Data Mining CMMSs: How to Convert Data into Knowledge

Larry Fennigkoh and D. Courtney Nanney

***Larry Fennigkoh**, PhD, PE, CCE, is a professor of biomedical engineering at the Milwaukee School of Engineering in Milwaukee, WI. Email: fennigko@msoe.edu*

***D. Courtney Nanney**, BSBME, CCE, CLSGB, is national quality manager of clinical engineering physical asset services at Catholic Health Initiatives in Louisville, KY. Email: courtneynanney@catholichealth.net*

Abstract

Although the healthcare technology management (HTM) community has decades of accumulated medical device–related maintenance data, little knowledge has been gleaned from these data. Finding and extracting such knowledge requires the use of the well-established, but admittedly somewhat foreign to HTM, application of inferential statistics. This article sought to provide a basic background on inferential statistics and describe a case study of their application, limitations, and proper interpretation. The research question associated with this case study involved examining the effects of ventilator preventive maintenance (PM) labor hours, age, and manufacturer on needed unscheduled corrective maintenance (CM) labor hours. The study sample included more than 21,000 combined PM inspections and CM work orders on 2,045 ventilators from 26 manufacturers during a five-year period (2012–16). A multiple regression analysis revealed that device age, manufacturer, and accumulated PM inspection labor hours all influenced the amount of CM labor significantly ($P < 0.001$). In essence, CM labor hours increased with increasing PM labor. However, and despite the statistical significance of these predictors, the regression analysis also indicated that ventilator age, manufacturer, and PM labor hours only explained approximately 16% of all variability in CM labor, with the remainder (84%) caused by other factors that were not included in the study. As such, the regression model obtained here is not suitable for predicting ventilator CM labor hours.

Since the introduction of personal computers in the early to mid-1980s, the healthcare technology management (HTM) community has faithfully and rigorously documented the results of its medical device repair and maintenance activities. Sophisticated computerized maintenance management systems (CMMSs) have made these tasks considerably easier, efficient, and more reliable. Most of these systems allow users to generate a variety of customizable reports and summary metrics on the performance of maintenance activities. The classic and management value of such reports and metrics helps HTM departments schedule device maintenance and summarize, track, and report the volume and, especially, costs associated with maintenance activities to both department administration and regulatory agencies.

Despite the managerial necessity and value of these summary reports and metrics, virtually all CMMSs are limited to the use of descriptive statistics: means or averages, ratios or percentages, and perhaps some measures of dispersion (standard deviation, variance, or ranges). Although these descriptive statistics are important and appropriate for summarizing maintenance activity and costs, they are unable and unsuitable for use in interrogating data on a deeper level—a scientific process that leads to new knowledge. The analytical tools needed to convert or extract knowledge from data belong to inferential statistics.

The intent of this article was to review the fundamental concepts associated with these inferential statistical tools and, via a case study example, demonstrate their proper use and limitations in converting raw data into knowledge.

How Inferential Statistics Are Used

Figure 1 illustrates the classic way in which inferential statistics are used. Typically, the focus of interest is how the larger group (or population) of people (or parts or devices) will respond under specific conditions or the relationships that may exist within or among these conditions. Because measuring or testing everyone or all things within this larger group typically is not possible or practical, representative (and preferably random) samples are selected. Controlled test conditions (or treatment[s]; e.g., drug or placebo) are administered, the results are measured, and appropriate inferential statistical test(s) are applied.

By performing these steps in a reasonably controlled and correct manner, the investigator then is able to, with some quantifiable level of confidence, use the results from the sampled data to generalize or infer how the larger population might be expected to behave under the same or similar conditions.

Of course, this does not mean that all members of the population will always behave in the same way as those included in the samples. However, the likelihood that most will behave similarly tends to be greater than the likelihood that they will not. For example, studies from small samples that produce the same or similar results consistently are used to evaluate the effects of new drugs or to achieve mainstream acceptance of new surgical techniques.

The same concepts apply when looking for cause-and-effect relationships within data. Here, the assumption or hypothesis is that an as-yet-to-be-determined relationship may exist within the population of interest. If a cause-and-effect relationship is found within a representative and randomized sample, the investigator then may be able to infer that a similar relationship also would be present in the larger population.

