

Implementing online interventions in ICare: A biostatistical perspective

Dennis Görlich*, Andreas Faldum

Westfälische Wilhelms-Universität Münster, Institute of Biostatistics and Clinical Research, Schmeddingstraße 56, D-48149 Münster, Germany

ARTICLE INFO

Keywords:

Biostatistics
Study designs
Statistical analysis
Data management and protection
Randomized controlled trials

ABSTRACT

The implementation of research studies is a highly complex process. All decisions with respect to the study design impact the statistical analyses and interpretation of the results. Within the ICare research project (EU H2020 Grant agreement 634757) seven research trials are conducted to generate evidence on efficacy, effectiveness and the dissemination potential of online interventions targeting eating disorders, common mental health problems and resilience. Within the project a central biometrical unit was established to manage and coordinate data collection, processing and statistical data analysis. This allows for harmonized trial planning, conduct, data management processes and analysis strategies. The purpose of this article is to describe the common concepts underlying all seven ICare trials. This includes development of (adaptive sequential) study designs, handling of missing values, general data management and processing as well as data protection aspects.

1. Introduction

The planning, implementation and conduct of research studies involving human subjects is a highly complex process involving different areas of expertise. A research trials' main aim is to provide reliable evidence for the targeted research questions. To reach this overarching aim the trial is planned accordingly and a suitable study design is chosen. Evidence will usually be provided by the statistical analysis of the collected data. This statistical analysis should thus be planned appropriately. Researchers should be aware that all decisions concerning the study design will have an impact on the statistical analyses and the interpretation of results. Thus a biostatistician should be involved already in the planning phase of a trial. Large collaborative research projects usually produce a significant amount of heterogeneous research data from various data sources. This increases the need for reliable study designs in the context of the project. Within the ICare research project (EU H2020 grant agreement No. 634757), we implemented a central biometrical unit for all included trials (6 randomized trials and a non-randomized dissemination trial). The Institute of Biostatistics and Clinical Research (University of Münster, Germany) is the central partner for all aspects of statistical planning and data analysis, data management and data protection. This central unit will allow for harmonization of the relevant key infrastructures to improve data collection, data processing, analysis strategies and also aspects of data protection. In the ICare project, seven studies are conducted in six European countries targeting various mental health domains. The present article describes the principles and considerations with respect to

trial planning, data collection and data processing, as well as strategies for data analysis.

2. Study designs

ICare trials are (stratified), multi-center, multi-country, parallel group trials, aiming to prove efficacy, or effectiveness of the selected internet interventions. To develop study designs we incorporated the relevant guideline documents of the [International Consortium for Harmonization \(ICH\) E6 “Good Clinical Practice” \(1996\)](#), [ICH E8 “General Considerations for Clinical Trials” \(1997\)](#), [ICH E9 “Statistical principles” \(1998\)](#), and [ICH E10 “Choice of Control groups and Related Issues in Clinical Trials” \(2000\)](#). These guidelines primarily set the regulatory framework for clinical trials in the context of pharmaceuticals and medical products. However, they also provide general rules which can be translated to other domains involving studies with human subjects.

The ICare trials aim at various mental health domains, e.g., common mental health disorders such as anxiety and depression, substance use and adjustment disorders, eating disorders, problems and obesity, burden in carers, but also include strengthening healthy habits and resilience. Study designs for all trials have been developed in close collaboration with the respective researchers. During the planning phase, we defined the research hypotheses, selected the most appropriate outcome parameters, formulated statistical (null) hypotheses to prove the primary research questions, selected statistical test procedures and performed power calculations to determine the necessary

* Corresponding author.

E-mail addresses: dennis.goerlich@ukmuenster.de (D. Görlich), faldum.andreas@ukmuenster.de (A. Faldum).

<https://doi.org/10.1016/j.invent.2018.12.004>

Received 20 December 2017; Received in revised form 30 November 2018; Accepted 12 December 2018

Available online 11 January 2019

2214-7829/ © 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

sample sizes. For the details of each study please see the individual study protocols (Vollert et al., 2018; Nacke et al., 2018; Herrero et al., 2018; Weisel et al., 2018; Jones Bell et al., 2018; Musiat et al., 2018; Spencer et al., 2018). ICare trials target a variety of primary outcomes, such as short-term effects (e.g., reduction of scores at post-measurement), long-term effects (e.g., reduction of target scores at follow-up measurements). Also, time-to-event endpoints, e.g., time to symptom reduction, will be used for some of the primary confirmatory hypotheses. Suitable control conditions have been chosen (active control groups, or waiting-list conditions). All ICare trials follow a common measurement schedule, i.e. screening, baseline measurement (pre-intervention), post-intervention measurement, 6-month follow-up, and a 12-month follow-up measurement. Additionally, mid-intervention assessments are performed to identify potential mediator variables.

