Statistics for Biology and Health

Tomasz Burzykowski
Geert Molenberghs
Marc Buyse
Editors

The Evaluation of Surrogate Endpoints

With 57 Illustrations

Springer

Tomasz Burzykowski
Center for Statistics
Limburgs Universitair Centrum
3590 Diepenbeek
Belgium
tomasz.burzykowski@luc.ac.be

Geert Molenberghs
Center for Statistics
Limburgs Universitair Centrum
3590 Diepenbeek
Belgium
geert.molenberghs@luc.ac.be

Marc Buyse
International Drug Development Institute
1050 Brussels
Belgium
marc.buyse@iddi.com

Library of Congress Cataloging-in-Publication Data
The evaluation of surrogate endpoints / [edited by] Tomasz Burzykowski, Geert Molenberghs, Marc Buyse.
p. cm. — (Statistics for biology and health)
Includes bibliographical references and index.
ISBN 0-387-20277-3 (alk. paper)
1. Clinical trials. 2. Drugs—Testing. I. Molenberghs, Geert. II. Buyse, Marc E. III. Burzykowski, Tomasz. IV. Series.
R853.C55E935 2005
610′.72′4—dc22 2004059192

ISBN 0-387-20277-3 Printed on acid-free paper.

Printed in the United States of America. (EB)

9 8 7 6 5 4 3 2 1 SPIN 10942956

springeronline.com

Bożenie i Mikołajowi

Voor Conny, An en Jasper

A Monique, Céline et Mathieu

Preface

More than ever is there a strong drive to search for and evaluate potential surrogate markers and surrogate endpoints for randomized clinical trials. A successful surrogate endpoint is able to reduce follow-up trial time and/or to reduce the number of patients needed to establish a certain treatment effect. From a statistical perspective, Prentice's framework (1989), amplified by Freedman, Graubard, and Schatzkin (1992), was instrumental to start the debate as to how statistical validation or, more modestly formulated, statistical evaluation, of a potential surrogate endpoint could be undertaken. Much debate ensued, also in the light of the historic "accidents" with surrogates not carefully evaluated, and it is fair to say the surrogate marker debate has since been laden with a certain amount of skepticism.

Connected to his involvement in clinical trial methodology, Marc Buyse has always had a strong interest in the surrogate marker validation debate. In April 1994, Marc and Geert met at a *Drug Information Association* meeting in Bruges, at the time where Marc was thinking about the *relative effect* as a measure to supplement the *proportion explained*. One thing led to another and soon an LUC-based research team was formed, headed by the three of us, that, over the years, has encompassed fifteen members from various research institutes. The team has investigated a number of aspects of surrogate marker validation. A move was soon made from the so-called single-trial framework to a meta-analytic or hierarchical one, in line with ideas developed by Michael Hughes and Michael Daniels, and also by Mitch Gail and his co-workers. A lot of subsequent activity focused on finding appropriate hierarchical statistical models for various types of surrogate and true outcomes. Formulating such models is not always straightforward, let alone fitting them, and consequently the need arose to explore simplified modeling and fitting strategies, and the Bayesian framework was considered as a potential alternative. Also, as different models incorporate different association parameters, the need arose to try and unify the surrogate marker evaluation measures.

While doing this, an eye had to be kept on several important application areas, such as oncology, HIV, and mental health. Even though there is a common basis for surrogate marker validation across these areas, a good number of aspects are area specific. For example, it is fair to say that the speed of the developments in HIV is tremendous, compared to other therapeutic areas. In mental health, the delineation between true and surrogate

endpoints is not as clear as it would be in other areas. Finally, because surrogate marker evaluation takes place, to a large extent, in the development of medicinal product arena, the perspectives of the pharmaceutical industry and the regulatory authorities have to be taken into account in a proper fashion.

This text hopes to give an accessible synthetic account of the developments just sketched, giving proper credit to historical developments, providing a balance between statistical considerations of a modeling and computation nature, scientific considerations coming from the various therapeutic areas, and the positions taken by the pharmaceutical industry and the regulatory authorities. As in any scientific debate, different people approach surrogate marker evaluation with various degrees of comfort. We hope the current text does proper justice to all views, not just the editors' views.

Although a variety of authors have contributed to this book, we have chosen a strongly edited form to achieve a smooth flow. As far as possible, a common set of notations has been used by all authors. Ample cross-references between chapters are provided. The book should be suitable either to read a selected number of chapters or the integral text.

Tomasz Burzykowski (LUC, Diepenbeek)

Geert Molenberghs (LUC, Diepenbeek)

Marc Buyse (IDDI, Brussels, and LUC, Diepenbeek)

Acknowledgments

Over the years, our research team has published several papers on the subject, has communicated at conferences and has taught short courses and held workshops in a wide variety of locations worldwide. This has been done for various audiences, including statistical and biopharmaceutical audiences. We are sure that not only preparing for the various communications, but also the numerous discussions have had a beneficial impact on this book.

