Editorial

The Vicissitudes of Prospective Multihospital Surveillance Studies: The Israeli Study of Surgical infections

Robert W. Haley MD

Conducting research on the problem of infections occurring in hospitals is a difficult endeavor. Often hospital administrators and physicians are reluctant to participate; infections are difficult to define and detect; rates are artifactually influenced by patients' length of stay, intrinsic risk, and other difficult-to-measure characteristics; and the analytic techniques needed to untangle the mass of complexity are often arcane. These and other difficulties have led to a pervasive skepticism in the minds of the consumers of these research studies and to an expanding commitment of researchers to test the validity of their study designs and measurement techniques. The two-part article in this issue (pp 232-249) reports the experience of an Israeli research team who chose the prospective, multihospital collaborative design to study the problem of surgical wound infections (SWIs) in their country.[1-2]

At first glance the report is impressive: a two-part series, a subsection devoted to each element of the study design, the use of sophisticated, even newly developed, statistical techniques, and the demonstration of alarmingly high wound infection rates with large inter-hospital differences. These guarantee wide interest in and discussion of the report. Upon further study, however, the profound difficulties that have beset most previous prospective, multihospital collaborative studies of nosocomial infections begin to show through. The validation efforts, intended to quiet the reader's concerns, may stimulate and heighten them further. There are strengths in the study. However, instead of an elaborate nationwide project with extensive methodologic validation, the authors compile routinely collected surveillance data. Introduce a series of analyses with a statistical flourish.

The objectives of the study in this design phase centered around the intent to obtain a cross-sectional measurement of the interhospital differences in overall SWI rates; however, a different objective of describing the surgical patient population was the central focus of the analys[1,2] The purpose of the validation studies was to demonstrate that the infection rates were measured with a uniformity and accuracy that would justify interhospital comparisons.' Six main characteristics of the study design were evaluated.

1. Selection of Hospitals. The sampling universe was defined as all 19 hospitals in Israel with more than 40 general surgery beds. A 100% sample was chosen, but only 11 (58%) agreed to participate. No analyses comparing participating and nonparticipating hospitals were presented to determine the degree of external validity of' the partial sample. The availability of only 11 hospitals was certain to limit the analyses of interhospital differences; recent efforts to develop valid means of comparing hospitals' mortality rates, for example, have included thousands of hospitals with little success.:[3]

2. Selection of Patients. In each hospital 500 consecutively admitted, general surgery patients were to he enrolled. This is the number of patients needed to detect a 30% difference between two hospitals with expected

From the Division of Epidemiology, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas.

Address reprint requests to Robert W. Haley, MD, Division of Epidemiology, Department of Internal Medicine, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235.

https://doi.org/10.1086/645844 Published online by Cambridge University Press

overall rates of SWI in all general surgery patients of 10% with the usual levels of confidence ($\alpha = .05$ and power = .8). This sample size would be expected to be far too small for interhospital comparisons limited to subgroups of patients (eg, analyses by operation). Only by abandoning the interhospital objective and simply pooling all patients could interesting subgroups be analyzed.

In view of the large seasonal variations in SWI rates known to exist and later documented in this study, the decision to select consecutive admissions. rather than a random sample of a year's admissions, was unfortunate. This resulted in each hospital's having a different mix (9-14 months). Moreover, in the larger hospitals, each general surgery department (ward) was represented by patients selected from a different season, likely giving rise to seasonal artifacts among different types of operations if these were segregated by department. This strong source of confounding must he examined in all analyses.

3. Data Collection System. The data were collected by nurses employed by the hospitals. We are told of no payments or other Incentives for the nurses to remain diligent or for the hospitals to avoid distracting them from the task, both real-life problems that are of concern to anyone who has managed data collectors. Training consisted of "a series of epidemiologic exercises" adapted from a Master of Public Health curriculum given one day a week times 5, plus visits of undetermined length and content by centrally supervised nurses (CTNs) to give on-the-job training after the study was underway. To detect SWIs the nurses "joined ward rounds, dressing rounds, etc." We are given no assurance that the nurses actually attempted to observe all wounds daily; no measurements were made to determine how often they actually made these observations; nor are we informed of any procedures used when several surgeons rounded simultaneously, on different patients or when nurses were ill or on vacation. The all-important daily supervision of the nurses is not described. The lack of control over the process was portended by the premature termination of the study in hospital C and the failure of the nurse in another hospital to do the required home follow-ups dur-- ing the regular period of data collection.

