Synthesizing Stealthy Reprogramming Attacks on Cardiac Devices

Nicola Paoletti
Royal Holloway, University of London, UK

Zhihao Jiang
ShanghaiTech University, China

Md Ariful Islam
Texas Tech University, USA

Houssam Abbas, Rahul Mangharam
University of Pennsylvania, USA

Shan Lin, Zachary Gruber, Scott A. Smolka
Stony Brook University, USA

ACM Reference Format:
Nicola Paoletti, Zhihao Jiang, Md Ariful Islam, Houssam Abbas, Rahul Mangharam, and Shan Lin, Zachary Gruber, Scott A. Smolka. 2019. Synthesizing Stealthy Reprogramming Attacks on Cardiac Devices. In *Proceedings of CPS-SR 2019: 2nd Workshop on Cyber-Physical Systems Security and Resilience (CPS-SR 2019).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Paper appeared in: Nicola Paoletti, et al. Synthesizing Stealthy Reprogramming Attacks on Cardiac Devices. In *10th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, 2019.

EXTENDED ABSTRACT

An *Implantable Cardioverter Defibrillator* (ICD) is a medical device for the detection and treatment of potentially fatal arrhythmias such as ventricular tachycardia (VT) and ventricular fibrillation (VF). ICDs run embedded software that processes intracardiac signals, called *electrograms* (EGMs), to detect arrhythmias and deliver appropriate therapy in the form of electrical shocks. EGMs are of three types: *atrial and ventricular EGMs*, describing the local, near-field electrical activity in the right atrium and ventricle, respectively; and the *shock EGM*, a far-field signal that gives a global view of the electrical activity. See Figure 1.

ICD software implements so-called *discrimination algorithms* which comprise multiple discriminators for the detection and classification of arrhythmia episodes based on the analysis of EGM features such as ventricular intervals and signal morphology. In particular, the ICD algorithm needs to distinguish between potentially fatal Ventricular Tachy-arrhythmias (VT) and non-fatal Supra-Ventricular Tachy-arrhythmias (SVTs).

ICD discriminators feature a number of programmable parameters that, if adequately configured, ensure minimal rates of arrhythmia mis-classification [9]. In contrast, wrongly configured parameters can result in unnecessary shocks (*false positive* classification errors), which are painful and damage the cardiac tissue, and even worse can prevent required therapy (*false negatives*), leading to sudden cardiac death.

An ICD *reprogramming attack* is one that alters the device's parameters to induce mis-classification and inappropriate or missed therapy. Reprogramming attacks can significantly compromise patient safety, with high-profile patients being obvious targets (e.g. former US Vice President Cheney had his pacemaker's wireless access disabled to prevent assassination attempts [12]). Seminal work by Halperin et al. [7] demonstrated that ICDs can be accessed and reprogrammed by unauthorized users using off-the-shelf software radios. More recently, over half a million cardiac devices have been recalled by the FDA for security risks related to wireless communication [6], and researchers managed to gain control of a pacemaker/ICD by exploiting vulnerabilities in the device's remote monitoring infrastructure [13]. These incidents confirm that vulnerabilities in implantable cardiac devices exist, and a thorough investigation of cyber-attacks on ICDs is needed to improve device safety and security.

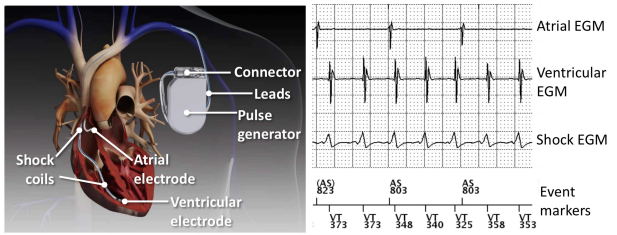


Figure 1: Left: illustration of a dual-chamber ICD. Right: sensed atrial, ventricular and shock electrograms. Event markers label sensed impulses (AS: atrial, VT: ventricular tachycardia) and corresponding intervals in milliseconds.

