USING REAL WORLD DATA FOR COVERAGE AND

PAYMENT DECISIONS:

THE ISPOR REAL WORLD DATA TASK FORCE REPORT

Task Force Co-Chair & Economic Outcomes Working Group Chair: **Louis P. Garrison, Jr., PhD**, Professor, University of Washington, Department of Pharmacy

Task Force Co-Chair: **Peter J. Neumann, Sc.D**., Director, Center for the Evaluation of Value and Risk in Health; Professor, Tufts University School of Medicine Institute, for Clinical Research and Health Policy Studies

Patient Reported Outcomes Working Group Chair: **Pennifer Erickson PhD**, Co-founder, O.L.G.A.

Clinical Outcomes Working Group Chair: **Deborah Marshall, PhD,** Director, Research & Development Health Economics & Outcomes Research**,** I3Innovus Research

Evidence Hierarchies Working Group Chair: **C. Daniel Mullins PhD,** Professor and Chair, Pharmaceutical Health Services, Research University of Maryland School of Pharmacy

Patient Reported Outcomes Working Group Chair: **Pennifer Erickson PhD**, Co-founder, O.L.G.A.

ABSTRACT

Background. Health decision makers involved with coverage and payment policies are increasingly developing policies that seek information on real world (RW) outcomes. Motivated by these initiatives, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) created a Task Force on "Real World" Data to develop a framework to assist health care decision-makers in dealing with "real world" data, especially related to coverage and payment decisions.

Methods. Task Force co-chairs were selected by the ISPOR Board of Directors. Co-chairs selected chairs for four working groups on: clinical outcomes, economic outcomes, patient-reported outcomes, and evidence hierarchies. Task Force members included representatives from academia, the pharmaceutical industry, and health insurers. The Task Force met on several occasions, conducted frequent correspondence and exchanges of drafts, and solicited comments on three drafts from a core group of external reviewers and from the ISPOR membership.

Results. We defined RW data as data used for decision making that are not collected in conventional randomized controlled trials (RCTs). We considered several characterizations: by type of outcome (clinical, economic, and patient-reported); by hierarchies of evidence (which rank evidence according to the strength of research design); and by type of data source (supplementary data collection alongside RCTs, large simple trials, patient registries, administrative claims database, surveys, and medical records). Our report discusses eight key issues: 1) the importance of RW data; 2) limitations of RW data; 3) the fact that the level of evidence required depends on the circumstance; 4) the need for good research practices for collecting and reporting RW data; 5) the need for good process in using RW data in coverage and reimbursement decisions; 6) the need to consider costs and benefits of data collection; 7) the ongoing need for modeling; and, 8) the need for continued stakeholder dialogue on these topics.

Conclusions. Real world data are essential for sound coverage and reimbursement decisions. The types and applications of such data are varied, and context matters greatly in determining the value of a particular type in any circumstance. It is critical that policy makers recognize the benefits, limitations, and methodological challenges in using RW data, and the need to consider carefully the costs and benefits of different forms of data collection in different situations.

ISPOR TASK FORCE ON REAL WORLD DATA

1. WHY A REAL WORLD DATA TASK FORCE?

1.1 Growing use of evidence syntheses and outcomes research

Health decision makers involved with coverage and payment policies are increasingly seeking information on real world outcomes on which to base their decisions. Many of them are developing policies that integrate evidence from different sources. These policies recognize the importance of evidence that goes beyond information collected during clinical development in randomized controlled trials (RCTs) required by regulatory authorities for marketing approval. It is broadly acknowledged that while RCTs provide a "gold standard" in the sense that they provide solid evidence of product efficacy under carefully controlled conditions, RCTs are carried out using selected populations under idealized conditions. In addition, they are expensive to conduct. Other sources of data can contribute in important ways to the evidence base (e.g., demonstrating how a drug works in populations or under conditions not studied in the trial, or relative to another drug not included in the study).

1.2 Policy developments

Recent policy initiatives highlight payers' attempts to collect and use such data. The Medicare Modernization Act (MMA) of 2003 illustrates the U.S. government's attempt to bolster

the evidence base on which private health plans participating in the new Medicare drug benefit can base their coverage decisions. The MMA contains a provision (Section 1013) calling on the Agency for Healthcare Research and Quality (AHRQ) to conduct research on the "outcomes, comparative clinical effectiveness, and appropriateness of health care, including prescription drugs"[1]. AHRQ has recently launched the Effective Health Care Program to synthesize, generate, and translate knowledge to aid stakeholders in grappling with the often difficult decisions they must make [2]. Many private payers are adopting the Academy of Managed Care Pharmacy (AMCP) Format, which calls for health plans to request formally that drug companies present a standardized "dossier," containing detailed information not only on the drug's efficacy and safety, but also on its projected effectiveness and economic value relative to alternative therapies.[3, 4]

Other efforts to review evidence systematically have also gained momentum, including those by private and public health plans. A notable initiative in the U.S. is the Drug Effectiveness Review Project (DERP), an alliance of fifteen state Medicaid programs and two nonprofit organizations, to synthesize and judge clinical evidence for drug class reviews.[5] Around the globe, national reimbursement authorities and the organizations that assist them, such as the National Institute for Health and Clinical Excellence (NICE), and the German Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) are struggling with similar issues.[6, 7] Both DERP and IQWiG emphasize RCTs for clinical evidence as opposed to synthetic modeling approaches, such as those used by NICE to project real world effectiveness.

