ISSN 2593-9157

An open-access journal of population research

REVUE Quetelet JOURNAL

Volume 11/1
2023

COVID-19 Mortality Statistics: A Comparative Study of Epidemiological Surveillance Data and Death Certificates in 2020 in Belgium

Catharina Vernemmen*[1], Reindert Ekelson[1], Jure Jurčević[1], Serge Nganda[1], Aline Scohy[1], Sara Dequeker[1], Koen Blot[1], Muhammet Savsin[2], Marie-Isabelle Joris[3], Hanna Masson[4], Mathieu Roelants[5], Peter Verduyckt[6], Elias Neirynck[7], Natalia Bustos Sierra[1]

[1] Department of Epidemiology and Public Health, Sciensano, 1050 Brussels, Belgium.
[2] Direction de l'Evaluation et de la Prospective, Agence pour une Vie de Qualité (AViQ), 6061 Charleroi, Belgium.
[3] Direction Promotion de la Santé, Prévention et Surveillance des Maladies, Bien-être et Santé, Agence pour une Vie de Qualité (AViQ), 6061 Charleroi, Belgium.
[4] Department of Care, Infection Prevention and Control, Government of Flanders, 1030 Brussels, Belgium.
[5] Department of Care, Policy Information and Data, Government of Flanders, 1030 Brussels, Belgium.
[6] Brussels-Capital Health and Social Observatory, 1040 Brussels, Belgium.
[7] Statistics Belgium (Statbel), 1000 Brussels, Belgium.

* corresponding author: Catharina.Vernemmen@sciensano.be

DEMO Centre de recherche en démographie
IACCHOS Institut d'analyse du changement dans l'histoire et les sociétés contemporaines

UCLouvain

DOI.10.14428/rqj2023.11.01.01 • ©2023 Catharina Vernemmen *et al.*

PUL PRESSES UNIVERSITAIRES DE LOUVAIN

Abstract – Résumé

Abstract

The COVID-19 pandemic necessitated the rapid establishment of the COVID-19 mortality epidemiological surveillance database (SURV) in Belgium due to a significant delay in availability of the cause of death database (COD). Understanding differences and limitations in both databases is crucial for contextualising COVID-19 mortality statistics. This study assesses SURV's data quality, raises awareness of differences and limitations in both databases, and proposes recommendations for future pandemic mortality surveillance. SURV and COD were linked probabilistically to explore overall coverages and discrepancies. Factors such as region and place of death, case classification, epidemic wave, age group, sex, and number of conditions, were analysed using logistic regression models. SURV identified 90% (n=19,801) of COVID-19-related deaths from COD (n=22,015). Coverage was higher in hospitals (98%, n=11,130 in SURV, n=11,335 in COD) and long-term care facilities (90%, n=8,602 in SURV, n=9,580 in COD) compared to deaths at home (5%, n=52 in SURV, n=1,057 in COD). However, 83.9% of SURV records listed COVID-19 as the underlying cause of death in COD, and 75.4% of COVID-19 deaths in COD were identified in SURV. Reduced COVID-19 activity and diagnostic uncertainty resulted in lower agreement between databases. Variations in data quality were observed across epidemic waves, regions, and healthcare facilities. In addition to reaching real-time objectives, SURV exhibited good data quality with limited discrepancies, but underreported COVID-19 deaths at home. Presuming neither database can be unequivocally considered as gold standard for COVID-19 mortality statistics, they provide valuable insights for policy formulation. Improving real-time mortality data collection is crucial, emphasising the need for effective collaboration among stakeholders.

Keywords: *COVID-19, Mortality, Epidemiologic Surveillance, Death Certificates, Cause of Death*

Résumé

La pandémie de COVID-19 a nécessité la mise en place rapide de la base de données de la surveillance épidémiologique de la mortalité COVID-19 (SURV) en Belgique en raison d'un retard important dans la disponibilité de la base de données sur les causes de décès (COD). Il est essentiel de comprendre les différences et les limites des deux bases de données pour contextualiser les statistiques de mortalité COVID-19. Cette étude évalue la qualité des données de la SURV, sensibilise aux différences et aux limites des deux bases de données et propose des recommandations pour la surveillance future de la mortalité durant une pandémie. Les données SURV et COD ont été reliées de manière probabiliste afin d'explorer les couvertures globales et les divergences. Des facteurs tels que la région et le lieu de décès, la classification de cas, la vague épidémique, le groupe d'âge, le sexe et le nombre de conditions ont été analysés à l'aide de modèles de régression logistique. La SURV a identifié 90 % (n = 19 801) des décès dus à la COVID-19 par rapport à la COD (n = 22 015). La couverture était plus élevée dans les hôpitaux (98 %, n = 11 130 dans la SURV, n = 11 335 dans la COD) et les établissements de soins de longue durée (LTCF) (90 %, n = 8 602 dans la SURV, n = 9 580 dans la COD) par rapport aux décès à domicile (5 %, n = 52 dans la SURV, n = 1 057 dans la COD). Cependant, 83,9 % des enregistrements de la SURV mentionnaient la COVID-19 comme cause sous-jacente de décès dans la COD, et 75,4 % des décès dus à la COVID-19 dans la COD ont été identifiés dans la SURV. L'activité réduite de la COVID-19 et l'incertitude du diagnostic ont entraîné une moindre concordance entre les bases de données. Des variations dans la qualité des données ont été observées entre les vagues épidémiques, les régions et les établissements de santé. En plus d'atteindre

les objectifs en temps réel, la SURV a présenté une bonne qualité de données avec des divergences limitées, mais a sous-déclaré les décès dus à la COVID-19 à domicile. En supposant qu'aucune des deux bases de données ne puisse être considérée sans équivoque comme l'étalon-or des statistiques de mortalité COVID-19, elles fournissent des informations précieuses pour la formulation des politiques. Il est essentiel d'améliorer la collecte de données sur la mortalité en temps réel, ce qui souligne la nécessité d'une collaboration efficace entre les parties prenantes.

Mots-clés : *COVID-19, Mortalité, Surveillance épidémiologique, Certificats de décès, Cause de décès*

1. Introduction

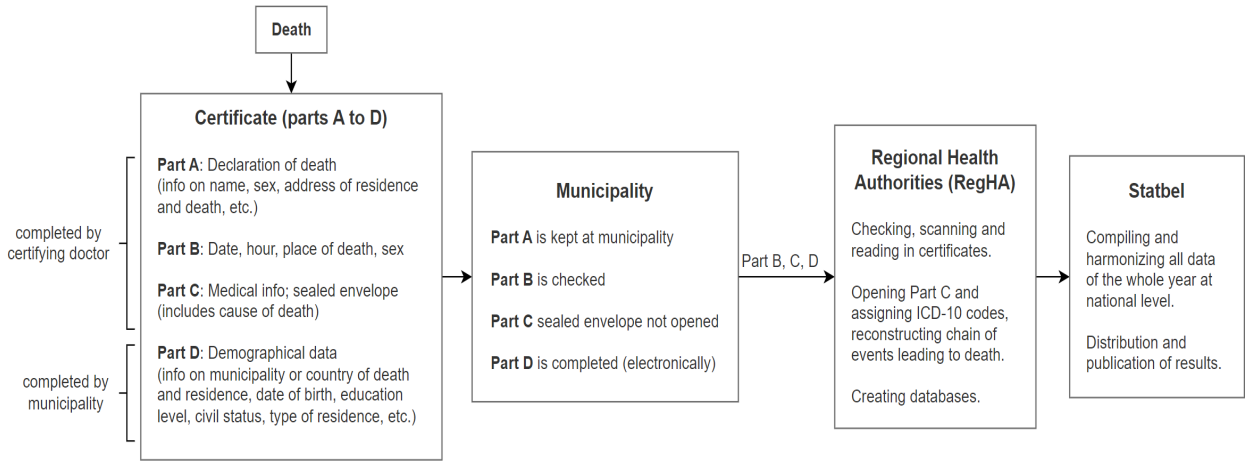
1.1.. Background

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first identified in Wuhan, China, in December 2019 and has since spread worldwide. By March 2020, the World Health Organization (WHO) declared the outbreak of COVID-19 as a pandemic. As of 16 August 2023, more than 760 million cases and 6.95 million deaths have been reported to the WHO (World Health Organization, 2023). Belgium's first COVID-19 case was reported on 3 February 2020, and the start of the epidemic was announced in March 2020 (Hope, 2020). To ensure the availability of timely and reliable data within the health emergency that COVID-19 represented, Sciensano, in collaboration with the health authorities, set up ad-hoc surveillances, including the COVID-19 mortality surveillance (Renard et al., 2021). This was needed as there were no alternatives to monitor COVID-19 mortality.

The first wave ended by 21 June 2020, and demonstrated a COVID-19 mortality rate amongst the highest worldwide (Johns Hopkins Coronavirus Resource Center, 2020). To some degree, these figures could be attributed to how COVID-19 deaths were counted in Belgium. In addition to deaths from laboratory-confirmed cases, all deaths from possible (i.e. meeting the clinical criteria) and radiologically-confirmed cases were also included in the ad-hoc COVID-19 mortality surveillance (Sciensano, 2020a). Furthermore, it aimed to be as complete as possible, including deaths registered in both a hospital and a long-term care facilities (LTCF) surveillance system or at home (Bustos Sierra et al., 2020; Renard et al., 2021). This inclusive approach provided results that were more in line with excess mortality figures both nationally and internationally (Karlinsky & Kobak, 2021; Molenberghs et al., 2022; Morgan et al., 2020; Wang et al., 2022). Ad-hoc COVID-19 mortality surveillance ended on 30 June 2023 upon termination of the exhaustive COVID-19 hospital surge capacity surveillance (Van Goethem et al., 2020) and the COVID-19 LTCF surveillance (Dequeker et al., 2020). In total, more than 4.8 million cases and 34,339 deaths have been reported to Sciensano, the Belgian Institute for Public Health (Sciensano, 2023d).

In Belgium, cause of death surveillance relies on paper death certificates completed by a certifying doctor. Notification of death to the National Register usually occurs within 14 days, but there is a 2-3-year delay between the date of death and the availability of cause of death data (Figure 1). The medical section of the death certificate (part C) goes to the respective Regional Health Authorities (RegHA). After digitalisation, nosologists (mortality medical coders) assign ICD-10 codes and follow WHO rules to reconstruct the chain of events leading to death. In cases where information on the death certificate is incomplete, the certifying doctor is contacted to provide the missing information. For suspicious deaths, coders have to gather legal investigation results. Once complete, the database for a whole calendar year is sent to the Belgian Statistical Office (Statbel), which compiles and harmonises the data at national level and distributes them for statistical and research purposes (Statbel, 2023).

Figure 1: Data flow diagram of death certificates in Belgium.



The two sources of COVID-19 mortality data – ad-hoc COVID-19 mortality surveillance and the cause of death database – each have distinct characteristics and benefits. The ad-hoc surveillance was able to provide timely and comprehensive data, which was crucial for real-time public health response and decision-making. On the other hand, the cause of death database, based on death certificates, offers detailed and standardized cause-of-death information. However, this process entails a substantial delay due to meticulous and time-consuming quality protocols. Investigating the differences in data collection methods, coverage, and timing between these two sources is important for understanding their respective strengths and limitations.

1.2. International context

The challenges encountered in monitoring and reporting COVID-19 mortality in Belgium were not unique. International experiences offered insights into COVID-19 mortality reporting complexities. Different countries employed various methods to monitor and report COVID-19 mortality, often resulting in disparities due to imperfect and diverse data collection systems (Garcia et al., 2021). In the United Kingdom, deaths within 28 days of a reported COVID-19 infection were used as primary metric for rapid pandemic monitoring. Throughout 2020 and 2021, between 80% and 90% of deaths reported within 28 days of a positive COVID-19 test also had COVID-19 mentioned on the death certificate. However, this proportion dropped below 50% from early 2022 onwards. With higher levels of immunity in the population and COVID-19 continuing to circulate in the community, it became not uncommon for people who died of other causes to also have a recorded COVID-19 infection around the time of death. Therefore, in 2023, the primary metric for COVID-19 death statistics changed to deaths with COVID-19 on the death certificate, which provides a more accurate measure of COVID-19 deaths over time, even though it involves a longer reporting delay (11 days compared to 2-3 days) (UK Health Security Agency 2023; Seghezzo et al. 2023).

France employed two simultaneous systems, SI-VIC and SurvESMS, to record COVID-19 deaths, with SI-VIC covering COVID-19 deaths of hospitalised patients and SurvESMS covering a broad range of institutions. An exhaustive analysis based on death certificates is available but with a delay of at least one year. In 2020, the combined reporting from the SIVIC and SurvEMSM system accounted for 64,600 COVID-19 deaths, which represented 93% of the 69,200 deaths from the exhaustive analysis based on death certificates (Clanché & Caserio-Schönemann, 2023).

The Netherlands faced a similar challenge in 2020, where COVID-19 deaths based on death certificates were available only after several weeks. Instead, they relied on COVID-19 death numbers based on non-mandatory notifications to the GGD (Municipal Health Service) from hospitals, nursing homes (NH), and general practitioners. However, these numbers excluded persons who died from COVID-19 but who were not tested. Retrospective analyses revealed that the number of notified COVID-19 deaths was lower compared to the numbers based on death certificates. By the end of February 2021, there were 15,818 COVID-19 deaths counted based on the notifications to GGD, representing about 58% of the 27,056 deaths counted based on the death certificates (Rijksoverheid, 2021b, 2021a).

These international experiences highlight that providing highly accurate estimates of COVID-19 deaths in real-time is a complex challenge. However, it is important to note that most of the presented systems performed effectively in real-time monitoring in 2020.

1.3. Objectives

In this paper, we assessed the data quality of the Belgian COVID-19 mortality epidemiological surveillance database and conducted a comparative analysis between this database and the cause of death register of death certificates for the year 2020. Our aim was to raise awareness of the differences and limitations of both COVID-19 mortality sources, and propose recommendations for mortality surveillance for other public health threats that need real-time monitoring.

2. Methods

2.1. Data sources and preparation

2.1.1. COVID-19 mortality epidemiological surveillance database

Sciensano received notification of COVID-19 deaths from hospitals, LTCF (including NH, service flats for elderly people, institutions for persons with disabilities, etc.), and general practitioners. Criteria for including COVID-19 deaths in the surveillance were based on the ECDC (European Centre for Disease Prevention and Control) and WHO guidelines (World Health Organization, 2020b), with minor modifications (Renard et al., 2021). Every notification contained information on the date of death, date of birth, sex, place of death (hospital, NH, at home, etc.), type of residence (community-dwelling individuals, residents of NH, residents of psychiatric institutions, etc.), postal code of place of death, postal code of place of residence, and case classification (laboratory-confirmed, radiologically-confirmed, possible case).

Sciensano harmonised, cleaned, and pooled all records into one final national database: the COVID-19 mortality epidemiological surveillance database (SURV). During the initial stages of the epidemic, not all variables were available across all data sources, and implausible or missing values were sometimes encountered. Moreover, due to limited testing capacity and prioritisation of testing in hospitals in the early stages of the pandemic, some deaths were probably erroneously attributed to COVID-19 or not reported. It was only since April/May 2020 that testing capabilities expanded to include confirmed cases in nursing homes (Dequeker et al., 2023). Nevertheless, as time progressed, data across all sources became more homogeneous. Retrospective updates continuously improved the database's quality throughout the epidemic (Sciensano, 2023b).

All SURV records with a date of death between 7 March 2020 (first COVID-19 death) and 31 December 2020 were retained, encompassing 19,801 records. Date of birth was missing for 1,296 SURV records (6.5%) (Table A-1 in Appendix A). The

vast majority (81%) of these were records received from NH in Wallonia because, at that time, their questionnaire asked only the year of birth. A retrospective analysis in 2022-23 recovered about two-thirds of the missing date of birth. Year of birth was missing for only a very small number of records (n=25, 0.13%). All other variables had a very low degree of missingness among the SURV records.

