The Validation Of COVID-19 Information In The Pharmacoepidemiological Research Database of Spain's Public Health System Data by Vaccination Status

Oliver Astasio, MD, PhD, Belén Castillo-Cano, MSc, Beatriz Sánchez Delgado, MSc, PharmD, Fabio Riefolo, PhD, Rosa Gini, PhD Elisa Martín-Merino, PhD, PharmD

Purpose
To validate COVID-19 information records in The Pharmacoepidemiological Research Database for Public Health System (BIFAP) of Spain.

Methods
The recorded COVID-19 cases in primary care or positive test registries (gold-standard) were identified among vaccinated patients against COVID-19 infection and their matched unvaccinated controls, between December 2020 and October 2021. The sensitivity, specificity, positive (PPV) and negative (NPV) predictive values were estimated for primary care records.

Results
Among 21,702 patients with positive tests and 20,866 with recorded COVID-19 diagnoses, the sensitivity, specificity, PPV and NPV were, respectively, 79.98 percent, 99.95 percent, 80.24 percent, and 99.94 percent among vaccinated, and 78.67 percent, 99.96 percent, 84.51 percent and 99.94 percent among controls.

Conclusions
Primary care COVID-19 diagnosis recorded in BIFAP showed that sensitivity was similar and PPV was slightly lower among vaccinated than unvaccinated controls. Among the elderly, COVID-19 diagnosis was less recorded. These findings permit the design of informed algorithms for performing COVID-19-related studies.

Keywords: COVID-19, primary care, validation, predictive values; misclassification, measurement errors, electronic health records, vaccination status

Key Points

1. Data on SARS-CoV-2 tests, vaccination and primary care (PC) consultations were rapidly unified in one of the most populated Spanish healthcare 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) that was lower among ≥70 years old than younger patients, probably influenced by the different healthcare settings.
3. PPV for COVID-19 diagnoses in PC was high and more predictive among unvaccinated and oldest people, probably due to be at-high risk of complications.
4. Specificity of COVID-19 diagnoses was very high.
5. This validation helps understand under- or over- estimations of associated vaccine effectiveness and develop informed algorithms to detect true COVID-19 outcomes in future studies.

Summary

Does the Spanish-collected primary care data about patients suffering from COVID-19 reflect 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 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.

Introduction

The SARS-CoV-2 pandemic triggered the need to rapidly share patient-level data across different healthcare institutions, giving them vital importance to promptly monitor pandemic-setting evolution, as well as conditionally approve 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 (Base de datos para la Investigación Farmacoepidemiológica en el Ámbito Público or 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].

A Spanish royal decree regulates the epidemiological surveillance network by making mandatory the case reporting of specific diseases to national authorities. COVID-19 was a mandatory notifiable disease during the pandemic. 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.

The pandemic data unification allowed the execution of different COVID-19 vaccination studies and the production of significant real-world data evidence during the last years[2,3]. 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[4,5,6]

COVID-19 information linked from different data sources may not always overlap and must be evaluated for identification of true cases for research. The data regarding COVID-19 diagnosis in some sources[3] have a positive predictive value (PPV) between 81percent and 94 percent of the true cases depending on the calendar period, whereas there was a sensitivity of 94.4 percent among all episodes. The implication of this could be substantial. For instance, if PPV were different between vaccinated- and unvaccinated-compared groups, the estimations of vaccines effectiveness would be confounded.

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[7]. Research using primary care (PC) databases required 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 study, 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 BIFAP[1]

- Data about COVID-19 diagnosis, birth year, sex, and COVID-19 vaccination of around 13.7 million patients (7.4 million of them aged ≥18 years) were obtained from the public PC source for four geographical regions (Aragón, Asturias, Castilla y León, and Murcia). The recorded episodes of COVID-19 diagnosis were identified through SNOMED (Systematized Nomenclature of Medicine) codes, as reported in Table 1. SNOMED codes were mapped to COVID-19 diagnosis codes that were introduced in 2020 into the International Classification of Primary Care ICPC-2[8] and the International Classification of Diseases ICD-9[9] used in PC settings.

- Positive test due to COVID-19 infections were tracked from a COVID-19 registry linked to PC data 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 gold standard.

