DEPARTMENT OF ECONOMICS

WORKING PAPER SERIES

2016-14

McMASTER UNIVERSITY

Department of Economics
Kenneth Taylor Hall 426
1280 Main Street West
Hamilton, Ontario, Canada
L8S 4M4

http://www.mcmaster.ca/economics/

A Standardized Method for the Evaluation of Adherence to Practice Guidelines

Stephanie Thomas, Ph.D.

McMaster University

Department of Economics

Abstract

Practice guidelines are widely used in medical settings as a means of improving efficiency and quality of care by aligning service provision with evidence of what is effective. The objective of this work is to propose a methodology for the effective evaluation of the match of clinical practice data with a practice guideline. The proposed methodology uses a combination of existing analytical techniques which minimize the need for the analyst to specify a functional form for the process generating the clinical data. The methodology is illustrated in an application to a set of field data on the supplemental oxygen administration decisions of volunteer medical first responders. The result is a methodology for evaluation of guideline adherence which leverages existing patient care records and is generalizable across clinical contexts. In addition, the results are visually intuitive, supporting communication across diverse audiences.

Keywords: adherence to guidelines, health care utilization, performance evaluation, medical first responders
JEL: C12, C18,C52, I1, I18
Special thank you to Professor David Feeny for insightful comments.

1 Introduction

Unnecessary use of medical care is a major concern for health care system administrators and patients. Not only can such care be costly, but it can also have negative health consequences. A recent article by Nelson (2015) highlights the cost and health consequences of unnecessary care in the setting of American hospitals, yet the issue extends to public health care systems and private practices alike. The Health Council of Canada (2009) cites that health care spending in Canada doubled between 1997 and 2007, with 48% of the increase attributable to increased use of health care services, but that the value of this greater service use is not fully understood. Practice guidelines are often advocated as a means of guiding practitioners in making appropriate care decisions to improve the quality of care, improve outcomes, and avoid unnecessary usage. Evaluation of adherence to such guidelines has, however, often been done in limited study-specific settings involving large randomized clinical trials. This paper proposes a standardized methodology for evaluation of adherence to guidelines using existing administrative data records. The aim of this work is to facilitate greater undertaking of such evaluations, and therefore improve the information available upon which to base efficient care decisions. The method relies on a combination of basic analytical techniques and state-of-the-art data-driven analysis which leverages existing patient care data and has several attractive features. The techniques used require a minimum of subjective decisions on the part of the analyst and are widely applicable to contexts in which assessment of practice against a standard is required. The results are visually intuitive and can be easily communicated across diverse audiences. These features are particularly important for a standardized framework of guideline evaluation and should foster rapid adoption across health care sectors and hence improve the effectiveness of of clinical practice

guidelines in improving the efficiency and quality of care in health care systems.

Total expenditure on health care represents a substantial proportion of GDP among OECD member nations. Yet, in the US it is estimated that approximately 30% of care given to patients can be classed as 'unnecessary', meaning that it did not serve to improve patient health outcomes Nelson (2015). Understandably, reducing this share of unnecessary care and improving health system efficiency has become a priority concern for established research groups such as the Institute of Medicine (2006) and Health Council of Canada (2009), as well as smaller non-governmental groups such as the Lown Institute, exclusively concerned with 'Right Care' (The Lown Institute, 2016).

Medical practice guidelines are widely used to improve practitioner performance with the aim of improving patient care and promoting cost-effectiveness. Yet McGlynn et al. (2003) find that patients receive only about 50% of the recommended care according to existing practice guidelines in the US. In Canada it has been found that nearly one quarter of seniors on public drug programs use a drug which has been identified as inappropriate (Canadian Institute for Health Information, 2014). The evaluation of practitioner adherence to evidence-based practice guidelines is particularly important in efforts to improve quality of care, and has been carried out in a variety of ways. While many studies rely on randomized controlled trials, these can be costly and time consuming to administer, as well as being highly specific to the particular clinical setting or health care guideline studied. Methods which make use of existing administrative patient care data have the potential to offer substantial insights about the performance of current health care systems and areas for improvement. Existing data can potentially be calibrated to offer insights on adherence in a much more readily accessible and constantly updated format, however. The standardized framework proposed in this paper

provides a starting point for a richer analysis of adherence to clinical practice guidelines using existing data sources which can easily be adapted to a broad range of health system settings and guidelines.

Adherence to guidelines is addressed in various ways in the literature. Systematic reviews of the effectiveness of practice standards or factors influencing effectiveness have been conducted. Barbui et al. (2014) reviews practice standards in the setting of mental health care, Thomas et al. (1999) in the setting of professions allied to medicine. Flodgren et al. (2013) review standards to prevent device related infections, while Fiander et al. (2015) investigate standards to improve the use of electronic health records. Flodgren, Pomey, Taber, and Eccles (2011) collect studies about the effect of printed computerized reminders upon compliance with practice standards and Arditi et al. (2012) study the effect of inspections. O'Brien et al. (2007) survey the effect of educational outreach visits on compliance. Nearly all authors cite low quality of evidence as key obstacles in drawing firm conclusions about the impact of interventions to improve practitioner compliance or health impacts of such interventions. An ongoing adherence monitoring system using the framework proposed in this paper would serve to harness existing data and overcome many issues related to lack of data encountered in previous studies, while offering a clear gauge of the level of adherence to a practice guideline. Outside of an RCT framework, adherence is also determined by modeling an outcome and comparing the result to the guideline. An example of this is presented by Askildsen, Holmås, and Kaarboe (2011), who model patient waiting times for surgery in Norway using a regression framework, comparing estimated wait times to ranges suggested by practice guidelines.[1] Dimakou, Parkin, Devlin,

[1] Adherence also enters analysis as an explanatory variable in a regression framework, for example by Andritsos and Tang (2014) who use a composite index of the fraction of agreement with clinical guidelines for a condition as a factor explaining the geometric mean of total in-patient stay for cardiac diagnoses. These authors argue that increased adherence leads to reduced resource use, but the effects are small.

and Appleby (2009) examine the impact of government waiting time targets upon patient wait-times for surgery in the NHS. These authors take a hazard function approach to analysis and find that peaks in the probability of admission coincide with government mandated targets. The case by case variety in evaluation strategies encountered makes it difficult to compare and implement similar studies of adherence across jurisdictions, even for similar clinical processes, limiting the usefulness of such information in improving health system performance. The analytical strategy used in the framework proposed in this paper aims to provide a standardized, but flexible, approach to guideline adherence evaluation, therefore improving the ability of health system administrators to accurately assess and compare performance against established practice guidelines and to test and identify performance enhancing interventions.

The proposed framework for evaluation of adherence to guidelines serves to improve comparability across studies by suggesting a standard presentation of results. The first stage of analysis relies on a simple classification matrix approach which provides a basic overview of adherence and important additional information on non-adherence, splitting non-adherent cases into instances of under-treatment and over-treatment in accordance with a guideline. Summary measures of the classification matrix are provided and can easily be constructed from existing data given that results are reported in a classification matrix. In the second stage of analysis a state-of-the-art regression approach which relies on data and not on assumptions about functional form is used to generate a profile of estimated decisions. Adherence to the guideline is then assessed by comparing the estimated profile of decisions to the practice guideline by identifying key indicators within this profile. Because the estimation procedure is carried out independently of information about the guideline, the estimated profile of decisions serves as a test for the

adherence to a guideline within a set of clinical observations. In addition, this lends to a naturally intuitive visual presentation of the results with the guideline and estimated profile of decisions plotted on the same axes.

Along with the analytical method proposed, this study contributes to the literature on adherence with an exploration of non-adherent decisions. Most studies of guideline compliance focus attention on the proportion of correct applications of a guideline. This work establishes a reference method and reasoning for including the evaluation of decisions which do not adhere to guidelines as well. Using a simple classification matrix, non-adherent decisions can be categorized as either over-treatment or under-treatment according to the guideline. Reducing over-treatment decisions represents potential future cost savings, while reducing under-treatment decisions represents potential future improvements in care. Non-adherent decisions are often not investigated in studies of guideline adherence. The reporting of this information can provide important clues to understanding the effectiveness of a guideline.

In the language of Tugwell, Bennett, Sackett, and Haynes (1985), this paper represents a single element of a continuous system for assessing the value of any health care intervention on burden of illness. The framework proposed here relies on patient level administrative data to assess practitioner compliance with an established guideline. This work fits into the broader literature on health system assessment and should expand the range of health system monitoring to a variety of guidelines using existing administrative data as well as improving the comparability and reliability of the results.

The framework will be developed and illustrated with the assistance of an application to a data set consisting of the decisions of volunteer medical first responders (MFRs) to administer supplemental oxygen to patients encountered during regular service duties. Section 2 describes the data. Sections 3 and 5

present descriptions of each stage of the framework followed by applications to the data in Sections 4 and 6. The medical and economic impacts of the decisions made by MFRs are examined with the results indicating that overall, adherence is poor. There is a tendency towards over-treatment rather than failure to treat medically necessary cases. This effect is more pronounced for serious incidents. Fortunately, in economic terms the impact of these over-treatment decisions is low and in medical terms, potentially beneficial.

2 Data

An illustration of the proposed framework is provided using a data set consisting of the observed oxygen administration decisions of volunteer MFRs as they carried out regular duties in a large metropolitan area over the years 2010-2014. MFRs come from a variety of backgrounds, some with prior experience with oxygen administration and others without. Extensive mandatory training is provided at no cost to MFRs accepted into the organization. All MFRs are well informed that any deviations either below or above the standard of care practice guidelines outlined in their training qualify as a breach of duty and the legal ramifications of such breaches are discussed in depth.[2] Given the voluntary nature of the organization, additional group training sessions have been favoured to direct enforcement of guidelines through penalties imposed upon individual MFRs.

It is important to note that for legal purposes any care provided outside of the pre-defined scope of practice qualifies as a breach of duty. Thus, while in uniform with the organization a physician, nurse or paramedic is expected to perform to the standard of care defined by the organization and not that

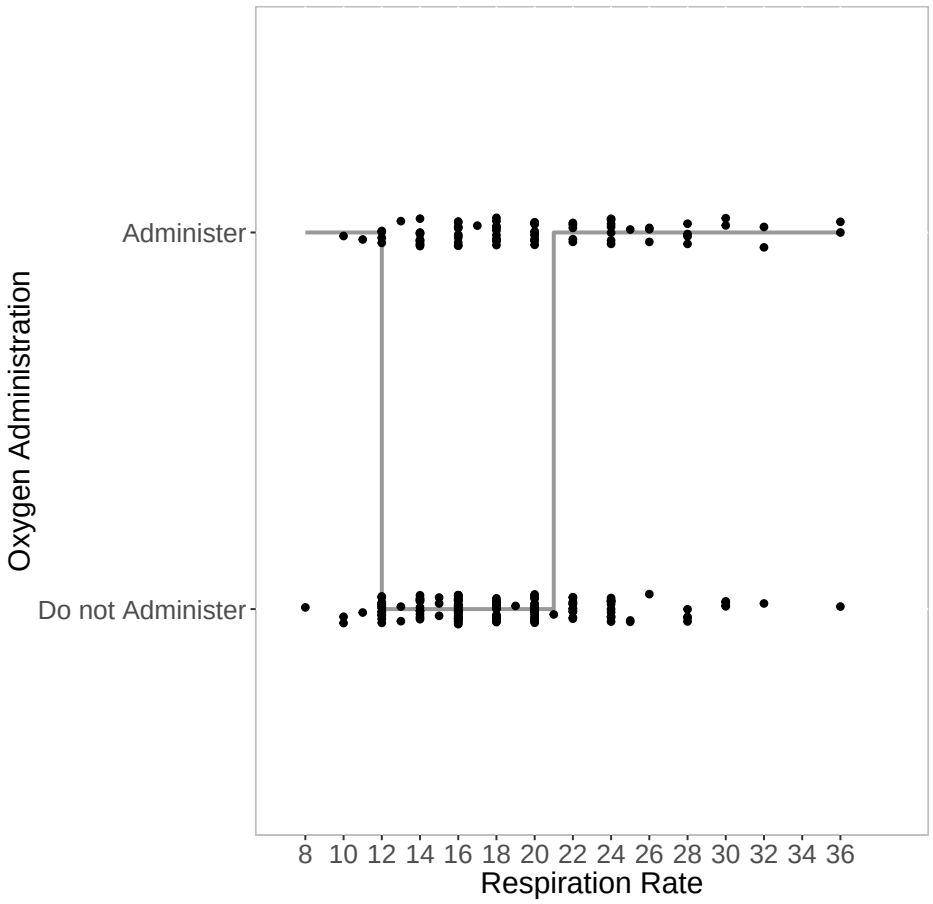
[2]Responders are made aware that they may be called to testify in court for any treatment administered, that they can be represented by the lawyers of the organization, and that ultimately they carry the legal burden to correctly follow the standard of care outlined by the organization.

which their medical training might dictate. It is therefore expected that practice guidelines are strictly adhered to in the field. However, supplemental oxygen is rarely harmful for otherwise healthy patients and may even be used as a comfort measure to provide both patients and responders with the sense that they are doing something to improve a situation, regardless of medical benefit. For this reason non-adherence to the guideline is informative, and more specifically, the manner in which this non-adherence occurs. Administering oxygen which is not medically necessary is unlikely to expose a patient to a health risk, while not administering oxygen when the guideline dictates doing so may have substantial adverse medical consequences.[3]

The medical practice guideline in place was developed and issued as a standing order by the organization's Provincial Medical Director. Proper assessment of respiration rate and administration of supplemental oxygen is taught to all MFRs during training. The guideline requires MFRs to administer supplemental oxygen if an adult patient exhibits a respiration rate less than 12 or greater than 20 breaths per minute. Plotting the expected oxygen administration decisions over respiration rates ranging from 0-36 results in a straight line at 1 ('Administer') with a step down to 0 ('Do not Administer') at 12 and a step up to 1 ('Administer') at 20 and is illustrated in Figure 1. With perfect compliance the points, which represent the observed administration decisions (jittered slightly), would align with the horizontal portions of the line.

