td>
</tr>
<tr>
<td>NAND</td>
<td>0.03</td>
<td>is_public</td>
<td>0.01</td>
<td>NOCL</td>
<td>0.02</td>
</tr>
<tr>
<td>VREF</td>
<td>0.03</td>
<td>is_nested</td>
<td>0.01</td>
<td>CREF</td>
<td>0.02</td>
</tr>
<tr>
<td>NOA</td>
<td>0.01</td>
<td>method_length</td>
<td>0.01</td>
<td>NOA</td>
<td>0.02</td>
</tr>
<tr>
<td>NEXP</td>
<td>0.03</td>
<td>nested_depth</td>
<td>0.01</td>
<td>MINC(class)</td>
<td>0.02</td>
</tr>
<tr>
<td>method_length</td>
<td>0.02</td>
<td>assertion_no</td>
<td>0.01</td>
<td>method_length</td>
<td>0.01</td>
</tr>
<tr>
<td>NOCL</td>
<td>0.02</td>
<td>getter_percent</td>
<td>0.01</td>
<td>nested_depth</td>
<td>0.01</td>
</tr>
</tbody>
</table>
assertion related metrics are the second significant category right after test directness (in Table 8 in Section 4.1). While in the feature importance analysis, assertion related metrics mostly rank after the top 5 (shown in Tables 10 and 11). To further investigate the reason behind the dramatic changes of ranks for assertion related metrics, we analyse the correlations between test directness (i.e., direct_test_no and test_distance) and assertion related metrics (i.e., assertion_no, assertion-McCabe and assertion_distance). Looking at the correlation results between test directness and assertion related metrics in Table 12, the major reason is that test directness and assertion related metrics are almost collinear in the prediction model (where $|\rho| > 0.87$). To put simply, there are almost no tests without assertions for the six subjects. If the method has a direct test, then the corresponding assertion no. is always greater than 1. Therefore, the ranks of assertion related metrics are not as high as we had initially expected in the feature importance analysis.

Moreover, we would like to put our observations into perspective by comparing our results with the work of Zhang et al. (2018), where they have constructed a similar Random Forest model to predict the result of killable mutant based on a series of features related to mutants and tests. The metrics that are common to their model and ours are Cyclomatic Complexity (COMP), Depth of Inheritance Tree (DIT), nested_depth, Number of Subclasses (NSUB), and method_length. Only two metrics in their study, i.e., method_length (in 6th place) and nested_depth (in 10th place) appear in our top 15 (Column “all” in Table 11). Especially COMP which ranks nine in their results is not in our top 15. There are multiple reasons for the difference in results: (i) we consider a much larger range of metrics, which provide a better explanatory power (statistically speaking) than the one in their paper; (ii) our goal is to determine patterns in production and test code that may prevent killing some mutants while Zhang et al. (2018) predict if a mutant is killable (aka different prediction target and different granularity level). Besides, as we see later (next section), we can use our model to determine common anti-patterns with proper statistical methods. (iii) the subjects used in our experiment are different from theirs. For example, in project java-apns-apns=0.2.3 (Column “4” in Table 11), COMP appears among the top 15.

5. RQ4 code refactoring

Our goal is to investigate whether we can refactor away the observability issue that we expect to hinder tests from killing mutants and thus to affect the mutation score. In an in-depth case study, we manually analysed 16 code fragments to understand better the interaction between testability/observability metrics that we have been investigating, and the possibilities for refactoring.

Our analysis starts from the combined model, which as Table 9 shows, takes the leading position among the models. We then apply Principal Component Analysis (PCA) (Wold et al., 1987) to perform feature selection, which, as Table 13 shows, leaves us with 36 features (or metrics). Then, as discussed in Section 3, we build a decision tree based on those 36 metrics using J48 (shown in Fig. 4), and select the top 6 leaves (also called end nodes) in the decision tree for further manual analysis as potential refactoring guidelines. We present the top six anti-patterns in Table 14.

