# The Validation Of Covid-19 information In The Pharmacoepidemiological Research Database for Public Health System by vaccination status

Oliver Astasio[1], Belén Castillo-Cano[2], Beatriz Sánchez Delgado[2], Rosa Gini[3], Fabio Riefolo[4], and Elisa Martin Merino[2]

[1]Instituto de Investigacion Sanitaria Hospital Clinico San Carlos
[2]Agencia Espanola de Medicamentos y Productos Sanitarios
[3]Agenzia Sanitaria Regionale
[4]Teamit Institute Partnerships Barcelona Health Hub

December 23, 2022

**Abstract**

**Purpose** To validate Covid-19 information records in The Pharmacoepidemiological Research Database for Public Health System (BIFAP), commonly used for pharmacoepidemiological research in Spain. **Methods** The recorded Covid-19 cases in primary care (PC) or positive test registries (gold-standard) were identified among vaccinated patients against SARS-CoV-2 infection of any age. They were matched with unvaccinated controls by birth year, vaccination date, region, and sex, between December 2020-October 2021. The sensitivity (SE), specificity (SP), positive (PPV), negative (NPV) predictive values, and date accurateness were estimated for PC by vaccination status and age brands. **Results** Among 21,702 patients with positive tests and 20,866 with recorded Covid-19 diagnoses, the SE, SP, PPV, and NPV were, respectively, 79.98%, 99.95%, 80.24% and 99.94% among vaccinated, and 78.67%, 99.96%, 84.51% and 99.94% among controls. For those aged [?]70 years old, SE (71.15-72.85%) was lower while PPV (84.68-88.04%) was higher compared to <70 years old participants. 94.12% of the total true positive cases (N=17,191) were recorded within ±5 days from the date of the test result. **Conclusions** PC Covid-19 diagnosis recorded in BIFAP showed high validation parameters. SE was similar and PPV was slightly lower among vaccinated than unvaccinated controls. Correction of vaccines effectiveness estimates by such misclassification is recommended. Data shows the influence of age. Among the elderly, Covid-19 diagnosis was less recorded but when recorded is more accurate than among younger patients. These findings permit the design of informed algorithms for performing Covid-19-related research.

## The Validation Of Covid-19 information In The Pharmacoepidemiological Research Database for Public Health System by vaccination status

Oliver Astasio [1], Belén Castillo-Cano[2], Beatriz Sánchez Delgado [2], Rosa Gini[3], Fabio Riefolo[4], Elisa Martín-Merino[2*]

1- Clinical Pharmacology Dept, Hospital Clínico San Carlos, IdISSC, Madrid (OA was affiliated to this hospital when the study was performed)

2-Pharmacoepidemiology and Pharmacovigilance Division. Spanish Agency of Medicines and Medical Devices (AEMPS). Madrid, Spain

3- Agenzia Regionale di Sanità (ARS), Italy

4- Teamit Institute, Partnerships, Barcelona Health Hub, Barcelona 08025, Spain

*Correspondence to:

Elisa Martín-Merino; Calle Campezo 1, Edif. 8, 28028, Madrid; Orcid: 0000-0002-3576-8605

Short title: Validation of Covid-19 information in primary care

**Abstract**

**Purpose**

To validate Covid-19 information records in The Pharmacoepidemiological Research Database for Public Health System (BIFAP), commonly used for pharmacoepidemiological research in Spain.

**Methods**

The recorded Covid-19 cases in primary care (PC) or positive test registries (gold-standard) were identified among vaccinated patients against SARS-CoV-2 infection of any age. They were matched with unvaccinated controls by birth year, vaccination date, region, and sex, between December 2020-October 2021. The sensitivity (SE), specificity (SP), positive (PPV), negative (NPV) predictive values, and date accurateness were estimated for PC by vaccination status and age brands.

**Results**

Among 21,702 patients with positive tests and 20,866 with recorded Covid-19 diagnoses, the SE, SP, PPV, and NPV were, respectively, 79.98%, 99.95%, 80.24% and 99.94% among vaccinated, and 78.67%, 99.96%, 84.51% and 99.94% among controls. For those aged [?]70 years old, SE (71.15-72.85%) was lower while PPV (84.68-88.04%) was higher compared to <70 years old participants. 94.12% of the total true positive cases (N=17,191) were recorded within +-5 days from the date of the test result.

