Medizinische Fakultät
der
Universität Duisburg-Essen

Aus dem Institut für Medizinische Informatik, Biometrie und Epidemiologie
(IMIBE)

Non-interventional studies - limitations, challenges, and opportunities

I n a u g u r a l d i s s e r t a t i o n
zur
Erlangung des Doktorgrades
Doctor of Philosophy (PhD)
durch die Medizinische Fakultät
der Universität Duisburg-Essen

Vorgelegt von
Marjan Amiri
aus Teheran, Iran
2021

DOI: 10.17185/duepublico/74733
URN: urn:nbn:de:hbz:464-20210923-083820-2

Dekan: Herr Univ.-Prof. Dr. med. J. Buer

1. Gutachter: Herr Prof. Dr. med. M. C. Michel

2. Gutachter: Herr Univ.-Prof. Dr. med. D. Dobrev

Tag der mündlichen Prüfung: 6. August 2021

List of publications

Amiri, M., Deckert, M., Michel, M. C., Poole, C. & Stang, A. (2021), Statistical inference in abstracts of three influential clinical pharmacology journals analyzed using a text-mining algorithm, British Journal of Clinical Pharmacology (Accepted for publication). doi: 10.1111/bcp.14836

Amiri, M., Schneider, T., Oelke, M., Murgas, S., Michel, M.C. (2021) Factors associated with initial dosing for up-titration of propiverine and treatment outcomes in overactive bladder syndrome patients in a non-interventional setting. Journal of Clinical Medicine, 10: 311. doi:10.3390/jcm10020311.

Amiri, M., Murgas, S., Stang, A., & Michel, M. C. (2020). Do overactive bladder symptoms and their treatment-associated changes exhibit a normal distribution? Implications for analysis and reporting. Neurourol Urodyn, 39: 754-761. doi:10.1002/nau.24275

Kuklik, N., Stausberg, J., **Amiri, M.**, & Jöckel, K. H. (2019). Improving drug safety in hospitals: a retrospective study on the potential of adverse drug events coded in routine data. BMC Health Serv Res, 19 (1): 555. doi:10.1186/s12913-019-4381-x

Table of contents

1. Introduction

1.1. Evidence based medicine

High quality evidence is the key to sound decision making in medicine. Evidence-based medicine is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al., 1996). During the clinical development of new drugs, randomized clinical trials (RCT) provide evidence about its safety and efficacy, which are generated from a controlled study setting. Typically, RCT are performed according to Good Clinical Practice (GCP) guidelines, which results in high data quality and internal validity, but due to extensive lists of inclusion and exclusion criteria, they have a more restricted external validity. Accordingly, RCT do not produce real world evidence (RWE).

In the light of using digital technologies in health care systems, electronic health data is becoming ubiquitous. By development of statistical software packages and data mining tools, aggregation and analysis of these big data is feasible. Currently, Real World Data (RWD) is used in different stages of the drug development. Their application ranges from go-no go decisions for development (e.g., by recognizing the unmet medical needs, evaluating the burden and epidemiology of the disease), improving the efficiency of clinical trials (e.g., by generating hypotheses for testing, and assessing trial feasibility) to market access and finally in post-marketing surveillance activities (Breckenridge et al., 2019, Galson, 2016). Regulatory authorities have been always interested in collecting RWD. Mainly by spontaneous adverse event reporting systems; the Eduravigilance data bank developed by European Medicine Agency (EMA), the Yellow Card Scheme from Medicine and Healthcare products Regulatory Agency (MHRA) in United Kingdom (UK) and the Adverse Event Reporting System of Food and Drug Administration (FDA) in United States of America (USA)(Breckenridge et al., 2019). Another way is by requesting a Post-authorization Study (PAS). They are all ways that regulatory authorities use to monitor the safety of the marketed medicinal products.

The FDA defines RWD as "the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources" (FDA, 2016). RWD

can either be collected from the existing data sources such as Electronic Healthcare Records (EHRs), claims and billing activities, product and disease registries, patient-generated data including in home-use settings, data gathered from other sources that can inform on health status, such as mobile devices or by generating completely new data e.g., by conducting post-marketing observational studies. RWD sources can be used to generate RWE. FDA defines RWE as "the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD. They can be generated by different study design or analyses, including but not limited to, randomized trials (e.g., large simple trials, pragmatic trials), and observational studies (prospective and/or retrospective)"(FDA, 2016).

1.2. Randomized clinical trials versus non-interventional studies

RCTs are the gold standard for evaluating the efficacy and safety of medicinal products, particularly for obtaining marketing authorization. The evidence gained from RCTs has high internal validity due to randomization and, most often, blinding, and typically following GCP guidelines. However, this evidence is based on the limited information from a highly selective and homogeneous study population, collected in an experimental setting for a relative short time. Therefore, evidence from RCT has limited external validity, which also limits insight about the cost effectiveness of a treatment. Further obstacles of RCTs are high cost and difficulty to recruit patients, particularly in oncology or rare diseases trials, where conducting RCT may face ethical issues. Thus, there is an increasing recognition that RWE needs to complement the evidence from RCTs.

While the evidence from RCTs is usually gained from interventional trials, RWE is mostly generated from non-interventional studies (NIS), which are observational studies. The term "trial" in RCT implies the experimental character of them, whereas the term "observational study" indicates that the use of the medicinal product is only observed under current clinical practice. In the context of a NIS, a treatment is not administered for study purpose but rather based on medical consideration and there is no random allocation (EMA, 2011). To avoid disturbing the real word setting, measures not applied in routine clinical practice are typically avoided, which includes non-routine diagnostics prior to treatment initiation and additional tests during and after treatment. Therefore, there are concerns about the data quality of NIS mainly because of missing or incomplete data and

hidden biases due to cofounding effects of underlying risk factors such as concomitant medications or diseases.

It is important to notice that RCT and NIS have different objectives. The aim of RCT is to evaluate the “efficacy” of a medicinal product under study conditions, whereas NIS measures the “effectiveness”, which is an evaluation of the beneficial effect under real life condition in a heterogeneous patient population (Singal et al., 2014).

1.3. Non-interventional studies

1.3.1. Definition

Article 2 (c) of Directive (DIR) 2001/20/EC defines NIS as “a study where the medicinal product is prescribed in the usual manner in accordance with the terms of the marketing authorization. The assignment of the patient to a particular therapeutic strategy is not decided in advance by a trial protocol but falls within current practice and the prescription of the medicine is clearly separated from the decision to include the patient in the study. No additional diagnostic or monitoring procedures shall be applied to the patients and epidemiological methods shall be used for the analysis of collected data”(European Parliament, 2001). The upcoming clinical trial Regulation (REG) No.536/2014 in article 2(2) (4) defines a NIS as "a clinical study other than a clinical trial" (European Parliament, 2014). A decision tree is provided in the regulation No. 536/ 2014 Question and Answers document and helps to decide if a study considered as NIS in the sense of this new clinical trial regulation (European Parliament, 2021) (Table 1).

The European Network of Centers for Pharmacoepidemiology and Pharmacovigilance (ENCePP) states that “the current definition of NIS allows room for interpretation; as a result studies have been delayed or have not been conducted at all “ due to differences in the interpretation of the definition (EMA, 2011). As mentioned by the module VIII of the Good Pharmacovigilance Practices (GVP), “NIS are defined by the methodological approach used and not by its scientific objectives” (EMA, 2017) and sometimes it is difficult to distinguish the interventional or non-interventional nature of them, especially when additional monitoring or testing is included in the study. Furthermore, ENCePP

highlighted that “current clinical practice regarding the diagnosis, intervention, and follow-up of a specific medical problem can vary between healthcare professionals and can differ depending on the setting (e.g. outpatient vs. inpatient clinics , district hospital vs. teaching hospital and between EU member states)”(EMA, 2011). For example, the competent authority in Finland considers the inclusion of the quality of life questionnaire not as a current practice and considers studies including them as interventional, Netherland also considers very long questionnaires as not a routine practice (Ramirez, 2015). Therefore, there is a need for a harmonized definition of NIS. According to ENCePP, a diagnostic, monitoring or therapeutic procedure can be considered as current practice if at least one of the following criteria is fulfilled (EMA, 2011):

- Routinely performed by a proportion of healthcare professionals
- Performed according to evidence-based medicine criteria
- Defined in guidelines issued by a relevant medical body
- Mandated by regulatory and/or medical authorities
- Reimbursed by the national or private health insurance

Table 1. Decision tree to establish whether a study is non-interventional (Adapted from the regulation No. 536/ 2014 draft Question and Answers version 3) (European Parliament, 2021)

	A	B	C	D	E	F
A clinical trial of a medicinal product?					A non-interventional study?	A low intervention clinical trial?
Is a medicinal product administered before or during the start of the clinical trial	Is it a medicinal product (i)?	Is it not a medicinal product?	What effects of the medicine are you looking for?	Why are you looking for those effects?	How are you looking for these effects?	Is the product authorized in any EU Member State?
If a medicinal product is administered before the start of the clinical trial, and it falls under current practice (vii), please go to column E. If a medicinal product is administered before the start of the clinical trial and it falls not under current practice, column E is excluded.	If you answer no to all the questions in column A, the activity is not a clinical trial on a medicinal product.	If you answer yes to the question below in column B the activity is not a clinical trial on a medicinal product.	If you answer no to all the questions in column C the activity is not a clinical trial under the scope of Regulation EU No 536/2014.	If you answer no to all the questions in column D the activity is not a clinical trial under the scope of Regulation EU No 536/2014.		
If a medicinal product is administered after the start of the clinical trial, please go to column A.	If you answer yes to any of the questions below go to column B.	If you answer no to this question below go to column C	If you answer yes to any of the questions below go to column D.	If you answer yes to any of the questions below go to column E.		

	A.1. (ii) Is it a substance or combination of substances presented as having properties for treating or preventing disease in human beings? A.2. Does the substance function as a medicine? i.e., can it be administered to human beings either with a view to restoring, correcting, or modifying physiological functions by exerting a pharmacological, immunological, or metabolic action or to making a medical diagnosis or is otherwise administered for a medicinal purpose? A.3.Is it an active substance in a pharmaceutical form?	B.1.(iii) Are you only administering any of the following substances? - human whole blood[1] - Human blood cells -Human plasma -A food product (iv) (including dietary supplements) not presented as medicine. -A cosmetic product (v) - A medical device	C.1. To discover or verify/compare its clinical effects? C.2. To discover or verify/compare its pharmacological effects, e.g., pharmacodynamics? C.3. To identify or verify/compare its adverse reactions? C.4. To study or verify/compare its pharmacokinetics, e.g., absorption, distribution, metabolism, or excretion?	D.1. To ascertain or verify/compare the efficacy (vi) of the medicine? D.2. To ascertain or verify/compare the safety of the medicine?		

(i) Cf. Article 1(2) of Directive 2001/83/EC, as amended

(ii) Substance is any matter irrespective of origin e.g. human, animal, vegetable or chemical that is being administered to a human being.

(iii) This does not include derivatives of human whole blood, human blood cells and human plasma that involve a manufacturing process.

(iv) Any ingested product which is not a medicine is regarded as a food. A food is unlikely to be classified as a medicine unless it contains one or more ingredients generally regarded as medicinal and indicative of a medicinal purpose.

(v) The Cosmetic Directive 76/768/EC, as amended harmonises the requirements for cosmetics in the European Community. A "cosmetic product "means any substance or preparation intended for placing in contact with the various external parts of the human body (epidermis, hair system, nails, lips and external genital organs) or with the teeth and mucous membranes of the oral cavity with the view exclusively or principally to cleaning them, perfuming them or protecting them in order to keep them in good condition, change their appearance or correct body odours.

(vi) Efficacy is the concept of demonstrating scientifically whether and to what extent a medicine is capable of diagnosing, preventing or treating a disease and derives from EU pharmaceutical legislation.

(vii) Assignment of patients to a treatment group by randomisation planned by a clinical trial protocol cannot be considered as current practice

1.3.2. Study types and applications

NIS are one of the useful tools to collect RWD. NIS may be retrospective with the secondary use of existing data or prospective with primary collection of data.

ENCePP gives examples of retrospective NIS (EMA, 2011):

- purely observational review/research of a database
- retrospective review of records, where all events of interest have happened e.g., case-control, cross-sectional and purely retrospective cohort studies
- studies in which the prescriber later become investigator, but the prescription has already occurred, e.g., retrospective data collection from individual medical records at the site of investigator

ENCePP gives examples of prospective NIS (EMA, 2011):

- registries in which the data collected derive from routine clinical care
- studies which evaluate patterns of the usage of medicines
- drug utilization studies including potential off-label use
- measuring the effectiveness of risk management measures
- measuring effectiveness of therapeutic interventions in current practice
- health outcome assessments
- long-term extension studies with patient follow up beyond trial protocol specified time for observation and active collection of additional data – such as death or event free survival

Another application of NIS is in PAS, e.g., in the form of post-authorization safety study (PASS). DIR 2001/83/EC Art 1(15) defines PASS as "any study relating to an authorized medicinal product conducted with the aim of identifying, characterizing or quantifying a safety hazard, confirming the safety profile of the medicinal product, or of measuring the effectiveness of risk management measures" (EMA, 2017). The main purpose of PASS is to evaluate the safety and benefit-risk profile of a medicine and support regulatory decision-making (EMA, 2017).

Marketing Authorization Holder (MAH) can initiate, manage or finance a PAS voluntarily or pursuant to a request by a National Competent Authority (NCA) in member state or EMA according to DIR 2001/83/EC or REG No 726/2004/EC. Voluntary PASS studies are normally part of the risk management activities. A PASS may be interventional or non-interventional. According to GVP Module VIII, for a PASS to be considered as NIS the requirement mentioned under the definition of NIS should be cumulatively fulfilled (EMA, 2017).

1.3.3. Regulatory framework in Europe

Current clinical trials in the EU are conducted according to DIR 2001/20/EC. REG No.536/2014, which is in force since 16 June 2014, will replace this directive in the future. It is announced that the new regulation will be applicable six month after the European Commission publishes the confirmation of the full functionality of the Clinical Trial Information System (CTIS) (EMA, no date). CTIS contains the centralized EU portal and databank for clinical trials, which provides a central clinical trial application submission to NCA and Ethics Committee (EC), and will harmonize the clinical trial process and requirements (EMA, no date). However, NIS are out of scope of DIR 2001/20/EC and the up-coming regulation with the justification that they have lowest risk for the study participant. However, as in clinical trials, these studies should be conducted in accordance with the 1964 Helsinki declaration and its later amendments. Moreover, the Guidelines for Good Pharmacoepidemiology Practice (GPP) should be followed for such studies (ISoP, 2016). Currently there is no harmonized regulation or guidance for submission and conduction of NIS in Europe and each member state has its own regulations and requirements. As NIS include also post-marketing studies, pharmaceutical association have also published a code of conduct for such studies to control their scientific purpose and provide guidance for their conduction (Ramirez, 2015). Therefore, to ensure the patient safety and to generate high quality data, the national laws and local regulations, guidelines and code of conduct must be followed while planning and conducting NIS.

An approval from a NCA is not required for a NIS in most member states, with the exception of Finland and Denmark (Ramirez, 2015). However, a favorable opinion of an EC is required; but there are exceptions, e.g., in Denmark and Austria, and in France "Advisory Committee on Research Information Processing" is responsible for ethical

evaluation of NIS (Ramirez, 2015). There is no central EC responsible for the ethical review of the NIS protocol. NIS are currently approved by local and sometimes regional ECs each requiring different documents, which creates an administrative burden (de Lange et al., 2019, Ramirez, 2015). Sometimes the approval or notification of the data protection authorities are also required for NIS (Ramirez, 2015). Currently the REG 2016/679 on personal data protection is in force in Europe and must be followed (European Parliament, 2016). There is no need for clinical trial insurance to cover NIS (exception: Belgium) (Ramirez, 2015).

There is no central submission procedure or registration obligation for NIS in an EU database, except for non-interventional PASS that are imposed as an obligation by a NCA or the EMA. It is not mandatory to register a voluntary non-interventional PASS, but as registration will support the same level of transparency, scientific and quality standards in all NIS, it is recommended in GVP Module VIII to register them as well (EMA, 2017). The EU PAS (EU PAS) register of EMA is a free, publicly available electronic database to register non-interventional PASS. According to a webinar conducted by the EMA, in total 2040 NIS are registered in EU PAS register to date of 5 March 2021(EMA, 2021). The EMA's Pharmacovigilance Risk Assessment Committee (PRAC) is responsible to evaluate their study protocol and their results. The EMA provides a guidance document for the protocol and final study report of a non-interventional PASS (EMA, 2012). DIR 2010/84/EU on pharmacovigilance and safety reporting (Article 107) requires EMA "to make public the protocols and abstracts of results of imposed non-interventional PASS public" (European Parliament, 2010). Chapter 4 of the DIR 2010/84/EU provides further details about conduct of such studies. The Commission Implementing REG (EU) No 520/2012 provides the format for the study protocol, abstracts and final reports in Annex III (European Parliament, 2012). Furthermore, date of study registration in the EU PAS register is mentioned here also as one of the milestones to be mentioned in the final study report of imposed non-interventional PASS (European Parliament, 2012).

1.3.4. Advantages, limitations, and challenges

NIS have considerable external validity as they collect data from large number of participants from normal treatment settings, which allows studying the patient populations that were excluded from RCTs (e.g., pregnant women, children, elderly and patients with

comorbidities and co-medications). This allows the subgroup analysis and multi-variate statistical approaches to explore the role of factors associated with efficacy and safety and provide useful information about the effectiveness and safety of the marketed medicinal products (Michel et al., 1998, Michel et al., 2000, Michel et al., 2007). Moreover, compares to large RCTs, NISs are easier and less expensive to perform (no clinical trial insurance is needed).

On the other hand, NIS have some methodological challenges. First, as with any big data, there are concerns about data quality and hidden biases. Different statistical approaches might be attempted to address the missing data or to adjust for the cofounding factors, however some residual confounding is likely to remain (Zeng et al., 2019). Second, lack of randomization, blinding and a control group limit their internal validity. Therefore, it is difficult to ensure the representativeness of the study population (selection bias), although this has to be weighed against the selection bias by in- and exclusion criteria applied in RCT. Some methodological changes may increase the internal validity of a NIS. For example, initial attempts have been made to include control groups (Michel et al., 2013), randomization-like elements (Schneider et al., 2014) and data source verification (Michel et al., 2004) into the design of prospective NIS. For retrospective NIS, randomization can be practiced by limiting the inclusion and analysis to new users of the medicine, the so-called new/incident user design. An incident user design follows the patients from the day that they have started an intervention, which may reduce chronology biases but with the tradeoff of reducing the study precision (Johnson et al., 2013, Ray, 2003). Third, given to their complexity and heterogeneity in regulation, conducting a global NIS is challenging if not impossible (Claudot et al., 2009, Ramirez, 2015). Finally, NIS are mostly initiated, managed, and financed by commercial sponsors such as the MAH, but recently non-commercial sponsors such as academic institutions also show interest in conducting NIS. However, the limited available regulatory guidance on NIS is applicable only to industry sponsored studies, which makes the conduct of such studies more challenging for academic institutes.

The pharmaceutical industry has been accused of abusing NIS as a marketing tool as many NIS lacked a valid scientific question (Gale, 2012, Morton et al., 2016, von Jeinsen and Sudhop, 2013). A recent cohort study showed that physician's participation in a NIS

sponsored by pharmaceutical, is associated with a more frequent prescription of the investigated medicinal product by them (Koch et al., 2020). A survey of the notifications sent to regulatory agencies found that post-marketing NIS do not improve the drug safety surveillance as the sample sizes are too small to detect rare adverse drug reactions and the confidentiality clauses affect the adverse event reporting of the participating physicians (Spelsberg et al., 2017). However, such comments reflect on the use of the tool NIS but not necessarily on what can be achieved using it. Moreover, a lack of power to detect rare adverse drug reactions does not necessarily limit their usefulness for the improvement of drug safety because NIS can provide evidence on the propensity of risk groups to exhibit known and more frequent adverse events, e.g., those with certain co-medications or co-morbidities (Michel et al., 1998).

1.4. Statistical reporting in evidence-based medicine

The next step after data collection is to present and interpret them and finally make decisions based on study findings (Smith, 2020). Statistical analysis and proper presentation of statistical results is a crucial part in both RCT and NIS. Even with high quality data, poor statistical analysis approaches, result in incorrect conclusions about the study results. False positive results in RCT results in entry of a product into the market that does not provide benefit to the patient, and false negative results may prevent a potentially beneficial product to come into the market. The same can be expected from false results of a NIS, a false positive about the safety of medicinal product can for example lead to market withdrawal, and false negative may impose life-threatening risks to patients. The Network for Enhancing the Quality and Transparency of Health Research (EQUATOR) provides reporting guidelines for different study types in evidence-based medicine. These guidelines "provide checklists, flow diagram or explicit text to guide authors in reporting a specific research developed using explicit methodology"(EQUATOR, No date). Consolidated Standards of Reporting Trials (CONSORT) and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) are two key guidelines for reporting the findings from RCTs and observational studies, respectively. These guidelines can also be endorsed by different journals and assist the peer reviewers and editors in their decision-makings. However, recommendations in guidelines, are merely

recommendation and not rules (Diong et al., 2018). Therefore, they guarantee neither that all authors will adhere to them, nor that all authors will agree with them.

Data analysis has become faster and easier by development and growing popularity of statistics software such as – R, SAS, and SPSS; however, many researchers still misunderstand and misinterpret some basic statistical concepts (Diong et al., 2018, Smith, 2020). Hirst and Altman state that "poor reporting of research studies hindering its utilisation in clinical practice and further research plagues the medical literature. This is unethical, wasteful of scarce resources and event potentially harmful" (Hirst and Altman, 2012). Sometimes the study findings on the same question do not align with each other, this has been observed in different study types ranging from clinical trials to epidemiological studies (Ioannidis, 2005). The underlying reason could be some basic statistical errors in the presentation or interpretation of the results (Lang, 2004, Ioannidis, 2005). Descriptive statistics and inferential statistics are the main part of the statistical analysis. "Descriptive statistic is the science of summarizing or describing the data, while inferential statistic is the science of interpreting the data in order to make estimates, hypothesis testing, predictions, or decisions from sample to the targeted population"(Shein-Chung Chow; Jen-Pei Liu, 2013). Demographics, baseline and treatment outcome parameters are usually summarized by using simple sample descriptive statistics such as mean, median and mode and dispersion of data like standard deviation, range or interquartile range (Shein-Chung Chow; Jen-Pei Liu, 2013). Reporting mean with standard deviation is intuitive to many, but it assumes that the data is normally distributed, which further impacts on the choice of the statistical test. Therefore, it is important to determine if the data is normally distributed, before presenting them.

The next step after presenting the data is to analyze and interpret them. A difference might be observed by comparing the descriptive statistics of the baseline and treatment outcome parameters. However, the observed difference might be due to chance alone. Therefore, before drawing conclusions about the observed differences (effects), we need to ascertain that it is not due to chance (Shein-Chung Chow; Jen-Pei Liu, 2013). The frequentist approaches of the data analysis involve either testing a pre-specified hypothesis or making an estimate of effect size about an intervention in the study population and then infer the results to the targeted population with the same characteristics, the so-called

statistical inference. The rejection or failure to reject a null hypothesis is the first thing, and sometimes the only thing, addressed in many statistics textbooks and courses under the rubric of „statistical inference". Such inferences are often drawn without explicit reference to the value of the test statistic, the chosen alpha level, or the p-value. Criticism of dichotomizing the data with p-value is old and there are warnings against decision-making based on categorized p-values (Boring, 1919). Moreover, in RCTs or NIS with large sample size , even small effects of questionable clinical relevance might be statistically significant, and therefore it is necessary to see if the results are clinically meaningful (Ioannidis, 2005).

2. Objectives and methods

This dissertation presents the results of three observational NIS and two systemic reviews with three main objectives. **First**, to understand the value of NIS to complement the evidence from RCT to improve the effectiveness of the treatment when considering their strengths and limitation. **Second**, to evaluate the potential use of the routinely collected inpatient data in the context of the NIS study design to support the drug safety surveillance. **Third**, to elaborate on two basic statistical reporting issues as an implication for analysis and reporting of research in evidence-based medicine. To address the overall objectives of this work, three NIS studies were used as examples and we conducted two systemic reviews of publications in PubMed database (https://pubmed.ncbi.nlm.nih.gov/). Moreover, the discussion section of each of the three NIS explicitly discusses the limitations related to the interpretation of the data being analyzed.

Among the included NIS, two of them were observational studies with prospective data collection conducted as part of the obligation of pharmaceutical companies in Germany for ongoing monitoring of the safety and tolerability of propiverine. Propiverine is an antimuscarinic drug used in adults who have Overactive Bladder syndrome (OAB). The symptoms of OAB are the urgency to urinate, frequent urination or being unable to hold urine. Two doses of propiverine extended release (30 and 45 mg/d) are available for treatment of OAB symptoms. This allows adjusting the dose to obtain the optimal therapeutic effect and to control the adverse effects. According to its prescribing information, the recommended dose for the treatment of OAB is 30 mg once daily. If the dose is well tolerated and clinical effect is not improved sufficiently, the dose may be increased to 45 mg. This dosing recommendation is based on the results of the clinical studies and dose adjustment might be needed for especial subgroup of patients. As some patients were excluded from clinical trials, dosing information about some patient groups are missing e.g., patients with moderate to severe renal or hepatic impairment. Moreover, prediction of the optimal dose for individual patient is difficult and is influenced by factors like age, genotype, comorbidity, co-medications, etc. To our knowledge, no RCT studied which factors are associated with the initial dosing decision. Moreover, it is possible that factors associated with the initial dosing decision also impact the decision to increase the dose and, thereby, the overall treatment outcome. The two NIS used for this research purpose

had similar designs. Study I was primarily designed to explore the effect of different starting dose and dose adjustment after 4 weeks with a treatment duration of 12 weeks. Study II was designed to explore the effect of additional material (information sheet about OAB and mode of action of the drug) on efficacy and tolerability and on premature discontinuation during a treatment period of 12 weeks and allowed extension of observation for up to 24 weeks in a subgroup. We evaluated the potential of these two NIS to complement former RCTs. To this end, we performed a post-hoc analysis of their datasets, using multivariate analysis approach to determine the factors associated with the initial dosing and up-titration decision of propiverine for the treatment of OAB symptoms. Furthermore, we analyzed how dosing relative to other factors affects treatment outcome.

The third NIS included in the context of this work was an observational study with retrospective data collection, funded by the Federal Ministry of Education and Research in Germany and conducted by academic investigators. In this project a secondary analysis of the routinely collected in-patient data sampled from four non-academic hospitals in Germany were performed and the recorded in-patient diagnosis were evaluated regarding causal relationship to any drug and whether the event was preventable. Here we examined the potential use of the routinely collected inpatient data in the context of NIS study design to complement the existing drug surveillance systems.

Lastly, this thesis covered two basic issues in statistical reporting that might influence the validity of the evidence from studies in medical science. One of the concepts that we considered is the choice in measure of center (mean, median, and mode) and dispersion of data (standard deviation, range, or interquartile range). To this end, we used parts of the two NIS datasets on propiverine to explore the presence or absence of normal distribution of OAB parameters urgency, incontinence, frequency and nocturia and treatment-associated changes thereof. Furthermore, we performed a systemic review of original studies reporting on at least one OAB symptoms published in four leading urology journals in 2016-2017. The second concept that we elaborated on is the reliance on p-value and preference of authors for statistical testing rather than estimating. For this purpose, we conducted a systemic review of the abstracts of the publications in three influential clinical pharmacology journals using a text-mining algorithm and estimated the time trend in the prevalence of reporting statistical inferences.

The remainder of this thesis reads as follows. Third section is the cumulative part of the dissertation, which contain the publications under this thesis with further details about the background, materials, methods, and the results of each project. The fourth section is the discussion part of the dissertation, in which the overall findings of the thesis are discussed. In the outlook section, the future perspectives of NIS are given and finally the thesis ends with a summary.

3. Cumulative part of the dissertation

3.1. Do overactive bladder symptoms and their treatment-associated changes exhibit a normal distribution? Implications for analysis and reporting.

Received: 15 October 2019 | Accepted: 30 December 2019

DOI: 10.1002/nau.24275

ORIGINAL CLINICAL ARTICLE

Neurourology Urodynamics ICS SUFU WILEY

Do overactive bladder symptoms and their treatment-associated changes exhibit a normal distribution? Implications for analysis and reporting

Marjan Amiri[1,2] | **Sandra Murgas**[3] | **Andreas Stang**[1,4] | **Martin C. Michel**[5]

[1]Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, Essen, Germany

[2]Center for Clinical Trials Essen (ZKSE), University Hospital Essen, Essen, Germany

[3]Apogepha Arzneimittel GmbH, Dresden, Germany

[4]Department of Epidemiology, School of Public Health, Boston University, Boston, Massachusetts

[5]Department of Pharmacology, Johannes Gutenberg University, Mainz, Germany

Correspondence
Martin C. Michel, Department of Pharmacology, Johannes Gutenberg University, Langenbeckstr. 1, Geb. 708, 1. OG, 55131 Mainz, Germany.
Email: marmiche@uni-mainz.de

Funding information
Apogepha; Innovative Medicines Initiative 2 Joint Undertaking, Grant/Award Number: 777364

Abstract

Aims: To explore the use of means vs medians (assuming or not the presence of normal distribution) in studies reporting overactive bladder syndrome symptoms and to test for normal distribution of basal values and treatment-associated changes thereof in two large noninterventional studies.