Here, we take a partial decision tree to demonstrate how we generate rules (shown in Fig. 5). In Fig. 5, we can see that there are three attributes (marked as an ellipse) and four end nodes or leaves (marked as a rectangle) in the decision tree. Since we would like to investigate how code refactoring increases mutation score (RQ4), we only consider the end nodes labelled with “LOW” denoting mutation score $< 0.5$. By combining the conditions along the paths of the decision tree, we obtain the two rules for “LOW” end nodes (as shown in the first column of the table in Fig. 5). For every end node, there are two values attached to the class: the first is the number of instances that correctly fall into the node, the other is the instances that incorrectly fall into the node. The accuracy in the table is computed by the number of correct instances divided by that of total instances. As mentioned earlier, we select the top 6 end nodes from the decision tree, where the end nodes are ranked by the number of correct instances under the condition accuracy $\geq 0.8$.

After selecting the rules, the first author of this paper has conducted the main task of the manual analysis. If there were any questions during the manual analysis, the attempts of refactoring or adding tests are discussed among all the authors to reach an agreement. In our actual case study, we manually analysed 16 cases in total. Due to space limitations, we only highlight six cases in this paper (all details are available on GitHub (Zhu, 2019)). We will discuss our findings in code refactoring case by case.
Table 12: Spearman results of test directness vs. assertions in terms of rho.

<table>
<thead>
<tr>
<th>(rho)</th>
<th>assertion_no</th>
<th>assertion-McCabe</th>
<th>assertion_distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct_test_no</td>
<td>0.9604</td>
<td>0.9472</td>
<td>0.9334</td>
</tr>
<tr>
<td>test_distance</td>
<td>-0.8707</td>
<td>-0.8707</td>
<td>-0.8707</td>
</tr>
</tbody>
</table>

Fig. 5. Demo of rule generation.

```java
/**
 * Draws the value label just below the center of the dial.
 * @param g2 the graphics device.
 * @param area the plot area.
 */
protected void drawValueLabel(Graphics2D g2, Rectangle2D area) {
    g2.setFont(this.valueFont);
    g2.setPaint(this.valuePaint);
    String valueStr = "No value";
    if (this.dataset != null) {
        Number n = this.dataset.getValue();
        if (n != null) {
            valueStr = this.tickLabelFormat.format(n.doubleValue()) + " " + this.units;
        }
    }
    float x = (float) area.getCenterX();
    float y = (float) area.getCenterY() + DEFAULT_CIRCLE_SIZE;
   TextUtils.drawAlignedString(valueStr, g2, x, y, TextAlign.TOP_CENTER);
}
```

**Listing 4:** plot.MeterPlot::drawValueLabel (Case 1)

### 5.1. Case 1: plot.MeterPlot::drawValueLabel from JFree Chart

This case (shown in Listing 4) is under anti-pattern **Rule 1:** test_distance > 5 & (loop (loop)) ≤ 0 & is_nested = 0 & is_public = 0 & XMET > 4 & (loop) ≤ 0 & NOCL ≤ 9 & non-void_percent ≤ 0.42. In total, there are 5 mutants generated from this method (shown in Table 15). All 5 mutants survive the test suite.

**Code changes.** We start with test_distance > 5 which means there is no direct test for this method. Accordingly, we add one direct test (shown in Listing 5). However, Mutant 4 and 5 cannot be killed because the DrawValueLabel(...) method is void. In particular, this means that the changes in the state caused by the TextUtils.drawAlignedString() method (line 1158) cannot be assessed. This is indicated by non-void_percent ≤ 0.42 in Rule 1. We then refactor the method to have it return Rectangle2D (shown in Listing 6). Also, we improve the direct test for this method in Listing 5 by adding a new test method (shown in Listing 7) to avoid the assertion roulette test smell (Moonen et al., 2008; Palomba et al., 2016). By refactoring the method to non-void and adding a direct test, all previously surviving mutants are now successfully killed.

### 5.2. Case 2: axis.SymbolAxis::drawGridBands from JFree Chart

This case (shown in Listing 8) is under **Rule 2:** test_distance > 5 & (loop (loop)) ≤ 0 & is_nested = 0 & is_public = 0 & XMET > 4 & (loop) ≤ 0 & NOCL > 9. In total, 4 mutants are generated from this method (see Table 16). None of the mutants are killed.