In this paper, we present a formal approach for the automated synthesis of ICD reprogramming attacks that are both *effective*, i.e., lead to fundamental changes in the required therapy, and *stealthy*, i.e., involve minimal changes to the nominal ICD parameters. Stealthy attacks are therefore difficult to detect and even if detected, would most likely be attributed to a clinician's error in configuring the device. We follow a model-based approach, as the attacks are not evaluated on the actual hardware but on a model of the ICD algorithm. We focus on the *Rhythm ID* algorithm implemented in Boston Scientific (BSc) ICDs, one of the principal ICD manufacturers, which was compiled from device manuals and the medical literature [4, 15].

Below we provide an overview of the ICD algorithm, attack model, solution method, and results. See [10] for more details.

ICD Discrimination Algorithm

Figure 2 illustrates the *Rhythm ID* algorithm implemented in BSc ICDs. The algorithm consists of a number of discriminators arranged in a decision tree-like structure, where each discriminator depends on one or more programmable parameters. Leaves of the

CPS-SR 2019, April 15, 2018, Montreal, Canada
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

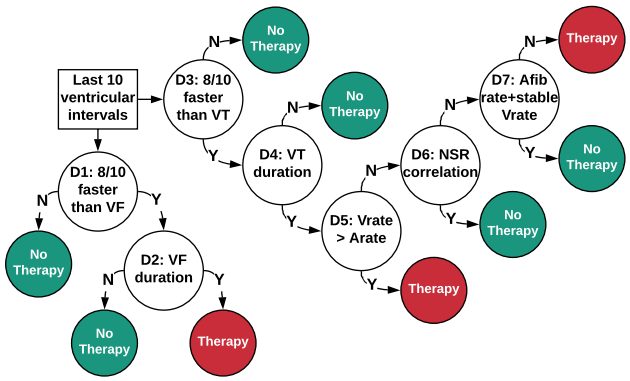


Figure 2: Discrimination tree of the Boston Scientific *Rhythm ID* algorithm. White nodes denote discrimination criteria.

Name	Description	**Nominal** (Programmable)
VF_{th} (BPM)	VF detection threshold	**200** (110 : 5 : 210, 220 : 10 : 250)
VT_{th} (BPM)	VT detection threshold	**160** (90 : 5 : 210, 220)
$AFib_{th}$ (BPM)	AFib detection threshold	**170** (100 : 10 : 300)
VFdur (s)	Sustained VF duration	**1.0** (1 : 0.5 : 5, 6 : 1 : 15)
VTdur (s)	Sustained VT duration	**2.5** (1 : 0.5 : 5, 6 : 1 : 15, 20 : 5 : 30)
$NSRcor_{th}$	Rhythm Match score	**0.94** (0.7 : 0.01 : 0.96)
stb (ms^2)	Stability score	**20** (6 : 2 : 32, 35 : 5 : 60, 70 : 10 : 120)

Table 1: Parameters of the Rhythm ID algorithm, including nominal and programmable values [4]. AFib: atrial fibrillation. $n : k : m$ denotes the sequence $n, n + k, n + 2k, \ldots, m$. Thresholds are programmed in BPM (beats per minute) but the algorithm employs the corresponding time duration.

tree determine whether or not therapy is delivered during the current cardiac cycle. The parameters of *Rhythm ID* are in Table 1.

The algorithm is executed at each ventricular event, which marks the end of the corresponding cardiac cycle, and works as follow. If at least eight out of the last ten ventricular intervals (i.e., the time between two consecutive ventricular beats) are shorter than the programmable threshold VF_{th} (discriminator **D1**), then onset of VF is detected, in which case, the algorithm checks if the VF episode persists, i.e., if the arrhythmia lasts for time VFdur (**D2**). If VF persists, therapy is given. A similar logic is applied to detect the onset and persistence of VT (discriminators **D3** and **D4**), but using different parameters, VT_{th} and VTdur. The main difference is that, if VT lasts for time VTdur (**D4** is true), before delivering therapy the algorithm ensures that the episode is not mistaken for SVT by checking discriminators **D5**-**D7**. See [10] for more details.

