CHAPTER 3

Data Acquisition and Management

Data play a key role in testing scientific theories or hypotheses and form the backbone of scientific inference. The different steps of research should be monitored carefully, and research design should include built-in safeguards to ensure the quality, objectivity, and integrity of research data. This chapter addresses ethical issues pertaining to data acquisition and management, including hypothesis formation, research design, data collection, data analysis, data interpretation, data storage, and data sharing.

Scientific research is the systematic attempt to describe, explain, and understand the world. Though different disciplines study different aspects of the natural world, they share some common methods that are designed to produce objective knowledge by subjecting hypotheses or theories to rigorous tests (see table 3.1). Ideas that cannot be tested, such as metaphysical theories, ideological claims, and private intuitions, are not scientific. Some (but not all) tests involve experiments. In an experiment, a researcher attempts to reduce the number of variables and control the conditions in order to understand statistical or causal relationships between variables or parameters. For an experiment to be rigorous, a researcher must describe it in enough detail that other researchers can obtain similar results by repeating the experimental conditions (Cheny 1993; Kirk 1995; Kitcher 1993; Popper 1959; Resnik 2007; Shamoo and Annau, 1987).

All test results in science, whether from controlled experiments, field observations, surveys, epidemiological studies, computer models, or meta-analyses, should be open to public scrutiny and debate. Peer review, with some limitations, is one of science's most important methods

because it promotes the public scrutiny of hypotheses, theories, and test results (see chapter 7 for further discussion). Once a hypothesis or theory becomes well established, it may be said to be a "fact." For example, the idea that the sun is the center of the solar system is now accepted as a fact, but it was a hypothesis during the time of Copernicus (1542 [1995]). Well-established generalizations, such as Newton's laws of motion and the ideal gas laws, are known as laws of nature (Giere 1991; Hempel 1965; Popper 1959; Resnik 1998a).

Data (or recorded observations) play a key role in testing scientific theories or hypotheses and form the backbone of scientific inference. Data can take many forms, including observations recorded by scientists in laboratory notebooks, field notes, entries into electronic notebooks or spreadsheets, outputs from machines (such as optical scanners, gas chromatographs, or automated DNA sequencers), photographs, x-rays, video or audio recordings, transcribed interviews, digital images, computer printouts and databases, historical documents, and case reports in clinical trials. Data include the primary (or original or source) data, which are drawn directly from the experiment or test. These include entries in a laboratory notebook, field notes, computer printouts, photographs, machine outputs, and so on. Secondary (or derived) data are data based on primary data, such as spreadsheets derived from entries into laboratory notebooks, or figures or diagrams based on machine outputs. Data are different from research materials, such as chemical reagents, biological samples (blood, tissue, urine, etc.), cell lines, slides, gels, rocks, laboratory animals, and others. To illustrate the difference between data and materials, consider a project to sequence an organism's genome. The materials would include biological samples from the organism (blood, tissue, cells, etc.). The data would include the genomic information derived from these materials (i.e., deoxyribonucleic acid [DNA] sequences, such as GTTAGATTCCA, etc.). In this chapter we will examine various ethical issues pertaining to the acquisition and management of data that arise at different stages of research.

PROBLEM SELECTTION

Ethical issues arise at the very first stage of research, because problem selection is often affected by funding and politics (Resnik 2007, 2009a). Though most scientists choose problems based on their curiosity and professional interests, research costs a great deal of money, and scientists usually end up working on problems that sponsors are willing to pay for. In the private sector, profit plays a major factor in funding decisions.

Table 3.1. STAGES OF SCIENTIFIC RESEARCH

1.	Select a problem or question to investigate.
2.	Review the relevant literature.
3.	Propose a hypothesis to solve the problem or answer the question.
4.	Design and plan experiments or other procedures to test the hypothesis.
5.	Collect and record data.
6.	Analyze data.
7.	Interpret data.
8.	Disseminate results.

A drug company will consider how a research project is likely to affect its bottom line in making funding decisions. For example, a company may prefer to fund research on a drug for a common illness, such as hypertension, than to fund research on a treatment for a rare disease, because it will make more money from treating a common illness. Though some private companies, such as Bell Laboratories, have sponsored basic research, most focus on applied research pertaining to their products or services. Scientists who conduct research for private companies will need to come to terms with their research agendas.

Politics can affect funding in many different ways. First, politics usually impacts a funding agency's research priorities. The National Institutes of Health (NIH), for example, establishes different priorities for research on different diseases, such as cancer, HIV/AIDS, and so forth, and for research pertaining to different areas of study, such as aging, mental health, and allergies. Second, politics sometimes impacts specific research projects. For example, since the 1980s the U.S. government has banned the use of federal funds for research on human embryos. In 2001, the G. W. Bush administration imposed significant restrictions on research involving human embryonic stem cells, though the Obama administration lifted some of these restrictions in 2009 (Obama 2009; Resnik 2009a). In 2003, a congressional committee held a hearing on 198 NIH-funded research projects on human sexuality, HIV/AIDS, and drug abuse that the Traditional Values Coalition said was a waste of the taxpayers' money (Resnik 2009a). Third, disease-specific advocacy bolstered by lobbying from the pharmaceutical industry influences the area of research as well as the amount of funding to a specific area. Though we believe that public funding of research should be as free as possible from politics, we call attention to this issue so that

scientists can be mindful of how the choice of a research topic may impact one's ability to receive funding. We will discuss funding issues again in chapters 5, 7, and 12.

LITERATURE SEARCH

The literature search can be an important early step in the overall project. This is an important step for the investigator because it can save a great deal of time and money by eliminating a flawed objective or a hypothesis. It can also help researchers to learn whether their projects may make an original or worthwhile contribution or whether they merely repeat previous work or would result in knowledge that has little value. A literature search can also help researchers learn about previously used methods, procedures, and experimental designs and can place the project's experimental design and protocol within the known realities of the subject matter. A thorough literature search can allow researchers to give proper credit to others who have already worked in the area. Failing to acknowledge other relevant work is arrogant and self-serving and is a type of plagiarism or serious bias if one knowingly or unknowingly claims to be the originator of someone else's idea (Resnik 1998b; Shamoo 1992).

