
Lawrence E. Barker, Elizabeth T. Luman, Mary M. McCauley, and Susan Y. Chu

From the National Immunization Program, Centers for Disease Control and Prevention, Atlanta, GA.

Received for publication April 16, 2002; accepted for publication July 19, 2002.

Eliminating health disparities in vaccination coverage among various groups is a cornerstone of public health policy. However, the statistical tests traditionally used cannot prove that a state of no difference between groups exists. Instead of asking, “Has a disparity—or difference—in immunization coverage among population groups been eliminated?,” one can ask, “Has practical equivalence been achieved?” A method called equivalence testing can show that the difference between groups is smaller than a tolerably small amount. This paper demonstrates the method and introduces public health considerations that have an impact on defining tolerable levels of difference. Using data from the 2000 National Immunization Survey, the authors tested for statistically significant differences in rates of vaccination coverage between Whites and members of other racial/ethnic groups and for equivalencies among Whites and these same groups. For some minority groups and some vaccines, coverage was statistically significantly lower than was seen among Whites; however, for some of these groups and vaccines, equivalence testing revealed practical equivalence. To use equivalence testing to assess whether a disparity remains a threat to public health, researchers must understand when to use the method, how to establish assumptions about tolerably small differences, and how to interpret the test results.

epidemiologic methods; ethnic groups; hypothesis testing; immunization; statistics; vaccination; vaccines

Abbreviations: DTP, diphtheria and tetanus toxoids and pertussis; NIS, National Immunization Survey; TOST, two one-sided test.

Eliminating differences in immunization coverage that occur by gender, race/ethnicity, income, geographic location, and other factors is a cornerstone of public health policy in the United States (1). Tremendous progress has been made. Evidence comes from survey research, which played a crucial role in identifying immunization disparities, quantifying them, and charting the progress of efforts to eliminate them. Today, survey research is used to investigate whether equivalence has been achieved among groups. This task is more challenging than many people realize.

To determine whether equivalence has been achieved, the question “Has a disparity been eliminated?” is often asked. To answer this question, a test for difference between groups may be conducted. If the test reveals a statistically significant difference, the answer to the question is “no.” Of course, the public health importance of the difference must be assessed; if the goal of “no difference” remains a priority, limited public health resources might be directed toward determining and correcting the cause of a disparity that no longer threatens public health.

The first step in avoiding this problem is to consider how the question about disparity is being framed. Instead of asking whether disparity—or difference—in vaccination coverage among population groups has been eliminated, we might instead ask, “Has practical equivalence been achieved?”

No sample survey can be used to prove that a difference has been eliminated, for two reasons. First, even if no significant difference was found in a given sample, increasing the power of the study would probably reveal some difference, however small. Second, in the unlikely event that a difference in vaccination coverage were totally eliminated, proving it would be impossible without the use of an error-free census, not a sample. In any sample, a difference smaller than the sample’s margin of error will be undetectable.
Questioning equivalence—instead of looking for a lack of difference—is more than just semantics. Questioning equivalence reveals the most appropriate way to use a sample survey to assess disparities in vaccination coverage, and it is operationally different from looking for a difference. Equivalence testing, a process mandated by the Food and Drug Administration to determine whether a generic drug performs as well as the brand-name drug (within a previously established, tolerable difference), may address the limitations of testing for public health disparities in the traditional manner (2–5).

DIFFERENCE TESTS VERSUS EQUIVALENCE TESTS

Limitations of difference tests

Difference tests have been widely used to answer questions about whether a disparity has been successfully addressed; however, these tests are subject to well-known limitations, and the results are sometimes misinterpreted (4, 6, 7). In tests of difference, analysts test the null hypothesis that the groups under consideration do not differ. If the analysis reveals a statistically significant difference between groups, the null hypothesis of no difference is rejected. Here, analysts sometimes call attention to a difference, even though it may be too small to be meaningful. This problem plagues large samples, in which differences too small to be of public health concern are often statistically significant.

