Chapter 3. Results

Over 500 articles presented data addressing some aspect of the study questions. To address the first question related to autopsy-detected errors in clinical diagnosis, we reviewed 225 English-language studies. Abstracts for 34 additional articles in languages other than English were reviewed, but none added to or substantially changed the information already available in the English-language literature. For other study questions, review of the non-English abstracts suggested a possible benefit from translating one article. A full translation was obtained for this Danish article presenting data on the impact of legislation on autopsy rates.[71]

Four basic methodological issues significantly impact the ability to address this question about autopsy-detected error rates in clinical diagnosis, but have received little to no attention in the autopsy literature. These issues and the literature that addresses them are summarized below.

The Autopsy as a Diagnostic Test

Quality of the Autopsy Procedure

Audits of the quality of autopsy reports by panels of reviewers have shown substantial deficiencies in terms of the report content and interpretation of autopsy findings. For instance, an audit of 104 perinatal autopsies in northern England reported low compliance with minimum autopsy reporting standards proposed by the Royal College of Pathologists. [72] These standards consisted of a 100 point scoring system (to achieve a maximum score of 600 points) for six factors identified in the report: body measurements (crown to rump or heel, head circumference, foot length, and body weight); organ weights (lung, heart, liver, brain); quoting of normal values; histological report on main organs; radiology report; microbiology report; and other relevant findings (e.g., cytogenetics). The minimum acceptable score based on the Royal College of Pathologists guidelines was set at 250 total points. Only 51% of cases met or exceeded the minimum score and adequate interpretative comments were recorded in only 49%. Also importantly, the reported autopsy rate of 45% for this study was well below the College's recommended rate of 75% and most of the autopsy cases were classified as Coroner's cases.[72]

An audit of 314 perinatal autopsies in Wales similarly reported achievement of minimum reporting standards in only 54% of cases and showed a clear difference in quality scores for autopsies performed at a regional pediatric referral center, where only 8% failed to meet the minimum score compared to smaller community hospitals in which 72% failed to meet the minimum score.[73, 74] In this case, the authors conducted a follow-up study and noted substantial improvements in autopsy quality as a result of general educational efforts but also in the context of a shift in autopsies to a regional center.[75]

The quality of adult autopsies has received less attention. Only one study explicitly addressed autopsy quality for adult cases, and this study focused on the completeness of autopsy requests and the timeliness of autopsy reports.[76] Two other studies provide some information relevant to autopsy quality in adults.[77, 78] In one series of 1000 adult autopsies, the authors listed 73 cases (7.3%, 95% CI: 5.8-9.1%) as too poorly performed or reported to include in their analysis of diagnostic errors.[77] In another study of 111 autopsies, the authors stated that in 3 cases (2.7%, 95% CI: 0.7-8.3%) deficiencies in autopsy performance resulted in failure of the autopsy to answer important clinical questions.[78]

While appropriate documentation is certainly a component of healthcare quality and performance, caution must be used in interpreting overall quality from the quality of record keeping. Similar problems exist for documenting or reporting different components within the patient medical record including: clinical discharge summaries, history and physical examinations, operative reports, and pharmacy and clinic notes.[79-89] More importantly, the level of completeness of such reports or summaries may effect medical decision making and patient care. Nonetheless, retrospective evaluation of this type of data as a surrogate for measuring quality often occurs, and the potential limitations must be acknowledged when drawing conclusions about the quality of care, whether for clinical activities or autopsy performance.

Accuracy of the Autopsy

Like most complex procedures involving multiple observational and cognitive elements, the autopsy almost certainly has an error rate of its own,[68] despite its role as a diagnostic standard. All diagnostic tests attempt to capture the "true diagnosis," and test characteristics such as accuracy are defined in terms of this true diagnosis. For a "gold standard," assessing the extent to which the test falls short of capturing the true diagnosis, presents practical difficulties because confirmatory tests generally rely on the gold standard for their own accuracy.

Cases in which the autopsy fails to provide a definite diagnosis convey some sense of the limitations or potential for inaccuracy on the part of the autopsy. A number of studies mention in passing the frequency with which adequately performed autopsy fails to establish a definitive diagnosis related to the cause of death (immediate, intervening or underlying). As shown in Appendix Table 1, technically adequate autopsies generally establish causes of death in all but 1-5% of cases, although some studies have reported slightly higher proportions. One of the outlier results indicating a substantially higher rate of persistent uncertainty after autopsy derives from data gathered in 1958.[90] Three other reports with rates of persistent uncertainty after autopsy that exceed 20% focused on perinatal deaths, [91-93] but one study of adult deaths after cardiac surgery also reported persistent uncertainty after autopsy in 25% of cases.[94]

It is important to realize that even fairly small error rates in any presumed diagnostic standard can significantly distort estimates of the sensitivity of clinical performance.[95] Table 1 (adapted from Saracci[68]) illustrates this effect using a hypothetical series of 2000 autopsies in which the investigators report the agreement between clinical and autopsy diagnosis of lung cancer related to the cause of death.

Acknowledging that no test provides the "true diagnosis", Table 1 compares the sensitivity of clinical diagnosis measured against the theoretic "true diagnosis" with that obtained when measured against the observed autopsy diagnosis. The values in parentheses correspond to the observed autopsy findings (rather than the "true diagnosis"), with the specific values in parentheses generated by assuming that 36 (2%) of the 1800 non-lung cancer cases were misclassified by autopsy as deaths due to lung cancer.

Using the true diagnosis as the diagnostic standard, the sensitivity of clinical diagnosis for detecting lung cancer as the cause of death is 180/(180+20) = 0.90. When the observed autopsy results are taken as the diagnostic standard, the same sensitivity is 182/(182+54) = 0.77. Thus, even a relatively small autopsy misclassification rate of 2% would reduce the observed sensitivity of clinical diagnosis for lung cancer from its "true" value of 90% to an apparent value of 77%.

This illustration assumes that all of the misclassified cases have the same distribution as those cases already in the left-hand column. In fact, it could be that none of these cases would be diagnosed clinically as lung cancer, so that the observed sensitivity would correspond to 180/236=76%. At the other extreme, all of these misclassified cases might be diagnosed clinically as lung cancer, resulting in an observed sensitivity of 216/236=91%. Given this range of values, the conclusion remains that small rates of misclassification at autopsy can produce significant changes in the apparent sensitivity of clinical diagnosis.

Precision of the Autopsy

In the absence of data on the accuracy of the autopsy, one can still ask about its reproducibility – i.e., the extent to which pathologists agree in their identification of the main autopsy diagnoses. Multiple studies within the pathology literature address the reproducibility for diagnoses in surgical pathology and cytology. These studies suggest a wide range of values for measures of agreement from excellent to poor, but the majority falls in the range conventionally regarded as indicating moderate to substantial agreement. [96-107] The range of values varies depending on the specific test evaluated (e.g., cervical cytology, [96, 97] prostate [98, 99] or liver biopsies[101]) and the evaluation setting (e.g., general pathologists versus specialists[98, 99]).

The above studies do not assess the extent to which pathologists confronted with the results of the diverse investigations that make up a given autopsy independently arrive at the same conclusions. Despite an extensive search, we found only three studies that address the issue of reproducibility among pathologists in interpreting autopsy findings.[108-110] The oldest of these studies[110] compared the judgments of three reviewers in assigning a principal cause of death based on the findings presented in 50 autopsy reports. Because the three reviewers were third-year medical students (trained to abstract autopsy reports and code death certificates for the purposes of this study), we did not regard the briefly summarized data on inter-rater agreement, presented in passing in this study of the accuracy of death certificates, as indicative of agreement among pathologists in ordinary clinical practice.

A second study[109] involved the analysis of a series of 39 patients who died in an intensive care unit and underwent autopsy specifically to address pulmonary pathology. The study excluded patients infected with HIV, bilateral chest tubes or bilateral pleural infection. Four pathologists (who were blinded to clinical and microbiological data) independently examined postmortem lung biopsies from these patients with the purpose of judging the presence or absence of pneumonia. The authors reported these independent judgments as well as the diagnoses derived from the application of pre-selected criteria for the histologic diagnosis of pneumonia.[111]

Using their own judgments, the four pathologists exhibited unanimity in the diagnosis of 30 (77%) patients—7 with pneumonia and 23 without. The authors reported a kappa score for the overall group, but also provided enough data to allow calculation of the more meaningful kappa scores for the different possible pairs of observers (Table 2).

As seen in Table 2, the different possible pairs of pathologists exhibited “substantial” (kappa=0.6-0.8) to “near perfect”(kappa=0.8-1.) agreement, with only 1 pair falling in the “moderate” range (kappa=0.4-0.6). Six months later, one pathologist re-reviewed the same biopsies and reclassified only 2 patients (one from “pneumonia” to “no pneumonia” and the other vice versa), corresponding to a kappa score of 0.82. However, the overall prevalence of pneumonia as determined by each of the four pathologists varied from 18% to 38%. Moreover,

this study assessed agreement with respect to a specific question with a dichotomous answer (pneumonia present or absent). For the performance of an autopsy, pathologists must choose from among multiple possible diagnoses, rather than determining the presence or absence of a single condition.

The third study[108] is the only one to address agreement among pathologists with respect to the autopsy as a whole – i.e. in determining principal underlying diseases and causes of death. This study involved four pathologists examining 35 autopsies. The authors reported excellent (near perfect) agreement for all pathologist pairs independently determining the principal disease (i.e., underlying cause of death), with kappa values between 0.83 and 0.97. Specifically, there were four cases that generated disagreement in the principal underlying disease. In three of these cases, only one pathologist disagreed with the other three, while in the fourth case two pathologists shared one diagnosis, while two shared a different diagnosis.

For assignments of the immediate cause of death, the authors reported kappa values ranging from 0.43-0.75. Importantly, this level of agreement discounts the roughly 30% of cases for which two of the pathologists could not decide on an immediate cause of death, since in these cases agreement was not measured. Listings of minor diseases showed much less agreement.