2.1.2. Cause of death database

The cause of death database (COD) based on death certificates of the year 2020 became available from Statbel in February 2023. It included any death of a person with legal residence in Belgium. Available variables included date of death, date of birth, sex, place of death, type of residence, NIS 5[1] of place of death, NIS 5 of place of residence, underlying cause of death (UCOD), and a string variable with multiple cause codes that captures all conditions specified on the death certificate. The multiple cause codes do not differentiate between underlying, immediate, intermediate, or associated causes of death. In COD, a COVID-19 death (or death *due* to COVID-19) was defined as a death having either the ICD-10 code U07.1 (COVID-19, virus identified) or U07.2 (COVID-19, virus not identified) listed as the UCOD. On the other hand, a death *with* COVID-19 was defined as a death having U07.1 or U07.2 appear in the multiple causes of death variable, but not as the UCOD.

All records with a date of death between 29 February 2020 and 31 December 2020 were retained in COD, culminating in 107,591 COD records. The seven days extra at the start, compared to SURV, were to take into account possible errors in the date of death around the start of SURV. None of the records had missing values for any of the variables of interest.

2.2 Linkage

2.2.1. Linkage methods

Considering the presence of erroneous and missing data within SURV, and the absence of a shared unique identifier, a probabilistic linkage approach was used to merge the two databases at individual level. We implemented a Fellegi-Sunter probabilistic record linkage model using the expectation-maximization (EM) algorithm and accounting for missing data (under the MAR (missing at random) assumption) (Enamorado et al., 2019). The strength of a particular match between SURV and COD was assessed using probability criteria (Gu et al., 2003; Potz et al., 2010). The open-source R software package, *fastLink* (Fast Probabilistic Record Linkage with Missing Data) (Enamorado et al., 2020), was used in this research.

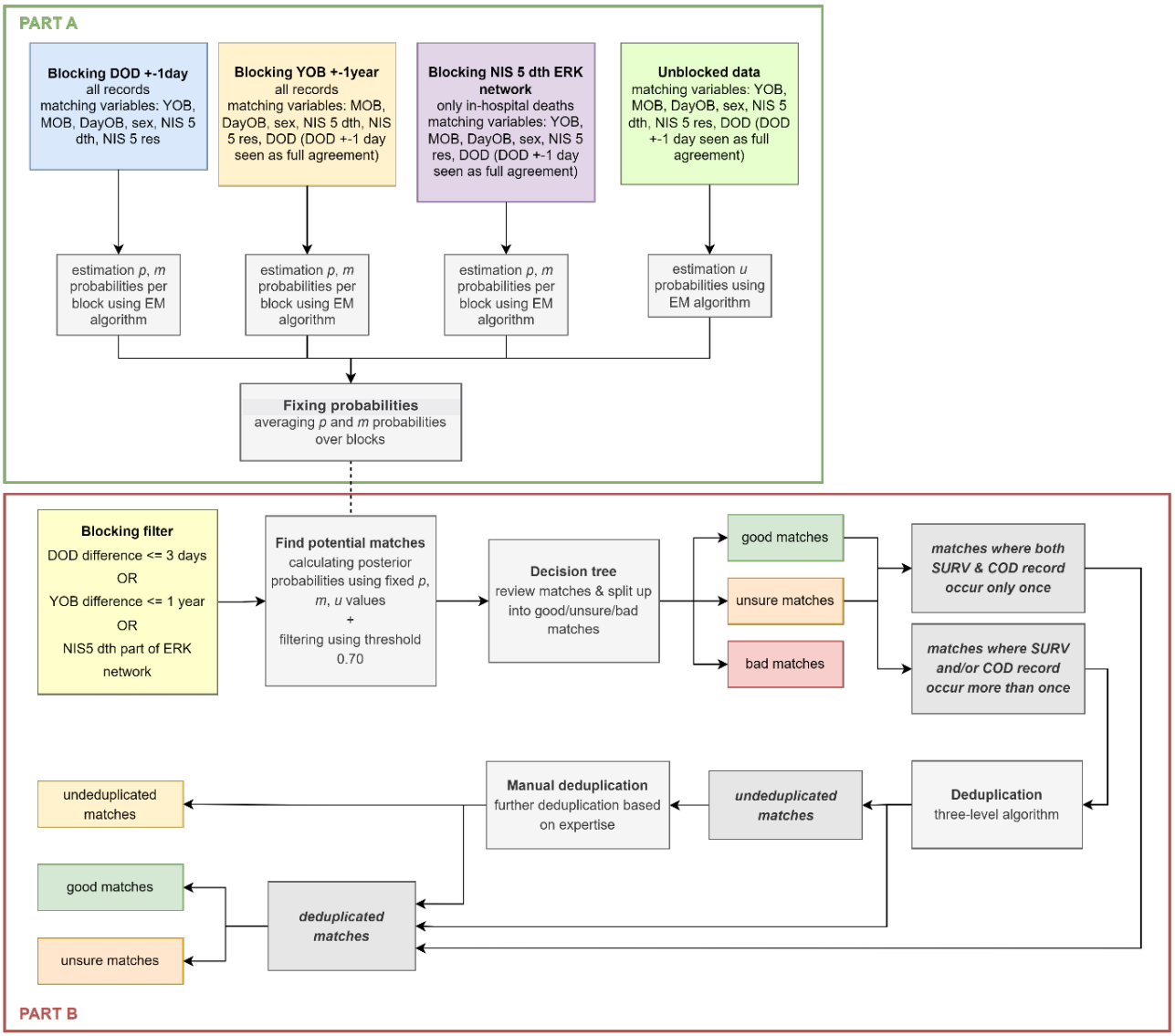
2.2.2..Linkage parameters

At the heart of probabilistic record linkage are *m* and *u* probabilities. An *m* probability is seen as an indicator of data quality and signifies the probability that a matching variable agrees given that the pair of records is a true match. A *u* probability, conversely, represents the probability of agreement purely by chance among matches (Blakely & Salmond, 2002). Based on Bayes' theorem, the match probability (posterior probability) for each record pair was calculated using all *m* and *u* values, and the prior probability *p* of a match. These parameters were estimated using the EM algorithm. We set fixed parameters to ensure consistency in posterior probability calculations across all record pairs (Figure 2, part A). The *m* and *p* probabilities were based on running the EM algorithm on blocked data and averaging over all blocks. Blocking reduces the number of record pairs to be processed.

[1] The NIS 5 code is a numeric code for the geographical areas in Belgium where the first digit identifies the province, the second digit identifies the arrondissement within the province, and the last three digits identify the municipality within that arrondissement. One NIS 5 code includes one or multiple postal codes.

For example, in blocking by year of birth ±1 year, only record pairs where the year of birth was within 1 year were compared. The three blocking passes involved one using date of death ±1 day, one using year of birth ±1 year, and one where all NIS 5 of death codes coming from the same hospital network were allowed. Each time, the date of death ±1 day was seen as full agreement (i.e. as if they had the same date of death), since we knew that the date of death in SURV records was sometimes one day off. The *u* probability of each matching variable was fully based on running the EM algorithm on the whole dataset, i.e. without blocking (Fellegi & Sunter, 1969). A mean prior probability *p* of 0.0036 was calculated based on this method. The estimated average *m* and *u* probabilities are shown in Table 1.

Figure 2: Pipeline of the linkage methodology used to match the COVID-19 mortality epidemiological surveillance database with the cause of death database, Belgium, 2020.



Note: COD: cause of death database; DayOB: day of birth; DOD: date of death; ERK: identifier of a hospital network; m: probability that a matching variable agrees given that the pair of records is a true match; MOB: month of birth; NIS 5 dth: NIS 5 of place of death; NIS 5 res: NIS 5 of place of residence; p: prior probability of a match; SURV: COVID-19 mortality epidemiological surveillance database; u: probability that a matching variable agrees given that the pair of records is not a true match; YOB: Year of birth.

Table 1: Set of fixed m and u probabilities, linkage between the COVID-19 mortality epidemiological surveillance database with the cause of death database, Belgium, 2020.

	matching variable						
	YOB	MOB	DayOB	Sex	NIS 5 dth	NIS 5 res	DOD
Average u probability	0.029	0.084	0.033	0.500	0.011	0.006	0.013
Average m probability	0.849	0.807	0.843	0.798	0.981	0.963	0.917

Note: DayOB: day of birth; DOD: date of death; m: probability that a matching variable agrees given that the pair of records is a true match; MOB: month of birth; NIS 5 dth: NIS 5 of place of death; NIS 5 res: NIS 5 of place of residence; u: probability that a matching variable agrees given that the pair of record is not a true match; YOB: Year of birth.

2.2.3. Finding potential matches

Once all parameters were fixed, potential matches between SURV and COD were found by calculating the posterior probabilities for each record pair and using a threshold of 0.70 to filter out all (very) unlikely matches. However, not every SURV record was compared to every COD record. A broad blocking filter was installed prior to calculating posterior probabilities (Figure 2, part B). Only record pairs where the date of death was within three days, or the year of birth was within one year, or the NIS 5 of death was part of the NIS 5 of death codes associated with the hospital network, were retained. This way, the number of total comparisons was drastically reduced, while minimising the number of missed correct matches.

Not every potential match with a posterior probability of 0.70 or higher was automatically considered a decent match. However, given the impracticality of manually reviewing each potential match and the associated elevated risk of inconsistency, an alternative approach was sought. A decision tree was created to categorise the potential matches into ‘good’, ‘unsure’, and ‘bad’ matches. This tree evolved through the examination of a subset of potential matches, taking into account various factors. For instance, cases with disagreeing NIS 5 of residence, yet neighbouring, were seen as better potential matches compared to cases with non-neighbouring NIS 5 of residence codes. These nuances were not captured by the original probabilistic linkage method but were important in the decision-making process to link the best probable records. The full decision tree can be found in Figures B-1 and B-2 in Appendix B.

2.2.4. Deduplication

Subsequently, matches that required further deduplication were separated from the rest of the matches. A three-level deduplication algorithm was constructed, as detailed in Appendix B. However, not all unduplicated matches could be disentangled, either because it was by definition not possible (e.g. for one SURV record there were two COD records found with the exact same values), or because the suggested matches were about equally as likely to be the correct match. The unduplicated matches were sent for manual deduplication, which was primarily based on experience.

Ultimately, three datasets were formed: deduplicated ‘good’ matches, deduplicated ‘unsure’ matches, and ‘unduplicated’ matches. These datasets were redistributed into five matching quality groups: ‘perfect’ match (identical values for all matching variables), ‘good’ match (disagreement on one or more of the matching variables but still regarded as a decent match), ‘unsure’ match (doubt if it is a ‘good’ match due to more extensive disagreement on matching variables), ‘deduplication issues’ (no one-to-one match found; connected subset of SURV and COD records could not be disentangled), no potential matches found (no COD record found for the SURV record; excluding SURV records that are part of the group with deduplication issues).

2.3. Comparison indicators

2.3.1. Data quality of the COVID-19 mortality epidemiological surveillance database

The quality of SURV data was assessed through coupling with COD, using the custom-designed probabilistic record linkage methodology. COD values for the matching variables served as the gold standard. First, a global comparison was made between the number of 'perfect', 'good', and 'unsure' matches between SURV and COD, along with the number of 'unresolved duplicates' and the number of 'unmatched' records. This comparison was also conducted in a stratified manner: by place of death (in hospital, NH, other places), by region of death (Flanders, Wallonia, Brussels), and by epidemic wave (first wave, interwave, second wave)[2].

Then, to shed light on the typical data entry errors occurring in SURV, the differences in values of the used matching variables were analysed for all 'good' and 'unsure' matched records.

2.3.2. Comparison between the COVID-19 mortality epidemiological surveillance database and the cause of death database

Comparing the COVID-19 mortality between SURV and COD for the year 2020 was conducted in two separate types of analyses: a global comparative analysis and several integrated comparative analyses, enabled by one-to-one matching. We started from the fundamental premise that, while the COD values of the matching variables serve as the gold standard for the data quality assessment, neither the UCOD variable of COD nor the SURV can be unequivocally considered the gold standard for COVID-19 mortality statistics.

- 2.3.2.1. Global comparative analysis

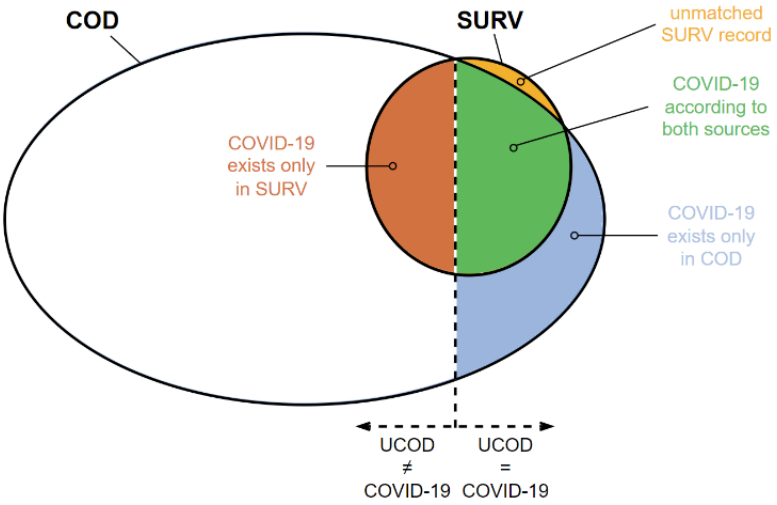
Here, an exploratory analysis was performed separately for each database, i.e. without using one-to-one matching. Consequently, the numbers obtained did not necessarily pertain to the same individuals. The analysis involved comparing the total and monthly number of COVID-19 deaths in 2020, analysing the distribution of deaths according to place of death (in hospital, LTCF, at home) and region of death (Flanders, Wallonia, Brussels), as well as comparing the numbers per type of residence (community-dwelling individuals, LTCF residents, other or unknown residents), and examining all other conditions existing in COD via ICD-10 codes when COVID-19 (U07.1 or U07.2) was indicated as a UCOD. This last part was solely based on COD, not SURV, as this type of information was not available in SURV. All deaths reported through SURV were assumed to be deaths *due* to COVID-19 by design of the surveillance system.

- 2.3.2.2. Integrated analysis – comparative analysis

The integrated comparative analysis was performed at the individual level. Each SURV record, which could be matched to a COD record, was supplemented with information about the UCOD, along with other conditions if mentioned. Overall, four mutually exclusive categories could be defined: COVID-19 deaths recorded in both databases, COVID-19 deaths solely present in SURV (where another UCOD was mentioned in COD), COVID-19 deaths solely present in COD (with no corresponding SURV record found), and unmatched SURV records (Figure 3). The numbers and percentages for each of the four categories were reported for the overall dataset, as well as for the specified places and regions of death. Additionally, COVID-19 mortality rates by sex were computed for each of the four categories individually, as well as for the scenarios involving exclusive use of either SURV or COD, and combined data from all categories.

[2] First wave: 7 March 2020 to 21 June 2020; Interwave: 22 June 2020 to 30 August 2020; Second wave: 31 August 2020 to 14 February 2021 (here 31 December 2020 as cut-off point).

Figure 3: Concept of defining four categories after linkage between the COVID-19 mortality epidemiological surveillance database with the cause of death database, Belgium, 2020.



Note: COD: cause of death database from the death certificates; SURV: COVID-19 mortality epidemiological surveillance database; UCOD: underlying cause of death.

- 2.3.2.3. Integrated analysis – probability analysis

To further quantify the matter, two multiple logistic regression models were constructed. The first model focused on exploring influencing factors on the probability of a SURV record having a corresponding record in COD indicating COVID-19 as UCOD (COVID-19 according to both sources category versus COVID-19 existing only in SURV). The factors of interest were the place of death (in hospital, LTCF, at home), region of death (Flanders, Wallonia, Brussels), epidemic wave (first wave, interwave, second wave), case classification according to SURV (laboratory-confirmed case, radiologically-confirmed case, or possible case), sex (male, female), age group (<65, 65-74, 75-84, 85+), and the number of unique conditions mentioned in COD (including underlying, immediate, intermediate, and associated causes of death). In addition, three potential interaction effects were investigated, based on their expected relevance: place of death with region of death, place of death with wave, and region of death with wave. All data pertaining to these variables were derived from SURV, except the number of conditions mentioned in COD.