An anonymized set of 5 years worth of treatment decisions of volunteer MFRs form the data set used in this study. The international volunteer organization is sub-divided into national and provincial sub-units, and further sectioned into municipal level units. MFR training is standardized at the provincial level. The municipal level MFR unit covered by the data operates within a

[3]It is worth noting here that patients refuse oxygen at times. Patient wishes are respected at all times.



Solid lines represent the recommendation. Points represent observed decisions.

Figure 1: Recommended and observed oxygen administration decisions by respiration rate.

large Canadian metropolitan area, providing first-aid at local sporting events, concerts, festivals and public functions. Only fully certified MFRs are qualified to wear the full uniform of the organization and to provide first-aid. MFRs are also trained to complete a Patient Care Report (PCR) in the event of delivery of any form of first aid to a patient. PCRs are filed with the unit chief at the end of each duty shift.

The database includes a total of 898 medical-encounter records containing information on patient age, type of medical situation encountered by the MFRs, administration of oxygen, vital signs including respiration rate, and whether

or not Emergency Medical Services (EMS) were contacted. All data were anonymized.[4] Reports missing a year of birth of the patient were excluded (71 cases), as were the records of patients under the age of 18 (271 records) because a different guideline applies to these cases. As well, a single outlier which appeared to be a recording error in the respiration rate was dropped (1 record). Of these remaining 555 cases, 315 cases were missing a recorded respiration rate (these cases are discussed in more detail in Subsection 2.1. These exclusions resulted in 240 complete cases.

Table 1 displays the variables considered in the analysis of supplemental oxygen administration behaviour. The use of supplemental oxygen is recorded simply as 0 if it was not used and 1 if it was used. MFRs monitor vital signs throughout an incident, up to 5 times for a single patient. RR1 is respiration rate recorded at the first vital signs check. Respiration Rate values in the range of 12-20 respirations per minute are considered normal for adults. EMS refers to whether or not emergency services were called to the scene (0 if not called and 1 if called). This variable acts as an indicator of call seriousness and is included because, during training, potential responders are told to call EMS in any situation serious enough to warrant the use of supplemental oxygen. It is therefore expected that the correlation between O2 and EMS would be very high, yet this is not the case. The Pearson correlation coefficient is 0.48 indicating a fair amount, but not complete, correlation; deviations can be explained by the finding that there is substantial administration of oxygen to less serious patients.

2.1 Missing vital signs

In the original set of observations, 315 of the available 555 observations (57%) are missing information on patient vital signs. For patients with minor in-

[4]Reclassification ensured at least 5 observations per cell.

Variable Set

Variable	Min	Max	Description
Year	2010	2014	Year of incident
O2	0	1	Supplemental Oxygen Used No=0, Yes=1
RR1	8	36	Respiration Rate recorded for Vitals Check
EMS	0	1	Emergency medical services called No=0, Yes=1

Table 1: Variables used in analysis of supplemental oxygen decisions.

juries vital signs are often not collected by MFRs. Typically this is because a vital sign check would seem invasive when all that is required is very basic first aid. This is not the only reason for missing vital signs, however. The patient might have refused, or for a very small proportion of cases, the call might have been too serious to record vital signs prior to the arrival of EMS, for example. The majority are non-serious cases of minimal first-aid. If these cases with missing vital signs are excluded the sample is substantially biased towards more serious patients. Therefore the missing records could have a substantial impact on the estimate of overall adherence to guidelines within this unit of the organization if a large proportion of cases without vital signs recorded are attributable to patients with normal range respiration rates who were correctly not administered oxygen in accordance with the guideline.
In order to improve the representativeness of the sample, missing data values were imputed. For the purposes of imputation, all respiration rates associated with cases in which vital signs were not recorded were assumed to fall within normal range. Based on this assumption, missing respiration rates were imputed by taking a random draw of the normal-range respiration rates. Personal communication with the municipal organization's unit Chief in 2015 confirmed this as an acceptable representation of the vast majority cases for which respiration rates were not recorded. Although this is not the only reason for a missing record, it is the most common. Replacing the missing respiration rates with 315 random draws (with replacement) from the pool of

observed normal range respiration rates increases the set of complete records to 555 observations. All analysis is undertaken using this augmented set of 555 observations. Analysis using the original data without the imputed values is contained in Appendix Section A for reference. The results lead to the same overall conclusions.

Figure 2 provides a visual summary of the key variables used in the analysis. The first pane highlights the unbalanced nature of oxygen administration decisions; there are nearly double the number of observations for 'Do not Administer' as there are for 'Administer'. The second pane highlights the fact that most treatment given by MFRs is not serious enough to warrant a call to EMS. In the third pane the Epanechnikov kernel density of respiration rates (RR1) is presented. The location of the mean of this distribution to the right of the mode suggests that this variable is not normally distributed. A Shapiro-Wilk test confirms this result.

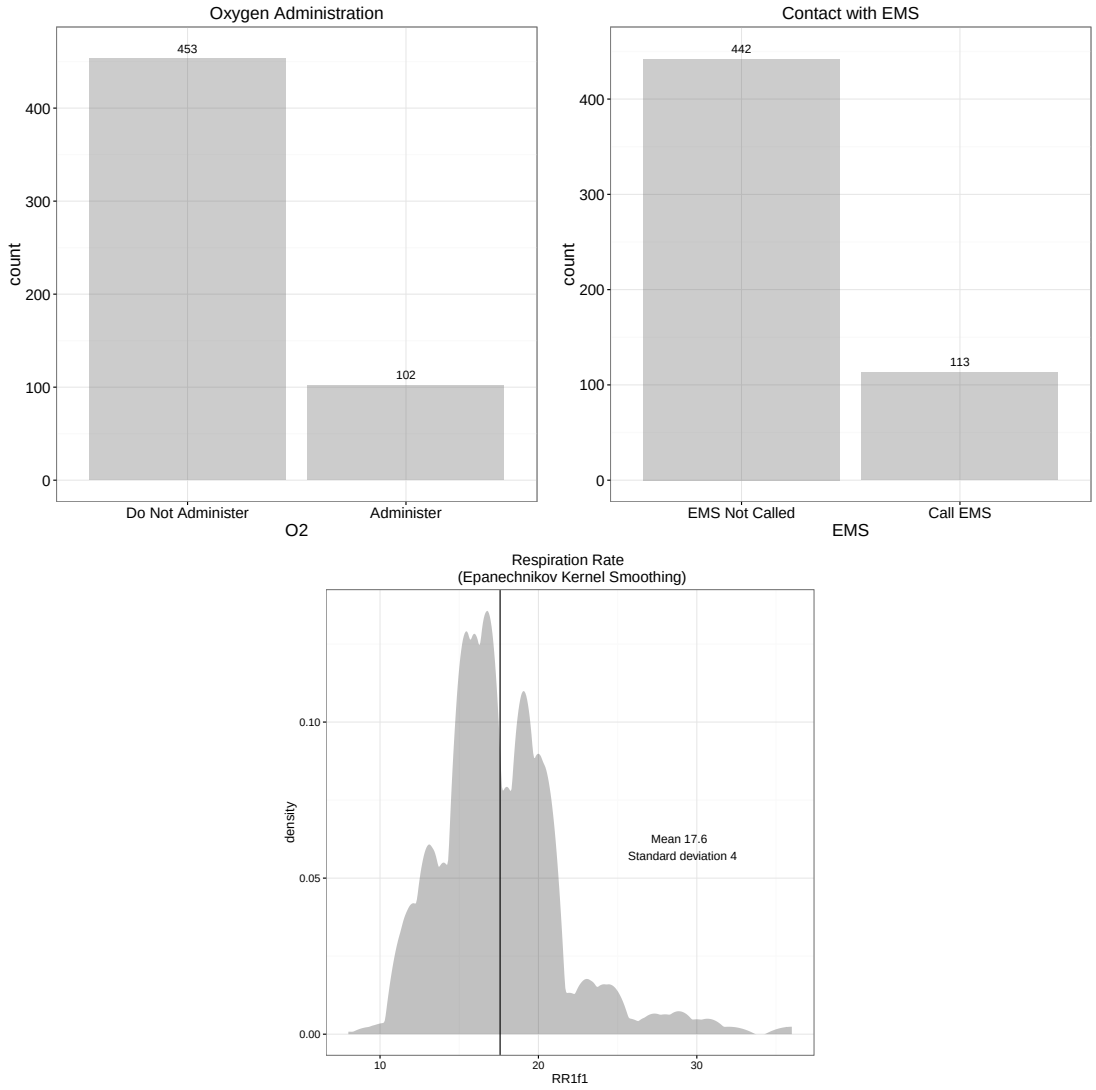
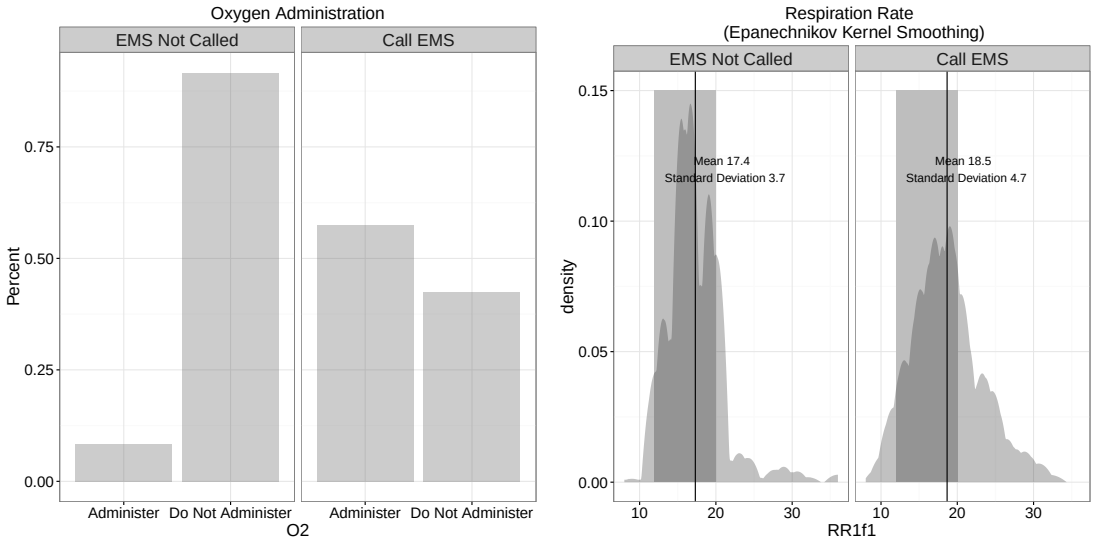


Figure 2: Illustrated data summary.

2.2 Contact with emergency services

In addition to the overall summary data reported in Figure 2, Figure 3 divides the data across categories of EMS Contact: 'EMS Not Called' and 'Call EMS'. Contact with Emergency Medical Services (EMS) is an indicator of the seriousness of an incident. A chi-square test for independence rejects the notion that Oxygen Administration, Respiration Rate and EMS calls are independent. Oxygen Administration is found to have significant correlation

with EMS contact (0.51 with a p-value of 0.0000 using the Pearson product-moment correlation), but not so much as to represent multicollinearity.
MFRs are trained to respond in serious emergencies according to procedures set out by the medical director of the organization. Acceptance to the role of MFR is dependent on the ability of candidates to follow such procedures accurately and this training is reviewed frequently. Serious cases receive much attention in training, and there is little to no room for mistakes in the administration of oxygen or making contact with EMS in such cases. Failure to apply oxygen when it is necessary can have substantial adverse health consequences for patients. In examinations, failure to apply oxygen when needed disqualifies an MFR candidate. In less serious cases patients demonstrate less frequent need for oxygen or EMS assistance. A difference in adherence across serious and non-serious cases is possible if more training for serious incidents affects adherence rates when such cases are encountered in the field. Figure 2 demonstrated that in the field the majority of incidents (approximately 80% of cases) are non-serious. MFRs thus also acquire experience in dealing with non-serious incidents which may also influence adherence rates across levels of case seriousness. Unfortunately, the characteristics of individual MFR experience were not available in this data set and so contact with EMS serves as the only means of evaluating the effect of scene seriousness on guideline adherence. As well, patient outcomes following treatment are unavailable so the health impacts of MFR services are unknown.
Figure 3 presents the results by EMS contact sub-groups. The first pane makes clear that when contact with emergency services is not made, oxygen is usually not administered. The second pane highlights a skew towards normal range respiration rates when EMS Services are not contacted. In this pane the Epanechnikov kernel smoothed density of respiration rates shows a marked difference in distributions across EMS contact levels. Normal range respira-



EMS is emergency medical services.

Figure 3: Illustrated data summary by EMS contact sub-group.

tion rates are highlighted in the darkened grey box.[5] These results inform the inclusion of EMS contact as an important marker of call seriousness, as illustrated by the difference in respiration rates, and also as an influence upon oxygen administration decisions, as shown by the relative percentages of administration decisions.

3 Method stage 1

The first stage of assessment involves simply sorting each recorded oxygen administration decision according to whether or not these decisions adhere or do not adhere to the guideline, based on the respiration rate associated with the decision. A classification matrix provides a concise means of summarizing the results. Each cell of a classification matrix represents the number of observations which satisfy the conditions specified in the row and column. The basic format is illustrated in Table 2.

Table 2 labels the totals in each cell. True negatives for example, are data

[5]Figure A.2 presents the result without the imputed values. The difference remains.