Methods: Systematic review of all original studies reporting on at least one overactive bladder syndrome symptom published in four leading urology journals in 2016 to 2017. Testing of the normal distribution of urgency, incontinence, frequency, and nocturia in two large noninterventional studies (n = 1335 and 745).

Results: Among 48 eligible articles, 86% reported means (assuming a normal distribution), 6% medians (not making this assumption), and 8% a combination thereof. Baseline values for all four symptoms and treatment-associated alterations thereof deviated from a normal distribution ($P < .0001$ in all cases). Means overestimated basal value and absolute changes thereof as compared with medians, for example, basal number of incontinence episodes in study 1 5.1 vs 4. Differences between means and medians for percentage changes of symptoms were small and did not consistently favor means over medians.

Conclusions: Dominant reporting of means implies the assumption of a normal distribution of overactive bladder syndrome symptoms but our data from two noninterventional studies do not support this assumption. We recommend that basal values and absolute symptom changes should be reported as medians and subjected to nonparametric analysis; means may be appropriate for the reporting of percentage changes of symptoms.

KEYWORDS
data analysis, normal distribution, overactive bladder syndrome, propiverine

1 | INTRODUCTION

The overactive bladder syndrome (OAB) has a high prevalence[1] and adversely impacts on the well-being of

Abbreviations: IQR, interquartile range; NIS, noninterventional study; OAB, overactive bladder syndrome; RCT, randomized controlled trial; SD, standard deviation.

the afflicted patients[2] and their family members.[3] Studies on OAB are mostly based on quantification of the four key symptoms of urgency, incontinence, frequency, and nocturia,[4] which typically are captured from voiding diaries. When looking at published OAB studies, we noticed that some investigators report baseline symptoms and treatment-associated changes thereof as means, whereas others report them as medians. The use of means appears intuitive to many, but it assumes that the data come from a population exhibiting a normal (Gaussian) distribution. A normal distribution is characterized by a unimodal, symmetrical distribution with ~68% of all values falling within ± 1 standard deviation (SD). In contrast, medians provide a useful description of the central tendency of a unimodal distribution that is not normal. For distributions that have more than one peak (multimodal distributions), both measures of central tendency are inappropriate. Whether a normal distribution exists also has implications on the choice of reported error bars and statistical tests. The SD and parametric tests such as *t* tests are only appropriate if a normal distribution can be assumed.

Against this background, we have systematically extracted information from four major urology journals reporting on OAB studies to explore whether they reported means or medians of OAB symptoms and whether they provided a justification for this choice. Moreover, we have used data from two large noninterventional studies (NIS) in OAB treatment to determine whether it is justified to assume a normal distribution. We have compared means and medians to determine whether using one yields a systematic over- or underestimation as compared with the other. Such calculations were made for the four OAB symptoms of urgency, incontinence, frequency, and nocturia and for treatment-associated alterations thereof; potential differences in reporting between studies with and without industry involvement was a secondary exploratory aim of our analysis.

2 | PATIENTS AND METHODS

2.1 | Present reporting practice in peer-reviewed publications

We conducted a systematic review of the original studies published in 2016 and 2017 in four major urology journals (BJU International, European Urology, Journal of Urology, Neurourology, and Urodynamics) that reported on at least one of the OAB symptoms of urgency, incontinence, frequency, and nocturia. We used the PubMed database (www.ncbi.nlm.nih.gov/pubmed) to identify relevant studies. In our PubMed search performed in May and June 2018, we entered the "name of journal" and "overactive bladder" in the search field and applied the date filter 2016.01.01-2017.12.31 for either print or online publication. To minimize selection bias, we included all original studies written in English and reporting on at least one of the four OAB symptoms within this period, including those that did not study OAB. We included randomized and observational studies investigating a medical or a surgical intervention including implanting devices as well as studies comparing groups of patients without studying an intervention. Our systemic review excluded preclinical studies, reviews, editorials, and letters. However, we did not use any filter regarding the sample size of the reported study. MA extracted the studies' primary and secondary endpoints and their respective measurement methods from the full text of each article. Moreover, she checked whether the authors provided a justification for their choice of reporting means or medians. In addition, she extracted the authors' affiliations and study funding sources to explore whether a study had involved industry. Individual studies considered in our analysis and a PRISMA flow chart are listed in the Online Supporting Information.

2.2 | Clinical data from two NIS

We used data from two NIS of a similar design and performed in 2012 and 2014 for a post hoc analysis. They included 1335 and 745 patients, respectively, and hereafter named study 1 and 2. Both studies had been performed with approval from the ethical committee of the state board of physicians in Saxony, Germany (Sächsische Landesärztekammer EK-BR-14/12-1 and EK-BR-18/14-1). While the analyses occurred after the studies had been completed, the statistical analysis plan had been prespecified before any analysis related to the normality of data distribution. Both NIS asked participating physicians to document baseline data and treatment outcomes for patients receiving propiverine ER based on the physician's judgment to treat their OAB symptoms. The planned duration of observation was 12 weeks with planned visits at baseline and after ~4 and 12 weeks. According to the applicable prescribing information, the starting dose could be 30 mg or 45 mg once daily and could be adapted during the duration of the studies. Data were collected on standardized case record forms. Based on study protocol, OAB symptom intensity assessment was based on voiding diaries but in line with the noninterventional character of the study, the length of observation period per assessment period in the diary was left to the discretion of the physician.

Our analysis of the baseline data was based on all patients having an entry for a given parameter. Patients not exhibiting a given symptom at baseline (value of 0) and those with medically implausible values (urgency > 50, frequency > 40,

nocturia > 20, and incontinence > 30) were excluded from the analysis for that symptom; this affected four patients each for urgency and frequency, one for nocturia, and one for incontinence in study 1 and none in study 2. Our analysis of the treatment data was based on a subgroup of the baseline cohort: to minimize heterogeneity based on dosing decision and duration of follow-up, this included only patients with a starting dose of 30 mg, having a recorded value at the 12-week time point and no change of administered dose during the observation period. We chose this group because it represents the majority of patients. We evaluated treatment-associated reductions in symptom frequency effects as delta (baseline – 12-week value) and as percentage reduction; for mathematical reasons, we calculated the latter only for patients with a baseline value other than 0 for the respective parameter. Missing data for one parameter did not exclude the use of other parameters from the same patient.

Based upon a reviewer's suggestion, we have performed a post hoc analysis to determine whether professional statisticians had been involved in the analysis of the data of the published papers. For this purpose, we checked the published manuscripts whether any author listed an affiliation to a statistics, biostatistics, or epidemiology department. Furthermore, we contacted each corresponding author to ask for the involvement of a professional statistician. If either was positive, we assumed the involvement of a statistician. Moreover, we asked corresponding authors whether to their knowledge a professional statistician had been involved as part of the manuscript evaluation by the journal. For either assessment, we compared numbers of articles reporting means, medians, or a combination thereof in an exploratory manner.

We tested for normal distribution using the D'Agostino and Pearson K2 omnibus test. To assess the impact of a lack of normal distribution, we compared means and medians. We performed all data analysis using the Prism software (version 8.2.1; GraphPad, La Jolla, CA). As we report only on a subset of both clinical studies, full data including those on tolerability will be presented in a subsequent report.

3 | RESULTS

3.1 | Present reporting practice in peer-reviewed publications

We retrieved and reviewed 183 articles in total, 16, 27, 39, and 101 articles from BJU International, European Urology, Journal of Urology, Neurourology, and Urodynamics, respectively (see Online Supporting Information). Forty-eight papers were eligible for inclusion in our analysis. Most of the articles (86%) reported means, only a few medians (6%) and some both means and medians (8%). All studies with industry involvement reported means; among those without industry involvement, corresponding numbers were 75% means, 11% medians, and 14% a combination thereof. The articles reporting means typically showed SD, standard error, or confidence interval error bars, which were internally consistent in their assumptions of a normal distribution. In contrast, those reporting medians showed error bars as interquartile ranges (IQR), which were also internally consistent in not assuming a normal distribution.

Only four studies (8%) provided information on testing for normal distribution within their data set: two studies reported having used the Shapiro-Wilk's test[5,6] and one the Kolmogorov-Smirnov test and a distribution histogram[7]; they did not disclose the results of normality testing but presented data as medians. One study showed a histogram for end-of-treatment values that clearly showed a lack of normal distribution but nonetheless reported means.[8] Of note, some of these studies included only 58[6] and 132 subjects,[7] indicating that they were probably too small to allow robust conclusions on the normal distribution. Four other studies (8%) made statements on normality but did not mention on which analysis this was based. One of them claimed to have applied parametric and nonparametric tests to parameters with and without normal distribution, respectively but did not disclose which applied to which parameter and concomitantly reported means and medians for all OAB symptoms.[9] Three others claimed normal distribution without showing supporting data and reported mean values.[10–12] The studies providing some justification for the choice of means vs medians reported both baseline symptoms and treatment responses but did not differentiate their assumptions related to a normal distribution for the two assessments.

In a post hoc analysis, we explored the impact of the involvement of a professional statistician in the data analysis on reporting. Twenty-one of 26 papers with available information had involved a statistician; they reported means in 16, medians in three and a combination thereof in two cases. Five of 26 papers reported not having involved a statistician; they all reported means. Of the 22 papers without information about the involvement of a statistician, 20 reported means and two a combination of means and medians. Authors of eight papers reported that the referee comments included specific feedback on statistical analysis; the published papers reported means in six and a combination of means and medians in two cases. Responding authors of four papers stated that no specific statistical review was provided; their published papers all reported means.

TABLE 1 Baseline severity of OAB symptoms (episodes per 24 hours)

	Study 1			Study 2		
Symptom	**n**	**Mean ± SD**	**Median (IQR)**	**n**	**Mean ± SD**	**Median (IQR)**
Urgency	1151	10.5 ± 5.9	10 (6; 14)	621	10.0 ± 5.5	10 (6; 13)
Incontinence	785	5.1 ± 3.9	4 (2; 7)	418	5.5 ± 3.9	5 (2; 7)
Frequency	1308	13.6 ± 4.4	13 (11; 16)	730	13.2 ± 4.2	13 (10; 15)
Nocturia	1269	3.4 ± 1.6	3 (2; 4)	706	3.5 ± 1.7	3 (2; 4)

Note: Patients not exhibiting urgency, incontinence, or nocturia were excluded for that parameter. All four parameters differed from a normal distribution in the D'Agostino and Pearson K2 omnibus test at $P < .0001$.
Abbreviations: IQR, interquartile range; OAB, overactive bladder syndrome.

3.2 | Clinical data from two NIS

The distribution of symptom intensity (episode frequency) differed significantly from normality for all four OAB symptoms in both studies ($P < .0001$ for all parameters; Table 1). As an example, a graphical representation of the distribution of intensity based on frequency data is shown as Figure 1 indicating that the data exhibited a unimodal, but not symmetrical distribution. Accordingly, means were systematically higher than the corresponding medians in both studies. The means as surrogate values for medians overestimated the medians in the two data sets by 0.5 and 0.0 for urgency, 1.1 and 0.5 for incontinence, 0.6 and 0.2 for frequency, and 0.4 and 0.5 for nocturia.

The distribution of treatment responses expressed as absolute changes of episode frequency differed significantly from normality in both NIS ($P < .0001$ for all parameters; Table 2). Like baseline values (see above), means as surrogate measures for the more appropriate medians of absolute differences systematically overestimated treatment effects on urgency (1.1 and 0.5), incontinence (0.7 and 0.9), and frequency (0.8 and 0.8) in both studies; however, there was no overestimation of changes in nocturia episodes (0 and −0.3).

The distribution of treatment responses expressed as the relative difference in episode frequency (percentage of week 12 related to baseline measurement) also differed significantly from normality in both NIS ($P < .0001$ for all parameters; Table 3). However, differences were small and, if anything, means underestimated symptom changes relative to medians for all four symptoms.

4 | DISCUSSION

4.1 | Critique of methods

Our systematic review was based on four leading journals in the OAB field. These journals were chosen assuming that leading journals may have published papers of the highest applicable standards. To minimize selection bias, the chosen years of publication were systematically screened for each journal.

Our analysis of clinical data was based on two NIS, not on randomized controlled trials (RCT). This was done because NIS tend to be larger than RCT, and a robust

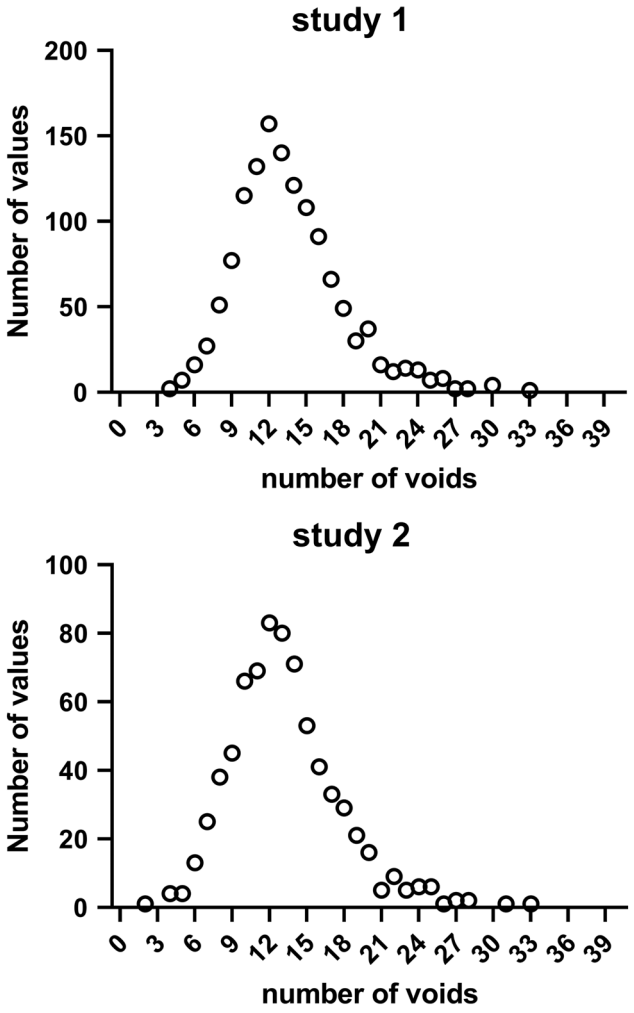


FIGURE 1 Distribution of basal micturition frequency (number of voids/24 hours) in studies 1 and 2. The highest observed values in both studies were 33

TABLE 2 Absolute reductions of OAB symptoms (delta of episodes per 24 hour) after 12 weeks of treatment in the subgroup continuously receiving 30 mg/day

	Study 1			Study 2		
Symptom	**n**	**Mean ± SD**	**Median (IQR)**	**N**	**Mean ± SD**	**Median (IQR)**
Urgency	627	7.1 ± 5.2	6 (3; 10)	335	6.5 ± 4.6	6 (3; 9)
Incontinence	414	3.7 ± 3.0	3 (2; 5)	218	3.9 ± 3.3	3 (1; 5)
Frequency	740	5.8 ± 3.7	5 (3; 7)	415	4.8 ± 3.5	4 (3; 6)
Nocturia	727	2.0 ± 1.9	2 (1; 3)	401	1.7 ± 1.5	2 (1; 2)

Note: All four parameters differed from a normal distribution in the D'Agostino and Pearson K2 omnibus test at $P < .0001$.
Abbreviations: IQR, interquartile range; OAB, overactive bladder syndrome.

assessment of normality requires a large sample to be analyzed.

The data sets being analyzed had been collected as part of the obligation of pharmaceutical companies in Germany to provide ongoing collection on the safety and tolerability of a medicine (German Arzneimittelgesetz). While our data represent post hoc analyses, they were based on a statistical analysis plan that had been finalized before data were inspected related to the goals of our investigation to avoid bias during the analysis process. Several statistical tests are available to test for deviation from a normal distribution. We had selected the D'Agostino and Pearson K2 omnibus test as a primary outcome measure. However, three other tests for normality (Anderson-Darling, Shapiro-Wilk, and Kolmogorov-Smirnov) consistently confirmed the deviation from normality for each parameter in each study (data not shown).

Previous NIS on the use of muscarinic receptor antagonists in the treatment of OAB have typically reported mean values, for instance, based on treatment with darifenacin,[13] solifenacin,[14] and tolterodine.[15,16] Mean baseline intensity of symptoms and treatment-associated improvements thereof in the previous four and in the present two studies were comparable, indicating that we have used data sets that are representative for the overall population of patients with OAB seeking medical treatment in a real-life setting. This is in line with the general observation that all muscarinic receptor antagonists have comparable efficacy.[17,18] However, it was greater than observed in most RCT,[17,18] at least in part because RCT typically includes a single-blind placebo run-in period before establishing baseline symptoms. In conclusion, our clinical analyses are based on data sets comparable with those of many other NIS and, therefore, our clinical findings may be generalizable.

4.2 | Present reporting practice in peer-reviewed publications

Our results show that reporting of distributions of symptoms or changes of symptoms related to OAB are statistically inconsequently and therefore potentially misleadingly described. Most authors used means, implicitly assuming a normal distribution of OAB symptoms. However, most of the authors reporting means did not provide any justification for their assumption of a normal distribution. With one exception,[8] the few articles providing a justification for assuming a normal distribution did not reference or show the data supporting it. Although it can be assumed that pharmaceutical companies sponsoring a clinical trial have professional statisticians on staff, industry-sponsored studies consistently used means and failed to

TABLE 3 Relative reduction of OAB symptoms (percentage reduction of episodes per 24 hour) after 12 weeks of treatment in the subgroup continuously receiving 30 mg/day

	Study 1			Study 2		
Symptom	**n**	**Mean ± SD**	**Median (IQR)**	**n**	**Mean ± SD**	**Median (IQR)**
Urgency	627	71 ± 29	75 (56; 92)	335	65 ± 30	67 (50; 90)
Incontinence	414	82 ± 30	100 (67; 100)	218	70 ± 39	78 (50; 100)
Frequency	740	41 ± 17	42 (30; 50)	415	36 ± 20	36 (25; 46)
Nocturia	727	59 ± 29	60 (50; 75)	401	48 ± 39	50 (33; 67)

Note: All four parameters differed from a normal distribution in the D'Agostino and Pearson K2 omnibus test at $P < .0001$.
Abbreviations: IQR, interquartile range; OAB, overactive bladder syndrome.

provide a justification for this. The involvement of a professional statistician was associated with the choice of measures of central tendency (mean or median): while medians were reported only when a statistician was involved, studies with such involvement in most cases also reported means. When a statistical evaluation had apparently been part of the peer review process of a manuscript, the proportion of publications containing medians was somewhat higher although the majority of reports still were limited to means. Based on these post hoc analyses, it appears that the involvement of a statistician in the data analysis and/or peer review of the manuscript made it more likely that medians were reported but even in those cases means were more common.

The heterogeneity in reporting of OAB parameters and the lack of providing data underlying it (except one of 48 studies) shows that an analysis to determine the validity of the assumption of a normal distribution is necessary. For variables that can only provide non-negative values, a simple heuristic makes clear that reporting means and SD is inappropriate: if the mean minus 1 to 2 SD's predicts negative values for the 95% range of the data, the assumption of normality cannot be correct as the descriptive statistic predicts values that are impossible (eg, present study 1: urgency: mean 10.7 and SD 6.6).

4.3 | Clinical data from two NIS

When a given symptom at baseline does not exhibit a normal distribution, the absolute (after minus before treatment) or relative (%) difference of symptoms can still exhibit a normal distribution. Our systematic literature review revealed only one study that showed a non-normal distribution of the treatment responses assessed as difference[8] but did not disclose similar data at baseline. Our analysis consistently found evidence of a deviation from normality across two large data sets, four OAB symptoms and for baseline data and treatment responses. This is in line with the limited data from other investigators focusing on treatment responses only.[8] Therefore, we conclude that it cannot necessarily be assumed that OAB diary data and their improvement upon treatment exhibit a normal distribution in the general population. It follows, that the distribution of OAB symptoms has to be checked carefully before authors can decide which measure of central tendency (means vs medians) and spread (SD vs IQR) are reasonable.

A true normal distribution is rare for any parameter in biomedical research. Therefore, the more relevant question is whether the extent of deviation from normality is large enough to make use of means and parametric null hypothesis tests assuming such normality misleading. To explore this, we have compared the means and medians of baseline values and treatment responses for each parameter in each study. According to our data, the difference between mean and median baseline values was 0.5 and 0.0 for urgency, 1.1 and 0.5 for incontinence, 0.6 and 0.2 frequency, and 0.5 and 0.5 for nocturia. For absolute treatment effects, that is, the difference between the number of symptoms, it was 1.1 and 0.5 for urgency, 0.7 and 0.9 for incontinence, 0.8 and 0.8 for frequency, and 0 and −0.3 for nocturia. These differences are clinically relevant because they are comparable with the difference in treatment effects between muscarinic antagonists and placebo in the reduction of incontinence and frequency episodes, which according to meta-analyses are < 1 and < 1.5 per day, respectively.[18] They are also comparable with reported differences between the β_3-adrenoceptor agonist mirabegron and placebo for a number of incontinence or micturition episodes (0.44 and 0.62, respectively)[19] and for comparisons between active treatments.[17] Similarly, the minimum noticeable change in incontinence episodes as assessed by patients using a quality of life tool was reported to be 3 per week,[20] that is, about 0.43 episodes/24 hours. The differences between mean and median in baseline or treatment-induced change of incontinence episodes in our two studies exceeded this threshold for being detectable by patients. Therefore, we conclude that the unsupported assumption of a normal distribution of OAB symptoms is not only theoretically flawed but also leads to an overestimation of symptom intensity and treatment improvements that are comparable with or greater than the placebo-corrected effect of muscarinic antagonists and greater than what has been reported to be noticeable by patients.

A different situation may exist for relative treatment effects (the difference between the number of symptoms after and before the therapy, divided by the number of symptoms before the therapy) expressed in percent. Although percentage changes also exhibited a deviation from a normal distribution, the resulting differences between means and medians typically were smaller and, if anything, means of percentage changes underestimated treatment effects. These differences most likely are not clinically relevant when compared with differences between active treatment and placebo[18,19] or to minimal differences noticeable by patients.[20] Therefore, it appears justifiable to report percentage changes of OAB symptoms as means. Some studies report baseline-adjusted absolute improvements of treatment,[21,22] which is conceptually the same as percentage improvements of treatment effects. Therefore, means of baseline-adjusted treatment effects may also be an acceptable way of

reporting efficacy data despite the formal deviation from the normal distribution. In a more general vein, not making assumptions on normal distribution in the absence of robust supporting data is the safer option for analysis of data.

One may expect that regulatory authorities such as the European Medicines Agency or the US Food and Drug Administration with their vast knowledge of treatment studies would have issued guidance on the use of means vs medians. However, the available guidance documents from both agencies related to OAB do not address the use of means vs medians or parametric vs nonparametric statistical analysis. However, the European Medicines Agency has issued general guidance that data should be checked for normality of distribution of reported variables and that analysis and presentation of the data should be based on this.[23]

5 | CONCLUSIONS

We conclude that most investigators report means of OAB symptoms and treatment-induced changes thereof, implying the assumption that these parameters exhibit a normal distribution. Our data shows that this assumption and the reporting of means may be scientifically unjustified and may potentially result in misinterpretations of study results on OAB symptoms. The difference between means and medians for OAB symptoms and symptom differences is likely to be of clinical relevance. Relative improvements of OAB symptoms (percentages changes) may be an exception from this, that is, even if there is no true normal distribution the consequences for parameter estimates are minimal. While our data are based on the use of propiverine, we propose that they may also be applicable to other treatments including muscarinic receptor antagonists as a class, β_3-adrenoceptor agonists, onabotulinum toxin A or behavioral or other nonpharmaceutical interventions.

ACKNOWLEDGMENT

This study is supported by Apogepha (the employer of authors Sandra Murgas) and Innovative Medicines Initiative 2 Joint Undertaking (grant no. 777364); this Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA.

ORCID

Andreas Stang http://orcid.org/0000-0001-6363-9061
Martin C. Michel http://orcid.org/0000-0003-4161-8467

REFERENCES

1. Coyne KS, Sexton CC, Thompson CL, et al. The prevalence of lower urinary tract symptoms (LUTS) in the USA, the UK and Sweden: results from the epidemiology of LUTS (EpiLUTS) study. *BJU Int*. 2009;104:352-360.
2. Coyne KS, Kvasz M, Ireland AM, Milsom I, Kopp ZS, Chapple CR. Urinary incontinence and its relationship to mental health and health-related quality of life in men and women in Sweden, the United Kingdom, and the United States. *Eur Urol*. 2012;61:88-95.
3. Coyne KS, Matza LS, Brewster-Jordan J, Thompson C, Bavendam T. The psychometric validation of the OAB family impact measure (OAB-FIM). *Neurourol Urodyn*. 2010;29:359-369.
4. Abrams P, Cardozo L, Fall M, et al. The standardisation of terminology of lower urinary tract function: report from the standardisation sub-committee of the International Continence Society. *Neurourol Urodyn*. 2002;21:167-178.
5. Kubota Y, Hamakawa T, Osaga S, et al. A kit ligand, stem cell factor as a possible mediator inducing overactive bladder. *Neurourol Urodyn*. 2018;37:1258-1265.
6. Blais A-S, Nadeau G, Moore K, Genois L, Bolduc S. Prospective pilot study of mirabegron in pediatric patients with overactive bladder. *Eur Urol*. 2016;70:9-13.
7. Azuri J, Kafri R, Ziv-Baran T, Stav K. Outcomes of different protocols of pelvic floor physical therapy and anti-cholinergics in women with wet over-active bladder: a 4-year follow-up. *Neurourol Urodyn*. 2017;36:755-758.
8. Martina R, Kay R, Abrams P, van Maanen R, Ridder A. A clinical perspective on the analysis and presentation of the number of incontinence episodes following treatment for OAB. *Neurourol Urodyn*. 2016;35:728-732.
9. Abulseoud A, Moussa A, Abdelfattah G, Ibrahim I, Saba E, Hassouna M. Transcutaneous posterior tibial nerve electrostimulation with low dose trospium chloride: could it be used as a second line treatment of overactive bladder in females. *Neurourol Urodyn*. 2018;37:842-848.
10. Drake MJ, Chapple C, Esen AA, et al. Efficacy and safety of mirabegon add-on therapy to solifenacin in incontinent overactive bladder patients with an iInadequate response to initial 4-week solifenacin monotherapy: a randomised double-blind multicentre phase 3B study (BESIDE). *Eur Urol*. 2016;70:136-145.
11. Borch L, Hagstroem S, Kamperis K, Siggaard CV, Rittig S. Transcutaneous electrical nerve stimulation combined with oxybutynin is superior to monotherapy in children with urge incontinence: a randomized, placebo controlled study. *J Urol*. 2017;198:430-435.
12. Koschorke M, Leitner L, Sadri H, Knüpfer SC, Mehnert U, Kessler TM. Intradetrusor onabotulinumtoxinA injections for refractory neurogenic detrusor overactivity incontinence: do we need urodynamic investigation for outcome assessment? *BJU Int*. 2017;120:848-854.
13. Schneider T, Marschall-Kehrel D, Hanisch JU, Michel MC. Do gender, age or life style factors affect responses to antimuscarinic treatment in overactive bladder patients? *Int J Clin Pract*. 2010;64:1287-1293.
14. Witte LPW, Peschers U, Vogel M, de la Rosette JJMCH, Michel MC. Does the number of previous vaginal deliveries affect overactive bladder symptoms or their response to treatment? *LUTS*. 2009;1:82-87.

15. Michel MC, Schneider T, Krege S, Goepel M. Do gender or age affect the efficacy and safety of tolterodine? *J Urol*. 2002;168:1027-1031.
16. Michel MC, de la Rosette JJ, Piro M, Schneider T. Comparison of symptom severity and treatment response in patients with incontinent and continent overactive bladder. *Eur Urol*. 2005;48:110-115.
17. Novara G, Galfano A, Secco S, et al. A systematic review and meta-analysis of randomized controlled trials with antimuscarinic drugs for overactive bladder. *Eur Urol*. 2008;54:740-764.
18. Reynolds WS, McPheeters M, Blume J, et al. Comparative effectiveness of anticholinergic therapy for overactive bladder in women. A systematic review and meta-analysis. *Obstet Gynecol*. 2015;125:1423-1432.
19. Cui Y, Zong H, Yang C, Yan H, Zhang Y. The efficacy and safety of mirabegron in treating OAB: a systematic review and meta-analysis of phase III studies. *Int Urol Nephrol*. 2014;46:275-284.
20. Homma Y, Koyama N. Minimally clinically important change in urinary incontinence detected by a quality of life assessment tool in overactive bladder syndrome with urge incontinence. *Neurourol Urodyn*. 2006;25:228-235.
21. Wagg A, Dale M, Tretter R, Stow B, Compion G. Randomised, multicenre, placebo-controlled, double-blind crossover study investigating the effect of solifenacin and oxybutynin in elderly people with mild cognitive impairment: the SENIOR study. *Eur Urol*. 2013;64:74-81.
22. Chapple CR, Dvorak V, Radziszewski P, et al. A phase II dose-ranging study of mirabegron in patients with overactive bladder. *Int Urogynecol J*. 2013;24:1447-1458.
23. European Medicines Agency. ICH Topic E9. Statistical principles for clinical trials; 1998. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf. Accessed December 5, 2019.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Amiri M, Murgas S, Stang A, Michel MC. Do overactive bladder symptoms and their treatment-associated changes exhibit a normal distribution? Implications for analysis and reporting. *Neurourology and Urodynamics*. 2020;39:754–761. https://doi.org/10.1002/nau.24275

3.2. Factors associated with initial dosing for up-titration of propiverine and treatment outcomes in overactive bladder syndrome patients in a non-interventional setting.

