Different Views

Conjecture versus Empirical Data: A Response to Concerns Raised about the National Preventive Dentistry Demonstration Program

STEPHEN P. KLEIN, PHD, HARRY M. BOHANNAN, DMD, ROBERT M. BELL, PHD, JUDITH A. DISNEY, DMD, AND RICHARD C. GRAVES, DDS, DRPH

Introduction

The April 1985 issue of this Journal contained an article that summarized the procedures and results in the National Preventive Dentistry Demonstration Program (NPDDP).[1] This article, and several other NPDDP reports, were reviewed by Dr. Theodore Rebich[2] (at the request of this Journal) and by an ad hoc committee of the American Public Health Association (APHA).[3] An American Association of Public Health Dentistry (AAPHD) committee also commented on certain project findings and procedures.[4]

The reviewers did not always agree with each other or with the project team as to the study's strengths and limitations. Some of the disagreements stemmed from differences in professional opinion about the appropriateness of the NPDDP's procedures. Other disagreements arose because some reviewers had an incomplete understanding of these procedures and project findings. And finally, a few of the NPDDP's features legitimately can be criticized on purely theoretical grounds. However, empirical data demonstrate that these theoretical problems had, at most, a trivial impact on results.

Considerations of space and relevance to public health policy inhibit our responding to all of the reviewer's comments. Thus, the remainder of this article deals with just those concerns that might affect the interpretation of (and the confidence that can be placed in) the NPDDP's results regarding the cost and effectiveness of school-based preventive dental care.

Cost of Care

Three issues were raised regarding the NPDDP's estimates of the cost of providing preventive dental care:

- What resources are really required to operate a school-based fluoride mouthrinse program?
- Is it appropriate to have a treatment regimen in which all children receive both sealants and prophy/gel applications? and
- Are the NPDDP's cost results generalizable to typical school-based programs?

Fluoride Mouthrinse Costs

The NPDDP found that it cost about $3.29 per child per year (in 1981 dollars) to add fluoride mouthrinsing to a school's dental health education program. This estimate of direct (as distinct from total) costs is consistent with that obtained in other studies of actual expenses.[5] For example, the National Institute of Dental Research (NIDR) "cited $3.49 as the average annual cost per child" (in 1979 dollars) for providing the rinse in its multisite national demonstration program.[6] NPDDP's estimate is therefore slightly *lower* than that experienced in the NIDR program.

Rebich[2] believed the NPDDP estimate was unrealistically high. He said that in 1981, New York State had a mouthrinse budget of $40,000 to serve 65,000 children (i.e., about $0.62 per child). Rebich admits the $0.62 only covers supplies. It does not include labor costs, such as for the hygienist who administers the program, orders supplies, trains the teachers and other staff, and recruits new program sites. No data are provided about the extent to which the 65,000 children received the rinse according to protocol, which is an important concern given the limited opportunity the hygienist must of had for supervising the over 2,000 classrooms that were presumably served

As the APHA panel noted,[3] the NPDDP's direct treatment costs were not affected by the size of each site's staff and direct costs were "stripped of hidden research costs, excessive administrative burdens, fringe benefits, downtime, other indirect staff expenses, and other cost elements of field trials and demonstration programs." Neither Rebich nor the NPDDP included the additional cost of teacher and volunteer time in their estimates of the resources that are required to operate a rinse program.

Sealant and Prophy/Gel Costs

Rebich claimed that the NPDDP's use of sealants and prophy/gel in the same treatment regimen resulted in an unnecessary duplication of effort which in turn inflated costs.[2]

The NPDDP followed the standard procedures for both components. A non-fluoride prophylaxis paste was used to clean the fissured surfaces on the teeth to be sealed. Following sealant application, a fluoride paste was used to clean *all* the surfaces on all teeth prior to applying the fluoride gel. Both types of cleaning were necessary and the time needed for cleaning prior to sealant application was trivial compared to that required for the full prophylaxis of all surfaces. Moreover, while sealants were expected to prevent decay on occlusal surfaces, the prophy/gel component was expected to protect the proximal surfaces. Thus, combining the two components into one regimen was a reasonable strategy and not a duplication of effort.