An inadequate literature search in clinical research can lead to tragic results. Ellen Roche died while participating in an experiment designed to produce a mild asthma attack in healthy (nonasthmatic) volunteers at Johns Hopkins University. Roche inhaled hexamethonium, a blood pressure medication used in the 1950s and 1960s. Roche developed a cough and breathing difficulties and was put on a ventilator. She died because of extensive lung damage produced by the hexamethonium. An Office of Human Research Protections investigation of Roche's death determined that this tragedy probably could have been avoided if the principal investigator, Alkis Togias, had consulted articles published in the 1950s (and cited in subsequent publications) warning of lung damage due to inhaling the hexamethonium. Togias did a standard PubMed search on hexamethonium and consulted current textbooks, but this literature search did not include references from the 1950s (Savulescu and Spriggs 2002).

HYPOTHESIS FORMATION

After selecting a problem and reviewing the literature, researchers should formulate a hypothesis (or hypotheses or theories) to test. They may also

need to formulate aims or objectives for the research project. It is important for hypotheses, aims, and objectives to be testable, because an important part of the scientific method is subjecting hypotheses to rigorous tests. If a hypothesis is not testable, researchers may waste time and money collecting data that have no clear value. To ensure that a hypothesis is testable, researchers need to state it clearly, avoiding ambiguous terms. They also need to derive predictions from the hypothesis for different tests (e.g., "hypothesis H predicts chemical X will increase the rate of tumor formation in laboratory mice"). If the predictions occur, they may confirm the hypothesis; if they do not, they may disconfirm it. In the past, most research projects were hypothesis-driven, that is, researchers formulated hypotheses prior to gathering data. Today, research is often data-driven, that is, researchers formulate hypotheses after gathering data. For example, researchers in the fields of genomics and proteomics may analyze large datasets in order to discover statistical associations among different variables. In psychology, sociology and epidemiology researchers often conduct cross-sectional studies to obtain some baseline data pertaining to a population.

One of the problems with data-driven research is that it may lead researchers to make up post-hoc hypotheses to explain patterns in the data. To avoid this problem, researchers must ensure that hypotheses in data-driven research are testable. They may also want to conduct more tests to provide additional, independent evidence for or against their hypotheses.

RESEARCH DESIGN

The design of experiments is one of these crucial steps in preserving the integrity, quality, and objectivity of the research project. In this stage of research, scientists should clearly describe experiments or other tests (e.g., surveys, focus groups, etc.) they will perform, the materials they will use, and the procedures, methods, and protocols they will follow. They should also describe their plan for collecting, recording, and analyzing the data, including the use of any statistical methods. The research should be described in sufficient detail so that someone not involved in the project can evaluate it and repeat the work. The research design should be based on one's previous research, existing literature, laboratory manuals, and other appropriate sources. Researchers should follow appropriate standards in applying methods and should keep records of what methods they use and how they use them. During initial tests, researchers should use and identify standard (or well-established) methods, but they can modify

these methods to suit new experimental applications or testing procedures. It is important for researchers to note changes they make and to state the reasons for them. Furthermore, researchers should not make changes in the middle of a test or experiment, because this will bias or corrupt the data. All accidental changes, such as dropping a test tube, should be noted in the laboratory notebook. Researchers should not pick and choose among experiments or tests to achieve a desired result. However, they may do so if they recognize a variable inherent in the protocol that was not first recognized in earlier stages of the project. For example, in testing a new drug in humans, researchers may realize that an unanticipated side effect should be recorded and could therefore change the protocol and then design a new experiment that measures this side effect. However, researchers should record these decisions and discuss them in detail at the same time and place where the experiments are recorded, derived, or manipulated.

Because it is easy to employ research designs that tend to bias the data and results, scientists should be mindful of how biases may affect their work and they should take steps to minimize the potential for bias. Since biases may operate at a subconscious level, researchers may not even be aware that their studies may be flawed. Scientists, like all human beings, are susceptible to self-deception (Broad and Wade 1982 [1993]). Because it is not always possible to see the biases in one's own work, it is important to solicit critical feedback from colleagues prior to the initiation of a study and also after its completion. Biased research wastes time, money, and effort, and it can also involve the unnecessary use of human or animal subjects. Sound experimental design is also one of the key ethical principles of research with animals and human subjects (Irving and Shamoo 1993; Levine 1988). Because no amount of statistical analysis or interpretation can overcome a design flaw, data that result from a flawed design are virtually useless, and using them can be unethical (Irving and Shamoo 1993; Resnik 2000).

Since there are many different ways that biases can affect research, we cannot discuss them all here. We will, however, call attention to some common biases. First, sometimes the experimental conditions may affect the data. For example, an in vitro experiment to determine how a chemical affects cell signaling may be affected by many different factors that scientists may not be aware of, such as subtle changes in temperature, humidity, PH, electrolytes, impurities in the chemical, or the growth medium. Montagnier's misconduct allegation against Gallo (discussed in chapter 2) probably resulted from a vigorous HIV strain that contaminated different cell lines used by two researchers. Temperature, feeding,

and other living conditions (such as overcrowding) may affect how laboratory animals respond to stimuli. Environmental factors (such as privacy, time of day) may affect the answers that human subjects provide during interviews. Second, a poor statistical design may impact research outcomes. A common statistical problem is inadequate sample size. For example, if scientists find that there is no difference between how a potential neurotoxin affects 20 laboratory mice, as compared to 20 controls, the sample size may not be large enough to demonstrate that the chemical has no neurological effects. One may need a larger sample to reach conclusions that have statistical significance. Third, exclusion/inclusion criteria may bias outcomes in clinical research involving human subjects. For example, testing an erectile dysfunction drug on healthy male subjects aged 18–50 may overestimate the efficacy of the drug in the general population, because many of the men who take the drug will be over 50 and will have health problems (such as a decline in stimulus response due to obesity or hypertension or other unknown factors). Fourth, survey questions may introduce subtle biases. For example, a question like "Do you think that President Obama is not doing enough to deal with the country's crippling federal deficit problem?" may generate a different response from this question worded slightly differently, "Do you agree or disagree with the way President Obama is dealing with the federal deficit?" Fifth, private companies may intentionally introduce biases into their experimental designs in order to promote their economic interests (Crossen 1994; Porter 1993; Resnik 2007). We will discuss biases related to privately funded research in greater depth in chapters 5 and 9.

