Steps to Evaluate a Health Policy

D. Schwefel, W. Satzinger, P. Potthoff, J. John

Gesellschaft für Strahlen- und Umweltforschung mbH München (GSF)
Institut für Medizinische Informatik und Systemforschung (MEDIS)
Ingolstädter Landstraße 1, D-8042 Oberschleißheim, F. R. Germany

Published in: W. van Eimeren et alii (Eds.),
Third International Conference on System Science in Health Care.
Heidelberg (Springer-Verlag) 1984, pp. 837-840

Health policy: In simplified terms, a health policy may be characterized as follows:

1. Decision-makers, who (agent)
2. regard a situation as deficient or improvable, and are being (motive)
3. backed by popular support orinstitutional power, (power)
4. try to influence the behaviour of individuals (or the relationship between certain groups and individuals) (people involved)
5. in order to attain certain goals or objectives (goal)
6. that are relevant to health (care); they do this (health)
7. for a definable space (space)
8. at a certain point in time, or for a certain period of time (time)
9. using a set of instruments (instrument)
10. and incentives (incentives)
11. at a certain amount of money or energy, (cost)
12. within the framework of a persisting or changing political, economic and ecological structure (environment)
13. for the benefit of selected target groups (beneficiaries)
14. with or without paying regard to (in)direct effects or side-effects on others. (people affected)

Projects and programmes differ only gradually from a policy, the latter just being of a more general scope.

Evaluation: If evaluation is to aim at assessing the (degree of) goal attainment of a project, programme or policy, the following points have to be considered, provided that relevance and effectiveness, rather than precision and selectivity, are to be the primary concerns of a policy evaluation.

1. Multiplicity of goals: Focussing on one goal - and perhaps on a declared one only - will, in most cases, fail to hit the core of a policy, for the decision-makers (1) as well as the people involved (4) or affected (14) by those decisions, have, as a rule, different, even conflicting goals (5). Indirect effects and side-effects of the policy can only be assessed if all the different goals, expectations and fears as to its coming into force are identified and rated in terms of the implications for the people affected (14) and the beneficiaries (13) and with regard to the intended changes in health (care) (6). Identifying all, or at

least a representative sample of, effects and side effects as accurately as possible is a basic feature of scientific (as opposed to ‘pro-domo’) evaluations.

2. Strength of attributions: There has to be evidence of a (direct or indirect) change in behaviour of the people involved (4); if such change would even have occurred without the policy, it must not be attributed to it. To investigate whether a change may be attributed to a stimulus - and, while doing this, to keep the dynamics of the environment (12) in mind! - is an essential for any scientific evaluation. One problem is, of course, that the stimulus cannot always be defined with sufficient clearness in terms of space (7) and time (8).

Limits: In what follows, the evaluation of efficiency and the efficiency of the evaluation will not be considered. Only aspects of relevance and effectiveness of the evaluation can be addressed.

MULTIPLICITY OF GOALS

Starting point: In programmes, policies and even in projects, a large number of goals can usually be identified. This reflects - what should be self-evident - that in any such measure different interest groups are involved, and that they have different implicit and explicit goals; it also reflects the fact that the declared goals of a project may conflict with its actual effects and side-effects, and that these - like sudden interventions or secular trends - may outpace the original goals, thus making a one-criterion evaluation irrelevant (e.g. to decision-making). This is why, in scientific evaluations of policies, a systematic and comprehensive systems approach has to be chosen, which tries to empirically identify the (anticipated) relevant effects by taking into account all the essential (or at least representative) goals and intentions, even beyond the interests of the beneficiaries or the groups directly involved.

Evaluation of the Bavarian Contract as a case in point: “Bavarian Contract” stands for a policy agreement between statutory sick-funds and the association of practice-based physicians. This policy aims at containing health care cost by intensifying ambulatory care at the expense of hospital treatment. To evaluate this policy, we used a wide range of exploratory procedures to identify 323 (implicit) criteria of evaluation (i.e. goals, and dimensions of effects or effectiveness) included in 120 empirically assessed 'naive' models of effectiveness (i.e. hypothetical relationships between the criteria, as had been formulated by politicians, scientists and laymen). As some criteria were operationalized by means of different procedures (secondary analysis of insurance data, surveys with physicians, etc.), it was possible to perform mutual tests for reliability and validity, if at least one result could be viewed as valid.

Multi-criteria and multi-level evaluation: If one chooses a multicriteria approach to evaluation, comprising all or representative goals and effects, one is not only provided with a network of information that can be used to identify side-effects, but also with a chance to control many (potentially) intervening, competing and con- current factors; in addition, one also has some control variables at hand which can function as quasi-controls in a non-experimental design. The evaluation of the Bavarian Contract, for example, showed that technological developments in areas which are only indirectly connected with this policy, had outpaced the intended developments in the primary goal areas of the Contract. This suggests that it is risky to evaluate policies exclusively at the aggregate level. For this reason, not only a multi-criteria but also a multi-level approach has to be used, which enables us to examine, on parallel lines, developments at the micro and macro level (e.g. at the individual doctor's and at the regional level). Such an approach may draw our attention to ecological fallacies which, particularly in policy evaluations, may come in very easily.

