BMJ Open

Pulse oximetry and supplemental oxygen use in nationwide Veterans Health Administration hospitals, 2013–2017: a Veterans Affairs Patient Database validation study

Xiao Qing Wang,[1] Theodore Iwashyna,[1,2] Hallie Prescott,[1,2] Valeria Valbuena,[3] Sarah Seelye [2]

[1]Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA
[2]Ann Arbor VA Medical Center, Veterans Affairs Center for Clinical Management Research, Ann Arbor, Michigan, USA
[3]Department of Surgery, University of Michigan, Ann Arbor, Michigan, USA

Correspondence to
Dr Sarah Seelye;
sarah.seelye@va.gov

ABSTRACT

Objective Extraction and standardisation of pulse oximetry and supplemental oxygen data from electronic health records has the potential to improve risk-adjustment, quality assessment and prognostication. We develop an approach to standardisation and report on its use for benchmarking purposes.

Materials and methods Using electronic health record data from the nationwide Veteran's Affairs healthcare system (2013–2017), we extracted, standardised and validated pulse oximetry and supplemental oxygen data for 2 765 446 hospitalisations in the Veteran's Affairs Patient Database (VAPD) cohort study. We assessed face, concurrent and predictive validities using the following approaches, respectively: (1) evaluating the stability of patients' pulse oximetry values during a 24-hour period, (2) testing for greater amounts of supplemental oxygen use in patients likely to need oxygen therapy and (3) examining the association between supplemental oxygen and subsequent mortality.

Results We found that 2 700 922 (98%) hospitalisations had at least one pulse oximetry reading, and 864 605 (31%) hospitalisations received oxygen therapy. Patients monitored by pulse oximetry had a reading on average every 6 hours (median 4; IQR 3–7). Patients on supplemental oxygen were older, white and male compared with patients not receiving oxygen therapy (p<0.001) and were more likely to have diagnoses of heart failure and chronic pulmonary diseases (p<0.001). The amount of supplemental oxygen for patients with at least three consecutive values recorded during a 24-hour period fluctuated by median 2 L/min (IQR: 2–3), and 81% of such triplets showed the same level of oxygen receipt.

Conclusion Our approach to standardising pulse oximetry and supplemental oxygen data shows face, concurrent and predictive validities as the following: supplemental oxygen clusters in the range consistent with hospital wall-dispensed oxygen supplies (face validity); there are greater amounts of supplemental oxygen for certain clinical conditions (concurrent validity) and there is an association of supplemental oxygen with in-hospital and postdischarge mortality (predictive validity).

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study focuses on a large dataset of 2 765 446 hospitalisations at >130 different Veterans Affairs hospitals.
⇒ Detailed extraction and standardisation process for complex pulse oximetry and supplemental oxygen data in >17.5 million observations.
⇒ Standardisation process assessed for face validity, concurrent validity and predictive validity.
⇒ This standardisation was developed for use in pulse oximetry and supplemental oxygen outside the intensive care unit, and therefore did not include recordings of invasive mechanical ventilation, non-invasive positive pressure ventilation or heated high-flow nasal cannula.

INTRODUCTION

Supplemental oxygen is among the most administered medical therapies in hospitalised patients.[1-3] In many hospitals, supplemental oxygen can be administered and titrated without a physician order, in contrast to most medications.[4 5] New hypoxemia is a cardinal bedside sign of acute or impending decompensation as well as a marker of disease severity for several chronic comorbidities.[6]

Despite the importance of hypoxemia, analyses of hospital data often exclude data on the respiratory system. Respiratory rate is routinely collected as a vital sign, but may be recorded as '18' or '20' regardless of actual respiratory rate. Furthermore, respiratory rate is often influenced by fear, anxiety, pain and sleep, reducing its reliability as a measure of pulmonary function.[7] In contrast, pulse oximetry, which measures the fraction of haemoglobin in the blood-carrying oxygen, has become ubiquitous, but it can be interpreted only in the context of knowing how much supplemental oxygen a patient is on

at the same time.[8][9] Adding further complexity, there are diverse modes of oxygen delivery[10] and variable charting conventions within and across hospitals. Although bedside clinicians can routinely interpret such diversely formatted data, the absence of standardisation limits the ability of health services researchers to use large datasets to study processes of care and outcomes for conditions where clinicians routinely define severity in part in terms of degree of hypoxemia—conditions such as pneumonia, acute exacerbation of chronic obstructive pulmonary disease (COPD) and congestive heart failure.

Extraction and standardisation of pulse oximetry and supplemental oxygen data from electronic health record data has the potential to improve risk-adjustment, quality assessment and prognostication. We therefore sought to identify the ways in which pulse oximetry and supplemental oxygen use were recorded in the nationwide Veterans Health Administration (VA) system, develop an approach to standardise these data and report on its use for benchmarking purposes. This study makes possible an infrastructure for future studies to explore the advantages and potential overuse of pulse oximetry. We focus on the data systems used in VA hospitals outside of the intensive care unit (ICU) and operating room (OR) context—the so-called hospital 'floors'—because a single system is used nationwide for recording vital signs on the floors, while each VA hospital may use one of several different ICU and OR data systems, including paper charting during the period under study.

METHODS
Study setting and data source
The nationwide VA healthcare system is an integrated system that provides comprehensive medical care to Veterans and operates >130 acute care hospitals. All VA hospitals use the same electronic health record platform, the Computerized Patient Record System,[11] which archives data to a single central repository—the Corporate Data Warehouse (CDW).[12][13] In this study, we extracted vital sign data from the CDW for all 2 765 446 hospitalisations in the Veteran's Affairs Patient Database (VAPD) for Veterans followed from 1 January 2013 through 31 December 2017; construction of the VAPD has been described previously.[14][15] Comorbidities were assessed using Elixhauser comorbidities,[16] and diuretic use was identified from Bar Code Medication Administration data within the VAPD, as previously described. Mortality data were extracted from the CDW's Vital Status File, which contains dates of death from sources including the VA's Beneficiary Identification Records Locator Subsystem File, Medicare Vital Status File and the Social Security Administration Death Master File.

Extracting pulse oximetry and supplemental oxygen
All numeric pulse oximetry measurements are archived in the CDW Vital Signs Domain, where they can be readily used for analysis.[11–13] However, because supplemental oxygen values are hand-entered by healthcare staff, with raw values such as '3.0 L/min %', '2.0 L/min', or '35%', additional cleaning and standardisation by clinicians was required. (See online supplemental appendices 1,2 and tables 1,2A,B for a step-by-step description of our approach to cleaning and standardisation.)