However, if the analysis does not reveal a statistically significant difference between groups, the null hypothesis must stand—it cannot be rejected. But failing to find difference is not proof of similarity. As was noted above, failing to find a difference may just indicate a sample size too small to detect a difference.

Equivalence tests

In equivalence testing, the null hypothesis is formulated so that the statistical test is proof of similarity: it states that the groups differ by more than a tolerably small amount, \( \Delta \). The alternative hypothesis is that the groups differ by less than \( \Delta \)—that is, they are similar. Thus, if one rejects the null hypothesis, one may correctly state that the groups are similar. A more complete discussion of equivalence testing appears elsewhere (8–10).

In this paper, we demonstrate equivalence testing using data from the National Immunization Survey (NIS). As recently as 10 years ago, disparities in childhood vaccination coverage contributed to a resurgence of measles in the United States. In particular, coverage was low among minority children in inner cities (11). Since then, gaps have narrowed (12). Using data from 2000, we show that racial/ethnic groups which differ in vaccination coverage by an amount that is “statistically significant” under the traditional approach are, for some comparisons, equivalent at a level of precision permitted by the survey. Defining tolerable levels of difference is a necessary precursor to applying equivalence testing to assess equity. Below, we discuss implications for defining tolerable levels of difference in this area of public health.

METHODS

The National Immunization Survey

The NIS is conducted annually by the Centers for Disease Control and Prevention to obtain estimates of vaccination coverage rates for noninstitutionalized US children aged 19–35 months. The NIS is a random-digit-dialing survey of households with age-eligible children, followed by a mail survey of the children’s vaccination providers to validate vaccination information. In general, analysis of NIS data is limited to children whose providers respond.

The weighting estimation methodology of the NIS adjusts vaccination coverage estimates for household nonresponse, households with multiple telephone lines or without telephones, and provider nonresponse. NIS methods have been described in detail elsewhere (13–15).

Using SUDAAN, version 7.5 (16), we calculated point estimates of coverage and their standard deviations by race/ethnicity for selected vaccines; these statistics formed the basis for both difference testing and equivalence testing. We conducted pairwise comparisons of immunization coverage among Whites, the largest group and (usually) the group with the greatest immunization coverage, and the largest minority racial/ethnic groups: Blacks, Hispanics, and Asians. Other groups, such as American Indians and Pacific Islanders, were not considered because of small numbers of children in those groups in the NIS.

Statistical analysis

Difference testing. Traditional tests of differences were conducted comparing coverage among minority groups with that among Whites. In statistical tests, the probability of incorrectly rejecting the null hypothesis (that is, the \( \alpha \) level) is often set at 0.05. Therefore, we constructed 95 percent confidence intervals \((1 – \alpha)100\text{ percent} = 95\text{ percent} \) for differences in coverage between two groups. A difference is statistically significant if the 95 percent confidence interval for the difference in coverage does not include 0; that is, if (point estimate of coverage white – point estimate of coverage minority) \( \pm 1.96(\text{standard deviation white}^2 + \text{standard deviation minority}^2)^{0.5} \) does not include 0.

In difference testing, the null hypothesis is “no difference.” Thus, \( \alpha \) is the probability of type I error (concluding that a difference exists when, in fact, none does). Reducing the chance of this error increases \( \beta \), the chance of type II error (the probability that the populations’ coverages will be found not to differ when the true difference is not zero).