Aloka Chakravarty, author of Chapter 3, would like to express appreciation for the support and encouragement received from Drs. R.T. O'Neill, C. Anello, M.F. Huque, G.Y.H. Chi, S. Machado, H.M. Hung, G. Chen, G. Soon, R. Sridhara, and P.L. Yang.

Ross L. Prentice, author of Chapter 19, expresses gratitude for support from National Institutes of Health grant CA53996.

Michael D. Hughes, author of Chapter 17, gratefully acknowledges support from National Institutes of Health grants AI24643, AI38855, and AI41110.

Tomasz Burzykowski and Geert Molenberghs gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data."

Tomasz, Geert, and Marc

Diepenbeek, September 2004

Chapter Authors

Ariel Alonso Abad
Limburgs Universitair Centrum, Diepenbeek, Belgium

Tomasz Burzykowski
Limburgs Universitair Centrum, Diepenbeek, Belgium

Marc Buyse
International Drug Development Institute, Brussels, Belgium
Limburgs Universitair Centrum, Diepenbeek, Belgium

Aloka Chakravarty
Food and Drug Administration, Rockville, MD, U.S.A.

José Cortiñas Abrahantes
Limburgs Universitair Centrum, Diepenbeek, Belgium

Laurence Freedman
Bar-Ilan University, Ramat Gan, Israel

Mitch Gail
National Cancer Institute, National Institutes of Health, Bethesda, MD, U.S.A.

Helena Geys
Limburgs Universitair Centrum, Diepenbeek, Belgium

Michael D. Hughes
Harvard School of Public Health, Boston, MA, U.S.A.

Annouschka Laenen
Limburgs Universitair Centrum, Diepenbeek, Belgium

Geert Molenberghs
Limburgs Universitair Centrum, Diepenbeek, Belgium

Ross L. Prentice
University of Washington, Seattle, WA, U.S.A.

Didier Renard
Eli Lilly & Company, Mont Saint Guibert, Belgium

Arthur Schatzkin
National Cancer Institute, National Institutes of Health, Bethesda, MD, U.S.A.