Attack Model

Our method, illustrated in Figure 3, synthesizes device parameters that are optimal with respect to the effectiveness-stealthiness tradeoff (i.e., lie along the corresponding Pareto front). We deem an attack effective when it compromises at least one decision of the discrimination algorithm to introduce false negatives (FN), i.e., prevent a required therapy during VF/VT, or false positives (FP), i.e., introduce inappropriate therapy during SVT. These are called *FN attacks* and *FP attacks*, respectively. Stealthiness depends on the clinician's ability to detect the attack, and thus, we are interested in finding malicious parameters that exhibit small deviations from the clinical settings of the victim's ICD, changes that are difficult for the clinician to notice or that can be mistaken for human error.

Reprogramming attacks are synthesized in an offline *training phase*, which allows the attacker to synthesize malicious parameters with optimal effectiveness and stealthiness with respect to a set of *training EGM signals*. We formulate this problem as one of multi-objective optimization, and solve it using *optimization modulo theories* (OMT) techniques [3], an extension of SMT for finding models that optimize given objectives. OMT is uniquely suited to solve this problem, because the problem is combinatorial in nature (parameters can be configured from a finite set of values), and is also constrained by the behavior of the ICD algorithm, which can be adequately encoded as SMT constraints.

To evaluate how the attack generalizes with previously unseen signals, which mimic the unknown EGM of the victim, we *validate* the parameters synthesized in the training phase *using a disjoint test dataset*.

We employ the method of [8] to generate synthetic EGMs with prescribed arrhythmia, a method based on combining simulations of a timed-automaton model of the electrical conduction system with true EGM morphologies obtained from real patients [5]. This allows the attacker to synthesize malicious parameters tailored to the victim's cardiac condition. We call such attacks *condition-specific*. We also consider more generic datasets that include signals for different arrhythmias (*condition-agnostic attacks*), suitable when the attacker has little knowledge of the victim's condition.

Real-world attacks. Our approach does not provide an exhaustive recipe for ICD attacks, as the actual algorithms on-board devices usually contain more decision branches than we have chosen to model, and indeed more than is described in the open literature. To conduct a real-world attack, the attacker must know 1) the ICD model of the victim, so that it can select the appropriate discrimination algorithm to use in the training phase, and 2) the device's communication protocol in order to change the parameter settings. Halperin et al. [7] show how 1) and 2) can be obtained from real devices. Further, the radio antenna transmitting the attack signals must be physically close to the victim.

Countermeasures. A possible countermeasure is to store a copy of the physician-programmed values both in a hospital database and in a secure memory location on the device. The currently programmed values are regularly checked against the stored, golden values. Any discrepancy leads to an alarm. A more general countermeasure is to secure device access through an authentication token (smart card, NFC device, etc.) that shares a secret key with the device [14]. Finally, a simple attack detection method would be to alert the patient (e.g., with a beep) whenever a communication happens with the device [7].

OMT Encoding

Formally, we describe *effectiveness* of an attack as the proportion of training signals where an FN attack (preventing required therapy) or an FP attack (delivering inappropriate therapy) occurs. *Stealthiness* is define as the distance between the reprogrammed and the default parameters. Since ICD parameters can be only programmed to a finite set of values (see Table 1), we quantify the distance between

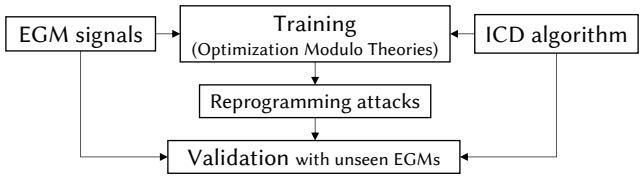


Figure 3: Overview of our method for synthesis of stealthy reprogramming attacks on ICDs.

two parameters as the number of programmable values separating them. Other definitions of distance are equally supported.