In general, the nature of the research question being addressed and the available data determine the statistical test that should be used. Table 1 provides a summary of the more common inferential tests and their primary uses. Potential applications to CMMS data include identifying:

- Cause-and-effect relationships. For example, and for the case study used here, what effect does ventilator preventive maintenance (PM) labor hours have on unscheduled corrective maintenance (CM) labor hours?
- Significant and meaningful trends in time series data. For example, is the apparent positive (or negative) trend in repair volume over the past three years statistically significant?
- Meaningful differences among averages across different levels in similar categories. For example, does a statistically significant difference in repair costs exist among five different infusion pumps?

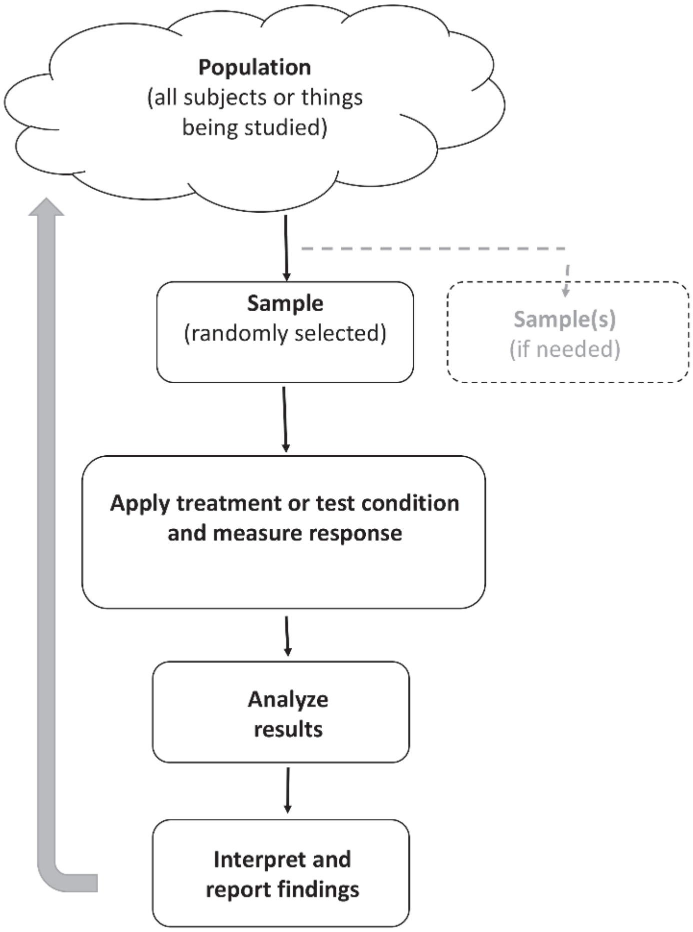

Figure 1. General nature of the research process and use of inferential statistics

Inferential Statistical Test	Common Applications
Regression analysis	1) Predicting a response variable based on one or more independent or predictor variables. 2) Identifying cause-and-effect relationships. 3) Evaluating the significance of trends in time series data.
Logistic regression analysis	Predicting the likelihood of a dichotomous (or binary) event based on one or more independent variables.
Fixed-effects ANOVA	Determining the effects of one or more categorical independent variables on a response variable.
Repeated-measures ANOVA	Determine the effects of one or more independent variables on a response variable measured from the same subject (or object) multiple times.
Student's *t* test	1) Comparing the difference between two independent means. 2) Comparing the difference between mean values from the same subject (or object). 3) Comparing a sample mean with a specific value.

Table 1. Common inferential statistical tests and applications. Abbreviation used: ANOVA, analysis of variance.

Inferential statistical tools are fundamental to scientific inquiry because of their ability to test hypotheses or claims about how a population or larger group of people, parts, or devices might behave.

- Meaningful comparisons with other or published benchmarking metrics. For example, are the cost-of-service ratios across all facilities within a health system statistically significantly different from other published benchmarks?

Hypothesis Testing

Inferential statistical tools are fundamental to scientific inquiry because of their ability to test hypotheses or claims about how a population or larger group of people, parts, or devices might behave. Although such tests cannot predict with absolute certainty the nature of this behavior, they provide a measure of chance or probability that the behavior will occur.

In this regard, all inferential statistics involve the testing of a hypothesis. Even though the conclusion offered by such tests is binary in nature (i.e., they force a "yes" or "no" decision as whether to reject the null hypothesis), there are actually four possible outcomes that must be considered. An understanding of these four conditions is crucial to the proper use and interpretation of the results from any inferential statistical test(s). The misuse and/or misunderstanding of these outcomes often makes interpretation of such tests confusing, incomplete, or, worse, simply wrong.