Early in the study the CTNs combined training with quality control (QC) measurements by simultaneously reviewing Two to three patients selected "at random" (convenience sample) at each visit. Analysis of' these QC data found that the data collected by the hospitals' nurses and the CTNs agreed closely (<1% errors). With only ten cases reviewed per nurse, however, there is little information to reach any firm conclusions about their accuracy in the complete patient sample. One could argue about the representativeness of these QC data collected only when the CTN was present, while she was simultaneously conducting training. and only in the initial stages of the study before the effects of fatigue and distraction would have been expected to diminish the quality of data collection.

4. Diagnosis of Wound Infection. Apparently wishing to be more compulsive than having nurses apply uniformly the two diagnostic algorithms (pus; nonpurulent drainage plus any two of systemic antibiotics. local wound treatment or pure culture), the authors requested that the nurses' handwritten descriptions of' the wounds be mailed to the central office where four physicians alternated making the diagnoses. Of what value was the physicians' clinical acumen when they could not see the patients and were limited to the nurses' descriptions? This approach would seem merely, to have introduced an unnecessary layer of variability mto the process.

In fact, this criticism is borne out by the results of the duplicate reading system used to measure the reliability of the diagnostic system (Table 5 in part I). In analyzing the duplicate reads, the authors pooled the various pairings into an overall estimate of kappa, Fleiss' measurement of agreement among raters.[a] Such pooling is usually considered appropriate only if kappa for each of the pairings had been roughly the same (homogeneity of agreement). A close examination of Table 5, however, shows that, far from homogeneous, 10 of the 11 (91%) disagreements over the classic definition (#1) involved physician 2. Given the simple algorithmic decisions involved ("pus" versus "no pus"), one would have expected extremely high agreement. Instead, the inescapable conclusion is that physician 2 used a different version of the classic definition than the other three physicians used. The unfortunate decision of the authors to pool the ratings (Appendix 1 of Part I) led them to conclude that the errors were due to some ill-defined imprecision in the classic definition ("pus"!), when in fact they were due to physician error in applying the simple diagnostic rule. Even if their system had given a high level of reliability ("repeatability") it would not have measured the validity (correct clarification of infected and noninfected patients). The latter can only be done by comparing the final diagnoses to some "gold standard:"" In defense of this omission, however, the conceptual and economic difficulties of measuring validity are well known.

5. Definition of Wound Infection. 'I'he authors devoted a great deal of analysis to demonstrating the desirability of using a "broader definition" of wound infection in addition to the "classic definition." While appearing to be a mere side benefit of the study, this issue is really crucial to the future analyses of' the database, for only by including the larger number of SWIs obtained through the broader definition do the overall SWI rates approach the 10% level assumed in the patient sample-size determination.~

The broader definition was supported by two analyses. First, the authors presented the analysis of' physician agreement on each definition, using the pooled kappa statistics, and concluded that the broader definition had greater reliability ("repeatability") than the classic definition. As shown above, the apparent low level of reliability of' the first definition was really due to misapplication of the classic definition by one of the four physicians.

Second, they designed a rather ingenious analysis to compare the presumed prolongation of' stay fi-om infection in patients with SWI by the classic definition and those with SWI added by the broader definition. While the matching approach overestimates prolongation of stay," one might expect the bias to be roughly equal (and

https://doi.org/10.1086/645844 Published online by Cambridge University Press

thus offsetting) when matching is used to compare the morbidity of two groups of patients with suspected SWI. Nevertheless, one would have been more comfortable if the details of' matching of patients with "high-risk diagnoses" were reported. By further inflating the estimates of prolongation of stay,[6] ineffective matching would tend to obscure the difference between the two definition groups.