Equivalence testing. The two one-sided test (TOST) procedure, the most basic form of equivalence testing used to compare two groups, has been studied extensively (3, 5, 7). To test for equivalence, confidence intervals for the difference between two groups must be defined, just as with difference testing. In a TOST analysis, a \((1 – 2\alpha)100\text{ percent} \) confidence interval is constructed. A careful explanation of why one defines a \((1 – 2\alpha)100\text{ percent confidence interval, not } (1 – \alpha)100\text{ percent, appears elsewhere (5). This TOST analysis, as in the difference analysis, we select } \alpha = 0.05. \) Thus, in the TOST analysis, we constructed 90 percent confidence intervals \((1 – 2 \times 0.05)100\text{ percent} = 90 \text{ percent} \) for differences in vaccination coverage between two groups. As was noted above, failing to find a statistically significant difference in vaccination coverage estimates for household nonresponse, households with multiple telephone lines or without telephones, and provider nonresponse. NIS methods have been described in detail elsewhere (13–15).

Using SUDAAN, version 7.5 (16), we calculated point estimates of coverage and their standard deviations by race/ethnicity for selected vaccines; these statistics formed the basis for both difference testing and equivalence testing. We conducted pairwise comparisons of immunization coverage among Whites, the largest group and (usually) the group with the greatest immunization coverage, and the largest minority racial/ethnic groups: Blacks, Hispanics, and Asians. Other groups, such as American Indians and Pacific Islanders, were not considered because of small numbers of children in those groups in the NIS.

Statistical analysis

Difference testing. Traditional tests of differences were conducted comparing coverage among minority groups with that among Whites. In statistical tests, the probability of incorrectly rejecting the null hypothesis (that is, the \( \alpha \) level) is often set at 0.05. Therefore, we constructed 95 percent confidence intervals \((1 – \alpha)100\text{ percent} = 95\text{ percent} \) for differences in coverage between two groups. A difference is statistically significant if the 95 percent confidence interval for the difference in coverage does not include 0; that is, if (point estimate of coverage white – point estimate of coverage minority) \( \pm 1.96(\text{standard deviation white}^2 + \text{standard deviation minority}^2)^{0.5} \) does not include 0.

In difference testing, the null hypothesis is “no difference.” Thus, \( \alpha \) is the probability of type I error (concluding that a difference exists when, in fact, none does). Reducing the chance of this error increases \( \beta \), the chance of type II error (the probability that the populations’ coverages will be found not to differ when the true difference is not zero).

Equivalence testing. The two one-sided test (TOST) procedure, the most basic form of equivalence testing used to compare two groups, has been studied extensively (3, 5, 7). To test for equivalence, confidence intervals for the difference between two groups must be defined, just as with difference testing. In a TOST analysis, a \((1 – 2\alpha)100\text{ percent} \) confidence interval is constructed. A careful explanation of why one defines a \((1 – 2\alpha)100\text{ percent confidence interval, not } (1 – \alpha)100\text{ percent, appears elsewhere (5). This TOST analysis, as in the difference analysis, we select } \alpha = 0.05. \) Thus, in the TOST analysis, we constructed 90 percent confidence intervals \((1 – 2 \times 0.05)100\text{ percent} = 90 \text{ percent} \) for differences in vaccination coverage between two groups. As was noted above, failing to find a statistically significant difference in vaccination coverage estimates for household nonresponse, households with multiple telephone lines or without telephones, and provider nonresponse. NIS methods have been described in detail elsewhere (13–15).

Using SUDAAN, version 7.5 (16), we calculated point estimates of coverage and their standard deviations by race/ethnicity for selected vaccines; these statistics formed the basis for both difference testing and equivalence testing. We conducted pairwise comparisons of immunization coverage among Whites, the largest group and (usually) the group with the greatest immunization coverage, and the largest minority racial/ethnic groups: Blacks, Hispanics, and Asians. Other groups, such as American Indians and Pacific Islanders, were not considered because of small numbers of children in those groups in the NIS.
TABLE 1. Assessing differences and equivalencies in early childhood immunization coverage by self-reported race/ethnicity, National Immunization Survey, 2000