Journal of
Clinical Medicine

MDPI

Article

Factors Associated with Decisions for Initial Dosing, Up-Titration of Propiverine and Treatment Outcomes in Overactive Bladder Syndrome Patients in a Non-Interventional Setting

Marjan Amiri [1,2], Tim Schneider [3], Matthias Oelke [4], Sandra Murgas [5] and Martin C. Michel [6,*]

1 Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, 45130 Essen, Germany; marjan.amiri@uk-essen.de
2 Center for Clinical Trials Essen (ZKSE), University Hospital Essen, 45130 Essen, Germany
3 Praxisklinik Urologie Rhein-Ruhr, 45468 Mülheim, Germany; t.schneider@pur-r.de
4 Department of Urology, St. Antonius Hospital, 48599 Gronau, Germany; matthias.oelke@st-antonius-gronau.de
5 Apogepha, 01309 Dresden, Germany; s.murgas@apogepha.de
6 Department of Pharmacology, Johannes Gutenberg University, 55131 Mainz, Germany
* Correspondence: marmiche@uni-mainz.de; Tel.: +49-6131-179346

Abstract: Two doses of propiverine ER (30 and 45 mg/d) are available for the treatment of overactive bladder (OAB) syndrome. We have explored factors associated with the initial dosing choice (allocation bias), the decision to adapt dosing (escalation bias) and how dosing relative to other factors affects treatment outcomes. Data from two non-interventional studies of 1335 and 745 OAB patients, respectively, receiving treatment with propiverine, were analyzed post-hoc. Multivariate analysis was applied to identify factors associated with dosing decisions and treatment outcomes. Several parameters were associated with dose choice, escalation to higher dose or treatment outcomes, but only few exhibited a consistent association across both studies. These were younger age for initial dose choice and basal number of urgency and change in incontinence episodes for up-titration. Treatment outcome (difference between values at 12 weeks vs. baseline) for each OAB system was strongly driven by the respective baseline value, whereas no other parameter exhibited a consistent association. Patients starting on the 30 mg dose and escalating to 45 mg after 4 weeks had outcomes comparable with those staying on a starting dose of 30 or 45 mg. We conclude that dose escalation after 4 weeks brings OAB patients with an initially limited improvement to a level seen in initially good responders. Analysis of underlying factors yielded surprisingly little consistent insight.

Keywords: propiverine; dose-titration; overactive bladder syndrome; allocation bias; escalation bias

Citation: Amiri, M.; Schneider, T.; Oelke, M.; Murgas, S.; Michel, M.C. Factors Associated with Decisions for Initial Dosing, Up-Titration of Propiverine and Treatment Outcomes in Overactive Bladder Syndrome Patients in a Non-Interventional Setting. *J. Clin. Med.* **2021**, *10*, 311. https://doi.org/10.3390/jcm10020311

Received: 23 December 2020
Accepted: 12 January 2021
Published: 15 January 2021

1. Introduction

Muscarinic receptor antagonists (antimuscarinics) are the mainstay of treatment of the overactive bladder (OAB) syndrome [1,2]. Several of the clinically available members of this drug class are available in multiple dose strengths, for instance darifenacin [3], fesoterodine [4], propiverine [5] and solifenacin [6]. This allows increasing the dose in patients with an insufficient improvement of symptoms and decreasing it in those with bothersome adverse drug reactions (ADRs). While it appears intuitive to assume that a greater dose will have greater effects, clinical experience demonstrates that this is not necessarily the case. This is expected because the law of mass action postulates that a greater dose of antagonist/inhibitor will have a greater effect only until a ceiling limit is reached due to maximum occupancy of the molecular target of the drug; for instance, dose escalation of the antidepressant paroxetine has been shown not to improve symptoms because the standard dose already occupies most of the serotonin transporters [7]. On the

other hand, the optimal position within the dose–response curve of a given drug, i.e., where an optimal ratio between efficacy and tolerability is achieved, can differ between patients [8]. Prediction of the optimal dose for a given patient is difficult because factors, such as age and comorbidity [9], genotype [10] or, for antimuscarinics, concomitant medication also acting on muscarinic receptors [11], may pharmacodynamically affect the drug response. Moreover, depending on the specific metabolic pathways responsible for the elimination of a given antimuscarinics, pharmacokinetic factors, including age, renal function and genotype of drug-metabolizing enzymes, may also play a role [12].

The prescribing information of most antimuscarinics for the treatment of OAB defines a recommended starting dose with the option to increase the dose if tolerability is acceptable and greater efficacy is needed. Accordingly, the effect of dose escalation has been studied for several antimuscarinics including fesoterodine [13–17] and solifenacin [18,19]. A variation of this theme has been to study the effects of escalating from a low dose of one drug to the standard dose of another one within the same drug class [20] or to model the effects of dose escalation based on findings from previous trials [21]. These studies demonstrate that the decision to increase the dose was often but not always associated with the extent of improvement of OAB symptoms during the initial treatment period; however, which OAB symptoms were primarily associated with that decision differed considerably between studies. Some reports also proposed that baseline symptom severity, body weight and gender (based social/cultural roles and personal identity) may be associated with the decision for dose escalation [15], but this was not found consistently. Interestingly, the degree of symptom improvement in the initial treatment period has also been found to be associated with the decision to increase dosage within the placebo arm of controlled dose-escalation studies [22].

Propiverine is an antimuscarinic for the treatment of the OAB syndrome that differs from other members of this drug class because the compound and some of its metabolites additionally have inhibitory effects on L-type Ca^{2+}-channels [23,24], which play a role in the control of bladder smooth muscle tone [25]. Moreover, propiverine extended release (ER) has two approved starting doses of 30 and 45 mg once daily, i.e., the higher dose is not only available as part of dose escalation. As most other antimuscarinics have only one approved starting dose, it has not been reported, to our knowledge, which factors are associated with the initial dosing decision. Moreover, it is possible that factors associated with the initial dosing decision also impact the decision to increase the dose and, thereby, the overall treatment outcome. Against this background, we report two non-interventional studies (NISs) of similar design in which patients with OAB syndrome were treated with either dose of propiverine for a planned observation period of approximately12 weeks and the possibility for dose-adjustment after about 4 weeks. While reporting on the primary outcomes of both studies, we have also applied multivariate analysis (general linear models) in a post-hoc approach to explore three questions:

Which factors are associated with initial dosing decision (allocation bias)?

Which factors are associated with a decision for dose escalation after 4 weeks of treatment (escalation bias)?

How much of differential efficacy of the two dose strengths can be attributed to greater dose and how much to other factors associated with dosing choice?

2. Materials and Methods

Two NISs with a similar design were performed. Study I was primarily designed to explore the effects of different starting doses and dose adjustment after 4 weeks with a treatment duration of 12 weeks, i.e., specifically designed for the purpose of the present analyses. Study II was designed to explore the effect of additional material (information sheet about OAB and mode of action of the drug) on efficacy and tolerability and on premature discontinuation during a treatment period of 12 weeks and allowed extension of observation for up to 24 weeks in a subgroup. Otherwise, its design was very similar to that of study I. Therefore, we have used both studies to address our research questions. Our

exploratory approach considers study I as hypothesis-generating and study II as exploring the robustness of the findings from study I but not as formally hypothesis-testing. To keep the manuscript readable, data related to the primary aim of study II and some other outcomes are shown in the Supplementary Materials. We have previously used parts of both datasets to explore the presence or absence of a normal distribution of the OAB parameters urgency, incontinence, frequency and nocturia and of treatment-associated alterations thereof [26]. We now report the primary analysis of efficacy and tolerability of the full dataset of both studies. Both studies were based on §67, 3 of the German Drug Act and had been approved by the ethical committee of the state board of physicians in Saxony, Germany (Sächsische Landesärztekammer EK-BR-14/12-1 and EK-BR-18/14-1).

Based on their non-interventional character, both studies lacked formal inclusion or exclusion criteria other than the Summary of Product Characteristics. Rather, 456 and 158 participating physicians in studies I and II, respectively, were asked to systematically document their observations in patients who were to be treated with propiverine ER (30 or 45 mg once daily) based on the physician's medical judgment. Three visits were planned during the observation period, visit 1—one at baseline (initiation of propiverine prescription and selection of starting dose), visit 2—after about 4 weeks (possibility of dose adaptation) and visit 3—after about 12 weeks (study end); study II allowed us to extend the observational period to 24 weeks. The first visit of the first patient was recorded on 1.7.2010 and the last visit of the last patient was on 23.1.2013 in study I and on 14.1.2014 and 15.7.2015 in study II.

During each visit, parameters related to OAB and decisions on subsequent dosing were recorded. Demographics, comorbidities and comedications were additionally recorded at baseline. Global tolerability (rated as very good, good, sufficient or insufficient) was assessed by the patient and the physician at visits 2 and 3. Data on ADR were collected throughout the entire study. Post-void residuals (PVRs) were considered as additional safety parameter if available (recording not mandatory to maintain the non-interventional character of the studies). The safety and tolerability analysis included all patients who had taken at least one drug dose and had at least one physician contact thereafter. The efficacy analyses included all patients who had OAB-related data at baseline and at least one time point thereafter.

In line with the non-interventional character of the study, the protocol did not specify whether OAB-related data were collected from voiding diaries or from patient recollection, but the applicable German guideline at the time the studies were performed recommended recording of voiding diaries [27]. Categorical data (e.g., gender or dose) are shown as % of the respective population. Continuous data are expressed as means ± SD (age, height, weight, body mass index (BMI)) or as median with inter-quartile range (IQR (reported as lower and upper quartile separated by a ";"); OAB duration and daily episodes of urgency, incontinence, voids and nocturia), depending on whether variability was considered to exhibit a normal distribution [26]. Despite deviating from a normal distribution, OAB parameters are also shown as means ± SD to facilitate comparison with previously reported studies. Patients with medically implausible values (urgency > 50, incontinence > 30, frequency > 40, and nocturia > 20 episodes/24 h) were excluded from the analysis for that symptom and visit; this affected four patients each for urgency and frequency, one each for incontinence and nocturia in study I and none in study II.

Data handling and statistical analyses were performed by Bioconsult GmbH (Rickenbach, Switzerland), a contract research organization, based on a statistical analysis plan developed by the authors and using SAS version 9.4 (SAS Institute, Cary, NC, USA). The demographic data and OAB parameters at baseline and subsequent visits for the efficacy population are shown descriptively. Univariate analyses compared (a) demographics and baseline data in patients starting at the 30 mg and 45 mg dose at visit 1, (b) such data plus initial OAB symptom changes at visit 2 in patients starting at 30 mg and either staying on that dose or increasing it to 45 mg, and (c) treatment outcomes at visit 3 for patients who started and stayed on 30 mg, who started on 30 mg and increased to 45 mg and who started

on 45 mg and stayed on that dose. These data were analyzed for descriptive p-values using the Kruskal–Wallis test.

Multivariate analyses (general linear models) were applied to identify factors associated with initial dosing decisions and dose escalation decision as well as the relative roles of dose and other factors in treatment outcomes. Variables were incorporated into the model if $p < 0.03$ and removed if $p > 0.35$ in the next step; the procedures were stopped when only variables with $p < 0.35$ were retained in the model. The model for the initial dosing choice included gender, age, body weight, height, BMI, duration of OAB and baseline values for urgency, incontinence, frequency and nocturia as potential explanatory variables. Precision of parameter estimates in the general linear models is indicated by their standard error (SE). Treatment-associated changes of a symptom were determined only for patients exhibiting that symptom at baseline, e.g., for incontinence episodes only in those having $\geq$ 1 incontinence episode; however, not exhibiting one symptom at baseline did not preclude analyses of other symptoms of the same patient. The model for factors associated with dose escalation included the same variables and additionally the OAB parameters after 4 weeks at visit 2. The model for treatment outcomes also included the same variables and additionally the dose after visits 1 and 2.

Based on the exploratory character of the study and in line with recent recommendations [28], no hypothesis-testing statistical analysis was applied. Therefore, all reported p-values should be interpreted as descriptive only. Rather, we have considered study I in line with its primary aim as hypothesis-generating and study II to check for robustness of our findings. Overall reporting follows the STROBE guidelines for cohort studies (https://strobe-statement.org).

3. Results

3.1. Patient Flow and Baseline Data

A total of 1335 and 745 patients participated in studies I and II, respectively. Demographics of both studies, baseline symptoms, documented prior interventions with implications for lower urinary tract function, comorbidities and comedications, as well as overall patient flow, are shown in the Supplementary Materials.

3.2. Descriptive Analysis of Treatment Outcomes

While no specific instructions on time of administration were given, most patients in study I (919/1120; 82.1%) reported taking propiverine in the morning. After 4 weeks of treatment, clinically meaningful improvements (median with IQR and mean $\pm$ SD in parentheses) were observed in the overall efficacy population with a reduction in urgency episodes by 4 (2; 7; 5.1 $\pm$ 4.6), in incontinence episodes by 2 (1; 4; 2.8 $\pm$ 3.1), in micturitions by 4 (2; 6; 4.2 $\pm$ 3.3), and in nocturia episodes by 1 (1; 2; 1.5 $\pm$ 1.3). Between weeks 4 and 12, the overall cohort of patients reported additional improvements in urgency episodes by 1 (0; 3; 1.8 $\pm$ 2.6), in incontinence episodes by 1 (0; 1; 0.9 $\pm$ 1.6), in frequency by 1 (0; 2; 1.4 $\pm$ 1.9), and in nocturnal voids by 0 (0; 1; 0.5 $\pm$ 0.9). Thus, overall improvements from baseline to week 12 were improvements of urgency episodes by 6 (3; 10; 6.9 $\pm$ 5.2), in incontinence episodes by 3 (1; 5; 3.7 $\pm$ 3.3), in frequency by 5 (3; 8; 5.7 $\pm$ 3.7), and in nocturnal voids by 2 (1; 3; 2.0 $\pm$ 1.4). Treatment effects in the various groups are summarized in Figure 1 based on medians and IQR and in the Supplementary Materials for means $\pm$ SD. Corresponding data for study II were comparable and are shown in the Supplementary Materials.

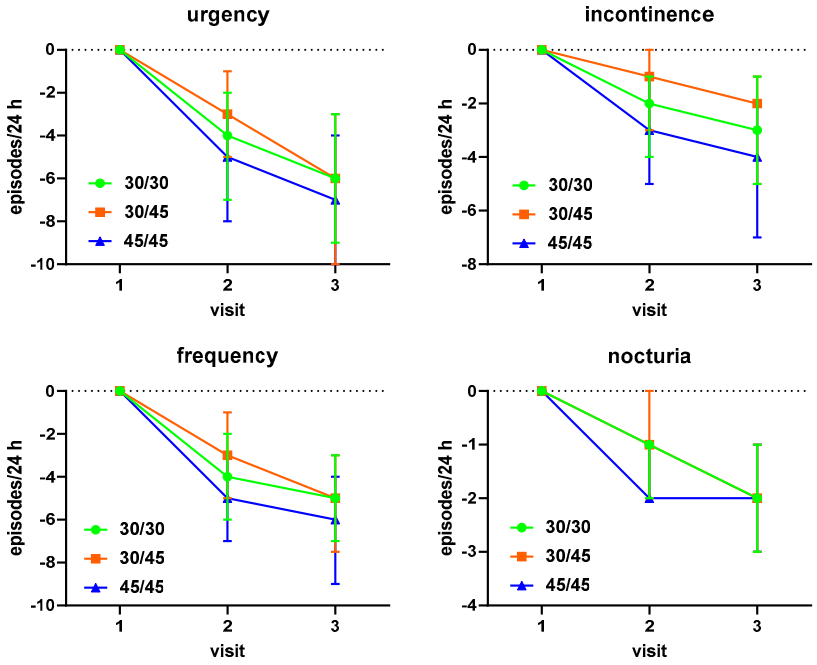


Figure 1. Intra-individual change of overactive bladder (OAB) symptoms in the cohorts of patients starting on 30 mg and staying on that dose (30/30), starting on 30 mg and escalating to 45 mg at visit 2 after about 4 weeks (30/45) and starting and staying on 45 mg until study end after about 12 weeks (45/45) in study I. Data are shown as medians with IQR. Means ± SD are shown for comparison in Supplementary Materials. Patients not exhibiting a given symptom at baseline were excluded from the analysis of that symptom; specifically, 364 subjects reported no incontinence at baseline. Corresponding data for study II are shown in the Supplementary Materials.

3.3. Factors Associated with Dosing Decision at Visit 1

Demographic and OAB-related baseline data of patients initially receiving 30 or 45 mg propiverine are summarized in Table 1 for study I. The two groups had a very similar gender distribution, mean age and height, but those starting on 45 mg had slightly greater mean body weight (+2.9 kg). The 45 mg group also had greater baseline symptom severity, which was most notable for number of incontinence episodes and micturitions with medians being 1–2 episodes greater than in the 30 mg group. Data for study II were comparable and are shown in the Supplementary Materials.

After exclusion of subjects due to missing values for the response or explanatory variables, the logistic regression analysis included 606 and 147 patients from study I, starting at the 30 and 45 mg dose, respectively, and 294 and 132 subjects from study II. A younger age was the only variable associated with a higher starting dose in both studies with a $p < 0.05$. Additionally, both a greater basal number of incontinence and nocturia episodes were associated with a higher starting dose in study I and a longer duration of OAB and greater number of micturitions in study II (Table 2).

3.4. Factors Associated with Dosing Increase at Visit 2

The dose of propiverine was adjusted in some patients continuing treatment at visit 2 (patient disposition figures for studies I and II are shown in the Supplementary Materials). In study I, 84% of patients having started with a dose of 30 mg remained on that dose (30/30 group), whereas 16% were switched to the 45-mg dose (30/45 group). In addition, 88% of patients having started with a dose of 45 mg remained on that dose (45/45 group), and 12% reduced it to 30 mg; the latter group was not considered further as it was deemed too small to allow meaningful analysis. Corresponding data for study II are shown in the Supplementary Materials.

Table 1. Demographic and OAB-related baseline variables in patients starting treatment with a propiverine dose of 30 or 45 mg/d.

	Initial 30 mg	Initial 45 mg	*p*-Value
n	1021	239	
Demographic parameters			
Gender, % female/male	64.3/32.6	64.3/32.9	
Previous OAB treatment, %	33.7	45.0	
Age, years	65.8 ± 13.1	65.4 ± 12.5	0.3820
Height, cm	169.0 ± 7.9	169.9 ± 8.3	0.1543
Weight, kg	77.4 ± 14.2	80.3 ± 17.0	0.0184
BMI, kg/m^2	27.1 ± 4.3	27.9 ± 5.5	0.1187
OAB-related parameters			
OAB duration, months	12.1 (3.9; 34.6)	13.2 (4.7; 35.1)	0.3343
Urgency episodes/24 h	9 (6; 13) 10.0 ± 5.8	10 (7; 14) 10.8 ± 5.8	0.0251
Incontinence episodes/24 h	6 (4; 23) 4.7 ± 3.7	8 (5; 30) 6.2 ± 4.4	<0.0001
Urinary frequency/24 h	13 (11; 16) 13.3 ± 4.2	14 (11; 17) 14.3 ± 4.4	0.0004
Nocturia episodes/24 h	3 (2; 4) 3.4 ± 1.6	3 (3; 5) 3.7 ± 1.7	0.0008

Data are shown as % of patients for gender (does not add up to 100 due to missing values), as means ± SD for continuous demographic parameters, and medians with inter-quartile range (IQR) of OAB-related parameters (means ± SD only shown to facilitate comparison with previous reports). Descriptive *p*-values for the difference between groups are from univariate analysis using unpaired, two-tailed Kruskal–Wallis tests. The analysis of the OAB symptoms included only patients that had a documented dose and a measured value at baseline other than 0.

Table 2. Factors associated with starting dose (30 vs. 45 mg) in a logistic regression analysis of studies I and II. Data are the reported parameter for maximum likelihood estimate with its standard error (SE) and descriptive *p*-value based on Chi square tests. Parameters with a $p > 0.3$ were not retained in the model and are not shown.

Parameter	Estimate ± SE	*p*-Value	Estimate ± SE	*p*-Value
	Study I		Study II	
Gender, female	−0.225 ± 0.140	0.1081	-	-
Age, years	0.017 ± 0.008	0.0335	0.027 ± 0.010	0.0067
Weight, kg	0.083 ± 0.070	0.2338	−0.132 ± 0.093	0.1556
Height, cm	−0.120 ± 0.069	0.0820	0.010 ± 0.090	0.2706
BMI, kg/m^2	−0.261 ± 0.198	0.1875	0.347 ± 0.271	0.2013
OAB duration, months	-	-	−0.011 ± 0.002	<0.0001
Urgency/24 h	-	-	0.032 ± 0.023	0.1606
Incontinence/24 h	−0.107 ± 0.024	<0.0001	-	-
Micturitions/24 h	-	-	−0.238 ± 0.071	0.0008
Nocturia/24 h	−0.126 ± 0.060	0.0348	-	-

Table 3 shows demographic and OAB-related data at baseline and after 4 weeks of treatment in the 30/30 and the 30/45 group of study I. Patients with dose escalation were slightly taller (2.3 cm) and heavier (4.2 kg). Prior to the start of treatment, they had more daily urgency, incontinence, micturition and nocturia episodes. These differences

were maintained after 4 weeks of treatment and became even greater except for nocturia. Corresponding data for study II are shown in the Supplementary Materials.

Table 3. Demographic and OAB-related baseline variables and after 4 weeks in patients starting treatment with a propiverine dose of 30 mg/d and either staying on that dose after 4 weeks or increasing it 45 mg/d.

	Stay on 30 mg	**Increase to 45 mg**	***p*-Value**
n	789	160	
Demographic parameters			
Gender, % female/male	66.2/33.8	60.0/40.0	
Previous OAB treatment, %	31.1	50.6	
Age, years	65.6 ±13.1	67.0 ± 13.0	0.2313
Height, cm	169.0 ± 7.7	171.3 ± 7.9	0.0010
Weight, kg	77.0 ± 13.9	81.2 ± 14.4	<0.0001
BMI, kg/m^2	27.0 ± 4.2	27.6 ± 4.6	0.1403
OAB-related parameters at baseline			
OAB duration, months	11.2 (3.6; 31.3)	13.6 (5.0; 51.9)	
Urgency episodes/24 h	9 (5; 13) 9.6 ± 5.6	11 (6.5; 15) 11.7 ± 6.6	<0.0001
Incontinence episodes/24 h	4 (2; 6) 4.6 ± 3.4	4 (2; 7) 5.4 ± 4.7	<0.0001
Urinary frequency/24 h	13 (10; 15) 13.1 ± 4.1	14 (12; 16) 14.3 ± 4.1	<0.0001
Nocturia/24 h	3 (2; 4) 3.3 ± 1.5	4 (3; 4.5) 3.7 ± 1.6	<0.0001
OAB-related parameters after 4 weeks			
Urgency episodes/24 h	3 (2; 6) 4.4 ± 3.6	7 (4; 11) 8.0± 5.4	<0.0001
Incontinence episodes/24 h	1 (0; 2) 1.8 ± 2.1	2 (1; 5) 3.6 ± 3.6	<0.0001
Urinary frequency/24 h	8 (7; 10) 8.8 ± 2.8	11 (9; 13) 11.2 ± 3.4	<0.0001
Nocturia/24 h	2 (1; 2) 1.8 ± 1.1	2 (2; 3) 2.7 ± 1.2	<0.0001

Data are shown as % of patients for gender, as means ± SD for continuous demographic parameters, and medians with IQR of OAB-related parameters (means ± SD only shown to facilitate comparison with previous reports). Descriptive *p*-values for the difference between groups are from univariate analysis using unpaired, two-tailed Kruskal–Wallis tests. The analysis of the OAB symptoms included only patients that had a documented dose and a measured value at baseline other than 0.

After exclusion of subjects due to missing values for the response or explanatory variables, the logistic regression analysis included 834 and 161 patients from study I staying at the 30 mg dose or increasing to 45 mg, respectively, and 435 and 59 subjects from study II. A greater number of urgency episodes at baseline and a greater number of incontinence episodes after 4 weeks of treatment were associated with dose escalation in both studies with a $p < 0.05$; however, greater height was associated only in study I and greater number of nocturia episodes were only associated in study II (Table 4).

Table 4. Factors associated with staying at the starting dose of 30 mg vs. increasing to a dose of 45 mg in a logistic regression analysis taking both OAB parameters at baseline and after 4 weeks into consideration.

Parameter	Estimate ± SE	*p*-value	Estimate ± SE	*p*-Value
	Study I		Study II	
Age, years	-	-	1.420 ± 1.387	0.3059
Weight, kg	−0.017 ± 0.009	0.0611	−0.17 ± 0.012	0.1444
Height, cm	−0.036 ± 0.018	0.0413	-	-
Urgency/24 h baseline	−0.100 ± 0.036	0.0049	-0.192 ± 0.054	0.0004
Incontinence/24 h baseline	−0.087 ± 0.053	0.0998	0.108 ± 0.078	0.1659
Micturitions/24 h baseline	−0.069 ± 0.056	0.2185	-	-
Nocturia/24 h baseline	−0.160 ± 0.119	0.1776	−0.311 ± 0.155	0.0450
Urgency/24 h 4 weeks	−0.067 ± 0.037	0.0669	-	-
Incontinence/24 h 4 weeks	−0.126 ± 0.058	0.0300	−0.201 ±00.084	0.0165
Micturitions/24 4 weeks	-	-	-	-
Nocturia/24 h 4 weeks	-	-	−0.549 ± 0.227	0.0153

Data are reported parameter estimate for maximum likelihood estimate with its standard error (SE) and descriptive *p*-value based on Chi square tests. Parameters with a $p > 0.3$ were not retained in the model and are not shown.

3.5. Factors Associated with Treatment Outcomes

As it cannot necessarily be assumed that factors associated with improvement of one symptom are the same as for other symptoms, we have explored factors associated with improvement (greater improvement = small symptom episode frequency) separately for each OAB parameter with the delta between value after 12 weeks of treatment as dependent and respective basal value as independent variable. These logistic regression analyses considered the 30/30, 30/45 and 45/45 groups. Demographics and baseline values for the 30/30 and 30/45 groups in study I are shown in Table 3 and those for the 45/45 group in Table 1. Values of OAB parameters after 4 weeks of treatment are shown in Table 3 for the 30/30 and the 30/45 group; for the 45/45 group, they were 5 (5.1 ± 3.9) for urgency, 2 (2.6 ± 3.1) for incontinence, 9 (9.5 ± 3.3) for frequency and 2 (2.1 ± 1.2) for nocturia (corresponding data for study II are shown in the Supplementary Materials).

After exclusion of subjects due to missing values for the response or explanatory variables, the logistic regression analysis included 699 patients from study I and 396 from study II. The only variable consistently associated with treatment outcome for any OAB symptom was the baseline value of the same symptom, i.e., baseline urgency for overall improvement of urgency (Table 5) and baseline incontinence for overall improvement of incontinence (Table 6); logistic regression results for frequency and nocturia are shown in the Supplementary Materials. Compared to the 45/45 group, being in the 30/30 or 30/45 group was associated with smaller treatment-associated improvements for urgency and incontinence in study I, but not in study II (Tables 5 and 6 and Supplementary Materials); dosing regimen had no statistically significant effects on treatment-associated improvements of frequency and nocturia in either study (Supplementary Materials).

Table 5. Factors associated with overall improvement of urgency (12 weeks vs. baseline) in a logistic regression analysis taking demographics, OAB parameters at baseline, duration of condition and dose level into consideration.