**Conclusions**

PC Covid-19 diagnosis recorded in BIFAP showed high validation parameters. SE was similar and PPV was slightly lower among vaccinated than unvaccinated controls. Correction of vaccines effectiveness estimates by such misclassification is recommended. Data shows the influence of age. Among the elderly, Covid-19 diagnosis was less recorded but when recorded is more accurate than among younger patients. These findings permit the design of informed algorithms for performing Covid-19-related research.

**Keywords:** Covid-19; Primary Care; Validation; PPV; misclassification; measurement errors; electronic records.

**Key Points**

1. SARS-CoV-2 tests and vaccination information were rapidly linked to one of the more populated Spanish Primary Care (PC) databases (BIFAP) with the purpose to study the effectiveness and safety of Covid-19 vaccines.
2. Covid-19 diagnoses in PC showed high sensitivity to detect true infections (i.e. positive tests).
3. Specificity of Covid-19 diagnoses was very high.
4. Sensitivity was lower among [?]70 years old than younger patients, probably due to the differential healthcare settings, while PPV was higher.
5. Based on this validation, informed algorithms to detect true Covid-19 outcomes and adjustment of vaccine effectiveness could be developed and applied in future studies.

**Plain Language Summary (PLS)**

Is the Spanish collected primary care data about patients suffering from Covid-19 imaging the real pandemic situation in Spain? Patients' healthcare records are, in an anonymized form, used for different research purposes. Covid-19 data has been widely used to study pandemic and vaccination campaign effects, guiding authorities' decisions in this regard. Validating whether the recorded Covid-19 diagnoses reliably reflect the true positive laboratory tests is fundamental to trust the performed research outcomes. Herein, we

demonstrated that Covid-19 diagnoses in the Spanish public primary care records are truly associated with infection-positive tests, especially for patients >70 years old, and that most of the patients with positive tests also have a diagnosis of infection in primary care. Thus, the Spanish data on Covid-19 is a valid research tool.

## Purpose

The SARS-CoV-2 pandemic triggered the need to rapidly share patients-level data across different healthcare institutions, being them of vital importance to promptly monitoring pandemic setting evolution, as well as conditionally approved Covid-19 vaccines' safety and effectiveness in different world countries through real-world-data evidence. In Spain, several efforts have been invested among public healthcare institutions to merge patients' information through the creation of Common Data Models (CDM) in order to facilitate and guarantee timely pharmacoepidemiology research related to Covid-19 matters. To this extent, a clear example of the work performed in Spain is given by the Spanish Pharmacoepidemiological Research Database for Public Health System (BIFAP) database, a single integrated electronic health record (eHR) system, able to link and merge patient information from several Spanish regional data sources with different settings.

[1,2].

A Spanish royal decree regulates the epidemiological surveillance network by making mandatory the case reporting of specific diseases to the competent national authorities[3]. Covid-19 was a mandatory notifiable disease during the pandemic emergency. Since 2020, primary care eHRs directly gathered by BIFAP have been merged in a CDM with SARS-Cov-2 positive laboratory tests, and hospital and intensive care unit (ICU) admissions of external healthcare institutions. This pandemic data unification allowed the execution of different Covid-19 vaccination studies and the production of significant real-world data evidence during the last two years[4,5]. Thus, the eHR CDM creation has been crucial for studying and understanding Covid-19-related matters on the population, undoubtedly supporting important urgent national authorities' decisions about public health measures[6,7].

While significant advantages have been achieved by using the CDM strategy in terms of promptly available outcomes with large population sizes, further validation studies to quantify the risk of data bias due to case misclassification in the performed pharmacoepidemiology studies are needed[8]. Research using primary care (PC) databases requires practical definitions based on the information recorded to identify Covid-19 and, more in general, defining validation parameters would be a useful tool for correctly designing future studies. In the current work, we aimed to estimate and describe the validation parameters of the collected SARS-Cov-2 disease information among vaccinated patients and their unvaccinated controls in BIFAP.

## Methods

### Data sources and Covid-19 information

Patients' data from the Spanish public National Health System (SNS) data sources were linked and unified in the BIFAP one[9,10].

