Implications for Interpreting Implant Registry Recommendations: A Computer Simulation

INTRODUCTION:
Implant data registries provide a useful means of tracking and understanding factors associated with specific surgical outcomes. Some have employed detection algorithms to identify specific implant designs with higher than expected failure rates. While implant registries can be used to retrospectively identify devices and practices associated with inferior outcomes, there is great public interest in the use of registry data as an “early warning system” to detect and eliminate inferior devices, thus reducing the number of clinical failures. However, little is known about the number of cases that must be examined to achieve acceptable sensitivity and specificity in identifying successful and unsuccessful devices, and the impact of controlled release of devices to the marketplace to minimize the occurrence of early failures. In this study we have developed a computer simulation of a large implant registry to address the following research questions:

1. How many cases must be examined to detect poorly performing implants?
2. How is device performance affected if a new device is only made available to a select group of surgeons?

METHODS:
A synthetic implant registry was created consisting of 10,000 patients who underwent joint replacement performed by 100 different surgeons using 5 different implants. Hazard functions representing the relative risks for revision associated with individual patients, surgeons, and implants were derived from the literature and the annual reports of implant registries. The cumulative revision rates (CRR values) of the 5 hypothetical implants were fixed at nominal values of 10%, 15%, 20%, 25%, and 30% at 15 years post operation. These corresponded to values that were 1, 1.5, 2, 2.5, and 3 times greater than the average failure rate of devices of the same type (e.g. a cementless THR or TKR) which was set at 10%. Procedure outcomes were modeled as the product of the relative risk values for the implant, surgeon, and patient. The outcome was defined as unsuccessful (ie requiring revision) if the overall relative risk of failure exceeded a fixed threshold corresponding to a CRR of 10% at 15 years.

Using this construct, patients, surgeons, and implants were randomly matched in an iterative fashion to create 2,000 registries. For each registry, implant failure rates were calculated for a range of sample sizes of procedures - from as few as the first 15 procedures performed with an implant to the full 2,000 as well as sample sizes in between.

To evaluate the efficacy of current defective implant detection methods, an implant was designated as defective at 2,000 procedures if the failure rate was greater than 20% (more than twice the expected average) and acceptable if it was less than or equal to 20%. This classification is consistent with Stage 1 of the statistical screening procedure adopted by the Australian Joint Replacement Registry. Sensitivities and specificities of the identification criteria at different sample sizes were calculated once the data from all 2,000 registries had been considered. Scenarios where patients were randomly matched with varying subsets of the surgeon pool were also considered.

RESULTS:
When patients and surgeons were randomly matched and procedures were randomly ordered, identification of poorly performing implants (sensitivity) and adequately performing implants (specificity) was shown to vary based on the average failure rate of the implant.

Regardless of sample size, failure rates for the different implants were 10%, 15%, 20%, 25%, and 30% respectively, when averaged across iterations. However, small sample sizes of observed procedures showed greater variability across iterations than larger ones (17-42% failure at 100 procedures vs. 21-38% at 200 procedures, 24-37% at 500 procedures, and 26-34% at 1000 procedures for a 30% average implant).

With respect to sensitivity, the number of cases that were required to correctly identify an implant with 2x the average failure rate was 1350 for 90% correct identification and 2,000 cases for 95%. The same levels of sensitivity were achieved for a 2.5x implant when 140 and 200 procedures had been performed and for a 3x implant with only 45 and 70 procedures. The same levels of specificity were achieved for a 2x implant after 1850, and 1950 procedures.

When patients were randomly matched with surgeons from subsets of the overall surgeon pool, observed failure rates were shown to be highly dependent on surgeon skill. Restricting the surgeon pool to the 50% of surgeons that 2x more perform above failure rates dramatically implant performance. When the surgeons with the top 25% of failures were excluded from the simulation, the average failure rate of all implants modeled remained below the threshold of 20%.

DISCUSSION:
1. Our results demonstrate that the ability of registry data to correctly identify poorly-performing implants is highly dependent on the number of procedures observed, the true failure rate of the implant, and surgeon skill.
2. These findings suggest that early data showing an implant’s failure rate to be slightly below or slightly above the acceptable failure rate should be viewed with caution, and they highlight the importance of additional review prior to public identification of a device as poorly performing.
3. The effectiveness of registry-based detection of defective implants is highly dependent on the clinical scenario to which it is applied. Registry data for a defective implant may falsely look acceptable if highly skilled surgeons are recruited to perform the initial procedures.

SIGNIFICANCE:
Little is known about how patient-surgeon pairing and the chronological ordering of the resulting procedures affect the ability of implant registry detection methods to identify poor performing implants. This study explores these relationships and provides a framework for how to interpret registry recommendations.