Even under the assumption that the approach is sound, the authors' counterintuitive conclusion that "the inclusion of the patients with discharge other than pus could he tolerated because their length of stay in hospital was the same as that for the group with pus" is not completely supported by the results. In fact, those SWIs defined by the classic definition had an average prolongation of stay of 11.4 days versus 9.3 days for those added by the broader definition (Table 6 of part I). The authors cited the P value of 0.10 from a two-tailed test (no prior assumption about which group was expected to have a greater prolongation) as the evidence that no real difference existed. A priori, however, the hypothesis at issue is whether the patients with infection added by the broader definition have a lower average prolongation of stay due to the misclassification of some uninfected patients among the infected ones (a one-tailed hypothesis). By the one-tailed test, the P value would have been 0.05. Thus, the results suggest the common-sense assumption, that the broader definition adds some infected and some uninfected patients. Moreover, by allowing surgeons' culturing practices into the broader definition, the misclassification might well produce a serious information bias.' Knowledge of' the usefulness of the broader definition awaits the study of its validity (sensitivity and specificity).

6. Hospital Discharge Policy. In attempting to demonstrate that differences in length of stay did not bias the hospitals' SWI rates (a counterintuitive hypothesis), the authors analyzed the correlation between hospitals' average length of stay and their rates of SWIs that appeared after discharge (note an analysis of hospitals, not patients). The identification of the postdischarge infections by telephone interviews with patients is open to criticism, but assuming it to give a reasonable proxy for the true rate, the results were very interesting. The correlation coefficient was -0.48, indicating a moderately strong inverse association: hospitals with longer average lengths of stay tended to have lower rates of postdischarge SWIs. The authors partially dismissed the finding, however, by citing the P value of 0.22 as nonsignificant and referring to the need for larger samples of patients in the future. 111 fact, with only I I hospitals in the correlation analysis, to dismiss a P value of 0.22 is certainly to risk committing a type II error. The remedy in this case would have been more hospitals in the analysis rather than more patients per hospital. Thus, average length of stay remains a source of' interhospital variation that must be controlled in comparisons of hospitals.

As has been true of prospective, multihospital surveillance studies in the past, the results of the Israeli study proved to be far more interesting and useful than the validation efforts. Had the authors not chosen to indulge in the methodologic analyses (and even though they did), there is no reason that the database should not prove fully as useful as its predecessors, all of which suffered the same vicissitudes to one degree or another. The ultimate usefulness, however, will depend heavily on the extent to which the authors take seriously the interhospital biases that clearly remain in their data.

The tantalizing initial view of the data presented in part II gives potentially important insights into the practices and risks of surgery in Israel.[2] At first glance, as the authors pointed out, the SWI rates appear very high in comparison with those seen in the United States for comparable operations. Rather than automatically ascribing the difference to "the variability in the methods of' measurement" used in previously published studies, the authors may also consider the variabilities in their own study. For example, a prime candidate for explaining their higher rates is the use of the broadened definition of SWI. In future analyses the authors might consider analyzing the SWI rates separately by the two definitions to provide for direct comparisons with previously published studies, virtually all of which used the classic definition only.

As for residual interhospital variations in SWI rates not explained by the models, the authors must not discount the potential for important interhospital differences in the sensitivity of their diagnoses of wound infection that was unmeasured and uncontrolled, particularly in the latter stages of' the study. As in other prospective studies, these could, and probably do, explain much of the residual interhospital differences. Interestingly, in the part I I descriptive article the authors largely abandoned the original objective of analyzing interhospital differences, appropriately so in light of the limitations of the database, and turned to the more productive patient-level analyses of risk factors that might lead to risk reduction. Possibly future analyses will focus on surgeon-specific SWI rates by wound class, or other appropriate intrinsic risk index, that have been used in programs that reduced SWI risks in other countries.[8-18]

Perhaps the most encouraging aspect of this study is the fact that, by focusing on the long-known association of SWI with surgical drains and the extremely high rate of drain use by Israeli surgeons, a randomized trial of drain use has been undertaken there. Despite the consistency of this association in many observational studies, it remains an open question how much of the association is the causal effect of putting a drain in a wound and how much is due to surgeons preferentially putting drains in wounds that have a high probability of becoming infected anyway. 111 the randomized trial, one assumes that the authors will measure the further characteristics of drain use importantly related to infection risk but not measured in the present study-namely, drain placement (through the wound versus a separate stab hole) and type of drain (Penrose versus suction drain)."'