<table>
<thead>
<tr>
<th>Vaccine</th>
<th>White coverage estimate</th>
<th>95% confidence interval from difference testing</th>
<th>90% confidence interval from equivalence testing (δ = 5%)†‡</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Black, Hispanic, Asian</td>
<td>Black, Hispanic, Asian</td>
</tr>
<tr>
<td>3 DTP*</td>
<td>95.0 (0.31)§</td>
<td>92.1 (0.82) 93.3 (0.61) 95.0 (1.2)</td>
<td>1.2, 4.6 0.4, 3.0 –2.5, 2.5</td>
</tr>
<tr>
<td>4 DTP*</td>
<td>84.3 (0.46)</td>
<td>76.4 (1.2) 78.6 (0.99) 84.9 (1.9)</td>
<td>5.3, 10.4 3.6, 7.8 –4.4, 3.2</td>
</tr>
<tr>
<td>Polio*</td>
<td>90.6 (0.41)</td>
<td>86.8 (0.99) 87.9 (0.82) 92.7 (1.4)</td>
<td>1.7, 5.9 0.9, 4.5 –4.9, 0.7</td>
</tr>
<tr>
<td>MMR*</td>
<td>91.5 (0.36)</td>
<td>87.9 (0.99) 90.0 (0.71) 89.6 (2.2)</td>
<td>1.7, 5.5 –0.1, 3.1 –2.5, 6.3</td>
</tr>
<tr>
<td>Hib*</td>
<td>94.7 (0.31)</td>
<td>92.8 (0.66) 91.1 (0.71) 91.9 (1.6)</td>
<td>0.5, 3.3 2.1, 5.1 –0.4, 6.0</td>
</tr>
<tr>
<td>Hep B*</td>
<td>91.4 (0.36)</td>
<td>89.2 (0.87) 88.2 (0.77) 90.5 (1.7)</td>
<td>0.3, 4.0 1.5, 4.9 –2.6, 4.4</td>
</tr>
<tr>
<td>Var*</td>
<td>66.4 (0.61)</td>
<td>67.6 (1.5) 70.2 (1.1) 77.2 (2.2)</td>
<td>–3.8, 1.9 –6.2, –1.3 –15.4, –6.2</td>
</tr>
</tbody>
</table>

* NIS, National Immunization Survey; 3 DTP, three or more doses of diphtheria and tetanus toxoids and pertussis vaccine; 4 DTP, four or more doses of diphtheria and tetanus toxoids and pertussis vaccine; Polio, three or more doses of poliovirus vaccine; MMR, one or more doses of measles-mumps-rubella vaccine; Hib, three or more doses of Haemophilus influenzae type b vaccine; Hep B, three or more doses of hepatitis B vaccine; Var, one or more doses of varicella vaccine.

† 95% confidence interval for the difference in coverage between Whites and the minority group (Whites minus minority group). Statistically significant differences are printed in boldface type. A difference is significant if (point estimate of coverage_white – point estimate of coverage_minority) ± 1.96(standard deviation_white2 + standard deviation_minority2)0.5 does not include 0.

‡ 90% confidence interval for the difference in coverage between Whites and the minority group (Whites minus minority group). Equivalencies are printed in boldface type. The groups are equivalent if (point estimate of coverage_white – point estimate of coverage_minority) ± 1.645(standard deviation_white2 + standard deviation_minority2)0.5 is contained in the interval with endpoints –5% and +5%.

§ Numbers in parentheses, standard deviation.

percent). In a TOST analysis, we reject the null hypothesis that the groups differ by at least Δ and declare two groups similar at the α = 0.05 level, if the 90 percent confidence interval for the difference in coverage is completely contained in the interval with endpoints –Δ and +Δ. That is, the groups are similar if (point estimate of coverage_white – point estimate of coverage_minority) ± 1.645(standard deviation_white2 + standard deviation_minority2)0.5 is contained in the interval with endpoints –Δ and +Δ.

In equivalence testing, the null hypothesis is “a difference of Δ or more.” Thus, α is the probability of concluding that the populations differ by less than Δ when, in fact, the difference is Δ or more. Similarly, β is the probability that the populations’ coverages will be found to differ by at least Δ when the true difference is less than Δ.