Data about Covid-19 diagnosis, birth year, sex, and Covid-19 vaccinations of around 13.7 million patients (7.4 of them aged [?]18 years) were obtained from the public PC source. The recorded episodes of **Covid-19 diagnosis** were identified through SNOMED (Systematized Nomenclature of Medicine) codes, as reported in Table 1. Covid-19 diagnosis codes were introduced in 2020 into the coding schemes used in BIFAP (i.e. the International Classification of Primary Care ICPC-2[11], the International Classification of Diseases ICD-9[12] and mapped to Snomed-CT).

**Positive test due to Covid-19 infections** (regardless of the symptoms, severity, or the result of the Covid-19 infection) were tracked from a Covid-19 registry on the date of the testing result. Infections might be confirmed through positive PCR, antigens, or any other confirmatory criteria established by clinical protocols whose definition is out of the scope of the current study. Herein, Covid-19 positive tests were the study gold standard.

BIFAP has been previously validated for research in pharmacoepidemiology, including the estimations of the precision of both several clinical outcomes [2,13] and vaccination records[14]. BIFAP is fully funded by the Spanish Agency on Medicines and Medical Devices (AEMPS), belonging to the public Department of Health, and is maintained with the collaboration of the participant Spanish regions.

The study protocol was approved by the BIFAP Scientific Committee (Reference Number 02_2021).

**Study design and Covid-19 case ascertainment**

A validation study of Covid-19-related data identified in two study cohorts (Covid-19 vaccinated or unvaccinated control individuals) was performed as designed in the study protocol[14]. In BIFAP, individuals of any age were included when vaccinated against Covid-19 (time0) during the study period, from 27 December 2020 to October 2021. The corresponding unvaccinated controls were matched 1:1 based on time0, birth year, sex, and region. All the study participants were free of prior SARS-CoV-2 infection.

In the study cohorts, the Covid-19 outcomes described above were identified during the study period (i.e. between time0 and the latest available data, death date, or study end date).

**Validation parameters estimation**

Using as gold standard the Covid-19 positive laboratory tests (main analysis), we estimated the sensitivity (SE), specificity (SP), positive (PPV), and negative (NPV) predictive values as well as the accuracy of the diagnosis date recorded by the PC physicians in the patients' clinical histories.

Parameters were estimated by vaccination status (i.e. vaccinated or control), age brand (<70 or [?]70 years old), and sex (female or male).

**Results**

Out of 3.80 million pairs of vaccinated and controls study participants, 21,702 had a positive test and 20,866 had a recorded Covid-19 episode (18,926 [90.7%] of them were recorded using two different Covid-19 diagnosis codes, see Table 1).

Table 2 shows the validation parameters of tracked Covid-19 cases stratified by vaccination status and age. Considering Covid-19 diagnosis codes, SE was similar among vaccinated (79.98%) and unvaccinated (78.67%) patients or among women (79.23%) and men (79.20%). However, differences appeared amongst age groups, i.e. SE ranged from 82.09 to 79.64% for younger subjects aged <70 years old and from 71.15 to 72.85% for older patients ([?]70 years old) among vaccinated and unvaccinated controls, respectively. PPV was lower among vaccinated (80.24%) than unvaccinated (84.51%) subjects and also lower among <70 years old (79.29%, vaccinated-73.95%, unvaccinated) than [?]70 years old (84.68%, vaccinated-88.04%, unvaccinated) individuals.

When recorded codes for suspected Covid-19 or contact with Covid-19 cases were included in the analyses, PPV decreased to 44.00% among vaccinated and to 57.62% among unvaccinated, while the other predictive values remained similar to their exclusion results (data not shown in tables). Regarding the accuracy of the Covid-19 diagnosis date records, 94.12% of true positive cases were recorded within 5 days (in a median value of 0 days) from the confirmatory positive laboratory test.

**Conclusions**

Overall, the recorded Covid-19 diagnoses in BIFAP PC eHR showed very high sensitivity in detecting confirmed SARS-CoV-2 infections and very high specificity to track non-real cases of the disease, both among vaccinated and their unvaccinated control group. The estimated predictive values suggest certain differential misclassification of the Covid-19 records and timing of infection when identified based on SNOMED codes in BIFAP or with laboratory positive tests. Quantifying such misclassification may be used to correct associated absolute (i.e. incidences) and relative risks (at least in unvaccinated vs vaccinated individuals). For instance, studies that aim to estimate Covid-19 vaccines' effectiveness can take advantage of those differential measurement errors.