3. Adaptive group sequential study designs in ICare

Still the majority of trials are planned as fixed one-stage trials, i.e. the analysis of the primary outcome will be performed after all participants have been recruited and completed the assessments. These classical statistical designs are fixed in the sense, that the preplanned sample size will be recruited and the collected data will only be analysed after data collection is complete. No changes to the design should be made during the ongoing trial to prevent unwanted bias and inflation of the trial's significance level. However, often a more flexible analysis strategy is desirable to decide on trial conduct early. Adaptive group sequential study designs, introduced by Bauer (Bauer, 1989; Bauer and Köhne, 1994), allow for such flexibility. Following preplanned interim analyses, the study can be stopped for efficacy or futility, or the study continues and the design can be changed based on the interim results. The decision to stop for efficacy or futility is made based on a predefined threshold. If the trial is not stopped adaptations can be made to e.g., the number of study arms, sample size, or switch of hypotheses. The study result is obtained by combining the results of the study stages. The planned interim analyses have an impact on the properties of the statistical analysis. In particular, multiplicity is introduced into the trial and needs to be treated appropriately. The adaptive combination of trial stages is e.g. possible if the test statistics of the stages are independent. In the case of a typical outcome measure, such as a post-score, independence is reached by recruiting two independent waves of participants. For time-to-event outcomes, independent test statistics are not straight forward, but can be acquired by computing the independent increments. Nowadays, a huge variety of adaptive designs are available for different study designs and outcome measures. Adaptive designs are more widely and increasingly used in clinical research (Bhatt and Mehta, 2016; Gerß et al., 2015; Hatfield et al., 2016), but – to our current knowledge – have not yet been applied to online intervention trials. Adaptive designs were considered for all trials but only implemented in two trials (ICare everybody-Plus trial (Vollert et al., 2018) and the ICare Healthy Teens@School trial (Jones Bell et al., 2018)) due to practical reasons (e.g. sample size consideration). Both trials were planned with one interim analysis leading to two-stage trials. Fig. 1 shows the exemplary conditional error curve of one of these trials. The displayed curve is the conditional error function (CEF). It summarizes the possible decisions at the interim analysis time point and final analysis based on the analyses' p -values. For $p_1 \leq \alpha_1$ (the p -value p_1 of the first stage) an early stop is attained, due to proven study effect, e.g. a benefit in the primary outcome over an active control arm. For $p_1 > \alpha_1$ the study continues. The CEF represents the threshold for the p -value of the second stage. If p_2 is smaller or equal than the CEF (depending on p_1) the study was successful and a significant effect of the intervention has been shown within the final analysis. For values of p_2 greater than the CEF, no significant result was achieved. The curve also shows that a large p -value in the first stage can still lead to a significant overall study result, if the effect observed in the second stage is very strong.

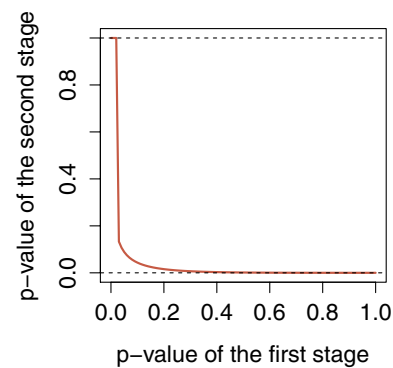


Fig. 1. Conditional error function of an adaptive group sequential design with two stages.

A variety of adaptive study designs and conditional error functions is available for trial planning. For an in-depth explanation see e.g. Wassmer and Brannath (2016). We chose to use optimal delta designs (Wang and Tsiatis, 1987), by optimizing the average sample size under the alternative hypothesis. These designs were developed for trials with equally sized stages as planned in ICare. Based on the planning parameters the adaptive designs within ICare require ca. 7% - 8% more participants, compared to one-stage trials, assuming both stages will be performed as planned. Early stopping and sample size recalculation can lead to a reduced overall sample size. Sample size recalculations at the interim analysis time point will be performed using the inverse normal method as proposed by Lehman and Wassmer (1999). For trial planning, sample size recalculations and the adaptive analyses the ADDPLAN Software will be used (Wassmer, 2006; Addplan Inc., an Aptive Solutions company).

4. Participant allocation

ICare trials are randomized controlled trials, except for the everybody dissemination trial (Nacke et al., 2018). The latter trial was planned as dissemination trial in a stratified, but not-randomized design. Validated randomization lists are generated for each trial based on permuted block randomization with fixed (confidential) block length. A balanced randomization ratio (1:1 or 1:1:1) is planned for ICare trials. Although permuted block randomization increases the chance for break of the allocation concealment and, by this, potentially adding bias, we prioritized the beneficial balancing properties (cf. Lachin et al., 1988; Matts and Lachin, 1988). To prevent selection bias to a certain degree randomization is organized centrally and independently from the trial personnel (principal investigator, therapists, etc.). For the selection of appropriate block sizes potential existence of chronological bias (time trends) was considered (Tamm and Hilgers, 2014) as well as the potential break of allocation concealment at the end of each block. The lists are generated using validated SAS scripts. Trials are stratified by country and potentially additional stratification variables (e.g., risk group). Randomization is provided via an online platform that allocates newly recruited participants to the study arms based on the pre-computed randomization lists. The implemented procedure maintains allocation concealment and by this reduces bias. During the analysis phase we will assess whether the randomization procedures worked as intended. For the everybody trial, which is planned as non-randomized dissemination study, the allocation to the study arms is based on stratification variables Body mass index (BMI), binge eating, and weight shape concerns (Nacke et al., 2018).

5. Planned statistical analyses

For all trials, statistical analysis plans (SAP) will be written before data analysis (as suggested by ICH E9 guideline). This should prevent

