NATIONAL EVALUATION SYSTEM FOR HEALTH TECHNOLOGY

NESTcc Research Methods Framework– *A Practical Guide to RWE for Medical Devices*

A Report of the Research Methods Subcommittee of the NEST Coordinating Center – An initiative of MDIC

National Evaluation System for health Technology Coordinating Center (NESTcc) Methods Framework

Subcommittee Members:

- *Naomi Aronson; Formerly Blue Cross Blue Shield Association*
- *Jesse A. Berlin, ScD (Chair); School of Public Health and Center for Pharmacoepidemiology and Treatment Science, Rutgers University*
- *Mitchell Krucoff, MD; Duke University Medical Center/Duke Clinical Research Institute(DCRI)*
- *Didier Morel, PhD; Becton, Dickinson, and Company*
- *Scott Snyder, PhD; Cook Research Incorporated*
- *Mwanatumu S. Mbwana, PhD, RAC (US, EU); MDIC/NESTcc***
- *Graeme L. Hickey, PhD; Medtronic*
- *Shumin Zhang, MD, ScD; Johnson & Johnson*
- *Ulka Campbell, PhD; Aetion Inc.*

Additional Contributors:

- *Jordan Hirsch, MHA; Formerly MDIC/NESTcc***
- *Sarah Merlino, MPH; MDIC/NESTcc***
- *Panagiotis Mavros, PhD, CERobs Consulting LLC**
- Mary Beth Ritchey, PhD, FISPE, CERobs Consulting LLC*

** These contributors were compensated for their contributions and editing work on this publication. Their involvement reflects their professional expertise and was provided under a contractual agreement. Their contributions do not reflect the views or opinions of their respective institutions, organizations or employers.*

*** These contributors participated as MDIC/NEST employees.*

For Sub-committee members: Unless otherwise noted, sub-Committee members have voluntarily contributed their personal time to this publication. Their contributions reflect their independent views and experience, free from funding or other benefits that may influence the content of this document. Participation was in a personal capacity and does not reflect view of their respective institutions, organizations or employer.

Names of Committee Members who are federal employees have been temporarily excluded from this draft document. We are awaiting further guidance from the agency regarding the "Pause on Issuing Documents and Public Communication" issued by certain federal agencies.

Funding Disclosures

As part of its commitments outlined in the 2023 Medical Device User Fee Amendments (MDUFA V), the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) provides User Fee funds to the National Evaluation System for health Technology Coordinating Center (NESTcc) in the Medical Device Innovation Consortium (MDIC) to: i) support the development of RWD resources to facilitate appropriate access for research studies; ii) convene experts to develop best practices and, advance innovative

methodology approaches with respect to RWE development and analysis. Funding for this work was made in part possible by FDA of HHS from industry user fees administered through a financial assistance award (FAIN# U01FD006292) to the MDIC. The contents are those of the author(s) and/or presenter(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.

Table of Contents

Preface

In 2012, FDA announced its vision for a medical device program to "quickly identify problematic devices, accurately and transparently characterize and disseminate information about device performance in clinical practice, and efficiently generate data to support premarket clearance or approval of new devices and new uses of currently marketed devices[1]." Soon thereafter, a multi-stakeholder effort began to establish the National Evaluation System for Health Technology (NEST) conducting much of its work acting as a Coordinating Center (NESTcc) bringing diverse stakeholders to the table.

NESTcc was established in 2016 with funding from the United States Food and Drug Administration (FDA) through a U01 Cooperative Agreement funded in part under the Medical Device User Fee and Modernization Act (MDUFA). Per MDUFA IV and MDUFA V commitments, NESTcc operates under the guidance of a Governance Committee to help ensure its work is in the best interest of the entire Medical Device Ecosystem, including health systems, patient groups, industry, clinicians, payers, and regulators. Its work is intended to help solve the unique challenges of using real-world data (RWD) to generate real world-evidence (RWE) in the study of medical devices.

NESTcc aims to support the sustainable generation and use of timely, reliable, and cost-effective RWE throughout the medical device total product lifecycle (TPLC), using high-quality RWD that are analyzed using robust methodological standards. Stakeholders across the medical device ecosystem, stand to benefit from improved use of RWD generated in the course of clinical care and everyday life to produce valid RWE. Opportunities include increased patient awareness of device safety issues, efficient and low-cost evidence generation for regulatory review and reimbursement purposes, and improved patient and provider ability to make care decisions based on robust evidence.

In 2018, NESTcc's Governing Committee commissioned two Subcommittees to develop Frameworks on Data Quality and Research Methods to support the development of high-quality RWE studies of medical devices. The Subcommittees included representatives from health systems, NESTcc Network Collaborators, medical device manufacturers, and the FDA. These original frameworks built upon existing work and utilized members' knowledge from similar initiatives like PCORnet, Sentinel, and MDEpiNet. They aimed to guide medical device ecosystem stakeholders in collaborating with NESTcc to ensure high-quality data and research methodology.

The first versions of both the Research Methods Framework and the Data Quality Framework were released in February 2020. The Data Quality Subcommittee reconvened in late 2020 to begin revisions to this Framework based on stakeholder feedback and lessons learned from 21 NESTcc RWE Test-Cases that were chosen through an Open Call Process. These test-cases explored the feasibility for medical device ecosystem stakeholders to work with RWD sources and NESTcc's initial set of Network Collaborators, and to identify areas where NESTcc could play a role in reducing transaction costs [e.g., contracting, Institutional Review Board (IRB) approvals, data sharing agreements, publication policies]. Descriptions of the 21 NESTcc Test-Cases are available on the NESTcc website[2]. Some test-cases progressed beyond feasibility leading to an FDA-approved label extension[3] [the first using solely a comparative EHR database RWE study for a label extension approved by FDA's Center for Device and Radiological Health (CDRH)] and postmarketing safety surveillance studies[4].

The test-cases also revealed strengths and limitations of specific data sources and the challenges

involved in creating datasets suitable for conducting regulatory-grade research[5-7]. NESTcc has now pivoted from the "test-case" environment to one of "implementation" using what was learned under MDUFA IV. One primary goal of the implementation case projects is the further development of the NEST Mark™ review approach to evaluation of RWD. The NEST Mark approach is designed to help de-risk the use of RWD for supporting regulatory filings using well-defined processes for evaluation of relevance and reliability of RWD specifically to generate RWE. As part of this process, NESTcc applies a systematic and consistent approach to evaluate essential data quality and study design elements from FDA's Guidance Documents and builds on the NEST Frameworks. This leads to a NEST Mark review report enhancing confidence that covered RWD are relevant and reliable to meet scientific and regulatory objectives for medical devices.

These next versions of the Research Methods and Data Quality Frameworks incorporate NESTcc's knowledge gained from test cases and early NEST Mark implementations, along with the subcommittee's extensive RWD experience. They include a broader range of data sources for quality assessment, consider recent RWD guidance documents, and offer additional RWE examples and best practices. They reflect the evolution in RWE innovation, aiming to provide a comprehensive resource for medical device ecosystem stakeholders.

On behalf of NESTcc, we would like to extend our heartfelt gratitude to each and every one of both our current and past subcommittee members for their incredible dedication and invaluable contributions. Their selfless commitment to creating these frameworks has provided an essential resource for professionals working with RWD and RWE in the medical device ecosystem. The expertise and hard work of the subcommittee will not only advance the field but will also pave the way for improved patient outcomes.

Jesse Berlin, ScD, School of Public Health and Center for Pharmacoepidemiology and Treatment Science, Rutgers University

Paul Coplan, ScD, MBA, FISPE, Johnson & Johnson MedTech Epidemiology & RWD Science

Adam Donat, JD, MS, Quest Diagnostics

Richard Smith, MBA, Senior Vice President, NEST

Jill Dreyfus, PhD, MPH, Senior Director of Evidence Generation, NEST

References

1. Gottlieb S, Shuren JE. Statement from FDA Commissioner Scott Gottlieb, M.D. and Jeff Shuren, M.D., Director of the Center for Devices and Radiological Health, on FDA's updates to Medical Device Safety Action Plan to enhance post-market safety. Food and Drug Administration. Accessed 12/22/2023, 2023. https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-and-jeff-shuren-md-director-center-devices-and-2
2. NEST Test Cases. https://nestcc.org/test-cases/
3. Dhruva SS, Zhang S, Chen J, et al. Safety and Effectiveness of a Catheter With Contact Force and 6-Hole Irrigation for Ablation of Persistent Atrial Fibrillation in Routine Clinical Practice. *JAMA Netw Open*. Aug 1 2022;5(8):e2227134. doi:10.1001/jamanetworkopen.2022.27134
4. Frankenberger EA, Resnic FS, Ssemaganda H, et al. Evaluation of intervertebral body implant performance using active surveillance of electronic health records. *BMJ Surg Interv Health Technol*. 2022;4(1):e000125. doi:10.1136/bmjsit-2021-000125
5. Blake HA, Sharples LD, Harron K, van der Meulen JH, Walker K. Probabilistic linkage without personal information successfully linked national clinical datasets. *J Clin Epidemiol*. Aug 2021;136:136-145. doi:10.1016/j.jclinepi.2021.04.015
6. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health*. May 14 2014;11(5):5170-207. doi:10.3390/ijerph110505170
7. Gottlieb S, Shuren JE. Statement from FDA Commissioner Scott Gottlieb, M.D. and Jeff Shuren, M.D., Director of the Center for Devices and Radiological Health, on FDA's updates to Medical Device Safety Action Plan to enhance post-market safety. Food and Drug Administration. Accessed 12/22/2023, 2023. https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-and-jeff-shuren-md-director-center-devices-and-2

Introduction

In 2018, NESTcc's Research Methods Subcommittee was tasked with developing a pragmatic methodological framework or "living playbook" that could be used by all stakeholders across the NESTcc medical device ecosystem in designing, executing and evaluating the validity of research studies using RWD. The Research Methods Framework was also intended to highlight device-specific considerations in benefit/risk studies based on both observational and experimental designs. The first version of the framework was posted on the NESTcc website in February 2020.

The Research Methods Framework emphasized two key principles in the design of RWE studies: (1) pre-specification of study design to avoid the fact or appearance of cherry-picked design decisions or selective data mining; and (2) close attention to potential confounders and how they will be addressed in the study design. The Framework organized its content into the form of a study protocol to provide step-by-step advice for readers planning medical device studies.

The NESTcc Research Methods Subcommittee consisted of a diverse range of stakeholders with academic, regulatory, and industry methodological expertise on constructs of study design and statistical methods. The Framework is intended to provide robust design and analytic principles for regulatory science, best practice recommendations, for any human subjects study endeavoring to quantify benefit/risk or safety outcomes causally associated with the use of medical devices. A complementary document offering a broader overview of such principles in incorporating external data, the External Evidence Methods Framework, was made publicly available at https://mdic.org/project/external-evidence-methods-framework/.

Organization and Approach

In the current version, the Subcommittee has tried to extend the scope of the framework to include more granular and pragmatic steps in the development of a RWE study. Expanding on numerous aspects of the original, this document is intended to convey the "living and learning" core character of these NESTcc frameworks as facilitative tools to help investigators design high quality research programs using RWD. The current version thus adds substantial detail and medical device-specific examples and addresses methodological gaps in the original Framework. It adds particular emphasis on two areas: (1) devising methods explicitly based on unique study design needs for RWD involving medical devices, and (2) approaches for combining multiple sources of data, such as hospital EHR data with registry or licensed third party claims data, fundamental to the establishment of "coordinated registry networks (CRTNs).[1]"

Scope

Devices: All types of medical devices including therapeutic, diagnostic and imaging, in vitro diagnostics, implantables, wearables, and software. Most current examples are based on therapeutics because the vast majority of studies conducted and published are in those areas. As experience grows in other areas, such as diagnostics, the framework will continue to be updated with additional examples.

Stage of Development: Research and development studies spanning the entirety of the total product life cycle (TPLC) are in scope, including early feasibility/first-in-human, pre- and post-market approval, post-market surveillance, health technology assessment (HTA) and reimbursement.

Risk category: All risk categories are included (Class I, II and III devices).

Study/data types: A wide range of studies, including both prospective data collection and retrospective data curation designs, are in scope as long as they involve the use of RWD, for example:

- Observational studies using administrative data (insurance claims or EHRs) where there is an exposure of interest AND a comparator
- Registry-based studies (prospective and retrospective)
- RWE studies to supplement clinical trial data through linking registries and/or databases
- A single-arm clinical trial that is compared with an external control group identified from another source, which could be a registry, claims data, EHRs, or past clinical trials
- A single-arm clinical trial in which the outcomes are compared to an Objective Performance Criterion

Included are methods applied to retrospective data curated from existing data collection infrastructure (e.g., claims and EHRs), where the data elements are generally predefined, compared with methods applied to prospective data collection studies, where the Health Care Practitioner (HCP) investigators largely have control over what and how data are collected on their patients. In some situations, e.g., a randomized trial that incorporates data from EHRs, a hybrid approach might be proposed.

Historically, randomized controlled (experimental) trials (RCTs) have not incorporated RWD, hence in the current framework detailed methodologies are not focused on this area. However, there is growing interest in hybrid study designs where including RWD capture for prospective RCTs of medical devices may add quality and efficiency to the conduct of RCTs per se, as described in this framework. This may be an area to be expanded in the living framework process at a later time.

As in the first Framework, this version does not focus on data quality and reliability but assumes that the data proposed in the protocol have been evaluated for completeness and accuracy for use in medical device evaluation. Data quality, along with issues related to data governance, characteristics, capture, transformation, and curation are covered by the NESTcc Data Quality Framework. Issues related to validation of coding, i.e., does the event or exposure contained in the database accurately represent the actual patient diagnosis or experience are covered in this document.

Audience

While this framework provides useful information for all stakeholders interested in medical

devices and RWE, the document was written with three primary audiences in mind: (1) manufacturers who are likely to sponsor clinical studies on their products, (2) researchers who are likely to design and execute those studies, and (3) regulators and payers who will review and utilize the results of those studies for regulatory and reimbursement decision making. **It is crucial to establish open, candid, and thorough conversations among these various groups as early as possible in the conceptualization and design of a proposed study. If the study is being done for regulatory or for reimbursement purposes, the sponsor should speak with the relevant federal regulatory or payors agency in the very early stages of planning. Explicit discussions will generally help ensure that the methodology of, and analytic plan for the completed study, can provide the necessary information for decision-making.**

Subcommittee's primary considerations in framework update

The Subcommittee adopted two principles during their deliberations: (1) pre-specification, and (2) justification for the method(s) used to control confounding. As a first step in developing the Research Methods Framework, the Subcommittee created a protocol template, which builds upon existing bodies of work and uses the Subcommittee members' knowledge and experience from similar initiatives, including the U.S. FDA, the Medical Device Epidemiology Network (MDEpiNet), PCORnet, and Sentinel. The Framework is intended to promote pre-specification of as much detail as possible prior to data analysis to be transparent regarding what was and was not pre-specified when presenting findings. The Subcommittee noted that the data supporting medical device evaluations could be retrospectively or prospectively collected; the data may be from EHRs, registries, insurance claims data, patients, or a combination of these sources. A critical strategy in bolstering the validity of RWE, however, is pre-specification. Pre-specification of study design features and of analytical strategies will help reduce selective reporting of study results.

The second principle adopted by the Subcommittee related to how confounders, variables related to both medical device use and outcomes, will be controlled for in the study design, analysis, or both. Randomization, which can control for both measured and unmeasured confounders, is one approach. In the absence of randomization, regression, matching, or other statistical tools attempt to provide statistical control of the measured confounders. For this reason, the template developed applies to both randomized and non-randomized designs. Pre-specification of study design features and of analytical strategies will help reduce selective reporting of study results.

Preliminary Considerations in Developing RWE for Medical Devices

This Framework focuses attention on Medical Devices and RWE, both of which present special challenges. This section outlines some of those challenges and discusses how to address them at the beginning of the study design process.

1. **Consider both the advantages and limitations of RWE versus a clinical trial, or of incorporating RWE into a clinical trial (which may, in turn, generate further RWE), to reach study objectives.**

There are several main reasons to consider RWD vs. a clinical trial with uniquely constructed case report forms. If retrospective analysis of an already existing large, high quality data set is "fit for purpose" for a study defining benefit/risk or safety outcomes, there may be substantial time and cost savings compared to designing and executing a prospective clinical trial. If such data are questionable in quality (high missingness on some key variables, limited long term follow up…) then imbedding a prospective trial into an existing data infrastructure with an emphasis on quality may still realize "hybrid" efficiencies compared to designing and executing a unique clinical trial structure from the ground up. Key considerations thus include the need for increased sample size, potential to collect long term outcomes, pragmatic experience in actual clinical practice (which might be more relevant to informing clinical decision-making), diverse data sources for wider applicability of results, ethical and pragmatic concerns with patient randomization, and investigator blinding and savings in cost and time required to reach a conclusion. The recently updated FDA guidance on "Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices"[2] has provided a comprehensive list of the potential uses of RWD/RWE for regulatory submissions:

- "To generate hypotheses to be tested in a clinical study;
- As a historical control, an informative alternative sampling prior distributions in a Bayesian analysis of a clinical trial, or as one source of data in a hierarchical model or a hybrid data synthesis;
- As a concurrent control group or as a mechanism for collecting data to support marketing authorization when a registry, EHR, claims data, or some other systematic data collection mechanism exists;
- As a mechanism for re-training artificial intelligence/machine learning-enabled medical devices;
- To generate evidence to identify, demonstrate, or support the clinical validity of a biomarker or clinical outcome assessment;
- To generate (primary) clinical evidence to support marketing authorization (e.g., HDE, PMA, 510(k) or De Novo request);
- To generate evidence directly by the subject device to provide new information on safety or effectiveness;
- To generate evidence to support a determination on whether the subject device meets the statutory criteria for a Clinical Laboratory Improvement Amendments (CLIA) waiver (e.g., CW and Duals);
- To generate evidence to support the interpretability of the primary clinical evidence (e.g., to demonstrate that the study population for an investigation conducted outside the United States (OUS) is representative of the U.S. population, or to provide context for an adverse event observed in the clinical study);
- To generate evidence to support a petition for reclassification of a medical device under section 513(e) or (f)(3) of the FD&C Act;
- To generate evidence for expanding the labeling of a device to include additional indications for use or to update the labeling to include new information on safety and effectiveness;
- To generate evidence for postmarket surveillance. Through ongoing surveillance, signals

are at times identified that suggest there may be a safety issue with a medical device. RWE may be generated using RWD to refine these signals for purposes of informing appropriate corrective actions and communication;
- To conduct post-approval studies that are imposed as a condition of device approval or to potentially preclude the need for 522 submissions; and
- To provide postmarket data in lieu of some premarket data, consistent with FDA's policy on balancing premarket and postmarket data collection."

RWD also have limitations, which will be addressed in more detail below. These include statistical challenges around causal inference (adequate control of confounding and avoiding other sources of bias), value not always being recognized (likely related to statistical challenges and other limitations of the data), and methods not universally understood by decision makers.

2. **Consider the specific purpose (e.g., regulatory, reimbursement and informing clinical practice guidelines) in determining the type of research study to conduct.** There are different evidentiary needs based on the stage of device development (e.g., early feasibility/first-in-human, new device for new indication vs. existing approved device for label expansion or iteration of approved devices, and surveillance of approved devices) and the perceived risk of the device. Such diverse device assessments may warrant different study designs, varying endpoints and have different use cases. For instance, for a medical device to be coverable by Medicare, evidence is needed demonstrating that the device is reasonable and necessary for the diagnosis or treatment of illness or injury or to improve the functioning of a malformed body member in patients who are representative of the affected Medicare beneficiary population. This entails evidence generation supporting improvements in clinically meaningful patient health outcomes. Alternatively, manufacturers may be interested in understanding long-term performance of a device, or researchers may focus on understanding the benefits of a marketed device compared to other devices. Study features for device evaluation are likely to differ across specific stages of the device's lifecycle. While this document does not discuss in detail design features specific to the device stage, examples are provided that highlight generation of RWE throughout the device lifecycle.

3. **Recognize that data elements in clinical trials and RWE studies may be different.** For the most part (excluding registries, which are also a form of RWD), existing RWD sources contain data that are collected for a purpose other than clinical research. Consequently, they may not match exactly the elements desired in a prospective clinical trial protocol. Instead of a purely clinical outcome, it may, for example, be necessary to substitute an outcome that is associated with a claim code (i.e., proxy variable). Such surrogate outcomes may lack precision as a representation of the outcome of interest and may be subject to bias in how they are routinely collected and coded. In such cases, depending on the regulatory purpose and the strength of evidence required, separate studies may be required to validate the surrogate data elements. Tests or other evaluation procedures may not be performed routinely in clinical practice, or may be performed according to perceived need, which can increase the risk that

these assessments are not representative of the entire sample of patients being analyzed. In contrast, clinical trials usually define the requirements for medical testing and procedures, and subsequent data collection, to ensure consistency across patients and study sites. In an effort to reduce bias and unnecessary variability, clinical trials often use central (core) laboratories and adjudication committees. Thus, while sites may report severity of a lesion or the results of a device as a continuous numerical measure in RWD, the variability in that measure is likely to be high compared to a standard operating procedure (SOP) driven and validated core laboratory. In a randomized trial design, such variability may be sufficiently mitigated by randomization to yield an interpretable result – or it may not. A hybrid design might include sending the images to a core laboratory while gaining the efficiency of collecting other variables directly from the incorporation of RWD.

4. **Consider device-specific RWD issues that may limit the ability to identify a sufficient exposure sample.**

   Most medical devices are associated with procedures, which are typically coded in EHR or health insurance claims data using standard coding systems. Even when a code correctly identifies a procedure, it is unlikely to identify the brand of device for comparative studies such as non-inferiority or superiority studies. For example, we may know from claims data that a patient had a total knee replacement, but we generally would not know the specific device that was implanted. In this case, other forms of administrative or clinical data must be used, for example, hospital chargemaster, supply chain systems, or physician notes. These are typically more resource intensive, as they require chart reviews, term searches, or natural language processing (NLP) algorithms. These are imperfect methods in that it is never known exactly the proportion of exposures that were identified. Bar code scanners, on the other hand, may directly acquire device identifiers with the potential to parse the output into an accessible RWD system such as electronic health records. This solution could provide direct digital truth that is complementary to the other RWD available, and also demonstrate a version of complementary information and interoperability across two digital systems fundamental to the construct of registry networks for RWD.

