Toward a Comprehensive, Integrated Framework for Advanced Statistical Analyses of Observational Studies (19w5198)

Michal Abrahamowicz (McGill University, Montreal, Canada, Contact Organizer),
Marianne Huebner (Michigan State University, East Lansing, MI, USA),
Willi Sauerbrei (Medical Center, University of Freiburg, Freiburg, Germany)

June 2-7, 2019

1 The STRATOS initiative — motivation, mission, structure and main aims

The overall objective of the 19w5m198 BIRS Workshop was to further boost and consolidate the research activities of the international initiative for STRengthening Analytical Thinking for Observational Studies (STRATOS) (www.stratos-initiative.org). Therefore, to facilitate understanding of the rather unique character and goals of the Workshop, we begin this Report with a short overview of the STRATOS initiative, including its *raison d'être*, overall mission, internal structure and main objectives.

Lack of statistical expertise is now recognized as a significant brake on scientific progress across a wide range of empirical sciences. For instance, an influential article in The Economist (19/10/2013), 'Unreliable research: Trouble at the lab.' commented: *"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."*

According to the general paradigm of modern sciences, statistical analysis methods are key to translate raw empirical data into new insights in, and deeper understanding of, complex processes affecting human health, the economy, environment, and many other phenomena studied in different branches of science. Yet, the complexity of such processes, and of the observable data they generate, create numerous analytical challenges. In the 21st century, parallel progress in the theory of mathematical statistics and computational resources and technology led to dynamic developments in statistical methodology, resulting in a large number of increasingly complex, ever more flexible, statistical techniques and models that allow researchers to account for several complexities frequently encountered in analyzing real-life data. Unfortunately, many of these important developments are ignored in every-day practice of data analysis, including analyses reported in influential publications in high-impact medical or social sciences journals. Consequently, the design and analysis of recent, often complex and costly, observational studies of human health and welfare often exhibit serious weaknesses. This leads to misleading inferences, which may, in turn, adversely affect the effectiveness or safety of different treatments, social or economic policy programs etc.

Formulating and overcoming these formidable challenges requires a well structured, highly interactive collaboration between a large, international group of statistical experts, whose research combines develop-

ment of new methodology with collaborative research on real-life applications and whose joint, complementary expertise covers different sub-areas of statistical research. Indeed, such need arises due to a combination of (i) ever increasing complexity and variety of analytical challenges encountered in the majority of important observational studies, together with (ii) the increasing trend for narrow specialization, necessary to achieve cutting-edge novel developments in modern statistics. Together, these two trends imply that no single university-based (bio-)statistics department and no team of (bio-)statisticians (working in even the largest private or public research institutes or, e.g., pharmaceutical companies) is able to ensure state-of-the-art expertise regarding even a reasonable fraction of analytical challenges encountered in real-life applications. At present, particular challenges are being addressed by leading authorities in different areas of statistical research, but little effort is invested in combining the results of these separate developments and ensuring their material impact on the practice of data analysis. This situation provided the motivation for, and the driving vision behind, the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative. The STRATOS Initiative was launched in 2013 at the 34th conference of the International Society of Clinical Biostatistics (ISCB). It is connected to this society and had dedicated invited sessions or mini-symposia at almost all ISCB annual meetings in 2013–19 (with the only exception in 2017). After an Invited STRATOS session at the bi-annual meeting of the International Biometric Society (IBS) in 2016, the STRATOS initiative was invited to publish a series of short papers in the Biometric Bulletin, the official newsletter of IBS.

The initiative brings together leading, internationally recognized, methodological experts in several areas of statistics essential for the analyses of observational studies, who — at present — are grouped in nine Topic Groups (TGs), each focusing on a different, highly relevant area of statistical research (see Table 1 below). The experts' joint, largely complementary, knowledge allows the initiative to address complex analytical challenges in the design and analysis of observational studies, by developing, validating and comparing state-of-the-art methods for various topics. To increase the impact of our endeavors on empirical research, individual topic groups are working to summarize our findings by developing practical guidance regarding 'best practice' to address a particular set of analytical challenges (e.g. handling of measurement error or dealing with right-censored time-to-event data). The guidance covers such practical issues as e.g. the awareness of potential pitfalls due to inappropriate use of 'conventional' methods, the choice of appropriate, validated analytical methods able to overcome specific challenges, and software that can be used to implement these advanced methods.

We are entering the era of 'big data' with automated collection of very large amounts of data and the paradigm of empirical sciences shifting toward 'data science'. However, 'big data' will not help answering such essential prognostic or etiology questions if researchers use designs and statistical methods which are unsuitable, e.g. by being unable to account for the complexity of the underlying dynamic multi-factorial processes. Therefore, it is of central importance to gain knowledge about advantages and disadvantages of alternative statistical approaches, and their dependence on the data structure. Equally important is to develop, validate and explain to end-users state-of-the-art methods that can address frequent limitations (e.g. missing data, measurement errors, unmeasured confounding) and complexities of the data (e.g. non-linear relationships of continuous variable with the outcome, time-varying effects, mediation).

One of the fundamental objectives of the STRATOS approach is to develop guidance for data analysts and researchers with different levels of statistical training, skills and experience. Specifically, we have identified three levels of statistical knowledge, each of which would require a somewhat different type of targeted guidance, and we have outlined the main criteria to be used when developing guidance aimed at the analysts at each level. Initially, we are working to derive guidance for experienced statisticians ('level 2'), which requires work on state-of-the-art methodology for each specific topic group. For each topic considered (see next section) several analytical strategies have been proposed in the statistical literature, but knowledge about their properties and relative strengths and weaknesses is often insufficient, as meaningful comparisons are rare and evidence-based guidance are lacking. For more details see Sauerbrei *et al*, (*Statistics in Medicine*, 2014, 33, 5413–5432), the STRATOS website http://www.stratos-initiative.org/ and short articles in the *Biometric Bulletin* (available on the STRATOS website).

