Cancer Treatment and Research Communications 25 (2020) 100253

Contents lists available at ScienceDirect

Cancer Treatment and Research Communications

journal homepage: www.sciencedirect.com/journal/cancer-treatment-and-research-communications

# Simplified, standardized methods to assess the accuracy of clinical cancer staging

Dolly Y. Wu [a,*], Ann E. Spangler [b], Dat T. Vo [b], Alberto de Hoyos [c], Stephen J. Seiler [d]

[a] Volunteer Services, University of Texas Southwestern Medical Center, Dallas, TX, United States; California Institute of Technology, Pasadena, CA, United States
[b] Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, United States
[c] Department of Thoracic Surgery, University of Texas Southwestern Medical Center, Dallas, TX, United States
[d] Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX, United States

## ARTICLE INFO

*Keywords:*
Cancer staging
Accuracy
Monitoring methods
Quality
Lung cancer
Registry data
SEER

## ABSTRACT

*Background:* Hospitals lack intuitive methods to monitor their accuracy of clinical cancer staging, which is critical to treatment planning, prognosis, refinements, and registering quality data.

*Methods:* We introduce a tabulation framework to compare clinical staging with the reference-standard pathological staging, and quantify systematic errors. As an example, we analyzed 9,644 2016 U.S. National Cancer Institute SEER surgically-treated non-small cell lung cancer (NSCLC) cases, and computed concordance with different denominators to compare with incompatible past results.

*Results:* The concordance for clinical versus pathological lymph node N-stage is very good, 83.4 ± 1.0%, but the tumor length-location T-stage is only 58.1 ± 0.9%. There are intuitive insights to the causes of discordance. Approximately 29% of the cases are pathological T-stage greater than clinical T-stage, and 12% lower than the clinical T-stage, which is due partly to the fact that surgically-treated NSCLC are typically lower-stage cancer cases, which results in a bounded higher probability for pathological upstaging. Individual T-stage categories Tis, T1a, T1b, T2a, T2b, T3, T4 invariant percent-concordances are 85.2 ± 9.7 + 10.3%; 72.7 ± 1.6 + 11.3%; 46.6 ± 1.8 + 10.9%; 54.6 ± 1.6 – 20.5%; 41.6 ± 3.3 – 0.1%; 54.7 ± 2.8 – 24.1%; 55.2 ± 4.7 + 2.6%, respectively. Each percent-concordance is referenced to an averaged number of pathological and clinical cases. The first error number quantifies statistical fluctuations; the second quantifies clinical and pathological staging biases. Lastly, comparison of over and under staging versus clinical characteristics provides further insights.

*Conclusions:* Clinical NSCLC staging accuracy and concordance with pathological values can improve. As a first step, the framework enables standardizing comparing staging results and detecting possible problem areas. Cancer hospitals and registries can implement the efficient framework to monitor staging accuracy.

*Abbreviations*

| | |
|---|---|
| SEER | U.S.A. National Cancer Institute Surveillance, Epidemiology, and End Results Program |
| NSCLC | non-small cell lung cancer |
| cTNM, pTNM | clinical and pathological TNM anatomic group stage of the cancer, respectively |
| cT, pT | T stage of the cancer based on clinical and pathological evaluation, respectively |
| *i* | generic index referring to any one of the stage categories or subcategories |
| cTNM-*i* | category i of the clinical group stage, e.g. cIA, cIB, cIIA, cIIB, etc. |
| pTNM-*i* | category i of the pathological group stage, e.g. pIA, pIB, pIIA, pIIB, etc. |
| cT-*i* | category i of the clinical T stage, e.g. cTx, cTis, cT1a, etc. |
| pT-*i* | category i of the pathological T stage, e.g. pTx, pTis, pT1a, etc. |
| TNM7, TNM8 | 7th and 8th editions of cancer staging manuals by American Joint Committee on Cancer (AJCC) and Union for International Cancer Control (UICC) |
| CT | Computed tomography scan |
| PET | Positron emission tomography scan |
| EMR | patient's electronic medical records |

* Corresponding author.
*E-mail addresses:* dolly.wu@utsouthwestern.edu, dollywu@alumni.caltech.edu (D.Y. Wu).

