Sauerbrei, W; Abrahamowicz, M; Altman, DG; le Cessie, S; Carpenter, J; STRATOS initiative, ; , COLLABORATORS; Abrahamowicz, M; Andersen, PK; Altman, D; Becher, H; Binder, H; Blettner, M; Bodicoat, D; Bossuyt, P; Carpenter, J; Carroll, R; Chadha-Boreham, H; Collins, G; De Stavola, B; Duchateau, L; Evans, S; Freedman, L; Gail, M; Goetghebeur, E; Gustafson, P; Harrell, F; Huebner, M; Jenkner, C; Kipnis, V; Kuechenhoff, H; le Cessie, S; Lee, K; Macaskill, P; Moodie, E; Pearce, N; Quantin, C; Rahnenfuehrer, J; Royston, P; Sauerbrei, W; Schumacher, M; Sekula, P; Stefanski, L; Steyerberg, E; Therneau, T; Tilling, K; Vach, W; Vickers, A; Wacholder, S; Waernbaum, I; White, I; Woodward, M (2014) STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. Statistics in medicine, 33 (30). pp. 5413-32. ISSN 0277-6715 DOI: https://doi.org/10.1002/sim.6265

Downloaded from: http://researchonline.lshtm.ac.uk/1883927/

DOI: 10.1002/sim.6265

Usage Guidelines

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/

Special Issue Paper

Statistics in Medicine

Received 18 December 2013, Accepted 20 June 2014 Published online 30 July 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6265

STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative

Willi Sauerbrei,[a*†] Michal Abrahamowicz,[b] Douglas G. Altman,[c] Saskia le Cessie,[d] and[‡] James Carpenter[e] on behalf of the STRATOS initiative

The validity and practical utility of observational medical research depends critically on good study design, excellent data quality, appropriate statistical methods and accurate interpretation of results. Statistical methodology has seen substantial development in recent times. Unfortunately, many of these methodological developments are ignored in practice. Consequently, design and analysis of observational studies often exhibit serious weaknesses. The lack of guidance on vital practical issues discourages many applied researchers from using more sophisticated and possibly more appropriate methods when analyzing observational studies. Furthermore, many analyses are conducted by researchers with a relatively weak statistical background and limited experience in using statistical methodology and software. Consequently, even 'standard' analyses reported in the medical literature are often flawed, casting doubt on their results and conclusions. An efficient way to help researchers to keep up with recent methodological developments is to develop guidance documents that are spread to the research community at large.

These observations led to the initiation of the strengthening analytical thinking for observational studies (STRATOS) initiative, a large collaboration of experts in many different areas of biostatistical research. The objective of STRATOS is to provide accessible and accurate guidance in the design and analysis of observational studies. The guidance is intended for applied statisticians and other data analysts with varying levels of statistical education, experience and interests.

In this article, we introduce the STRATOS initiative and its main aims, present the need for guidance documents and outline the planned approach and progress so far. We encourage other biostatisticians to become involved.
© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

Keywords: observational studies; guidance for analysis; level of statistical knowledge

1. Introduction

The validity and practical utility of medical research depend critically on good study design, excellent data quality, appropriate statistical methods and accurate interpretation of results. From this perspective, the inherent complexity of the processes involved in disease occurrence, progression and treatment, together with the increasing volume and complexity of data collected in medical studies, pose important

[a] *Center for Medical Biometry and Medical Informatics, Medical Center – University of Freiburg*
[b] *Department of Epidemiology and Biostatistics, McGill University, Montréal, QC H3A 0G4, Canada*
[c] *Centre for Statistics in Medicine, University of Oxford, Oxford OX3 7LD [Correction added on 15 August 2014, after first online publication: Postcode OX1 2JD has been corrected to OX3 7LD], U.K.*
[d] *Department of Clinical Epidemiology and Department for Medical Statistics and Bioinformatics, Leiden University Medical Centre, 2333 ZA Leiden, The Netherlands*
[e] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine, and MRC Clinical Trials Unit, Kingsway, London, U.K.*
[*] *Correspondence to: Willi Sauerbrei, Center for Medical Biometry and Medical Informatics, Medical Center – University of Freiburg*
[†] *E-mail: wfs@imbi.uni-freiburg.de*
[‡] *Correction added on 15 August 2014, after first online publication: placement of 'and' corrected.*
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

conceptual and analytical challenges. In response, the ever-growing statistical research community continues to develop and refine new methods, each aimed at addressing specific problems encountered in real-life data analyses. The richness and the complexity of the new methodology being published, every month, in several dozen statistical journals makes it impossible for a single statistician, or even a group of collaborators or members of a given institution, to keep pace with these developments. Further, the level of sophistication attained by statistical methodology at the beginning of the 21st century requires ever more specialization. As a result, many theoretical and applied statisticians are experts in only a few selected areas. However, data collected and analyzed in large medical research projects often require a number of analytical challenges to be successfully addressed simultaneously, each of which calls for a different state-of-the-art statistical method.

Equally important is the serious shortage of experienced statisticians [1], culminating in many analyses being conducted by researchers with a relatively weak statistical background and limited experience in using statistical methodology. Consequently, even 'standard' analyses reported in the medical literature often have major weaknesses, casting doubt on their results and conclusions. For example, dichotomization of a continuous variable or restricting multivariable analyses to subjects with no missing data ('complete records analysis') are still popular in spite of many statistical publications that have demonstrated serious limitations of these approaches and have proposed appropriate user-friendly methods. Often, even papers with a significant amount of statistical content have no statistician as a co-author. Similarly, many medical publications give an impression that available computer programs were used with only the most rudimentary grasp of the analytical properties of the methods being applied or the conditions necessary for their valid use. Those performing such analyses would greatly benefit from accessible guidance on important assumptions, potential problems, alternative analytical methods and information about available software.

Observational studies typically create more complex and varied analytical challenges than randomized clinical trials (RCTs). They use a wide range of designs (cohort, case–control, nested case–nested case–control, case–cohort, case series, etc.), data sources (e.g. large administrative data sets, retrospective and prospective clinical data), and address a rich variety of research questions. Each of these aspects has its own important implications for the analysis. For example, observational studies that seek to estimate the causal effect of an intervention, either in the absence of or to complement RCTs, typically have less rigorous design and data collection protocols than RCTs. This implies analytical complexities related to, for example, (i) imbalance of risk factor distributions due to non-random treatment allocation; (ii) the resulting risk of confounding bias, including confounding by indication and time-varying confounding; (iii) substantial variability in duration and intensity of, as well as compliance with, prescribed treatments; and (iv) coarsened data, due, for example, to measurement errors and missing observations. To address these challenges specialized, sophisticated methods such as causal graphs, structural models with outcome regression and inverse probability weighting, G-estimation, instrumental variables, principal stratification, multiple imputation and measurement error models are being used. Furthermore, the robustness of our inferences with respect to the assumptions made should be addressed. Diagnostic and prognostic studies also have their own particular challenges, even more when high-dimensional predictor spaces present themselves.

Obviously, each analysis does not throw up all the issues that may arise. Thus, it is important to keep the aims of the study in mind when deciding on suitable analysis strategies. The research question should determine the general approach to design and analysis and guide the researchers when deciding which methods to use and how to interpret their results. Even for experts, it is clear that several different types of analysis approach are possible and that each strategy raises some difficult methodological issues.

Because many methodological challenges are common to observational studies in different substantive areas, one efficient way to help individual data analysts or research teams to keep up with recent methodological developments is by disseminating guidelines or guidance, developed by experts in the relevant methodology. For example, for RCTs, several guidelines discuss the way data should be analyzed [2, 3] and how the results should be reported at different phases of drug development and approval process. Specifically, the CONSORT statement for the reporting of phase III RCTs represents the first, most comprehensive and influential set of reporting guidelines [4, 5], which continue to have a major impact on the practice of analysis and reporting of modern RCTs. Following CONSORT, several guidelines for reporting of many types of observational studies have been developed, for example, [6–8]. Some of them have been accompanied by explanation and elaboration articles, which explain each issue and provide examples [9–11]. Educational papers that focus on *specific* issues in the design and analysis of observational studies have appeared in statistical, epidemiological and medical journals (e.g. tutorials in

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

Statistics in Medicine, the Practice of Epidemiology section of the American Journal of Epidemiology or the Statistics in Oncology series in the Journal of Clinical Oncology). However, a *comprehensive* guidance series dealing with a variety of topics relevant to the design and analysis of observational studies, written and endorsed by panels of renowned specialists and general researchers, and developed through a common, integrated approach, is still lacking. The strengthening analytical thinking for observational studies (STRATOS) initiative aims to fill this gap.

The remainder of this article introduces the STRATOS initiative and its main aims (Section 2), and then presents our arguments for the need for guidance documents in Section 3. The STRATOS approach is outlined in Section 4, and progress so far is summarized in Section 5, with concluding remarks in Section 6.