2. TASK FORCE OBJECTIVES AND SCOPE

2.1 Objectives

Motivated by the MMA and other efforts, and recognizing the lack of a framework for considering real world data in coverage and reimbursement decisions, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) created a Task Force on "Real World" Data. The mission of the Task Force was to develop a framework to assist health care decision-makers in dealing with "real world" data and information in "real world" health care decision-making, especially related to coverage and payment decisions.

2.2 Scope

Defining "real world" data. From the outset, the Task Force grappled with the definition and appropriate characterization of "real world" (RW) data. It seemed self-evident that RW outcomes data should come from real world situations. Part of the Task Force's charge was to consider uses and limitations of evidence not obtained from randomized controlled trials. On the other hand, it was also clear that decision-making is a highly integrative process of synthesizing information from different sources—both "laboratory" and real world.

Some Task Force members questioned the appropriateness of the term "real world data" in the first place. In the end, we decided to adhere to the term, not only because it reflected the

charge from the ISPOR Board, but also because the term has gained currency in some policy circles.

We settled on a definition that reflects data *used for decision making that are not collected in conventional controlled randomized trials.* This is not to say that data from RCTs are irrelevant or not used by decision makers: indeed, they remain the critical foundation for almost all initial coverage and payment decisions. For if there is not a belief in the plausibility of the underlying biological mechanism or hypothesis, why should anyone seek further evidence of effectiveness or cost impact in the real world? Yet, *efficacy* evidence in a particular group or subgroup is typically insufficient to project the size of the *effectiveness* impact in the population that would actually use a product. Decision makers therefore seek additional types and sources of data.

Data vs. evidence. Our Task Force also deliberated distinctions between the terms, "real world *data*" and "real world *evidence*." Some in our group favored the latter term, or at least raised questions about whether we meant evidence when we employed the term data. The notion was that "data" conjures the idea of simple factual information, whereas "evidence" connotes the organization of the information to inform a conclusion or judgment. Evidence is generated according to a research plan and interpreted accordingly, whereas data is but one component of the research plan. Evidence is shaped, while data simply are raw materials and alone are non-informative.

In the end, we adhered to the term, "real world data" for reasons noted above. However, we try to remain sensitive to the distinctions throughout, and emphasize that thorough knowledge of the quality of the data source is necessary when deciding whether a specific set of data can be used to answer a research question or shed light on a decision.

Focus on coverage and payment decisions. We focus our report on coverage and payment (i.e., both pricing and reimbursement) decisions. This sets it apart from other contexts, such as direct patient care or regulatory approval for marketing.

We recognize the tension that sometimes arises between what is good for an individual patient and what is good for the population at large. Evidence-based medicine has been defined as "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients."[8] While physicians retain a pivotal role in patient care decisions, other bodies such as pharmacy and therapeutic (P&T) committees and national reimbursement bodies (e.g., in Canada, Australia, and most European countries) are taking on greater importance in medical decision making.

Drugs vs. other interventions. While most of our examples are drawn from the world of prescription drugs, our report has implications for all types of interventions, including drugs, devices, procedures, and health programs.

U.S. vs. global focus. The motivation for the Task Force came largely from the MMA and U.S. policy arena, but debates about types and sources of data are, of course, present around the world—as suggested in the different approaches taken, for example, by NICE and IQWiG. Though many of our examples are taken from U.S. contexts, our findings are global in reach.

2.3 Task Force Participation

Task Force co-chairs were selected by the ISPOR Board of Directors. Co-chairs selected chairs for four working groups on: clinical outcomes, economic outcomes, patient-reported outcomes, and evidence hierarchies. The first three groups correspond to the three major areas

of outcomes research that contribute to the studies of comparative effectiveness and economic evaluation sought by decision makers. The working group on evidence hierarchies addresses the cross-cutting, foundational issue of the quality of evidence. This paper reflects an integration of the various working group reports.

Task Force members included representatives from academia, the pharmaceutical industry, and health insurance plans, and brought varied experiences in using real world data. An outline and draft report developed by the panel were presented at the 2005 International and European ISPOR Meetings, respectively. The manuscript was then submitted to a Reference Group for review and comment. The task force met on several occasions, conducted frequent correspondence and exchanges of drafts by electronic mail, and solicited comments on three drafts from a core group of external reviewers and more broadly from the membership of ISPOR.