Ziv Shkedy
Limburgs Universitair Centrum, Diepenbeek, Belgium

Franz Torres Barbosa
Limburgs Universitair Centrum, Diepenbeek, Belgium

Tony Vangeneugden
Tibotec, Mechelen, Belgium

Contents

Preface **vii**

Acknowledgments **ix**

Chapter Authors **xi**

1 Introduction **1**

1.1 The Concept of a Surrogate Endpoint 1

1.2 Why Is There Reservation Toward the Use of Surrogate Endpoints? 2

1.3 Why the Use of Surrogate Endpoints Is Still Being Considered? 3

1.4 Validation of Surrogate Endpoints 5

2 Setting the Scene **7**

2.1 Historical Perspective 7

2.2 A Regulatory Agencies Perspective 8

2.3 Main Issues 10

3 Regulatory Aspects in Using Surrogate Markers in Clinical Trials **13**

3.1 Introduction and Motivation 13

3.1.1 Definitions and Their Regulatory Ramifications . . . 13

3.1.2 Support for Surrogates 14

3.1.3 Criteria for Surrogate Markers To Be Used in Drug Development 15

3.1.4 Surrogate Markers and Biomarkers 17

3.2 Surrogate Markers in Regulatory Setting 19

3.2.1 Fast Track Program – A Program for Accelerated Approval 19

3.2.2 Subpart H and Its Relevance to Surrogate Markers . 21

3.3 Use of Surrogate Markers in Anti-viral Drug Products . . . 28

3.3.1 Crixivan: A Case Study 31

3.3.2 Viramune: A Case Study 32

3.4 Use of Surrogate Markers in Anti-cancer Drug Products . . 35

3.4.1 Doxil: A Case Study 37

3.5 Use of Surrogate Markers in Cardiovascular Drug Products 40

3.5.1 Anti-hypertensive Drugs 42

3.5.2 Anti-platelet Drugs 43

3.5.3 Drugs for Heart Failure 43

3.5.4 Drugs for Angina and Silent Ischemia 43

3.5.5 Ventricular Arrhythmias 44

3.5.6 The CAST Experience: A Case Study in Ventricular Arrhythmia 44

3.6 Statistical Issues Related to Accelerated Approval 45

3.7 Surrogate Markers at Other Phases of Drug Development . 50

4 Notation and Motivating Studies 53

4.1 Notation 53

4.2 Key Datasets 53

4.2.1 Ophthalmology: Age-related Macular Degeneration Trial 53

4.2.2 Advanced Ovarian Cancer: A Meta-analysis of Four Clinical Trials 55

4.2.3 Corfu Study in Advanced Colorectal Cancer: Two Clinical Trials 56

4.2.4 Four Meta-analyses of 28 Clinical Trials in Advanced Colorectal Cancer 58

4.2.5 Advanced Prostate Cancer: Two Clinical Trials . . . 59

4.2.6 Meta-analysis of Five Clinical Trials in Schizophrenia 60

4.2.7 An Equivalence Trial in Schizophrenia 63

4.2.8 A Clinical Trial in Cardiovascular Disease 65

4.2.9 Acute Migraine: A Meta-analysis of 10 Clinical Trials 65

5 The History of Surrogate Endpoint Validation 67

5.1 Introduction 67

5.2 Prentice's Definition and Criteria 68

5.2.1 Definition 68

5.2.2 Prentice's Criteria 69

5.2.3 Example 69

5.2.4 Theoretical Foundations of Prentice's Criteria 71

5.3 Proportion of Treatment Effect Explained by a Surrogate . 73

5.3.1 Example 74

5.3.2 Properties of the Proportion of Treatment Effect Explained by a Surrogate 75

5.4 Relative Effect and Adjusted Association 75

5.4.1 Example 77

5.4.2 Properties of Relative Effect and Adjusted Association and Further Problems 77

5.5 Further Problems with Single-trial Validation Measures . . 79

5.6 Discussion 81

6 Validation Using Single-trial Data: Mixed Binary and

Continuous Outcomes **83**

6.1 Introduction 83

6.2 Ordinal Endpoints 84

6.2.1 The Dale Model 84

6.2.2 Application: A Cardiovascular Disease Trial 86

6.3 Mixed Continuous and Binary Endpoints 87

6.3.1 A Probit Formulation 87

6.3.2 A Plackett-Dale Formulation 89

6.3.3 Application: A Cardiovascular Disease Trial 90

6.3.4 Application: Age-related Macular Degeneration . . . 91

6.4 Discussion 92

7 A Meta-analytic Validation Framework for Continuous Outcomes **95**

7.1 Introduction 95

7.2 A Meta-analytic Approach 95

7.2.1 Trial-level Surrogacy 97

7.2.2 Individual-level Surrogacy 101

7.2.3 A New Approach to Surrogate Evaluation 101

7.3 Single-trial Measures versus Multi-trial Measures 102

7.4 Computational Issues 104

7.4.1 Initial Simulation Study 104

7.4.2 Simplified Modeling Strategies 106

7.4.3 Additional Simulation Study 110

7.5 Case Studies 111

7.5.1 Age-related Macular Degeneration Study (ARMD) . 112

7.5.2 Advanced Colorectal Cancer 114

7.5.3 Advanced Ovarian Cancer 116
7.6 Discussion 117
7.7 Extensions 120

8 The Choice of Units 121

8.1 Introduction 121
8.2 Model Description and Setting 122
8.3 Modeling Strategies 124
8.3.1 Strategy I: Two-level Only 125
8.3.2 Strategy II: Three Levels, Fixed Effects 126
8.3.3 Strategy III: Three Levels, Random Effects 126
8.4 A Simulation Study 128
8.4.1 Simulation Settings 128
8.4.2 Simulation Results, Equal Trial- and Center-level Association 130
8.4.3 Simulation Results, Unequal Trial- and Center-level Association 133
8.5 Analysis of Schizophrenia Trials 138
8.6 Concluding Remarks 140

9 Extensions of the Meta-analytic Approach to Surrogate Endpoints 143

9.1 Introduction 143
9.2 The Normal Model 145
9.2.1 Precision of Estimates of $\boldsymbol{\delta_0}$ Based on the Meta-analytic Approach 147
9.3 Flexibility of the Marginal Approach 148
9.4 Discussion 150

10 Meta-analytic Validation with Binary Outcomes 153

10.1 Introduction 153

10.2 Model Formulation 153

10.3 Parameter Estimation 154

10.4 Acute Migraine: A Meta-analysis of Ten Clinical Trials . . . 159

10.5 Concluding Remarks 160

11 Validation in the Case of Two Failure-time Endpoints 163

11.1 Introduction 163

11.2 Meta-analytic Approach: The Two-stage Model 164

11.2.1 Bias in the Estimation of Measures of Surrogacy . . 167

11.2.2 Prediction of Treatment Effect on the True Endpoint 170

11.3 Analysis of Case Studies 172

11.3.1 Advanced Ovarian Cancer: Four Clinical Trials . . . 173

11.3.2 Advanced Colorectal Cancer: Two Clinical Trials . . 177

11.4 The Choice of the First-stage Copula Model 180

11.5 A Simulation Study 182

11.5.1 Parameter Settings 183

11.5.2 Summary Conclusions 185

11.6 Alternatives to the Two-stage Modeling 187

11.7 Simplified Modeling Strategies 189

11.8 Discussion 193

12 An Ordinal Surrogate for a Survival True Endpoint 195

12.1 Introduction 195

12.2 A Meta-analytic Approach: The Two-stage Model 196

12.3 Analysis of Case Study 201

12.3.1 Descriptive Analysis 201
12.3.2 Analysis of Four-category Tumor Response 204
12.3.3 Analysis of Binary Tumor Response 210
12.4 Discussion 214

13 A Combination of Longitudinal and Survival Endpoints 219

13.1 Introduction 219
13.2 Joint Modeling Approach 220
13.2.1 Model and Notation 220
13.2.2 Measures of Surrogacy 222
13.3 Application to Advanced Prostate Cancer Data 223
13.4 Discussion 229