		Observed Decisions	
		0	1
Guideline	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Table 2: Classification matrix guide.

points in which no action was taken and the guideline suggested no action should be taken. These decisions are thus adherent to the guideline. Similarly, decisions in the cell row labelled 1 and column labelled 1 indicate decisions in which action was taken and the guideline suggested that action should be taken. These decisions thus also adhere to the guideline. In terms of oxygen administration decisions the possible categorizations entail:

True Positive: The guideline recommends oxygen administration and oxygen was administered.

True Negative: The guideline recommends no oxygen administration and oxygen was not administered.

False Negative: The guideline recommends oxygen administration and oxygen was not administered ('under administration').

False Positive: The guideline recommends no oxygen administration and oxygen was administered ('over administration').

The classification matrix approach thus provides information, not only on adherence to guidelines, but on non-adherence as well. Estimates of over and under provision of care with reference to a guideline can be made easily.

3.1 Summary measures of correct classification

Table 3 presents a set of measures which can be used to summarize the classification matrix. Because the classification matrix is unit-free these measures are generally comparable across applications. It is important to note, how-

Measure	Formula
True Positive Rate	$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$
True Negative Rate	$\text{TNR} = \frac{\text{TN}}{\text{TN+FP}}$
False Positive Rate	$\text{FPR} = \frac{\text{FP}}{\text{TN+FP}}$
Correct Classification Ratio	$\text{CCR} = \frac{\text{TP+TN}}{\text{Total}}$
Balanced Correct Classification Ratio	$\text{bCCR} = 0.5(\text{TPR} + \text{TNR})$
Area Under the Curve	$\text{AUC} = \int_{-\infty}^{\infty} TPR(t)FPR'(t)dt$
Cohen's kappa	$\kappa = \frac{p_0 - p_e}{1 - p_e}$

TN= True Negatives, TP = True Positives, FN= False Negatives, FP = False Positives, and Total is the sum of all entries in the classification matrix: TP+TN+TN+FN.
t is the threshold used for sorting observations.
$p_e = \frac{(\text{TN+FN})}{\text{Total}} * \frac{(\text{TN+FP})}{\text{Total}} + \frac{(\text{FP+TP})}{\text{Total}} * \frac{(\text{FN+TP})}{\text{Total}}$
$p_0 = \frac{\text{TP+TN}}{N}$

Table 3: Classification matrix summary measures.

ever, that not all the measures listed in Table 3 are equally suitable for all data sets. For example, the Correct Classification Ratio (CCR) is the proportion of adherent decisions in the data and is therefore a basic summary measure of adherence. The CCR does not account for the fact that the more frequent outcome has a greater probability of adhering to the guideline simply by chance. Straube and Krell (2014) also note that this measure is biased when the data are unbalanced, which, as noted in Section 2 is the case for the illustrative application.

Four metrics are suggested by Straube and Krell (2014) to improve the informative capacity of classification matrix summarization on the basis of being insensitive to class imbalance: The Balanced Correct Classification Ratio (bCCR), the Geometric Mean (G-mean), Area Under the Receiver Operator

Characteristics Curve (AUC) and d-prime. The bCCR and AUC metrics are applied here because these are both simple to understand and insensitive to class-imbalance. The bCCR is related to the CRR, and improves upon the CCR by taking into account class imbalance. If the the sample of data used were balanced the bCCR with be equivalent to the CCR. The AUC is another measure of the degree of match between observed outcomes and the guideline. The AUC is a proportional measure which ranges from 0.5 to 1.00 and is commonly used to evaluate the match between predictions and observations over a broad a range of scientific settings, thus making this measure accessible to a variety audiences. An approximate classification of the level of match for AUC values is presented in Table 4.

AUC Value	Performance
0.5 - less than 0.6	Fail
0.6- less than 0.7	Poor
0.7 - less than 0.8	Fair
0.8 - less than 0.9	Good
0.9 - 1.00	Excellent

Source: Tape (2015).
AUC is area under the receiver operator characteristics curve.

Table 4: Classification of AUC values.

At this stage of analysis outcomes are binary (0 for 'Do Not Administer' decisions and 1 for 'Administer' decisions) so the threshold (t) is irrelevant, but Section 6 will discuss the AUC in greater detail, making use of this measure with relevant threshold values. Jeni, Cohn, and De La Torre (2013) show that the AUC is insensitive to skewness, where $skewness = \frac{FN+TP}{TN+FP}$, but that this measure can mask poor performance of a classifier. Fortunately, such masking is not an issue at this stage of analysis due to the binary nature of the outcomes.[6]

[6]A classifier is a model which predicts outcomes. These measures are often used in eval-

Cohen's κ (Cohen, 1960) is used to assess the amount of agreement between the guideline and the observations beyond that which would be achieved by chance. Jeni et al. (2013) show that Cohen's κ is sensitive to both the degree of skewness and rate of misclassification in simulation experiments. However, for rates of skewness between 0.1 and 10 and a misclassification rate of 1%, the accuracy of Cohen's κ is near 95%, so this measure will be quite accurate for many data sets. While no single metric dominates in terms of providing information, Cohen's κ is important for summarizing the rate of adherence to a guideline net of chance.

4 Application: Stage 1 adherence results

This section reports the Stage 1 results of the standardized analysis of the guideline evaluation framework applied to the data described in Section 2. As previously mentioned, these data showed substantive differences across sub-groups of EMS contact, therefore the analyses are displayed for each sub-group.

Table 5 presents the classification matrix from the sub-group defined by non-contact with EMS. These cases required less serious first-aid treatment. Non-applications of oxygen feature prominently in this sub-group, with a skewness measure of 0.09. This level of skewness implies that the use of skew-insensitive classification matrix summary metrics are appropriate.

Non-adherent decisions represent a very small share of the total decisions for this sub-group and there is virtually no difference in the type of non-adherent decisions across sub-groups. The equal distribution of non-adherent decisions across categories of under- and over- administration provides no evidence to

uating model performance, i.e. the ability of a predictive model which generates values between 0 and 1 to match with the observed outcomes which are 0 or 1. This stage of the analysis simply evaluates the match of the observed outcomes, which are 0's and 1's and thus do not require a threshold for sorting, with the guideline recommendations (0 or 1).

support the claim that MFRs have a tendency towards administering oxygen to non-emergency patients as a comfort measure any more than they have a tendency towards neglecting to administer oxygen to patients who demonstrate need for it based on respiration rate in non-emergency situations.

		Observed Oxygen Use		
		Not Administered	Administered	Total
Guideline	Not Administered	376	30	406
	Administered	29	7	36
	Total	405	37	442

Cells indicate the number of recorded cases which agree (disagree) with the respiration rate guideline for administration of supplemental oxygen.

Table 5: Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is not called.

Table 6 tells a different story for the sub-group defined by EMS contact. This sub-sample is more balanced, with a skewness measure of just 1.35 which indicates a very slightly larger share of 'Administer' decisions than 'Do Not Administer'.[7] Non-adherent decisions in this sub-group represent a larger absolute share of decisions than in the non EMS contact sub-group (49% vs 13%), and this difference is significant using a two sample test for equality of proportions with continuity correction (p-value of 0). As well, the type of non-adherence exhibited within this sub-group is significantly biased towards administering oxygen when the guideline recommends against it (p-value of 0, using a two-sample test for equality of proportions with continuity correction). This result combined with the overall greater proportion of administration decisions made in this sub-group suggests that when faced with a serious scene MFRS tend towards administering oxygen regardless of the guideline.

[7]Skewness values in the range of [0.1-10] are symmetric around 1. Thus, a skewness value of 9.14 would be an equivalent degree of skewness towards 'Administer' decisions as the skewness towards 'Do Not Administer' decisions observed in the sub-group defined by non-contact with EMS.

		Observed Oxygen Use		
		Not Administered	Administered	Total
Guideline	Not Administered	37	44	81
	Administered	11	21	32
	Total	48	65	113

Cells indicate the number of recorded cases which agree (disagree) with the respiration rate guideline for administration of supplemental oxygen.

Table 6: Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is called.

The classification matrix summary measures presented in Table 7 suggest that the guideline is not well followed. The substantial impact of skewness is illustrated in the EMS-not-called sub-group as the CCR drops by 36% moving to the skew-insensitive bCCR. Across sub-groups the skew-insensitive metrics bCCR and AUC are virtually identical. The AUC results suggest failure to adhere with the guideline. Cohen's κ suggests that only 12% of decisions above and beyond chance are adherent to the guideline in the EMS-not-called sub-group, and only 9% of decisions are adherent to the guideline in the EMS-called sub-group. In the EMS-not-called sub-group the level of skewness suggests that Cohen's κ may even be slightly overstated.

Measure	EMS Not Called	EMS Called
TPR	18.92%	32.31%
TNR	92.84%	77.08%
FPR	7.16%	22.92%
CCR	86.65%	51.33%
bCCR	55.88%	54.7%
AUC	0.56	0.55
Cohen's κ	11.9%	8.62%

Table 7: Classification matrix summary measures by EMS contact sub-group.

4.1 The economic consequences of non-adherence

Guidelines are expected to encourage the appropriate use of medical resources. This section estimates the annual cost to deliver supplementary oxygen to patients and the potential savings from greater adherence to the administration guideline. The organization reports annual oxygen costs of $400 per year. Over 2010-2014 a total of 898 patients were treated. Approximately 30% of these cases were children under the age of 18 (271 cases), which follow a different oxygen administration guideline. Attributing 70% of oxygen costs to adults implies an annual cost of $280. Costs attributable to each sub-group are obtained by simply dividing the costs in proportion to the number of treated patients in each sub-group.

Table 8 reports the estimated costs of oxygen per patient treated with oxygen. Training in correct oxygen delivery is a substantial part of MFR training and oxygen delivery equipment represents approximately half of the total amount of equipment required to be carried by MFRs while on duty. Table 8, however, reflects that despite the substantial training and equipment devoted to oxygen administration, overall oxygen use in infrequent. In total the sample contains 37 instances of oxygen use when EMS was contacted and 65 total cases when EMS was not contacted, as reported in Tables 6 and 5. The mean annual number of patients treated is estimated by taking the mean of the number of patients in each year from 2010-2014. Only patients treated with oxygen incur costs attributable to oxygen, [8] so costs per patient are simply total costs per year divided by the average number of treated patients per year. The table reflects that oxygen was administered to patients both when EMS was called and when EMS was not called. Since responders may behave differently when faced with more serious situations, such as those requiring contact with EMS, these sub-groups are presented in separate columns.

[8] There is insufficient information to justify different costs across levels of adherence.

Measure	EMS Not Called	EMS Called
All Cases		
Treated Patients	7.4	13
Total Costs	$ 101.57	$ 178.43
Cost/patient	$ 13.73	$ 13.73
Guideline Non-Adherent Cases		
Over-Administration		
Treated Patients	6	8.8
Unjustified Costs	$ 82.38	$ 120.82
Under-Administration		
Untreated Patients	5.8	2.2
Unjustified Savings	($ 79.63)	($ 30.21)
Perfect Guideline Adherence		
Treated Patients	7.2	6.4
Total Cost	$ 98.86	$ 87.87
% Savings under Perfect Adherence	3%	51%

Table 8: Annual cost of oxygen delivery by EMS contact sub-group.

To estimate costs under perfect adherence, the mean number of patients per year is simply multiplied by the cost per treated patient to obtain a figure for the average cost all treated patients. The cost of over-treatment (over-administration) is then deducted from the total cost figure for each sub-group and under-treatment added. The percentage of actual cost that will be saved under perfect adherence is equal to 100 times the ratio of the difference between the actual cost and the cost with perfect adherence to the actual cost. It cost donors and community sponsors $13.73 to administer oxygen to a patient over the five years 2010 through 2014. If adherence to the guidelines was perfect a 3% reduction in expenditures would have been realized for non-serious cases, and a 51% reduction in expenditures would have been realized for serious cases. Bearing in mind that only 20% of all patient care reports reviewed were serious enough to warrant contact with EMS (21% per year on average), efforts to encourage greater adherence may be better directed at other activities within the organization. In addition to representing only

a minor level of costs, a large proportion of non-adherent decisions occur in serious situations and constitute over-administration. In serious emergencies supplemental oxygen has potentially life-saving benefits and has a low risk of adverse medical events associated with administration. The benefits of such 'better safe than sorry' over-administration likely outweigh the costs for the organization. The continuous training to recognize serious emergencies approach rather than strict guideline enforcement appears to serve emergency patients who require oxygen rather well, while carrying minimal negative impacts on patients who do not need oxygen. In this situation advocating for strict guideline adherence may have a negative impact on oxygen administration practices. This is because a strict policy of guideline adherence could very well crowd-out the MFR's focus on accurate scene assessment and patient well-being which informs administration of oxygen in a serious situation and redirecting attention towards recalling and implementing guidelines. The reaction might be seen as a type of 'crowding-out' of focus; Frey and Jegen (2001) give examples of the empirical relevance of such effects. The potential for stricter adherence strategies to induce unanticipated negative reactions warrants careful examination of the behavioural context in which a guideline is situated and should ideally be carried out prior to implementation.

5 Method stage 2

The classification matrix approach is a useful first stage in evaluation, however many studies take a regression approach to evaluating guideline adherence. When there are more than 2 or 3 covariates the classification matrix approach can become unwieldy, as a set of matrices must be computed for each level of each covariate. Stage 2 of the analysis presents a state-of-the-art regression framework. This framework removes the need to select an appropri-

ate parametric form for estimation, eliminating errors arising due to model misspecification. The framework also has variable selection embedded, as covariates which do not influence the outcome are eliminated from the regression. Section 6.1 compares the performance of the regression framework proposed to various alternatives using the AUC as a metric for comparison. In economics, applications involving evaluation of binary outcome data, such as the 'Administer' and 'Do Not Administer' outcomes in the example, rely heavily upon parametric smoothing approaches such as probit, logit, or linear regression, sometimes with higher order terms. This preference is reflected in the health economics literature. An EconLit search of 'Guideline*' and 'Medical' returned 115 articles. After sorting for guideline adherence 30 were retained and reviewed. The main analysis styles applied were survival analysis, probit regression, logit regression and OLS or variants thereof (fixed effects or random effects). The selection of a particular regression model is often dictated by the type of data encountered, and the guideline itself. As will be illustrated in Section 6.1 below, these regression models can inaccurately identify distinct patterns within a set of observations which appear to match with the profile of guideline adherent decisions. Because of this, I propose a regression approach that applies nonparametric conditional density estimation to assess the pattern of observed treatment decisions. This approach generates a profile of estimated probabilities of administering oxygen over the range of respiration rates for each level of EMS contact. The profile of estimates may take on any form, and so may or may not follow a pattern reflective of the guideline itself. The guideline need not even be known at the outset of applying this regression strategy.