2. Overview of the STRATOS initiative and its aims

2.1. Motivation

The STRATOS collaboration was initiated to help researchers cope with the methodological complexity arising from the broad range of issues potentially thrown up by observational studies. The over-arching objective is to provide accessible and accurate guidance for data analysts with different levels of statistical education and interests, taking into account the differences in their training and experience. Furthermore, analysis aims differ widely and influence the analysis strategy and the amount of work intended to be put in. In many projects, a state-of-the-art analysis is desired, whereas in others, researchers are satisfied with a relatively simple and straightforward approach.

Initiated and led by Willi Sauerbrei, preliminary ideas were presented to the Epidemiology subcommittee of the International Society for Clinical Biostatistics (ISCB) in 2011. The initiative is closely linked to ISCB and was formally launched in August 2013, at a dedicated mini-symposium during the 34th annual ISCB meeting in Munich. This paper in part presents a summary of the presentations and discussions at the mini-symposium and provides a template for the development of STRATOS, as well as an invitation to contribute.

2.2. Overall aims

STRATOS brings together applied and methodological experts in topics relevant to the analyses of observational studies. The experts' joint, largely complementary knowledge will allow us to address analytical issues in the design and analysis of observational studies. The STRATOS initiative seeks to address these issues by developing guidance documents, which will be both published in peer-reviewed journals and also made available on the STRATOS website with a moderated forum. This will provide the opportunity for researchers to submit their comments on the guidance. As for any guidance document, we will focus on 'typical' situations and 'generic' analytical issues, while recognizing that every study has several important 'individual' components. Thus, individual end-users of the documents will need to assess if and how the specific aspects of their studies may affect the relevance of the 'generic' recommendations' and/or require some modifications of the proposed approaches.

2.3. Three levels of expertise

One of the fundamental objectives of the STRATOS approach is to develop guidance documents for data analysts and researchers with different levels of statistical training, skills and experience. These documents will, of course, also be consistent with each other. Table I identifies the three levels of statistical knowledge for which the guidance documents are targetted and outlines the main criteria to be used when developing the guidance documents aimed at the analysts at each level, and the corresponding focus of our efforts.

2.4. Topic groups

Guidance document development will be carried out under the oversight of the STRATOS Steering Group (SG), by separate Topic Groups (TGs), each comprising experts in a given area of statistical methodology, alongside applied researchers who may represent future end-users of the STRATOS documents. The oversight provided by the STRATOS SG will ensure a broadly unified common framework, so that the resulting guidance is accessible and practically relevant. This will be achieved by those TG chairs who

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

Table I. Guidance for different levels of statistical knowledge.

Level 1
Low statistical knowledge
We have to assume that most analyses are carried out by analysts at that level. It is important to point out weaknesses of approaches that are often used despite of problems (e.g. categorizing continuous variables in the analysis; complete case analysis if some variables have missing values) and to propose methods that may not be optimal or state of the art, but which are easy to use and which are still acceptable from a methodological point of view. Required software should be generally available.
Level 2
Experienced statistician
Here we should point to methodology which is perhaps slightly below state of the art, but doable by every experienced analyst. We should refer to advantages and disadvantages of competing approaches, point to the importance and implications of underlying assumptions, and stress the necessity of sensitivity analyses. If these issues are well understood it is most likely that a sensible analysis strategy is chosen for the specific question. Sufficient guidance about software plays a key role that this approach is also used in practise.
Level 3
Expert in a specific area
To improve statistical models and to adapt them to complex problems in reality researches develop new and more complicated approaches. However, usually it is unclear whether the use of such an approach has relevant advantages in practise. Most often, advantages are presented in a small number of examples and in specific situations, but a more systematic comparison to the state-of-the-art is missing. Software requires specific knowledge and is not generally available.
This level would give an overview of recent research with statements about possible advantages and disadvantages of the approaches. It could help to identify important weaknesses when using level 2 proposals in more specific situations. It will certainly help to identify areas needing more methodological research and would trigger the development of software for more general use.

are also members of the SG. They will also help to ensure that the documents produced by the TGs are mutually complimentary.

2.5. Short-term aims

In the next 2 years, we will focus on a relatively few 'generic' topics that the SG has identified as highly relevant to a wide range of empirical observational studies. Seven initial TGs have been established, and these are described in Section 5. Further topics are being actively considered for inclusion. Each of the TGs should review the current literature and arrive at a consensus on the use of the appropriate methods, taking software availability into consideration . To this end, each TG will provide an overview of the current state of practice and try to identify any relevant existing guidance documents. Then, building on the literature, the TG members will attempt to reach consensus regarding the guiding principles and formulate specific recommendations. The recommendations will be supported by a range of examples, involving both real-life and simulated data, which will illustrate some pitfalls of inappropriate analyses and the potential advantages of the recommended methods.

The final STRATOS guidance documents, developed by each TG, should be (i) of acknowledged quality; (ii) accessible to end-users with various levels of statistical expertise; and (iii) readily available. To achieve the first of the aforementioned criteria, the TG should consist of renowned experts in all sub-areas of the particular topic. To meet the second criterion, we propose to write separate documents for three different levels of statistical expertise, as elaborated earlier (see also Table I). The third criterion requires publications of the main guidance documents in high impact journals, coupled with the use of online materials through a website, which will serve both as a repository for extended discussions and examples, and also as a forum for continuing discussion of the guidance proposals in the research community.

2.6. Long-term aims

Further stages of the STRATOS guidance development will build on the experience gained while working on the topics selected for the next two years. On one hand, the process tested and refined in the first stage of the initiative will be applied in order to develop similar guidance regarding several additional analytical challenges often encountered in observational studies, including more specialized topics whose relevance may be limited to a subset of study designs or statistical methods. On the other hand, while initial guidance documents will be developed *separately* for each of the selected topics, at the later

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

stages, a different TG will collaborate to gradually develop more comprehensive guidance, which will *simultaneously* address several closely related issues. For example, many studies exploring the role of continuous risk or prognostic factors, or explanatory variables, have to deal simultaneously with (i) measurement errors; (ii) missing data; and (iii) need to select the functional forms for modeling the effect of each variable on the outcome. Furthermore, the 'optimal' approaches to address each of the three issues may depend on how the other two issues are handled. Yet, in current practice, even if researchers use sophisticated methods to deal with one of these interrelated issues, the other issues are often ignored or handled using simplistic methods. Discussions across STRATOS TG members from each of the three fields will aim to identify a more comprehensive set of analytical strategies that will jointly address all the three issues.

3. Why do we need support and guidance documents?

The future will be characterized by an ever increasing ocean of data and statistical methodology that provides the tools for navigating this information to inform scientific and societal advances [1]. However, many analytical methods, even those in wide use, have serious weaknesses raising concerns about potentially incorrect results and conclusions. The impact of insufficient statistical knowledge in the research community, arising from lack of guidance and training in statistical methods, is strongly emphasized in a heavily cited article entitled 'Why most published research findings are false' [12]. Severe problems caused by lack of statistical expertise are also stressed in the public press. A recent article in *The Economist* entitled 'Unreliable research: Trouble at the lab.' [13] states:

> Scientists' grasp of statistics has not kept pace with the development of complex mathematical techniques for crunching data. Some scientists use inappropriate techniques because those are the ones they feel comfortable with; others latch on to new ones without understanding their subtleties. Some just rely on the methods built into their software, even if they don't understand them.

The poor reporting of clinical research studies in journal articles has long been a concern, leading to notable improvements during the last two decades [14]. Several years later, reporting guidelines for observational studies have also been developed, such as STROBE for epidemiological studies [8], STARD for diagnostic test accuracy studies [9] and REMARK for biomarker studies [7]. However, there has been little broad-based guidance on the appropriateness and utility of statistical methods to be used in the *analyses* of observational studies. Several recent publications, including some from individual members of the STRATOS initiative, address *specific* analytical issues. However, we are not aware of any *comprehensive* series of documents comparing the most relevant available strategies and providing sufficient guidance for many methodological and statistical issues frequently encountered in the analyses of observational studies. Thus, there is a clear need for the STRATOS initiative.

In the following section, we briefly illustrate some important weaknesses in statistical analyses used in current applied research. Then, we outline the challenges related to a rapid development of statistical methodology and their implications for applied research.

3.1. Weaknesses of many analyses

Every experienced statistician has encountered weaknesses of design or analyses in published studies, and often would have advocated a different analysis approach. The following weaknesses are typical:

- design [inappropriate/inefficient]
- choice of statistical methods [inappropriate/inefficient/outdated]
- use of the chosen methods [misapplication of valid method]
- interpretation [misinterpretation of *p*-value/over-confidence in results/misleading interpretation of parameter estimates/bias/confounding]
- reporting [inadequate details of methods and results].

Although some methodological errors relate to the failure to grasp some quite complex or indeed subtle statistical issues, there are widespread problems in the application of quite simple methods.

Reviews of the methodology used in published studies are far more common for randomized trials than observational studies. Also for observational studies, there have been fewer systematic studies of the weaknesses of analyses than studies of the completeness of reporting. Many examples of problems of reporting are given in the detailed 'Explanation and Elaboration' papers of reporting guidelines [9–11, 15].