Table 1. STRATOS Topic Groups and co-Chairs

	Topic Groups	Co-Chairs
TG1	Missing Data	James Carpenter, Kate Lee
TG2	Selection of variables and functional forms in multivariable analysis	Georg Heinze*, Aris Perperoglou*, Willi Sauerbrei*
TG3	Initial Data Analysis	Marianne Huebner*, Saskia le Cessie*, Werner Vach*
TG4	Measurement Error and Misclassification	Laurence Freedman*, Victor Kipnis*
TG5	Study Design	Suzanne Cadarette*, Mitchell Gail*
TG6	Evaluating Diagnostic Tests and Prediction Models	Ewout Steyerberg*, Ben van Calster*
TG7	Causal Inference	Els Goetghebeur*, Ingeborg Waernbaum
TG8	Survival Analysis	Michal Abrahamowicz*, Per Kragh Andersen*, Terry Therneau*
TG9	High dimensional Data Analysis	Lisa McShane*, Joerg Rahnenfuehrer*

* Indicates Participants of the BIRS 19w5198 Workshop.

2   Objectives of the Workshop

Overall Objectives:
The two, closely inter-related, overarching objectives of the 19w5m198 BIRS Workshop were to (i) further boost and consolidate the research activities of the STRATOS Topic Groups, and to (ii) identify and initiate new interdisciplinary collaborations between experts in different areas of statistical methodology, regrouped in different Topic Groups. (Both goals built on the earlier success of the 2016 BIRS Workshop 16w5091 that provided the first opportunity for a large group of 38 STRATOS members, and seven research trainees, to meet in-person, exchange ideas, jump-start several joint articles writing and develop an operational plan for further development and internal homogenization of the scientific and knowledge translation endeavors.)

In particular, during the 2019 Workshop, regarding objective (i), individual Topic Groups proposed recommendations to address the main analytical challenges within their area of expertise. Then, regarding objective (ii), based on these recommendations, we have started designing comprehensive strategies to deal *simultaneously* with several problems likely to be encountered in real-life empirical studies. By providing a unique opportunity for in-person discussions between experts from 14 countries on 3 continents, the Workshop was essential to initiate and largely facilitate such inter-disciplinary discussions and developments.

Specific Goals:
The 4 inter-related specific goals of the workshop were:

1. To provide an overview of the methods, related to the area of expertise of individual Topic Groups, applied in the current empirical studies and identify the priorities for improving the methodological quality of such studies;

2. To identify methodological challenges, within the area of expertise of Topic Group, that require further validation or comparison of new or existing methods, and outline the analytical work or simulation studies necessary to provide reliable evidence supporting specific approaches and demonstrating the limitations of other methods, in the spirit of 'neutral comparison studies' (see the following article https://doi.org/10.1002/bimj.201700129);

3. To identify analytical challenges at the cross-roads of the interests of various Topic Groups, and to initiate new between-Groups collaborations on the new developments necessary to address such complex challenges in order to approximate the complexity of large real-life empirical observational studies, as well as comprehensive simulations necessary to assess and validate the resulting new methods;

4. To develop the uniform format, criteria, and general content for the integrated STRATOS-based guidance documents.

3 Overview of the Workshop participants and activities

3.1 Participants

Despite a few late cancellations we had no problems to fill all the 42 available places at BIRS. With 42 participants from 14 countries, spread over 3 continents, the meeting had truly a global character. Five research trainees, including 4 PhD students and 1 post-doc fellow, attended, in addition to one Research Associate. Among the 42 participants, 15 (36%) were women. Because of private reasons, the two co-chairs of Topic Group 1 (Missing data) could not attend and, therefore, this topic played only a marginal role in the 2019 Workshop.

List of participants and their affiliations (* indicates trainees):

1. Abrahamowicz, Michal; McGill University, CANADA
2. Ambrogi, Federico; Università degli Studi di Milano, ITALY
3. Andersen, Per Kragh; University of Copenhagen, DENMARK
4. Baillie, Mark; Novartis, SWITZERLAND
5. Beauchamp, Marie-Eve* (F); Research Institute, McGill University Health Centre, CANADA
6. Becher, Heiko; University Medical Center Hamburg-Eppendorf, GERMANY
7. Benner, Axe; German Cancer Research Center, GERMANY
8. Cadarette, Suzanne (F); University of Toronto, CANADA
9. Carroll, Orlagh* (F); London School of Hygiene and Tropical Medicine, UK
10. De Bin, Riccardo; University of Oslo, NORWAY
11. Deffner, Veronika* (F); Ludwig-Maximilians-Universität München, GERMANY
12. Didelez, Vanessa (F); Leibniz Institute for Prevention Research and Epidemiology - BIPS, GERMANY
13. Freedman, Laurence; Gertner Institute for Epidemiology, ISRAEL
14. Gail, Mitchell; National Institutes of Health, USA
15. Goetghebeur, Els (F); University of Ghent, BELGIUM
16. Gustafson, Paul; University of British Columbia, CANADA
17. Harrell, Frank; Vanderbilt University, USA
18. Heinze, Georg; Medical University of Vienna, AUSTRIA
19. Huebner, Marianne (F); Michigan State University, USA
20. Joly, Pierre; Université de Bordeaux, FRANCE
21. Kammer, Michael*; Medical University of Vienna, AUSTRIA
22. Keogh, Ruth (F); London School of Hygiene and Tropical Medicine, UK
23. Kipnis, Victor; National Cancer Institute, USA
24. le Cessie, Saskia (F); Leiden University Medical Center, The NETHERLANDS
25. McShane, Lisa (F); U.S. National Cancer Institute, USA
26. Pang, Menglan* (F); McGill University, CANADA
27. Perperoglou, Aris; Global Medicines Development, R&D, UK
28. Pohar-Perme, Maja (F); University of Ljubljana, SLOVENIA
29. Rahnenführer, Jörg; TU Dortmund University, GERMANY
30. Rauch, Geraldine (F); Charite - Universitatsmedizin Berlin, GERMANY
31. Sauerbrei, Willi; Medical Center - University of Freiburg, GERMANY
32. Schmid, Matthias; Universität Bonn, GERMANY