The guideline instructs administration of oxygen to adult patients when respiration rates are less than 12 breaths per minute, no administration of oxygen when respiration rates are in the normal range of 12 to 20 breaths per

minute, and administration when respiration rates exceed 20 breaths per minute. Plotting the guideline over a range of respiration rates results in a solid line at 'Administer' with a single downward step the 'Do Not Administer' at a respiration rate of 12, and a single step upwards, back to 'Administer' at a respiration rate of 20, as was shown in Figure 1. If the observations reflect this guideline it is expected that the estimates generated by the regression will show a region of concentrated change in the negative direction at 12 breaths per minute and a region of concentrated change in the positive direction at 20 breaths per minute. These changes in the estimates are summarized by the gradients (instantaneous rates of change). The minimum and maximum valued gradients of the estimates are used to identify the respiration rates where the most concentrated changes occur in the profile of estimates for each sub-group, and over all encountered respiration rates. The identified respiration rates form 'candidate steps' which can then be compared to the guideline 'steps' of 12 and 20. Bootstrapped confidence intervals are used for this comparison.

Nonparametric conditional density estimation makes as few distributional assumptions about the data as possible. In a regression context, this means that issues of model misspecification arising due to incorrect parametric assumptions about errors are avoided. Because no form is specified, it also means that the patterns unmasked by such estimation may serve as the basis for a hypothesis test. Estimates resulting from nonparametric estimation techniques may be better able to suggest evidence for or against guideline adherence than parametric alternatives alone. Three other approaches to evaluation are illustrated for comparison: an Empirical approach which plots simple proportions, a parametric regression which fits a quadratic specification of the relationship between respiration rate, call seriousness and oxygen administration; and a search approach which simply chooses the guideline of best

fit based on maximization of the AUC.

Nonparametric conditional density estimation is undertaken in the statistical environment R (R Core Team, 2015) using the *np* package (Hayfield and Racine, 2008). All that is required is to first compute optimal bandwidths and secondly to generate the estimated values. All interactions between covariates are automatically accounted for and irrelevant covariates are smoothed out asymptotically. It is important to note that this method does not deliver scalar coefficients as parametric regression does because the estimates are not confined to taking on a particular form (such as a linear form in linear regression). This means that statistical tests on parameter values (i.e. tests on coefficients) are generally out of the question and motivates the use of 'candidate steps' described above.

Smoothing nonparametrically relies on the observations rather than a prespecified functional form for the resulting estimates. An equation with respiration rate (*RR1*) and contact with emergency medical services (*EMS*) entering as explanatory variables defines the relationship and fitting is done via the estimation of optimal bandwidths for kernel conditional density estimation. All interaction effects are automatically incorporated. Nonparametric conditional density estimation was described first by Stone (1977) and more recently described by Hall, Racine, and Li (2004). Appendix B provides greater detail of the estimation strategy. All that is required for implementation is to enter the formula:

$$\texttt{O2} \sim \texttt{RR1} + \texttt{EMS}, \tag{1}$$

in the R computer package (R Core Team, 2015), where O2 is the binary outcome of the oxygen administration decision made by an individual MFR. This formula is used first in the determination of the optimal bandwidth sizes.

The routine which determines the bandwidth automatically accounts for interactions between *RR1* and *EMS*. Having computed the bandwidths, the estimates and gradients are then generated over the range of values encountered in the data. Once a profile of estimates is obtained, the identification of candidate steps is carried out by simply selecting the respiration rates associated with the lowest and highest gradients. The main aim of this methodology is to determine whether observations are reflective of a pre-set practice guideline. Generating a profile of estimates offers insight into the shape of the responses. Within the profile of estimates, identifying regions of greatest change in the negative and positive directions provides candidate steps which can now be compared to the guideline. For guidelines involving single steps up or down see Thomas (2016).

Comparison with the guideline proceeds by creating confidence intervals by bootstrapping. This method simply reconstructs the estimates using a resample of the original observations of the same size and with replacement. The process is repeated 1000 times and the 5th and 95th percentiles of the results taken as the boundaries of the 90% confidence interval. This avoids the need to construct a bootstrap sample which is consistent with the null hypothesis in order to carry out a full nonparametric hypothesis test, which can be a complex task, especially for non-standard estimators (MacKinnon, 2007). Support for a particular step matching with the guideline is offered if the guideline falls within the bounds of the confidence interval around a step.

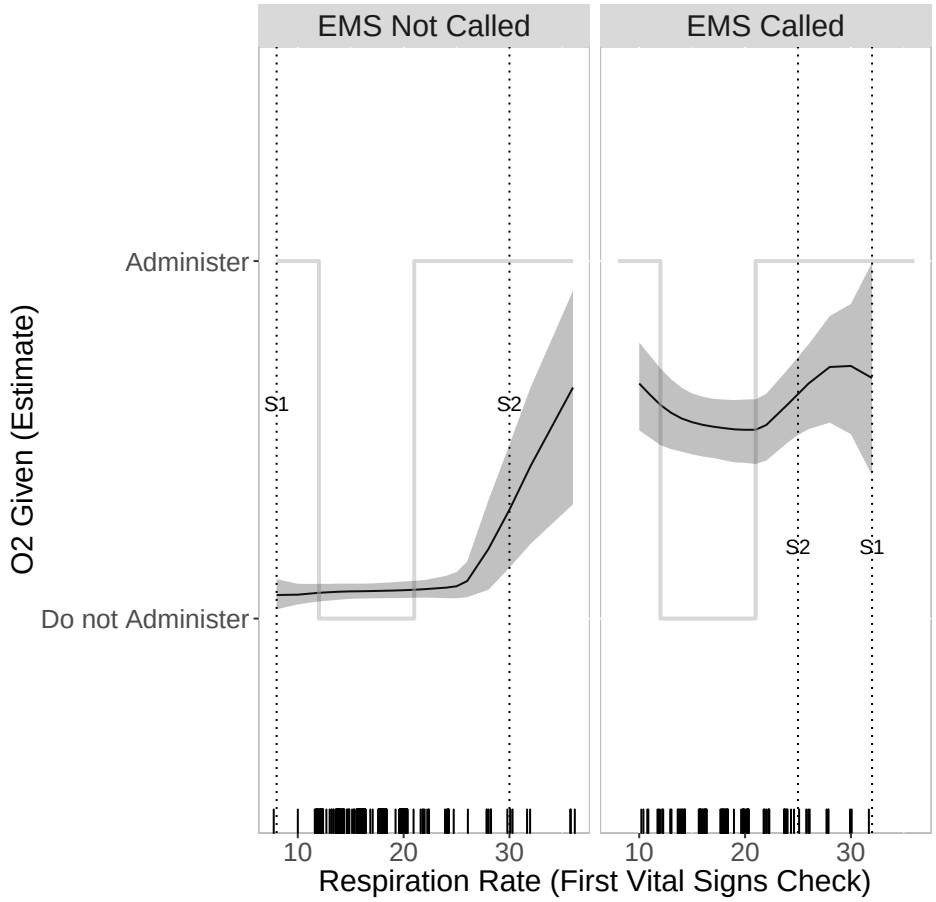
6 Application: Stage 2 adherence results

Figure 4 presents the results of the nonparametric strategy, the two 'candidate steps', and the practice guideline, along with the bootstrapped 90% confidence bounds. The results illustrate a partial effects surface, because the

nonparametric routine fits a multi-dimensional surface. The partial effects are therefore 'slices' through the three dimensional prediction object (*RR1*, *EMS* and the prediction $\hat{O}2$) along each ridge defining whether or not EMS was called and mapped over the range of all *RR1* recorded. All interactions are accounted for automatically and the prediction surface can take on any shape. All estimates are projected over the full range of possible values of respiration rates encountered in the data set.

Figure 4 makes clear that there is a difference in the probability of using supplemental oxygen depending on the sub-group defined by whether or not the situation warranted a call to EMS. The pattern which would suggest guideline adherence should have candidate Step 1 and Step 2 values in order S1, S2 and near respiration rates of 12 and 20. Neither sub-group thus exhibits a pattern indicating guideline adherence. Visual inspection reveals that there is a lower overall use of oxygen in the 'EMS Not Called' sub-group. As well, there appears to be a slight indication of a U-shape in the cases where EMS was called, but the candidate steps do not identify this pattern at all. In 'EMS-called' cases the first step, the step downward to 'Do Not Administer', is located at a higher respiration rate (32) than the upward step to 'Administer', which is located at the respiration rate of 25. In 'EMS Not Called' cases the estimates slope upwards very slowly.

To test the precision of the candidate steps bootstrapped confidence intervals are constructed. Table 9 reports the steps, confidence intervals, guideline values and gradients for for Step 1 and Step 2 for each sub-group. The confidence intervals for Step1 and Step 2 are nearly as wide as the data range in all cases except Step 2 in the 'EMS Not Called' sub-group, which occurs at the upper bound with greater precision than the other candidates. In both sub-groups the Step 1 and Step 2 confidence intervals overlap, indicating that two distinct steps are not identified. This is evidence against the finding of adher-



Guideline is the grey line. Candidate step 1 (S1) and step 2 (S2) are dotted lines. Bootstrapped 90 percent confidence interval is the shaded area. Bars along the x-axis indicate frequency of observations.

Figure 4: Estimated probability of oxygen administration by respiration rate and EMS contact sub-group using the Nonparametric approach.

ence to the guideline in both serious and non-serious cases, despite the fact that the guideline falls within the bounds of the confidence intervals in all but the Step 2 'EMS Not Called' case.

The overall result is consistent with the results of Stage 1. This Stage 2 regression approach indicates that there is little to no guideline adherence. What Stage 2 adds is the visual intuition of how decisions differ across subgroups. The weakness of adherence to the guideline is striking under this representation. In the 'No Call' case oxygen use is less frequent, but increases

EMS	Step		Lower ci	Upper ci	Guideline	Gradient
EMS Not Called	Step 1	8	8	36	12	-0.0024
EMS Not Called	Step 2	30	28	36	20	0.0524
EMS Called	Step 1	32	10	32	12	-0.0377
EMS Called	Step 2	25	22	32	20	0.0357

Table 9: Candidate steps and bootstrapped 90 percent confidence intervals using the nonparametric strategy.

at very high respiration rates. In the 'EMS-called' sub-group there is an overall greater estimated use of oxygen, slightly higher at low and high respiration rates, but not in a pattern reflective of the guideline. The areas of most concentrated change in the estimates are not distinct, as demonstrated by bootstrapped confidence bounds.

6.1 Relative performance of stage 2

In order to assess the relative performance of the nonparametric strategy used in Stage 2, I undertook comparisons with three other strategies for condensing observations into estimates. For each strategy the profile of estimates with their associated bootstrapped confidence bands is presented. The ability of each model to correctly predict the data is evaluated using the AUC metric. Next, the location of the largest changes in the negative and positive directions are identified and 90% confidence intervals generated. These results are evaluated for the presence of distinct candidate steps and their match with the guideline.

The 'empirical' strategy simply plots the proportion of 'Administer' decisions at each respiration rate. The 'linear' strategy fits an ordinary least squares regression to the data, making use of a quadratic form and interaction effects. The 'search' method evaluates the fit of every possible specification of the guideline to the data. While the nonparametric AUC results may not always be largely different from the comparisons, the nonparametric strategy

requires the input of just one calculation and is thus a more straight-forward and efficient approach to analysis. In contrast, the search approach requires the analyst to specify each possibility and to calculate every possible outcome. The estimates also serve as a visual test for the presence of the guideline.