3   TYPES AND SOURCES OF REAL WORLD DATA

3.1 Characterizing RW data

There are several ways in which one might characterize real world data. One is by type of outcome: clinical, economic, and patient-reported outcomes (the focus of three of our Working Groups). An advantage of this approach is that it corresponds to the way in which many decision makers conceive of data. A downside is that it provides broad categories, each of which combines many types and sources of evidence.

A second characterization involves traditional hierarchies of evidence (the focus of our fourth Working Group), which rank evidence according to the strength of the research design. Typically, data from RCTs sit atop the hierarchy followed by data from non-randomized intervention studies, followed by epidemiological studies and so forth.[9] Evidence hierarchies provide a useful ranking based on the rigor of the research design; however, they do not provide a complete picture of RW data. The results from many RCTs are not generalizable to a broader population. Conversely, a well-conducted observational study may prove highly useful in certain situations provided that potential biases have been adequately addressed. Indeed, some would argue that observational data can often provide more relevant evidence regarding patient outcomes in actual clinical practice than can a registration RCT.

Finally, one might consider RW data by types of data sources. The value of this classification is that it identifies tangible sources of information. A potential drawback is that it represents a simplification that does not capture important design issues within each source of evidence.

Each of the three characterizations provides a different perspective on RW data. Collectively, we believe they provide a useful portrait of the strengths, weaknesses, and complexities inherent in the topic.

3.2 Types of Outcomes

Clinical outcomes. Clinical outcomes include biological measures of morbidity (e.g., blood pressure, cholesterol level, symptoms, and side effects) and mortality. Clinical outcomes include both surrogate (intermediate) and long-term measures. For purposes of this discussion,

we differentiate clinical outcomes from health outcomes and especially patient-reported

outcomes (PROs) and health-related quality of life (HRQoL).

Much of the data collected in phase III registration trials involves clinical outcomes.

Clinical outcome data are also found in many other sources, such as patient registries or

observational databases. A key issue, discussed in greater detail in Section 3.4, involves the

validity of clinical data from such sources.

Economic outcomes. "Economic outcomes" are narrowly defined here to include

estimates of medical and non-medical resource utilization and their associated costs. Such data

are used to project the expected cost of an intervention in the real world—e.g., in the numerator

of a cost-effectiveness ratio. As discussed below, many sources of RW data are useful in

providing use and cost information. Of course, we recognize that the term economic outcomes

can be more broadly construed since economic evaluations consider both benefits and costs, but

keep this narrow definition in the interest of continuity with the original charter of the task force.

Patient-reported outcomes/quality of life. Patient-reported outcome is the term adopted

by the FDA and internationally to encompass any report coming directly from patients about a

health condition and its treatment, including symptoms, functional status, HRQoL, treatment

satisfaction, preference and adherence. Researchers have long recognized that self reports of

outcomes related to disease, injury, treatment or policy are important, because they provide the

only direct voice that an individual has in the health decision making process. This information

has been widely used for decision making in the U.S. since the mid-1950s with the establishment

of the annual national health surveys; other countries have subsequently developed similar data

collection mechanisms. The Centers for Medicare and Medicaid Services's (CMS) adoption of the Medicare Health Outcome Survey is recognition of the importance of PRO data for understanding outcomes of the Medicare program. As people live longer with chronic conditions, PROs have become increasingly important to pharmaceutical manufacturers in assessing the impact of emerging chronic treatments and in communicating the benefits of these drug treatments in label and promotional claims.

Interested parties have actively debated diverse conceptual and methodological issues related to all types of PROs, resulting in the publication of various "best practices" documents.[10, 11] The level of understanding of both the scientific rigor and its application to the reporting of these subjective measures has improved. Like any science, however, questions remain. FDA and regulatory agencies globally have recently developed statements to guide the development and use of these measures, especially by the pharmaceutical industry in the drug approval process.[12, 13]

3.3 Evidence hierarchies

Historically, evidence hierarchies have been linked to "evidence-based medicine" (EBM). The thrust of the EBM movement is to ground clinical practice in rigorous research. EBM proponents emphasize that traditional medical practice incorporated local practices and expert opinion that were not tested in controlled studies.[9] They stress the need for clinical researchers to document all study protocols, utilize appropriate analytical techniques, and strive for internal consistency.[14] Studies are to be considered externally valid when findings are generalizable beyond local clinical practices. A scientific body of evidence became reliable and

generalizable when similar results were reported by different researchers across a range of study designs and patient populations. For these reasons, RCTs were placed at the top of the evidence hierarchy.

Decision makers, however, quickly recognized the impracticality of basing all of medicine on RCTs. For one thing, RCTs are expensive. For another, even the best RCT reflects a limited controlled experiment that may not generalize to populations, settings, or conditions not reflected in the trial. The need for non-RCT information became apparent, raising the question of how to grade information that by definition was of "poorer quality?"

