Journal of Clinical Epidemiology 170 (2024) 111337

Journal of Clinical Epidemiology

COMMENTARY

Considerations for the use of biochemical laboratory registry data in clinical and public health research

Lasse M. Obel[a,b,*], Kasper Adelborg[c,d], Anton Pottegård[e], Henrik T. Sørensen[d], Mads Nybo[a,b]

[a]*Department of Clinical Biochemistry, Odense University Hospital, Odense, Denmark*
[b]*Department of Clinical Research, University of Southern Denmark, Odense Denmark*
[c]*Department of Clinical Biochemistry, Gødstrup Regional Hospital, Herning, Denmark*
[d]*Department of Clinical Epidemiology, Department of Clinical Medicine, Aarhus University, Aarhus University Hospital, Aarhus, Denmark*
[e]*Clinical Pharmacology, Pharmacy and Environmental Medicine, University of Southern Denmark, Odense, Denmark*

Accepted 25 March 2024; Published online 30 March 2024

Abstract

Objectives: To inform researchers of central considerations and limitations when applying biochemical laboratory-generated registry data in clinical and public health research.

Study Design and Setting: After review of literature on registry-based studies and the utilization of clinical laboratory registry data, relevant paragraphs and their applicability toward the creation of considerations for the use of biochemical registry data in research were evaluated. This led to the creation of an initial ten considerations. These were elaborated, edited, and merged after several read-throughs by all authors and discussed thoroughly under influence by the authors' personal experiences with laboratory databases and research registries in Denmark, leading to the formulation of five central considerations with corresponding items and illustrative examples.

Results: We recommend that the following considerations should be addressed in studies relying on biochemical laboratory-generated registry data: why are biochemical laboratory data relevant to examine the hypothesis, and how were the variable(s) utilized in the study? What were the primary indications for specimen collection in the study population of interest? Were there any pre-analytical circumstances that could influence the test results? Are data comparable between producing laboratories and within the single laboratory over time? Is the database representative in terms of completeness of study populations and key variables?

Conclusion: It is crucial to address key errors in laboratory registry data and acknowledge potential limitations.

Keywords: Registry-based; Guideline; Laboratory registry; Biochemical data; Considerations; Methodology; Laboratory data; Pitfalls

1. Introduction

Laboratory data covera wide variety of measurements and observations including biochemical test results, microbiological culture outcomes, histopathological findings, and radiological imaging reports. The ongoing automation of a wide variety of equipment and analyses has facilitated the laboratory production and thereby vastly increased the production of test results. This has led to storage of large amounts of laboratory-generated data in health-care registries with enormous potential in research settings.

Although reporting guidelines for randomized controlled trials, observational research, and diagnostic accuracy studies have been developed, no guidelines address the potential and important limitations when utilizing laboratory-generated registry data in research. It is therefore crucial to address these limitations as use of pre-existing laboratory-generated data is expanding rapidly, also among nonlaboratory researchers [1].

This commentary highlights the most common pitfalls when using biochemical laboratory-generated registry data in clinical and public health research and introduces five considerations for researchers to consult to improve quality, accuracy, and transparency of their studies.

2. Generation of laboratory data

Laboratory data are generated constantly during the daily practice at hospitals and general practitioners. Upon

All authors listed above take responsibility for the reliability and freedom from bias of data presented and interpreted in this paper.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

* Corresponding author. Department of Clinical Biochemistry, Odense University Hospital, J. B. Winsløwsvej 4, 5000 Odense C, Denmark.

E-mail address: Lasse.Mollegaard.Obel@rsyd.dk (L.M. Obel).

https://doi.org/10.1016/j.jclinepi.2024.111337

What is new?

Key findings

- Laboratory registry data represent unique data resources for research; However, without sufficient knowledge of inconsistencies in laboratory registries, studies utilizing these data may produce biased results and misleading conclusions.

What this adds to what is known?