Parameter	Estimate ± SE	*p*-Value	Estimate ± SE	*p*-Value
	Study I		Study II	
Gender, female	0.026 ± 0.009	0.0056	0.011 ± 0.414	0.3954
Age, years	0.026 ± 0.009	0.0056	0.011 ± 0.012	0.3739
Weight, kg	−0.123 ± 0.081	0.1592	0.019 ± 0.112	0.8628
Height, cm	0.114 ± 0.081	0.1592	−0.028 ± 0.109	0.7973
BMI, kg/m^2	0.340 ± 0.235	0.1475	−0.039 ± 0.327	0.9049
OAB duration, months	0.012 ± 0.002	<0.0001	0.005 ± 0.003	0.0677
Urgency/24 h	−0.714 ± 0.025	<0.0001	−0.842 ± 0.034	<0.0001
Incontinence/24 h	−0.003 ± 0.035	0.9282	0.210 ± 0.043	<0.0001
Micturitions/24 h	−0.008 ± 0.037	0.8364	0.029 ± 0.054	0.5928
Nocturia/24 h	−0.043 ± 0.008	0.6264	−0.087 ± 0.115	0.4483
Dose 30/30 *	0.109 ± 0.297	0.7138	−0.428 ± 0.328	0.1917
Dose 30/45 *	0.998 ± 0.373	0.0076	0.343± 0.490	0.4483

p-values for gender relate to male and those for dose level relate to the 45/45 group as reference; *: term not uniquely estimable.

Table 6. Factors associated with overall improvement of incontinence (12 weeks vs. baseline) in a logistic regression analysis taking demographics, OAB parameters at baseline, duration of condition and dose level into consideration.

Parameter	Estimate ± SE	*p*-Value	Estimate ± SE	*p*-Value
	Study I		Study II	
Gender, female	−0.012 ± 0.178	0.2662	−0.331 ± 0.261	0.2060
Age, years	0.012 ± 0.005	0.0182	0.003 ± 0.008	0.7324
Weight, kg	−0.048 ± 0.046	0.3025	−0.051 ± 0.071	0.4701
Height, cm	0.042 ± 0.046	0.3741	0.043 ± 0.069	0.5347
BMI, kg/m^2	0.130 ± 0.132	0.3257	0.156 ± 0.206	0.4498
OAB duration, months	0.008 ± 0.001	<0.0001	0.003 ± 0.002	0.1493
Urgency/24 h	−0.008 ± 0.014	0.5685	−0.056 ± 0.021	0.0086
Incontinence/24 h	−0.765 ± 0.020	<0.0001	−0.657 ± 0.028	<0.0001
Micturitions/24 h	−0.018 ± 0.021	0.3851	0.011 ± 0.035	0.7559
Nocturia/24 h	−0.032 ± 0.050	0.3851	0.022 ± 0.072	0.7655
Dose 30/30 *	−0.103 ± 0.168	0.5384	−0.247 ± 0.208	0.2359
Dose 30/45 *	0.487 ± 0.211	0.0211	−0.128 ± 0.306	0.6758

p-values for gender relate to male and those for dose level relate to the 45/45 group as reference; *: term not uniquely estimable.

3.6. Safety and Tolerability

Safety and tolerability were assessed in three ways: Firstly, PVR was unchanged (study I: 10 (10; 30) ml at baseline and after 12 weeks of treatment; study II: 20 (0; 38) ml at baseline and 20 (0; 39.5) ml after 12 weeks of treatment). Urinary retention was reported for no patient in study I and one patient in study II.

Second, 324 patients (24.3%) in study I reported a total of 461 ADR with dry mouth (19.6%) and constipation (6.0%) mentioned most frequently. The incidence of ADR was comparable between dose regimens. A total of 145 patients (10.9%) discontinued treatment during or after the treatment period in study I (48 due to ADR, 38 insufficient efficacy, 33 based on patient wish, 14 because of being symptom-free, 8 due to other and 3 due

to unknown reasons). This included 71 patients up to visit 2 with similar incidence with a starting dose of 30 and 45 mg (57/1069, 5.3% vs. 13/249, 5.2%). It also included 74 patients up to visit 3 (41/875, 4.7% received the 30 mg dose and 33/371, 8.9% the 45 mg dose). Reasons for discontinuation were ADR (n = 48), insufficient efficacy (n = 38), patient wish (n = 33), having become symptom free (n = 14), and other reasons (n = 12); no information on the reason for discontinuation was recorded for three patients. Patients with discontinuation of treatment after the planned observation period of 12 weeks were included in the calculations of treatment effects.

In study II, 163 patients (21.9%) reported 231 ADR with dry mouth (16.1%) and constipation (5.5%) mentioned most frequently and mostly being rated as mild. The capture of treatment discontinuation at visit 3 was not comparable with study I, because of the voluntary extension after 12 weeks. At visit 3, treatment discontinuation was observed in a total of 83/745 patients (11.1%), 5.2% due to ADRs, 1.9% insufficient efficacy, 0.9% based on patient's own decision, 0.8% because of being symptom-free and 0.8% due to other reasons. The incidence at visit 3 regarding the starting dose was similar (30 mg: 62/531, 11.7% vs. 18/200, 9.0%).

Third, global tolerability was rated by the patient at study end as very good, good, sufficient or insufficient in 41.3%, 47.2%, 10.3% and 1.3% for propiverine 30 mg and in 38.6%, 49.6%, 9.6% and 2.2% for propiverine 45 mg in study I. In study II, the rating for propiverine 30 mg was 43.0%, 45.7%, 7.6% and 3.7% for very good, good, sufficient or insufficient and correspondingly for propiverine 45 mg 36.3%, 54.0%, 6.8% and 3.0% at study end.

4. Discussion

The present analyses attempted to address three questions:

Which factors are associated with initial dosing decision (allocation bias)?

Which factors are associated with a decision for dose escalation after 4 weeks of treatment (escalation bias)?

How much of differential efficacy of the two dose-strengths can be attributed to greater dose and how much to factors associated with the dosing decision?

While the second question had previously been addressed in several studies with other muscarinic antagonists, the first and third question, to the best of our knowledge, are addressed here for the first time, most likely because antimuscarinics other than propiverine do not allow a choice of starting doses according to their respective prescribing information.

4.1. Critique of Methods

NISs differ from randomized, controlled trials (RCTs) in several ways. As NISs typically lack a placebo or other comparator group, they do not allow direct conclusions on the efficacy or tolerability of a drug; however, they describe what realistically can be expected to occur in routine clinical practice. For instance, the treatment of lower urinary tract symptoms (LUTSs) is known to exhibit a strong placebo effect [29]. Accordingly, antimuscarinics as a class have little effect on nocturia relative to placebo in RCT [30], but improvements of nocturia upon treatment with antimuscarinics in OAB patients [31,32] or α_1-adrenoceptor antagonists in male LUTS patients [33,34] have consistently been reported from NISs, as also observed in the present study. On the other hand, NISs lack the strict inclusion and exclusion criteria typical for RCTs, which leads to a less artificial study population. Thus, RCTs have a greater internal, whereas NISs can have a greater external validity.

In line with their non-interventional character, NISs have fewer rules on how data are to be captured than RCTs. While the applicable German guideline at the time the studies were performed recommended assessing OAB symptoms based on voiding diaries [27], we have no data regarding whether this has been implemented consistently.

Our analyses are based on two NISs that had a very similar design. There were three differences: Study II formally had a distinct primary aim (effect on information material),

had an option to extend the observation to 24 weeks, and was performed later in time (2014–2015 as compared to 2010–2013). With regard to the latter, it is important to note that no major changes had occurred between the two time frames in the overall German healthcare system, or in general, regarding recommendations related to the treatment of OAB syndrome. Thus, we did not identify major differences in treatment approaches or overall treatment outcomes and had expected to obtain very similar findings in both studies. Nonetheless, we decided not to pool the data but rather to analyze them in parallel. The main reason was that this would enable us to see whether analytical outcomes would be consistent in two distinct samples of the OAB syndrome population.

Our previous systematic review of the literature has demonstrated that the majority of studies are reporting means ± SD of OAB symptoms; while this implies the assumption of a normal distribution of these parameters, almost all studies failed to provide evidence in this regard [26]. On the other hand, our own analyses of the databases underlying the present manuscript found that distribution of OAB symptoms clearly deviates from normality and that means and medians differ systematically [26]. Therefore, we primarily report medians and IQR for OAB symptoms and have applied non-parametric statistical tests (which do not assume a normal distribution); however, to facilitate comparison with published data from other studies, we also report means ± SD.

The prevalence of reported comorbidities and comedications in the two studies reported here is higher than in previously reported NISs in OAB syndrome from Germany with propiverine [35] or other antimuscarinics such as darifenacin [32], fesoterodine [36], solifenacin [37] or tolterodine [31]. Whether this reflects a more careful documentation by physicians participating in the present study or a shift in patient populations over time cannot be determined with certainty based on the present data. However, overall improvements in symptoms were comparable in the present and previous NIS with muscarinic antagonists in OAB patients, confirming the conclusion from the previous meta-analysis of the RCT with these drugs [2]; comparable improvements have also been reported from a NIS with propiverine in male LUTS [35]. All of these aspects should be kept in mind in the interpretation of the present data.

4.2. Factors Associated with Initial Dosing

Propiverine ER differs from other muscarinic antagonists in having two approved starting doses. This allowed us to explore factors associated with starting dose selection by participating physicians, a question not addressed in any previous study. While subjects starting with the 45 mg dose had a slightly greater body weight and symptom severity, our logistic regression analyses identified age as the only variable consistently associated with the choice of starting dose. Thus, younger patients were more likely to start at the 45 mg. This may reflect the idea that young subjects may be less vulnerable to ADR of antimuscarinics [38]. On the other hand, factors such as duration of OAB syndrome, number of incontinence, voiding and nocturia episodes were associated with starting dose in one, but not the other, of the two studies. Moreover, all associations of these explanatory variables with the starting dose were of moderate strength only. This highlights the importance of testing for consistency of findings across multiple databases.

4.3. Factors Associated with Dose Escalation

Several previous studies have explored factors associated with a decision for dose escalation during treatment. This included both RCTs [15,17,18] and non-randomized studies [13,14,19]. In the present analyses, the number of urgency episodes at baseline and the change in incontinence episodes after 4 weeks of treatment were the only parameters consistently associated with the decision to increase the dose from 30 to 45 mg. In contrast, height, basal number of nocturia episodes and change thereof were associated with dose escalation in one, but not the other, study. Greater baseline symptoms, although not necessarily nocturia as in the present studies, had also been reported to be associated with a decision for dose escalation in some previous studies with other antimuscarinics [13,17,18],

but this was not confirmed in others [15]. Similarly, smaller OAB symptom improvements in the initial treatment period were associated with a decision to increase the dose in some studies [13,14,18] but not in others [15,17]. Adding to this heterogeneity in findings, dose escalation was associated with greater baseline values and/or smaller initial improvements sometimes for all OAB symptoms, but sometimes such as also observed in the present study for only selected OAB symptoms. Interestingly, a smaller initial treatment response was also reported to be associated with sham dose escalation within the placebo arm of RCT [22]. An inconsistency of reported associations with dose escalation also applies to factors other than OAB symptoms, including age ([14] and present studies), BMI ([14,15] and present studies), OAB duration ([17] and present studies), the presence of incontinence [13,19], symptom scores [13,15], and previous use of antimuscarinics [17]. Based on the overall evidence, we conclude that both greater baseline OAB severity and smaller improvements upon initial treatment tend to be associated with a decision for dose escalation. However, each of these associations appears to be too weak to be robustly detected across studies; this applies even more so if individual OAB symptoms are considered.

4.4. Factors Associated with Treatment Outcomes

While an understanding of allocation and escalation bias is interesting mechanistically, the clinically more relevant question is which factors are associated with treatment outcomes at study end. In the present studies, reductions in OAB symptoms were of comparable extent in the 30/30, 30/45 and 45/45 mg cohorts; if anything, the 45/45 cohort tended to have slightly greater and the 30/45 cohort slightly smaller efficacy as compared to the 30/30 group. For a more quantitative analysis, we have specifically looked at improvements of each of the four OAB parameters in separate logistic regression analyses. As previously reported from the NIS with other antimuscarinics [32], the strongest and most consistent contributor to improvement of a given symptom was the baseline value of that symptom. This reflects that the four symptoms are only moderately correlated to each other [31,39] and that a high baseline value for a given symptoms allows for a large reduction upon treatment. However, compared to the 45/45 group, starting and staying at 30 mg or escalating from 30 to 45 mg tended to have only minor and inconsistent effects achieving a p-value < 0.05 only for the 30/45 group for urgency and incontinence in study I but not for the other two outcome parameters within that study or for any outcome parameter in study II. While other studies have not applied such complex models to overall treatment outcomes, to the best of our knowledge, they have reported that patients with and without dose escalation achieved similar improvements of OAB symptoms at study end [13,14,18], although some reported slightly smaller improvements in escalators than in non-escalators [15].

5. Conclusions

Our analyses identified younger age as the only factor consistently associated with a greater starting dose, but the relative impact of age was small. Several drugs allow for a dose escalation if greater efficacy is desired and tolerability is adequate. While the overall evidence points to both greater baseline symptom intensity and smaller initial improvement, either effect is apparently too small to be detected consistently across studies, particularly when individual OAB parameters are considered. Most important, from a clinical perspective, is the observation in the present and in most previous studies that dose escalation in patients with an insufficient initial efficacy results in a symptom improvement comparable to that observed in patients exhibiting a greater initial improvement and staying on the lower dose or in those starting and staying on the higher dose. These data support the concept that dose escalation helps to achieve meaningful symptom improvements in patients where tolerability of the lower dose allows for a dose increase. In this regard, both dose strengths of propiverine had similar tolerability and discontinuations after the initial treatment period were similar with the starting doses of 30 and 45 mg.

Supplementary Materials: The following are available online at https://www.mdpi.com/2077-0383/10/2/311/s1, Supplementary file containing additional data related to both studies, data for study II where analogous data for study I are presented in main manuscript and data related to the primary aim of study II (exploration of role of additional material), Table S1: Demographic and OAB-related baseline variables in studies I and II; Table S2: Demographic and OAB-related baseline variables in patients starting treatment with a propiverine dose of 30 or 45 mg/d in study II; Table S3: Demographic and OAB-related variables at baseline and after 4 weeks in patients starting treatment with a propiverine dose of 30 mg/d and either staying on that dose after 4 weeks or increasing it 45 mg/d; Table S4: Factors associated with overall improvement of frequency (12 weeks vs. baseline) in a logistic regression analysis taking demographics, OAB parameters at baseline, duration of condition and dose level in to consideration; Table S5: Factors associated with overall improvement of nocturia (12 weeks vs. baseline) in a logistic regression analysis taking demographics, OAB parameters at baseline, duration of condition and dose level into consideration; Figure S1: Patient disposition in study I; Figure S2: Patient disposition in study II; Figure S3: Intra-individual change of OAB symptoms in the cohorts of patients starting on 30 mg and staying on that dose (30/30), starting on 30 mg and escalating to 45 mg at visit 2 after about 4 weeks (30/45) and starting and staying on 45 mg until study end after about 12 weeks (45/45) in study I; Figure S4: Intra-individual change of OAB symptoms in the cohorts of patients starting on 30 mg and staying on that dose (30/30), starting on 30 mg and escalating to 45 mg at visit 2 after about 4 weeks (30/45) and starting and staying on 45 mg until study end after about 12 weeks (45/45) in study II; Figure S5: Intra-individual change of OAB symptoms in the cohorts of patients starting on 30 mg and staying on that dose (30/30), starting on 30 mg and escalating to 45 mg at visit 2 after about 4 weeks (30/45) and starting and staying on 45 mg until study end after about 12 weeks (45/45) in study II.

Author Contributions: Conceptualization, S.M. and M.C.M.; methodology, M.A., S.M., M.C.M.; formal analysis, M.A. and M.C.M.; resources, S.M.; writing—original draft preparation, M.A. and M.C.M.; writing—review and editing, T.S., M.O. and S.M.; visualization, M.A. and M.C.M.; supervision, M.C.M.; project administration, S.M. and M.C.M. All authors have read and agreed to the published version of the manuscript.

Funding: The underlying studies and the APC were funded by Apogepha (Dresden, Germany).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of State Board of Physicians of Saxony, Germany (Sächsische Landesärztekammer EK-BR-14/12-1 and EK-BR-18/14-1).

Informed Consent Statement: Patient consent was waived due to non-interventional character of the study with only pseudonymized information being transferred for study purposes.

Data Availability Statement: Data are owned by Apogepha (Dresden, Germany). They will be made available to qualified non-commercial investigators upon request to the corresponding author.

Acknowledgments: We gratefully acknowledge the biometrical and statistical support by Stephan Bucher (Bioconsult GmbH, Rickenbach, Switzerland and thank all participating physicians and patients.

Conflicts of Interest: M.A. does not report a conflict of interest. T.S. is a lecturer for Allergan, Apogepha, Pfizer and Takeda. M.O. is a speaker and/or trial participant for Apogepha, Astellas, Dr. Willmar Schwabe, Ferring GSK, Omega Pharma, Pfizer, Pierre Fabere, SAJA Pharmaceuticals and SUN Pharmaceuticals. S.M. is an employee of Apogepha. In the urology field, M.C.M. is a consultant and/or lecturer to/for Apogepha, Astellas, Willmar Schwabe, Ferring, GSK and Velicept; he is also a shareholder of Velicept. Employees of the funder recruited participating physicians and distributed and collected case record forms. S.M. worked on the project as part of her employment by the funder. Other than that, the funder had no role in the analysis or interpretation of the data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Novara, G.; Galfano, A.; Secco, S.; D'Elia, C.; Cavalleri, S.; Ficarra, V.; Artibani, W. A systematic review and meta-analysis of randomized controlled trials with antimuscarinic drugs for overactive bladder. *Eur. Urol.* **2008**, *54*, 740–764. [CrossRef] [PubMed]
2. Reynolds, W.S.; McPheeters, M.; Blume, J.; Surawicz, T.; Worley, K.; Wang, L.; Hartmann, K. Comparative effectiveness of anticholinergic therapy for overactive bladder in women. A systematic review and meta-analysis. *Obstet. Gynecol.* **2015**, *125*, 1423–1432. [CrossRef] [PubMed]
3. Steers, W.D.; Corcos, J.; Foote, J.; Kralidis, G. An investigation of dose titration with darifenacin, an M_3-selective receptor antagonist. *BJU Int.* **2005**, *95*, 580–586. [CrossRef] [PubMed]
4. Khullar, V.; Rovner, E.S.; Dmochowski, R.; Nitti, V.; Wang, J.; Guan, E. Fesoterodine dose response in subjects with overactive bladder syndrome. *Urology* **2008**, *71*, 839–843. [CrossRef] [PubMed]
5. Jünemann, K.P.; Hessdörfer, E.; Unamba-Oparah, I.; Berse, M.; Brünjes, R.; Madersbacher, H.; Gramatté, T. Propiverine hydrochloride immediate and extended release: Comparison of efficacy and tolerability in patients with overactive bladder. *Urol. Int.* **2006**, *77*, 334–339. [CrossRef]
6. Cardozo, L.; Lisec, M.; Millard, R.; van Vierssen Trip, O.B.; Kuzmin, I.; Drogendijk, T.E.; Huang, M.; Ridder, A.M. Randomized, double-blind placebo-controlled trial of the once-daily antimuscarinic agent solifenacin succinate in patients with overactive bladder. *J. Urol.* **2004**, *172*, 1919–1924. [CrossRef]
7. Ruhe, H.G.; Booij, J.; van Weert, H.C.; Reitsma, J.B.; Fransen, E.J.F.; Michel, M.C.; Schene, A.H. Evidence why paroxetine dose-escalation is not effective in major depressive disorder: A randomized-controlled trial with assessment of serotonin transporter occupancy. *Neuropsychopharmacology* **2009**, *34*, 999–1010. [CrossRef]
8. Michel, M.C.; Staskin, D. Understanding dose titration: Overactive bladder treatment with fesoterodine as an example. *Eur. Urol. Suppl.* **2011**, *10*, 8–13. [CrossRef]
9. Michel, M.C.; Barendrecht, M.M. Physiological and pathological regulation of the autonomic control of urinary bladder contractility. *Pharmacol. Ther.* **2008**, *117*, 297–312. [CrossRef]
10. Fenech, A.G.; Billington, C.K.; Swan, C.; Richards, S.; Hunter, T.; Ebejer, M.J.; Felice, A.E.; Ellul-Micallef, R.; Hall, I.P. Novel polymorphisms influencing transcription of the human CHRM2 gene in airway smooth muscle. *Am. J. Respir. Cell Mol. Biol.* **2004**, *30*, 678–686. [CrossRef]
11. Ancelin, M.L.; Artero, S.; Portet, F.; Dupuy, A.M.; Touchon, J.; Ritchie, K. Non-degenerative mild cognitive impairment in elderly people and use of anticholinergic drugs: Longitudinal cohort study. *Br. Med. J.* **2006**, *332*, 455–458. [CrossRef] [PubMed]
12. Witte, L.P.W.; Mulder, W.M.C.; de la Rosette, J.J.M.C.H.; Michel, M.C. Muscarinic receptor antagonists for overactive bladder treatment: Does one fit all? *Curr. Opin. Urol.* **2009**, *19*, 13–19. [CrossRef] [PubMed]
13. Wyndaele, J.J.; Goldfischer, E.R.; Morrow, J.D.; Gong, J.; Tseng, L.J.; Choo, M.S. Patient-optimized doses of fesoterodine improve bladder symptoms in an open-label, flexible-dose study. *BJU Int.* **2011**, *107*, 603–611. [CrossRef] [PubMed]
14. Cardozo, L.; Hall, T.; Ryan, J.; Bitoun, C.E.; Kausar, I.; Darekar, A.; Wagg, A. Safety and efficacy of flexible-dose fesoterodine in British subjects with overactive bladder: Insights into factors associated with dose escalation. *Int. Urogynecol. J.* **2012**, *23*, 1581–1590. [CrossRef] [PubMed]
15. Wagg, A.; Darekar, A.; Arumi, D.; Khullar, V.; Oelke, M. Factors associated with dose escalation of fesoterodine for treatment of overactive bladder in people >65 years of age: A post hoc analysis of data from the SOFIA study. *Neurourol. Urodyn.* **2015**, *34*, 438–443. [CrossRef]
16. Wyndaele, J.J.; Schneider, T.; MacDiarmid, S.; Scholfield, D.; Arumi, D. Flexible dosing with fesoterodine 4 and 8 mg: A systematic review of data from clinical trials. *Int. J. Clin. Pract.* **2014**, *68*, 830–840. [CrossRef]
17. Goldman, H.B.; Oelke, M.; Kaplan, S.A.; Kitta, T.; Russell, D.; Carlsson, M.; Arumi, D.; Mangan, E.; Ntanios, F. Do patient characteristics predict which patients with overactive bladder benefit from a higher fesoterodine dose? *Int. Urogynecol. J.* **2018**, *30*, 239–244. [CrossRef]
18. Cardozo, L.; Amarenco, G.; Pushkar, D.; Mikulas, J.; Drogendijk, T.; Wright, M.; Compion, G.; Group, S.S. Severity of overactive bladder symptoms and response to dose escalation in a randomized, double-blind trial of solifenacin (SUNRISE). *BJU Int.* **2013**, *111*, 804–810. [CrossRef]
19. Chun, J.-Y.; Song, M.; Han, J.-Y.; Na, S.; Hong, B.; Choo, M.-S. Clinical factors associated with dose escalation of solifenacin for the treatment of overactive bladder in real life practice. *Int. Neurourol. J.* **2014**, *18*, 23–30. [CrossRef]
20. Shim, M.; Kim, J.K.; Bang, W.J.; Lee, Y.S.; Cho, S.T.; Cho, J.S.; Joo, K.J.; Hyun, J.S.; Kim, B.H.; Lee, J.B.; et al. Efficacy and safety of dose escalation in male patients with overactive bladder showing poor efficacy after low-dose antimuscarinic treatment: A retrospective multicenter study. *Investig. Clin. Urol.* **2020**, *61*, 600–606. [CrossRef]
21. Cardozo, L.; Khullar, V.; El-Tahtawy, A.; Guan, Z.; Malhotra, B.; Staskin, D. Modeling dose-response relationships of the effects of fesoterodine in patients with overactive bladder. *BMC Urol.* **2010**, *10*, 14. [CrossRef] [PubMed]
22. Staskin, D.R.; Michel, M.C.; Sun, F.; Guan, Z.; Morrow, J.D. The effect of elective sham dose escalation on the palcebo response during an antimuscarinic trial for overactive bladder symptoms. *J. Urol.* **2012**, *187*, 1721–1726. [CrossRef] [PubMed]
23. Wuest, M.; Hecht, J.; Christ, T.; Braeter, M.; Schoeberl, C.; Hakenberg, O.W.; Wirth, M.P.; Ravens, U. Pharmacodynamics of propiverine and three of its metabolites on detrusor contraction. *Br. J. Pharmacol.* **2005**, *145*, 608–619. [CrossRef]

24. Zhu, H.L.; Brain, K.L.; Aishima, M.; Shibata, A.; Young, J.S.; Sueishi, K.; Teramoto, N. Actions of two main metabolites of propiverine (M-1 and M-2) on voltage-dependent L-type Ca^{2+} currents and Ca^{2+} transients in murine urinary bladder myocytes. *J. Pharmacol. Exp. Ther.* **2008**, *324*, 118–127. [CrossRef] [PubMed]
25. Frazier, E.P.; Peters, S.L.M.; Braverman, A.S.; Ruggieri, M.R., Sr.; Michel, M.C. Signal transduction underlying control of urinary bladder smooth muscle tone by muscarinic receptors and ß-adrenoceptors. *Naunyn Schmiedebergs Arch. Pharmacol.* **2008**, *377*, 449–462. [CrossRef]
26. Amiri, M.; Murgas, S.; Stang, A.; Michel, M.C. Do overactive bladder symptoms and their treatment-associated changes exhibit a normal distribution? Implications for analysis and reporting. *Neurourol. Urodyn.* **2020**, *39*, 754–761. [CrossRef]
27. Dimpfl, T.; Kölbl, H.; Peschers, U.; Petri, E.; Gauruder-Burmester, A.; Höfner, K.; Schultz-Lampel, D.; Tamussino, K.; Heidler, H.; Schär, G. *The Overactive Bladder*; AWMF: Frankfurt, Germany, 2010.
28. Michel, M.C.; Murphy, T.J.; Motulsky, H.J. New author guidelines for displaying data and reporting data analysis and statistical methods in experimental biology. *J. Pharmacol. Exp. Ther.* **2020**, *372*, 136–147. [CrossRef]
29. van Leeuwen, J.H.S.; Castro, R.; Busse, M.; Bemelmans, B.L.H. The placebo effect in the pharmacologic treatment of patients with lower urinary tract symptoms. *Eur. Urol.* **2006**, *50*, 440–453. [CrossRef]
30. Cornu, J.N.; Abrams, P.; Chapple, C.R.; Dmochowski, R.R.; Lemack, G.E.; Michel, M.C.; Tubaro, A.; Madersbacher, S. A contemporary assessment of nocturia: Definitions, epidemiology, pathophysiology and management. A systematic review and meta-analysis. *Eur. Urol.* **2012**, *62*, 877–890. [CrossRef]
31. Michel, M.C.; de la Rosette, J.J.M.C.H.; Piro, M.; Schneider, T. Comparison of symptom severity and treatment response in patients with incontinent and continent overactive bladder. *Eur. Urol.* **2005**, *48*, 110–115. [CrossRef]
32. Schneider, T.; Marschall-Kehrel, D.; Hanisch, J.U.; Michel, M.C. Do gender, age or life style factors affect responses to anti-muscarinic treatment in overactive bladder patients? *Int. J. Clin. Pract.* **2010**, *64*, 1287–1293. [CrossRef] [PubMed]
33. Schneider, T.; de la Rosette, J.J.M.C.H.; Michel, M.C. Nocturia—A non-specific but important symptom of urological disease. *Int. J. Urol.* **2009**, *16*, 249–256. [CrossRef] [PubMed]
34. Michel, M.C.; Schumacher, H.; Mehlburger, L.; de la Rosette, J.J.M.C.H. Factors associated with nocturia-related quality of life in men with lower urinary tract symptoms and treated with tamsulosin oral controlled absorption system in a noniInterventional study. *Front. Pharmacol.* **2020**, *11*, 816. [CrossRef] [PubMed]
35. Oelke, M.; Murgas, S.; Baumann, I.; Schnabel, F.; Michel, M.C. Efficacy of propiverine ER with or without α-blockers related to maximum urinary flow rate in adult men with OAB: Results of a 12-week, multicenter, non-interventional study. *World J. Urol.* **2011**, *29*, 217–223. [CrossRef] [PubMed]
36. Schneider, T.; Arumi, D.; Crook, T.J.; Sun, F.; Michel, M.C. An observational study of patient satisfaction with fesoterodine in the treatment of overactive bladder: Effects of additional educational material. *Int. J. Clin. Pract.* **2014**, *68*, 1074–1080. [CrossRef]
37. Michel, M.C.; Wetterauer, U.; Vogel, M.; de la Rosette, J.J.M.C.H. Cardiovascular safety and overall tolerability of solifenacin in routine clinical use. *Drug Saf.* **2008**, *31*, 505–514. [CrossRef] [PubMed]
38. Wolff, G.F.; Kuchel, G.A.; Smith, P.P. Overactive bladder in the vulnerable elderly. *Res. Rep. Urol.* **2014**, *6*, 131–138. [CrossRef]
39. Michel, M.C.; Oelke, M.; Goepel, M.; Beck, E.; Burkart, M. Relationships among symptoms, bother, and treatment satisfaction in overactive bladder patients. *Neurourol. Urodyn.* **2007**, *26*, 190–195. [CrossRef]

3.3. Improving drug safety in hospitals: a retrospective study on the potential of adverse drug events coded in routine data.