**Code changes.** It is clear that this method is private, thus, it is impossible to call this method from outside the class directly. We first refactor this method from private to public. This is revealed by is_public = 0 in Rule 2.
Table 13
Selected feature by PCA.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Condition</th>
<th>Density</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>is_public</td>
<td>(cond)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is_static</td>
<td>(cond(cond))</td>
<td></td>
<td>LMET</td>
</tr>
<tr>
<td>is_void</td>
<td>(cond(loop))</td>
<td></td>
<td>NLOC</td>
</tr>
<tr>
<td>is_nested</td>
<td>(loop)</td>
<td></td>
<td>VDEC</td>
</tr>
<tr>
<td>method_length</td>
<td>(loop(cond))</td>
<td></td>
<td>TDN</td>
</tr>
<tr>
<td>nested_depth</td>
<td>(loop(loop))</td>
<td></td>
<td>NAND</td>
</tr>
<tr>
<td>direct_test_no</td>
<td>non-void_percent</td>
<td></td>
<td>LOOP</td>
</tr>
<tr>
<td>test_distance</td>
<td>getter_percent</td>
<td></td>
<td>MOD</td>
</tr>
<tr>
<td>assertion_no</td>
<td>assertion-McCabe</td>
<td></td>
<td>NOPR</td>
</tr>
</tbody>
</table>

Table 14
Top six anti-patterns from J48 decision tree 4.

<table>
<thead>
<tr>
<th>Rule no.</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>test_distance &gt; 5 &amp;&amp; (loop(loop)) ≤ 0 &amp;&amp; is_nested = 0 &amp;&amp; is_public = 0 &amp;&amp; XMET &gt; 4 &amp;&amp; (loop) ≤ 0 &amp;&amp; NOCL ≤ 9 &amp;&amp; non-void_percent ≤ 0.42</td>
</tr>
<tr>
<td>2</td>
<td>test_distance &gt; 5 &amp;&amp; (loop(loop)) ≤ 0 &amp;&amp; is_nested = 0 &amp;&amp; is_public = 0 &amp;&amp; XMET &gt; 4 &amp;&amp; (loop) ≤ 0 &amp;&amp; NOCL &gt; 9</td>
</tr>
<tr>
<td>3</td>
<td>test_distance &gt; 5 &amp;&amp; (loop(loop)) ≤ 0 &amp;&amp; is_nested = 0 &amp;&amp; is_public = 0 &amp;&amp; NOCL ≤ 4 &amp;&amp; NOCL &gt; 0 &amp;&amp; is_static = 0 &amp;&amp; getter_percent ≤ 0.01 &amp;&amp; HBUG ≤ 0.02 &amp;&amp; method_length &gt; 3</td>
</tr>
<tr>
<td>4</td>
<td>test_distance &gt; 5 &amp;&amp; (loop(loop)) ≤ 0 &amp;&amp; is_nested = 0 &amp;&amp; is_public = 1 &amp;&amp; NOCL &gt; 4 &amp;&amp; (cond) ≤ 0 &amp;&amp; is_static = 0 &amp;&amp; LMET ≤ 1 &amp;&amp; HBUG &gt; 8 &amp;&amp; NOPR &gt; 5 &amp;&amp; is_void = 1</td>
</tr>
<tr>
<td>5</td>
<td>test_distance ≤ 5 &amp;&amp; is_void = 1 &amp;&amp; nested_depth ≤ 0 &amp;&amp; NOS &lt; 2 &amp;&amp; assertion-density ≤ 0.14 &amp;&amp; MOD &gt; 1</td>
</tr>
<tr>
<td>6</td>
<td>test_distance ≤ 5 &amp;&amp; is_void = 1 &amp;&amp; nested_depth ≤ 0 &amp;&amp; NOS &gt; 2 &amp;&amp; assertion-density ≤ 0.22 &amp;&amp; CREF &lt; 1 &amp;&amp; XMET &gt; 0</td>
</tr>
</tbody>
</table>

Table 15
Summary of mutants from Listing 4 (Case 1).