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

Some common weaknesses in the methodology of observational studies have been discussed for many years [16, 17]. In the remainder of this section, we illustrate the range of issues by briefly summarizing the findings of some systematic reviews of specific methodological issues.

Specific study designs

A review of 125 reports of ecological studies in six major epidemiology journals revealed frequent important deficiencies [18]. Only 18% of the studies pre-specified the ecologic units; in 23 of 36 papers which standardized the outcomes for age or sex, the investigators failed to adjust for these potential confounders, and many investigators (49%) did not sufficiently consider the possibility of cross-level bias.

A detailed study of 37 matched case–control studies found that fewer than half (16/37; 43%) were analyzed with proper statistical techniques [19].

Categorizing continuous variables

Severe problems caused by categorization of continuous variables have been well-known for many years [20]; yet reviews of published prognostic factor studies in oncology show that almost all studies reported results for dichotomized marker values [21–23]. Further, variation of cutpoints across studies hinders a sound comparison and aggregation of results across studies. Altman *et al.* [20] found 19 different cutpoints for the marker SPF in breast cancer patients, and a review of p53 in bladder cancer found that definitions of positive p53 staining cut-off values ranged from 1% to 75% [24].

Interaction

Comparison of results from independent data (e.g. sub-groups) should be examined by direct statistical comparison of effects by a single test of interaction, not by comparing *p*-values from two separate tests [25]. Yet, among 157 behavioral, systems and cognitive neuroscience articles in five top-ranking journals, in which the authors described at least one situation in which the interaction test should be used, 79 (50%) used the incorrect procedure [26].

Multivariable models

Multivariable models are widely used in observational studies, as some degree of confounding is almost always expected. Yet multivariable models are often not used in situations that cry out for them. For example, only 15% of 184 studies on prognostic markers for acute pancreatitis used multivariable analyses [27].

An approach that is not recommended, namely, using statistical significance in univariate analysis as a pre-screening test to select variables for inclusion in the multivariate model, was applied in 48% (21/43) of a sample of multivariable models in oncology studies [22].

Of 39 risk prediction models in diabetes, 21 (49%) were developed by categorizing all continuous risk predictors. The treatment and handling of missing data was not reported in 16 studies (41%) [28].

Survival analysis

Several statistical problems are common in the application of multivariable methods in survival analyses. For example, 19% (127) of 682 observational studies in clinical journals that used a survival analysis included covariates not measurable at baseline and incorrectly modeled them as time-fixed covariates [29]. Among 42 articles in top-ranking cancer journals that employed Cox proportional hazards (PH) model to assess mortality, only 2 (5%) reported that they verified the crucial PH assumption [30].

Missing data

In 2007, Hippisley-Cox *et al*. [31] published QRISK, a new cardiovascular disease risk score based on an open prospective cohort study derived from routine electronic health records. At baseline, only 40% of participants had total serum cholesterol and 30% high density lipoprotein. However, following established practice in cardiovascular research, the ratio of serum cholesterol to high density lipoprotein levels was used in the prognostic modeling.

The authors used multiple imputation for the missing data. Their published risk prediction model showed no adjusted association between the incidence of cardiovascular disease and the cholesterol ratio (e.g. adjusted OR for Men 1.001, 95% CI 0.999, 1.002) a fact that attracted critical attention (http://

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

www.bmj.com/content/335/7611/136?tab=responses). Subsequently, the authors clarified that the analysis restricted to the subset of complete records showed a clear adjusted association between serum cholesterol ratio and cardiovascular risk. This was later confirmed using a revised, improved imputation procedure [31].

This experience motivated Sterne *et al.* to provide some practical advice on multiple imputation in this setting, along with reporting guidelines [32]; the need for this in the research community is illustrated by the fact that this has attracted over 600 citations. This has in turn motivated further research [33], which has pinpointed the likely cause of the original problem (instability in imputing a ratio as denominator values obtains close to zero), and how researchers might avoid it in future.

Summary

These examples illustrate the potential benefits of the STRATOS initiative. Readily available guidance at an appropriate level could have saved the situations, either through helping the original analysts or by alerting the peer reviewers to the potential issues. Furthermore, as in the missing data case, carefully presented guidance highlights the limits of current knowledge and stimulates research that addresses these.

3.2. Rapid development of statistical methodology and the necessity of guidance

In recent decades, statistical methodology has developed at an ever increasing rate. The exponential trend in the improvement of computer facilities can be viewed as the key driver of this development. Nowadays, it is possible, for example, to assess properties of complex estimation and inferential methods, and to compare complex model building strategies under a variety of alternative assumptions and sample sizes, using large-scale simulation studies. Computationally, intensive approaches including resampling methods such as the bootstrap and Bayesian modeling using MCMC allow investigations that were impossible even a decade ago and offer practical solutions to analytically intractable problems. In some fields, machine learning techniques provide interesting alternatives to more traditional model-based approaches. A wealth of new statistical software packages, often freely available through the Internet, allow a rapid implementation of novel statistical methods and comparison of the results of alternative approaches. However, many sensible improvements, potentially highly relevant to the data at hand, are ignored in practical statistical analyses. On the other hand, the ever increasing number of user-friendly programs creates an 'embarras de choix' and increases the risk of unwarranted use of the available methods in real-life analyses where the underlying assumptions are badly violated. Indeed, while in the past three decades, the Cox PH model has become by far the most popular statistical tool for survival analyses, the crucial PH assumption is seldom tested [29] even if it is frequently violated by several important prognostic factors, across a range of different cancer sites, for example, [34–36].

In many areas of science, empirical data are analyzed with the aim of deriving a model to help assess or test the associations between relevant variables. Statistical models are, of course, always a simplification of real life processes. Yet, following the dictum that "all models are wrong, but some are useful" [37], to better account for the complexities of the empirical data and processes, statisticians continue developing new and more complicated approaches. Obviously, expert knowledge is required to identify the novel method(s) able to address specific analytical challenges of a given empirical study, understand the importance and implications of the underlying assumptions and apply these novel models, using available statistical software.

For a specific example, consider the Framingham Heart Study, the flagship example of a longitudinal, prospective observational cohort study [38–40]. A researcher who wishes to analyze these data to explore a potential association between an 'emerging', not yet well established, continuous risk factor for cardiovascular disease (CVD), will have to simultaneously address several complex issues, for example, the need to (i) select functional forms for both the risk factor of interest and several inter-correlated covariates, representing well-established CVD risk factors; (ii) account for time-varying effects of variables, which are not monitored continuously but at 2-year intervals, that is, decide how to aggregate past measures and/or whether to use most recent or baseline or lagged (e.g. by 2 or 4 years) values; (iii) account for interval-censored data (outcomes established only at the time of clinic visits, at 2-year intervals); (iv) correct for measurement errors; (v) decide whether and how to impute missing data; and (vi) account for time-varying confounding.

For many topics, the current state of recommendations is well described by the experience from a systematic review on the relatively simple issue of investigating clinical heterogeneity in systematic reviews.

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

Gagnier *et al.* [41] summarize ' Though the consensus was minimal, there were many recommendations in the literature for investigating clinical heterogeneity in systematic reviews. Formal recommendations for investigating clinical heterogeneity in systematic reviews of controlled trials are required.' Findings such as this illustrate the desirability and the necessity of producing high quality guidance documents, on the basis of consensus reached by a large group of experts, and of wide dissemination of such guidance, with the ultimate aims of enhancing the quality of future research.

3.3. Lack of guidance and its consequences

As illustrated earlier, many published analyses use inappropriate methods, which may lead to biased estimates, inaccurate inference and possibly incorrect conclusions. Sometimes methods employed may be acceptable but—by not making use of the recent state-of-the-art developments—they may be inefficient and may fail to discover some salient features of the data at hand. Even experienced analysts and experts in specific fields of statistical research will sometimes be confronted with analytical issues requiring methods developed in other fields that they are unfamiliar with.

Furthermore, we have to assume that *published* studies likely represent a non-random selection of the very large number of analyses conducted every day, some of which are not accepted for publication *because of* inappropriate statistical methods and/or incorrect interpretation of results. In addition, many analyses are never summarized in a manuscript and submitted for publication, often because the authors are aware of the substantial weaknesses of their analyses but they have no access to the required statistical expertise. This is often the case with clinical data analyzed by medical students and residents. To address this, level 1 guidance documents would be most helpful. Moreover, even interdisciplinary research teams that collaborate on 'large' medical studies may lack statistical expertise regarding state-of-the-art methods necessary to address some specific analytical challenges. In such cases, limitations of the design and analyses may result in a waste of potentially valuable data and possibly affect the chance of publishing the results. On the other hand, if the reviewers of journals also lack expertise in the relevant field of statistics, a study that employs seriously flawed methods could be published. In all aforementioned situations, data analysts and researchers stand to gain substantially from guidance tailored to their experience and requirements. Such guidance will give the confidence in conducting suitable analyses and in critiquing the statistical analysis in articles submitted for peer review.