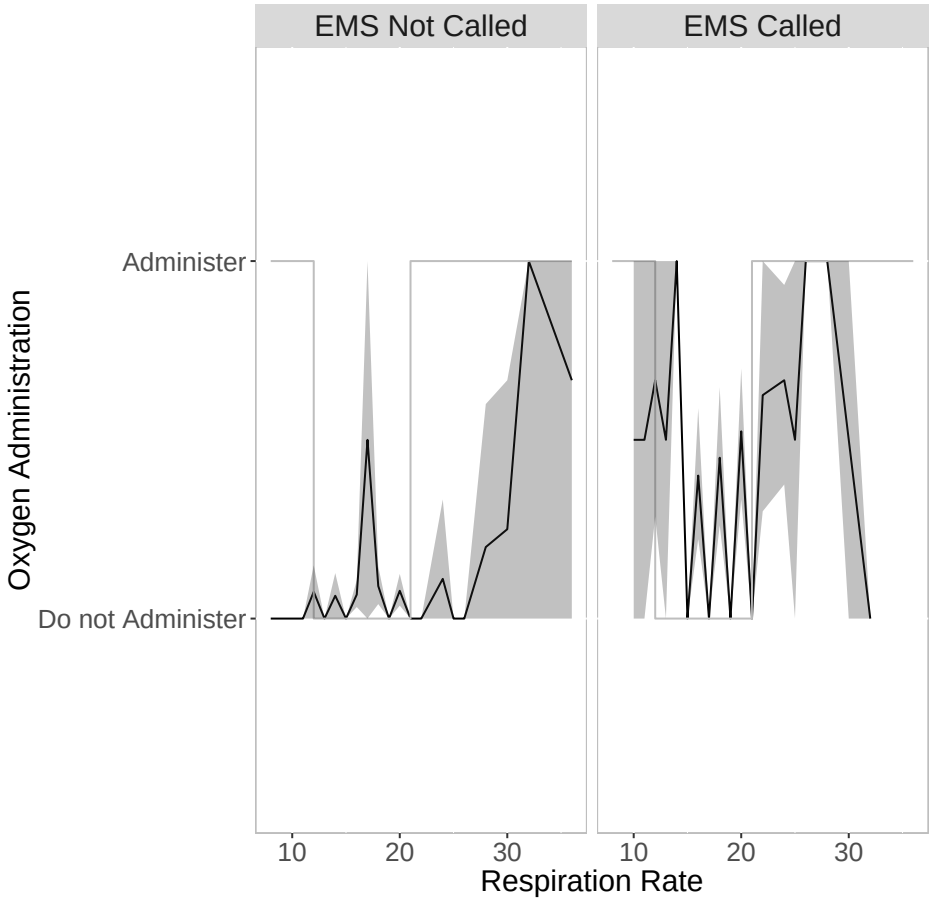
The strategies applied here can be thought of as approaches to smoothing observations. Smoothing attempts to overcome the inherent noisiness of data collected in the field. For example, the empirical strategy produces a single estimate for each level of respiration rate in each sub-group by plotting the proportion of cases in which oxygen was used over the recorded respiration rates. While some support for the reflection of a guideline is visible in the profile of estimates, which are noisy, greater smoothing could make the results clearer. For single step patterns, the main models encountered in the literature were Probit and Logit which are smooth approximations to a heavyside step function. In this case both a step down and a step up are expected so instead of a step function an approximation to a boxcar function is more appropriate and is done here by including higher order terms in a linear regression framework. This results in a U-shaped profile of estimates over respiration rate. A third approach considers each possible specification of the guideline, choosing the specification with the best fit according to the greatest AUC. The nonparametric strategy outlined in Section 5 smooths, but is also responsive enough to allow for sharp changes akin to the discrete changes defining the steps downward and upward within the estimates if the data reflect such a pattern. The main point is that such a pattern need not be defined at the outset of the analysis.

6.1.1 Empirical strategy

As already mentioned, the empirical strategy simply plots the proportion of 'Administer' decisions for each sub-group over the range of respiration rates. The result is a noisy profile of estimates. The selection of candidate steps proceeds as outlined in Section 5. In cases where multiple minimum and maximum gradient values are encountered the median value respiration rate is reported. Table 10 reports the candidate steps and confidence intervals. All candidates except Step 2 in the 'EMS Not Called' sub-group fall within the 90% confidence interval. The intervals of Step 1 and Step 2 overlap in both sub-groups offering evidence against two uniquely defined steps. Figure 5 illustrates the result. The profiles of estimates are very noisy, masking any potential steps.

EMS	Step		Lower ci	Upper ci	Guideline	Gradient
EMS Not Called	Step 1	18.0	18	36	12	-0.4091
EMS Not Called	Step 2	32.0	10	26	20	0.7500
EMS Called	Step 1	15.0	11	15	12	-1.0000
EMS Called	Step 2	22.0	10	26	20	0.6250

Table 10: Candidate steps of the empirical strategy and bootstrapped 90 percent confidence intervals.



Proportion of 'Administer' outcomes is solid black line and the guideline is the grey line in each panel. Bootstrapped 90 percent confidence interval is the shaded area.

Figure 5: Predicted administration of oxygen using the Empirical approach by respiration rate and EMS contact sub-group.

6.1.2 Linear approach: regression with higher order terms

One approach to obtaining smoother profiles of estimates than those of the Empirical approach is to specify a linear model with higher order terms which result 'U' shaped profiles of estimates in each sub-group. Here fitting is carried out using a simple Ordinary Least Squares approach with the form [9]

$$\hat{O2} = \alpha + \beta_1 \text{EMS} + \beta_2 \text{RR1} + \beta_3 \text{RR1}^2 + \beta_4 \text{RR1} * \text{EMS}, \quad (2)$$

[9] Since this is a binary outcome a Probit model with an index function based on the quadratic specification of equation 2 is also appropriate. The results are virtually identical and omitted here for simplicity.

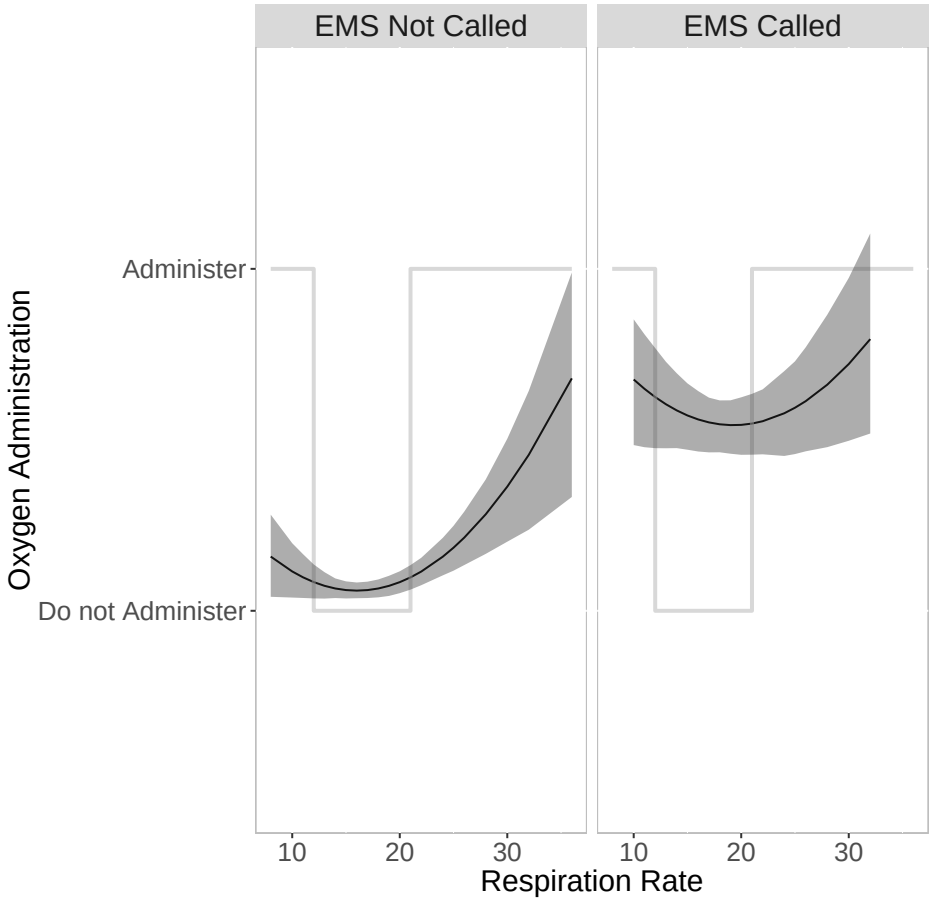
where α represents an intercept, EMS is an indicator of call seriousness equal to 1 if EMS services were contacted and 0 other wise, and RR1 and $RR1^2$ the recorded respiration rate and respiration rate squared and $RR1 * EMS$ the interaction of respiration rate and contact with EMS.[10] The regression estimates, presented in Table 11, suggest that all covariates except the interaction term are significant at the 2% level or less. A Ramsey RESET test for correct specification fails to reject the null of a correctly specified model (p-value 0.7). Figure 6 plots the estimates from fitting Equation 2. Two clear u-shapes are apparent in each of the levels of call seriousness. As well, there is a marked difference in the proportion of oxygen use across levels of call seriousness: more serious calls have a 47% higher chance of oxygen administration. Overall, the unadjusted R^2 value of this regression is 0.29, suggesting that variation in respiration rate and call seriousness contributes to just 29% of the variation in the decision to apply oxygen.

	Estimate	Std.Error	t.value	Pr(>|t|)
Intercept	0.4569	0.19	2.35	0.02
RR1	-0.0497	0.02	-2.57	0.01
RR1^2	0.0016	0.00	3.33	0.00
EMS=1	0.6628	0.15	4.41	0.00
RR1*EMS=1	-0.0101	0.01	-1.27	0.21

Table 11: Linear regression estimates with interaction and squared terms.

The drawback of this type of line fitting is that while there are two 'U' patterns they are not indicative of the discrete changes suggested by the guideline. Two discrete steps need to be identified in order to effectively compare these results to the guideline. Using the same method outlined in Section 5 leads to very precise estimates of candidate steps, identifying the boundary

[10]The same specification with the addition of a cubic term for *RR1* was also run, none of the coefficients associated with *RR1* were significant while the adjusted coefficients of determination were the same under both specifications (0.28 with added cubic and squared terms, and 0.28 with only the added squared term). The additional squared and cubic terms are highly collinear and the model is over-fitted.



Estimated outcome is solid black line and guideline is grey line in each panel. Bootstrapped 90 percent confidence interval is the shaded area. Estimates include interaction and squared terms.

Figure 6: Predicted administration of oxygen using the linear approach by respiration rate and EMS contact sub-group.

values as the candidate steps in 90% of the cases. But these boundary values are identified as candidates due to the specification of the model, and so may or may not be reflective of the raw observations. Under the linear model the largest gradients always occur at the upper and lower bounds regardless of the data at hand. In this case these bounds do not contain the guideline values of 12 for Step 1 or 20 for Step 2. However, for a guideline at the lower and upper bounds the guideline will always match the steps identified with the linear method purely by coincidence. As well, the size of the gradients at the

candidate steps indicate that the size of the change in the profile of estimates is not very large. Essentially the method precisely identifies weak candidates. See Table 12 for the candidate steps, confidence intervals and gradients.

EMS	Step		Lower ci	Upper ci	Guideline	Gradient
EMS Not Called	Step 1	8.0	8	8	12	-0.0249
EMS Not Called	Step 2	36.0	36	36	20	0.0621
EMS Called	Step 1	10.0	10	10	12	0.0333
EMS Called	Step 2	32.0	32	32	20	0.0271

Table 12: Candidate steps of the linear strategy and bootstrapped 90 percent confidence intervals.

6.1.3 Search approach: the best fit guideline

As the extreme alternative to the linear regression with higher order terms the contrasting strategy is to assume that the estimates in fact take on the same profile as the guideline, but with unknown locations of the steps. This form is pre-specified by identifying each combination of respiration rates in the set of all possible combinations of respiration rates, with each set representing a candidate guideline with two steps. For each sub-group, calculating the AUC for each candidate against the observations, the maximum AUC defines the best candidate, which is a combination of a Step 1, downwards, a Step 2, upwards. In this case there are 210 unique combinations with Step 1 occurring before Step 2, and 1 maximum AUC is identified. Figure 7 illustrates the result for each sub-group.

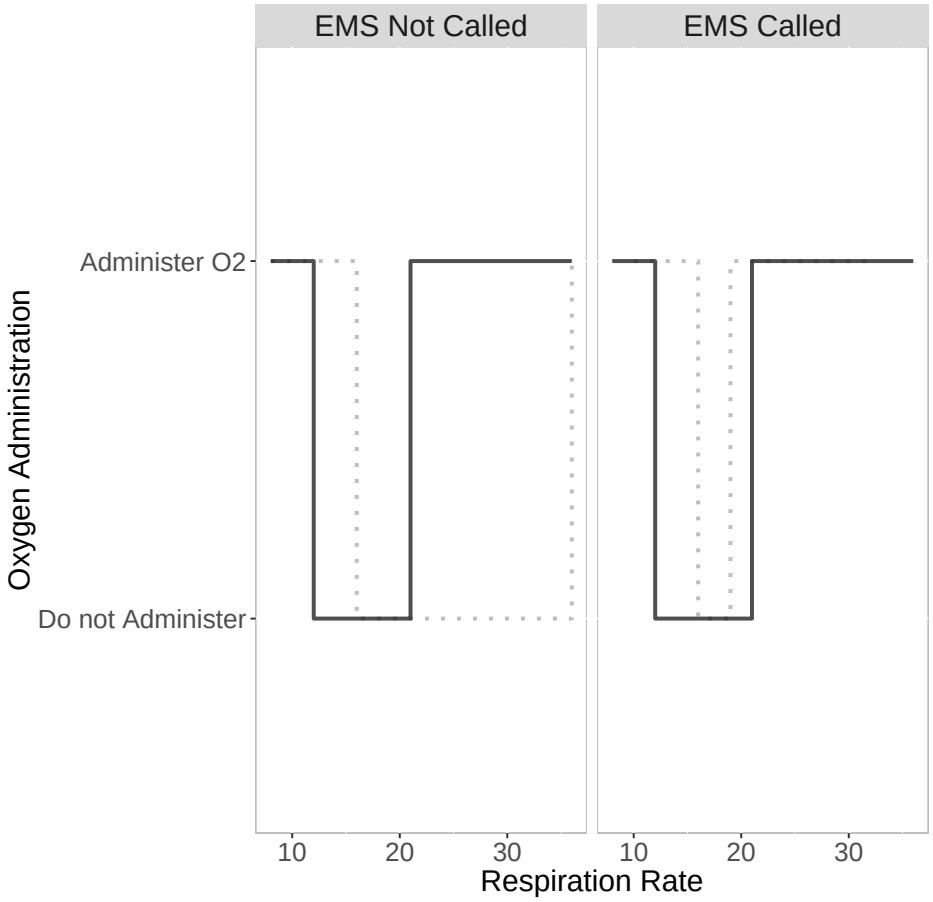
This method is a simple way to search for a candidate, but relies on the observations only in the sense of finding the best fit relative to other candidates. It is possible that the observations do not suggest such a pattern at all, in which case searching for the guideline by specifying the guideline is counter productive in determining what happened in the field.

