Cancer Drug Discovery and Development

Manuel Hidalgo
Elizabeth Garrett-Mayer
Neil J. Clendeninn
S. Gail Eckhardt *Editors*

Principles of Anticancer Drug Development

Humana Press

Cancer Drug Discovery and Development

Series Editor: Beverly A. Teicher, Genzyme Corporation
Framingham, MA, USA

For other titles published in this series, go to
www.springer.com/series/7625

Principles of Anticancer Drug Development

Edited by

Manuel Hidalgo, MD, PhD

Universidad CEU San Pablo, Spain

S. Gail Eckhardt, MD

University of Colorado at Denver, USA

Elizabeth Garrett-Mayer, PhD

*Medical University of South Carolina,
South Carolina, USA*

Neil J. Clendeninn, MD, PhD

CANAID, Inc., Hanalei, USA

Springer

Editors
Manuel Hidalgo, MD, PhD
Department of Oncology
School of Medicine
CEU San Pablo University
Madrid, Spain
and
Centro Integral Oncológico Clara Campal
(CIOCC)
Madrid, Spain
and
Gastrointestinal Clinical Research Unit
Centro Integral Investigaciones Oncológicas
(CNIO)
C/ Melchor Fernández Almagro 3
Madrid, Spain
mhidalgo@cnio.es

S. Gail Eckhardt, MD
Professor and Division Head,
Medical Oncology Stapp
Harlow Chair in Cancer Research
University of Colorado at Denver
Aurora, CO 80045
USA
Gail.Eckhardt@UCDenver.edu

Elizabeth Garrett-Mayer, PhD
Hollings Cancer Center
Medical University of South Carolina
Charleston, SC 29425
USA
garrettm@musc.edu

Neil J. Clendeninn, MD, PhD
CANAID, Inc
Drug Development Consultant
96714 Hanalei
USA
cybermad@msn.com

ISBN 978-1-4419-7357-3 e-ISBN 978-1-4419-7358-0
DOI 10.1007/978-1-4419-7358-0
Springer New York Dordrecht Heidelberg London

While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Blurb

A practical guide to the design, conduction, analysis and reporting of clinical trials with anticancer drugs.

Preface

The development of cancer drugs, from the preclinical studies to final randomized clinical trial is a science per se. The process involves multiples distinct steps and requires the participation of multiple individuals with unique expertise such as toxicologist, pharmacologists, pathologists, statisticians, clinicians and ethics and regulatory experts. In addition, it involves several different organizations such as academic center, pharmaceutical industry and governmental organizations. Teaching and learning this process is not a simple task. The editors are seasoned clinical investigators with different backgrounds who spend considerable time teaching student and junior colleagues the nuts and bolts of drug development both at their own institutions but also through active participations in workshops and seminars on the topic. It became apparent to us that there are no sources in which the basis and principles of drug development are concisely summarize. This book has been written to fill that gap and to provide a guide for the beginners of drug development as well as a consultation manual for more advance drug developers. It is intended to provide a practical tool for the design, conduction, analysis and reporting of a clinical trial as well as to establish a developmental plan for a new agent.

The book is organized into five parts – all of them written by experts and renowned authors who have done a great job putting the chapters together. Part I summarizes basic concepts in biostatistics and in clinical and analytical pharmacology that are needed to understand the clinical drug development process. Part II provides a comprehensive summary of preclinical studies that are required before a medical agent can be tested in humans. Part III deals with clinical trial design from phase I to phase III as well as with correlative studies in clinical trials including the more classic pharmacokinetics and the newer molecular imaging and tissue biomarkers. Part IV is an important section that outlines the FDA requirement for testing and approving a drug for cancer treatment. Part V focuses on more specific descriptions of developmental strategies for the different classes of anticancer agents ranging from conventional cytotoxic agents to molecularly targeted agents. The final section outlines the resources and perspective of the National Cancer Institute.

We expect this book to be a night table manual and guide for those interested in the complex but rewarding field of anticancer drug development and the place to get started when training in this field. We also hope that this text book would be useful to our peer teachers in drug development.

Manuel Hidalgo
S. Gail Eckhardt
Elizabeth Garrett-Mayer
Neil J. Clendeninn

Contents

Part I

1 Basic Biostatistics for the Clinical Trialist 3
Elizabeth G. Hill and Elizabeth Garrett-Mayer

2 Fundamental Concepts in Clinical Pharmacology 37
Daniel L. Gustafson and Erica L. Bradshaw-Pierce

3 Bioanalytical Methods in Clinical Drug Development 63
Walter J. Loos, Peter de Bruijn, and Alex Sparreboom

Part II

4 Preclinical Models for Anticancer Drug Development 89
Edward A. Sausville

Part III

5 Phase I Clinical Trials with Anticancer Agents 117
Stephen Leong, Justin Call, Alex Adjei, and Wells Messersmith

6 Phase II Trials with Anticancer Agents 141
Hui K. Gan, J. Jack Lee, and Lillian L. Siu

7 Phase III Clinical Trials with Anticancer Agents 163
Wendy R. Parulekar and Daniel J. Sargent

8 Pharmacokinetic Studies in Early Anticancer Drug Development 189
Alex Sparreboom and Sharyn D. Baker

9 Pharmacodynamic Studies in Early Phase Drug Development 215
D. Ross Camidge, Robert C. Doebele, and Antonio Jimeno

10 Prediction of Antitumor Response 257
Fred R. Hirsch and Yu Shyr

11 Imaging Studies in Anticancer Drug Development 275
David A. Mankoff

Part IV

12 **Role of the US Food and Drug Administration in Cancer Drug Development** 305
Ann T. Farrell, Ramzi N. Dagher, and Richard Pazdur

Part V

13 **Early Clinical Trials with Cytotoxic Agents** 335
M.J.A. de Jonge and Jaap Verweij

14 **Challenges and Successes in Developing Effective Anti-angiogenic Agents** 347
Laura Q.M. Chow and S. Gail Eckhardt

15 **Targeted Therapeutics in Cancer Treatment** 403
Colin D. Weekes and Manuel Hidalgo

16 **Cancer Chemoprevention** 463
Christopher H. Lieu, William N. William Jr, and Scott M. Lippman

17 **Combined Modality Therapy in Cancer Management** 483
David Raben and Kyle Rusthoven

18 **Cancer Vaccines** 519
Daniel Laheru

19 **Optimising the Development of Antibodies as Treatment for Cancer** 535
Craig P. Carden, Hendrik-Tobias Arkenau, and Johann S. de Bono

20 **Oligonucleotide Therapeutics** 569
Cy A. Stein, Britta Hoehn, and John Rossi

21 **Anticancer Drug Development in Pediatric Patients** 589
Lia Gore and Margaret Macy

22 **Clinical Trials in Special Populations** 603
S. Percy Ivy, Merrill J. Egorin, Chris H. Takimoto, and Jeannette Y. Wick

Part VI

23 **NCI-Sponsored Clinical Trials** 631
Andriana Papaconstantinou and Janet E. Dancey

Index 659

Contributors

Alex A. Adjei, MD, PhD, FACP
Department of Medicine, Roswell Park Cancer Institute, Elm &
Carlton Streets,
Buffalo, NY 14263, USA
alex.adjei@roswellpark.org

Dr. Hendrik-Tobias Arkenau, MD, PhD
The Medical Professorial Unit, Prince of Wales Medical School, University
of New South Wales, Level 1, South Wing, Edmund Blacket Building,
Avoca Street, Sydney, NSW, 2031, Australia
ht.arkenau@unsw.edu.au

Sharyn D. Baker, PharmD, PhD
Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital,
262 Danny Thomas Place, CCC Room I5308, Mail Stop 313, Memphis,
TN, 38105-3678, USA
sharyn.baker@stjude.org

Erica L. Bradshaw-Pierce, PhD
Department of Pharmacokinetics, Dynamics and Metabolism, Pfizer Global
Research and Development, 10646 Science Center Drive, San Diego, CA 92121
Erica.Pierce@pfizer.com

Justin Call, MD
Medical Oncology, Developmental Therapeutics Program/GI Malignancies,
University of Colorado Cancer Center, Mail Stop 8117, PO Box 6511, Aurora,
CO, 80045, USA
justin.call@uchsc.edu

D. Ross Camidge, MD, PhD
Developmental Therapeutics and Thoracic Oncology Programs, Clinical Thoracic
Oncology Program, University of Colorado Cancer Center, Aurora, CO, USA
ross.camidge@udenver.edu

Craig P. Carden, MBBS, FRACP
Drug Development Unit, Section of Medicine, The Institute
of Cancer Research, The Royal Marsden Hospital NHS Trust,
Sutton, London, SM2 5PT, UK
craig.carden@icr.ac.uk

Laura Q.M. Chow, MD, FRCPC
Division of Medical Oncology, Department of Medicine,
University of Washington, 825 East lake Avenue East (SCCA) MS:
G4-940, Campus Box 358081, Seattle, Washington, USA 98109-1023
lchow@seattlecca.org; lqchow@u.washington.edu; laurachow@hotmail.com

Ramzi N. Dagher, MD
Worldwide Regulatory Strategy, Global Regulatory, Pfizer Inc., 50 Pequot Avenue MS6025-C5141, New London, CT, 06320, USA; Oncology Business Unit, Pfizer Inc., 50 Pequot Avenue MS6025-C5141, New London, CT, 06320, USA
ramzi.dagher@pfizer.com

Janet E. Dancey, MD, FRCPC
Investigational Drug Branch, Cancer Therapy Evaluation Program,
Division of Cancer Treatment and Diagnosis, National Cancer Institute,
6130 Executive Blvd, EPN 7131, Rockville, MD 20892, USA
danceyj@ctep.nci.nih.gov

Peter de Bruijn, BSc
Laboratory of Translational Pharmacology, Department of Medical Oncology,
Erasmus University Medical Center, Groene Hilledijk 301, PO Box 5201,
3008 AE, Rotterdam, The Netherlands
p.debruijn@erasmusmc.nl

Dr. Johann S. de Bono, MD, FRCP, MSc, PhD
Section of Medicine and Drug Development Unit, Institute of Cancer Research,
Royal Marsden Hospital, Downs Road, Sutton, Surrey, SM2 5PT, UK
jdebono@icr.ac.uk

M.J.A. de Jonge, MD, PhD
Department of Medical Oncology, Erasmus University Medical Center (Daniel den Hoed Kliniek), P.O. Box 5201, 3008 AE, Rotterdam, The Netherlands
m.dejonge@erasmusmc.nl

Robert C. Doebele, MD
Division of Medical Oncology, Department of Medicine, University of Colorado Cancer Center, University of Colorado at Denver Anschutz Medical Campus, Aurora, CO, USA
robert.doebele@ucdenver.edu