Furthermore, in cases where a SURV record existed but the UCOD was not documented as COVID-19, a list of the UCODs was created to identify possible reasons for the mismatch, considering both the ICD-10 code and chapter level of the ICD-10 UCOD.

The second model aimed to explore influencing factors on the probability that a COVID-19 death as UCOD in COD has a corresponding record in SURV (COVID-19 according to both sources category versus COVID-19 existing only in COD). The factors of interest remained consistent with the first model, including the three aforementioned interaction terms. However, in this model, all values were obtained from COD.

For both models, purposeful selection was used as the model-building strategy (Zhang, 2016). To aid in the interpretation of the statistical results, estimated probabilities and estimated average probabilities, along with their 95% confidence intervals, were calculated and plotted.

2.4. Data analysis

The analysis focussed on COVID-19 deaths in 2020, extracted from the 14 July 2023 version of the final COVID-19 death database in Belgium (Sciensano, 2023a). All analyses and graphs were created using R version 4.2.3 (R foundation, Vienna, Austria). The *UpSetR* library was used for Figure 4 (Conway et al., 2017).

2.5. Ethical statement

Sciensano is legally entitled to surveillance activities related to public health in Belgium (Moniteur Belge, 2018). Collection of aggregated data was performed within lawful grounds of the General Data Protection Regulation (GDPR). The study protocol in LTCF was approved by the Ghent University ethical committee (23/10/2020, BC-08065) (Vandael et al., 2022). All data are stored on a secured server at Sciensano.

3. Results

3.1. Data quality of the COVID-19 mortality epidemiological surveillance database

Out of the 19,801 SURV records, 14,280 (72.1%) were 'perfectly' matched to a COD record (Table 2). An additional 4,682 (23.6%) were considered 'good' matches, and 228 (1.2%) as 'unsure' matches. For 64 SURV records (0.3%), there were still 'de-duplication issues' that could not be resolved. No potential matches were found for 547 (2.8%) SURV records.

In-hospital deaths were matched 'perfectly' for 75.9% of the records, while this figure was slightly lower for NH deaths (67.9%), and dropped to 28.8% for deaths in other places. However, this difference was less pronounced when comparing all the 'matched' groups combined (perfect, good and unsure) with the 'non-matched' group: 97.4% of in-hospital deaths were matched, compared to 96.7% for in-NH deaths, and 74.5% for deaths in other places.

Regional variations were apparent, with Flanders having a 'perfect' match rate of 82.0% compared to 62.6% in Wallonia and 62.4% in Brussels. However, these differences decreased when considering all the 'matched' groups together, resulting in matches for 98.6% in Flanders, 94.9% in Wallonia, and 96.3% in Brussels.

The proportion of 'perfect' matches increased from 71.5% in wave 1 to 72.7% in wave 2. At the same time, the proportion of records for which no match was found slightly decreased from 3.4% in wave 1 to 2.8% in wave 2.

Looking at the proportions of all types of matches, 80.4% of all matches, where the SURV record had Flanders as the region of death, NH as the place of death, and date of death in the year 2020, were matched perfectly to a COD record (Table 3). The proportion of 'perfect' matches for deaths in NH in Flanders was comparable between the first and second wave (around 80%) but dropped somewhat during the interwave period (73.2%). These proportions were much lower in Wallonia because of missing day and month of birth values for a proportion of SURV records coming from NH. For Brussels, these proportions lay between those of Flanders and Wallonia. At the same time, for in-NH deaths, the proportion of SURV records with no match at all was lowest in Flanders (1.1%), followed by Brussels (1.9%), and Wallonia (6.3%).

Table 2: Number of COVID-19 mortality epidemiological surveillance records matched or not to a cause of death record, by the quality of the match, with stratification by place of death, by region of death, and by epidemic wave, Belgium, 2020.

		COVID-19 in SURV matched to a COD record				COVID-19 in SURV not matched to a COD record			Total
		'Perfect' match	'Good' match	'Unsure' match	Total	Deduplication issues	No potential matches found	Total	
Overall		14,280 (72.1%)	4,682 (23.6%)	228 (1.2%)	19,190 (96.9%)	64 (0.3%)	547 (2.8%)	611 (3.1%)	19,801 (100%)
Stratified by place of death	in NH	5,783 (67.9%)	2,241 (26.3%)	210 (2.5%)	8,234 (96.7%)	39 (0.4%)	245 (2.9%)	284 (3.3%)	8,518 (100%)
	in hospital	8,453 (75.9%)	2,377 (21.4%)	12 (0.1%)	10,842 (97.4%)	19 (0.2%)	269 (2.4%)	288 (2.6%)	11,130 (100%)
	other*	44 (28.8%)	64 (41.8%)	6 (3.9%)	114 (74.5%)	6 (3.9%)	33 (21.6%)	39 (25.5%)	153 (100%)
Stratified by region of death	Flanders	7,995 (82.0%)	1,602 (16.4%)	14 (0.2%)	9,611 (98.6%)	8 (0.1%)	127 (1.3%)	135 (1.4%)	9,746 (100%)
	Wallonia**	4,624 (62.6%)	2,187 (29.6%)	203 (2.7%)	7,014 (94.9%)	54 (0.7%)	324 (4.4%)	378 (5.1%)	7,392 (100%)
	Brussels	1,661 (62.4%)	893 (33.5%)	11 (0.4%)	2,565 (96.3%)	2 (0.1%)	96 (3.6%)	98 (3.7%)	2,663 (100%)
Stratified by epidemic wave	wave 1	6,913 (71.5%)	2,285 (23.6%)	147 (1.5%)	9,345 (96.6%)	44 (0.5%)	280 (2.9%)	324 (3.4%)	9,669 (100%)
	interwave	242 (74.5%)	69 (21.2%)	4 (1.2%)	315 (96.9%)	0 (0.0%)	10 (3.1%)	10 (3.1%)	325 (100%)
	wave 2	7,125 (72.7%)	2,328 (23.7%)	77 (0.8%)	9,530 (97.2%)	20 (0.2%)	257 (2.6%)	277 (2.8%)	9,807 (100%)

*including at home, collectivities such as service flats for elderly people, psychiatric institutions, revalidation centres, etc.
**German-speaking Community included in Wallonia

Note: Row-wise percentages are given between brackets. First wave: 7 March 2020 to 21 June 2020; Interwave: 22 June 2020 to 30 August 2020; Second wave: 31 August 2020 to 14 February 2021 (here 31 December 2020 as cut-off point); COD: cause of death database from the death certificates; SURV: COVID-19 mortality epidemiological surveillance database.

Among the 4,682 'good' and 228 'unsure' matches (4,910 in total, Table 2), SURV errors were mainly in NIS 5 of residence, date of birth, and date of death (Figure 4). Date of death differed in 1,167 matches, with or without additional differences in other matching variables. The most prevalent disagreement was on NIS 5 of residence (1,634 matches), with 62% involving neighbouring codes. In case the NIS 5 of residence codes were known (93%), but different and non-neighbouring, the majority (83%) were labelled as NH residents. There were a considerable number of differences in day and month of birth (1,290 and 1,422, respectively). However, the majority of those differences were due to a missing day or month of birth in SURV

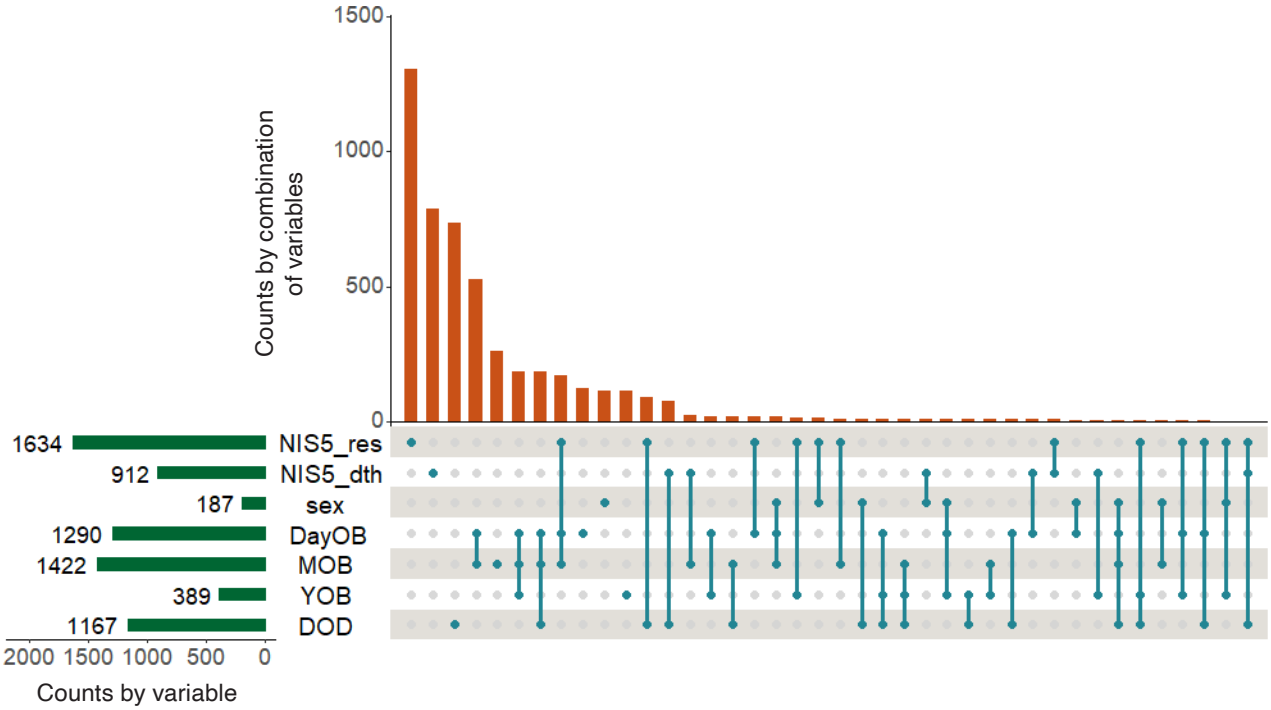
(79% and 72%, respectively). Another 17% of matches with a difference in month of birth were instances where the SURV record had the first of January as date of birth, whereas the COD record had the first of July as date of birth. Out of the 1,142 SURV records with a missing day and month of birth (and no other missing variables), 918 (80%) were matched to a COD record. For about half of the matches where the date of death did not correspond, SURV had a date of death one day later than COD, while in 33%, it was a day earlier. Discrepancies in NIS 5 of death were common, but 94% of those were coming from hospitals that had SURV and COD NIS 5 of death codes that were part of the same hospital network. Sex and year of birth had fewer differences (187 and 389, respectively).

Table 3. Proportions of all types of matches between COVID-19 mortality epidemiological surveillance database with the cause of death database, per region of death, per place of death, and epidemic wave, Belgium, 2020.

		FLANDERS				WALLONIA				BRUSSELS				BELGIUM			
		NH	Hosp	Other	All	NH	Hosp	Other	All	NH	Hosp	Other	All	NH	Hosp	Other	All
Perfect	2020	80.4	84.3	40.0	82.0	47.1	73.1	11.4	62.6	66.5	60.3	61.5	62.4	67.9	75.9	28.8	72.1
	wave 1	81.3	84.6	33.3	82.3	46.5	72.5	12.5	59.8	67.1	61.9	42.9	63.9	67.9	76.0	24.3	71.5
	interwave	73.2	85.8	NA	81.7	35.0	70.8	NA	63.0	100.0	70.2	0.0	71.9	65.9	77.7	0.0	74.5
	wave 2	79.5	84.0	54.5	81.8	48.0	73.5	9.1	64.9	64.1	57.9	100.0	59.5	67.9	75.9	38.8	72.7
Good	2020	18.3	14.4	34.3	16.4	38.6	23.6	51.4	29.6	30.9	34.9	30.8	33.5	26.3	21.4	41.8	23.6
	wave 1	17.6	13.5	35.4	15.9	37.4	23.7	54.2	30.4	30.4	34.0	42.9	32.6	25.9	21.0	44.7	23.6
	interwave	21.4	14.2	NA	16.6	40.0	26.4	NA	29.3	0.0	22.8	100.0	21.9	24.4	19.8	100.0	21.2
	wave 2	18.9	15.1	31.8	17.0	40.0	23.5	45.5	28.9	32.9	36.7	0.0	35.7	26.9	21.7	34.7	23.7
Unsure	2020	0.2	0.0	0.0	0.1	6.8	0.1	8.6	2.7	0.7	0.3	0.0	0.4	2.5	0.1	3.9	1.2
	wave 1	0.1	0.0	0.0	0.1	8.5	0.1	8.3	4.1	0.6	0.2	0.0	0.4	2.9	0.1	3.9	1.5
	interwave	1.8	0.0	NA	0.6	15	0.0	NA	3.3	0.0	0.0	0.0	0.0	4.9	0.0	0.0	1.2
	wave 2	0.4	0.0	0.0	0.2	4.6	0.1	9.1	1.6	0.9	0.4	0.0	0.5	1.8	0.1	4.1	0.8
Dedup.	2020	0.1	0.1	0.0	0.1	1.3	0.3	8.6	0.7	0.0	0.1	0.0	0.1	0.5	0.2	3.9	0.3
	wave 1	0.0	0.2	0.0	0.1	1.5	0.6	10.4	1.2	0.0	0.0	0.0	0.0	0.5	0.3	4.9	0.5
	interwave	0.0	0.0	NA	0.0	0.0	0.0	NA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	wave 2	0.1	0.0	0.0	0.1	1.0	0.0	4.5	0.4	0.0	0.2	0.0	0.2	0.4	0.1	2.0	0.2
No match	2020	1.1	1.2	25.7	1.3	6.3	3.0	20.0	4.4	1.9	4.4	7.7	3.6	2.9	2.4	21.6	2.8
	wave 1	1.0	1.7	31.2	1.6	6.1	3.1	14.6	4.6	1.9	3.9	14.3	3.1	2.8	2.6	22.3	2.9
	interwave	3.6	0.0	NA	1.2	10.0	2.8	NA	4.3	0.0	7.0	0.0	6.2	4.9	2.5	0.0	3.1
	wave 2	1.1	0.8	13.6	1.0	6.4	2.9	31.8	4.2	2.1	4.8	0.0	4.2	3.0	2.3	20.4	2.6

Note: denominators are the sum of all types of matches, by the specified combination of region of death, place of death, and epidemic wave. Hosp: in hospitals;
NH: in nursing homes; NA: not applicable (denominator is zero); Dedup.: deduplication issues.

Figure 4: Number of matches (including only 'good' (4,682) and 'unsure' (228) matches) where a matching variable (green bars) or a specific combination of variables is disagreeing (blue lines, orange bars) between the COVID-19 mortality epidemiological surveillance database and the cause of death database, Belgium, 2020.



DayOB: day of birth; DOD: date of death; NIS5_dth: NIS 5 of death; NIS5_res: NIS 5 of residence; MOB: month of birth; YOB: year of birth.

Note: The horizontal green bars represent the total number of matches where the specific variable was different between both databases, with or without additional differences in other variables. Blue vertical lines represent a specific combination of variables disagreeing between both databases. Orange bars represent the corresponding counts of the combinations of disagreeing variables.