Overall, it is likely that pathologists exhibit substantial agreement in generating the principal autopsy diagnoses, although the literature contains little data addressing this issue.[1]

Reproducibility of Judgments of Autopsy-detected Errors in Clinical Diagnosis

Numerous studies report group consensus for important discrepancies between clinical and autopsy diagnoses related to the cause of death.[2, 51, 54, 55, 116-133] Unfortunately, only five studies[55, 117, 124, 130, 131] mention the issue of reproducibility for these judgments of diagnostic discrepancies. In one of these studies,[55] three investigators independently classified each case, with a fourth reviewer for cases with discordant classifications. Reclassification occurred for 20% of Class I and II errors, but the authors provided no further details on the consensus process nor did they calculate a kappa score or other measure of inter-rater agreement. In another study,[124] the authors report that in 25 of 338 autopsies (7.4%, 95% CI: 4.9-10.9%) two reviewers could not reach consensus, so classification of clinical-autopsy discordances required a panel of peer reviewers. The other three studies[117, 130, 131] provided too little detail regarding the assessment of agreement between reviewers to permit meaningful assessments of the reproducibility of judgments pertaining to diagnostic errors detected at autopsy.

Because these five studies provide so few data on the frequency of differences in opinion between reviewers assessing clinical-autopsy discrepancies, we reviewed studies from the medical error literature that provide data on comparable assessments. Several well-known studies have examined the reproducibility of peer review classifications of hospital deaths as "preventable" or "not preventable." [134-137] These studies are relevant to the present question because the definition of preventability lies at the heart of the two most commonly used classifications of discrepancies between clinical and autopsy diagnoses.[51, 70]

Assessments of reviewer agreement from the medical error literature indicate that achieving anything beyond fair or modest inter-rater reliability for judgments about error or preventability

[1] Of note, classification of forensic deaths almost certainly has greater disagreement. In a study of 198 physician medical examiners, participants exhibited substantial variability in classifying the manner of death (homicide, suicide, accident, natural, undetermined) for the majority of the 23 scenarios presented.[114,115]

for major adverse outcomes requires reviewer training and, in some cases, dropping reviewers that consistently yield divergent judgments.[136] In the most recent of these studies (and the one which most specifically focused on the issue of reproducibility), the authors noted that "if one reviewer rated a death as definitely or probably preventable, the probability that the next reviewer would rate that case as definitely not preventable (18%) was actually slightly higher than the probability that the second reviewer would agree with the first (16%)." [137]

Summary of the autopsy as a diagnostic test

- The quality of the autopsy has received little systematic study except in the case of perinatal autopsies, where deficiencies in the quality of reporting appear to be common.
- In at least 1-5% of cases, diagnostic uncertainty persists despite technically adequate autopsy. Classification errors affecting autopsy diagnoses at even this relatively small rate can substantially distort estimates of the performance of clinical diagnosis.
- Only one small study has specifically assessed the reproducibility of autopsy diagnoses among pathologists. This study showed excellent inter-rater agreement for the principal diagnosis, but data on agreement for other major diagnoses were not presented.
- The reproducibility of judgments about errors in clinical diagnosis as indicated by autopsy findings has received almost no attention. Extrapolation from studies of inter-rater reliability for peer review assessments of case notes in the healthcare quality and medical error literature suggests that reproducibility of such judgments is likely to be fair to moderate at best.

Autopsy Detection of Clinically Unsuspected Diagnoses

Quantitative Analysis of Diagnostic Error Rates

Notwithstanding the existing deficiencies in standardizing a systematic means of using the autopsy as a measure of clinical diagnostic performance, the extensive literature on clinically missed diagnoses detected at autopsy nevertheless bears attention from clinicians and others interested in measures of clinical quality.

We identified 50 studies reporting diagnostic error rates that met our inclusion criteria. These 50 studies reported a total of 61 distinct autopsy series, as many of them reported more than one observation period or presented data from more than one institution. (Thus, if a study reported autopsy data for two hospitals, we presented these separately whenever the original paper presented sufficient detail to allow this.) Appendix Tables 2-4 present the salient features of these autopsy series using the three classification systems adopted for this study. Studies in each table are listed chronologically (with respect to the period of study, not date of publication)

within patient population categories, with studies of more general patient populations listed first, followed by studies of more specific patient populations (adult medical, surgical, pediatric, etc.). Appendix Figures 3 & 4 displays the error rates and 95% confidence intervals for the studies listed in Appendix Tables 2 & 3. (A comparable figure was not created for the third definition of error, major ICD disease discrepancies, because of the small number of studies involved.)

One would expect diagnostic error rates to decrease over time due to medical advances, and thus would expect an inverse correlation between error rates and study period. Appendix Figures 5 & 6 present time trends for Class I and major error rates, respectively. Inspection of these figures indicates no clear changes in error rates over time, but it is possible that trends are obscured by differences in case mix, country effects and changes in autopsy rates. The latter possibility is of considerable interest, as clinicians generally believe that autopsy cases represent precisely those cases most likely to have unexpected findings detected at autopsy. Thus, true improvements in clinical diagnosis might be offset by increased selection of autopsy cases by clinicians. We addressed this possibility as well as the effects of case mix and study country by performing logistic regression using a model in which study period, autopsy rate, case mix (using 7 different categories) and country (U.S. vs. non-U.S.) were included as predictors of error rates.

The results of the quantitative analysis are presented in detail in Appendix B and summarized below. Briefly, we performed logistic regression for each of the three definitions of error (Class I, major and discrepancies in major ICD disease categories). In each case, the analysis began with a model that included all predictors—study period, autopsy rate, case mix (general adult inpatients, medical patients, surgical parents, pediatric, etc.) and country (U.S., non-U.S.). Hospital teaching status was not included as a predictor in any of the models, because too few studies involved non-teaching hospitals and because the nature of the teaching status (academic medical center, community teaching vs. tertiary referral center, presence of residents across a broad range of specialties) was usually unclear.

The results generated by the complete model were compared with models in which one variable at a time was dropped. For Class I errors, the differences between the results for the different models did not achieve statistical significance, so the mean error rate could be calculated from the model that used time as the only predictor for the probability of a Class I error. However, because the contributions of these other factors (autopsy rate, country, case mix) were still noticeable (even if not statistically significant) and were clearly plausible, we used the more complete model to compute the mean error rate and the range of error rates shown in Table 3. This decision seemed especially appropriate given that, in the analysis for "major errors," time and autopsy rate both proved to have a statistically significant effect on the observed error rate.

Mean Diagnostic Error Rates and Relationships to Time and Autopsy Rate

The relationships between error rates and both time and autopsy rate were statistically significant for major errors, but not Class I error rates. Nonetheless, as shown below (Tables 4 and 5), the effects of these factors are noticeable enough for both error definitions so that a mean error rate *per se* is not meaningful in the absence of stipulated values for time and autopsy rate. Thus, the "means" listed in Table 3 are not true means. Rather, they represent the probability of an error for the base values of time, autopsy rate, case mix and country used in the regression model. Time was centered on 1980 for Class I and major errors and 1975 for major ICD

discrepancies. For case mix, the base category consisted of general inpatients (or general adult inpatients), and the reference value for country was defined as the U.S. Autopsy rates were centered on the mean autopsy rate for the studies included in the model.

Using the base values for case mix, country, autopsy rate (the overall mean autopsy rate of 44.3%), and time (1980, the midpoint of the period of analysis), the rate of autopsy-detected diagnostic errors was 10.2% (95% CI: 6.7-15.3%) and 25.6% (95% CI: 20.8-31.2%) for Class I and major errors, respectively. Restricting the analysis to U.S. institutions yielded similar point estimates and almost entirely overlapping confidence intervals—11.2% (95% CI: 6.9-17.5%) and 24.0% (95% CI: 17.3-32.3%) for Class I and major errors, respectively.

Error Rates as a function of Time and Autopsy Rates

The expected inverse correlation between error rate and study period (i.e., the more recent the study the lower the error rate) was observed for Class I and major errors, though this relationship was statistically significant only for the latter. Specifically, Class I errors decreased at a rate of 26.2% per decade relative to the base prevalence in 1980 ($p=0.10$). The probability of a major error exhibited a relative decrease of 28.0% (95% CI: 9.8-42.6%) per decade.

One would expect the rate of decrease for Class I errors and major errors to be similar. As this was, in fact, observed (with relative decreases of 26.2% and 28.0% per year for Class I and major error rates, respectively), it is likely that the decrease in Class I errors is real, and that the statistically non-significant result reflects inadequate power. In fact, the 95% confidence interval for this relationship between Class I errors and time extends from a 48.8% decrease to a 6.3% increase (i.e., the 95% CI lies predominantly on the side of a relative decrease), reinforcing the interpretation that the time trend for Class I errors is real.

The expected inverse correlation between error rate and autopsy rate (i.e., the higher the autopsy rate, the lower the error rate) was also observed, though this relationship was again statistically significant only for major errors. Specifically, for every 10% increase in autopsies, the Class I error rates exhibited a relative decrease of 7.8% ($p=0.2$) and major errors decreased by 12.0% ($p=0.0003$) relative to the base error rate.

In contrast with the other two definitions of errors, discrepancies in major ICD disease categories (between clinical and autopsy diagnoses) showed an increase over time, and this relationship was statistically significant. Specifically, the error rate increased by roughly 28% per year ($p<0.0001$). We had included this definition because, in principle, it represents a relatively objective type of error and because several of the studies using this definition were generally well designed and had high autopsy rates. Unfortunately, these positive features were likely overwhelmed by the heterogeneous effects introduced by variations in coding practices in different countries over different time periods.

Although the impacts of time and autopsy rate were statistically significant only for major errors, the cumulative effects were noticeable for both error types. Table 4 shows Class I error rates for varying autopsy rates in 4 different years. These results were calculated using the regression model, as shown in Appendix B. Table 5 shows the same trends for major errors. These trends are displayed graphically (with a greater range of values) in Appendix Figures 7 & 8.

The results of table 4 and 5 strongly suggest a trend toward decreased error rates over time, though the magnitude of the decrease is modest. These results corroborate the findings of the

single study[55] assessing trends in autopsy-detected error rates over time in which all time periods had high and approximately equal autopsy rates.

Using the regression model and standard error to calculate 95% confidence intervals (as shown in Appendix B) we can estimate the error rates expected in an academic medical center in the U.S., where one might find an autopsy rate of 20%. (In fact, many academic centers have even lower rates than this). The specific error rates to be expected with an autopsy rate of 20% are 6.7% (95% CI: 3.9-9.2%) and 20.3% (95% CI: 16.7-24.5%) for Class I and major errors respectively.

Publication Bias

Publication bias tends to operate in such a way that 'negative' results are less likely to appear in the peer review literature. For studies evaluating therapeutic interventions, this bias results in an under-representation of studies indicating no benefit for the intervention.[138, 139] Autopsy studies do not involve an intervention. Nonetheless, a study showing a low rate of important errors in clinical diagnosis might be regarded as analogous to a 'negative study'. One can imagine such a bias occurring, if, as with intervention studies, studies showing no significant errors at autopsy are considered uninteresting.