The APHA panel felt that the NPDDP may have been "over zealous" in providing sealants and prophy/gel to all children in Regimens 1 and 3.[3] They apparently believe it is more appropriate to target expensive preventive care on just the children who are most likely to be at highest risk to dental

Address reprint requests to Stephen P. Klein, PhD, Rand Corporation, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138. Dr. Bell is also with Rand Corporation; Drs. Bohannan, Disney, and Graves are with the University of North Carolina in Chapel Hill. This manuscript, submitted to the Journal September 25, 1985, was revised and accepted for publication December 17, 1985.

Editor's Note: See also related Technical Report p 434, article p 453, and editorial p 369 this issue.

decay. Such a strategy might very well reduce total program expenditures. However, because of economies of scale and the added expense of identifying high-risk children, the cost per child treated would most likely increase (especially when necessary indirect expenses are considered). Thus, the costs presented in our reports are likely to underestimate rather than overestimate the *per child* expenditures for sealants and prophy/gel applications in a targeted, school-based program.

The APHA panel might have qualified its endorsement of targeting preventive dental care had it considered the empirical data on this strategy. Although a small percentage of children account for a large percentage of the decay, no one has yet been especially successful in identifying who these children are *before* they get decay.[7] Moreover, the children in the NPDDP who appeared to be at highest risk to decay did not derive substantially more benefit from the preventive measures tested than did children in general. Thus, while we support the concept of targeting preventive dental care, we hesitate to recommend it until its utility has been demonstrated.

Generalizability

The APHA panel noted that under certain conditions, it might be possible for school-based dental public health programs to operate more efficiently than the NPDDP. We do not argue with this position. However, our goal was not to determine the lowest costs that might be achieved by the most expert manager under ideal conditions. Instead, we focused on the costs (and effectiveness) that programs are *likely* to experience.

In keeping with this orientation, the treatment regimens tested were established by a distinguished panel of school, public health, and dental professionals, including the head of NIDR's national caries program. Staffing patterns were fixed by state dental practice laws. Site staff were hired locally and paid at prevailing rates. All the site supervisors were experienced in managing public health programs in schools. Volunteers were used when they were available.

In short, the NPDDP took several steps to ensure that its estimates of direct costs were based on the expenses that were likely to be incurred under normal operating conditions. And in that regard, the APHA panel acknowledged that the NPDDP's "cost report appears to be a true and valid reflection of the costs of the program's services."[3] Policy makers also should keep in mind that necessary indirect costs tended to be about 100 per cent of the direct costs. Thus, total costs are likely to be substantially higher than those presented in our article.

Unit and Method of Analysis

The NPDDP used the child as the unit of analysis and adjusted the size of treatment effects by means of a weighted regression technique. This approach permitted investigating the effectiveness of the various treatment procedures for all children as well as the efficacy of targeting preventive care on high-risk children. In addition, this statistical approach was adopted only after preliminary analyses indicated its appropriateness.

The APHA panel felt the data should have been analyzed using a very different approach. They preferred to use the school rather than the child as the unit of analysis. They also preferred to split the Grade 1+2 cohort into two groups and use a DMFS (decayed, missing, filled surfaces) square root transformation and a repeated measures ANOVA. The APHA and NPDDP approaches have different advantages and limitations. However, the real question for policy makers is: "Did the choice of statistical approach make a difference?"

The APHA panel's own answer to this question is a resounding NO! The panel's statisticians reanalyzed the data using *all* of the methods they preferred. Their conclusion from this reanalysis was that "*The results remained the same.*"

The APHA panel further noted that using the school rather than the child as the unit of analysis greatly reduced the sample size for the statistical tests that were conducted. This situation led the panel to express some reservation about whether the NPDDP had enough statistical power to detect a small rinse effect when the school is used as the unit of analysis.

The issue of adequate statistical power to detect small mouthrinse effects does not have to be left to conjecture. It can be tested empirically. Specifically, did the reduction in sample size that came from using the school as the unit of analysis lead to large standard errors and thereby wide confidence intervals around observed effects? Or, did the confidence intervals remain narrow even when the school was used as the unit of analysis?

Figure 1 in our article[1] shows the confidence intervals are narrow even when the *school* is the unit of analysis. For example, the interval for Cohort 1+2 at nonfluoridated sites suggests that the classroom procedures (which included the rinse) prevented between 0 and 1 surface in four years! None of the intervals for the classroom procedures went above one full surface.