14 Repeated Measures and Surrogate Endpoint Validation 231

14.1 Introduction 231
14.2 The Model 232
14.3 Variance Reduction Factor 233
14.4 Validation from a Canonical Correlation Perspective 237
14.4.1 Relationship Between VRF, θ, and R^2_{indiv} 239
14.5 R^2_Λ and the Likelihood Reduction Factor: A Unifying Approach Based on Prentice's Criteria 241
14.5.1 The Measure R^2_Λ 241
14.5.2 Relationship Between R^2_Λ and θ_P 243
14.5.3 The Likelihood Reduction Factor 244
14.6 Analysis of Case Studies 246
14.6.1 Study in Schizophrenia 246
14.6.2 Age-related Macular Degeneration Trial 249
14.7 Discussion 251

15 Bayesian Evaluation of Surrogate Endpoints 253

15.1 Introduction 253

15.2 Bivariate Models for Meta-analytic Data 255

15.2.1 The Two-stage Model of McIntosh (1996) 255

15.2.2 The Model of van Houwelingen, Arends, and Stijnen (2002) 256

15.3 Models for the Validation of Surrogate Endpoints Using Meta-analytic Data 257

15.3.1 The Hierarchical Bayesian Model of Daniels and Hughes (1997) 257

15.3.2 The Two-stage Model of Buyse *et al.* (2000a) 258

15.4 A Hierarchical Bayesian Model for the Validation of Surrogate Endpoints 261

15.5 Analysis of Case Studies 262

15.5.1 Age-related Macular Degeneration (ARMD) Trial . . 263

15.5.2 Advanced Ovarian Cancer 266

15.6 Simulation Study 268

15.7 Discussion 269

16 Surrogate Marker Validation in Mental Health 271

16.1 Introduction 271

16.2 Mental Health 274

16.2.1 Mental Health and Schizophrenia 275

16.3 Surrogate Endpoint Validation Criteria 277

16.3.1 Prentice's Criteria 278

16.3.2 Freedman's Proportion Explained 279

16.3.3 Relative Effect and Adjusted Association 279

16.3.4 Hierarchical Approach 280

16.3.5 Variance Reduction Factor and Likelihood Reduction Factor 282

16.4 Analysis of Case Studies 282

16.4.1 A Meta-analysis of Trials in Schizophrenic Subjects 282

16.4.2 An Equivalence Trial in Schizophrenic Patients . . . 290

16.5 Discussion 292

17 The Evaluation of Surrogate Endpoints in Practice: Experience in HIV 295

17.1 Introduction and Background 295

17.2 Framework for Evaluating Surrogacy 297

17.3 Defining the True Endpoint 298

17.4 Defining the Potential Surrogate Endpoints 299

17.5 Prognostic Value of HIV-1 RNA and CD4 Cell Count . . . 300

17.6 Prognostic Value of Changes in HIV-1 RNA and CD4 Cell Count 301

17.7 Association of Differences Between Randomized Treatments in Their Effects on Markers and Progression to AIDS or Death 307

17.7.1 Regression Approach 307

17.7.2 Results from the Meta-analysis 310

17.8 Discussion 318

18 An Alternative Measure for Meta-analytic Surrogate Endpoint Validation 323

18.1 Introduction 323

18.2 The Use of the Trial-level Validation Measures 324

18.3 Surrogate Threshold Effect 327

18.3.1 Normally Distributed Endpoints 327

18.3.2 Other Distributions 330

18.4 Analysis of Case Studies 332
18.4.1 Advanced Colorectal Cancer 333
18.4.2 Advanced Ovarian Cancer 335
18.5 An Extension of the Concept of a Surrogate Threshold Effect 336
18.5.1 Application to the Advanced Ovarian Cancer Data . 338
18.6 Discussion 338

19 Discussion: Surrogate Endpoint Definition and Evaluation 341

19.1 Introduction 341
19.2 Surrogate Endpoint Definition 341
19.3 Surrogate Endpoint Evaluation 343
19.4 Treatment Effect Prediction 345
19.5 Discussion 347

20 The Promise and Peril of Surrogate Endpoints in Cancer Research 349

20.1 Introduction 349
20.2 When Are Surrogates Appropriate? 350
20.3 Identifying Surrogate Endpoints for Cancer 350
20.4 Validating Surrogate Markers 352
20.5 The Logic of Cancer Surrogacy 353
20.6 Can Surrogate Validity Be Extrapolated from One Exposure to Another? 355
20.7 Epithelial Hyperproliferation: A Case Study 356
20.8 Evaluating Potential Surrogate Endpoints 357
20.8.1 Is the Surrogate Associated with Cancer? 358
20.8.2 Is E Associated with S? 359

20.8.3 Does S Mediate the Link Between E and T? 360
20.9 Surrogates That Are Likely To Be Valid 361
20.10 Measurement Error 363
20.11 Conclusion 364

References **367**

Index **401**

1

Introduction

Geert Molenberghs, Marc Buyse, and Tomasz Burzykowski

1.1 The Concept of a Surrogate Endpoint

One of the most important factors influencing the duration and complexity of the process of developing new treatments is the choice of the endpoint, which will be used to assess the efficacy of the treatment. Two main criteria to select the endpoint are its sensitivity to detect treatment effects and its clinical relevance to goals of the study (Fleming 1996). The relevance depends on, for example, whether evidence for biological activity of a drug is sought (as in Phase II trials) or whether a definitive evaluation of clinical benefit to patients has to be made (as in Phase III trials). For instance, in life-threatening diseases, such as cardiovascular diseases or cancer, the endpoint relevant for definitive evaluation of a treatment typically is survival.