S. Gail Eckhardt, MD
Professor and Division Head, Medical Oncology Stapp Harlow Chair in Cancer Research University of Colorado at Denve, Aurora, CO 80045, USA
Gail.Eckhardt@UCDenver.edu

Merrill J. Egorin, MD, FACP
University of Pittsburgh Cancer Institute, Room G27E, Hillman Research Pavilion
5117 Centre Avenue, Pittsburgh, PA, 15213-1863, USA
egorinmj@upmc.edu

Ann T. Farrell, MD
Division of Hematology Products, Office of Oncology Drug Products (OODP),
Center for Drug Evaluation and Research, US Food and Drug Administration,
Silver Spring, MD, 20993-0002, USA
ann.farrell@fda.hhs.gov

Hui K. Gan, MBBS, FRACP, PhD
Drug Development Programme, University Avenue, Room 5-224, Princess
Margaret Hospital 610, Toronto, ON, M5G 2M9, Canada
huikgan@gmail.com

Elizabeth Garrett-Mayer, PhD
Hollings Cancer Center, Medical University of South Carolina,
Charleston, SC 29425, USA
garrettm@musc.edu

Lia Gore, MD, FAAP
The Children's Hospital, Center for Cancer and Blood Disorders and
Developmental Therapeutics Program, University of Colorado Cancer Center,
Box B115 13123 East 16th Avenue, Aurora, CO, 80045, USA
Lia.Gore@ucdenver.edu

Daniel L. Gustafson, Ph.D
Department of Clinical Sciences, Colorado State University, ACC226, Veterinary
Teaching Hospital, 300 West Drake Road, Fort Collins, CO, 80523-1620, USA;
Pharmacology Core, CU Cancer Center, Colorado State University, ACC226,
Veterinary Teaching Hospital, 300 West Drake Road, Fort Collins, CO, 80523-1620,
USA; CSU Animal Cancer Center, Colorado State University, ACC226, Veterinary
Teaching Hospital, 300 West Drake Road, Fort Collins, CO, 80523-1620, USA
Daniel.Gustafson@colostate.edu

Manuel Hidalgo, MD, PhD
Department of Oncology, School of Medicine, CEU San Pablo University,
Madrid, Spain; Centro Integral Oncológico Clara Campal (CIOCC), Madrid,
Spain; Gastrointestinal Clinical Research Unit, Centro Integral Investigaciones
Oncológicas (CNIO), C/ Melchor Fernández Almagro 3, 28029, Madrid, Spain
mhidalgo@cnio.es

Elizabeth G. Hill, PhD
Biostatistics Core, Hollings Cancer Center, Medical University of South
Carolina, 86 Jonathan Lucas Street, Room 118D, MSC 955, Charleston, SC,
29425 – 9550, USA
hille@musc.edu

Fred R. Hirsch, MD, PhD
Department of Medical Oncology, University of Colorado Cancer Center, Aurora,
CO 80045, USA
Fred.Hirsch@ucdenver.edu

Britta Hoehn, PhD
City of Hope National Medical Center, 1450 E. Duarte Road, Duarte,
CA, 91010, USA
brittahoehn@gmx.de

S. Percy Ivy, MD
Investigational Drug Branch, Cancer Therapy Evaluation Program, Division
of Cancer Treatment and Diagnosis, National Cancer Institute, 6130 Executive
Blvd, Suite 7131, Rockville, MD 20852, USA
ivyp@ctep.nci.nih.gov

Antonio Jimeno, MD, PhD
Medical Oncology, University of Colorado Cancer Center, Mail Stop 8117,
PO Box 6511, Aurora, CO, 80045, USA;Developmental Therapeutics/
Pharmacodynamic Laboratory, Developmental Therapeutics, Head and Neck
Cancer and Stem Cell Programs, University of Colorado Cancer Center, Mail
Stop 8117, PO Box 6511, Aurora, CO 80045, USA
Antonio.Jimeno@ucdenver.edu

Dan Laheru, MD
Skip Viragh Center for Pancreas Cancer Clinical Research and Patient Care,
The Sol Goldman Pancreatic Cancer Research Center, The Sidney Kimmel
Comprehensive Cancer Center Bunting-Blaustein, The Johns Hopkins University
School of Medicine, CRB Room G89, 1650 Orleans Street, Baltimore,
MD 21231, USA
laherda@jhmi.edu

J. Jack Lee, PhD, MS, DDS
Division of Quantitative Sciences, Department of Biostatistics, The University
of Texas M.D. Anderson Cancer Center, 1400 Pressler Street, Unit Number:
1411 Room Number: FCT4.6012, Houston, TX, 77030, USA
jjlee@mdanderson.org

Stephen Leong, MD
Medical Oncology, Developmental Therapeutics Program/GI Malignancies,
University of Colorado Cancer Center, Mail Stop 8117, PO Box 6511, Aurora,
CO, 80045, USA
Stephen.Leong@UCHSC.edu

Christopher H. Lieu, MD
Department of Thoracic/Head and Neck Medical Oncology, Unit 432,
The University of Texas MD Anderson Cancer Center, Houston, TX,
77030-4009, USA

Scott M. Lippman, MD
Department of Thoracic/Head and Neck Medical Oncology,
The University of Texas MD Anderson Cancer Center,
1515 Holcombe Blvd., Unit 432, Houston, TX
77030-4009, USA
slippman@mdanderson.org

Walter J. Loos, Ph.D
Laboratory of Translational Pharmacology, Department of Medical Oncology,
Erasmus University Medical Center, Groene Hilledijk 301, PO Box 5201, 3008
AE, Rotterdam, The Netherlands
w.loos@erasmusmc.nl

Margaret Macy, MD
The Children's Hospital, Center for Cancer and Blood Disorders,
13123 East 16th Avenue, Aurora, CO 80045, USA
Macy.margaret@tchden.org

David A. Mankoff, MD, PhD
Department of Radiology, Seattle Cancer Care Alliance, G2-600, 825 Eastlake
Avenue East, Seattle, WA 98102, USA
dam@u.washington.edu

Wells Messersmith, MD, FACP
GI Medical Oncology Program, University of Colorado Cancer Center, Mail Stop
8117, 12801 East 17th Avenuem, Aurora, CO 80045, USA; Division of Medical
Oncology, University of Colorado Cancer Center, Mail Stop 8117, 12801 East
17th Avenue, Aurora, CO, 80045, USA
wells.messersmith@ucdenver.edu

Andriana Papaconstantinou, PhD
Technical Resources International Inc., 6500 Rock Spring Drive, Suite 650,
Bethesda, MD, 20817, USA
adpapacon@gmail.com

Wendy R. Parulekar, MD, FRCPC
NCIC Clinical Trials Group, Queen's University, Kingston, ON, Canada;
Department of Oncology, Queen's University, Kingston, ON, Canada
wparulekar@ctg.queensu.ca

Richard Pazdur, MD
Office of Oncology Drug Products, Center for Drug Evaluation and
Research, U.S. Food and Drug Administration, Silver Spring, MD,
20993-0002, USA
richard.pazdur@fda.hhs.gov

David Raben, MD
Department of Radiation Oncology, University
of Colorado Denver, Anschutz Cancer Pavilion, MS F-706, 1665 N. Ursula St.,
Suite 1032, Aurora, CO 80045-0510, USA
david.raben@uchsc.edu

John Rossi, PhD
Department of Molecular and Cellular Biology, Dean, Irell and Manella
Graduate School of Biological Sciences, Beckman Research Int. of City of Hope,
1500 East Duarte Road, Duarte, CA, 91010, USA
jrossi@coh.org

Kyle Rusthoven, MD
Department of Radiation Oncology, University of Colorado Health Sciences
Center, 1665 N. Ursula St., Suite 1032, Denver, CO, 80045-0508, USA
Kyle.Rusthoven@UCHSC.edu

Daniel J. Sargent, PhD
Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN,
55905, USA
Sargent.Daniel@mayo.edu

Edward A. Sausville, MD, PhD
Marlene and Stewart Greenebaum Cancer Center, University of Maryland, 22 S.
Greene Street, Room S9D07, Baltimore, MD, 21201, USA
esausville@umm.edu

Yu Shyr, Ph.D
Cancer Research, Vanderbilt University School of Medicine, Nashville, TN, USA;
Division of Cancer Biostatistics, Vanderbilt University School of Medicine,
Nashville, TN, USA; Cancer Biostatistics Center, Vanderbilt-Ingram Cancer Center,
2220 Pierce Avenue, 571 Preston Building, Nashville, TN 37232-6848, USA
Yu.Shyr@Vanderbilt.edu

Lillian L. Siu, MD
Division of Medical Oncology and Hematology, Princess Margaret Hospital,
University of Toronto, 610 University Avenue, Suite 5-718, Toronto,
ON, Canada, M5G 2M9
lillian.siu@uhn.on.ca

Alex Sparreboom, PhD
Department of Medical Oncology, Erasmus MC, Rotterdam, The Netherlands
and
Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital,
Memphis, TN, USA
alex.sparreboom@stjude.org

Cy A. Stein, MD, PhD
Department of Oncology, Albert Einstein-Montefiore Cancer Center, Montefiore Medical Center, 111 E. 210 Street, Bronx, NY 10467, USA
cstein@montefiore.org

Chris H. Takimoto, MD, PhD
Translational Medicine, Ortho Biotech Oncology R&D, 145 King of Prussia Road, Mail Stop RA-2-2, Radnork, PA, 19087, USA
CTakimot@its.jnj.com

Colin D. Weekes, MD, PhD
Developmental Therapeutics Program/GI Oncology, University of Colorado Health Science Center, Mail Stop 8117 RC1 South, Rm 8123 12801 E. 17th AvenueAuroraP.O. Box 6511, CO, 80045, USA; Shana Spears, Denver, CO, USA
colin.weekes@ucdenver.edu.

Jeannette Y. Wick
Pharmaceutical Management Branch, Cancer Therapy Evaluation Program, National Cancer Institute, Rockville, MD, USA
wickj@ctep.nci.nih.gov

William N. William Jr, MD
Department of Thoracic/Head and Neck Medical Oncology, Unit 432, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030-4009, USA

Jaap Verweij, MD, PhD
Dept. of Medical Oncology, Erasmus University Medical Center, PO Box 2040, 3000 CA, Rotterdam, The Netherlands; Gravendijkwal 230, 3015 CE, Rotterdam, The Netherlands
j.verweij@erasmusmc.nl

Part I

Chapter 1
Basic Biostatistics for the Clinical Trialist

Elizabeth G. Hill and Elizabeth Garrett-Mayer

1.1 Introduction

The purpose of this chapter is to acquaint the reader with some typically used biostatistical principles and methods in anticancer drug development for summarizing and analyzing data. Understanding and properly interpreting statistics is critically important for drug development. Each stage of development, ranging from preclinical studies to phase III clinical trials, utilizes some form of statistical analysis whether it is as simple as the calculation of a mean or as complex as a longitudinal model with a complicated correlation structure. Proper statistical design and analysis will be critical for making valid inferences and moving to the next phase of research.