Obviously, in many situations, it may neither be possible nor necessary to perform a 'state-of-the-art' analysis. Firstly, guidance documents are lacking, and there might be no consensus concerning a 'start-of-the-art' analysis. Secondly, this could be due to limitations of the available data, insufficient sample size or restricted resources. Nevertheless, it will still be essential to provide arguments and evidence to distinguish between acceptable and seriously flawed analytical strategies and analyses.

In conclusion, the complexity of the novel, powerful methods described in theoretical statistical literature, combined with often very limited guidance on key issues that are vital in practice, frequently discourages data analysts from utilizing more sophisticated, and possibly more appropriate methods, or—conversely—may result in an unwarranted use of the available methods, both traditional and novel, and incorrect interpretation of their results. To improve the situation, experts in specific methodological areas need to work together to develop comprehensive guidance for different practical issues in statistical analysis. This will require international collaboration, in order to identify the relevant methods, including those recently developed, to assess the underlying assumptions and their practical implications, as well as—if required—to evaluate and compare their formal properties and performance. Our initiative will foster quicker evaluation of new methodology, as well as its dissemination to the broader community of researchers analyzing real data, and its adoption to practice.

4. Concept, structure and the general approach of the STRATOS initiative

4.1. Structure and governance

The internal structure of the STRATOS research team has been designed so as to facilitate the complex, multi-task multi-stage process of the guidance development, revisions and dissemination. In general, the two-layer structure includes the SG, which will oversee and manage the overall process, and several TGs, each of which will focus on a specific analytical issue, or a subset of closely inter-related issues, highly relevant for a wide range of observational studies (Section 5). SG members will participate in TGs corresponding to their areas of expertise and research interests, and individual members of the STRATOS team may participate in more than one TG. The SG will identify issues to be tackled by individual TGs

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

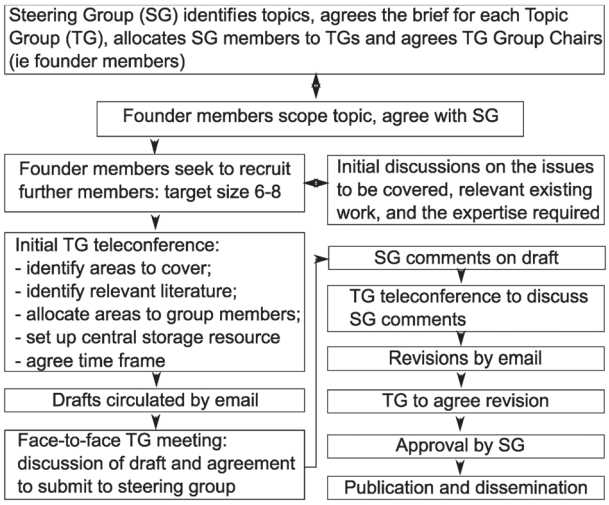

Figure 1. STRATOS structure and the initial road map.

and will help selecting suitable chairpersons, including one from the SG, set priorities, take strategic decisions and coordinate the work of different TGs.

Currently, the idea is to establish a 'headquarter' for the initiative in Freiburg, Germany, and to have another center on a different continent. Establishing these centers is important for coordinating and supporting the initiative at various levels, such as creating and maintaining a website, coordinating applications for funding, efficient networking and organization of future meetings and events, and helping with literature searches.

4.2. Key stages in guidance development

Figure 1 outlines the initial 'plan of attack' and identifies crucial junctions of the proposed 'road map' leading toward the achievement of the STRATOS aims, including major stages of the development of the guidance by individual TGs, and their interactions with the SG. Section 5 identifies TGs that have been already created and illustrates their work in progress, to provide more insight into the underlying concepts, type of challenges to be faced, general way of thinking and the resulting process of development, revision and refinement of the guidance documents.

4.3. Phase I: developing guidance for knowledge level 2

One of the fundamental principles underlying STRATOS is our recognition of the necessity to develop different guidance documents for data analysts and researchers with different levels of statistical training, skills and experience (Table I). It is essential that the documents be consistent with each other.

Each TG will start by developing guidance aimed primarily at level 2 statistical knowledge (Table I), that is, suitable for experienced data analysts, typically with at least an MS degree in (bio)statistics and adequate hands-on experience of real-life data analyses. Level 2 statisticians are generally able to implement relatively complex methods, provided they are directed toward the appropriate software and given guidance regarding the data structures and conditions necessary for the method to be valid. They can also accurately interpret the results of novel methods, given pertinent examples. On the other hand, they are not experts and therefore may feel uncomfortable about taking complex decisions regarding the choice between alternative cutting-edge methods, adaptation of the recently developed methods to new data structures or study designs, or combining different methods into a single, multi-faceted analysis. Consequently, level 2 guidance will focus on the identification of the appropriate, well tested and documented, and easily implementable methods that avoid major pitfalls of the simplified 'conventional analyses' but may still fall somewhat short of the state-of-the-art methodology already existing in the statistical literature.

Often, many alternative approaches will have been proposed for a specific task in the literature. Obviously, deriving comprehensive guidance documents requires agreement between the group of 'experts' participating in a given TG, who may adhere to different schools of thought, and have different (subjective) preferences for, as well as different experiences with, alternative methods. Such variation

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

may be even more marked if some of the experts have actually developed some of the methods under consideration. Therefore, an important requirement, and challenge, for the SG, is to derive some general agreement about the due process to be used in the assessment of alternative statistical approaches, and the criteria according to which their suitability and usefulness for specific applications can be judged in practice. Such discussions will also help to identify questions requiring further primary research. To illustrate, while the panel of experts from a TG may rule out some methods as definitely sub-optimal, while some other methods may be determined to be clearly superior, there will be situations where either systematic comparisons between the competing methods are not available or their results are inconclusive. Thus, a further function of guidance documents will be to point out methods that perform about equally well, to some extent allowing for selection based on individual preference and/or plausibility of the underlying assumptions in the context of a particular empirical study.

Section 4.5 briefly outlines how the results will be disseminated through educational activities, links with other societies and the website with a moderated forum.

4.4. Phase II: extending to knowledge levels 1 and 3

Once level 2 guidance has been developed, and approved by both the relevant TG and the SG, and the process of dissemination begun, the TGs will turn their attention to adapting and extending their recommendations to the other two levels of statistical knowledge. In particular, level 1 guidance will target data analysts with a relatively weak statistical background, and limited practical experience, such as medical students or residents, or epidemiologists who completed only a few basic courses in applied statistics. Level 1 analysts tend to choose the method to be applied in a given study either based on the availability of the software, for example, a program widely used in their research lab or by other members of their team, or by a copy-cat strategy, often resulting in non-discriminating adaptation of the method found in recent publications from the same substantive area, or published in the 'target' clinical journal. Both these merry-go-around strategies lead to uncritical propagation of some popular methods, without any serious consideration of either the underlying assumptions or the salient aspects of the study design or data structure, specific to a given study, which may invalidate the chosen method.

Accordingly, one of the primary aims of level 1 guidance documents will be the eradication of major errors. This will be illustrated by real-life and, if useful, simulated examples and accompanied by recommendations on the readily available, user-friendly methods that can help in identifying and/or avoiding major problems and biases, while not requiring in-depth understanding of the method. Accessibility of the software and ease of interpretation of the results will be important criteria for level 1 recommendations, implying that some more powerful/efficient, yet more complex methods, recommended in level 2 documents, may not be recommended at level 1. However, the guidance documents will be regularly updated and improved, so that gradually some of the level 2 recommendations will be moved to level 1, especially if the software becomes more user-friendly.

In parallel, level 2 documents will provide a starting point from which to develop the level 3 guidance, aiming at very experienced data analysts, usually with a PhD in (bio-)statistics, and active researchers. We expect the level 3 documents to be less 'restrictive' or 'categorical', and more nuanced, in their recommendations. For example, level 3 guidance may reflect, when relevant, the existence of alternative, possibly competing approaches designed to address the same or similar analytical challenges and the fact that there may be insufficient evidence regarding their advantages and disadvantages, or how their relative performance may depend on some design parameters or characteristics of the data at hand. Indeed, one important output of the level 3 guidance will be to identify the need for further research on the evaluation and development of various methods. To this end, the original TG may be expanded by including additional experts actively involved in development and refining of the cutting-edge methods in a respective area of statistical research.

4.5. Contact with related projects and scientific societies

To achieve its long-term aims, STRATOS will develop and gradually refine a multi-stage guidance development and evaluation process, as outlined in Figure 1. In these endeavors, we will partly build on the experience of the groups that have developed reporting guidelines in observational studies and clinical trials, as their experience of forming and successfully working with multi-disciplinary international teams will be highly relevant.

We emphasize the need to ensure both the direct relevance of the issues being addressed and of the STRATOS outputs to the wider community of data analysts working in different fields of empirical

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

research, and the efficient dissemination of the guidance documents to this community. Collaboration with well-established scientific organizations will play an essential role. We are linked to the ISCB, but we intend to establish formal links with other international societies and research organizations, including the International Biometric Society (IBS), International Society for Pharmacoepidemiology (ISPE), International Epidemiological Association (IEA), American Epidemiological Association and other, ongoing initiatives that typically focus on more specialized issues or fields of study, many of which also involve STRATOS members (for example, EQUATOR http://www.equator-network.org/ and PCORI http://www.pcori.org/).