- Although reporting guidelines for randomized controlled trials, observational research, and diagnostic accuracy studies have been developed, no guidelines address the potential and important limitations when utilizing laboratory-generated registry data in research.
- We introduce five central considerations for researchers who extract and analyze routine biochemical laboratory-generated data from registers, with the aim of improving the quality, accuracy, and transparency of future studies.

What is the implication and what should change now?

- We encourage researchers to consult the considerations in the planning phase of studies involving laboratory-generated data to ensure that the research question can be examined with high-quality data and with high internal validity. Further, we hope that journal editors and reviewers can use the considerations as a guide when evaluating papers relying on routine biochemical laboratory-generated data.

request from medical doctors, biological specimens are obtained and undergo analyses. Test results are reported through laboratory information systems for physicians to access and use. In some countries, these test results are transferred directly into research registries [1—4]. To ensure accurate biochemical laboratory test results, several internal and external quality control systems have been implemented in laboratories [5]. Thus, the overall quality of biochemical data must be considered high.

The generation of laboratory data can be divided into three phases: a preanalytical phase (patient preparation, specimen collection, transportation, and sample processing), an analytical phase (specimen testing), and a postanalytical phase (reporting and storage of test results) [6]. Each phase consists of multiple steps, and each step may introduce test-specific errors leading to a biased test result.

3. Considerations for biochemical laboratory-generated registry data in research

To identify relevant considerations and limitations when applying biochemical laboratory-generated registry data in clinical and public health research, the ‘European Medicines Agency’s Guideline on Registry-based Studies ‘(2021b, European Medicines Agency/426390/2021) and the ‘Danish Manual for Using Clinical Laboratory Information System Research Databases for Research Projects’ were reviewed for inspiration. Relevant paragraphs and their applicability toward the creation of considerations for the utilization of biochemical registry data were evaluated by L.M.O., K.A.D., and M.N., leading to the creation of an initial ten considerations. These were elaborated, edited, and merged after several read-throughs by all authors and discussed thoroughly under influence by the authors’ personal experiences with laboratory databases and research registries in Denmark, leading to the formulation of the presented five central considerations. Comments and relevant examples are provided for each consideration followed by summary key points.

In general, when a potentially significant bias is observed and the discrepancy cannot be explained using already available information and data, the specific laboratory producing the data should be identified and contacted for clarification. Also, a research group utilizing laboratory registry data could often benefit from the knowledge of a clinical biochemist in the planning phase of a study.

3.1. Consideration 1: why are biochemical laboratory data relevant to examine the hypothesis, and how were the variable(s) used in the study?

Biochemical laboratory registry data can be used in various ways in research, for example, to validate certain discharge diagnosis codes [7], to define a study population [8], for risk-stratification [9], as an exposure or covariate variable [10], or as an outcome parameter [11]. Laboratory test results can also be used to elucidate a trend in usage of a given test, which may provide important knowledge for health-care decision makers and department chairs, for example, to provide guidance for health-care resource planning, to optimize co-ordering patterns of analyses profiles, or to evaluate retesting intervals between analysis of different biomarkers [12,13].

In the planning phase of a research project, it should be assessed whether appropriate data are available to test a given hypothesis or to address a specific research question. Data acquired from laboratory-generated registers rarely come with a detailed description of the indication and how data were collected, analyzed, and recorded. When deciding if biochemical laboratory registry data should be used in a study, relevant considerations include whether the available data are the most suitable, or if other accessible data should be preferred depending on the intended use and overall research question. As an example, instead of using biomarker data as surrogate or proxy markers for medical treatment or to define specific diagnoses, it may be more accurate to rely on redeemed prescriptions or validated diagnosis codes.