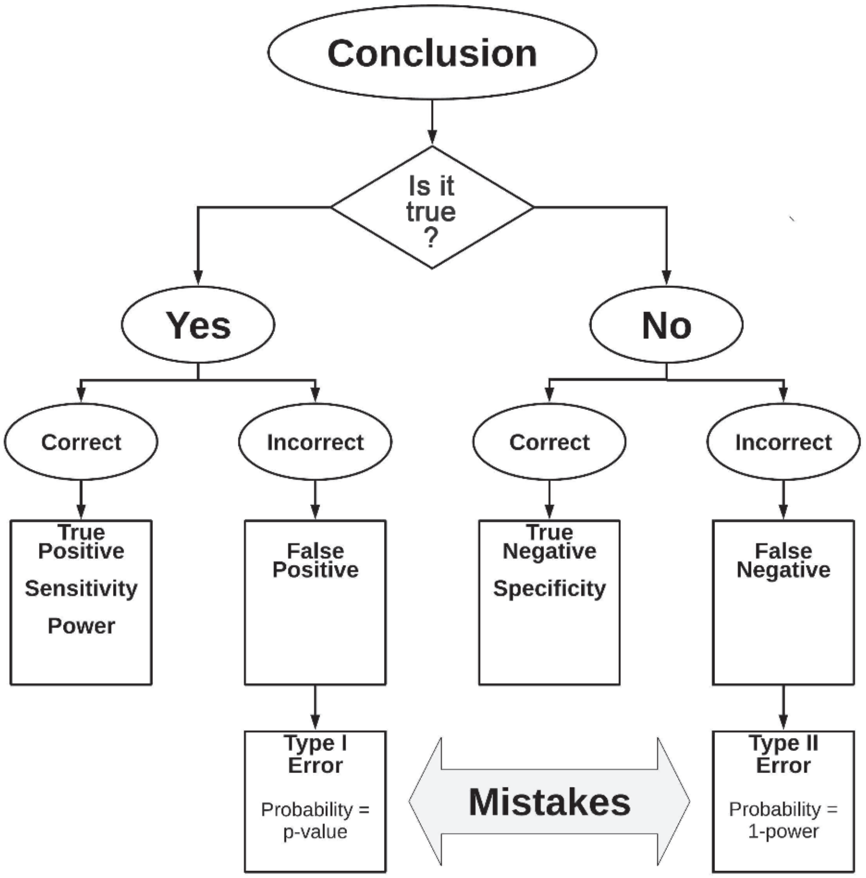

Figure 2. Possible outcomes with statistical hypothesis testing

As shown in Figure 2 and for any given hypothesis test, one of two potentially correct or "true" outcomes can result:

1. True positive. A true positive is concluding that a significant or meaningful difference (effect, relationship, or trend) exists when there really is one. The probability of correctly making such a conclusion is based on the statistical power of the test. Here, statistical power is the "ability to find significant differences when they really exist,"[1] as well as the probability of correctly rejecting a false null hypothesis. As a probability, it also ranges from 0 to 1 and often is expressed as a percentage; the larger the number, the more powerful the test. (In the context of a diagnostic screening test, a true positive also is synonymous with the sensitivity of the test.)
2. True negative. A true negative is concluding that no significant or meaningful difference (effect, relationship, or trend) exists when there really is none. (Similarly, with diagnostic screening tests, the true negative is synonymous with specificity.)

   Similarly, two potential errors or mistakes can occur in t=he interpretation of statistical results (as well as with diagnostic screening tests):

1. False positive (or type I error). A false positive is concluding that a significant difference (effect, relationship, or trend) exists when there really is none. The probability of making such a mistake is the *P* value of the test. More precisely, "the *P* value is also the probability of obtaining a value of the test statistic as large as or larger than the one computed from the data when in reality there is no difference."[2] Somewhat conventionally, investigators often set the acceptable level of making a type I error to 0.05 (i.e., being wrong five times out of 100). (This threshold could just as readily be set at 0.001 if the cost of being wrong is especially severe.)

2. False negative (type II error). A false negative is concluding that no significant difference (effect, relationship, or trend) exists when there really is one (but it could not be detected). Because investigators never really know when they may have made such an interpretative mistake, they must (or should) rely on the accompanying measures of chance or probability in assessing their likelihood of doing so. In this regard, the chance of committing a type II error is equal to the difference of "1 – statistical power."

   For example, if your statistical test returns a relatively large *P* value and you conclude no significant effect, but the power of your test is only 30%, you would have a 1 – 0.30 (0.7 or 70%) chance of being wrong and having committed a type II error. The conclusions or claims associated with these tests may now (and should) be further qualified based on the nature of these probabilities. In essence, any time a statistical claim of no significance is made, the astute observer also should seek to learn the power of the test. Only then may the strength of such a claim be fully evaluated.