It often appears, however, that the most sensitive and relevant clinical endpoint, which will be called the "true" endpoint throughout this text, might be difficult to use in a clinical trial. This can happen if the measurement of the true endpoint:

- is costly (for example, to diagnose "cachexia," a condition associated with malnutrition and involving loss of muscle and fat tissue, expensive equipment measuring content of nitrogen, potassium, and water in patient's body is required);
- is difficult (for example, involving compound measures such as typically is the case in quality of life or pain assessment);
- requires a long follow-up time (for example, survival in early-stage cancers);
- requires a large sample size due to a low incidence of the event (for

example, short-term mortality in patients with suspected acute myocardial infarction).

In such cases, use of the true endpoint increases the complexity and/or the duration of research. To overcome these problems, a seemingly attractive solution is to replace the true endpoint by another one, which is measured earlier, more conveniently, or more frequently. Such "replacement" endpoints are termed "surrogate" endpoints (Ellenberg and Hamilton 1989).

Note that several related but somewhat distinct terms are in use, such as surrogate endpoint, surrogate marker, or biomarker. *Surrogate endpoint* has the connotation of replacement of the true endpoint in a clinical study by another one. A *marker* on the other hand is an outcome, a measurement, or a set of measurements that is indicative for a variable or a general concept. For example, a number of blood, urine, and other measurements can be used to detect environmental stress in living organisms. Although there are common aspects in the evaluation of surrogate endpoints and markers, the contexts are different. In this book, we will largely focus on surrogate endpoints, with a lot of emphasis on randomized clinical trials.

1.2 Why Is There Reservation Toward the Use of Surrogate Endpoints?

Because of the possible benefits for the duration of a clinical trial, surrogate endpoints have been used in medical research for a long time (Ellenberg and Hamilton 1989, Fleming and DeMets 1996). Table 1.1 presents several examples. The use of the surrogate endpoints presented in Table 1.1 was based on an established *association* between them on the one hand and the corresponding true endpoints on the other hand. However, the mere existence of an association between a candidate surrogate endpoint and the true endpoint is not sufficient for using the former as a surrogate. As Fleming and DeMets (1996) put it, "a correlate does not make a surrogate." What is required is that the effect of the treatment on the surrogate endpoint reliably predicts the effect on the true endpoint. Unfortunately, partly due to the lack of appropriate methodology, this condition was not checked in the early attempts to use surrogates. Consequently, for most of the surrogates mentioned in Table 1.1, it was found that their use, at least in some applications, led to erroneous, or even harmful, conclusions. A review of several such examples is given by Fleming and DeMets (1996). Probably the best known case is the approval by the Food and Drug Administration (FDA) in the United States of the use of three drugs: encainide, flecainide, and moricizine. The drugs were approved based on the fact that

TABLE 1.1. *Examples of surrogate endpoints used in medical research.*

Disease	Endpoints	
	Surrogate	True
Early stage cancer	Time to progression	Survival time
Advanced cancer	Tumor response	Survival time
Osteoporosis	Bone mineral density	Bone fracture
Ophthalmology (glaucoma)	Intraocular pressure	Long-term visual acuity
Chronic granulomatous disease	Superoxide production	Serious infection
	Ability to kill bacteria	Serious infection
Cardiovascular disease	Ejection fraction	Myocardial infarction
	Blood pressure	Stroke, survival time
	Arrythmias	Survival time
HIV infection	CD4 counts; viral load	Development of AIDS, survival time

they were shown to effectively suppress arrythmias. It was believed that, because arrythmia is associated with an almost fourfold increase in the rate of cardiac-complication-related death, the drugs would reduce the death rate. However, a clinical trial conducted after the drugs had been approved by the FDA and introduced into clinical practice showed that in fact the death rate among patients treated with encainide and flecainide was more than twice the one among patients treated with placebo (The Cardiac Arrhythmia Suppression Trial (CAST) Investigators 1989). An increase of the risk was also detected for moricizine.

This and other examples of unsuccessful replacement of true endpoints led to the scepticism about usefulness of surrogate endpoints. Consequently, negative opinions about the use of surrogates in the evaluation of treatment efficacy have been voiced (Fleming 1996, Fleming and DeMets 1996, DeGruttola *et al.* 1997).