A number of groups have developed evidence hierarchies over the years that reflect the primacy of data from RCTs, and grade other types of evidence by the rigor of the research design. For example, the hierarchy adopted by AHRQ grades evidence in order from most to least rigorous as follows:[15] 1) systematic reviews and meta-analyses of RCTs; 2) non-randomized intervention studies; 3) observational studies; 4) non-experimental studies; 5) expert opinion.

The U.S. Preventive Services Task Force (USPSTF) took the approach a step further recognizing that decision makers require information not only on the rigor of the research design but also on the magnitude of the net benefit in support of a particular technology or health service.[16] Recent work developed by the Scottish Intercollegiate Guidelines Network and the Oxford Center for Evidence-Based Medicine also recognized limitations of traditional grading systems.[17] Strict use of evidence hierarchies may not account for the methodological quality of studies or may fail to reflect the overall strength of the evidence base. Users may misinterpret the grade of recommendation or they may fail to properly weigh lower grade recommendations.[18]

Other groups maintain their own evidence grading and classification systems that combine judgments about evidence quality with judgments about the usefulness of the intervention.[19]

3.4 Sources of Real World Data

RW data can also be categorized by type of data source. Our Task Force defined six such sources: 1) supplements to traditional registration RCTs; 2) large simple trials (also called practical clinical trials); 3) registries; 4) administrative data; 5) health surveys; and 6) electronic health records and medical chart reviews.

Supplements to RCTs. To provide additional data alongside standard clinically focused RCTs, researchers often gather information on variables such as PROs,medical resource use, and costs. Such efforts can add valuable evidence on treatment patterns for common events, e.g., such as the doses of drugs used to treat rejection in kidney transplantation.[20]

Limitations to such data are also well known: their primary aim is to measure a key clinical efficacy endpoint in a carefully limited population and clinical setting. Furthermore, trials are not usually powered statistically to measure precisely the probability of rare adverse or other events and hence are of limited use in measuring the associated resource utilization and costs. RCTs are generally conducted over a shorter time frame than what is relevant for determining the overall clinical and economic impact of an intervention, and resource use is often protocol-driven.

Practical clinical trials. Large simple trials (also called practical or pragmatic clinical trials) involve prospective, randomized assignment but aimed at larger more diverse real world population.[21] Practical clinical trials have the important strength of randomization, which minimizes bias in the estimation of treatment effects. These trials are by design larger than conventional RCTs. For this reason, they are more likely to have sufficient power to capture significant differences in key outcomes of interest, such as hospitalizations.

Because the focus is on obtaining policy-relevant outcomes, costs and cost-effectiveness are more likely to be central endpoints, and the results can be more readily generalized to the relevant treatment population than those obtained from conventional RCTs: costs are less likely to reflect protocol-driven health care use; well-documented variations in resource use across various ethnic, racial, age groups and genders can be better captured by opening the trial to a more diverse population; people more at risk for adverse events are less likely to be excluded from the trial, and the related economic effects are more likely to be captured; and resource use and costs are more likely to reflect those observed in community-based settings where most people obtain their care, especially since study drugs in phase III trials are generally provided for free.

However, the large size of a practical clinical trial increases the cost of data collection and raises some concerns about the quality of data collected. Costs are increased not only because a larger number of patients are enrolled, but also because a larger number of settings are involved. Some of the issues raised by economic data collection within practical clinical trials are: identification of where subjects receive care may be more difficult (less in a closed system); data collection systems of community-based settings may be less sophisticated than those of academic settings (for example, more likely to use paper rather than electronic records, thus

increasing the likelihood of data entry errors); there is more likely to be a lack of standardization in financial and billing systems across different settings of care; and more study coordinators will be involved in the data collection effort.

Registries. Registries are prospective, observational cohort studies of patients who have a particular disease and/or are receiving a particular treatment or intervention. They can be used for understanding natural history, assessing or monitoring real-world safety and effectiveness, assessing quality of care and provider performance, and assessing cost-effectiveness..[22]

Registries involve prospective data collection of clinical, economic, and PRO information, and are increasingly relying on real-time data capture. They typically include a larger and more diverse group of patients than what is generally studied in phase III RCTs; therefore, they better reflect real-world patients, management practices, and outcomes. Patients are often followed over a longer timeframe, allowing for an assessment of longer-term outcomes. Most registries have very few, if any, required visits, evaluations, or procedures; therefore, the treatment patterns reflect the everyday clinical decision-making that is most relevant to providers and payers. Disease registries enable providers and payers to gain insight into the most cost-effective treatment approaches.