data-driven decisions and the introduction of bias. The SAP defines the planned primary confirmatory and key secondary statistical analyses and all relevant additional information about the trial. In particular, the purpose of the trial, the definition of the study collectives (intention-to-treat, per-protocol, as-treated) as well as criteria to identify these collectives, and the statistical hypotheses that will be tested in the planned analyses. Furthermore, detailed information about derived variables, e.g., definition of events, calculations for questionnaire scores, etc., are given. For each analysis, the SAP also defines the statistical tests that will be used to analyze the defined statistical hypotheses, the chosen significance levels, description of the multiple testing problem (if applicable), and sensitivity and safety analyses. In ICare trials a variety of sensitivity analyses will be applied that assess the robustness of the primary result under varying assumptions (e.g., normality), methods (e.g., parametric vs. non-parametric testing), collectives (e.g., per-protocol), and stratification methods (pooled vs. stratified analysis). Also, missing data which occurs in any clinical research trial (cf. O'Neill and Temple, 2012) is a major problem that should be addressed appropriately (O'Neill and Temple, 2012; Guideline on Missing Data in Confirmatory Clinical Trials, EMA, 2011). A main source of missing data in online interventions is attrition (cf. Eysenbach, 2005; Watson et al., 2018). Potential reasons for study discontinuation are e.g. reported by Fernández-Álvarez et al. (2017). In the ICare project a strategy to prevent missing data will be implemented as well as appropriate analysis strategies will be defined. Details how this problem is handled within the ICare project are given below. Subsequently, the secondary outcome parameters will be analysed using suitable statistical methods. Applied methods needs to take into consideration the longitudinal design of the trial. Therefore, we will mainly use generalized linear mixed models, or generalized estimation equations, which can model repeated measurements appropriately. Exploratory data analyses that will be performed with ICare data after the primary and key secondary analyses are finished should also be planned appropriately and described in (additional) SAP.

For each trial power and sample size calculations have been performed. The trial's sample sizes were based on the statistical properties of the chosen primary outcome measures, the statistical tests and considerations about multiplicity. In each trial the family-wise type-I error rate is controlled, if necessary, in a strong sense using appropriate methods (Bender and Lange, 2001; ICH E9, 1998) like Bonferroni correction. All trials ensure a power of at least 80% for the planned effect sizes. Sample sizes were corrected for the expected rate of study dropout. Effect size estimates are based on published literature. For details on the study specific effect size estimates and power calculations please see the individual study protocols (Vollert et al., 2018; Nacke et al., 2018; Herrero et al., 2018; Weisel et al., 2018; Jones Bell et al., 2018; Musiat et al., 2018; Spencer et al., 2018). Sample sizes were calculated using the following software: G*Power (Faul et al., 2007), PASS (Hintze, 2014; PASS 13. NCSS, LLC. Kaysville, Utah, USA. www.ncss.com) and ADDPLAN (Wassmer, 2006; ADDPLAN Version 6. Addplan, Inc., an Aptive Solution company).

6. Prevention and handling of missing data

Missing data is a major problem in clinical trials (cf. O'Neill and Temple, 2012). The impact of missing data on quantitative research can be serious, leading to biased parameter estimates, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings (Baraldi and Enders, 2010; Carpenter and Kenward, 2012; Enders, 2017; Dong and Peng, 2013). Therefore, missing data should be considered at each stage of a trial: design, conduct, analysis, and reporting (Bell et al., 2014). The prevention of missing data should be one major aim in clinical trials. Little et al. (2012) proposed a number of actions to be taken to prevent missing data in clinical trials. They identified eight points concerning the design stage, and eight points concerning the active trial phase that should be

considered to prevent missing data. Here, we summarize the most relevant aspects in the context of an online intervention setting as suggested by Little et al. (2012).

During the design stage it can be beneficial to target a population that has an intrinsic motivation to participate in the trial especially to participate in an online program which can be unguided and, if guided, often does not provide any personal contact to a guide beside feedback messages or online chats. Furthermore, it should be a standard procedure to apply a test phase before start of randomization, to optimize all aspects of study conduct, to e.g. prevent technical problems during the intervention as well as during data collection. Also, the choice and design of an online intervention can also have an effect on drop out and therefore impacts the statistical analysis, due to missing data. E.g. Fernández-Álvarez et al. (2017) report in a qualitative study that participants who discontinued a trial would prefer a more flexible treatment regimen (individualization). In the same analysis, also, technical aspects (usability, technical problems) were reported as factors leading to increased dropout. E.g. the ICare Prevent trial (Weisel et al., 2018) provides a transdiagnostic individual tailored online training. With respect to technical problems we will analyze our study data, by relating documented technical problems on the ICare platforms to the respective dropout rates. With respect to the assessment schedule it can be helpful to shorten the follow-up period for the primary outcome. By this the increased drop-out at later assessment time points can be avoided. Nevertheless, this procedure cannot be applied if long-term effects of online interventions need to be analysed. Also avoiding outcome measures with a high risk of missing data such as very long questionnaires can prevent attrition. In the everyBody Plus trial (Vollert et al., 2018) as well the ICare Prevent trial (Weisel et al., 2018) such outcomes, i.e. time to symptom reduction and time to onset, respectively, were included. Statistical methods for time-to-event outcomes (e.g. Kaplan-Meier estimates, Log-Rank tests, or Cox-proportional hazard regression) have the benefit that participants without complete data can be included into the analysis with their last known observation (right censoring), appropriately. By this all relevant observations can be taken into account for the analysis. Furthermore, the intention-to-treat principle is not violated, since all randomized participants are included into the analysis. Furthermore, Little et al. suggest for the active trial phase to set acceptable target rates for missing data and monitor the progress of the trial with respect to these targets. In case these rates are not met, e.g. either actions to improve motivation to provided data, or increased recruitment efforts (in accordance with an ethical approval) may be taken. Also, if the degree of missing data is substantial the statistical analysis plan might need to be amended. In the ICare project we apply a regular data quality monitoring which also reports on missing data. Incentives, such as vouchers, provided in an ethically sound way, could also increase adherence and prevent missing data. Researchers should consider limiting the burden and inconvenience of data collection on the participants (e.g. by integrating data collection neatly into the online system) and make the study experience as positive as possible. In the ICare projects, beside the necessary study-specific questionnaires, additional questionnaires (e.g. PHQ9 and GAD7) have been selected as common measures, to reach comparability between the trials. These questionnaires can be seen as additional burden to the participants, but are an integral part of the project related to a separate work package and research questions spanning all studies. Finally, training of investigators and study staff to keep participants in the trial is important, regardless of whether they continue to receive the assigned online program. Within the ICare trials, all investigators, trainers and coaches are trained to motivate participants to stay in the trials.

