Chapter No .9

IVHM Assessment Metrics

"What's measured, improves"

— Peter F. Drucker

1.1 Measuring IVHM Effectiveness

Metrics let us quantify the effectiveness of a proposed solution in a standardized manner to aid in decision making and/or process improvement. There are several stakeholders of IVHM who have different user objectives and IVHM performance must be measured to assess effectiveness towards fulfillment of their respective functional requirements such as identified in (Jennions, 2011). Naturally, the development and use of metrics is largely guided by various user objectives of these stakeholders. System level requirements can be generally grouped under *safety, cost, and performance* categories. The goal of IVHM system is understood to be able to maintain system performance by ensuring safety and minimizing additional costs through improvements in reliability and availability. Therefore, various IVHM scenarios are largely divided into two top level goals from stakeholder perspectives - (1) *maintenance management*, and (2) *contingency management*, which, otherwise, are also sometimes viewed as relating to (1) *cost management*, and (2*) safety management*. Therefore, all stakeholders of IVHM fall under either Cost or Safety categories. As pointed out in (Wheeler *et al.*, 2009), depending upon the role of various stakeholders in the IVHM lifecycle these stakeholders and their objectives are further categorized into three broad categories in: (1) Operations (program manager, plant/fleet manager, mission

operator, command and control, and maintainer, owner, insurer), (2) Regulatory (policy makers), and (3) Engineering (designer, researcher, OEMs). Several metrics have been designed that cater to these different user groups. However, the discussion here will be limited to metrics relevant to engineering development and operations of IVHM.

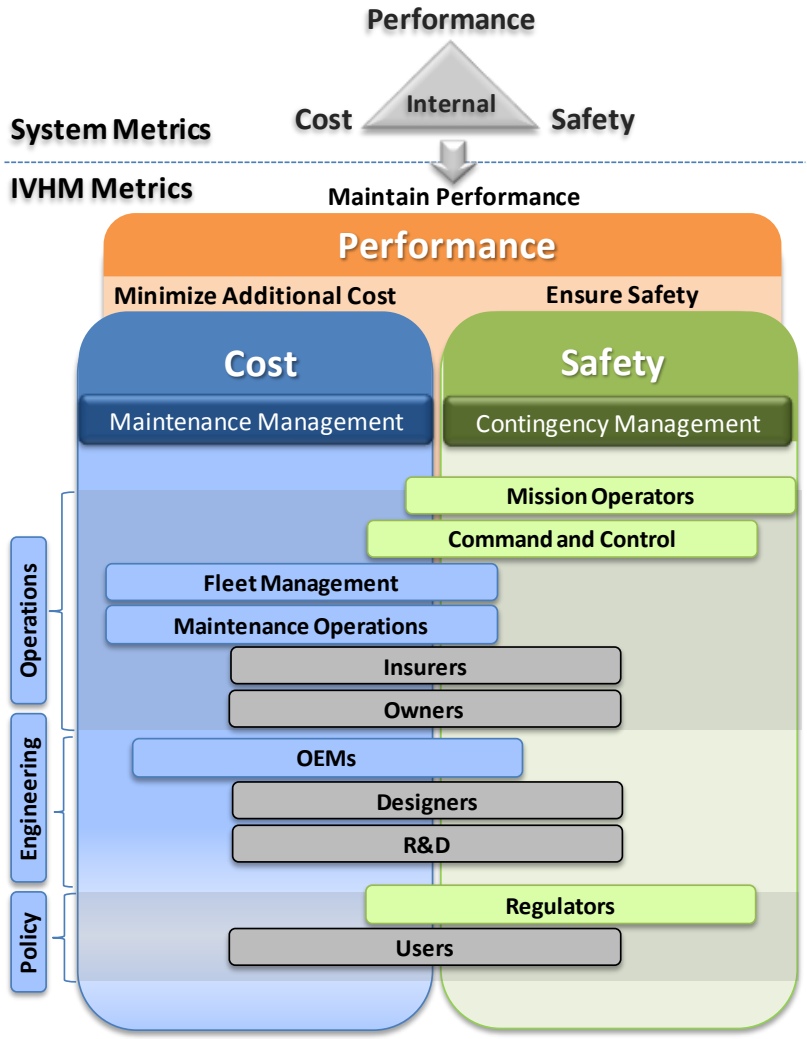


Figure 1.1 Various IVHM stakeholders and their user objectives mapped on to high level goals of cost and safety.

1.2 IVHM Metrics

As identified in the previous section, metrics quantify the attributes expected from the implemented IVHM system and, therefore, vary for different stakeholders. In this section we

discuss these different types of metrics and list some that are commonly found to be useful in the IVHM community. First, some of the high level metrics that evaluate the overall performance of an IVHM system are presented (Dibsdale, 2012). These metrics, as presented in Table 1.1, directly reflect the interests of various stakeholders of IVHM.

Table 1.1 High Level IVHM Effectiveness Metrics.

	Metric	Description/Usage
Effectiveness	Overall Value Added	It is a general metric that defines IVHM value based on a general set of individual metrics measuring effectiveness of key factors for a stakeholder, e.g. recovery time saved (gain in availability) by using IVHM and hence the avoidance of cost due to downtime. Value can also be established by quantifying improvements in traditional metrics used for system reliability as discussed in Section 1.2.1.
Design	Failure Coverage	It is specified as the ratio of failures covered by IVHM to total number of significant failures that are candidates for IVHM
	System Coverage	It is specified as the proportion of significant components/subsystems covered by IVHM. It requires a system to be partitioned into several independent modules with significant criticality.
Cost	IVHM Implementation & Operational Costs	This measure quantifies the time and cost of installing an IVHM system on a vehicle. Installation costs include labor costs, costs for commissioning and training, licensing costs, cost of hardware and software etc. This in turn also includes the costs of computational complexity as reflected in hardware costs and power requirements (see Section 1.2.2). Furthermore, These measures can be weighed against potential benefits of IVHM as discussed in Section 1.2.3.
IVHM Function	Warning Times	This is the class of measures that quantifies the efficacy in terms of timeliness of IVHM output towards avoiding dire consequences. Warning times should long enough to accommodate: - Amount of time necessary to avoid secondary damage from a primary failure mode - Amount of time necessary to plan and act upon remedial actions - Amount of time necessary to carry out planned outages and minimize spare holdings
	Diagnostic & Prognostic Performance	The overall performance of IVHM heavily depends on the performance (accuracy, precision, timeliness, etc.) of underlying diagnostic and prognostic algorithms as discussed in Section 1.2.4.

