2017/03/07

バリデーションスタディーの実施方法に関するガイドラインについて

「日本における傷病名を中心とする
レセプト情報から得られる指標のバリデーション
に関するタスクフォース」

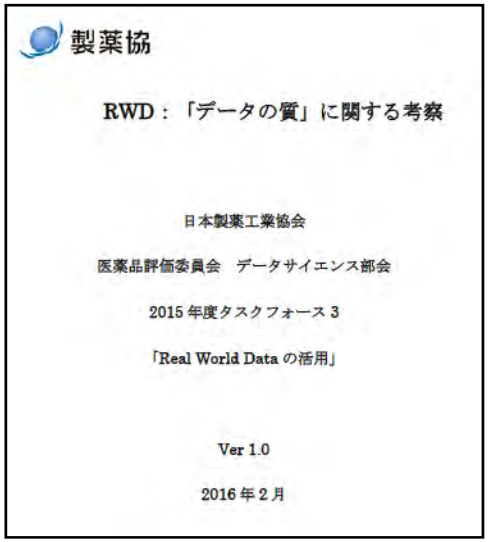
塩野義製薬　小林典弘

結論的には

- 久保田先生の結論と同様に、バリデーションスタディーの実施方法をメインにとりあげたガイドラインは見当たらなかったが、バリデーションに言及したガイドラインは存在した
  - Validity（研究そのもの、DB、変数など）については多くのガイドラインで言及されているが、そのValidityの確認方法については特定していないものが多い
  - 診断名の正確性に関する報告方法についての論文も存在した

主に確認した資料

2015年度の製薬協の活動にて、Real World Dataの「データの質」に関する報告書を作成する過程で収集した各国の薬剤疫学関係のガイドラインや論文（収集期間：2015/5-7月頃）から、バリデーションに関する記述について確認（DB研究に特化したGLに限らず）


製薬協

RWD：「データの質」に関する考察

日本製薬工業協会

医薬品評価委員会　データサイエンス部会

2015 年度タスクフォース 3

「Real World Data の活用」

Ver 1.0

2016 年 2 月

http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/rwd_qualityl.html

特に着目したガイドライン

グレーのものは久保田先生ご報告済みのもの

FDA

1. Guidance for Industry and FDA Staff: Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data

EMA

2. Guideline on good pharmacovigilance practices (GVP): Module VIII – Post-authorisation safety studies (Rev 1)
3. ENCePP: Guide on Methodological Standards in Pharmacoepidemiology (Revision 4)

PMDA

4. 医療情報のデータベース等を用いた医薬品の安全性評価における薬剤疫学研究の実施に関するガイドライン(初版)

その他

5. Guidelines for Good Database Selection and use in Pharmacoepidemiology Research
6. GPS – Good Practice in Secondary Data Analysis: Revision after Fundamental Reworking

前出の報告書ではこれらの記述しかしていないが、実際にはもっと多くのガイドラインをチェックしていた

EMA: GVP Module VIII
（2016/8にRev2に）

- Validation Studyの方法論について特別に言及しているものではないが、Validation Study自体については言及している

- VIII. Appendix 1. Methods for post-authorisation safety studies
  - App1.2. Data sources

- 研究に最適なデータソースを選択する上で考慮すべき事項の一つとして、アウトカムの正確性について言及

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf

EHRやAdministrative data由来のDB

（前略）In addition, these databases may not have the detailed and accurate information needed for some research, such as validated diagnostic information or laboratory data, and paper-based medical records should be consulted to ascertain and validate test results and medical diagnoses. Depending on the outcome of interest, the validation may require either a case-by-case approach or just the review of a random sample of cases. Other key aspects may require validation where appropriate. There are many databases in place for potential use in pharmacoepidemiological studies or in their validation phase.