Because registries do not involve random assignment to treatment, care must be taken in analyzing and interpreting the results due to the inherent limitations of observational studies. There is no guarantee that patient groupings are comparable: therefore, registries may not be suitable to test hypotheses, but are useful to generate them. Furthermore, there are limitations in terms of the amount of data that can be collected, and because visit schedules are not required, data cannot necessarily be obtained at fixed intervals. Registries sometimes include study sites

that are not experienced in conducting research, and without appropriate oversight, data integrity could be in question. However, the use of real-time data capture is likely to improve data monitoring and integrity. Registries are, in some cases, established to collect post-marketing safety data, either in response to specific safety concerns or to fulfill regulatory obligations established as a condition of marketing approval.

Claims databases. Administrative data (typically retrospective or real time, if possible) are collected primarily for reimbursement, but contain some clinical diagnosis and procedure use with detailed information on charges. Claims databases lend themselves to retrospective longitudinal and cross-sectional analyses of clinical and economic outcomes at patient, group, or population levels. Such analyses can be performed at low overall cost and in a short period of time. Given the sheer size of claims databases, researchers can identify outcomes of patients with rare events more easily, assess economic impact of various interventions, and gain insight into possible association between interventions and outcomes.

Administrative claims databases can prove very useful in measuring resource use and costs, provided some basic principles are met. A clear research question needs to be defined and addressed by an appropriate design from a well-defined perspective. Available statistical tools can be used to help control for some of the potential biases. Methods and results should be reported in a clear and transparent fashion, so that other researchers are able to understand and reproduce the analyses.

It is worth noting that with appropriate Institutional Review Board (IRB) approvals and patient consent, some health plan databases can be linked with PROs, lab results, medical

records, and physician surveys—effectively becoming a blend of traditional registries and claims data.

Beyond challenges posed by privacy issues, the validity of retrospective claims database analyses has been challenged on several fronts: data quality (missing data, coding errors--whether random or "intended"--and the lack of comprehensive data across health care settings); the lack of or very limited clinical information on inpatient stays, health outcomes, health status, and symptoms; limited validation; absence of a population denominator; and the lack of distinction between costs and charges. Of course, the large size of these databases may be able to overcome the issue of missing data if they are missing at random. If data quality can be ascertained and privacy issues addressed, then treatment selection bias in the sample is the most common and challenging methodological issue. Estimates of the effects and costs can be biased due to a correlation between unobserved factors associated with treatment selection and outcomes, such as baseline health status.

Health surveys. Health surveys are designed to collect descriptions of health status and well being, health care utilization, treatment patterns, and health care expenditures from patients, providers, or individuals in the general population. Health surveys typically collect information on representative individuals in the target population, whether patients, physicians or general population, and are methodologically rigorous, for example, relying on complex sample survey designs. With these designs, surveys can provide information about all members of the target population, not just those who are participating in a given RCT, or members of a particular health plan. As a result, health survey data can make unique contributions about generalizability of treatments and their impacts and about use of and expenditures for health services.

The major limitation of health survey data for initial coverage and reimbursement decisions is the lack on relevant data on specific products. Survey data are also subject to issues of subjectivity and recall bias.

Electronic health records (EHRs) and medical chart review. Finally, we note that electronic health records (and other technologies capturing real-time clinical treatment and outcomes) are important sources for RW data for a wide range of clinical settings throughout the world. The expansion of electronic data capture is essentially lowering the cost of the medical chart reviews that have been widely used in the past to produce specific information on the RW use of specific tests or drugs for particular conditions. EHRs—such as the UK General Practice Research Database--contain more detailed, longitudinal information including disease-specific symptoms at the personal level and should greatly expand the use of this type of information. However, transforming the information for research purposes requires high-end statistical analysis tools and remains a challenge.

4   KEY FINDINGS

4.1 Recognizing the importance of RW data

We start with a strong affirmation of the need for RW data. As we have emphasized, RCTs have many advantages: their prospective design, pre-specified well-defined endpoints, randomization and control groups, and blinding all work to provide unbiased measures of impact

in the trial population; however, this strong internal validity can limit their external validity and generalizability about which interventions work best when implemented in different settings.

Decision makers rely on multiple sources of real world data that must be integrated or synthesized in some fashion. While RCTs remain the gold standard for demonstrating clinical efficacy in restricted trial settings, other designs contribute to the evidence base. In some situations, RW data may provide clear advantage for understanding outcomes of treatment, for example, for patients excluded from trials, patients in actual clinical practice settings (vs. research settings), and patients whose treatment is not determined by trial protocol or practice guidelines.