A large proportion of missing data is due to study drop-out or loss to follow-up. We adjusted the recruitment numbers within the ICare trials to cope with a significant amount of drop-out (between 20% and 40% based on prior experiences, please see also the individual study protocols) to maintain sufficient power for the primary analysis.

Nevertheless, further strategies will be necessary to deal with the collected study data, in case the preplanned rates are exceeded. In the ICare project, we will develop appropriate strategies to handle missing data within the statistical analysis. To prevent data-driven decisions a blinded data review step will be performed, without knowledge of the randomized study arm, by an (optimally) independent statistician, to assess missing data patterns. Multiple imputation (cf. Rässler et al., 2013), full information maximum likelihood (Hartley and Hocking, 1971) and expectation maximization (Dempster et al., 1977) methods will be considered. E.g. selection of auxiliary variables (i.e. covariables) will be following objective methods (cf. Collins et al., 2001; Enders, 2017; Pedersen et al., 2017) and advantages as well as disadvantages of the individual methods needs to be considered carefully (Dong and Peng, 2013). Given the active research in the field we will closely monitor the most recent state-of-the-art and apply the appropriate methods based on the observed types of missingness. The final decision on the optimal analysis strategy will be taken within the blinded review step. Statistical procedures to implement the mentioned methods exists in statistical software packages, e.g. PROC MI for general multiple imputation or the 'fiml' method in PROC CALIS for confirmatory factor model in the SAS Software (SAS Institute Inc., Cary, NC, USA).

The underlying mechanisms that lead to the missing data needs to be considered to select appropriate methods (Newman, 2014). Following Rubin's classification (Rubin, 1976; Little and Rubin, 2002) data is missing completely at random (MCAR) when a participant's probability to produce missing data does not depend on the observed data, nor on the unobserved (missing) data. Under MCAR we can expect increased standard errors (Enders, 2017), but it is unlikely that significant bias is introduced to the data. Data is missing-at-random (MAR) when the probability for missing data depends on observed data. MAR is assumed by the most prominent statistical methods (multiple imputation, maximum likelihood) for treating missing values. Often list-wise deletion is used during statistical analysis, which usually will yield biased results, if data are MAR (Molenberghs et al., 2014). Finally, data can be missing-not-at-random (MNAR). Here the probability for missing data is dependent on the unobserved (missing) data itself. Examples could be subgroups of ICare participants who intentionally do not disclose their household income (e.g., the participants from the lower income categories). The identification of the different missing data categories is not always feasible. While MCAR e.g., can be tested by Little's multivariate test (Little, 1988; Little and Schenker, 1995) there is no direct test for MAR. Under MNAR conditions, variables in the dataset are inadequate predictors of missingness undermining the effectiveness of the usual imputation methods, including multiple imputation. Obtaining accurate estimates from an MNAR mechanism requires sophisticated modeling approaches (Enders, 2010; Enders, 2017). Mainly Enders reviews different models for longitudinal data that can be broadly categorized in a selection model approach and a pattern mixture model approach. While the former comprises models that combine regression models for the outcome parameter and a set of regression models that predict missingness, the latter approach stratifies subset of the study cohort by missingness pattern. Due to the fact that the assumptions (e.g. normality) which needs to be taken for the analysis (either for the selection model approach or the pattern mixture approach) are not known before trial start these approaches can only be considered within the exploratory analyses of ICare data.

7. Data management and data protection

Data collection is done on the online platforms provided by Minddistrict (Minddistrict Ltd., Amsterdam, The Netherlands). The online platform integrates the delivery of the intervention programs and data collection by means of online questionnaires. Participants can seamlessly switch between the online program and data collection. The collected research data consists mainly of questionnaire data, socio-demographic information, health-economic questions, but will also

contain log file information about intervention usage and further "touchpoint" data (e.g. amount of messages written), documenting platform usage. Data can be exported from the Minddistrict platform in an accessible format, but require subsequent data processing. To define the relevant aspects of the project data management, a data management plan (DMP) has been developed. The DMP covers all processing steps that are necessary to generate analyzable datasets from the raw data exports. All data processing steps will be performed in the SAS software (SAS Inc., Cary, NJ, USA) as main data management platform. This includes, data import, processing of nested data structures, data annotation/meta data (coding), computations (e.g., questionnaire scores), data validation, and log file analysis (user journeys). Repeated processing steps will be programmed as SAS macros, allowing for easy and consistent data processing across all ICare studies.

Study meta-data will be collected using a separate item database. Meta-data will contain information about the collected data, i.e., variable names, measuring scale, category coding, units of measurements (e.g., cm, kg). This meta-data can be used e.g., to generate code books. Additionally, we will use the meta-data to detect inconsistencies and develop harmonization strategies (e.g., common coding) before merging data of different countries or trials.

The ICare data protection plan aims at providing a unified protection level for personal data of the study participants. Data needs to be protected from data collection until data analysis, i.e. at all data processing steps. Before the project's data protection plan was established, all ICare partners participated in consortium-internal survey. The survey covered information about the types of personal data (name, email address, ...) and which items are considered critical for pseudonymisation and anonymisation. Also, the possibility to collect informed consent online was considered, as well as storage durations, as well as additional requirements (e.g. duration of data storage). The survey resulted in an overview on data protection rules in the different countries, e.g. personal information that needs to be deleted in case of anonymization, or if an online consent (compared to a paper based consent) is possible in general. The results of the survey allowed us to tailor a data protection plan, which takes national differences into account. With the introduction of the General Data Protection Regulation (Regulation (EU) 2016/679) a harmonized legislation is active, which is also adhered to within the ICare project.