Marketing authorisation holders should select the best data source according to validity (e.g. completeness of relevant information, possibility of outcome validation) and efficiency criteria (e.g. time span to provide results). External validity should also be taken into account.（後略）

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf

ENCePP: Guide on Methodological Standards in Pharmacoepidemiology

(2016/7にRev5に)

- 4.2.1.2. Assessment of outcomes
  - 全般的に、アウトカムを特定する方法（コードの選び方や複数のコードを使ったアルゴリズムなど）について述べており、そのvalidationについては議論していない
  - In some cases, initial plausibility checks or subsequent medical chart review will be necessary.
- 4.2.1.4. Validation
  - In healthcare databases, the correct assessment of drug exposure, outcome and covariate is crucial to the validity of research.
  - Completeness and validity of all variables used as exposure, outcomes, potential confounders and effect modifiers should be considered. Assumptions included in case definitions or other algorithms may need to be confirmed. For databases routinely used in research, documented validation of key variables may have been done previously by the data provider or other researchers. Any extrapolation of previous validation should, however, consider the effect of any differences in variables or analyses and subsequent changes to health care, procedures and coding. A full understanding of both the health care system and procedures that generated the data is required.

  といったことは述べているが、Validation Study自体の方法論については示していない

http://www.encepp.eu/standards_and_guidances/documents/ENCePPGuideofMethStandardsinPE_Rev5.pdf

ENCePP Checklist for Study Protocols

- ENCePP Checklist for Study Protocolsにも、validation studyに関する項目が含まれているが、方法論についての説明はなし

Section 6: Outcome definition and measurement	**Yes**	**No**	**N/A**	**Section Number**
6.1 Does the protocol specify the primary and secondary (if applicable) outcome(s) to be investigated?				
6.2 Does the protocol describe how the outcomes are defined and measured?				
6.3 Does the protocol address the validity of outcome measurement? (e.g. precision, accuracy, sensitivity, specificity, positive predictive value, prospective or retrospective ascertainment, use of validation sub-study)				
6.4 Does the protocol describe specific endpoints relevant for Health Technology Assessment? (e.g. HRQoL, QALYs, DALYS, health care services utilisation, burden of disease, disease management)				

http://www.encepp.eu/standards_and_guidances/documents/ENCePPChecklistforStudyProtocols.docx

Guidelines for Good Database Selection and use in Pharmacoepidemiology Research

- 基本的には、二次利用によるDB研究をする際の、研究に適したDBの特徴・性質について記したもので、validation studyについての言及はあるが、その方法については深く述べていない
- **5.5.1 External validation:** It may be possible to validate database records against external documents such as clinical notes or death certificate registries. Although this is expensive and time consuming, often necessitating a sample approach, it can be essential in certain settings as in claims data where 'working' rather than final diagnoses may be recorded. False negatives, as well as positives, should be considered. Where the electronic medical record is the source record, other validation methods should be used such as comparison of rates in the database with external figures, such as mortality statistics or meta-analyses from large clinical trials or review of the individual record for appropriate treatment/procedures. The positive predictive value of the algorithms/methods used to identify key outcomes can be assessed.
- ここでいう"External validation"とは、「外的妥当性の確認」の意味合いより、そのDB以外のデータソースと比較する確認方法全般を意図している

https://www.pharmacoepi.org/pub/1c2a306e-2354-d714-5127-9fd12e69fa66

ちなみに

- **5.5.3 Internal validation:** The interdependence of variables within a case may be examined. Do various items of information contradict each other, or does one variable highlight an omission elsewhere such as when there is an administrative record of death but no cause of death captured? It can be automatic and logical, such as testing that there has been no medical intervention after death. Often, simple logistic testing has to suffice as a proxy for data capture completeness, for example, where one would expect a series of tests or visits associated with certain conditions, or a diagnosis captured before prescription for a specific therapy.
- 「内的妥当性の確認」といった意味ではなく、DBに含まれる複数の項目の値を使ったロジカルチェック的な意味合い

- 5.5.2 はLogical checksとなっているが、こちらは単一のデータ項目のみで可能なロジカルチェック（Null check, Range checkなど）について記している

https://www.pharmacoepi.org/pub/1c2a306e-2354-d714-5127-9fd12e69fa66

GPS – Good Practice in Secondary Data Analysis: Revision after Fundamental Reworking

- 特に記載なし
- あえて挙げれば…
- Recommendation 6.4 – Data quality
  - The reliability and validity of the data used should be tested on the basis of available information. It is important to ensure external validation of critical features in the context of primary surveys e.g. for sub-populations.

http://dgepi.de/fileadmin/pdf/leitlinien/gps-version2-final_ENG.pdf

その他の資料での記載

Guidelines for Good Pharmacoepidemiology Practices (GPP)

- validation studyに関する言及はあるが、方法論についての記載はなし
  - The GPP are intended to apply broadly to all types of pharmacoepidemiologic research, including feasibility assessments, validation studies, descriptive studies, as well as etiologic investigations, and all of their related activities from design through publication.
  - Any procedures to be used to validate diagnosis should be described.
  - If data are validated, validation methods should mentioned,.e.g. review of ICD codes.