The confidence intervals remained narrow because there was far less variability between school means within a regimen than there was between children in that regimen. In other words, the standard deviation was much smaller when the school rather than the child was used as the unit of analysis. This reduction in standard error size occurred because each school mean was based on several children and was therefore a far more reliable indicator of effect size than was the score on an individual child. Thus, the reduction in sample size that resulted from using the school as the unit of analysis was offset largely by the reduction in standard deviation that also resulted from this change. In short, the empirical data show that regardless of whether the child or the school was used as the unit of analysis, the NPDDP had adequate statistical power to detect small mouthrinse effects if they were present.

Some of the concerns raised about the NPDDP's procedures were apparently based on factual errors. For example, the APHA panel said "The use of the school as the unit for analysis requires the use of more than one school at each site on each treatment regimen."[3] Having more than one school is indeed desirable, but it is certainly not *required* to measure a treatment's overall (main) effect. The APHA panel also said that "Examination of the allocation scheme for this experiment showed that there were very few cases in which more than one school was used for a treatment at a given site."[3] In fact, there was more than one school in over 70 per cent of the 108 possible combinations of treatment regimen, site, and cohort. Rebich made similar errors.

School Assignments

Schools were assigned randomly to treatment regimens at Chattanooga and Billerica. These were the first two sites to enter the program. A formula was used to make the assignments at the remaining eight sites. The APHA panel

expressed some concern about these latter assignments. However, the results at the first two sites, where there was random assignment, were very similar to the results at the remaining sites.

The only difference in results between one of the first two and last eight sites that was of any potential practical import was with Cohort 5 at Chattanooga. In this group, the combination of classroom procedures, including fluoride mouthrinsing, prevented *less* decay than at any of the other fluoridated sites.

As the APHA committee noted, the assignments at the last eight sites were done with a mechanistic formula that strove to achieve balance across regimens on three factors:

- the number of children that were examined at each school prior to the initiation of the treatment procedures at their site,
- the mean level of dental decay at each school on these baseline dental examinations, and
- each school's percentage of minority students.

As a result of this formula, all the treatment regimens had very similar baseline scores, which in turn had the highly desirable consequence of reducing the need to make major statistical adjustments in measuring treatment effects (which is why the formula was employed).

Only the three variables described above were used in making the school assignments. Given the way the assignments were made, no other factor, such as teacher enthusiasm for the program, was or could have been considered in the assignment process. In addition, none of the reviewers suggested how the assignment formula could have biased results or hypothesized why the findings at the two sites where there was random assignment were so similar to the results at each of the sites where formula assignments were made. Thus, the school assignment issue is moot due to the findings at the first two sites, the consistency of findings across all 10 sites, and the mechanistic method that was used to make assignments at the last eight sites.

Control Groups

In keeping with a basic principle of research design, the NPDDP measured the effectiveness of different treatment regimens by comparing the amount of decay that developed during the program in these regimens with the amount that developed in a longitudinal control group. The APHA committee strongly supported the NPDDP's use of longitudinal control groups for this purpose, especially given the biases that would otherwise have been introduced by the now well-documented secular decline in decay levels.[8] On the other hand, Rebich felt that we should *not* have used a longitudinal control group because it could have been biased by possible Hawthorne effects.

We agree with the APHA committee's stand on this issue. However, this matter also is moot because in addition to its longitudinal control groups, the NPDDP also had cross-sectional control groups at both the beginning and end of the project. Chapter 5 in our report on treatment effects is devoted to a discussion of the analyses with these groups.[9] These analyses demonstrate the fallacy of using cross-sectional comparisons to measure the effectiveness of preventive dental care. They also show that the children who were not examined until the end of the study (and did not participate in any treatment regimen) had the same amount of dental decay as the children in the longitudinal control group. Thus, there were no Hawthorne effects.

Treatment Regimens

The APHA committee raised a theoretically valid concern about the NPDDP always using the education package in the same classrooms that received the mouthrinse. If the presence of this package had a deleterious effect on the rinse's effectiveness, then the committee is correct in stating that this feature of the design could have led to underestimating the mouthrinse's effectiveness. However, if the education package had a positive impact on the rinse's effectiveness, then the design was biased in favor of the rinse.