There is not a uniformly accepted way to report the amount of supplemental oxygen administered. There are two complexities here. The first is reporting: different institutions have different conventions of reporting supplemental oxygen when the same therapy is administered. The second is technical: oxygen can be delivered at a fixed flow rate or at a fixed percentage of inspired oxygen. Standard nasal cannula delivers a fixed flow, but the percentage inspired will vary as a function of the patient's respiratory mechanics. The greater the inspiratory flow generated by the patient, the more likely the flow of oxygen delivered will be overcome, and the fraction of inspired air that is oxygen (FiO_2) received will be reduced by entrapment of room air. Fixed percentage inspired oxygen devices (eg, Venturi masks) attempt to overcome this variation by delivering high gas flow rates that are less likely to be overcome, although there is variation in effective FiO_2 delivered depending on a patient's respiratory pattern, positioning and anatomy. We therefore used a common clinical 'rule of thumb' formula to approximate the oxygen flow in litres per minute (LPM) as a function of the fixed inspired oxygen fraction (FiO_2, when that is used), as LPM = $(FiO_2 - 0.21)/0.03$, to standardise all modes of supplemental oxygen delivery as if they were fixed flow devices. We standardised to LPM rather than to FiO_2 because the imprecision of fixed flow devices for FiO_2 received is generally known clinically and may thereby prevent false precision in the interpretation of the data.

Evaluation of validity
We assessed the validity of SpO_2 and supplemental oxygen data using a three-criteria approach to validation: face validity, concurrent validity and predictive validity. Validity describes the accuracy of a given method to measure a given parameter. When a method uses measurements that closely reflect the true value of a given parameter, the method is considered valid.[17][18] There are several types of validity by which a measurement can be examined, which is particularly useful when a gold standard comparison is not available. Face validity is established when a method 'appears' to measure the parameter of interest. Concurrent validity is established when higher levels of the measurement are more common in populations where the higher levels of the true parameter are expected to be more common. Predictive validity is established when the measurement is associated with future outcomes in a manner consistent with the true parameter. In the case

Table 1 Seven general forms for recording the amount of supplemental oxygen and our approach to standardisation

Type	Form	Example	Number of values in CDW vital signs domain	Percent of total values (%)	Number of hospitals at which this format was found	Approach to standardisation
1	nn%	35%	712 568	4.1	129	Interpreted as fixed percentage inspired oxygen device; converted to LPM according to equation LPM = (FiO$_2$–0.21)/0.03
2	nnL/min	2 L/min	6 522 303	37.2	129	Interpreted as nn LPM fixed flow
3	L/min%	L/min%	9 461 987	54.0	42	Interpreted as zero LPM
4	L/minnn%	L/min95%	20 041	0.1	37	Excluded because nn were not clinically interpretable values
5	nn L/min%	2.0 L/min%	461 408	2.6	42	Interpreted as nn LPM fixed flow
6	nn L/min xx% where xx%=Pulse Ox reading at same time	2 L/min99% when pulse ox=99%	232 679	1.3	129	Interpreted as nn LPM (eg, 2 LPM)
7	nn L/min xx% where xx% not equal to pulse Ox reading at same time	2 L/min99% when pulse ox=93%	126 698	0.7	129	Excluded
Total			17 537 684			

See also online supplemental appendix.
CDW, Corporate Data Warehouse; LPM, litres per minute.

of this investigation and pulse oximetry, we consider the following.

Face validity
Clinical experience suggests that (1) pulse oximetry readings will be overwhelmingly 88% and higher because values lower than this indicate clinically significant hypoxemia that clinicians will promptly correct, (2) supplemental oxygen is most commonly administered in whole LPM and (3) supplemental oxygen is most often titrated to even numbers. We assessed these elements by looking at the distribution of pulse oximetry and supplemental oxygen values. We further assessed face validity by examining consecutive triplets of supplemental oxygen readings within 24 hours to assess the stability of values and the frequency at which the second reading had a similar flow rate as the first and third.

Concurrent validity
We assessed this by testing for greater amounts of supplemental oxygen in patients likely to need supplemental oxygen based on their clinical situation—for example, those with cardiopulmonary disease, those receiving diuretic therapy and those whose hospitalisation involves ICU stays.

Predictive validity
We examined whether greater amounts of supplemental oxygen use on the first day of hospitalisation were associated with higher rates of in-hospital mortality, and whether greater amounts of supplemental oxygen on the last day of hospitalisation were associated with higher rates of mortality in the subsequent year. To assess each of these, we used logistic regression analysis with patient-level random effects to examine the association between supplemental oxygen and mortality. Patient-level random effects were included to account for cases in which multiple admissions occurred during the study period.

Data management and analyses
Data management and analyses were conducted using SQL and SAS (SAS Institute), V.9.4. The analytic code is available online to allow others to extract and clean pulse oximetry and supplemental oxygen values as we have (https://github.com/CCMRcodes/PulseOx).

Data are presented as mean values, medians and IQRs at the hospitalisation-day and patient-day level. We assessed the proportion of a hospitalisation with pulse oximetric coverage as the time between the first and last pulse oximetry reading during a hospitalisation (in hours), divided by the time from admission to discharge (in hours).

Table 2 Pulse oximetry measurements

	Receiving supplemental oxygen and pulse oximetry	Receiving pulse oximetry but not supplemental oxygen
Hospitalisations with pulse oximetry, N (%)	864 605 (31.3)	1 836 317 (66.4)
Number of pulse oximetry measurements during hospitalisations, if at least one		
Mean	26	15
Median	17	10
IQR	10–31	6–18
Number of pulse oximetry measurements per day, if at least one		
Mean	4.0	3.5
Median	4	3
IQR	2–5	2–4
Number of pulse oximetry measurements per calendar day, if at least one N (%)	5 716 352	7 709 751
One time	513 532 (9%)	843 037 (11%)
Two times	939 268 (16%)	1 592 787 (21%)
Three times	1 363 715 (24%)	2 114 788 (27%)
>4 times	2 899 837 (51%)	3 159 139 (41%)
Hours between pulse oximetry measurements within a hospitalisation		
Mean	5.9	5.9
Median	4	5
IQR	3–7	3–8
Proportion of hospitalisation with pulse oximetry monitoring, by hour		
Mean	0.72	0.63
Median	0.78	0.66
IQR	0.63–0.88	0.50–0.81

Figure 1 Histogram of pulse oximetry values for recorded saturations, with a majority of observations between 90 and 100. Values are used to identify potential data errors and to confirm data validation.

in an ICU, and 391 330 days on which there were transitions into an ICU.

Patterns of pulse oximetry and supplemental oxygen data

The amount of supplemental oxygen that patients received were reported in one of seven forms (table 1). Of the 17 537 684 supplemental oxygen readings in the CDW, every hospital used at least four different reporting conventions; indeed, not a single hospital relied on only one reporting method. Data demonstrating the rationale for each of these interpretations are presented in the supplementary file (online supplemental appendix 1, figure 1).

Among 2 700 922 (97.7%) hospitalisations with one or more pulse oximetry values, 1 836 317 (66.4%) hospitalisations received a pulse oximetry reading without

Patient and public involvement

Research was completed without patient or public involvement; patients were not invited to contribute to the design of the study, interpretation of results, or the writing or editing of the article.