など

http://www.pharmacoepi.org/resources/guidelines_08027.cfm

他の資料での記載

A Checklist for Retrospective Database Studies—Report of the ISPOR Task Force on Retrospective Databases

- validationに対する意識はあるものの、validation studyについては言及なし（チェックリストの目的のためか？）
- Investigators attempting to identify group(s) of persons with a particular disorder (Alzheimer's disease) that has some diagnostic or coding uncertainty should provide a rationale and, when possible, cite evidence that a particular set of coding (ICD-9-CM, CPT-4, Drug Intervention) criteria are valid. Ideally, this evidence would take the form of validation against a primary source but more often will involve the citation of previous research. When there is controversial evidence or uncertainty about such definitions, the investigator should perform a sensitivity analysis using alternative definitions to examine the impact of these different ways of defining events.

https://www.ispor.org/workpaper/research_practices/A_Checklist_for_Retroactive_Database_Studies-Retrospective_Database_Studies.pdf

他の資料での記載
Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide

Validity of Key Data Definitions

Validity assessment of key data in an investigation is an important but sometimes overlooked issue in health care research using secondary data. There is a need to assess not only the general definition of key variables, but also their reliability and validity in the particular database chosen for the analysis. In some cases, particularly for data resources commonly used for research, other researchers or the organization may have validated outcomes of health events (e.g., heart attack, hospitalization, or mortality). Creating the best definitions for key variables may require the involvement of knowledgeable clinicians who might suggest that the occurrence of a specific procedure or a prescription would strengthen the specificity of a diagnosis. Knowing the validity of other key variables, such as race/ethnicity, within a specific dataset is essential, particularly if results will be described in these subgroups.

Ideally, validity is examined by comparing study data to additional or alternative records that represent a "gold standard," such as paper-based medical records. We described in the Administrative Data section above how validity of diagnoses associated with administrative claims might be assessed relative to paper-based records. EHRs and non–claims-based resources do not always allow for this type of assessment, but a more accommodating validation process has not yet been developed. When a patient's primary health care record is electronic, there may not be a paper trail to follow. Commonly, all activity is integrated into one record, so there is no additional documentation. On the other hand, if the data resource pulls information from a switch company (an organization that specializes in routing claims between the point of service and an insurance company), there may be no mechanism to find additional medical information for patients.

In those cases, the information included in the database is all that is available to researchers.

https://www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-Observational-CER-130113.pdf

他の資料での記載

Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide

Administrative Data

When using these claims for research purposes, the validity of the coding is of the highest importance. This is described in more detail below. The validity of codes for procedures exceeds the validity of diagnostic codes, as procedural billing is more closely tied to reimbursement. Understandably, the motivation for coding procedures correctly is high. For diagnosis codes, however, a diagnosis that is under evaluation (e.g., a medical visit or a test to "rule out" a particular condition) is indistinguishable from a diagnosis that has been confirmed. Consequently, researchers tend to look for sequences of diagnoses, or diagnoses followed by treatments appropriate for those diagnoses, in order to identify conditions of interest. Although Medicare requires an appropriate diagnosis code to accompany the procedure code to authorize payment, other insurers have looser requirements. There are few external motivators to code diagnoses with high precision, so the validity of these codes requires an understanding of the health insurance system's approach to documentation.

Investigators using claims data for CER should validate the key diagnostic and procedure codes in the study. There are many examples of validation studies in the literature upon which to pattern such a study.

Additional codes are available in some datasets - for example, the "present on admission" code that has been required for Medicare and Medicaid billing since October 2007 - which may help in further refinement of algorithms for identifying key exposures and outcomes.

https://www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-Observational-CER-130113.pdf

おまけ：前頁のmany examples の引用論文

- Segal JB, Powe NR. Accuracy of identification of patients with immune thrombocytopenic purpura through administrative records: a data validation study. Am J Hematol. 2004 January;75(1):12-7.
- Stein BD, Bautista A, Schumock GT, et al. The validity of ICD-9-CM diagnosis codes for identifying patients hospitalized for COPD exacerbations. Chest. 2012 Jan;141(1):87-93.
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3251268/pdf/110024.pdf
- Tollefson MK, Gettman MT, Karnes RJ, et al. Administrative data sets are inaccurate for assessing functional outcomes after radical prostatectomy. J Urol. 2011 May;185(5):1686-90.