Among the benefits of RW data is that they can provide:

- Estimates of effectiveness rather than efficacy in a variety of typical practice settings;

- Comparison of multiple alternative interventions (e.g. older vs. newer drugs) or clinical strategies to inform optimal therapy choices beyond placebo comparators;

- Estimates of the evolving risk-benefit profile of a new intervention, including long-term (and rare) clinical benefits and harms;

- Examination of clinical outcomes in a diverse study population that reflects the range and distribution of patients observed in clinical practice;

- Results on a broader range of outcomes (e.g., PROs, HRQoL, and symptoms) than have traditionally been collected in RCTs (i.e., major morbidity and short-term mortality);

- Data on resource use for the costing of health care services and economic evaluation;

- Information on how a product is dosed and applied in clinical practice and on levels of compliance and adherence to therapy;

- Data in situations where it is not possible to conduct an RCT (e.g., narcotic abuse);

- Substantiation of data collected in more controlled settings;

- Data in circumstances where there is an urgency to provide reimbursement for some therapies because it is the only therapy available and may be life-saving;

- Interim evidence--in the absence of RCT data--upon which preliminary decisions can be made;

- Data on the net clinical, economic, and PRO impacts following implementation of coverage or payment policies or other health management programs (e.g., the kind of data CMS expects to collect under its coverage with evidence development policy).[23]

4.2 Recognizing the limitations of RW data

We also recognize important limitations of RW data. For all non-randomized data, the most significant concern is the potential for bias. Retrospective or prospective observational or database studies do not meet the methodological rigor of RCTs, despite the availability of sophisticated statistical approaches to adjust for selection bias in observational data (covariate adjustment, propensity scores, instrumental variables etc.). Observational studies need to be evaluated rigorously to identify sources of bias and confounding, and adjusted for these prior to estimating the impact of interventions on health outcomes. Observational or database studies may also require substantial resources.

4.3 The level of evidence required depends on the circumstance

The complexity of data collection underscores the fact that the level of evidence required in any circumstance will relate to the question at hand. It is important to recognize the variable quality of all data (whether prospective or retrospective, or experimental or observational). The extent to which data provide good or bad evidence depends on the research design, the quality of the information collected, and how the data are used. The optimal solution will depend on the circumstances. Decisions typically rely on multiple sources, and are best thought of as conditional, to be revisited as additional evidence is generated.

Ongoing work in evidence hierarchies recognizes this complexity. As noted, new approaches for grading evidence depend on the quality of the evidence and the magnitude of the net benefit. The approaches are more explicit about which studies provide stronger evidence in support of the use of a particular health service or drug.

In all likelihood, we need clinical trials that are more practical and "real world data" that are more statistically rigorous from a design and analysis standpoint.[24] For the collection of economic data alongside RCTs, this suggests that data collection efforts focus on 'big ticket' items, rather then trying to capture a complete picture of resource use. To ensure that the data collected in these trials are of high quality, a greater emphasis is needed on centralized training of study coordinators and records abstractors and on the centralized quality control mechanisms.

The high cost associated with data collection may further necessitate linkage of trial data to claims data, which requires collection of sensitive patient identifiers such as Social Security numbers. This may raise additional patient confidentiality concerns with Institutional Review Boards and may require collaboration and cooperation of a government agency, or private insurers to ensure patient-level claims can be made available.

4.4 The need for good research practices for collecting and reporting RW data

Our review underscores the need for good practices for collecting and reporting RW data. In terms of data collection, it is important that efforts follow well-established research practices. These include posing well-defined questions, specifying timeframes for the duration of data collection, conducting periodic monitoring to ensure quality and responsiveness to research questions, and limiting sample sizes to the minimum necessary. These good practices should also ensure that informed consent and human subject protections are in place.

Registries, for example, should be carefully planned, beginning with clear and achievable objectives and extensive clinical input. A protocol or study plan should guide the conduct of the registry and the data analysis. Data collection tools should be carefully designed and tested, and site coordinators thoroughly trained to ensure data quality. Quality control mechanisms should be in place to identify and resolve study issues. There is also a need to balance research needs with privacy concerns as it becomes easier to link data across multiple sources.

Researchers should draw inferences from observational data with caution. For example, a widely recognized challenge with the analysis of real world data is the need to correct for sample selection bias—defined as the bias introduced by unobserved variables that are correlated with both treatment selection and patient outcomes (e.g., unobserved illness severity). On the other hand, it is less well recognized that selection bias is just one of a variety of factors including missing variables, measurement error, incorrect functional form and two-way causation ('structural simultaneity') that can introduce bias into treatment estimates. All of these problems have the common statistical result of introducing a correlation between the treatment variable and the error term, which is the source of the bias. Overcoming this "endogeneity bias"

is the most common and vexing challenge that arises in the analysis of real world data. Statistical tests for endogeneity bias exist, along with methods for its correction although operationalizing these methods is often a significant challenge: in many circumstances, our ability to minimize this bias is limited by the lack of exogenous variation in the available data.[25, 26]

4.5 Good process in using RW data in coverage and reimbursement decisions

We also recognize the importance of the processes implemented by decision makers in using RW data in coverage and reimbursement. Observers point to several conditions for establishing good process, including transparency (the decision and the rationale for making them must be publicly accessible) and relevance (there must be a reasonable explanation for a decision's rationale).[27] They should also be fair in the sense that RW data will be used in similar fashion across technologies, or if situations demand a different rationale, the circumstances or principles would be known. Decisions should not be "bureaucratically arbitrary," or based on reasons that people do not view as meaningful or just.