When deciding how to use the laboratory registry data, it is important to consider challenges associated with the specific applied methodological approach. As an example, if a study population of interest is stratified according to a certain biomarker level, the most extreme observations are likely to become less extreme upon subsequent measurements, that is, the biomarker levels will be expected to regress toward the mean of the background population [14]. If regression toward the mean is not accounted for in the statistical analyses, for example, through covariance analysis, the decrease/increase of the biomarker could be wrongfully recognized as solely due to the investigated intervention. Also, if the researcher chose to dichotomize the variable, it is typically on the expense of doses-response analyses. Another consideration is defining the study's starting point; if relevant patients are excluded due to death caused by the investigated disease prior to the starting point, or if the outcome of interest cannot occur at the beginning of the study period, the study may be susceptible to immortal time bias [15].

Therefore, it is essential to clearly communicate to the reader how the laboratory test data was employed in the study, for what purpose(s), what time windows were used, and thoroughly discuss potential limitations. Similar as for other design considerations, the use of proper diagrams to convey the applied study design is recommended [16].

3.1.1. Key Points #1

When deciding if biochemical laboratory registry data should be applied in a study, relevant considerations include whether the available data are the most suitable or if other accessible data should be preferred. If laboratory registry data are utilized, it is crucial to consider challenges associated with the applied methodological approach, and it should be evident to the reader how and for what purpose(*s*) the laboratory data were used in the study.

3.2. Consideration 2: what were the primary indications for specimen collection in the study population of interest?

The underlying clinical reasoning for requesting the analyses of interest in the study population should be evaluated carefully: was the biomarker measured as part of routine testing, to monitor a patient in an acute or chronical setting, or as part of a diagnostic procedure? When extracting laboratory registry data for research purposes, the researcher risks to encounter confounding by indication, that is, the potential observed association between a given intervention and outcome is misleading due to the presence of an underlying factor related to the indication for specimen collection [17]. For example, most newly diagnosed cancer patients have platelet counts measured, but during the course of their disease, platelet counts are likely to be measured only if the patient is treated with chemotherapy, if a bleeding or thrombosis episode occurs, or if an oncology department measures platelet counts as part of its routine outpatient program. Platelet counts measured at these different scenarios represent different indications for sample collection and would in most cases not be comparable. Thus, to identify potential confounding by indication, it is important to consider possible trajectories of measurements in relation to the underlying reason for specimen collection in regard to the research question addressed — not to be confused with confounding by severity or selection bias [18].

Researchers are strongly encouraged to explore reasons for both having and not having a biomarker measured, and the interpretation of the research study should include such considerations. If the researcher encounters confounding by indication, proper matching between cases and controls, propensity score analysis, adjustment for relevant confounders, and sensitivity analyses, amongst others, are highly relevant to assure valid results.

3.2.1. Key points #2

To account for possible confounding by indication, it is important to consider possible trajectories of measurements in relation to the underlying clinical reason for specimen collection in the study population of interest. Researchers are strongly encouraged to explore reasons for having a biomarker measured, and the interpretation of the research study should include such considerations.

3.3. Consideration 3: were there any preanalytical circumstances that could influence the test results?

The researcher should consider if the biomarker in question could be affected by improper preanalytical conditions. Some biomarkers are more likely to be prone to preanalytical uncertainties than others.

Fasting (absence of food, physical exercise, and medication) prior to sample collection is indicated for several analyses, but unfortunately, fasting is in general poorly defined [19]. When relevant, a researcher should ensure that (any necessary) patient preparation was defined correctly. If information regarding preceding fasting is not provided, data must be evaluated carefully. Although separate nomenclature, properties, and units (NPUs, elaborated under consideration #4) codes exist for fasting measurements, access to medical records is often necessary to ensure fasting, which in most cases is not feasible. As an example, baseline prolactin values are regularly required prior to treatment with certain antipsychotic medications. As physiological parameters (including exercise and stress) can cause hyperprolactinemia, the patient should not perform heavy exercise (physical fasting) before sample collection.

Moreover, patients suspected of having drug-induced hyperprolactinemia should discontinue the medication for 3 days before retesting the prolactin level (medication fasting). However, based on pre-existing laboratory data, it is difficult to ensure that these demands were complied with.