（上記の論文で引用されてはいないが、こちらも細かく記載がされている）

- Training, Quality Assurance, and Assessment of Medical Record Abstraction in a Multisite Study
  - https://academic.oup.com/aje/article/157/6/546/74983/Training-Quality-Assurance-and-Assessment-of

https://www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-Observational-CER-130113.pdf

久保田先生からご報告済み

報告の観点からの資料における記載
The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement

GUIDELINES AND GUIDANCE

The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement

Eric I. Benchimol[1,2]*, Liam Smeeth[3], Astrid Guttmann[2,4], Katie Harron[3], David Moher[5], Irene Petersen[6], Henrik T. Sørensen[7], Erik von Elm[8‡], Sinéad M. Langan[3‡]*, RECORD Working Committee[¶]

RECORD
REporting of studies Conducted using Observational Routinely-collected Data

Home | Discussion | RECORD Group | Links | Members | Contact

Information

- News
- RECORD Checklist
- Publications
- Commentaries
- Endorsements
- Aims and Methods
- Consensus Meeting (Workshop)
- Acknowledgements
- Get Involved

What is RECORD?

REporting of studies **C**onducted using **O**bservational **R**outinely-collected **D**ata (**RECORD**) is an international collaborative which will develop reporting guidelines for studies conducted using routinely-collected health data (such as health administrative data, electronic medical record data, primary care surveillance data, and disease registries).

RECORD was developed with the input from stakeholders who use routinely-collected health data, ranging from health researchers, physicians, and journal editors, all of whom hold differing specializations across all aspects of health care.

As an extension of the existing STROBE guidelines (**ST**rengthening the **R**eporting of **OB**servational studies in **E**pidemiology), it is our overall goal to enhance transparency by providing researchers with the minimum reporting requirements needed to adequately convey the methods and results of their research.

http://www.record-statement.org/

久保田先生からご報告済み

報告の観点からの資料における記載
The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement

Methods (Participants)

- RECORD ITEM 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided.
- RECORD ITEM 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided.

Methods (Variables)

- RECORD ITEM 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these codes or algorithms cannot be reported, an explanation should be provided.

報告の観点からの資料における記載

Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data

ELSEVIER Journal of Clinical Epidemiology 64 (2011) 821–829

REVIEW ARTICLES

Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data

Eric I. Benchimol[a,b,c,d,e,f,g,*], Douglas G. Manuel[a,h,i,j,k], Teresa To[a,c,d], Anne M. Griffiths[b,e], Linda Rabeneck[a,d,l], Astrid Guttmann[a,d,e,m]

Abstract

Background and Objectives: Validation of health administrative data for identifying patients with different health states (diseases and conditions) is a research priority, but no guidelines exist for ensuring quality. We created reporting guidelines for studies validating administrative data identification algorithms and used them to assess the quality of reporting of validation studies in the literature.

Methods: Using Standards for Reporting of Diagnostic accuracy (STARD) criteria as a guide, we created a 40-item checklist of items with which identification accuracy studies should be reported. A systematic review identified studies that validated identification algorithms using administrative data. We used the checklist to assess the quality of reporting.

Results: In 271 included articles, goals and data sources were well reported but few reported four or more statistical estimates of accuracy (36.9%). In 65.9% of studies reporting positive predictive value (PPV)/negative predictive value (NPV), the prevalence of disease in the validation cohort was higher than in the administrative data, potentially falsely elevating predictive values. Subgroup accuracy (53.1%) and 95% confidence intervals for accuracy measures (35.8%) were also underreported.

Conclusions: The quality of studies validating health states in the administrative data varies, with significant deficits in reporting of markers of diagnostic accuracy, including the appropriate estimation of PPV and NPV. These omissions could lead to misclassification bias and incorrect estimation of incidence and health services utilization rates. Use of a reporting checklist, such as the one created for this study by modifying the STARD criteria, could improve the quality of reporting of validation studies, allowing for accurate application of algorithms, and interpretation of research using health administrative data.