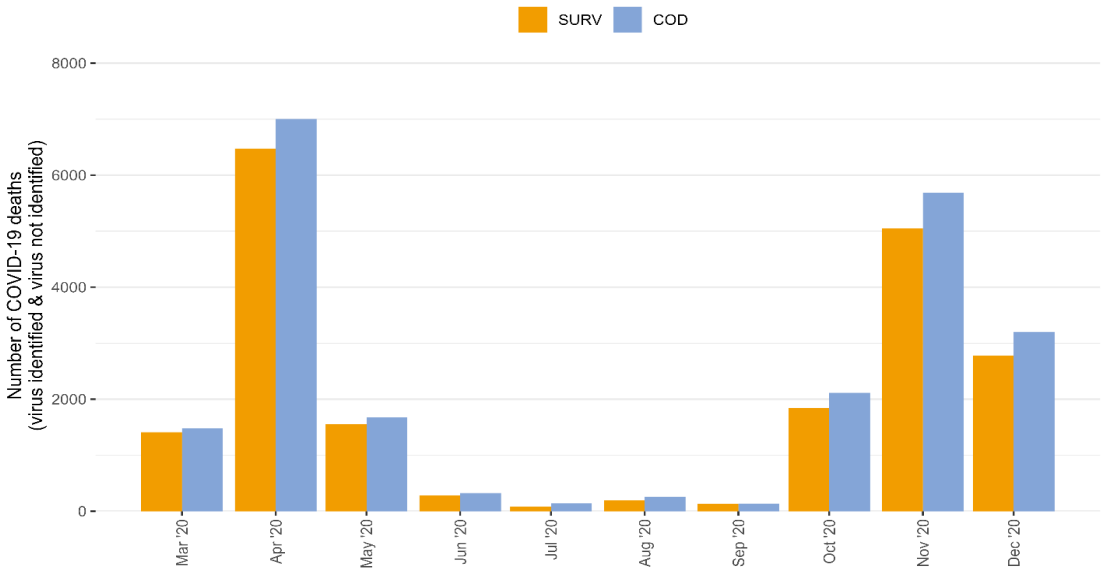
For example, the fourth orange bar represents the number of matches where the specific combination of day and month of birth disagreed, but all other matching variables agreed. In other words, these represent the intersecting numbers for each combination. The fourth green bar represents the total number of matches where the day of birth was different between both databases, with or without any additional differences in the other variables.

3.2. Comparison between both data sources

3.2.1. Global comparative analysis

In 2020, SURV reported 19,801 COVID-19 deaths, while COD recorded 22,015 (2,214 more). This indicates that SURV achieved a global coverage of 90%. Additionally, there were 847 deaths *with* COVID-19 in COD, but these were excluded from the counts. Notably, the numbers reported by COD consistently exceeded those reported by SURV each month (Figure 5). Relatively speaking, the largest disparity occurred in July, where SURV reported 81 COVID-19 deaths compared to the 140 deaths reported by COD.

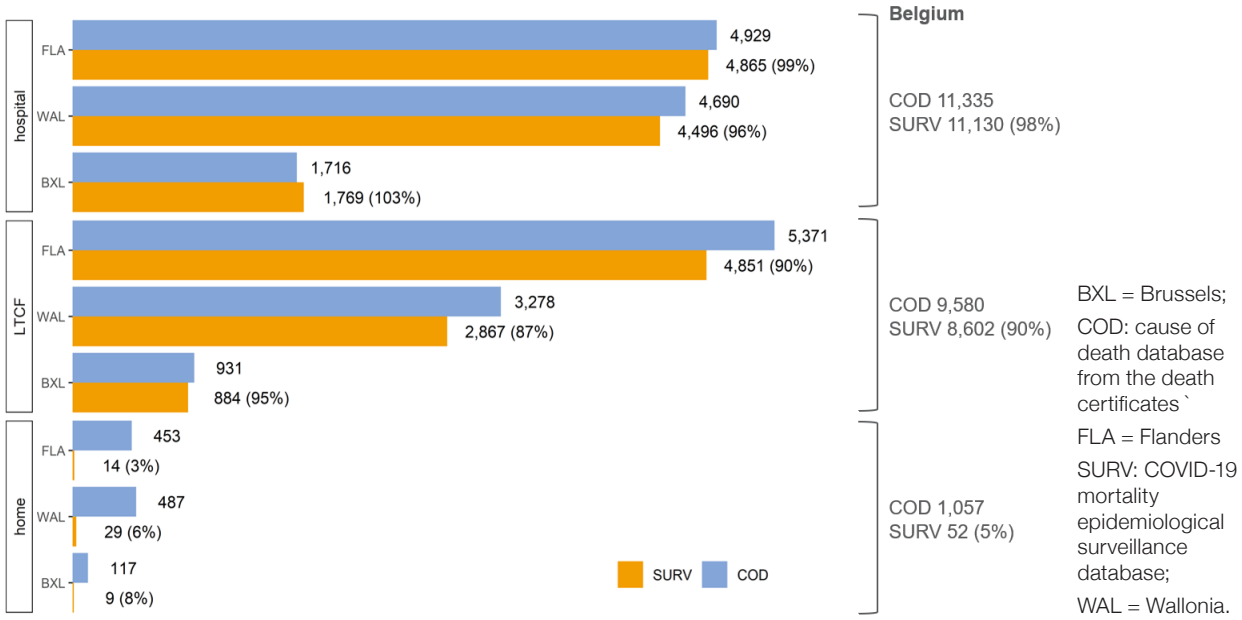
Figure 5: Monthly number of COVID-19 deaths in the COVID-19 mortality epidemiological surveillance database and the cause of death database, Belgium, 2020.



Note: COD: cause of death database from the death certificates; SURV: COVID-19 mortality epidemiological surveillance database.

There was high coverage via hospital (98%) and LTCF (90%) surveillances, but low coverage for COVID-19 deaths that occurred at home (5%), despite surveillance being in place (Figure 6). Regionally, SURV reported slightly more deaths that occurred in hospitals in Brussels, compared with the COD, due to methodological considerations in SURV.

Figure 6: Number of COVID-19 deaths per place of death and region of death in the COVID-19 mortality epidemiological surveillance database and the cause of death database, Belgium, 2020.



Note: Unknown and 'other' places of death are not shown (17 SURV records and 43 COD records);

The distribution of COVID-19 deaths among LTCF residents differed substantially between SURV and COD, as SURV reported 11,694 (59.1%) deaths among LTCF residents (mainly from NH), 6,258 (31.6%) among community-dwelling individuals, and 1,849 (9.3%) among other or unknown types of residents. In contrast, COD reported 9,951 (45.2%) COVID-19 deaths among LTCF residents (underestimating LTCF residents by 1,743 deaths), 11,727 (53.3%) community-dwelling individuals (+5,469 compared to SURV), and 337 (1.5%) other or unknown types of residents (-1,512 compared to SURV).

When COVID-19 was indicated as UCOD, the most frequently co-occurring ICD-10 codes were J18 (pneumonia) (42.2%), R09 (other symptoms and signs involving the circulatory and respiratory systems, e.g. asphyxia and hypoxemia, pleurisy, respiratory arrest, abnormal sputum) (28.6%), and J96 (respiratory failure) (24.5%) (Table 4).

Table 4: Highest-frequency conditions with their corresponding ICD-10 codes listed on the death certificate in case COVID-19 was indicated as the underlying cause of death, Belgium, 2020.

Condition	ICD-10 code	No. (% of 20,723*)
Pneumonia, organism unspecified	J18	8,744 (42.2)
Other symptoms and signs involving the circulatory and respiratory systems	R09	5,929 (28.6)
Respiratory failure, not elsewhere classified	J96	5,080 (24.5)
Heart failure	I50	1,809 (8.7)
Unspecified dementia	F03	1,797 (8.7)
Essential (primary) hypertension	I10	1,357 (6.5)
Cardiac arrest	I46	1,190 (5.7)
Other chronic obstructive pulmonary disease	J44	1,185 (5.7)
Atrial fibrillation and flutter	I48	1,156 (5.6)
Malaise and fatigue	R53	1,156 (5.6)
Other general symptoms and signs	R68	1,083 (5.2)
Chronic kidney disease	N18	1,082 (5.2)
Chronic ischaemic heart disease	I25	905 (4.4)
Alzheimer's disease	G30	876 (4.2)
Type 2 diabetes mellitus	E11	854 (4.1)

* Total number of records in the cause of death database from the death certificates having U07.1 (virus identified) or U07.2 (virus not identified) as underlying cause of death, and at least one other condition mentioned on the death certificate.

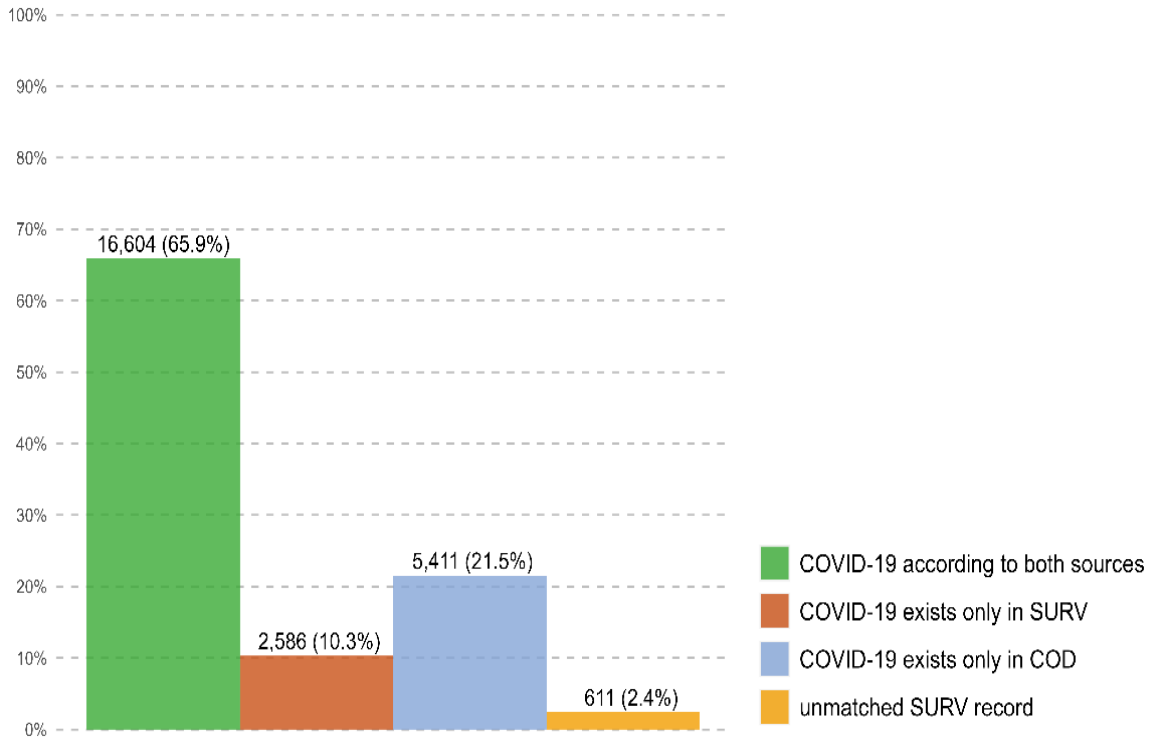
***Note**: conditions appearing less than 4.0% on death certificates are not shown.*

3.2.2. Integrated analysis – comparative analysis

Out of the 19,801 SURV records, 19,190 (96.9%) were matched with a COD record (Table 2). The total number of COVID-19 deaths reported by either SURV, COD, or both, i.e. encompassing all four categories, amounts to 25,212 (Figure 7). Among these, 16,604 (65.9%) COVID-19 deaths were recorded in both databases, 2,586 (10.3%) solely in SURV, 5,411 (21.5%) solely in COD, and 611 (2.4%) were unmatched SURV records.

For the majority of SURV records (83.9%, 16,604 out of 19,801), the UCOD of the matched COD record was also COVID-19, and for 13.1% (2,586) of them, the UCOD was another ICD-10 code (presented in Table 5). Of these 2,586 COVID-19 deaths existing only in SURV, 1,364 occurred in hospitals, 1,210 in LTCF, 10 at home, and 2 were reported from other or unknown places of death. Uncertainty persists for 3.1% (611) SURV records that could not be matched to COD. From the perspective of COD, 75.4% of its COVID-19 deaths (16,604 out of 22,015) were identified in SURV, whereas 24.6% (5,411) did not exist in SURV. Of the 5,411 deaths existing only in COD, 1,823 occurred in hospitals, 2,643 in LTCF, 920 at home, and 25 were reported from other or unknown places of death.

Figure 7: Overall distribution of four categories after linkage between the COVID-19 mortality epidemiological surveillance database with the cause of death database, Belgium, 2020.



Note: COD: cause of death database from the death certificates; SURV:COVID-19 mortality epidemiological surveillance database.

The COVID-19 mortality rate per 100,000 when considering all SURV records is 171.2 for males and 172.7 for females (Table 5). In contrast, when considering all COD records, it amounts to 189.3 for males and 193.3 for females. If all COVID-19 records reported by either SURV, COD, or both were taken into account, the COVID-19 mortality rate would equate to 214.9 for males, and 222.9 for females.

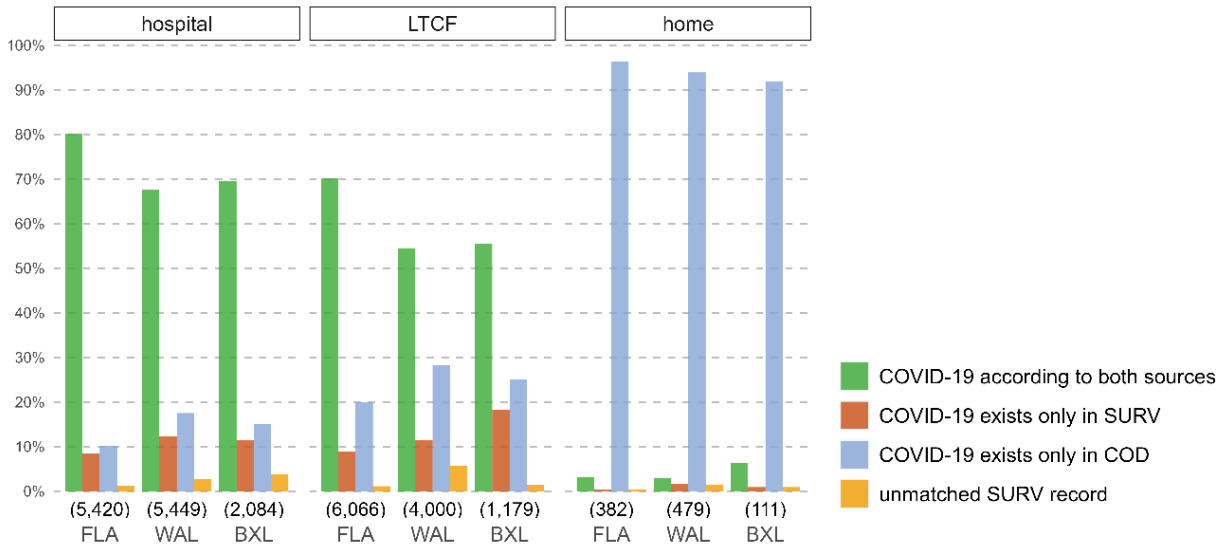
Table 5 : COVID-19 mortality rates per 100,000 by category and scenario, stratified by sex, Belgium, 2020.

	COVID-19 mortality rate per 100,000 (95% CI)	
Category	Males	Females
[1] COVID-19 according to both sources	145.6 (142.5; 148.8)	143.0 (139.9; 146.1)
[2] COVID-19 exists only in SURV	20.6 (19.4; 21.7)	24.3 (23.1; 25.6)
[3] COVID-19 exists only in COD	43.7 (42.0; 45.4)	50.3 (48.4; 52.1)
[4] Unmatched SURV record	5.0 (4.4; 5.6)	5.3 (4.7; 5.9)
Scenario	Males	Females
All SURV records ([1] + [2] + [4])	171.2 (167.8; 174.6)	172.7 (169.3; 176.0)
All COD records ([1] + [3])	189.3 (185.7; 192.9)	193.3 (189.7; 196.8)
All categories combined ([1] + [2] + [3] + [4])	214.9 (211.1; 218.7)	222.9 (219.1; 226.8)

Note: The midyear population of Belgium in 2020 was used to calculate the COVID-19 mortality rates; 17 out of 611 unmatched SURV records had an unknown sex and were removed.