RESULTS

The VAPD consists of a total of 15 437 270 patient-days for 2 765 446 hospitalisations at 134 VA hospitals from 2013 to 2017. Among all hospitalisations, 2 700 922 (98%) hospitalisations had at least one pulse oximetry reading, and 864 605 (31%) hospitalisations received oxygen therapy. Pulse oximetry data were present for 12 369 844 days not
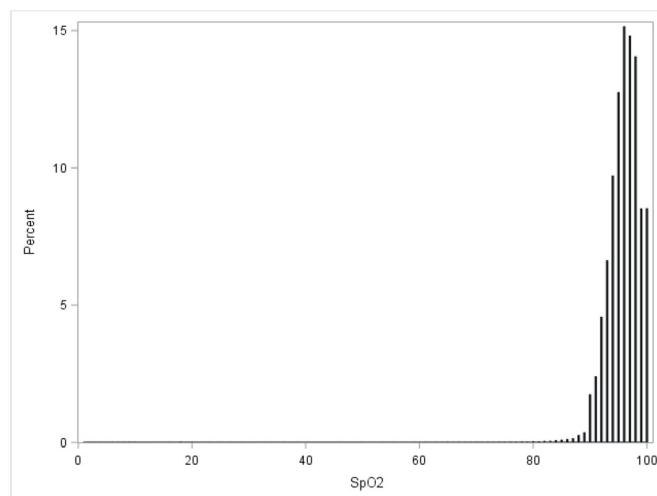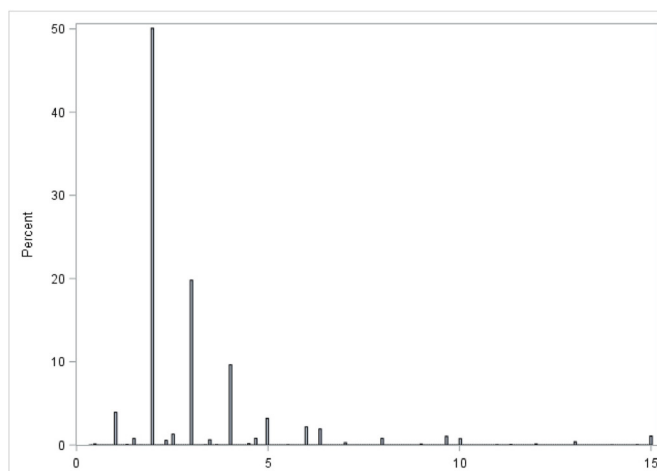

Figure 2 Histogram of supplemental oxygen rates (in LPM) among those receiving at least some oxygen, with a majority receiving between 2 LPM and 4 LPM. About 15 LPM is top coded to include a small number of patients recorded as receiving >15 LPM of oxygen. LPM, litres per minute.

Table 3 Patient and hospitalisation characteristics of hospitalisations with supplemental oxygen therapy versus hospitalisations without supplemental oxygen therapy

	Hospitalisations receiving supplemental oxygen (n=864605)	Hospitalisations not receiving supplemental oxygen (n=1900841)	P value
Unique patients, N	504523	994630	
Age, year, median (IQR)	68 (63–77)	66 (58–72)	<0.0001
Male, N (%)	825985 (95.5)	1789807 (94.2)	<0.0001
Race, N (%)			<0.0001
White	664981 (76.9)	1325297 (69.7)	
Black	135123 (15.6)	437930 (23.0)	
Other	64501 (7.5)	137614 (7.2)	
Comorbidities			
Heart failure (CHF), N (%)	253083 (29.3)	277616 (14.6)	<0.0001
Chronic pulmonary disease, N (%)	347550 (40.2)	334212 (17.6)	<0.0001
Ever treated in ICU during hospitalisation	203740 (23.6)	244346 (12.8)	<0.0001
Ever on mechanical ventilation	31971 (3.7)	15906 (0.8)	<0.0001
Ever received diuretics	414488 (47.9)	559017 (29.4)	<0.0001
Length of hospitalisation, median (IQR)	5 (3–8)	3 (2–5)	<0.0001
In-hospital mortality*	29885 (3.5)	15502 (0.8)	<0.0001
30-day mortality*	68273 (7.9)	50830 (2.7)	<0.0001
One-year postdischarge mortality*	243891 (28.2)	291012 (15.3)	<0.0001

*All-cause mortality, rate is per hospitalisation.
CHF, congestive heart failure; ICU, intensive care unit.

supplemental oxygen. For these hospitalisations, the number of pulse oximetry readings was median 10 (IQR: 6–18) (table 2). There were median 3 (IQR: 2–4) pulse oximetry readings per patient-day. On average, patients received a pulse oximetry reading every 6 hours (median 5; IQR 3–8), with 63% of the hospital stay (defined in hours) under pulse oximetry monitoring (median 66; IQR 50–81). The distribution of pulse oximetry readings is shown in figure 1, with 99% reading at ≥88%, supporting face validity of the data.

As part of our validation process for face validity, we examined patients with triplets of at least three consecutive supplemental oxygen values during a 24-hour period to ensure that supplemental oxygen values were clinically stable and that changes in LPM were reasonable. The amount of supplemental oxygen for patients with at least three consecutive values recorded during a 24-hour period fluctuated by median 2 LPM (IQR: 2–3). Furthermore, 81% of such triplets showed the same level of oxygen receipt throughout.

Patterns of supplemental oxygen use

Between 2013 and 2017, 864605 (31.3%) hospitalisations received supplemental oxygen (table 2). Patients on supplemental oxygen received more frequent pulse oximetry monitoring during the hospitalisation (mean 26; median 17; IQR 10–31 vs mean 15; median 10; IQR 6–18 for patients not on oxygen) and throughout the day (mean 4.0; median 4; IQR 2–5 vs mean 3.5; median 3; IQR 2–4 for patients not on oxygen) (table 2). Patients on supplemental oxygen had a pulse oximetry reading on average every 6 hours (median 4; IQR 3–7). As shown in figure 2, oxygen was typically given in integer—often even integer—flow rates, consistent with the face validity of the data.

Consistent with the concurrent validity of the data, patients who received oxygen therapy at VA hospitals were older compared with patients who did not receive oxygen therapy (table 3, p<0.001). They were also more likely to have diagnoses of heart failure and chronic pulmonary diseases (p<0.001). Patients with hospitalisations that included stays in the ICU as well as those on mechanical ventilation were also more likely to receive oxygen therapy recorded in the CDW Vital Signs domains (hence from the general care floors) at some point during their hospitalisation (p<0.001).

In an unadjusted model with patient-level random effects, greater amounts of supplemental oxygen use on the first day of hospitalisation were associated with higher rates of in-hospital mortality (OR 1.16 per LPM-equiv, 95% CI 1.15 to 1.16, p<0.001). Similarly, greater amounts of supplemental oxygen on the last day of the hospitalisation were associated with higher rates of 1 year

postdischarge mortality (OR 1.55 per LPM-equiv, 95% CI 1.54 to 1.56, p<0.001). These are consistent with predictive validity of the data.