<table>
<thead>
<tr>
<th>ID</th>
<th>Line no.</th>
<th>Mutator</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1146</td>
<td>Removed call to java.awt.Graphics2D::setFont</td>
<td>SURVIVED</td>
</tr>
<tr>
<td>2</td>
<td>1147</td>
<td>Removed call to java.awt.Graphics2D::setPaint</td>
<td>SURVIVED</td>
</tr>
<tr>
<td>3</td>
<td>1149</td>
<td>Negated conditional</td>
<td>SURVIVED</td>
</tr>
<tr>
<td>4</td>
<td>1151</td>
<td>Negated conditional</td>
<td>SURVIVED</td>
</tr>
<tr>
<td>5</td>
<td>1157</td>
<td>Replaced float addition with subtraction</td>
<td>SURVIVED</td>
</tr>
</tbody>
</table>

Listing 6: Refactoring of Listing 4 (Case 1)

```java
protected Rectangle2D drawValueLabel(Graphics2D g2, Rectangle2D area) {
    g2.setFont(this.valueFont);
    g2.setPaint(this.valuePaint);
    String valueStr = "No value";
    if (this.dataset != null) {
        Number n = this.dataset.getValue();
        if (n != null) {
            valueStr = this.tickLabelFormat.format(n.doubleValue()) + " " + this.units;
        }
    }
    float x = (float) area.getCenterX();
    float y = (float) area.getCenterY() + DEFAULT_CIRCLE_SIZE;
    return TextUtils.drawAlignedString(valueStr, g2, x, y, TextAnchor.TOP_CENTER);
}
```

Code changes. Starting with test_distance > 5, we add a direct test for this method (shown in Listing 12), which works perfectly to kill the mutant.

5.3. Case 3: builder.IDKey::hashCode from Apache Commons Lang

This case (shown in Listing 11) is under Rule 3: test_distance > 5 && (loop(loop)) ≤ 0 && is_nested = 0 && is_public = 1 && NOCL ≤ 4 && NOCL > 0 && is_static = 0 && getter_percent ≤ 0.01 && HBUG ≤ 0.02 && method_length > 3. Only one mutant is generated for this method: a mutant that replaces the return value with (x == 0 ? 1 : 0). This mutant survives.
@Test
public void testDrawValueLabelArea() {
  MeterPlot p1 = new MeterPlot(new DefaultValueDataset(1.23));
  BufferedImage image = new BufferedImage(3, 4, BufferedImage.TYPE_INT_ARGB);
  Graphics2D g2 = image.createGraphics();
  Rectangle2D drawArea = p1.drawValueLabel(g2, area);
  assertEquals(0.5, drawArea.getCenterX(), 0.01);
  assertEquals(18.8671875, drawArea.getCenterY(), 0.01);
  assertEquals(15.0, drawArea.getHeight(), 0.01);
  assertEquals(64.0, drawArea.getWidth(), 0.01);
}

Listing 7: Improved direct test for Listing 4 (Case 1)

/**
 * Similar to @link Color#darker().
 * The essential difference is that this method maintains the alpha-channel unchanged<br>
 * @param paint a (Code Color)
 * @return a darker version of the (Code Color)
 */
private static Color darker(Color paint) {
  return new Color((int)(paint.getRed() * FACTOR),
                   (int)(paint.getGreen() * FACTOR),
                   (int)(paint.getBlue() * FACTOR), paint.getAlpha());
}

Listing 8: axis.SymbolAxis::drawGridBands (Case 2)

public static Color darker(Color paint) {
  return new Color((int)(paint.getRed() * FACTOR),
                   (int)(paint.getGreen() * FACTOR),
                   (int)(paint.getBlue() * FACTOR), paint.getAlpha());
}

Listing 9: Refactoring of Listing 8 (Case 2)

@Test
public void testDarker() {
  Color paint = new Color(10, 20, 30);
  Color darker = PaintAlpha.darker(paint);
  assertEquals(7, darker.getRed());
  assertEquals(14, darker.getGreen());
  assertEquals(21, darker.getBlue());
}

Listing 10: Direct test for Listing 8 (Case 2)

particular change applied is the removal of the call to AbstractCategoryPlot::drawOutline. The original test suite did not kill the mutant.