Hospitals exhibited a high level of agreement between SURV and COD, with smaller and comparable proportions of records existing either only in SURV or only in COD (Figure 8). Compared to hospitals, LTCF displayed somewhat lower levels of agreement between SURV and COD, with increased proportions of records existing only in COD, while the proportions of records existing only in SURV remained similar (except for Brussels). In the case of COVID-19 deaths occurring at home, there was negligible reporting through SURV in every region of death, whereas COD reported 1,057 deaths.

Figure 8: Overall distribution of four categories, per place and region of death, after linkage between the COVID-19 mortality epidemiological surveillance database with the cause of death database, Belgium, 2020.



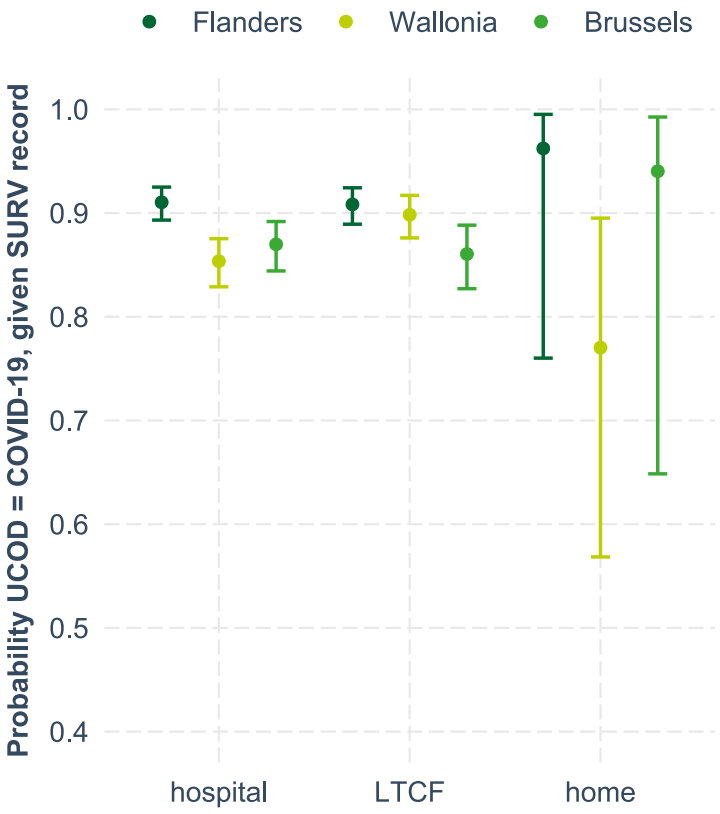
Note: Total sample numbers are provided between brackets.
BXL = Brussels; COD: cause of death database from the death certificates;
FLA = Flanders; LTCF: long-term care facilities;
SURV: COVID-19 mortality epidemiological surveillance database; WAL = Wallonia.

In terms of the regions of death, Flanders showed higher levels of agreement between SURV and COD for both hospitals and LTCF (80.2% and 70.1%). Wallonia and Brussels shared similar agreement proportions for LTCF (54.5% and 55.4% respectively), but their disagreement bars varied.

3.2.3. Integrated analysis – probability analysis

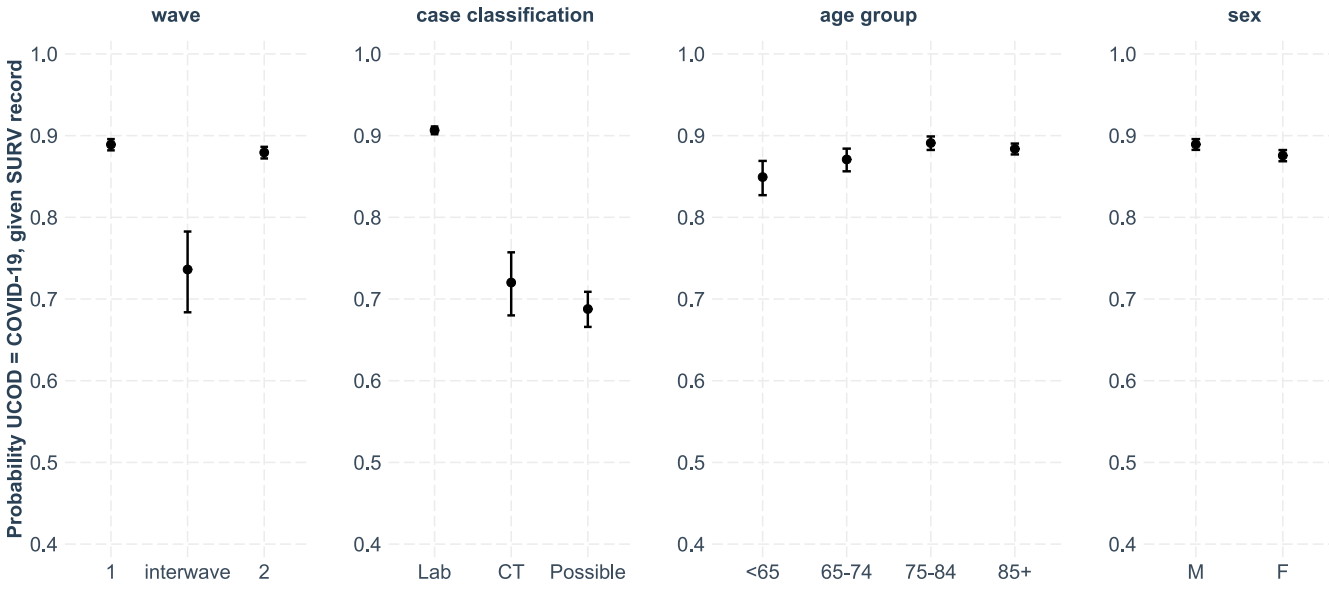
For the factors influencing the probability of a SURV record having a corresponding record in COD indicating COVID-19 as UCOD, the final model did not include the number of conditions mentioned in COD, nor the interaction terms place of death with wave and region of death with wave (Table C-1 in Appendix C). Once a COVID-19 death was reported through SURV, the chance that COD also indicated COVID-19 as UCOD was close to 90%, regardless of the origin of the death (Flanders/Wallonia/Brussels; hospital/LTCF/home) (Figure 9, Table C-2 in Appendix C).

Figure 9: Estimated probabilities that a COVID-19 death in the COVID-19 mortality epidemiological surveillance database has a corresponding record in the cause of death database indicating COVID-19 as underlying cause of death, per place of death, region of death and epidemic wave, multiple logistic regression model 1, Belgium, 2020.



Probabilities were high and comparable between waves 1 and 2, but a drop of about 15% was observed during the interwave period (Figure 10). Case classification appeared to have a significant effect, with a drop of about 20% observed when it concerned radiologically-confirmed cases or possible cases compared to laboratory-confirmed cases. Age groups did have a significant effect, albeit rather small, with a slightly lower chance for the age group under 65 years. Sex had a very minor effect on the outcome.

Figure 10: Estimated probabilities that a COVID-19 death in the COVID-19 mortality epidemiological surveillance database has a corresponding record in the cause of death database indicating COVID-19 as underlying cause of death, per epidemic wave, case classification, age groups and sex, multiple logistic regression model 1, Belgium, 2020.



Note: CT: radiologically-confirmed case; F: female; Lab: laboratory-confirmed case; M: male; SURV: COVID-19 mortality epidemiological surveillance database; Possible: possible case; UCOD: underlying cause of death.

Based on this first model, only case classifications and epidemic waves appeared to have somewhat of an influence on the reason why some SURV records were not found as COVID-19 deaths in COD.

When COVID-19 deaths existed only in SURV (n=2,586), UCODs presented a diverse range of different causes of death, with a category of other ill-defined and unspecified causes of mortality (6.0%) (which incorporates 2% lost certificates), pneumonia with unspecified organism (5.4%), unspecified dementia (5.0%), heart failure (4.2%), other chronic obstructive pulmonary disease (3.4%) and Alzheimer's disease (2.8%) (Table 6). Nevertheless, these percentages were relatively low: considering the top eleven causes, they accounted for 39% of the 2,568 records existing only in SURV.

However, when considering the ranking at ICD-10 chapter level, diseases of circulatory system (I00-I99) explained 24.7% of the COVID-19 deaths existing only in SURV, followed by diseases of the respiratory system (J00-J99, 17.1%), (malignant) neoplasms (C00-D48, 13.3%), and symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99, 10.6%) (Table 7). In addition, 18% of COVID-19 deaths existing only in SURV were deaths *with* COVID-19 in COD.

For the second analysis, factors influencing the probability that a COVID-19 death in COD has a corresponding record in SURV, the final model did not include age groups as covariate, nor the interaction term place of death with region of death (Table D-1 in Appendix D).

Table 6: Highest-frequency conditions with their corresponding ICD-10 codes listed on the death certificate in case of COVID-19 mortality epidemiological surveillance records not having COVID-19 as underlying cause of death in the death certificate, Belgium, 2020.

Condition	ICD-10 code	No. (% of 2,586*)
Other ill-defined and unspecified causes of mortality	R99	156 (6.0)
Pneumonia, organism unspecified	J18	140 (5.4)
Unspecified dementia	F03	129 (5.0)
Heart failure	I50	109 (4.2)
Other chronic obstructive pulmonary disease	J44	87 (3.4)
Alzheimer's disease	G30	72 (2.8)
Malignant neoplasm of bronchus and lung	C34	71 (2.7)
Stroke, not specified as hemorrhage or infarction	I64	71 (2.7)
Acute myocardial infarction	I21	61 (2.4)
Chronic ischaemic heart disease	I25	60 (2.3)
Unspecified fall	W19	53 (2.0)

* Total number of COVID-19 mortality surveillance records not having COVID-19 (U07.1 or U07.2) as underlying cause of death.

Note: conditions appearing less than 2.0% are not shown.

Table 7: Highest-frequency ICD-10 chapters in case of COVID-19 mortality epidemiological surveillance records not having COVID-19 as underlying cause of death in the death certificate, Belgium, 2020.

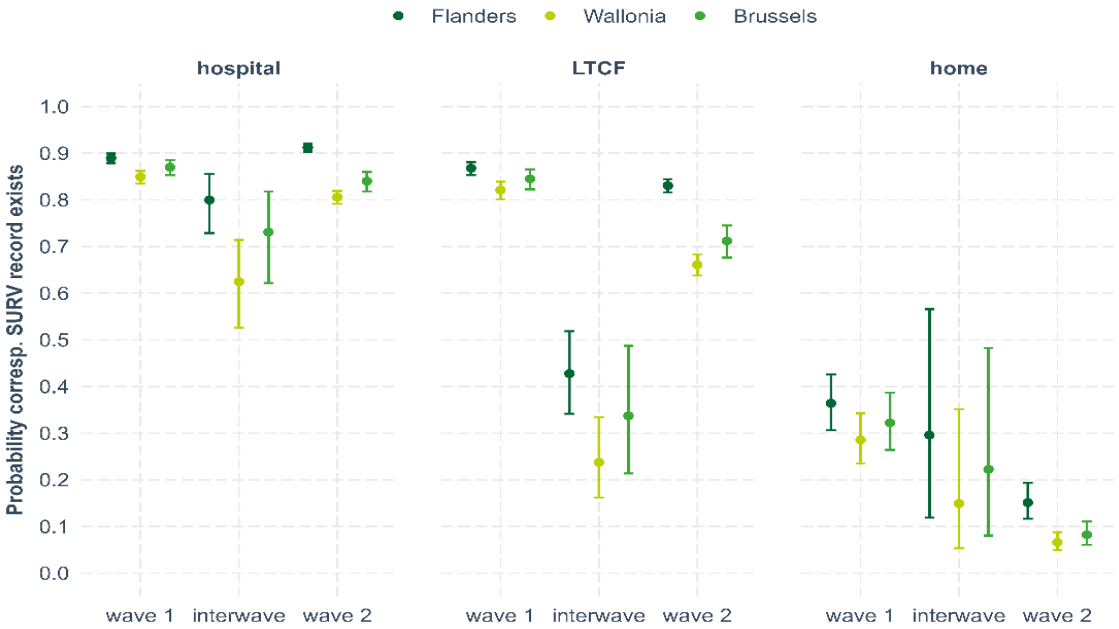
Chapter	No. (% of 2,586*)
Diseases of the circulatory system (I00-I99)	639 (24.7)
Diseases of the respiratory system (J00-J99)	443 (17.1)
Neoplasms (C00-D48)	344 (13.3)
Symptoms, signs and abnormal clinical and lab findings, not elsewhere classified (R00-R99)	274 (10.6)
Mental and behavioural disorders (F00-F99)	152 (5.9)
Diseases of the digestive system (K00-K93)	144 (5.6)
Diseases of the nervous system (G00-G99)	137 (5.3)
External causes of morbidity and mortality (V01-Y98)	128 (4.9)

* Total number of COVID-19 mortality epidemiological surveillance records not having COVID-19 (U07.1 or U07.2) as underlying cause of death according to the cause of death database.

Note: chapters appearing less than 4.0% are not shown.

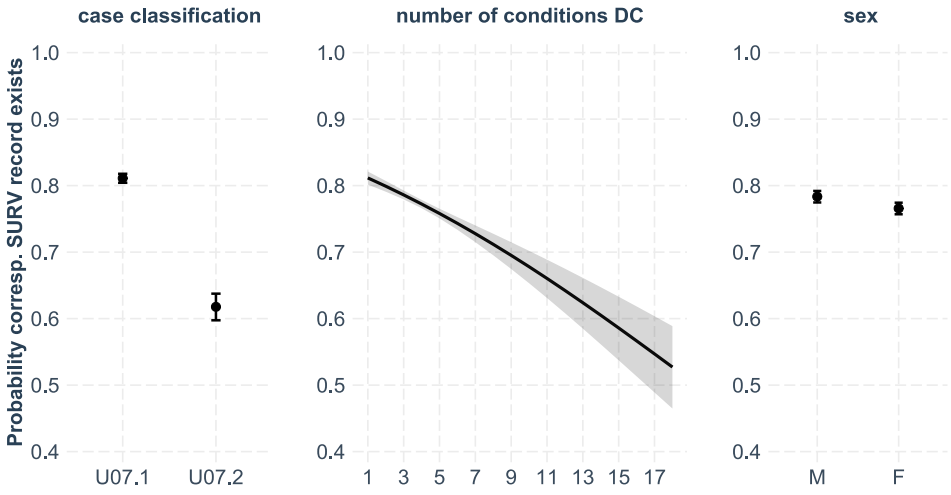
In-hospital deaths had overall high probabilities, mainly in waves 1 and 2 (Figure 11). When considering all hospitals in Belgium, there was on average 83% chance of having a corresponding SURV record, which was higher compared to the 75% for all LTCF in Belgium (Table D-2 in Appendix D). Substantially lower probabilities were observed for deaths occurring at home. A clear drop was observed for the interwave period, which was even more outspoken for LTCF compared to hospitals. The probabilities for LTCF did not seem to fully recover back to their values in wave 1, especially for Wallonia and Brussels.

Figure 11: Estimated probabilities that a COVID-19 death in the cause of death database has a corresponding record in the COVID-19 mortality epidemiological surveillance per place of death, region of death and epidemic wave, multiple logistic regression model 2, Belgium, 2020.



Whether the virus was identified (U07.1) or unidentified (U07.2) appeared to have a considerable effect on the probability that a corresponding SURV record exists (Figure 12). A drop of almost 20% was observed when it concerned an unidentified virus. The effect of sex on the outcome did not appear to be substantial. Furthermore, when only one condition was coded (the UCOD), the probability of having the corresponding SURV record was highest at approximately 80%. However, as the number of listed conditions increased, the probability decreased.

Figure 12: Estimated probabilities that a COVID-19 death in the cause of death database has a corresponding record in the COVID-19 mortality epidemiological surveillance per case classification, number of conditions mentioned on the death certificate and sex, multiple logistic regression model 2, Belgium, 2020.