In conclusion, the Israeli Study of Surgical Infections is a welcome addition to the literature of nosocomial infection epidemiology. It will likely be remembered both for what it is and what it is not. It is *not* a model for assessing

https://doi.org/10.1086/645844 Published online by Cambridge University Press

the methodologic validity of an epidemiologic study, an example of clarity in scientific writing, or a perfect database in which interhospital differences can be assumed to be free of bias. But then, no study is perfect! It is a reputable attempt to collect data in the difficult setting of collaborating hospitals, as good as previous efforts in other countries, hopefully a means to solving serious SWI problems in Israel, and the springboard for a potentially important randomized trial of drain usage.

REFERENCES

1. Simchen E, Wax Y, Pevsner B, et al: The Israeli study of surgical infections (ISSI): I. Methods for developing a standardized surveillance system for a multicenter study of surgical infections. *Infect Control Hosp Epidemiol* 1988; 9:232-240.
2. Simchen E, Wax Y, Pevsner B: The Israeli study of surgical infections (ISSI): II. Initial comparisons between hospitals, with special focus on hernia operations. *Infect Control Hosp Epidemiol* 1988; 9:241-249.
3. Health Care Financing Administration: *Medicare Hospital Mortality Information, 1986*. Washington, DC, Government Printing Office (GPO #017-060-00206-9), 1987.
4. Fleiss JL: *Statistical Methods for Rates and Proportions*, ed 2. New York, John Wiley and Sons, 1971.
5. Quade D, Lachenbruch PA, Whaley FS, et al: Effects of misclassifications on statistical inferences in epidemiology. *Am J Epidemiol* 1980; 111:503-515.
6. Haley RW, Schaberg DR, Von Allmen SD, et al: Estimating the extra charges and prolongation of hospitalization due to nosocomial infections: A comparison of methods. *J Infect Dis* 1980; 141:248-257.
7. Kleinbaum DG, Kupper LL, Morgenstern H: *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, California, Lifetime Learning Publications, 1982, pp 220-241.
8. Brewer GE: Studies in aseptic technic. With a report of some recent observations at the Roosevelt Hospital. *JAMA* 1915; 64:1369-1372.
9. Cruse PJE, Foord R: The epidemiology of wound infection—A 10-year prospective study of 62,939 wounds. *Surg Clin North Am* 1980; 60:27-40.
10. Condon RE, Schulte WJ, Malangoni MA, et al: Effectiveness of a surgical wound surveillance program. *Arch Surg* 1983; 118:303-307.
11. Olson M, O'Connor M, Schwartz ML: Surgical wound infections. A 5-year prospective study of 20,193 wounds at the Minneapolis VA Medical Center. *Ann Surg* 1984; 199:253-259.
12. Haley RW, Culver DH, White JW, et al: The efficacy of infection surveillance and control programs in preventing nosocomial infections in US hospitals. *Am J Epidemiol* 1985; 121:182-205.
13. Mead PB, Pories SE, Hall P, et al: Decreasing the incidence of surgical wound infections. *Arch Surg* 1986; 121:458-461.
14. Borst M, Collier C, Miller D: Operating room surveillance: A new approach in reducing hip and knee prosthetic wound infections. *Am J Infect Control* 1986; 14:161-166.
15. Collier C, Miller DP, Borst M: Community hospital surgeon-specific infection rates. *Infect Control* 1987; 8:249-254.
16. Nyström B: Hospital infection control in Sweden. *Infect Control* 1987; 8:337-338.
17. Gil-Egea MJ, Pi-Sunyer MT, Verdaguer A, et al: Surgical wound infections: Prospective study of 4,468 clean wounds. *Infect Control* 1987; 8:277-280.
18. Onesko KM, Wienke EC: The analysis of the impact of a mild, low-iodine, lotion soap on the reduction of nosocomial methicillin-resistant *Staphylococcus aureus:* A new opportunity for surveillance by objectives. *Infect Control* 1987; 8:284-288.
19. Alexander JW: Bacteriologic comparison of closed suction and Penrose drainage. *Am J Surg* 1984; 147:699.