One direct product of such collaborative links will take the form of educational sessions and mini-symposia at the annual meetings of the respective societies. Indeed, the STRATOS initative was launched at a half-day mini-symposium on the last day of the ISCB meeting in Munich, in August 2013. The fact that this was attended by more than 100 researchers shows the potential impact of such endeavors. Unknown to the current STRATOS members, most of whom focus primarily on medical biostatistics, some relevant guidance documents may have been developed by statisticians working in other fields of empirical research. Further, by connecting STRATOS with the leading societies that bring together statisticians and analysts from a broad bio-medical spectrum, we may identify some already existing guidance documents that will feed into the work of the STRATOS TGs. Finally, another important aspect of the collaboration will involve interactive links from the respective societies' websites to the STRATOS website, which (as discussed earlier) will provide a moderated forum for discussions around, and ultimately refinement of, the guidance documents. From this perspective, we emphasize that STRATOS will strive for a high degree of transparency in its work, with the website as a key resource for our initiative. Obviously, to ensure high quality, it has to be managed and controlled by the initiative.

5. Topics to start

The SG of the initiative has decided to start with seven topics (Table II). We intend that each TG will have about 6 to 10 members, of which at least one is also a member of the SG. Each group will work differently, but the SG members will try to ensure coherence concerning the methodological approaches

Table II. Topics and members.

	Topic group		Chairs and further members
1	Missing data	Chairs:	James Carpenter
		Members:	Els Goetghebeur, Kate Lee, Rod Little, Kate Tilling, Ian White
2	Selection of variables and functional forms in multivariable analysis	Chairs:	Michal Abrahamowicz, Willi Sauerbrei
		Members:	Harald Binder, Frank Harrell, Patrick Royston
3	Descriptive and initial data analysis	Chairs:	Marianne Huebner, Saskia le Cessie, Werner Vach
		Members:	Maria Blettner, Danielle Bodicoat
4	Measurement error and misclassification	Chairs:	Raymond Carroll, Laurence Freedman
		Members:	Paul Gustafson, Victor Kipnis, Helmut Küchenhoff, Len Stefanski
5	Study design	Chairs:	Mitchell Gail, Neil Pearce
		Members:	Doug Altman, Gary Collins, Luc Duchateau, Stephen Evans, Peggy Sekula, Sharom Wacholder, Mark Woodward
6	Evaluating diagnostic tests and prediction models	Chairs:	Petra Macaskill, Ewout Steyerberg, Andrew Vickers
		Members:	Patrick Bossuyt, Gary Collins
7	Causal inference	Chairs:	Els Goetghebeur, Erica Moodie
		Members:	Bianca De Stavola, Saskia le Cessie, Ingeborg Waernbaum

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

and the output. The SG will set out the general principles for the TGs to work in. In this section, we provide short summaries of work-in-progress from the seven TGs. In further updates (to be posted on the website), they will be extended and gradually harmonized across TGs. Most of the TGs will be further expanded by inclusion of additional members.

Future stages of STRATOS will involve further extensions aimed at both combining activities of different TGs to develop guidance for researchers who have to simultaneously address issues discussed by relevant TGs and add new TGs to tackle additional analytical challenges. For example, recently, we have decided to start with a TG on survival analysis, and start of further TGs is under discussion.

5.1. TG1: missing data

Missing data are ubiquitous in observational studies, and the simple solution of restricting the analyses to the subset with complete records will often result in bias and loss of power. The seriousness of these issues for resulting inferences depends on both the mechanism causing the missing data and the form of the substantive question and associated model.

The methodological literature on methods for the analysis of partially observed data has grown substantially over the last 20 years, for example, [42, 43] and references therein, such that it may be hard for analysts to identify appropriate (but not unduly complex) methods for their setting. Our aim is to draw on both the existing advice, for example, [32, 44–46] and the expertise of the TG1 members, to provide practical guidance that will lead to appropriate analysis in standard observational settings, while giving principles that can inform analysis plans for less common substantive models.

To achieve this aim, the TG will describe a set of principles for the analysis of partially observed observational data and illustrate their application in a range of settings, ranging from simple summaries of single variables through regression models, models for hierarchical and longitudinal data and models to adjust for time varying confounding.

Specifically, we aim to

- assist analysts in understanding the nature of the additional assumptions inherent in the analysis of partially observed data;
- describe, in a range of settings, the implications of these assumptions for analyses that restrict to the subset of complete records;
- detail the range of methods available for improving on a complete records analysis, including the EM and related algorithms, multiple imputation, inverse probability and doubly robust methods; and
- provide guidance on the utility and pitfalls of each approach, bearing in mind the importance of software availability for most applied researchers.

In particular, we will delineate how the various methods relate to each other and in particular when they are likely to give similar answers.

Because the data at hand cannot definitively identify the missing data mechanism, exploring the robustness of inferences to departures from the primary assumption about the missing data mechanism is important in many applications. We will discuss how to frame assumptions for such sensitivity analyses and practical approaches to analyses under these assumptions.

Consistent with the STRATOS initiative, the group will not focus narrowly on missing data (which is itself but an extreme form of coarsened data, such as measurement error) but place its guidance firmly in the context of appropriate design and statistical methods for the inferential question at hand. Rather than providing a recipe book, our aim is to foster understanding of the key principles, so that methods can be chosen and applied with confidence. As a result of our work, it is inevitable that key areas requiring further research will emerge, and we anticipate that scoping their nature will also be a useful contribution to future research in this area.

5.2. TG2: selection of variables and functional forms in multivariable analysis

In multivariable analysis, it is common to have a mix of binary, categorical (ordinal or unordered) and continuous variables that may influence an outcome. While TG6 considers the situation where the main task is predicting the outcome as accurately as possible, the main focus of TG2 is to identify influential variables and gain insight into their individual and joint relationship with the outcome. Two of the (interrelated) main challenges are selection of variables for inclusion in a multivariable explanatory model and choice of the functional forms for continuous variables [47, 48].

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

In practice, multivariable models are usually built through a combination of (i) a priori inclusion of well-established 'predictors' of the outcome of interest and (ii) a posteriori selection of additional variables, based often on arbitrary, data-dependent procedures and criteria such as statistical significance or goodness-of-fit measures. There is a consensus that all of the many suggested model building strategies have weaknesses [49] but opinions on the relative advantages and disadvantages of particular strategies differ considerably.

The effects of continuous predictors are typically modeled by either categorizing them (which raises such issues as the number of categories, cutpoint values, implausibility of the resulting step-function relationships, local biases, power loss, or invalidity of inference in case of data-dependent cutpoints) [50] or assuming linear relationships with the outcome, possibly after a simple transformation (e.g. logarithmic or quadratic). Often, however, the reasons for choosing such conventional representation of continuous variables are not discussed and the validity of the underlying assumptions is not assessed

To address these limitations, statisticians have developed flexible modeling techniques based on various types of smoothers, including fractional polynomials [51, 52] and several 'flavors' of splines. The latter include restricted regression splines [47, 53], penalized regression splines [54] and smoothing splines [55]. For multivariable analysis, these smoothers have been incorporated in generalized additive models.

Various examples illustrate that such smoothers can yield new insight into the role of continuous variables [52, 56]. However, further practical guidance is urgently needed, necessitating extended investigations of analytical properties and systematic comparisons between alternative methods. TG2 will start with a comprehensive review of methodological, medical and econometrics literature to (i) identify and assess methods currently used in practice, (ii) find any published guidelines on selection of variables and their functional forms, and (iii) find systematic simulation-based comparisons of alternative techniques, especially in multivariable analyses [57]. Part (iii) may lead to new comparative simulation studies and provide building blocks for evaluation of new techniques by simulation.

We aim to develop consensus-based tentative recommendations, initially for level 2 expertise, under some simplifying assumptions about the data structure. Recommendations will address accuracy, efficiency, transportability, ease of implementation and interpretability, in wide range of applications [48]. Furthermore, we aim to develop systematic guidance for using splines in applications, similar to existing guidelines for fractional polynomials [52]. Longer-term goals include evaluation of and recommendations for computationally intensive variable selection algorithms that incorporate shrinkage and resampling techniques, collaborations with other TGs to account for such complexities as missing data, measurement errors, time-varying confounding or issues specific to modeling continuous predictors in survival analyses [58].

5.3. TG3: descriptive and initial data analysis

The initial steps of all data analyses consist of checking consistency and accuracy of the data, describing and exploring the study sample and preparing the data for further analyses. It is crucial that this is performed before embarking on complex analyses.