Most cancer centers and pharmaceutical companies will have a biostatistics group or division. The role of these biostatisticians is to support cancer research by assisting with study design, statistical analysis, and presentation of results. It is highly recommended that, in addition to understanding the basic statistical principles utilized in oncology, oncology drug developers utilize the biostatisticians in their institution or company and consider and treat them as part of the research team. It is well-known that drug development cannot be done independently and requires a host of experts: biostatistical expertise is critical to valid and efficient research, from basic science to preclinical research to clinical trials.

1.2 Example

The most effective way to demonstrate statistical methods used in drug development is by example. In the sections that follow, we present selected results from a phase II study of the farnesyltransferase inhibitor tipifarnib in patients with acute

E. Garrett-Mayer (✉)
Division of Biostatistics and Epidemiology, Hollings Cancer Center, Medical University of South Carolina, 86 Jonathan Lucas Street, Room 118G, Charleston, SC 29425, USA
e-mail: garrettm@musc.edu

M. Hidalgo et al. (eds.), *Principles of Anticancer Drug Development*, Cancer Drug Discovery and Development, DOI 10.1007/978-1-4419-7358-0_1,

myelogenous leukemia (AML) [3]. Briefly, farnesyltransferase inhibitors (FTIs) are potent and selective inhibitors of intracellular farnesyltransferase (FTase) which is an enzyme that catalyzes the transfer of farnesyl moiety to the cysteine terminal residue of a substrate protein. A number of intracellular proteins are substrates for prenylation via FTase (including Ras). Interruption of prenylation may prevent substrates from maturation which may result in inhibition of cellular events that depend on the function of those substrates. FTIs have been shown to be "unselective," targeting proteins involved in different pathways, despite the initial presumption that FTase inhibition would specifically target posttranslational processing of Ras. The use of FTIs is expanding in the treatment of hematological malignancies, especially in AML but also in other leukemias, myelodysplastic syndromes, and myeloproliferative disorders in large part due to their oral bioavailability and acceptable toxicity profile. The FTI tipifarnib (Zarnestra, R115777) had been shown to have in vitro activity against a wide range of malignancies. Based on promising preliminary results in phase I testing, a multicenter phase II study was developed in poor-risk previously untreated AML patients.

One-hundred fifty-eight patients with poor-risk AML were enrolled between March 2001 and December 2005 and followed in this single-arm study. "Poor-risk" was defined as having at least one of the following (1) age $\geq$65; (2) adverse cytogenetic profile; (3) AML arising from an antecedent hematologic disorder; or (4) therapy-related AML. Patients received 600 mg oral tipifarnib twice daily for 21 days followed by a rest period of up to 42 days to allow for peripheral blood count recovery. Additional cycles were administered (up to a total of four) if stable disease or clinical response were observed. Clinical endpoints of interest included complete and partial remission, duration of complete remission (CR), overall survival, and tolerability. The authors also explored correlative endpoints including farnesylation of the surrogate protein HDJ-2 (measured at baseline and day 8) and normalized baseline expression levels of ERK and AKT phosphorylation and their relation to clinical outcomes. Note that only 15% (24 of 158) of AML patients had evaluable data for ERK and AKT expression. Additional information relating to the rationale, study design, and measurement of correlative endpoints can be found in the original article by Lancet et al. [3].

1.3 Aims, Endpoints, and Data Analysis

The distinction between a study's *aims* and *endpoints* is often unclear to cancer researchers, and subsequently these terms are commonly confused with one another. In most cancer research, a plan is developed where specific aims or goals are stated. In the AML example described below, the study's primary aim was to define the antileukemic activity of tipifarnib (the investigational agent) in adults with poor-risk, previously untreated AML [3]. Secondary study aims described evaluating expression levels of ERK and AKT, and examining their ability to predict response to agents such as tipifarnib. For each aim, there is a corresponding endpoint (or outcome) used for that aim's quantitative assessment. An endpoint should be measurable in the sense that it can be observed and recorded for each individual in the study, whether the individual be mouse, patient, or cell line. For example, a measure of tipifarnib's efficacy is its ability to induce a CR, and so, in the AML example, the investigators evaluate patients' CR status (achieved or not) as the primary study endpoint. Whether or not a patient has a CR is determined by clinical criteria outlined in the study's design, and CR status is recorded for each study subject. The endpoint is summarized across subjects using *data analysis*, resulting in both an estimated CR rate, as well as a measure of that estimate's uncertainty. Data analysis facilitates the averaging of information across subjects, resulting in measures (here, estimated CR rate and its associated uncertainty) that provide an objective assessment of a study's primary aim.

To summarize, for any research project, the aims need to be clearly stated. For each aim, an endpoint of interest is identified which provides for a quantitative assessment of the aim. At the completion of the study, inference is made regarding the study's aim via formal statistical analysis methods, some of which are described in this chapter.

1.4 Variable Types

Biostatisticians in drug development generally refer to their data as comprising a set of "variables" because in most instances, the numerical observable data varies across subjects, where a subject may be a cell line, a mouse, a patient, etc. It is important to distinguish between several different kinds of variables and to specify, when defining a variable, how it is measured. As an example, we are often interested in gene expression. However, it is not always clear how gene expression is measured: in a given study, the researcher could be using two categories of expression (expressed vs. not expressed), or she could be using a numeric value of expression that can take any value within a specified range. These two different types of variables are treated differently in statistical analyses.

1.4.1 Continuous Variables

Continuous variables can take any value within a given (and potentially wide) range of values. In our tipifarnib AML study, an example of a continuous variable is expression of phosphorylated AKT. Although it is does not have a wide distribution, the values of baseline AKT in our study range from 0.03 to 2.11, as shown in Fig. 1.1a. Another commonly utilized continuous variable is age. Note that age can be measured in fine increments, such as weeks, days, or even minutes. Age measured in days or weeks is commonly used in animal studies; however, for clinical applications age expressed in years is generally the preferred metric.

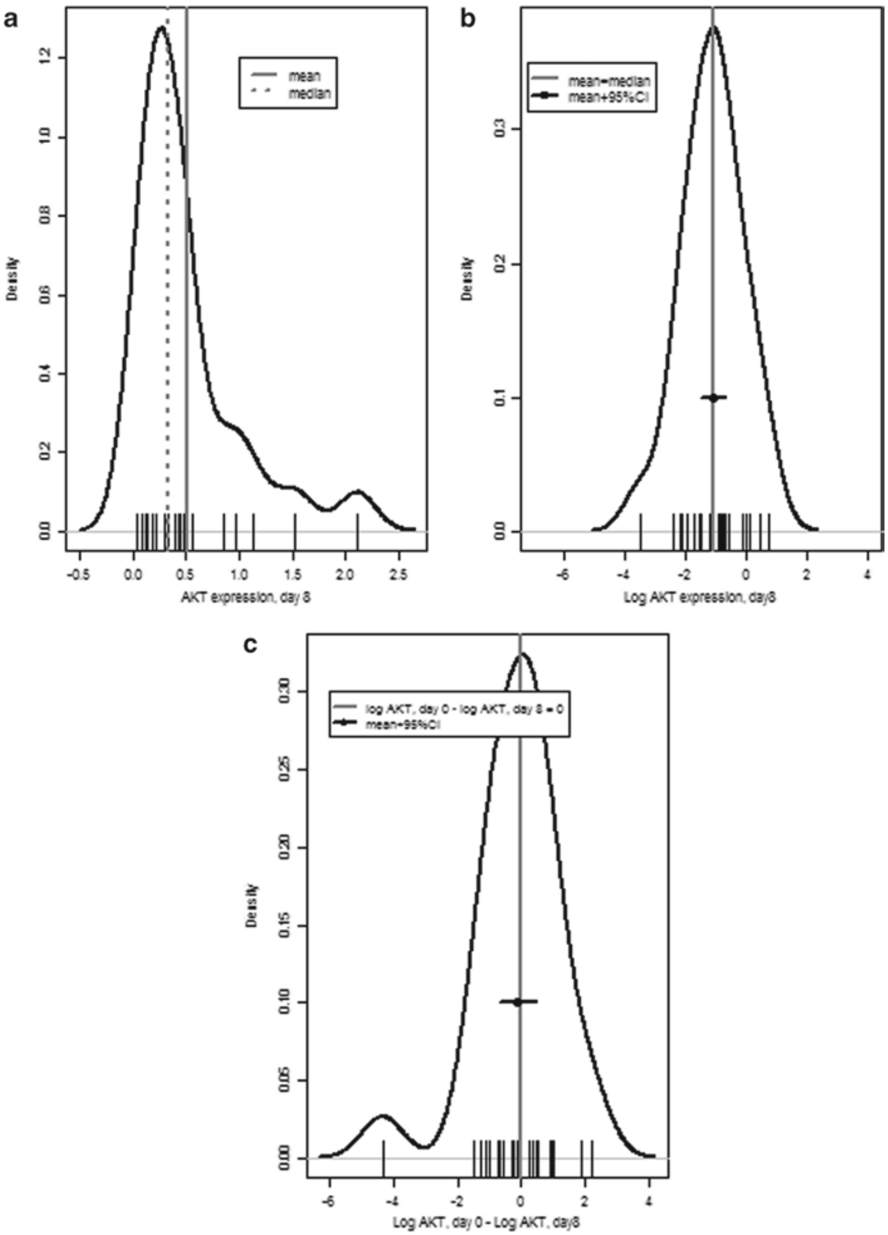

Fig. 1.1 (**a**) Density plot of day 8 AKT expression values. *Tick marks* along *x*-axis indicate observed data points. Mean and median values are indicated by *vertical solid* and *dashed gray lines*, respectively. (**b**) Density plot of day 8 AKT expression values on the log scale. Mean and median are the same and indicated by *vertical solid gray line*. *Horizontal black line* at height of 0.1 indicates the 95% confidence interval for the true day 8 log AKT expression value. (**c**) Density plot of difference between day 0 and day 8 log AKT expression. *Horizontal black line* at 0.1 shows 95% confidence interval for the difference. *Vertical gray line* is plotted at 0, indicating no difference in values

1.4.2 Categorical Variables

Categorical variables have several categories to which an individual may belong. Categorical variables with only two categories are called binary or dichotomous and examples could include gender (with categories of male and female), mutant (with categories of mutant vs. wild-type), or clinical response (with categories of nonresponders and responders). More than two categories are possible as well: a variable with three categories could be genotype, with levels defined as homozygous dominant, homozygous recessive, and heterozygous. Note that there is no specific ordering to genotype: it could be coded numerically with 1 = homozygous dominant, 2 = heterozygous, and 3 = homozygous recessive. Or, the numeric assignments could be transposed without any loss of interpretation. This implies that genotype is a *nominal* categorical variable. Another common example of a nominal categorical variable is race, which can take a number of categories, but the numeric values assigned are irrelevant.