Code changes. Based on test_distance > 5, we add one direct test (as shown in Listing 14) for this method to kill the surviving mutant.

5.5. Case 5: builder.ToStringStyle::setUseShortClassName from Apache Commons Lang

This case (shown in Listing 15) is under Rule 5: test_distance ≤ 5 && is_void = 1 && nested_depth ≤ 0 && NOS ≤ 2 && assertion-density ≤ 0.14 && MOD > 1. In this case, a single (surviving) mutant is generated that removes the call to builder.ToStringStyle::setUseShortClassName.

Code changes. We can see that Rule 5 is different from the previous rule in that test_distance is less than 5, while in
/*<p>Sets whether to output short or long class names.</p>*/

**Listing 15:** builder.ToStringStyle::setUseShortClassName (Case 5)

```java
@Test
public void testSetUseShortClassName() {
    assertTrue(STYLE.isUseShortClassName());
    STYLE.setUseShortClassName(false);
    assertFalse(STYLE.isUseShortClassName());
    STYLE.setUseShortClassName(true);
    assertTrue(STYLE.isUseShortClassName());
}
```

**Listing 16:** Additional assertions for Listing 15 (Case 5)

```java
/**
 * @param max Maximum number of evaluations.
 *
 * Construct the exception.
 */

public TooManyEvaluationsException(Number max) {
    super(max);
    getContext().addMessage(LocalizedFormats.EVALUATIONS);
}
```

**Listing 17:** exception.TooManyEvaluationsException::<init> (Case 6)

```java
public List<Localized> getMsgPatterns() {
    return msgPatterns;
}
```

**Listing 18:** Refactoring of Listing 17 (Case 6)

```java
@Test
public void testMsgPatterns() {
    final int max = 12345;
    final TooManyEvaluationsException e = new TooManyEvaluationsException(max);
    final String msg = e.getLocalizedMessage();
    Assert.assertTrue(e.getContext().getMsgPatterns().contains(LocalizedFormats.EVALUATIONS));
}
```

**Listing 19:** Additional assertion for Listing 17 (Case 6)

As such, our first step is to add a getter for msgPatterns (shown in Listing 18). In **Rule 6**, we can see that is_void = 1 is the underlying cause since void methods could be difficult to test if no getters for private fields exist.

To kill the surviving mutant, we add one extra assertion (in a new test method) to examine the changes in msgPatterns (in Listing 19). This action is also partly evidenced by assertion-density ≤ 0.22 in **Rule 6**. As assertion-density denotes the ratio between the total number of assertions in direct tests and the lines of code in direct tests, low assertion-density is a sign of insufficient assertions in the direct tests to detect the mutant.

5.7. RQ4 summary

Based on all 16 cases that we analysed (available in our GitHub repository Zhu, 2019), we found that our code observability metrics can lead to simple refactorings that enable to kill mutants that were previously not being killed. Ultimately, this leads to an increase of the mutation score:

- Most cases can be easily fixed by adding direct tests if useShortClassName.<init>
- Most cases can be easily fixed by adding assertions if test_distance ≤ 5.
- Private methods must be refactored to protected/public for testing (indicated by is_public = 0).
- Three void methods had to be refactored to non-void (indicated by is_void = 1 and non-void_percent ≤ 0.42).
- One void method needed an additional getter because a private field was changed (indicated by is_void = 1).

5.8. Discussion

From the findings of RQ4, we can see that some code refactorings break OO design principles (Booch, 2006). For instance, we suggest to change the access modifier from private to protected/public to kill the mutants; this violates the idea of Encapsulation, the ability to protect some components of the object from external entities (Booch, 2006). A new hypothesis emerges from our study: the trade-off between OO design principles and testing and hence software testability (Suri and Singhani, 2015). The main concepts of OO design are centred around the features of Data abstraction, Encapsulation, Inheritance, Polymorphism, and Dynamic binding. However, some factors such as Encapsulation and Inheritance could increase the complexity of OO systems and
Table 16
Summary of mutants from Listing 8 (Case 2).