Another important consideration is the time of specimen collection, especially considering drug monitoring. Time as a variable (hours; minutes; seconds) is frequently available in laboratory register data [1]. As an example, treatment with low molecular weight heparin is often monitored by anti-Xa measurements, but if the blood sample is not drawn 4 hours after administration, the recommended therapeutic indices cannot be applied. Additionally, some biomarker levels vary significantly depending on the time of the day for specimen collection, for example, growth hormone and cortisol levels peak during nighttime and in the morning, respectively. Thus, to evaluate and compare a given biomarker among patients included in a study, time of specimen collection may be of importance. When relevant for a specific biomarker, such considerations should be addressed and taken into account in the statistical analysis and interpretation of the results. In studies relying on drug monitoring, it should often be acknowledged as a limitation if the time from medication to specimen collection is not available.

Lastly, sample stability is crucial for many analytes and can be hampered by temperature, agitation, and time. Most laboratories have strict control of temperature and duration from sampling to analysis, but also sample transportation is a potentially relevant factor due to the increasing use of pneumatic tubes systems, which can cause hemolysis in the blood samples [20]. Some patient populations have higher risk of misleading test results caused by improper transportation, for example, pseudo-hyperkalemia in leukemic patients. Hence, if the biomarker of interest is fragile or if the laboratory data is derived from a patient population where misleading test results could be of significance, it should be considered if storage and transportation from patient to laboratory could influence test results.

3.3.1. Key points #3

Consider if the biomarker in question could be affected by improper pre-analytical circumstances including inadequate patient preparation, time from medication to specimen collection, normal physiological variance of the biomarker level during the day, and inappropriate sample transportation.

3.4. Consideration 4: are data comparable between producing laboratories and within the single laboratory over time?

Biochemical data can be recorded as text, continuous, dichotomous, or categorical variables. Typically, laboratory data registers include NPU codes [21]. The NPU system was established in 1960 and represents a standardized terminology when reporting laboratory test results. These codes are often used in research settings to identify relevant test results in a register. Most biochemical analyses are assigned a specific NPU code, but few biomarkers are still reported with unique national or local codes. In countries not operating with the NPU system codes, the Logical Observation Identifiers Names and Codes are typically implemented. In addition to NPU codes, laboratory data registers typically include patient, laboratory, and requester identification codes, specimen type (blood, urine, cerebrospinal fluid), time of sampling, components measured in the sample, and numeric and/or text results including reference values [1,22].

When merging biochemical data for research purposes, analyses principles, units, and report format of test results must be consistent and comparable over time, both within and between laboratories. It is important to note that NPU codes do not account for differences in equipment used or analyses principles. These are very likely to differ when data are extracted from more than one laboratory, but also within the same laboratory over time. As an example, various methods are used to estimate low-density lipoprotein cholesterol levels encompassing direct measurement and computation using the Friedewald or other equations. However, it is well documented that low-density lipoprotein levels vary significantly between these different methods [23,24]. Hence, the researcher should be careful when merging data from producing laboratories utilizing different analysis principles. Of note, food-fasting before low-density lipoprotein cholesterol level estimation are likely to differ significantly between fasting and non-fasting patients (Consideration #2) [25].

To assure comparability between laboratories and over time, it is essential to perform thorough preliminary data assessments focusing on differences in analyses results (mean/median), units, reference values, and outliers/clusters stratified by proper time intervals and by each laboratory from where data were obtained. When relevant, results from external quality control programs of the participating accredited laboratories could be evaluated to ensure that the analysis of interest remains comparable over time.

3.4.1. Key points #4

Analysis principles, units, and report format of test results must be consistent and comparable over time, both within a single laboratory and between all producing laboratories. It is crucial to perform thorough preliminary data assessments focusing on differences in analysis results, units, reference values, and outliers/clusters stratified by proper time intervals and by each laboratory from which data has been obtained.