On the other hand, one could imagine publication bias operating in this case in the opposite direction, with studies showing high error rates less likely to appear. In fact, the bias need not operate at the level of publication *per se*. Institutions with particularly high error rates might choose not to submit such results for publication out of concern for the scrutiny or criticism it might bring. Alternatively, institutions with such high error rates might be less likely to conduct performance evaluation studies of this type in the first place.

To evaluate publication bias in this review, we constructed a funnel plot (Appendix Figures 9 & 10). Because trim-and-fill techniques are more appropriate for assessing systematic reviews of studies reporting odds ratios[140-142] and not proportions (such as rates of diagnostic errors), we assessed publication bias using the graphical assessment of funnel plots.[141, 143, 144] These plots have as their x- and y-axes the reported error rates and autopsy rates, respectively. In the absence of publication bias, such plots have the appearance of an inverted funnel.

In Appendix Figure 9a, there is a fairly definite appearance of a half-funnel. As it is the right half that is missing (i.e., the half with higher than average error rates), it would appear that publication bias is operating against studies reporting higher error rates. This appearance likely reflects the inclusion of studies reporting errors among different patient populations, which undermines the expectation of a symmetric funnel distribution. When only studies reporting general inpatients are included, the distribution of error rates is more symmetric (Appendix Figure 9b). The funnel plot for all studies reporting major errors (Appendix Figure 10a) shows greater symmetry than the plot in Appendix Figure 9a distribution, but it still appears that the right half of the plot has fewer data points. When inclusion is limited to general inpatient samples only (Appendix Figure 10b), there is still some asymmetry. In this case, it appears that there is an under-representation of large studies with low error rates and small studies with high error rates.

The analysis for publication bias is substantially limited by the complexities introduced by differences in case mix between studies and the competing directions in which publication bias might operate for autopsy studies, in contrast to most systematic topics. Overall, though, it appears unlikely that publication has substantially affected the results presented in this report.

Even if substantial publication bias were present, it is reasonable to expect this publication bias to remain stable over time, in which case the observed trend of a decrease in published error rates over time would still be meaningful. Finally, the results of the regression analysis indicating trends towards decreased error rates over time are corroborated by the single study[55] that assessed trends in autopsy-detected error rates over time and maintained high (and approximately equal) autopsy rates throughout the time periods assessed.

Summary of the autopsy as a diagnostic test

- The Class I error rate for the base time (1980), autopsy rate, case mix and country in a multivariate model was 10.2% (95% CI: 6.7-15.3%).
- The probability of the autopsy detecting a major error in a given case was 25.6% (95% CI: 20.8-31.2%).
- Restricting the analysis to U.S. institutions yielded similar point estimates and almost entirely overlapping confidence intervals for both Class I and major errors.
- These rates decreased by approximately 25% per decade over the 40-year observation period, though these decreases were statistically significant only for major errors.
- The combined effects of study period and autopsy rate on error rates are still noticeable, such that the overall rates of Class I and major errors do appear to be decreasing over time, though likely less so than expected by clinicians.
- Some publication bias may operate against studies reporting high error rates, but this bias is unlikely to account for the modest trend toward decreased diagnostic errors over time.

Clinical Selection of Cases for Autopsy

If the above analysis had shown a decrease in error rates over time by adjusting for decreases in autopsy rates, there would have been indirect support for the hypothesis that unchanged error rates reflect increased selection of cases by clinicians. In other words, as fewer autopsies are performed, the proportion of cases that presented diagnostic challenges to clinicians may increase, offsetting any improvements in overall diagnostic performance. Such an effect was not demonstrable in the above analysis, but is not ruled out by these results. Thus, we looked for studies specifically addressing the issue of clinical selection of cases for autopsy.

One study commonly cited as evidence that clinical selection does not occur simply demonstrated that major diagnostic groups did not differ between autopsied and non-autopsied cases.[57] Thus, what this study really shows is that patients with neoplastic diseases are no more or less likely to undergo autopsy than patients with cardiovascular disease, infectious diseases,

and so on. Even in this limited sense of selection, other studies have shown a bias towards certain diagnostic categories (e.g., trauma and “ill-defined conditions”) and a bias against other diagnostic categories (e.g., infectious and neoplastic) in referral for autopsy.[14]

More important than the diagnostic categories for principal causes of death is the level of diagnostic certainty, and whether or not clinicians refer cases for autopsy primarily in uncertain cases. Addressing this question requires assessing diagnostic certainty using a chart-based tool or prospectively eliciting from clinicians their level of diagnostic certainty prior to autopsy results becoming available.

In the only example we identified involving the first of these two approaches, Dhar et al formally assessed “diagnostic certainty” among deaths in a neonatal intensive care unit (ICU).[124] The authors defined three levels of diagnostic certainty ranging from “completely certain” (Level 1), in which the diagnosis was based on the definitive diagnostic standard attainable during life and on genetic, microbiologic or biochemical testing, to “probable, possible or suspected” (Level 3). As shown in the table 6, clinicians obtained autopsies for a greater percentage of patients with Level III than Level II, which in turn had a greater percentage of autopsies than Level I cases. Thus, the decision to obtain autopsies was clearly impacted by the level of certainty, even in this group of neonatal deaths with a higher overall autopsy rate than seen in most other patient populations.

Stronger evidence for the effect of clinical selection on the autopsy decision comes from a small number of studies conducted prospectively with the goal of assessing clinical certainty prior to autopsy. These studies are discussed in detail below; their results are summarized in Appendix Table 5.

Prospective Studies Addressing Autopsy Selection by Clinicians

1. Heasman MA, Lipworth L: Accuracy of certification of cause of death. In: General Register Office, ed. Studies on medical and populations subjects, No. 20. London: HMSO, 1966.

This prospective study conducted by government departments of health and vital statistics in the United Kingdom involved 75 British hospitals, including 23 in London and 52 scattered throughout England and Wales. The 9,501 autopsies included in this study consisted of adult inpatients who died during a 6-month period in 1959.

For each autopsy request, a clinician filled out a “dummy death certificate,” designed specifically for this study and independent of the usual death certificate. This form asked the clinician requesting autopsy to state the differential diagnosis at the time of death (prior to autopsy results) and indicate the cause of death as "fairly certain," probable" or "uncertain." Table 7 summarizes the autopsy rates and rates of clinical-autopsy diagnostic disagreement in terms of clinicians’ stated confidence in their diagnoses.

As expected, the less certain clinicians were regarding their diagnoses, the more likely they were to obtain an autopsy and the higher the chance of detecting an error in clinical diagnosis.

2. i) Britton M. Diagnostic errors discovered at autopsy. *Acta Med Scand* 1974;196:203-10. ii) Britton M. Clinical diagnostics: experience from 383 autopsied cases. *Acta Med Scand* 1974;196:211-9.

This study of 383 adult inpatient autopsies conducted at Serafimerlasarettet University Hospital in Sweden from 1970-71 is notable for its prospective design and 96% autopsy rate.

The author defined errors in clinical diagnosis in terms of discrepancies in ICD major disease categories. Error rates were then analyzed in terms of patient age and the certainty of clinical diagnosis as prospectively stated by clinicians prior to autopsy. Table 8 displays results for 333 of the 383 cases; the 50 cases identified by clinicians as "unknown" were not included, since (by definition for this study) no clinical diagnoses had been assigned to these cases prior to autopsy.

The study author reported comparisons of each italicized cell to the adjacent cell (vertically or horizontally) as statistically significant ($p<0.05$). The relevant comparisons for the totals were also significant (i.e., the total error rate for patents < 70 years of age was significantly lower than for older patients, and the total error rate for "fairly certain" cases was significantly lower than the error rate for the "probable" cases). Overall, the prevalence of errors in the cases in which the clinical diagnosis was only "probable" was almost double that seen in the "fairly certain" group, even when age is taken into account.

3. i) Cameron HM, McGoogan E: A prospective study of 1152 hospital autopsies: I. Inaccuracies in death certification. *J Pathol* 1981; 133: 273-83.
ii) Cameron HM, McGoogan E: A prospective study of 1152 hospital autopsies: II. Analysis of inaccuracies in clinical diagnoses and their significance. *J Pathol* 1981; 133: 285-300.

This study of inpatient autopsies from hospitals in the South Lothian District in Scotland from 1975-77 also asked clinicians to record their assessment of the likelihood of their clinical diagnoses as "fairly certain," "probable," or "uncertain." As shown in table 9, confirmation of clinical diagnoses correlated with clinical confidence.

4. Source: Cameron HM, McGoogan E, Watson H: Necropsy: a yardstick for clinical diagnoses. *Br Med J* 1980; 281: 985-8.

The authors conducted a follow-up to the previous study in which they worked with clinicians to try to increase the autopsy rate at one of the institutions from the previous study. As before clinicians were asked to record their confidence in the clinical diagnoses, but were also asked to state whether or not an autopsy would normally have been obtained (i.e., in the absence of the initiative to increase autopsy frequency).

Table 10 illustrates the impact of clinical selection on autopsy performance and diagnostic error rates. As shown in table 9, the confirmation of clinical diagnosis occurred more commonly in cases identified by clinicians as fairly certain." The following year, the autopsy rate doubled, with the result that more "certain" or "fairly certain" cases underwent autopsies, and as a result the confirmation rate increased (and corresponding error rate decreased). Interestingly, though, confirmation rates did not differ between the two groups defined by clinicians' assessment of the need for autopsy. Cases identified as unlikely to undergo autopsy in the absence of the study had the same error rate as the cases for which clinicians' stated they would pursue autopsy even without the study. A similar result was obtained in the study summarized next.

5. Hartveit F. Clinical and Post-Mortem Assessment of the Cause of Death. *J Pathol* 1977;4:193-210.

In this Norwegian study from The University of Bergen, clinicians were asked to identify cases as clinically "certain" or "uncertain," but were also asked prospectively to record their impression of the autopsy as "essential," "desirable," or "of little interest." The results of this study are presented in a very complicated manner, but table 11conveys one of the main findings of the study, namely that clinicians' opinion of the necessity of an autopsy does not derive solely from their confidence in the clinical diagnosis. Even when clinicians regarded the diagnosis as

clinically uncertain, autopsy was regarded as essential in only 45%. Also, though not indicated in the table below, autopsies in children were more likely to be regarded as essential despite the fact that clinical certainty about the diagnosis was also higher in children.