Steps in Mining a CMMS Database

Identifying the Study Population

As applied to a typical CMMS, the population of interest may be all devices within a particular category (e.g., infusion pumps, defibrillators), from a specific manufacturer, or from a specific facility in a larger health system. After this study population has been defined, it should be understood that any results obtained from the analysis can only be generalized back to this specific population. For example, if a study only involves ventilators, we could not conclude or claim that similar results apply to infusion pumps.

Formulating the Research Question(s)

Establishing a clear, statistically testable research question is an essential first step in determining how a dataset is interrogated or experimental data are analyzed. For the case study results presented here, the question was, "What are the effects of ventilator age, manufacturer, and accumulated PM labor hours on unscheduled CM labor hours?"

In the above question, the terms that typically follow the word "of" represent the study's predictor or independent variables and are what the investigator will be manipulating, controlling, or allowing into the study. Here, the age of the ventilators (in months) at the time of CM service, manufacturer, and accumulated hours of PM are the predictor variables. (For the actual data subset used, "ventilator" was the generic device name used within the CMMS. As a result, this subset may contain a variety of different types of ventilators.) The term (or terms) that follows the word "on" in the above question typically represents the response or dependent variable in the study. In essence (and alternatively), the following question is being asked of the data: "Is the amount of unscheduled CM labor influenced by ventilator age, manufacturer, or amount of PM labor?" The process by which we go about answering such a question is how we convert mere raw data into knowledge.

What also is implied within the wording of the research question—and fundamental to statistical hypothesis testing—is the beginning presumption that none of the predictors will have a statistically significant effect on the response variable (ventilator CM labor hours). Also referred to as the null hypothesis, this is analogous to the presumption of innocence in a court of law. Legally, as in science, it is up to the experiment and the resulting data (or the prosecution) to argue against this null condition and prove otherwise.

Establishing a clear, statistically testable research question is an essential first step in determining how a dataset is interrogated or experimental data are analyzed.

Identifying the Appropriate Statistical Test(s)

After the research question and its associated independent and dependent variables have been properly defined, the most appropriate statistical test(s) to be used most likely will be determined. Because the question used in this case study has multiple predictor variables (ventilator age, manufacturer, PM labor hours) and a numerical response variable (CM labor hours), multiple regression analysis becomes a natural first statistical method of choice. If we only wanted to compare CM labor hours from two different manufacturers over the same study period, for example, then an unpaired Student's *t* test would be a good first choice.

Handling the Data

Because virtually all CMMS databases may be inherently "messy" (i.e., contain missing or incorrect failure codes, parts costs, inflated labor hours), some level of data cleansing should be done after a subset has been identified and extracted. Although it may be time consuming, individual device records should be reviewed and corrected if possible; especially if the CMMS has not already prevented incorrect or incompatible entries. The value of doing such a data-cleansing "deep dive" also offers the promise of better PM program management.[3] This natural messiness, however, does not preclude the effective and meaningful use of inferential statistical tools. In fact, and one of the by-products of multiple regression analysis (and analysis of variance [ANOVA]) that makes it such an elegant and appealing analytical tool is its ability to quantify and partition out the variability in our response variable that remains unexplained. It is in this unexplained variability, or statistical error, where the cumulative effects of messy data end up. In knowing such a quantity (expressed as a percentage), we then have an additional and extremely useful metric for giving perspective to and assessing the ultimate value of our statistical results.

In addition to the necessity for having reasonably "clean" data, most statistical software packages will require correct formatting of the data before meaningful tests can be performed. For the regression analysis used here, a stacked data format is required. Simply, this involves having the response variable and each independent variable in separate columns and each individual work order or PM inspection in separate rows.

Running Statistical Test(s) and Interpreting Results

To provide foundation and context to the results of any inferential statistical test(s), investigators often will (and should) include an accompanying summary table describing the nature or demographics of their study population and representative sample(s). Summary statistics for the case study described here are shown in Table 2.

The results and output from the statistical software (Minitab version 18; Minitab, State College, PA) are shown in Table 3.