Table 13 shows the results. Sometimes the bootstrap samples used to gen-

erate confidence intervals result in the identification of multiple maxima, in such cases the median value of the identified steps is reported (e.g. the median of the Step 1 values associated with the same maximal AUC value was taken as the result of a particular bootstrap sample.). In the same manner as all other bootstrap intervals the 5th and 95th percentiles of the all 1000 of the resulting bootstrapped steps form the lower and upper confidence interval values for each of the steps calculated from the original data set. When EMS was called the candidates are within the bounds of the confidence intervals. When EMS was not called the steps occur outside the bounds of the confidence regions, suggesting that this calculation method did not identify significant candidates. In both levels of contact with EMS the confidence intervals overlap, suggesting that the steps are so imprecise that they do not reflect two distinct steps. This strategy hinges on specifying the distinctive step pattern which the guideline predicts, but does not indicate the true profile of estimates, only the optimal placement of the candidate guideline relative to all possible placements.

EMS	Step		Lower ci	Upper ci	Guideline	Gradient
EMS Not Called	Step 1	15.0	16	18	12	-1.0000
EMS Not Called	Step 2	36.0	18	25	20	1.0000
EMS Called	Step 1	14.0	8	18	12	-1.0000
EMS Called	Step 2	19.0	16	30	20	1.0000

Table 13: Candidate steps of the search strategy and bootstrapped 90 percent confidence intervals.



Back solid lines represent the guideline. Grey dotted lines represent the optimal AUC outcome.

Figure 7: Optimal guideline using the search approach by respiration rate and EMS contact sub-group.

6.1.4 Comparison

In order to assess the performance of each approach, the goodness of fit of the profile of estimates with the observations is considered using the area under the receiver operator characteristics curve (AUC). The AUC is chosen because it is largely insensitive to imbalance in the proportion of positive to negative outcomes in the data [11] and easy to calculate.

In order to assess the match of the predictions, which range from [0,1], with the observations, which are binary (either a 0 or a 1), a threshold is used to

[11] See Figure 1 in Jeni et al. (2013).

classify the predictions into 0 and 1 categories. Once categorized, the True Positive Rate (TPR) and False Positive Rate (FPR) are calculated. The process is repeated for each possible threshold value and the results plotted, forming a Receiver Operator Characteristics Curve (ROC) which describes the ability of the predictions to accurately match the observations at hand. Strategies which do not match the observations result in a 45° line extending from the origin while approaches with a high degree of accuracy curve towards the upper left of the plot. The area under the ROC curve (AUC) takes on a value between 0.5 and 1, with higher values indicating better performance.

Approach	AUC
Empirical	84.51
Linear	79.60
Search: EMS Called	54.78
Search: EMS Not Called	55.72
Nonparametric	83.14

Table 14: AUC values of each approach

Table 14 presents the results of each approach, calculated with the 'pROC' package in R (Robin et al., 2011). The empirical and nonparametric perform similarly, followed by the linear and search approach.[12] In particular, according to Tape (2015), the search approach fails to match the observations. The Nonparametric approach exhibits the best performance of any of the smoothing approaches (i.e. excluding the Empirical approach) and is a good fit with the observations. Two insights are gained from assessing the AUC values in this way. The first is that even the best fit 'rectangular-u' fails to describe the observations. The second is that, among the smoothing approaches, the Nonparametric approach describes the observations the best, and required no input from the analyst about the 'shape' of the estimates, unlike the search and

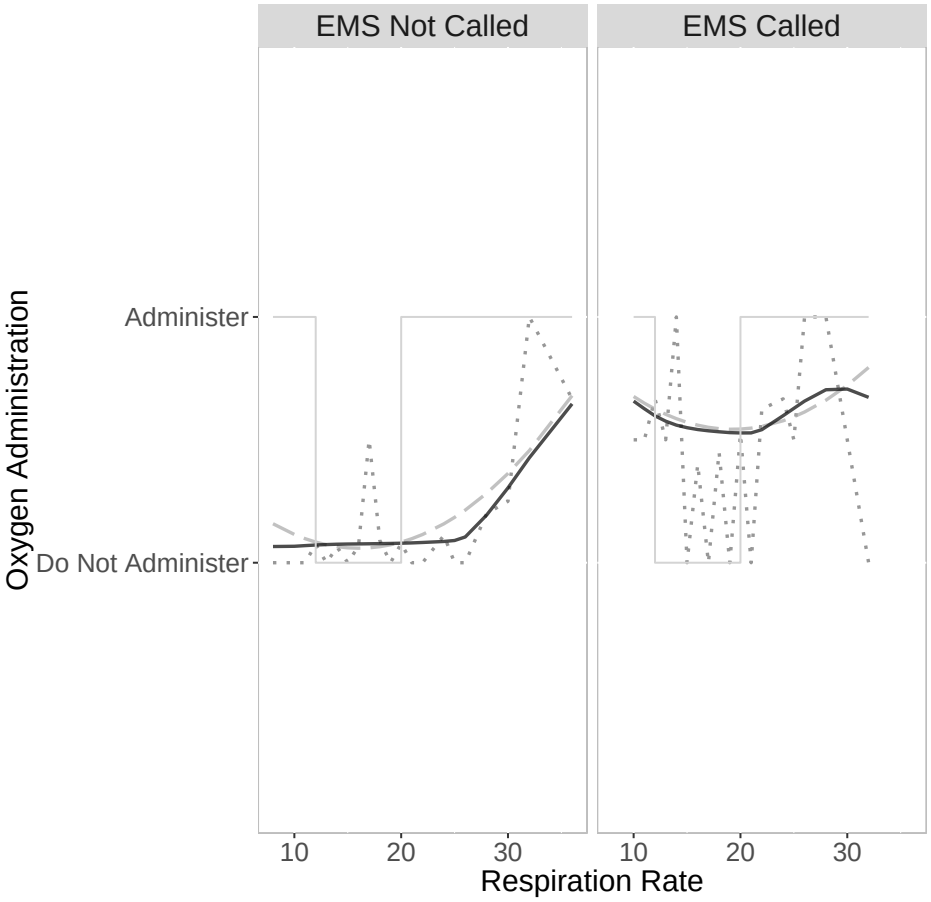
[12]The AUC is calculated over the entire range of estimates for the linear and nonparametric strategies. The AUC for each sub-group is presented for the search method due to the manner in which this strategy is formed: by choosing the guideline which maximizes the AUC in each sub-group.

linear methods. This makes nonparametric conditional density estimation an excellent approach for exploratory data analysis, and especially useful for assessing the behaviour of observations independently of presumptions about guidelines. The results of a bootstrapping test of the differences between the AUCs for the Empirical, Linear and Nonparametric strategies, presented in Table 15 confirms this result. The nonparametric strategy has a significantly greater AUC than the linear strategy but is not significantly different from the Empirical strategy at the 10% level of significance.

Approach	Empirical	Linear
Empirical		
Linear	0.00496	
Nonparametric	0.22954	0.05781

Table 15: P-values of bootstrap tests of differences in areas under ROC curves.

Layering the estimates of each approach, the result in Figure 8 is visually striking. While the linear approach fails to reject a null of misspecification, the better performing nonparametric estimates give no indication of a u-shape in the case of non-serious cases and only a slight u shape for serious cases. Finally, using the information in Tables 9, 10, 12 and 13, a visual summary of the identified candidate steps can be constructed. Figure 9 presents the two candidate steps for each strategy along with the confidence intervals associated with each step as a shaded rectangular area. The height of each confidence area is scaled to the height of the gradient associated with the candidate step. The figure thus summarizes both the precision and the degree of change associated with each candidate step. Confidence areas which are narrower indicate more precise estimates, while areas which are wider indicate less precision. Confidence areas which are taller indicate stronger gradients while shorter areas indicate smoother declines in the profile of estimates. In all cases distinct confidence areas between Step 1 and Step 2 which each con-



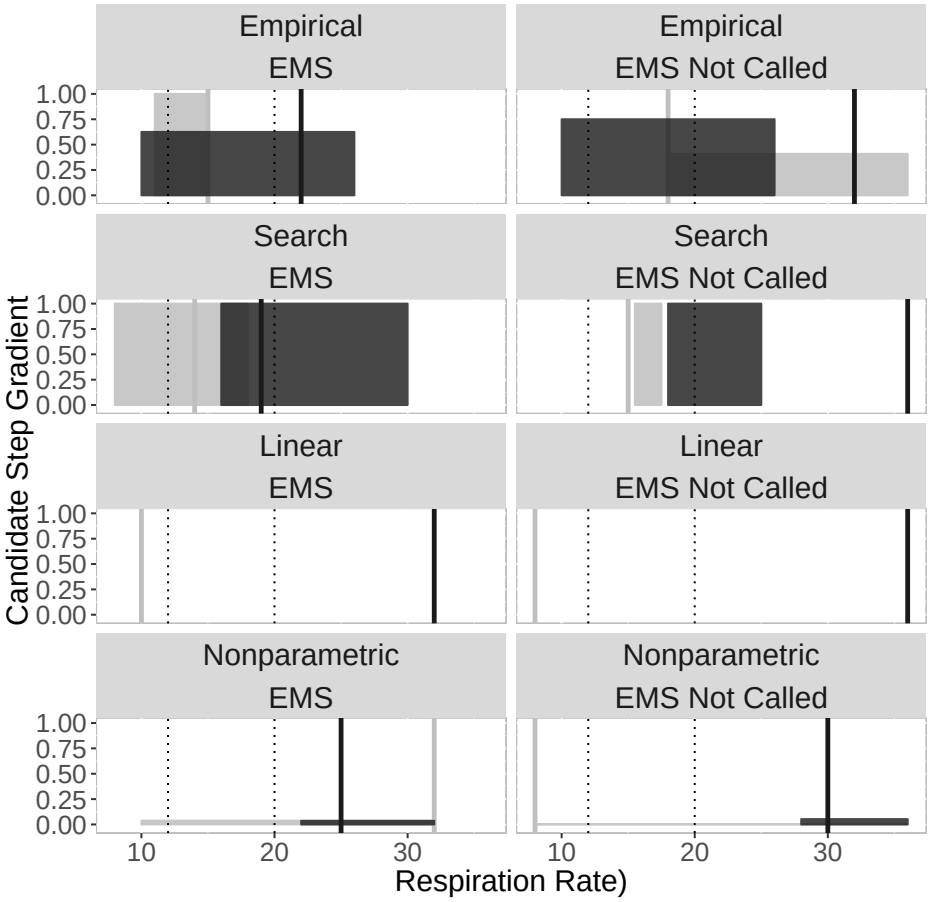
Guideline is the solid grey line in each panel. Empirical estimates are dotted grey lines, linear estimates are the dashed grey lines and nonparametric estimates are the solid black lines.

Figure 8: Predicted administration of oxygen for smooth approaches by respiration rate and EMS contact sub-group.

tain the guideline steps indicate support for the guideline.

Taking a closer look at the values of the gradients provides insight about the strength of the candidate steps. Gradients capture the change in the outcome at each evaluation point. Gradients for binary data can range from [-1,1] since the largest steps possible are from 0 to 1 ('Do Not Administer' to 'Administer') or 1 to 0 ('Administer' to 'Do Not Administer'). The ideal pattern for a match with the guideline would be defined by a gradient of -1 for Step 1 and 1 for Step 2 which is what is observed for the search approach due to the

complete specification of the estimates. The search approach demonstrated the worst fit to the observations, however, and so the finding of such strong gradients is spurious. The empirical steps are next largest, and are indicative of the jagged path of the estimates, these estimates were the most sensitive to noise. The linear gradients are very small, due to the specification of the form of the estimates and the least sensitive to noise. The linear approach indicated two well-defined 'U' patterns in the estimates, but the strength of these step candidates is weak. The Nonparametric approach suggests an inverted U with a weak upward step and a stronger downward step for the serious cases. For the non-serious cases the Nonparametric approach suggests a nearly non-existent first step and a mild second step. Since the Nonparametric approach exhibited the best match with the observations it provides the strongest evidence against the guideline pattern in the observations for both serious and non-serious cases. The classification matrix results alone gave the impression that there was a failure of adherence to the guideline in the data. Figures 8 and 9 give a visual representation of the difference in adherence across sub-groups.



Height of confidence region corresponds to absolute size of the gradient at the candidate. Guideline steps are shown as as dotted lines. Lighter grey shading corresponds to the confidence interval of the first candidate step, darker grey shading to the second step.

Figure 9: Candidate steps and confidence intervals by approach and EMS contact sub-group.

7 Conclusion and discussion

This paper presents a new standardized methodology for assessing adherence to a clinical practice guideline. The framework employs both a classification matrix and nonparametric conditional density estimation approach. Several metrics for summarizing the classification matrix are considered, with the area under the receiver operator characteristics curve (AUC) and Cohen's κ being favoured. Observations are smoothed using state-of-the-art nonpara-

metric conditional density estimation. Candidates for discrete changes in the profile of estimates are identified using the gradients of the estimates. Candidate discrete changes are then evaluated for distinctness from each other, and for agreement with the discrete changes suggested by the guideline using bootstrapped confidence intervals.