Keywords: Health administrative data; Misclassification bias; Diagnostic accuracy; Sensitivity and specificity; Predictive values; Health services research; Epidemiology

「Health Administrative Dataを対象としたバリデーションスタディーの質」を評価する時の、報告ガイドラインの開発と使用

J Clin Epidemiol. 2011 Aug;64(8):821-9

Table 1
Data collection tool with extraction results reported in the appropriate columns as percentages

Checklist criteria	Yes (%)	No (%)	Uncertain	Not applicable (%)
Title, keywords, abstract				
1. Identifies article as study of assessing diagnostic accuracy?	94.1	5.9		
2. Identifies article as study of administrative data?	97.4	2.6		
Introduction				
3. States disease identification and validation as one of the goals of study?	93.4	6.6		
Methods				
Participants in validation cohort				
4. Describes validation cohort? (cohort of patients to which reference standard was applied)	98.9	1.1		
4a. Age?	49.1	50.6		
4b. Disease?	95.2	2.2		
4c. Severity?	17.3	48.0		34.7
4d. Location/jurisdiction?	90.8	9.2		
5. Describes recruitment procedure of validation cohort?	98.2	0.7		
5a. Inclusion criteria?	94.8	3.3		
5b. Exclusion criteria?	45.4	52.4		
6. Describes patient sampling? (random, consecutive, all, etc.)	91.5	7.4		
7. Describes data collection?	88.9	7.0		
7a. Who identified patients and ensured selection adhered to patient recruitment criteria?	74.2	14.0		10.7
7b. Who collected data?	64.6	22.5		12.2
7c. A priori data collection form?	59.0	5.2		14.4
7d. How was disease classified?	78.6	14.8		
8. Was there a split sample (i.e., revalidation using a separate cohort)?	11.4	88.2		
Test methods				
9. Describe number, training and expertise of persons reading reference standard?	46.1	23.6		29.5
10. If >1 person reading reference standard, is kappa quoted?	11.4	30.6	13.3	44.7
11. Were the readers of the reference (validation) test blinded to the results of the classification by administrative data for that patient? (e.g., Was the reviewer of the charts blinded to how that chart was billed?)	19.2	9.2	42.4	29.2
Statistical methods				
12. Describes methods of calculating/comparing diagnostic accuracy?	83.4	16.6		

J Clin Epidemiol. 2011 Aug;64(8):821-9

Results				
Participants				
13. Report when study done, start/end dates of enrollment	80.8	17.3		
14. Describe number of people who satisfied inclusion/exclusion criteria?	83.4	14.0		
15. Study flow diagram?	17.0	83.0		
Test results				
16. Reports distribution of disease severity?	19.2	46.5		34.3
17. Report cross-tabulation of index tests by results of reference standard	80.4	19.2		
Estimates				
18. Reports at least 4 estimates of diagnostic accuracy? (Estimates reported in included studies)	36.9	63.1		
18a. Sensitivity	67.2	32.8		
18b. Specificity	49.8	50.2		
18c. PPV	63.8	36.2		
18d. NPV	32.1	67.9		
18e. Likelihood ratios	3.3	96.7		
18f. Kappa	29.2	70.8		
18g. Area under the ROC curve/c-statistic	7.0	93.0		
18h. Accuracy/agreement	26.6	73.4		
19. Was the accuracy reported for any subgroup? (e.g., age, geography, different sexes, and so on)	53.1	46.9		
20. If PPV/NPV reported, does ratio of cases/controls of validation cohort approximate prevalence of condition in the population?	21.8	42.1		36.2
21. Reports 95 CIs for each of above?	35.8	63.8	0.4	
Discussion				
22. Discusses the applicability of the findings?	96.3	3.7		

Abbreviations: PPV, positive predictive value; NPV, negative predictive value; ROC, receiver operating characteristic; CI, confidence interval.

- STARDのチェックリストを参考に、著者のうちの4名で独自改変
- Administrative dataを対象に、患者を抽出するためのアルゴリズムをバリデーションスタディーした時の、診断名の正確性を記述した論文を集め、システマティックレビューを実施（271報）

J Clin Epidemiol. 2011 Aug;64(8):821-9