1.3 Why the Use of Surrogate Endpoints Is Still Being Considered?

It will be clear from the previous section that the very mention of surrogate endpoints has always been very controversial. However, not all early applications were failures. For example, the dramatic surge of the AIDS epidemic, the impressive therapeutic results obtained early on with zidovudine, and the pressure for an accelerated evaluation of new therapies have all led to, first, the use of CD4 blood count and then, with the advent of highly active antiretroviral therapy (HAART), viral load as endpoints that replaced time to clinical events and overall survival (DeGruttola *et al.*

1995), in spite of some concerns about their limitations as surrogates for clinically relevant endpoints (Lagakos and Hoth 1992).

Generally, before a new drug can be accepted for the use in clinical practice, its efficacy and safety needs to be rigorously assessed in a series of clinical trials. This process of testing a new therapy can (and, in fact, does) take many years. At the same time, the number of candidate biomarkers and ultimately the number of surrogate endpoints based upon them is increasing dramatically. Indeed, an increasing number of new drugs have a well-defined mechanism of action at the molecular level, allowing drug developers to measure the effect of these drugs on the relevant biomarkers (Ferentz 2002). There is also increasing public pressure for new, promising drugs to be approved for marketing as rapidly as possible, and such approval will have to be based on biomarkers rather than on some long-term clinical endpoint (Lesko and Atkinson 2001). The pressure can become especially high in a situation where rapidly increasing incidence of a disease can become a serious threat to public health. As an illustration of this trend toward early decision-making, recently proposed clinical trial designs use treatment effects on a surrogate endpoint to screen for treatments that show insufficient promise to have a sizeable impact on survival (Royston, Parmar, and Qian 2003). Last but not least, if the approval process is shortened, there will be a corresponding need for earlier detection of safety signals that could point to toxic problems with new drugs. It is a safe bet, therefore, that the evaluation of tomorrow's drugs will be based primarily on biomarkers, rather than on the longer-term, harder clinical endpoints that have dominated the development of new drugs until now.

In conclusion, because surrogate endpoints can shorten the duration of the process, their use does constitute an attractive option. Thus, although many would like to avoid surrogate endpoints altogether, sometimes surrogates will be the only reasonable alternative, especially when the true endpoint is rare and/or distant in time.

Another reason to shorten the duration of the process of testing new therapies may be related to new discoveries in medicine and biology, which create a possibility for development of many potentially effective treatments for a particular disease. In such a situation, a need to cope with a large number of new promising treatments that should be quickly evaluated with respect to their efficacy might appear. As a matter of fact, this can already be observed happening in oncology, as the increased knowledge about the genetic mechanisms operating in cancer cells led to the proposal of qualitatively new approaches to treat cancer. An example is found in the use of a genetically modified virus that selectively attacks p53-deficient cells, sparing normal cells (Heise *et al.* 1997). It is known that for several cancers, mutations of the p53 gene are quite common. For instance, in head and

neck tumors they are detected in 45-70% of the cases (Khuri *et al.* 2000), whereas in pancreatic tumors, this is about 60% of the cases (Barton *et al.* 1991). Consequently, in these cancers the injection of the virus in the tumor might result in the eradication of the cancer cells without affecting normal cells. In fact, clinical trials investigating the efficacy of such a treatment have already been started, showing promising results (Von Hoff *et al.* 1998, Khuri *et al.* 2000, Lamont *et al.* 2000, Nemunaitis *et al.* 2001). With the results of the human genome mapping now available (International Human Genome Sequencing Consortium 2001, Venter *et al.* 2001), development of even a larger spectrum of treatments aimed at disease mechanisms present at the gene level might be expected.

From a practical point of view, shortening the duration of a clinical trial also limits possible problems with non-compliance and missing data, which are more likely in longer studies, and therefore increases effectiveness and reliability of the research.

Finally, an important area of potential application of surrogate endpoints is the assessment of safety of new treatments. Duration and sample size of clinical trials aimed at development of new drugs are usually insufficient to detect rare or late adverse effects of the treatment (Dunn and Mann 1999, Jones 2001). The use of surrogate endpoints (for toxicity-related clinical endpoints) might allow one to obtain information about such effects even during the clinical testing phase.

All of these reasons apply to the current state of research on novel treatments. Despite the failed past attempts, it is therefore difficult to abandon the idea of using surrogate endpoints altogether.

1.4 Validation of Surrogate Endpoints

Nevertheless, the failed past attempts to use surrogate endpoints do make it clear that, before deciding on the use of a candidate surrogate endpoint, it is of the utmost importance to investigate its validity. (The term validity is used here in a broad sense, and not in the narrow, well-defined psychometric sense, even though there is a relationship between both, see also Chapter 16.) Consequently, formal methods allowing for validation are required. Such methods have become the subject of intensive research over the past decades. In this volume, the results of this research, as well as some novel concepts and techniques, will be presented.