6. Landefeld CS, Chren MM, Myers A, Geller R, Robbins S, Goldman L. Diagnostic yield of the autopsy in a university hospital and a community hospital. *N Engl J Med* 1988; 318: 1249-54.

Investigators in this prospective case-control study conducted daily reviews of in-hospital deaths in order to contact clinicians prior to receipt of autopsy results. Patients undergoing autopsy were paired with the next adult who died in the same hospital without undergoing autopsy.

For each death, regardless of whether or not autopsy was eventually performed, clinicians were asked to record their assessment of the probability that the autopsy would reveal a major undiagnosed cause of death (i.e., a Class I or II finding). The questionnaire asked for an absolute probability (percentage from 0-100) and a qualitative estimate (5 point scale: “much more likely than usual,” more likely than usual,” “as likely as usual,” “less likely than usual,” much less likely than usual.”) The chi-square test for linear trend revealed no significant relationship between the physicians' estimated probability of Class I or II errors and the observed prevalence of such findings. The relationship between qualitative estimates and unexpected findings was not presented, but was reported as non-significant as well.

The studies discussed above generally bear out the intuitively plausible claim that clinicians’ diagnostic error rates occur more commonly in cases that clinicians identify as diagnostically uncertain. However, important diagnostic errors were still found in cases clinicians rated as diagnostically certain. Moreover, the relationship between clinicians’ reported confidence in antemortem diagnoses and autopsy-detected error rate is complex. On the one hand, some studies show that clinicians can identify levels of diagnostic certainty that correlate with the rate of unexpected findings at autopsy. On the other hand, clinicians show little ability to predict the utility of the autopsy. One possible explanation is that the decision to request autopsy is not completely determined by clinicians’ clinical certainty. Another possibility is that, in the context of a study explicitly targeting the accuracy of clinical diagnosis, clinicians reflected more carefully on the confidence of the clinical diagnosis (or even downplayed it) than they would routinely or when considering the decision to pursue autopsy. Probably both explanations play a role.

Summary of the autopsy as a diagnostic test

- The relationship between clinicians' reported confidence in antemortem diagnoses and autopsy-detected error rate is complex.
- To some extent, clinicians can identify levels of diagnostic certainty.
- Cases with greater uncertainty are more likely to go for autopsy and, in some studies, are more likely to have diagnostic errors.
- However, even in cases clinicians identify as "certain" with respect to confidence in diagnosis, the autopsy reveals major errors in roughly 5-15% of cases.
- Moreover, when asked specifically to predict unexpected findings or to state if the autopsy would be important/useful in a particular case, clinicians do not appear able to predict the findings (or utility) of postmortem findings.

Changes in Diagnostic Errors Over Time

The above sections have indicated that rates of diagnostic errors detected at autopsy have shown little change over time, and that this lack of change likely reflects an effect other than decreased autopsy rates and increased clinical selection. It is possible that, while the rates of errors have remained fairly constant, the nature of the errors has changed. For instance, cirrhosis, bacterial pneumonia, and tumors were commonly identified as missed clinical diagnoses in autopsy studies prior to 1965. [51, 145] Subsequent studies were more likely to reveal previously uncommon diagnoses such as systemic fungal infections, but also pulmonary emboli.[51, 145]

The small numbers of specific diagnoses in each autopsy study (e.g., 1 missed aortic dissection, 5 missed pulmonary emboli, 2 missed cases of pneumonia) and inconsistent reporting between studies preclude formal analysis of trends in the nature of the missed diagnoses. Further complicating such analysis is the question of whether to report that, for example, 5 of the 100 autopsies were cases of missed pulmonary emboli or that, of the 10 cases of pulmonary emboli detected at autopsy, 5 were clinically missed. Most studies did not report sufficient data to explore these two ways of summarizing error rates for specific diseases and assessing trends over time.

While "newer diseases" such as opportunistic infections have undoubtedly increased in recent decades and account for some of the misdiagnoses detected at autopsy, clinically missed diagnoses continue to include common diagnoses such as myocardial infarction, pulmonary embolism, bowel perforation, and other common conditions.

Impact of the Autopsy in Detecting of Unsuspected Complications of Care

The autopsy has traditionally played a prominent role in morbidity and mortality rounds, in which clinicians review cases with adverse outcomes for potential quality problems. Unfortunately, even in surgical departments, where these rounds have retained their traditional focus to a greater extent than in many departments of medicine, only a fraction of the complications on a given surgical service typically result in discussion.[146] Moreover, many clinicians may assume that autopsies simply confirm recognized complications and do not detect unsuspected ones.

Because of the recent interest in medical error and patient safety,[147, 148] we specifically looked for studies that reported the proportion of autopsies that detected clinically unsuspected complications of care. As shown in Appendix Table 6, roughly 1-5% of autopsies disclose unsuspected complications of care. This range has somewhat less variability than do studies of error rates in general (i.e., few studies show substantially higher rates). Unfortunately, these data were usually mentioned in passing in these studies, so the extent to which these complications contributed to deaths (and even the extent to which they were truly unsuspected) was often unclear. For this reason, and because of the heterogeneity of the case mixes in the relatively small sample of studies reporting the relevant data, we did not pool estimates of rates of autopsy-detection of unsuspected complications of care.

Nonetheless, it is interesting to note that this complication rate is similar to the rate of serious complications of care noted in population based studies such as the Harvard Medical Practice Study[134] and the more recent Colorado-Utah Study.[136] In other words, despite the various selection forces operating on the autopsy decision, the rate at which it reveals complications of care is comparable to the rates detected by chart-based studies.

The Autopsy and Malpractice

Fear of litigation is often cited as a reason for the decline in autopsy rates, i.e., the presumption that clinicians fear that autopsy detected errors will lead to malpractice claims. Pathologists have offered anecdotal evidence to the contrary, i.e., that autopsy findings can resolve potential malpractice claims by definitively refuting alleged diagnoses or complications.[149-151] Only one study of diagnostic discrepancies detected at autopsy directly addressed the relationship between postmortem findings and medicolegal exposure for clinicians and healthcare organizations. This study[152] of diagnostic errors in 176 autopsies was conducted at the University of Pittsburgh Medical Center in 1994. In addition, the authors reviewed all cases after the two-year statute of limitations on malpractice suits (in Pennsylvania) had expired. Only one malpractice lawsuit had been filed, and the intent to proceed to litigation in that case had been declared prior to the patient's death.

A second study explicitly addressed the issue of the role of the autopsy in malpractice suits. Unfortunately, all 15 cases reviewed in this study[153] involved autopsies requested after the filing of suit. Thus, the information from this small series does not address the issue of a possible increase in medicolegal exposure due to routine autopsy performance.

A third survey study assessed the contribution of the autopsy to loss control/ risk management.[154] The author's principal interest lay in determining the extent to which non-

forensic autopsies increased or reduced institutional medicolegal exposure. The survey of pathologists at 183 teaching institutions had a response rate of only 31%, and the survey instrument appeared quite informal. In response to the main survey question, 36 of 58 respondents (62%) indicated that autopsies reduce overall institutional medicolegal losses.

Summary of the autopsy as a diagnostic test

- Only two studies have addressed the relationship between routine autopsy performance and medicolegal exposure.
- One of these studies indicated no increased risk of malpractice as a result of autopsy findings.
- The other study, a survey of pathologists' impressions of the relationship between the autopsy and risk management activities, suggested that the autopsy may reduce legal exposure more often than it increases it.
- Overall, the relative frequencies with which the autopsy generates information that is legally harmful or helpful to an institution remain unclear and deserve further study.

The Autopsy as a Performance Measure for Clinical Diagnosis

The findings reviewed in the previous section indicate that, for a given patient who has died, there is a roughly 10% chance that autopsy will detect a clinically significant misdiagnosis. Moreover, clinicians' ability to predict cases likely to have unexpected findings at autopsy is at best weak to moderate. On an individual level, these data are quite compelling – i.e., from the point of view of an individual clinician deciding whether or not to pursue autopsy in the case of an individual patient death. But what about at the hospital or population level? To what extent do clinicians miss certain diagnoses in all patients, not just patients who die? As outlined in the Objectives section, this question can be understood as asking what proportion of clinically important diagnoses are first detected at autopsy.

Despite the extensive literature on diagnostic errors detected at autopsy, this question remains largely unanswerable with the existing literature, especially the autopsy literature *per se.* This gap in the literature does not result from the problem of selection bias – i.e., that clinicians may request autopsies precisely because they are uncertain about clinical diagnoses, so that autopsied cases are not representative of all deaths. Rather, this basic question remains largely unanswerable because the prevalence of clinically missed diagnoses at autopsy is not the same as the prevalence of diagnostic errors among all clinical cases, which is the true error rate of interest.

A typical study of diagnostic errors found at autopsy reports the number of clinically missed diagnoses and often lists the particularly common missed diagnoses (e.g., myocardial infarction, pulmonary embolism, systematic fungal infection, perforated bowel, and so on.) With only one exception,[155, 156] these studies include no information relevant to estimating the relevant

denominator, namely all cases of myocardial infarction, pulmonary embolism, fungal infection, and so on. If patient deaths were a random sample of all patients, error rates from autopsies alone would provide reasonable estimates of error rates for clinical diagnosis. However, precisely because some deaths represent diagnostic failures (not just treatment failures), autopsies do not provide a random sample with which to gauge the performance of clinical diagnosis.

Because of the importance of this issue, the one study from the autopsy literature that addresses the overall performance of clinical diagnosis is discussed in detail below. Before doing so, we address a point that will arise repeatedly in the discussion of estimates for clinical diagnostic performance, namely how to handle the non-autopsied deaths.

Overall Performance of Clinical Diagnosis

Methodological Note

Even if a study provides data on the patients with a given diagnosis who left the hospital alive, estimating the sensitivity of clinical diagnosis is hampered by the possibility of a significant number of unidentified cases with this diagnosis among the non-autopsied deaths. Thus, if a study has an autopsy rate of 20%, we need to have a reasonable approximation for the prevalence of errors among the 80% of deaths that did not undergo autopsy.

In developing this approximation, we need to consider the possibility that non-autopsied deaths might have even more errors than observed among autopsied cases. For instance, clinicians might have a bias against referring cases with errors for autopsy (e.g., to avoid medicolegal exposure), resulting in an under-representation of diagnostic errors among cases referred for autopsy. This possibility stands diametrically opposed to the view many clinicians have, which is to assume that the few (non-forensic) cases sent for autopsy are requested precisely because of diagnostic uncertainty.