The data in this paper represent a previously un-examined sub-population of volunteer non-physician emergency health practitioners. Adherence to clinical practice guidelines has previously been suggested to differ substantially between physicians and non-physicians, with non-physicians often adhering more strictly to practice guidelines Higuchi et al. (2012). In a volunteer setting adherence to guidelines impacts the quality of service provided and may very well affect volunteer satisfaction and individual confidence in carrying out their duties. Higuchi et al. (2012) studied non-physician and non-volunteer practitioners (nurses) in Timor-Leste and found that guidelines were well adhered to, and had the qualitative effect of increasing confidence in the appropriateness of care provided. Both quality of care delivered and retention of volunteer resources are critical in the setting investigated in this work. In this paper volunteers tended towards over-administration of a medical therapy with a relatively low-risk of adverse health consequences. This may imply that over-administration provided responders with a sense of helping or that their focus was upon patient well-being rather than strict adherence to the guideline. The method for dealing with missing observations also results in estimates which tend towards over estimation of over-administration, since some patients who genuinely needed oxygen but did not have a respiration rate recorded had a normal range respiration rate applied to their case. Further work in this area might include improving data collection processes. As well, administration of qualitative surveys of volunteer satisfaction and confidence in their treatment with respect to the guide-

line as well as behavioural experiments studying the effects of different training techniques upon adherence may be helpful in better understanding the decisions of this particular subset of practitioners. As well, further studies of the potential for crowding out behavior should be undertaken prior to implementation of any program which would enforce guideline adherence more strictly. In this setting it may be that a policy of strict guideline enforcement could shift MFR attention away from accurate scene assessment and overall patient well being, and towards guidelines in a manner which would be unhelpful for both patients, and the organization. This is because in many emergency situations over-administration of oxygen is in fact desirable and the consequences of over-administration small. In non-emergency settings oxygen administration is also undesirable but the costs of greater guideline adherence were found to be negligible.

This work contributes to efforts to improve quality of medical care by providing a framework for evaluating the adherence to clinical practice guidelines. Insights about over- and under- administration of a medical therapy are obtained and the framework enables estimates of the financial impact of improved guideline adherence to be made. In most cases, no new data collection is required to implement the framework. The framework fits into the larger system of health care system assessment detailed by Tugwell et al. (1985), and would readily integrate into the system of appropriateness assessment suggested by Brook (2009). Such integration would provide a constantly updated measure of adherence to guidelines which would be comparable across regions and levels of the health care system.

References

Andritsos, D. A. and Tang, C. S. (2014). "Linking Process Quality and Resource Usage: An Empirical Analysis". In: *Production and Operations Management* 23.12, pp. 2163–2177.

Arditi, C., Rege-Walther, M., Wyatt, J., Durieux, P., and Burnand, B. (2012). "Computer-generated reminders delivered on paper to healthcare professionals; effects on professional practice and health care outcomes". In: *Cochrane Database Syst Rev* 12.12.

Askildsen, J. E., Holmås, T. H., and Kaarboe, O. (2011). "Monitoring prioritisation in the public health-care sector by use of medical guidelines. The case of Norway". In: *Health economics* 20.8, pp. 958–970.

Barbui, C., Girlanda, F., Ay, E., Cipriani, A., Becker, T., and Koesters, M. (2014). "Implementation of treatment guidelines for specialist mental health care." In: *The Cochrane database of systematic reviews* 1.

Brook, R. H. (2009). "Assessing the appropriateness of care - its time has come". In: *Journal of the American Medical Association* 302.9, pp. 997–998.

Canadian Institute for Health Information (2014). *Drug Use among Seniors on Public Drug Programs in Canada, 2012.* Ottawa,ON.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

Dimakou, S., Parkin, D., Devlin, N., and Appleby, J. (2009). "Identifying the impact of government targets on waiting times in the NHS". In: *Health care management science* 12.1, pp. 1–10.

Epanechnikov, V. A. (1969). "Non-Parametric Estimation of a Multivariate Probability Density". In: *Theory of Probability & Its Applications* 14.1, pp. 153–158. DOI: 10.1137/1114019. eprint: http://dx.doi.org/10.1137/1114019. URL: http://dx.doi.org/10.1137/1114019.

Fiander, M., McGowan, J., Grad, R., Pluye, P., Hannes, K., Labrecque, M., Roberts, N., Salzwedel, D. M., Welch, V., and Tugwell, P. (2015). "Interventions to increase the use of electronic health information by health care practitioners to improve clinical practice and patient outcomes". In: *Cochrane Database of Systematic Reviews* 3.

Flodgren, G., Pomey, M.-P., Taber, S. A., and Eccles, M. (2011). "Effectiveness of external inspection of compliance with standards in improving healthcare organisation behaviour, healthcare professional behaviour or patient outcomes". In: *Cochrane Database Syst Rev* 11.

Flodgren, G., Conterno, L., Mayhew, A., Omar, O., Pereira, C. R., and Shepperd, S. (2013). "Interventions to improve professional adherence to guidelines for prevention of device-related infections". In: *Cochrane Database Syst Rev* 3.

Frey, B. S. and Jegen, R. (2001). "Motivation crowding theory". In: *Journal of economic surveys* 15.5, pp. 589–611.

Hall, P., Racine, J. S., and Li, Q. (2004). "Cross-validation and the estimation of conditional probability densities". In: *Journal of the American Statistical Association* 99, pp. 1015–1026.

Hayfield, T. and Racine, J. S. (2008). "Nonparametric Econometrics: The np Package". In: *Journal of Statistical Software* 27.5.

Health Council of Canada (2009). *Value for Money: Making Canadian Health Care Stronger*. Tech. rep. Toronto: Health Council, p. 52.

Higuchi, M., Okumura, J., Aoyama, A., Suryawati, S., and Porter, J. (2012). "Application of Standard Treatment Guidelines in Rural Community Health Centres, Timor-Leste". In: *Health policy and planning* 27.5, pp. 396–404.

Institute of Medicine (2006). *Medicare's Quality Improvement Organization Program: Maximizing Potential (Series: Pathways to Quality Health Care)*. Washington, DC: The National Academies Press. URL: http://www.nap.edu/

catalog/11604/medicares-quality-improvement-organization-program-maximizing-potential-series-pathways-to.

Jeni, L., Cohn, J. F., and De La Torre, F. (2013). "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics". In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, pp. 245–251.

Li, Q. and Racine, J. S. (2003). "Nonparametric estimation of distributions with categorical and continuous data". In: *Journal of Multivariate Analysis* 86.2, pp. 266 –292.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton University Press.

MacKinnon, J. G. (2007). *Bootstrap Hypothesis Testing*. Working Papers 1127. Queen's University, Department of Economics. URL: https://ideas.repec.org/p/qed/wpaper/1127.html.

McGlynn, E., Asch, S., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., and Kerr, E. (2003). "The Quality of Health Care Delivered to Adults in the United States". In: *New England Journal of Medicine* 348.26, pp. 2635–2645. URL: http://dx.doi.org/10.1056/NEJMsa022615.

Nelson, B. (2015). "Waste: Unnecessary Overuse of Medical Care Causes Both Waste and Harm". In: *The Hospitalist* 19.6:1, pp. 23–27.

O'Brien, M., Rogers, S., Jamtvedt, G., Oxman, A., Odgaard-Jensen, J., Kristoffersen, D. T., Forsetlund, L., Bainbridge, D., Freemantle, N., Davis, D., et al. (2007). "Educational outreach visits: effects on professional practice and health care outcomes". In: *Cochrane database syst rev* 4.4.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Bioinformatics* 12, p. 77.

Stone, C. J. (1977). "Consistent Nonparametric Regression". In: *The Annals of Statistics* 5.4, pp. 595–620.

Straube, S. and Krell, M. M. (2014). "How to evaluate an agent's behavior to infrequent events? Reliable performance estimation insensitive to class distribution". In: *Frontiers in computational neuroscience* 8.43.

Tape, T. G. (2015). *The Area Under an ROC Curve*. URL: `http://gim.unmc.edu/dxtests/ROC3.htm` (visited on 11/23/2015).

The Lown Institute (2016). *About Us*. URL: `http://lowninstitute.org/home/vision-mission-history/` (visited on 01/06/2016).

Thomas, L., Cullum, N., McColl, E., Rousseau, N., Soutter, J., and Steen, N. (1999). "Guidelines in professions allied to medicine". In: *The Cochrane Database of Systematic Reviews* 1.

Thomas, S. (2016). *Playing by the rules? Agreement between predicted and observed binary choices.* Department of Economics Working Paper 2016-12. McMaster University. URL: `https://www.economics.mcmaster.ca/research/department-working-papers`.

Tugwell, P., Bennett, K., Sackett, D., and Haynes, R. (1985). "The measurement iterative loop: a framework for the critical appraisal of need, benefits and costs of health interventions". In: *Journal of chronic diseases* 38.4, pp. 339–351.

Appendices

A Observations without imputed missing values

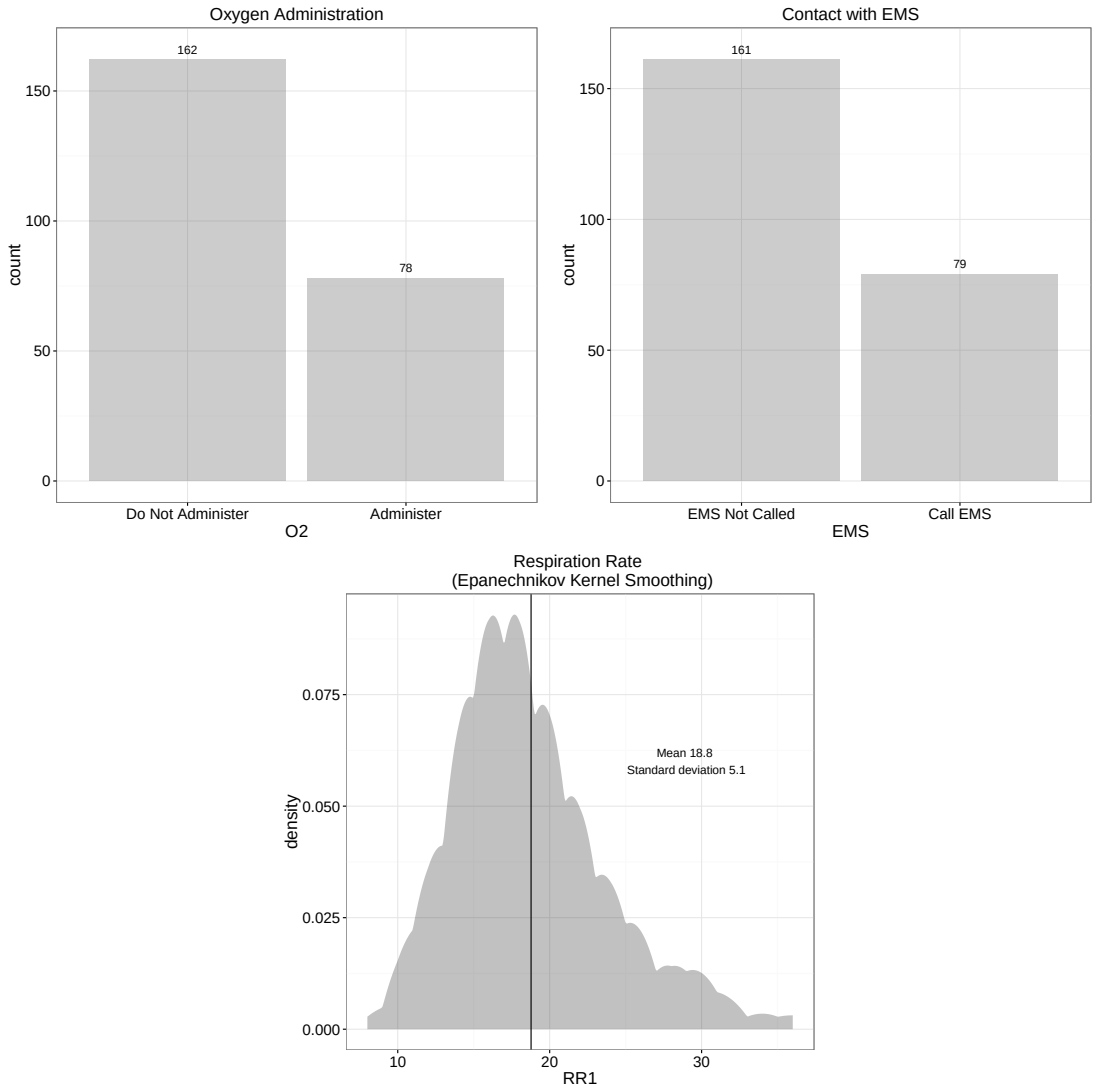


Figure A.1: Illustrated data summary for data without imputed missing values.

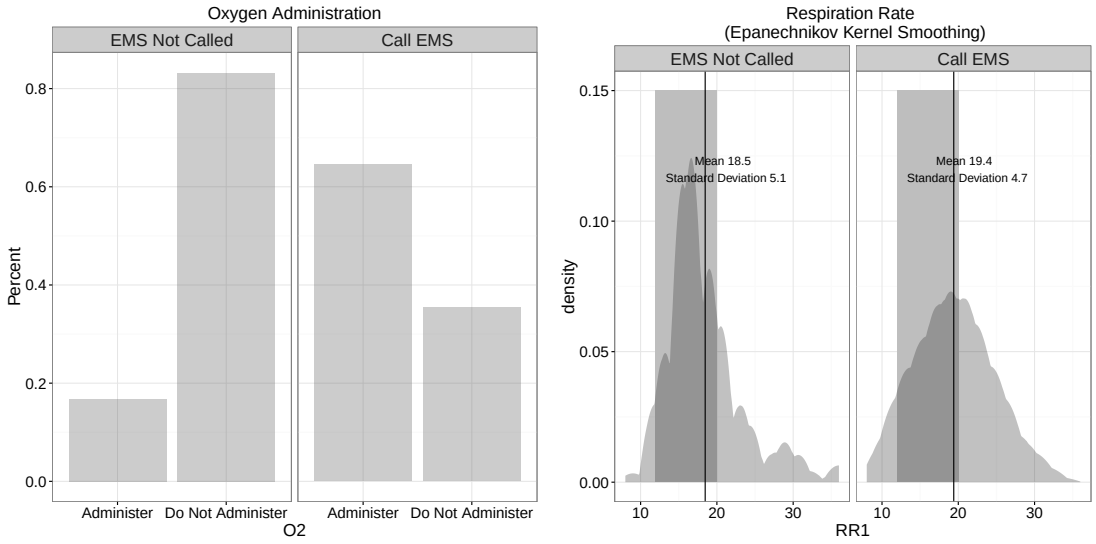


Figure A.2: Illustrated data summary by EMS contact sub-group for data without imputed missing values.