   In summary, these issues can impact the quality, size and appropriateness (i.e., do we have the right patients with the right device) of the sample that can be identified and may suggest that multiple sources and extensive data validation steps may be required to meet the required evidentiary standard, adding cost and complexity to a study.

Study Protocol

The planning of a study, whether a randomized trial or an observational study, should involve the construction of a detailed document prospectively indicating how the study will be conducted. Unfortunately, this step is not always taken. This document, the study protocol, pre-specifies fundamental features of study design that are precisely defined at an early stage, prior to study subject enrollment (for primary data collection) or prior to analysis of the data (for existing data sources). The organization of the Framework follows a typical Study Protocol, and is organized as

follows:

1. Background: Disease, available therapies, and device potential benefit/risk
2. Device description
3. Study-specific objectives
4. Target population, source for patient recruitment, and time period of interest
5. Study population and patient selection
6. Validation of key study variables
7. Outcomes: primary, secondary, exploratory, procedural, and device
8. Patient exposure to the device
9. Study design
10. Study procedures
11. Required sample size
12. Study registration
13. Interim analyses, decision rules, and oversight
14. Statistical analysis plan (SAP)
15. Study reporting
16. Future work

Because details of study design may evolve, depending on questions of feasibility that can arise during study conduct (e.g., participants might be unable to provide certain information, sample sizes may be reduced because of lack of availability of specific data elements, a primary endpoint might change, prior to data analysis, motivated by the publication of other studies), the study protocol may need to be updated after its initial completion. This should ALWAYS be accomplished through formal protocol amendment processes, which should be posted to whatever registry was used to register the initial protocol, and which is required for notification of regulatory authorities if relevant (e.g., IDE studies) and for IRB review.

References or Supporting Literature

1. Krucoff MW, Sedrakyan A, Normand SL. Bridging Unmet Medical Device Ecosystem Needs With Strategically Coordinated Registries Networks. JAMA. 2015 Oct 27;314(16):1691-2. PMID: 26302152
2. US 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. December 19, 2023. *https://www.fda.gov/regulatory-information/search-fda-guidance-documents/draft-use-real-world-evidence-support-regulatory-decision-making-medical-devices*

1. BACKGROUND: DISEASE, AVAILABLE THERAPIES, AND DEVICE POTENTIAL BENEFIT/RISK

The protocol introduction should provide background sufficient to understand the underlying disease or condition, available standard of care therapy and outcomes. The background should consider patient impact, including disease burden, the safety and effectiveness of currently available therapies, gaps in the pathophysiologic understanding of the disease or its treatments, and how the proposed device might improve outcomes or address unmet medical needs. The background should also describe the device

(including predicate devices) and associated procedures, the known and potential device effects based on the underlying anatomy, disease pathology, physiology, and duration of exposure. There should be a clear statement of the proposed benefits and risks of the device relative to those of the underlying disease and the current therapies.

Thus, the background provides the quantitative and qualitative information necessary to understand intended use, and indications for use* of the device (see Box 1A); the study objective; the rationale for the proposed study design, and the adequacy of the planned clinical and statistical analyses. Such background information is often derived from real-world sources, including health insurance claims data, EHRs, and registry data.

Overall, the goal of the background information is to demonstrate that based on the information presented, there is a justified rationale for conducting the study, that the study objective is reasonable and achievable, and that both ethical equipoise and sufficient safety oversight exist

BOX 1A:
reSET is software (an app) intended to provide cognitive behavioral therapy, as an adjunct to a contingency management system, for patients 18 years of age and older who are currently enrolled in outpatient treatment under the supervision of a clinician.[1]
reSET is indicated as a 12-week (90 days) prescription-only treatment for patients with Substance Use Disorder (SUD), who are not currently on opioid replacement therapy, who do not abuse alcohol solely, or who do not abuse opioids as their primary substance of abuse.
reSET is intended to increase abstinence from a patient's substances of abuse during treatment and increase retention in the outpatient treatment program.[2]

(as appropriate to the design) in order to proceed with an appropriately designed study.

* Intended use is defined by FDA medical product regulations. The FDA issued an amended final rule August 2, 2021. "The words intended uses … refer to the objective intent of the persons legally responsible for the labeling of an article (or their representatives)." The intended use is the label claim, what the sponsor says the product will do, as accepted by the FDA in clearing or approving that product to enter interstate commerce. The indications for use describe the reasons or situations for using the device. Not all labels include indications for use.

1.1 General Principles to Follow

A. A description of the **disease target**, its incidence, prevalence, natural history, and patient impact
B. A summary of the **currently available therapy** or therapies including:
    I. The known risks and benefits in specified patient populations (separately for intended use and indications for use within the study, if different)
    II. The critical assessment of evidence
    III. The known outcomes
    IV. The rationale for selection of comparator therapy, or the performance standard, for the investigational protocol
C. An assessment of the **underlying/unmet need** for the therapy proposed – why the device isneeded and where the device fits in, including discussions of:
    I. The pathophysiologic rationale for development of the device including identification ofgaps or insufficiencies with current therapy

II. The experience with existing cleared (e.g., predicate) or approved devices, drugs,biologics, or combination products, or other standard of care treatments
III. The anatomic rationale for development of the device
IV. The known and potential new risks that might result from use of the device
V. The known and new clinical benefits that might result from use of the device

D. A summary of the reports of prior investigations, including but not limited to a summary of the **literature, clinical experience, and investigations**, relevant to the clinical study, including discussions of:
   I. Why the proposed clinical study is needed based on the absence or limitations of existing pre-clinical or clinical data
   II. Rationale for the mechanism of device performance resulting in clinical benefit (effectiveness)
   III. Rationale for the use of RWE to address the gap in previous evidence generation
   IV. Assessment of the potential benefits and risks of the medical device
   V. Safety profile for the procedure and device (expected adverse events)
   VI. Primary clinical benefit and likelihood of demonstrating statistical certainty of clinical benefit related to device effectiveness

Box 1B illustrates in a table the application of the general principles in the introductory section of three studies.

1.2 References or Supporting Literature

1. Kawasaki S, Mills-Huffnagle S, Aydinoglo N, Maxin H, Nunes E. Patient- and Provider-Reported Experiences of a Mobile Novel Digital Therapeutic in People With Opioid Use Disorder (reSET-O): Feasibility and Acceptability Study. JMIR Form Res. 2022 Mar; 6(3): e33073.
2. Campbell ANC, Nunes EV, Miele GM, Matthews A, Polsky D, Ghitza UE, Turrigiano E, Bailey GL, VanVeldhuisen P, Chapdelaine R, Froias A, Stitzer ML, Carroll KM, Winhusen T, Clingerman S, Perez L, McClure E, Goldman B, A. Crowell AR Design and methodological considerations of an effectiveness trial of a computer-assisted intervention: An example from the NIDA Clinical Trials Network. Contemp Clin Trials. 2012 Mar; 33(2): 386–395.
3. Campbell ANC, Nunes EV, Matthews AG, Stitzer M, Miele GM, Polsky D, Turrigiano E, Walters S, McClure EA, Kyle TL, Wahle A, Van Veldhuisen P, Goldman B, Babcock D, Stabile PQ, Winhusen T, Ghitza UE.Internet-Delivered Treatment for Substance Abuse: A Multisite Randomized Controlled Trial. American Journal of Psychiatry. Published Online**:**1 Jun 2014https://doi.org/10.1176/appi.ajp.2014.13081055
4. A Remote, 9-week Insomnia Treatment Trial to Collect Real World Data for a Digital Therapeutic (DREAM) (ClinicalTrials.gov Identifier: NCT01355939).
5. Thorndike FP, Berry RB, Gerwien R, Braun S, Maricich YA. Protocol for Digital Real-world Evidence trial for Adults with insomnia treated via Mobile (DREAM): an open-label trial of a prescription digital therapeutic for treating patients with chronic insomnia. *J Comp Eff Res*. 2021

May;10(7):569-581. doi: 10.2217/cer-2021-0004. Epub 2021 Mar 8.
6. Miller et al. Impact of Powered and Tissue Specific Endoscopic Stapling Technology; Clinical and Economic Outcomes of Video-Assisted Thoracic Surgery Lobectomy Procedures: A Retrospective, Observational Study. Adv.Ther. (2018) 35:707-723.
7. Henrikson et al. Antibacterial Envelope is Associated with Low Infection Rates After Implantable Cardioverter-Defibrillator and Cardiac Resynchronization Therapy Device Replacement. JACC: Electrophysiology (2017) 3: 1158-67.
8. Wimmer et al. Effectiveness of Arterial Closure Devices for Preventing Complications with

BOX 1B: Examples of Background Information from 3 published studies

	Miller et al. (2018)[3]	Henrikson et al. (2017)[4]	Wimmer et al. (2016)[5]
A. Disease Target	Adult cancer patients undergoing video assisted thoracic surgery (VATS) lobectemy. Stapling is critical for pulmonary vessels and affects complications, recovery, and resource use	Infections in high-risk patients undergoing implantable cardioverter defibrillator (ICD) or cardiac resynchronization therapy (CRT) implantation.	Vascular complications related to arterial access are a cause of mortality, morbidity and cost in patients undergoing percutaneous intervention (PCI).
B. Current Therapy	Powered stapling may be superior to manual stapling, especially if designed specifically for thoracic procedures.	Infections are a source of substantial mortality, morbidity, and cost. Systemic antibiotics can reduce infections.	Various strategies have been used incorporating several areas of patient care: vascular closure devices, micropuncture techniques; ultrasound guided vascular access; arterial access site for PCI (transfemoral vs. transradial) and optimizing pharmacologic therapy during PCI.
C. Need	Newer power staplers are designed specifically for thoracic procedures.	TRYX antibacterial envelope is impregnated with minocycline and rifampin.	The effect of arterial closure devices (ACD) is controversial. A meta-analysis suggested increased risk of complications with ACD; randomized trial found noninferiority between ACDs and manual compression; large observational studies have generally favored ACDbut observational studies may be confounded because in practice ACD use is determined by numerous individual factors.
D. Safety and Effectiveness	Powered staplers for gastric bypass were associated with less bleeding and lower hospital costs than manual stapling.	Nonrandomized retrospective studies report TRYX 60-100% relative risk reduction for implantation infection.	The study outcome was a safety-related outcome. In an analysis using instrumental variables, ACDs were associated with a 0.40% absolute risk reduction in vascular access site complications (95% confidence interval, 0.31-0.42). They also had negative control outcome, which suggested good control of confounding in their main analysis.
E. Literature Summary	Similar studies are needed for lobectomy, especially devices designed to be tissue specific.	The purpose of this study is to assess TRYX performance in a large, prospective population.	This study uses an instrumental variable approach and hypothesizes that ACDS would not be associated with a clinically meaningful reduction in complications or hospitalization.
F. Mechanism of Benefit	Clear mechanistic explanation of device performance to clinical benefit is lacking.	The antibiotic impregnated envelope is the mechanism to reduce infection.	A mechanistic explanation of benefit is not relevant to this study. Mechanical reduction of the arteriotomy puncture with a closure device could mechanistically translate into earlier stability of the arteriotomy site, potentially reducing patient discomfort with earlier ambulation and potentially enhancing patient safety overall by reducing re-bleeding with hypertension, coughing or other sources of vascular strain. This mechanistic construct may be confounded if access is achieved in arteries with intrinsic disease that renders the closure system mechanism (e.g., sutures, glue, etc.) unstable or even deleterious to the access site.

Percutaneous Coronary Intervention: An Instrumental Variable Analysis. Circ Cardiovasc Interv. (2016): 9(4).

2. DEVICE DESCRIPTION

A detailed description of the device(s) being evaluated should be included in the protocol. The Device Description is separated from the Section on the Patient Exposure to the Device, which defines how the

device will be identified and incorporated into the study design and analysis. This section highlights relevant information to describe each important component, ingredient, or material that will be in contact with tissues or body fluids of the study subject (Boxes 2A and 2B provide examples of device descriptions). If the device is marketed already, specify the brand/manufacturer and model number of the device; if more than one generation of the device is used, specify all models. If Unique Device Identifiers (UDIs) are available, those should be included with this description. If numerous brands/manufacturers or models of devices will be included in the study without further differentiation (e.g., study of the device class), then provide a listing of all included devices within this section.

BOX 2A:

In its Section 5 System Description and Intended Use, the Clinical Investigation Plan of the study registered on CT.gov as NCT02758301 contains a description of the device (The Reveal LINQTM is a programmable device that continuously monitors a patient's ECG and other physiological parameters) and its medical indication for use (Patients with clinical syndromes or situations at increased risk of cardiac arrhythmias and patients who experience transient symptoms such as dizziness, palpitation, syncope and chest pain that may suggest a cardiac arrhythmia) followed by a detailed description of the system being investigated with a list of all its components in a referenced table (Table 6), including the model number (e.g. LNQ11, SW026), the component (e.g. Reveal LINQTM Insertable Cardiac Monitor, Incision and Insertion Tools, 2090 Programmer), the manufacturer (here Medtronic) and the Investigational vs Market-Released nature of the component (in this example the only investigational component is the RAMWare which once downloaded into the device change the device status from Market-Released to Investigational).[1]

BOX 2B:

The Device Description section (5.1) of the protocol of the study (AMPLATZER™ Amulet™ Observational Post-Market Study) registered on CT.gov as NCT02447081 represent a more literal example. It focuses on the physical description of the device, providing a figure of its shape and dimensions. A link to the Investigator's Brochure provides a reference to the IFU documentation which indicates information related to the device mode of action (Section 2 Background and Justification) or clinical benefits (Section 3 Risks and Benefits).[2]

UDIs can be used to identify many devices, in particular implantable devices. The FDA has published guidance for the creation of UDIs (21 CFR 801.20). The system is intended to make it possible to rapidly and definitively identify a device and some key attributes that affect its safe and effective use (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/unique-device-

identification-system-form-and-content-unique-device-identifier-udi). The Global Unique Device Identification Database (GUDID), a database administered by the FDA, contains key device identification information submitted to the FDA on devices that have UDIs and can serve as a reference catalog. Device identification information in the GUDID is available to the public through the AccessGUDID (https://accessgudid.nlm.nih.gov/).

2.1 General Principles to Follow

A. A description of the device sufficient for understanding should include:
    I. The device and its components, accessories, including the model and manufacturer number(s) and a unique device identifier (UDI), when available.
    II. The device mode of action and population of intended use;
    III. Unique features of the device designed to mitigate risks or enhance performance or clinical benefits;
    IV. Results of pre-clinical testing for relevant bench tests, animal studies, computational modeling, biocompatibility, potential hypersensitivity, toxicity, sterilization, and manufacturing;
    V. Sizing requirements and technical training for clinical insertion or implantation of devices. This might be accomplished by a reference to the Instructions for Use (IFU) developed for specific devices.
    VI. Characterization of the expected device performance over time;
    VII. For each component, list its status (e.g., investigational, market released)

2.2 References or Supporting Literature

1. Reveal LINQ™ Heart Failure (LINQ HF). Study NCT02758301. Clinical Investigation Plan: https://clinicaltrials.gov/ProvidedDocs/01/NCT02758301/Prot_SAP_000.pdf
2. AMPLATZERTM AmuletTM Observational Post-Market Study. NCT02447081. Study Protocol: https://clinicaltrials.gov/ProvidedDocs/81/NCT02447081/Prot_SAP_000.pdfhttps://clinicaltrials.gov/ProvidedDocs/81/NCT02447081/Prot_SAP_000.pdf

3. STUDY-SPECIFIC OBJECTIVES

The protocol of a medical device study should contain unambiguous statements of its objectives alignedwith its overall purpose, such as assessing the feasibility of the device, supporting a future premarket approval, expanding the indication of an approved device, or conducting postmarket studies for its intended stakeholders. Stakeholders include patients, patient organizations, clinicians, regulators, industry scientists, academic researchers, journals (for peer review), and payors. The objectives must be relevant, specific, based on measurable quantities, and attainable within a reasonable timeframe (Boxes 3A and 3B provide examples of how study-specific objectives are defined, based on clinical justification for risks and benefits and translated into outcomes with corresponding measurement types). The objectives are typically organized in order of decreasing importance. A study objective may be operationalized by inclusion of statistical hypotheses, although this is not obligatory. A description of the key outcomes of

interest and basis for making conclusions, however, should be included. The choice of the primary objective(s) is important and should be made explicit; secondary and exploratory objectives should be identified as such.

3.1 General Principles to Follow

A. State the overall purpose of the study and correspondingly specific objectives following the **SMART principle** (Specific, Measurable, Attainable, Relevant, and Time-Framed) organized to:
    I. Include **rationale for primary objective**, including the scientific and clinical bases used to choose the objective and to formulate specific hypotheses
        a. If there are multiple primary objectives, justify each
    II. Include **rationale for secondary objectives**
B. Specify, for devices consisting of **multiple components** (a "system"), if the system is the devicebeing assessed or if a specific component is being assessed for each objective
C. For each study objective, precisely **define the outcome measure(s)** from which clinically meaningful effects interms of risks relative to benefits can be derived, and clearly specify the type of measurement(e.g., binary, time to event)[1,2]

BOX 3A: DEFINING STUDY SPECIFIC OBJECTIVES: Comparative Effectiveness Multicenter Trial forAdhesion Characteristics of Ventral Hernia Repair Mesh.[3] This **observational study** compares the benefits, harms, and comparative effectiveness of intraperitoneal barrier-coated and non-barrier-coated ventral hernia repair (VHR) mesh in reducing adhesions, adhesion-related complications, and adhesiolysis sequelae in actual patient subpopulations and clinical circumstances. *Specific Aim 1*: To evaluate and comparethe adhesion characteristics of intraperitoneal barrier-coated versus non-barrier-coated mesh during abdominal re-exploration after prior ventral hernia repair. *Specific Aim 2*: To evaluate and compare the adhesion-related complications and adhesiolysis-related complications of intraperitoneal barrier-coated versus non-barrier-coated mesh during abdominal re-exploration after prior ventral hernia repair. These aims are "translated" intoone single primary outcome (Mesh adhesiolysis time: Mesh surface area [Time Frame: Intraoperatively (day 1)].

BOX 3B: STUDY AIMS: Transcatheter aortic valve replacement versus surgical valve replacement in intermediate-risk patients: a propensity score analysis (ClinicalTrials.gov Identifier: NCT01314313). The observational study aims to report one-year outcomes with SAPIEN 3 TAVR in intermediate-risk patients and then uses a prespecified propensity score analysis to compare these outcomes with those for similar patients given surgical aortic valve replacement in the PARTNER 2A randomized trial. The prespecified propensity analysis allows for meaningful comparisons between the two groups.[4]

3.2 References or Supporting Literature

1. Weinstein EJ, Ritchey ME, Lo Re V, 3rd. Core concepts in pharmacoepidemiology: Validation of

health outcomes of interest within real-world healthcare databases. Pharmacoepidemiol Drug Saf 2023 Jan;32(1):1-8. doi: 10.1002/pds.5537.
2. Velentgas P, Dreyer NA, Nourjah P, et al. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. AHRQ. 2013;12(13) Chapter 6-EHC099. Agency for Healthcare Research and Quality (Rockville, MD). https://www.ncbi.nlm.nih.gov/books/NBK126186/.
3. Comparative Effectiveness Multicenter Trial for Adhesion Characteristics of Ventral Hernia Repair Mesh (ClinicalTrials.gov Identifier: NCT01355939 / 2011-02112 1KM1CA156708-01 (U.S. NIH Grant/Contract).
4. Thourani, et al. Transcatheter aortic valve replacement versus surgical valve replacement in intermediate-risk patients: a propensity score analysis. Lancet 2016; 387: 2218-2225.

4. TARGET POPULATION, SOURCE FOR PATIENT RECRUITMENT, AND TIME PERIOD OF INTEREST

A description of the population to which the results of the study will apply (i.e., target population) should be provided. This population should be well depicted through the use of inclusion and exclusion criteria in identifying appropriate study patients. In principle, the research "participants" (whether they are actively enrolled in a study or contained in an existing data source) should closely reflect the population of intended use, the population of indicated use, or sub-group(s) of interest in certain postmarket studies.

Additionally, the source of patient recruitment should be described, and if appropriate, the experience of the physicians or device operators. Other factors that should be considered are the setting in which the device is indicated in routine practice, and the previous and current treatments of the patients being enrolled. For example, if the target population is adults with peripheral vascular disease (PVD) and a national disease registry exists for PVD, patients could be recruited and randomized from patients enrolled in the registry. Alternatively, a local registry comprised of health record data for adults with PVD collected during routine clinical care could serve as the basis for an observational study or as a source of recruitment for a clinical trial. Using RWD sources for patient identification implies that the RWD has relevant fields to the study design's inclusion and exclusion criteria.

In addition, the time period of interest should be specified, to support rationalizing the selection of the data source(s) and design decisions (such as choice of comparator) described later in the protocol.

4.1 General Principles to Follow

A. Factors to consider and specify in describing the **target population** should include:
    I. Disease state or health condition under study (e.g., previously untreated, failed prior treatment, measurable disease)
        a. Descriptors might include stage and severity of the condition, duration of the condition, existence (or exclusion of) specific comorbidities (e.g., diabetes), age of the population (e.g., adult vs. pediatric, adults restricted to certain age ranges), or geographic

region, etc. There may also be important anatomic descriptors defining patients in whom a specific device is indicated, such as calcification, tortuosity, etc.