Data collection and processing in the ICare trials, as a scientific research project, is only possible after participants have given informed consent. The following data protection principles are applied by ICare partners: (i) Lawful and transparent data processing. Data is only be collected after informed consent is given; (ii) Data is collected to specified, explicit, and legitimate purposes. Participants agree that the collected data is used in the scientific analysis of the ICare trials. Additionally, anonymized data can be used for subsequent scientific research purposes; (iii) Data collection is limited to necessary and relevant data items. No unnecessary data will be collected, and all collected data items will be used in one of the planned analyses, i.e. either within the primary confirmatory, sensitivity, or exploratory analyses. By this we adhere to the data minimization principle as describes in data protection legislation; (iv) Data accuracy is maintained, i.e. incorrect data will be rectified and all data processing steps will be carefully validate to no introduce inconsistent data sets; (v) Identification of participants (e.g., by email addresses) will only be possible as long as necessary. After the end of the trial (plus a defined period of time for query management) all identifying data (personal information) will be deleted further anonymisation steps will be discussed and applied if necessary; (vi) data should be protected against (vi.a) unauthorized access. This is maintained by a role-concept at platform level and further technical measures active after data export. (vi.b) Accidental loss is prevented. Effective backup procedures at all data storage locations are in action. (vi.c) Destruction or damage is prevented. This is mainly guaranteed by protection systems in the respective data storage locations, e.g. fire-extinguishing systems in server

rooms. Security of all computers that can access ICare data needs to be guaranteed. I.e. regular virus scans, and updates of operation systems and other software are necessary. As part of the technical and organizational measures (TOM) that are implemented in ICare participants and ICare therapists log on to the platform using their personal login credentials, preventing unwanted access. Personal data is only visible for the participant and the associated therapist (in guided interventions). Access to the platform is protected by encrypted communication channels. A pseudonymization number is assigned to participants within the platforms. This pseudonymization is used to identify participant data in the different datasets during data processing. On some platforms participants can choose to take part as anonymous users using a self-chosen pseudonym name. E.g. this feature allows participants to keep a degree of confidentiality e.g. in group discussions between participants and therapists (if offered). If anonymous users are allowed on a platform was decided individually for each trial and country.

During online interventions, and thus also in the ICare trials, the direct personal face-to-face contact between therapist and participant often does not exist. Thus, the collection of a legitimate consent needs to be organized appropriately, such that the legitimacy can be guaranteed. If an online consent is allowed this is procedure is used in ICare. Otherwise, a paper-based consent, sent via mail to the study coordinator, will be collected. If the consent is given together with other information (e.g., a study information document) the consent on data processing needs to be clearly identifiable as such. Participants must understand the content of the informed consent; thus, it should be phrased in clear and easy language in an easily accessible form. Benefits and obligations of the trial protocol are described clearly. Some trials offer incentives for fully answered questionnaires. These were mentioned during the ethical review of trials. Ethics approvals were obtained for all trials and study centers. Trial participants can withdraw their consent at any time during the running trials without any disadvantages.

8. Data quality

Data quality is of enormous importance for the validity of every research project. Low-quality data can lead to either wrong results or, potentially complete loss of participant data for the analyses. To produce high-quality evidence and reliable results we established a common data quality monitoring workflow. Data quality issues can emerge from various sources that can be classified into quality impairment in (i) accuracy/validity, (ii) composition and organization, (iii) completeness, (iv) transparency, and (v) timeliness.

We will subsequently discuss these quality domains, as active in the ICare project, sketch potential problems, and describe actions to prevent quality loss.

Accuracy measures if study data correctly represents the attributes of a real-world object or events. Inaccurate data, i.e. false information, can be introduced into data sets either accidentally (e.g., while entering data) or on purpose. Detecting inaccurate data within the source data is almost not possible within ICare trials, since this primary data is entered remotely, i.e. directly on the online platform by the participants. Wrong information thus cannot be compared to any kind of source documentation. Derived data sets, e.g., processed data, can however, be compared against the source data. To some degree, inaccurate data can be detected and prevented by defining validation rules that either restrict the range of valid values (e.g. minimum and maximum) or use context information, e.g., from other data items. A typical example is the question “Are you currently pregnant?” that cannot be answered by male participants or conditional questions that depend on an online calculated questionnaire score. Similarly, validation rules using context information about the respective items can be constructed to identify potential data quality impairment.

Data quality problems that emerge from composition and data organization can also lead to wrong study results. Data composition and

organization refers to structural aspects, e.g., longitudinal vs. horizontal data layout, but also to meta-data, e.g., the coding of categorical variables and the measurement units of continuous variables. These issues mainly emerge during data collection and data processing.

All ICare trials use the same technical platform leading to a homogenous data management workflow. During data processing, several datasets need to be merged. This can either be done by adding variables or new cases to an existing data set. The former situation can lead to inconsistencies by assigning the new data to wrong participant data rows. This can efficiently be prevented by always using suitable identifying variables, such as participant ID and potentially additional variables, such as the respective assessment time point. Adding new cases to a data set can lead to inconsistencies due to incompatible data types. Different trials might contain identically named variables, but coding or measurement units differ. We can detect such problems using an ICare item database (compare [Section 7 Data management and data protection](#)). Before datasets can be merged, the affected variables need to be “homogenized”, i.e., values need to be mapped and/or transformed.