33. Schmidt, Carsten; University Medicine of Greifswald, GERMANY

34. Shaw, Pamela (F); University of Pennsylvania, USA

35. Steyerberg, Ewout; Erasmus MC, The NETHERLANDS

36. Therneau, Terry; Mayo Clinic, USA

37. Vach, Werner; University Hospital Basel, SWITZERLAND

38. Van Calster, Ben; KU Leuven, BELGIUM

39. Wallace, Michael; University of Waterloo, CANADA

40. Wallisch, Christine (F); Medical University of Vienna, AUSTRIA

41. van Houwelingen, Hans; Leiden University Medical Center, The NETHERLANDS

42. van Smeden, Maarten; Leiden University Medical Center, The NETHERLANDS

3.2   Overview of the presentations and discussions

On Monday morning, we started with a summary of the recent developments of the STRATOS Initiative. During the week, we had several keynote talks by internationally recognized authorities. Frank Harrell (TG2) and Per Kragh Andersen (TG8) spoke about issues which were not directly related to the current work of any topic group, but which could be relevant for the future of STRATOS. Mitchell Gail, one of the chairs of the Design group (TG5) spoke about potential collaborations of the Design group with several other TGs. This talk was highly relevant for jump-starting several new inter-TG collaborations, one of the main aims of the Workshop. The Measurement Errors topic group (TG4) had recently made much progress and we decided dedicating a 90-minute plenary session to presentations from TG4. In four talks, Laurence Freedman, Victor Kipnis, Ruth Keogh and Pam Shaw summarized the results from the last year and presented directions for new research. During the first two days of the Workshop, each of 7 other TGs represented at the Workshop summarized their progress and plans for future in 45-minute plenary presentations. TG presentations also included thoughts about potential collaborations with other TGs, which created the basis for further discussions that initiated several joint efforts between topic groups (more details below in section 5 on new inter-TG collaborations). We had also more detailed presentations from two panels (talks from Lisa McShane and Saskia le Cessie from the Simulation Panel and Mark Baillie presented ideas of the new Visualization Panel).

Progress of the Glossary Panel was outlined in video presentation from Martin Boeker (Freiburg, Germany), an expert in medical informatics. In addition, we had shorter presentations from the Data Set and the Knowledge Translation panels.

To allow more detailed discussions and time for outlining the content of future manuscripts within TGs, as well as discussions between members of different TGs and/or panels, we also decided dedicate a substantial amount of time to separate meetings in smaller groups. These discussions often continued in the evenings. Results of these small groups discussions were presented and discussed on Friday morning, in two general sessions that summarized the conclusions of the BIRS meeting and provided an outlook for main activities planned for the next two years.

4   Summary of progress and plans by Topic Groups

4.1   Overview

Over the course of several plenary meetings, spread across the 5 days of the BIRS Workshop, each of the eight Topic Groups (TGs) represented at the Workshop, and selected cross-cutting Panels, presented their recent progress in the research activities leading toward the ultimate goal of developing the guidance regarding the choice, evaluation and implementation of state-of-the-art statistical methods relevant for addressing analytical challenges frequently encountered in their specific areas of interest. (Please see Table 1 above for the overview of the topics being the main focus of different TGs). (The exception was TG1 'Missing data', that – because of a number of conflicting responsibilities by its members – was represented only by a single member, and a PhD student who gave a talk on missing data). These presentations and updates helped inform

and learn not only about the successes obtained so far, but also about the outstanding challenges, and the potential solutions that need to be evaluated to provide rigorous evidence regarding which methods may work, and under what conditions. All members benefitted immensely from face-to-face meetings, made possible by the BIRS grant. This was vital for establishing new long-term research collaborations, both within individual TGs, and – especially – between members of different TGs who have identified a number of the complex analytical challenges that will require combining the expertise developed within particular groups (outlined in more detail below). Such in-person meetings will also largely facilitate future continuation of these new collaborations, which – for the next two years - will rely almost exclusively on email and teleconferences.

4.2 Progress and Future Plans of individual Topic Groups

TG2: Selection of variables and functional forms in multivariable analysis

Discussions and interactions

All three co-chairs and six members of the topic group were present, as well as 2 trainees (1 post-doctoral fellow and 1 PhD student). This excellent representation led to productive conversations and initiation of several future projects and manuscripts (see below). Frank Harrell delivered a plenary talk on *"Controversies in Predictive Modeling, Machine Learning, and Validation."*

Research: current

The meeting allowed the plans of the group to be consolidated as follows:

- State-of-the-art paper on variable selection has been completed. It was decided to upload a copy at arxiv.org and the paper was submitted for publication in *Diagnostic and Prognostic Research.*

- Geraldine Rauch presented the current status of the manuscript on *'Systematic review of statistical series in medical journals'*. Following the BIRS meeting, a protocol has been finalized in a collaboration between Geraldine Rauch and Georg Heinze, and associates and was submitted as publication to a medical journal.

- Marie-Eve Beauchamp and Michal Abrahamowicz presented the current status of the collaborative paper, involving several TG2 members, on *'Systematic review of methods used for modeling of continuous variables and variables selection in medical and epidemiological journals'*. The protocol and evaluation criteria have been finalized, and currently data are independently extracted by two reviewers, one in Oxford, UK and one at McGill, in Canada.

Research: future

- A Topic Group meeting, to start a new collaboration on systematic evaluation and comparison of various spline-based approaches for selecting functional forms, will take place in Bonn, Germany on November 21-22, 2019. Presentations at the STRATOS-oriented Invited Sessions at ISCB 2020 (Krakow, Poland) and IBC 2020 (Seoul, South Korea) by TG2 members are planned.

- The group has identified a set of seven important open problems in their "State-of-the-art" paper. Members are expected to express their interested in leading relevant work. Further discussions will take place in the meeting in Bonn.

- Variable selection: Georg Heinze would lead on a new TG2 paper on variable selection, that will partly build on his previous work. Protocol for a simulation study is currently in preparation.

- Splines: Aris Perperoglou to lead a follow-up paper, building on the published TG2 paper on the review of spline packages in R.

- Procedures for variable & functional form selection: Willi Sauerbrei will take the lead for the new collaborative TG2 paper that will address both challenges simultaneously.

- Post-selection shrinkage: There is some preliminary work by Michael Kammer (Georg Heinze's student). Results will be included in Michael Kammer's PhD thesis.

- 'What is an appropriate stability measure?' Daniela Dunkler has submitted a session proposal for CEN2020 conference in Berlin on 'Variable selection and model instability', including contributions from Ewout Steyerberg (TG6), and Georg Heinze (TG2), Riccardo de Bin (TG9), Christine Wallisch.

- A new manuscript entitled 'When to prespecify a model and in which cases to consider variable selection?' was proposed by Frank Harrell.