Typically, the amount of RW data available for the initial coverage and reimbursement decision is very limited: instead, effectiveness projections are modeled from efficacy data. Good process would encourage the subsequent gathering of RW data, in part by updating coverage and reimbursement decisions based on those data.

Processes should also allow opportunity for stakeholder participation. Different stakeholders (physicians, patients, family caregivers, payers, hospitals, regulatory agencies,

employers, manufacturers, and researchers) may have different perspectives on the use and value

of RW data. Ideally, there will be a mechanism for challenge and change, which contributes to

democratic governance.

4.6 The need to consider the costs and benefits of data collection

Two critical questions are how much RW data should be collected and who should pay

for it.. Evidence costs money. Inevitably, there are questions about whether resources devoted

to the endeavor are worthwhile. There is a need to prioritize decisions about RW data such that

the benefits of collecting additional information are expected to outweigh the costs.

The tool of "value-of-information (VOI) analysis" offers a formal approach to deciding

when and what types of data to collect. Formal use of decision analysis and VOI analysis can

help determine whether an intervention should be adopted, whether additional evidence to

further inform that decision is worth gathering, and what kind of information is of greatest

value.[28]

VOI analysis evaluates the extent to which new evidence might improve expected

benefits by reducing the chance for error, and compares that improvement to the cost of the

information. Evidentiary considerations will depend on the particular circumstances of a

decision (the consequences of an error, what can be learned from additional evidence, how new

knowledge will change and improve the option identified as optimal), rather than on

predetermined specifications about the type of study design (e.g., RCT). Explicit consideration

is given to the potential positive health outcomes forgone due to delays in coverage as well as

the potential adverse consequences of too rapid uptake when the risk-benefit ratio is highly

uncertain.

4.7 The need for modeling

A previous ISPOR Task Force on modeling studies emphasized that the purpose of

modeling is to "structure evidence on clinical and economic outcomes in a form that can help

inform decisions. . . ."[29] They stated: "Models synthesize evidence on health consequences and

costs from many different sources, including data from clinical trials, observational studies,

insurance claim databases, case registries, public health statistics, and preference surveys."

Our field has adopted bio-clinical cost-effectiveness models as an integrative framework,

incorporating biological, clinical outcomes, and economic data into a decision-making structure.

Within this general structure, there are several different types of possible pharmacoeconomic

evaluations, such as cost-consequences, cost-utility, and cost-benefit analyses. Such models and

analyses are the primary vehicle for combining RCT and RW data to support coverage and

reimbursement decision-making. At the same time, expanded use of RW data in assessing post-

launch cost-effectiveness is needed to update the modeling projections made to support initial

coverage and reimbursement decisions.

In terms of good research practice, the previous task force emphasized the need for

transparency of assumptions, including the point that "conclusions are conditional upon the

assumptions and data on which the model is built."

4.8 The need for ongoing dialog

Finally, our review highlights the need for ongoing stakeholder dialog on all of these issues. Implicit in much of our discussion is a central policy question about the appropriate role of the public sector in producing and judging evidence. Who collects and evaluates RW data, who pays for these efforts, and what magnitude of resources is provided are key questions for policy makers worldwide. There is no general regulatory requirement for collecting real world data.

In the U.S., CMS has called for data that reflect real world practice. In addition, FDA is requiring the implementation of mandatory registries in instances where there is any concern regarding long-term safety of a therapy. The opportunity for funding observational studies of therapy is limited compared to traditional epidemiological studies and therapeutic trials. It is critical that we have an intelligent and vigorous public discussion on these and other issues.

5. CONCLUSIONS

Real world data are essential for sound coverage, payment, and reimbursement decisions. The types and applications of such are varied, and context matters greatly in determining the value of a particular type of evidence in any circumstance. Different study designs can provide useful information in different situations. Randomized controlled trials remain the gold standard for demonstrating clinical efficacy in restricted trial setting, but other designs—such as observational registries, claims databases, and practical clinical trials—can contribute to the evidence base needed for coverage and payment decisions.

It is critical that policy makers recognize the benefits, limitations, and methodological challenges in using RW data, and the need to carefully consider the costs and benefits of different forms of data collection in different situations. In this report, we have attempted to identify the salient issues involved in real world data. We acknowledge many challenges ahead and view this effort as a starting point for future debates.

POSTSCRIPT

We received and reviewed over 70 comments from ISPOR members on our draft report, which was posted on the ISPOR website in the summer of 2006. Space does not allow us to print or address all of them here, though they are posted at

http://www.ispor.org/workpaper/RWD_TF/MemberComments.pdf . One recurring theme was a perceived need for more guidance on precisely how decision makers will actual use real world data after the launch of a product. Others themes included calls for more discussion of limitations of various aspects of RW data collection, calls for more attention to the uses of electronic health records, and comments that our report could have used more emphasis on international data collection, and to medical devices. Although we made numerous changes to this document in light of these comments, we still refer the interested reader to them, as several of the important broader themes and limitations they raise should help to define the agenda for the next steps in this continuing inquiry.