COLLECTING, RECORDING, AND STORING DATA

After one has designed a research project, the next step is to collect, record, and store the data. Scientists should keep accurate records of all aspects of the research project, including data, protocols, and methods (including any changes); drafts of manuscripts; and correspondence with institutional officials, funding agencies, and journal editors. Good scientific record keeping is important for ensuring the quality and integrity of research for numerous reasons. First, good record keeping is essential for conducting your own research. Members of the research team need access to records to conduct experiments or tests, analyze data, make reports, draft manuscripts, and so on. Second, good record keeping is important for authentication of your work by outside parties, such as peer reviewers, or scientists who want to reanalyze your data or

replicate or build on your work. Third, good record keeping is crucial for investigating allegations of research misconduct and other problems. Indeed, good scientific records can be your best defense against a misconduct allegation. If other scientists have trouble replicating your results, they may suspect that you have committed misconduct if you keep poor records (see the Imanishi-Kari case, discussed in chapter 2). Fourth, detailed and accurate record keeping is essential for proving ownership of legal claims related to patents and copyrights (we discuss intellectual property in greater depth in chapter 8). Fifth, good record keeping is legally required for research that is submitted to the Food and Drug Administration (FDA) and other regulatory agencies. Sixth, good record keeping is needed for accurate auditing and quality assurance. Although this may sound strict to some, we believe that research records can be viewed as quasi-legal documents analogous to medical records, business inventories, or investment accounts (Shamoo 1989, 1991a, 1991b).

Although different disciplines, laboratories, and research groups have different record-keeping styles and formats, the following guidelines apply generally. First, records should be accurate and thorough. Records should include what was done (i.e., data and results), how it was done (i.e., methods and materials), when it was done, why it was done, who did it, and the next steps. Records should be signed and dated. If laboratory notebooks are used, all additive information directly relevant to the raw data, such as derived data, tables, calculations, or graphs, should be either done directly in the notebook or taped thoroughly on an adjacent page in the notebook. If this is not feasible, files can be used; providing clear identification of the data and the page where the data were derived from is essential. Ideally, entries should also be signed (or initialed) and dated. If electronic notebooks are (or other types of electronic records) are used, these should include an electronic date trail to allow for accurate information concerning the identity of individuals who enter data and entry time (Schreier et al. 2006).

Second, records should be well organized. Researchers should be able to keep track of their records and know where and how they are kept. A laboratory data notebook should be bound and the pages numbered consecutively. Loose-leaf notebooks are hazardous and may tempt a beginning researcher or technician to tear off pages with mistakes. If electronic notebooks are used, these should be properly filed and linked to other records. Large collaborative projects involving different laboratories, research groups, or institutions sometimes hire a data manager to help organize and keep track of all the data (Schreier et al. 2006).

Third, records should be clear, legible, and recorded in the language collectively used by the research group (e.g., English). All entries in a laboratory notebook should be made legibly with permanent, nonerasable ink. Researchers should draw a line through a mistaken entry, without making it completely illegible, and should not use correction fluid (Schreier et al. 2006).

Fourth, records should be secure. Paper records should be stored in a secure place. Electronic records should be protected against unauthorized use or hacking. Access to research records should be restricted to members of the research team or to institutional officials who have the right to review them. Although many researchers take data with them when they change jobs, we strongly recommend that research institutions keep copies of all raw data while allowing individuals to have copies. Some universities follow the example of private industry and treat research data as the property of the institution. Keeping the data within the institution is important so that future interested parties can check the original data against derived data, graphs, or published results (Schreier et al. 2006).

Fifth, records should be backed up as a safeguard against destruction or theft. Data recorded on older formats (such as computer diskettes) should be transferred to newer formats (such as CDs or computer servers) so that they will be readable (Schreier et al. 2006).

Sixth, records should be kept for the appropriate length of time. Keeping records for seven years from the time of a last expenditure report or publication is a good general rule, although some records (such as FDA-regulated research records) may need to be kept longer (National Academy of Sciences 1994; National Institutes of Health 2008b; Shamoo and Teaf 1990). In the event that a federal agency or sponsor audits and inquires about the data, data storage should be automatically extended for the needed length of time.

Although it is important to store research records and materials, storage introduces problems of space allocation. Some of these problems can be handled by transferring records to digital formats, but computer storage space may also be limited. Few universities have provisions for storing records or materials in centralized facilities. We recommend that research institutions develop archives for records and materials and require researchers to make deposits on a regular basis. The federal government can and should provide funding to develop resources for data storage, such as GenBank, which stores genomic data. Researchers who create banks for storing biological samples may have to deal with space allocation issues.

Seventh, supervisors, mentors, laboratory directors, and other scientists who are responsible for mentoring students or leading research groups have responsibilities related to good record keeping. Supervisors

and mentors should instruct students on how to keep good records for their research projects. They should have regular meetings with members of the research team to review data and address concerns. Laboratory directors should provide some record-keeping rules within their laboratory (Schreier et al. 2006).

Eighth, quality assurance procedures should be used to correct errors related to data entry or processing. Primary data are usually processed through many stages, depending on the type of research, before they are presented as graphs, charts, or tables or in a publishable form. As data are processed, the risk of introducing (intentional or unintentional) biases, adjustments, or errors increases (Grinnell 1992; Shamoo 1989, 1991a, 1991b).