Data completeness is a major issue in all types of studies. Missing data often leads to the complete exclusion of participants from data analysis (compare [Section 6 Prevention and handling of missing data](#)), which can lead to biased results (cf. ICH Guidelines). Incompleteness in datasets can have various reasons, e.g., data loss during data collection or processing, missing values (not provided by participants), data destruction due to external events (e.g., fire). Data loss during data collection can only be solved on a technical basis. At all data processing steps, data loss has to be prevented, i.e., in the ICare project during data entry on the online platforms, data export and data processing within the data management software. Missing values, i.e., data not provided by participants, often cannot be prevented. Participants will be informed about the trials and explicitly reminded that the ICare questionnaires are important for the underlying research questions. Nevertheless, online interventions usually show a much higher drop-out rate compared with clinical trials among hospitalized patients. Participants can easily stop participating in trials due to the remote setting or high questionnaire burden. The main intention to include a substantial number of questionnaires is to characterize study participants as completely as possible to ensure exhaustive and comprehensive interpretation of study results. Several statistical imputation methods to cope with missing data exist (see above). While the power of a trial will be maintained after data imputation, the application of imputation methods to data sets with high percentage of missing data is often also questionable, since the results will strongly depend on the chosen imputation method. We will closely monitor data completeness as part of the regular data quality reports. This will also help to identify data loss due to technical reasons within the online platform. The affected data items or questionnaires will inform trial coordinators about potential technical problems and/or critical items. In severe cases, changes to the provided electronic questionnaires needs to be made. As a quantitative measure, the percentage of participants who provided data at a specific assessment time point will be analysed. Additionally, the percentage of participants who provided specific questionnaires (e.g., the outcome measures) will also be provided. Another quantitative measure is the participant's individual data completeness ratio. We compute this ratio as relative frequency of provided items from all expected items per participant. This individual completeness ratio will be analysed as continuous data quality outcome.

Transparency is a data quality dimension which is not directly measurable, but can help to identify data quality issues. All data processing steps should be traceable and changes within the dataset should be documented. The origin of each data item (original data, or derived variables) needs to be known. By informing all persons who will work with study data (raw data, processed, analyzable datasets) about the data management and processing steps misunderstandings can be prevented with respect to data analysis and interpretation. This also

comprises the provision of data books containing information (e.g. coding information) about the items in a specific dataset.

Timeliness has two aspects that need to be considered for data quality. The first aspect refers to the trial's assessment schedules. Planned assessment time points (in ICare trials: pre, mid, post, 6 months, and 12 months) should be adhered to. Thus, deviance from this schedule is also an adherence indicator. The comparability of data can be threatened when the schedule is not adhered to, especially, when delays are not distributed at random, but are associated with a certain participant characteristic. A second aspect of timeliness is the timely provision of datasets to reach the project goals. This mainly affects the timely export from the database and the preparation of analyzable data sets.

Within the ICare data quality assessment the only a selection of variables will be monitored in a risk based approach. A variable will be monitored for data quality based on its importance in reaching the project goals. Quality impairments of these variables may threaten the project's success. For each trial the selection of monitored variables will be defined from the following data items: Primary and key secondary outcome variables, the health-economic questionnaires (Client Service Receipt Inventory: CSRI; Beecham and Knapp, 2001), moderator and mediator variables, and measures of adherence.

Overall, the implemented data quality analyses in the domains accuracy, composition and organization, completeness, and timeliness resembles a central monitoring, i.e. analyses can be made based on the exported datasets during the running trials at the central biometric unit. Data quality reports will be provided to the ICare principle investigators in a regular manner and can be taken into account to assess potential problems in the running ICare trials.

9. Conclusion and outlook

In this article, we discussed different aspects of ICare research trials and elaborated on the design decisions within the ICare project. We foster harmonized data management, central monitoring, data protection and data analysis strategies. ICare trials implement actions to prevent missing data during the planning stage and during the active trial phase. All aspects of a study design have an impact on the planned analyses and the interpretation of a trial. Thus, the involvement of a biostatistician in all trial phases (planning, implementation, conduct, analysis) is beneficial and often mandatory even in non-pharmaceutical trials. Recent reviews on adaptive designs in clinical trials shows a clear trend towards applying these designs more often (Hatfield et al., 2016; Bothwell et al., 2018). The two ICare trials implementing adaptive designs will be among the first trials targeting online interventions. Adaptive group sequential trials designs can be used to decide on a trial result early and, in the case of trial continuation, to adapt various aspects of a trial design, e.g. sample size, but also statistical hypotheses. For the two specific ICare trials we decided that an early information on a potential study effect would be most beneficial. As potential adaptation only a sample size recalculation was preplanned.

In clinical research the application of guidelines significantly improves the quality of research results, by proper planning and reporting. Finally, by applying the actions as described in this article we will achieve high data quality and also high quality scientific results to provide reliable evidence within online intervention research.

Competing interests

The authors declare that they have no competing interests.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 634757.

Author contributions

DG and AF wrote the manuscript and approved the final version.