DISCUSSION

In this study, we examine 2.7 million hospitalisations across >100 hospitals over 5 years. We show that pulse oximetry and supplemental oxygen data can be usefully extracted from a large, unstandardised national database of inpatient vital sign recordings. The extracted data have plausible face, content and predictive validities, suggesting their potential usefulness in analyses. We find that pulse oximetry and supplemental oxygen readings can be used in secondary data analyses to identify non-ICU patients at risk of adverse outcomes, such as in-hospital mortality and one year postdischarge mortality. We also find that pulse oximetry and supplemental oxygen readings can be used in retrospective studies to define clinically relevant subgroups of patients with disease for targeted study, including pneumonia patients with a new inpatient oxygen requirement and patients with a newly hypoxemic acute exacerbation of COPD. We believe these results demonstrate that valid pulse oximetry and supplemental oxygen data can be recovered from the VA electronic health record despite the diverse and unstandardised ways in which data are recorded by bedside clinicians. These data can be considered for use in large-scale, system-wide risk-adjustment, assessments of processes of care and in defining patient subgroups of interest.

While efforts at ensuring a standard approach to recording these data—as has been done by creating separate systolic, diastolic, and mean arterial blood pressure fields in the recording of that vital sign—may be of value in the future, science using pulse oximetry data in the VA need not necessarily wait for widescale adoption. The work necessary to recover these data for large-scale analyses may also serve as a caution to those designing electronic health records, because even for seemingly simple things, bedside clinicians often carry diverse representations of information, and those diverse representations will find their way into free-text fields unless (and perhaps even if) there is a clear and easy-to-use indication of the standardised, structured approach for recording the data.

The face validity of the data is supported by the clustering of pulse oximetry values in a plausible physiological range that is compatible with life. Face validity is also supported by intervals that align with conventional shift durations and vital sign measurement timing, with integer values of 2–15 LPM—values which most hospital wall-dispensed oxygen supplies provide. The concurrent validity of supplemental oxygen rates is supported by greater amounts of supplemental oxygen in clinical conditions for which supplemental oxygen use is more common. The predictive validity of the supplemental oxygen data is shown by its association with in-hospital and postdischarge mortality.

Our study should be interpreted in the context of some limitations. First, our analysis did not include direct validation of the data against the ground truth of what patients were receiving, as this was not feasible. Individual clinician notes typically cover shifts, days, or even longer time-periods, whereas pulse oximetry and supplemental oxygen can change second to second, and therefore cannot be confirmed via chart review. Second, our approach to data extraction and standardisation was developed for use outside of the ICU, and hence did not include recordings of invasive mechanical ventilation, non-invasive positive pressure ventilation, or heated high flow nasal cannula. Future work is needed to expand our methods to include these advanced modes of respiratory support.

In conclusion, the recording of pulse oximetry and supplemental oxygen within the VA varies across sites. However, our approach to standardising these data yielded results with face, concurrent, and predictive validity, indicating that the data are valid and can be used for secondary data analysis as well as for operational purposes within the VA.

ORCID iD
Sarah Seelye http://orcid.org/0000-0001-5098-7470

REFERENCES

1 Cousins JL, Wark PAB, McDonald VM. Acute oxygen therapy: a review of prescribing and delivery practices. *Int J Chron Obstruct Pulmon Dis* 2016;11:1067–75.

2 Blakeman TC. Evidence for oxygen use in the hospitalized patient: is more really the enemy of good? *Respir Care* 2013;58:1679–93.

3 Grainge C. Breath of life: the evolution of oxygen therapy. *J R Soc Med* 2004;97:489–93.

4 Devoe NC, Kyriazis P, Eltanbedawi A. An audit of oxygen supplementation in a large tertiary hospital - we should treat oxygen as any other drug. *Hosp Pract* 2020:1–4.

5 Small D, Duha A, Wieskopf B, *et al*. Uses and misuses of oxygen in hospitalized patients. *Am J Med* 1992;92:591–5.

6 Bedoya AD, Bhavsar NA, Adagarla B, *et al*. Unanticipated respiratory compromise and unplanned Intubations on general medical and surgical floors. *Respir Care* 2020;65:1233–40.

7 Lovett PB, Buchwald JM, Stürmann K, *et al*. The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage. *Ann Emerg Med* 2005;45:68–76.

8 DeMeulenaere S. Pulse oximetry: uses and limitations. *J Nurse Pract* 2007;3:312–7.

9 Gruber P, Kwiatkowski T, Silverman R, *et al*. Time to equilibration of oxygen saturation using pulse oximetry. *Acad Emerg Med* 1995;2:810–5.

10 Pruitt WC, Jacobs M. Breathing lessons: basics of oxygen therapy. *Nursing* 2003;33:43–5.

11 Fihn SD, Francis J, Clancy C, *et al*. Insights from advanced analytics at the Veterans health administration. *Health Aff* 2014;33:1203–11.

12 US Department of Veterans Affairs. 172VA10P2: VHA corporate data Warehouse, VA. 79 FR 4377.

13 Wiitala WL, Vincent BM, Burns JA, *et al*. Variation in laboratory test naming conventions in EHRs within and between hospitals: a nationwide longitudinal study. *Med Care* 2019;57:e22–7.

14 Wang XQ, Vincent BM, Wiitala WL, *et al*. Veterans Affairs patient database (VAPD 2014-2017): building nationwide granular data for clinical discovery. *BMC Med Res Methodol* 2019;19:94.

15 Vincent BM, Wiitala WL, Burns JA, *et al*. Using Veterans Affairs corporate data Warehouse to identify 30-day hospital readmissions. *Health Serv Outcomes Res Method* 2018;18:143–54.

16 Elixhauser A, Steiner C, Harris DR, *et al*. Comorbidity measures for use with administrative data. *Med Care* 1998;36:8–27.

17 Thanasegaran G. Reliability and validity issues in research. *Integ Dissemin* 2009;4:35–40.

18 Price P, Jhangiani R, Chiang I. *Research methods of psychology*. 2nd Canadian Edn, 2015.

Pulse oximetry and supplemental oxygen supplement

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

BMJ Open

Appendix 1. Extracting and cleaning supplemental oxygen

Our goal was to extract and standardize supplemental oxygen data from the VA's Corporate Data Warehouse (CDW). We used data from two fields in the CDW's Vital Signs Domain: (1) the supplemental oxygen text field ('SupplementalO2') and (2) the vital results numeric field ('VitalResultNumeric') (see Table 1).

As illustrated in Table 1, the raw data for supplemental oxygen were recorded using a variety of notations (e.g. 3.0 l/min %, l/min %, 2 l/min, 35%). We binned these notations into 7 separate forms following a standardization process shown in Figure 1. In total, we extracted 17,537,727 unique supplemental oxygen values from the VA's CDW for the 2013-2017 Veteran's Affairs Patient Database (VAPD) and placed each value into one of 7 smaller datasets according to the value of the supplemental oxygen recording and the vital results numeric field.

In the remaining sections of the appendix, we provide our standardization strategy for each of these 7 forms. Standardization of the supplemental oxygen text field resulted in a usable variable for clinical research: O2_LPM (supplemental oxygen measured in liters per minute).