Another class of categorical variables is *ordinal* variables, where there are a fixed (and relatively small) number of categories, but the ordering is meaningful. Common examples of ordinal variables in clinical cancer research are cancer stage, performance status, or grade. For example, there are discrete values assigned to cancer stage and the ordering of the categories is meaningful: stage 2 is higher than stage 1, and stage 3 is higher than stage 2.

1.4.3 Time-to-Event Variables

The predominant clinical outcomes in cancer research are time-to-event variables: overall survival (time to death), progression-free survival (time to progression or death), and disease-free survival [time to relapse or recurrence (or death)]. Time-to-event variables are defined by the occurrence of an event. In a clinical trial, the time from enrollment until death is used to measure overall survival. At first glance, this may seem to be a continuous variable because it is a time that can be measured in very small increments. However, most time-to-event variables in cancer research have the additional characteristic that they can be *censored*, meaning that some of the individuals under study may not experience the event during the time course of the experiment or trial. In the example of overall survival, patients who do not die by the end of the study are considered censored at the time at which they were last known to be alive.

1.4.4 Variable Transformation

In many cases, variables will naturally take one form, but be transformed to another for convenience. For example, age is a continuous variable, but for the purposes of analysis and interpretation, it may make more sense to create a new variable with three age categories, such as <40 years, 40–65 years, and >65 years. This may have some utility, as mentioned, for interpretation, but some information about age is lost. Specifically, when using the categorical example of age above, two patients whose ages are 66 and 91 are considered equivalent with respect to age.

Other common transformations are applied to continuous variables to reduce skewness or asymmetry in their distributions. Skewness can create problems in data analysis by allowing a few data values that are relatively extreme (i.e., outliers) to have substantial influence on inferences. An example of skewness can be seen in Fig. 1.1a where a density plot of AKT at day 8 is shown. Notice that most of the data lies close to the left side of the plot toward 0; however, there are a few points scattered to the right. This distribution is called *right skewed* (or positively skewed) because it has a long right *tail*. To symmetrize the distribution of day 8 AKT, we can apply a logarithmic transformation, shown in Fig. 1.1b, that results in "pulling in" the right tail and making the distribution look more bell shaped. Notice how the points that may have been considered outliers in Fig. 1.1a would no longer be labeled as outliers after this transformation.

1.5 Data Description and Displays

In statistical practice, there is an important distinction between a *parameter* and a *statistic*. A parameter is a quantity whose true value is unknown and is the measure we are trying to estimate. For example, in theory there is a true CR rate to tipifarnib in poor-risk AML patients. This could be determined by treating every poor-risk AML patient with tipifarnib and observing their response. This approach is, of course, impractical. Instead, we collect data on a sample of patients from the identified population, and construct a statistic (or estimate) as our best guess of the true parameter's value. Statisticians are also concerned with an estimated parameter's uncertainty – how much faith do I have that the estimate represents the truth? – and so accompanying each estimated statistic is a measure (usually an interval) describing a range of values consistent with the data within which the true parameter could lie. Thus, the parameter is the unknown value we are trying to make inferences about, and the statistic and its associated uncertainty are quantities calculated based on sample data.

1.5.1 Continuous Variables

Continuous variables have a number of summary statistics used to describe their distributions which generally fall into two common types: statistics to summarize the center of the distribution, and those to describe the data's variability or spread. Statistics used to summarize the center are usually the mean and the median.

The mean is simply the arithmetic average, calculated by adding up all the observed values of the variable and dividing by the number of values. The median is the middle observation (or 50th percentile) and can be found by sorting the data from lowest to highest and identifying the value in the middle of the sorted list. In the case where there is an even number of values, the median is the average of the middle two data points. In our AML example, the mean AKT expression on day 8 is 0.51 and the median is 0.33, as shown in Fig. 1.1a. In the case of skewed data, this is a common result: the median and the mean are different. The mean will be quite sensitive to skewness and extreme values, while the median will not be sensitive. In Fig. 1.1a, notice that the median is closer to the bulk of the data points while the mean tends to be displaced in the direction of the outliers. When describing the center of skewed data, the median is often preferred. Now consider Fig. 1.1b where a log transform of AKT on day 8 has been applied. Because the data is symmetric (i.e., it is not skewed), the mean and the median are almost the same (in this example they are the same to four decimal places).

Variability is most commonly measured by the range and the standard deviation (SD). The range is the difference between the largest and smallest values, but instead it is common to report the minimum and maximum values for a particular variable. The SD is a one-number summary that describes how far the data tend to deviate from the mean. In the AKT example in Fig. 1.1a, the range is 0.03–2.11 and the SD is 0.51. The standard error is a related measure of variability and will be described later when confidence intervals are discussed.

Another measure of spread of the data is the interquartile range (IQR). Recall that the median is the middle data point, or the 50th percentile. Using the same approach of sorting the data, we can identify the 25th and the 75th percentiles of the data. The IQR is defined as the 75th percentile minus the 25th percentile. For expression of AKT at day 8, the 25th and 75th percentiles are 0.20 and 0.53, resulting in an IQR of $0.53-0.20=0.33$. As is the case with the range, it is common practice to report the 25th and 75th percentiles rather than their difference.

At least as important as the summary statistics used to quantify the center and spread of continuous variables are data displays that show the overall distribution. There are various ways to display the distribution of a variable, one of which (a density plot) is shown in Fig. 1.1. Other common plots are boxplots, histograms and, dotplots. Figure 1.2 demonstrates each of these plots for the distribution of age in the AML clinical trial example.

The boxplot (for age, shown in Fig. 1.2a), also known as a box and whisker plot, emphasize quartiles of the distribution and its skewness. The lower and upper limits that define the box are the 25th and 75th percentiles. The line crossing the middle of the box indicates the location of the median (i.e., the 50th percentile). As noted above, the IQR is the distance between the 25th and 75th percentile. In Fig. 1.2, the 25th, 50th, and 75th percentiles are 69, 74, and 78, and the IQR is $78-69=9$. The upper whisker is the line drawn from the box out to the smallest data point within 1.5 times the IQR from the 75th percentile. In the age distribution in Fig. 1.2a, the 75th percentile is 78 and 1.5 times the IQR is $9\times 1.5=13.5$. Hence, the whisker could be drawn as far as $78+13.5=91.5$. However, the largest

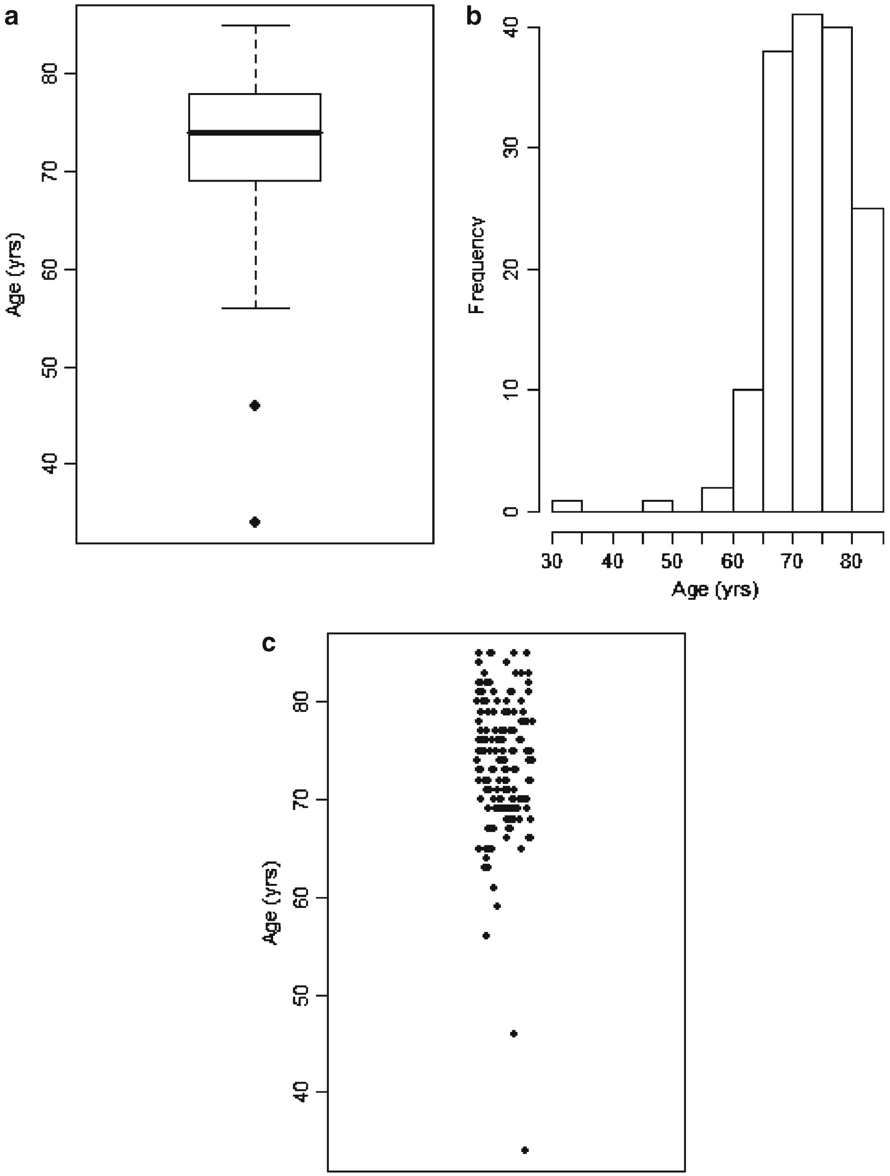

Fig. 1.2 Graphical displays of age for 158 high-risk AML patients treated with tipifarnib. (**a**) Boxplot. (**b**) Histogram. (**c**) Dotplot

observed age in the study is 85 so the whisker stops at 85. The lower whisker is defined in an analogous way and here, the 25th percentile minus 1.5 times the IQR is $69-13.5=55.5$, and so the whisker is drawn to 56, the largest observed value above 55.5. There are, however, two outliers: age values of 34 and 46. These are indicated using individual points in the plot. That is, any points that fall outside the allowed limits for the whiskers are plotted using individual points. To interpret a boxplot, we look at the location of the median in relation to the upper and lower quartiles and the length of the whiskers. These comparisons provide information about the relative symmetry vs. skewness of the data, and also provide a range where the bulk of the data lie: we know the middle 50% of the data lies within the extent of the box, and that the remaining 50% are above and below the box. The whiskers and outliers provide information about the tails of the distribution. A boxplot with one whisker that is significantly longer than the other implies skewness.