We formalize the behavior of the BSc discrimination algorithm in the framework of Satisfiability Modulo Theories (SMT) [1]. ICD Parameters are represented as uninterpreted constants in the SMT encoding, and parameter synthesis corresponds to finding a satisfiable assignment to those constants, i.e., a so-called model. Among the satisfiable parameters, we find those attaining the optimal effectiveness-stealthiness trade-off by solving an OMT problem, i.e., an extension of SMT for finding models that optimize given objectives [3]. Since we are interested in analyzing the behavior of the algorithm offline over a fixed set of EGM signals, we can pre-compute for each signal the non-linear operations underlying some of the discriminators, resulting in an SMT encoding in the decidable theory of quantifier-free linear integer real arithmetic (SMT QF_LIRA).

The behavior of the algorithm for the j-th signal is described by a sequence of symbolic states $s_{j,0}, \ldots s_{j,N_j}$, one for each cardiac cycle, where N_j is the number of cycles in the j-th signal. The evolution of the discrimination algorithm over the training signals is characterized by the following formula (inspired by bounded model checking [2]):

$$\mathsf{paramRanges} \wedge \bigwedge_{j=1}^{|S|} \left(\mathsf{Init}(s_{j,0}) \wedge \bigwedge_{k=0}^{N_j-1} T(k, s_{j,k}, s_{j,k+1}) \right) \quad (1)$$

where paramRanges is a predicate describing the programmable values of the ICD parameters (see Table 1); $\mathsf{Init}(s_{j,0})$ is the predicate for constraining the initial state of the algorithm, and $T(k, s_{j,k}, s_{j,k+1})$ is the transition relation determining from the current state and cardiac cycle, the admissible states of the algorithm at the next cycle. See [11] for more details about the formalization of the reprogramming attack and its full SMT QF_LIRA encoding.

Results

We evaluate our approach by synthesizing attacks for 19 different arrhythmias (i.e., *condition-specific* attacks), as well as more generic attacks (*condition-agnostic*) that are suitable when the attacker has little knowledge of the victim's condition. Our results demonstrate that some arrhythmias are particularly vulnerable, as only minor changes to the detection thresholds are sufficient to prevent the required therapy.

For the synthesis of condition-specific attacks, we synthesize Pareto-optimal parameters using a training set of 100 signals for each arrhythmia. We validate the attacks with test sets of 50 signals per arrhythmia (disjoint from the training sets). We classify the 19 arrhythmias into two categories, VT and SVT, depending on whether or not the corresponding signals require ICD therapy under nominal parameters. In particular, we have 8 VT arrhythmias (subject to FN attacks) and 11 SVT arrhythmias (subject to FP attacks). For condition-agnostic attacks, we consider two attacks for generic VT and SVT arrhythmias, using training sets of 200 EGMs randomly sampled among the 8 VT-like arrhythmias and the 11 SVT arrhythmias, respectively. We validate the two attacks with disjoint test sets of 100 signals.

Condition-specific attacks. Figure 4 shows the Pareto-optimal fronts for a selection of representative arrhythmias (see [11] for the full set of plots and synthesized parameters). The synthesized attacks attain validation scores[1] that are either positive or very close to zero, indicating that the attacks generalize well with unseen data and, thus, would have comparable effectiveness on the unknown EGM of the victim.

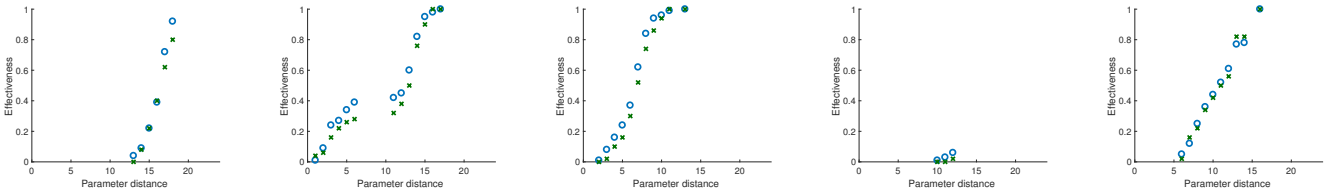
Our method can derive effective FN attacks for all VT arrhythmias, but not all attacks are comparably stealthy (see Figure 4). For instance, for arrhythmia 10 a parameter distance of 7 ensures that the attack is effective with half of the training signals, while for arrhythmia 17, the same effectiveness level is obtained only at a distance of 11 from the nominal parameters (worse stealthiness).