It is highly unlikely that the education package had either a large enough negative *or* positive impact on the rinse's effectiveness that it would change major results. The NPDDP's estimate of the rinse's effectiveness is identical to that obtained by NIDR in its 17-site national demonstration project after the NIDR data are adjusted for the secular decline in decay.[1,10] In addition, the APHA panel's hypothesized mechanism for the rinse's low effectiveness (a pervasive lack of teacher commitment to the education package) is not consistent with survey results.[11] At all 10 sites, teacher responses to mailed questionnaires showed they were generally supportive of all of the program's components and would like to see them continued. This was especially true at the lower elementary school grade levels.

The APHA panel was critical of the NPDDP providing classroom components (including the rinse) to Cohort 5 children for only two years. They correctly point out that extending the rinse to four years (as was done with Cohort 1+2) might have increased Cohort 5's four-year rinse effect.

The rinse was not given beyond the *sixth* grade because of the difficulties associated with providing it in junior and senior high schools. These problems are documented in other studies of school-based fluoride mouthrinsing,[12] including NIDR's national demonstration project.[13] This literature indicates that in most school systems, logistical, resource, and attitudinal considerations make it extremely difficult to ensure adequate compliance with the rinse protocol beyond the sixth grade. Thus, the NPDDP results with Cohort 5 reflect what is likely to occur in practice rather than what might be achievable under ideal or unique conditions.

The NPDDP's conclusions about the rinse's low effectiveness were based mainly on results with children in Cohort 1+2 at nonfluoridated sites. Even though these children received the rinse for four years and benefited the most from this procedure, it prevented decay on an average of only one-tenth of one tooth surface per child per year.

Examination Procedures

Two concerns were raised about the NPDDP's examination procedures:

- differences in decay levels between fluoridated and nonfluoridated sites could have been due in part to the examiners' knowing a site's fluoridation status, and
- having a child seen by one examiner at baseline and another examiner at the end of the study could have theoretically biased results.

Empirical data discount both concerns. The alleged bias stemming from examiners knowing whether they were at a fluoridated or nonfluoridated site was not borne out by the Wichita and Hayward data. At baseline, project staff believed that Wichita had an essentially nonfluoridated water supply and that Hayward had been continually fluoridated for several years before the project began. These beliefs were

based on statements in these sites' written requests to participate in the project.

Chemical analyses of Wichita's water supply and discussions with Hayward's Water Department that were conducted after the baseline examinations were completed revealed that both assumptions were wrong. However, as noted in our baseline examination report,[14] mean decay levels at these sites were consistent with their water fluoridation histories rather than with the examiners' erroneous beliefs about those histories. These findings suggest the examiners were not biased, but simply "called them as they saw them."

The APHA panel stated that the procedures the NPDDP used to form examiner/child pairings could have theoretically affected results. This is true if *all* of the following three conditions are present:

- some examiners are more likely than others to classify a marginally decayed surface as carious,
- this tendency has a major impact on the number of surfaces an examiner considers carious, and
- this tendency is consistent over time.

The NPDDP conducted a two-part empirical test of this issue. Part 1 investigated the extent to which there were systematic differences between examiners within a given year in their tendency to classify a surface as carious. Part 2 explored whether any differences between examiners were consistent over time.

The results from Part 1 demonstrated that within a given year, less than 1 per cent of the variance in examination scores could be attributed to systematic differences between examiners.[15] Part 2 compared the number of examination errors that occurred when pairings were and were not maintained across years. In this context, an examination error is indicated by a "reversal," i.e., a surface that is classified as affected by decay on one examination and as free from decay on a subsequent examination (because the decay process is assumed to be largely irreversible). If examiners remain consistent in their tendency to classify a surface as carious, there would be many more reversals between examinations when examiner/child pairings were changed than when they were maintained.

The APHA panel noted that there were just as many reversals when examiner/child pairings were maintained over time as there were when a child had different examiners. The APHA panel further noted that "the extremely small systematic differences among the examiners were not highly consistent across study years."

These findings provide unequivocal evidence that the examiner/child pairings had no impact on estimates of regimen effects or the power and precision of the statistical analyses that were conducted. Thus, while we agree with the APHA panel about there being a "theoretical possibility of bias" with the examiner assignment process, the empirical data show that this problem did not occur in the NPDDP.