BIFAP has been previously validated for research in pharmacoepidemiology, including the estimations of the precision of both, clinical outcomes[10,11] and vaccination records[12]. 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 (3.805.279 COVID-19 vaccinated and unvaccinated control individuals) was performed as designed in the study protocol[12]. In summary, individuals of any age were included when vaccinated against COVID-19 (time0) during the study period, from December, 27 2020 to October, 31 2021. The corresponding unvaccinated controls were matched 1:1 based on the date of the first vaccination of the vaccinated pair, birth year, sex, and region. All the study participants were free of prior SARS-CoV-2 infection. Follow-up was until the end of the study period (October, 31 2021) or until diagnosis of COVID-19.

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).

Statistical Analysis

Using as gold standard the COVID-19 positive laboratory tests (main analysis), we estimated the sensitivity, specificity, 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 band (<70 or ≥70 years old), and sex (female or male). The results of the study were calculated using STATA v.16.1.

Results

Out of 3.80 million pairs of vaccinated and controls study participants (mean age: 53.4 years), 21,702 had a positive test and 20,866 had a recorded COVID-19 episode (18,926 [90.7 percent percent] of them were recorded using two 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, sensitivity was similar among vaccinated (79.8 percent) and unvaccinated (78.7 percent) patients or among women (79.2 percent) and men (79.2 percent). However, differences appeared amongst age groups, i.e. sensitivity ranged from 82.1 percent to 79.6 percent for subjects aged <70 years old and from 71.2 percent to 72.9 percent for older patients (≥70 years old) among vaccinated and unvaccinated controls, respectively. PPV was lower among vaccinated (80.2 percent) than unvaccinated (84.5 percent) subjects and also lower among <70 years old (79.3 percent, vaccinated-84.0 percent, unvaccinated) than ≥70 years old (84.7 percent, vaccinated-88.0 percent, unvaccinated) individuals. Specificity was ≥99.94 percent over all groups.

When recorded codes for suspected COVID-19 or contact with COVID-19 cases were included in the analyses, PPV decreased to 44.0 percent among vaccinated and to 57.6 percent 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 records, COVID-19 of true positive cases were recorded within five days (in a median value of zero days) from the confirmatory positive laboratory test.

Conclusions

During the fourth and fifth SARS-CoV-2 epidemiological waves with incidences ranging between 21 (October 2021) and 800 (August 2021) cases per 100,000 inhabitants in Spain in 14 days as reported by the public institutions[13] , the recorded COVID-19 diagnoses in BIFAP PC EHRs showed high sensitivity in detecting confirmed SARS-CoV-2 infections and very high specificity to track non-cases of the disease, both among vaccinated and their unvaccinated control group. The estimated predictive values suggested 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 permits to understand potential under- or over-estimations in the associated absolute (i.e. incidences; considering that up-to 30 percent of cases could be missed if only primary care diagnosis are collected) and relative risks (at least in unvaccinated vs vaccinated individuals, considering that confirmation seems slightly different among them) of COVID-19 episodes.

On the other hand, we do not recommend 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 sensitivity values remained similar, those records' inclusion strongly decreased the PPV, especially among vaccinated individuals, increasing the probability to include misdiagnosed cases of SARS-CoV-2 infections. This misclassification may be due to frequent PC physician 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, in studies interested only in PC 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 17.9 percent to 28.8 percent 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 i.e., people with a positive test and/or a clinical diagnosis (given challenges in accessing testing and/or primary care during the pandemic) or only positive laboratory tests.

Concerning age, PC records' sensitivity 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 sensitivity 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 patients living in nursing homes. They receive in-house medical attention directly from the nursing homes' experts, thus, may not visit their PC physician to communicate the COVID-19 infection. Nursing homes' cases of COVID-19 are not systematically collected by the BIFAP data source. Other cofactors that may justify the sensitivity 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 percent and 14 percent, among vaccinated and unvaccinated, respectively, more accurate than the younger group. This variation could be led to different reasons such as more frequent testing 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 test have been recorded within five 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. The study[3] showed that the PPV of COVID-19 diagnoses codes across all participating data sources was between 81.2 percent and94.1 percent (variability depends on the considered time period), values almost close to our PPVs of 80.2 percent and 84.5 percent among vaccinated and unvaccinated, respectively, whereas the sensitivity was reported to 94.4 percent, which is a higher value than our estimations of ≈79 percent in both vaccinated and unvaccinated groups. The differences in sensitivity 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. Thus, the parameters observed in our study may mainly be used to interpret studies performed in the same data source and period and may not be generalisable to other contexts or settings.

Some limitations must be acknowledged.

Race, ethnicity and other demographic characteristics potentially associated with unequal burden of COVID-19 were not available to assess any differential parameters among them.