ACKNOWLEDGMENTS

References

1. Center for Medicare and Medcaid Services. The Medicare Prescription Drug, Improvement, and Modernization Act. P.L. 108-173. 12-8-2003. Washington, D.C.

2. Effective Heath Care Program. Agency for Healthcare Research and Quality . 3-28-2006. Accessed: 7-13-2006. Available at: http://effectivehealthcare.ahrq.gov/aboutUS/index.cfm.

3. Neumann PJ. Evidence-based and value-based formulary guidelines. Health Aff (Millwood) 2004; 23(1):124-134.

4. AMPC Format for Formulary Submissions, Version 2.1. Academy of Managed Care Pharmacy . 2005. Accessed: 7-13-2006. Available at: http://www.amcp.org/amcp.ark?c=pr&sc=link.

5. Drug Effectiveness Review Project . 2006. Accessed: 7-13-2006. Available at: http://www.ohsu.edu/drugeffectiveness/reports/final.cfm.

6. National Institute for Health and Clinical Excellence . 7-13-2006. Accessed: 7-13-2006. Available at: http://www.nice.org.uk/.

7. IQWIG. German Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen . 7-4-2006. Accessed: 7-13-2006. Available at: http://www.iqwig.de/.

8. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996; 312(7023):71-72.

9. Woolf SH. Evidence-based medicine and practice guidelines: an overview. Cancer Control 2000; 7(4):362-367.

10. Wild D, Grove A, Martin M et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005; 8(2):94-104.

11. Acquadro C, Berzon R, Dubois D et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. Value Health 2003; 6(5):522-531.

12. Guidance for Industry Patient Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. US Department of Health and Human Services, Food and Drug Administration . 2006. Accessed: 7-13-2006. Available at: http://www.fda.gov/cber/gdlns/prolbl.pdf.

13. Szende A, Leidy NK, Revicki D. Health-related quality of life and other patient-reported outcomes in the European centralized drug regulatory process: a review of guidance

documents and performed authorizations of medicinal products 1995 to 2003. Value Health 2005; 8(5):534-548.

14.	Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. JAMA 1995; 274(22):1800-1804.

15.	United States Department of Health and Human Services.Agency for Health Care Policy and Research. Acute Pain Management: Operative or Medical Procedures and Trauma. 92-0023, 107. 1993. Rockville, MD, ACHPR. Clinical Practical Guideline No 1.

16.	Harris RP, Helfand M, Woolf SH et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001; 20(3 Suppl):21-35.

17.	Scottish Intercollegiate Guidelines Network . 6-7-2006. Accessed: 7-17-2006. Available at: http://www.sign.ac.uk/.

18.	Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ 2001; 323(7308):334-336.

19.	Clinical Statements and Guidelines. American College of Cardiology Foundation . 7-10-2006. Accessed: 7-13-2006. Available at: http://www.acc.org/clinical/statements.htm.

20.	O'Sullivan AK, Thompson D, Drummond MF. Collection of health-economic data alongside clinical trials: is there a future for piggyback evaluations? Value Health 2005; 8(1):67-79.

21.	Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003; 290(12):1624-1632.

22.	Effective HealthCare, DEcIDE: Registries for Evaluating Patient Outcomes. Agency for Healthcare Research and Quality . 7-17-2006. Accessed: 7-18-2006. Available at: http://effectivehealthcare.ahrq.gov/decide/registryOutline.cfm.

23.	Guidance for the Public, Industry, and CMS Staff: National Coverage Determinations with Data Collection as a Condition of Coverage: Coverage with Evidence Development. Centers for Medicare and Medicaid Services. 7-12-2005. Accessed: 7-23-2006. Available at: https://www.cms.hhs.gov/mcd/ncpc_view_document.asp?id=8.

24.	Glasgow RE, Magid DJ, Beck A, Ritzwoller D, Estabrooks PA. Practical clinical trials for translating research to practice: design and measurement recommendations. Med Care 2005; 43(6):551-557.

25.	D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998; 17(19):2265-2281.

26. Crown WH, Obenchain RL, Englehart L, Lair T, Buesching DP, Croghan T. The application of sample selection models to outcomes research: the case of evaluating the effects of antidepressant therapy on resource utilization. Stat Med 1998; 17(17):1943-1958.

27. Daniels N, Sabin J. Setting Limits Fairly: Can We Learn to Share Medical Resources? New York: Oxford University Press; 2002.

28. Claxton K, Cohen JT, Neumann PJ. When is evidence sufficient? Health Aff (Millwood) 2005; 24(1):93-101.

29. Weinstein MC, O'Brien B, Hornberger J et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies. Value Health 2003; 6(1):9-17.