2

Table 1. CDW vital signs domain

VitalResult	VitalResultNumeric	Systolic	Diastolic	Supplementa lO2
95	95.0000	.	.	3.0 l/min %
94	94.0000	.	.	l/min %
94	94.0000	.	.	l/min %
93	93.0000	.	.	l/min %
98	98.0000	.	.	l/min %
93	93.0000	.	.	l/min %
92	92.0000	.	.	l/min %
94	94.0000	.	.	l/min %
100	100.0000	.	.	l/min %
93	93.0000	.	.	3.0 l/min
92	92.0000	.	.	2 l/min
100	100.0000	.	.	5 l/min
100	100.0000	.	.	l/min %
95	95.0000	.	.	l/min %
95	95.0000	.	.	2.0 l/min
97	97.0000	.	.	3 l/min
91	91.0000	.	.	3 l/min
91	91.0000	.	.	l/min %
100	100.0000	.	.	35%
94	94.0000	.	.	3 l/min
99	99.0000	.	.	l/min %
98	98.0000	.	.	l/min %
96	96.0000	.	.	l/min %
96	96.0000	.	.	l/min %
94	94.0000	.	.	l/min %
65	65.0000	.	.	l/min %
97	97.0000	.	.	2 l/min

3

Figure 1. Consort flow diagram of supplemental oxygen text fields, 2013-2017
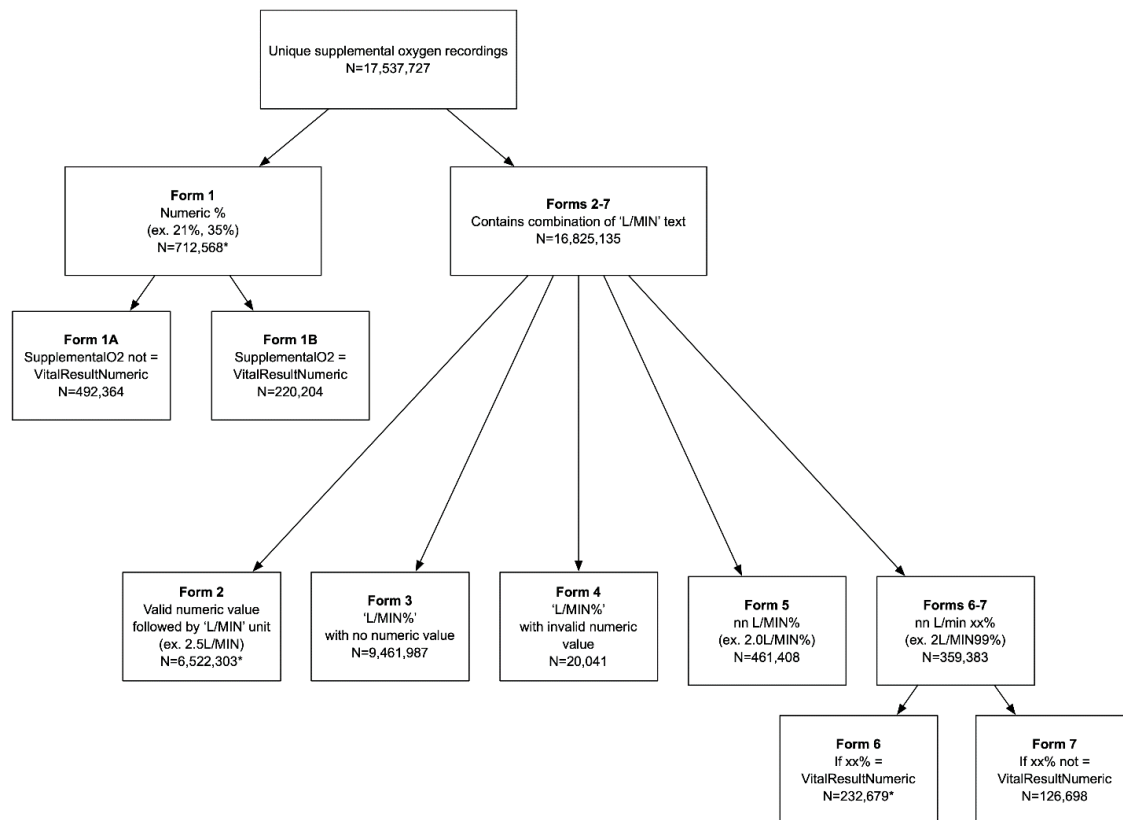
Figure 1 Legend

Form 1 consisted of 712,568 supplemental oxygen recordings that contained a numeric value followed by the '%' character (e.g., 21%, 35%, 55%, 95%). For these observations, we used the formula LPM = (FiO2 - 0.21) / 0.03 to obtain liters per minute of oxygen (O2_LPM). Because Form 1 contained values in the supplemental oxygen field that were not always consistent with values in the VitalResultNumeric field (Tables 2a and 2b), we used two approaches for standardization. When supplemental oxygen values differed from VitalResultNumeric values (N=492,364), we considered the supplemental oxygen field to be the valid measure of FiO2, (Table 2a). However, when supplemental oxygen values were equal to the VitalResultNumeric values (N=220,204), we considered these to be SpO2 values that were miscoded in the SupplementalO2 field and excluded them (Table 2b).

Form 2 consisted of 6,522,303 supplemental oxygen recordings with a numeric value followed by the 'L/MIN' unit (e.g. 2.5L/MIN). These recordings reflected the number of liters per minute of oxygen that patients were receiving at the time of the recording and were considered valid numeric supplemental oxygen values. Values in Form 2 required the numeric digits to be extracted from the string of text characters (e.g. separating '2.5' from 'L/MIN').

Form 3 consisted of 9,461,987 supplemental oxygen recordings that contained the text characters 'L/MIN%' with no numeric value. This form was coded as '0' to indicate that patients were not receiving supplemental oxygen.

Form 4 contained 20,041 observations with the text character ''L/MIN%' and a numeric value. However, the numeric value for this form represented an invalid supplemental oxygen recording (e.g. L/MIN95%) and was dropped.

Form 5 included 461,408 valid numeric supplemental oxygen recordings followed by 'L/MIN%' (e.g. 2.5L/MIN%). For these observations, we extracted the numeric values from the string characters (e.g. separating '2.5' from 'L/MIN%').

Form 6 consisted of 232,679 observations in which the supplemental oxygen value was recorded in the form 'xx%' and was consistent with the numeric value recorded in the 'VitalResultNumeric' field. Supplemental oxygen values in this form appear to contain values for both pulse oximetry and Supplemental oxygen in the text string. When the number before a '%' unit was equal to the value in the VitalResultNumeric field, then we extracted the numeric digits before the '%' unit and coded them as Supplemental oxygen values.

Form 7 consisted of 126,698 observations in which 'xx%' was linked to a numeric value that was not consistent with the numeric value in the 'VitalResultNumeric' field. Supplemental oxygen recordings in this form appear to contain values for both pulse oximetry and supplemental oxygen in the text string. Unlike form 5, when the number before a '%' unit was NOT equal to the VitalResultNumeric value, these were considered incoherent data and were dropped.

* A total of 24 observations were dropped from Form 1 due to invalid liters per minute of oxygen of less than 0 (O2_LPM<0). 13 observations were dropped from Form 2 due to missing values for liters per minute of oxygen. 6 observations were dropped from Form 6 due to data entry errors in the text character field for SupplementalO2. Exclusions are noted in the data extraction and standardization code of Appendix 2.