A histogram is another popular data display tool for continuous variables. It bins the data into a number of categories and then plots the number of observations in each bin vs. each category. This is not the same as a bar chart which is more general. The y-axis of a histogram provides either the frequency or proportion of observations in a bin. (The same is not true of a bar chart.) In Fig. 1.2b, age is plotted in bins with widths of 5 years. Like the boxplot, this figure provides information about the skewness, range, and center of the data. For age, we see that there is evidence of left-skewness (due to the left tail) and there are two outliers with values below 50 years. Note that the size of the bins can alter the interpretation of histograms. Most statistical software packages have a default algorithm for determining the width and number of bins. However, this varies across packages and altering the bin width can lead to different inference.

The dotplot is a very simple tool to see all of the raw data of a variable. It is most useful in situations where the sample size is relatively small, and there is interest in looking at a particular continuous variable across subgroups. Figure 1.2c shows the dotplot for age. As in Fig. 1.2a, b, two outliers are notable and some left-skewness is seen. Notice also that the data are "jittered" horizontally, facilitating visualization of overlapping data points. Failure to add *noise* to the plot makes it is impossible to tell how many data points are represented by a single symbol. In our example of age, although there are 158 patients in the study, there are only 28 unique values of age so that jittering the points is imperative to displaying all the data in a figure such as Fig. 1.2c.

1.5.2 Categorical Variables

Categorical variables are summarized using tabulations of counts and proportions or percentages. In the case of a binary variable, such as gender, the proportion of male individuals provides all the information necessary to summarize its distribution. For categorical variables with three or more categories, proportions and counts per category are used. Table 1.1 shows a tabulation of counts and percentage of AML patients in each of three response categories.

Table 1.1 Distribution of response in the tipifarnib AML study

	N	%
Complete remission	22	13.9
Partial remission/hematologic improvement	15	9.4
Nonresponse[a]	121	76.6
Total	158	100

[a] Includes stable disease, progressive disease, and not evaluable for response

1.5.3 Time-to-Event Variables

Recall that time-to-event variables are characterized by censoring when some of the individuals under study do not experience the event of interest by the end of the study, or are lost to follow-up before the event has occurred. As a result, means, SDs, and other summary statistics appropriate for continuous variables are not valid. However, the median applies in situations where a large fraction of the individuals have incurred the event. Other time-to-event summary statistics used include the estimated survival fraction at a given time point. For example, in the AML example, the median survival among patients who did not respond to treatment was only 3.6 months, while the median survival among patients who had a CR was 14.4 months. The estimated fraction of patients alive at 12 months are 13% and 66% in nonresponders and responders, respectively.

Note that the fraction surviving and median survival are *not* calculated using the methods described in previous sections for continuous or categorical variables. Censoring needs to be accounted for and the most common approach for estimating these summary statistics is by using the product-limit estimator, also known as the Kaplan–Meier estimator. Without providing great detail, the fraction of patients without the event is estimated at each time point, accounting for how many patients are still at risk of experiencing the event at that time point (called the *risk set*). Patients are removed from the risk set when they have had an event or are censored. This approach is very commonly accepted and explained in greater detail in [5, 6].

Figure 1.3 demonstrates overall survival in our AML study, where patients are defined by three categories (1) complete remission (CR), (2) partial remission and hematologic improvement (PR/HI), and (3) nonresponse (NR). The display is called a Kaplan–Meier plot because the Kaplan–Meier estimates of overall survival are shown. Notice that the curve for each group is a *step-function* relating time, t, to the fraction of individuals alive at time t, denoted $S(t)$. Each step represents a time at which one or more patients has had the event of interest, and so the curve steps down, indicating a lower survival fraction at that point. On the survival curves, in addition to the steps indicating when events occurred, there are tick marks indicating the times at which patients who do not experience the event are censored. This provides information as to what time the patient left the risk set, and is important for understanding the censoring patterns and the number of patients still under study at any given time.

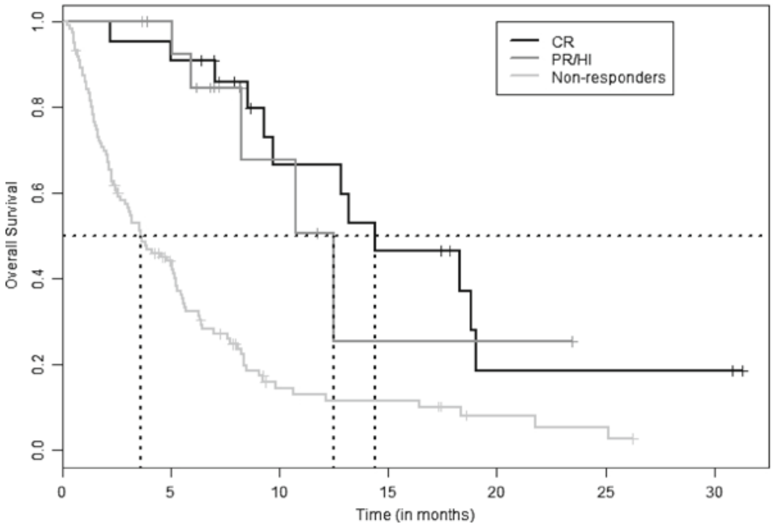

Fig. 1.3 Kaplan–Meier curve of overall survival for patient experiencing CR (*black solid line*), partial remission/hematologic improvement (*gray solid line*), and nonresponse (*light gray solid line*). Median survival is indicated by *dotted black lines*. *Vertical tick marks* on survival curves show censoring times

The survival curve always begins at $S(t) = 1$ for $t = 0$, indicating everyone under study is at risk of the event at time 0. The median survival for a group can be found by drawing a horizontal line at $S(t) = 0.5$ and evaluating the time where it intersects the survival curve. In Fig. 1.3, the horizontal line at 0.5 intersects the curves at 3.6, 12.5, and 14.4 months for the CR, PR/HI, and NR groups, respectively. Other estimates, such as the 12-month survival, can be found be drawing vertical lines up from a particular time of interest. Twelve-month survival estimates for these three groups are 66, 51, and 13% for the CR, PR/HI, and NR groups, respectively.

1.5.4 Confidence Intervals

Confidence intervals provide information about a likely range of values for a given parameter of interest, such as a mean expression level or a true response rate. Confidence intervals are created for many parameters of interest and provide a measure of the estimated parameter's precision. We most often see 95% confidence intervals, but 90 and 99% confidence intervals are also fairly common. A 95% confidence interval is an interval which we are 95% certain contains the true value of the parameter. For example, the mean of log AKT at day 8 is −1.1 and the range of log AKT at day 8 is −3.6 to 0.8. The estimated 95% confidence interval for log AKT at day 8 is (−1.5, −0.7). This means that we are 95% confident that the true mean of log AKT at day 8 lies somewhere between −1.5 and −0.7. Note that the

95% confidence interval provides inference about the mean: it does not provide information about a likely range of values that we might observe for *individuals* in the population. This can be noted by looking at the 95% confidence interval in Fig. 1.1b, which is shown as a small horizontal black line at a height of 0.1. The confidence interval is relatively narrow compared to the range of the data.

The width of the confidence interval depends on three things (1) the variability of the data in the sample (i.e., patient heterogeneity with regard to the variable of interest), (2) the level of confidence desired (e.g., 95%), and (3) the sample size. The variability of the data in the sample will depend on the patient population you choose. For example, the variability in PSA (prostate-specific antigen) values in a sample of healthy male volunteers will be much smaller than the variability of PSA in a sample of men with relapsed prostate cancer. We expect that men without prostate cancer will all have PSA values in the range of 0–4 ng/ml, while men with refractory prostate cancer will have PSA values ranging anywhere from 4 ng/ml into the tens of thousands. This latter group will have much greater variability which will affect our confidence in a mean estimate. As noted previously, the level of confidence chosen is most often 95%, but in some cases, a 90 or 99% confidence level is justified when we are satisfied with less confidence while gaining a narrower range, or require greater confidence at the expense of a wider interval.

The width of the confidence interval also depends directly on the sample size: as the sample size increases, the width of the confidence interval decreases. This is intuitive: the more information we collect, the more certainty we have in our estimate. For formulas for construction of confidence intervals, see [1, 5, 6].

1.5.5 Confidence Intervals for Means and Differences in Means

In the previous section, the 95% confidence interval for mean day 8 log AKT was presented and is also shown in Fig. 1.1b. In addition to the mean value at day 8, we may also be interested in the mean of the *difference* between log AKT at day 0 and day 8. By calculating a confidence interval for this difference, we obtain both a range of reasonable values for the difference as well as evidence to support or refute the hypothesis that the difference differs from zero. A common use of the confidence interval is to test whether a difference in means is equal to zero: if 0 is not within the 95% confidence interval we conclude that the difference differs meaningfully from 0. This is an example of the duality between confidence intervals and hypothesis testing (hypothesis testing is described in Sect. 1.6 of this chapter).

Using the difference in log AKT between days 0 and 8 as an example, we construct a 95% confidence interval by taking the difference between the day 0 and the day 8 log AKT values resulting in a single calculated difference per patient. The data for this is shown in Fig. 1.1c, where a density plot is shown in addition to the observed differences along the bottom of the figure. The estimated mean difference is –0.11 and the 95% confidence interval for the mean difference is (–0.68, 0.46), indicated by the horizontal black line at a height of 0.10. This implies that we are 95% confident that the true average difference lies somewhere between –0.68 and 0.46. The vertical gray line indicates a difference of 0 and is the location at which there is no difference between the day 0 and day 8 values. Notice that the 95% confidence interval overlaps this vertical line suggesting that the mean difference between the day 0 and day 8 log AKT values does not differ from 0.