II. In some situations, the target population will be defined by having had (or being about to have) a particular procedure (e.g., implantation of a total knee replacement), regardless of the specific device implanted.
    a. Specify **Source of Patient Recruitment or Patient Data**
III. Describe clinical centers that will be enrolling participants (for prospective, primary data collection) or the clinical setting(s) for patient encounters or other data capture (retrospective or prospective use of existing data source)
IV. Describe principal investigator proficiencies and experience required to participate in the study. Describe overall site characteristics (patient/procedural volume, technology on site, research staff) required to participate in the study.
V. For a complex device with an operator learning curve the study may need to be limited to certain sites (e.g., high-volume centers with highly experienced operators who are specialized and trained., In this setting it may be noted in the protocol that generalizability of results to smaller centers or those with less experienced operators in real world practice may be limited, and may create the need for further studies with a broader spectrum of enrolling sites and operator experience. Similar considerations and even high-level plans for subsequent data collection in a broader set of centers may also be important if early site selection lacks diversity including underserved communities or racial disparities.
VI. Provide a high-level description of steps taken to assess data quality as described in the NESTcc Data Quality Framework in terms of the ability to identify the target population in a relevant and reliable (reproducible) manner.
VII. Data reuse: It describes secondary use of existing data for a new research purpose. If mechanisms are planned to share data for analyses beyond the primary study SAP, these intentions and mechanisms should be included in the study protocol's SAP overview. Mechanisms of specific interest may include issues of data governance or oversight to ensure that data reuse, reanalysis or publications represent good science—particularly with regard to bias in retrospective analyses-- and thus produce responsible, ethical health information.

    Unconscious (or conscious) biases may influence methodological choices in subtle ways that could yield "hoped for" results. Wang and colleagues[3] don't fully share the skepticism around reuse of data, particularly in the context of assessing safety. They present numerous examples in the pharmaceutical setting of appropriate reuse, particularly in the context of safety assessment.

VIII. When using unstructured data (e.g., imaging, physician notes), one should consider the reliability and quality of unstructured EHR data and the methods utilized (NLP, machine learning) and the appropriateness of using it.

BOX 4: EXAMPLE OF TARGET POPULATION

This study used hospital billing records contained in the Premier Hospital Database (PHD). [1,2] The PHD contains complete clinical coding, hospital cost, and patient billing data from more than 700 hospitals throughout the United States. Although the database excludes federally funded hospitals, the hospitals included are nationally representative based on bed size, geographic region, location (urban/rural) and teaching hospital status (Premier Applied Sciences. 2020). The database contains a date-stamped log of all billed items by cost-accounting department including medications; laboratory, diagnostic, and therapeutic services; and primary and secondary diagnoses for each patient's hospitalization. The database also provides demographic, payer, and device information.

***Population**: The study setting was inpatient admissions for video-assisted thoracoscopic surgery (VATS) lobectomy identified within the Premier database. The study population comprised of patients ≥18 years of age undergoing elective VATS lobectomy during a hospital admission between January 1, 2012, and September 30, 2016, for whom the endoscopic surgical stapler used during the index hospitalization could be identified from hospital administrative records as either powered or manual and with respect to manufacturer (Ethicon/Johnson & Johnson; Medtronic/Covidien).*

Subject Selection: Inclusion Criteria:

Underwent VATS lobectomy during a hospital admission between January 1, 2011, and September 30, 2016

The first observed hospital admission for VATS lobectomy during this period was designated as the index hospital admission

Aged ≥ 18 years or older at the time of the index admission

Subject Selection: Exclusion Criteria:

Had missing data on hospital supply, room and board, operating room, or total hospital costs

Were transferred from another institution

Had a non-elective VATS lobectomy

A stapler used during the index hospitalization could not be identified as either powered or manual and with respect to manufacturer (Ethicon/Johnson & Johnson; Medtronic/Covidien)

Both powered and manual staplers were used during the index hospitalization (these patients were excluded from the study because of inability to assign them to one of the study groups: powered stapler group or manual stapler group)

da Vinci EndoWrist surgical staplers were used during the index hospitalization

B. Time Period of Interest

I. As has been strongly highlighted by the pandemic's impact on active research protocols, time delays in enrollment or follow up out-of-window may provide multiple sources of subtle bias, and hence undermine data quality. Statistical methods should be carefully reconsidered in the event of

significant time delays.

II. In some fast-moving device sectors, time delays in enrollment may also span an iteration or change in the medical device platform under study – either in the commercially available control arm or in the version of the test article used early vs later in the trial. Landmark analyses or other structured approaches may be considered to understand the potential impact of such device version evolution during the course of a single study.

4.2. References or Supporting Literature

1. Miller DL, Roy S, Kassis ES, Yadalam S, Ramisetti S, Johnston SS. Impact of Powered and Tissue-Specific Endoscopic Stapling Technology on Clinical and Economic Outcomes of Video-Assisted Thoracic Surgery Lobectomy Procedures: A Retrospective, Observational Study. Adv Ther 2018;35:707-23.
2. Premier Applied Sciences. Premier Healthcare Database: Data that Informs and Performs. March 2, 2020. https://products.premierinc.com/downloads/PremierHealthcareDatabaseWhitepaper.pdf
3. Wang SV, Kulldorff M, Glynn RJ, Gagne JJ, Pottegård A, Rothman KJ, Schneeweiss S, Walker AM. Reuse of data sources to evaluate drug safety signals: When is it appropriate? Pharmacoepidemiol Drug Saf 2018 Jun;27(6):567-569. doi: 10.1002/pds.4442. Epub 2018 Apr 27.

5.STUDY POPULATION AND PATIENT SELECTION

The research cohort (whether they are actively enrolled in a study or are extracted from an existing data source) should be well characterized and supported by specifically stipulated inclusion and exclusion criteria (Box 5). For regulatory science or submissions, the degree to which the study cohort closely reflects the population of intended use (e.g., “on label” device use), a specific sub-population, off-label use, or a mixture of these should be clearly depicted.

5.1 General Principles to Follow

Factors to consider and specify in describing the study population should include:

A. **Disease state** under study should adopt all possible measures to minimize data collection bias (e.g., standardized structured data capture, with harmonized definitions) to minimize misclassification of inclusion and exclusion criteria, and to reduce missing data. Understanding the disease state in the context of the overall patient is essential for device studies, especially for class III devices, to ascertain potential sources of device-related risk. Key descriptors might include:
   I. Stage and severity of the condition at time of the index (study) procedure
   II. Duration of the condition at time of the index (study) procedure
   III. Critical anatomic features of relevance to device use (for instance calcification or tortuosity of blood vessels, chamber size, bone density…)

IV. Relevant information related to previous relevant device exposures (previous hip implant; previous valve surgery, etc.)
V. The study device (which might be a class of devices, e.g., replacement hips, or might be a specific device, e.g., a specific manufacturer and model of hip) may sometimes be used to define the population (e.g., women who have a specific brand and type of breast implant).
VI. Baseline descriptors (age, sex/gender, race, blood pressure (BP), heart rate (HR), laboratories, body surface area (BSA), etc.)
VII. Co-morbidities (previous relevant device procedures, diabetes, hypertension, renal failure) not excluded by the protocol
VIII. Concomitant medications at the time of the index (study) procedure

B. Though sometimes necessary, careful consideration should be applied to the consistency of inclusion/exclusion criteria that primarily rely on research team judgements about the patient. Frequently encountered examples in device trial designs include:
   I. Inability to provide informed consent
   II. Likelihood of non-compliance with protocol medications, follow up timelines, etc.
   III. Presence of a limited life-span (12 months or less; 6 months or less; etc.) due to other health related issues
C. Denote the time frame and modality of **assessment where relevant** for each inclusion and exclusion criterion, e.g.,:
   I. Is the criterion assessed by the patient history (e.g., smoking history), by an imaging modality (abnormal ECG or echocardiogram), or by a laboratory value (creatinine/glomerular filtration rate (GFR))
   II. Is a lab test, diagnostic test, or imaging study required within 1 year, 1 month, 1 week, etc. of the index procedure

D. If criteria are met by data extraction from an existing data repository (registry, claims, etc.), each criterion's operational definition should include:
   I. The structured codes and semantic structure of text used to identify occurrence
   II. The algorithm by which the identification will be applied
   III. If natural language processing (NLP) or other machine learning (ML) is used to define a criterion, then a full description of the NLP or ML development should be provided

BOX 5: EXAMPLE OF STUDY DESIGN AND PARTICIPANTS

Transcatheter aortic valve replacement versus surgical valve replacement in intermediate-risk patients: a propensity score analysis (ClinicalTrials.gov Identifier: NCT01314313).[1] In this analysis the authors used populations from the PARTNER 2 SAPIEN 3 intermediate risk observational study[2] and the PARTNER 2A randomised trial (NCT01314313).[3] These two prospective multicentre studies enrolled patients with symptomatic severe aortic stenosis who were considered to be at intermediate risk for 30 day surgical mortality. Risk status was evaluated by a Heart Team that included cardiac surgeons. Patients were deemed intermediate risk via clinical assessment or if their Society of Thoracic Surgeons (STS) score was 4% or higher. In those with an STS score lower than 4%, the Heart Team deemed the patient intermediate risk if they had risk factors not present within the predictive score (e.g., liver disease, frailty, and pulmonary hypertension).

In PARTNER 2A, patients were randomly assigned to receive either surgical valve replacement or TAVR using SAPIEN XT; here the patients assigned to surgery[3] were included in propensity score analysis. In the SAPIEN 3 study, all TAVR patients who were eligible to receive a valve had mandated multidetector computed tomography (MDCT) analyzed by the study core laboratory and were presented on a conference call in which a screening committee reviewed imaging and clinical data and approved patients prior to enrolment.

Inclusion and exclusion criteria for the SAPIEN 3[2] and PARTNER 2A[3] studies were the same. Key exclusion criteria were a congenitally bicuspid aortic valve, severe aortic regurgitation, left ventricular ejection fraction lower than 20%, severe renal insufficiency, and estimated life expectancy of less than 2 years. Patients with noncomplex coronary disease requiring

5.2 References or Supporting Literature

1. Thourani, et al. Transcatheter aortic valve replacement versus surgical valve replacement in intermediate-risk patients: a propensity score analysis. Lancet 2016; 387: 2218-2225.
2. Kodali S, Thourani VH, White J, et al. Early clinical and echocardiographic outcomes after SAPIEN 3 transcatheter aortic valve replacement in inoperable, high-risk and intermediate-risk patients with aortic stenosis. Eur Heart J 2016;37:2252-62.
3. Leon MB, Smith CR, Mack MJ, et al. Transcatheter or Surgical Aortic-Valve Replacement in Intermediate-Risk Patients. N Engl J Med 2016;374:1609-20.

6. VALIDATION OF KEY STUDY VARIABLES

When using existing data such as administrative claims and electronic health records, assuring the validity of operational definitions used to measure key study variables – device(s) of interest, key subgroup characteristics, key confounders, primary endpoints – is essential to the relevance, reliability, and interpretability of the study results. “Measurement validity” refers to the extent to which a measure accurately represents the intended underlying construct. Measurement validity cannot be assured without evidence of data reliability and relevance within the specific patient

population and setting under study. When RWD sources are mined using data extraction programming, the validity of the programming output is a critical dimension of validation. Prior evidence of validity may be relied upon if a robust argument about the transportability of such evidence to the current research context can be made. The approaches to generating evidence of validity are described below.

Key study variables may be defined by demographics or other stand-alone structured data fields, device use, procedures, diagnoses, medications, or some combinations of these variables. For a given variable, a code list or algorithm can be developed based on diagnoses, procedures, medications, diagnostic tests or their results, patient-reported symptoms or diagnoses, or some combinations of these variables, and constructed through literature search and review of the existing code list or algorithm, knowledge of clinical workflow, consultation with clinicians with experience in diagnosing and treating the target disease, and consultation with coding and database experts. Attention should be paid to code description and its specificity, code position (primary [principal] or secondary), setting of care (inpatient or outpatient), and timing (admission or discharge). In the U.S., structured coding systems such as the International Classification of Diseases (ICD)[2] and Current Procedural Terminology (CPT)[3] are widely accepted medical nomenclature used to classify diseases and report medical, surgical, and diagnostic procedures and services for processing claims reimbursement. The transition of ICD-9-CM diagnosis and procedure codes to ICD-10-CM/PCS took effect on October 1, 2015. Drugs are identified and reported by the National Drug Code (NDC), a unique, 3-segment, 11-digit number. These structured coding systems are also commonly used for conducting research using administrative claims and electronic health record data.[4]

Coding for devices, procedures, and medications may be considered sufficiently reliable based on face validity when the underlying data entry is scrutinized for errors for processing claims reimbursement. The coding accuracy and completeness for diagnoses may be more uncertain since specific diagnosis codes may not be available all the time or may be changed over time. Also, diagnoses determined in the outpatient setting may be less scrutinized for errors than those determined in the inpatient hospital setting, and diagnosis codes assigned for the purposes of reimbursement may not reflect incidence (or event occurrence) of a condition.

Based on the intended use of algorithms, either internal or external validation can be used for confirmation. Internal validation evaluates the ability of the algorithm to accurately identify the diagnosis reported by treating clinicians (i.e., treating clinicians' judgment).[5] External validation evaluates the ability of the algorithm to accurately identify the true disease or condition[5] determined by rigorous reference standards such as adjudication by clinicians, disease registries, or clinician response to a questionnaire confirming the diagnosis.[6] The choice of the reference standard depends on the type of data elements for validation and the availability of data.

6.1 General Principles to Follow

Factors to consider and specify in data element validation should include:

A. External validation of **key study variable operational definitions** consists of determining cases (and, possibly non-cases) based on an algorithm to a reference standard within a large enough

sample to provide adequate precision of the estimate and then comparing performance of the algorithm to the reference standard via statistical measures of positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity

I. Conduct validation under a pre-specified study plan, including development of structured abstraction forms, with input from clinicians and data source subject matter experts
II. Train data abstractors and assess agreement between abstractors to ensure quality of data abstraction
III. Pre-specify and justify the optimal threshold for algorithm performance – when performance is below the prespecified threshold, the algorithm should be updated accordingly
    a. There are no universally accepted optimal cut-offs; statistical methods such as Youden's index can be used in some cases to determine optimal cut-off values from receiver-operating characteristic (ROC) curve analysis
    b. Conventionally, researchers have targeted PPV values of at least 80%, however a target of 95% PPV or higher may be appropriate if small amounts of misclassification would substantially bias study results

B. The **intended use of the algorithm** should be made clear so that the most relevant performance measure is selected
    I. If the goal is to estimate incidence, sensitivity should be prioritized
    II. If the goal is to compare outcomes that are expected to occur in only a small proportion of subjects, the PPV of the outcome is generally prioritized
C. When **variations of an algorithm** are equally plausible and the most reliable operational definition is unclear, performance should be considered for each, and the sensitivity evaluated against the impact to PPV or specificity
    I. It may be difficult to ascertain the exact date for measures that require a specific start date (e.g., incident outcome, start of device exposure) – studying severe or acute events associated with the condition reduces potential for misclassification because they are more likely to be captured within the data source
    II. It may be unclear whether to include more detail within an algorithm (e.g., using admission codes with and without laboratory measurements, unstructured data), especially if a larger data source contains only structured data (or limited laboratory and imaging data) while a smaller data source could include clinician notes (or labs and imaging) within the operational definition – assessing the validity of the simpler algorithm can inform the utility of the simpler approach
D. Potential reference standards include medical records (including imaging and laboratory results), disease registries, disease surveys (completed by clinicians or patients), a panel of clinical experts who have experience in diagnosing and treating the target disease, prescription-dispensing records, and supply chain records (for device use). The choice of the reference standard depends on the type of data elements for validation and the availability of data. These reference standards can, in many situations, be useful for population definition, outcomes, and covariates.
    I. The clinician chart review is considered a "gold standard" for evaluating

algorithms for identifying device use, health outcomes, populations, and covariates. However, medical charts are not always accessible to researchers and manual chart review can be time-consuming and labor-intensive.

II. Traditional validation studies (see Box 6B), particularly those including manual chart review, can be both time consuming and expensive. As we've noted, there are also questions about transportability, e.g., can an algorithm that has been developed in a health insurance claims database and successfully validated be applied equally confidently to an EHR system or even another claims database. We've also seen that some aspects of validity, notably sensitivity, can be difficult or impossible to assess in an affordable manner. To address these challenges, Swerdel and colleagues,[7] working in the context of the Observational Health Data Sciences and Informatics (OHDSI) collaboration, have developed a tool that can be used quickly and relatively inexpensively, in any database, without relying on external sources of information (see Box 6C).[7]

III. When the reference standard is imperfect for classifying the health outcome of interest, researchers should consider statistical correction methods,[8] or utilization of the Delphi technique among a group of clinical experts.

BOX 6A: EXAMPLE OF CONDUCTING VALIDATION OF A HEALTH OUTCOME OF INTEREST

A recent validation study examined the PPV of three ICD-9-based and three ICD-10-based coding algorithms to identify prosthetic joint infection (PJI) following total knee arthroplasty (TKA) within the U.S. Veterans Health Administration (VHA) using a clinician chart review approach.[9] A random sample of 80 potential PJI cases was identified using each algorithm within VHA data and stratified by annual TKA procedure volume at each site. A sample size of 80 patients for each algorithm was estimated to be sufficient for examining the PPV with a narrow 95% confidence interval (CI) width of ±10%, assuming a PPV of 80.0%. Medical records of those sampled patients were reviewed by two infectious diseases clinicians independently to adjudicate PJI events. Two algorithms comprising an ICD-9 or ICD-10 PJI hospital discharge diagnosis following a TKA code in combination with a current procedural terminology (CPT) code for knee X-ray and a CPT code for a relevant surgical procedure or microbiological culture had a PPV of 75.0% (95% CI: 64.1%–84.0%] for the ICD-9 PJI algorithm and 85.0% (95% CI: 75.3%–92.0%) for the ICD-10 PJI algorithm.

E. Sampling methods should ensure that the validation study sample represents the target population, including geographic regions, setting of care, or sub-diagnoses, with the similar prevalence of the health outcome of interest.

I. If the goal is to compare outcomes that are expected to occur in only a small proportion of subjects, the prevalence should be similar within the study population and reference standard, since PPV is affected by prevalence.

II. Sampling should not be affected by the study question, e.g., do not sample outcome stratified by exposure status.

BOX 6B: PHEVALUATOR PROCESS

a. The process involves creating a cohort of subjects who are very highly likely to have a health condition of interest. They refer to this as "noisy" positives, with "noisy" used here to mean that, although they are very highly likely to have the condition, this is not a true gold standard. They call this the xSpec cohort, meaning extremely specific. This cohort then defines the outcome cohort to be used with a patient-level prediction model. A typical approach to developing the xSpec cohort would be to include subjects who have multiple codes in the database for the condition of interest. This might be 5 or more codes for acute conditions, e.g., myocardial infarction, or it might be 10 or more codes for chronic conditions.

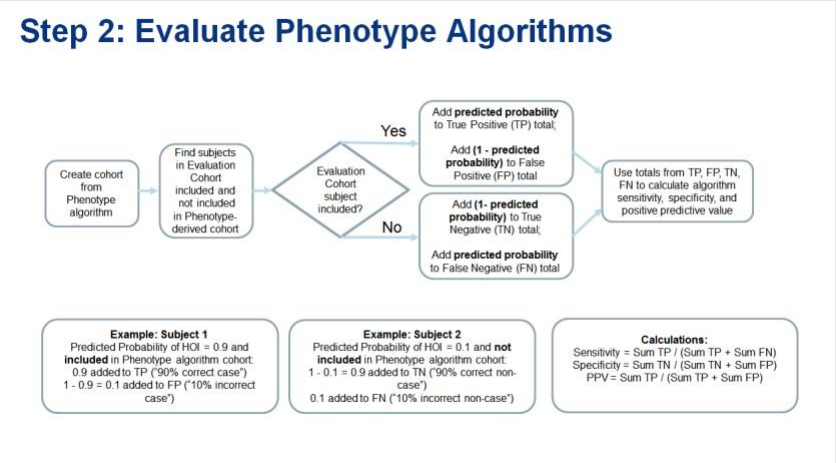
b. The process also defines a corresponding noisy negatives cohort, which is created by taking a random sample of subjects in the database who have no evidence of the condition of interest. To provide some assurance of the absence of the condition, they use a very sensitive definition, e.g., 1 or more codes for that condition, then exclude any subjects who enter that sensitive cohort from the noisy negatives.

c. The xSpec and noisy negative cohorts are then combined to form the target cohort for the patient-level prediction modeling. They use a LASSO regularized regression approach to patient-level prediction, but other modeling approaches could be used instead. The LASSO approach starts with all the data in the subject's record. The results of the model, i.e., after winnowing down the predictors, are a set of coefficients for the included characteristics that are used to discriminate between those with and without the condition of interest.

d. The next step is to create an evaluation cohort, i.e., a large group of randomly-selected subjects, typically up to 100,000, that is used to evaluate the phenotype algorithms. The extracted data include the same covariates as those used in the target cohort during the predictive model creation step. They then apply the model to the evaluation cohort, producing predicted probabilities for the condition of interest. These predicted probabilities are saved for use in the next step.

e. The next step is to conduct a formal evaluation of the phenotype algorithms. Every subject in the evaluation cohort (just described) should be eligible to be included in the cohort developed from the phenotype algorithm. The algorithm is applied to identify those subjects who are positive, according to the algorithm. The figure shows how the predicted probabilities for subjects are used, based on whether or not the subjects in the evaluation cohort are also identified as positive by the phenotype algorithm.

BOX 6C: PHEVALUATOR PROCESS



f. Essentially, if the algorithm included a subject from the evaluation cohort, i.e., the algorithm considered the subject a "positive", the predicted probability for the phenotype gets added to the True Positives value and one minus the predicted probability for the phenotype gets added to the False Positives value. If the algorithm did not include a subject from the evaluation cohort, i.e., the algorithm considered the subject a "negative", one minus the predicted probability for the phenotype for that subject is added to the True Negatives value and the predicted probability for the phenotype is added to the False Negatives value.

g. The sensitivity, specificity, and positive and negative predictive values can then be estimated from the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These statistics are generated using the predicted probabilities. Examples of the calculations are shown in the diagram.

h. Swerdel and colleagues[7] compared their findings for acute myocardial infarction with published studies that estimated PPVs for acute myocardial infarction. The PheValuator gave lower estimates of performance of the same algorithms evaluated in the publications. However, an advantage to using PheValuator is that multiple algorithms may be tested on each database to determine relative advantages and disadvantages of each algorithm. In contrast, using validation results from published algorithms is limited to the specific algorithm tested. If changes to the algorithm are needed, the published results can no longer be directly applied. For example, if a proposed study requires a limitation on patient history, such as no prior statin use, the results from earlier validation studies that did not apply that limitation could be very different from the performance characteristics of the algorithm to be used in the proposed study. PheValuator also allows examination of the impact of added algorithm elements on performance. For example, Swerdel and colleagues[7] found that including a diagnosis code from a hospital in-patient visit improved the PPV for acute myocardial infarction with only a small impact on sensitivity, while the same change in algorithm for atrial fibrillation produced only a moderate gain in PPV with a large impact on sensitivity.