3.5. Consideration 5: is the database representative in terms of completeness of study populations and key variables?

Information on key data must be recorded for the majority of the population enrolled in a study. Based on clinical and epidemiological knowledge, the researcher should attempt to quantify how many individuals are expected to have a specific biomarker of interest measured within the study's duration. If significantly fewer individuals were identified than expected, potential explanations may comprise use of wrong NPU codes for a required biomarker, differences in coding policies between laboratories, chronologically and/or geographically incomplete databases, or presentation of data as text with no quantitative results. In cases of incomplete laboratory registry data, the study may be left truncated, that is, individuals who have encountered the specific exposure or event of interest at the beginning of the study-period are not certain to be included in the registry [22].

It is equally important to be aware of inherent missing laboratory registry data. For example, measurements from point-of-care tests performed at the general practitioners (such as blood glucose, hemoglobin, and C-reactive protein) are rarely reported to registries and could therefore skew the eligible data in the laboratory registers due to a selected patient population with more severe disease, that is, those with measurements performed during hospitalization.

The above-mentioned pitfalls should be evaluated for potential impact on results from the study and included in the interpretation of the data. If the register is deemed incomplete, it is equally important to reconsider consideration number 2; why are the biomarkers of interest only available for some individuals.

3.5.1. *Key points #5*

Missing information on key laboratory registry variables must be held to a minimum in the target population. Potential biases include use of wrong identification codes or differences in coding policies between laboratories, chronologically and/or geographically incomplete databases, or presentation of data as text with no quantitative results in the registry. Also, inherent missing laboratory registry data should be considered.

4. Discussion

Biochemical laboratory data represent unique data resources for clinical and public health research. The five considerations introduced include items that are essential to address in studies relying on biochemical laboratory-generated registry data to ensure that study findings are presented in a transparent way. This will provide the reader with information allowing assessment of whether conclusions are supported by the data and methods applied. Although this report focuses on biochemical laboratory registry data, the listed considerations are likely to be applicable to studies utilizing any kind of laboratory data.

We encourage researchers to consult the considerations in the planning phase of studies involving laboratory-generated data to ensure that the research question can be examined with high-quality data and with high internal validity. Further, we hope that journal editors and reviewers can use the considerations as a guide when evaluating papers relying on routine biochemical laboratory-generated data.

Ethics statement

Not applicable/relevant to this report.

CRediT authorship contribution statement

Lasse M. Obel: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Kasper Adelborg:** Writing – original draft, Writing – review & editing, Conceptualization, Methodology. **Anton Pottegård:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Henrik T. Sørensen:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Mads Nybo:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors report no conflicts of interest regarding the study presented in this paper. K.A.D. is now an employee of Novo Nordisk A/S.

References

[1] Arendt JFH, Hansen AT, Ladefoged SA, Sorensen HT, Pedersen L, Adelborg K. Existing data sources in clinical epidemiology: laboratory information system databases in Denmark. Clin Epidemiol 2020;12:469–75.

[2] Kuiper JG, Bakker M, Penning-van Beest FJA, Herings RMC. Existing data sources for clinical epidemiology: the PHARMO database network. Clin Epidemiol 2020;12:415–22.

[3] Lee PC, Kao FY, Liang FW, Lee YC, Li ST, Lu TH. Existing data sources in clinical epidemiology: the taiwan national health insurance laboratory databases. Clin Epidemiol 2021;13:175–81.

[4] Kanazawa N, Tani T, Imai S, Horiguchi H, Fushimi K, Inoue N. Existing data sources for clinical epidemiology: database of the national hospital organization in Japan. Clin Epidemiol 2022;14: 689–98.

[5] Libeer JC. Role of external quality assurance schemes in assessing and improving quality in medical laboratories. Clin Chim Acta 2001;309(2):173–7.

[6] Plebani M, Laposata M, Lundberg GD. The brain-to-brain loop concept for laboratory testing 40 years after its introduction. Am J Clin Pathol 2011;136:829–33.