2

Setting the Scene

Geert Molenberghs, Marc Buyse, and Tomasz Burzykowski

2.1 Historical Perspective

Often, the most clinically relevant endpoint, that is, the "true" endpoint, is difficult to use in a clinical trial. In cancer trials, for instance, survival is still regarded as the ultimate endpoint of interest, but it may lack sensitivity to true therapeutic advances, it may be confounded by competing risks and second-line treatments, and it is observed late, which results in long delays before new drugs can be approved. In such cases, a seemingly attractive solution is to replace the true endpoint by another one, which might be measured earlier, more conveniently, or more frequently. As stated in Chapter 1, such "replacement" endpoints are termed "surrogate" endpoints.

Before a surrogate can replace a true endpoint, it should be *validated* or *evaluated*. Merely establishing a correlation between both endpoints is not sufficient (Baker and Kramer 2003). Several formal methods for this purpose have already been proposed (Prentice 1989, Freedman, Graubard, and Schatzkin 1992, Daniels and Hughes 1997, Buyse and Molenberghs 1998, Buyse *et al.* 2000a, Gail *et al.* 2000). With the statistical methods available, it ought to be possible to conduct a formal investigation on the quality of various endpoints used as surrogates in clinical practice. Such an investigation can shed light on the feasibility of the use of these endpoints and guide the regulatory agencies, for example, in the choice of the endpoints that can be used for accelerated approval of investigational drugs. Of course, as stated earlier, a quantitative evaluation is important but is by no means the only component in the decision process leading to the replacement of the true endpoint by the surrogate one. Several parties are involved, including the regulatory agencies (Section 2.2) and the industry developing a medicinal product.

2.2 A Regulatory Agencies Perspective

The need to develop new drugs and treatments as quickly as possible has become acute nowadays. Regulatory agencies from around the globe, in particular in the United States, in Europe, and in Japan, have reacted to this challenge through various provisions and policies.

In the United States, there are mechanisms available for accelerated approval based on surrogate endpoints, in order to reduce the time to review an application for indications with no known effective therapy and for providing access to patients for unapproved drugs. Accelerated approval (sometimes referred to as "conditional approval" or "Subpart H") refers to an acceleration of the overall development plan by allowing submission of an application, and if approved, marketing of a drug on the basis of surrogate endpoints while further studies demonstrating direct patient benefit are underway. Accelerated approval is limited to diseases where no effective therapies exist and is based on a surrogate endpoint likely to predict clinical benefit.

The recent recommendation of the Food and Drug Administration (FDA) for accelerated approval of investigational cancer treatments states that

> "FDA believes that for many cancer therapies it is appropriate to utilize objective evidence of tumor shrinkage as a basis for approval, allowing additional evidence of increased survival and/or improved quality of life associated with that therapy to be demonstrated later"

(Food and Drug Administration 1996). This marks a departure from the traditional requirements for new cancer treatments to show survival or disease-free survival benefits prior to being granted market approval (Fleming *et al.* 1994, Cocchetto and Jones 1998). If the achievement of a complete remission has indeed a major impact on prognosis in hematological malignancies (Armitage 1993, The International Non-Hodgkin's Lymphoma Prognostic Factors Project 1993, Kantarjian *et al.* 1995), the relationship between tumor response and survival duration is far less clear in solid tumors, even though the shrinkage of metastatic measurable masses has long been the cornerstone of the development of cytotoxic therapies (Oye and Shapiro 1984). In the United States, response rate has been used as a surrogate for patient benefit for accelerated approval and as a component of full approval for some hormonal and biological products. Among them are docetaxel for second-line metastatic breast cancer, irinotecan for second-line metastatic colorectal cancer, capecitabine for refractory metastatic breast cancer, liposomal cytarabine for lymphomatous meningitis, and temozolo-

mide for second-line anaplastic astrocytoma. Two drugs received accelerated approval for supplemental indications: liposomal doxorubicin for refractory ovarian cancer and celecoxib for polyp reduction in familial adenomatous polyposis.

In the European Union, there is a different "accelerated approval" mechanism. The European legislation allows for granting a marketing authorization under "exceptional circumstances" where comprehensive data cannot be provided at the time of submission (e.g., because of the rarity of the disease) and provided that the applicant agrees to a further program of studies that will be the basis for post-authorizations review of the benefit/risk profile of the drug. Although this primarily refers to situations where randomized clinical trials are lacking, it applies equally well to absence of data on a particular endpoint. According to the European Agency for the Evaluation of Medicinal Products (EMEA) guideline for the evaluation of anticancer agents, the choice of endpoints should be guided by the clinical relevance of the endpoint and should take into account methodological considerations. Possible endpoints for phase III trials in oncology include progression-free survival, overall survival, response rate (and duration), and symptom control/quality of life. The guideline also states that if objective response rate is used as the primary endpoint, compelling justifications are needed and normally additional supportive evidence of efficacy in terms of, for example, symptom control is necessary (Committee for Proprietary Medicinal Products 2001). Thus, where justified, the use of surrogate endpoints in oncology is possible although it may require confirmation of efficacy in the post-authorization phase, e.g., by confirming an effect on the true endpoint or in confirmatory trials. The initial EMEA experience with antineoplastic and endocrine therapy agents has shown that in the majority of cases, approval was indeed obtained based on a surrogate endpoint such as objective response rate. This was the case, e.g., for docetaxel in second-line (monotherapy) metastatic breast cancer, liposomal doxorubicin in AIDS-Kaposi sarcoma, and paclitaxel in second-line AIDS-Kaposi sarcoma. Topotecan was approved in second-line metastatic ovarian cancer based on response rate and progression-free survival, and temozolomide was approved in recurrent glioblastoma and recurrent anaplastic astrocytoma based on progression-free survival. Thus, the European system is coming close to an accelerated approval system like in the United States perhaps with more flexibility.