6.2 References or Supporting Literature

1. 510(k) premarket notification: https://www.accessdata.fda.gov/cdrh_docs/pdf17/K173860.pdf
2. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005;40:1620-39.
3. Dotson P. CPT(®) Codes: What Are They, Why Are They Necessary, and How Are They Developed? Adv Wound Care (New Rochelle) 2013;2:583-7.
4. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016;375:2293-7.
5. Nicholson A, Tate AR, Koeling R, Cassell JA. What does validation of cases in electronic record databases mean? The potential contribution of free text. Pharmacoepidemiol Drug Saf 2011;20:321-4.
6. Weinstein EJ, Ritchey ME, Lo Re V, 3rd. Core concepts in pharmacoepidemiology: Validation of health outcomes of interest within real-world healthcare databases. Pharmacoepidemiol Drug Saf 2023 Jan;32(1):1-8. doi: 10.1002/pds.5537.
7. Swerdel JN, Hripcsak G, Ryan PB, PheValuator: Development and Evaluation of a Phenotype Algorithm Evaluator. J Biomed Inform 2019 September ; 97: 103258. doi:10.1016/j.jbi.2019.103258.
8. Umemneku Chikere, Chinyereugo M, Kevin J Wilson, A Joy Allen, and Luke Vale. 'Comparative Diagnostic Accuracy Studies with an Imperfect Reference Standard – a Comparison of Correction Methods'. BMC Medical Research Methodology 21, no. 1 (2021): 1–12. https://doi.org/10.1186/s12874-021-01255-4
9. Weinstein EJ, Stephens-Shields A, Loabile B, et al. Development and validation of case-finding algorithms to identify prosthetic joint infections after total knee arthroplasty in Veterans Health Administration data. Pharmacoepidemiol Drug Saf 2021;30:1184-91.

7. OUTCOMES: PRIMARY, SECONDARY, EXPLORATORY, PROCEDURAL, AND DEVICE

The primary outcome (note that the terms "outcome" and "endpoint" are used interchangeably throughout) is the primary study protocol objective. In some settings the primary outcome may be a composite endpoint, combining multiple specific outcomes into a single variable. Sometimes, two or more primary outcomes may be of interest. For instance, for joint replacement, the primary outcome may be both time to revision and one-year pain assessed by a questionnaire.

Secondary outcomes provide additional information that may relate to clinically important subgroups or that may reflect additional clinical considerations as measures of benefits or risks of the device. If the primary outcome in an oncology study is overall survival (OS), the secondary outcome may be progression-free survival (PFS). Exploratory outcomes are to explore endpoints useful for generating new hypotheses for future confirmatory research efforts.

Procedural data are taken from the index procedure using or implanting the device. Procedural data of interest are features that are likely to affect meaningful aspects of device performance, safety, or clinical outcome. Specific data elements characterizing the procedure vary greatly depending on the device and disease state.

Device-related outcomes include 1) specific measures of device performance (does the device do what it is mechanistically designed to do) and 2) assessments of the balance of risks/benefits related to the device use, its clinical effectiveness and safety. High risk devices such as permanently implantable devices will typically be expected to provide greater clinical benefits over alternative care options that have less risk.

Device performance, effectiveness and safety measures all may be multidimensional in that performance may relate to biomaterials, design features, manufacturing tolerances, operator proficiency, patient selection criteria, disease characteristics, anatomic variations, lesion variations, or adjunctive therapies. When appropriate, involving patients in identifying outcome measures that are directly relevant to their experience of the condition (patient preference, patient reported outcomes metrics) should be considered.

In observational studies, the inclusion of a (negative) control outcome, defined as an outcome unaffected by exposure to the device of interest, can strengthen the study design. This approach is also known as testing falsification hypotheses. While such outcomes in observational studies cannot unequivocally prove the absence of bias between treatment arms, they can test a putative mechanism of bias (Box 7A, 7B, and 7C). Justification for the choice of control outcomes should be supplied. The association between the device and the control outcome should adopt the same analytical procedure used to assess the association between the device and the study's primary outcome.

Sometimes the endpoint measured in clinical trials (e.g., pain) might not be captured in existing data sources, such as claims data. In such situations, an alternative, available endpoint reflecting similar benefit relevant to patients (e.g., reoperation) may be identified and justified in order to realize other advantages (e.g., larger cohorts efficiently accessed) of using the data source.

7.1 General Principles to Follow

A. Primary, Secondary, and Exploratory Outcomes:

I. Provide clear definitions of primary, secondary, and exploratory outcomes and method(s) of outcomes assessment

a. Provide criteria for precise definition (including standard codes such as ICD-10-CM) and objective classification of the outcome that is sufficient for statistical analysis to address the research question (i.e., define the endpoint); include the type of assessments made, the timing of those assessments, the tools used, and indicate how multiple assessments within an individual are to be combined

b. Denote how relevance and reliability of the endpoint was determined; if using existing data, specify validation parameters for target population in data source[1]

c. If endpoint adjudication is required, describe process for classifying potential cases, including case definitions, the number of adjudicators, and resolution of conflicting decisions, as well as the level of adjudicator independence (from each other and from the study sponsor) and their qualifications

d. Independent core laboratories with established SOPs may enhance consistency and freedom from bias in some device studies. Core laboratories should be considered when imaging, unique laboratory or other diagnostic surrogates reflect important mechanistic or clinical outcomes related to the device being

studied
    e. Characterize the misclassification rate associated with the endpoints; describe approaches to reducing misclassification and characterizing the impact of any outcome misclassification
    f. Describe measures adopted to minimize data collection biases (e.g., standardized structured data capture, with harmonized definitions) and measures taken to reduce missing data
II. Specify the type of variable for each endpoint (e.g., binary, failure time, categorical)
III. Describe the rationale for using composite or surrogate outcomes, and considerations for interpretation of results
IV. Specify and justify timepoints of endpoint data collection, including any windows of measurement time
V. Describe what outcomes, if any, were discussed or prioritized with input from patients
VI. Consider the role of patient reported outcomes in the context of the study objectives

B. **Procedural Outcomes:**
I. List **specific procedural outcomes** (including standard codes such as CPT, if applicable); these may include procedure time, physiological and biological data captured as part of the procedure, and other procedure-specific data
    a. Capture procedural details (approach, length, etc.), success (was intended device successfully implanted), and complications (related to access, approach or acute device malfunction)
II. Describe if the data are standardized (e.g., are the data routinely available in a similar format across systems) [Refer to Data Quality Framework]
III. Characterize the expected completeness of data capture

C. **Device Outcomes:**
I. Device performance should be reported commensurate with the manner of exposure of the patient to the device. For example, a single exposure to an atherectomy device, a dosing series of therapeutic radiation exposures, and a permanently implantable device should be characterized over time relevant to the patient exposure.
II. For implantable devices, aspects of device performance may change over time; thus, clearly identify which features of the device will be measured.
    a. Initial ability of the device to perform as intended may be eroded over time, through wear and tear, materials failures, battery depletion, infection, or temporal changes in the implant site.
    b. Indicate if and how both short- and long-term device outcomes are collected
III. Adverse events are often reported by investigators as to whether or not they are related to the device being studied. Independent adjudication of "device related" by experts based on the documentation available is highly

recommended to avoid bias in device studies. However, such evaluations of relatedness often remain subjective.

D. Patient-Reported Outcome Measures (PROMs):

I. Provide details on what instruments are used, e.g., Kansas City Cardiomyopathy Questionnaire (KCCQ), Short-Form Survey questionnaires (e.g., SF-12, SF-36), EQ-5D, Oswestry Disability Index (ODI), including details of whether it is validated in the disease setting of interest (and if so, year validated).

II. Provide details on how the PROMs are captured, e.g., visual analogue scale (VAS), picture scale, Likert scale, and whether it is self-administered, interviewer-administered, etc.

III. If success is based on a PROM, clearly provide the definition of "clinically meaningful" deltas, e.g., KCCQ improvement by ≥10 points.

E. **Control Outcomes** in observational studies (falsification hypotheses):

I. Describe why the outcome is highly **unlikely** to be causally related to the device or comparator

II. Generally, when using negative controls, it is important to demonstrate that the confounders of the association between the device and the control endpoint are the same as the confounders of the association between the device and the primary study endpoint. Another approach that has been used recently is use of a large (30-40) number of negative controls and examine the distribution of those negative control results[2]

F. **Types of complex outcomes:**

I. **Recurrent events** record events that can occur multiple times over the study period, and which the investigator would like to take into account (e.g., (re-)hospitalization for heart-failure)

a. If recurrent events are planned to be incorporated into the analysis, how will investigators capture all events?

b. How will the treatment effect be reported, e.g., incidence, regression coefficient?

c. Are the recurrent events expected to be correlated?

II. **Competing risks** are when multiple causes of the same event can preclude observation of the cause-specific event of interest (e.g., when the outcome is death due to myocardial infarction, death due to any other cause would be a competing risk).

a. How will the treatment effect be reported, e.g., cumulative incidence, regression coefficient?

b. If the event of interest is non-terminal (e.g., readmission), but the patients might be truncated by a terminal event, then semi-competing risks might be applicable[3]

III. **Repeated measurements** are common (e.g., pain score recorded at baseline, 1, 3, 6, 12, 18-months post-surgery for device implantation), but additional

consideration needs to be given when all outcome measurements are used as opposed to a single endpoint (e.g., change at 12-months from baseline).

   a. Clarify if observation times were pre-specified or not, e.g., was left ventricular ejection recorded at unplanned visits?
   b. Will all measurements be used for outcome analysis (e.g., using a linear mixed model), or just one follow-up time point (e.g., analysis of covariance)?

IV. Composite outcomes include multiple components and are frequently used as primary endpoints to increase event rates, thus improving trial efficiency, as long as the different events could plausibly be affected by treatment.

   a. Secondary analysis of individual components should be performed to assess whether the treatment effect is in the same direction for each endpoint.
   b. How will the composite be defined: time to first event, occurrence of any event, a weighted approach, a prioritized endpoint (e.g., Finkelstein-Schoenfeld approach)?

Single component outcomes that are not impacted by external events provide the simplest approach for the conduct of a study. However, in many cases outcomes are more complex, more nuanced, and require additional consideration. More complex outcomes can require more sophisticated analytic approaches. Moreover, these outcomes can be specified as primary, secondary, or exploratory outcomes.

BOX 7A: CONTROL OUTCOME

To assess the effectiveness of arterial closure devices (ACD) for preventing complications with percutaneous coronary intervention (PCI),[4] undertook a retrospective analysis using the CathPCI Registry from 2009-2013 at 1,470 sites across the U.S. The primary outcome was defined as vascular access site complications in patients undergoing transfemoral PCI. The control outcome was non- access site bleeding. It was found that the use of ACDs was associated with a modest absolute risk reduction in vascular access site complications. Absolute differences in non- access site bleeding were negligible, suggesting acceptable statistical control of confounding in the comparison with regard to the study primary endpoint.

BOX 7B: CONTROL OUTCOME

To evaluate a strategy of active surveillance of a national cardiovascular registry for assessment of the postmarketing safety of an implantable vascular-closure device,[6] conducted a prospective, propensity-matched analysis of the safety of the Mynx vascular closure device in comparison with alternative approved vascular closure devices using data from the CathPCI Registry of the National Cardiovascular Data Registry. The primary outcome was any vascular complication, which was a composite endpoint that comprised access-site bleeding requiring treatment, access-site hematoma requiring treatment, retroperitoneal bleeding, or any vascular complication requiring intervention. Secondary safety endpoints included access-site bleeding requiring treatment (a component of the primary outcome) and postprocedural blood transfusion. The risk of the primary and secondary outcomes as well as each component of the primary outcome was all significantly higher in the Mynx device group than that in the alternative vascular closure device group. The outcome of postprocedural contrast-induced nephropathy, which was not expected to differ among various vascular-closure devices, was included as a negative control for a post-hoc analysis to assess the robustness of the primary findings. The risk of contrast-induced nephropathy was slightly but non-significantly higher in the Mynx device group. The authors stated that this indicates the possibility of a small amount of residual risk imbalance between the two study groups.

For additional information on data capture for outcomes, please refer to the Data Quality Framework.

7.2 References or Supporting Literature

1. Weinstein EJ, Stephens-Shields A, Loabile B, et al. Development and validation of case-finding algorithms to identify prosthetic joint infections after total knee arthroplasty in Veterans Health Administration data. Pharmacoepidemiol Drug Saf 2021;30:1184-91.
2. Schuemie MJ. Hripcsak G, Ryan PB, Madigan D, Suchard MA. Robust empirical calibration of p-values using observational data. Stat Med 2016;35(22):3883-3888.
3. Haneuse, Sebastien, and Kyu Ha Lee. 'Semi-Competing Risks Data Analysis'. Circulation: Cardiovascular Quality and Outcomes 9, no. 3 (2016): 322–31.
4. Wimmer NJ, Resnic FS, Mauri L, et al. Comparison of Transradial Versus Transfemoral Percutaneous Coronary Intervention in Routine Practice. Journal of the American College of Cardiology. 2013;22(62):2147-2150. http://www.onlinejacc.org/content/62/22/2147.
5. McFadden E, Lay-Flurrie S, Koshiaris C, Richards GC, Heneghan C. The Long-Term Impact of Vaginal Surgical Mesh Devices in UK Primary Care: A Cohort Study in the Clinical Practice Research Datalink. Clin Epidemiol. 2021;13:1167-1180. Published 2021 Dec 31. doi:10.2147/CLEP.S333775.
6. Resnic FS, Majithia A, Marinac-Dabic D, et al. Registry-Based Prospective, Active Surveillance of Medical-Device Safety. N Engl J Med 2017;376:526-35.

8. PATIENT EXPOSURE TO THE DEVICE

Exposure assessment may vary based on the types of devices that are being studied. For example,

a device that is implanted may have a different exposure measurement compared to a device that is used to perform a procedure. The latter involves time-limited exposure (Box 8) while with the former, exposure could be lifelong. Exposure definitions should be as specific and detailed as needed to reliably address both mechanistic and clinical research questions. For studies in which detailed device information is collected de novo, the device or procedure to which patients are exposed should be known exactly.

Additionally, assessment of when exposure might change for the specific device and plans to capture when and how exposure changed are critical. For example, an implanted device may be removed; knowing when this occurred and why it occurred is essential in device evaluation. The schedule of exposure assessments (patient or device) should be directly mapped to the study objectives.

As noted earlier in this Framework, lack of availability of standardized UDI codes (or, usually, any UDI codes) can make it difficult to identify specific devices, especially across different institutions that utilize different coding systems. Most EHR and insurance claims databases do not currently contain UDI information. With considerable effort, some health systems, exemplified by the Building UDI into Longitudinal Data for Medical Device Evaluation (BUILD) Initiative and other related projects, have been able to link UDI codes to EHRs, to facilitate research.[1-5]

8.1 General Principles to Follow

A. Define any **induction** (time between device use and expected time of primary outcome initiation) or latent (time from outcome initiation to outcome detection such as malignant tumor initiation to detection) **periods**.
B. Describe the units for exposure measurement

    I. Indicate if exposure is "any" (randomized to new implant or received new implant) versus duration of exposure (e.g., number of days since breast implant date, drug dosage released from drug-coated balloon or drug-eluting stent)
    II. Describe whether multiple exposures are inherent to the clinical situation. For instance, if multiple stents are implanted in a single procedure in a single patient, describe if the measurements to be made are for each patient-stent or for the first stent only

C. Describe the precision with which exposure will be measured; this includes the data source, misclassification error, and measurement error
    I. Specify how the device or "system," for devices consisting of multiple components, will be identified within the RWE data source (e.g., model number, UDI) and the specificity of information regarding the device use (e.g., anatomic location) (also see NESTcc Data Quality Framework).
    II. Describe approaches to reducing misclassification and assessing the impact of any device misclassification.
D. Describe the approach to confirming exposure to the device under study
E. Identify specific clinical or surgical aspects that may narrow or broaden the definition of the

exposure (e.g., anterior approach for hip replacement)

F. As noted in the section on Target Population, provide information on the training and **experience of device operator/surgical team, as appropriate to the study type**. For instance, do surgeons require 25 hours of training or 15 cases to be proficient for the device? (also, see "Roll-in subjects" in the Section on the Statistical Analysis Plan0) In some devices, the reduced procedure time may result in potentially less radiation exposure and radiopaque contrast injections to the patients.
G. Include **dose** of exposure (where relevant), **changes** in exposure status, and exposure to **other devices** (if multiple devices are used for the same procedure) that may impact the performance of the device being evaluated. For instance, whether a balloon pre-dilatation is used or not during a percutaneous coronary intervention procedure likely would affect the performance of a (bare-metal/drug-eluting) coronary stent for improving coronary luminal diameter in patients with symptomatic heart disease.
H. Define the comparator. When the research question pertains to a time-varying exposure, allowances for real-world switching between the exposed and comparator groups should be described; if the comparator group consists of unexposed patients who may become exposed during the study period, the index date should be defined in a way that avoids immortal time bias.
I. Protect against immortal time bias if that is a possibility. According to Suissa (2008)[6]: "Immortal time is a span of cohort follow-up during which, because of exposure definition, the outcome under study could not occur." The usual situation in which immortal time can occur is when a subject has to remain alive and free from the event of interest to meet the exposure definition. Suissa provides several examples, including the two studies of heart transplants published around 1970 giving rise to much of the subsequent discussion of this bias (Box 8B). See Suissa and Dell'Aniello (2020)[7] for further examples.

BOX 8: TIME-LIMITED EXPOSURE

Impella Ventricular Support Systems (Impella 2.5, 5.0, CP, LD) (FDA PMA Number: P140003/S004). Four devices, Impella 2.5, Impella CP, Impella 5.0, and Impella LD catheters, are in conjunction with the Automated Impella Controller and are temporary ventricular support devices. The devices are intended for short term use (< 4 days for the Impella 2.5 and Impella CP, and ≤ 6 days for Impella 5.0 and LD) and are indicated for the treatment of ongoing cardiogenic shock that occurs immediately (< 48 hours) following acute myocardial infarction or open-heart surgery as a result of isolated left ventricular failure that is not responsive to optimal medical management and conventional treatment measures. The original PMA for the Impella 2.5 system was approved and indicated for temporary use (< 6 hours). To support indication expansion (from < 6 hours to 4-6 days depending on the device models), the primary clinical study (ISAR-SHOCK) was a randomized clinical trial with two arms (intra-aortic balloon pump arm vs Impella 2.5 arm). Supplemental data and analyses from the Impella registries (U.S. Impella Registry and AB5000 Registry) were provided to demonstrate real world use for the patient population. In addition, PMA Post-Approval Study was conducted using cVAD registry to evaluate the safety and effectiveness of Impella devices in a real-world population.[8-10]

8.2 References or Supporting Literature

1. Drozda JP Jr, Dudley C, Helmering P, Roach J, Hutchison L. The Mercy unique device identifier demonstration project; implementing point of use product identification in the cardiac catheterization laboratories of a regional health system. Healthcare 2016;4:116-119. doi:10.1016/j.hjdsi.2015.07.002. Originally published on-line July, 2015.
2. Drozda JP Jr, Roach J, Forsyth T, Helmering P, Dummitt B, Tcheng JE. Constructing the informatics and information technology foundations of a medical device evaluation system: a report from the FDA unique device identifier demonstration. J Am Med Inform Assoc 2018;25:111-120. Originally published on-line May 3, 2017. Available at https://doi.org/10.1093/jamia/ocx041.
3. Wilson NA, Tcheng JE, Graham J, Drozda JP Jr. Advancing Patient Safety Surrounding Medical Devices: A Health System Roadmap to Implement Unique Device Identification at the Point of Care. Med Devices (Auckl). 2021;14:411-421. https://doi.org/10.2147/MDER.S339232
4. Tcheng JE, Nguyen MV, Brann HW, Clarke PA, Pfeiffer M, Pleasants JR, Shelton GW, Kelly JF. The medical device unique device identifier as the single source of truth in healthcare enterprises – roadmap for implementation of the clinically integrated supply chain. Med Devices (Auckl). 2021;14:459-67. Available at https://doi.org/10.2147/MDER.S344132.
5. Dhruva SS, Zhang S, Chen J, Noseworthy P, Doshi AA, Agboola K, Herrin J, Jiang G, Yu Y, Cafri G, Collison Farr K, Ervin K, Ross JS, Coplan P, Drozda JP. Safety and Effectiveness of a Catheter With Contact Force and 6-Hole Irrigation for Ablation of Persistent Atrial Fibrillation in Routine Clinical Practice. JAMA Netw Open. 2022 Aug 1;5(8):e2227134. doi: 10.1001/jamanetworkopen.2022.27134. Available at https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2795250
6. Suissa, S., Immortal Time Bias in Pharmacoepidemiology, American Journal of Epidemiology, Volume 167, Issue 4, 15 February 2008, Pages 492–499, https://doi.org/10.1093/aje/kwm324
7. Suissa S, Dell'Aniello S. Time-related biases in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2020;29(9):1101-1110. doi:10.1002/pds.5083
8. Summary of safety and effectiveness data (FDA SSED) of P140003/S004. https://www.accessdata.fda.gov/cdrh_docs/pdf14/p140003s004b.pdf.
9. Seyfarth, et al. A Randomized Clinical Trial to Evaluate the Safety and Efficacy of a Percutaneous Left Ventricular Assist Device Versus Intra-Aortic Balloon Pumping for Treatment of Cardiogenic Shock Caused by Myocardial Infarction. J Am Coll Cardiol. 2008; 52(19):1584-1588.
10. Vetrovec, et al. The cVAD registry for percutaneous temporary hemodynamic support: A prospective registry of Impella mechanical circulatory support use in high-risk PCI, cardiogenic shock, and decompensated heart failure. Am Heart J. 2018; 199:115-121.