1.2.1 Reliability Analysis Metrics

Reliability analysis stems from statistical aggregates from historical data or carefully designed experiments to collect statistics about failure rates of certain components, which can then be used to assess overall system reliability, design tolerances, and indirectly predict system life. While not

a focus of this chapter, it is important to mention that reliability analysis is a very mature field and much progress has been made in its application to the optimization of aftermarket operations. Reliability centered maintenance (RCM) is the basis for much of how assets are maintained today and it is important to draw a contrast between RCM and IVHM in the econtect of performance assessment.

In RCM component failure is analyzed in great detail on the basis of experimental and field data to determine inspection, repair, and replacement intervals for all categories of components. The techniques of RCM form the basis for all aerospace maintenance practices, as embodied in MSG-3 guidelines, which are the way FAA recommends operators maintain their transport aircraft (FAA, 1997). However, much of the reliability analysis is based on past data (historical usage or lab testing), which is where IVHM differentiates itself being focused on condition based health management. Use of condition based techniques calls for different metrics to assess IVHM effectiveness, which are the focus of this chapter. However, it is noteworthy that reliability metrics can still be utilized to assess the additional value in terms of cost-benefit or the risks of using an IVHM system over the baseline system performance without IVHM since an effective IVHM system must improve these metrics as well. Therefore a brief introduction to these metrics is provided here with some representative examples. Such metrics are generally useful for *operators, maintainers, designers,* and *regulatory* stakeholders. More discussion on reliability based metrics can be found in (IEEE, 2002). Broadly speaking there are two types of reliability metrics as described below (Wood, 2001).

Constant Rate Reliability Metrics

Constant Rate Reliability metrics are commonly referred to as Mean Life metrics. They represent a good approximation of the flat region of the reliability bathtub curve (see Figure 1.2).

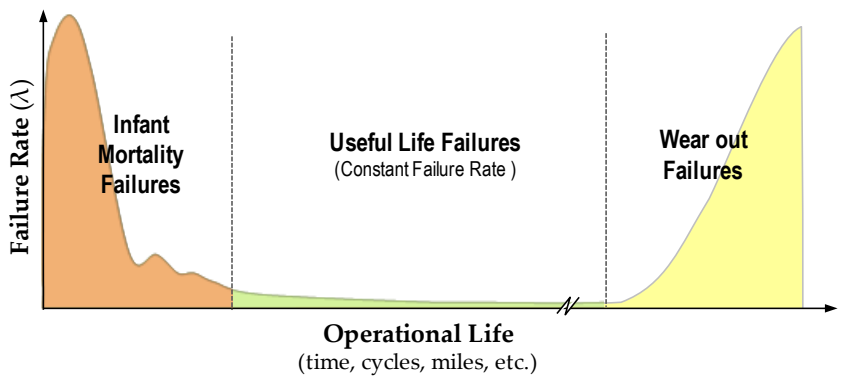


Figure 1.2 Reliability bath tub curve.

Generally expressed as MTB*xx* (Mean Time Between *xx*), where *xx* is the domain (industry) specific quantity of interest such as number of failures. MTBF (Mean time between failures) usually assumes an exponential distribution, which makes them an equivalent constant rate (λ) metrics where the rates are either measured from field data or computed from aging experiments. It expresses observed failure rate computed as *total failures divided by total population operating time.* MTB*xx* metrics are also sometimes customized by the reference scales involved in any application such as number of cycles (MCBF), or distance in miles (MMBF) for automotives. Some other variations of these metrics are MTBSC (Mean Time Between Service Call) and MTBMA (Mean Time Between Maintenance Action) that express perception of failure rates, or MTBWC (Mean Time Between Warranty Claim) that can be used for pricing warranties or setting warranty reserves. These metrics can be expressed as failures per hour, failures per year, or Annualized Failure Rate (AFR). Another variation of these metrics is Part Return/Repair Rate, expressed as total parts replaced divided by total population operating time. It is an indirect measure of failure rate when failure analysis is not available. Likewise, other metrics such as MTBF/MTBUR ratio are used that measures the ratio between how long a component is expected to last based on reliability estimates and how long it is actually used before replacing it. IVHM should enable the reduction of this ratio by allowing components to be used longer (Teal & Larsen, 2003). Therefore, although these metrics have been traditionally used in the field of

reliability, they can still be used to track improvements due to IVHM system by quantifying the before and after difference.

Probability of Success Metrics

Expressing a probability that a system performs a required function under stated condition for a stated period of time is useful when systems do not show constant failure rates and hence specifying mean times is not sufficient. Some examples of such metrics include Probability of Mission Success/Failure, Probability of Loss of Control, etc. Traditional methods normally finesse this by adding margins of safety, but this is not cost effective in the long run. Alternatively, measuring the percentage of population that survives a specific duration is also a useful measure of performance in some cases. These metrics are usually duration dependent, i.e. depend on the length of the mission. These are often specified in terms of percentiles of the distributions. For instance Lx-Life in mechanical systems specifies the number of hours after which at least $x\%$ of the population would have failed. Some metrics in the automotive industry are specified along with confidence levels, e.g. R96C90 meaning 96% Reliability with 90% Confidence. These metrics can be used to assess the effectiveness of IVHM system by quantifying improvement in both reliability and confidence about failure free operation during planning and decision making based on health information.