The situation is somewhat different in Japan. Objective response rate has played there the central role for oncology drug approvals where cytotoxic drugs can be approved based on tumor shrinkage in phase II studies, as defined in the guideline issued in 1991. The initial approval of a drug is considered to be conditional on a subsequent re-examination of the safety and efficacy of the drug at something like four to ten years after marketing

authorization. At least two independent randomized trials with survival as an endpoint need to be conducted in a post-marketing setting and results need to be made available at the time of re-examination.

At the international level, the International Conference on Harmonization (ICH) Guidelines on Statistical Principles for Clinical Trials state that

> "In practice, the strength of the evidence for surrogacy depends upon (i) the biological plausibility of the relationship, (ii) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (iii) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome"

(ICH Guidelines 1998). As such, it is close in spirit to the procedures proposed by the U.S., European, and Japanese regulatory authorities.

A detailed regulatory perspective is provided in Chapter 3.

2.3 Main Issues

Taking into account the arguments developed in the Introduction and earlier in this chapter, it is difficult to abandon the idea of using surrogate endpoints altogether, in spite of the failed attempts, described in the Introduction. However, it has also been stated, and this is in line with the regulatory authorities' policies, that there is a need for formal evaluation as an important component of the decision whether or not a surrogate endpoint can be used. Prentice (1989) formulated a definition of surrogate endpoints, as well as operational criteria for validating a surrogate endpoint. Freedman, Graubard, and Schatzkin (1992) introduced the concept of *proportion explained*, which was meant to indicate the proportion of the treatment effect mediated by the surrogate. Buyse and Molenberghs (1998) decomposed the proportion explained further into the *relative effect* and *adjusted association*, and argued in favor of using these quantities instead. The aforementioned proposals, reviewed in Chapter 5, were formulated under the assumption that the validation of a surrogate is based on data from a single randomized clinical trial.

This leads to problems with untestable assumptions and too low statistical power. To overcome these problems, the combination of information from several groups of patients (multi-center trials or meta-analyses) was suggested by Albert *et al.* (1998). It was subsequently implemented by Daniels

TABLE 2.1. *Examples of possible surrogate endpoints in various diseases (Abbreviations: AIDS = acquired immune deficiency syndrome; ARMD = age-related macular degeneration; HIV = human immunodeficiency virus).*

Disease	Surrogate endpoint	Type	Final endpoint	Type
Resectable solid tumor	Time to recurrence	Censored	Survival	Censored
Advanced cancer	Tumor response	Binary	Time to progression	Censored
Osteoporosis	Bone mineral density	Longitudinal	Fracture	Binary
Cardiovascular disease	Ejection fraction	Continuous	Myocardial infraction	Binary
Hypertension	Blood pressure	Longitudinal	Coronary heart disease	Binary
Arrhythmia	Arrhythmic episodes	Longitudinal	Survival	Censored
ARMD	6-month visual acuity	Continuous	24-month visual acuity	Continuous
Glaucoma	Intraoccular pressure	Continuous	Vision loss	Censored
Depression	Biomarkers	Multivariate	Depression scale	Continuous
HIV infection	CD4 counts + viral load	Multivariate	Progression to AIDS	Censored

and Hughes (1997), Buyse *et al.* (2000a) and Gail *et al.* (2000), among others. The meta-analytic framework is introduced in Chapter 7.

Statistically speaking, the surrogate endpoint and the clinical endpoint are realizations of random variables. As will be clear from the formalisms developed in Chapter 7, interest needs to focus on the joint distribution of these variables. The easiest situation is where both are Gaussian random variables. This is, however, seldom the case, because the surrogate endpoint and/or the clinical endpoint are often realizations of non-Gaussian random variables. Table 2.1 shows a number of settings that can occur in practice. Thus, grouped by type of endpoint, one can encounter:

- Binary (dichotomous): biomarker value below or above a certain threshold (e.g., CD4+ counts over 500/mm3) or clinical "success" (e.g., tumor shrinkage).
- Categorical (polychotomous): biomarker value falling in successive, ordered classes (e.g., cholesterol levels <200 mg/dl, 200–299 mg/dl, 300+ mg/dl) or clinical response (e.g., complete response, partial response, stable disease, progressive disease).
- Continuous (Gaussian): biomarker (e.g., log-PSA level) or clinical measurement (e.g., diastolic blood pressure).