9. STUDY DESIGN

A study protocol for a controlled trial or an observational study should include a detailed description of the design features used to evaluate the medical device. Applying sound strategies, observational studies may be designed to emulate randomized controlled trials,[1] but careful attention to bias vulnerability should be considered. Ultimately there may be more than one possible valid and efficient study design that are appropriate in different situations.[2] As examples:

- Pragmatic trials: these are randomized trials but often with broad entry criteria and without the rigorous focus on adherence that usually characterize tightly controlled randomized studies. The idea is to get the benefits of randomization but in a setting that more closely resembles clinical practice.

- Multi-arm cohort studies: More than two exposure (device / procedure) groups are followed over time for predefined outcomes.
- Case-control studies: patients with a particular outcome (e.g., need for reoperation related to breast implants) and exposure history (e.g., which type of implant) is assessed retrospectively.
- Single-arm cohorts: generally following a group with a specific device or procedure without a concurrent, parallel, comparison group with data collected under the same protocol. These may be extensions of particular arms from randomized trials. They may also include external control groups, in which data are drawn from a different source from the main study, and an attempt is made to identify a similar group of patients from another setting. The single arm may also be compared to an objective performance criterion or a performance goal that has been established historically.
- Self-controlled designs: there are a variety of these designs, but in all of these each patient serves as their own control. This design applies in situations in which the exposure is transitory.

The study question(s) of interest should be established first, and then the data source(s) and study design most appropriate for addressing these questions should be determined. Fundamental features required include the number and type of comparison groups, blinding, outcomes (primary, secondary, procedural, device etc.), if a controlled trial, or a pragmatic randomized control trial (as mentioned above), the experimental unit of randomization, and how randomization will occur. Sometimes hospitals or clinics or surgical services are randomized, rather than individual patients.

Additional aspects associated with device evaluations related to the effects of the device operator, the device procedure, and the complexity of the device should also be considered. The choice of study design will depend upon the ability to reduce bias, ethical issues, the practicality of executing the design, data quality, data availability, and the objectives of the study.

Because details of study design may evolve, depending on questions of feasibility that can arise during study conduct (e.g., participants might be unable to provide certain information, sample sizes may be reduced because of lack of availability of specific data elements, a primary endpoint might change, prior to data analysis, motivated by the publication of other studies), the study protocol may need to be updated after its initial completion. Prospectively incorporating adaptive options into a protocol is generally far more robust for final interpretative analysis than are mid-course corrections due to un-anticipated situations during an actively enrolling trial. Protocol changes should ALWAYS be accompanied by a formal protocol amendment, which should be posted to whatever registry was used to register the initial protocol, IRBs and FDA notifications may also be mandated. We note that the potential for later changes does not obviate the need to establish a protocol PRIOR to study initiation.

9.1 Specific Design

Characterize the specific study design, the number and type of treatment arms, and whether blinding is used to mask treatment (for pragmatic randomized controlled trials). For studies using existing databases (health insurance claims or EHRs), masking of outcomes during assessment of

covariate balance is encouraged, and the time period(s) for the analysis need to be prespecified and clearly justified.

A. General Principles to Follow

I. Describe and justify the choice of design as precisely as possible, using standard descriptors (e.g., “a 2-group parallel sham-controlled fully blinded randomized trial,” “a primary data collection observational cohort study,” “a case-control study”)
    a. If a primary data collection study, provide rationale for using randomization (controlled) or for not using randomization
II. Define the primary study objective (e.g., superiority, non-inferiority, equivalence, comparison with an objective performance criterion [OPC] / performance goal [PG] in a single arm study, descriptive study)
III. If a randomized trial, describe and justify treatment allocation
    a. If unequal allocation, discuss why the choice was made, given the tradeoff in statistical efficiency when the allocation ratio is other than 1:1.
IV. If an observational study and utilizing matching, describe number of matched sets, size of matched sets, and whether a fixed or variable ratio of one group to the other is used.
V. If adopting a machine learning approach to adjust for differences between participants in different treatment groups, details on the creation of training, validation, and test sets should be provided and justified.
VI. If performing a comparative study on existing data, describe whether a new-user design (also known as an incident user design) is used.[3,4] For example, in studies of catheter ablation for arrhythmia, study eligibility criteria require the exclusion of patients who had cardiac ablation procedures prior to the index ablation procedure so that new users of ablation procedures can be identified.[5] A new-user design decreases biases such as confounding since new users between study comparison groups are likely to be more similar in disease state than if also including prevalent users. Prevalent users, by virtue of having survived to the point of study enrollment, could be different from new users.
VII. Include a study design schema indicating as relevant, the index date, baseline period for covariate evaluation, continuous enrollment requirements, length of follow-up, censoring criteria, and study time period. For guidance with regard to development of a study design diagram,[6] "Graphical Depiction of Longitudinal Study Designs in Health Care Databases"

BOX 9A:
Study NCT02577887 was a prospective, non-randomized, multi-center observational study designed to evaluate the diagnostic capabilities, indications, MRI scanning capabilities and clinical outcomes of patients implanted with a Magnetic resonance imaging (MRI) compatible SJM pacemaker with a standard bradycardia pacing indication. Any patient that received an Assurity MRI, Endurity MRI pacemaker (or newer version) in the EMEA region or any patient that receives an Accent MRI, Assurity MRI, Endurity MRI or similar SJM MRI compatible device in the Asia-Pac region was eligible for enrollment in the study if they met the inclusion/exclusion criteria. The protocol then describes the subject follow-up (subjects followed for 12-months after pacemaker implant, 6 and 12-months post-implant and during any unscheduled follow-up visit with details on remote follow-up with Merlin.net) and data collection process (Electronic Data Capture system). The objectives of the study were clearly defined as the characterization of complication rates in the general pacemaker patient population (primary) and the characterization of the impact of usage of advanced features in pacemaker on the clinical outcomes and of MRI scanning capabilities and rates in pacemaker patients by country (secondary).[7]

BOX 9B:
Study NCT01805154 was a worldwide, multicenter, non-randomized registry/observational study. The study was set to enroll a minimum of 1500 Cardiac Resynchronization Therapy (CRT) patients from up to 70 centers worldwide. The goal of the study was to have a maximum of 500 CRT non-responder patients identified and assessed. Patients who were successfully implanted with St. Jude Medical CRT-D/P devices were eligible for enrollment in the study up to 30 days post implant of the device. The follow-up for all enrolled patients was every 3 months for 12 months after implant and Merlin.net remote follow-up was optional but encouraged. The protocol also contains a clear description of the study purpose (rate of response to CRT and understanding of the treatment and management strategies of non-responders to CRT). The decision algorithm to classify response to CRT based on the Clinical Composite Score (CCS) is detailed as part of the follow-up description.[8]

B. References or Supporting Literature

1. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758-64. doi: 10.1093/aje/kwv254. Epub 2016 Mar 18. PMID: 26994063; PMCID: PMC4832051.
2. Taur SR. Observational designs for real-world evidence studies. Perspect Clin Res. 2022 Jan-Mar;13(1):12-16. doi: 10.4103/picr.picr_217_21. Epub 2022 Jan 6. PMID: 35198423; PMCID: PMC8815667.
3. Johnson ES, Bartman BA, Briesacher BA, Fleming NS, Gerhard T, Kornegay CJ, Nourjah P, Sauer B, Schumock GT, Sedrakyan A, Stürmer T, West SL, Schneeweiss S. The incident user design in comparative effectiveness research. Pharmacoepidemiol Drug Saf. 2013 Jan;22(1):1-6. doi: 10.1002/pds.3334. Epub 2012 Oct 1. PMID: 23023988.
4. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in

pharmacoepidemiology: historical foundations and contemporary application. Curr Epidemiol Rep. 2015;2(4):221-228. doi:10.1007/s40471-015-0053-5. Epub 2015 Sep 30. PMID: 26954351; PMCID: PMC4778958.
5. Dhruva SS, Zhang S, Chen J, Noseworthy P, Doshi AA, Agboola K, Herrin J, Jiang G, Yu Y, Cafri G, Collison Farr K, Ervin K, Ross JS, Coplan P, Drozda JP. Safety and Effectiveness of a Catheter With Contact Force and 6-Hole Irrigation for Ablation of Persistent Atrial Fibrillation in Routine Clinical Practice. JAMA Netw Open. 2022 Aug 1;5(8):e2227134. doi: 10.1001/jamanetworkopen.2022.27134.
6. Schneeweiss S, Rassen JA, Brown JS, Rothman KJ, Happe L, Arlett P, Dal Pan G, Goettsch W, Murk W, Wang SV. Graphical Depiction of Longitudinal Study Designs in Health Care Databases. Ann Intern Med. 2019 Mar 19;170(6):398-406. doi: 10.7326/M18-3079. Epub 2019 Mar 12. PMID: 30856654.
7. Protocol of Advanced Bradycardia Device Feature Utilization and Clinical Outcomes II (BRADYCARE II). https://clinicaltrials.gov/ProvidedDocs/87/NCT02577887/Prot_SAP_000.pdf.
8. ADVANCE CRT Registry. ADVANCE Cardiac Resynchronization Therapy Registry. https://clinicaltrials.gov/ProvidedDocs/54/NCT01805154/Prot_ICF_000.pdf.

9.2 Blinding (Masking)

The treatment that a study subject receives may be blinded to all or some individuals involved in thestudy, including subjects, investigators, outcome assessors, and data analysts. To the extent possible,whether a randomized or observational study, proper blinding is encouraged. For example, it might be possible to blind data analysts to study group (e.g., by identifying "device A vs device B." In some studies it may also be important to blind independent event adjudicators to treatment assignment (e.g., in adjudicating whether an MI or a stroke occurred.)

A. General Principles to Follow

I. Describe who is blinded, when they are blinded, procedures used to blind, and when the blind will be broken
    a. Rationale for lack of blinding of investigators, participants, outcome evaluators, or statisticians should be provided; other strategies to conceal treatment allocation, outcome data, and covariates should be described
    b. In observational studies, researchers should remain blinded to all endpoints until the estimation of the treatment assignment mechanism is adequate (good balance on observable characteristics between treatment arms and sufficient overlap of treatment arms). For example, if propensity scores are used to control for confounding, the propensity scores should be estimated and comparison of covariate balance between treatments when fitting the propensity scores should be performed while the investigators are still blinded to endpoints. See the material in the Section on Statistical Analysis Plan for further details.

c. If outcomes will be adjudicated by experts, they should be blinded with regard to patients' treatment status.

II. Procedures used to maintain the blind should be included in the protocol

9.3 Units of Randomization, Observation, and Analysis

Units of randomization and observation are the unit that is randomized and the unit of outcome measurement, respectively. Often, the unit of randomization is the individual subject. However, for logistical reasons, the unit of randomization could be larger, such as randomly assigning families rather than individuals to receive treated versus untreated nasal tissues. Conversely, the unit of randomization could be "smaller" than the participant, such as randomizing the right limb to receive a device and the left limb to the comparison treatment. In the limb example, the unit of observation is the "person-limb" given outcomes are measured on each limb within a participant, a distinction that must be specified throughout study procedures as well as statistical analyses. Even in an observational study, units of analysis need to be carefully considered. In the limb example, the fact that there are two limbs per person needs to be considered in the analysis. In addition to these limb-level analyses, it is often useful to perform a person-level analysis, with an appropriately defined person-level outcome.

A. General Principles to Follow

I. Provide a precise definition of the randomization unit, including the rationale for the specific choice of unit

II. Include a clear definition of the unit of observation and analysis and the rationale for the choice[1]

B. References or Supporting Literature

1. van Schie P, van Bodegom-Vos L, Zijdeman TM IQ Joint study group, et al Effectiveness of a multifaceted quality improvement intervention to improve patient outcomes after total hip and knee arthroplasty: a registry nested cluster randomised controlled trial BMJ Quality & Safety 2023;32:34-46.
2. Cafri G, Wang W, Chan PH, Austin PC. A review and empirical comparison of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. Statistical Methods in Medical Research 2019, Vol. 28(10–11) 3142–3162.
3. Armstrong RA. Statistical guidelines for the analysis of data obtained from one or both eyes. Ophthalmic Physiol Opt 2013, 33, 7–14. doi: 10.1111/opo.12009.

BOX 9C:

This methodologic paper addresses multiple issues related to confounding in studies of medical devices but focuses on the particular issue of clustered data. The motivating example uses data from the Kaiser Permanente Total Joint Replacement Registry, specifically analyzing elective total hip replacements performed between 1 January 2003 and 31 December 2015. The treatment of interest was the use of implants with a ceramic femoral head from a single manufacturer (Biolox Delta, femoral head model Articul/EZE, made by Depuy Synthes in Warsaw, Indiana). The comparison group had femoral implants with the same model but made of metal.

The outcome for the study was time to failure of any component of the device for any reason, i.e., not just failure of the femoral head. This outcome captures failures of other components that might somehow be a consequence of the type of femoral head. This is a commonly-used outcome in post-market surveillance studies of orthopedic devices. A patient leaving the health insurance plan and death were treated as censoring events, in the absence of which implants were censored at the end of the study period (the end of 2015).

The authors first consider confounding at the individual patient level and the device level. Specific confounders included indication for surgery, age, BMI, race, gender, overall patient health, diabetes, characteristics of the implantation process (operative year, surgical approach), and device-level confounders (model names and other characteristics of the types of shells, liners, and stems used). They then consider two issues related to clustering: the estimation of variance of treatment effects and control of cluster-level confounding. At the cluster level, they focus on what they term observation-invariant cluster (surgeon) confounders. These are characteristics of the surgeon that, as the name implies, don't vary from patient to patient within the same cluster, an example of which is the education level of the surgeon prior to beginning his or her job. This is in contrast to characteristics that might vary across patients, e.g., some form of post-employment education.

The authors discuss the question of conditional vs marginal inferences. Conditional inferences are based on making comparisons within surgeon, whereas a marginal model considers variance estimation but does not condition on surgeon. The authors found that methods incorporating clustering into the estimation of variance performed better than those that ignored clustering. As the within-cluster correlation and the cluster size increase, the unadjusted standard error underestimates the true standard error. They also found that methods that account for cluster confounding were the least biased.[2]

BOX 9D

Armstrong (2013)[3] reviewed published papers describing ophthalmology studies, specifically addressing how data analyses dealt with the lack of statistical independence between eyes from the same patient. Measurements from the right and left eye of an individual are likely to be correlated, but most typical statistical tests assume observations in a sample are independent. They reviewed publications from three optometry journals during the period 2009–2012.

Of the 230 articles reviewed, 148/230 (64%) dealt with the lack of independence by collecting data only from one eye per patient and 82/230 (36%) collected data from both eyes. Various strategies were used to select the single eye in the 148 one-eye articles: the right eye, left eye, a randomly selected eye, the better eye, the worse or diseased eye, or the dominant eye. In the 82 two-eye articles, the analysis utilized data from: (1) one eye only, (2) both eyes separately (ignoring the correlation), (3) both eyes taking into account the correlation between eyes, or (4) both eyes using one eye as a treated or diseased eye, the other acting as a control.

Armstrong goes on to outline a variety of statistical methods for the analysis of eye-related data, starting with a discussion of the intraclass correlation coefficient. His Table 3 provides brief descriptions of those approaches, with an emphasis on analysis of variance, which can estimate within-subject and between-subject variability (the so-called components of variance). He also addresses ways to take advantage of the correlation between eyes in the experimental design. A simple approach, in a randomized trial, is to randomly assign each eye within a patient to a given treatment. Paired tests are then used to make the treatment comparison, and these tests are more statistically powerful compared with allocating individuals to treatments.

From a practical perspective, Armstrong suggests that if only one eye is to be included and both eyes are eligible, then the eye should be chosen randomly (as opposed to only the right or only the left, or only the dominant eye, etc.). Most importantly, he notes that "Investigators should clearly describe the design of their study, provide a rationale for their choice of one or both eyes, the selection criteria applied if one eye is chosen, and describe the appropriate data analysis."

His focus is on clinical trials, but these same principles apply to *any* situation in which there might be data from both eyes measured separately. In general, when considering a study RWD, Armstrong's work highlights points to consider when multiple measurements are captured on the same patient, whether eyes, limbs, lesions, or vessels. These considerations should begin with the study design (randomization or random selection) and end with the appropriate analysis methodology that accounts for the potential dependence in measurements.[3]

9.4 Mechanism of Treatment Assignment

This is the manner by which a treatment (device A versus B) is assigned (usually done for randomized studies, but it's possible to have an interventional study in which treatments are actually assigned in a non-random manner, (e.g., using an alternating sequence: A, B, A, B, ...) or administered (observational study) to a unit when there is more than one treatment option. In randomized trials, the treatment assignment mechanism is described as known because the investigators have control of the process. In non-randomized studies, treatment administration is

not assigned randomly; drivers for administration (e.g., more severe disease) must be hypothesized, measured, and controlled to minimize confounding.

A. General Principles to Follow

I. For randomized studies, characterize and justify the **treatment assignment mechanism**, including:
    a. Whether it is a fixed or adaptive randomization
    b. Whether randomization is centralized
    c. Describe stratification variable(s) such as center, operator, etc.
    d. Describe choice of a fixed or random block size and justify choice
    e. Indicate how and by whom assignment will be communicated (in-person, phone, web, etc.)
    f. Indicate who will know the allocation and when it will be known
II. For observational studies, characterize treatment administration and indicate **how confounding will be controlled**:
    a. For example, describe variables that will be used to estimate the probability of treatment administration(e.g., the propensity score). If adopting machine learning (an algorithm), describe the process. See the material in the Section on the Statistical Analysis Plan for further details on how this can be accomplished.
III. Describe procedures used to determine comparability of units in the treatment arms (e.g., standardized mean differences).
IV. Provide attrition for the number of participants: approached or identified, eligible, provided consent (if required), and included in study as depicted in a **CONSORT diagram**.
V. Describe how the treatment assignment mechanism (randomized studies) or characterization of treatment administration (observational studies) will be handled when competing products enter the market while assessing a medical device.

9.5 Other Covariates

A. General Principles to Follow

The following aspects should be pre-specified in the protocol:

I. Subgroups: Define and justify covariates describing groups of participants for which the device effect may vary. If continuous variables are dichotomized (or otherwise transformed into categorical variables), provide a justification for the choice of category boundaries.
II. Confounding: Define (continuous vs categorical) and justify covariates that may impact treatment administration and outcomes in observational

designs.

a. Approaches to control confounding, such as propensity score methods, (see Section on the Statistical Analysis Plan for details).

III. Causal diagrams are useful to identify the appropriate variables (either inclusion or exclusion of those covariates) for controlling confounding in the analysis. These diagrams, such as Directed Acyclic Graphs (see Figure), depict the **hypothesized** causal and confounding relationships among the variables potentially relevant to a given question),[1] and help avoid adjustment for: (1) intermediate variables or collider variables (i.e., variables caused by both exposure and outcome) for which control can negatively impact valid effect estimation (overadjustment), and (2) variables for which control has no impact on valid effect estimation but may affect its precision (unnecessary adjustment).[2]

IV. If covariates are not pre-specified, justify the approach to selecting variables for inclusion in statistical analyses (e.g., empirical variable selection). If adopting machine learning approaches, pre-specification of the procedure to implement the algorithm should be detailed.

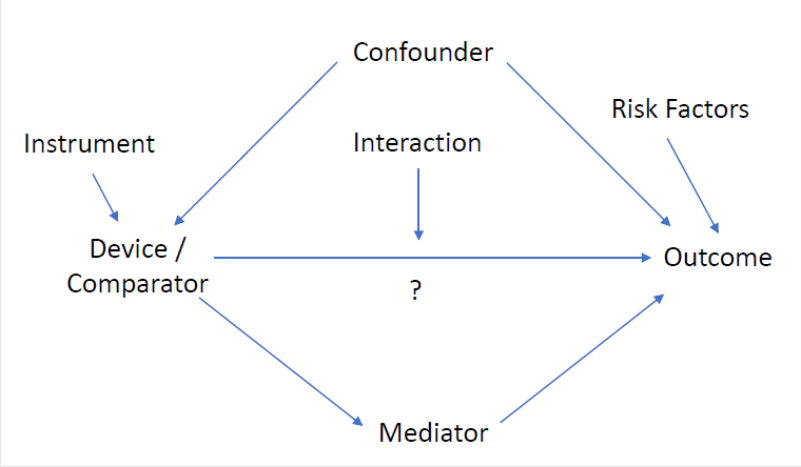
V. If categorizing covariates, provide the rationale for the choice of categories **and** ensure that the category definitions are not based on how the definition influences the estimated treatment effect.

VI. Characterize the completeness, quality, validity, and replicability of the covariates.

a. For additional information on completeness, quality, validity, and replicability of data, please refer to the Data Quality Framework

i. See the material in the Section on the Statistical Analysis Plan further details on how propensity scores can be estimated

Figure 1. Directed acyclic graph to identify potential data elements and assess which are key for the study question[3]



and applied.

B. References or Supporting Literature

1. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research.Causal diagrams for epidemiologic research. Epidemiology 1999 Jan;10(1):37-48. PMID: 9888278
2. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009 Jul;20(4):488-95. doi: 10.1097/EDE.0b013e3181a819a1. PMID: 19525685; PMCID: PMC2744485.
3. US 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. December 19, 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/draft-use-real-world-evidence-support-regulatory-decision-making-medical-devices.

10. STUDY PROCEDURES

For randomized studies and observational studies involving primary data collection, a clear description of how the study will be conducted ("study procedures") should be included in the protocol. Information should be provided regarding how patients are approached and consented (if required), how randomization will be conducted, how data will be collected and verified (source data verification), what constitutes subject withdrawal or discontinuation, definitions of protocol deviations and how these will be treated.