Evaluation Strategy

The APHA panel correctly stated that the NPDDP was "conceived by dentists, designed largely by dentists (with important contributions from Rand), and managed by dentists. The data were entered and analyzed chiefly by a non-dental, independent group of researchers."[3] The panel then said that Rand's independent status was "destroyed" because it also was "involved in the planning process and participated actively in data collection as well as data analysis and interpretation..."[3]

Being involved in the planning process and participating actively in data collection activities do not destroy an evaluator's independent status. In fact, in large scale demonstration programs like the NPDDP, it is essential that the evaluator remain actively involved in the project in order to make sure necessary data are collected properly. Moreover, Rand's active involvement in data collection increased rather than decreased the credibility of the project's findings. For example, one should have far less confidence in the validity of the results on examiner reliability if the dental team that hired, trained, and supervised the examiners was also responsible for gathering and analyzing the data on these examiners' consistency.

It is naive to believe the necessary evaluation activities could have been conducted properly without a close working relationship between the evaluator and the dental team. It also is apparent that this relationship had no impact on the independence of Rand's conclusions regarding the cost and effectiveness of the preventive measures studied.

As the APHA panel noted, the dental team proposed the project to the Robert Wood Johnson Foundation. This proposal argued that providing various combinations of preventive procedures in schools was a very effective and inexpensive strategy for substantially reducing an assumed high prevalence of decay among children. However, the fact that Rand and the dental team coordinated their separate activities during the planning and data collection phases of the project did not inhibit Rand from refuting all of the dental team's assumptions regarding the cost and effectiveness of preventive dental care. This is hardly the outcome that would be obtained if Rand's independence had been destroyed.

Conclusions

Most of the reviewers' concerns about the NPDDP are based on highly improbable scenarios and purely theoretical considerations. Empirical tests of the potential problems demonstrate they had little or no impact on NPDDP results. For example:

- The results were the same regardless of whether the NPDDP or the APHA panel's statistical approach was used to analyze the data.
- The fact that the confidence intervals remained narrow regardless of whether the child or school was used as the unit of analysis shows the NPDDP had adequate statistical power to detect small mouthrinse effects if they were present.
- Results at the first two sites where there was random assignment of schools to regimens were the same as the results at the last eight sites where a mechanistic formula was used to make the assignments. Thus, the choice of assignment procedure had no effect on results or conclusions.
- There was almost no sign of some examiners being more likely than other examiners to classify a surface as carious. In addition, even the slight tendencies that were present within a given year were not consistent across years. Thus, maintaining or not maintaining examiner/child pairings over time had no impact on the size of estimated treatment effects.
- Baseline differences in decay levels between fluoridated and nonfluoridated sites corresponded to the actual fluoridation histories at these sites rather than any beliefs the examiners may have had about these histories.

Other concerns about the NPDDP stem from misunderstandings about the project's procedures. For instance, the APHA panel thought there was usually only one school per site per treatment regimen when there were actually two or more schools in 72 per cent of the 108 combinations of site,

regimen, and cohort. Similarly, Rebich believed the NPDDP's procedures and cost allocation rules inflated estimates of treatment expenses whereas the empirical results of other studies (and the AAPHD panel) suggest that the NPDDP may have underestimated costs.

Criticism of the NPDDP seems peculiarly concerned with possible bias *against* the preventive measures tested. However, the American Fund for Dental Health, which originated the study, clearly believed in the efficacy of these techniques. The Robert Wood Johnson Foundation, which supported the project, also was predisposed toward demonstrating their effectiveness, and the independent research organization that evaluated the results (Rand) had no stake in the outcomes.