In contrast, FP attacks on SVT arrhythmias are not all equally successful. For arrhythmia 5 we can find parameters with 100% effectiveness as well as stealthy attacks that e.g. are able to affect almost 40% of the signals with a distance of only 5. For arrhythmias 2 and 15 we obtain parameters with nearly 100% effectiveness but with poor stealthiness. Some EGMs turned out to be difficult to attack: for arrhythmia 11 the strongest attack affects only 6% of the signals and, for arrhythmia 9, no Pareto-optimal attacks exist but the trivial one that leaves the nominal parameters unchanged.

The reason why VT arrhythmias are easier to attack is that it takes only a minor increase to the VT and VF detection thresholds (parameters $\mathrm{VF_{th}}$ and $\mathrm{VT_{th}}$) to make the ICD mis-classify a tachyarrhythmia episode. On the other hand, $\mathrm{VF_{th}}$ and $\mathrm{VT_{th}}$ must be reprogrammed to very low values in order for the ICD to classify a slow heart rate as VT/VF and induce unnecessary therapy. This is not always possible because in SVT arrhythmias, the heart rate is often below the lowest programmable values for $\mathrm{VF_{th}}$ (110 BPM) and $\mathrm{VT_{th}}$ (90 BPM). We remark that these results are *provably correct* because OMT is *guaranteed* to find Pareto-optimal attack parameters, when they exist.

Figure 5 compares nominal and reprogrammed parameters over an execution of the BSc algorithm at the start of a VF episode, using an EGM from arrhythmia 10. With nominal parameters, VF duration starts after the last 8/10 ventricular intervals faster than VF (see marker 1 in Fig. 5) and ends after an interval is found below the VF threshold (see marker 2). A new VF duration can start right away, ending this time with a therapy (marker T). Here, the reprogramming attack sets $\mathrm{VF_{th}}$ = 240 BPM (250 ms), $\mathrm{VF_{th}}$ = 185 BPM (325 ms), and VTdur = 7 s. With the higher VF threshold, the attack leads to marking the VF episode as VT, triggering VT duration (marker 3). VT duration ends with one interval found below the

[1]The validation score is the average deviation of the attack effectiveness between training and test data.



(a) Cond. 2 (SVT) (b) Cond. 5 (SVT) (c) Cond. 10 (VT) (d) Cond. 11 (SVT) (e) Cond. 17 (VT)

Figure 4: Pareto fronts for a selection of condition-specific reprogramming attacks (see [11] for the full set of arrhythmias). Blue dots: Pareto front obtained with training signals. Green crosses: effectiveness of the synthesized parameters on the test signals.

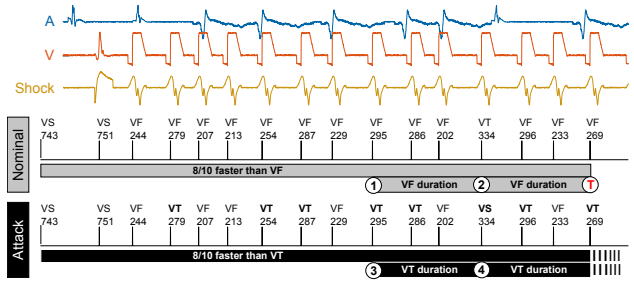
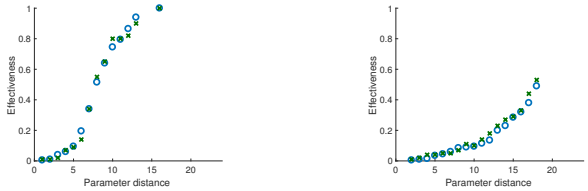


Figure 5: Execution of BSc ICD algorithm with nominal and attack parameters on atrial (A), ventricular (V), and shock EGMs from arrhythmia 10. Markers are: VF – sensed ventricular fibrillation, VT – tachycardia, and VS – normal rate. Intervals are in milliseconds. See text for a detailed explanation.