1.2.2 Computational Performance Metrics

With advancements and sophistication of vehicle systems' IVHM approaches, resoners and models are growing in complexity, therefore demanding high computational efficiency and performance as key engineering requirements. There is often a competing relationship between time and computational resources available versus the desired fidelity and performance expected from an IVHM system. Computational performance metrics help specify the hardware requirements or otherwise specify constraints within which the software must work, yet still

satisfy algorithmic performance requirements, hence influencing the engineering design and implementation costs. Note that some of the following metrics (Table 1.2) depend on the hardware chosen to implement the IVHM function.

Table 1.2 Computational Performance Metrics.

Metric	Description	Unit
Computational Complexity	Generally described using "Big O" notation (O(n),O(n^2), O(n^3), O(n^m), etc.) it describes the amount of time needed for an algorithm to run (independent of a software and hardware implementation), as a function of the size of the input.	Millisec, sec
CPU Time	Measures the time a CPU spends executing software to specify combined performance of an algorithm, its software implementation, and the hardware on which it is run.	Millisec, sec
Elapsed Time	Measures total CPU time including wait times for input/output (I/O)	millisec, sec
Memory Size	Specifies memory (RAM, Flash and/or disk space) requirements. It is a key metric for embedded applications, such as for flight computer in aerospace that usually have very limited space available	Bits, Bytes
Data Rate	Measures how many samples per time unit the software can handle for a given hardware configuration. This is particularly important in assessing software capability to handle real-time application	Bits/sec (bps)

1.2.3 Cost Benefit Metrics

A common use of IVHM performance data is to perform a cost-benefit analysis (CBA) of implementing an IVHM system. Clearly, the cost-benefit equation is significantly affected by how effective the IVHM system will be. Therefore, the cost-benefit metrics try to quantify the performance of IVHM such that they can be easily converted into monetary terms to be used in a cost-benefit analysis. Some examples of such metrics are provided below.

Table 1.3 Cost-Benefit Assessment Metrics.

Metric	Description
Return on Investment	A return on investment on health management is generally realized as cost avoidance through timely maintenance and possibly prevention of downtime

(ROI) (Sandborn, 2012)	or lost hardware over the life of the system. $ROI = \frac{C_{\mathrm{u}} - C_{\mathrm{HM}}}{I_{\mathrm{HM}}}$, where C_{u} is life-cycle cost of the system when managed using unscheduled maintenance C_{HM} is life-cycle cost of the system when managed using HM approach I_{HM} is investment in implementing and managing HM, which is further obtained as the sum of C_{NRE} (non-recurring HM costs), C_{REC} (recurring HM costs), and C_{INF} (HM infrastructure cost).
Net Present Value (NPV) (Michael J. Roemer, 2012)	NPV is used to project the costs for a specified time period in the future. Present value of the costs expected to be incurred in future are estimated by incorporating the discount rates. In the IVHM context it is calculated as $NPV = C_{F\|\neg HM} - C_{F\|HM} - C_{HM}$, where $C_{F\|\neg HM}$ is present value of the total expected consequential cost of a failure without doing maintenance $C_{F\|HM}$ is the present value of total expected consequential cost of a failure after doing maintenance C_{HM} is present value of the cost of maintenance
Payback Period	Payback period is the time for which a system owner would have to wait until the benefits realized from an IVHM system break even with the value adjusted costs. It is only after the payback period net potential benefits may be realized. A short payback period is preferred.
Technical Value (Dzakowic & Valentine, 2006; Vachtsevanos *et al.*, 2006)	Costs and benefits associated with the low level algorithmic performance can be aggregated to be used in a cost-benefit analysis. The benefits achieved through accurate detection, fault isolation and prediction of critical failure modes are weighed against the costs associated with false alarms, inaccurate diagnoses/prognoses, and resource requirements of implementing and operating specific techniques $TV = P_f(D \cdot \alpha + I \cdot \beta) - (1 - P_f)(P_D \cdot \varphi + P_i \cdot \theta)$ P_f : Probability of a failure mode D : overall detection confidence metric α : savings realized by fault detection in advance I : overall isolation confidence metric β : savings realized by isolating a fault in advance P_D : false positive detection metric ϕ : cost of false positive detection P_i : false positive isolation metric θ : cost of false positive isolation
Total Value	Describes the value of a PHM technology in a particular application as the summation of the benefits it provides over all the failure modes that it can diagnose or prognose less the implementation cost, operation and maintenance cost, and consequential cost of incorrect assessments. $V_{TOTAL} = \sum_{i=1}^{FM} TV_i - A - O - (1 - P_c) \cdot \delta$ *A*: acquisition and implementation costs *O*: lifecycle operational and maintenance cost P_c: computer resource requirements *δ*: cost of computer systems

1.2.4 Algorithm Performance Metrics

Successful implementation of IVHM technologies heavily relies on the intelligent algorithms that interpret health monitoring data and reason about the health of the vehicle systems. Therefore, a lot of effort has been put into evaluating and refining these algorithms in the recent past. However, the notion of using fairly sophisticated algorithms to implement PHM concepts have been around in civil aviation since at least the late seventies when engine and aircraft data became available via ACARS transmission to the ground. A number of metrics often customized for specific applications are available in the literature, a sampling of which is provided in this section. This section is further divided into discussions specific to performance evaluation of diagnosis and prognosis, which are the core algorithmic components at the heart of IVHM systems.