10.1 Assessment Schedule

NOTE: references to timing of data collection most obviously apply to interventional and observational studies using primary data collection, but in some instances, there will be analogous decisions when using existing data. The protocol should describe efforts to minimize loss to follow-up, such as a plan to escalate follow-up contacts.

In an observational study, the options are likely to be limited by what was done in usual clinical practice. The protocol should state the expected frequency of assessments of interest in clinical practice, including whether they are made on the basis of a suspected underlying condition. For example, blood pressure might typically be measured as part of any routine clinical office visit. Measuring troponin levels would likely depend on suspicion of a myocardial infarction. The protocol should (as relevant) describe any direct patient contact for follow-up and specify any additional data sources that may be used to supplement follow-up information (e.g., state registry of vital statistics).

All protocols, regardless of study type, should clearly define loss to follow-up.

A. General Principles to Follow

I. Specify **timing of patient evaluation** and justify the schedule, including:
   a. Baseline measurements related to patient characteristics, clinical history, and prognostic factors
   b. Measure baseline primary outcome if goal is to measure change

c. If using patient reported outcomes, it is important to collect one or more baseline outcomes
d. Specify that any baseline data must be measured or have occurred prior to treatment exposure
e. Provide rationale for both short-term (e.g., 30 days) outcomes such as length of stay, intensive care unit duration of stay, acute complications related to access or device, and late outcomes. For primary data collection, the scheduled assessments should be based on expectations of timing of safety events or expected benefits – is the device performing safely and having the desired effect? For existing data, the timing component of the outcome definition may be determined by whatever occurs (occurred) in clinical practice.

II. If assessing change, then describe the schedule of assessments and justify the need to measure repeatedly
III. Pre-specify a list of potential adverse effects and justify the frequency of assessment

Describe efforts that should be undertaken by study investigators to minimize loss to follow-up (for primary data collection studies), and clearly define the circumstances in which a study subject is considered lost to follow-up.

10.2 Informed Consent

Consent involves informing the patient or study participant what the study involves, why it is important,what is required of the participant, and who to contact in the event of a question, among other items. Itis a critical feature of clinical trials and a growing area in observational studies (expected for primary data collection). Use of secondary data may sometimes require participant consent or an IRB waiver. The Department of Health & Human Services has placed informed consent policies on the Office for Human Research Protections' website.[1] ISPE also includes a brief discussion of informed consent in the context of pharmacoepidemiology studies. For unique device research settings such as mechanical circulatory devices for patients in shock, the severity of the syndrome and compromise of patient mentation may require special statutory boundaries defining exemption from informed consent (EFIC).

A. General Principles to Follow

I. If no consent is required, provide rationale and supporting documents from the relevant Institutional Review Board or Research Ethics Committee
II. If required, consent should be obtained prior to subject enrollment
III. The consent process in special circumstances (e.g., subject unable to read or write, emergency treatments) should be described
IV. Include a statement indicating if vulnerable populations, e.g., children, are included and the process for obtaining consent
V. Provide an explanation of the research (e.g., risks, benefits, study completion, study discontinuation) using language that is non-technical and

understandable to the subject in a separate informed consent form (ICF), if required

VI. Provide ample time for the subject to read and understand the informed consent and to ask questions, receive answers, and consider participation

VII. Obtain dated signature acknowledging that his/her participation is completely voluntary

B. References or Supporting Literature

1. ISPE Guidelines for good pharmacoepidemiology practice (GPP). Pharmacoepidemiology and Drug Safety. 2016; 25:2-10.

10.3 Source Data Verification / Study Monitoring Plans

Study monitoring, including source data verification, is essential not only for the protection of human subjects, but also for the conduct of high-quality studies. Appropriate monitoring plans help ensure protection of the rights, welfare, and safety of the human subjects, and the quality of the study data pursuant to Good Clinical Practice (GCP) standards. Reasons for study monitoring include protocol compliance, and to ensure that data accuracy and completeness are maximized. Modern approaches, including risk-based strategies, should be considered, and would likely provide substantial efficiencies in ensuring data reliability.

In certain RWE applications, such as secondary analysis of existing data, monitoring functions may be more limited to assessment of data quality (e.g., missingness, values outside of viable range) and potential pathways to address such issues (see the NESTcc Data Quality Framework).

A. Site-Based and Central Data Monitoring

I. Describe the process for site-based and central data quality monitoring including members and how data issues will be resolved.

II. Describe data query, resolution, and final documentation processes including audit trail technology consistent with the electronic records, electronic signatures – scope and application portion of FDA Part 11 compliance.

10.4 Protocol Deviation Handling

For interventional studies and observational studies involving primary data collection, describe what types of deviations are anticipated, strategies to avoid them, and how the deviations will be handled in the study/analysis.

A. General Principles to Follow

I. Describe procedures in place to **minimize the inclusion of ineligible subjects** in the study

II. For interventional studies, provide procedures to **minimize the number of assessments** made outside of a follow-up interval, unless those are medically necessary.

III. Treatment crossovers are generally treated as protocol deviations in device trials. For example, if we have a registry-based randomized trial of Transcatheter aortic valve replacement (TAVR) vs. surgical aortic valve replacement (SAVR), and a SAVR patient receives TAVR, then that would constitute a protocol deviation, with the subject excluded from a per-protocol analysis. In an observational study, based on data arising from healthcare practitioners' decisions during the course of real-world clinical practice, treatment crossover isn't really defined, since there is no protocol-defined intervention, so is not considered as a protocol deviation. We also note that for a pharmaceutical trial, in contrast, switching therapies in the course of the trial might better be viewed as an adherence issue, which can be addressed analytically.
IV. All deviations from the final protocol, as well as their implications for study findings, should be fully described in the final study report.
V. For observational studies involving secondary analysis of existing healthcare data, the final study report should describe any departures in the conduct of pre-specified analyses from the specifications described in the protocol or statistical analysis plan, as well as their implications for study findings.

11. REQUIRED SAMPLE SIZE

The determination of sample size is a critical component of the design of a clinical study, whether randomized (Box 11A) or observational (Box 11B). If the sample size is too small, firm conclusions are unlikely to be inferred or results might be obtained by chance. On the other hand, an excessively large sample size would be wasteful and unethical and could lead to a statistically significant finding for an effect that may appear relevant without being clinically meaningful. For studies involving existing databases, a sample size calculation based on power may be unnecessarily prohibitive against conducting a study to address an important question,[1] although it's often helpful to reframe the calculations for those studies in terms of the magnitude of detectable treatment differences given a known sample size. For all studies, a clinically meaningful target difference or effect size (or non-inferiority margin) should be used as the basis for the sample size calculations. Determining the meaningful effect size can be challenging.[2] In practice, the study sample size is determined based on several design parameters and follows a set of statistical principles. Not all study designs require that sample size be fixed before the beginning of the study. In a group sequential design or an adaptive design, the eventual sample size depends on the trajectory of outcome data. In these designs, a stopping rule is used rather than a fixed sample size, which will then fluctuate based on the results of interim analyses. Nonetheless, the same basic statistical principles apply. If an observational study using existing data is planned, the sample size availability will generally be determined by the data source and dates defining the beginning and ending of the study period. One might also plan a study of existing data but not implement the analysis until sufficient sample size is available to meet the statistical power requirements. Often, it is difficult to compute exact sample sizes from prespecified power calculations due to the complexity of measurement error, selection bias, and confounding. Simulation studies are one

potential way to address such complexities. For studies in which high quality real world historical patient-level data already exist (e.g., professional society device procedure registries), randomized trials may be able to use RWE as a way to inform prospective RCT study design, possibly including providing an estimate of anticipated treatment effect.

11.1 General Principles to Follow

A. Indicate the type of **study design**:
    I. Fixed sample size
    II. Group sequential or adaptive (see interim analysis and stopping rule topic, Section on the Interim Analysis, Decision Rules, and Oversight)
B. Indicate approach to **evaluation**:
    I. If an estimation approach is adopted, provide and justify assumptions regarding widths of confidence intervals (i.e., precision) and anticipated effect size
    II. If a hypothesis testing approach is adopted, specify null and alternative hypotheses (basis for margin for a non-inferiority test), method of testing, test statistic, anticipated effect size (with justification), target power, and type I error rate/significance level
    III. Justify the selection of one-sided versus two-sided confidence intervals (or one-sided vs two-sided hypothesis test)
C. Indicate and justify additional features of the study that might influence sample size:
    I. Adjustment for multiplicity (e.g., hierarchical testing or simultaneous confidence intervals)
    II. Adjustment for clustering (e.g., center effects)
    III. Approach to controlling for confounding variables
    IV. Prevalence/incidence rates (reference and control cohorts)
    V. Accounting for missing data
    VI. Correction for loss to follow-up, treatment discontinuation, or other forms of

BOX 11A: SAMPLE SIZE CALCULATION (RANDOMIZED STUDY)

Ultrasonic pulsed echo imaging system: For the qualitative and quantitative evaluation of vascular morphology in the coronary arteries and vessels of the peripheral vasculature (FDA 510(k) Number: K173860). The iFR-SWEDEHEART study was a multicenter, randomized, controlled, open-label clinical trial using the Swedish Coronary Angiography and Angioplasty Registry for enrollment. The 12-month Kaplan-Meier estimates of the primary endpoint (all cause death, non-fatal MI, unplanned revascularization) will be compared in the two treatment groups by taking the difference of these estimates and calculating a one-sided upper 95%-confidence interval limit. The sample size estimation is based on formula of Makuch and Simon.[3] The non-inferiority (NI) limit is chosen to 3.2%, which corresponded to a noninferiority margin for the hazard ratio of 1.40 that was based on the anticipated event rate in the FFR group. An assumed endpoint event risk in the FFR-group is 0.08 compared to the risk in iFR-group of 0.076, which is equal to a 5% relative risk reduction. The noninferiority test is accepted if the upper 95%-confidence limit is less than 3.2%. This test requires a sample size of N=2000 to achieve 85% power.[4]

BOX 11B: SAMPLE SIZE AVAILIBILITY (OBSERVATIONAL STUDY)

Transcatheter aortic valve replacement (TAVR) comparing with surgery in intermediate-risk patients (ClinicalTrials.gov Identifier: NCT01314313). In the SAPIEN 3 observational study, 1077 intermediate-risk patients at 51 sites in the U.S. and Canada were assigned to receive TAVR with the SAPIEN 3 valve (952 [88%] via transfemoral access) between 17 February 2014 and 3 September 2014. In this population the all-cause mortality and incidence of stroke, re-intervention, and aortic valve regurgitation at 1 year after implantation were assessed. The 1-year outcomes in this population were compared to intermediate-risk patients treated with surgical valve replacement from the PARTNER 2A trial between 23 December 2011 and 6 November 2013, using a prespecified propensity score analysis to account for between-trial differences in baseline characteristics.[5]

censoring

11.2 References or Supporting Literature

1. Hernán MA. Causal analyses of existing databases: no power calculations required. J Clin Epidemiol. 2021 Aug 27:S0895-4356(21)00273-0. doi: 10.1016/j.jclinepi.2021.08.028.
2. Cook JA et al. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomized controlled trial. BMJ 2018;363:k3750.
3. Makuch RW, Simon RM. Sample size considerations for non-randomized comparative studies. J Chronic Dis 1980; 33: 175-81.
4. Götberg, et al. Instantaneous wave-free ratio versus fractional flow reserve to guide PCI. NEJM 2017; 376:1813-1823.
5. Thourani, et al. Transcatheter aortic valve replacement versus surgical valve replacement in intermediate-risk patients: a propensity score analysis. Lancet 2016; 387: 2218-2225.

12. STUDY REGISTRATION

Registration of randomized trials and non-randomized interventional trials (e.g., on www.clinicaltrials.gov) prior to conduct is standard practice and is required by publication policies at major journals and by governmental regulations (referred to as the "Final Rule"). Trial registration helps prevent selective analysis and reporting of endpoints. As an example, when trial results for the primary endpoint are not favorable, and secondary endpoints are favorable, registration allows the reader to make an informed judgment about the appropriateness of the reporting and the validity of the emphasis on secondary endpoints, if those endpoints become the focus of a publication.

Recommendations for registering observational study protocols are increasing with the goal of enhancing reproducibility and credibility (Box 12A and 12B). The benefits of posting randomized clinical trials protocols for public access have been discussed before[1] and some of these benefits would also certainly be valid for observational studies.[2,3] (Because www.clinicaltrials.gov,[4] was not designed to accommodate observational studies, registration is challenging. Other venues more suitable for observational studies are available.

One option is the Heads of Medicines Agencies-European Medicines Agency (HMA-EMA) Catalogue of RWD studies, which recently replaced the European Union electronic Register of Post Authorisation Studies (EU PAS Register))[5] An alternative recently made available is the International Society for Pharmacoeconomics and Outcomes Research and International Society for Pharmacoepidemiology (ISPOR/ISPE) registry.[6] At the current time, we believe that clinicaltrials.gov is better known than the ISPOR/ISPE website, so there may be advantages to being able to post a study on both sites and link the two listings. Currently, the Open Science Framework, hosting the ISPOR/ISPE registry, allows users to search both clinicaltrials.gov and the RWE registry simultaneously but the two sites are not dynamically linked unless a cross-link is provided during their respective registrations on either sites, which implies duplication of data entry. The ISPOR/ISPE site allows posting of full protocols, which may have an embargo date, before which the protocol would not be publicly available. Pre-specification and publication for all studies is strongly encouraged, will make the best evidence available, will assure a high degree of transparency, and will reduce ethical questions of conflict of interest.

12.1 General Principles to Follow

A. Trials should be registered on a publicly available website prior to enrolling the first patient, with no exceptions; observational studies could also be registered.

BOX 12A: EXAMPLES OF REGISTERED STUDIES WITH PUBLISHED RESULTS

1. One-Year Results From the SURPASS Observational Registry of the CTAG Stent-Graft With the Active Control System - \ (NCT03286400)
2. Antibacterial Envelope Is Associated With Low Infection Rates After Implantable Cardioverter-Defibrillator and Cardiac Resynchronization Therapy Device Replacement: Results of the Citadel and Centurion Studies - (NCT01043861/NCT01043705)
3. REPLACE DARE (Death After Replacement Evaluation) score: determinants of all-cause mortality after implantable device replacement or upgrade from the REPLACE registry - (NCT00395447)
4. Incidence and standardised definitions of mitral valve leaflet adverse events after transcatheter mitral valve repair: the EXPAND study - (NCT03502811)
5. Dedicated plug based closure for large bore access –The MARVEL prospective registry - (NCT03330002)
6. Actions elicited during scheduled and unscheduled in-hospital follow-up of cardiac devices: results of the ATHENS multicentre registry - (NCT01073449)

BOX 12B: OBSERVATIONAL REGISTRY CHARACTERIZING THE CTAG DEVICE WITH ACTIVE CONTROL

(SURPASS)(ClinicalTrials.gov registration number NCT03286400 - as can be accessed here: https://clinicaltrials.gov/ct2/show/study/NCT03286400). This study was first submitted on CT.gov on September 18, 2017, posted on September 18, 2017, recruitment started in October 2017 and the study's status was modified to Completed on October 23, 2019 for an actual study completion date recorded as October 9, 2019. The last update was posted on December, 17 2020. Protocol and SAP documents are available on ClinicalTrials.gov in the Study Documents section and the results were published by Torsello et al. In 2020.[7] Traceability of recorded changes made to the study information is done through a dedicated page: https://clinicaltrials.gov/ct2/history/NCT03286400

12.2 References or Supporting Literature

1. Chan AW, Hróbjartsson A. Promoting public access to clinical trial protocols: challenges and recommendations. *Trials*. 2018 Feb 17;19(1):116. doi: 10.1186/s13063-018-2510-1.
2. Chavers S, Fife D, Wacholtz M, et al. Registration of Observational Studies: Perspectives from anIndustry-Based Epidemiology Group. *Pharmacoepidemiology and Drug Safety*. October 2011;20(10):1009-1013. *https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.2221.*
3. Willians RJ, Tse T, Harlan WE and Zarin DA. Registration of observational studies: Is it Time? CMAJ 2010. DOI:10.1503/cmaj.092252
4. ClinicalTrials.gov. https://clinicaltrials.gov/ Accessed April 26, 2024.
5. European Union HMA-EMA catalogue of real-world data studies.https://catalogues.ema.europa.eu/catalogue-rwd-studies. Accessed May 16, 2024.
6. International Society for Pharmacoeconomics and Outcomes Research and International Society for Pharmacoepidemiology (ISPOR/ISPE) Registry. https://osf.io/registries Accessed April 26, 2024.

7. Torsello GF, Argyriou A, Stavroulakis K, Bosiers MJ, Austermann M, Torsello GB; SURPASS Registry Collaborators. One-Year Results From the SURPASS Observational Registry of the CTAG Stent-Graft With the Active Control System. J Endovasc Ther. 2020 Jun;27(3):421-427. doi: 10.1177/1526602820913007. Epub 2020 Mar 20. PMID: 32193990; PMCID: PMC7288855.

13. INTERIM ANALYSIS, DECISION RULES, AND OVERSIGHT

Use of an independent Data Safety Monitoring Board/Committee (DSMB/DSMC/or sometimes DMC) may not only ensure human subject safety but also reduce bias in study management. In the context of observational studies that include primary data collection, a version of the DSMB could be implemented for similar reasons. Some registries, and other RWE programs, even if no additional procedures are implemented, have Scientific Oversight Committees or Steering Committees to help make objective decisions about process, publication plans, authorship, and methodology. This level of oversight can be particularly important for industry-sponsored studies, where the objectivity provided by an external group can help by providing independent and objective scientific direction.

13.1 General Principles to Follow

A. **Data Safety Monitoring Boards/Committees**
   I. Describe the **charge** of the data safety monitoring committee, members and theirexpertise, frequency of meetings, and procedures in the DSMC Charter
   II. Describe the **processes for providing unblinded** data tables to independent committees without undermining central study integrity (indicate who is blinded to what information and when treatment assignments are revealed)
   III. Provide a description for periodicity of data review and formal approach to stoppingrule(s)

B. Interim Analyses
   I. Define operational procedures for the committee interpreting interim analyses (SteeringCommittee, Data Safety Committee, etc.)
   II. Define the purpose of any interim analyses (e.g., early stopping for futility, for efficacy,for safety, for adaptive designs, or potential mid-course corrections)
   III. Describe and justify the number and frequency of analyses
      a. If stopping rules are part of a specific dynamic study design, describe rules for stopping for futility, efficacy, or continuing and how sample size is impacted
      b. Pre-specify rule for stopping for safety
      c. Provide clinical and statistical justification for stopping rules
   IV. Describe and justify sample size, type I error, and alpha spending functions, and how the interim analyses impact the sample size needed for the primary outcome

13.2 References or Supporting Literature

1. Bruno A, Durkalski VL, Hall CE, et al. The Stroke Hyperglycemia Insulin Network Effort (SHINE) Trial Protocol: A Randomized, Blinded, Efficacy Trial of Standard vs. Intensive Hyperglycemia Management in Acute Stroke. International Journal of Stroke. 2014;9(2):246-251. https://doi.org/10.1111/ijs.12045.
2. SAP of Stroke Hyperglycemia Insulin Network Effort: The SHINE Trial. https://clinicaltrials.gov/ProvidedDocs/69/NCT01369069/SAP_001.pdf.
3. Lerner, et al. (2011) Design of the Circulation Improving Resuscitation Care (CIRC) Trial: A new state of the art design for out-of-hospital cardiac arrest research, Resuscitation https://doi.org/10.1016/j.resuscitation.2010.11.013.

BOX 13A: EXAMPLE FOR STOPPING RULES IN AN ADAPTIVE DESIGN USING O'BRIEN AND FLEMING GUIDELINES

The Stroke Hyperglycemia Insulin Network Effort (SHINE) trial protocol: a randomized, blinded, efficacy trial of standard vs. intensive hyperglycemia management in acute stroke (ClinicalTrials.gov Identifier: NCT01369069). The sample size estimate was based on data from the two NIH funded pilot trials, as well as other relevant acute stroke trials (see references 11-14 above). These data supported an estimate of 25% favorable outcome rate in the control group. The minimal clinically relevant absolute difference in favorable outcome between the two treatment groups was estimated to be 7% (control group = 25%; intervention group = 32%). The study is therefore powered to detect an absolute 7% difference in favorable outcome between the groups. The study design includes four interim analyses for both efficacy and futility of the primary outcome (after 500, 700, 900, and 1,100 patients complete the study) and a final analysis for a total of five planned analyses of the primary outcome. Including a 3% non-adherence rate and the four interim analyses, approximately 1,400 randomized patients are needed to provide 80% power with a two- sided type I error rate of 0.05.[1,2]

BOX 13B: ***Example of a Group Sequential Double Triangular Test to Monitor the effectiveness of a mechanical chest compression device***

BOX 13B: EXAMPLE OF A GROUP SEQUENTIAL DOUBLE TRIANGULAR TEST TO MONITOR THE EFFECTIVENESS OF A MECHANICAL CHEST COMPRESSION DEVICE

Design of the Circulation Improving Resuscitation (CIRC) Trial: a new state of the art design for out-of-hospital cardiac research (ClinicalTrials.gov Identifier: NCT NCT00597207). The study was designed to test for superiority or non-inferiority of a mechanical chest compression device against standard chest compression in a real-world ambulatory setting. The mechanical chest compression device was approved for use. A DSMB was formed to monitor study integrity as well as the group sequential testing procedure, while maintaining blinding of the accumulating results to the study sponsor. The

14. STATISTICAL ANALYSIS PLAN (SAP)

The statistical analysis plan (Boxes 14A to 14D) provides the detailed description of **all statistical analyses** to be conducted once the data are available. The **contents of the SAP in the protocol** are often less detailed than the final SAP, which might be a separate document. Ideally, the SAP should be approved prior to enrollment of the first study subject (for trials and observational primary data collection studies) or initiation of the primary analyses (in secondary data collection studies). Revision of the SAP should be minimized and occur before finalizing the analytic dataset ("database lock") and unblinding of the data or conducting any analysis. All SAP updates, especially those occurring after database lock or unblinding, should be accompanied by a listing of the modifications and rationale in the study report.