The drive to obtain high quality data should start long before data collection and include a careful database design with variable definitions, plausibility checks, date checks and a well-planed system for identifying likely data for errors and resolving inconsistencies. Cleaning data especially when integrating multiple data sources should be carried out systematically and carefully [59]. After being (reasonably) confident that the data are error-free, the next step is to become familiar with the collected data and examine it for consistency of data formats, number and patterns of missing data, distributions of continuous variables, (e.g. skewness, variation) and frequencies of categorical variables, checking group labels [60, 61]. The inclusion and exclusion criteria in the process of selecting the subset of data to be analyzed in the study should be described along with an overview of missing measurements and follow-up data [8]. Both raw data and the final data set need to be saved. A general rule is that a statistical analysis plan is made and agreed upon before the data collection starts and that it should not be altered without agreement of the project SG. This should reduce the extent of data dredging or hypothesis fishing leading to false positive studies [62]. It is important that the complete initial data analysis process is transparent and that researchers document all steps for reproducibility [63].

One of the aims of the initial data analysis is to provide a clear description of the data in tables and figures. This can be done in many different ways: summary statistics can be reported for the total population or for subgroups; continuous variables can be summarized by means and standard deviations, by

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

medians and percentiles or by categorizing them. Small groups of categorical variables can be combined. Medical papers often have a descriptive table of the data. In certain instances, for example when data are missing not completely at random, summary statistics of the study sample do not unbiasedly represent population characteristics. In this case, one has to decide whether the descriptive table should include corrections for the missing data.

Another step is preparation of the data for more advanced analyses. In this step decisions have to be made about the way variables are used in further analyses. Variables can be used in their raw form, they may be transformed or categorized, they may be rescaled or standardized, they may be used as single variables, combined to summary scores or as more complex functions, e.g. as ratios [64]. Further, procedures to handle missing values and outliers should be clarified.

This topic group aims to provide guidance for all of the above mentioned steps in the initial data analysis. We will discuss advantages and limitations of different approaches. Recommendations will be given based on an overview of existing literature and feedback from experienced researchers.

5.4. TG4: measurement error and misclassification

Measurement error and misclassification, hereafter MEM, in covariates and responses occurs in many observational studies and some experimental studies. See [65, 66] and [67] for textbook treatments of the literature.

The MEM can be taken as an extreme version of a missing data problem, where the true data affected by error are missing in all or the vast majority of subjects. This special feature requires special methodology and means of thinking.

It is well-known that MEM in predictors can result in biased parameter estimates and a loss of statistical power, sometimes a startling loss of power. Not accounting for MEM can lead to over-optimistic power calculations and result in failed studies. It is less well-known that measurement error can, in some circumstances, lead to incorrect inferences and hypothesis testing; that is, it is not always the case that MEM results in a bias toward the null.

We will provide guidance into the following series of topics. The listing is not in order of importance.

- When will MEM likely affect the validity of statistical analysis?
- When will MEM result simply in a bias toward the null, so that parameter estimates are biased by hypothesis testing for null effects is valid but less powerful than if there were no MEM? Conversely, when will MEM result in bias that is not toward the null, so that incorrect inferences and conclusions will result?
- How can one design a study in the presence of MEM both to assess their extent and to provide properly tuned sample sizes for sufficient statistical power?
- What software is available to make sample size calculations when there is measurement error or misclassification in the key predictor?
- What is the role of Bayesian analysis in MEM?
- What software is available for a measurement error analysis? We will give advice on available software for measurement error analysis. Here is a sampling.
    - In SAS, the CALIS procedure, SAS macros available from the National Cancer Institute (http://appliedresearch.cancer.gov/diet/usualintakes/)programs from Iowa State University (http://www.cssm.iastate.edu/software/cside.html) and programs from the Harvard School of Public Health (http://www.hsph.harvard.edu/donna-spiegelman/software/) are available. In R, there are the 'decon' and 'simex' packages. Stata also has programs for regression calibration
- What analyses are possible if some variables are subject to measurement error and others are subject to misclassification?
- What can be done if measurement error or misclassification is known to exist but data are not available in a current study to assess the impacts of these errors?
- How can one perform variable selection when some predictors are subject to MEM?
- What can be done if there is MEM in the response variable?

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

5.5. TG5: study design

Appropriate and valid study design is crucial for the valid conduct of observational studies [8]. Observational studies can make a key contribution to establishing causal relationships, in combination with other types of evidence (e.g. mechanistic studies, animal studies) [68]. They may be particularly valuable when randomized trials are impractical and/or unethical, for example, for exposures such as smoking and lung cancer [68]. Some observational studies observe and measure associations without any attempt to infer causality, for example, prognostic studies relating to biomarkers. Other studies may seek solely to have unbiased estimates of some phenomenon, such as the prevalence of some condition in a community.

The appropriateness of any study design thus depends on the research question, in the context of the current state of theory and knowledge, the availability of valid measurement tools and the proposed uses of the information to be gathered. Although, in theory, certain study designs are better than others, in practice, the validity of a study design is highly topic and context-specific [69]. Hierarchies of observational study designs are often proposed (usually with cohort studies at the top, followed by case–control studies, cross-sectional studies, etc). However, their relative validity represents a continuum, rather than a dichotomy, and 'less valid' study designs may yield valuable information in some instances [69].

It is highly unusual for a single observational study to deliver definitive results. Assessing the epidemiologic evidence almost always involves a process of triangulation across studies in different populations, using a variety of study designs, investigators and methods. No individual study can be perfect or can deliver a definitive answer. Rather, the aim of a particular study is to contribute to the pool of knowledge for a particular issue.

With these considerations in mind, there are a number of important considerations with regard to design of observational studies:

1. What is the question that one wants to answer? What does this imply about the basic design?
2. What is the most efficient study design?
3. What is the most appropriate (and available) study source population and risk period?
4. How can the design assist in an effort to control for potential confounding?
5. How can design assist in assuring reliable exposure assessment?
6. How can the design assist in assuring reliable and complete disease ascertainment?
7. What are the best methods to assure completeness of the data from sampled subjects?
8. What is the role (if any) of subgroup analysis, and what does this imply for the study design?
9. What are the specific issues involved in the design of studies of prognosis (prognostic factors and prognostic modeling studies) and diagnosis?

Appropriate and valid study design involves a context-specific balance between these competing considerations [70]. Whatever study design is used, it is also important to have the capacity to conduct sensitivity analyses, for example, by gathering information on the likely extent of uncontrolled confounding and measurement error.

This TG will review these issues and will produce guidelines (but not inflexible rules) for weighing up these considerations when designing an observational study.

5.6. TG6: evaluating diagnostic tests and prediction models

Methods and measures are well established to evaluate the accuracy of a diagnostic test for classifying individuals according to whether they do, or do not, have a condition of interest [71, 72]. A test may be intrinsically binary, ordinal or continuous, with a reference standard ('gold standard') used to define whether patients have the target condition. For a binary test, the most commonly reported measures of test accuracy are sensitivity and specificity. These estimate the probability that the test result is correct, conditional on the disease status of an individual. Other commonly reported measures include likelihood ratios, and positive and negative predictive values of the test. For ordinal and continuous tests, a receiver operating characteristic (ROC) curve is typically used to represent the trade-off in sensitivity and specificity as the test threshold varies, with the area under the ROC curve used as a global measure of test accuracy or discrimination. Sensitivity, specificity and other measures are often reported at chosen cut-point(s) for test positivity. These methods underpin diagnostic test evaluation, but they are not necessarily well applied or interpreted. Indeed, there are concerns about the value of sensitivity and specificity as metrics.

It is widely accepted that test performance is likely to vary according to the context in which the test is used, for instance, where it lies in the clinical pathway. This has clear implications for the (potential) role

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

of the test, for example, a replacement for an existing test, an 'add-on' test or a triage test [73]. Evaluation of the potential role of a test has implications for study design as well as methods for test comparisons and assessing the gain in using tests in combination [74].

In practice, a test is generally used in conjunction with other information such as the age of the patient, their sex, symptoms, clinical signs and possibly the results of other tests when making a diagnosis. Multivariable logistic regression is commonly used to develop a model to predict the presence of disease, thereby utilizing all available relevant information for diagnostic decision support. Prediction modeling is especially important also in the area of prognosis, where survival modeling such as Cox regression is frequently used to predict the probability of a given outcome (e.g. mortality) in the future, on the basis of the profile of an individual in terms of prognostic factors, test results, biomarkers and so on. Model performance is usually assessed in terms of overall predictive performance, discrimination (ability to classify individuals correctly into two outcome categories) and calibration (agreement between the predicted probabilities and observed outcomes) [75]. While these are generally regarded as the key criteria for assessing model performance, specific methods and measures vary occasionally producing inconsistent or conflicting results. This is especially evident when assessing the incremental gain of adding a test or biomarker to a model [76, 77].

Diagnostic tests and model predictions are imperfect. Thus, there is a potential for harm as well as benefit in terms of decisions regarding (further) investigations, treatment and prognosis for individuals. Even though an evaluation may indicate good diagnostic accuracy or model performance, evaluation of clinical utility requires determining whether decisions based on the test or model improve patient outcomes. This can be performed either by decision analysis (net benefit) [75] or by prospective cost effectiveness analysis [78].