Some recent studies indicate that academic researchers are not doing a good job of record keeping. In a survey of 1,479 researchers funded by the NIH (2007a), Martinson et al. (2005) found that the most prevalent (27.5%) self-reported inappropriate behavior was "inadequate record-keeping." Moreover, one in ten had withheld details in publications, used an inadequate experimental design, or dropped data. At 90 major research institutions, 38% of research integrity officers reported encountering problems with research records during misconduct inquiries and investigations, which often delayed investigations or made them impossible to complete (Wilson et al. 2007). In a survey conducted at the National Institute of Environmental Health Sciences (NIEHS), 31% of 243 researchers said that they had encountered poor record keeping at the NIEHS (Resnik 2006).

DATA ANALYSIS

The analysis of data in modern science involves the application of various statistical techniques, such as correlation, regression, analysis of variance (ANOVA), t-tests, and chi-square tests. These techniques provide a way of drawing inductive inferences from data and distinguishing any real phenomena or effects from random fluctuations. A responsible researcher will make every attempt to draw unbiased inferences from data. Statistical practices vary a great deal across different disciplines. Most fields have accepted practices for data analysis, and it is prudent for researchers to follow these norms (Resnik 2000). There is nothing inherently unethical in the use of unconventional statistical methods. It is important, however, to be forthright in clearly stating the method of analysis, why it is being used, and how it differs from others. It is unethical to fail to disclose

important information relevant to the data analysis, such as assumptions made concerning populations or parameters or computer programs used (Resnik 2000).

Given the complexities of data analysis, it is easy to introduce biases or other errors in the analysis and to misrepresent the data (Bailar 1986). The failure to provide an honest and accurate analysis of the data can have as significant an impact on research results as recording data improperly. Moreover, research indicates that statistical errors are fairly common in science (DeMets 1999). Thus, this step is crucial for ensuring the objectivity, integrity, and quality of research. Some aspects of data analysis that raise ethical concerns include excluding outliers, imputing data (i.e., using a statistical method to fill in missing data), editing data, analyzing databases for trends and patterns (or data mining), developing graphical representations of the data, and establishing the statistical and practical significance of the data. While none of these areas of data analysis are inherently deceptive, biased, or unethical, researchers must be sure to follow good statistical practices and honestly describe their statistical methods and assumptions to avoid errors in data analysis (American Statistical Association 1999). Intentionally misrepresenting the data can be regarded as a type of misconduct (Resnik 2000).

A problem common in many research disciplines is deciding whether to report and analyze all of the data collected as part of a research project. It may be the case that not all of the data collected as part of a study are relevant to the overall results. For example, some data may be corrupted due to human or experimental error. Some data may be statistical outliers (generally two standard deviations from the mean) that may skew the analysis. Sometimes researchers may decide to take a project in a different direction so that not all of the data collected will be relevant to the results. Researchers may also conduct small pilot studies to establish the feasibility of a larger study. Honesty is an important ethical concern when one is deciding whether to report or analyze all of the data. As noted in chapter 2, exclusion of data that impact one's overall results is a type of misconduct known as falsification. However, reporting and analyzing all of the data collected as part of a study may also be problematic if the data are not relevant. Researchers must use good judgment when dealing with these issues.

Consider the case of the physicist Robert Millikan (1868–1953), who won the Nobel Prize in Physics in 1923 for measuring the smallest electrical charge (i.e., the charge on an electron). Millikan's famous oil-drop experiment involved spraying oil drops through electrically charged plates. When a drop was suspended in the air, the electrical force pulling up on

the drop was equal to the force of gravity pulling it down. Millikan was able to determine the charge on an electron by calculating these forces. In his paper describing this experiment, Millikan said that he had reported all the data. However, the science historian Gerald Holton (1978) examined Millikan's laboratory notebooks and found that Millikan did not report 49 out of 189 observations (26%) that were marked as "poor" in the notebook. Though some commentators, such as Broad and Wade (1982), have argued that Millikan's work was fraudulent, a plausible explanation for his conduct is that he had a good understanding of his experimental apparatus and was therefore able to determine when it was not working properly. The "poor" results he excluded may have involved oil drops that were too big or too small for accurate measurements. However, one could argue that Millikan should have discussed this issue in the paper and that it was dishonest to claim that he had reported all the data.

Another area of concern is the treatment of digital images, such as pictures of proteins from gel electrophoresis or cell structures. Computer programs, such as Photoshop, can enhance the quality or clarity of digital images. In some cases, researchers have manipulated images in order to deceptively change the image to produce a desired result. To deal with this potential problem, many journals have adopted requirements for the submission of images for publication (Couzin 2006). Journals usually require researchers to submit the original images so they can be compared to the enhanced images. The Office of Research Integrity has special instructions on its website for forensic tools to detect fraud in images (Office of Research Integrity 2007b). Researchers should be aware of and use these tools when necessary. While it is acceptable to use image-manipulation technologies to make it easier for researchers to perceive patterns in an image, it is not acceptable to manipulate an image in order to mislead or deceive other researchers. The *Journal of Cell Biology* has adopted the following guidelines, which we endorse:

> No specific feature within an image may be enhanced, obscured, moved, removed, or introduced. The grouping of images from different parts of the same gel, or from different gels, fields, or exposures must be made explicit by the arrangement of the figure (i.e., using dividing lines) and in the text of the figure legend. If dividing lines are not included, they will be added by our production department, and this may result in production delays. Adjustments of brightness, contrast, or color balance are acceptable if they are applied to the whole image and as long as they do not obscure, eliminate, or misrepresent any information present in the original, including backgrounds. Without any background information, it is not possible to see exactly how much of the

original gel is actually shown. Non-linear adjustments (e.g., changes to gamma settings) must be disclosed in the figure legend. All digital images in manuscripts accepted for publication will be scrutinized by our production department for any indication of improper manipulation. Questions raised by the production department will be referred to the Editors, who will request the original data from the authors for comparison to the prepared figures. If the original data cannot be produced, the acceptance of the manuscript may be revoked. Cases of deliberate misrepresentation of data will result in revocation of acceptance, and will be reported to the corresponding author's home institution or funding agency. (*Journal of Cell Biology 2007*)