From the previous section, it appears that clinicians do not avoid sending cases with diagnostic uncertainty for autopsy, so over-representation of errors among the non-autopsied cases is unlikely. On the other hand, clinicians have a very limited ability to predict the detection of diagnostic errors at autopsy. Thus, the assumption of random distribution between autopsied and non-autopsied cases represents a reasonable lower bound for estimating the sensitivity of clinical diagnosis. In other words, if a series with an autopsy rate of 20% reported 5 Class I errors, the upper bound for the sensitivity of clinical diagnosis would consider these as the only "false negatives" for clinical diagnosis. Thus, one way (admittedly still somewhat arbitrary) to estimate the lower bound, the false negatives would include 5 times this number to account for the possibility that the 80% of non-autopsied deaths included the same proportion of errors as the autopsied cases.

The single autopsy study including prevalence data for principal diagnoses among patients who did not die

i) Cameron HM, McGoogan E: A prospective study of 1152 hospital autopsies: I. Inaccuracies in death certification. *J Pathol* 1981; 133: 273-83.

ii) Cameron HM, McGoogan E: A prospective study of 1152 hospital autopsies: II. Analysis of inaccuracies in clinical diagnoses and their significance. *J Pathol* 1981; 133: 285-300.

The authors of the two companion articles[155, 156] describing this study reported the total number of patients discharged (alive or dead) with the principal diagnoses highlighted among the autopsy cases, permitting estimation of the sensitivity and specificity of clinical diagnosis.

Using these data to estimate diagnostic performance presents problems, as we do not know the accuracy of the clinical diagnoses in patients except in the case of the patients who went to autopsy. However, reasonable estimates are possible for TB and appendicitis because the ultimate clinical diagnoses for these conditions are often unequivocal (microbiologic results, findings at laparotomy and subsequent pathology). For other of these diagnoses, it is difficult to make comparable calculations because the certainty of these clinical diagnoses exhibits greater heterogeneity, with many diagnoses reflecting a general clinical impression rather than definitive results from specific investigations (e.g., pulmonary angiography, liver biopsy).

Pulmonary TB. In 1976 (the middle study year), 138 patients were discharged alive with pulmonary TB as the principal diagnosis and 5 patients died with this as the main diagnosis. Among the 15 patients who died in the 3-year study period with a principal clinical diagnosis of pulmonary TB, autopsy confirmed this diagnosis in 7 cases and rejected it in 8. In an additional 7 deaths, autopsy disclosed pulmonary TB as the main diagnosis or the immediate cause of death. As a rough approximation, then, it appears we can assume that the patients who die with a clinical diagnosis of pulmonary TB are approximately equal in number to the missed deaths due to pulmonary TB. This estimate ignores patients who died but did not undergo autopsy (75% of deaths in this study). Unfortunately, we do not know if clinicians are more or less likely to

autopsy cases of TB. If we discount the possibility of a bias *against* autopsying such cases,[2] then a reasonable lower bound for clinical diagnosis is that in which the diagnosis of TB is evenly distributed between autopsied and non-autopsied deaths. With an autopsy rate of 25%, this assumption implies an additional 3*7=21 clinically undetected cases. Thus, the lower bound for the sensitivity of clinical diagnosis for pulmonary TB can be estimated as 138/(143 + 21) = 84%.

This estimate ignores patients discharged alive with a missed diagnosis of pulmonary TB, but presumably most of these patients are either subsequently diagnosed or eventually counted among the deaths. The above estimate also assumes that all of the patients discharged alive in fact had pulmonary TB. The rate of false positive diagnoses for pulmonary TB is low, as the false positive rate for TB culture is extremely low. In fact, most false positives consist of culture-negative cases treated on clinical grounds. Assuming even 10% false-positives, the sensitivity of clinical diagnosis for pulmonary TB remains relatively unchanged at 124/150 = 83%.

Acute appendicitis. This example illustrates the complete disconnect between the "error rates" and sensitivity values presented in autopsy studies as compared with the actual sensitivity of clinical diagnosis. Focusing on the autopsy cases only, the "sensitivity" of clinical diagnosis for acute appendicitis would be calculated as 3/4=75%. In fact, though, the above data suggest sensitivity closer to 98%.

---

[2] In principle, clinicians might have a bias against referring cases with errors for autopsy (e.g., to avoid medicolegal exposure). This possibility stands in opposition to the view of many clinicians that cases are sent for autopsy precisely because of diagnostic uncertainty. From the previous section, it appears that the truth lies somewhere in between these two extremes. Clinicians may think they are sending for autopsy those cases most likely to have diagnostic errors, but it turns out that have only a limited ability to identify such cases correctly. Thus, the assumption of random distribution between autopsied and non-autopsied cases may be a reasonable approximation in the context of estimating prevalence of clinically missed cases.

Given standard practice in the developed world, patients discharged alive with the diagnosis of appendicitis have undergone surgery (as opposed to empiric antibiotic treatment without operation) so the diagnosis has been confirmed. Autopsies were performed on all four cases of patients who died before going to surgery. Thus, for appendicitis, we can calculate the sensitivity of clinical diagnosis quite precisely. True positives = 441 (alive) + 4 (dead) – 1 overturned = 444. Since one case was missed, the total number of cases is 444 + 1, so that the sensitivity of clinical diagnosis is 444/445 = 99.8%. Again, however, we have to take into account the possibility of missed cases among non-autopsied deaths. The 25% of deaths that underwent autopsy contained 1 clinically missed case. If we assume that the other 75% of deaths included 6 cases (i.e., same proportion as observed among autopsied deaths), the sensitivity would still be 444/451 = 98%.

This sensitivity estimate has a wide confidence interval (due to the small number of autopsied cases and the autopsy rate of only 25%), so the true sensitivity may be lower. Moreover, the above estimate does not address the detection of the diagnosis in a timely enough fashion to avoid perforation. Cases in which clinicians recognized the diagnosis only after perforation has occurred represent potential quality problems. Also, the specificity of the diagnosis of acute appendicitis is an important performance measure in practice, as a high sensitivity should not be achieved by submitting a large number of patients to unnecessary laparotomy. These two issues are discussed in the section below, in which appendicitis is one of the five specific conditions reviewed.

Sensitivity of Clinical Diagnosis for Five Target Conditions

Because this one study provides the only data from the autopsy literature *per se* relevant to assessing the extent to which the autopsy provides a valid measure of clinical diagnostic performance, we conducted ancillary searches from the general clinical literature. Our goal consisted of finding population level studies in which, through a combination of autopsy for patients who die and unequivocal diagnostic testing or long-term clinical follow-up for patients who do not, error rates for clinical diagnosis can be established.

Given the breadth of the potentially relevant literature, our search may not be completely exhaustive. However, we found at least one study providing an estimate of the sensitivity of clinical diagnosis in detecting:

- acute pulmonary embolism among patients presenting to the hospital
- cases of myocardial infarction among patients presenting to the hospital with chest pain
- acute appendicitis
- acute dissection of the aorta
- active tuberculosis

Data addressing the performance of clinical diagnosis for these five conditions are presented below and summarized in Appendix Table 7.

Acute Pulmonary Embolism

Many general autopsy series list pulmonary embolism as a significant percentage of the major clinically missed diagnoses. [2,3,51,54,55,70,116,119,122,129,131,152,155-160] Additionally, multiple

autopsy studies have specifically assessed the prevalence of clinically significant (but clinically undetected) PE at autopsy.[161-173] The most recent of these studies is representative of the general problem with using such data to assess the performance of clinical diagnosis. Pineda et al[173] reviewed 778 autopsy reports from 1991-96 at a teaching hospital and identified 67 patients with PE as the primary or major cause of death. Review of the clinical records in these 67 cases of fatal PE indicated that the diagnosis had been suspected clinically in only 30 (45%) patients. The authors compared this prevalence of clinically missed fatal PE with data from previous studies,[161, 162, 171] and concluded that the clinical suspicion of fatal PE has shown only minimal improvement over time.

This conclusion involves an important misconception that illustrates a problem common to many of the studies in the autopsy literature. The benchmark for the sensitivity of clinical diagnosis of PE is generally taken to be 96%,[174, 175] as the implied false negative rate of 4% corresponds to the prevalence of PE among patients with normal lung ventilation-perfusion scans. [176] No benchmark exists, however, for the sensitivity of clinical diagnosis in detecting fatal PE. One obvious reason for the lack of such a benchmark is that clinicians do not attempt to diagnose fatal PE; they attempt to diagnose PE among patients before they die.

In fact, it is not clear what the ideal sensitivity for the clinical detection of fatal PE ought to be (or, equivalently, what is an acceptable "miss rate" is for fatal PE). As stated more generally in the introduction, patients who die of PE generally fall into one of two categories – treatment failures and diagnostic failures. Contrary to the expectation implied by the authors of many autopsy studies of fatal PE, [161, 162, 171, 173] improvements in care need not reduce the relative proportion of the latter. For instance, if the number of patients in the first category (treatment failures) decreases over time (due to earlier and more aggressive anticoagulation), one might even expect an increase in the relative proportion of diagnostic failures.[3]

In addition to the above misconception, the rate of clinically missed fatal PE bears little connection to the rate of greater interest, namely the sensitivity of clinical diagnosis in detecting acute PE among all patients. Assessing the overall sensitivity of clinical diagnosis for PE in routine practice requires estimating a ratio that has as its numerator the "true positives" recognized during life (regardless of whether or not they subsequently died). The denominator consists of all cases of PE, i.e., true positives plus false negatives. The false negatives include those cases of PE detected only at autopsy plus the additional clinically missed cases among patients who died but not did not undergo autopsy.

Reasonable assumptions about the prevalence of PE among non-autopsied cases combined with published data permit an estimate of this numerator and denominator. For instance, the International Cooperative Pulmonary Embolism Registry included 2454 patients with suspected or confirmed PE, with 2110 cases confirmed by autopsy, angiography or venous ultrasound combined with a high clinical suspicion.[177] In 61 cases, the diagnosis was detected only at autopsy, so the clinically detected "true positives" included 2049 cases. This estimate ignores the clinical false positives among the 2,049 clinically detected cases. It also ignores the additional true positives among the 344 remaining patients with less clear-cut diagnoses of PE. Applying rates of false positive and false negative diagnoses from large clinical studies[175, 176] suggests that

[3] It is also possible that advances in diagnosis would allow more cases to be recognized than can be successfully treated, resulting in an increase in treatment failures relative to diagnostic failures. The main point is that the behavior of diagnostic error rates (among patients who die) over time is complex, so that it is not clear to what extent one ought to expect a decrease over time in the proportion of fatal PEs that were clinically missed.

these two factors roughly cancel each other out (as they affect the estimate in opposite directions), but this is a rough approximation.