1.5.6 Confidence Intervals for Proportions and Comparisons of Proportions

The interpretation of confidence intervals remains generally the same, regardless of the parameter of interest. In the case of proportions, we use a different method for estimating the confidence interval, but nonetheless it has the same meaning. In the AML tipifarnib example, 22 of 158 patients, or 14%, had a CR. The report of this statistic will be better understood by providing a 95% confidence interval which will convey, in addition to our observed remission rate, a range of likely true remission rates if this treatment approach were applied in general to the high-risk AML population (consistent with those patients in our trial). As described above, the confidence interval width depends on the sample size and our level of confidence. It also depends on the variability of remissions in the population but, in the case of proportions, the variability is determined by the true remission proportion. For example, if the true proportion is near 1 (or 0) most of the subjects will (or will not) experience a remission, and therefore the variability in remissions is low. Conversely, variability in the event is highest for true proportions near 0.5. The 95% confidence interval for the true CR rate to tipifarnib in AML patients is (0.09, 0.20).

The confidence interval for the true CR rate is somewhat asymmetric. We may have expected the observed remission rate of 0.14 to lie in the middle of the interval, but this is not the case, and the asymmetry is not due to rounding. As estimates of proportions get close to 0 or 1, the corresponding confidence intervals become increasingly asymmetric. For example, only three patients experienced a partial remission, yielding an observed partial remission rate of 0.02 with a 95% confidence interval of (0.004, 0.05). Here, the distance between the estimated remission rate and the interval's upper bound is roughly twice the distance from the estimated rate to the lower bound.

The second thing to notice about the 95% confidence interval for the CR rate to tipifarnib is its fairly narrow width of 0.11 ($0.11 = 0.20 - 0.09$). As mentioned above, the width of the interval depends on the sample size. If our sample size had been only 50, an observed CR rate of 0.14 would have a 95% confidence interval of (0.06, 0.27), for a width of 0.21. And, if we had a much larger sample size of 400, the width would be only 0.07.

Often confidence intervals for proportions are created using approximate approaches. These approximations work very well under two conditions (1) the sample size is reasonably large, and (2) the proportion is not close to 0 or 1. It is difficult to provide rules determining *reasonably large* and *not close to 0 or 1* because they depend on each other. For example, a proportion of 0.90 is not very close to 1 if the sample size is 1,000 but it would be considered close to 1 if the sample size were only 20. But, in general, almost all statistical software packages can generate *exact* confidence intervals, so reliance on approximations is not necessary, although it is still very commonly seen.

There are other parameters that we use to compare proportions in different subpopulations. For example, odds ratios or relative risks are often used for quantifying the risk or benefit associated with an exposure or treatment. Of the AML patients treated with tipifarnib who had at least three cycles of treatment 41.2% (14/34) had a CR, compared to only 6.5% (8/124) among patients who did not complete three or more cycles. We can use an odds ratio to represent this difference in CR by cycles. Specifically, the odds ratio for CR in this example is 10.15: the odds of experiencing a CR for patients who are able to complete three cycles of treatment is ten times that of the odds for patients who had fewer than three cycles. We calculate this by taking the odds of a CR in patients with three or more cycles (41.2/58.8% = 0.700) and dividing it by the odds of CR in patients with fewer than three cycles (6.5/93.5% = 0.069). The 95% confidence interval for the odds ratio is then used to determine if there is an association between the exposure (i.e., three or more cycles) and the outcome (CR). In this case, the 95% confidence interval for the odds ratio for CR in patients with and without at least three cycles of tipifarnib is (3.4, 31.2). Recall that an odds ratio of 1 indicates no association between CR and three or more cycles of treatment. Because this confidence interval does not contain one, we can conclude that there appears to be a significant association between receiving three or more cycles of tipifarnib and CR. However, note that this is a rather simplistic analysis where we have not adjusted for additional confounders that may play a role in a patient's ability to receive three or more cycles. It is likely that patients least able to tolerate treatment also have other factors making them less likely to respond to treatment.

1.5.7 Confidence Intervals for Time-to-Event Parameters

The interpretation of the confidence interval for time-to-event parameters, such as median survival or 12-month survival, is the same as for other types of parameters. Recall that the estimated median survival for AML patients who had a CR to tipifranib was 14.4 months and the median survival in patients who were nonresponders was 3.6 months. The associated 95% confidence intervals for median survival (in months) are (9.7, Inf) and (2.9, 5.2), where "Inf" represents a bound of infinity. This is not uncommon: in cases where there are relatively few patients on study when the median survival is achieved and the survival does not drop dramatically below 0.5 by the end of the study, the upper limit of the 95% confidence interval may be infinite. To interpret this, we would state that we are 95% confident that the true median survival in AML patients who achieved a CR is greater than 9.7 months. Similarly, we are 95% confident that the true median survival in nonresponders lies somewhere between 2.9 and 5.2 months.

1.6 Hypothesis Testing

1.6.1 From Research Question to Statistical Hypothesis

In the Lancet study, 10 of 75 (13%) poor-risk AML subjects with an unfavorable cytogenic profile achieved a CR. A natural question to ask is how this rate compares to CR rates in the same patient population receiving standard treatment. Is the tipifarnib rate better? Worse? Different? Because subjects in the Lancet study received only tipifarnib, there is no internal comparative arm. However, investigators frequently use published or historical rates in comparable patient populations to compare treatments in single-arm studies. For example, Leith et al. [4] conducted a study of elderly AML patients in which they investigated the association between patient cytogenics and response. They report a 21% CR rate among elderly poor-risk AML patients in response to standard chemotherapy (standard-dose cytosine arabinoside and daunomycin + rhG-CSF). For illustrative purposes, we use the CR rate reported by Leith et al. [4] as the historical CR rate with which to compare the tipifarnib rate.

To quantify the relationship between the tipifarnib and chemotherapy CR rates, we use *hypothesis testing*, a statistical approach that allows us to draw conclusions from sample data and infer to the entire population. Hypothesis testing begins with a statement of "no effect," appropriately called the *null hypothesis* (H_0). For the current example, our null hypothesis states that the tipifarnib CR rate is equal to the historical chemotherapy rate. Specifically, we write H_0: $p_{\text{tipifarnib}}=0.21$, where $p_{\text{tipifarnib}}$ is the true CR rate among elderly poor-risk tipifarnib-treated AML patients with unfavorable cytogenics. A second statement, called the *alternative hypothesis* (H_1 or sometimes H_A), summarizes the research question of interest and is phrased in contrast to H_0. Here, a reasonable alternative hypothesis states the tipifarnib rate is different from the historical chemotherapy rate and is written as H_1: $p_{\text{tipifarnib}}\neq 0.21$. The latter hypothesis is called a *two-sided alternative* and captures in a single statement two *one-sided alternatives*, specifically (1) the tipifarnib rate is better than the chemotherapy rate ($p_{\text{tipifarnib}}>0.21$) and (2) the tipifarnib rate is worse than the chemotherapy rate ($p_{\text{tipifarnib}}<0.21$). A two-sided alternative is appropriate when there is no reason to assume a priori that the effect of the new treatment will be better or worse than that of the standard treatment.

1.6.2 Evaluating Evidence Through p-values

A 13% CR rate is smaller than the 21% published chemotherapy rate, but this difference may be a chance occurrence, that is, an observation not attributable to tipifarnib treatment. Is there sufficient evidence in the data to allow us to rule out random variation as an explanation for the observed tipifarnib rate? Stated another way, if the true CR rate among this subgroup of elderly poor-risk tipifarnib-treated AML patients is the same as the historical chemotherapy rate of 21%, how unusual is an observation of 10 CRs in 75 subjects?

To answer this question, we need an understanding of the distribution of the frequency of observed CRs under conditions specific to this study. These study-specific conditions refer to the composition of the study population (elderly poor-risk tipifarnib-treated AML patients with unfavorable karyotype); the sample size of 75 subjects; and the null-hypothesized tipifarnib CR rate of 21%. Under these conditions, how many CRs in 75 subjects should we expect to observe? How variable is the number of CRs in 75 subjects? Addressing these questions requires repeating the study many times under identical conditions and observing the number of CRs for each repetition. A more practical approach is to conduct a computer-based simulation in which a virtual "coin," with probability of a head equal to 0.21, is tossed 75 times and the number of heads observed. Tossing a coin and recording heads or tails is like observing a patient from our study population following treatment with tipifarnib and deciding if the subject has or has not experienced a CR. Using a computer, the simulation can be repeated thousands of times in a matter of seconds, rapidly providing information pertaining to variability needed to address our hypothesis.

Figure 1.4a shows a bar chart of the number of CRs observed in 100 repetitions of our simulated study. From this graph we note that 10 CRs were recorded in four of the 100 repetitions, or 4% of the simulations. If we could conduct our simulation an infinite number of times, the bars would "smooth out" and we would observe a bar chart like the one shown in Fig. 1.4b. Figure 1.4b shows the *exact sampling distribution* of the number of CRs in 75 subjects, where the true CR rate is 21%. Here the bar heights are probabilities, where the probability of an outcome is loosely defined as the long-term proportion of simulations in which that outcome is observed. The probability of 10 CRs in 75 subjects is 0.031.

To summarize how unusual an observation is, statisticians generally sum the probabilities of all outcomes *at least as extreme as the one observed*, where the "extremeness" of an event is measured by how probable it is relative to the observed outcome. In this case, nine or fewer CRs are extreme events since each is less probable than the observed outcome of 10 CRs in 75 subjects. For the same reason, 22 or more CRs are considered extreme. The bars corresponding to extreme events are shaded dark gray in Fig. 1.4b, and each bar's height (probability) is no greater than the height of the bar corresponding to the observed outcome of 10. Therefore, the probability of observing 10 CRs in 75 subjects, or any observation at least as extreme, is found by summing the heights of the dark gray bars in Fig. 1.4b and is equal to

$$\underbrace{\Pr(10\text{ CRs})}_{\text{probability of observed event}} + \underbrace{\Pr(9\text{ CRs}) + \Pr(8\text{ CRs}) + \cdots + \Pr(1\text{ CR}) + \Pr(0\text{ CRs})}_{\text{probability of events at least as extreme as, and to the ``left'' of the observed event}}$$

$$+ \underbrace{\Pr(22\text{ CRs}) + \Pr(23\text{ CRs}) + \cdots + \Pr(74\text{ CRs}) + \Pr(75\text{ CRs})}_{\text{probability of events at least as extreme as and to the ``right'' of the observed event}} = 0.12.$$

This probability is an example of a *p-value*. A *p*-value is always calculated assuming that the null hypothesis is true – in this case that the true tipifarnib CR rate is 21% – and represents the probability that the observed result or one more extreme is a random event. If a *p*-value is small – usually less than 0.05 – we eliminate random chance as an explanation for the observed results and *reject the null hypothesis*. A finding for