Diagnostic Performance Metrics

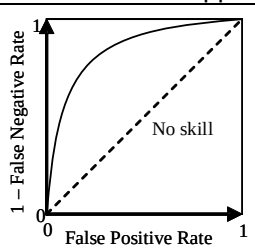
Diagnostics is one of the key elements in implementing an IVHM function, which strives to accomplish an early fault detection, fault isolation (pinpointing the root cause), and fault identification (estimating the severity of the isolated fault). There are several flavors to diagnostic activities such as on-board vs. off-board assessments, real-time vs. long term trending, component level vs. system level reasoning, etc. However, the key point to note here is that despite the granularity and approach of diagnostic methods used the performance assessment methods share common attributes and, therefore, can be largely divided into four categories of metrics for diagnosis.

1. *Diagnostic Performance Statistics*: Performance statistics of system level diagnostic assessments collected over time where multiple fault modes may be present on a variety of subsystems or components.
2. *Accuracy and Precision*: Accuracy and precision in detecting faults and estimating severity that determine the confidence levels in diagnostic assessment due to presence of noise, variability, and other uncertainties in the system.
3. *Robustness of Detection*: Robustness or sensitivity to various factors (environmental, operational, design, etc.) that could change diagnostic outcomes.
4. *Response Time*: Timeliness in assessments while maintaining accuracy and precision to facilitate responsive and useful post diagnostic reasoning actions.

A list of representative metrics for each of these four types of diagnostic metrics is provided in Table 1.4 with a brief description and a list of original sources where these metrics have been reported.

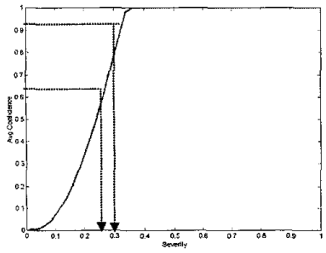
Table 1.4 Diagnostic Performance Metrics.

<table>
<tr><th>Metric</th><th>Explanation and Usage</th></tr>
<tr><td colspan="2">Diagnostic Performance Statistics (Rate) Metrics</td></tr>
<tr><td>Confusion Matrix</td><td>
<table>
<tr><th rowspan="3">Predicted System S ate</th><th colspan="3">Actual System State</th></tr>
<tr><th></th><th>Non-faulty</th><th>Faulty</th></tr>
<tr><th>Non-faulty</th><td>a</td><td>b</td></tr>
<tr><th></th><th>Faulty</th><td>c</td><td>d</td></tr>
</table>
A confusion matrix is often constructed to document outcomes (correct or incorrect) and subsequently analyze performance of a diagnostic system. Once populated this matrix is then used to derive a variety of performance indicators commonly used in the industry. Performance is often specified as absolute correct or incorrect outcomes, or weighted to express as rates over a period of time. Some examples are given below.
</td></tr>
<tr><td>Detection Accuracy</td><td>The ratio of correctly classified cases (i.e., fault detected when the system was actually faulty, and no fault detected when the system was non-faulty) to the total number of scenarios to assess overall accuracy of the system.
$\mathrm{DA} = \dfrac{a + d}{a + b + c + d}$</td></tr>
</table>

<table>
<tr><td>Fault Detection Rate</td><td>The ratio of cases where a fault is detected out of the total number of scenarios when the system was actually faulty.
$\mathrm{FDR} = \dfrac{c+d}{b+d}$</td></tr>
<tr><td>False Negative Rate</td><td>The ratio of the number of times a fault is missed (goes undetected) to the number of scenarios when the system was actually faulty. Low false negatives are preferred to avoid downtime due to missed failures.
$\mathrm{FNR} = \dfrac{b}{b+d}$</td></tr>
<tr><td>False Positive Rate</td><td>The ratio of the number of times a fault is detected to the number of scenarios when the system was actually non-faulty. Low false positive rates are desired to avoid unnecessary maintenance actions.
$\mathrm{FPR} = \dfrac{c}{a+c}$</td></tr>
<tr><td>False Positive Confidence Metric (M. J. Roemer et al., 2005)</td><td>Measures the frequency and upper confidence limit associated with false anomaly detection by a diagnostic tool. Calculation of the false confidence metric is based on the false positive function that is stated in the following equation:
$F(c) = n(c)/N$
where,
$n(c)$ – number of false positive detection events with confidence > c,
N = number of opportunities to detect a normal operating condition</td></tr>
<tr><td>Receiver Operating Characteristic (ROC)</td><td> 

ROC gives a comprehensive overview of the tradeoff between false positives (FP) and false negatives (FN). The ideal curve would have zero false positives and zero false negatives. Such a curve cannot realistically be achieved for real-world problems and therefore, a balance must be achieved based on cost and safety priorities in a given application.</td></tr>
<tr><td>Power of Test</td><td>Power of test estimates the overall bias of a diagnoser towards FPs or FNs over the entire range of the ROC curve. It is computed by calculating the area under the ROC curve, which represents the probability that a diagnoser favors false positives over false negatives in general.</td></tr>
<tr><td>Isolation Performance Matriix</td><td>Similar to confusion matrix above a higher fidelity confusion matrix may be constructed for a system that not only detects faults but also isolates various fault modes. Absolute or rate metrics can be constructed for isolation performance as well. In the matrix shown below all elements are weighted by total number of cases, which can also be seen as probabilities if large number of cases is observed.
<table>
<tr><td rowspan="6">Predicted S ate</td><td colspan="5">Actual State</td></tr>
<tr><td>Component</td><td>Mode$_1$</td><td>Mode$_2$</td><td>…</td><td>Mode$_n$</td></tr>
<tr><td>Mode$_1$</td><td>p_{11}</td><td>p_{21}</td><td>…</td><td>p_{n1}</td></tr>
<tr><td>Mode$_2$</td><td>p_{12}</td><td>p_{22}</td><td>…</td><td>p_{n2}</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>Mode$_n$</td><td>p_{1n}</td><td>p_{2n}</td><td>…</td><td>p_{nn}</td></tr>
</table></td></tr>
<tr><td>Isolation Classification Rate</td><td>The rate of correct classification. This metric equals to the sum of elements along the diagonal.</td></tr>
</table>