At first glance, the above results suggest that the sensitivity of clinical diagnosis could be as high as 2049/2110=97%. However, this estimate does not take into account missed PEs among non-autopsied deaths. If we assume an average autopsy rate of 20% (the participating centers were all teaching hospitals), then the number of clinically missed cases could be as high as 305, rather than 61 (i.e., assuming the same proportion of missed cases among non-autopsied and autopsied deaths). The resulting estimate for sensitivity is 2049/2354=87%. Overall, then it appears that the sensitivity of clinical diagnosis for detecting PE falls within the range of 87-97%.

This range suggests that the performance of clinical diagnosis may fall short of the ideal benchmark of 96%. Nonetheless, by taking into account all cases of PE (even with the above pessimistic assumptions) the estimated performance of clinical diagnosis is strikingly different from the picture suggested by autopsy data alone, in which 30-50% of cases are clinically missed.

Only one autopsy study[172] attempted to assess the performance of clinical diagnosis of PE among all patients and not just those who die. The authors reviewed autopsy data from their institution during the period in which their institution participated in the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) study.[176] The authors reviewed all autopsies for cases of PE and also analyzed data from PIOPED patients at their institution, with the primary objective of estimating the prevalence of acute PE among hospitalized patients. Unfortunately, this study did not take into account the roughly 45% of all patients with suspected PE who were not eligible for inclusion in PIOPED, nor the roughly 50% of eligible patients who declined participation or the 10-40% of patients (depending on the institution) who consented but were not selected for assessment of the sensitivity and specificity of ventilation-perfusion scans.[176] These exclusions make assessments of the sensitivity of clinical diagnosis for PE based on these data much more speculative than the estimate generated above from the ICOPER data.

Acute Myocardial Infarction

Many general studies of diagnostic errors detected at autopsy report cases of myocardial infarction (MI) as a condition contributing to death but unsuspected antemortem. [2, 3, 51, 54, 55, 70, 116, 119, 122, 129, 131, 155, 156, 159, 160] Other studies have specifically addressed clinicopathologic discrepancies in the diagnosis of myocardial infarction. [178, 179] As in the discussion of PE above, the true question of interest is the sensitivity of clinical diagnosis in detecting MI among all patients, rather than the sensitivity for fatal MI.

Analyses of large cohort studies (e.g., the Framingham Study),[180-187] have allowed estimates of the prevalence of unrecognized MI. As reviewed recently, [188] these studies suggest that clinically unrecognized myocardial infarction accounts for at least 25% of all myocardial infarctions. The high prevalence of unrecognized MI has clear implications for public health strategies and clinical practice in terms of screening and diagnosing patients. Importantly, though, these data combine clinically silent MI (i.e., patients who experience asymptomatic MI) with symptomatic, but undiagnosed patients. As the former group accounts for the majority of unrecognized MI,[188] these data do not address the issue of errors in clinical diagnosis, as many of these patients do not present to medical attention.

More appropriate for the assessment of the sensitivity of clinical diagnosis for acute MI are the several large studies of myocardial ischemia among patients presenting to the hospital with acute symptoms.[189-192] The three more recent of these studies[189-191] indicate that 2-4% of patients who present to the hospital with symptoms related to acute MI are clinically missed and inappropriately discharged from the emergency department.

Acute Appendicitis

Appendicitis is a common condition, with a lifetime risk of 8.6% for males and 6.7% for females. Approximately 12% of males and 23% of females undergo appendectomy (with the difference in rate due in large part to more frequent opportunities for incidental appendectomies among women).[193] Clinical performance measures for acute appendicitis have typically included rates of perforation (reflecting delayed diagnosis and increased complications) and rates of normal appendix at laparotomy (reflecting unnecessary exposure of patients to risks of surgery).[194] With the introduction of helical CT as a sensitive and specific test for acute appendicitis,[195, 196] it seems reasonable to hypothesize that rates of appendiceal perforation (reflecting delayed diagnosis) and normal appendices discovered at laparotomy should have decreased in recent years.

One recent study[197] used procedure codes from a state hospital discharge database to identify patients undergoing appendectomy from 1987-1998. Among 63,707 non-incidental appendectomy patients, 84.5% had appendicitis (25.8% with perforation) and 15.5% had no associated diagnosis of appendicitis. Adjusting for important demographic features (age, gender) showed that the population-based incidence of unnecessary appendectomy and of appendicitis with perforation had not changed significantly over time. While many new technologies take time to diffuse into routine practice, ten years is a fairly long time period, especially for an easily implemented diagnostic test (a type of computed tomography scan) affecting such a common condition.

Alternatively, the lack of a time trend might be due to inaccuracies in administrative data. While such inaccuracies are well-known, there is no reason to suppose that they significantly changed during this study period, so that time trend analyses may be quite accurate. Moreover, another recent study reported similar findings using clinical data. This study[194] included two cohorts, one of which consisted of 1,026 patients undergoing appendectomy. The authors reported a normal appendix in 10.5% of cases, with a range of 4.7-19.5% for the 12 institutions participating in the study.

This study[194] included a second cohort of consecutive patients presenting to the emergency department (ED) with abdominal pain. The 1,118 patients identified for this cohort included 44 patients who ultimately proved to have appendicitis. Focusing on physicians initially assessing these patients (in the office or ED), the sensitivity of clinical diagnosis was 81.4%, but with an institutional range from 72.2-89.4%. Moreover, perforation was observed in the concurrent appendectomy cohort in 20.3% of cases with a range from 6.9-33%.

Another recent study did show reductions in appendiceal perforation and false positive rates (i.e., normal appendix at laparotomy).[198] This study may accurately reflect institutional expertise with appendiceal CT and/or the development of an efficient protocol for the work-up of suspected appendicitis. On the other hand, this study did not include all patients presenting with abdominal pain, so that patients who subsequently re-presented (i.e., false negatives) to another hospital would not be captured.

Acute Dissection of the Aorta

General autopsy series commonly list cases of missed aortic dissection among the major diagnostic errors.[2, 3, 116, 120, 199] A number of older studies from the clinical literature report 50% or more cases as clinically missed.[200-202] One of the first of the more modern studies[203] provides a misleadingly high estimate of clinical sensitivity because, as the authors note, the series was predominantly clinical, with less systematic attempts to identify cases detected only at autopsy. Three other recent studies more comprehensively look for cases detected only at autopsy.[204-206] The most recent of these studies [206] focused only on clinical diagnosis or suspicion by physicians in the emergency department, so this study was not considered further.

One study[204] reports the clinical findings in a series of 235 patients with aortic dissection seen at the Mayo Clinic in Rochester from 1980-90. The diagnosis was confirmed by surgery (162), autopsy (27) or radiologic testing without surgery (47); iatrogenic dissections (e.g., cases that occurred during vascular surgery or catheterization) were excluded. This series included 59 patients referred from outside facilities with the radiologic diagnosis of aortic dissection. Including these patients significantly overestimates sensitivity of clinical diagnosis, because clinically missed cases from these referring institutions are not included. Among the 176 patients presenting initially to the Mayo Clinic, 17 cases were not identified until autopsy, suggesting a sensitivity of 159/176= 90%. This estimate does not take into account missed cases among non-autopsied deaths. The study does not report the autopsy rate. If we use published data on autopsy rates in Olmsted county[323], we can estimate the autopsy rate as approximately 30%. As a rough estimate, then, the non-autopsied deaths together with the autopsied cases included as many as 56 (i.e., 17/0.3) clinically missed aortic dissections, suggesting the sensitivity of clinical diagnosis of 159/215=74%. Thus, the sensitivity of clinical diagnosis for dissection of the aorta has 90% as its upper bound (if the non-autopsied deaths contained no missed cases) and as its lower bound 74% (if the non-autopsied deaths contained the same proportion of missed cases as detected among the autopsied patients).

Another article[205] describes the clinical findings and outcomes for 258 patients with aortic dissection from 1966 to 1986. (This series includes the patients reported in an earlier study from the same institution, so this earlier article is not discussed.[207]) In this series of 258 patients with 259 dissections, 69 cases were identified only at autopsy (including 58 acute Type A dissections.) Thus, the sensitivity of clinical diagnosis in this series was no higher than 190/259= 73%. Again, this estimate is an upper limit because it does not include missed cases among non-autopsied deaths. If we attribute an autopsy rate as high as that attributed above (i.e., 30%), then the total number of missed cases would be 69/0.3=230, for an overall sensitivity of 90/(190+230)=45%. If the institution involved had a lower autopsy rate, the estimated sensitivity would be even lower. Thus, this study[205] suggests that clinical diagnosis detects no more than 73% of cases and possibly as few as 45%.

Active Tuberculosis

Many general autopsy series list tuberculosis (TB) among the clinically significant missed diagnoses first detected at autopsy.[116, 127, 129, 131, 156, 208-211] Studies specifically assessing the prevalence of TB among autopsied patients have reported that roughly 50% of cases were not detected antemortem.[212-217]

The difference between the proportion of diagnostic failures among fatal cases and the failure rate for clinical diagnosis for all cases is particularly striking for an eminently treatable infectious disease such as TB. As discussed previously, the single autopsy study reporting diagnoses for patients discharged alive suggested a sensitivity of clinical diagnosis in detecting pulmonary TB of greater than 83%.[155, 156] We identified two other studies providing population level data on all diagnoses of TB. [218, 219] The more recent study [219] reviewed all cases of TB reported to the San Francisco Department of Health from 1986-95. Among 3102 reported cases of TB, 120 (3.9%) met the definition for diagnosis after death. (This definition included patients who literally were diagnosed after death with postmortem tissue cultures, but also patients who were not receiving therapy with more than one agent at the time of death.) The earlier study at the national level reported approximately the same result, with 5.1% of cases meeting the same definition of diagnosis after death.

Unfortunately, neither of these results takes into account missed cases among non-autopsied death. Moreover, we do not know the rate of autopsy for the catchment areas involved. Given the variation in autopsy rates between institutions and the predominance of non-teaching hospitals (typically with low autopsy rates) it would be unlikely for the effective autopsy rate to be higher than 10% (see national autopsy rates in Appendix Figure 2). As before, regarding missed cases of TB as evenly distributed between autopsied and non-autopsied deaths provides a reasonable lower bound for the sensitivity of clinical diagnosis. With the non-autopsied cases thus including as many as 9*120=1080 additional cases, the sensitivity of clinical diagnosis for TB could be as low as 39%.