References

- Baraldi, A.N., Enders, C.K., 2010. An introduction to modern missing data analyses. *J. Sch. Psychol.* 48 (1), 5–37. <https://doi.org/10.1016/j.jsp.2009.10.001>.
- Bauer, P., 1989. Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie* 20 (4), 130–148.
- Bauer, P., Köhne, K., 1994. Evaluation of experiments with adaptive interim analyses. *Biometrics* 50 (4), 1029. <https://doi.org/10.2307/2533441>.
- Beecham, J., Knapp, M., 2001. Costing psychiatric interventions. In: Thornicroft, G. (Ed.), *Measuring Mental Health Needs*, Gaskell, 2nd edition. pp. 200–224.
- Bell, M.L., Fiero, M., Horton, N.J., Hsu, C.-H., 2014. Handling missing data in RCTs; a review of the top medical journals. *BMC Med. Res. Methodol.* 14 (118), 1–8. <https://doi.org/10.1186/1471-2288-14-118>.
- Bender, R., Lange, S., 2001. Adjusting for multiple testing - when and how? *J. Clin. Epidemiol.* 54 (4), 343–349. [https://doi.org/10.1016/S0895-4356\(00\)00314-0](https://doi.org/10.1016/S0895-4356(00)00314-0).
- Bhatt, D.L., Mehta, C., 2016. Adaptive designs for clinical trials. *N. Engl. J. Med.* 375 (1), 65–74. <https://doi.org/10.1056/NEJMr1510061>.
- Bothwell, L.E., Avorn, J., Khan, N.F., Kesselheim, A.S., 2018. Adaptive design clinical trials: a review of the literature and ClinicalTrials.gov. *BMJ Open* 8, e018320. <https://doi.org/10.1136/bmjopen-2017-018320>.
- Carpenter, J.R., Kenward, M.G., 2012. Multiple imputation and its application. In: *Multiple Imputation and its Application*. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119942283>.
- Collins, L.M., Schafer, J.L., Kam, C.-M., 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* 6 (4), 330–351. <https://doi.org/10.1037/1082-989X.6.4.330>.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39 (1), 1–38. <https://www.jstor.org/stable/2984875>.
- Dong, Y., Peng, C.-Y.J., 2013. Principled missing data methods for researchers. *Springerplus* 2 (1), 222. <https://doi.org/10.1186/2193-1801-2-222>.
- Enders, C.K., 2010. *Applied Missing Data Analysis*. The Guilford Press.
- Enders, C.K., 2017. Multiple imputation as a flexible tool for missing data handling in clinical research. *Behav. Res. Ther.* 98, 4–18. <https://doi.org/10.1016/j.brat.2016.11.008>.
- European Medicines Agency, 2011. *Guideline on Missing Data in Confirmatory Clinical Trials*, CPMP/EWP/1776/99 Rev. 1.
- Eysenbach, G., 2005. The law of attrition. *J. Med. Internet Res.* 7 (1), e11. <https://doi.org/10.2196/jmir.7.1.e11>.
- Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39 (2), 175–191. <https://doi.org/10.3758/BF03193146>.
- Fernández-Álvarez, J., Díaz-García, A., González-Robles, A., Baños, R., García-Palacios, A., Botella, C., 2017. Dropping out of a transdiagnostic online intervention: a qualitative analysis of client's experiences. *Internet Interv.* 10, 29–38. <https://doi.org/10.1016/j.invent.2017.09.001>.
- Gerß, J., Eveslage, M., Faldum, A., Schmidt, R., 2015. Modern study designs and analysis methods in clinical research. *Z. Rheumatol.* 74 (2), 106–1012. <https://doi.org/10.1007/s00393-014-1450-8>.
- Hartley, H.O., Hocking, R.R., 1971. The analysis of incomplete data. *Biometrics* 27 (4), 783–823. <https://doi.org/10.2307/2528820>.
- Hatfield, I., Allison, A., Flight, L., Julious, S.A., Dimairo, M., 2016. Adaptive designs undertaken in clinical research: a review of registered clinical trials. *Trials* 17 (1), 150. <https://doi.org/10.1186/s13063-016-1273-9>.
- Herrero, R., Mira, A., Cormo, G., Etchemendy, E., Banos, R., García-Palacios, A., Ebert, D.D., Franke, M., Berger, T., Schaub, M., Görlich, D., Jacobi, C., Botella, C., 2018. An internet based intervention for improving resilience and coping strategies in university students: Study protocol for a randomized controlled trial. *Internet Interv.* <https://doi.org/10.1016/j.invent.2018.03.005>.
- International Commission on Harmonization, 1996. *Harmonised tripartite guideline: guideline for good clinical practice E6(R1)*. In: *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*. European Medicines Agency.
- International Commission on Harmonization, 1997. *Harmonized tripartite guideline: general considerations for clinical trials, E8*. In: *International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use*, pp. 1–13. <https://doi.org/10.1016/j.ygyno.2007.12.032>.
- International Commission on Harmonization, 2000. *Harmonised tripartite guideline: choice of control group and related issues in clinical trials*. In: *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*. European Medicines Agency.
- International Conference on Harmonisation, 1998. *ICH Harmonised Tripartite Guideline for Statistical Principles for Clinical Trials, E9*. Brookwood Medical Publications. [https://doi.org/10.1002/\(SICI\)1097-0258\(19990815\)18:15<1903::AID-SIM188>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1097-0258(19990815)18:15<1903::AID-SIM188>3.0.CO;2-F).
- Jones Bell, M., Zeiler, M., Herrero, R., Kuso, S., Nitsch, M., Etchemendy, E., Fonseca-Baeza, S., Oliver, E., Adamcik, T., Karwautz, A., Wagner, G., Banos, R., Botella, C., Görlich, D., Jacobi, C., Waldherr, K., 2018. Healthy teens @ school: evaluating and disseminating transdiagnostic preventive interventions for eating disorders and obesity for adolescents in school settings. *Internet Interv.* <https://doi.org/10.1016/j.invent.2018.02.007>.