On the other hand, we do consider not recommendable the inclusion of codes for suspected SARS-CoV-2 infection or contact with the virus in the definitions of Covid-19 outcomes. In fact, while SE values remained similar, those records' inclusion strongly decreased the PPV, especially among vaccinated individuals, increasing the probability to include unreal cases of SARS-Cov-2 infections. This misclassification may be due to frequent GP consultations of those individuals or other unknown reasons.

The validation parameter of Covid-19 cases in PC and its accuracy, herein provided, can be potentially used as a supportive design tool for outcome definitions in other studies. For example, or studies interested only in primary care consultations, when a decision should be taken over including only Covid-19 events linked to positive test results (to increase the PPV), or whether using Covid-19 diagnoses regardless of any associated positive laboratory test. This latter case may not include up to one-third (from 17.91 to 28.85% among vaccinated and unvaccinated) of individuals with Covid-19, especially for the elderly group ([?]70 years old). , Alternatively, for studies interested in all infection regardless of the setting, whether using both types of records or only positive laboratory tests.

Concerning age, PC records' SE for the detection of Covid-19 cases was lower among the oldest patients ([?]70 years old), especially those vaccinated, while PPV was higher in this group compared to <70 years old participants. The identified differences in SE across the different ages may be due to the tendency of [?]70 years old patients of seeking medical attendance directly at the hospital. Another point that should be taken into account is related to the eldest patients living in nursing homes. They receive in-house medical attention directly from the nursing homes' experts, thus, may not visit their GP to communicate the Covid-19 infection. Nursing homes' cases of Covid-19 are not collected by the BIFAP data source. Other cofactors that may justify the SE differences in identifying Covid-19 cases between the two age categories above/below 70 years old are, among others, the higher number of elders experiencing the infection during long stays in the hospital for other reasons or when receiving special care directly at their own home and may also die of Covid-19. These cases might not be correctly tracked by the BIFAP data sources and could explain the higher numbers of losses when compared to the <70 years old population.

Differently, our results suggest that if the Covid-19 diagnosis is recorded in the PC clinical registries, the PPV of those aged [?]70 years old is 5% and 14%, among vaccinated and unvaccinated, respectively, more accurate than the younger group. This difference could be led to different reasons such as more frequent testings of Covid-19 cases due to more clear infection symptoms in the eldest population. We also observed that the accuracy of the infection diagnosis date in BIFAP was also high since almost all Covid-19 positive laboratory tests have been recorded within 5 days in PC registries. This is of fundamental importance when time-window analyses are needed to evaluate if and when taking preventative measures and decisions, such as promoting large vaccination campaigns for specific age categories.

Finally, comparing our study with an already-published work on Covid-19 diagnosis validation carried out in the national medical product safety surveillance program funded by the Food and Drug Administration (FDA) in 2020, we can highlight comparable results. S. Kluberg et al.[5] showed that the PPV of Covid-19 diagnosis codes across all participating data sources was between 81.2-94.1% (variability depends on the considered time period), values almost close to our PPVs of 80.24% and 84.51% among vaccinated and unvaccinated, respectively, whereas the SE was reported to 94.4%, which is a higher value than our estimations of [?]79% in both vaccinated and unvaccinated groups. The differences in SE among the two works can be the result of our chosen study cohorts (which, in our case, have been selected according to the characteristics of the vaccinated patients and may not represent the entire BIFAP population), diverse healthcare settings (population-based versus claim data sources), or diverse healthcare systems, age, socioeconomic status or geographical areas of the covered populations, healthcare data recording habits, or virus epidemiology.

In the BIFAP data source, the tracked Covid-19 diagnoses in primary care records have high validation parameters with a low misclassification of their timing. Both Covid-19 vaccination status and old age of the patients influenced the recordings of infection diagnoses and the accuracy of their timing. Thus, the PPV in primary care should be a parameter to be taken into account in Covid-19 research studies. These findings reinforce the reliability of using the linked healthcare registries to BIFAP clinical histories as a source of

data for performing observational studies on SARS-CoV2 infection.