BMC Health Services Research

RESEARCH ARTICLE Open Access

Improving drug safety in hospitals: a retrospective study on the potential of adverse drug events coded in routine data

Nils Kuklik[1,2*], Jürgen Stausberg[1], Marjan Amiri[2] and Karl-Heinz Jöckel[1]

Abstract

Background: Adverse drug events (ADEs) that occur during hospitalization are an ongoing medical concern. Systematic strategies for ADE identification are lacking. The aim of this study was to evaluate the potential to identify adverse drug events caused by medication errors (preventable ADEs, pADEs), and previously unknown adverse drug reactions (ADRs or non-preventable ADEs, npADEs) in inpatients by combining diagnosis codes in routine data with a chart review.

Methods: Diagnoses of inpatients are routinely coded using the International Classification of Diseases, 10th Revision (ICD-10). A total of 2326 cases were sampled from routine data of four hospitals using a set of ICD-10 German Modification ADE codes. Following a chart review, cases were evaluated in a standardized process with regard to drug relation and preventability of events.

Results: By chart review, 1302 cases were classified as hospital-acquired and included in the evaluation. This yielded 1285 cases indicating an ADE. 96.8% of ADEs (1244 ADEs) were classified as known npADEs, only three cases as suspected previously unknown npADEs, one case as event after drug abuse. A total of 37 ADEs were classified as preventable (2.9% of all ADEs) by identifying a medication error as probable cause. The prevalence of pADEs varied considerably between included ADE codes, with hemorrhagic diathesis due to coumarins and localized skin eruptions showing the highest rates (8.7 and 9.1%, respectively). Most frequent medication errors were non-compliance to a known allergy, and improper dose.

Conclusions: When focusing on specific ADE codes, routine data can be used as markers for npADEs and medication errors, thus providing a meaningful complement to existing drug surveillance systems. However, the prevalence of medication errors is lower than in former studies on the frequency of pADEs.

Keywords: Routine data, ICD-10, Adverse drug event, Medication error, Drug safety

Background

Patients often experience adverse drug events (ADEs) during hospitalization [1, 2]. Such inpatient ADEs pose a considerable health and economic burden on the patients as well as on the health care system [3–5]. A significant number of inpatient ADEs are caused by medication errors and can be prevented (pADEs) [1, 6, 7]. By release of the action plans for improvement of medication safety by the Federal Ministry of Health in Germany, various measures have been implemented and promoted over the past decade in order to prevent and identify ADEs [8]. In developed countries, hospitals increasingly use clinical decision support systems and computerized physician order entry systems to reduce prescription errors [9, 10].

To further improve drug safety it is crucial to overcome the lack of systematic detection and reporting of non-preventable adverse drug events (npADEs) and pADEs, and to perform an ongoing root cause analysis in order to identify factors that contribute to errors in hospitals [11, 12]. Spontaneous reporting systems and critical incident reporting systems (CIRS) for reporting ADEs are internationally established, but they suffer

* Correspondence: nils.kuklik@uk-essen.de
[1]Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg-Essen, Hufelandstr. 55, 45147 Essen, Germany
[2]Centre for Clinical Trials Essen (ZKSE), University Hospital Essen, University of Duisburg-Essen, Essen, Germany

from acceptance problems in daily routine [13]. Although the total number of spontaneous reports in Germany has been increasing for several years, the number is still low and the increase is mainly a result of higher reporting rates from pharmaceutical companies and patients [14].

Important data sources in hospitals are the diagnoses of inpatients routinely coded in Germany with the ICD-10 German Modification (ICD-10-GM) [15]. The codes are part of the hospital routine data, which are transmitted promptly to sickness funds and annually to the Institute for the Hospital Remuneration System as a standardized data set. Various studies have identified and validated ICD-10 codes as high-precision markers for the identification of ADEs (so-called ADE codes) [16–19]. It was further reported that 50% of inpatient ADEs are coded as disease in the routine data [19], including between 7 and 12% [18–20] that are coded as drug-related disease. Despite this moderate sensitivity, given the high precision and nationwide availability of ADE codes, routine data could complement existing pharmacovigilance systems and thereby contribute to the improvement of drug safety in hospitals.

Therefore, the aim of this study was to evaluate the potential of utilizing ADE codes encoded in routine data as a complementary drug safety source by identifying a) preventable ADEs including causes and contributing factors of medication errors, and b) previously unknown non-preventable ADEs, those that are not listed in the Summary of Product Characteristics (SmPC). The results of the study could stimulate the use of routine data as a pharmacovigilance resource.

Methods

Definitions

The following definitions are used allowing a clear distinction between non-preventable and preventable ADEs [21, 22]: an adverse drug event (ADE) is any harmful incident resulting from medical intervention related to a drug. ADEs are subdivided into non-preventable ADEs (npADEs) defined as harmful and unintended reactions to a drug after its appropriate use (adverse drug reaction), and preventable ADEs (pADEs), defined as harm to the patient due to errors in the drug treatment process. The definition of a npADE is consistent with the definition of an adverse reaction in ICD-10-GM, version 2018: adverse reaction of a drug that has been correctly prescribed and properly administered [15].

Study design and database

We conducted a retrospective, multicenter, observational study with an explorative approach using secondary data analysis to identify preventable and non-preventable harm in inpatients. Hospital discharge data from four full-service, non-academic hospitals in Germany of the calendar years 2015 and 2016 were used. The hospitals are located in cities with a population between 45,000 and 180,000 and together operate 2300 beds and treat 109,000 inpatients. The routine data contain inpatient conditions coded by ICD-10-GM with one principal diagnosis and several additional diagnoses. The principal diagnosis is defined as “that condition established after study to be chiefly responsible for occasioning the admission of the patient to the hospital for care”, whereas additional diagnoses are defined as “all conditions that coexist at the time of the principal diagnosis, or that develop during the hospital stay” [15]. Since the focus of this study was on hospital-acquired ADEs (nosocomial conditions acquired during hospitalization) and as by definition hospital-acquired events cannot be assigned as principal diagnosis, only additional diagnoses were included. However, because additional diagnoses include comorbidities present at admission as well as hospital-acquired complications, and because the ICD-10-GM does not contain a Present on Admission (POA) indicator, the time of occurrence of corresponding events was determined during the chart review process.

Sample selection

In previous studies of the authors, the general suitability of ICD-10 codes for ADE identification was investigated by calculating prevalence, precision and sensitivity of ICD-10 additional diagnosis codes [19, 23]. Based on these results, ADE codes were selected for evaluation in this study if they a) indicate predominantly hospital-acquired events, b) have been validated as codes representing ADEs with high precision (positive predictive values 68 to 94%, see [19]), and c) occurred more frequently compared to other ADE codes. One ADE code identifies one inpatient stay and is defined as one observational unit (hereinafter called “case”). In each hospital, all cases in each code group that fulfilled inclusion criteria were independently retrieved from the respective routine data, resulting in a sample of 2326 cases. A case was identified in the hospital information system by the patient identifier linked to the ICD-10 code representing the ADE. Then, the patient chart was retrieved either in electronic or in paper-based format. Table 1 shows the included codes and the number of cases.

Data extraction and evaluation

Data for the analysis were recorded and evaluated in a multi-level, standardized procedure (Fig. 1). First, experienced personnel with medicinal and pharmaceutical background performed the chart review and documentation of event characteristics at the hospital’s site after completing a two-month study-specific training. To assure the data quality in the chart review process,

Table 1 Included ICD-10-GM codes and number of cases

Code/Code group	Term	*N*
A04.7	Enterocolitis due to *Clostridium difficile*	362
D68.33; D68.34; D68.35	Hemorrhagic diathesis due to coumarins (vitamin K antagonists)/due to heparins/due to other anticoagulants	488
D69.52; D69.53	Secondary thrombocytopenia: Heparin induced thrombocytopenia type I/II	114
I95.2	Hypotension due to drugs	563
K52.1	Toxic gastroenteritis and colitis	155
L27.0; L27.1	Generalized/localized skin eruption due to drugs and medicaments	506
N99.0	Postprocedural renal failure	138
Total		2326

multiple on-site visits were carried out. Starting point for a chart review was the diagnosis of one of the ADE codes listed in Table 1. After identification of the event in the medical record, data regarding the time point of the event, the relationship of the event to a drug if reported by the hospital staff, patient characteristics, drugs taken before event, patient's known allergies and comorbidities, and source of information (physician or nurse) were extracted from the medical charts and the hospital information system. Data were recorded on a

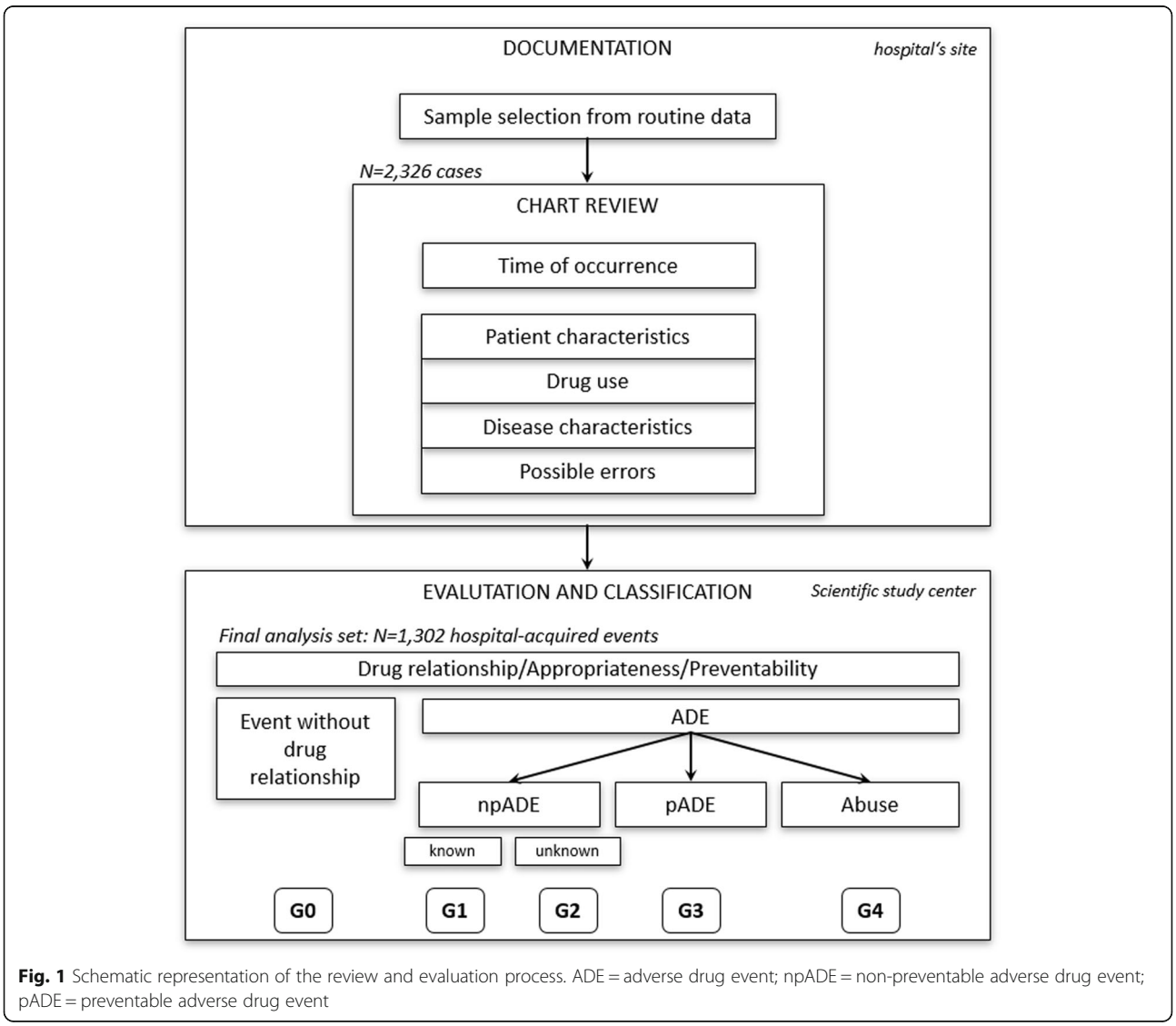


Fig. 1 Schematic representation of the review and evaluation process. ADE = adverse drug event; npADE = non-preventable adverse drug event; pADE = preventable adverse drug event

standardized data collection form. The parameters to be collected to identify and characterize a medication error were adopted from the taxonomy of medication errors from the National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP, USA) [24].

At the scientific study center, the completed data collection forms were evaluated. All cases that occurred during hospitalization were classified according to type of adverse event. The underlying root cause of the occurrence of the event was assessed and related drugs to the ADE were assigned by taking into account the drug relationships recorded by the hospitals and by inspecting the SmPCs of administered drugs. A preventability assessment of the ADEs was performed to distinguish pADEs from npADEs by considering errors recorded by the hospital staff and by comparing collected data with the SmPC's application instructions. A clinical decision support system (Software ID DIACOS PHARMA; ID Information und Dokumentation im Gesundheitswesen GmbH, Berlin) was used to support the assessment process. The following groups were defined to categorize the ADEs.

- **G0 – event without drug relationship:** Adverse event for which no related drug therapy was identified.
- **G1 – known non-preventable ADE:** ADE after proper use without indication of a medication error; ADE listed in the SmPC.
- **G2 – suspected previously unknown non-preventable ADE:** ADE after proper use without indication of a medication error; ADE not listed in the SmPC.
- **G3 – preventable ADE:** ADE with medication error as probable cause for reported event.
- **G4 – drug abuse:** ADE because of drug abuse by the patient.

No personnel of the study center was involved in the data extraction process at the hospital's site. At the scientific study center, cases were evaluated and categorized by author MA. All assessments were reviewed by author NK. For cases assigned to G2 and G3, a final consensus agreement was achieved by authors JS and NK. Absolute and relative frequencies and exact 95% confidence intervals (CI) were calculated using SAS (SAS Institute Inc., Release 9.4).

Results

By chart review, data from 2326 cases were extracted. Of the reviewed cases, 1328 cases encoded events that occurred during hospitalization (57.1% of 2326), 747 cases represented events present at admission, and 251 with unknown onset date. Overall, 26 cases were excluded from the set of hospital-acquired events due to incomplete data or implausibility. Therefore, 1302 cases were included in the final analysis and assigned to groups G0 - G4 (Table 2).

Most of the cases were confirmed as ADE (G1-G4). A small percentage of the codes enterocolitis due to Clostridium difficile (A04.7-), and postprocedural renal failure (N99.0) represented events without drug relationship (G0). While 90 to 100% of cases across all codes were classified as known npADEs, only three cases were classified as suspected previously unknown npADEs, i.e. the ADE was not listed in the SmPC: Eliquis (active ingredient: Apixaban) associated with toxic gastroenteritis and colitis (K52.1), and Valoron (active ingredient: Tilidine; two cases) associated with localized skin eruption (L27.1).

A total of 37 cases (2.9% of all 1285 ADEs) represented pADEs. pADEs were identified in association with the ADE codes hemorrhagic diathesis due to coumarins (D68.33), hypotension due to drugs (I95.2), generalized and localized skin eruption (L27.0 and L27.1), and postprocedural renal failure (N99.0). Among pADEs, the codes D68.33, L27.0, and L27.1 showed the highest rates. One case with the ADE code hypotension due to drugs (I95.2) was related to drug abuse by the patient.

Out of the 37 cases with pADEs, 28 medication errors were related to skin eruptions. The non-compliance to a known allergy (27 cases) was the most frequent type of medication error (Table 3). Of these, 24 cases were associated with allergies to antibiotics. Improper dosing was rarely observed (seven cases). Possible causes and contributing factors could only be identified for a small proportion of medication errors.

Discussion

In our study, codes of the ICD-10-GM (ADE codes) were analyzed to assess their potential for the detection of pADEs and previously unknown npADEs. As observed in the preceding validation study [19], the selected ADE codes represented high-precision markers for drug-related conditions that, with the exception of hemorrhagic diatheses, by the majority developed during hospitalization. These codes are thus suitable for the analysis of hospital-acquired ADEs.

The evaluation of the ADE codes revealed no evidence of medication errors in the vast majority of cases. Only 2.9% of all ADEs (G1-G4) were classified as probable consequences of medication errors and therefore as preventable (pADEs). However, the prevalence of pADEs varied significantly between ADE codes, ranging from 0 to 9.1%. In particular, higher rates were found for the ADE codes hemorrhagic diathesis associated with administration of vitamin K antagonists (8.7%), and skin eruptions (9.1%), mostly due to antibiotics. Both drug groups are frequently reported in association with

Table 2 Classification of ADEs per ADE code: absolute and relative frequencies

Group	N[%] cases per ADE code												
	A04.7	D68.33	D68.34	D68.35	D69.52	D69.53	I95.2	K52.1	L27.0	L27.1	N99.0	Total	95% CI
G0 - event without drug relationship	15 [7.1]	0	0	0	0	0	0	0	0	0	2 [1.9]	17 [1.3]	0.8–2.1
G1 - known npADE	195 [92.9]	42 [91.3]	7 [100]	12 [100]	9 [100]	41 [100]	401 [99.0]	87 [98.9]	209 [93.7]	138 [89.6]	103 [96.2]	1244 [95.6]	94.3–96.6
G2 – previously unknown npADE	0	0	0	0	0	0	0	1 [1.1]	0	2 [1.3]	0	3 [0.2]	0.1–0.7
G3 - pADE	0	4 [8.7]	0	0	0	0	3 [0.7]	0	14 [6.3]	14 [9.1]	2 [1.9]	37 [2.8]	2.0–3.9
G4 – event after drug abuse	0	0	0	0	0	0	1 [0.3]	0	0	0	0	1 [0.1]	0–0.4
Total	210	46	7	12	9	41	405	88	223	154	107	1302	

npADE non-preventable adverse drug event, *pADE* preventable adverse drug event, *CI* confidence interval

hospital-acquired medication errors [7, 25, 26]. Former studies found higher rates of hospital-acquired pADEs compared to the results presented. For example, a prospective study at two hospitals in the Netherlands reported a rate of 5% inpatient pADEs [22], a prospective study in the UK found a pADE rate of 52%, and classified 47% of them as “possible” and 5% as “definite” [25]. One meta-analysis reviewing eight prospective studies from the years 1994–2010 [6] assessed 45% of all hospital-acquired ADEs to be preventable, whereas another meta-analysis evaluating nine prospective and retrospective studies from the years 2006 to 2014 [7] reported 32% pADEs. However, differences in methodology and study population complicate the comparison of the results. A continuous improvement of quality standards in the drug therapy process and a more frequent use of electronic systems contribute to a reduction of preventable adverse events [9, 10]. This might explain the rarity of pADEs determined in this study, indicating a possible overestimation of the burden of medication errors in the current discussion on drug safety in the inpatient setting. However, considering the total number of inpatients in Germany and high percentage of ADE codes, rates of pADEs as determined in this study still demonstrate the ongoing relevance of drug safety improvement.

Possible causes and contributing factors of medication errors could only be determined in a few cases. Hospital staff related human factors such as heavy workload, transmission errors between documents, and communication deficits could be identified. To increase the patient’s safety, a systematic root cause analysis of medication errors at hospitals is essential in order to identify conditions in which medication errors are favored, to initiate structural changes to remedy them, and to define and optimize specific workflows. These measures have received increasing attention in recent years, for example through implementation of CIRS [27] or the formulation of standard operating procedures [28]. In total, three cases of suspected previously unknown npADEs were identified. The low number and lack of information on the actual frequency of previously

Table 3 Types, causes and factors of medication errors

ADE code	Type of medication error	*N*	Causes/Factors	*N*
D68.33	Wrong time of administration	1	–	
	Improper dose	3	Heavy workload	1
I95.2	Improper dose	3	Verbal miscommunication	1
L27.0	Contraindication, known allergy (antibiotic)	14	Transcription error Written miscommunication	2 1
L27.1	Contraindication, known allergy (antibiotic)	10	Transcription error Written miscommunication	2 1
	Contraindication, known allergy (analgesic)	2	Verbal miscommunication	1
	Contraindication, known allergy (heparin)	1	–	
	Improper dose	1	–	
N99.0	Contraindication, comorbidity	1	–	
	Contraindication, drug-drug interaction	1	–	
	Total	37		

ADE adverse drug event

unknown ADEs in hospitals hampers a final qualitative assessment of the usability of routine data in this context. Therefore, the potential of routine data for the detection of previously unknown npADEs cannot be conclusively derived. A validation of the prevalence having a larger sample is recommended.

Limitations in the interpretation of the presented results can be discussed at different levels. Generally, routine data have only a moderate sensitivity for inpatient ADEs. As reported in the preceding validation study, 50% of hospital-acquired ADEs were coded as disease in the routine data, from which a subgroup of 12% was coded as drug-associated disease [19]. A possible impact of under-reporting of ADEs in routine data on the rate of pADEs was not verified in this study. It can be argued that clinical personnel may be reluctant to code events related to medication errors and that there is a lack of information in the source data. On the other hand, this effect may be compensated because the severity and relevance of pADEs may in turn lead to higher coding rates. Therefore, taking into account the impact of under-reporting of pADEs but also the tendency to code ADEs with high severity more frequently, there is no evidence that the sensitivity of ADE codes indicating medication errors is lower than of ADE codes encoding non-preventable ADEs. The suspected medication errors and previously unknown npADEs identified in this work are distributed over a small set of ADE codes. Although the most frequent ADE codes were included in the analysis, it is not easily possible to generalize the prevalence rates determined in this study to other codes. The hospitals in this study have no specific characteristics. The evaluation based on nationwide uniform ICD-10 codes that are coded according to standardized guidelines [15]. Therefore, a generalization of the results to other hospitals in Germany is reasonable. However, due to possible structural differences in different countries with regard to the pharmacovigilance infrastructure, a generalizability to other countries is only possible to a limited extent. Data on the frequencies of additional diagnoses in Germany show that unspecific codes are regularly used to code events [23]. This includes codes such as T78.4 “Other and unspecified allergy” and T88.7 “Unspecified adverse effect of drug or medicament” - codes which do not directly identify the underlying event and which were therefore excluded from the study. Further studies are necessary to validate the impact of these codes on the rate of hospital-acquired pADEs.

Conclusion

Detection of pADEs and previously unknown npADEs in everyday clinical practice is a major challenge in healthcare. Our study confirmed the potential of utilizing ADE codes encoded in routine data as a complementary drug safety source. Furthermore, our data indicated that pADEs occur less frequently than expected. The majority of npADEs were mentioned in the SmPCs of related drugs.

The Drug Commission of the German Medical Association is currently developing a reporting system to systematically collect and evaluate medication errors within the framework of the spontaneous reporting system for ADRs [29]. To address the under-reporting of ADEs, additional strategies to collect drug safety data are needed. Having a comprehensive and standardized acquisition, routine data can be effectively used as a complementary data source to detect medication errors. Our results demonstrate that the majority of ADEs coded in routine data are known npADEs. However, using routine data as markers for pADEs in combination with chart review is reasonable when focusing on specific ICD-10 codes. In a study from South Korea, ADE codes from nationwide routine data have been used as a basis to evaluate drug safety following the realization of an electronic drug prescription system [30]. Furthermore, pADEs coded in routine data can provide important information for systematic prospective quality assessments in hospitals in order to implement preventive, risk-reducing measures in hospital management. One important step towards greater use of routine data in drug safety is the identification of further, suitable ADE codes [31]. The implementation of a POA indicator in the German version of the ICD-10, a more strict specification of medication error coding in routine data, and not least raising awareness of ADE coding in hospitals can further increase the potential of routine data within the framework of drug safety.

Abbreviations

ADE: Adverse Drug Event; ADR: Adverse Drug Reaction; CIRS: Critical Incident Reporting System; ICD-10-GM: International Classification of Diseases, 10th Revision, German Modification; npADE: non-preventable Adverse Drug Event; pADE: preventable Adverse Drug Event; POA: Present on Admission; SmPC: Summary of Product Characteristics

Acknowledgements

The authors thank the participating hospitals for providing the routine data and for performing the chart review.

Authors’ contributions

NK, JS and KHJ designed the study. MA, JS and NK performed the data evaluation and interpretation. NK performed the statistical analysis and drafted the manuscript. All authors participated in the critical revision of the manuscript and approved the final manuscript for submission.

Funding

The work was funded by the Federal Ministry of Education and Research (BMBF, funding code 01GY1328). This authority played no role in the collection, analysis and interpretation of the data or in the decision to submit the manuscript for publication.

Availability of data and materials

The data that support the findings of this study are available from the participating hospitals but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly

available. Data are however available from the authors upon reasonable request and if no interests of the participating hospitals are affected.

Ethics approval and consent to participate

The study was conducted in accordance with national law and the 1964 Helsinki declaration and its later amendments, and according to the recommendations of the guidelines on Good Epidemiological Practice [32]. Ethical approval was given by the institutional review board of the University Duisburg-Essen.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 2 January 2019 Accepted: 30 July 2019
Published online: 08 August 2019

References

1. de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review. Qual Saf Health Care. 2008;17:216–23.
2. Martins ACM, Giordani F, Rozenfeld S. Adverse drug events among adult inpatients: a meta-analysis of observational studies. J Clin Pharm Ther. 2014; 39:609–20.
3. Hug BL, Keohane C, Seger DL, Yoon C, Bates DW. The costs of adverse drug events in community hospitals. Jt Comm J Qual Patient Saf. 2012;38:120–6.
4. Jha AK, Larizgoitia I, Audera-Lopez C, Prasopa-Plaizier N, Waters H, Bates DW. The global burden of unsafe medical care: analytic modelling of observational studies. BMJ Qual Saf. 2013;22:809–15.
5. Rottenkolber D, Hasford J, Stausberg J. Costs of adverse drug events in German hospitals-a microcosting study. Value Health. 2012;15:868–75.
6. Hakkarainen KM, Hedna K, Petzold M, Hagg S. Percentage of patients with preventable adverse drug reactions and preventability of adverse drug reactions--a meta-analysis. PLoS One. 2012;7:e33236.
7. Laatikainen O, Miettunen J, Sneck S, Lehtiniemi H, Tenhunen O, Turpeinen M. The prevalence of medication-related adverse events in inpatients-a systematic review and meta-analysis. Eur J Clin Pharmacol. 2017;73:1539–49.
8. Arzneimittelkommission der deutschen Ärzteschaft: Aktionspläne zur Verbesserung der Arzneimitteltherapiesicherheit in Deutschland [in German]. www.akdae.de/AMTS/Aktionsplan/index.html. Accessed 08 Jul 2019.
9. Prgomet M, Li L, Niazkhani Z, Georgiou A, Westbrook JI. Impact of commercial computerized provider order entry (CPOE) and clinical decision support systems (CDSSs) on medication errors, length of stay, and mortality in intensive care units: a systematic review and meta-analysis. J Am Med Inform Assoc. 2017;24:413–22.
10. Varghese J, Kleine M, Gessner SI, Sandmann S, Dugas M. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. J Am Med Inform Assoc. 2017;0:1–10.
11. Carnovale C, Mazhar F, Pozzi M, Gentili M, Clementi E, Radice S. A characterization and disproportionality analysis of medication error related adverse events reported to the FAERS database. Expert Opin Drug Saf. 2018; 17:1161–9.
12. Morrison M, Cope V, Murray M. The underreporting of medication errors: a retrospective and comparative root cause analysis in an acute mental health unit over a 3-year period. Int J Ment Health Nurs. 2018;27:1719–28.
13. Alatawi YM, Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration adverse event reporting system (FAERS). Expert Opin Drug Saf. 2017;16:761–7.
14. Dubrall D, Schmid M, Alešik E, Paeschke N, Stingl J, Sachs B. Frequent adverse drug reactions, and medication groups under suspicion—a descriptive analysis based on spontaneous reports to the German Federal Institute for Drugs and Medical Devices from 1978–2016. Dtsch Arztebl Int. 2018;115:393–400.
15. German Institute for Medical Documentation and Information (DIMDI): ICD-10, German Modification (ICD-10-GM). https://www.dimdi.de/dynamic/de/klassifikationen/icd/icd-10-gm/. Accessed 08 Jul 2019.
16. Hodgkinson MR, Dirnbauer NJ, Larmour I. Identification of adverse drug reactions using the ICD-10 Australian modification clinical coding surveillance. J Pharm Pract Res. 2009;39:19–23.
17. Hohl CM, Karpov A, Reddekopp L, Doyle-Waters M, Stausberg J. ICD-10 codes used to identify adverse drug events in administrative data: a systematic review. J Am Med Inform Assoc. 2014;21:547–57.
18. Hougland P, Xu W, Pickard S, Masheter C, Williams SD. Performance of international classification of diseases, 9th revision, clinical modification codes as an adverse drug event surveillance system. Med Care. 2006;44: 629–36.
19. Kuklik N, Stausberg J, Jöckel KH. Adverse drug events in German hospital routine data: a validation of international classification of diseases, 10th revision (ICD-10) diagnostic codes. PLoS One. 2017;12:e0187510.
20. Hohl CM, Kuramoto L, Yu E, Rogula B, Stausberg J, Sobolev B. Evaluating adverse drug event reporting in administrative data from emergency departments: a validation study. BMC Health Serv Res. 2013;13:473.
21. World Health Organization (WHO): Conceptual Framework for the International Classification for Patient Safety 2009. www.who.int/patientsafety/taxonomy/icps_full_report.pdf. Accessed 08 Jul 2019.
22. Dequito AB, Mol PG, van Doormaal JE, Zaal RJ, van den Bemt PM, Haaijer-Ruskamp FM, et al. Preventable and non-preventable adverse drug events in hospitalized patients: a prospective chart review in the Netherlands. Drug Saf. 2011;34:1089–100.
23. Stausberg J, Hasford J. Drug-related admissions and hospital-acquired adverse drug events in Germany: a longitudinal analysis from 2003 to 2007 of ICD-10-coded routine data. BMC Health Serv Res. 2011;11:134.
24. National Coordinating Council for Medication Error Reporting and Prevention (NCC-MERP): Taxonomy of Medication Errors. www.nccmerp.org/sites/default/files/taxonomy2001-07-31.pdf. Accessed 08 Jul 2019.
25. Davies EC, Green CF, Taylor S, Williamson PR, Mottram DR, Pirmohamed M. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. PLoS One. 2009;4:e4439.
26. Ducharme MM, Boothby LA. Analysis of adverse drug reactions for preventability. Int J Clin Pract. 2007;61:157–61.
27. Hubertus J, Piehlmeier W, Heinrich M. Communicating the improvements developed from critical incident reports is an essential part of CIRS. Klin Padiatr. 2016;228:270–4.
28. Leotsakos A, Zheng H, Croteau R, Loeb JM, Sherman H, Hoffman C, et al. Standardization in patient safety: the WHO high 5s project. Int J Qual Health Care. 2014;26:109–16.
29. 'Aus der UAW-Datenbank' - Nebenwirkungen durch Medikationsfehler [in German]. Dtsch Arztebl 2016; 113: A-1948/B-1636/C-1624.
30. Lee J, Noh Y, Lee S. Evaluation of preventable adverse drug reactions by implementation of the nationwide network of prospective drug utilization review program in Korea. PLoS One. 2018;13:e0195434.
31. Amelung S, Meid AD, Nafe M, Thalheimer M, Hoppe-Tichy T, Haefeli WE, et al. Association of preventable adverse drug events with inpatients' length of stay-a propensity-matched cohort study. Int J Clin Pract. 2017;71:e12990.
32. Hoffmann W, Latza U, Terschuren C, GSf E. Guidelines and recommendations for ensuring good epidemiological practice (GEP) -- revised version after evaluation. Gesundheitswesen. 2005;67:217–25.