TG3: Initial data analysis

Discussions and interactions

Four members, including the co-chairs, of the topic group were present, and one member joined via conference call for a group meeting. All were engaged in discussions and plans with other TG's (TG2, TG4, TG5, and TG9) as initial data analysis (IDA) is relevant for most other TGs, as outlined below, in the section on Between Groups Collaborations.

Research: current

- The manuscript on a systematic literature review on IDA reporting was discussed and revised. It has been approved by the STRATOS Publication Panel and has been submitted to *BC Medical Res. Meth.*

- Initial Outcome Data Analysis (IODA). Werner Vach described a randomized trial with several outcome variables and analyses that could be termed "Initial Outcome Data Analysis" (IODA). We discussed what might be considered IODA and other examples were shared.

- Saskia le Cessie developed a module with videos on IDA to be included in a MOOC by the University of Leiden. This course will run on the *Coursera* platform.

Research: future

As discussed below, several projects have been identified with other TGs. In addition:

- Carsten Schmidt obtained permission to use data from the large cohort Study of Health in Pommerania (SHIP). This will lead to a manuscript illustrating the IDA concepts and will serve as an example of reporting. In addition, recommendations for automatization of IDA can be developed.

- A conceptual paper on IDA in the context of multivariable regression analyses will be led by Werner Vach.

- New project on IDA for longitudinal studies will be led by Lara Lusa.

- TG3 will work with Mark Baillie and the STRATOS Data Visualization Panel to develop examples and recommendations for enhancing visual displays of the IDA results.

TG4: Measurement errors and misclassification

Discussions and interactions

Seven members, including the co-chairs, of the topic group were present, as well as 2 trainees (1 postdoctoral fellow and 1 PhD student). Current Research and progress was presented jointly in a plenary session. Veronika Deffner will be setting up a website for TG4, in line with the other TGs. Explorations for future joint papers with TG3, TG5, and TG8 took place, as outlined below.

Research: current

- A paper on a literature survey of the use of methods to adjust for measurement error in four areas of epidemiology was published in *Annals of Epidemiology*.

- A paper in *Statistics in Medicine* is under revision, after a favorable first review, which will be a guidance paper (in two parts) on measurement error and misclassification of variables in epidemiology, aimed at biostatisticians (level 2-3 paper).

- A paper on the general problem of measurement errors in epidemiology will be submitted to the journal *Significance*. This paper is written for the researchers without formal statistical background (level 1 paper).

- TG4 is conducting research into methods of handling the Berkson measurement error that occurs, for example, when prediction equation variables, or aggregate variables, are used in regression analyses.

Research: future

As discussed below, several future interdisciplinary collaborative projects have been identified with different other TGs. In addition, the following topics will be addressed in future TG4 research:

- Handling error-prone variables that are truly continuous but have been categorized for data analyses;

- A case-study paper on how to account and correct for measurement errors in practice (with real-life examples);

- A guidance paper for nutritional epidemiologists on handling measurement errors in nutritional epidemiology.

TG5: Study design

Discussions and interactions

Both co-chairs were in attendance, and they concentrated on discussions about collaborations with other Topic Groups. This focus was in agreement with the Workshop Organizers who had invited Mitch to give a Plenary talk designed to promote collaborations between TG5 and other TGs. Mitch and Suzanne had discussions with members of several TGs about joint projects, resulting in plans for projects with TGs 3 and 4. For more details see section 5 below. Discussions about approaches to increase the TG5 membership, and the related criteria, played also an important role.

Research: current

The first TG5 paper has been recently accepted for publication:

1. Final corrections of the revised version of an overview paper on study design were discussed among the TG5 members. The revised paper was resubmitted and will be published in *BMJ Open*:
Gail MH, Altman DG, Cadarette SM, Collins G, Evans SJW, Sekula P, Williamson E, Woodward M, for the STRATOS initiative (STRengthening Analytical Thinking for Observational Studies). Design Choices for Observational Studies of the Effect of Exposure on Disease Incidence. *BMJ Open* (In Press).

2. In addition, the short paper introducing TG5 to the members of the International Biometric Society was completed and published in *Biometric Bulletin*:
Gail M, Cadarette SM on behalf of TG5. STRengthening Analytical Thinking for Observational Studies (STRATOS): Introducing the Study Design Topic Group (TG5). *Biometric Bulletin* 2019; 36(2): 12-13.

Research: future

Interactions, during the BIRS Workshop, and plans for future collaborative projects with STRATOS TG3 and TG4 are described below, in the section 5 on Between-Groups Collaborations. In addition to the joint projects with other TGs, members of TG5 started work on two new papers, with the following preliminary titles:

- *'How to choose a design to study associations between prescription medications and risk of osteo-porotic fracture.'* (This paper will focus on controlling for 'confounding by indication', which represents a paramount challenge for design of observational studies of medication effects. It will target a clinical audience).

- *'Design issues in prognostic studies.'*

TG6: Evaluating diagnostic tests and prediction models

Discussions and interactions

Three members attended, including the two co-chairs. Several topics were discussed, including some that had active involvement of the TG6 members, while not published under the umbrella of STRATOS. Potential new TG6 members were approached at the BIRS meeting.

Research: current

Three TG6 papers, being prepared for publication, have been discussed:

- "Performance assessment of survival models: a review". D McLernon and various STRATOS members: B van Calster, M van Smeden, E Steyerberg, T Therneau (TG8).

- "Myths about risk thresholds in prediction models". Laure Wynants, Maarten van Smeden, David J. McLernon, Dirk Timmerman, Ewout W. Steyerberg, Ben Van Calster. (This paper is now available at https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1425-3).

- "Calibration: the Achilles heel of predictive analytics". Ben Van Calster; David J McLernon; Maarten van Smeden; Laure Wynants; Ewout W Steyerberg. *BMC Medicine*.

Research: future

In collaboration with TG8, two future review papers on: 1/ performance assessment of competing risk and 2/ dynamic risk prediction models are planned.

TG7: Causal inference

Discussions and interactions

Four members were in attendance, including one of the two co-chairs. One of the main focus of the discussions were the Knowledge Translation (KT) activities of the TG7, in addition to initiation of collaborations with other STRATOS TG's (outlined below). TG7 created several short courses taught in an international setting. The output generated so far, consists of a TG7 website, www.ofcaus.org, that contains educational material on the concepts and methods of causal inference (using the potential outcomes framework). The website gives free access to presentation slides, practical assignments with questions and answers, data analysis problems with solutions and software code in R, SAS and stata, and source code for a 'simulation learner.'