	$\text{ICR} = \sum_{i=1}^{n} p_{ii}$
Isolation Misclassification Rate	The rate of misclassification computed as the sum of all off-diagonal elements. $\text{IMR} = \sum p_{ij} \ \forall i \neq j$
Size of Isolation Set	Often times a diagnostic system does not converge on a single isolated fault but a number of possible faults with corresponding probabilities. In such cases the size of isolation set becomes a metric. A smaller size implies less ambiguity and hence preferable.
Accuracy and Precision Measures	
Error	Error defines the basic notion of deviation from desired output. Applications where estimates of fault dimension are computed, such as crack length or bearing spall size, typically use estimation error as a metric of diagnostic performance. $\Delta = Measured\ Ground\ Truth - Diagnostic\ Estimate$ Most accuracy based metrics are derived directly or indirectly from error as defined below.
Error Based Metrics (Hyndman & Koehler, 2006)	Several metrics are defined to aggregate errors (Δ) into absolute or weighted measures. Aggregates can be computed over multiple systems (*l*) (e.g. in a fleet scenario) or over multiple test cases (*i*) for a single system over time. Some representative examples are given here, but many more metrics exist that can be useful in different cases. Mean Squared Error (MSE) $MSE(i) = \frac{1}{L}\sum_{l=1}^{L}\Delta^l(i)^2$ Mean Absolute Error (MAE) $MAE(i) = \frac{1}{L}\sum_{l=1}^{L}\left\|\Delta^l(i)\right\|$ Root Mean Squared Percentage Error (RMSPE) $RMSPE(i) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left\|\frac{100\Delta^l(i)}{r_*^l(i)}\right\|^2}$ Mean Absolute percentage Error (MAPE) $MAPE(i) = \frac{1}{L}\sum_{l=1}^{L}\left\|\frac{100\Delta^l(i)}{r_*^l(i)}\right\|$
Precision Based Metrics (Hoaglin *et al.*, 1983)	Precision in estimation is estimated by computing spread of the error around the true value. Once again, there are several variants of dispersion measures, some of which are mentioned here to exemplify. Users pick their favorite ones based on desired robustness and acceptable computational complexity in computing these measures.
Sample Standard Deviation	Sample standard deviation measures the dispersion/spread of the error with respect to the sample mean of the error. Technically this metric is restricted to the assumption of normal distribution of the error, which may not always be a correct assumption. $S(i) = \sqrt{\frac{\sum_{l=1}^{n}\left(\Delta^l(i) - M\right)^2}{n-1}}$ where *M* is the sample mean of the error
Mean absolute deviation from the sample median (MAD)	A more robust estimator of the dispersion/spread of the estimation error is MAD. $MAD(i) = \frac{1}{n}\sum_{l=1}^{n}\left\|\Delta^l(i) - M\right\|$ where $M = \underset{l}{median}(\Delta^l(i))$ and median is the $\frac{\text{n}+1}{2}\text{th}$ order statistic

Robustness Metrics In presence of noise and other uncertainties diagnostic algorithms can be tuned to adjust detection thresholds and confidence levels per user requirements. Therefore, confidence level and detection thresholds can be seen as tuning parameters in specifying risk acceptance.	
Detection Thresholds (Orsagh *et al.*, 2002; M. J. Roemer *et al.*, 2005)	 This metric quantifies the ability of a diagnostic algorithm in terms of the minimum ground truth severity an algorithm can detect with a designated confidence level and is represented on the detection success function as shown in the figure. Confidence of 67% and 95% (corresponding to one and two standard deviations of a normal distribution) are used to calculate the detection threshold metric score as: $Detection\ Threshold = 1 - S(c)$ where S(c) = ground truth severity at confidence = c
Detection Sensitivity (Bartys *et al.*, 2006; Orsagh *et al.*, 2002; M. J. Roemer *et al.*, 2005)	A diagnostic system should detect anomalies over the full range of operating conditions such as loads, speeds, noise, duty cycles, etc. This metric measures the difference between the success functions of a diagnostic tool under two levels of conditions. $Condition\ Sensitivity = 1 - \sqrt{\int_0^1 \left(C_1(s) - C_2(s)\right)^2 ds}$ where, $C_1(s)$ = success function at condition 1 $C_2(s)$ = success function at condition 2 s = severity
Detection Stability (Orsagh *et al.*, 2002; M. J. Roemer *et al.*, 2005)	This metric quantifies the fluctuation in confidence values that occur over the fault transition by integrating the peak to peak difference at each point on the success function as stated here. Basically, a confidence level that fluctuates wildly is difficult to interpret and therefore undesirable. $Stability = 1 - \sqrt{\int_0^1 \left(C_H(s) - C_L(s)\right)^2 ds}$ where, $C_H(s)$ = maximum value of the success function at severity s, $C_L(s)$ = minimum value of the success function at severity s, s = severity
Detection Stability Factor (Kurtoglu *et al.*, 2008)	In presence of noise a diagnostic output can fluctuate especially when fault is less severe. In such cases decisions can be made more robust by quantifying the level of stability of the detection signal, specified as duration for which the outcome doesn't change as a percentage of total time elapsed since the first detection signal. Thresholds may be set for detection stability factors for a corrective action to be initiated.
Response Time Metrics It is critical in many applications that diagnostics is quick enough to enable an actionable decision in a timely manner. In such cases once can measure time taken to carry out a particular diagnostic function (Bartys *et al.*, 2006; Kurtoglu *et al.*, 2009; Poll *et al.*, 2011). Such metrics are generally useful during design and testing while tuning the diagnostic system and requires the knowledge about true fault appearance time, which is never available in real applications. Faults can be injected to a system during testing and diagnostic response times may be measured.	
Time to Detect	The period of time from the beginning of fault injection to the moment when fault was detected for the first time.