In summary, the overall performance of clinical diagnosis may diverge significantly from that suggested by autopsy results alone. Unfortunately, few data assess the performance of clinical diagnosis (i.e., answering the question of how often cases of a given diagnosis remain undetected during a patient's life), and studies generally come from the clinical literature, rather than the autopsy studies. (Finding such studies thus involved extensive searches outside the originally targeted autopsy literature.)

Among the five conditions for which we found relevant data, the sensitivity of clinical diagnosis substantially exceeded the performance suggested by autopsy studies for clinical detection of PE and acute MI. The sensitivity of acute appendicitis is also relatively high, but not necessarily as high as many clinicians may believe. For the other two diagnoses—TB and aortic dissection—clinical or population level studies confirm the findings of the autopsy literature, with a substantial number of cases clinically missed. The example of acute appendicitis also illustrates that advances in diagnostic testing may not always translate into advances in overall diagnostic sensitivity.

Even for the two conditions for which clinical diagnosis appears to perform relatively well, it is possible that these high sensitivities overstate clinical performance. By focusing on the identification of target conditions such as PE and MI, clinicians may miss other important conditions once these target diagnoses are ruled out. Cohort studies of patients being investigated for MI or PE do not evaluate the extent to which other diagnoses are missed as a result of focusing on the identification of these diagnoses. A patient who is correctly identified as not having an MI or PE counts as a success, regardless of whether or not the patient eventually receives a diagnosis other than "non-cardiac chest pain" or "unexplained dyspnea."

Logistic Barriers to using the Autopsy for Performance Measurement

The use of autopsy data—whether as conventionally reported error rates (proportion of autopsies with diagnostic errors) or as rates with the true denominator of interest (all cases treated during the observation period) —presents formidable practical problems in terms of clinical performance measurement. Only a small percentage of patients admitted to the hospital die, only a minority of these patients undergo autopsy, and only in a minority of these cases are clinically important misdiagnoses. Moreover, only a portion of these misdiagnoses represent true "errors" (and therefore quality problems), as many represent atypical presentations—in fact, cases in which the clinicians were frankly unsure of the diagnosis prior to death were counted as "errors" in most of the autopsy studies—or deliberate decisions not to pursue aggressive diagnostic work-ups (e.g., in chronically ill patients approaching the end of life). The opportunity to detect true quality problems using the autopsy data at a single institution is thus quite small and error rates derived solely from autopsy data are unlikely to generate statistically meaningful measures for comparing or benchmarking institutional performance.

Summary of the autopsy as a diagnostic test

- The overall performance of clinical diagnosis may diverge significantly from that suggested by autopsy results alone.
- Assessing the performance of clinical diagnosis requires supplementing autopsy data with discharge diagnoses or other follow-up information concerning patients who do not die.
- We found studies providing data on the performance of clinical diagnosis for only five conditions: acute MI, PE, aortic dissection, acute appendicitis, and active tuberculosis. Among these five conditions, clinical diagnosis exhibited substantial variation in sensitivity, with satisfactory values only for acute MI and to a lesser extent PE. Moreover, these results may overstate the performance of clinical diagnosis as they focus on the dichotomous outcome of correct recognition or exclusion of a specific target condition, rather than successful diagnosis of whatever condition the patients may have.
- The existing literature indicates that antemortem diagnosis likely does not perform as well as generally believed by clinicians. Unfortunately, effective use of autopsy data for institutional performance measurement is hampered by problems with small numbers and the difficulties of obtaining the true error rates of interest (diagnostic errors in all cases). At the level of the health care system, though, autopsy data could likely be combined with other clinical data to provide meaningful estimates of the performance of clinical diagnosis in routine practice.

Impact of the Autopsy on Clinical Performance Improvement

Two examples from the medical error and patient safety literature illustrate the type of study required to demonstrate a benefit for performance improvement. In a prospective study aimed at reducing errors in radiographic interpretation by emergency physicians, example errors were used to create a teaching file.[220] Moreover, a formal protocol for interpreting radiographs was adopted, and attending radiologists read all films within 12 hours to provide feedback and quality control. Subsequent outcome measures showed a significant decrease in the rate of significant missed diagnoses from 3% (95% CI: 2.8-3.2%) to 1.2% (95% CI: 1.03% to 1.37%). Further process redesign achieved a further reduction to 0.3% (0.26% to 0.34%). On a larger scale (comparable to what would be involved in an autopsy performance improvement project), a program at an Australian hospital [221] showed a reduction in adverse events over an 8-year period, attributable to an intensive system for incident reporting and conducting root cause analyses of serious events.

No intervention study has examined the impact of autopsy-detected errors in clinical diagnosis on subsequent clinical performance. Given the absence of any studies of the impact of postmortem findings, the current literature provides no direct evidence for or against an impact of autopsy-detected errors on performance improvement at the level of individual practitioners or institutions. This is not to say that there is no valid role for the autopsy in relation to clinical practice or performance improvement, but instead reveals that the current literature is insufficient to address these issues.

Dissemination of findings from studies of autopsy-detected errors may, however, exert an impact on clinical practice at a broader level. While such an impact is plausible, the existing literature does not lend itself to analyzing trends in specific misdiagnoses, and certainly does not allow for inferences about causality. For instance, cirrhosis, bacterial pneumonia and tumors were commonly missed by clinicians and thus detected as major errors in autopsy studies prior to 1965. [51, 145] Subsequent studies were more likely to reveal previously uncommon diagnoses such as systemic fungal infections and pulmonary emboli,[51, 145] suggesting some improvements in clinical practice. Similarly, even though specific tumor misdiagnoses continue to occur,[222]the recognition of the presence of a malignancy has almost certainly improved over time.[51, 145, 223] These improvements, though difficult to document with existing data, plausibly result from general dissemination of the results of autopsies in the medical literature, even if local impacts remain unclear.

In the event that some connection could be shown between the identification of diagnostic errors at autopsy and subsequent performance, it is unclear how best to utilize the autopsy as a performance improvement tool at the institutional level. Discussions in the literature have generally assumed that the persistence of significant rates of diagnostic errors at autopsy implies that autopsy rates must be increased. However, the ability to measure the benefit of investing resources in such efforts is currently not feasible. Studies would be needed to establish optimal ways of utilizing the information generated by autopsies as currently performed or, at the least, to document the effectiveness of current methods.

Use of the Autopsy as an Institutional Performance Measure

Putting aside the possibility of improvement and focusing only on measurement of diagnostic performance still presents logistical problems related to statistical power. For instance, one might consider using institutional Class I error rates as a benchmark for clinical diagnostic performance. As shown in Appendix Table 2, even studies including 100 autopsies have fairly wide confidence intervals. Achieving this number of autopsies generally required one year at a high autopsy rate or several years with an autopsy rate in the 20-30% range. Since most of the hospitals in these studies are medium to large teaching hospitals, the time required for community non-teaching hospitals would almost certainly be longer, as these hospitals tend to be smaller and to have lower autopsy rates.

Taking into account overall diagnostic performance does not substantially improve the power to detect differences in diagnostic performance. Consider, for instance, a relatively large hospital with 20,000 admissions per year.[4] Based on national discharge data,[224] this hospital would have a crude mortality rate of 2%, so the total number of deaths would be 400. Even, if we attribute to this hospital a reasonably high (by contemporary standards) autopsy rate of 25%, so that 100 autopsies would be performed in one year, only we would expect to observe 5 Class I errors among these autopsies (using the regression model to predict Class I error rate for a U.S. institution with autopsy rate = 0.25 and year = 2000, as shown in Appendix B), for an error rate of 5% (95% CI: 1.9-11.8%).

Error rates for specific diagnoses have even wider confidence intervals. Suppose the above 5 Class I errors included two missed cases of major PE. Based on national hospital discharge data,[224] we can estimate a hospital with 20,000 annual discharges[e] to include 40 cases of clinically diagnosed PE. Putting aside the complicating issue of false positive clinical diagnoses, we would thus estimate clinical diagnosis as having false negative rate of 2/42 or 4.8% for the detection of PE, with a 95% confidence interval extending from 0.8% to 17.4%. Taking into account the possibility of missed cases among non-autopsied death widens the confidence interval even further. As an approximation, we can regard the non-autopsied deaths as including the same proportion of Class I errors as the autopsied deaths. Thus, with a 25% autopsy rate, there might be a total of 4*2=8 missed PEs. The upper limit for the 95% CI for the error rate of 8/48 would then be 31%.

Autopsy Detected Diagnostic Errors at the Health System Level

The previous sections have outlined the weak connection between autopsy-detected errors and overall clinical performance and the lack of any evidence for an impact of autopsy-detected errors on performance improvement. The prevalence of autopsy-detected errors in clinical diagnosis do, however, have major implications for vital statistics and other epidemiologic data, and these are discussed below.

---

[4] A hospital of this size would be just below the top 5% in terms of patient volume, as the 95th percentile for annual hospital discharges in 1997 began at 21,556 (based on data available from 2,349 hospitals in 19 states in the Hospital Costs and Utilization Project[224]). The median number of discharges in 1997 was 4,818, and 75% of hospitals had fewer than 10,000 admissions. Thus, the confidence intervals for most hospitals would be even wider than those shown here.

Accuracy of Vital Statistics

Vital statistics such as mortality data provide a fundamental source of death information for different demographic groups (by age, gender, race, geographic area) across the country. One of the few sources of such data are the vital statistics collected by the National Center for Health Statistics.[225] These data permit characterization of the leading causes of death, calculation of life expectancy, and comparisons of mortality trends with other countries.

The National Vital Statistics System compiles these national mortality data from death certificates. In the United States, state laws require completion of death certificates for all deaths, and federal law mandates national collection and publication of deaths and other vital statistics data.[226] Importantly, though, these activities all depend on the accuracy of death certificates.