Research: current

Saskia le Cessie gave a plenary talk on implementation, code, guidance, as well as potential and new statistical insights derived from the simulation learner. This is a large-scale, real-world based, simulated experiment in R that not only generates the 'observed data' but, for every subject, simulates also a set of possible alternative exposures with their potential outcomes. It allows to illustrate concepts, target estimands and statistical techniques with their practical properties for causal effect estimation in a unique way that has proven to be a great learning tool.

Research: future

Future joint collaborative projects were identified with TG8 and TG4, respectively, as described below.

TG8: Survival Analysis

Discussions and interactions

Six of the nine current TG8 members participated in the BIRS Workshop, including all three co-chairs, in addition to two trainees (1 PhD student and one Research Associate). Per Kragh Andersen gave a Plenary talk for all meeting participants. His talk focused on the methodological issues, as well as on their practical advantages in real-life applications, related to use of pseudo-values in survival analysis. Terry Therneau gave a plenary talk in which he summarized the recent TG8 progress and plans for future. Excellent representation of TG8 members (from 6 different countries on two continents) at the BIRS Workshop largely facilitated in-person exchanges and led to productive discussions. An after-hours TG8 meeting was held during the BIRS Workshop, in addition to regular program, to discuss both the ongoing research and plans for future TG8 activities. In addition, as outlined below, two small-group meetings with other STRATOS TG's were held to initiate and outline future collaborative projects involving TG8 members.

Research: current

The main focus of the within-group discussions was on the final steps of the revisions of the TG8 review/tutorial paper that provides guidance on the fundamental issues and analytical challenges typically encountered in survival analysis (or 'time-to-event' analyses), in the classic setting of single-event studies (e.g. of all-cause mortality). This paper, co-authored by all nine TG8 members, is oriented toward researchers with solid general statistical training who, however, do not have in-depth expertise in survival analysis (considered 'level 2' paper according to the STRATOS criteria). The BIRS meeting gave an excellent opportunity to exchange ideas regarding some specific details of the content of this first TG8 paper, mathematical notation and – above all – the final choice, and presentation, of the real-life examples that will clearly illustrate different methodological challenges, and the methods recommended to avoid the potential pitfalls. During the BIRS meeting, a consensus was reached regarding these outstanding issues, and the manuscript is currently undergoing the final revision and will be submitted for publication in Statistics in Medicine in December 2019.

Research: future

As a logical extension of the discussions about the first TG8 paper, the group members present in Banff had also an opportunity to consider different options for how to address several further, more 'advanced' analytical challenges. These discussions, together with earlier discussions among six TG8 members present at the IBC 2018 meeting in Barcelona, identified several issues that are also highly relevant for time-to-event analyses of observational data, but – largely because of the space limitations – could not be adequately covered in the initial TG8 'tutorial', outlined above. These future activities will aim at providing further guidance regarding such, more complex topics as, for example (in random order): (i) competing risks and modeling of multi-state transitions, (ii) alternatives to the ubiquitous Cox's proportional hazards model (Accelerated Failure Time and Additive Hazards models), (iii) interval-censored data, (iv) modeling of time-varying covariates (lagged or cumulative effects, impact of sparse, irregular measurements), and (v) net survival methods to correct for unknown causes of death. It was generally agreed upon that different (though partly overlapping) subgroups of the TG8 members may collaborate more closely on each specific topic, and preliminary ideas regarding who may take a lead on particular future manuscripts were also considered. It was decided to start developing concrete plans for the future papers, to deal with some of the above issues, in the Winter/Spring of 2020. These plans will be then ironed out during an in-person meeting of the most TG8 members during or after the 41st International Society for Clinical Biostatistics (ISCB) meeting in August 2020 (several TG8 members are either Invited Speakers or members of the Scientific Program Committee for the ISCB 2020 conference).

In parallel, TG8 started discussions about new inter-group collaborative projects with the STRATOS TG7 (Causal Inference), TG5 (Predictive Models) and TG4 (Measurement Errors), as outlined below.

TG9: High dimensional data

Discussions and interactions

Six members, including the two co-chairs, were in attendance. and an overview paper for high dimensional data was discussed. Key issues were presented by Lisa McShane in a Plenary presentation to the full STRATOS group at BIRS. Details of the development, and of the future content, of the dedicated TG9 website, to be linked to the STRATOS main website, was agreed upon.

Research: current

- The overview paper for the high dimensional data group was further refined. Data analysis pipelines for a wide variety of omics data with detailed worked examples would serve as guidance to practitioners. While this is geared toward "omics" data, other types of high-dimensional data (e.g., medical records databases with large numbers of variables per individual) are potential examples. Real-life examples of how and why standard statistical approaches can "break down" in high dimensional data settings will be important to include.

- A paper on the topic of simulation of high dimensional data, in collaboration with the STRATOS Simulation Panel, is planned to address those aspects and challenges of designing and carrying out complex simulations that are particular for high dimensional data.

Research: future

- Needs already identified include better guidance and methods for design of studies involving high-dimensional data and methods for simulation of high-dimensional data. Better methods to simulate realistic high-dimensional data will be critical for assessing performance of new analysis methods and for comparison of existing methods and are therefore essential to many aspects of TG9's work.

- Design of studies is particularly challenging given the often broad aims of these studies, e.g., identification of clusters, and likely complex distributions of these data.

- Several members of TG9 will be actively involved in a STRATOS-wide project to evaluate machine learning methods and better understand their strengths and weaknesses compared to more conventional statistical methods.

5 New Collaborative research projects involving Joint Efforts of different Topic Groups

In addition to enhancing the aforementioned research activities of individual Topic Groups (TG), the BIRS Workshop created a unique opportunity to identify, stimulate and foster new 'inter-disciplinary' collaborations between members of different TG's. Indeed, as reflected in the Workshop's title, one of our major focuses was on *integrating* recent developments made in different areas of statistical research of major relevance for the analyses of complex observational studies. This aspect of the Workshop drew directly on the rich range of the expertise of the members of separate TG's, who – in most cases – had no earlier opportunities to work together on common collaborative projects. Figure 1 below summarizes different new between-TG's collaborative links created during the BIRS Workshop. Later, in this section, we outline the scope of 10 new between-groups collaborations, and provide a brief summary of respective discussions.