Time to isolate	The period of time from the beginning of the fault injection to when the isolation convergence set, i.e., final diagnostic output of an algorithm, is obtained.

Prognostics Metrics

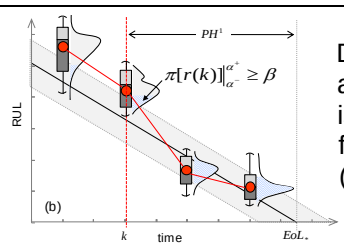
Prognostics is the process of predicting a system's remaining useful life by estimating the time at which the system will no longer perform its intended function within desired specifications. It is one of the advanced functions that can significantly improve IVHM performance by warning the system operators of impending failures to avoid downtimes, avoid secondary damages, and allowing sufficient time to plan and execute required MROs. However, predicting the future with limited information is not a trivial task and leads to uncertainties in the predictions. Furthermore, assessing performance of prognostics algorithms is an acausal problem as it would require the knowledge of the true End of Life (EoL) time, which is clearly not possible, since this will occur sometime in the future. Therefore, most prognostics performance metrics have been developed for offline performance assessment, and relatively few exist for online applications.

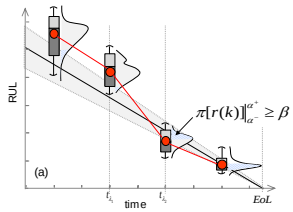
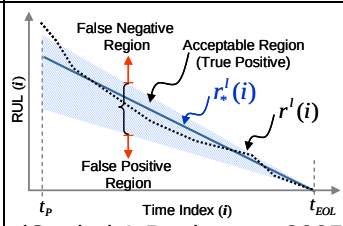
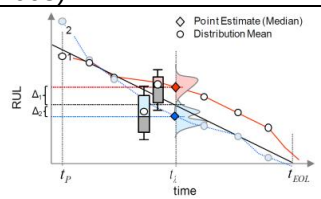
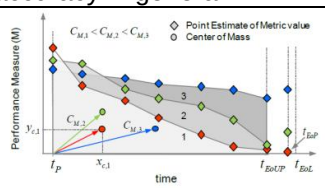
In order to use metrics suggested in Table 1.5 some characteristics in prognostics implementation are assumed. A prognostic routine may predict system/component health from the very beginning or may be triggered by an event such as detection or isolation of a critical fault. Upon trigger the prognostic algorithm starts predicting the Remaining Useful Life (RUL) according to a predefined End of Life (EoL) criteria or failure threshold, such as maximum allowable crack length, stiffness loss in a structure, etc. Furthermore, prognostic estimates should be continuously updated as more information about system's health becomes available through time. Nominally, a prognostic algorithm is expected to produce improved predictions as time progresses and this improvement is an important metric to track as well. Last but not the least, predicting into the future inherently involves uncertainties which must be accounted for. Several approaches have used probability distributions of RUL to represent such uncertainties and metrics should take

these distributions into considerations wherever needed. These metrics can be simplified as needed to evaluate point predictions instead of distributions if that happens to be the preferred prognostic output for various reasons, however a more general form of the metrics is presented here.

It is noteworthy that most of these metrics combine the intent of measuring accuracy and precision together with time tracking of performance by plotting the predicted RUL vs. operational time. In an ideal case accurate predictions must fall on the -45° straight line representing ground truth, but that's rarely the case. Any deviation from the ground truth can be quantified or bounded by acceptable error bounds (α) around ground truth line. RUL distributions can be represented as error bars showing mean, median, or percentiles as needed. Further, for better estimates one could estimate the total probability mass of the distributions that fall within desired accuracy bands and compare with a threshold specified by β to classify a prediction into good or bad categories. More details on these metrics and use cases can be found in (Saxena *et al.*, 2010).

Table 1.5 Prognostic Performance Metrics.

Offline Prognostic Performance Metrics	
Metric	**Description and Usage**
Prognostic Horizon	 Determines the maximum advance warning an algorithm can provide with desired confidence. It is defined as the maximum predicted RUL (or the first prediction) that falls within desired accuracy (α) bounds. An algorithm with larger PH is preferred. $PH = t_{EoL} - t_{prediction} \mid first\ prediction \in desired\ accuracy\ (\alpha)$

α-λ Performance		Characterizes how well an algorithm's performance (accuracy and precision) improves as EoL approaches. Improvement is measured agaainst desired accuracy level (*α*) expressed as a percentage of RUL at any specified time (λ). $\alpha-\lambda\ Accuracy=\begin{cases}1 & \text{if a prediction falls inside the } \alpha \text{ a}\\ 0 & \text{otherwise}\end{cases}$
Prognostic False Alarm Rate		This metric is an extension from classical false positive and false negative metrics used for diagnostics performance where an early prediction is classified as false positive and a late prediction as false negative. Therefore, the α accuracy bounds mentioned above define FP and FN regions on an RUL vs. time graph. (Goebel & Bonissone, 2005; Leao *et al.*, 2008; Luna, 2009; Saxena *et al.*, 2008)
Relative Accuracy		Quantifies how accurate an algorithm is at any given time (λ) relative to the EoL. It is computed as RUL error normalized by actual RUL for any given time. It captures the notion that accuracy of prediction becomes more critical as EoL approaches. Further, a cumulative relative accuracy may be computed by aggregating the RAs at different time instants to specify algorithm's accuracy in general.
Convergence Rate		Quantifies the rate of convergence of prognostic performance improvement as predictions are updated with time. Convergence can be computed for any metric of interest such as accuracy, precision, etc. This metric quantifies the convergence rate by computing the center of mass of the area under the predicted trajectories for respective algorithms. A faster convergence rate indicates faster performance improvement and hence more time available for making actionable decisions for IVHM.
Sensitivity (Vachtsevanos *et al.*, 2006)	Sensitivity is a measure of how sensitive a prognostic algorithm is to input changes or external disturbances. Can be assessed against any performance metric of interest. Δ*M* is the distance measure between two successive outputs for metric M's value and Δ_{input} is a distance measure between two successive inputs, e.g. failure threshold, noise level, available sensor set, sampling rate, etc. Sensitivity may be aggregated over similar sustems (units under test (*l*)). $S(i)=\frac{1}{L}\sum_{l=1}^{L}\left\{\frac{\Delta M^l(i)}{\Delta_{input}}\right\}$	
Online Prognostic Performance Metrics		
RUL Online Precision Index (RUL-OPI) (Orchard *et al.*, 2009)	RUL-OPI quantifies and tracks the precision of predicted RUL distributions by quantifying the length of 95%confidence bounds ($\mathrm{CI}(i)$) normalized by the predicted RUL ($r(i)$) at any given time instant. An algorithm with a high index (close to 1) is preferred, which indicates high precision or narrow confidence bounds.	