5

Table 2a. Different values for VitalResultNumeric and Supplemental oxygen

vital_date	year	VitalResultNumeric	SupplementaIO2	SpO2	O2_LPM	incoherent
06/10/2016	2016	88.0000	50%	88	9.6666666667	0
12/20/2014	2014	97.0000	23%	97	0.6666666667	0
01/24/2015	2015	97.0000	35%	97	4.6666666667	0
02/14/2015	2015	93.0000	50%	93	9.6666666667	0
06/06/2013	2013	94.0000	50%	94	9.6666666667	0
09/19/2015	2015	94.0000	35%	94	4.6666666667	0
05/30/2014	2014	95.0000	50%	95	9.6666666667	0
02/28/2017	2017	97.0000	21%	97	0	0

Table 2b. Same values for VitalResultNumeric and Supplemental oxygen

vital_date	year	VitalResultNumeric	SupplementaIO2	SpO2	O2_LPM	incoherent
10/20/2015	2015	97.0000	97%	97	25.333333333	1
02/21/2013	2013	96.0000	96%	96	25	1
10/01/2013	2013	97.0000	97%	97	25.333333333	1
09/02/2014	2014	96.0000	96%	96	25	1
01/14/2016	2016	95.0000	95%	95	24.666666667	1
02/25/2016	2016	92.0000	92%	92	23.666666667	1
03/16/2016	2016	97.0000	97%	97	25.333333333	1
06/01/2015	2015	97.0000	97%	97	25.333333333	1
03/13/2013	2013	98.0000	98%	98	25.666666667	1
06/03/2013	2013	97.0000	97%	97	25.333333333	1
06/18/2013	2013	96.0000	96%	96	25	1
05/12/2014	2014	97.0000	97%	97	25.333333333	1
08/22/2013	2013	99.0000	99%	99	26	1
05/07/2014	2014	100.0000	100%	100	26.333333333	1

6

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

BMJ Open

Appendix 2. Data extraction and standardization code of pulse oximetry and supplemental oxygen

SQL code for data extraction from CDW

```
SELECT a.Sta3n, a.PatientSID, a.vitalSignTakenDateTime, a.VitalResultNumeric , a.SupplementalO2, a.VitalTypeSID,
       B.VitalType, c.patienticn
into #NewPulseOx20132014_SW20200609
FROM [Src].[Vital_VitalSign] as   A
left JOIN [CDWWORK].[Dim].[VitalType] as  B ON A.VitalTypeSID =B.VitalTypeSID
left join  Src.SPatient_SPatient c on a.patientsid=c.patientsid
WHERE (a.VitalSignTakenDateTime >= '20170101' and a.VitalSignTakenDateTime < '20180101') /*times pulled*/
 AND  (a.CohortName='Cohort20200427') /*new cohort updated*/
AND (
/*1) pull all where SupplementalO2 not null, type=PulseOX and VitalResultNumeric > 0 and VitalResultNumeric <= 100 */
(a.SupplementalO2 <> 'NULL' and  b.VitalType='PULSE OXIMETRY' and a.VitalResultNumeric > 0 and a.VitalResultNumeric <= 100)

 /*OR 2) also pull pulseox, and SupplementalO2 is NULL and VitalResultNumeric > 0 and VitalResultNumeric <= 100*/
OR (a.SupplementalO2 is NULL and b.VitalType='PULSE OXIMETRY' and a.VitalResultNumeric > 0 and a.VitalResultNumeric <= 100)
 )
```

SAS code for data cleaning and standardization

Step 1. Before data cleaning, standardize the supplemental oxygen field by changing all its values into capital letters and remove the extra spaces.

```
/*clean SupplementalO2 to take out extra space and turn everything into CAPS*/
DATA PulseOx20132017_v2 (compress=yes rename=SupplementalO2_v3=SupplementalO2); /*n=17469574*/
SET PulseOx20132017_v1;
SupplementalO2_v2=upcase(SupplementalO2); /*turn all units into uppercase*/
SupplementalO2_v3=compress(SupplementalO2_v2);  /*removes all blanks*/
drop SupplementalO2_v2 SupplementalO2;
/*drop the original SupplementalO2 field and rename SupplementalO2_v3 as SupplementalO2*/
RUN;
```

Step 2. Clean Supplemental oxygen field by separating them out into different forms.

Form 1 Code

```
PROC SQL;
CREATE TABLE pulse_noLmin (compress=yes) AS
SELECT *
FROM PulseOx20132017_v2
WHERE SupplementalO2 not like '%L/MIN%';
QUIT;

PROC MEANS DATA=pulse_noLmin MIN MAX MEAN MEDIAN Q1 Q3;
VAR VitalResultNumeric;
RUN;

PROC FREQ DATA=pulse_noLmin order=freq;
TABLE SupplementalO2;
RUN;

/*remove the % sign and change to numeric value and look at descriptive*/
DATA pulse_noLmin2 (compress=yes);
SET pulse_noLmin;
supple_char=compress(SupplementalO2,'%');
supple_num=input(supple_char, 3.);
RUN;
```

7

```
/*check if supple_num NE (not equal) VitalResultNumeric*/
data checking;
set pulse_noLmin2;
if supple_num NE VitalResultNumeric;
run;

PROC FREQ DATA=pulse_noLmin2;
TABLE supple_num;
RUN;

PROC MEANS DATA=pulse_noLmin2 MIN /*0*/ MAX/*100*/ MEAN /*58.5*/ MEDIAN /*50*/ Q1 /*35*/ Q3 /*95*/;
VAR supple_num;
RUN;

/*use Jack's conversions to get SpO2 and O2 (L/MIN), call this form1_20132017.
create new fields: SpO2=VitalResultNumeric, O2_LPM, and incoherent*/
DATA form1_20132017_test (compress=yes);
SET pulse_noLmin2;
SpO2=VitalResultNumeric;
O2_LPM=((supple_num/100)-0.21)/0.03; /*turn into % first*/
if O2_LPM < 0 then delete; /*on 3/2/20 Jack said it's okay to delete*/
RUN;

PROC MEANS DATA=form1_20132017_test MIN MAX MEAN MEDIAN Q1 Q3;
VAR VitalResultNumeric SpO2 O2_LPM;
RUN;

PROC FREQ DATA=form1_20132017_test order=freq;
TABLE O2_LPM;
RUN;

/*there's form 1A and 1B: in general, SupplementalO2 should not equal VitalResultNumeric, Jack said this is
incoherent on 6/17/20*/
DATA form1A_20132017 (compress=yes)
form1B_20132017 (compress=yes);
SET form1_20132017_test;
if supple_num=VitalResultNumeric then incoherent=1;
 else incoherent=0;
if incoherent=0 then output form1A_20132017;
if incoherent=1 then output form1B_20132017;
keep patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM incoherent;
RUN;


DATA form1A_20132017 (compress=yes);
retain patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM  incoherent;
set form1A_20132017;
RUN;

DATA form1B_20132017 (compress=yes);
retain patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM  incoherent;
set form1B_20132017;
RUN;
/********************************** END OF FORM 1 CODES **********************************/
```