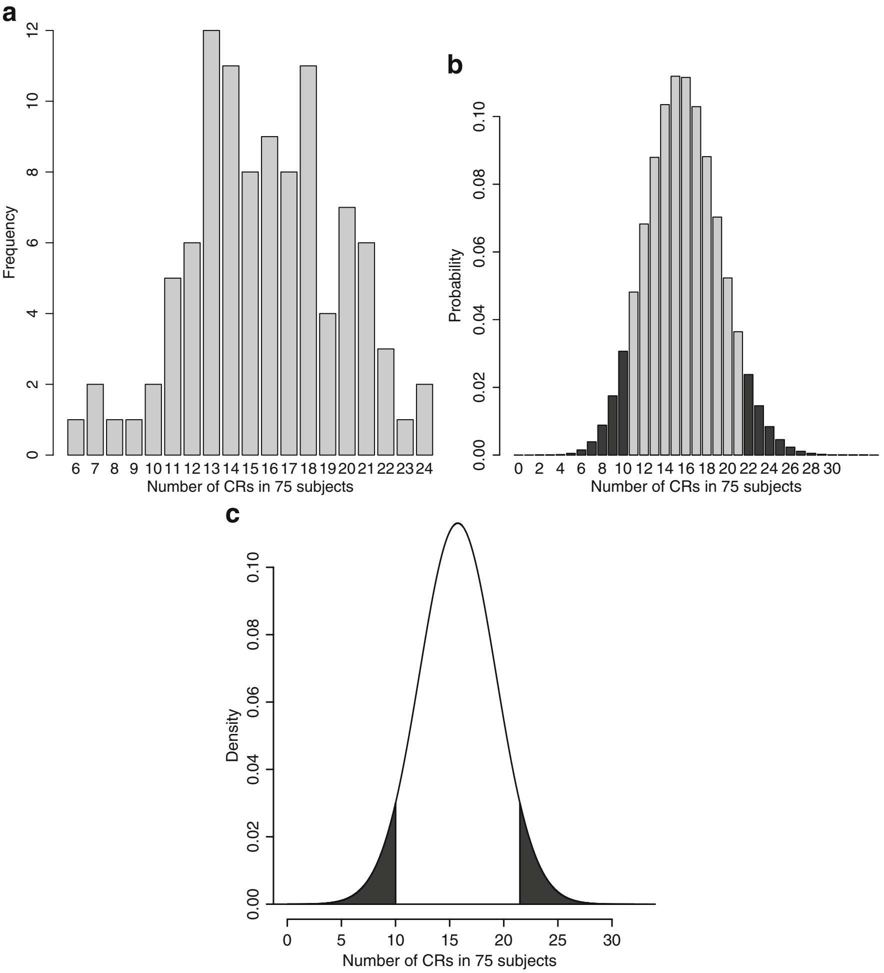

Fig. 1.4 (**a**) Bar chart of the frequency of the number of CRs in 75 subjects for 100 simulations assuming the true CR rate is 21%. (**b**) Exact sampling distribution of the number of CRs in 75 subjects, assuming the true CR rate is 21%. Each bar's height is the probability of observing the number of CRs indicated on the *horizontal axis*. (**c**) Normal approximation to the exact sampling distribution of the number of CRs in 75 subjects, assuming the true CR rate is 21%

which the *p*-value is smaller than 0.05 is said to be *statistically significant* or simply *significant*. A *p*-value greater than or equal to 0.05 indicates that, under the null hypothesis, the observed result is not so unlikely – the event could occur by chance 5% of the time or more – and we *fail to reject the null hypothesis*. Such findings are called *nonsignificant*. For our example, the *p*-value is 0.12, which is greater than 0.05. We therefore fail to reject the null hypothesis and conclude that the data provide insufficient evidence to claim that the true CR rate for elderly poor-risk tipifarnib-treated AML patients with unfavorable karyotype differs meaningfully from 21%.

On a final note, statisticians often compute *p*-values using an approximation to the exact sampling distribution of a statistic. This alleviates the need to construct exact sampling distributions that change as conditions vary from one problem to the next. This is similar to the approximation discussed in Sect. 1.4.2 for confidence intervals and is shown in Fig. 1.4c, the normal approximation density curve.

1.6.3 Types of Errors

With each decision one can make concerning the null hypothesis – reject or fail to reject – there is a corresponding potential error or mistake. If the null hypothesis is rejected when in fact it is true, this is called a *type I error*. On the other hand, failing to reject the null hypothesis when in fact it is false is called a *type II error*. The probability of a type I error is represented by the Greek letter alpha (α) and the probability of a type II error is represented by the Greek letter beta (β).

The layout in Table 1.2 displays the interpretations of type I and type II errors in the context of the hypothesis test of the tipifarnib CR rate. In this example, a type I error occurs if we conclude that the tipifarnib CR rate differs from the historical chemotherapy rate when it really is not different; a type II error occurs if we conclude that the tipifarnib rate is the same as the chemotherapy rate when it really is different. In drug discovery, we are typically more concerned with type I errors since rejecting the null hypothesis in error means a nonefficacious drug may advance to larger, more expensive (e.g., randomized) trials and, more importantly, patients will receive a drug that offers no additional clinical benefit. Phase II trials are often designed with $\alpha = \beta = 0.10$ and phase III trials with $\alpha \leq 0.05$ and $\beta \leq 0.20$.

1.7 Common One- and Two-Sample Tests

1.7.1 Comparing Proportions

The hypothesis test highlighted in Sect. 1.6 is called a *one-sample test of a proportion*. This test is appropriate when interest surrounds relating a true but unknown proportion to a reference value. The test accounts for sampling variability inherent in

Table 1.2 Type I and type II errors in the context of the hypothesis test of the tipifarnib CR rate

	Truth	
Decision	Tipifarnib CR rate	
Tipifarnib CR rate	Same as chemotherapy CR rate	Different from chemotherapy CR rate
Same as chemotherapy CR rate	No error ($1-\alpha$)	Type II error (β)
Different from chemotherapy CR rate	Type I error (α)	No error ($1-\beta$)

estimating the unknown proportion but treats the reference value as a constant. For our example, the historical chemotherapy rate was actually constructed from an observation of 11 CRs in 52 subjects. Treating the 21% reference rate as "truth" ignores the fact that it was estimated from sample data, and therefore subject to sampling variability.

An alternative test that accounts for the sampling variability in both the estimated tipifarnib and chemotherapy CR rates is a *two-sample test of proportions* whichis most commonly seen in randomized studies. Here the null hypothesis is H_0:$p_{\text{tipifarnib}} = p_{\text{chemotherapy}}$, where $p_{\text{tipifarnib}}$ is defined in Sect. 1.5.1, and $p_{\text{chemotherapy}}$ is the true but unknown CR rate for elderly poor-risk chemotherapy-treated AML patients. The corresponding two-sided alternative is H_0: $p_{\text{tipifarnib}} \neq p_{\text{chemotherapy}}$. The test is conducted based on binary data from the two sets of patients under comparison. In this example, we construct the test based on a comparison of 10 CRs in 75 subjects with 11 CRs in 52 subjects. The *p*-value for this test is 0.17. This is slightly larger than the *p*-value for the corresponding one-sample test because the two-sample test incorporates the uncertainty associated with estimating the chemotherapy CR rate in addition to the tipifarnib rate. The interpretation of the two-sample test is also slightly different. Our conclusion for the one-sample test was the true tipifarnib CR rate did not differ significantly from 21%. Here we conclude that the true tipifarnib CR rate does not differ significantly from the true chemotherapy CR rate, whatever that rate may be.

1.7.2 Comparing Means

Comparisons between groups of continuous data are commonly constructed based on the relative locations of the data distributions' centers, and the most common measure of central tendency is the mean. For example, to compare baseline levels of ERK phosphorylation (p-ERK) between responders and nonresponders, appropriate null and alternative hypotheses are H_0: $\mu_{\text{resp}} = \mu_{\text{nonresp}}$ and H_1: $\mu_{\text{resp}} \neq \mu_{\text{nonresp}}$, where μ_{resp} is the true but unknown average baseline level of p-ERK among responders, and μ_{nonresp} is the true but unknown average baseline level of p-ERK among nonresponders. Here responders are defined as subjects achieving either a complete or partial remission, or a hematologic improvement. Both the null and alternative hypotheses can be expressed equivalently based on a difference in means – that is, H_0: $\mu_{\text{resp}} - \mu_{\text{nonresp}} = 0$ vs. H_1: $\mu_{\text{resp}} - \mu_{\text{nonresp}} \neq 0$ – and the test is carried out in a manner similar to that described in Sect. 1.5.2. Specifically, the test is based on estimates of the average baseline p-ERK levels among the eight responders and 21 nonresponders for whom biologic correlative data are available. The mean (SD) p-ERK levels for responders and nonresponders are 0.55 (0.75) and 0.36 (0.27), respectively. Additionally, it is important to understand characteristics of the distribution of the difference in sample means. Under assumptions that p-ERK baseline levels are normally distributed and independently sampled from both groups, and that the variance of p-ERK levels is unknown, the shape of the distribution of the difference in sample means is unimodal and symmetric – much like a normal distribution. However, the distribution of the difference has "heavier" tails than a normal distribution, which is to say that extreme differences are more likely to occur than would be the case had the distribution been normal.

The distribution of the difference in sample means under the stated assumptions is known as the *t-distribution*, and the corresponding hypothesis test is called a *two sample t-test*. (A *one-sample t-test* also exists and is appropriate when testing a true but unknown mean against a reference value.) In conducting any hypothesis testing, it is important to first evaluate how well underlying assumptions are satisfied. When assumptions are violated, the resulting inference is potentially compromised. As illustrated in Fig. 1.1a, day 8 p-ERK levels are positively skewed. Baseline p-ERK levels similarly violate the normality assumption (figure not shown). The mean and SD for baseline p-ERK levels provide additional evidence that approximate normality is not satisfied. A normally distributed variable has the property that 68% of its values fall within 1 SD of the mean, 95% within 2 SDs, and 99% within 3. Notice that for responders, 1 SD to the left of the mean $(0.55-0.75)$, and for nonresponders 2 SDs to the left of the mean $(0.36-2\times0.27)$, results in implausible values – p-ERK expression levels cannot be negative.

What can be done? As illustrated in Fig. 1.1b, a logarithmic transformation of day 8 p-ERK levels alleviates the distribution's skewness resulting in a more symmetric, approximately normal shape. A logarithmic transformation of baseline p-ERK levels induces the same approximate normality (figure not shown). We therefore conduct a two-sample *t*-test on the *log-transformed* baseline p-ERK values. The null and alternative hypotheses are similarly stated, but μ_{resp} and μ_{nonresp} now represent the true but unknown average *log* baseline p-ERK levels among responders and nonresponders, respectively. The *p*-value associated with this test is 0.72 leading to the conclusion that average log baseline p-ERK levels do not differ significantly between responders and nonresponders. Using properties of logarithms,[1] an equivalent conclusion is the average baseline p-ERK expression ratio comparing responders to nonresponders is not significantly different from 1.