For this analysis, we demonstrate equivalence testing using Δ = 0.05. This value was selected for this demonstration because NIS data for 2000 can detect equivalences at ∆ = 0.05 between Whites and the other races/ethnicities and vaccine series presented here. For some vaccines in some contexts, this value might have public health relevance. However, for other vaccines in other contexts, an appropriate value might be lower or higher. As we discuss below, if researchers apply the equivalence testing method in studies that conclude with public-health recommendations, the critical choice of Δ should depend on public health issues and will vary with the context.

RESULTS

Difference and equivalence in coverage

When we examine point estimates of vaccination coverage among young children by race/ethnicity, percentages are generally higher among Whites and Asians (table 1). Smaller standard deviations for estimates of coverage among Whites show less variance in the data, which reflects the larger number of Whites sampled.

For difference tests, boldface type in table 1 indicates that these groups differ from non-Hispanic Whites; that is, the null hypothesis of no difference is rejected. For vaccines for which coverage differs between two groups, the confidence intervals of the differences do not contain zero (figure 1, left column). Traditional tests of differences show that, compared with Whites, coverage among Blacks is significantly lower for all vaccines listed, except varicella vaccine.

Among Hispanics, coverage is significantly higher for varicella vaccine and lower for all other vaccines except measles-mumps-rubella vaccine. Compared with Whites, Asians have significantly higher coverage for one dose of varicella vaccine.

For equivalence tests, boldface type in table 1 indicates similarity; that is, the null hypothesis of a difference greater than ±Δ is rejected. For vaccines for which coverage is similar between two groups, confidence intervals for the difference are completely contained in the interval with endpoints –5 percent and +5 percent (figure 1, right column). Within Δ = 5 percent, Blacks in comparison with Whites have similar coverage for ≥3 doses of diphtheria and tetanus toxoids and pertussis (DTP) vaccine, ≥3 doses of Haemophilus influenzae type b vaccine, ≥3 doses of hepatitis B vaccine, and one dose of varicella vaccine (table 1). Among Hispanics, coverage is equivalent to that among Whites for ≥3 doses of DTP, ≥3 doses of poliovirus vaccine, one dose of measles-mumps-rubella vaccine, ≥3 doses of Haemophilus influenzae type b vaccine, and ≥3 doses of hepatitis B vaccine. Among Asians, coverage is equivalent to that
FIGURE 1. Differences and equivalencies in early childhood immunization coverage by self-reported race/ethnicity for three minority groups versus Whites, National Immunization Survey, 2000. The horizontal T-shaped bars depict confidence intervals. Confidence intervals for statistically significant differences (95% confidence intervals) do not include zero. Confidence intervals for equivalencies (90% confidence intervals) are contained in the interval with endpoints $-5\%$ and $+5\%$. Confidence intervals marked with an asterisk go beyond $\pm 10\%$. 3 DTP, three or more doses of diphtheria and tetanus toxoids and pertussis vaccine; 4 DTP, four or more doses of diphtheria and tetanus toxoids and pertussis vaccine; Polio, three or more doses of poliovirus vaccine; MMR, one or more doses of measles-mumps-rubella vaccine; Hib, three or more doses of Haemophilus influenzae type b vaccine; Hep B, three or more doses of hepatitis B vaccine; Varicella, one or more doses of varicella vaccine.
among Whites for ≥4 doses of DTP and ≥3 doses each of DTP, poliovirus vaccine, and hepatitis B vaccine.

**Comparisons of test results**

All four possible combinations of results arise in this comparison of difference testing and equivalence testing. Both difference tests and equivalence tests must be interpreted carefully and correctly.