<table>
<thead>
<tr>
<th>ID</th>
<th>Line no.</th>
<th>Mutator</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>165</td>
<td>Mutated return of Object value for org/jfree/chart/util/PaintAlpha::darker to ( if (x != null) null else throw NO_COVERAGE</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>166</td>
<td>Replaced double multiplication with division</td>
<td>NO_COVERAGE</td>
</tr>
<tr>
<td>3</td>
<td>167</td>
<td>Replaced double multiplication with division</td>
<td>NO_COVERAGE</td>
</tr>
<tr>
<td>4</td>
<td>168</td>
<td>Replaced double multiplication with division</td>
<td>NO_COVERAGE</td>
</tr>
</tbody>
</table>

hence hinder testing and testability (Suri and Singhani, 2015). Existing literature (Mouchawrab et al., 2005; Singh and Saha, 2010; Zhou et al., 2012; Nazir et al., 2010) has already addressed this dilemma. Mouchawrab et al. (2005) pointed out that increasing the size of the inheritance hierarchy could increase the cost of testing due to dynamic dependencies. Singh and Saha’s work (Singh and Saha, 2010) has shown that Inheritance and Polymorphism increase testing effort and lower software testability. All the works above indicate that there is a trade-off between OO design features and software testability. Currently, it is up to practitioners to balance the two perspectives themselves, depending on the requirements of software and their preferences.

In the context of mutation testing, a similar trade-off between OO design features and the ease of killing mutants exists. In this study, we relate the ease of killing mutants to the testability and observability. In Section 5.7, we found that a simple strategy to kill all the mutants is to write additional direct tests and/or assertions. However, some OO design features related to Encapsulation, such as the private access modifier (see Listing 8), increase the difficulty to add a direct test. Also, the void return type prevents killing the mutants generated from the immediate states that cannot propagate to the output (see Listing 4). As such, a very important note here is that our refactoring recommendations listed in Section 5.7 are centred around the anti-patterns based on the testability and observability; they do not take OO design principles into consideration. The recommendations attempt to help developers in understanding the cause of the low mutation score considering testability and observability, but not all surviving mutants are due to test quality.

Take Listing 8 for instance. The developer found the mutation score of this method is low, and our tool shows the low mutation score is mainly due to private access control modifier. Then, the developer can decide to ignore the surviving mutants if he cannot break Encapsulation based on the requirement. Or if this method is critical and must be well-tested according to the document, he may alter the access control modifier from private to protected/public to kill the mutants. Whether the developers make use of these testability and observability recommendations depends on their choices with regard to either (1) adding test cases (Beller et al., 2015b, 2019a) (2) refactoring the production code to kill the mutants, or (3) ignoring the surviving mutants.

6. Threats to validity

External validity. Our results are based on mutants generated by the operators implemented in PIT. While PIT is a frequently used mutation testing tool, our results might be different when using other mutation tools (Kurtz et al., 2016). Concerning the subject systems selection, we choose six open-source projects from GitHub; the selected projects differ in size, the number of test cases, and application domain. Besides, as mentioned in Section 3, the large number of methods with low number of mutants is an unavoidable bias in our experiment. The reason is partly due to the optimisation mechanism of PIT (Coles, 2019c) and partly due to a large number of short methods in those projects. Nevertheless, we do acknowledge that a broad replication of our study would mitigate any generalisability concerns even further.

Internal validity. The main threat to internal validity for our study is the implementation of the Mutation Observer tool for the experiment. To reduce internal threats to a large extent, we rely on existing tools that have been widely used, e.g., WEKA, MATLAB, and PIT. Moreover, we carefully reviewed and tested all code for our study to eliminate potential faults in our implementation. Another threat to internal validity is the disregard of equivalent mutants in our experiment. However, this threat is unavoidable and shared by other studies on mutation testing that attempt to detect equivalent mutants or not (Grünn et al., 2009; Mirshokraie et al., 2013). Moreover, we consider equivalent mutants as a potential weakness in the software (reported by Coles Coles, 2019b, slide 44–52); thereby, we did not manually detect equivalent mutants in this paper.

Construct validity. The main threat to construct validity is the measurement we used to evaluate our methods. We minimise this risk by adopting evaluation metrics that are widely used in research (such as recall, precision, and AUC), as well as a sound statistical analysis to assess the significance (Spearman’s rank-order correlation).