In the BIFAP data source, the tracked COVID-19 diagnoses in PC 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 PC 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-CoV-2 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 Comité de ética de la investigación con medicamentos regional de la Comunidad de Madrid (CEIm-R) with the reference Number BIFAP_02_2021.

Authors

Oliver Astasio, MD, PhD, is a physician in clinical pharmacology and PhD in Biomedical Investigation at Complutense University in Madrid. At the time of the study, Astasio was affiliated with the clinical pharmacology department at the Hospital Clínico San Carlos' Health Research Institute in Madrid and external expert at the Spanish Agency of Medicines and Medical Devices. At this time, he is medical advisor in Novartis pharmaceutical for haematology diseases.

Belén Castillo-Cano, **MSc,** is working in the pharmacoepidemiology and pharmacovigilance division at the Spanish Agency of Medicines and Medical Devices in Madrid. She is collaborating as a junior biostatistician on different projects with the BIFAP team. At this time, she is studying for a PhD in technology in the department of computer science, applied mathematics and statistics at Girona University. She has a Bachelor's degree of mathematics at the University of Almería and a Master's degree of statistics at the University of Granada.

Beatriz Sanchez-Delgado, mailto:is a pharmaco-epidemiologist in BIFAP at the Spanish Agency of Medicines and Medical Devices in Spain. She has a bachelor of pharmacy from the University of Salamanca, Spain and a Masters degree on both

International Public Health (Queen Margaret University in Edimburgh, UK) and pharmacoepidemiology and pharmacovigilance (Alcala University in Madrid, Spain)

Fabio Riefolo, **PhD**, worked on the development of cholinergic nervous system drugs at the University of Milan (Italy) and Wuerzburg (Germany) during his Master's thesis in pharmaceutical chemistry & technology. He obtained a Ph.D. in medicinal chemistry at the Institute for Bioengineering of Catalonia in Barcelona (Spain), working on cardiovascular diseases and neurological disorders and was also a researcher for the Biomedical Research Networking Centre in Bioengineering, Biomaterials, and Nanomedicine (CIBER-BBN). From a post-doc position at IBEC, he moved to Teamit (Barcelona, Spain) as scientific study manager and regulatory science advisor, expanding his knowledge of medicines development and their regulatory roadmap, from preclinical to clinical regulation, post-marketing authorization studies based on real-world-evidence (participation in several HMA-EMA-registered studies), medical device regulation, and working in various healthcare-related public European proposals.

Rosa Gini, is a data scientist focused on secondary use of EHRs for pharmacoepidemiology, epidemiology and health services research. Her specific interest is in developing culture and tools for accurate, reliable, transparent, and fast generation of evidence to support health policy making at a national, European, and international level. In ARS Toscana, the Regional Agency for Public Health of Tuscany, she is the head of the pharmacoepidemiology unit and conducts methodological studies, providing expertise for studies using real-world evidence on the use and safety of medicines and vaccines on an international distributed networks of databases.

Elisa Martín-Merino, PhD, PharmD (emartinm@aemps.es), is a senior pharmacoepidemiologist at the Spanish Agency of Medicines and Medical Devices in Madrid, Spain. She earned her PhD in preventive medicine and public health, with her doctoral research focusing on assessing the risk of acute coronary syndrome associated with the use of non-steroidal anti-inflammatory drugs in a field study. Martín-Merino has actively contributed to pharmacoepidemiological research studies aimed at evaluating potential adverse reactions to medications used by individuals in real-world settings- outside the controlled context of clinical trials. Additionally, she is interested in studying the precision of electronic health records for research on medication use and its effects.

References

1. “BIFAP Base de Datos Para La Investigación Farmacoepidemiológica En El Ámbito Público.”. Accessed January 28, 2021. http://bifap.aemps.es/.

2. Brown CA, Londhe AA, He F, Cheng A, Ma J, Zhang J, Brooks CG, et al. 2022. “Development and Validation of Algorithms to Identify COVID-19 Patients Using a US Electronic Health Records Database: A Retrospective Cohort Study,” no. May: 699–709.

3. Kluberg SA, Hou L, Dutcher SK, Billings M, Kit B, Toh S, Dublin S,, et al. 2022. “Validation of Diagnosis Codes to Identify Hospitalized COVID-19 Patients in Health Care Claims Data.” Pharmacoepidemiology and Drug Safety 31 (4): 476–80. https://doi.org/10.1002/pds.5401.