14.1 General Principles to Follow

Factors to consider and specify in the SAP include:

A. If the SAP is a separate document, it must include a description of the study objective, design, procedure, endpoints, and analysis population detailed in the final protocol to provide enough context for interpretation and implementation of the SAP
B. Define all variables used in development of datasets and conduct of analyses 0; describe computation of derived variables
C. Define study success criteria
D. Provide information about specific datasets, how they will be derived, where they will be stored, and what analyses are planned
E. Provide description of all statistical models, statistical hypotheses, tests, and estimation for:
    I. Analyses of primary, secondary, exploratory, procedural, device, and safety outcomes
    II. Interim analyses
    III. Subgroup analyses defined by baseline variables
    IV. Poolability analyses (by study site, and/or by region or data source). The question is whether it's appropriate to combine results from multiple centers, regions, or data sources. This is partly a clinical question: are the procedures, ancillary care, definitions of outcomes, definitions of exposure, etc., similar enough across these entities that it makes clinical sense to combine? This question can also be addressed analytically, using heterogeneity tests (interaction terms in a statistical model that test whether the estimate of the association between treatment and outcome is sufficiently similar across these entities to be combined).
    V. Any models or algorithms used in development of summary scores (e.g., propensity scores, disease risk scores), prediction scores, or variable selection procedures (e.g., NLP/MI-based algorithms, regression selection approaches), including the procedures used, the requirements for variable inclusion/exclusion, tuning parameters for the algorithm,

and subsequent integration into analyses of outcomes

VI. Sensitivity and supportive analyses including the features addressed and assumptions made

VII. Specific method(s) for assessing/handling missing data

VIII. Misclassification analyses evaluating impact of potential bias, heterogeneity, and error of measurement

F. Specify the intended statistical software

G. Summarize data sources:

I. Registry or database (timeframe of data collected)

II. Number of eligible subjects (selected by inclusion/exclusion criteria)

III. All potential baseline covariates and confounders available with details on how definitions might be different between data sources

H. Definition and justification of target population and study samples

I. Randomized controlled trials and real-world studies may define different study populations for analyses; a research question may be addressed with different study populations as long as both are relevant and reliable to address the question

II. Effectiveness and safety endpoints may be analyzed with different study populations

III. Clinically meaningful study populations for analyses may include Intention-to-treat (ITT), modified ITT (mITT), As-treated (AT), Per-protocol (PP), Complete set (CS), etc. The ICH E9(R1) Estimands Framework may be useful in this context and is being adopted by regulatory agencies.[1]

I. Roll-in subjects (i.e., subjects treated during a surgeon training period prior to enrollment of the primary cohort as a means to address a possible learning-curve effect) and/or crossover subjects should be analyzed separately in addition to the commonly defined target populations

J. If multiple endpoints and/or hypotheses were proposed for statements being proposed for inclusion in labeling (e.g., clinically relevant secondary endpoints), a plan for adjustment for multiplicity should be prespecified (e.g., gatekeeping procedure)

K. Describe and justify any interim analysis plan and its impact on statistical design and operating characteristics

L. If a noninferiority design is proposed, justify the acceptable or tolerable clinical margin

M. Specify details of supportive analyses. For example, if using machine learning or variable selection procedures for observational studies, specify the algorithm that will be adopted. If using Bayesian Additive Regression Trees, describe the number of trees, the size of the cross-validation samples, and the prior distributions for the number of variables. For regression selection approaches, indicate what procedures will be used, the requirements for variables to enter or to exit, etc. along with any tuning parameters.

N. Provide details of approaches to control confounding, such as propensity score methods, (see Section on the SAP).

O. Approaches to Control Confounding:

In studies generating RWE, randomization may not be feasible. Thus, any comparison of outcomes between treatment arms is potentially subject to confounding, which can lead to biased treatment effect estimates. Confounding may be addressed via study design (see Section on the Study Design) and analysis. There are numerous approaches to control for confounding at the analysis stage. Common approaches are:

- Restriction to those with specific characteristics, e.g., excluding people with certain comorbidities
- Stratification by subject characteristics, generally involving calculation of a summary estimate of association, e.g., stratification by age group, calculating a measure of association within each stratum, then producing a combined weighted estimate across strata
- Multivariable regression models, where specified covariates are included as adjustment terms in calculating a measure of association
- Balancing score adjustment, e.g., propensity scores
- Matching, e.g., based on the Mahalanobis distance, exact matching, or coarsened-exact matching

Propensity score-based confounding adjustment is perhaps the most common analytic approach used in recent years. Given the complexities of this approach, additional details should be included in the SAP specifically for this method, denoting: modeling approach and selection criteria for development of propensity score, assessment of covariate balance, propensity score adjustment in outcome analyses, and planned approach if propensity score adjustment is not reliable.

Considerations for development and utilization of propensity score-based confounding adjustment include:

I. Describe propensity score modeling and estimation:
    a. Definition of propensity score
    b. Provide statistical methods such as logistic regression, random forest method, etc.
    c. Specify model selection criteria including trimming of the estimated propensity scores, mitigation of extreme propensity score values, and if and how unmatched subjects will be analyzed
    d. Include a plan for how missing data on baseline covariates will be handled
    e. Specify baseline covariate balance assessment methods based on the selected propensity score model, the methods may include goodness of fit tests, averaged standardized (absolute) mean differences, and graphical diagnostics (e.g., box plots)
II. A two-stage **outcome-free study design**[2-4] should be used where feasible
    a. The first stage of study design is similar to designing traditional randomized controlled trials or single arm studies
    b. For the second stage of study design, identify in advance independent statisticians who are blinded to the outcome data and who will develop the propensity score models
    c. Propensity score estimates and values for later adjustment (stratification IDs,

strata weights, inverse-probability weights, matching IDs, etc.) should be stored as the results of the second stage of study design, and linkable to outcome data for the final analyses. These results should be reviewed (e.g., by regulators) prior to conducting the comparative analysis.

III. Specify propensity score adjustment method(s) to be used, including details on any methodological parameters:
    a. Propensity score covariance adjustment (other adjustment covariates, whether a non-linear transformation was used for the propensity score term, etc.)
    b. Propensity score stratification (number of strata)
    c. Propensity score matching (caliper, algorithm, number of matched pairs, with or without replacement, whether a paired or independent samples analysis will be performed, etc.)
    d. Propensity score weighting (e.g., Inverse probability weighting with or without trimming, average treatment effect on the treated (ATT) weighting, stabilization, etc.)
IV. Considerations of the proposed propensity score adjustment method should be detailed, for example,
    a. If the intent is to avoid excluding subjects and/or altering the target population, stratification and average treatment effect on the treated (ATT) weighting might be preferred
    b. When extreme propensity scores (very close to 0 or 1) are not expected to be an issue and/or the proportion of subjects with extreme propensity scores may not be significant, inverse-probability treatment weighting might be preferred
    c. A matching approach might be preferred when matched subjects are expected to adequately represent the target population
V. Rather than prespecifying a single method for using propensity scores (e.g., matching, stratification, others), it has been suggested[5] that one applies multiple methods during the preliminary analysis blinded to outcomes, then select one method that minimizes imbalances across measured covariates. The search for the best design would prespecify the decision-making process for which method will be selected, rather than selecting a method without knowledge of relative performance of several methods.
VI. In case reasonable propensity score adjustment cannot be achieved, a secondary analysis plan should be prespecified before unblinding of any outcomes data, with fallback options potentially including:
    a. Performance goal for single arm study
    b. Continuing enrollment of current study
    c. Use of other data sources
VII. When approaches other than propensity score adjustment are used, details of the method applied, including justification of how confounding is adequately dealt with should be reported

Propensity score methods can present a variety of challenges and limitations under certain circumstances. For example, limited overlap in covariate distributions may result in substantial loss in sample size arising from from unmatched or extremely weighted treated or control units. In the context of rare exposures (small Ns), failures of model convergence may cause issues with the estimation of the propensity score. Propensity scores also do not allow the user to flexibly assign the priority of balance to covariates that may possess greater or lesser prognostic value relative to outcomes. [6-8]

Two optimization-based covariate balancing techniques are currently being explored as promising alternatives to propensity score matching and weighting methods, specifically: cardinality matching (CM) and stable balancing weights (SBW). [9-10] CM and SBW are unique in that they use the principles of optimization to directly target covariate balance. Specifically, they allow researchers to pre-specify desired covariate balancing constraints – such as maximum allowable standardized mean differences for covariates between two or more groups – and explicitly solve an optimization problem to yield the mathematically-guaranteed largest matched sample (CM) or weights of minimum variance (SBW). Pre-specified balancing constraints may also be applied to each covariate separately to allow for tighter balance on covariates known to have greater prognostic value, and they can be applied flexibly to target balance of various moments of a distribution or even exact/perfect marginal or conditional distributional balance. CM and SBW also allow researchers to flexibly match or weight, including precisely specified covariate distributions (e.g., an external control). Finally, CM and SBW also enable matching-adjusted indirect comparisons when individual patient data are available for only one group under comparison. These modern techniques, which have been enabled by recent advances in the science of optimization and computing power, warrant attention and thorough evaluation relative to pre-existing covariate balancing techniques.

Specific considerations for addressing misclassification and missing data include:

A. Describe measures adopted to minimize data collection biases (e.g., standardized structured data capture, with harmonized definitions) and to assess the potential impact of any remaining misclassification
    I. Aspects of potential misclassification that may affect assessment of the effects of the device under study include, but are not limited to: baseline covariates (e.g., extreme and therefore erroneous values in lesion length, incorrect disease stage), received treatments (e.g., coding errors leading to inaccurate distinction of drug-eluting stent vs. bare metal stent), duration of device exposure (see also Section on the Patient Exposure to the Device of this document), and measurement of outcomes (e.g., cause of censoring not captured properly, misdiagnosis)
    II. Information from other medical record data (see also NESTcc Data Quality

Framework) may be implemented to reduce the misclassification rate of the RWD involved in the study, i.e., additional data sources, which can be linked to the primary data, can sometimes be used to confirm (or not) the classification used in the primary dataset

III. Measurement performance metrics (e.g., sensitivity, specificity, kappa value) assessing alignment/consistency between RWD and other adjudicated medical record data may be based on a validation study (see Section on the Validation of Key Study Variables on validation studies) and should be reported. These measurement characteristics can be used in subsequent statistical approaches to determine robustness of findings to misclassification

IV. Statistical methods for assessing misclassification may include simple bias analysis,[6,11] probability bias analysis,[12] Bayesian bias analysis,[13] modified maximum likelihood, multiple imputation,[14] or regression calibration[15] that evaluates the impact of potential bias, heterogeneity, and error of measurement.

B. Provide information regarding missing data

I. Summarize the proportion of missing data for each study outcome and baseline covariate based on the target populations (e.g., ITT, mITT, AT, PP, CS, etc.), by devices (when comparators involved), or by data sources

II. Provide any strategies to potentially identify patterns in the missing data (e.g., missed visits associated with adverse events; or missed visits associated with poor clinical outcomes recorded at earlier follow-ups.)

III. Generally, it is good practice to compare subjects with missing data with those without missing data, with respect to baseline covariates and outcome variable trajectories up to the last observed value[16]

IV. It's potentially helpful to think of missing data in the contexts of formal, prospective data collection compared with what happens in usual care. In a clinical trial, when blood pressure measurements are collected on a protocol-specified schedule, data will be missing when someone misses a visit. It's then important to understand the reason(s) for missed visits. In claims or EHR data, blood pressures are collected according to routine clinical care, and the presence of blood pressure measurements may depend on the clinician having a specific reason to take the measurement, e.g., someone who is older, has cardiovascular disease, and a history of hypertension might be more likely than a younger, healthier person, to have a recorded blood pressure measurement.

C. Acceptability of the amount of missing data should be assessed on a case-by-case basis either considering the proportion of missing data alone or using a novel approach such as the fraction of missing information[12,13]

D. Indicate the planned treatment of missing data, associated assumptions (e.g., missing completely at random (**MCAR**), missing at random (**MAR**), or missing not at random (**MNAR**),[19] and how the associated assumptions will be validated.

I. Although MCAR and MAR are different conceptually, under either assumption, standard multiple imputation can be used, which allows missing data to be handled in a statistically valid manner.[20]
II. To illustrate the principles, consider blood pressure measurements. MCAR means that the people with missing blood pressure are a random subset of all people in the dataset, i.e., the distribution of missing blood pressure values (if they had been measured) would look the same as the distribution of measured blood pressure values. MAR means that the missing values of blood pressure may differ systematically from the measured values, but the differences can be explained by other variables.
    - For example, blood pressure measurements in claims or EHR databases, older people and those with cardiovascular disease are more likely to have their blood pressure measurements taken and recorded. But they are also more likely to have high blood pressure than younger people without cardiovascular disease, who are less likely to have measurements taken and recorded. Given information on age and cardiovascular disease, blood pressure could still be **MAR**. Stratifying on age and cardiovascular disease status will tend to reduce differences between the missing and the measured blood pressure measurements. Within the stratum of young people with no cardiovascular disease, missing and measured blood pressures are less likely to differ than without the stratification. The same argument goes for all strata defined by age and cardiovascular disease. It's this logic that allows the use of methods that assume **MAR** (because within strata, missing and measured blood pressures are likely to have similar distributions. If there are other factors, e.g., sex, that can influence blood pressure, these can be incorporated into the imputation process
III. A systematic review of data missingness and methods for handling missing data can be found in Yan, Lee, Li, 2009.[21]
IV. When the missing data may not be due to randomness such as MCAR and MAR, the missing mechanism may be referred to as MNAR, where the chance of a measurement being missing may be dependent on the unobserved value itself. For example, higher blood pressures may be measured less completely if participants drop out of the study because of an ineffective intervention. On a case-by-case basis, analyses may be based on further assumptions and clinical insights about the distribution of unobserved values. In addition, tipping point analyses may be considered because the method does not depend on the mechanism of missing data.[21]

E. Include clinical assessments for the importance of covariates associated with study outcomes and consider clinical imputation when it is appropriate.
    I. For example, using imaging to assess medical device integrity, such as stent

fracture, from a later follow-up to impute an earlier follow-up.) and statistical imputation methods (e.g., multiple imputation or likelihood approaches) for missing study outcomes and/or baseline covariates, and any planned imputation models that could be pre-specified.

F. Use tipping point analyses for study outcomes (i.e., sensitivity analyses that test the robustness of the conclusions to various assumptions about the missing data. Generally, these ask how different the subjects with missing data would need to be to overturn the original conclusions.)[21]

BOX 14A: STATISTICAL ANALYSIS PLAN

XIENCE Family of Everolimus Eluting Coronary Stents: Indication expansion to include patients with diabetes mellitus where registry data is a primary source of clinical real-world evidence. Four historical studies (SPIRIT IV (ClinicalTrials.gov Identifier: NCT00307047), SPIRIT PRIME (NCT00916370), XIENCE V USA 5K, and XIENCE V USA 3K (NCT00676520) and two external registry databases (Cleveland Clinic and the Wake Forest Baptist Medical Center) were included to support the indication expansion for approved XIENCE Family of Stents under an FDA PMA supplement: P070015. The two external databases are real-world observational registries which are part of the National Cardiovascular Data Registry CathPCI registry (https://cvquality.acc.org/NCDR-Home/registries/hospital-registries/cathpci-registry). A Bayesian hierarchical model was utilized to analyze the primary endpoint of target vessel failure (TVF) at 12-months, defined as a composite of cardiac death, target-vessel myocardial infarction (TVMI), or ischemia driven target vessel revascularization (ID-TVR). The TVF rate was also tested against a prespecified performance goal (PG) of 14.8% (expected rate 8.6% plus an absolute margin of 6.2%). Section X of the FDA SSED for P070015/S128 and P110019/S075 also provided the summary of study design and primary clinical studies that were analyzed in line with the general principles to follow.[22]

BOX 14B: SAP INCORPORATING PROPENSITY SCORE ADJUSTMENT

IN.PACT Admiral Paclitaxel-Coated Percutaneous Transluminal Angioplasty (PTA) Balloon Catheter: The indication was expanded to treat in-stent restenotic (ISR) lesions in superficial femoral or popliteal arteries. Data from the clinical study ("DCB ISR Cohort") were retrospectively compared to standard PTA data ("PTA ISR Comparator") from 23 US sites employing propensity score adjustment. For the DCB ISR Cohort, patients were treated in the IN.PACT Global Study between June 6, 2012 and December 16, 2013 at 31 sites outside the U.S. (OUS). A total of 164 DCB subjects from this study met the inclusion criteria. For the PTA ISR Comparator, patients were treated at the U.S. sites from the Society of Vascular Surgery (SVS) Vascular Quality Initiative (VQI) registry between 2011 and 2014. More than 500 patients were screened for eligibility and a total of 153 PTA subjects met the inclusion criteria. A propensity score analysis was performed using clinically relevant baseline characteristics pre-specified as the covariates in the propensity score model. All the 20 covariables were included in the propensity score calculation except TASC lesion type, which was excluded due to a missing data rate that exceeded the prespecified cutoff (20%). For each variable with missing values (<20%), a gender-specific imputation was performed by replacing the missing values of the variable with the gender-specific median observed value within each group. The primary analysis set was based on the intent-to-treat (ITT) principle. All subjects enrolled through the selection process specified in the SAP were included as ITT subjects. To analyze the treatment differences between the DCB ISR Cohort and PTA ISR Comparator groups in the clinical/safety endpoints such as TLR, a propensity-quintile-stratified Cox proportional hazards model was employed, with time to event as the dependent variable and treatment group as the independent variable. The superiority of DCB ISR Cohort on the 12-month primary effectiveness endpoint of target lesion revascularization compared to the PTA ISR Comparator (10.13% vs. 35.92%, p<0.001) was demonstrated in a prespecified, propensity score-adjusted analysis.[23]

BOX 14C: SAP INCORPORATING PROPENSITY SCORE ADJUSTMENT

da Vinci® Xi and X Surgical Systems: Indication expansion to include ventral hernia repair where registry data is the primary source of clinical real-world evidence. Data from the Americas Hernia Society Quality Collaborative (AHSQC) Registry was used in a submission sought clearance for a labeling modification to include "Ventral Hernia Repair" (VHR) procedures under the cleared "general laparoscopic surgical procedures" indication for use of the da Vinci Xi and X Surgical Systems. As part of this submission, a propensity score matched analysis was performed comparing da Vinci and laparoscopic non-complex VHR (without myofascial release). The analysis included data from the AHSQC registry for procedures that occurred between July 7, 2013, and January 1, 2017. The total number of non-complex VHR procedures in the registry during this time period was 873 and 1,961 for the robotic-assisted and laparoscopic cohorts, respectively. One-to-one propensity score matching algorithm was used to identify comparable groups of patients and to adjust for potential selection bias that could result from surgeon choice of repair approach. Only those demographic and surgical characteristics that were known preoperatively were used as covariates in a logistic regression model. Furthermore, outcome data were not analyzed prior to the development of the propensity score model and selection of matched subjects. Because less than 3% of observations were missing covariate values used in the propensity score model, a single imputation strategy, specifically, single imputation with Multivariate Imputation by Chained Equations (MICE) was utilized. Patients in the laparoscopic group were matched to the patients in the robotic-assisted group based on the logit transformation of the propensity score, and the difference in scores between matches was not allowed to exceed 0.2. Matched data included 615 patients in each treatment group. A standardized mean difference of less than 0.1 was considered excellent balance and between 0.1 and 0.2 was considered acceptable. Other matching diagnostics used included hypothesis testing, butterfly plots, and empirical cumulative distribution function plots. The submission also included a propensity score matched comparison between robot-assisted cohort and open complex VHR.[24,25]

BOX 14D: DETAILED DESCRIPTION OF PROPENSITY SCORES AND WEIGHTING

In the study on the Association of Uterine Perforation and IUD Expulsion With Breastfeeding Status at the Time of IUD Insertion and Postpartum Timing of IUD Insertion in Electronic Medical Record Databases (NCT03754556) Confounding was controlled through the use of propensity scores and overlap weights, as described in Section 4.1.5 of the SAP. The SAP then details how the overlap weights will be computed and how extreme weights will be assessed "although extreme weights are not expected to be an issue with overlap weights because the overlap weights are bounded. The overlap weights put the strong focus on those patients with the highest overlap in their propensity scores and therefore avoid the need for trimming the population.". The document also provides a threshold for balance: "The exposure groups are considered balanced if the standardized difference is less than 0.20 (generally considered small)." In both cases, references are provided.[26]

14.2 References or Supporting Literature

1. ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principals for clinical trials. 2020. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf. Accessed on April 24, 2024.
2. Yue, L.Q. (2007) Statistical and Regulatory Issues with the Application of Propensity Score Analysis to Nonrandomized Medical Device Clinical Studies. Journal of Biopharmaceutical Statistics, 17(1), pp. 1-13. .
3. Yue, Lilly Q, Nelson Lu, and Yunling Xu. 'Designing Premarket Observational Comparative Studies Using Existing Data as Controls: Challenges and Opportunities'. Journal of Biopharmaceutical Statistics 24, no. 5 (2014): 994–1010. .
4. Lu N, Xu Y, Yue LQ (2020). Some Considerations on Design and Analysis Plan on a Nonrandomized Comparative Study Using Propensity Score Methodology for Medical Device Premarket Evaluation. Statistics in Biopharmaceutical Research 12(2): 155-163. https://doi.org/10.1080/19466315.2019.1647873
5. Cafri G, Coplan P, Zhang S, Berlin JA. Selecting an Optimal Design for a Non-Randomized Comparative Study: A Comment on "Some considerations on design and analysis plan on a nonrandomized comparative study utilizing propensity score methodology for medical device premarket evaluation" (letter). Statistics in Biopharmaceutical Research 2022;14:262-264.
6. Shiba K, Kawahara T. Using propensity scores for causal inference: pitfalls and tips. J Epidemiol. 2021 Aug 5;31(8):457-463. doi: 10.2188/jea.JE20210145.
7. Ross ME, Kreider AR, Huang YS, Matone M, Rubin DM, Localio AR. Propensity Score Methods for Analyzing Observational Data Like Randomized Experiments: Challenges and Solutions for Rare Outcomes and Exposures. Am J Epidemiol. 2015 Jun 15;181(12):989-95. doi: 10.1093/aje/kwu469. Epub 2015 May 20. PMID: 25995287.
8. Li H, Wang C, Chen WC, Lu N, Song C, Tiwari R, Xu Y, Yue LQ. Estimands in observational studies: Some considerations beyond ICH E9 (R1). Pharm Stat. 2022 Sep;21(5):835-844. doi: 10.1002/pst.2196.
9. Zubizarreta JR, Paredes RD, Rosenbaum PR. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. Ann. Appl. Stat. 2014;8(1):204-231. DOI: 10.1214/13-AOAS713.
10. Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association. 2015;110(511):910-922. https://doi.org/10.1080/01621459.2015.1023805.
11. Berkson J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. Int J Epidemiol. 2014;43(2):511-515. doi:10.1093/ije/dyu022
12. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of

misclassified binary variables. Int J Epidemiol. 2005;34(6):1370-1376. doi:10.1093/ije/dyi184
13. Dellaportas P, Stephens DA. Bayesian analysis of errors-in-variables regression models. Biometrics. 1995:1085-1095.

14. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol. 2006;35(4):1074-1081.
15. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. The American Journal of Clinical Nutrition Volume 65, Issue 4, April 1997, Pages 1179S-1186S.
16. Austin, P. C., White, I. R., Lee, D. S. and van Buuren, S. (2021) Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Canadian Journal of Cardiology. 37. 1322-1331.
17. Madley-Dowd, P., Hughes, R., Tilling, K. and Heron J. (2019) The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinica Epidemiology. 110. 63-73.
18. Jakobsen, J. C., Gluud, C., Wetterslev, J and Winkel, P. (2017) When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology. 17:162
19. Little, R., Rubin, D. (1987). Statistical Analysis With Missing Data. New York: John Wiley & Sons.
20. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? Int J Epidemiol. 2014 Aug;43(4):1336-9. doi: 10.1093/ije/dyu080. Epub 2014 Apr 4. PMID: 24706730; PMCID: PMC4121561.
21. Yan X., Lee S., Li N. (2009) Missing Data Handling Methods in Medical Device Clinical Trials, Journal of Biopharmaceutical Statistics, 19:6, 1085-1098, DOI: 10.1080/10543400903243009
22. Summary of safety and effectiveness data (FDA SSED) for P070015/S128 and P110019/S075. https://www.accessdata.fda.gov/cdrh_docs/pdf7/P070015S128B.pdf. Accessed April 26, 2021.
23. Summary of safety and effectiveness data (FDA SSED) of P140010/S015. https://www.accessdata.fda.gov/cdrh_docs/pdf14/P140010S015B.pdf. Accessed April 26, 2021.
24. 510(k) premarket notification: https://www.accessdata.fda.gov/cdrh_docs/pdf17/K173585.pdf Accessed April 26, 2021.
25. LaPinska, M., Kleppe, K., Webb, L. et al. Robotic-assisted and laparoscopic hernia repair: real-world evidence from the Americas Hernia Society Quality Collaborative (AHSQC). Surg Endosc 35, 1331–1341 (2021).
26. Statistical analysis plan. Study on the Association of Uterine Perforation and IUD Expulsion With Breastfeeding Status at the Time of IUD Insertion and Postpartum Timing of IUD Insertion in Electronic Medical Record Databases – A Postmarketing Requirement for Mirena (APEX IUD). https://cdn.clinicaltrials.gov/large-docs/56/NCT03754556/SAP_001.pdf. Accessed April 26, 2024.

15. STUDY REPORTING

Throughout this Framework, we have emphasized the need for transparency in describing research methods for studies of medical devices and the motivation for those methods. The same need applies with respect to reporting studies when they are completed. A commonly stated principle is that a study report should contain enough detail that someone with knowledge of the appropriate methods could reproduce the study, if they were provided with the data. To that end, we will simply state that whatever details are important to the development of a study protocol, should also go

into the study report.

The material we present in this section focuses on preparation of peer-reviewed publications. We note that, in many situations, there will also be a full study report prepared for various stakeholders. If anything, such full reports should have even more detail than we describe below, but the specifics will depend on the needs of the stakeholder. For some documents, an Executive Summary is also provided, which would summarize key points at a high level for those stakeholders who don't need the full details. There is also increasing emphasis on plain language dissemination of study methods and findings aimed at patients and healthcare providers. What we describe below should be viewed as a minimum set of requirements.

To help accomplish the goal of full transparency, there are published reporting guidelines for various types of studies. For RCTs, the standard reporting guideline is CONSORT,[1] which lays out the important information that should be included in a study report. The checklist[1] and a longer explanatory paper[2] are widely used by both industry and academic scientists and this version includes RCTS that incorporate RWD, either as a data source for the randomized trial itself, or to provide an additional external control, beyond the randomized control group that is part of the trial.

For observational studies, one relevant guideline is the STROBE statement, which is also published as a checklist[3,4] and a separate, much longer, explanatory document.[5,6] More recently, an ISPOR/ISPE taskforce published guidance on reporting of observational studies, aimed at improving transparency and reproducibility. That work was further enhanced by publication of the StaRT-RWE structured template for planning and reporting observational studies.[7] This focuses mainly on studies of pharmaceutical products, but the principles will also apply to studies of medical devices. A very detailed approach to reporting of studies conducted using routinely-collected health data is presented in the RECORD Statement.[8] The Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network[9] is a useful resource that provides a single, general site for finding the appropriate guideline for any given type of study. The FDA CDRH also published a guidance on the content and format of the study protocol and interim and final reports of post-approval studies imposed by device premarket approval application order.[10]

The report should include the statistical analysis actually performed. Additional analyses (not specified in the protocol) performed after the conduct of the primary/inferential analysis should be clearly reported as post-hoc. Full accounting of statistical analyses, both protocol-specified and post hoc, is essential. Any changes in what was planned should, in principle, have been captured by protocol amendments. For example, if evolving analytical results raise new questions prior to initiation of the primary (inferential) analyses, the additional analyses should be included in a protocol amendment. If any details were inadvertently missed, or were added too late, or if plans could not be realized because of technical issues, e.g., failure of a particular statistical procedure to converge (to produce estimates of associations), the study report should indicate these deviations and explain the reasons for them. Sometimes, reviewers for a journal, for a health authority, or for a payer, might have specific requests for additional / different analyses from those that were originally planned and carried out. Again, such discrepancies should be reported as not having been specified in the original protocol. The shift from what is being planned (i.e. protocol and statistical analysis plan) to what was actually done, will primarily involve a change in verb tenses, from "what we are planning to do" to "what we did in practice."

15.1 General Principles

Potential structure for reporting on studies used to generate RWE.

1. Background on diseases and current approaches to treatment
2. Description of the device
3. Study-specific objectives
4. Target population and patient selection
5. Outcomes: primary, secondary, procedural, and device
6. Device exposure and outcome schedules
7. Study design including comparison treatments/devices, blinding, and control of confounders
8. Study procedures
9. Required sample size
10. Study registration
11. Monitoring plans
12. Statistical analysis plan
13. Results
14. Discussion and Conclusions
15. References
16. Note all changes from the original protocol within the final results and reporting

An important point is that the recommendations appearing in these guideline documents focus on how to report observational research. They are not meant to direct the design or conduct of such research, nor are they intended to serve as tools for evaluating the quality of observational research. However, clarity of reporting is essential to allow thorough evaluation of the quality of the research.

Presenting all the protocol details outlined in this Framework, and the corresponding sections of a study report, will be too much detail for most publications in most medical journals. Providing online supplements for articles would be one solution orfull study reports can be posted on a journal website. This detailed information can also be included in study registration sites.

15.2 References or Supporting Literature

1. Schulz KF, Altman DG, Moher D, for the CONSORT GroupCONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332 doi: 10.1136/bmj.c332.
2. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869 doi: 10.1136/bmj.c869.
3. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Link to the STROBE checklist: https://www.strobestatement.org/
4. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)statement: guidelines for reporting observational studies. Lancet. 2007 Oct 20;370(9596):1453-7. PMID:

180647.
5. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. PLoS Med 2007;4(10):e297. doi:10.1371/journal.pmed.0040297.
6. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Int J Surg. 2014 Dec;12(12):1500-24. doi: 10.1016/j.ijsu.2014.07.014. Epub 2014 Jul 18. PMID: 25046751.
7. Wang SV, Pinheiro S, Hua W, Arlett P, Uyama Y, Berlin JA, Bartels DB, Kahler KH, Bessette LG, Schneeweiss S. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. BMJ 2021;372:m4856.
8. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015 Oct 6;12(10):e1001885. doi: 10.1371/journal.pmed.1001885. PMID: 26440803; PMCID: PMC4595218.
9. Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network. https://www.equator-network.org. Accessed April 24, 2024.
10. US Food and Drug Administration Center for Devices and Radiological Health. Guidance for Industry and Food and Drug Administration Staff: Procedures for Handling Post-Approval Studies Imposed by Premarket Approval Application Order. October 7, 2022. https://www.fda.gov/media/71327/download.

16. FUTURE WORK

The methodological considerations presented include many of the basic components of a protocol for medical device evaluation and evidence generation in human subjects. In an era of rapid innovation in device technology, producing better, safer medical devices, the challenges of developing high quality evidence of benefit/risk and safety – particularly for high risk and permanently implantable devices – remain substantial.

Prominent in the recommendations in this document, and central to the NESTcc mission, is the rapidly evolving landscape of real world electronic health information systems. In particular, the progressive adoption of data structure standards and interoperable structure across these systems has truly opened the door to incorporating real world infrastructure with high quality RWD into the most informative and efficient means of developing device evidence ever in public health history. But while the door has opened, device research has only put its toe across the threshold – there are many operational gaps that remain between where we are and where most stakeholders in the device ecosystem would like to go with the efficiencies, costs and timelines of device research and development programs.

In this, the second such recommendations document from the NEST coordinating center, there is much that is new, but still a very heterogeneous admixture of classical research methodologies along with more novel approaches leveraging what is currently available in real world infrastructure and RWD to support evidence development. In this dynamic time, we here outline some of the directions of future work, which includes both changes in how future research may be conducted and potential future changes to policies influencing this work.

An important step in facilitating evidence generation will be adoption of a structured protocol template to harmonize RWE generation. Other resources, such as the HARmonized Protocol Template to Enhance Reproducibility (HARPER) developed by the joint ISPE and ISPOR task force, provide further recommendations about communicating scientific decisions through a common text, tabular, and visual structure.[1,2] As the HARPER template was primarily intended for real-world studies of medicinal products, its principles should be considered in light of any issues specific to studying medical devices, such as operator experience with a device.

A second important element will be to create and maintain a library of modules for RWE studies, including the definition of target populations, interventions, and outcome measure; including the information on data linkages. The library should maintain the validated coding algorithms based on ICD diagnosis and procedure codes, and CPT/HCPCS codes for common target populations and disease conditions. The library also should systematize the rapidly progressing field of clinical outcome assessments (COAs), including the commonly used instruments and their psychometric properties. COAs describe "how a person feels, functions, or survives and can be reported by a health care provider, a patient, a non-clinical observer (such as a parent), or through performance of an activity or task in the evaluation of medical devices".[3] COAs directly measure the impact on outcomes that are important to patients, families, and clinicians –and that therefore are also a priority for regulatory science. The use of COAs is growing due to the availability of digital devices to capture outcomes data electronically. Collection of COAs needs careful consideration of timing (preoperative, postoperative, and time intervals after procedures). Evaluation of COAs usually requires a comparison group. A simple pre-/post-comparison is generally difficult to interpret without the context of a control group. Additional considerations are development, selection, and adoption of instruments to collect and analyze COA data for interpretation.[4,5]

Another high priority is the development of reliable interoperability pathways, establishing data linkages across complementary electronic health systems to eliminate key data gaps for device evaluation and evidence generation. For example, developing coordinated registry networks (CRNs) has provided valuable registry data sources specifically styled for device evidence development. The SEER-Medicare Linkage has served to fill data gaps and generated insights in cancer research. Linking EHR data to claims data would strengthen long-term follow-up, linking EHR data to mortality data such as the National Death Index would help better capture the mortality endpoint, and linking EHR or claims data to registry data could help provide more detailed information on disease- or device-specific data. Data tokenization, an emerging area, replaces patient identifying information with encrypted tokens that are unique to each person; therefore enabling the linkage of patient data from various sources to generate a longitudinal view of a patient's journey in a fully de-identified manner.[6] Data tokenization links multiple sources of data while preserving the patient privacy. Given that implantable devices tend to be used for longer duration and that the risk may change over time due to wear and tear of materials, data linkages will support follow-up studies with long-term outcomes such as 5-year, 10-year, and 20-year risk and performance.

A key factor for success of RWE studies is to make it universal practice to enter UDIs into EHR systems, claims forms, and registries —a process that could potentially be integrated with direct digital tools requiring very little added human hands-on effort. UDIs and model numbers are

critical for unambiguous identification of device usage. However, UDIs are not yet widely adopted and available in the EHR systems and claims databases. The task is to promote the implementation of standard practice to make the device description, such as UDIs, model numbers, and brand names readily available in the EHR systems and claims databases. The NESTcc has established the UDI Center (,[7] which has rich information on this topic. A complementary document (A Playbook for Health System Unique Device Identifier Implementation at the Point of Care) has been published.[8]

Finally, areas for future research include emerging statistical methods, integration of clinical trials into routine clinical care and EHR systems and furthering our understanding of how RWE studies may predict clinical trial results.

Future work will focus on study designs and statistical methods that incorporate both randomized and observational approaches for balancing treatment arms[9,10] and construction of external controls for contextualizing single arm trials (Medical Device Innovation Consortium External Evidence Methods (EEM) Framework, 2021). More research is needed on statistical methods to detect and reduce potential bias in observational studies (e.g., the use of multiple negative controls to calibrate p-values[11]). These steps are needed because it is not always feasible to conduct a randomized trial. In some instances, there might be a small randomized comparator, but more power or longer follow-up are needed. It will be important to assess the robustness of conclusions drawn from RWE studies, in terms of whether they can provide valid information in addition to, or instead of, randomized trials. There are published comparisons between randomized trials and observational studies of the same topics. The DUPLICATE Demonstration Project[12-14] has generated insights into the relationship between RWE study potential for replication of randomized trials. To enhance the reproducibility and confidence in the credibility of evidence from studies using RWD, continued efforts are needed in deepening our understanding of why some RWE studies succeed while others fail to generate comparable results to randomized trials.[15] A promising approach is to emulate a hypothetical randomized target trial to evaluate the effect of the treatment of interest.[16] The target trial emulation specifies the key components of the protocol of the randomized target trial, including eligibility criteria, treatment strategies, treatment assignment procedures, follow-up period, outcome of interest, causal contrast(s) of interest (intention-to-treat effect, per-protocol effect), and analysis plan. Then the researchers emulate the randomized target trial using RWD. Numerous studies have demonstrated the ability of this target trial emulation approach to approximate the results from well conducted randomized controlled trials.[17]

Embedding elements of clinical trials, such as randomization, administration of study intervention, and data acquisition, into routine clinical care and EHR systems reduces duplication of trial and care activities and promotes the development of a learning health care system, where research will inform practice and practice will inform research. This can support better decision making, treatment options, and outcomes for patients. However, integrating interventional clinical trials into health care settings is challenging and complex, and operational direction is needed. Therefore, the Clinical Trials Transformation Initiative (CTTI) conducted in-depth interviews with study designers and implementers, gathered case studies, and created a set of draft recommendations to facilitate the fit-for-purpose integration of randomized, interventional trial elements into clinical care; including, but not limited to, trials of drugs,

devices, and biologics intended for regulatory review.[18]

Building prediction models using artificial intelligence or predictive analytics also poses challenges.[19,20] It is important to provide details on the model development and justify the model specifications, including training and validation sets for transparency and generalizability, and validate models broadly in multiple different datasets. For example, a recent external validation study questioned a widely used Epic sepsis prediction model in the U.S. hospitals.[21] This model was internally developed and validated by Epic using 405,000 patient encounter data across three health systems during 2013-2015; however, "only limited information is publicly available about the model's performance, and no independent validations have been published" prior to this external validation study.[21] Since the publication of this external validation study, Epic has revised this model and is now recommending the training of the model on a hospital's own data before its clinical use.[22]

In closing, priorities for future work are universal adoption of a structured protocol template to harmonize RWE generation; creation and maintenance of a library of modules for RWE studies, including the information on data linkages; and universal practice to report UDIs. Areas for future research are emerging statistical methods, integration of clinical trials into routine clinical care and EHR systems, and insight into how RWE studies may predict clinical trial results.

16.1 References or Supporting Literature

1. Wang SV, Pinheiro S, Hua W, Arlett P, Uyama Y, Berlin JA, Bartels DB, Kahler KH, Bessette LG, Schneeweiss S. STaRT-RWE: structured template for planning and reporting on the implementation of real-world evidence studies. BMJ 2021;372:m4856.
2. Wang SV, Pottegård A, Crown W, Arlett P, Ashcroft DM, Benchimol EI, Berger ML, Crane G, Goettsch W, Hua W, Kabadi S, Kern DM, Kurz X, Langan S, Nonaka T, Orsini L, Perez-Gutthann S, Pinheiro S, Pratt N, Schneeweiss S, Toussi M, Williams RJ. HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force. Pharmacoepidemiol Drug Saf. 2023 Jan;32(1):44-55. doi: 10.1002/pds.5507. Epub 2022 Oct 10. PMID: 36215113; PMCID: PMC9771861. https://onlinelibrary.wiley.com/doi/10.1002/pds.5507
3. US Food and Drug Administration Center for Devices and Radiological Health (CDRH) Patient Science and Engagement Program. Clinical Outcome Assessments (COAs) in Medical Device Decision Making. https://www.fda.gov/about-fda/cdrh-patient-science-and-engagement-program/clinical-outcome-assessments-coas-medical-device-decision-makinghttps://www.fda.gov/about-fda/cdrh-patient-science-and-engagement-program/clinical-outcome-assessments-coas-medical-device-decision-making
4. US Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research. Guidance for Industry and Food and Drug Administration Staff, And Other Stakeholders: Principles for Selecting, Developing, Modifying, and Adapting Patient-Reported Outcome Instruments for Use in Medical Device Evaluation. January 2022. https://www.fda.gov/media/141565/downloadhttps://www.fda.gov/media/141565/download
5. US Food and Drug Administration Center for Drug Evaluation and Research, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support

Labeling Claims. December 2009. https://www.fda.gov/media/77832/downloadhttps://www.fda.gov/media/77832/download
6. Bernstam EV, Applegate RJ, Yu A, Chaudhari D, Liu T, Coda A, Leshin J. Real-World Matching Performance of Deidentified Record-Linking Tokens. Appl Clin Inform 2022;13:865-73. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9474266/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9474266/
7. NEST UDI Center. https://nestcc.org/udi-center/. Accessed April 26, 2024.
8. NEST. A Playbook for Health System Unique Device Identifier Implementation at the Point of Care. https://nestcc.org/nestcc-udi-playbook. Accessed April 26, 2024.
9. Cafri G, Coplan P, Zhang S, Berlin JA. Selecting an Optimal Design for a Non-randomized Comparative Study: A Comment on "Some Considerations on Design and Analysis Plan on a Nonrandomized Comparative Study Utilizing Propensity Score Methodology for Medical Device Premarket Evaluation". Statistics in Biopharmaceutical Research. 2021:1-3.
10. Lu N, Xu Y, Yue LQ. Some Considerations on Design and Analysis Plan on a Nonrandomized Comparative Study Using Propensity Score Methodology for Medical Device Premarket Evaluation. Statistics in Biopharmaceutical Research. 2020;12(2):155-63.
11. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med 2014;33:209-18.
12. Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, Schneeweiss S. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. Clin Pharmacol Ther 2020;107:817-26.
13. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, Pawar A, Bessette LG, Lee H, Garry EM, Gautam N, Schneeweiss S. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. Circulation 2021;143:1002-13.
14. Wang SV, Schneeweiss S, Initiative R-D. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. JAMA. 2023;329(16):1376-1385. https://jamanetwork.com/journals/jama/article-abstract/2804067
15. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther 2017;102:924-33.
16. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016;183:758-64.
17. Gomes M, Latimer N, Soares M, Dias S, Baio G, Freemantle N, Dawoud D, Wailoo A, Grieve R. Target Trial Emulation for Transparent and Robust Estimation of Treatment Effects for Health Technology Assessment Using Real-World Data: Opportunities and Challenges. Pharmacoeconomics 2022;40:577-86.
18. CTTI
19. US Food and Drug Administration Center for Drug Evaluation and Research. Guidance for Industry and Food and Drug Administration Staff: Software as a Medical Device (SAMD): Clinical Evaluation. December 8, 2017. https://www.fda.gov/media/100714/downloadhttps://www.fda.gov/media/100714/download
20. US Food and Drug Administration Center for Drug Evaluation and Research. Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. January 2021. https://www.fda.gov/media/145022/downloadhttps://www.fda.gov/media/145022/download.
21. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, Pestrue J, Phillips M,

Konye J, Penoza C, Ghous M, Singh K. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med 2021;181:1065-70.

22. Epic overhauls popular sepsis algorithm criticized for faulty alarms. https://www.statnews.com/2022/10/03/epic-sepsis-algorithm-revamp-training/https://www.statnews.com/2022/10/03/epic-sepsis-algorithm-revamp-training/. Accessed April 26, 2024.

NEST

NATIONAL EVALUATION SYSTEM FOR HEALTH TECHNOLOGY

CONTACT INFORMATION

For more information, please contact NESTcc at nestcc@mdic.org

NEST

NATIONAL EVALUATION SYSTEM FOR HEALTH TECHNOLOGY

Learn more: www.nestcc.org

Phone: (202) 559-2938

Email: nestcc@mdic.org