https://doi.org/10.1016/j.ctarc.2020.100253

Available online 1 December 2020

## Introduction

Cancer stage describes the spread and severity of a patient's primary tumor; staging accuracy is clearly necessary. Treating physicians perform the clinical cancer staging that guides initial treatment planning, prognosis, monitoring and future treatment refinement. Among cancers, lung cancer is the leading cause of cancer deaths. For non-small cell lung cancer (NSCLC), pathological staging and tumor size measurements have historically been considered more accurate and are the reference standard by which the clinical values are compared [1-6]. However, 80% of NSCLC are not treated surgically and do not have pathological staging so that accurate clinical assessment, alone, is critical. Yet, hospital staff lack intuitive, simple analysis tools they can use on hospital data to assess the quality, and to quantify and detect any inaccuracies of clinical staging at their facility.

Previously, we published a tabulation framework to quantify concordance [7]. Our current objectives are to explain and expand the techniques, and present a more-universal way to report numerical results for "accuracy" (i.e. clinical and pathological staging concordance) to avoid an ambiguity that traditionally existed. Additionally, we introduce definitions to quantify systematic errors that cause discordance so that hospitals can track their progress if they adjust their staging techniques. We provide "how-to" details, examples, and explain the arithmetic motivation for the framework. Further, if cancer hospitals report cancer staging results using such methods, it will be easier to compare study results or, alternatively, aggregate results among hospitals or studies to decrease statistical errors to better understand cancer or clinical practice trends. Cancer registrars may also use the techniques to expeditiously monitor staging data quality.

The framework applies to any component of TNM-stage and also to other types of cancers, but we focused on NSCLC T-stage, which measures the greatest dimension ("length") of the primary tumor and its location-invasion properties, partly because N-staging accuracy issues have been studied already [1, 2]. In addition, T-stage categories have been redefined in the staging manuals, whereas the N-stage categories have remained the same. Clinical T-stage is usually determined from computed tomography images (CT). T-stage affects treatment options, surgical methods, and validates imaging algorithms for tumor delineation, detection and measurements. T-stage category cutoffs have been continuously refined by AJCC and UICC over the decades, to continually refine future prognosis and treatment.

The U.S. National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) recently made available a 2016 TNM7 dataset that includes more data fields. The methods presented herein based on the seventh-edition TNM7-staging are also applicable to the present eighth-edition TNM8-staging [8-11]. More importantly, individual hospitals wishing to benchmark their own results against SEER's are more likely to have sufficient surgically-treated NSCLC number of cases for TNM7 than TNM8. Our updated numerical TNM7 results reflect our new definitions and also additional data-selection criteria.

## Materials and methods

In 1990 in Reference [38], staging concordance was assessed by tabulating the number of cases for each stage category in a matrix format, pathological versus clinical cancer stage. Their matrix also presented percent-agreement for each stage category, but their study had only 103 total cases such that the results would have had very large statistical fluctuation errors had the errors been reported. After 1990, some studies turned to or included Kappa index to quantify agreement, which required statisticians to help calculate or interpret the significance of the results. We instead build on the 1990 approach because the arithmetic is simple and efficient, and accessible to hospital staff who are not statisticians.

We also expand the original approach and quantify statistical and systematic errors. For example, statistically, 103 cases are small and subsequent study-participant results may fluctuate. As for systematic errors, they are also critical in assessing the significance of study results. They represent persistent, repeatable errors associated with some bias, such as measuring tumor size with an overstretched ruler. With respect to cancer results, the magnitude of the statistical and systematic errors indicates whether the results are measurably significant, and thus clinically meaningful. In addition, quoting the errors provides a common terminology to compare results from different studies to determine whether they are significantly different or actually the same within the margin of errors.