	$I(i) = e^{-\left(\frac{\sup\{CI(i)\} - \inf\{CI(i)\}}{r(i)}\right)}$
Dynamic Standard Deviation (DStd) (Olivares *et al.*, 2013)	DStd quantifies the stability of predictions within a time window (Δ). Variance between individual predictions made within the time window is computed. The metric is normalized to a range [0,1] using the logistic function (φ) for easy comparisons. $DStd = \varphi\left(Var\left(E\left\{EOL \mid y_{1:j}\right\}\right)_{j\in\Delta}\right)$
Critical-α Performance Measure (Olivares *et al.*, 2013)	Looking from the perspective of actionable decision making this measure computes the critical percentile (α) of an RUL distribution that would define a Just-in-time-point (JITP) for that application. JITP must always occur before actual failure and hence the value of this metric lies in interval (0,0.5] and should be maximized to avoid unnecessary conservatism in decision making. $\alpha_{crit} = \arg\max_{\alpha}\left\{JITP_{\alpha\%}(k_{pred}) \le EOL\right\}; \quad \forall k_{pred} \in [1, EOL]$

1.3 The Challenge of Data Availability

Performance of an IVHM system can be evaluated using various metrics described in the previous section, however, it must be noted that the usefulness of the output from these metrics depends on the quality of data on which these metrics are applied. In order to perform validation and verification (V&V) of specific PHM algorithms or technologies, various types of data sources with different levels of fidelity can be utilized. From the perspective of a PHM algorithm designer, the most desirable data sets would come from either seeded fault testing, accelerated mission testing or actual field data from the operational system where detailed loading profiles are known and selected parameters were monitored that are correlated with the progression of the failure. However, these types of data sets are typically expensive to obtain and hence represent less than 10% of the data that is used for performing a comprehensive V&V program. In order to supplement these more comprehensive data sets, subsystem or component rig tests can be performed that are more cost effective but limited in applicability and realism. Also, dynamic simulation models with different levels of fidelity can be used to simulate fault signatures and provide data at any desired sampling rate. There are obviously many benefits to the simulation-

based approach but many drawbacks as well. The Figure below details these pros and cons of these various data sources and their applicability to PHM V&V and metrics evaluation.

Table 1.6 Data Source Applicability to IVHM Metrics

Method	Description	Benefit	Drawback
Simulation Models and Fault Generation	Signals, noise and fault signatures simulated mathematically	Ideal for generalizing (varying) signals and noise; partially overcomes small sample problem	Data may not always be realistic; model may be expensive to develop to reliably simulate.
Component and LRU Fault and Failure Tests	Use data collected on component or subscale test rigs.	Real seeded fault, natural fault, and fault progression data available at little cost to program	Limited applicability for actual system fault observability.
Subsystem/System Fault Characterization Tests	Tests on stands using actual subsystem hardware	Somewhat realistic data; usually well controlled and documented with ground truth information.	Seeded faults may not be entirely realistic of natural fault.
Limited Vehicle Data Evaluation	Dedicated missions for normal and off-normal test and evaluation	Realistic data of normal and some performance faults.	Expensive; doesn't cover full range of potential conditions; no critical faults or progression.
Technology Maturation Field Program	Data collection and tests during vehicle service	Covers full range of vehicle mission, operational, and environmental conditions	Limited to opportunistic fault occurrence. Delays use and implementation in field

Hence it is argued that performance evaluations should consider the effects of data quality while interpreting the results and their applicability towards actual stakeholders' objectives. Metrics should be chosen carefully and an adequate thought must be given to what their outputs really represent and what is the extent of their validity towards V&V activities.

1.4 Conclusions

This chapter discusses various performance measures that are used to assess the effectiveness of an IVHM system. It is argued that metrics are developed to measure effectiveness against the user objectives of various stakeholders; therefore, whatever matters is what should be measured. Examples of overall IVHM effectiveness measures point towards broad categories of value and cost, which are further broken down into specific metrics. A special focus is given to algorithmic performance metrics since diagnostic and prognostic performance lie at the core of how effective

an IVHM system can be. Considering the wide variety of metrics available for assessing effectiveness it is suggested that this list be used as a representative set of basic metrics that can be further augmented with more customized versions that cater to a specific stakeholder's needs. Last but not the least, metrics convey message based on what data they are applied on. Therefore, it is not sufficient to just choose the right metrics but also to decide the right set of measurements and parameters that must be tracked. In the absence of real and/or adequate field data clever approaches are necessary that cover the gaps generated by unavailability of such data some of which can be implemented by choosing the right metrics.

References

Bartys, M., Patton, R., Syfert, M., Heras, S. d. l., & Quevedo, J. (2006). Introduction to the DAMADICS actuator FDI Benchmark Study. *Control Engineering Practice, 14*, 577-596.