DATA INTERPRETATION

If all researchers interpreted the data in the same way, science would be a dry and dull profession. But this is not the case. Many important and heated debates in science, such as research on firearm violence, studies of intelligence tests, and studies of global warming, involve disputes about the interpretation of data. Sometimes an important discovery or advance in science occurs as the result of a new interpretation of existing data. Of course, challenging a standard interpretation of the data is risky: Those who challenge the existing paradigm either go down in flames or win the Nobel Prize. Most challenges to the existing paradigm turn out to be wrong. But those few times that the new interpretation is correct can change and advance our knowledge in a revolutionary fashion. For example, Peter Mitchell won the Nobel Prize for his chemiosmotic theory. He advanced the notion that a proton gradient across the mitochondrial membrane is the driving force to synthesize adenosine triphosphate (ATP) from adenosine diphosphate (ADP) and inorganic phosphate. The chemiosmotic theory was originally considered heresy because it contradicted the long-held theory of a phosphorylated intermediate for the synthesis of ATP.

The path of a trailblazer is full of hazards. Most researchers, despite their image as being open-minded and liberal, resist new ideas and stick to generally accepted standards. Although revolutions do occur in science, most research conforms to the model of "normal" science—science that falls within accepted standards, traditions, and procedures (Kuhn 1970). It is often the case that researchers who have new interpretations are scoffed at before their ideas are accepted. For example, the idea of continental drift was viewed as ludicrous, as was the idea that a bacterium could cause ulcers. However, if researchers can find new ways of interpreting data, they should be encouraged to do so. Their new interpretations

will be more readily accepted (or at least considered) if they properly acknowledge the existing paradigm (Resnik 1994).

Even within the existing paradigm, the interpretation of the same data can take very different pathways, none of which are likely to be unethical. As we discussed in chapter 2, there is an important distinction between misconduct and disagreement. Just because one researcher disagrees with another's interpretation does not mean that one of them is being dishonest (Resnik and Stewart 2012). It is especially important for researchers with new interpretations to be even more careful about documenting and leaving a thorough paper trail of their data, so that other researchers will be able to understand their interpretations and not dismiss them as resulting from fraud or error. Ensuring the integrity of research data does not mean straitjacketing the investigator's creativity and latitude in introducing new ideas and interpretation. However, prudence suggests that all interpretations of data should be consistent with the existing knowledge. If the interpretation of new data is inconsistent with existing knowledge, an honest discussion of the differences is in order.

One common ethical problem with data interpretation is what we will call "overreaching." Researchers overreach when they claim that their data are more significant or important than they really are. This problem often occurs with industry-funded pharmaceutical research (Resnik 2007). For example, suppose that a study shows that a new analgesic medication is 2% more effective at reducing arthritis pain compared to acetaminophen and 4% more effective than aspirin. However, the new medication also increases systolic and diastolic blood pressure by 10% in about 30% of the people who take it. Because its patent has not expired, the new medication will be much more expensive than acetaminophen or aspirin. The researchers would be overreaching if they claimed that the new medication is superior to acetaminophen and aspirin, because the medication brings a marginal improvement in pain relief but has some dangerous side effects. Overreaching can be an ethical problem in clinical research if it causes physicians to prescribe new medications to their patients without considering costs or side effects (Angel 2004). Overreaching can be a significant issue in research with public policy implications if it supports unwise decisions.

PUBLISHING DATA

We discuss publication issues in more detail in chapters 5, 7, and 9. For now, we simply note that researchers have an obligation to disseminate work for the obvious reason that science cannot advance unless researchers

report and share results. Dissemination can include publication in peer-reviewed journals, monographs or other books, and web pages, as well as presentations at professional meetings. The important ethical consideration is that research should be disseminated to colleagues and the public for scrutiny and review. Indeed, researchers who receive grants from the government or private funding agencies are usually required to specify a plan for disseminating their research in the grant proposal and to report to the agency about publications that result from the grant (Grinnell 1992). However, researchers who work for business and industry or the military often sign agreements to not publish results or to withhold publication until they obtain approval from management (Blumenthal 1997; Gibbs 1996). For instance, researchers working for the tobacco industry did not publish their work on nicotine's addictive properties for many years (Resnik 1998b). Pharmaceutical companies have also suppressed data pertaining to their products (Resnik 2007).

SHARING DATA AND MATERIALS

As noted in chapter 1, openness is a key principle in research ethics. Scientists should share data, results, methods, and materials to (a) promote the advancement of knowledge by making information publicly known; (b) allow criticism and feedback as well as replication; (c) build and maintain a culture of trust, cooperation, and collaboration among researchers; and (d) build support from the public by demonstrating openness and trustworthiness. While openness is considered by many people to be a fundamental part of academic research and scholarship, the real world of research does not always conform to this ideal. Although researchers share data within the same team of collaborators working on a common project, they rarely share data with noncollaborators and often do not welcome requests to share data with other researchers in the field, much less with people from outside the research community. The resistance to data sharing is especially high among researchers who have concerns about intellectual property, such as potential patents or trade secrets, but resistance is also high among researchers who want to protect their own interests in claiming priority (to be first) for discoveries or publishing original research.

Several recent studies have documented problems with data sharing in biomedical science. In a survey by Campbell et al. (2002) of academic geneticists concerning their experiences with data withholding, 47% stated that at least one of their requests to share data or research materials related to published research had been denied in the last three years; 28% reported that

they had been unable to confirm published research due to refusals to share data or materials; and 12% said that they had denied a request to share data or materials. Of those who refused to share data or materials, 80% said they refused because sharing required too much effort; 64% said they refused to share to protect someone else's ability to publish; and 53% wanted to protect their own ability to publish (Campbell et al. 2002). Another survey (Blumenthal et al. 2006) found that 32% of biomedical researchers had engaged in some type of data withholding during the last three years and that data withholding is common in the biomedical sciences.