Electronic healthcare databases share common challenges, including the accurate identification of healthcare outcomes of interest for observational studies. Considering the evolving fundamental role of real-world data and healthcare databases, the validation process, to what this study contributes, is crucial for assuring the quality and accuracy of the produced evidence in pharmacoepidemiology studies.

**ACKNOWLEDGEMENTS**

This study is based on data from the "Pharmacoepidemiological Research Database for Public Heatlh System" (BIFAP) in Spain.

BIFAP is a public program for independent research financed by the Spanish Agency of Medicines and Medical Devices (AEMPS). The results, discussion and conclusions of this work are only of the authors and do not represent in any way the position of the AEMPS on this subject. The authors would like to acknowledge the excellent collaboration of the primary care physicians (general practitioners/paediatricians), and patients taking part the primary care records as well as the support from the regional health administrations providing BIFAP data.

**CONFLICT OF INTEREST**

Authors declare they do not have conflict of interest in the publication of this article.

**ETHICS STATEMENT**

The study protocol was approved by the Ethical Committee COMITE DE ETICA DE LA INVESTIGACION CON MEDICAMENTOS REGIONAL DE LA COMUNIDAD DE MADRID (CEIm-R) with the reference Number BIFAP_02_2021).

**References**

1. http://www.bifap.org/. Accessed September 1, 2022. http://www.bifap.org/

2. Macia-Martinez MA, Gil M, Huerta C, et al. Base de Datos para la Investigacion Farmacoepidemiologica en Atencion Primaria (BIFAP): A data resource for pharmacoepidemiology in Spain. *Pharmacoepidemiol Drug Saf* . 2020;29(10):1236-1245. doi:10.1002/pds.5006

3. Sanidad M de. *Boletin Oficial Del Estado Del 12 de Mayo de 2020* .; 2020. https://www.boe.es/boe/dias/2020/05/12/pdfs/BOE-A-2020-4933.pdf

4. Brown CA, Londhe AA, He F, et al. Development and Validation of Algorithms to Identify COVID-19 Patients Using a US Electronic Health Records Database : A Retrospective Cohort Study. 2022;(May):699-709.

5. Kluberg SA, Hou L, Dutcher SK, et al. Validation of diagnosis codes to identify hospitalized COVID-19 patients in health care claims data.*Pharmacoepidemiol Drug Saf* . 2022;31(4):476-480. doi:10.1002/pds.5401

6. Dodd C, Gardarsdottir H, Huerta C, Van Puijenbroek E, Kildemoes W.*Background Rates of Adverse Events of Special Interest for Monitoring COVID-19 Vaccines Title Background Rates of Adverse Events of Special Interest for Monitoring of COVID-19 Vaccines Protocol Version Identifier 1.1 Date of Last Version of Protocol* .; 2020.

7. Klungel O. *Title Evaluation of Assumptions of SCRI in Simulation Studies. Protocol Version Identifier 2.1 Date of Last Version of Protocol* .; 2021.

8. Seeger JD, Jonsson M, Layton JB, Clarke TC. Considerations of Misclassification and Confounding on COVID-19 Vaccines Effectiveness Studies - A Vaccine SIG Endorsed Symposium. In: *International Conference of PharmacoepidemiologyPharmacoepidemiology* . Vol 67. ; 2022.

9. Centro Nacional de Epidemiologia. COVID-19 en Espana. Published 2022. Accessed June 14, 2022. https://cnecovid.isciii.es/covid19/#evolucion-pandemia

10. Bosca JE, Cano J, Ferri J. *Covid-19 En Espana Durante 2021* .; 2022. Accessed June 14, 2022. https://documentos.fedea.net/pubs/dt/2022/dt2022-01.pdf

11. Oxford University Press. *ICPC-2. International Classification of Primary Care.* Second Edi.

12. World Health Organization. WHO IRIS: International Classification of Diseases. ninth revi. www.who.int/iris/handle/10665/39473

13. Martin-Merino E, Martin-Perez M, Castillo-Cano B, Montero-Corominas D. The recording and prevalence of Inflammatory bowel disease in girls' primary care medical Spanish records. *Pharmacoepidemiol Drug Saf* . 2020;29(11):1440-1449. doi:10.1002/pds.5107

14. Martin-Merino E. Real-world effectiveness of different COVID-19 vaccines in Spain: a cohort study based on public electronic health records (BIFAP).