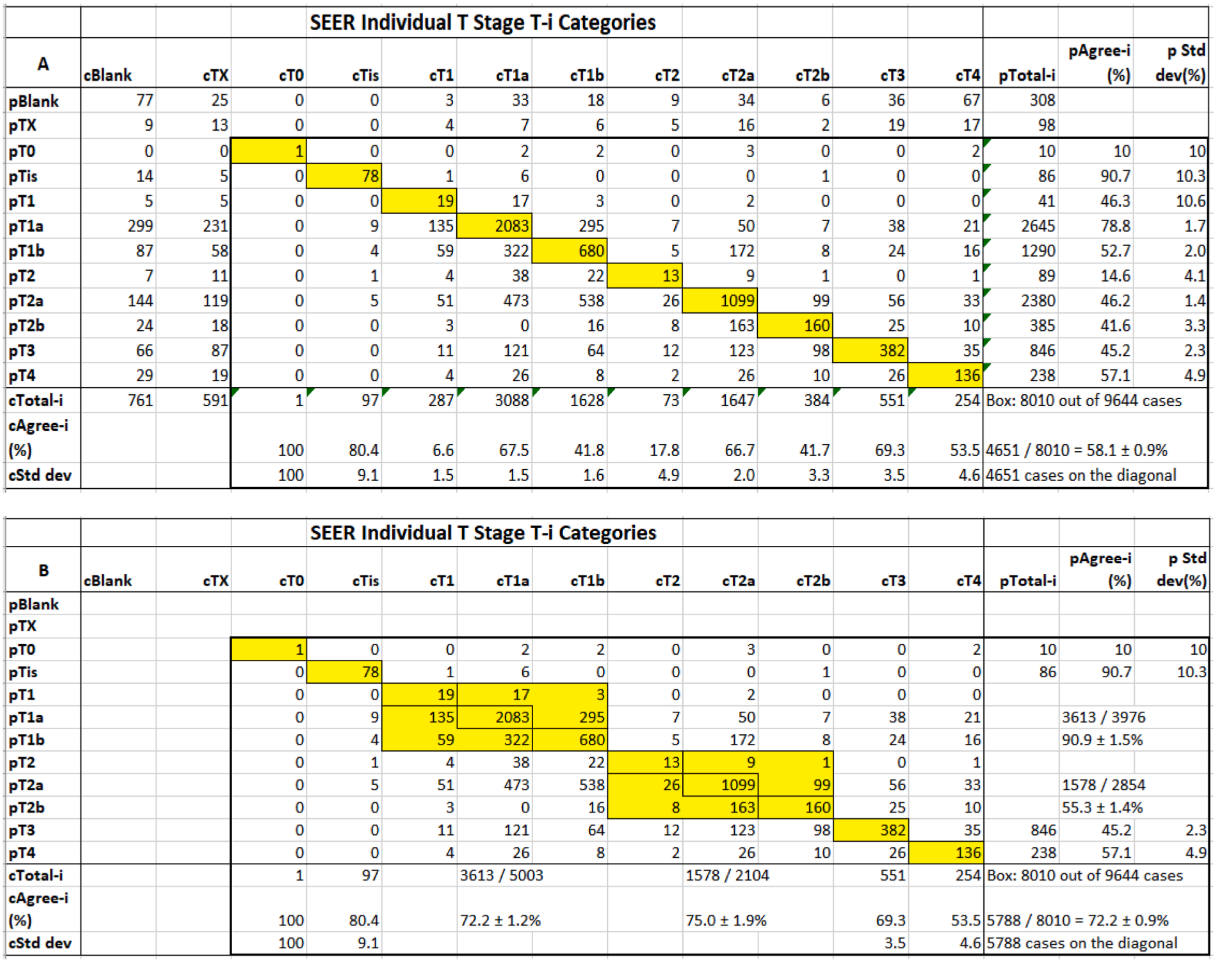
Our "tabulation" framework comprises tabulating the number of cancer cases in different categories (stage or tumor characteristics), quantifying percentage of stage agreement, and statistical and systematic errors, and looking for patterns or spikes in the tables. The arithmetic consists only of basic functions (i.e. addition, division, square-roots) and the simple equations may be coded as entries in certain cells of an Excel spreadsheet template, equations which may be repeatedly re-used by entering new case numbers in the data fields of the template. Fig. 1 and Table 3 are the two example Excel templates used in this article, the former to determine stage concordance, and the latter to correlate staging concordance with different patient/tumor characteristics.

We applied the framework to SEER's NSCLC cases as an example. We retrospectively analyzed TNM7 data for benchmarking purposes because large, curated TNM8 datasets are not yet available, but the tumor length and N-staging results are applicable to either staging system. We required resection as the initial treatment, without neoadjuvant therapy [7]. For the present analysis, we added a delay-to-surgery requirement of less than four months, a choice that is based on the staging guidelines diagnosis requirements [8, 9]. Four months also minimizes tumor growth biases [12]. The selection criteria yielded 9, 644 cases. Less than 2% were metastatic. Appendix Fig. S1 shows the distributions for histology, grade, laterality, primary site, age, and stage for the selected cases.