Although refusals to share data and materials appear to be common, especially in biomedical sciences, some organizations have adopted policies that require researchers to share data and materials following publication. Many government granting agencies, such as the NIH and National Science Foundation (NSF), encourage or require researchers to share data and materials. The NIH expects intramural and extramural researchers to share data as widely and freely as possible (National Institutes of Health 2003). The NIH also has policies that encourage or require funded researchers to share reagents and model organisms (e.g., transgenic animals). The NIH also requires researchers to state their plans to share data, reagents, or organisms in their grant applications or to explain any proposed restrictions on sharing (National Institutes of Health 1998a, 2003). The NIH has a genome-wide association studies (GWAS) policy that establishes a repository for all GWAS data obtained with NIH funding (National Institutes of Health 2009c).

Many scientific journals have also created policies that require researchers to share supporting data or materials as a condition of publication. Many journals have websites where researchers can deposit data and other supporting materials that do not appear in a published article. For example, *Science* requires researchers to share data and materials. The journal asks researchers to deposit large databases on a publicly available website prior to publication and to share data and materials after publication (*Science* 2007).

While the progress of science thrives on sharing data and materials as soon as possible, there are some legitimate reasons to refuse to share data or materials, at least temporarily, such as the following:

1. To protect a researcher's interests in publishing articles from the data or materials. If a researcher collects data or develops materials for a project, she should not have to share the data or materials until she is ready to publish, since sharing prior to publication may impact her ability to publish. But once a researcher has published, she has an obligation to share. A difficult question arises when a researcher has acquired

a large database and hopes to publish a series of papers from the database. Should the researcher be required to share the whole database as soon as she publishes the first paper from it? If she must share the whole database with other investigators, this could jeopardize her ability to publish other papers from it, because the other investigators might beat her to it. One way of handling this dilemma is to allow a researcher to publish a specific number of papers from her database before releasing the entire database to the public. Another solution is for researchers with databases to collaborate with other researchers when they share data, so that they both can receive publication credit. Difficult questions also can arise with sharing research materials, since sharing materials with others can jeopardize one's prospects of publishing articles based on those materials. Also, if the materials are in limited supply and cannot be re-created, then researchers must decide how to allocate the materials. For example, a blood sample is a limited quantity—once it has been used up, it is gone. To protect their own ability to use the sample in research, investigators need to decide carefully whom to share it with.

2. To protect intellectual property claims. Sometimes investigators are conducting research that may be patentable. Sharing data or other information related to the research prior to submitting a patent application can jeopardize the patent. Thus, researchers may refuse to share data in order to protect potential patents. It is important for society to protect patent rights to stimulate invention and private investment in R&D (Resnik 1998b). We discuss intellectual property issues in more depth in chapter 8.
3. To protect a researcher's reputation. Researchers may not want to share data because they are not ready to present it to the public. They may need to do quality-control checks on the data or analyze it. A researcher may fear that his reputation could be damaged if he publishes data prematurely and there are problems with it. Charles Darwin [1809–1882] waited more than 20 years to publish his theory of evolution by natural selection so that he could solidify the arguments and evidence in favor of the theory and anticipate objections.
4. To protect confidential information pertaining to human subjects (discussed in more depth in chapter 11), trade secrets (discussed in more depth in chapter 5), or national security (discussed in more depth in chapter 12).
5. To avoid wasting time, effort, and money. Sometimes it takes a great deal of time, effort, or money to share data or materials with other researchers. There are significant costs with answering requests, shipping

materials, taking care of animals, and synthesizing chemicals. One way of dealing with this problem is to deposit data on a public website or to license a private company to make data or materials available to other researchers. Whenever data or materials are shared, a reasonable fee can be charged to cover the costs of sharing.

6. To avoid being hassled by industry or political interest groups. Sometimes industry representatives will request data in order to reanalyze the data or reinterpret the results. For example, if a study finds that exposure to a pesticide increases the risk of Parkinson's disease, the manufacturer of the pesticide might want to acquire the data to reanalyze it or challenge the study. Political interest groups, such as animal rights activists, may also request data or other information to harass or intimidate researchers. While these requests can sometimes be legitimate attempts to advance scientific knowledge, they often are not.

Researchers who are considering refusing a request to share data or materials should use their good judgments to make an ethical choice. While the default ethical standard should be to share data and materials as soon as possible once research is completed, researchers may decide not to share data and materials in cases where other concerns (such as protecting confidentiality or career interests) outweigh the ethics of openness.

In the United States, if federally funded researchers refuse to share data (or other information), outside parties may still be able to obtain the data under the Freedom of Information Act (FOIA). Other countries, such as the United Kingdom, have similar laws. FOIA allows the public to gain access to recorded information gathered or generated using federal funds, including scientific research records. To gain access to information under FOIA, one must send a request in writing to the head of the appropriate federal agency asking for the records that are sought. One must also specify the records being sought and explain why they are being sought. The agency should respond to this request within 20 days by sending the documents, promising to send the documents within a reasonable time, or explaining why they cannot be sent. The agency may charge a reasonable fee for sending the records. There are some exceptions to FOIA: Agencies can refuse to share records pertaining to national security or foreign relations, agency rules or practices, confidential business information, information related to personal privacy, some types of law enforcement records, and information pertaining to the supervision of financial institutions. Federal authorities have determined that some of these exceptions apply to federally funded scientific research. For example,

researchers do not have to disclose confidential information pertaining to human subjects. They also do not have to disclose information protected by trade secrecy law, including information pertaining to potential patents (U.S. Department of Justice 2007).