When transformation fails to induce normality, a test based on the ranked data is an alternative to the two-sample *t*-test. If a variable's distribution for one group is centered at a larger value relative to a second group, data sampled from the first group will likely have larger ranks than data sampled from the second group. This is the idea behind the *Wilcoxon rank-sum test*, an example of a *nonparametric test*. A nonparametric test is one that makes no assumption about the form of the sample data's distribution. The Wilcoxon rank-sum test is the nonparametric equivalent to the two-sample *t-test*. If we use the Wilcoxon rank-sum test to assess the association between baseline p-ERK level and response status, the *p*-value is 0.65, and our

[1] Since $\log(A)-\log(B)=\log(A/B)$, then the equation $\log(A)-\log(B)=0$ is equivalent to $\log(A/B)=0$. Exponentiation of both sides of the latter leads to the equivalent expression, $A/B=1$.

conclusion is the same – baseline p-ERK levels do not differ significantly between responders and nonresponders.

1.7.3 The Chi-Square Test

Tipifarnib is a FTase inhibitor so an important component of the Lancet study was the assessment of FTase inhibition. In addition to exploring FTase inhibition in AML isolates, Lancet study investigators examined inhibition in buccal (cheek) mucosa samples to determine if inhibition could be detected in normal tissue. The investigators report FTase inhibition failure in AML isolates from 14 of 57 (25%) subject samples but in only four of 49 (4%) normal tissue samples. In their discussion, the authors postulate that this difference potentially indicates a patient subpopulation with FTase posttranslational modification or possibly an alteration in drug accumulation, and may identify a patient cohort unlikely to benefit from tipifarnib.

Table 1.3a shows a two-by-two table of the observed distribution of FTase inhibition status (yes or no) by sample type (AML isolate or normal tissue from buccal mucosa). This table is an example of a *contingency table* and is used to display the joint distribution of categorical variables. Usually, interest surrounds understanding the association (if any) between the row and column variables. Consistent with hypothesis testing strategies already presented, we construct a test of association under a null condition; we assume sample type and FTase inhibition status are independent in the sense that a sample's origin – AML isolate or normal tissue – does not influence FTase inhibition.

Table 1.3 Observed (A) and expected (B) frequencies of FTase inhibition status (yes or no) by sample type (AML isolate or buccal mucosa)

	Farnesyltransferase inhibition status		
Sample type	Yes	No	Total
A: Observed			
AML isolate	43	14	57
Buccal mucosa	45	4	49
Total	88	18	106
B: Expected			
AML isolate	47.3	9.7	57
Buccal mucosa	40.7	8.3	49
Total	88	18	106

Expected frequencies are derived based on probability laws that assume independence between row and column variables

Using laws of probability, we derive a table of the frequencies we would *expect* to see if the variables under consideration really were independent.[2] Table 1.3b shows the expected cell frequencies under an assumption of independence between sample type and FTase inhibition status. We test the variables' independence based on how far the observed table deviates from that expected under independence. Such a test is called a *chi-square test*. Its name derives from the property that the statistic used to measure the discrepancy between the observed and expected tables has a distribution that can be approximated by a chi-square distribution, provided the sample size is large. The chi-square test based on Table 1.3a, b has a *p*-value of 0.025, indicating FTase inhibition status differs significantly by sample type.[3]

A chi-square test for a two-by-two contingency table is equivalent to the two-sample test of proportions discussed in Sect. 1.6.1. However, chi-square tests apply more generally to tests of association between categorical variables with any number of levels. For example, we may be interested in knowing if response differs meaningfully across levels of Eastern Cooperative Oncology Group (ECOG) performance status (PS). We define response status (response or nonresponse) as in Sect. 1.6.2 – responders achieved complete or partial remission, or a hematologic improvement, while nonresponders had progression, stable disease, or were inevaluable. ECOG PS has three levels (0, 1, and 2) based on patient eligibility requirements. Table 1.4a shows the two-by-three contingency table for the joint distribution of response status and ECOG PS. The expected frequencies under independence are displayed in Table 1.4b. The corresponding chi-square test has a

Table 1.4 Observed (A) and expected (B) frequencies of response status (response vs. nonresponse) by ECOG PS (0, 1, or 2)

	ECOG performance status			
Response status	0	1	2	Total
A: Observed				
Response	12	23	2	37
Nonresponse	25	68	11	104
Total	37	91	13	141
B: Expected				
Response	9.7	23.9	3.4	37
Nonresponse	27.3	67.1	9.6	104
Total	37	91	13	141

Expected frequencies are derived based on probability laws that assume independence between row and column variables

[2] For independent events, A and B, Pr(A and B) = Pr(A) × Pr(B). For example, if we assume FTase inhibition and sample type are independent, then from Table 1.3a the probability a sample comprises AML isolates and exhibits FTase inhibition is $(57/106) \times (88/106) \approx 0.446$. Therefore, out of 106 total samples, we expect $0.446 \times 106 \approx 47.3$ to be AML isolates exhibiting FTase inhibition – assuming independence. The remaining cells in Table 1.3b are derived in a similar manner.

[3] In this example, we assume FTase inhibition levels in AML isolates and buccal samples from the same patient are uncorrelated.

p-value of 0.46, and we conclude there is no significant association between response status and ECOG PS.

1.7.4 Fisher's Exact Test

The approximate chi-square distribution of the statistic measuring the discrepancy between the observed and expected frequencies is based on large-sample asymptotic. When sample sizes are small, an alternative test of independence is *Fisher's exact test*. The test of independence between sample type and FTase inhibition status has a *p*-value of 0.037 based on Fisher's exact test. For response status and ECOG PS, Fisher's exact test yields a *p*-value of 0.53. Both examples result in equivalent inference compared to their corresponding chi-square tests discussed in Sect. 1.7.3. Had the tests conflicted, the more conservative finding (i.e., the one least in support of rejecting independence) would be reported or some would argue to report the finding of the exact test.

1.7.5 Testing Paired Data

The hypothesis tests discussed thus far rely on an assumption that the data are independently sampled. Examples of data that violate this assumption are as follows: measures sampled from the same subject over time, for example, serum cytokine concentrations measured at baseline, week 1, week 4, and week 8 of a study; cluster-correlated measures, for example, standardized test scores of school-aged children from classrooms sampled from selected elementary schools in a state; and repeated measures, for example, visual acuity measures from the left and right eyes of the same subject.

In the Lancet study, p-ERK levels were measured at baseline and at day 8. Is there a meaningful change in p-ERK levels from baseline? Although it may seem natural to assess the significance of the change in p-ERK using a two-sample *t*-test, the baseline and day 8 measures from the same subject do not represent independently sampled values. One remedy to this violation is to construct differences from the paired observations, resulting in a collection of independent measures of change. We construct the difference, *d*, from log-transformed p-ERK values, with each subject contributing a single value $d=\log(\text{p-ERK}_{\text{day 8}})-\log(\text{p-ERK}_{\text{baseline}})$. If the differences are meaningfully different from zero, we conclude change from baseline to day 8 in log p-ERK levels is significant. The corresponding null and alternative hypotheses are H_0: $\Delta=0$ vs. H_1: $\Delta\neq0$, where Δ is the true mean difference in day 8 and baseline log p-ERK values. A test of the null hypothesis is accomplished using the one-sample *t*-test described in Sect. 1.7.2. Here, the *p*-value is 0.10, and we conclude there is not a significant change in log p-ERK from baseline to day 8. As described in Sect. 1.7.2, the conclusion of no significant difference in log p-ERK levels is equivalent to a conclusion that the ratio of day 8 to baseline p-ERK levels does not differ significantly from 1.

1.7.6 Comparing Survival Times

In Sect. 1.4.3, we described survival endpoints as the predominant clinical outcome in cancer trials. The most common test to compare survival experiences between groups is the *log-rank test*. We consider a two-group comparison here, but the test easily extends to multiple groups. Consider two groups with corresponding survival functions $S_1(t)$ and $S_2(t)$. The log-rank test tests the null hypothesis H_0: $\mathrm{S}_1(t)=\mathrm{S}_2(t)$ for all times, t, vs. the alternative H_1: $S_1(t) \neq S_2(t)$ for at least one time, t. The test is constructed from differences in the observed and expected number of deaths at each failure (death) time, under the null hypothesis that survival is the same in each group. The resulting test statistic has an approximate chi-square distribution.

Figure 1.5 shows Kaplan–Meier estimates of overall survival for subjects younger than 75 years and subjects 75 years or more. The log-rank test yields a highly significant result with a p-value of 0.000022. (Most publications print very small p-values as being less than some threshold, typically $p < 0.001$, as indicated

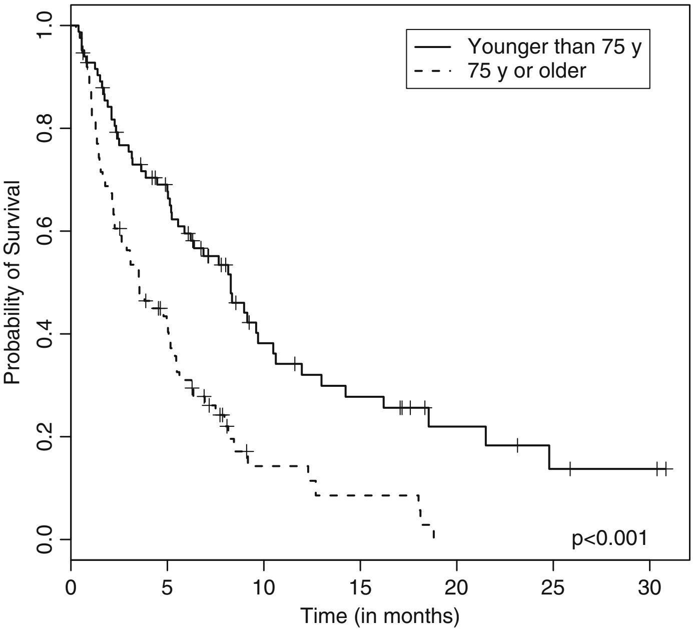

Fig. 1.5 Kaplan–Meier estimates of the probability of survival comparing subjects younger than 75 years to those 75 years and older. The p-value corresponds to the log-rank test of the null hypothesis that the survival curves are equal