Note: DC: death certificate; F: female; M: male; SURV: COVID-19 mortality epidemiological surveillance database; U07.1: virus identified; U07.2: virus not identified.

4. Discussion

4.1. Data quality of the COVID-19 mortality epidemiological surveillance database

The proportion of 'perfect' matches (72.1%) between SURV and COD, along with the 23.6% 'good' matches, indicates a robust linkage between both databases. This demonstrates an overall 95.7% reliability in the administrative data quality manually reported by healthcare professionals for COVID-19 deaths in SURV in 2020. This level is remarkable given the real-time requirements and the context of a public health emergency.

However, the presence of 'unsure' matches and 'unmatched' SURV records can be explained both a posteriori by our linkage methodology and a priori, either by data collection errors, manual data entry errors, or other unknown reasons.

Differences in match quality were observed based on the place and region of death, with higher 'perfect' match rates for in-hospital deaths (75.9%) compared to in-NH deaths (67.9%) and deaths in other places (28.8%), and with Flanders having a somewhat higher overall match rate compared to Brussels and Wallonia. These disparities may be attributed to variations in reporting practices and data entry methods, as SURV consisted of nine different data flows, with each their own individual tools for reporting COVID-19 deaths (Renard et al., 2021). Hospitals benefited from a single data collection tool, which resulted in fewer regional differences, unlike surveillance systems in LTCF and other places, which each had four different systems. This highlights the importance of harmonized reporting systems. A harmonized approach to data collection and reporting could significantly improve data quality and reliability.

Matching quality improved in wave 2 compared to wave 1, along with a slight decrease in the percentage of 'unmatched' SURV records in wave 2. In a context where patient load was still high, and healthcare professionals not recovered from wave 1, this improvement might be attributed to potential improvements in adherence to registration procedures. Between 2020 and 2023, Sciensano conducted numerous retrospective investigations among hospitals and LTCF to improve the COVID-19 mortality data for 2020, particularly for the first wave. These investigations suggest that, without these efforts, the match quality for wave 1 would have been lower (Sciensano, 2023b).

The most prevalent disagreements in SURV were on NIS 5 of residence. However, a majority were neighbouring codes, suggesting the possibility of minor human errors in the postcodes entered into SURV. For NH residents we assumed that their postal code of residence was that of the NH itself. However, around 20-25% of NH residents may still be officially domiciled at their former homes for various reasons, which could also explain these mismatches (Surkyn, 2020).

Matches with a disagreeing NIS 5 code of death were also quite prevalent, but 94% of these disagreements can be attributed to the design of the surveillance system in hospitals. Typically, within a hospital network, the principal site reports all COVID-19 deaths for all sites together in SURV and the NIS 5 code of the principal site is considered the NIS 5 code of death. This phenomenon is especially apparent in Brussels, where there are proportionally a lot more hospitals with multiple sites covering various NIS 5 codes of death compared to Flanders and Wallonia. As a result, about 21% of the in-hospital COVID-19 deaths in Brussels actually took place at one of the other hospital sites with a different NIS 5 code. In contrast, for Flanders, this was the case for only about 4%, and for Wallonia, it was approximately 7%. Discrepancies relating to date of birth, which could lead to unmatched records or deduplications issues, may be linked either to manual encoding errors in SURV and in death certificates, or to questionnaires which initially asked only the year of

birth and whose complete date of birth could not be retrieved retrospectively, or to unknown complete date of births on identity cards, a pattern that is more commonly observed among migrant populations. In this case, the first of July is often applied in COD by Statbel.

The few key demographic and administrative variables available in SURV have proven essential for providing epidemiological information on the severity of COVID-19 and, a posteriori, for enabling effective record linkage with COD, all while minimising the administrative reporting burden for field personnel.

Overall, these findings highlight the need to enhance data collection tools within healthcare institutions and improve reporting interfaces for centralisation purposes during the setup of ad hoc surveillance systems. Moving forward, there is a need for comprehensive, flexible, and harmonised reporting practices, particularly in crisis situations requiring rapid deployment of surveillance capabilities. This includes exploring alternatives to reduce the manual encoding of administrative data.

4.2. Comparison between both data sources

4.2.1. Global Analysis

The global comparative analysis revealed that SURV achieved a high coverage of 90% in 2020. A comparison with excess mortality suggested that SURV may have underestimated COVID-19 deaths (Bustos Sierra et al., 2021). The global comparison with death certificates confirms this discrepancy. However, these differences come with various nuances, which became apparent through one-to-one matching.

Hospitals showed higher agreement levels between both databases, likely due to their mandatory participation (Royal Decree of 30.04.2020; Moniteur Belge 2020). Notably, SURV reported slightly more in-hospital deaths in Brussels (103%), which can be explained by a hospital located in Brussels reporting data from another substantial site located in Wallonia within the same network. In contrast, for LTCF, the level of agreement was slightly lower, though still remarkable given that there was no legal obligation for LTCF to participate in COVID-19 surveillance. However, NH were strongly encouraged to participate and were contacted in case of non-participation by the RegHA. Overall, 1,529 NH out of the 1,542 NH (99%) participated at least once in COVID-19 surveillance in 2020 (Vandael et al., 2022).

However, a clear gap in reporting through the surveillance system was identified for COVID-19 deaths occurring at home, despite regional procedures (for general practitioners and for dealing with the death of a patient with COVID-19) (Sciensano, 2020b, 2022). Possible explanations include a lack of reminders, workload, and assumptions that severe cases would be hospitalised. Nevertheless, we have noticed inconsistencies in COD records for individuals who died at home, as some of these cases were also recorded in SURV with a listed place of death as a hospital or NH, creating a discrepancy in the reported place of death between the two databases.

4.2.2. Integrated Analysis – Comparative Analysis

Deeper underlying differences were discovered through the integrated comparative analysis. The coverage of SURV decreased from 90% to 83.9%.

COVID-19 deaths not reported in SURV (n=5,411) may be explained by various factors, including the heavy workload faced by healthcare professionals and institutions, staff shortages, internal organisation challenges, fatigue, disinterest in statistical reporting, and the voluntary nature of the survey, and the emotional challenge mainly for professionals in LTCF encoding statistical data for individuals who have tragically passed away.

ICD mortality coding followed the same algorithm in all regions but may have

inflated COVID-19 death counts in COD, as the WHO technical note for mortality certification was followed (World Health Organization, 2020a). This guidance stipulated that COVID-19 should be recorded on the death certificate by the certifying doctor for all individuals in whom the disease caused, was presumed to have caused, or contributed to death. During the ICD mortality coding process, COVID-19 was removed from UCOD in cases of non-natural death (e.g. trauma) and myocardial infarction. Nevertheless, COVID-19 needed to be counted independently of pre-existing conditions that were suspected of triggering a severe course of COVID-19. Given the intense public health requirements for data, COVID-19 was not considered as an obvious consequence of anything else (similar to coding rules applied for influenza). These instructions were to be followed regardless of whether they were considered medically accurate or not. This approach, while ensuring consistency in reporting COVID-19 deaths, may lead to potential overestimation of COVID-19 deaths in COD.

After merging the regional COD databases, Statbel excluded individuals who were not part of the Belgian legal population (such as tourists, asylum seekers, undocumented individuals, etc.) from the official national COD database. This was anticipated in the final database of SURV, where 55 in-hospital COVID-19 deaths whose postal code of residence was abroad, were excluded. Conversely, this was not done for NH surveillance data, which could have included some COVID-19 deaths, excluded in COD, for NH likely located in border areas.

The 2,586 COVID-19 deaths in SURV not reported as UCOD in COD are less evident to justify. Explanations include issues in data collection (e.g. misunderstandings of inclusion criteria in SURV and misunderstanding the guidance to fill a death certificate for COVID-19), death certificate completion (e.g. hesitations to list COVID-19 as UCOD for possible cases, and the complexity of differentiating a death *due* to COVID-19 from a death *with* COVID-19), legibility problems (e.g. illegible or blank causes of death on death certificates), death certificate processing (e.g. the medical part never reaching RegHA), and coding challenges (e.g. 61 COVID-19 deaths reported in SURV had myocardial infarction as UCOD but also COVID-19 in the other list of conditions, which could be a consequence of a posteriori coding guidelines).

The COVID-19 mortality rates per sex illustrate the implications of using one database over the other. When relying on SURV alone, the COVID-19 mortality rate stands at 171.2 for males and 172.7 for females. However, using COD results in higher rates, with figures of 189.3 for males and 193.3 for females. Considering all COVID-19 deaths, whether from SURV, COD, or both, yields even higher rates of 214.9 for males and 222.9 for females. The true COVID-19 mortality rates likely lie somewhere in between these figures.

4.2.3. Integrated Analysis – Probability Analysis

From the first probability analysis, we observed that once a COVID-19 death was reported through SURV, the chance that COD also indicated COVID-19 as UCOD was very high (around 90%), regardless of the origin of death (Flanders/Wallonia/Brussels; hospitals/LTCF/home).

When a COVID-19 death in SURV was not indicated as COVID-19 in UCOD, our analysis revealed that such deaths were often reported as other ill-defined and unspecified causes of mortality, or as pneumonia with an unspecified organism, which could be compatible with a COVID-19-related death. However, on a chapter level, diseases of the respiratory system ranked second to diseases of the circulatory system. Yet, it is known that patients suffering from COVID-19 are also at a higher risk of cardiovascular complications that could precipitate their death (Brogi et al., 2022; Goyal et al., 2021). In some cases, where both cancer and COVID-19 were present, cancer was deemed the primary cause of death, overshadowing COVID-19.

Furthermore, 10.6% of cases were categorised under 'symptoms, signs and abnormal clinical and laboratory findings that are nowhere else classified', which includes unknown causes of death, and administrative errors, such as lost death certificates. Lastly, 18% of records existing only in SURV were deaths *with* COVID-19 in COD, possibly due to inclusion criteria misunderstandings in SURV, death certificate completion discrepancies, or our linkage methodology. But it remains quite unclear why for a relatively small percentage of them, they ended up in this category.

The second probability analysis revealed that a higher number of conditions listed in COD led to lower probabilities of having a corresponding SURV record for a COVID-19 record in COD. This reflects that the presence of multiple comorbidities makes it more difficult to pinpoint the exact UCOD, and subsequently reduces the chances that there will be a COVID-19 SURV record.

This probability was also influenced by the place and region of death. Hospitals displayed a 83% average of having a matching SURV record, followed by LTCF at 75%, and at-home deaths at 15%. This may be partly explained by differences in operational structures. Hospitals, with mandatory participation and reminders for timely data reporting, had more streamlined reporting processes. In contrast, LTCF are more diverse, decentralised, and less supported to participate in sustained long-term surveillance. Additionally, an investigation by Sciensano in 2021 targeting data from hospitals may have contributed to the higher hospital probabilities. Flanders typically had slightly higher probabilities than Brussels and Wallonia, revealing regional differences, yet the underlying factors driving these variations remain largely unknown and require further investigation.

During the calmer interwave period in July and August 2020, SURV underestimated COVID-19 deaths, likely due to fatigue and a diminished sense of urgency after an intense first wave, the perceived futility of reporting low numbers, and staffing reductions due to holidays and absenteeism.

Case classification, as revealed by both probability analyses, heightened the chance of disparities between data sources, with a 20% drop when it concerned an unidentified virus (U07.2) or radiologically-confirmed/possible cases. Case classification was rarely mentioned on death certificates, making coding choices difficult. If there was no information on case classification, the place and date of death guided the decision. If the person had died in a hospital, the death was automatically coded as U07.1. In Wallonia, for deaths that occurred outside of the hospital and when COVID-19 was mentioned without any test information, U07.2 was assigned up until 1 May 2020. For deaths occurring after that date, U07.1 was assigned. Statbel adopted this logic for Flanders and Brussels to ensure comparability. Deaths of radiologically-confirmed cases in SURV were found in COD under both ICD-10 codes.

Neither SURV nor COD can be regarded as the gold standard for COVID-19 mortality statistics. In 2020, SURV underestimated COVID-19 deaths in general by approximately 10%, with a notable relative underestimation of at-home and interwave deaths. In addition to the ability to provide data in real-time, SURV offered a more comprehensive overview of the place of death, with in-hospital and in-LTCF deaths reported by the respective surveillances. This design ensured a high degree of certainty regarding the place of death, and, to a somewhat lower degree, the type of residence. SURV remains available in the open data of Sciensano covering the period March 2020 to 30 June 2023.

On the other hand, filling in death certificates is seen as an administrative burden by the certifying doctor, who may not always be the treating physician. There is also no formal dedicated training for death certificate completion. Determining the exact cause of death is often not that simple, especially for older individuals with multiple comorbidities, particularly during the initial stages of the epidemic when COVID-19 was not well understood yet. SURV and COD were constructed based on the WHO definition of a COVID-19 death for surveillance and for medical

certification (death certificates) respectively. The broad definition for the medical certification could inherently have led to an inflation of COVID-19 death counts in COD. This database provided awareness about the number of at-home COVID-19 deaths, it missed some COVID-19 deaths, possibly overestimated others, and was less precise about place of death and NH residents. Nonetheless, COD is the official death statistics for Belgium and is used for research requiring linkages with external databases.

4.3. Linkage methodology

We developed a custom-designed probabilistic record linkage method for matching SURV with COD. The method involved certain manual interventions and decisions guided by intuition and experience. Consequently, a certain level of bias is inherent to this approach, potentially leading to both false positives and false negatives in the final results.

Various deduplication methods exist to ensure a one-to-one merge, like selecting matches with the highest posterior probability or using linear programming to optimize log-likelihood over the entire set of record pairs (Jaro, 1989). However, we found that this first approach proved to be overly simplistic, especially when multiple possible matches exhibited similar posterior probabilities. The second approach offers a rigorous mathematical framework, but it does not scale well with the number of record pairs. This led to the need for a more nuanced strategy to resolve duplicates efficiently. We ultimately opted for a three-level deduplication algorithm, which struck a balance between the simplicity of choosing the highest posterior probability match and the computational demands of the linear programming approach.

4.4. Lessons learned and recommendations for specific real-time mortality surveillance

Our experience with the epidemiological surveillance of COVID-19 deaths underscores the important role of individual-level data in maintaining data quality and gaining a clear understanding of the profiles of the deceased. Particularly in the case of respiratory pathogens, coverage in different places of death beyond hospitals is essential for comprehensive surveillance. Experience with COVID-19 demonstrated the immense risks for LTCF. Belgium was fortunate to have initiated exhaustive surveillance in LTCF from the start of the epidemic. Additionally, the introduction of a ‘possible case’ classification for cases and deaths in the questionnaires helped identify the catastrophic situation in NH. Deaths of possible cases were included in the official statistics, both prospectively and retrospectively, from 30 March 2020. The highest number of these cases occurred during the first six weeks of the epidemic, coinciding with the deployment of laboratory tests outside of hospitals. In COD, the number of COVID-19 deaths at home, although proportionately lower than in hospitals and LTCF, revealed that 1,057 people died from COVID-19 in Belgium overall, compared to 52 in SURV, indicating a clear gap. Improved surveillance is necessary, especially in anticipation of potential future diseases characterised by more rapid lethality. We emphasise the importance of a carefully designed surveillance questionnaire, harmonised across all dataflows, and meticulous data reporting. In addition, having a flexible data collection approach that accommodates the evolving characteristics of an epidemic remained important for robust surveillance.

To strengthen precise surveillance of cause-specific mortality in real-time, investments in integrated monitoring systems are warranted to ensure readiness and stability during public health emergencies, by having a streamlined data flow available upfront. By anticipating and establishing necessary linkages in advance through

the national number, such as those for test results, vaccination records, socio-economic data, and death certificates, the need to hastily create ad-hoc solutions during emergencies is mitigated.