7. Related work

The notion of software testability dates back to 1991 when Freedman (1991) formally defined observability and controllability in the domain of software. Voas (1992) proposed a dynamic technique coined propagation, infection, and execution (PIE) analysis for statistically estimating the program’s fault sensitivity. More recently, researchers have aimed to increase our collective understanding of testability by using statistical methods to predict testability based on various code metrics. A prime example is the work of Bruntink and van Deursen (2006), who have explored the relationship between nine class-level object-oriented metrics and testability. To the best of our knowledge, no study uses statistical or machine learning methods to investigate the relationship between testability/observability metrics and the mutation score.

Mutation testing was initially introduced as a fault-based testing method which was regarded as significantly better at detecting errors than the covering measure approach (Budd et al., 1979). Since then, mutation testing has been actively investigated and studied, thereby resulting in remarkable advances in its concepts, theory, technology, and empirical evidence. For more literature on mutation testing, we refer to the existing surveys of DeMillo (1989), Offutt and Unch (2001), Jia and Harman (2011), Offutt (2011) and Zhu et al. (2018b). Here we mainly address the studies that concern mutant utility (Just et al., 2017), the efficacy of mutation testing. Yao et al. (2014) have reported on the causes and prevalence of equivalent mutants and their relationship to stubborn mutants based on a manual analysis of 1230 mutants. Visser (2016) has conducted an exhaustive analysis of all possible test inputs to determine how hard it is to kill a mutant considering three common mutation operators (i.e., relational, integer constants and arithmetic operators). His results show that mutant reachability, mutation operators, and oracle sensitivity are the key contributors to determining how hard it is to kill a mutant. Just et al. (2017) have shown a strong correlation.
between mutant utility and context information from the program in which the mutant is embedded. Brown et al. (2017) have developed a method for creating potential faults that are more closely coupled with changes made by actual programmers where they named “wild-caught mutants”. Chekam et al. (2018) have investigated the problem of selecting the fault revealing mutants. They put forward a machine learning approach (decision trees) that learns to select fault revealing mutants from a set of static program features. Jimenez et al. (2018) investigated the use of natural language modelling techniques in mutation testing. All studies above have enriched the understanding of mutation testing, especially its efficacy. However, the aim of our work is different from those studies, as we would like to gain insights into how code quality in terms of testability and observability affects the efforts needed for mutation testing, especially in how to engineer tests to kill more the mutants.

Similar to our study, there have been a few recent studies also investigating the relationships between assertions and test directness with mutation testing. Schuler and Zeller (2013) introduced checked coverage – the ratio of statements that contribute to the computation of values that are later checked by the test suite – as an indicator for oracle quality. In their experiment, they compared checked coverage with the mutation score, where they found that checked coverage is more sensitive than mutation testing in evaluating oracle quality. Huo and Clause (2016) proposed direct coverage and indirect coverage by leveraging the concepts of test directness with conventional statement coverage. They used the mutants as an indicator of the test suite effectiveness, and they found faults in indirectly covered code are significantly less likely to be detected than those in directly covered code. Zhang and Mesbah (2015) evaluated the relationship between test suite effectiveness (in terms of the mutation score) and the (1) number of assertions, (2) assertion coverage, and (3) different types of assertions. They found test assertion quantity and assertion coverage are strongly correlated with the mutation score, and assertion types could also influence test suite effectiveness. Compared to our studies, those works only addressed one or two aspect(s) of code observability in our study. We provide a complete view of the relationships between code observability and mutation testing.

The study most related to ours is that of Zhang et al. (2018)’s predictive mutation testing, where they have constructed a classification model to predict killable mutant result based on a series of features related to mutants and tests. In their discussion, they compared source code related features and test code related features in the prediction model for the mutation score. They found that test code features are more important than source code ones. But from their results, we cannot draw clear conclusions on the impact of production code on mutation testing as their goal is to predict exact killable mutant results. Another interesting work close to our study is Vera-Pérez et al. (2017)’s pseudo-tested methods. Pseudo-tested methods denote those methods that are covered by the test suite, but for which no test case fails even if the entire method body is completely stripped. They rely on the idea of “extreme mutation”, which completely strips out the body of a method. The difference between Vera-Pérez et al. (2017)’s study and ours is that we pay attention to conventional mutation operators rather than “extreme mutation”.