[7] Holland-Bill L, Christiansen CF, Ulrichsen SP, Ring T, Jorgensen JO, Sorensen HT. Validity of the international

classification of diseases, 10th revision discharge diagnosis codes for hyponatraemia in the Danish national registry of patients. BMJ Open 2014;4(4):e004956.

[8] Flaeng S, Nygaard S, Granfeldt A, Hvas AM, Sorensen HT, Thachil J, et al. Exploring the epidemiology of disseminated intravascular coagulation: protocol for the DANish disseminated intravascular coagulation (DANDIC) cohort study. BMJ Open 2022;12(7):e062623.

[9] Arendt JF, Pedersen L, Nexo E, Sørensen HT. Elevated plasma vitamin B12 levels as a marker for cancer: a population-based cohort study. J Natl Cancer Inst 2013;105:1799−805.

[10] Adelborg K, Nicolaisen SK, Hasvold P, Palaka E, Pedersen L, Thomsen RW. Predictors for repeated hyperkalemia and potassium trajectories in high-risk patients - a population-based cohort study. PLoS One 2019;14:e0218739.

[11] Thomsen RW, Nicolaisen SK, Hasvold P, Garcia-Sanchez R, Pedersen L, Adelborg K, et al. Elevated potassium levels in patients with congestive heart failure: occurrence, risk factors, and clinical outcomes: a Danish population-based cohort study. J Am Heart Assoc 2018;7(11):e008912.

[12] O'Sullivan JW, Stevens S, Hobbs FDR, Salisbury C, Little P, Goldacre B, et al. Temporal trends in use of tests in UK primary care, 2000-15: retrospective analysis of 250 million tests. BMJ 2018;363:k4666.

[13] Munk JK, Hansen MF, Buhl H, Lind BS, Bathum L, Jørgensen HL. The 10 most frequently requested blood tests in the Capital Region of Denmark, 2010−2019 and simulated effect of minimal retesting intervals. Clin Biochem 2022;100:55−9.

[14] Bland JM, Altman DG. Regression towards the mean. BMJ 1994; 308:1499.

[15] Yadav K, Lewis RJ. Immortal time bias in observational studies. JAMA 2021;325:686−7.

[16] Lund LC, Hallas J, Wang SV. Online tool to create publication ready graphical depictions of longitudinal study design implemented in healthcare databases. Pharmacoepidemiol Drug Saf 2021;30(7):982.

[17] Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. JAMA 2016;316:1818−9.

[18] Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. Am J Epidemiol 1999;149:981−3.

[19] Nybo M, Grinsted P, Jørgensen PE. Blood sampling: is fasting properly defined? Clin Chem 2005;51:1563−4.

[20] Nybo M, Lund ME, Titlestad K, Maegaard CU. Blood sample transportation by pneumatic transportation systems: a systematic literature review. Clin Chem 2018;64:782−90.

[21] Pontet F, Magdal Petersen U, Fuentes-Arderiu X, Nordin G, Bruunshuus I, Ihalainen J, et al. Clinical laboratory sciences data transmission: the NPU coding system. Stud Health Technol Inform 2009;150:265−9.

[22] Sørensen ST, Kristensen FP, Troelsen FS, Schmidt M, Sørensen HT. Health registries as research tools: a review of methodological key issues. Dan Med J 2023;70(4):A12220796.

[23] Maki KC, Grant JK, Orringer CE. LDL-C estimation: the perils of living with imperfection. J Am Coll Cardiol 2022;79:542−4.

[24] Mora S, Rifai N, Buring JE, Ridker PM. Comparison of LDL cholesterol concentrations by Friedewald calculation and direct measurement in relation to cardiovascular events in 27,331 women. Clin Chem 2009;55:888−94.

[25] Sathiyakumar V, Park J, Golozar A, Lazo M, Quispe R, Guallar E, et al. Fasting versus nonfasting and low-density lipoprotein cholesterol accuracy. Circulation 2018;137:10−9.