We will review diagnostic test evaluation in terms of methods, measures and study designs that are relevant to the assessment of a test in terms of its intended use. We will also review methods for the evaluation of prediction models for diagnosis and prognosis, with a particular focus on reclassification and approaches that assess clinical utility.

In the longer term, we will consider extending the aforementioned framework to address prediction models for differential diagnosis, dealing with missing data (e.g. incomplete verification), assessment of calibration, consistency with economically relevant outcomes (e.g. medical costs and quality adjusted life years) and impact of measurement error (e.g. error in the reference standard).

5.7. TG7: causal inference

The desire to draw causal inference from observed associations is age-old. The ensuing quest has contributed greatly to scientific progress. While simple association models have gradually gained in sophistication and their potential is typically well understood by practicing statisticians, causal questions and answers need an extra dimension of abstraction that calls for special care and caution. The move from association to causation is by no means trivial and requires assumptions not only about the observed data structure but also beyond the sampled data. Notorious examples, such as the hormone replacement therapy story, have taught us that lesson.

This TG sets out to provide guidance on the sequence of steps involved in causal inference. This includes phrasing the causal question, designing a sampling frame and/or selecting the observational data, formulating assumptions to justify specific causal effect estimators, reporting results and [79] conducting sensitivity analyses for untestable assumptions. Several formalisms and schools of thought have been developed over the past decades that have deepened our insight, expanded the tool kit available and made the questions we can hope to address more ambitious.

Data selection and corresponding assumptions on the data structure will determine the specific causal parameter we set out to find. When the causal effect of a single treatment regime is envisaged, one may attempt to mimic a randomized trial either by controlling for all necessary confounders or by relying on an instrumental variable. In either case, an estimator of the intention-to-treat effect, the per-protocol effect or as-treated effect may follow. When more ambitiously we aim to estimate the effect of a dynamic treatment regime, or sequence of treatment decision rules in response to covariates evolving over time, longitudinal data are needed and a more stringent set of assumption involve no unmeasured time-varying confounders or the equivalent of sequential randomization. Adjusting for time-varying confounders that are at the same time intermediate variables on the causal path from exposure to outcome is a special challenge and best achieved through an inverse probability weighting technique.

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

To understand the causal structure and assumptions, we are willing to impose that on a data problem, causal diagrams [80] and/or the formalism of potential outcomes can be very helpful. They can also point to estimators for the target parameter. Many different estimation techniques exist, and the terminology includes besides causal graphs the following: augmented inverse probability weighting (with stabilized weights), doubly robust procedures, G-computation, marginal structural models, (robust) multiple imputation, propensity scores, principal stratification and so forth. Some relevant references are [79–89].

Drawing this together, the TG aims to advise on the following:

- Classes of causal questions to consider with options of data structures
- Corresponding analysis techniques, their dependence on (untestable) assumptions and available software options. Design recommendations.
- Tools that help visualize and interpret assumptions, a basis for discussions with clinicians and study design considerations
- Pros and cons of specific estimation approaches in terms of the bias/variance trade of, transparency and ease of implementation, robustness and back-up interpretation when assumptions for causal inference fail.
- Pointers to tutorials and worked out case studies
- Bridges between the jargon and terminology used by different schools of thought
- Point to tools and tutorials for sensitivity analysis

6. Summary and discussion

Our aim in writing this article has been twofold: to motivate the STRATOS initiative and to introduce it to the statistical community, thereby to encourage researchers to consider becoming involved.

With regard to the former aim, we have argued that the frequency and seriousness of analytical errors in current applied research, together with continuing rapid methodological development, provide a strong motivation for STRATOS. This is especially the case when we consider that it is increasingly difficult for any research team to include the full range of expertise that may be required to meet the analytical challenges of observational studies.

With regard to the latter aim, we hope that the aims and scope of the TG will prompt colleagues to consider approaching the TG or the SG with a view to becoming involved in either the development or the evaluation of guidance. In particular, we anticipate that for each TG, the core members will need, from time to time, to call on specific advice from a broader pool of experts. The contributions of members of such a pool will be important but not a burdensome commitment. On the other hand, feedback from applied researchers will help enhancing both the accessibility of the guidance documents to end-users and their practical relevance.

As some of the examples discussed earlier illustrate, a guidance initative such as this—with its focus on advising on analysis, rather than reporting—provides a great opportunity for highlighting key areas where further research is needed. The comprehensive nature of the STRATOS initiative, combined with the structure proposed, has the important additional benefit that guidance in each area will be ultimately both developed and evaluated in relation to guidance in other areas. In consequence, we hope that the initiative will serve as an indicator of when additional methodological developments in a particular area are delivering increasingly marginal benefits, when viewed from the broader research perspective.

Dissemination is the oxygen for impact on practice. Alongside traditional publications, STRATOS will seek to pioneer innovative approaches through the use of the Internet to ensure that the guidelines are responsive to the needs of practitioners at each of the targeted levels. Indeed, the recognition of the need for different guidelines for each of these levels—under a common umbrella—is a key distinguishing component of STRATOS.

Wikipedia advises that STRATOS may refer (among other things) to a municipality in Greece, a type of computer and a high altitude skydiving mission. We hope we have made the case for setting out from our current location, computer in hand, for an exhilarating journey, with the ultimate aim to land in the STRATOSphere of methodological sophistication and practical relevance. Floreat STRATOS!

Members of the STRATOS initiative

The memebers of the STRATOS initiative are Michal Abrahamowicz (Canada), Per Kragh Andersen (Denmark), Doug Altman (UK), Heiko Becher (Germany), Harald Binder (Germany), Maria Blettner

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

(Germany), Danielle Bodicoat (UK), Patrick Bossuyt (the Netherlands), James Carpenter (UK), Raymond Carroll (USA), Harbajan Chadha-Boreham (Switzerland), Gary Collins (UK), Bianca De Stavola (UK), Luc Duchateau (Belgium), Stephen Evans (UK), Laurence Freedman (Israel), Mitchell Gail (USA), Els Goetghebeur (Belgium), Paul Gustafson (Canada), Frank Harrell (USA), Marianne Huebner (USA), Carolin Jenkner (Germany), Victor Kipnis (USA), Helmut Kuechenhoff (Germany), Saskia le Cessie (the Netherlands), Kate Lee (New Zealand), Petra Macaskill (Australia), Erica Moodie (Canada), Neil Pearce (UK), Catherine Quantin (France), Joerg Rahnenfuehrer (Germany), Patrick Royston (UK), Willi Sauerbrei (Germany), Martin Schumacher (Germany), Peggy Sekula (Germany), Len Stefanski (USA), Ewout Steyerberg (the Netherlands), Terry Therneau (USA), Kate Tilling (UK), Werner Vach (Germany), Andrew Vickers (USA), Sholom Wacholder (USA), Ingeborg Waernbaum (Sweden), Ian White (UK) and Mark Woodward (Australia).

Acknowledgements

We thank Saskia Motschall and Clemens Wachter for their assistance in preparing the manuscript and Frank Werner for his technical advice. Michal Abrahamowicz is a James McGill professor of Biostatistics at McGill University.

References

1. Davidian M, Louis TA. Why statistics? *Science* 2012; **336**(6077):12.
2. ICH harmonised tripartite guideline. Statistical principles for clinical trials. International conference on harmonisation E9 expert working group. *Statistics in Medicine* 1999; **18**(15):1905–1942.
3. European Medicines Agency. Guidelines, European Medicines Agency. http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000366.jsp™mid=WC0b01ac0580032ec4.
4. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA: The Journal of the American Medical Association* 1996; **276**(8):637–639.
5. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ (Clinical research ed.)* 2010; **340**:c332.
6. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Standards for reporting of diagnostic accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for reporting of diagnostic accuracy. *Clinical Chemistry* 2003; **49**(1):1–6.
7. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Statistics subcommittee of the NCI-EORTC working group on cancer diagnostics. Reporting recommendations for tumor marker prognostic studies (REMARK). *Journal of the National Cancer Institute* 2005; **97**(16):1180–1184.
8. Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. STROBE initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Epidemiology (Cambridge, Mass.)* 2007; **18**(6):800–804.
9. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, Vet HCW de, Lijmer JG. Standards for reporting of diagnostic accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical Chemistry* 2003; **49**(1):7–18.
10. Vandenbroucke JP, Elm E von, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. STROBE initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Medicine* 2007; **4**(10):e297.
11. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *PLoS Medicine* 2012; **9**(5):e1001216.
12. Ioannidis JP. Why most published research findings are false. *PLoS Medicine* 2005/08/01; **2**(8):e124.
13. Unreliable research: trouble at the lab. *The Economist* 19/10/2013.
14. Turner L, Shamseer I, Altman D, Schulz K, Moher D. Does use of the CONSORT statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *System Review* 2012; **1**:60.
15. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ (Clinical research ed.)* 2010; **340**:c869.
16. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer* 1994; **69**(6):979–985.
17. Rushton L. Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. *Occupational and Environmental Medicine* 2000; **57**(1):1–9.
18. Dufault B, Klar N. The quality of modern cross-sectional ecologic studies: a bibliometric review. *American Journal of Epidemiology* 1101–1107; **174**(10).
19. Niven DJ, Berthiaume LR, Fick GH, Laupland KB. Matched case-control studies: a review of reported statistical methodology. *Clinical Epidemiology* 2012; **4**:99–110.