Some scientists have objected to FOIA on the grounds that it could subject them to harassment from people who want to interfere with their work (Macilwain 1999). Although it is important for researchers to be free from harassment from industry representatives, political activists, or other parties, we do not think that researchers who receive public funds can be completely shielded from this threat. It is difficult to know in advance whether any particular request for information would be harassment of researchers. Without having this knowledge in advance, any policy short of answering all requests for data would be arbitrary and possibly biased.

Public access to federally supported research has reached the public domain in the past few years. In an editorial, the *New York Times* (2013) opined that if we (i.e., the public) paid for it we should have access to it. The U.S. government has instructed all federal agencies with more than $100 million in expenditures on research to have a plan submitted to the government on how it will provide public access to the data. Berman and Cerf (2013) have proposed that public access to the data should be the result of private-sector partnerships.

QUESTIONS FOR DISCUSSION

1. How would you characterize scientific research? In your opinion, what is the most crucial part of research?
2. How would you list the steps in carrying out research? Are there some steps you could skip? Why? Is there a particular order to doing the steps?
3. Can scientific research incorporate quality control and quality assurance methods? Would this stifle creativity or increase workload?
4. Can you give an example of how one might modify data to suit inappropriate goals in the steps of research?
5. Can you give an example of an experimental design that would bias the data?
6. What principles or rules do you follow related to research record keeping?
7. Do you keep good records? Could someone reproduce your work from your research records (laboratory notebook, etc.)?

8. How is a lab notebook like (or not like) a business or medical record?
9. Can you give an example of an ethical or scientific issue you have faced concerning data analysis or interpretation?
10. When would you be justified in refusing to share data?

CASES FOR DISCUSSION

CASE 1

A medical student has a summer job with a faculty mentor at a research university. The student is bright, hardworking, and industrious and hopes to publish a paper at the end of the summer. He is the son of a colleague of the mentor at a distant university. The student is working on a cancer cell line that requires three weeks to grow in order to test for the development of a specific antibody. His project plan is to identify the antibody by the end of the summer. The student has written a short paper describing his work. The mentor went over the primary data and found that some of the data were written on pieces of yellow pads without clearly identifying from which experiment the data came or the data. She also noticed that some of the experiments shown in the paper's table were repeated several times without an explanation as to why. The mentor was not happy about the data or the paper, but she likes the student and does not want to discourage him from a potential career in research.

- What is the primary responsibility of the mentor?
- Should the mentor write a short paper and send it for publication?
- Should the student write a short paper and send it for publication?
- If you were the mentor, what would you do?
- Should the mentor or her representative have paid more attention to the student's work during the course of the summer?

CASE 2

A graduate student at a research university finished her dissertation and graduated with honors. Her mentor gave the continuation of the project to a new graduate student. As usual, the mentor gave the entire laboratory notebook (or computer disk) to the new graduate student, who had to repeat the isolation of the newly discovered chemical entity with high-pressure liquid chromatography (HPLC) in order to follow up the chemical and physical characterization of the new compound. The new graduate student found that if he followed the exact method described in the laboratory notebooks and published by the previous student, he could obtain the new chemical

entity but not at the same HPLC location as published, but slightly shifted to the left, and there was a different peak at the location stated. However, the new student discovered that if the ionic strength is doubled, he could find the same chemical at the same location in accordance with the previous student's dissertation. The new student discussed with the mentor how he should proceed. The mentor replied, "Why make a fuss about it? Just proceed with your slightly different method and we can move on."

- What are the responsibilities of the new student? Should the new student refuse to accommodate the mentor's request?
- Should the new student have read more thoroughly the relevant laboratory notebooks prior to starting the experiment? Should there have been a paper trail of the error in the laboratory notebook? Do you think the error was intentional, and does it matter?
- If the laboratory notebook does not reveal the error, is it then misconduct? Does it indicate that a better recording of the data would have been helpful?
- Can you propose a reasonable resolution to the problem?

CASE 3

A new postdoctoral fellow in a genetic research laboratory must sequence a 4-kDa fragment. After the sequence, he is to prepare a 200-base unit to use as a potential regulator of a DNA-related enzyme. The 4-kDa fragment is suspected to contain the 200-base unit. The sequence of the 200-base unit is already known in the literature, but not as part of the 4-kDa fragment and not as a potential regulator. The fact that the 200-base unit is known is what gave the mentor the idea that it may have a functional role. The new postdoctoral fellow tried for three months to sequence the 4-kDa fragment, without success, and so simply proceeded to synthesize the 200-base unit without locating it within the fragment. After two years of research, the 4-kDa fragment appeared to play a key regulatory role in an important discovery, but at this time the mentor learned that the postdoc never sequenced the original 4-kDa fragment. The mentor could never find a "good" record of the attempts to sequence the 4-kDa fragment.

- What impression do you gather about how this mentor runs the laboratory?
- Should there be records of sequence attempts of the 4-kDa fragment?
- Are there reasons to suspect that data may have been fabricated?
- How should the mentor proceed?
- If you were the new postdoc, what steps you would take to ensure proper records of your work?

CASE 4

A graduate student prepared for her thesis a table showing that a toxic substance inhibits an enzyme's activity by about 20%. She has done 12 experiments. The mentor looked at the data and found that one of the data points showed an inhibition of 0% and that this point is the one that skewed the results to a low level of inhibition with a large standard of deviation. The mentor further determined with the student that the outlier is outside the mean by 2.1 times the standard derivation and that it is reasonable not to include it with the rest of the data. This would make the inhibition about 30% and thus make the potential paper more in line with other research results and hence more "respectable." The mentor instructed the student to remove the statistical outlier from the data.

- Should the student simply proceed with the mentor's instructions?
- Should the mentor have been more specific regarding what to do with the outlier? In what way?
- Can you propose a resolution? Should the outlier be mentioned in the paper?
- How should this laboratory handle similar issues in the future? Should each laboratory have an agreed-upon standard operating procedure, or SOP, for such a statistical issue?