- Lachin, J.M., Matts, J.P., Wei, L.J., 1988. Randomization in clinical trials: conclusions and recommendations. *Control. Clin. Trials* 9 (4), 365–374.
- Lehmacher, W., Wassmer, G., 1999. Adaptive sample size calculations in group sequential trials. *Biometrics* 55 (4), 1286–1290. <https://doi.org/10.1111/j.0006-341X.1999.01286.x>.
- Little, R.J.A., 1988. A test of missing completely at random for multivariate data with missing values. *J. Am. Stat. Assoc.* 83 (404), 1198–1202. <https://doi.org/10.1080/01621459.1988.10478722>.
- Little, R.J.A., Rubin, D.B., 2002. *Statistical analysis with missing data*, second edition. Wiley-Interscience.
- Little, R.J., Schenker, N., 1995. Missing data. In: Arminger, G., Clogg, C., Sobel, M. (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Plenum Press, New York, pp. 39–75.
- Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Frangakis, C., Hogan, J.W., Molenberghs, G., Murphy, S.A., Neaton, J.D., Rotnitzky, A., Scharfstein, D., Shih Weichung, J., Siegel, J.P., Stern, H., 2012. The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.* 367 (14), 1355–1360. <https://doi.org/10.1056/NEJMs1203730>.
- Matts, J.P., Lachin, J.M., 1988. Properties of permuted-block randomization in clinical trials. *Control. Clin. Trials* 9 (4), 327–344.
- Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A.A., Verbeke, G. (Eds.), 2014. *Handbook of Missing Data Methodology*. Taylor & Francis Ltd.
- Musiati, P., Potterton, R., Gordon, G., Spencer, L., Zeiler, M., Waldherr, K., Kuso, S., Nitsch, M., Adamcik, T., Wagner, G., Karwautz, A., Ebert, D.D., Dodd, A., Dooley, B., Harrison, A., Haselgrove, M., Sharpe, H., Smith, J., Tressler, R., Troop, N., Vinyard, C., Görlich, D., Beecham, J., Bonin, E., Jacobi, C., Schmidt, U., 2018. Web-based indicated prevention of common mental disorders in university students in four European countries – study protocol for a randomised controlled trial. *Internet Interv.* <https://doi.org/10.1016/j.invent.2018.02.004>.
- Nacke, B., Beintner, I., Görlich, D., Vollert, B., Schmidt-Hantke, J., Hütter, K., Taylor, C.B., Jacobi, C., 2018. everyBody–tailored online health promotion and eating disorder prevention for women; study protocol of a dissemination trial. *Internet Interv.* <https://doi.org/10.1016/j.invent.2018.02.008>.
- Newman, D.A., 2014. Missing data: Five practical guidelines. *Organ. Res. Methods* 17 (4), 372–411. <https://doi.org/10.1177/1094428114548590>.
- O'Neill, R.T., Temple, R., 2012. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. *Clin. Pharmacol. Ther.* 91 (3), 550–554.
- Pedersen, A.B., Mikkelsen, E.M., Cronin-Fenton, D., Kristensen, N.R., Pham, T.M., Pedersen, L., Petersen, I., 2017. Missing data and multiple imputation in clinical epidemiological research. *Clin. Epidemiol.* 9, 157–166. <https://doi.org/10.2147/CLEP.S129785>.
- Rässler, S., Rubin, D.B., Zell, E.R., 2013. Imputation. *Wiley Interdiscip. Rev. Comput. Stat.* 5 (1), 20–29. <https://doi.org/10.1002/wics.1240>.
- Rubin, D.B., 1976. Inference and missing data. *Biometrika* 63 (3), 581–592. <https://doi.org/10.2307/2335739>.
- Spencer, L., Schmidt-Hantke, J., Allen, K., Gordon, G., Potterton, R., Musiat, P., Hagner, F., Beintner, I., Görlich, D., Beecham, J., Bonin, E., Jacobi, C., Schmidt, U., 2018. A web-based intervention for carers of individuals with anorexia nervosa (we can): trial protocol of a randomised controlled trial investigating the effectiveness of different levels of support. *Internet Interv.* <https://doi.org/10.1016/j.invent.2018.02.005>.
- Tamm, M., Hilgers, R.D., 2014. Chronological bias in randomized clinical trials arising from different types of unobserved time trends. *Methods Inf. Med.* 53 (6), 501–510.
- Vollert, B., Beintner, I., Musiat, P., Gordon, G., Nacke, B., Schmidt-Hantke, J., Potterton, R., Spencer, L., Grant, N., Görlich, D., Schmidt, U., Jacobi, C., 2018. Using internet-based self-help to bridge waiting time for face-to-face outpatient treatment for bulimia nervosa, binge eating disorder and related disorders: study protocol of a randomized controlled trial. *Internet Interv.* <https://doi.org/10.1016/j.invent.2018.02.010>.
- Wang, S.K., Tsiatis, A.A., 1987. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 43 (1), 193–199. <https://doi.org/10.2307/2531959>.
- Wassmer, G., 2006. Planning and analyzing adaptive group sequential survival trials. *Biom. J.* 48 (4), 714–729. <https://doi.org/10.1002/bimj.200510190>.
- Wassmer, G., Brannath, W., 2016. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer International Publishing.
- Watson, N.L., Mull, K.E., Heffner, J.L., McClure, J.B., Bricker, J.B., 2018. Participant recruitment and retention in remote eHealth intervention trials: methods and lessons learned from a large randomized controlled trial of two web-based smoking interventions. *J. Med. Internet Res.* 20 (8), e10351. <https://doi.org/10.2196/10351>.
- Weisel, K., Zarski, A.-C., Berger, T., Krieger, T., Schaub, M.P., Moser, C.T., Berking, M., Dey, M., Botella, C., Banos, R., Herrero, R., Etchemendy, E., Riper, H., Cuijpers, P., Bolinski, F., Kleiboer, A., Görlich, D., Beecham, J., Jacobi, C., Ebert, D.D., 2018. Efficacy and cost-effectiveness of guided and unguided internet-based mobile-supported indicated prevention of depression and anxiety (ICare Prevent): a multi-country three-armed randomized controlled trial. *Internet Interv.* <https://doi.org/10.1016/j.invent.2018.04.002>.