(a) VT. V. score: -0.0032 (b) SVT. V. score: 0.0179

Figure 6: Pareto fronts for condition-agnostic reprogramming attacks. Legend is as in Figure 4.

reprogrammed VT threshold (marker 4). A new VT duration can start right away, but therapy is prevented due to the long VTdur.

Condition-agnostic attacks. Pareto fronts for the condition-agnostic attacks on VT and SVT, hereafter referred to as VT attack and SVT attack, are shown in Figure 6. The corresponding parameters are available in Tables 22 and 23 of [11]. These attacks attain very good validation scores, comparable to the condition-specific case, suggesting that our method can generalize well also when trained with heterogeneous arrhythmias. The Pareto front for the VT attack has a similar profile to the condition-specific ones: the effectiveness is poor for parameter distance below 5, it has a sharp increase between distance 5 and 10, growing slowly after that up to reaching 100% success at distance 16. On the other hand, the parameters for the SVT attack reach a maximum effectiveness of 49% at distance 18, compatibly with the fact that condition-specific attacks are reasonably successful only for a subset of SVT arrhythmias.

REFERENCES

[1] C. W. Barrett, R. Sebastiani, S. A. Seshia, and C. Tinelli. 2009. Satisfiability Modulo Theories. *Handbook of satisfiability* 185 (2009), 825–885.

[2] A. Biere et al. 1999. Symbolic model checking without BDDs. In *Tools and Algorithms for the Construction and Analysis of Systems*. 193–207.

[3] N. Bjørner, A. D. Phan, and L. Fleckenstein. 2015. *νZ*-An Optimizing SMT Solver. In *TACAS*, Vol. 15. 194–199.

[4] Boston Scientific Corporation. 2017. Implantable Cardioverter Defibrillator, reference guide (part number: 359407-003). (2017).

[5] Electrogram. 2018. Ann Arbor Electrogram Libraries. (2018). http://electrogram.com/

[6] Food and Drug Administration. 2017. Implantable Cardiac Pacemakers by Abbott: Safety Communication. (2017). https://www.fda.gov/safety/medwatch/safetyinformation/safetyalertsforhumanmedicalproducts/ucm573854.htm

[7] D. Halperin et al. 2008. Pacemakers and implantable cardiac defibrillators: Software radio attacks and zero-power defenses. In *IEEE Security and Privacy Symposium*. 129–142.

[8] Z. Jiang et al. 2016. In-silico pre-clinical trials for implantable cardioverter defibrillators. In *EMBC*. IEEE, 169–172.

[9] A. J. Moss et al. 2012. Reduction in inappropriate therapy and mortality through ICD programming. *New England Journal of Medicine* 367, 24 (2012), 2275–2283.

[10] N. Paoletti et al. 2018. Synthesizing Stealthy Reprogramming Attacks on Cardiac Devices. In *International Conference on Cyber-Physical Systems (ICCPS)*.

[11] N. Paoletti et al. 2018. Synthesizing Stealthy Reprogramming Attacks on Cardiac Devices. *CoRR* abs/1810.03808 (2018).

[12] A. Peterson. 2013. Yes, terrorists could have hacked Dick Cheney's heart. *Washington Post* (2013).

[13] B. Rios and J. Butts. 2018. Understanding and Exploiting Implanted Medical Devices. Black Hat USA conference. (2018).

[14] F. Xu et al. 2011. IMDGuard: Securing implantable medical devices with the external wearable guardian. In *IEEE Infocom*. 1862–1870.

[15] N. Zanker et al. 2016. Tachycardia detection in ICDs by Boston Scientific. *Herzschrittmachertherapie+ Elektrophysiologie* 27, 3 (2016), 186–192.