TG2 and TG3. Our meeting included 4 TG3 members (Mark Baillie, Marianne Huebner, Werner Vach, and Carsten Oliver Schmidt) and 5 TG2 members (Frank Harrell, Georg Heinze, Aris Perperoglou, Matthias Schmid, and Willi Sauerbrei). Two joint TG2-TG3 papers have been proposed: one more theoretical under the lead of Werner Vach (TG3) and one applications-oriented to be co-led by Marianne Huebner (TG3) and Georg Heinze (TG2). For the latter, five initial data analysis (IDA) topics are to be included (distribution of a single variable, associations between variables, missing data, measurement error, levels of measurements) and the paper will illustrate how these activities addressing these might impact the set-up of the multivariable regression analysis. Three data sets were proposed by MH, FH, and COS. Work on illustrating the IDA topics with these data sets will be divided between the participants. Georg Heinze and Marianne Huebner will
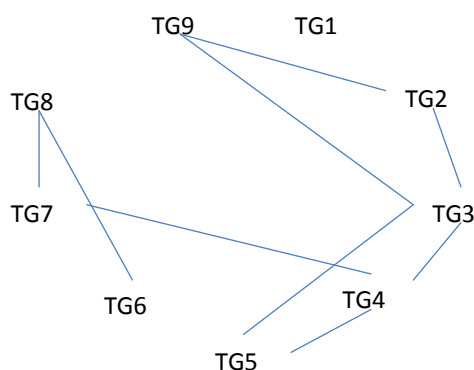
Figure 1: New between-TG's collaborative links

co-present the results at the Invited STRATOS Session at the 41st ISCB meeting in Poland, in August 2020.

TG2 and TG9 and Simulation panel. Issues of model-based simulation and data-based (plasmode) simulation were discussed. Michael Kammer (PhD student of Georg Heinze, TG2) presented his R package 'SimulatoR' which implements the ideas of e.g. [Binder et al 2013, Stat Med], and [Royston and Sauerbrei, 2008 monograph, Wiley], for model-based simulation of realistic data. The package will soon be made available on CRAN. Further discussions focused on issues specific to simulating high-dimensional data. Michal Abrahamowicz (TG2/TG8), a co-chair of the Simulation Panel, made some suggestions based on his experience, and it was decided to pursue this challenging theme in a future in-person meeting during the ISCB 2020.

TG3 and TG9. Marianne Huebner (TG3) and Joerg Rahnenfuehrer (TG9) worked on the TG9 overview paper to align initial data analysis (IDA) steps for high dimensional data with those already developed by TG3 for 'standard' settings, where the number of subjects largely exceeds the number of variables.

TG3 and TG4. Marianne Huebner, Werner Vach, and Carsten Schmidt from TG3 met with Laurence Freedman, Victor Kipnis, Pamela Shaw, Michael Wallace, Paul Gustafson, and Veronika Deffner from TG4 to discuss possible future collaborations and the need to develop a common approach to measurement errors. TG4 approaches to measurement error rely mostly on modeling such errors in the context of specific application areas, e.g. nutrition and physical activities, or validation studies. On the other hand, TG3, so far, has discussed examples of detecting possible measurement errors through investigating data properties. CS shared examples of examiner effects over time in the SHIP data (see TG3 report above). LF suggested that CS shall come up with a concrete proposal on which the two TGs can work together.

TG3 and TG5. An informal discussion of Carsten Schmidt (TG3) and Mitchell Gail (TG5) aimed at identifying and initiating potential collaborations concerning study design set up that will facilitate the IDA activities. CS has sent a draft with ideas on study design relevant issues for IDA approaches being promoted by TG3.

TG4 and TG5. Mitchell Gail and Suzanne Cadarette (TG5 co-chairs) met with Laurence Freedman, Victor Kipnis, Pamela Shaw, Michael Wallace, Paul Gustafson, and Veronika Deffner from TG4 to discuss the objectives and the content of a future joint paper on design issues related to measurement error. The focus will be on the introduction of previously proposed two-phase designs in case-cohort or nested case-control studies (e.g. by N. Breslow) to researchers who may be unfamiliar with these techniques, in the context of investigating the safety or comparative effectiveness of treatments in observational data. We decided to focus on validation studies where there is gold standard (or at least accepted) measurement against which to assess error. A motivating example might be the use of electronic health records (EHRs), which are subject to measurement errors in both covariates and outcomes. Validation studies within and EHR study might improve inference. Ideally, one would have a test example with validated measurements on all subjects. Using these data we could estimate the "true" model. Then we could see how well limited-size subsamples used

for validation led to estimates approximating the true model, and we could see how far off were the original estimates, based on the model fit to error-prone EHR data. Pamela Shaw will liaise with Mitchell Gail on taking this plan further.

TG7 and TG4. Members of TG7 (Causal inference) met with Ruth Keogh (TG4) to discuss potential issues of common interest to both groups. Two topics emerged for future collaborations:

- Guidance on design and analysis of observational population-based studies on the real-life effectiveness and safety of cancer treatments, and their impact on the patients' quality of life.

- Developing further insights and guidance on landmark analysis and its ability to allow for additional causal effect insights under well understood assumptions.

TG6 and TG8. Discussion between TG6 members (Ewout Steyerberg, Ben van Calster, Marteen van Smeden) and Terry Therneau (TG8 co-chair) revolved around a planned review paper on the new approaches and criteria for performance assessment of the predictive models in time to event analyses of right-censored data. The content and methods of the paper were outlined and the design of future simulations, to validate and compare alternative approaches, was considered. Additional TG8 members may join this collaborative project in future.

TG4 and TG8. Michal Abrahamowicz (one of the TG8 co-chairs), and his two trainees, met with TG4 members: Laurence Freedman, Victor Kipnis, and Pamela Shaw, as well as a PhD student of Ruth Keogh, to discuss possible new collaboration on correcting for measurement errors in the context of flexible modeling of the effects of time-varying covariates (TVC) in survival analysis. These discussions identified 2 different sources of potential measurement errors: (i) 'classic' errors in the observed TVC values, and (ii) errors specific to TVC's, related to sparse and/or irregular measurements during the follow-up time (e.g. at the time of clinic visits), each of which requires different analytical tools. Additional challenges may need to be addressed if the hazard does depend not only on the current TVC value but possibly also on its past values (due e.g. to the lagged or cumulative effects). As a preliminary step, the participants exchanged some ideas regarding whether and how the existing approaches to handling measurement errors, such as SIMEX or regression calibration, could be adapted to this complex setting. It was decided that Michal Abrahamowicz will update the interested TG4 members on the progress in implementing these ideas and evaluating them in preliminary simulations.