4. Bots SH, Riera-Arnau J, Belitser SV, Messina D, Aragón M, Alsina E, Douglas IJ, et al. 2022. “Myocarditis and Pericarditis Associated with SARS-CoV-2 Vaccines: A Population-Based Descriptive Cohort and a Nested Self-Controlled Risk Interval Study Using Electronic Health Care Data from Four European Countries.” Frontiers in Pharmacology 13: 1038043.

5. Willame C, Dodd C, Durán CE, Elbers R, Gini R, Bartolini C, Paoletti O, et al. 2023. “Background Rates of 41 Adverse Events of Special Interest for COVID-19 Vaccines in 10 European Healthcare Databases - an ACCESS Cohort Study.” Vaccine 41 (1): 251–62. https://doi.org/10.1016/j.vaccine.2022.11.031.

6. Riefolo F, Castillo-Cano B, Martín-Pérez M, Messina D, Elbers R, Brink-Kwakkel D, Villalobos F, et al. 2023. “Effectiveness of Homologous/Heterologous Booster COVID-19 Vaccination Schedules against Severe Illness in General Population and Clinical Subgroups in Three European Countries.” Vaccine 41 (47): 7007–18. https://doi.org/10.1016/j.vaccine.2023.10.011.

7. Seeger, JD, Jonsson, M, Layton, JB and Clarke, TC. 2022. “Considerations of Misclassification and Confounding on COVID-19 Vaccines Effectiveness Studies - A Vaccine SIG Endorsed Symposium.” In International Conference of Pharmacoepidemiology. Pharmacoepidemiology. 2022. Vol. 67.

8. Oxford University Press. 1998. “ICPC-2. International Classification of Primary Care.” Second Edi.

9. “World Health Organization. WHO IRIS: International Classification of Diseases: [9th] Ninth Revision, Basic Tabulation List with Alphabetic Index.” 1978. 1978. http://www.who.int/iris/handle/10665/39473.

10. Martín-Merino E, Martín-Pérez M, Castillo-Cano B, Montero-Corominas D. 2020. “The Recording and Prevalence of Inflammatory Bowel Disease in Girls’ Primary Care

Medical Spanish Records." Pharmacoepidemiology and Drug Safety 29 (11): 1440–49. https://doi.org/10.1002/pds.5107.

11. Maciá-Martínez MA, Gil M, Huerta C, Martín-Merino E, Álvarez A, Bryant V, Montero D; BIFAP Team. 2020. "Base de Datos Para La Investigación Farmacoepidemiológica En Atención Primaria (BIFAP): A Data Resource for Pharmacoepidemiology in Spain." Pharmacoepidemiology and Drug Safety 29 (10): 1236–45. https://doi.org/10.1002/pds.5006.

12. Martin-Merino, E, Seco-Meseguer, E, Castillo-Cano, B, Limia-Sanchez, A, Olmedo-Lucerón, C, Monge-Corella, S, and Larrauri, A. 2021. "Real-World Effectiveness of Different COVID-19 Vaccines in Spain: A Cohort Study Based on Public Electronic Health Records (BIFAP)." Spain. EU PAS Register (study EUPAS42668)

13. "Instituto de Salud Carlos III. Informes COVID-19. Informe No 103. Situación de COVID-19 En España a 3 de Noviembre de 2021.". Accessed April 6, 2023. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/Informes%20COVID-19/INFORMES%20COVID-19%202021/Informe%20nº%20103%20Situación%20de%20COVID-19%20en%20España%20a%203%20de%20noviembre%20de%202021.pdf

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 SARS-CoV-2 lab positive test.

SNOMED description	**SNOMED codes**	**Frequency**	**Percentage**
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)	882784691000119100130840008	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
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 SARS-CoV-2 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)	N. Positive test without PC record	Sensitivity of PC records	Specificity of PC records	PPV of PC records	NPV of PC records	Missing in PC overall positive test (%)
Vaccinated	10,439	10,381	8,330	2,051 (19.76%)	2,109	79.80	99.95	80.24	99.94	20.20
<70	8,248	8,540	6,771	1,769 (20.71%)	1,477	82.09	99.94	79.29	99.95	17.91
≥70	2,191	1,841	1,559	282 (15.32%)	632	71.15	99.97	84.68	99.93	28.85
Unvaccinated	11,263	10,485	8,861	1,624 (15.49%)	2,402	78.67	99.96	84.51	99.94	21.33
<70	9,657	9,156	7,691	1,465 (16.00%)	1,966	79.64	99.95	84.00	99.93	20.36
≥70	1,606	1,329	1,170	159 (11.96%)	436	72.85	99.98	88.04	99.95	27.15