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

20. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994; **86**(11):829–835.
21. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Jones DR, Heney D, Burchill SA. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *British Journal of Cancer* 2003; **88**(8):1191–1198.
22. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Medicine* 2010; **8**:20.
23. Mallett S, Timmer A, Sauerbrei W, Altman DG. Reporting of prognostic studies of tumour markers: a review of published articles in relation to REMARK guidelines. *British Journal of Cancer* 2010; **102**:173-180.
24. Malats N, Bustos A, Nascimento CM, Fernandez F, Rivas M, Puente D, Kogevinas M. Real FX. P53 as a prognostic marker for bladder cancer: a meta-analysis and review. *Lancet Oncology* 2005/09/01; **6**(9):678–686.
25. Matthews JNS, Altman DG. Statistics notes: interaction 2: compare effect sizes not P values. *BMJ* 1996; **313**:808.
26. Nieuwenhuis S, Forstmann BU, Wagenmakers E. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience* 2011; **14**(9):1105–1107.
27. Sigounas DE, Tatsioni A, Christodoulou DK, Tsianos EV, Ioannidis JPA. New prognostic markers for outcome of acute pancreatitis: overview of reporting in 184 studies. *Pancreas* 2011; **40**(4):522-532.
28. Collins GS, Mallett S, Omar O, Yu L. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Medicine* 2011; **9**:103.
29. van Walraven C, Davis D, Forster AJ, Wells GA. Time-dependent bias was common in survival analyses published in leading clinical journals. *Journal of Clinical Epidemiology* 2004; **57**(7):672–682.
30. Altman DG, Stavola BL de, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. *British Journal of Cancer* 1995; **72**(2):511–518.
31. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007; **335**(7611):136.
32. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ (Clinical research ed.)* 2009; **338**:b2393.
33. Morris TP, White IR, Royston P, Seaman SR, Wood AM. Multiple imputation for an incomplete covariate that is a ratio. *Statistics in Medicine* 2014; **33**(1):88–104.
34. Remontet L, Bossard N, Belot A, Estève J. French network of cancer registries FRANCIM. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine* 2007; **26**(10):2214–2228.
35. Quantin C, Abrahamowicz M, Moreau T, Bartlett G, MacKenzie T, Tazi MA, Lalonde L, Faivre J. Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *American Journal of Epidemiology* 1999; **150**(11):1188–1200.
36. Gray RJ. Flexible methods for analysing survival data using splines. *Journal of the American Statistical Association* 1992; **87**:942–951.
37. Box GE, Draper NR. *Empirical Model-building and Response Surfaces*. Wiley: New York, 1987.
38. Kannel WB, Cupples LA, D'Agostino RB. Sudden death risk in overt coronary heart disease: the Framingham Study. *American Heart Journal* 1987; **113**(3):799–804.
39. Sytkowski PA, Kannel WB, D'Agostino RB. Changes in risk factors and the decline in mortality from cardiovascular disease. The Framingham Heart Study. *The New England Journal of Medicine* 1990; **322**(23):1635–1641.
40. Benjamin EJ, Levy D, Vaziri SM, D'Agostino RB, Belanger AJ, Wolf PA. Independent risk factors for atrial fibrillation in a population-based cohort. The Framingham Heart Study. *JAMA: The Journal of the American Medical Association* 1994; **271**(11):840–844.
41. Gagnier JJ, Moher D, Boon H, Beyene J, Bombardier C. Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. *BMC Medical Research Methodology* 2012; **12**(1):111.
42. Fitzmaurice GM, Kenward MG, Molenberghs G, Tsiatis AA, Verbeke G. *Handbook of Missing Data*. CRC Press: New York, 2014.
43. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley and Sons Ltd: Chichester, 2002.
44. National Research Council (U.S.) *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press: Washington D.C, 2010.
45. Carpenter JR, Kenward MG, Goldstein H. In *Statistical Modelling of Partially Observed Data Using Multiple Imputation: Principles and Practice*, Tu Y, Greenwood D (eds). Springer: New York, 2012; 15–23.
46. Carpenter JR, Kenward MG. *Multiple Imputation and its Application*. John Wiley & Sons Ltd: Chichester, 2013.
47. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer: New York, 2001.
48. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine* 2007; **26**:5512–5528.
49. Miller A. *Subset Selection in Regression*. Taylor & Francis: Boca Raton, Florida, 2002.
50. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology (Cambridge, Mass.)* 1995; **6**(4):450–454.
51. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametic modelling. *Applied Statistic* 1994; **43**(3):429–467.
52. Royston P, Sauerbrei W. *Multivariable Model-building. A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Continuous Variables*. Wiley: Chichester, 2008.
53. Boer C de. *A Practical Guide to Splines* revised edn. Springer: New York, 2001.
54. Wood S. *Generalized Additive Models*. Chapman & Hall/CRC: New York, 2006.
55. Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman & Hall/CRC: New York, 1990.
56. Abrahamowicz M, Du Berger R, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. *American Journal of Epidemiology* 1997; **145**(8):714–729.

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

57. Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine* 2013; **32**(13): 2262–2277.
58. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine (Stat Med)* 2007; **26**(2):392–408.
59. van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Medicine* 2005; **2**(10):e267.
60. Chatfield C. Confessions of a pragmatic statistician. *Journal of the Royal Statistical Society. Series D (The Statistician)* 2002; **51**(Part 1):1–20.
61. Cox D, Donnelly C. Preliminary analysis. In *Principles of Applied Statistics*. Cambridge University Press: Cambridge, 2011.
62. George DS, Shah E. Data dredging, bias, or confounding. *BMJ* 2002; **325**(7378):1437–1438.
63. Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics* 2009; **3**(4):1309–1334.
64. Vach W. Transformation of covariates. In *Regression Models as a Tool in Medical Research*. Taylor & Francis Group: Boca Raton, FL, USA, 2013; 264–273.
65. Buonaccorsi JP. *Measurement Error: Models, Methods, and Applications*. Chapman & Hall/CRC: Boca Raton, Florida, 2010.
66. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective,* Second Edition. Chapman and Hall/CRC Press: Boca Raton, Florida, 2006.
67. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall/CRC: Boca Raton, Florida, 2004.
68. Hill AB. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* 1965; **58**:295–300.
69. Pearce N. Epidemiology in a changing world: variation, causation and ubiquitous risk factors. *International Journal of Epidemiology* 2011; **40**:503–512.
70. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology* 3rd ed. Lippincott Williams & Wilkins: Philadelphia, 2008.
71. Pepe M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
72. Zhou X, Obuchowski N, McClish D. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons Ltd: New York, 2002.
73. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006; **332**:1089–1092.
74. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. *Journal of Clinical Epidemiology* 2010; **63**(8):883–891.
75. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**(1):128–138.
76. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B[§]. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *European Journal of Clinical Investigation* 2012; **42**(2): 216–228.
77. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Statistics in Medicine* 2013; **32**(9):1467–1482.
78. Hunink MGM[§], Glasziou PP, Siegel JE, Weeks JC, Pliskin JS, Elstein AS, Weinstein MC. *Decision Making in Health and Medicine. Integrating Evidence and Values*. Cambridge University Press: Cambridge, UK, 2001.
79. Daniel RM, Cousens SN, Stavola BL de, Kenward MG, Sterne JAC. Methods for dealing with time-dependent confounding. *Statistics in Medicine* 2013; **32**(9):1584–1618.
80. Pearl J[§]. Causal diagrams for empirical research. (With discussion). *Biometrika* 1995; **82**(4):669–710.
81. Fischer-Lapp K, Goetghebeur E. Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials* 1999; **20**(6):531–546.
82. Gagne JJ, Polinski JM, Avorn J, Glynn RJ, Seeger JD. Standards for causal inference methods in analyses of data from observational and experimental studies in patient-centered outcomes research, 2012. *For: Patient-Centered Outcome Research Institute Methodology Committee*.
83. Hernan[§] MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**(5):561–570.
84. Hernan MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, Manson JE, Robins JM. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008; **19**(6):766–779.
85. Moodie EE, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. *Biometrics* 2007; **63**(2): 447–455.
86. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1):41–55.
87. Sterne JTK. G-estimation of causal effects, allowing for time-varying confounding. *The Stata Journal* 2002; **2**(2):164–182.
88. Valeri L, VanderWeele T. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods* 2003; **18**:164–182.
89. Vansteelandt S, Bowden J, Babanezhad MGE. On instrumental variables estimation of causal odds ratios. *Statistical Science* 2011; **26**(3).

[§] *Correction added on 15 August 2014, after first online publication: authorship of reference corrected.*

© 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.