Publisher's Note

3.4. Statistical inference in abstracts of three influential clinical pharmacology journals analyzed using a text-mining algorithm.

Note: This article has been accepted for publication in British journal of Clinical Pharmacology and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process. Due to page limits for this dissertation supplemetary tables has not been presented. For further information please refer to doi: 10.1111/bcp.14836.

Statistical inference in abstracts of three influential clinical pharmacology journals analyzed using a text-mining algorithm

Marjan Amiri, [1, 2] Markus Deckert, [3] Martin C. Michel, [4, 5] Charles Poole, [6] Andreas Stang, [1, 3, 7]

(1) Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

(2) Centre for Clinical Trials Essen (ZKSE), University Hospital Essen, University of Duisburg-Essen, Essen, Germany

(3) Center of Clinical Epidemiology; c/o Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

(4) Dept. of Pharmacology, Johannes Gutenberg University, Mainz, Germany

(5) Partnership for the Assessment and Accreditation of Scientific Practice, Heidelberg, Germany

(6) Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, CB #7435, 135 Dauer Drive, Chapel Hill, NC 27599-7435, USA

(7) School of Public Health, Department of Epidemiology, Boston University, Boston, USA

Corresponding author

Andreas Stang, MD, MPH; Center of Clinical Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, Germany; phone +49 201 723-77-289; fax +49 201 723-77-333; email: imibe.dir@uk-essen.de

Key words: Models, statistical; data interpretation; data analysis

Word count: 7653

Table count: 6

Figure count: 2

Abstract

Aim: To describe the trend in the prevalence of statistical inference in three influential clinical pharmacology journals

Methods: We applied a computer-based algorithm to abstracts of three clinical pharmacology journals published in 1976 to 2016 to identify statistical inference and its subtypes. Furthermore, we manually reviewed a random sample of 300 articles to access algorithm's performance in finding statistical inference in abstracts and as a screening tool for presence and absence of statistical inference in full text.

Result: The algorithm identified 59% (13,375/22,516 [mid p 95% CI, 59%-60%]) article abstracts with statistical inference. The percentage of abstracts with statistical inference was similar in 1976 and 2016, 48% (179/377 [mid p 95%CI, 42%-52%]) versus 49% (386/791 [mid p 95%CI, 45%-52%]). Statistical reporting pattern varied among journals. Among abstracts containing any statistical inference in the publications from 1976 to 2016 null-hypothesis significance testing was the most prevalent reported statistical inference. The algorithm had high sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for finding statistical inferences in abstract. While PPV for predicting the statistical inference in full text (including abstract, text, tables and figures) was high, NPV was low.

Conclusion:

Despite journal's editorials and statistical associations' guidelines, most authors focused on testing rather than estimation. In future, a better statistical reporting might be ensured by improving the statistical knowledge of authors and an addition of statistical guides to journals' instruction to authors to the extent that editors would like their statistical inference preferences to be incorporated into submitted manuscripts.

Introduction

Clinical pharmacology, like all other parts of drug research and development consists of a wide range of exploratory and confirmatory investigations. The former develop whereas the latter test hypotheses. The statistical investigation of hypotheses leads to either statistical significance testing (ST) or to null- hypothesis testing (NHT). ST and NHT are nowadays frequently mixed up or combined into a hybrid that has been called “null hypothesis significance testing” (NHST) [1-3]. The increasing popularity of NHST resulted in an increasing trend in reporting p-values [4].

Criticisms of dichotomizing p-values (i.e., of NHT) is old. As early as 1921, Boring warned against decisions based on categorized p-values [5]. Stigler concluded that there is ample evidence that the abuse of statistical significance even predated Fisher [6]. In 1988, the International Committee of Medical Journal Editors (ICMJE) updated its recommendation to authors in favor of additional reporting of confidence intervals (CI) [7]. Subsequently, a number of medical journals published editorials discouraging the use of p-values [8-10]. A count of over 300 warnings of limitations of ST, NHT and NHST in 2000 [11] was followed a year later by a list of 402 references [12], among which we found 89 in biomedical publications. Rothman described scientific inference as “a thoughtful process that pivots on measurements, whereas statistical significance testing is a mechanical process that debases measurements into the qualitative and sometimes misleading categories of ‘significant’ and ‘not significant’“ [13].

Recently, several high-impact publications have triggered a lively debate about dichotomization of p-values. In 2016, the American Statistical Association (ASA) published its first critical statement about the use of NHT in research since 1893, the year of its inception. It published six principles “that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community“. Some explanatory notes accompanied each principle. For example, for statement 3, the ASA stated “the widespread use of ‘statistical significance’ (generally interpreted as “$p \leq 0.05$”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process” [14]. More recently, ASA leadership issued an even stronger statement “that it is time to stop using the term ‘statistically significant’ entirely. Nor should variants such as ‘significantly different’, ‘$p < 0.05$’, and ‘nonsignificant’ survive, whether expressed in words, by asterisks in a table or in some other way. Regardless of whether it was ever useful, a declaration of ‘statistical significance’ has today become meaningless” [15].

Furthermore, methodologists of statistics and epidemiological methods published a paper that illustrates 25 misconceptions related to statistical tests, p-values, confidence intervals and statistical power [1]. In 2019, a paper entitled “Retire statistical significance” was published in Nature and accompanied by a published list of signatories to “Retire statistical significance” including 854 scientists from 52 countries, with 20 scientists (2.3%) from Germany [16] including one of the present authors (M.C.M.). Contemporary defenses of

statistical significance are comparatively rare and include a proposal to lower the level of significance to $\alpha=0.005$ [17].

These developments lead to hypotheses that NHT might be losing popularity, and that ST and CI reporting might be increasing, in recent years. A review of abstracts in epidemiological, medical, and psychiatry journals did find an increasing trend in CI reporting, while NHT reporting was the dominant statistical inference in abstracts containing statistical inference [18, 19]. However, to our knowledge data on the time trend in reporting statistical inferences in abstracts published in pharmacology journals are missing. As clinical pharmacology has a stronger track record of reporting CI as compared to experimental pharmacology [10], we explored how clinical pharmacology journals have approached the reporting of statistical inference over time. To this end, we performed a review of abstracts of publications published in 1976 to 2016 in three influential clinical pharmacology journals as an illustrative sample of the field and described their time trend in reporting statistical inference.

Material and Methods

Journal extraction

We systematically reviewed publications from three major clinical pharmacology journals with different impact factors (IF) based on 2017 journal citation reports (JCR) from the Web of Science® website (www.webofknowledge.com): *British Journal of Clinical Pharmacology* (Br J Clin Pharmacol, IF = 3.8), *Clinical Pharmacology and Therapeutics* (Clin Pharmacol Ther, IF = 6.5), and the *European Journal of Clinical Pharmacology* (Eur J Clin Pharm, IF = 2.7).

The website's source data table presents the number of citable items (research articles and reviews) and other items (editorials, letters, news items, and meeting abstracts) in the JCR year. We examined and compared trends in each journal's source data for the period 1976 through 2016 because changes of editorial policies and changes in statistical doctrines can influence reporting of statistical inference.

We also checked whether each journal is committed to the recommendations of the ICMJE and reviewed the ICMJE archives for changes in relevant recommendations. Furthermore, we asked the journal editors if they archive the instructions or other guidance they provide to authors to check relevant changes. In a post-hoc analysis, we examined their most recent online instructions to the authors available at their website (access date: 12.06.2020).

We searched for and retrieved all article abstracts from the PubMed database (http://www.ncbi.nlm.nih.gov/gov/pubmed/) on May 31, 2017 for the period 1976 through 2016. We started this review on 2017. Therefore, we extracted the data up to 2016. Because previous studies revealed that abstracts regularly appeared not earlier than in the mid-70s, the observation periods starts in 1976 [18, 19]. Raw data regarding the publication year 1976 and 2016 is provided in online supplement (Table S5 and S6). We then used a SAS text-mining algorithm that has been developed and validated in two previous projects [18, 19]. This algorithm automatically identifies the presence of statistical inference by employing the index function of SAS to search for the letter string "signif", "power", "inferior", "equiv", p-values, p-thresholds and CIs. In a post-hoc analysis, we categorized the abstracts according to the publication types, which are assigned by PubMed.

Definition of measures of statistical inference

We defined statistical inference as the presence of any of the following items: p-thresholds (e.g. $p \leq 0.01$, $p > 0.05$), p-values (e.g. $p = 0.03$, $p = 0.84$), significance terminology (e.g. "significant association", "significant difference"), terminology related to NHT ("power", "inferior", "equiv") and CIs. Based on the type of statistical inference reported in the abstracts, we classified them into the following sub-groups:

1. Significance testing only , "ST only": reporting only p-values such as 0.0348

2. Null- hypothesis testing only, “NHT only”: reporting only p-thresholds or at least one term that indicates null- hypothesis testing such as “signif”, “power”, inferior”, “equiv”),
3. Null- hypothesis testing and significance testing , “NHT & ST” : abstracts with null-hypothesis and significance testing regardless of the presence of CIs,
4. “any NHST”: abstracts with significance testing OR null-hypothesis testing regardless of the presence of CIs,
5. “any CI”: abstracts with any CI regardless of NHST,
6. “CI only”: abstracts with CI without any NHST.

Statistical methods

We estimated the prevalence of reporting any statistical inference and, among the abstracts with any statistical inference, the prevalence of different sub-groups as defined above. We looked for all instances of statistical inference and not merely those pertaining to a paper’s primary analysis – which often would be difficult to discern in any event. We performed all statistical analysis using SAS (version 9.4; SAS Institute INC., Cary, NC, USA). We estimated time trends by weighted nonparametric local regression smoothing [20, 21] with quadratic local polynomials and a smoothing parameter of 0.6, which means that 60% of the data in each local neighborhood was used for the smoothing procedure. Furthermore, we estimated 95% confidence bands for the smoothed trends.

Validation of text-mining algorithm

On 8 June 2017, we took a random sample of 300 articles (100 from each journal) published in the most recent calendar period (2012-2016) for review by a knowledgeable reviewer (M.A.). We selected 100 abstracts per journal because the expected precision of the proportions would be sufficient (e.g. for a proportion of 30%, the 95% confidence interval by expectation would be 21%-39%). A second reviewer (A.S.) randomly double-checked a sub-sample of the first reviewers’ assessment. They discussed disagreement in their evaluation, which potentially reduced the error rates. The aim was to compare the performance of the algorithm with the manual reading of the abstracts. Furthermore, we compared results of the algorithm that is based on abstracts only with the presence of statistical inference in full papers.

M.A. retrieved the full papers from the PubMed database (dates of retrieval: 8th-16th June 2017) and manually identified statistical inference in abstract and full text (text, tables and figures). The manual review of the full paper included the search for expressions like “signif”, “power”, “equiv”, and “inferior” by using Adobe Acrobat Reader’s find function. For each instance of each of these letter strings, we determined by a close reading of the abstract and the remainder of the paper if the word in which the letter string appeared was used statistically or substantively (e.g., by the presence of the adjective, “clinical” before the noun, “significance”). In cases that the expression was a citation by the authors, we also checked the cited paper. However, even after a close reading, the use of some terms remained ambiguous.

Therefore, we classified the use in abstracts of words containing these letter strings as “clearly statistical”, “clearly non-statistical” and “unclear non-statistical.” In reviewing the abstracts if the terminology was without any qualifier, we read the full text to classify them as either statistical or non-statistical. For example out of 25 abstracts with unclear “signif” terminology, 24 were used in statistical sense and 1 in non-statistical sense. However, in reviewing the full text it was difficult to assign some terminologies to the statistical or non-statistical groups, therefore we have a sub-category of “unclear non-statistical”. For the letter string “equiv” we checked if the author used the “Bioequivalence NHT” prescribed by the European Medicines Agency guideline. According to the Guideline on the Investigation of Bioequivalence, Committee for Medicinal Products for Human Use, European Medicines Agency, London, 2010, “The assessment of bioequivalence is based upon 90% confidence intervals for the ratio of the population geometric means (test/reference) for the parameters under consideration. This method is equivalent to two one-sided tests with the null hypothesis of bioinequivalence at the 5% level”[22].

For CI reporting, we only considered CIs for measures with null values (i.e., measures of comparison, association or effect). We then checked whether the null value (e.g., a hazard ratio of 1) was included within each CI. Moreover, from our random sample of 300 articles, we took a sub-sample of all “CI-only” articles. One knowledgeable reviewer (C.P.) reviewed them to identify papers employing implicit or covert NHT, which consists of emphasizing or treating special in some way those CIs that are accompanied by rejections of the null-hypothesis (e.g., by selectively reporting them and not others in the abstract). All reviewers were unaware of the outputs of the automatic classification of the abstract by the algorithm.

To assess the algorithm performance, we compared its output with the results of the manual review of the random sample of 300 publications. We calculated the median unbiased estimates and 95% confidence interval for sensitivity, specificity, positive and negative predictive values. We presented the median unbiased estimates as they provide more accurate estimation than maximum likelihood estimation, when the sample size is small [23].

For the interested readers, we present two samples of abstracts that published in year 1976 and 2016 (Table S7 and S8). We filtered the data set of the year 1976 and 2016 to identify those abstracts that the algorithm assessed as “any NHT=1”, “p-use=0” and “any CI=0” and took the first 20 abstracts that appeared after filtering. These supplementary tables show how the performance of the algorithm differs from a manual review.

Result

Algorithmic review of abstracts- period 1976-2016

The total number of article abstracts increased form 377 in 1976 to 791 in 2016. The annual number of abstract published between 1976 and 2016 varied across the included journals between 1976 and 2016 (Table 1). The algorithm identified 13,375 out of 22,516 (59%) abstracts that contained statistical inference (Table 1). The time trend showed that the annual percentage of abstracts containing statistical inference increased initially, reached a peak in the early 2000s and then declined again (Figure 1) so that the overall percentage of abstracts with statistical inference was similar in 1976 and 2016 (48% vs. 49%) (Table 1). The prevalence of statistical inference was high in some publication types like meta-analysis (85%), randomized controlled trial (77%), clinical trial (74%), and journal article (59%), whereas it was low in reviews (26%) and case reports (24%) (Table S2).

Statistical inference reporting pattern varied among journals. From 1976 to 2016, the percentage of statistical inferences in abstracts increased from 42% to 59% in *Br J Clin Pharmacol*, and from 47% to 60% in *Eur J Clin Pharm*, whereas in *Clin Pharmacol Ther* it decreased drastically from 49% to 25% (Table 1). In 2007, *Clin Pharmacol Ther* exhibited a marked drop from 85% in 2006 to 33% in statistical inference reporting (Table S1). *Eur J Clin Pharm* also showed a decreasing trend but to a different extent, from 68% in 2006 to 63% in 2007 (Table S1) (Figure 1). According the journal's source data characterization across the time span 1997-2016, in 2003, the proportion of "other publication" in each journal grew noticeably from none to about 80%, 40% and 9% in *Clin Pharmacol Ther*, *Br J Clin Pharmacol* and *Eur J Clin Pharm*, respectively (Table S4).

Overall, among abstracts containing statistical inference, NHST appeared most frequently. In the mid-1970s, nearly all abstracts reporting any statistical inference contained NHST, while CI was absent (Figure 2). In early 1990s, NHST showed a decreasing trend, while CI reporting grew in popularity, and its prevalence rose in all three journals (Figure 2). Reporting CI as the only means of statistical inference became more popular since early 2000 (Figure 2). From mid-1970s to early 1990s, the prevalence of p-values increased in *Br J Clin Pharmacol* and in *Clin Pharmacol Ther* and then dropped noticeably over the last decades (Figure 2A, 2B). The *Eur J Clin Pharm* showed a fluctuating p-value reporting pattern with a low in early 1990s and a peak in 2006, thereafter its prevalence decreased again (Figure 2C).

Algorithmic review of the most recent abstracts

Table 2 provides an overview of the prevalence of reporting statistical inferences in abstracts of publication years 2012-2016. Overall, about half of the abstracts contained statistical inference (1,603 out of 3,215 abstracts) and among abstracts with statistical inference, NHST was the most frequent subtype. Prevalence of NHST was higher than CI in abstracts, especially in *Clin Pharmacol Ther* (92% vs. 18%) followed by *Eur J Clin Pharm* (86% vs. 38%) and *Br J Clin Pharmacol* (80% vs. 50%). In *Br J Clin Pharmacol*, the prevalence of CIs was higher than p-

values (50% vs. 30%), whereas in *Clin Pharmacol Ther* and *Eur J Clin Pharm* prevalence of p-values was higher than CIs, (34% vs. 18%) and (41% vs. 38%), respectively. The Prevalence of reporting only CI was 20% in *Br J Clin Pharmacol*, 14% in *Eur J Clin Pharm* and 8% in *Clin Pharmacol Ther*.

The prevalence of statistical inference was 91% in meta-analysis, 73% in randomized controlled trial, 68% in clinical trial , 50% in journal article, 27% in reviews and 25% in case reports (Table S3).

Manual review of full paper- period 2012-2016

Overall 50% of all abstracts contained statistical inference (150 out of 300 abstracts), of which 80% was any NHST and 45% was any CI (Table 4). Prevalence of NHT in abstracts was 76%, whereas the prevalence of ST was 21% (Table 4). As it is presented in table 5, the evaluation of kind of usages of terminologies (substantive use versus statistical use), among abstracts that used significance terminology (overall 31%), 95% were used in a statistical sense and 12% in a substantive sense (e.g. clinical significance). Among the 8 abstracts using the term "power", 6 times it referred to statistical concepts and 3 times to non-statistical (e.g. "powerful"). Abstracts that contained the term "equiv" (n=11) or "inferior" (n=2) all used these terminologies in a statistical sense. The percentages of the statistical and non-statistical subcategories do not sum up to 100 as the term appeared more than once in some abstracts, once in statistical and once in non-statistical sense.

Table 3 shows the percentage of the statistical inference anywhere in full paper (abstract, text, table, or figure). Overall, 83% (265 out of 300) of all full texts contained statistical inference. The majority (98%) was NHST reporting. ST was always accompanied by other inferential statistics in the full paper. Only 2% of the papers reported only CI in the entire paper. After a detailed analysis of CI only papers, we did not detect covert use of NHT. *Clin Pharmacol Ther* had a lower percentage of statistical inference in the full paper (80%) compare to the other two. All papers in *Clin Pharmacol Ther* reported NHST, while none reported only CI in the whole paper. The percentage of the CI was lower in *Clin Pharmacol Ther* (38%) compared to *Br J Clin Pharmacol* (69%) and Eur J Clin Pharm (65%).

Table 4 summarizes the comparison of the abstract text with the remaining text of the full paper including tables and figures. It revealed that the remaining text had a higher percentage of statistical inferences than abstracts (88% vs. 50%). The prevalence of statistical inference was 88% in text (abstract excluded), 46% in tables and 30% in figures. Presentation of result with only one statistical inference subtype was more common in abstracts. For example, the prevalence of 'CI only', 'NHT only' and 'ST only' was 20%, 41% and 2% in abstracts versus 2%, 28% and 0% in the full text, respectively.

As shown in table 5, overall, 90% of all full texts including tables and figures (excluding the abstract) used "significance" related terminologies. The vast majority (90%) used this terminology in a statistical sense. In 51%, it was used in a non-statistical sense and in 4%, it

remained unclear. Furthermore, 38% of all full texts used the terminology “power”, of which 70% was in a statistical sense, 35% in a non-statistical sense and 2% remained unclear. Out of 54 full texts containing the term “equiv”, 32 was used them in statistical sense and 24 in non-statistical sense. Of the 14 paper with “inferior” terminology, 13 were statistical and 1 non-statistical. As each term was multiple times present in the text, the total percentages of the statistical and non-statistical terms do not sum up to 100.

Assessment of the performance of the algorithm compared with the manual review of the abstracts

Table 6 shows these assessments results. The algorithm had a high sensitivity and specificity in detecting statistical inferences in abstracts. Comparison of the automatic identification of statistical inferences against the manual review showed that the algorithm failed to detect three abstracts with CI, two with any NHT, and five with ST (false negative). Its positive predictive value (PPV) ranged between 92% (for ST) to 98% (for CI and any NHT). Among 300 abstracts, the algorithm detected six abstract with NHT, two with ST and one with CI whereas manual review did not find it (false positive). Its negative predictive value (NPV) ranged from 99% (for any NHT and CI) to 98% (for ST).

Prediction of the occurrence of subtypes of statistical inference in the complete manuscript based on the abstract-based algorithm

Of the 300-reviewed paper, there were 65, 116 and 27 with CI, NHT and ST respectively in both in full text (including text, table and figure) and in algorithmic search of abstracts (Table 6). Whereas the PPV of the occurrence of any NHT, ST and CI in the full text (including the abstract, text, tables and figures) is high, the NPV is low for all three measures, especially for the occurrence of NHT. As seen from the manual review of the entire paper, based on solely reading the abstracts without reading the entire paper it is difficult to estimate the occurrence of the statistical inference in the full paper. The algorithm presented a low sensitivity in predicting reporting style of statistical inference in the remainder of the paper, ranging from 23% to 45% for the occurrence of ST, NHT, and CI respectively (Table 6). The manual review of full papers showed that conclusions about the presence of statistical inference based on abstracts only may differ from those based on the remainder of the paper.

Discussion

The practice of statistical inference reporting in abstracts was heterogeneous among journals that we reviewed. Overall, half of the article abstracts contained statistical inferences. After an increase in the percentage of reporting statistical inference from 1976 to the early 2000s, the percentage fell again by 2016. Some publication types like editorials usually do not contain statistical inference, reviews including meta-analyses may be an exception from this. Review articles either discuss a topic in a qualitative way or may contain statistical inference, but typically, these reviews quote statistical inferences of publications from other authors and other journals. One might look at review articles from a different perspective in this regard. For instance, authors of review articles who report in their abstracts statistical inference on the part of original researchers are indicating that they, the reviewers, deem those inferences worthy of being highlighted in that way.

Statistical analysis may not be required for every type of original article. Some clinical pharmacology studies provide descriptive information on the mechanism of action of the drug, or pharmacokinetic and pharmacodynamics data. Analyzing whether work published by others is descriptive or exploratory post hoc is difficult. However, it is good statistical practice to report measures of statistical uncertainty of the estimated value presented in exploratory studies.

The time trend was accompanied by a remarkable increase in CI reporting. Since the early 1990s, the percentage of CIs in abstracts with some form of statistical inference has been continuously increasing and rose from 0% in 1976 to 53% of all abstracts by 2016. Systematic reviews of other types of journals from 1975 through 2014 showed similar time trends in the use of CIs [18]. In contrast, uptake of CIs was slower by psychology journals over almost the same period [19]. This shows the variation in acceptance of CIs by different medical journals and scientific communities.

However, NHST reporting has remained consistently ubiquitous. Thus, CIs are increasingly reported to supplement, rather than to supplant, NHST. Moreover, the manual review of 300 abstracts showed that, despite the regulatory guidance in favor of reporting precise p-values (ST) over p-threshold (NHT) [24], reporting p-thresholds was more common than precise p-values (76% any NHT versus 21% any ST). However, ASA discourages decisions “based only on whether a p-value passes a specific threshold” [14]. Diong highlighted “researchers will often opt for reporting practices that make their paper look like others in their field”[25]. The reporting of CIs might or might not reflect an emphasis on estimation over testing [14], as it easy to conduct NHT by inspecting CIs for the presence or absence of null values.

The proliferation of CI reporting in abstracts may reflect the influential update of ICMJE recommendation in 1988 in favor of additional reporting of CI and the subsequent editorial policy changes in a number of medical journals[9]. ICMJE encourages journals following them to incorporate ICMJE recommendations into their instruction to authors and to endorse in those instructions that they follow ICMJE. *Eur J Clin Pharm* and *Clin Pharmacol Ther* both

follow the ICMJE; however only *Eur J Clin Pharm* stated in its current author's instruction that the manuscripts must comply with the ICMJE guidelines.

While all three journals guide authors in respect of content of manuscript and submission process, only *Eur J Clin Pharm* included some points regarding the statistical reporting. However, all three journals advocate following the CONSORT guidelines for reporting the results of randomized clinical trials. *Br J Clin Pharmacol* mentioned also other reporting guidelines such as STROBE, PRISMA, SPIRIT, CARE, STRAD, CHEERS, etc. Moreover, A. Ring published an editorial in *Br J Clin Pharmacol* in 2017 [26] and in 2010 B. Smith in *Clin Pharmacol Ther* [27]*, which* provide guidance on statistical reporting and interpretation of study findings. There is a debate on how influential are the editorials in improving the statistical reporting.

While Diong et al showed that statistical reporting did not appreciably change in response to editorial advice offered in 2011 in journals of physiology and British journal of pharmacology [25], studies in public health and psychology journals showed greater response to editorial policies[28, 29]. Recommendations, even those in instructions for authors, are merely guidelines and not rules. Therefore, they guarantee neither that all authors will adhere to them, nor that all authors will agree with them. Therefore, future studies are needed to assess the trend in reporting statistical inferences before and after changes in the advisory committee policies in clinical pharmacology journal.

The comparison of the statistical reporting style in abstracts compared to the remaining text of the publications showed that the remaining text contained more frequently a mixture of reporting styles (NHT, ST, CI) than the abstract alone, which is in line with a previous study [4]. The ambiguous usage of „significance", „significantly, "and „significant "without qualifying it as substantive or statistical was present in abstract and full text. The same results were reached in former reviews [18, 19]. Moreover, the use of the term "significance" might have changed over time; however, inferring the intended meaning from the publication, itself can be complicated and it might require contacting authors, but even that is difficult, if not impossible. Future studies are needed. The rejection or failure to reject a null hypothesis is the first thing, and sometimes the only thing, addressed in many statistics textbooks and courses under the rubric of „statistical inference". Such inferences are often drawn without explicit reference to the value of the test statistic, the chosen alpha level, or the p-value. The implicit reference is widely understood, unless otherwise stated, to a two-sided test with alpha=0.05. The numbers are there even though widespread convention makes it possible, and customary, not to refer to them explicitly.

Taking a broader perspective and in line with a previous guideline paper [10], we would like to emphasize that presentation of statistical inference should not be separated from an indication of effect size: an observation may be statistically significant but have an effect size of questionable biological relevance; on the other hand, the observed effect size may be

sufficiently large to be of likely biological relevance but have a large p-value, which makes the finding inconclusive.