Often referred to as a summary or ANOVA source table, these results (somewhat unfortunately) are not customarily included in most scientific reports; rather, only their accompanying *P* values are reported. Nonetheless, the essence and interpretation of such tables include:

- The source. The source is a listing of each of the predictor variables used in the study (e.g., Age@CM, MFG, PM_hrs; see Table 3 footnote). Each predictor variable has a corresponding *P* value.
- *P* value. The *P* value is the number used to declare whether a predictor has a significant effect on the response variable. It is derived from its corresponding and calculated *F* value. Accordingly, and if the statistical threshold for claiming significance is 0.05, the interpretation of these *P* values include:
  - Regression ($P < 0.001$): at least one of the predictors is significant in explaining the number of ventilator CM labor hours.

Demographic	Notes/Quantities
Study period	January 2012 to December 2016
No. of ventilators	2,045
No. of manufacturers	26
No. of PM inspections	15,715
No. of PM inspection labor hours	21,093
No. of CM work orders	5,882
No. of CM labor hours	11,599

Table 2. Ventilator study summary statistics. Abbreviations used: CM, corrective maintenance; PM, preventive maintenance.

  - Device age at the time of its CM, PM labor hours, and device manufacturer all are significant predictors of CM labor hours ($P < 0.001$).
- A note of caution: Investigators should be particularly cautious bout *not* equating statistical significance with practical significance. Just because a statistical test reveals a significant mathematical effect does not automatically confer a practical, meaningful level of significance. To put *P* values in their necessary perspective, further inquiry and analysis of the statistical output often is necessary.

Lastly, the R^2 values reported in these "summary tables" can be particularly insightful. Expressed as a percentage and shown in Table 3, the R^2 value is the proportion of variability in CM labor hours that is explained by age, PM labor hours, and manufacturer. The value is defined as the ratio of the regression sum of squares to the total sum of squares (i.e., 19,564/12,4975 = 0.157).

At a particularly low 15.7%, this further implies that 1 – 0.157 (or approximately 84%) of all of the variability in CM labor hours was caused by factors that were not included in the regression analysis. Simply, and perhaps not surprisingly from a practical perspective, ventilator CM labor is being influenced by factors other than those included within this study. Key factors are therefore missing from this regression analysis. In addition to the relatively low R^2, the significant *P* value (<0.001) associated with the model's lack of fit further cautions against using the results of this analysis for any predictive purposes. Nonetheless, even when they may be counter to what we hope or expect, such findings still represent new knowledge, new insights, and a higher level of understanding than what we would have had if the study had not been performed.

Checking Assumptions of Statistical Tests

Although a full review of statistical assumption testing in general and regression studies in particular is beyond the scope of this article, all inferential statistical tests involve certain underlying assumptions. A particular test may be robust to violations of these assumptions; however, severe violations could lead to incorrect or misleading conclusions. Most statistical software packages make such testing relatively easy. A review by Williams et al.[4] is particularly applicable to those assumptions for regression analysis.

Conclusion

Regardless of the statistical outcome, scientific inquiry and analysis of raw data will always reveal something. At most, it may reveal a strong cause-and-effect relationship (which may become a new law). At the least, such interrogation provides new insight and motivation for further study. The bottom line is that the HTM community has the data and inferential statistical tools are available. It is in the merging of the two that we can convert mere data into knowledge.

References

1. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, NJ: Prentice Hall; 2009:156.
2. Glantz S. *Primer of Biostatistics*. 4th ed. New York: McGraw-Hill; 1997:104.
3. Sheffer J. Group's Deep Data Dive Helps Optimize Medical Equipment Maintenance. *Biomed Instrum Technol*. 2015;49(3):203–7.
4. Williams MN, Gómez Grajales CA, Kurkiewicz D. Assumptions of Multiple Regression: Correcting Two Misconceptions. Available at: http://pareonline.net/getvn.asp?v=18%26n=11. Accessed March 19, 2018.

Source	Degrees of Freedom	Sum of Squares	Mean Square	*F* value	*P* value
Regression	28	19,556	698.4	10.1	<0.001
Age@CM	1	1,365	1,364.7	19.8	<0.001
PM_hrs	1	13,595	13,594.8	197.1	<0.001
MFG	26	4,596	176.8	2.6	<0.001
Error	1,528	105,419	69.0	—	—
Lack of fit	1,509	105,196	69.7	5.9	<0.001
Pure error	19	223	11.7	—	—
Total	1,556	124,975	—	—	—

Table 3. Analysis of variance regression statistics for ventilator study: results and output from statistical software. Model summary R^2 = 15.7%. Abbreviations used: Age@CM, device age (in months) at the time of CM; CM, corrective maintenance; MFG, device manufacturer; PM, preventive maintenance; PM_hrs, PM labor hours.