Dibsdale, C. (2012). Asset Management and its Implications to IVHM. In I. K. jennions (Ed.), *Integrated Vehicle Health Management Business Case Theory and Practice* (Vol. 2, pp. 169). Warrendale, PA: SAE International.

Dzakowic, J. E., & Valentine, G. S. (2006). *Advanced Techniques For The Verification And Validation Of Prognostics & Health Management Capabilities.* Paper presented at the Machinery Failure Prevention Technologies Conference (MFPT 60), Virginia Beach, VA.

FAA. (1997). *Maintenance Review Board Procedures AC 121-22A*. US Department of Transportation.

Goebel, K., & Bonissone, P. (2005). *Prognostic Information Fusion for Constant Load Systems*. Paper presented at the 7th Annual Conference on Information Fusion.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding Robust and Exploratory Data Analysis*: John Wiley & Sons.

Hyndman, R. J., & Koehler, A. B. (2006). Another Look at Measures of Forecast

Accuracy. *International Journal of Forecasting, 22*(4), 679-688.

IEEE. (2002). IEEE Guide for Selecting and Using Reliability Predictions Based on IEEE 1413™ *1413.1 TM* (pp. 97): IEEE Standards Coordinating Committee 37.

Jennions, I. K. (2011). *Integrated Vehicle Health Management: Perspectives on an Emerging Field*. Warrendale, Pennsylvania USA: SAE International.

Kurtoglu, T., Mengshoel, O. J., & Poll, S. (2008). *A framework for Systematic Benchmarking of Monitoring and Diagnostic Systems.* Paper presented at the International Conference on Prognostics and Health Management, Denver, CO.

Kurtoglu, T., Narasimhan, S., Poll, S., Garcia, D., Kuhn, L., Kleer, J. d., . . . Feldman, A. (2009). *International Diagnosis Competition – DXC'09.* Paper presented at the 20th International Workshop on Principles of Diagnosis.

Leao, B. P., Yoneyama, T., Rocha, G. C., & Fitzgibbon, K. T. (2008). *Prognostics Performance Metrics and Their Relation to Requirements, Design, Verification and Cost-Benefit.* Paper presented at the International Conference on Prognostics and Health Management (PHM08), Denver CO.

Luna, J. J. (2009). *Metrics, models and scenarios for evaluating PHM effects on logistics support.* Paper presented at the Annual Conference of the Prognostics and Health Management Society, San Diego, CA.

Olivares, B. E., Muñoz, M. A. C., & Orchard, M. E. (2013). Particle-Filtering-Based Prognosis Framework for Energy Storage Devices With a Statistical Characterization of State-of-Health Regeneration Phenomena. *IEEE Transactions on Instrumentation and Measurement, 62*(2), 13.

Orchard, M. E., Tang, L., Goebel, K., & Vachtsevanos, G. (2009). *A Novel RSPF Approach to Prediction of High-Risk, Low-Probability Failure Events*. Paper presented at the Annual Conference of the Prognostics and Health Management Society (PHM09), San Diego, CA.

Orsagh, R. F., Roemer, M. J., Savage, C. J., & Lebold, M. (2002). *Development of Performance and Effectiveness Metrics for Gas Turbine Diagnostic Techniques.* Paper presented at the IEEE Aerospace Conference, Bigsky, MT.

Poll, S., Kleer, J. d., Abreu, R., Daigle, M., Feldman, A., Garcia, D., . . . Sweet, A. (2011). *Third International Diagnosis Competition – DXC'11.* Paper presented at the 22nd International Workshop on Principles of Diagnosis, 2011.

Roemer, M. J. (2012). Maintenance Cost Analysis - Leveraging Prognostic Information. In I. K. jennions (Ed.), *Integrated Vehicle Health Management Business Case Theory and Practice* (Vol. 2, pp. 169). Warrendale, PA: SAE International.

Roemer, M. J., Dzakowic, J., Orsagh, R. F., Byington, C. S., & G. Vachtsevanos, " in Proceedings of , 2005, paper 1344. (2005). *Validation and Verification of Prognostic and Health Management Technologies.* Paper presented at the IEEE Conference Aerospace, Bigsky, MT.

Sandborn, P. A. (2012). Making Business Case for Health Management - Return on Investment. In I. K. jennions (Ed.), *Integrated Vehicle Health Management Business Case Theory and Practice* (Vol. 2, pp. 169). Warrendale, PA: SAE International.

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008). *Metrics for Evaluating Performance of Prognostics Techniques*. Paper presented at the 1st International Conference on Prognostics and Health Management (PHM08), Denver, CO.

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Metrics for Offline Evaluation of Prognostic Performance. *International Journal of Prognostics and Health Management, 1*(1), 21.

Teal, C., & Larsen, B. (2003). *Technology Update II: Wire Systems Diagnostics & Prognostics*. Paper presented at the Digital Avionics Systems Conference.

Vachtsevanos, G., Lewis, F. L., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent Fault Diagnosis and Prognosis for Engineering Systems* (1st ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Wheeler, K. R., Kurtoglu, T., & Poll, S. (2009). *A Survey of Health Management User Objectives Related to Diagnostic and Prognostic Metrics*. Paper presented at the ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE), San Diego, CA.

Wood, A. P. (2001, 22 Jan 2001 - 25 Jan 2001). *Reliability-Metric Varieties and their Relationships.* Paper presented at the Reliability and Maintainability Symposium, 2001, Philadelphia, PA.

About the Author

This section, typically one to two pages, provides readers with information about your background and qualifications to write the book, and those of your co-authors if applicable. It is written in paragraph style (not a resume) and includes information about work experience, formal education degrees, professional certifications and affiliations, and other professional accomplishments such as awards and professional memberships. Along with a writeup, please provide a photo of yourself to accompany the text and to use in promoting the book.