8

Form 2 Code

```sas
/*select everything else (non % only) not in pulse_noLmin dataset, and do a frequency to see what units it entail*/
PROC SQL;
CREATE  TABLE pulse_lmin (COMPRESS=YES)  AS
SELECT  A.* FROM PulseOx20132017_v2  AS A
WHERE A.obs not IN (SELECT  obs  FROM pulse_noLmin);
QUIT;

PROC FREQ DATA=pulse_lmin order=freq;
TABLE  SupplementalO2;
RUN;

/*look at those only with L/MIN units*/
PROC SQL;
CREATE  TABLE pulse_lmin_only (compress=yes) AS
SELECT  *
FROM pulse_lmin
WHERE  SupplementalO2 like '%L/MIN' or SupplementalO2 like 'L/MIN';
QUIT;

/*check*/
PROC FREQ DATA=pulse_lmin_only order=freq;
TABLE  SupplementalO2;
RUN;

/*there were some types for LPM, remove 'L/MIN', turn into numeric to look at descriptive*/
DATA pulse_lmin_only2 (compress=yes);
SET pulse_lmin_only;
if SupplementalO2='3EL/MIN' then SupplementalO2='3L/MIN';
if SupplementalO2='5LTL/MIN' then SupplementalO2='5L/MIN';
if SupplementalO2='4LITERSL/MIN' then SupplementalO2='4L/MIN';
if SupplementalO2='3LPML/MIN' then SupplementalO2='3L/MIN';
if SupplementalO2='4LPML/MIN' then SupplementalO2='4L/MIN';
if SupplementalO2='2L.NCL/MIN' then SupplementalO2='2L/MIN';
if SupplementalO2='2LPML/MIN' then SupplementalO2='2L/MIN';
if SupplementalO2='3LTL/MIN' then SupplementalO2='3L/MIN';
if SupplementalO2='RAL/MIN' then SupplementalO2='0L/MIN';
SupplementalO2_char=compress(SupplementalO2,'L/MIN'); /*first drop the characters, removes 'L/MIN' in units*/
SupplementalO2_num=input(SupplementalO2_char, 3.); /*turn the numeric digits into numbers*/
RUN;

PROC FREQ DATA=pulse_lmin_only2 order=freq;
TABLE  SupplementalO2;
RUN;

PROC MEANS DATA=pulse_lmin_only2 MIN  MAX MEAN  MEDIAN  Q1 Q3;
VAR SupplementalO2_num VitalResultNumeric;
RUN;

/*call this form2*/
DATA form2_20132017_test  (compress=yes);
SET pulse_lmin_only2;
SpO2=VitalResultNumeric;
incoherent=0;
O2_LPM=SupplementalO2_num;
if SupplementalO2='RAL/MIN' then O2_LPM=0;  /*recode RAL/MIN =0, don't exclude*/
if O2_LPM =. then delete; /*3/2/20: Jack said it's ok to delete*/
RUN;
```

9

```
/*check*/
PROC MEANS DATA=form2_20132017_test  MIN MAX  MEAN  MEDIAN  Q1 Q3;
VAR  VitalResultNumeric O2_LPM SpO2;
RUN;
PROC FREQ DATA=form2_20132017_test  order=freq;
TABLE  incoherent O2_LPM;
RUN;


DATA form2_20132017 (compress=yes);
retain patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM incoherent;
SET form2_20132017_test;
drop SupplementalO2_char SupplementalO2_num;
RUN;


/*******************************  END OF FORM 2 CODES  *******************************/


Form 3 & 4 code

/*look at those only with 'L/MIN%' unit*/
PROC SQL;
CREATE  TABLE  pulse_lminpercent_only (compress=yes) AS
SELECT  *
FROM pulse_lmin
WHERE SupplementalO2 like 'L/MIN%';
QUIT;

PROC FREQ DATA= pulse_lminpercent_only order=freq;
TABLE  SupplementalO2;
RUN;

/*2/3/20: get the descriptive on X%*/
/*first, get the last 4 digits*/
DATA last_4char_pulse_lminpercent (compress=yes);
SET pulse_lminpercent_only;
last_4=substr(SupplementalO2,length(SupplementalO2)-3,4);
RUN;


PROC FREQ DATA=last_4char_pulse_lminpercent order=freq;
TABLE  last_4;
RUN;


/*from last character values, compress "MIN%" characters*/
DATA last_4char_pulse_lminpercent2 (compress=yes);
SET last_4char_pulse_lminpercent;
last_4char_v2=compress(last_4,'M');
last_4char_v3=compress(last_4char_v2,'I');
last_4char_v4=compress(last_4char_v3,'N');
last_4char_v5=compress(last_4char_v4,'%');
last_4num=input(last_4char_v5,3.);
RUN;


PROC FREQ DATA= last_4char_pulse_lminpercent2 order=freq;
TABLE  last_4num last_4char_v5;
RUN;


DATA form3232 (compress=yes);
SET last_4char_pulse_lminpercent2;
if last_4num NE .;
if last_4num = VitalResultNumeric then equal=1; else equal=0;
RUN;
```

10

```
PROC FREQ DATA=form3232 order=freq;
TABLE equal;
RUN;

PROC MEANS DATA=form3232 MIN MAX MEAN MEDIAN Q1 Q3;
VAR last_4num VitalResultNumeric;
RUN;

/*call this form3*/
DATA form3_20132017_test (compress=yes);
SET last_4char_pulse_lminpercent2;
if SupplementalO2='L/MIN%';
SpO2=VitalResultNumeric;
O2_LPM=0;
incoherent=0;
RUN;

/*check Saturation_LPM forms*/
PROC MEANS DATA=form3_20132017_test MIN MAX MEAN MEDIAN Q1 Q3;
VAR SpO2 O2_LPM;
RUN;
PROC FREQ DATA=form3_20132017_test order=freq;
TABLE O2_LPM SupplementalO2;
RUN;


DATA form3_20132017 (compress=yes);
retain patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM incoherent;
SET form3_20132017_test;
drop last_4 last_4char_v2-last_4char_v5 last_4num;
RUN;

/*2/24/20: Jack decided to label these as incoherent*/
DATA form4_20132017_test (compress=yes);
SET last_4char_pulse_lminpercent2;
if SupplementalO2 NE 'L/MIN%';
SpO2=VitalResultNumeric;
O2_LPM=.;
incoherent=1;
run;

/*Jack said he wants to look at distribution of the nn%*/
DATA form4_20132017 (compress=yes);
retain patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM incoherent;
SET form4_20132017_test;
RUN;

PROC MEANS DATA=form4_20132017 MIN MAX MEAN MEDIAN Q1 Q3;
VAR last_4num;
RUN;

proc sgplot data=form4_20132017 noautolegend;
histogram last_4num;
density last_4num;
run;
/********************************* END OF FORM 3 & 4 CODES *********************************/
```