Numerous studies from the United States,[14, 15, 114, 227-235] Canada,[11, 236, 237] Europe,[12, 16, 238-247] Australia,[248-250] New Zealand[251, 252] and elsewhere[253] document a high prevalence of inaccurate diagnoses on death certificates, using autopsy findings (sometimes supplemented by clinical records) as the diagnostic standard. These studies indicate that positive predictive values for causes of death listed on certificates range from values as low as 25% for some disease categories to a high of approximately 90%. [14, 228, 229, 232, 233, 254] Even without these studies, the results of the many studies of autopsy-detected errors in clinical diagnosis imply major inaccuracies for death certificates.[5]

Death certificates primarily reflect clinical diagnoses, because the majority of deaths do not undergo autopsy and because death certificates are typically completed before autopsy results become available and are seldom corrected in light of postmortem findings.[14, 254] Thus, vital statistics based on death certificates contain all of the errors in clinical diagnosis reviewed in the literature on autopsy-detected diagnostic errors. In fact, because the magnitude of these errors remains independent of impacts on prognosis, death certificates would be expected to have errors in the range we found for major errors, which in the regression model ranged from 8.1-23.8% in the year 2000, depending on the autopsy rate (Appendix Figure 8). Variations in coding of death certificates introduce additional inaccuracies,[14, 15, 110, 114, 227-229, 231-235] further increasing the error rate for death certificates. Clinicians unfamiliar with protocols for assigning causes of death undoubtedly contribute to much of this variation, but significant variation occurs even among pathologists and medical examiners.[114]

Several studies have suggested that errors in death certificates roughly cancel each other out. In other words, for a given disease category (e.g., infectious, neoplastic, cardiovascular), the false positives approximate the false negatives, resulting in little net change in major causes of death at a population level.[14, 15, 254] However, the specific disease categories that remain relatively unchanged are not the same across studies, and for some disease categories the changes have been large. Moreover, for most epidemiologic purposes, more than a major disease category is required (e.g., emphysema or myocardial infarction, not simply death due to respiratory disease or cardiovascular disease).

---

[5] Autopsy-detected errors also affect other major sources of national mortality data, such as the Nationwide Inpatient Sample.[224] These databases consist of discharge diagnoses for hospitalized patients, and therefore reflect the diagnoses contained in clinical records. Differences in coding practices for death certificates and hospital discharge records create discrepancies between these two sources of causes of death.[234] Moreover, neither source can exceed the accuracy of the clinical diagnoses they attempt to capture, which, as discussed above, have major errors in 20-30% of cases (Appendix Table 3).

Significance of inaccuracies in vital statistics

Mortality data, such as those contained in death certificates[225, 226] or large administrative databases,[224, 255] are commonly used in outcomes research.[256-266] Moreover, major decisions related to the allocation of healthcare research funds derive in part from estimates of disease burden.[267-271] The degree of correlation between disease burden and research funding depends on the specific measure—prevalence, incidence, mortality, disability-adjusted life-years and economic costs.[269, 271] Nonetheless, all of these measures depend on accurate vital statistics and other epidemiologic data derived from clinical diagnoses, both of which are known to contain major inaccuracies.

Given the prevalence of major errors in clinical diagnoses among deceased patients (Appendix Table 3), improvements in autopsy rates would be expected to produce substantial improvements in mortality data. These improvements would provide multiple tangible benefits to researchers and funding agencies.

Use of the Autopsy as a Surveillance Tool

One epidemiologic application of the autopsy consists of the use of incidental findings or 'necropsy surprises' in routine autopsies to gauge the prevalence of important chronic diseases. For instance, patients who die of conditions unrelated to gallstones or an abdominal aortic aneurysm provide a quasi-random sample of the general population from the perspective of attempting to measure the prevalence of these conditions. This approach to using the autopsy as an epidemiologic tool has been applied to diseases of the biliary tract and abdominal aorta, as well as several common forms of cancer.[17-19, 21, 22, 272, 273, 274]

Another epidemiologic application of the autopsy is in identifying or helping to characterize new and emerging diseases. Prominent examples of this role of the autopsy have consisted of infectious diseases,[275] such as Legionnaire's disease, AIDS, and pulmonary hantavirus.[23-26, 276-281] Outbreaks, such as the West Nile virus epidemic, and, more recently the use of anthrax a biological weapon, emphasize the importance of surveillance. Routine autopsies may detect cases that would otherwise escape investigation.

Table 1 – Sensitivity of Clinical Diagnosis Comparing Theoretic True Diagnosis with Autopsy Results as the Diagnostic Standard

		"True Diagnosis" (Observed Autopsy Diagnosis)		
		Lung Cancer	**Other Major Diagnosis**	**Total**
Clinical Diagnosis	**Lung Cancer**	180 (182)	90 (88)	270 (270)
	Other Major Diagnosis	20 (54)	1710 (1676)	1730 (1730)
	Total	200 (236)	1800 (1764)	2000 (2000)

Adapted from Saracci[58]

Table 2 – Inter-rater Agreement for Pathologists Assessing the Presence or Absence of Pneumonia at Autopsy

Observer pair	Observed Kappa (95% CI)*	Maximum possible Kappa†
A,B	0.72 (0.49-0.95)	0.83
A,C	0.65 (0.39-0.91)	0.65
A,D	0.52 (0.22-0.82)	0.52
B,C	0.81 (0.59-1.0)	0.81
B,D	0.66 (0.38-0.94)	0.66
C,D	0.84 (0.63-1.0)	0.84
D,D (6 months later)	0.82 (0.59-1.0)	1.0

* Cohen's kappa,[112, 113] with 95% confidence-interval limits calculated according to the Wilson efficient-score method, corrected for continuity

† Maximum possible kappa, given the observed marginal frequencies

Table 3 - Mean Error Rates and Correlations with Time, Autopsy Rates

Error Classification	Error rate* (95% CI)	Change in error rate with time**	Change in error rate with increased autopsy rate†
Class I	10.2% (6.7-15.3%)	26.2% relative decrease per decade (p=0.1)	7.8% relative decrease for each 10% increase in autopsy rate (p=0.2)
Major errors	25.6% (95% CI: 20.8-31.2%)	28% relative decrease per decade (p=0.0006)	12% relative decrease for each 10% increase in autopsy rate (p=0.0003)
Discrepant ICD Disease Categories	11.7% (9.7 -13.9%)	28% *increase* per year (p<0.0001)	1.4% decrease for each 5% increase in autopsy rate (p=0.1)

* The error rates listed here represent the probability of a diagnostic error at time zero (1980 for Class I and major errors, 1975 for ICD discrepancies), with the referent case mix category (general inpatients or adult inpatients), reference country (U.S.) and unweighted mean autopsy rate (44.3%) for included studies.

† Changes are all relative to the value for the base year and mean autopsy rate.

Table 4 - Class I Error Rates Computed for Varying Autopsy Rates in 4 Different Years

Autopsy Rate	1970	1980	1990	2000
5%	17.6%	13.6%	10.4%	7.9%
10%	17.0%	13.1%	10.0%	7.6%
20%	15.9%	12.2%	9.3%	7.1%
40%	13.9%	10.6%	8.0%	6.1%
80%	10.4%	7.9%	5.9%	4.5%
100%	9.0%	6.8%	5.1%	3.8%

Table 5 - Major Error Rates Computed for Varying Autopsy Rates in 4 Different Years

Autopsy Rate	1970	1980	1990	2000
5%	44.3%	36.4%	29.2%	22.9%
10%	42.7%	34.9%	27.9%	21.8%
20%	39.6%	32.1%	25.4%	19.7%
40%	33.7%	26.8%	20.8%	15.9%
80%	23.3%	17.9%	13.6%	10.2%
100%	19.0%	14.5%	10.8%	8.1%

Table 6 – Autopsy Rate Stratified by Formally Assessed Levels of Diagnostic Certainty for Deaths in a Neonatal Intensive Care Unit (as reported by Dhar et al [124])

Diagnostic Certainty	Autopsy Rate (95% CI)
Completely Certain (Level 1)	53.1% (41.7-64.1%)
Almost Certain (Level 2)	61.3% (56.2-66.2%)
Probable or suspected (Level 3)	73% (62.4-81.6%)

Table 7 – Autopsy Rate and Diagnostic Disagreements Stratified by Clinical Confidence (from Heasman and Lipworth[127])

Clinical Confidence	Total Number of deaths	Autopsy rate	Disagreement rate
Fairly certain	9,248	57%	16%
Probable	3,694	76%	33%
Uncertain	1,282	88%	50%
Not stated	393	67%	28%
Total	14,617	65%	25%

Table 8 – Number of Autopsies and Percent Diagnostic Errors Stratified by Clinical Confidence and Patient Age (from Britton[121, 122])

Clinical Diagnosis	Age < 70	Age≥ 70	Total
Fairly certain - # (% erroneous)	91 (15%)	91 (34%)	182 (25%)
Probable - # (% erroneous)	60 (33%)	91 (53%)	151 (45%)
Total - # (% erroneous)	151 (23%)	182 (43%)	333 (34%)

Table 9 – Autopsy Rates and Diagnostic Confirmation Rates Stratified by Clinical Confidence (from Cameron and McGoogan[155, 156])

Clinicians' Assessments of Principal Diagnosis	Percent of Cases	Confirmation of Principal Diagnosis at Autopsy
Fairly certain	47%	75%
Probable	35%	55%
Uncertain	16%	30%

Table 10 – Diagnostic Confirmation Rates Stratified by Clinical Confidence Before and After Intervention to Increase Autopsy Rate (from Cameron et al[123])

	Number of cases (Percent of cases with autopsy confirmation of principal clinical diagnosis %)	
	5/75-6/77 (Autopsy rate 30%)	6 months in 1978 (Autopsy Rate 63%)
Total # of autopsies	326	154
Main diagnosis "certain" or "fairly certain"	168 (52%)	144 (94%)
Autopsy confirmation of main diagnosis	182 (56%)	131 (85%)
Autopsy confirmation of causes of death	60 (18%)	90 (58%)
Autopsy would normally NOT be requested	-----	44 (86%)
Autopsy would normally be requested	-----	110 (85%)

Table 11 – Clinicians Assessment of Necessity in Cases Sent for Autopsy and Stratified by Clinical Confidence (from Hartveit[293])

Confidence in clinical diagnosis	Clinicians' Assessments of Autopsy Necessity		
	Essential	Desirable	Little Interest
Certain	22%	70%	8%
Uncertain	45%	54%	1%

Table 12 – Diagnostic Errors Detected by Autopsy in the Single Study Including Discharge Diagnoses for Patients who did not Die (from Cameron & McGoogan[155,156])

	Principal diagnoses and causes of death as determined by autopsy in 1975-7			Principal discharge diagnoses from same hospitals in 1976	
	Confirmed	Overturned	Missed	Alive	Died
Pulmonary TB	7	8	7	138	5
PE	44	35	99	140	18
Acute appendicitis	3	1	1	441	4
Bowel obstruction	6	3	0	159	13
Acute MI	198	58	51	1452	258
Cirrhosis	22	5	13	125	12

TB – tuberculosis; PE – pulmonary embolism; MI – myocardial infarction