Table 1. SNOMED description of Covid-19 diagnosis mapped to available ICPC/ICD-9 codes in primary care clinical histories and frequency of true positives found against test results.

| **SNOMED description** | **SNOMED codes** | **Freq.** | **Percent** |
|---|---|---|---|
| Coronavirus infection (disorder) | 186747009 | 10,249 | 49.12 |
| Disease caused by severe acute respiratory syndrome coronavirus 2 (disorder) | 840539006 | 8,677 | 41.58 |
| Diagnosis of COVID-19 infection confirmed by laboratory testing (disorder) | 63681000122103 | 1,740 | 8.34 |
| Pneumonia caused by Human coronavirus (disorder) | 713084008 | 107 | 0.51 |
| Pneumonia caused by severe acute respiratory syndrome coronavirus 2 (disorder) | 882784691000119100130840062 | 62 | 0.30 |
| Disease caused by Coronaviridae (disorder) | 27619001 | 20 | 0.10 |
| Polymerase chain reaction positive for severe acute respiratory syndrome coronavirus 2 (finding) | 62531000122108 | 7 | 0.03 |
| Asymptomatic severe acute respiratory syndrome coronavirus 2 infection (finding) | **189486241000119100** | 1 | 0.00 |

| **SNOMED description** | **SNOMED codes** | **Freq.** | **Percent** |
|---|---|---|---|
| Procedure for action related to case of disease due to SARS-CoV-2 (procedure) | 64121000122109 | 1 | 0.00 |
| Testing positive for IgG against SARS-CoV-2 (finding) | 64671000122103 | 1 | 0.00 |
| Outcome: case of COVID-19 still under follow-up (finding) | 63511000122107 | 1 | 0.00 |
| Positive result of rapid test for detection of IgM and IgG antibodies against SARS-CoV-2 in blood (finding) | 63621000122102 | 0 | - |
| Detection of severe acute respiratory syndrome coronavirus 2 (observable entity) | 871562009 | 0 | - |
| SARS-CoV-2 antigen testing positive (finding) | 64731000122108 | 0 | - |
| Secondary triage for severity level in patient with disease due to SARS-CoV-2 (procedure) | 64031000122106 | 0 | - |
| Diagnosis of COVID-19 infection confirmed by laboratory testing (disorder) | 63681000122103 | 0 | - |
| Detection of severe acute respiratory syndrome coronavirus 2 antigen (observable entity) | 871553007 | 0 | - |
| Positive serologic study for COVID-19 (finding) | 62951000122108 | 0 | - |
| Total | | 20,866 | 100.00 |

Table 2. Validation parameters of Covid-19 Codes recorded in primary care clinical histories using as gold-standard lab positive test.

| | N. Positive Covid test (gold-standard) | N. Covid Recorded in PC | N. in both sources (True positive) | N. recorded in PC without +test (% False positives) | Sensitivity of PC records | Specificity of PC records | PPV of PC records | NPV of PC records | M in ov po tes |
|---|---|---|---|---|---|---|---|---|---|
| **Vaccinated** | 10,439 | 10,381 | 8,330 | 2,051 (19.76%) | 79.98 | 99.95 | 80.24 | 99.94 | 20. |
| **<70** | 8,248 | 8,540 | 6,771 | 1,769 (20.71%) | 82.09 | 99.94 | 79.29 | 99.95 | 17. |
| [?]**70** | 2,191 | 1,841 | 1,559 | 282 (15.32%) | 71.15 | 99.97 | 84.68 | 99.93 | 28. |
| **Unvaccinated** | 11,263 | 10,485 | 8,861 | 1,624 (15.49%) | 78.67 | 99.96 | 84.51 | 99.94 | 21. |
| **<70** | 9,657 | 9,156 | 7,691 | 1,465 (16.00%) | 79.64 | 99.95 | 73.95 | 99.93 | 20. |
| [?]**70** | 1,606 | 1,329 | 1,170 | 159 (11.96%) | 72.85 | 99.98 | 88.04 | 99.95 | 27. |