		Observed Oxygen Use		
		Not Administered	Administered	Total
Guideline	Not Administered	105	20	125
	Administered	29	7	36
	Total	134	27	161

Cells indicate the number of recorded cases which agree (disagree) with the respiration rate guideline for administration of supplemental oxygen.

Table A.1: Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is not called for data without imputed missing values.

		Observed Oxygen Use		
		Not Administered	Administered	Total
Guideline	Not Administered	17	30	47
	Administered	11	21	32
	Total	28	51	79

Cells indicate the number of recorded cases which agree (disagree) with the respiration rate guideline for administration of supplemental oxygen.

Table A.2: Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is called for data without imputed missing values.

Measure	EMS Not Called	EMS Called
TPR	25.93%	41.18%
TNR	78.36%	60.71%
FPR	21.64%	39.29%
CCR	69.57%	48.1%
bCCR	52.14%	50.95%
AUC	0.52	0.51
Cohen's κ	3.78%	1.64%

Table A.3: Classification matrix summary measures by EMS contact sub-group for data without imputed missing values.

Measure	EMS Not Called	EMS Called
All Cases		
Treated Patients	5.4	10.2
Total Costs	$ 96.92	$ 183.08
Cost/patient	$ 17.95	$ 17.95
Guideline Non-Adherent Cases		
Over-Administration		
Treated Patients	4	6
Total Costs	$ 71.8	$ 107.7
Under-Administration		
Untreated Patients	5.8	2.2
Total Costs	$ 104.11	$ 39.49
Perfect Guideline Adherence		
Treated Patients	7.2	6.4
Total Cost	$ 129.23	$ 114.87
% Savings under Perfect Adherence	-33%	37%

Table A.4: Annual cost of oxygen delivery by EMS contact sub-group for data without imputed missing values.

B Nonparametric Estimation

The problem at hand is to estimate the conditional density function $g(y|x) = \frac{f(x,y)}{\mu(x)}$:

$$\hat{g}(y|x) = \frac{\hat{f}(x,y)}{\hat{\mu}(x)}, \tag{B.1}$$

The approach used in this paper is that described by Li and Racine (2007) where the numerator and denominator of the conditional probability function are described by:

$$\hat{f}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_\gamma(x,X_i)k_{\lambda_0}(y,Y_i) \tag{B.2}$$

$$\hat{\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} K_\gamma(x,X_i), \tag{B.3}$$

with $K_\gamma(x,X_i)$ and $k_{\lambda_0}(y,Y_i)$ representing kernel density functions.
For the purposes of this study the kernel suggested by Li and Racine (2003) will be used for the estimation of the unordered discrete variable *O2*:

$$\begin{aligned} k_{\lambda_0}(y,Y_i) &= l(Y_{is},y_s,\lambda_s) \\ &= \begin{cases} 1-\lambda_s & \text{if } Y_{is}=y_s \\ \frac{\lambda_s}{c_s-1} & \text{if } Y_{is}\neq y_s \end{cases}, \end{aligned} \tag{B.4}$$

where y_s can take on c_s ordered values $0,1,c_s-1$. If $\lambda_s = 0$ then $l(Y_{is},y_s,\lambda_s) = 1$ is an indicator function, and if $\lambda_s = \frac{c_s-1}{c_s}$, then $l(Y_{is},y_s,\frac{c_s-1}{c_s}) = \frac{1}{c_s}$, a constant. Thus the range for the smoothing parameter associated with *participate* is $[0,\frac{2-1}{2}=0.5]$.
$K_\gamma(x,X_i)$ is a product kernel, in this case composed of the kernel proposed by Li and Racine (2003) for the unordered levels of *EMS* and the kernel proposed by Epanechnikov (1969) for the continuous variable *RR1*.

$$K_\gamma(x,X_i) = W_h(x^c,X_i^c)L(x^d,X_i^d,\lambda), \tag{B.5}$$

where $\gamma = (h,\lambda)$ is a vector of continuous and discrete bandwidths in this case for *RR1* and *EMS*. The superscript *c* denotes the continuous variable *RR1* and *d* the discrete variable *EMS*. The Epanechnikov kernel used here is fur-

ther defined by:

$$W(u) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}u^2) & \text{if } u^2 < 5 \\ 0 & \text{otherwise} \end{cases} \quad \text{where } u = \frac{x^c - X_i^c}{h} \text{ and } h > 0, \tag{B.6}$$

and the Li and Racine kernel by:

$$L(x_i^d, x^d, \lambda) = \begin{cases} 1 \text{ if } |x_i^d - x| & = 0, \\ \lambda^{x_i^d - x} \text{ if } |x_i^d - x| & \geq 1 \end{cases}, \tag{B.7}$$

where λ must lie between 0 and 1.

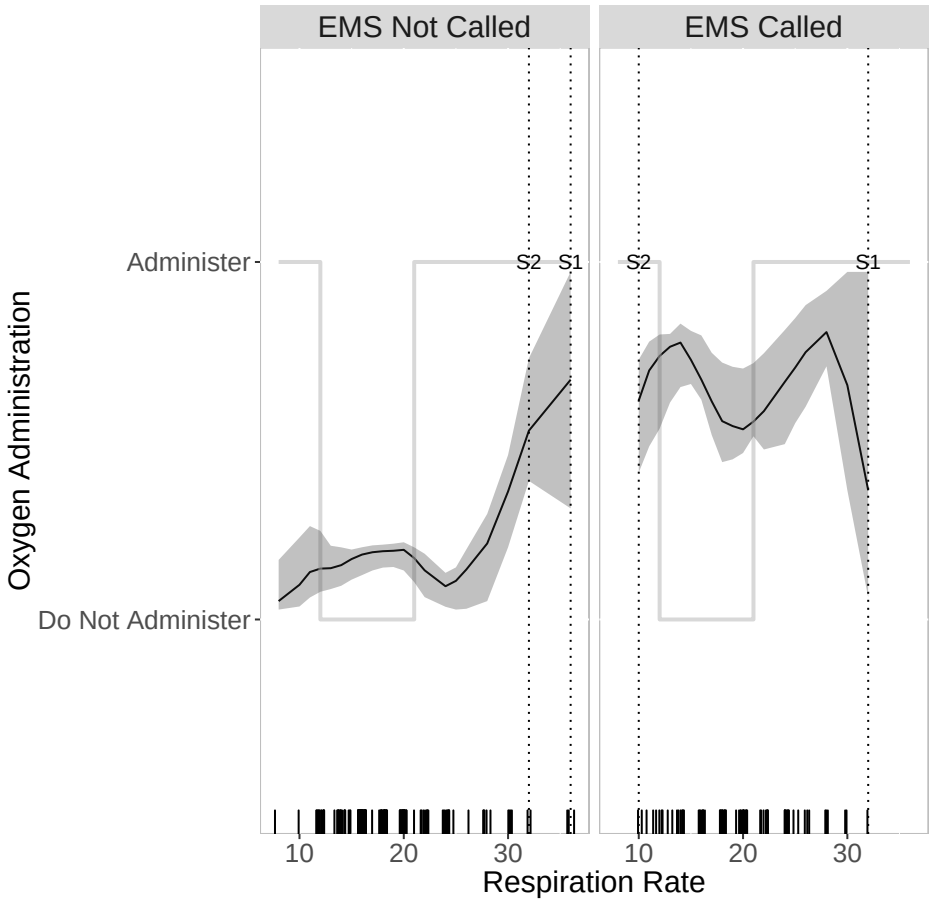
In order to estimate the nonparametric model, optimal bandwidths were determined using a least squares cross validation routine. This approach has the advantage that if a regressor is irrelevant [13] it will be smoothed entirely out of the regression asymptotically. Smoothing out in this case is demonstrated by a large bandwidth. Table B.1 provides the results. In both data sets all variables are relevant as these are well below their upper bounds.

Variable	Original	Imputed	Maximum.Range
O2	0.0285	0.0042	1
RR1	1.8510	3.3319	28
EMS	0.0000	0.0000	1

Table B.1: Bandwidths generated using least squares cross validation for data with and without imputed missing values.

C Nonparametric results with missing values

[13] Meaning that the regressor does not substantially affect the outcome.



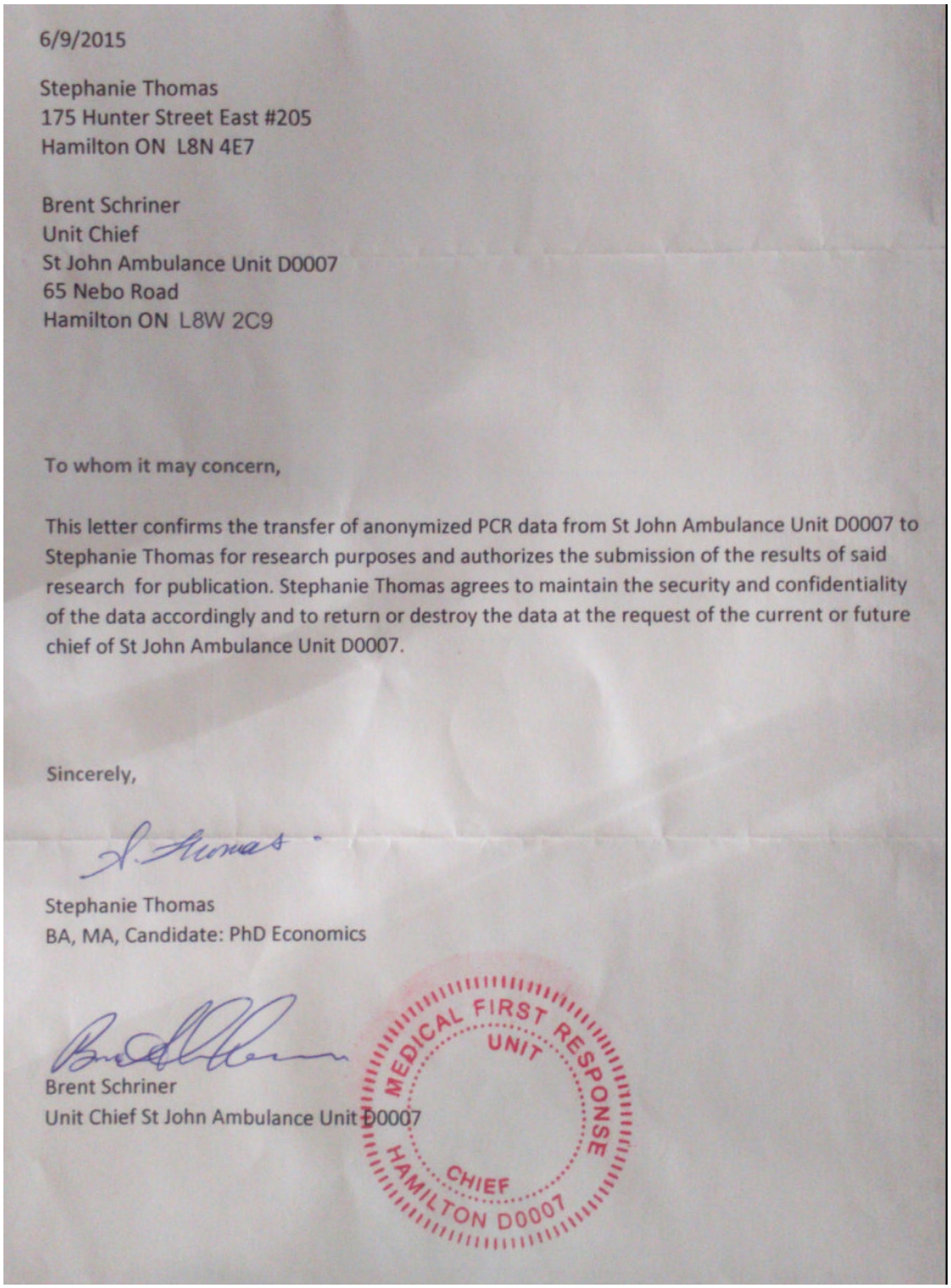
Guideline is the grey line. Candidate step 1 (S1) and step 2 (S2) are dotted lines. Bootstrapped 90 percent confidence interval is the shaded area. Bars along the x-axis indicate frequency of observations.

Figure C.1: Estimated probability of oxygen administration by respiration rate and EMS contact sub-group using the Nonparametric approach for data without imputed missing values.

EMS	Step		Lower ci	Upper ci	Guideline	Gradient
EMS Not Called	Step 1	36	16	36	12	-0.0900
EMS Not Called	Step 2	32	9	32	20	0.2066
EMS Called	Step 1	32	16	32	12	-0.3372
EMS Called	Step 2	10	10	28	20	0.1418

Table C.1: Candidate steps of the nonparametric strategy for data without imputed missing values and bootstrapped 90 percent confidence intervals.

D Data sharing agreement



6/9/2015

Stephanie Thomas
175 Hunter Street East #205
Hamilton ON L8N 4E7

Brent Schriner
Unit Chief
St John Ambulance Unit D0007
65 Nebo Road
Hamilton ON L8W 2C9

To whom it may concern,

This letter confirms the transfer of anonymized PCR data from St John Ambulance Unit D0007 to Stephanie Thomas for research purposes and authorizes the submission of the results of said research for publication. Stephanie Thomas agrees to maintain the security and confidentiality of the data accordingly and to return or destroy the data at the request of the current or future chief of St John Ambulance Unit D0007.

Sincerely,

Stephanie Thomas
BA, MA, Candidate: PhD Economics

Brent Schriner
Unit Chief St John Ambulance Unit D0007

MEDICAL FIRST RESPONSE UNIT
CHIEF
HAMILTON D0007