From 1976 through 2016, the percentage of statistical inference in abstracts decreased by 24% in *Clin Pharmacol Ther*, whereas this percentage increased by 17% in *Br J Clin Pharmacol* and by 13% in *Eur J Clin Pharm* respectively. As the reporting style of statistical inference in abstracts of *Clin Pharmacol Ther* varied markedly, we reviewed its editorials. In 2002, a new section called “perspectives” was introduced to keep the journals as an educational tool in the clinical pharmacology field [30] which had consequences: whereas the proportion of “other articles” (editorials and letters) was virtually zero before 2003, this proportion became 80% in 2003. These publication types barely contain statistical inferences. The change in editorial policy in 2003 may explain the delayed effect from 2006 through 2007, where the percentage of statistical inference in abstracts declined from 85% to 33% (Table S1). While more opinion and invited content papers may have been published during some periods and that may have led to reduced percentage of papers reporting statistical inference, that should not have affected the relative roles of reporting types of inference such as NHT, ST, NHST or CI.

Our review has some limitations. First, we described the time trends in abstracts of three journals, which do not necessarily reflect time trends in the remaining texts of publications and the selected journals might not be representative of all clinical pharmacology journals.

Second, this time trend provides a picture of the net or combined effect of several factors including journal policies, author preferences, developments in the scientific community and publication type and we cannot disentangle which factors played major roles in the observed time trend for the three journals that we studied.

Third, our analysis was descriptive and we estimated the three time trends separately and non-statistically. Our key finding was that statistical inference reporting differed between the three journals and varied over time. The latter could be analyzed using more sophisticated approaches like estimating the differences among the time trends, which would be a simple matter only if the trends were assumed to be linear on some scale. Another more sophisticated analysis might consist of a random-effects meta-analysis in which the one would estimate the „average time trend“ and the spread (variance) of time trends in a population of time trends in journals. However, we fail to see how this would change our key findings.

Lastly, the automated algorithm is unable to evaluate substantive versus statistical use of some terminologies, which are indicators of different types of statistical inference reporting resulting in overestimation of their prevalence. However, our manual review shows that this overestimation is small. Moreover, despite the fact that knowledgeable authors manually reviewed the abstracts, there remained some subjectivity in the rating of abstracts regarding reporting of statistical inference.

We believe that further studies by reviewing the full text of the most recent publications from larger number of clinical pharmacology journals are needed to make more certain interpretations.

Conclusion

The time trend showed that CI reporting have been increasing whereas NHST reporting have been decreasing in reviewed journals. Despite journals' editorials and statistical associations' guidelines, most abstracts in journals reviewed focused on NHT rather than on ST or estimation with CIs, even in recent years.

Our study estimated only the time trend on statistical inference reporting; however, there are many other statistical concepts and methods for which different approaches are taken and on which editors might wish have their preferences reflected in submitted articles by expressing those preferences in instructions for authors. Both authors and editors play an important role in the observed trend. The format required for submission of manuscripts differs from journal to journal. The authors need to prepare their manuscripts in the format specified by the journal they have chosen. In future, a better statistical reporting might be ensured by improving the statistical knowledge of authors [25, 31] and an addition of statistical guides to journals' instruction to authors. Some preclinical pharmacology journals have recently done so [10].

Acknowledgements

We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Conflict of interest

The authors declared no competing interests for this work.

Funding

Work in the lab of MCM related to data analysis is supported in part by the Innovative Medicines Initiative 2 Joint Undertaking (grant agreement no. 777364); this Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA).

Contributors

A.S., C.P. and M.D. designed the study; M.A., A.S. and C.P. involved in the review. M.A and A.S. performed the statistical analysis; C.P., and M.C.M. provided feedback on data analysis. M.A. drafted the manuscript. All authors participated in the critical revision of the manuscript and approved the final manuscript for submission.

Reference List

1. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology 2016; 31: 337-50.
2. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. European journal of epidemiology 2010; 25: 225-30.
3. Gigerenzer G SZ, Porter T, Daston L, Beatty J, Krüger L. . The empire of chance. How probability changed science and everyday life. . Cambridge: Cambridge University press, 1989.
4. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. Jama 2016; 315: 1141-8.
5. Boring EG. Mathematical vs. scientific significance. Psychological Bulletin 1919; 16: 335-38.
6. Stigler S. Fisher and the 5% level. CHANCE 2008; 21: 12-12.
7. Uniform requirements for manuscripts submitted to biomedical journals. International Committee of Medical Journal Editors. British medical journal (Clinical research ed) 1988; 296: 401-5.
8. Altman DG, Machin, David, Bryant, Trevor N. and Gardner, Martin J. Estimation with confidence. In: Statistics with confidence, 2nd Edition, edBMJ, London, 2000: 3-5.
9. Lang JM, Rothman KJ, Cann CI. That confounded P-value. Epidemiology (Cambridge, Mass) 1998; 9: 7-8.
10. Michel MC, Murphy TJ, Motulsky HJ. New Author Guidelines for Displaying Data and Reporting Data Analysis and Statistical Methods in Experimental Biology. Drug metabolism and disposition: the biological fate of chemicals 2020; 48: 64-74.
11. Anderson DR, K. P. Burnham, and W. L. Thompson. . Null hypothesis testing in ecological studies: problems, prevalence, and an alternative. . Journal of Wildlife Management 2000; 64: 912-23.
12. Thomson W. 402 Citations questioning the indiscriminate use of null hypothesis significance tests in observational studies. In, 2001: <https://www.gwern.net/docs/statistics/2001-thompson.html>.Accessed September 2015.

13. Rothman KJ. Significance questing. Annals of internal medicine 1986; 105: 445-7.
14. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 2016; 70: 129-33.
15. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond “p < 0.05”. The American Statistician 2019; 73: 1-19.
16. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019; 567: 305-07.
17. Ioannidis JPA. The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance. Jama 2019; 321: 2067-68.
18. Stang A, Deckert M, Poole C, Rothman KJ. Statistical inference in abstracts of major medical and epidemiology journals 1975-2014: a systematic review. European journal of epidemiology 2017; 32: 21-29.
19. Baethge C, Deckert M, Stang A. Tracing scientific reasoning in psychiatry: Reporting of statistical inference in abstracts of top journals 1975–2015. 2018; 27: e1735.
20. Cleveland WS, Devlin S, Grosse E. Regression by local fitting. J Econometrics 1988; 37: 87-114.
21. Cleveland WS, Grosse E. Computational methods for local regression. Stat Comput 1991; 1: 47-62.
22. European Medicines Agency . Committee for Medicinal Products for Human Use (CHMP), Guideline on the Investigation of Bioequivalence. CPMP/EWP/QWP/1401/98 Rev. 1/Corr **. In, London, 2010: <https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf>. Accessed July 2019.

23. Rothman KJ, Greenland S, Lash T. Median unbiased estimates. In: Modern Epidemiology, 3rd Edition, edRothman KJ, Greenland S, Lash T, Philadelphia: Lippincott Williams & Wilkins, 2008: 221, 55-56.
24. European Medicines Agency (EMA), ICH Topic E9. Statistical principles for clinical trials, (CPMP/ICH/363/96). In, London, 1998: <https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf>. Accessed July 2019.

25. Diong J, Butler AA, Gandevia SC, Heroux ME. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. PloS one 2018; 13: e0202121.
26. Ring A, Schall R, Loke YK, Day S. Statistical reporting of clinical pharmacology research. British journal of clinical pharmacology 2017; 83: 1159-62.
27. Pharmacology C, Team TE. Statistical Guide for Clinical Pharmacology & Therapeutics. 2010; 88: 150-52.
28. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. Psychological science 2004; 15: 119-26.
29. Sharpe D. Why the resistance to statistical innovations? Bridging the communication gap. Psychological methods 2013; 18: 572-82.
30. Reidenberg MM. Thoughts from the Editor. 2002; 71: 1-2.
31. Smith BP. Where Will Statistical Sciences for Clinical Pharmacology Be in 2030? Clinical pharmacology and therapeutics 2020; 107: 17-21.

Characterization of journals included in the review

Table1 Prevalence of number of abstracts in each journal included in the review of publication year 1976-2016

Journal	Total no. of abstracts in calendar year (n)			No. of any statistical inference in calendar year (%)			Abstracts per calendar year	
	1976-2016	1976	2016	1976-2016	1976	2016	n min	n max
Br J Clin Pharmacol	8380	31	345	5312 (63%)	13 (42%)	204 (59%)	31	345
Clin Pharmacol Ther	6732	187	244	3577 (53%)	91 (49%)	61 (25%)	101	244
Eur J Clin Pharmacol	7404	159	202	4486 (61%)	75 (47%)	121 (60%)	89	327
Overall	22516	377	791	13375 (59%)	179 (48%)	386 (49%)	74	305

Results of the algorithmic approach of papers published in most recent years 2012–2016

Table 2 Prevalence of reporting of statistical inference in abstracts of the publication years 2012–2016

	Br J Clin Pharmacol		*Clin Pharmacol Ther*		*Eur J Clin Pharmacol*	
	n	***%**	**n**	***%**	**n**	***%**
Total number of abstracts	1358		912		945	
Abstracts containing any statistical inference	757	56	262	29	584	58
Abstracts containing NHST	605	80	240	92	500	86
Abstracts containing any p-values	226	30	90	34	241	41
Abstracts containing CI	376	50	47	18	220	38
Abstracts containing only NHST	381	50	215	82	364	62
Abstracts containing only CI	152	20	22	8	84	14

Legend: NHST = null- hypothesis significance testing; CI = confidence interval.

*Note: The percentage of different types of statistical inference is calculated among the abstracts containing any statistical inference.

Results of manual review of 300 random samples of articles in the period 2012-2016

Table 3 Percentage of statistical inference anywhere in full paper (abstract, text, table, or figure) in validation sample (n=300)

Journal	Total	Any statistical inference		Percentages of statistical inference among papers containing any statistical inferences (%)							
	n	n	%	Any CI	Any ST	Any NHT	CI only	Any NHST	ST only	NHT only	ST & NHT
Br J Clin Pharmacol	100	93	93	69	43	98	2	98	0	24	43
Clin Pharmacol Ther	100	80	80	38	34	100	0	100	0	55	34
Eur J Clin Pharmacol	100	92	92	65	58	96	4	96	0	10	58
All journals	300	265	83	58	45	98	2	98	0	28	45

Legend: Any CI = any confidence interval; CI only = confidence interval only; any NHST = any null- hypothesis testing OR significance testing; ST only = significance testing only; NHT only = null- hypothesis testing only; ST& NHT= null- hypothesis testing AND significance testing*; Br J Clin Pharmacol = British Journal of Clinical Pharmacology; Clin Pharmacol Ther = Clinical Pharmacology &Therapeutic; Eur J Clin Pharmacol = European Journal of Clinical Pharmacology.*

Table 4 Percentage of statistical inference in abstract versus full- text (text, table, figures) in validation sample (n=300)

Abstracts only											
Journal	Total	Any statistical inference		Percentages of statistical inference among papers containing any statistical inferences (%)							
	n	n	%	Any CI	Any ST	Any NHT	CI only	Any NHST	ST only	NHT only	ST & NHT
Br J Clin Pharmacol	100	59	59	56	22	69	25	75	3	32	17
Clin Pharmacol Ther	100	23	23	30	13	83	17	83	0	61	13
Eur J Clin Pharmacol	100	68	68	41	24	79	16	84	1	41	19
All journals	300	150	50	45	21	76	20	80	2	41	17
Full texts only											
Br J Clin Pharmacol	100	93	93	69	43	98	2	98	0	24	43
Clin Pharmacol Ther	100	80	80	38	34	100	0	100	0	55	34
Eur J Clin Pharmacol	100	92	92	64	58	96	4	96	0	10	58
All journals	300	265	88	58	45	98	2	98	0	28	45

Legend: Any CI = any confidence interval; CI only = confidence interval only; any NHST = any null- hypothesis testing OR significance testing; ST only = significance testing only; NHT only = null- hypothesis testing only; ST& NHT= null- hypothesis testing AND significance testing; *Br J Clin Pharmacol = British Journal of Clinical Pharmacology; Clin Pharmacol Ther = Clinical Pharmacology &Therapeutic; Eur J Clin Pharmacol = European Journal of Clinical Pharmacology.*

Table 5 Percentage of the terms in abstracts and full text (text, table, Fig) of the 300 papers

Term	Abstract only				Text only (text, table, figure)			
	Term present in abstract n (%)*	unclear non-statistical n (%)*	Clear statistical n (%)*	Clear non-statistical n (%)*	Term present in text n (%)	unclear non-statistical n (%)*	Clear statistical n (%)*	Clear non-statistical n (%)*
Signif	94(31)	0	89(95)	11 (12)	270 (90)	12 (4)	244 (90)	138 (51)
Power	8(3)	0	6 (75)	3 (38)	115 (38)	2 (2)	81(70)	40(35)
Equiv	11 (4)	0	11 (100)	0	54 (18)	0	32 (59)	24(44)
Inferior	2 (1)	0	2(100)	0	14 (5)	0	13(93)	1 (7)

Note: The dominators for the relative frequencies in each column is 300.The percentages of the statistical and non-statistical subcategories do not sum up to 100 as the term appeared multiple times, once in statistical and once in non-statistical sense.

Table 6 Performance of the algorithm at identifying statistical inferences (N=300 random samples of articles published in the period 2012-2016)

	Statistical inference	Algorithm (n)				Median unbiased estimates of Sn, Sp, NPV &PPV and their mid-P 95% CI			
		Absent		Present					
		Manual (n)		Manual (n)					
		Absent	Present	Absent	Present	**PPV (%) [95% CI]**	**NPV (%) [95% CI]**	**Sn (%) [95% CI]**	**SP (%)[95% CI]**
Analysis 1[a]	**CI**	231	3	1	65	98 [93-100]	99 [97-100]	95 [88-99]	100 [98-100]
	Any NHT	180	2	6	112	98 [90-98]	99 [96-100]	98 [94-100]	97 [93-99]
	ST	266	5	2	27	92 [79-99]	98 [96-99]	84 [69-94]	99 [98-100]
Analysis 2[b]	**CI**	145	89	1	65	98 [93-100]	62 [56-68]	42 [35-50]	99 [97-100]
	Any NHT	39	143	2	116	98 [95-100]	21 [16-28]	45 [39-51]	95 [85-99]
	ST	178	93	2	27	93 [79-99]	66 [60-71]	23 [16-31]	99 [85-99]

[a] Analysis 1: Assessment at identifying statistical inference and its subtypes in abstracts (N=300)

[b] Analysis 2: Assessment at identifying statistical inference and its subtypes in papers, including but not limited to, their abstracts (N=300)

Legend: Sn = sensitivity; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; CI = confidence interval; NHT = null-hypothesis testing; ST = significance testing

Figure legend

Figure 1. Annual percentage of abstracts containing any statistical inference for the publication years 1976 through 2016

Figure 2. Annual percentage of different types of statistical inference among abstracts containing any statistical inference for the publication years 1976 through 2016

Br J Clin Pharmacol = British Journal of Clinical Pharmacology; Clin Pharmacol Ther = Clinical Pharmacology &Therapeutic; Eur J Clin Pharmacol = European Journal of Clinical Pharmacology.

Appendices

Supplementary Table S1. Overview of number of abstracts published in the included journals and the prevalence of abstracts that contain any statistical inference included in the review of publication years 1976-2016

Supplementary Table S2. Prevalence of number of abstracts categorized according to the publication type in each journal included in the review of

publication years 1976-2016

Supplementary Table S3. Prevalence of reporting of statistical inference in abstracts categorized into different publication types of publication years 2012–2016

Supplementary Table S4. Journal's source data information in years 1997-2016

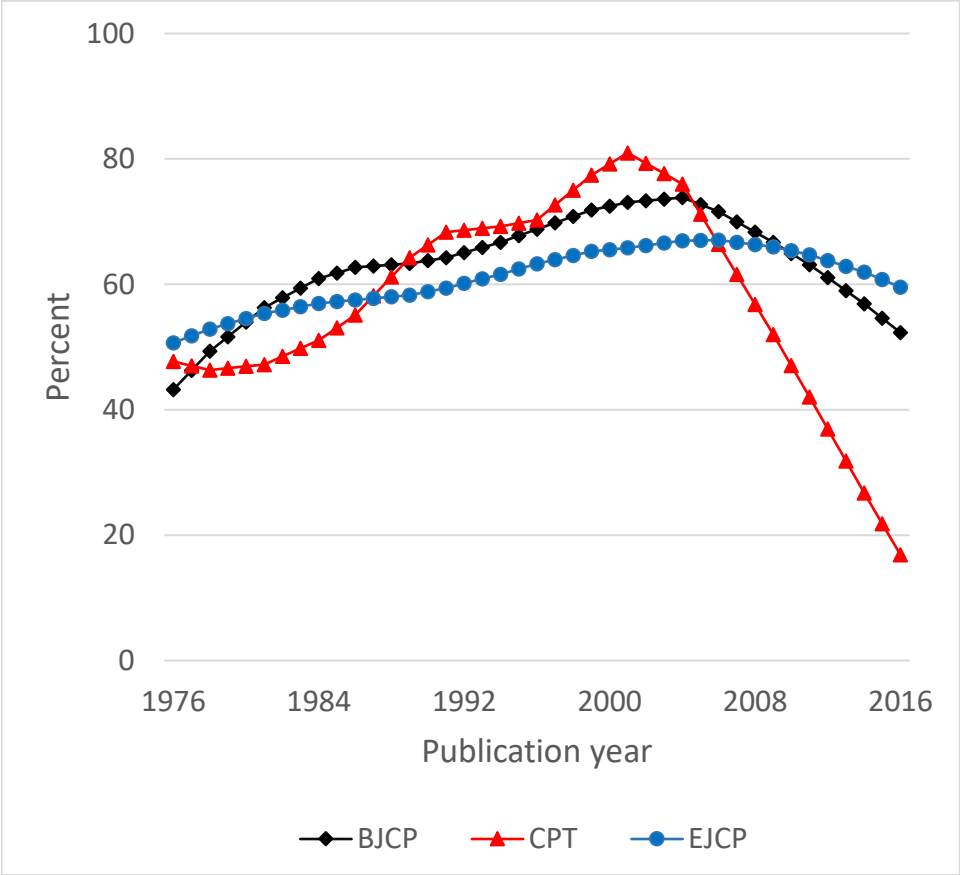
Supplementary Table S5. Raw data related to year 1976

Supplementary Table S6. Raw data related to year 2016

Supplementary table S7. Comparison of performance of the algorithm with manual review at identifying any NHT in a sample of abstracts published in 1976

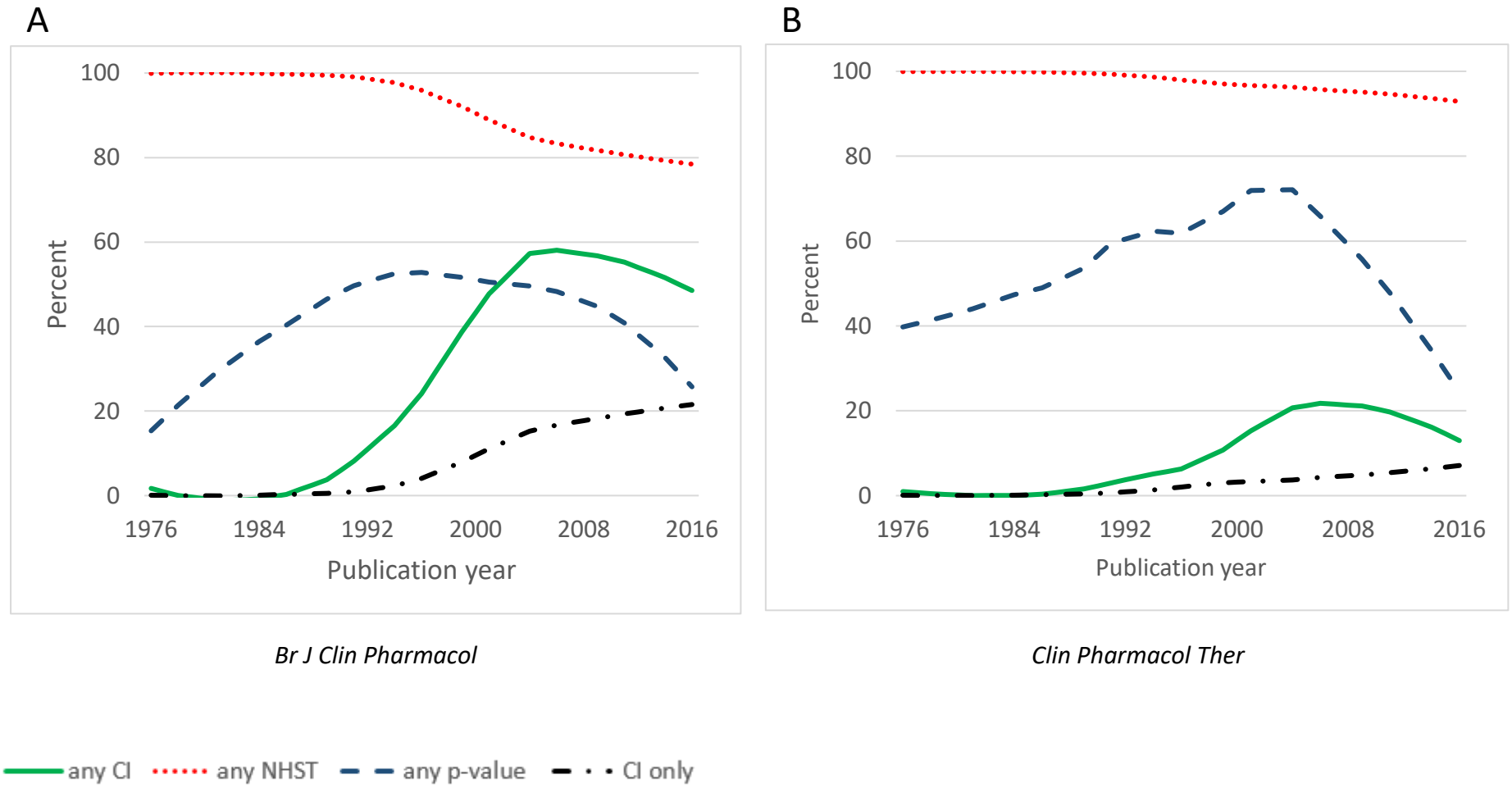
Supplementary table S8. Comparison of performance of the algorithm with manual review at identifying any NHT in a sample of abstracts published in 2016

Figure 1 **Annual percentage of abstracts containing any statistical inference for the publication years 1976 through 2016**



Legend: *BJCP= British Journal of Clinical Pharmacology; CPT= Clinical Pharmacology &Therapeutic; EJCP= European Journal of Clinical Pharmacology*;

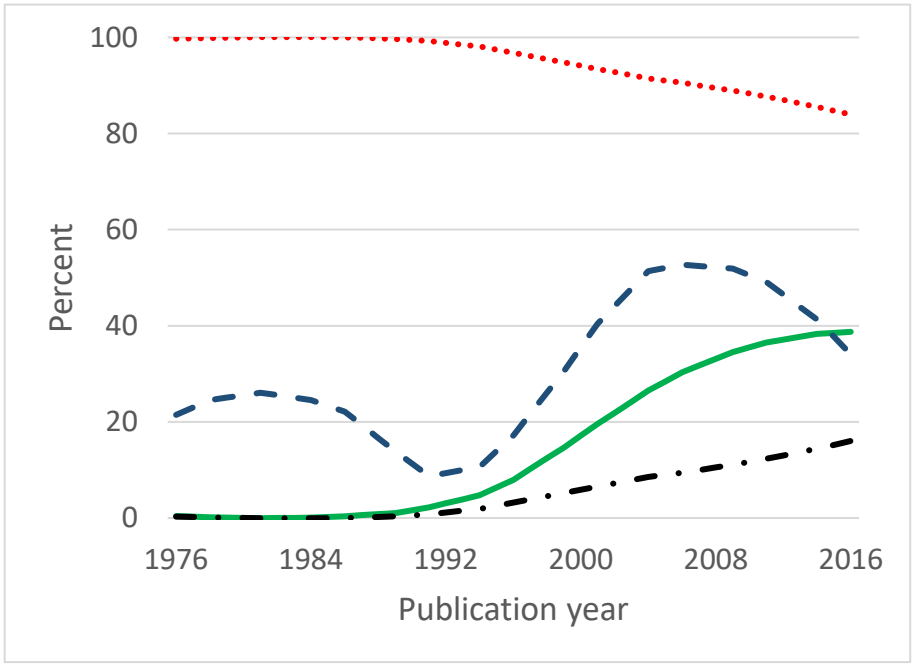
Figure 2 **Annual percentage of different types of statistical inference among abstracts containing any statistical inference for the publication years 1976 through 2016**



Br J Clin Pharmacol

Clin Pharmacol Ther

C



Eur J Clin Pharmacol

Legend: any CI—any confidence intervals; any NHST— null hypotheses significance testing includes any use of ST or NHT regardless of CI reporting; any p value —reporting of either ‘‘p equals’’ or p value thresholds regardless of CI reporting; CI only— reporting of confidence intervals without continuous or categorical p values and without significance terminology

4. Discussion

Our analysis of the two datasets of NIS on propiverine found age as the only factor consistently but weakly associated with the initial dosing decision. Physicians prescribed higher doses to younger patients. This may reflect the dose selection based on the lower probability of side effects of antimuscarinics in younger subjects (Michel et al., 2008). On the other hand, factors such as duration of OAB syndrome, number of incontinences, voiding and nocturia episodes were associated with starting dose in one NIS, but not in the other NIS. Our results also suggest that, the number of baseline urgency episodes and change in incontinence episodes after 4 weeks were the only factors that were consistently associated with the decision to increase the dose from 30 to 45 mg in both studies. In contrast, height, basal number of nocturia episodes and change thereof were associated with dose escalation in one NIS, but not in the other NIS. According to previous RCTs (Wyndaele et al., 2011, Goldman et al., 2019, Cardozo et al., 2013) severity of the OAB baseline parameters, although not necessarily nocturia episodes, were the predictors for dose escalation, but another RCTs conducted by (Wagg et al., 2015) did not confirmed it. Similarly, smaller improvements of OAB symptoms upon initial treatment were associated with the decision to increase the dose in some studies (Cardozo et al., 2012, Wyndaele et al., 2011, Cardozo et al., 2013) but not in others (Wagg et al., 2015, Goldman et al., 2019). Moreover, when studying which factors were associated with the treatment outcome, we found that 45/45 cohort had slightly greater, and the 30/45 cohort slightly smaller efficacy as compared to the 30/30 group. In line with another study comparing dose escalator with non-escalator showed slightly smaller improvement, our analysis showed that compared to 45/45 group, starting, and staying at 30 mg or escalating to 45 mg tended to have only minor effect on the overall treatment efficacy. However, some studies that did not apply complex models to overall treatment outcome found no difference in treatment outcomes of dose escalator and non-escalators (Wyndaele et al., 2011, Cardozo et al., 2012, Cardozo et al., 2013). Overall, our work showed the potential value of NIS to explore the factors associated with the initial dose selection by physicians under real life condition. The evidence generated from these two NIS support the concept that dose escalation helps to achieve clinically meaningful symptom improvements in patients where tolerability of the lower dose allows for dose increase. However, the study findings in this area are heterogeneous. To make a sound decision about the initial dose and further

dose escalation, it is important to check the consistency of the findings across multiple databases (RCTs and other observational studies).

In the context of the NIS with the retrospective data collection, by secondary analysis of the documented inpatient diagnosis in patient charts, we could classify only 2.9 % of all adverse drug events (ADE) coded as probable medication error and therefore as hospital acquired preventable adverse events (pADEs). Only for few cases, the cause for the medication errors could be identified, e.g., heavy workload, documentation errors and poor communication. In contrast, former studies reported higher prevalence of inpatient pADEs (Dequito et al., 2011, Davies et al., 2009). In total, three cases of suspected previously unknown non-preventable ADE (npADE) were identified. However, the low number and lack of information on the actual frequency of previously unknown ADE in hospitals hampers a final qualitative assessment of the potential of the routine data in this context. Therefore, evaluation of a larger sample is needed to assess the potential of routine data to detect the previously unknown npADE. On the other hand, our results showed that the majority of the ADEs coded in routine data are already mentioned in the summary of the product characteristics of the related drugs (SmPCs) which supports their potential use to provide a meaningful complement to existing drug surveillance system. Our study revealed that although the routine inpatient data could be used as markers for npADE and pADE (medication errors) but this a complex and challenging task. This is mainly because these data are not originally collected for the purpose of drug safety surveillance. Having a comprehensive and standardized acquisition, routine data can be effectively used as a complementary data source to detect medication error and might be one of the strategies that could be used in future to address the underreporting of the ADE.