8. Conclusion & future work

This paper aims to bring a new perspective to software developers helping them to understand and reason about the mutation score in the light of testability and observability. This should enable developers to make decisions on the possible actions to take when confronted with low mutation scores. To achieve this goal, we first investigate the relationship between testability and observability metrics and the mutation score. More specifically, we have collected 64 existing source code quality metrics for testability, and have proposed a set of metrics that specifically target observability. The results from our empirical study involving 6 open-source projects show that the 64 existing source code quality metrics are not strongly correlated with the mutation score (\(\rho < 0.27\)). In contrast, the 19 newly proposed code observability metrics, that are defined in terms of both production code and test cases, do show a stronger correlation with the mutation score (\(\rho < 0.5\)). In particular, test directness, test_distance, and direct_test_no stand out.

To better understand the causality of our insights, we continue our investigation with a manual analysis of 16 methods that scored particularly bad in terms of mutation score, i.e., a number of mutants were not killed by the existing tests. In particular, we have refactored these methods and/or added tests according to the anti-patterns that we established in terms of the code observability metrics. Our aim here is to establish whether the removal of the observability anti-patterns would lead to an increase in the mutation score. We found that these anti-patterns can indeed provide insights in order to kill the mutants by indicating whether the production code or the test suite needs improvements. For instance, we found that private methods (expressed as is_public=0 in our schema) are prime candidates to potentially refactor to increase their observability, e.g., by making them public or protected for testing purposes.

However, some refactoring recommendations could violate OO design principles. For example, by changing private to protected/public we increase observability, but we also break the idea of encapsulation. Therefore, we suggest developers make a choice between – (1) adding test cases, (2) refactoring the production code to kill the mutants, or (3) ignoring the surviving mutants – by considering the trade-off between OO design features and testability/observability.

To sum up, our paper makes the following contributions:

1. 19 newly proposed code observability metrics
2. A detailed investigation of the relationship between testability/observability metrics and the mutation score (RQ1–RQ3)
3. A case study with 16 code fragments to investigate whether removal of the anti-patterns increases the mutation score (RQ4)
4. A guideline for developers to make choices when confronting low mutation scores
5. A prototype tool coined MUTATION OBSERVER (openly available on GitHub (Zhu, 2019)) that automatically calculates code observability metrics

Future work. With our tool, and since the results are encouraging, we envision the following future work: (1) conduct additional empirical studies on more subject systems; (2) evaluate the usability of our code observability metrics by involving practitioners; (3) investigate the relations between more code metrics (e.g., code readability) and mutation score.

CRediT authorship contribution statement

Qianqian Zhu: Methodology, Software, Data curation, Visualization, Investigation, Validation, Writing - original draft, Resources. Andy Zaidman: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing - review & editing. Annibale Panichella: Methodology, Supervision, Writing - review & editing.
Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially funded by the Netherlands Organisation for Scientific Research (NWO) through the “TestRoots” project. Further funding came from the EU Horizon 2020 ICT-10-2016-RIA “STAMP” project (No.731529).

References


Qianqian Zhu received master’s degree from Imperial College London in 2014 and Ph.D. degree in Software Engineering from Delft University of Technology in 2020. Her research interests include software testing, debugging, software quality, fault predication, mutation testing, and static analysis.

Andy Zaidman received M.Sc. (2002) and Ph.D. (2006) from the University of Antwerp, Belgium. He is currently a full professor in software engineering at the Delft University of Technology, The Netherlands. His research interests are in software evolution and software testing.

Annibale Panichella received M.Sc. and Ph.D. from the University of Salerno, Italy, in 2010 and 2014, respectively. He is currently an Assistant Professor in the Software Engineering Research Group (SERG) at Delft University of Technology (TU Delft) in Netherlands. His research interests include security testing, evolutionary testing, search-based software engineering, textual analysis, and empirical software engineering.