11

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

BMJ Open

Form 5 Code

Select what's left:

```
/*combine pulse_lminpercent_only+pulse_lmin_only+pulse_noLmin2 and see what is not in those datasets, do
frequency check of those units*/
DATA all (compress=yes);
SET pulse_lminpercent_only pulse_lmin_only pulse_noLmin2;
RUN;

PROC SQL;
CREATE TABLE whatisleft (COMPRESS=YES)  AS
SELECT A.* FROM PulseOx20132017_v2 AS A
WHERE A.obs not IN (SELECT obs FROM work.all);
QUIT;

PROC FREQ DATA=whatisleft order=freq;
TABLE SupplementalO2;
RUN;

/*1) separate out the ##%*/ /*first, get the last 4 digits*/
DATA last_4charV1;
SET whatisleft;
last_4=substr(SupplementalO2,length(SupplementalO2)-3,4);
RUN;

/*look at list*/
PROC FREQ DATA=last_4charV1 order=freq;
TABLE last_4;
RUN;

/*2/3/20: look at descriptive for this 126,091 form with only nnL/MIN%*/
DATA LMINpercentonly_v1 (compress=yes);
SET last_4charV1;
if last_4 = 'MIN%';
RUN;

PROC MEANS DATA=LMINpercentonly_v1 MIN MAX MEAN MEDIAN Q1 Q3;
VAR VitalResultNumeric;
RUN;

/*get the first 4 characters*/
DATA LMINpercentonly_v2 (compress=yes);
SET LMINpercentonly_v1;
first_4=substr(SupplementalO2,1,4);
RUN;

PROC FREQ DATA=LMINpercentonly_v2 order=freq;
TABLE first_4;
RUN;
/*compress L,/,M*/
DATA LMINpercentonly_v3;
SET LMINpercentonly_v2;
first_4char_v2=compress(first_4,'M');
first_4char_v3=compress(first_4char_v2,'/');
first_4char_v4=compress(first_4char_v3,'L');
first_4num=input(first_4char_v4,3.);
RUN;

PROC MEANS DATA=LMINpercentonly_v3 MIN MAX MEAN MEDIAN Q1 Q3;
VAR first_4num VitalResultNumeric;
RUN;
```

12

```sas
/*call this form5*/
DATA form5_20132017_test (compress=yes);
SET LMINpercentonly_v3;
SpO2=VitalResultNumeric;
incoherent=0;
O2_LPM=first_4num;
RUN;

/*check*/
PROC MEANS DATA=form4_20132017_test  MIN MAX MEAN MEDIAN Q1 Q3;
VAR SpO2 O2_LPM;
RUN;

PROC FREQ DATA=form5_20132017_test  order=freq;
TABLE incoherent SpO2;
RUN;

DATA form5_20132017 (compress=yes);
retain patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM incoherent;
SET form4_20132017_test;
drop last_4 first_4 first_4num first_4char_v2-first_4char_v4;
RUN;
/*********************************** END OF FORM 5 CODES ***********************************/
```

Form 6 Code

```sas
/*Question: some last 4 characters have PRN%, NLT%, TO2%, exclude them?*/
PROC FREQ DATA=last_4charV1 order=freq;
TABLE last_4;
RUN;

/*compress 1 step at a time*/
DATA last_4charV2; /*n=6 were deleted*/
SET last_4charV1;
/*Jack ok with deleting these on 3/2/20*/
if last_4 in ('NNC%','INO%','NLT%','PRN%','TO2%') then delete;
last_4char_v2=compress(last_4,'M');
last_4char_v3=compress(last_4char_v2,'I');
last_4char_v4=compress(last_4char_v3,'N');
last_4char_v5=compress(last_4char_v4,'%');
last_4num=input(last_4char_v5,3.);
RUN;

PROC FREQ DATA=last_4charV2 order=freq;
TABLE last_4char_v5;
RUN;

PROC MEANS DATA=last_4charV2 MIN MAX MEAN MEDIAN Q1 Q3;
VAR last_4num;
RUN;

PROC FREQ DATA=last_4charV2;
TABLE last_4num;
RUN;

/*to see if there are any last_4num = VitalResultNumeric*/
DATA form_20132017_test;
SET last_4charV2;
```

```
If (last_4num NE .) and (last_4num = VitalResultNumeric);
RUN;

/*2/3/20: look at nn L/Min descriptive*/
DATA form_20132017_test2 (compress=yes);
SET form_20132017_test;
first_4=substr(SupplementalO2,1,4);
RUN;

PROC FREQ DATA=form_20132017_test2 order=freq;
TABLE first_4;
RUN;

DATA form_20132017_test3 (compress=yes);
SET form_20132017_test2;
first_4char_v2=compress(first_4,'M');
first_4char_v3=compress(first_4char_v2,'/');
first_4char_v4=compress(first_4char_v3,'L');
first_4num=input(first_4char_v4,3.);
RUN;

PROC MEANS DATA=form_20132017_test3 MIN MAX MEAN MEDIAN Q1 Q3;
VAR first_4num VitalResultNumeric;
RUN;

/*call this form6*/
DATA form6_20132017_test (compress=yes);
SET form_20132017_test3;
SpO2=VitalResultNumeric;
incoherent=0;
O2_LPM=first_4num;
RUN;

/*check*/
PROC MEANS DATA=form6_20132017_test MIN MAX MEAN MEDIAN Q1 Q3;
VAR O2_LPM SpO2;
RUN;

DATA form6_20132017 (compress=yes);
retain patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM incoherent;
SET form5_20132017_test;
drop last_4 last_4char_v2-last_4char_v5 last_4num first_4 first_4char_v2-first_4char_v4 first_4num;
RUN;

/********************************* END OF FORM 6 CODES **********************************/
```

Form 7 Code

```
/*look at those that last_4num not equal to vitalresultnumeric*/
DATA val_form;
SET last_4charV2;
if (last_4num NE .) and (last_4num NE VitalResultNumeric);
RUN;

/*call this form6*/
DATA form7_20132017_test (compress=yes);
SET val_form;
SpO2=VitalResultNumeric;
incoherent=1;
O2_LPM=.;
```

14

```
RUN;

/*check*/
PROC MEANS DATA=form7_20132017_test  MIN MAX MEAN MEDIAN Q1 Q3;
VAR O2_LPM SpO2;
RUN;
PROC FREQ DATA=form7_20132017_test  order=freq;
TABLE incoherent;
RUN;

DATA form7_20132017 (compress=yes);
retain patienticn PatientSID Sta3n VitalTypeSID VitalType obs vitalSignTakenDateTime vital_date year
VitalResultNumeric SupplementalO2 SpO2 O2_LPM incoherent;
SET form7_20132017_test
drop last_4 last_4char_v2-last_4char_v5 last_4num;
RUN;

/*********************************** END OF FORM 7 CODES ***********************************/
```

Step 3. Combine all of the valid forms to make up a complete dataset

```
/*combine all forms (cleaned and don't clean) and make sure the totals add up, dropped the incoherent forms*/
DATA PulseOx_20132017_07022020  (compress=yes);
SET DONT_CLEAN_FORM FORM1A_20132017 FORM2_20132017 FORM3_20132017 FORM5_20132017
FORM6_20132017;
RUN;
```

15