Leveraging existing systems for this purpose means that financial investments in digital infrastructure and personnel can be limited and justified by the added benefits for public health. Improved surveillance can lead to better disease prevention, more informed public health policy decisions, and ultimately, cost savings by preventing larger outbreaks and mitigating their impact.

The benefits of a well-maintained system during non-crises times include better routine public health surveillance, faster response times to emerging threats, and the ability to provide accurate, real-time data to policymakers.

Conclusion

This study is the first to conduct a comprehensive evaluation of the quality of reported COVID-19 deaths through the surveillance system and perform a comparative analysis with COD for 2020. Our research has shed light on the strengths and limitations of both databases, offering valuable insights into the intricate processes involved in capturing and documenting COVID-19-related deaths during a public health emergency.

Our findings confirmed SURV's overall reliability. The presence of 'unsure' matches and 'unmatched' SURV records highlighted the importance of meticulous data reporting and a well-designed surveillance questionnaire, with particular emphasis on key demographic and administrative variables. Data quality variations across regions and healthcare facilities may have stemmed from variations in reporting practices and data entry methods.

The global comparative analysis demonstrated an overall coverage of 90% by SURV compared to COD, with variations by setting. Hospitals displayed higher agreement than LTCF, and a clear gap existed in SURV for COVID-19 deaths that occurred at home.

The integrated comparative analysis delved deeper into the discrepancies between both databases, revealing that 83.9% of the SURV records had COVID-19 listed as the UCOD in COD, and 75.4% of COVID-19 deaths reported in COD were identified in SURV. These disparities are influenced by various factors, including the time period, diagnostic certainty, region and place of death.

Both databases have limitations but provide valuable insights for policy formulation, even though they cannot be considered the gold standard for COVID-19 mortality statistics. The quality of SURV depended on the willingness and motivation of data providers, resulting in decreased data quality and reporting during periods of reduced COVID-19 activity.

Accurately pinpointing the precise cause of death, particularly among elderly individuals with complex health profiles, remains intricate. The delayed availability of the official death certificate data by three years renders them unsuitable for real-time monitoring, and the WHO's broad definition of COVID-19 deaths raises concerns about potential overestimation.

Our study emphasizes the need for enhanced surveillance systems capable of adapting and responding efficiently to public health emergencies. Three paths are envisaged: (1) Ideally, we envision an automated registry based on electronic health records that reduces the workload for healthcare professionals. This registry should integrate monitoring in hospitals, LTCF, and the community through an interface

like HealthData (Sciensano, 2023c). Achieving this ideal scenario requires the development and implementation of a robust digital infrastructure at regional and federal health authority level for real-time data linkage and monitoring. (2) In the short term, efforts should focus on establishing a linkage between test databases and all-cause mortality data within a 28-day frame (similar to the UK experience). Moreover, this system can be utilised year-round to monitor COVID-19, but also other infectious diseases for which test data are available. (3) Digitising death certificates should be prioritized to improve the efficiency, reliability, and timeliness of cause-of-death data in Belgium. However, this initiative faces challenges, including high costs, coordination issues and data safety concerns.

To sum up, it is crucial to interpret COVID-19 mortality data carefully due to its nuanced nature. The insights gained from this study highlight the need to invest in methodologies for collecting data on cause-specific deaths in real-time in response to evolving public health challenges.

Glossary of main acronyms

COD: cause of death database

ECDC: European Centre for Disease Prevention and Control

SURV: epidemiological surveillance database

LTCF: long-term care facilities

NH: nursing homes

WHO: World Health Organisation

RegHA: regional health authorities

Statbel: Belgian Statistical Office

GGD: municipal health service

UCOD: underlying cause of death

ICD-10: WHO international classification of diseases, 10th version

GDPR: General Data Protection

References

Blakely, T., & Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predictive value. International Journal of Epidemiology, 31(6), 1246–1252. https://doi.org/10.1093/ije/31.6.1246

Brogi, E., Marino, F., Bertini, P., Tavazzi, G., Corradi, F., & Forfori, F. (2022). Cardiac complications in patients with COVID-19: a systematic review. Journal of Anesthesia, Analgesia and Critical Care, 2(1), 18. https://doi.org/10.1186/s44158-022-00046-7

Bustos Sierra, N., Bossuyt, N., Braeye, T., Haarhuis, F., Peeters, I., Proesmans, K., Renard, F., Scohy, A., Vanhaverbeke, M., Vermeulen, M., Vernemmen, C., & Van der Heyden, J. (2021). Excess mortality during the first and second wave of the COVID-19 epidemic in Belgium. https://www.sciensano.be/sites/default/files/be-momo_report_excess_mortality_during_the_covid-19_epidemic_in_belgium.pdf

Bustos Sierra, N., Bossuyt, N., Braeye, T., Leroy, M., Moyersoen, I., Peeters, I., Scohy, A., Van der Heyden, J., Van Oyen, H., & Renard, F. (2020). All-cause mortality supports the COVID-19 mortality in Belgium and comparison with major fatal events of the last century. Archives of Public Health, 78, 1–8.

Clanché, F., & Caserio-Schönemann, C. (2023). La surveillance de la mortalité en lien avec la COVID-19 au quotidien: résultats, forces et faiblesses. https://www.rencontressantepubliquefrance.fr/wp-content/uploads/2023/06/7-CLANCHE-CASERIO-SCHONEMANN.pdf

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics, 33(18), 2938–2940.

Dequeker, S., Callies, M., Catteau, L., Int Panis, L., Islamaj, E., Klamer, S., Latour, K., Pauwels, M., Vernemmen, C., Mahieu, R., Masson, H., Savsin, M., De Clercq, E., Thomas, M., Catry, B., & Vandael, E. (2023). COVID-19 Clusters in Belgian Nursing Homes: Impact of Facility Characteristics and Vaccination on Cluster Occurrence, Duration and Severity. Viruses, 15(1), 232. https://doi.org/10.3390/v15010232

Dequeker, S., Latour, K., Islamaj, E., Int Panis, L., Callies, M., Catteau, L., Catry, B., & Vandael, E. (2020). COVID-19 surveillance in residential institutions. https://www.sciensano.be/sites/default/files/20210111_protocol_covid-19_surveillance_in_residential_institutions_v4.2_final.pdf

Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. American Political Science Review, 113(2), 353–371.

Enamorado, T., Fifield, B., & Imai, K. (2020). FastLink: fast probabilistic record linkage with missing data 0.6. CRAN: Enamorado, Ted.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183–1210.

Garcia, J., Torres, C., Barbieri, M., Camarda, C.-G., Cambois, E., Caporali, A., Meslé, F., Poniakina, S., & Robine, J.-M. (2021). Differences in COVID-19 mortality: Implications of imperfect and diverse data collection systems. Population, 76(1), 35–72.

Goyal, P., Reshetnyak, E., Khan, S., Musse, M., Navi, B. B., Kim, J., Allen, L. A., Banerjee, S., Elkind, M. S. V, & Shah, S. J. (2021). Clinical characteristics and outcomes of adults with a history of heart failure hospitalized for COVID-19. Circulation: Heart Failure, 14(9), e008354.

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record linkage: Current practice and future directions. CSIRO Mathematical and Information Sciences Technical Report, 3, 83.

Hope, A. (2020). New coronavirus infection reported in Antwerp. https://www.brusselstimes.com/all-news/belgium-all-news/97873/new-coronavirus-infection-reported-in-antwerp

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, 84(406), 414–420.

Johns Hopkins Coronavirus Resource Center. (2020). Johns Hopkins University Mortality Analyses - COVID-19. https://coronavirus.jhu.edu/data/mortality

Karlinsky, A., & Kobak, D. (2021). Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. ELife, 10. https://doi.org/10.7554/eLife.69336

Molenberghs, G., Faes, C., Verbeeck, J., Deboosere, P., Abrams, S., Willem, L., Aerts, J., Theeten, H., Devleesschauwer, B., & Sierra, N. B. (2022). COVID-19 mortality, excess mortality, deaths per million and infection fatality ratio, Belgium, 9 March 2020 to 28 June 2020. Eurosurveillance, 27(7), 2002060.

Moniteur Belge. (2018). Loi portant création de Sciensano. https://www.ejustice.just.fgov.be/cgi/article_body.pl?language=fr&caller=summary&pub_date=18-03-21&numac=2018011241

Moniteur Belge. (2020). Arrêté royal concernant un flux d'information correct et en temps voulu sur les chiffres de patients COVID-19, la capacité de traitement dans les hôpitaux et les stocks de matériel de protection individuelle (numac: 2021032850). https://www.ejustice.just.fgov.be/cgi/article_body.pl?language=fr&caller=summary&pub_date=21-09-20&numac=2021032850

Morgan, D., Ino, J., Paolantonio, G. Di, & Murtin, F. (2020). Excess mortality: Measuring the direct and indirect impact of COVID-19. OECD Health Working Papers, 122. https://doi.org/10.1787/c5dc0c50-en

Potz, N., Powell, D., Lamagni, T. L., Pebody, R., Bridger, D., & Duckworth, G. (2010). Probabilistic record linkage of infection records and death registrations: a tool to strengthen surveillance. Statistical Communications in Infectious Diseases, 2(1).

Renard, F., Scohy, A., Van der Heyden, J., Peeters, I., Dequeker, S., Vandael, E., Van Goethem, N., Dubourg, D., De Viron, L., & Kongs, A. (2021). Establishing an ad hoc COVID-19 mortality surveillance during the first epidemic wave in Belgium, 1 March to 21 June 2020. Eurosurveillance, 26(48), 2001402.

Rijksoverheid. (2021a). Wat is het COVID-19-sterftecijfer op basis van doodsoorzaak? https://coronadashboard.rijksoverheid.nl/artikelen/wat-is-het-covid-19-sterftecijfer-op-basis-van-doodsoorzaak

Rijksoverheid. (2021b). Wat zegt het aantal meldingen van sterfgevallen? https://coronadashboard.rijksoverheid.nl/artikelen/wat-zegt-het-aantal-meldingen-sterfgevallen

Sciensano. (2020a). COVID-19 - Gevalsdefinitie en testing. https://covid-19.sciensano.be/nl/covid-19-gevalsdefinitie-en-testing

Sciensano. (2020b). Procédure pour les médecins généralistes en cas de suspicion de maladie COVID-19. https://covid-19.sciensano.be/sites/default/files/Covid19/ARCHIVE_COVID-19_procedure_GP_FR.pdf

Sciensano. (2022). Procédure pour la prise en charge du décès d'un patient atteint du COVID-19. https://covid-19.sciensano.be/sites/default/files/Covid19/COVID19_procedure_deaths_20220629_FR.pdf

Sciensano. (2023a). COVID-19. https://epistat.sciensano.be/covid/

Sciensano. (2023b). COVID19BE OPEN DATA CODEBOOK. https://epistat.sciensano.be/COVID19BE_codebook.pdf

Sciensano. (2023c). HealthData. https://healthdata.sciensano.be/en/home

Sciensano. (2023d, August 16). Belgium COVID-19 Dashboard. https://lookerstudio.google.com/embed/reporting/c14a5cfc-cab7-4812-848c-0369173148ab/page/ZwmOB

Seghezzo, G., Allen, H., Griffiths, C., Pooley, J., Beardsmore, L., Caul, S., Glickman, M., Clare, T., Dabrera, G., & Kall, M. (2023). Comparison of two COVID-19 mortality measures used during the pandemic response in England. International Journal of Epidemiology, dyad116. https://doi.org/10.1093/ije/dyad116

Statbel. (2023, August 18). Causes of death. https://statbel.fgov.be/en/themes/population/mortality-life-expectancy-and-causes-death/causes-death#documents

Surkyn, J. (2020). Oversterfte in de Belgische woonzorgcentra. https://interfacedemography.be/covid-19/oversterfte-in-de-belgische-woonzorgcentra/

UK Health Security Agency. (2023). Changes to the way we report on COVID-19 deaths. https://ukhsa.blog.gov.uk/2023/01/27/changes-to-the-way-we-report-on-covid-19-deaths/

Van Goethem, N., Vilain, A., Wyndham-Thomas, C., Deblonde, J., Bossuyt, N., Lernout, T., Rebolledo Gonzalez, J., Quoilin, S., Melis, V., & Van Beckhoven, D. (2020). Rapid establishment of a national surveillance of COVID-19 hospitalizations in Belgium. Archives of Public Health, 78(1), 121. https://doi.org/10.1186/s13690-020-00505-z

Vandael, E., Latour, K., Islamaj, E., Panis, L. I., Callies, M., Haarhuis, F., Proesmans, K., Devleesschauwer, B., Rebolledo Gonzalez, J., Hannecart, A., Mahieu, R., de Viron, L., De Clercq, E., Kongs, A., Hammami, N., François, J.-M., Dubourg, D., Henz, S., Catry, B., & Dequeker, S. (2022). COVID-19 cases, hospitalizations and deaths in Belgian nursing homes: results of a surveillance conducted between April and December 2020. Archives of Public Health, 80(1), 45. https://doi.org/10.1186/s13690-022-00794-6

Wang, H., Paulson, K. R., Pease, S. A., Watson, S., Comfort, H., Zheng, P., Aravkin, A. Y., Bisignano, C., Barber, R. M., & Alam, T. (2022). Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. The Lancet, 399(10334), 1513–1536.

World Health Organization. (2020a). Medical certification, ICD mortality coding, and reporting mortality associated with COVID-19. https://iris.who.int/bitstream/handle/10665/332297/WHO-2019-nCoV-Mortality_Reporting-2020.1-eng.pdf?sequence=1

World Health Organization. (2020b). Public health surveillance for COVID-19: interim guidance. https://iris.who.int/bitstream/handle/10665/333752/WHO-2019-nCoV-SurveillanceGuidance-2020.7-eng.pdf?sequence=1&isAllowed=y

World Health Organization. (2023, August 16). WHO Coronavirus (COVID-19) Dashboard. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/

Zhang, Z. (2016). Model building strategy for logistic regression: purposeful selection. Annals of Translational Medicine, 4(6).

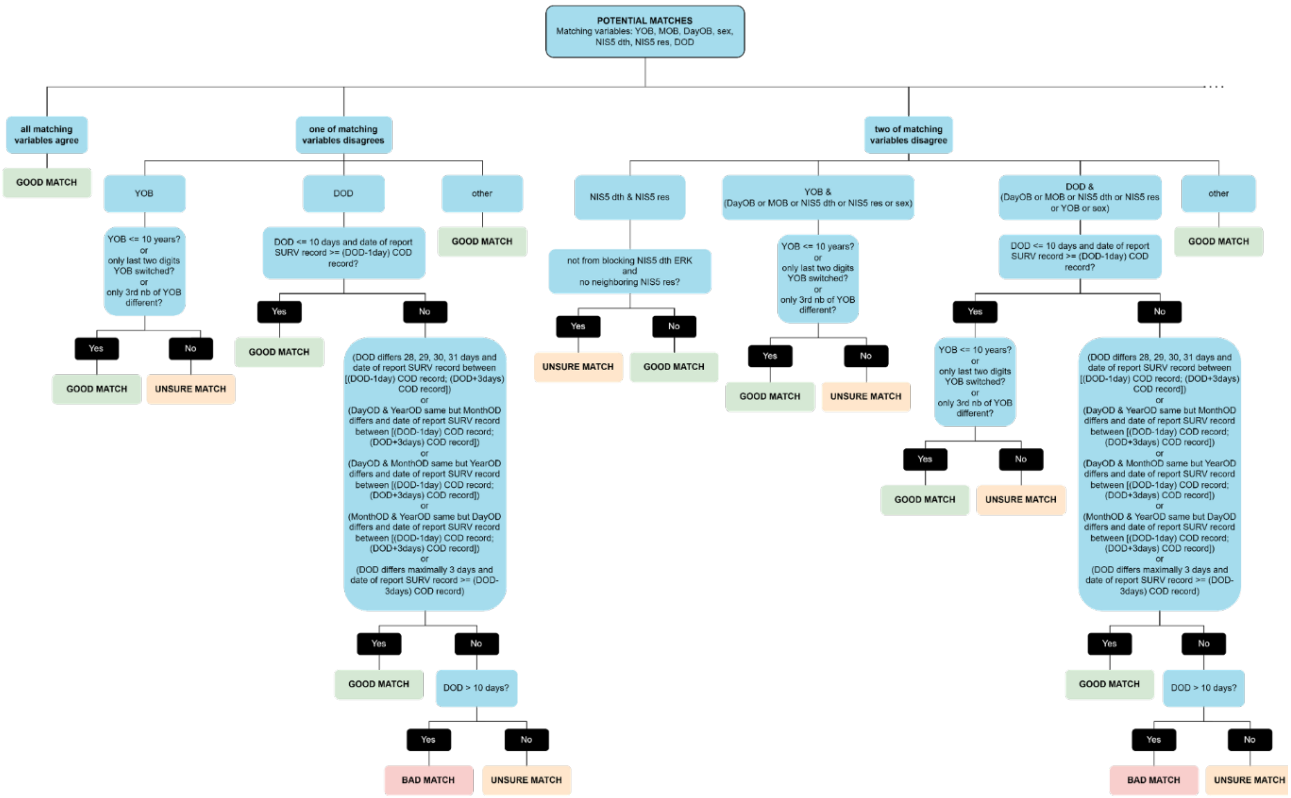
Appendix A

Table A-1: Variables of interest in the COVID-19 mortality epidemiological surveillance database with their formats, degree of missingness, and abbreviations used in the main text, Belgium, 2020.