Results: The procedures mentioned above - broad exploration of goals, multiple operationalizations, validation of indicators, and analyses at different levels - should be used to ensure that the evaluation is both valid and relevant.

VALIDITY OF ATTRIBUTIONS

Arguments: Wise politicians recognize trends before they pass them off as their policies. This is quite easy inasmuch as policies are often formulated in vague terms with respect to space, time, beneficiaries, people affected or involved, etc., and inasmuch as they fit into a general trend of development. For this reason, the problem of valid attribution is, at least for the evaluating scientists, but not always for the politicians, a particularly serious one. The following may be regarded as steps towards increasing the evidence of valid attributions:

1. With/without comparisons: Regions that are meant to be affected by a policy are compared to those that are not. Yet, the problem with such comparisons is, among other things, that a policy may spread beyond the spaces defined.

2. Before/after comparisons: The purpose is not only to pin down categorial differences between before and after, but also to identify developments of target versus residual or control variables occurring at different paces, as well as 'straw fire' effects which may occur even before the very policy has been implemented. Yet, such comparisons cannot but roughly estimate at what point a stimulus really emerges; effects may come about too early or very late.

3. Group comparisons: On the assumption that a policy, at least at its start, does not reach the whole coverage intended, differences in behaviour of groups (e.g. of physicians or patients) can be attributed to the policy, given that there are different degrees to which they know of, or are affected by, this policy. Yet, such comparisons can, in some cases, be misguided and misleading when - triggered by the easy availability of certain data - groups are investigated that have no real relevance to the policy under research.

4. Testing models: The better models of effectiveness can be theoretically formulated and empirically validated, the greater is the chance that attributions may be valid. Yet, at least in the case of the West German health care system, tests of this kind are difficult to perform since generalizations and conclusive theories about it are, on the whole, still lacking.

5. Consensus: Correspondence between different groups of observers on the plausibility of a certain attribution can be regarded as a confirmation. Yet, what is reflected by a consensus such as this, may be purposeful ideologies or false harmonies.

Of course, it has to be borne in mind that not a single one of those arguments can be cogent in and by itself. By consequence, is maximization of evidence the only way out? Does the correctness of an attribution depend on the number and/or strength of arguments in its favour?

Maximization of evidence: Usually, scientific evaluations do not start as soon as policies are being cogitated for the first time. Therefore, study designs comparable to those that are used in intervention and treatment studies, can - as a rule - not be chosen. What we must resort to is a net or system of information which is - even in the case of a most easily accessible data base - full of holes and will normally raise more questions than it can answer. For this reason, maximization of evidence in favour of a certain attribution by using different designs or arguments seems to be appropriate. The example of certificates of illness - to reduce the number and duration of sick-leave cases was one of the explicitly stated goals of the Bavarian Contract - can serve as an at least tentative illustration of this contention: With the help of quite a broad information system available to us, it was possible to test, if only partially, the first three arguments of attribution mentioned above.

Result: Only a broad information system, constituted mainly by routine and survey data and, to a considerably smaller extent, by official statistics, enables us to take into account the whole variety of goals that are embedded in the health policy to be evaluated; in addition, such a system - imperfect as it may be - will also help to formulate and test arguments for attributing certain effects to the policy scrutinized.

CONCLUSIONS

Necessarily, evaluation is an iterative process. Even if it may be true that asking too many questions may result in poor answers to each individual question, one should never a priori abstain from asking all those questions; for they might be decisive in securing the relevance of the evaluation by opening one's eyes to the variety of groups, acting in the field, and of possible interventions, parallel developments and side-effects; rather, one should resort to even finer nets or subsets of spezialized information whenever doubts and dubiosities come along. Even if it may be correct that developing good designs like randomization is highly desirable, one should make do with an explorative statistical analysis of some routine data, that might be of unsatisfactory quality, yet offer a retrospective time series, and can be supported by intelligent interpretations and other flanking data sources which make up for the inherent fallacies; after all, this procedure is appropriate because of the fact that, usually, an evaluation only starts after its subject has come into being, and that all important and measurable effects may be of a long-term or of a straw-fire quality. Even if it may be true that (quasi-)causal interpretations require hard tests, these tests can, by no means, guarantee the validity and strength of attributions, but might even make us unaware of supporting and complementary arguments. For all these reasons, it may seem most appropriate to evaluate a health policy by making use - in a exploratory, heuristic and sequential way - of as many criteria, designs, sources, levels, spaces and periods of time as possible.

The authors wish to thank the Bavarian Ministry of State for Labour and Social Affairs, the Association of Sick-fund Affiliated Physicians of Bavaria, and the Associations of Bavarian Statutory Sickfunds for their support.