TG7 and TG8. Most members of both TG7 (Causal Inference) and TG8 (Survival analysis) met to discuss topics of joint interest. Several topics were identified as promising, relevant and targets for future collaboration:

- A sequel of the original TG8 tutorial paper, where right censored survival times are considered as the key outcome, and all specifics and key complications for that setting are explained and similarly worked out, taking into account considerations relevant for causal inference . This joint project may represent an important step toward reconciliation of the survival analytical vs. causal inference research paradigms.

- An extension of the currently available methods that allow for more accurate modelling of the effects of measured time-varying confounders and time-varying exposures. The possibility of a future contribution, regarding this complex challenge, to the new book on survival analysis being written by Terry Therneau (TG8 co-chair) and co-authors was suggested.

- A further similar extension that will allow for mediation analysis.

- Development of a graph-based tool (DAG-oriented) and method to help researchers decide on the covariates they need to measure and adjust for in a given setting where the causal effect of an observational exposure is targeted, in time-to-event analyses, and the 'no unmeasured confounders assumption' will be relied upon.

6 Outline of the progress and future plans by selected STRATOS Panels

Eleven cross-cutting STRATOS panels have been created to coordinate the activities of different TGs, share best research practices, and disseminate the research tools and the results across the TGs. These panels address common issues such as creating a glossary of statistical terms, giving advice on how to conduct literature reviews or simulation studies, and setting publication policies for the initiative. The recommendations of the cross-cutting panels are intended to support, integrate and harmonize work within and across the TGs, and to increase transparency in producing guidance. The STRATOS website www.stratos-initiative.org provides an overview of all 11 panels.

Discussions about the structure, objectives and modus operandi of several panels played a key role in the first BIRS meeting, in 2016. Since then, several panel-specific issues have been resolved, so that during the 2019 Workshop, it was decided to focus more detailed panel presentations and discussions on the issues relevant mostly to three panels: the Simulation Panel, the (newly created) Visualization Panel, and the Glossary Panel. Below, we briefly summarize the goals of these three panels, and the relevant discussions during the 2019 BIRS Workshop.

Simulation Panel (SP)

Simulation studies are key to the work of all STRATOS TGs. They are essential to (i) validate some of the recently developed methods, (ii) compare how the relative advantages and disadvantages of alternative methods depend on the true underlying data structure, and (iii) at time, illustrate the pitfalls and potentially serious errors induced by conventional methods (frequently used in the applied research), especially when the underlying assumptions are violated. This panel develops and promotes principles for, and provides examples of, best practice for simulation studies.

Discussions during the BIRS workshop focused on drawing on the literature to identify the key principles, and find useful published examples, which will help TG members with the design, conduct, analysis and reporting of targeted simulation studies, addressing issues most relevant for their areas of expertise. This will help ensure that the conduct of simulation studies is as consistent as possible across TGs (especially when they touch on similar issues), and support the accessibility, transparency, and reproducibility principles, that are fundamental to all STRATOS research activities. Moreover, there were specific discussions on a project for a level 1 paper on simulation studies, and a potential collaboration with TG9 about the design of simulations for high-dimensional data.

Visualization Panel (VP)

The goal of quantitative science is to enable informed decisions and actions through a data-driven understanding of complex scientific questions. It is the role of any quantitative scientist (e.g. psychometrician, statistician, epidemiologist, etc.) to support this goal through (1) identification of the scientific question of interest (2) choice of the appropriate quantitative methods to address this question (experimental design, statistical or mathematical models, etc.), and (3) effective communication of the results. All of these aspects have to work in concert, including the 3rd challenge, which often receives less attention in the conduct of scientific studies. Yet, the ability of a scientific study to make an impact depends strongly on effective communication, and neglecting this essential component may be often the reason why many sophisticated investigations remain without a material impact. Effective visual communication is a core competency for the quantitative scientist. It is essential in every step of the quantitative workflow, from scoping to execution and communicating results and conclusions. With this competency, we can better understand data and influence decisions towards appropriate actions. Without it, we risk missing an opportunity to turn the research results into clear conclusions and evidence-based actions. These considerations motivated the STRATOS Steering Group to create, in 2018, a new Visualization Panel (VP), whose overall goal is to promote the use of good graphical principles for effective visual communication. The aim of the panel is to provide guidance and recommendations covering all aspects relevant for the statistical analyses, from the study design, to implementation of specific analytical methods, to the choice of most effective graphical displays and tools. Discussions during

the BIRS workshop focused on membership and direction of this newly formed Panel, with the aim to have a formal kick-off meeting around the end of 2019.

Glossary Panel (GP)

The status of the STRATOS glossary and a web-based user interface for its editing were reported.

The STRATOS glossary is based on the second edition of the "Dictionary for Clinical Trials" by Simon Day. All terms and definitions relevant to STRATOS have been imported from the original text source into a database. Links between dictionary entries have been conserved. Agreement has been given by NICE to use terms from their glossary (https://www.nice.org.uk/glossary). Relevant terms from there are currently being extracted and will be included into the database together with a source indicator (provenance) of each entry.

The current graphical user interface of the STRATOS glossary allows a limited number of editors to read and write proposals for changes and new entries. All users can see what the other editors proposed. They can individually add/comment/propose. The current editor user interface is technically based on Jupyter which allows for a rapid integration of new requirements. The current interface will not be the definitive end user interface.

The development of a STRATOS glossary is an important step for the overall STRATOS project. As a mid-term objective a machine-readable STRATOS terminology/ontology derived from the glossary should be developed integrating with already existing terminologies (e.g. from https://bioportal.bioontology.org/).

The next step will be to roll the database out to all STRATOS members, along with a little training/tutorial on how it can be used.