CASE 5

A social scientist is conducting an anonymous survey of college students on their opinions on various academic integrity issues. The survey is administered in four different sections of an Introduction to Sociology class. The survey includes 20 questions in which respondents can use a Likert scale to answer various questions: 1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = disagree, and 5 = strongly disagree. The survey also includes 10 open-ended questions that ask for respondents to state their opinions or attitudes. The social scientist distributes 480 surveys and 320 students respond. A graduate student helps the social scientist compile the survey data. When examining the surveys, the student encounters some problems. First, it appears that eight surveys are practical jokes. The persons filling out these surveys wrote obscene comments and for many questions added extra numbers to the Likert scale. Although some of the 20 Likert-scale questions in these surveys appear to be usable, others are not. Second, in 35 surveys, the respondents appeared to have misunderstood the instructions on how to use the Likert scale. They answered "5" on questions where it would seem that "1" would be the most logical answer, given their written comments. Third, on 29 surveys, the respondents wrote their names on the survey, when they were instructed not to do so.

- How should the researchers deal with these issues with their data?
- Should they try to edit/fix surveys that have problems?
- Should they throw away any surveys? Which ones?
- How might their decisions concerning the disposition of these surveys affect their overall results?

CASE 6

A pharmaceutical company conducts five small (20 subjects) phase I studies on a new drug to establish its safety in healthy individuals. Three of these studies had a p-value < 0.05, indicating significant results; two had a p-value > 0.05, indicating nonsignificant results. As it so happens, undesirable side effects were observed in both studies with the nonsignificant results but not in the other studies. The researchers report all their results to the FDA but they do not report all of these results in a publication. The publication only reports significant results.

- Are there any design problems with these studies?
- Is there an ethical responsibility to report all of the data? Would it make a difference if the subjects were not human (i.e., animals)?
- Is not reporting nonsignificant results falsification?
- What are the responsibilities of the researchers to this company, to themselves, and to society?
- Should there be a federal mandate to report all side effects to the public?

CASE 7

Dr. Heathcliff is a toxicologist testing the effects of an industrial compound that is used in manufacturing plastic food containers and that functions like testosterone in the body. The study involves two groups of laboratory mice: one is fed the compound each day, and a control group is not fed the compound. The main outcome measure is aggressive behavior, which is known to be linked to testosterone activity. He completes that study and finds that the animals fed the compound displayed significantly more aggressive behavior than the control group. He submits the paper for publication the following week. A technician assisting with the experiment discovers that the heating system in the area where animals are kept was malfunctioning one of the nights prior to their main set of observations, which took place that following morning. The temperature in the cages where the animals were kept was as much as 5°C higher than normal, according

to the maintenance crew. The technician informs Dr. Heathcliff about this problem, who replies that "the temperature probably didn't make any difference since the animals were at normal temperature in the morning. And besides, the control group was not unusually aggressive; only the experimental group displayed above normal aggression."

- Should Dr. Heathcliff inform the journal about this issue?
- Should he include information about the temperature issue in the methods or discussion section of the paper?
- Should he withdraw the paper?
- Should he repeat the experiments?
- Would it be unethical to publish the paper in its present form?
- Would this be misconduct (i.e., data fabrication or falsification)?
- What should the technician do?

CASE 8

A graduate student in physics is writing a thesis that develops a mathematical model of gamma ray bursts. The student conducts a literature review on the subject as a background to her research. In conducting this review, she searches various computer databases for articles and abstracts relevant to her work, for the past five years. She gathers many abstracts and papers. For much of the research, she reads only abstracts and not the full papers. Also, she does not include some of the important work on gamma ray bursts that took place more than five years ago.

- Should the graduate student read the full articles, not just abstracts?
- If she cites an article in a publication or in her thesis, should she read the full article?
- If she cites a book, should she read the full book or only the part that she cites?
- Should the graduate student include articles published more than five years ago?

CASE 9

Dr. Reno is a junior investigator who has just received her first major NIH grant. She has used NIH funds to create a transgenic mouse model to study depression. The mouse has a genetic defect that leads to an underproduction of serotonin. She has used the model to show how a compound found in an herbal medicine increases serotonin levels in mice and also produces effects associated with normal (i.e.,

nondepressed) behavior, such as normal levels of exercise and normal sleep patterns. Dr. Reno applies for a patent on this compound with the intent of eventually testing it on human subjects. She also publishes a paper in a high-impact journal describing her work with the mouse model. Almost immediately after the paper appears in print, she receives dozens of requests from researchers who would like to use her transgenic mice in their own research. Dr. Reno is flattered but also overwhelmed. She has barely enough mice for her own work, and she doesn't want to turn her laboratory into a factory for producing mice for someone else's research.

- How should Dr. Reno deal with these requests to share transgenic mice?
- Does she have an obligation to share the mice?

CASE 10

Drs. Kessenbaum and Wilcox are conducting a long-term, observational study of the health of pesticide applicators. The protocol calls for an initial health assessment, including a health history, physical exam, and blood and urine tests. The researchers will collect a DNA sample from cheek scrapings and collect dust samples from the applicators' clothing and hair and underneath their fingernails. After the initial health assessment, the applicators will complete yearly health surveys and undergo a full health assessment every four years. The researchers will follow the subjects for at least 25 years. Their work is funded by the NIH. Drs. Kessenbaum and Wilcox have been conducting their study for 15 years, and they have compiled an impressive database. They have already published more than a dozen papers from the database. Whenever they share data, they require researchers who request it to sign elaborate data-sharing agreements, which spell out clearly how the data will be used. The agreements also specify the kinds of studies that can be published using the data, which allows Drs. Kessenbaum and Wilcox to protect their interests in publishing on certain topics. In the past month, they have received some requests to access their database. One request has come from a pesticide company, another has come from a competing research team also studying the health of pesticide applicators, and another has come from a radical environmental group with an antipesticide agenda.

- How should Drs. Kessenbaum and Wilcox handle these requests to access their database?
- Should they refuse to share data with the pesticide company or the environmental group?
- Is it ethical to require people who request data to sign elaborate data-sharing agreements?