Given the limited quality of the large data generated through NIS, their statistical analysis and interpretation is challenging. While findings from NIS and RCT matched in many cases (Anglemyer et al., 2014, Benson and Hartz, 2000), they have been often criticized as not replicating the results of RCTs (Hemkens et al., 2016, Collins et al., 2020, Mehra et al., 2020). For example, one RCT studied the association of the postmenopausal hormone therapy (HRT) and heart disease by comparing the initiators (incident users) with non-initiator of HRT and concluded that, the initiators have higher risk as compare to non-initiator (Manson et al., 2003). On the other hand, one NIS designed to answer the

same question by comparing the prevalent (current) users with the never users of HRT, and found that the prevalent users have lower risks for heart disease versus never users (Grodstein et al., 2006). Most often, when the results of the NIS studies do not match with the RCT, the blame will be put on the NIS (Labrecque and Swanson, 2017); however, the underlying cause might be poor statistical knowledge or analysis approaches and not considering the sources of biases. For example, the non-replication reason in the above-mentioned studies (Manson et al., 2003, Grodstein et al., 2006) was not addressing the chronology bias due to incident users and prevalent users. Moreover, our work with propiverine shows that two studies of remarkably similar design can lead to distinct conclusions, apparently reflecting a limited robustness of findings and not a difference based on study design or methodology.

We examined two common statistical errors in published medical literature. Our study on the normality of the OAB parameters showed the heterogeneity in the choice of measure of center and dispersion of data in the descriptive reporting of OAB parameters. Most investigators reported means of OAB symptoms and treatment-induced changes thereof without justifying the rationale behind their decision, although these parameters are not normally distributed. However, it appeared that the involvement of a statistician in the data analysis and/or peer review of the manuscript made it more likely to report medians. The difference between mean and medians for OAB symptoms and symptom differences is likely of clinical relevance and may potentially result in misinterpretation of study results on OAB symptoms. Furthermore, our systemic review on the prevalence of the statistical inferences in the abstracts of the publications in selected clinical pharmacology publications showed that the practice of statistical inference reporting in abstracts was heterogeneous among journals that we reviewed. The time trend showed that among the publications with statistical inference reporting significance testing was more prevalent than estimating. Although the time trend was accompanied by a remarkable increase in confidence interval (CI) reporting, but reporting of CIs might or might not reflect an emphasis on estimation over testing (Wasserstein and Lazar, 2016), as it easy to conduct null hypothesis testing (NHT) by inspecting CIs for the presence or absence of null values. Thus, CIs are increasingly reported to supplement, rather than to supplant, null-hypothesis significance testing (NHST). In line with a previous guideline paper (Michel et al., 2020), we would like to emphasize that presentation of statistical inference

should not be separated from an indication of effect size. An observation may be statistically significant but have an effect size of questionable biological/clinical relevance; on the other hand, the observed effect size may be sufficiently large to be of likely biological relevance but have a large p-value, which makes the finding inconclusive. As it is more likely to obtain statistically significant results in a NIS with large study population(Ioannidis, 2005), caution need to be taken in reporting the casual inferences based on observational NIS study (Hernán, 2018, Collins et al., 2020). To make a causal inference from the results of a NIS, one need to emulate a target trial. The concept of the target trial emulation is defined as "the application of design principles from randomized trials to the analysis of observational data, with the aim of improving the quality of the observational epidemiology" (Labrecque and Swanson, 2017). Target trial emulation can help researchers to identify and avoid unnecessary biases (Labrecque and Swanson, 2017). However, emulating the core elements of gold standard RCTs is not always possible, e.g., we cannot emulate blinding (Labrecque and Swanson, 2017) and sometimes finding a target trial is difficult. There are many challenges for collecting and analyzing data using a NIS study design. "The pure "Gold" standard is unattainable. However there are several approaches to improve the post-study probability"(Ioannidis, 2005).Therefore, instead of focusing on the replication ability of the NIS, we should identify the sources of bias due to their study design. In future, a better statistical reporting might be ensured by improving the statistical knowledge of authors and applying a proper epidemiological method to adjust for the cofounders.

In conclusion, RCT and NIS are both valuable and each has its own advantages, opportunities, and limitations. "All scientific work is incomplete- whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge that we already have, or to postpone the action that it appears to demand at a given time" (Hill, 1965). The observational data from a NIS can expand the knowledge from the RCTs and add to their generalizability. Optimally designed NIS can provide useful information about the effectiveness and safety of the medicinal product and can be used as a tool to monitor their benefit risk balance. Ideally, NIS and RCT should be used together to provide a more comprehensive understanding of a treatment or disorder. Therefore, NIS

should be considered as complementary rather than alternative to RCT. In future, the development of standards by the regulators and other stakeholders would improve the data quality and prevent misuses of NIS.

Strength and weakness of this dissertation. This dissertation is based on the real NIS studies that were conducted by commercial as well as non-commercial sponsors. This increased the robustness of our overall conclusion about the potentials of the NIS. Nevertheless, given the lack of harmonization on the conduct and regulatory requirements for NISs, the results of this work are not generalizable to other NIS answering the similar questions.

5. Outlook

A lot has changed in the drug development in the last few years. The RWE is revolutionizing the pharmaceutical industry. Various stakeholders in healthcare such as regulators and payers are interested in RWE. The new clinical trial regulation introduced a new term, i.e., low-interventional clinical trial. Article 2 (2) (3) of REG 536/2014 defines it as "a clinical trial where the investigational medicinal products, excluding placebos, are authorized and are used in accordance with the terms of the marketing authorization; or its use is evidence-based and supported by published scientific evidence on their safety and efficacy"(European Parliament, 2014). Furthermore, EMA supports the development of medicines for life threatening or orphan diseases, where there are unmet medical needs and issue a conditional marketing authorization based on less completed or limited evidence on medicine if their benefits outweigh their risk. The application of RWE in medicine is currently extended to developing areas in drug development such as precision medicines (Breckenridge et al., 2019). It additionally provides opportunities in oncology and rare diseases where conduction of RCTs may face ethical issues or are not feasible at all (Breckenridge et al., 2019). This would contribute to more understanding of the disease also and could shift the focus of the medicine form the treatment to prevention.

NIS is one of the tools to collect RWE; however, there is no global harmonization in the way that they are conducted and analyzed (Breckenridge et al., 2019, Koch et al., 2020). Regulatory authorities and other stakeholders in health care system need to develop infrastructure and data standard to promote sound research by encouraging rigorous data collection, analysis, and reporting. In April 2019, FDA issued a labeling extension for Ibrance® (palbociclib) to treat male breast cancer in combination with endocrine therapy based on RWE (Wedam et al., 2020). The sources of this evidence were RWD from the EHRs, insurance claims and the safety information from the global safety database and post-marketing reports for palbociclib from FDA safety database (Wedam et al., 2020). This shows the importance and recognition of the RWD, but whether the use of RWE become the standard for the regulatory decision-making is still unclear.

6. Summary

Randomized clinical trials (RCTs) are the gold standard for evaluating efficacy and safety of medicinal products, particularly for obtaining marketing authorization. The evidence from RCTs has high internal validity due to randomization and, most often, blinding. However, their external validity is limited, as they collect data form a homogeneous study population, in an experimental setting for a relative short time. Therefore, there is an increasing recognition that real world evidence (RWE) needs to complement RCTs. Large RCTs are challenging mainly due to high cost and difficulties in patient recruitment, thus RWE is mostly generated from observational studies, which are non-interventional (NIS). NISs have considerable external validity as they collect data from large number of participants from normal treatment settings and are useful to collect real life information on drug safety. Nevertheless, absence of randomization and control group limit their internal validity and there are concerns about data quality and hidden biases.

In this dissertation, we studied the value of NIS to determine factors associated with the initial dosing and up-titration of propiverine and study how dosing relative to other factors affects treatment outcome on overactive bladder (OAB) symptoms. Data from 2 prospective NIS of 1335 and 745 OAB patients, respectively, receiving propiverine, were analyzed post-hoc, using multivariate analysis. Furthermore, as an example of a NIS with retrospective data collection, we evaluated the potentials of routinely collected inpatient data as a complementary source for existing drug surveillance systems. We sampled 2326 cases using a standardized procedure and evaluated them regarding drug relation and preventability of event. Lastly, this thesis covers two common basic statistical reporting issues, as an implication for analysis and reporting.

This work showed the potential of NIS to complement the evidence from RCT to improve the effectiveness of the treatment by providing guidance for choosing the correct dose. Moreover, we found out that routinely collected inpatient data in the context of a NIS study design could be used as markers for non-preventable adverse drug effects and medication errors, thus providing a meaningful complement to existing drug surveillance system. Finally, we highlighted the importance of a proper statistical analysis and reporting in the interpretation and reporting of the study finding in evidence-based medicine.

7. Zusammenfassung auf Deutsch

Randomisierte klinische Studien (RCTs) sind der Goldstandard für die Bewertung der Wirksamkeit und Sicherheit von Arzneimitteln. Die aus RCTs gewonnene Evidenz hat aufgrund der Randomisierung und der Verblindung eine hohe interne Validität. Ihre externe Validität ist jedoch begrenzt. Daher wird zunehmend anerkannt, dass reale Evidenz (Real World Evidence, RWE) RCTs ergänzen muss. Große RCTs sind vor allem wegen der hohen Kosten und der Schwierigkeiten bei der Patientenrekrutierung eine Herausforderung. Daher wird RWE meist aus Beobachtungsstudien generiert, die nicht-interventionell (NIS) sind. Sie haben eine beträchtliche externe Validität. Nichtsdestotrotz schränken die fehlende Randomisierung und die fehlende Kontrollgruppe ihre interne Validität ein und es gibt Bedenken hinsichtlich der Datenqualität und versteckter Verzerrungen.

In dieser Dissertation untersuchten wir den Wert der NIS, um Faktoren zu bestimmen, die mit der anfänglichen Dosierung und Aufdosierung von Propiverin zusammenhängen und wie die Dosierung im Verhältnis zu anderen Faktoren das Behandlungsergebnis bei Symptomen der überaktiven Blase (OAB) beeinflusst. Daten aus zwei prospektiven NIS von 1335 bzw. 745 OAB-Patienten, die Propiverin erhielten, wurden post-hoc mittels multivariater Analyse analysiert. Darüber hinaus haben wir die Potenziale von routinemäßig erhobenen stationären Daten als ergänzende Quelle für bestehende Systeme zur Überwachung der Arzneimittelsicherheit evaluiert. Wir haben 2326 Fälle nach einem standardisierten Verfahren stichprobenartig untersucht und im Hinblick auf die Medikamentenrelation und die Vermeidbarkeit des Ereignisses ausgewertet. Schließlich werden in dieser Arbeit zwei grundlegende Probleme der statistischen Berichterstattung behandelt, die sich auf die Analyse und das Berichtswesen auswirken.

Diese Arbeit zeigte das Potenzial der NIS, die Evidenz aus RCTs zu ergänzen, um die Wirksamkeit der Behandlung zu verbessern. Darüber hinaus haben wir herausgefunden, dass routinemäßig erhobene stationäre Daten im Rahmen eines NIS-Studiendesigns als Marker für nicht vermeidbare unerwünschte Arzneimittelwirkungen und Medikationsfehler verwendet werden können und somit eine sinnvolle Ergänzung zum bestehenden Arzneimittelüberwachungssystem darstellen. Abschließend wurde die Bedeutung einer korrekten statistischen Analyse und Berichterstattung bei der Interpretation und Darstellung der Studienergebnisse in der evidenzbasierten Medizin hervorgehoben.

8. Reference List

1. ANGLEMYER, A., HORVATH, H. T. & BERO, L. 2014. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev**,** Mr000034.

2. BENSON, K. & HARTZ, A. J. 2000. A comparison of observational studies and randomized, controlled trials. N Engl J Med, 342**,** 1878-86.

3. BORING, E. G. 1919. Mathematical vs. scientific significance. Psychological Bulletin, 16**,** 335-338.

4. BRECKENRIDGE, A. M., BRECKENRIDGE, R. A. & PECK, C. C. 2019. Report on the current status of the use of real-world data (RWD) and real-world evidence (RWE) in drug development and regulation. Br J Clin Pharmacol, 85**,** 1874-1877.

5. CARDOZO, L., AMARENCO, G., PUSHKAR, D., MIKULAS, J., DROGENDIJK, T., WRIGHT, M. & COMPION, G. 2013. Severity of overactive bladder symptoms and response to dose escalation in a randomized, double-blind trial of solifenacin (SUNRISE). BJU Int, 111**,** 804-10.

6. CARDOZO, L., HALL, T., RYAN, J., EBEL BITOUN, C., KAUSAR, I., DAREKAR, A. & WAGG, A. 2012. Safety and efficacy of flexible-dose fesoterodine in British subjects with overactive bladder: insights into factors associated with dose escalation. Int Urogynecol J, 23**,** 1581-90.

7. CLAUDOT, F., ALLA, F., FRESSON, J., CALVEZ, T., COUDANE, H. & BONAÏTI-PELLIÉ, C. 2009. Ethics and observational studies in medical research: various rules in a common framework. Int J Epidemiol, 38**,** 1104-8.

8. COLLINS, R., BOWMAN, L., LANDRAY, M. & PETO, R. 2020. The Magic of Randomization versus the Myth of Real-World Evidence. N Engl J Med, 382**,** 674-678.

9. DAVIES, E. C., GREEN, C. F., TAYLOR, S., WILLIAMSON, P. R., MOTTRAM, D. R. & PIRMOHAMED, M. 2009. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. PLoS One, 4**,** e4439.

10. DE LANGE, D. W., GUIDET, B., ANDERSEN, F. H., ARTIGAS, A., BERTOLINI, G., MORENO, R., CHRISTENSEN, S., CECCONI, M., AGVALD-OHMAN, C., GRADISEK, P., JUNG, C., MARSH, B. J., OEYEN, S., BOLLEN PINTO, B., SZCZEKLIK, W., WATSON, X., ZAFEIRIDIS, T. & FLAATTEN, H. 2019. Huge variation in obtaining ethical permission for a non-interventional observational study in Europe. BMC Med Ethics, 20**,** 39.

11. DEQUITO, A. B., MOL, P. G., VAN DOORMAAL, J. E., ZAAL, R. J., VAN DEN BEMT, P. M., HAAIJER-RUSKAMP, F. M. & KOSTERINK, J. G. 2011. Preventable and non-preventable adverse drug events in hospitalized patients: a prospective chart review in the Netherlands. Drug Saf, 34**,** 1089-100.

12. DIONG, J., BUTLER, A. A., GANDEVIA, S. C. & HEROUX, M. E. 2018. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. PLoS One, 13**,** e0202121.

13. EMA. 2011. European Medicine Agency, European Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Considerations on the definition of non-interventional trials under the current legislative framework ("clinical trials directive"2001/20/EC) [Online]. Available: http://www.encepp.eu/publications/documents/ENCePPinterpretationofnoninterventionalstudies.pdf [Accessed 11 March 2021].

14. EMA. 2012. European Medicines Agency, Guidance for the format and content of the protocol of non-interventional post-authorisation safety studies [Online]. Available: https://www.ema.europa.eu/en/documents/other/guidance-format-content-protocol-non-interventional-post-authorisation-safety-studies_en.pdf [Accessed March 2021].

15. EMA. 2017. European Medicine Agency, Guideline on good pharmacovigilance practices (GVP) Module VIII – Post-authorisation safety studies (Rev 3) [Online]. Available: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-gvp-module-viii-post-authorisation-safety-studies-rev-3_en.pdf [Accessed March 2021].

16. EMA. 2021. European Medicines Agency, European Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) webinar for Academia - Real world research on medicines [Online]. Available: https://www.ema.europa.eu/en/documents/presentation/presentation-encepp-webinar-academia-real-world-research-medicines-x-kurz-fsalvo-ema_en.pdf [Accessed 11 March 2021].

17. EMA. no date. European Medicine Agency, Clinical Trial Regulation [Online]. Available: https://www.ema.europa.eu/en/human-regulatory/research-development/clinical-trials/clinical-trial-regulation [Accessed March 2021].

18. EQUATOR. No date. The Network for Enhancing the Quality and Transparency of Health Research. What is a reporting guideline? [Online]. Available: https://www.equator-network.org/about-us/what-is-a-reporting-guideline/ [Accessed March 2021].

19. EUROPEAN PARLIAMENT. 2001. Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on

medicinal products for human use. [Online]. Official Journal of the European Union.Available: http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:121:0034:0044:EN:PDF [Accessed March 2021].

20. EUROPEAN PARLIAMENT. 2010. DIRECTIVE 2010/84/EU of the European Parliament and of the Council of 15 December 2010 amending, as regards pharmacovigilance, Directive 2001/83/EC on the Community code relating to medicinal products for human use [Online]. Official Journal of the European Union. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:348:0074:0099:EN:PDF [Accessed March 2021].

21. EUROPEAN PARLIAMENT. 2012. Commission Implementation Regulation (EU) No 520/2012 of 19 June 2012 on the performance of pharmacovigilance activities provided for in Regulation (EC) No 726/2004 of the European Parliament and of the Council and Directive 2001/83/EC of the European Parliament and of the Council [Online]. Official Journal of the European Union. Available: https://eur-lex.europa.eu/eli/reg_impl/2012/520/oj [Accessed March 2021].

22. EUROPEAN PARLIAMENT. 2014. Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC. [Online]. Official Journal of the European Union. Available: http://ec.europa.eu/health/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf [Accessed March 2021].

23. EUROPEAN PARLIAMENT. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [Online]. Official Journal of the European Union. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj [Accessed March 2021].

24. EUROPEAN PARLIAMENT. 2021. Clinical Trials Regulation (EU) No 536/2014 Draft Questions & Answers, Version 3 [Online]. Available: https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-10/regulation536_2014_qa_en.pdf [Accessed March 2021].

25. FDA. 2016. Food and Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices: draft guidance for industry and Food and Drug Administration staff. [Online]. Available: (http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM513027.pdf. opens in new tab). [Accessed 11 March 2021].

26. GALE, E. A. 2012. Post-marketing studies of new insulins: sales or science? Bmj, 344**,** e3974.

27. GALSON, S. A. G. S. 2016. Real-World Evidence to Guide the Approval and Use of New Treatments. NAM Perspectives. Discussion Paper, National Academy of Medicine, Washington, DC.

28. GOLDMAN, H. B., OELKE, M., KAPLAN, S. A., KITTA, T., RUSSELL, D., CARLSSON, M., ARUMI, D., MANGAN, E. & NTANIOS, F. 2019. Do patient characteristics predict which patients with overactive bladder benefit from a higher fesoterodine dose? Int Urogynecol J, 30**,** 239-244.

29. GRODSTEIN, F., MANSON, J. E. & STAMPFER, M. J. 2006. Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. J Womens Health (Larchmt), 15**,** 35-44.

30. HEMKENS, L. G., CONTOPOULOS-IOANNIDIS, D. G. & IOANNIDIS, J. P. 2016. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. Bmj, 352**,** i493.

31. HERNÁN, M. A. 2018. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. Am J Public Health, 108**,** 616-619.

32. HILL, A. B. 1965. THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION? Proc R Soc Med, 58**,** 295-300.

33. HIRST, A. & ALTMAN, D. G. 2012. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. PLoS One, 7**,** e35621.

34. IOANNIDIS, J. P. 2005. Why most published research findings are false. PLoS Med, 2**,** e124.

35. ISOP. 2016. Public Policy Committee, International Society of Pharmacoepidemiology, Guidelines for good pharmacoepidemiology practice (GPP). [Online]. Pharmacoepidemiology and Drug Safety 25(1):2–10. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.3891 [Accessed March 2021].

36. JOHNSON, E. S., BARTMAN, B. A., BRIESACHER, B. A., FLEMING, N. S., GERHARD, T., KORNEGAY, C. J., NOURJAH, P., SAUER, B., SCHUMOCK, G. T., SEDRAKYAN, A., STÜRMER, T., WEST, S. L. & SCHNEEWEISS, S. 2013. The incident user design in comparative effectiveness research. Pharmacoepidemiol Drug Saf, 22**,** 1-6.

37. KOCH, C., SCHLEEFF, J., TECHEN, F., WOLLSCHLÄGER, D., SCHOTT, G., KÖLBEL, R. & LIEB, K. 2020. Impact of physicians' participation in non-interventional post-marketing studies on their prescription habits: A retrospective 2-armed cohort study in Germany. PLoS Med, 17**,** e1003151.

38. LABRECQUE, J. A. & SWANSON, S. A. 2017. Target trial emulation: teaching epidemiology and beyond. Eur J Epidemiol, 32**,** 473-475.

39. LANG, T. 2004. Twenty statistical errors even you can find in biomedical research articles. Croat Med J, 45**,** 361-70.

40. MANSON, J. E., HSIA, J., JOHNSON, K. C., ROSSOUW, J. E., ASSAF, A. R., LASSER, N. L., TREVISAN, M., BLACK, H. R., HECKBERT, S. R., DETRANO, R., STRICKLAND, O. L., WONG, N. D., CROUSE, J. R., STEIN, E. & CUSHMAN, M. 2003. Estrogen plus progestin and the risk of coronary heart disease. N Engl J Med, 349**,** 523-34.

41. MEHRA, M. R., DESAI, S. S., KUY, S., HENRY, T. D. & PATEL, A. N. 2020. Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med, 382**,** e102.

42. MICHEL, M. C., BOHNER, H., KÖSTER, J., SCHÄFERS, R. & HEEMANN, U. 2004. Safety of telmisartan in patients with arterial hypertension : an open-label observational study. Drug Saf, 27**,** 335-44.

43. MICHEL, M. C., MEHLBURGER, L., BRESSEL, H. U., SCHUMACHER, H., SCHÄFERS, R. F. & GOEPEL, M. 1998. Tamsulosin treatment of 19,365 patients with lower urinary tract symptoms: does co-morbidity alter tolerability? J Urol, 160**,** 784-91.

44. MICHEL, M. C., MEHLBURGER, L., SCHUMACHER, H., BRESSEL, H. U. & GOEPEL, M. 2000. Effect of diabetes on lower urinary tract symptoms in patients with benign prostatic hyperplasia. J Urol, 163**,** 1725-9.

45. MICHEL, M. C., MINARZYK, A., SCHWERDTNER, I., QUAIL, D., METHFESSEL, H. D. & WEBER, H. J. 2013. Observational study on safety and tolerability of duloxetine in the treatment of female stress urinary incontinence in German routine practice. Br J Clin Pharmacol, 75**,** 1098-108.

46. MICHEL, M. C., MURPHY, T. J. & MOTULSKY, H. J. 2020. New Author Guidelines for Displaying Data and Reporting Data Analysis and Statistical Methods in Experimental Biology. Drug Metab Dispos, 48**,** 64-74.

47. MICHEL, M. C., OELKE, M., GOEPEL, M., BECK, E. & BURKART, M. 2007. Relationships among symptoms, bother, and treatment satisfaction in overactive bladder patients. Neurourol Urodyn, 26**,** 190-5.

48. MICHEL, M. C., WETTERAUER, U., VOGEL, M. & DE LA ROSETTE, J. J. 2008. Cardiovascular safety and overall tolerability of solifenacin in routine clinical use: a 12-week, open-label, post-marketing surveillance study. Drug Saf, 31**,** 505-14.

49. MORTON, S. C., COSTLOW, M. R., GRAFF, J. S. & DUBOIS, R. W. 2016. Standards and guidelines for observational studies: quality is in the eye of the beholder. J Clin Epidemiol, 71**,** 3-10.

50. RAMIREZ, I. 2015. Navigating the maze of requirements for obtaining approval of non-interventional studies (NIS) in the European Union. Ger Med Sci, 13**,** Doc21.

51. RAY, W. A. 2003. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol, 158**,** 915-20.

52. SACKETT, D. L., ROSENBERG, W. M., GRAY, J. A., HAYNES, R. B. & RICHARDSON, W. S. 1996. Evidence based medicine: what it is and what it isn't. Bmj, 312**,** 71-2.

53. SCHNEIDER, T., ARUMI, D., CROOK, T. J., SUN, F. & MICHEL, M. C. 2014. An observational study of patient satisfaction with fesoterodine in the treatment of overactive bladder: effects of additional educational material. Int J Clin Pract, 68**,** 1074-80.

54. SHEIN-CHUNG CHOW; JEN-PEI LIU 2013. Design and Analysis of Clinical Trials: Concepts and Methodologies,.

55. SINGAL, A. G., HIGGINS, P. D. & WALJEE, A. K. 2014. A primer on effectiveness and efficacy trials. Clin Transl Gastroenterol, 5**,** e45.

56. SMITH, B. P. 2020. Where Will Statistical Sciences for Clinical Pharmacology Be in 2030? Clin Pharmacol Ther, 107**,** 17-21.

57. SPELSBERG, A., PRUGGER, C., DOSHI, P., OSTROWSKI, K., WITTE, T., HÜSGEN, D. & KEIL, U. 2017. Contribution of industry funded post-marketing studies to drug safety: survey of notifications submitted to regulatory agencies. Bmj, 356**,** j337.

58. VON JEINSEN, B. K. & SUDHOP, T. 2013. A 1-year cross-sectional analysis of non-interventional post-marketing study protocols submitted to the German Federal Institute for Drugs and Medical Devices (BfArM). Eur J Clin Pharmacol, 69**,** 1453-66.

59. WAGG, A., DAREKAR, A., ARUMI, D., KHULLAR, V. & OELKE, M. 2015. Factors associated with dose escalation of fesoterodine for treatment of overactive bladder in people >65 years of age: A post hoc analysis of data from the SOFIA study. Neurourol Urodyn, 34**,** 438-43.

60. WASSERSTEIN, R. L. & LAZAR, N. A. 2016. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70**,** 129-133.

61. WEDAM, S., FASHOYIN-AJE, L., BLOOMQUIST, E., TANG, S., SRIDHARA, R., GOLDBERG, K. B., THEORET, M. R., AMIRI-KORDESTANI, L., PAZDUR, R. & BEAVER, J. A. 2020. FDA Approval Summary: Palbociclib for Male Patients with Metastatic Breast Cancer. Clin Cancer Res, 26**,** 1208-1212.

62. WYNDAELE, J. J., GOLDFISCHER, E. R., MORROW, J. D., GONG, J., TSENG, L. J. & CHOO, M. S. 2011. Patient-optimized doses of fesoterodine improve bladder symptoms in an open-label, flexible-dose study. BJU Int, 107**,** 603-11.

63. ZENG, C., DUBREUIL, M., LAROCHELLE, M. R., LU, N., WEI, J., CHOI, H. K., LEI, G. & ZHANG, Y. 2019. Association of Tramadol With All-Cause Mortality Among Patients With Osteoarthritis. Jama, 321**,** 969-982.

9. List of abbreviations

ADE	Adverse drug event
CI	Confidence interval
CONSORT	Consolidated Standards of Reporting Trial
CTIS	Clinical Trial Information System
DIR	Directive
EC	Ethics Committee
EHR	Electronic Healthcare Record
EMA	European Medicine Agency
ENCePP	European Network of Centers for Pharmacoepidemiology and Pharmacovigilance
EQUATOR	Network for Enhancing the Quality and Transparency of Health Research
EU	European Union
FDA	Food and drug administration
GCP	Good Clinical Practice
GPP	Good Pharmacoepidemiology Practice
GVP	Good Pharmacovigilance Practice
HRT	Hormone replacement therapy
MAH	Marketing Authorization Holder
MHRA	Medicine and Healthcare Products Regulatory Agency
NCA	National Competent Authority
NHST	Null hypothesis significance testing
NHT	Null hypothesis testing
NIS	Non-interventional study
npADE	non-preventable adverse drug event
OAB	Overactive Bladder
pADEs	Preventable adverse events
PAS	Post-authorization Study
PASS	Post-authorization Safety Study
PRAC	Pharmacovigilance Risk Assessment Committee
RCT	Randomised Clinical Trial
REG	Regulation
RWD	Real World Data
RWE	Real World Evidence
SmPCs	Summary of the product characteristics
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
UK	United Kingdom
USA	United States of America

10. Acknowledgements

I would like to thank my supervisors, Professor Dr. med. Martin C. Michel and Professor Dr. med. Andreas Stang, who provided an enormous amount of guidance and insight throughout my dissertation. Their detailed feedback on each one of my drafts and their willingness to answer all questions made this process substantially easier.

I would also like to thank Professor Charles Poole from university of North Carolina, who provided me with additional research literature. I am grateful to him for his treasured support that was influential in shaping my research methods and analyzing my results.

I would also like to express my sincere gratitude to Dr. Sandra Murgas from Apogepha for providing the data of their studies and making this dissertation possible. I am thankful to Prof. Dr. med. Tim Schneider and Prof. Dr. med. Dr. phil. M. Oelke for their contribution and insightful comments to my papers for this dissertation.

My gratitude extends to the Institute for Medical Informatics, Biometry and Epidemiology in Essen for giving me the opportunity to undertake my PhD studies at the medical faculty of University hospital Essen. I am thankful to all my current and former colleagues. I would like to thank my colleague Monika Grätsch for her emotional support and I would like to acknowledge her for providing me insights for data management and statistical analysis with her expertise in SAS.

I would also like to thank Professor Dr. med. Jürgen Stausberg, Professor Dr. med. Karl Heinz Jöckel, and Dr. Nils Kuklik. It was a great pleasure to be the part of the project UAE Detect under their supervision.

Lastly, I would like to thank my parents and friends for their unparalleled emotional support in this journey. Thank you for standing by my side and encouraging me.

11. Curriculum vitae

The curriculum vitae is not included in the online version for data protection reasons.