7   Further issues discussed at BIRS and plans for the future

7.1   STRATOS representation at future international conferences

Presentations at scientific conferences are one of the most important activities to disseminate knowledge developed within the STRATOS Initiative to a broader audience of analysts with varying levels of statistical education, experience and interests. Invited sessions, mini-symposia and courses are highly relevant to increase both the efficiency and the timeliness of dissemination of the STRATOS results. We discussed details of presentations in the near future (e.g. Mini symposium at the 40th ISCB meeting in July 2019) but also more detailed plans for future presentations (eg. Invited STRATOS Sessions at the two most important international biostatistical conferences in 2020: (i) the 41st ISCB conference in Poland (August) 2020, and (ii) the 30th IBC meeting in South Korea (July 2020), as well as mini-symposia at the CEN & GMDS 2020 meetings in Germany) and discussed to approach organizers of other meetings relevant for topics of STRATOS. After the BIRS Workshop we were invited to give a short presentation and present a poster (attached at the end) at the European Public Health Conference in November 2019.

7.2   STRATOS series in the *Biometric Bulletin*

After the invited session at the International Biometric Conference (IBC) in 2016 we were invited to write a series of short papers in the Biometric Bulletin, summarizing quarterly the news for members from the International Biometric Society (IBS). In 2019, IBS has approximately 6,000 members from 80 countries. We had started with an overview paper in Issue 3 of 2017, followed by a series of TG-specific short papers, in the consecutive issues in 2018 and 2019. During the BIRS Workshop, we discussed our interest to proceed with further STRATOS papers in the Biometric Bulletin and agreed that papers from several panels could also be very helpful to transfer STRATOS-generated knowledge to a wider community of for IBS members. Following on these discussions, more recently we have agreed with the IBS President and the Bulletin Editor to start with papers from STRATOS panels in issue 1/2020. The first two papers will be written by the Simulation and the Visualization Panels.

7.3 STRATOS involvement in the SISAQOL Consortium

Two weeks before the BIRS Workshop, the Consortium on Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data (SISAQOL, https://qol.eortc.org/projectqol/sisaqol/) has invited us to join a large grant application in the Innovative Medicines Initiative of the European Union. Viktor Kipnis (TG4) has agreed to review some of the related papers and gave a short presentation about the intended project. We discussed the potential interest for STRATOS and agreed having a first meeting with leading members of the SISAQOL Consortium in July 2019. Led and coordinated by members of the European Organisation for Research and Treatment of Cancer (EORTC) a grant application was submitted recently. Several STRATOS members lead a work package about recommendations from observational studies.

7.4 STRATOS paper on 'Methodological issues in medical research and patient care-critical appraisal of statistical and machine learning techniques'

Many stakeholders (researchers, companies, funders, the public, doctors, patients and more) consider 'Data Science' as a key part to improve issues related to patient care and health research. However, most people do not understand what's behind such complex concepts as 'Artificial Intelligence (AI)', personalized treatment, individual predictions, or causal effects. Even many researchers do not (fully) understand that AI related methods not only have an exciting potential but they also need to confront several important challenges. Aiming to provide more evidence supported knowledge about advantages and disadvantages of various techniques, during the Workshop we have proposed to write a STRATOS paper with the working title "Methodological issues in patient care and medical research - critical appraisal of statistical modeling and machine learning techniques". It was decided that Jörg Rahnenführer (TG9) and Matthias Schmid (TG2) will co-lead the development of this paper. In the four months after the BIRS Workshop, joint work on this complex paper has started and more than ten members from several STRATOS Topic Groups and Panels have joined the project and participated in discussions about its content and methods.

8 Summary of the BIRS Workshop activities and achievements

Activities:

Over the course of several plenary meetings, Topic Groups and selected Panels presented progress in identifying those issues within their areas of statistical expertise that need guidance, and in the steps undertaken to prepare and develop the guidance. This helped to inform and learn not only about the successes obtained so far, but also about the challenges having been encountered, and that solutions that either were already implemented and found to have worked or were planned to address these challenges. All members benefitted immensely from face-to-face meetings. This was vital for establishing new long-term research collaborations, the continuation of which will be almost exclusively by email, teleconferences and meetings of smaller groups.

Interactions between topic groups and establishing new cross-TG collaborative projects was one of the main themes of the BIRS Workshop. Topic groups came prepared with proposals of such projects and were able to connect in break-out sessions, which resulted in 10 concrete new collaborative projects (summarized in section 5 of this Report). In addition, to the inter-group collaborative plans, each TG discussed current and future STRATOS manuscripts and project management. In general, for each such collaboration, one or two leading researchers within the relevant TG have been identified, and all TG members who express an interest are invited to actively contribute to a given project. Input from the other TG members will be sought at various stages, e.g. before presentations and for revisions of manuscript drafts. Moreover, TG-specific websites with a common format were proposed in the interest of up-to-date information, activities, and resources. All participants expressed the strong wish having more regular meetings in smaller groups, and plans for several future TG meetings, joint meetings of the members from two or more TGs and/or one or two panels with somewhat overlapping mandates (e.g. Knowledge Translation vs Publication panels) were outlined during the Workshop small-group meetings and discussions.

Achievements:

The participants were unanimous in that the 2019 BIRS workshop was very successful in stimulating both: (i) further progress in research activities of individual Topic Groups (summarized in section 4 of this Report); and (ii) initiation of new, creative, multi-disciplinary inter-TG collaborative projects, tackling complex analytical challenges and issues that require joint expertise of the members of two or more different TG's (section 5 of the Report). Indeed, there was a general consensus that the BIRS Workshop created a unique opportunity for in-person exchanges and productive discussions between experts in different areas of statistics, and from 14 different countries on 3continents, without which such fruitful, dynamic interactions would not be possible. Each of the eight Topic Groups (TGs) represented at the Workshop made an impressive progress in developing and finalizing their respective research papers, as summarized in individual TG reports in section 4 above. Furthermore, the longer-term scientific yield of the BIRS Workshop activities will be reflected in many research papers designed or enhanced through the Workshop discussions (please see sections 4, 5 and 7.4 for more details on the manuscripts being prepared or planned through, respectively, (i) joint efforts of several members of specific TG's and (ii) new inter-TG collaborations initiated during the Workshop). Thus, the Workshop met its overarching objectives, outlined in section 2 above, of both enhancing and integrating the research activities of individual STRATOS Topic Groups. Given the achievements of the 2019 BIRS Workshop, there was also an overwhelming consensus that it will be essential to organize the next general meeting of the STRATOS team in about two years.