**Positive for difference and equivalence.** For some comparisons, the difference test shows difference, but the equivalence test shows similarity at \( \Delta = 0.05 \). This holds in seven of 21 tests shown in table 1 and figure 1. For example, for ≥3 doses of *Haemophilus influenzae* type b vaccine and ≥3 doses of hepatitis B vaccine, difference testing finds significant differences among Blacks versus Whites and among Hispanics versus Whites. However, in these instances, the equivalence test allows us to conclude that the groups are similar at \( \Delta = 0.05 \).

**Negative for difference and positive for equivalence.** When comparing Blacks and Whites for varicella vaccine coverage, difference testing does not show a difference, but we cannot conclude from this test that the groups are identical. The test merely shows that a difference has not been proven. If the sample size were large enough, a statistically significant difference would probably emerge. However, for the same comparison, equivalence testing permits us to conclude that the groups are similar regarding varicella vaccination, within the tolerance of the test.

**Positive for difference and negative for equivalence.** For coverage with ≥4 doses of DTP, the results from difference testing indicate that coverage is significantly different among Blacks versus Whites and among Hispanics versus Whites. However, the tests of equivalence do not reveal similarity. (Remember that in equivalence testing, failure to find similarity is not proof of difference.)

**Negative for difference and equivalence.** For measles-mumps-rubella vaccine coverage among Asians versus Whites, the difference test did not show difference, and similarity could not correctly be concluded from the difference test. The equivalence test did not show similarity and could not prove difference. In this case, neither the evidence against equality nor similarity was strong enough to reject the null hypotheses.

**DISCUSSION**

Equivalence testing permits investigators to pose and answer very precise questions about the levels of difference in vaccination coverage that remain between Whites and members of other racial/ethnic groups. As we have shown, for some comparisons, a difference test may reveal that groups differ by a statistically significant amount. However, equivalence testing for that particular comparison may reveal that the groups are, for practical purposes, equivalent.

Difference testing is rigorous and is appropriate for many applications—for example, to quantify disparities that adversely affect public health. However, statistically significant differences are not necessarily meaningful from a public health standpoint. Moreover, difference testing allows no conclusions about similarity between groups (we implicitly assume that only a test resulting in a decision to declare that the groups differ or that they cannot be proven to differ is being considered). Difference testing should be reserved for situations where the appropriate research question is, “Is there a disparity between groups that can be addressed with public health interventions?”

On the other hand, when interventions have succeeded in reducing disparities between groups, the appropriate research question is, “Is there practical equivalence among these groups?” In equivalence testing, as in difference testing, the method for showing that groups are similar is rigorous. With this method, the researcher must specify an acceptable difference between groups, \( \Delta \). We cannot reject the null hypothesis of a difference greater than \( \Delta \) unless the statistical evidence is strong. The sample size might limit the values of \( \Delta \) that are practical; if both \( \Delta \) and the sample size are small, it is unlikely that equivalence can be established.

For results of equivalence testing to be useful in assessing immunization coverage, \( \Delta \) must be carefully selected. For example, some small differences in coverage between groups may not increase the risk of morbidity and mortality from vaccine-preventable disease. Determining the magnitude of a tolerable difference will depend on many factors. For example, a tolerable difference for one vaccine might not be tolerable for another, depending on the epidemiology of the disease in question. A \( \Delta \) considered acceptable on a national scale might be unacceptable on a local scale.

As with difference testing, equivalence testing requires an adequate sample size. The NIS samples smaller numbers of minorities than of Whites, which can limit the ability of the data to detect similarity at small values of \( \Delta \). Oversampling minorities or combining data from multiple years could increase the sensitivity. Surveys with smaller sample sizes than that of the NIS might or might not provide sufficient discriminatory power to support equivalence testing with reasonable values of \( \Delta \).

Reducing disparities in immunization coverage should remain a top priority. However, even when health interventions succeed, differences will persist, to some small degree. Establishing the goal of practical equivalence, defining a tolerable difference that does not threaten public health, and using equivalence testing to prove that equity has been achieved could permit us to focus limited public health resources on identifying and addressing those disparities that might threaten public health.

**REFERENCES**