The NPDDP was not a perfect study. Few are and rarely does any study provide "absolute proof"[3] about a technique's effectiveness. Nevertheless, the APHA panel found the NPDDP's sealant results "so dramatic as to be compelling."[3]

The APHA panel also should have applied this same statement to all of the preventive measures tested. The same sites, design, examiners, and analytic procedures were used to assess the cost and effects of sealants as were used with the other measures. The same high degree of consistency of results across all 10 sites was present for these measures as they were for sealants. The empirical data, including the APHA's own analyses of these data, clearly demonstrate that problems that theoretically might have been created by certain NPDDP procedures simply did not materialize. In addition, some of the concerns about the NPDDP were based on such implausible scenarios and/or limited at best to such a trivial portion of the variance in dental decay that they do not merit further comment. The views of the APHA panel's one dissenting opinion fall in this category, which probably contributed to the other panelists refusing to endorse his opinions. We therefore anticipate even stronger endorsements of the NPDDP after other researchers review the detailed reports of its procedures and results,[9,11,14–17] and analyze the data in the public use copies of its computer files.[18]

We recognize the NPDDP's results provide disappointing news about the utility of certain types of preventive care, but disappointment should not cause public health dentistry to discount the NPDDP's findings. This is especially true with respect to the costs and effects of fluoride mouthrinsing because the NPDDP's results are so consistent with those obtained by NIDR in the only other large and recent multisite demonstration study of the rinse.

We also recognize there are differences in professional opinion about the procedures that should be used to analyze data, form treatment regimens, assign schools to regimens and examiners to children, etc. However, the empirical data show that the NPDDP's basic findings would not have changed even if all the suggested alternative approaches were adopted. Given this situation, we recommend that attention should now be directed toward exploring the implications of the NPDDP's results for public health policy.

REFERENCES

1. Klein SP, Bohannan HM, Bell RM, Disney JA, Foch CB, Graves RC: The cost and effectiveness of school-based preventive dental care. Am J Public Health 1985; 75:382–391.
2. Rebich T: School-based preventive dental care: a different view. Am J Public Health 1985; 75:392–394.
3. Sheps C, *et al*: Review of the National Preventive Dentistry Demonstration Program. Am J Public Health 1986; 76:434–447.
4. American Association of Public Health Dentistry: Comments on the National Preventive Dentistry Demonstration Program. J Public Health Dent 1985; 45:29–31.
5. Foch CB: The costs, effects, and benefits of preventive dental care: a literature review. N-1732-RWJ. Santa Monica, CA: Rand Corp, 1981.
6. General Accounting Office: Reducing tooth decay—more emphasis on fluoridation needed. Report to the Congress of the United States. Washington, DC: GAO, HRD-79-3, April 13, 1979.
7. Bibby BG, Shern RJ: Methods of Caries Prediction. Washington, DC: Information Retrieval, Inc, 1978.
8. Brunelle JA, Carlos JP: Changes in the prevalence of dental caries in US schoolchildren, 1961–1980. J Dent Res 1982; 61:1346–1351.
9. Bell RM, *et al*: Treatment effects in the National Preventive Dentistry Demonstration Program. R-3072-RWJ. Santa Monica, CA: Rand Corp, 1984.
10. Miller AJ, Brunelle JA: A summary of NIDR community caries prevention demonstration program. J Am Dent Assoc 1983; 107:265–269.
11. Klein SP, *et al*: Teacher and principal survey results in the National Preventive Dentistry Demonstration Program. N-1921-RWJ. Santa Monica, CA: Rand Corp, 1984.
12. Silversin JB, Coombs JA, Drolette ME: Adoption of dental preventive measures in United States schools. J Dent Res 1980; 2233–2242.
13. Eckhaus B, *et al*: Enrollment in and compliance with a community demonstration program of caries prevention for grades kindergarten through 12. J Public Health Dent 1982; 42:142–154.
14. Bell RM, *et al*: Results of baseline dental examinations in the National Preventive Dentistry Demonstration Program. R-2862-RWJ. Santa Monica, CA: Rand Corp, 1982.
15. Klein SP, *et al*: The reliability of clinical and radiographic examinations in the National Preventive Dentistry Demonstration Program. R-3138-RWJ. Santa Monica, CA: Rand Corp, 1984.
16. Foch CB, *et al*: Cost of treatment procedures in the National Preventive Dentistry Demonstration Program. R-3034-RWJ. Santa Monica, CA: Rand Corp, 1984.
17. Bohannan HM, *et al*: Examination procedures in the National Preventive Dentistry Demonstration Program. N-2187-RWJ. Santa Monica, CA: Rand Corp, 1984.
18. Carson S, *et al*: Guide to the National Preventive Dentistry Demonstration Program public use files. N-2188-RWJ. Santa Monica, CA: Rand Corp, 1985.