Variable	Abbreviation	Format	Missingness of variable (n=19,801)	
Date of birth	DOB	YYYY-MM-DD	1,296	(6.55%)
Year of birth	YOB	YYYY (integer)	25	(0.13%)
Month of birth	MOB	MM (integer)	1,296	(6.55%)
Day of birth	DayOB	DD (integer)	1,296	(6.55%)
Date of death	DOD	YYYY-MM-DD	0	(0.00%)
Sex	Sex	1 (male), 2 (female)	17	(0.09%)
NIS 5 of death	NIS 5 dth	5-digit	4	(0.02%)
NIS 5 of residence	NIS 5 res	5-digit	192	(0.97%)

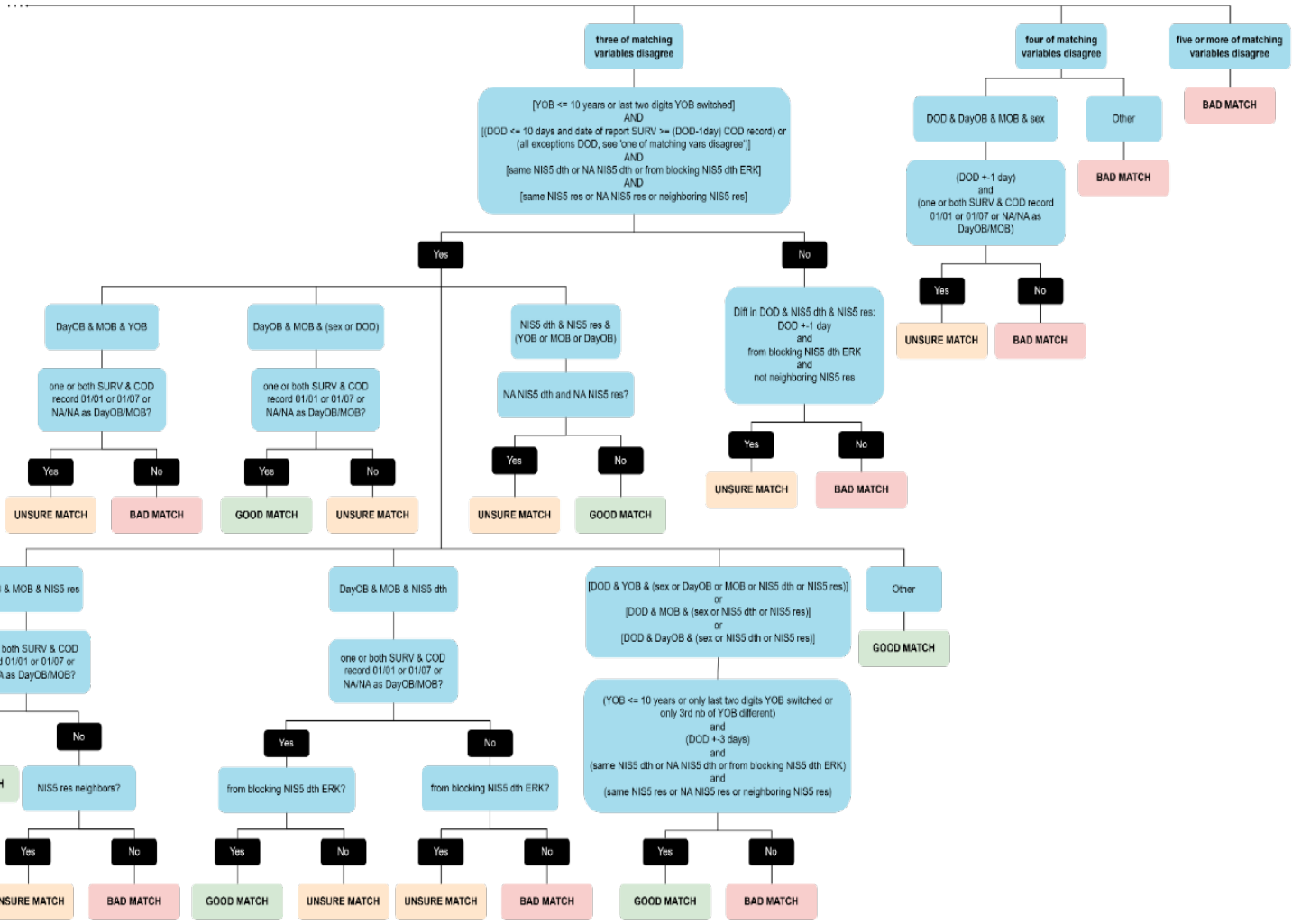
Appendix B

Figure B-1: Part 1 of the decision tree used to classify potential matches between the COVID-19 mortality epidemiological surveillance database and the cause of death database, Belgium, 2020.



Note: COD = cause of death database; DayOB = day of birth; DayOD = day of death; DOD = date of death; ERK = identifier of a hospital (network); MOB = month of birth; MonthOD = month of death; NIS5 dth = NIS 5 code of place of death; NIS5 res = NIS 5 code of place of residence; SURV = COVID-19 mortality epidemiological surveillance database; YOB = year of birth; YearOD = year of death

Figure B-2: Part 2 of the decision tree used to classify potential matches between the COVID-19 mortality epidemiological surveillance database and the cause of death database, Belgium, 2020.



Note: COD = cause of death database; DayOB = day of birth; DayOD = day of death; DOD = date of death; ERK = identifier of a hospital (network); MOB = month of birth; MonthOD = month of death; NIS5 dth = NIS 5 code of place of death; NIS5 res = NIS 5 code of place of residence; SURV = COVID-19 mortality epidemiological surveillance database; YOB = year of birth; YearOD = year of death.

Duplication

A three-level deduplication algorithm was constructed. The first level involved:

1. Split up the set of matches ('good' and 'unsure' together) into its smaller connected subsets. A connected subset can be seen as a bigraph where each node of the first set is a SURV record, and each node of the other set is a COD record. Vertices between both sets represent a possible match ('good' or 'unsure') between two records (Figure B-3 in Appendix B).

2. For every bigraph separately, retain all 'perfect' matches and discard matches for which the SURV or COD record was already part of a 'perfect' match. 'Perfect' matches may still not be fully unique, as one SURV record could have two 'perfect' matches with COD records.

3. Retain matches where only the sex is different or where only the date of death is one day different. Discard the other matches involving either the SURV or COD records.

4. Based on the leftover set of matches in the connected subset, choose matches having only one difference in their matching variables above matches having three differences in their matching variables.

The result of this first level of deduplication was a set of deduplicated matches and a set of unduplicated matches. The unduplicated matches underwent level two in the deduplication algorithm:

1. Split up matches into their connected subsets.

2. Run Winkler's linear programming solution (implemented in the *fastLink* R package) on each connected subset. Instead of using the posterior probabilities to minimise the cost function in this algorithm, choose simple weights indicating the quality of the match (3 = 'good' match, 2 = 'unsure' match). This avoids suboptimal linkage results in case multiple matches are present with similar posterior probabilities. Additionally, if the SURV record of the chosen match also has the same weight (e.g. 'good' match) with other COD records, retain all matches for that SURV record with that weight.

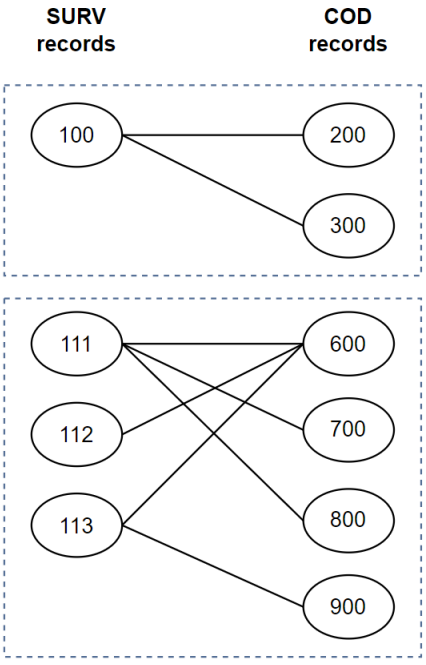
This again resulted in a set of deduplicated matches (added to the set of level one) and unduplicated matches. The unduplicated matches were sent to the third level of the deduplication algorithm:

1. Split up matches into their connected subsets.

2. For every connected subset, consider the unique identifier of COD records in every connected subset, and only keep the 'best' quality connections. If one unique identifier is matched with three SURV records, two labelled as 'good', and one labelled as 'unsure', the unsure match gets thrown out.

The resulting set of deduplicated matches was added to the set obtained after levels one and two.

Figure B-3: Two connected subsets (bigraphs) examples to match COVID-19 mortality epidemiological surveillance records with cause of death records, Belgium, 2020.



Note:
All COVID-19 mortality surveillance and cause of death records that were directly or indirectly connected were part of a connected subset. Numbers within each node represent fictitious unique record numbers.

Appendix C

Table C-1: Estimated regression parameters (B), standard errors (SE), and exponentiated estimates (exp(B)) with 95% confidence intervals (CI) of factors influencing the probability of a COVID-19 mortality epidemiological surveillance record having a corresponding record in the cause of death database indicating COVID-19 as underlying cause of death, multiple logistic regression model 1, Belgium 2020.

Independent variable	Reference category	Estimate (B)	SE	Exp(B)	95% CI for exp(B)
(Intercept)		2.32	0.10	10.16	8.38 – 12.37
Place of death	Hospital				
LTCF		-0.03	0.07	0.98	0.84 – 1.13
Home		0.92	1.06	2.51	0.46 – 46.76
Region of death	Flanders				
Wallonia		-0.56	0.07	0.57	0.50 – 0.65
Brussels		-0.42	0.09	0.66	0.55 – 0.78
Epidemic wave	Wave 1				
Interwave		-1.06	0.13	0.35	0.27 – 0.45
Wave 2		-0.09	0.05	0.91	0.82 – 1.00
Case classification	Laboratory-confirmed				
Radiologically-confirmed		-1.33	0.10	0.27	0.22 – 0.32
Possible case		-1.48	0.06	0.23	0.20 – 0.26
Sex	Male				
Female		-0.13	0.05	0.88	0.80 – 0.96
Age groups	<65				
65-74		0.18	0.10	1.20	0.98 – 1.46
75-84		0.37	0.09	1.45	1.21 – 1.73
85+		0.30	0.09	1.35	1.13 – 1.61
Place of death : region of death					
LTCF : Wallonia		0.44	0.10	1.55	1.28 – 1.89
Home : Wallonia		-1.48	1.16	0.23	0.01 – 1.64
LTCF : Brussels		-0.06	0.13	0.95	0.73 – 1.22
Home : Brussels		-0.06	1.52	0.94	0.03 – 27.43

Observations = 19,176 (14 observations with 'other' or unknown place of death and/or missing age removed)

Model fit

$\chi^2(14) = 1{,}075.37$; $p < .001$

Pseudo-R^2 (Cragg-Uhler) = 0.10

Pseudo-R^2 (McFadden) = 0.07

AIC = 14,108.82; BIC = 14,242.46

Hosmer and Lemeshow (GOF) test: $\chi^2(8) = 28.323$; $p < .001$

Table C-2: Estimated average probabilities that a COVID-19 death in the COVID-19 mortality epidemiological surveillance database has a corresponding record in the cause of death database indicating COVID-19 as underlying cause of death, for each place of death and region of death, multiple logistic regression model 1, Belgium, 2020.

	Estimated average probability [%]	95% CI
Place of death		
Hospital	88	87 – 88
LTCF	89	88 – 90
Home	91	77 – 97
Region of death		
Flanders	90	89 – 91
Wallonia	86	85 – 87
Brussels	85	84 – 87

Appendix D

Table D-1: Estimated regression parameters (B), standard errors (SE), and exponentiated estimates (exp(B)) with 95% confidence intervals (CI) of factors influencing the probability that a COVID-19 death in the cause of death database has a corresponding record in the COVID-19 mortality epidemiological surveillance database, multiple logistic regression model 2, Belgium 2020.

Independent variable	Reference category	Estimate (B)	SE	Exp(B)	95% CI for exp(B)
(Intercept)		2.39	0.07	10.95	9.64 – 12.44
Place of death	Hospital				
LTCF		-0.21	0.06	0.81	0.72 – 0.92
Home		-2.64	0.13	0.07	0.05 – 0.09
Region of death	Flanders				
Wallonia		-0.36	0.05	0.70	0.63 – 0.78
Brussels		-0.19	0.07	0.83	0.72 – 0.96
Epidemic wave	Wave 1				
Interwave		-0.70	0.21	0.49	0.33 – 0.75
Wave 2		0.26	0.07	1.29	1.12 – 1.49
Case classification	U07.1				
U07.2		-0.98	0.05	0.38	0.34 – 0.42
Sex	Male				
Female		-0.13	0.05	0.88	0.80 – 0.96
Number of conditions		-0.08	0.01	0.92	0.91 – 0.94
Place of death: wave					
LTCF: interwave		-1.47	0.24	0.23	0.14 – 0.37
Home: interwave		0.39	0.59	1.48	0.40 – 4.36
LTCF: wave 2		-0.55	0.08	0.58	0.49 – 0.67
Home: wave 2		-1.42	0.20	0.24	0.16 – 0.36
Region of death: wave					
Wallonia: interwave		-0.52	0.25	0.60	0.36 – 0.97
Brussels: interwave		-0.20	0.32	0.82	0.44 – 1.52
Wallonia: wave 2		-0.56	0.08	0.57	0.49 – 0.66
Brussels: wave 2		-0.50	0.11	0.61	0.49 – 0.76
Observations = 21,972 (43 with 'other' and unknown places of death removed)					
Model fit					
$\chi^2(17) = 3{,}431.69$; $p < .001$					
Pseudo-R2 (Cragg-Uhler) = 0.22					
Pseudo-R2 (McFadden) = 0.14					
AIC = 21,077.68; BIC = 21,221.63					
Hosmer and Lemeshow (GOF) test: $\chi^2(8) = 9.426$; $p = 0.31$					

Table Table D-2: Estimated average probabilities that a COVID-19 death in the cause of death database has a corresponding record in the COVID-19 mortality epidemiological surveillance, for each place of death, region of death, and epidemic wave, multiple logistic regression model 2, Belgium, 2020.

	Estimated average probability [%]	95% CI
Place of death		
Hospital	83	83 – 84
LTCF	75	74 – 76
Home	15	13 – 18
Region of death		
Flanders	82	82 – 83
Wallonia	71	70 – 72
Brussels	75	73 – 77
Epidemic wave		
Wave 1	81	80 – 82
Interwave	47	42 – 52
Wave 2	75	74 – 76