To compare clinical with pathological staging, we used percent-concordance (percent-agreement) to quantify clinical accuracy for each stage category, by tabulating the "n" number of cases in each clinical versus pathological stage category "i" in matrices, Fig. 1. The diagonal elements of the matrices are the number of cases where the pathology and clinical stage values agree. One issue is the choice of denominator to compute agreement, whether the percent-concordance is referenced to the total number of clinical cases in the category ("cTotal-i"), or to the total number of pathological cases instead ("pTotal-i"). The choice is different or unstated in publications, which causes ambiguity, and *potentially large discrepancy among different studies* [1, 3, 6, 13, 38]. For example, the percent-concordance for T1a in Fig. 1A is either pathological 78.8 ± 1.7% or clinical 67.5 ± 1.5%, which disagree by >3 standard deviations due to different denominators. Initially, the choice used by lung cancer staging advisory members (e.g. Goldstraw) was with respect to the pathological cases, pTotal-i [3, 38], but subsequent studies sometimes used the other choice (clinical staging) so that studies are not comparable, "apples-to-apples". Nevertheless, calculating percent-concordance values for both choices allows comparison with either type of past study results.

To avoid the ambiguity of one choice of denominator or another, we introduce an universal value and define an "invariant concordance" percentage by using the average of the pathological and clinical cases (avg(pTotal-i, cTotal-i)) as the denominator in the percent-concordance for each stage category i. For example for T1a, the invariant percent-concordance is 2083/ (3088+2645)/2 = 72.7 ± 1.6% (see the numbers in Fig. 1A). However, we summarize the results for all three choices (Table 2).

The statistical error in these percent-concordances is approximated by a binomial distribution for large n-number discrete cases (Poisson distribution) so that for calculation simplicity, the quoted standard deviation is taken as square-root(n) [14]. The concordance and



|  | SEER Individual T Stage T-i Categories |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | cBlank | cTX | cT0 | cTis | cT1 | cT1a | cT1b | cT2 | cT2a | cT2b | cT3 | cT4 | pTotal-i | pAgree-i (%) | p Std dev(%) |
| pBlank | 77 | 25 | 0 | 0 | 3 | 33 | 18 | 9 | 34 | 6 | 36 | 67 | 308 |  |  |
| pTX | 9 | 13 | 0 | 0 | 4 | 7 | 6 | 5 | 16 | 2 | 19 | 17 | 98 |  |  |
| pT0 | 0 | 0 | 1 | 0 | 0 | 2 | 2 | 0 | 3 | 0 | 0 | 2 | 10 | 10 | 10 |
| pTis | 14 | 5 | 0 | 78 | 1 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 86 | 90.7 | 10.3 |
| pT1 | 5 | 5 | 0 | 0 | 19 | 17 | 3 | 0 | 2 | 0 | 0 | 0 | 41 | 46.3 | 10.6 |
| pT1a | 299 | 231 | 0 | 9 | 135 | 2083 | 295 | 7 | 50 | 7 | 38 | 21 | 2645 | 78.8 | 1.7 |
| pT1b | 87 | 58 | 0 | 4 | 59 | 322 | 680 | 5 | 172 | 8 | 24 | 16 | 1290 | 52.7 | 2.0 |
| pT2 | 7 | 11 | 0 | 1 | 4 | 38 | 22 | 13 | 9 | 1 | 0 | 1 | 89 | 14.6 | 4.1 |
| pT2a | 144 | 119 | 0 | 5 | 51 | 473 | 538 | 26 | 1099 | 99 | 56 | 33 | 2380 | 46.2 | 1.4 |
| pT2b | 24 | 18 | 0 | 0 | 3 | 0 | 16 | 8 | 163 | 160 | 25 | 10 | 385 | 41.6 | 3.3 |
| pT3 | 66 | 87 | 0 | 0 | 11 | 121 | 64 | 12 | 123 | 98 | 382 | 35 | 846 | 45.2 | 2.3 |
| pT4 | 29 | 19 | 0 | 0 | 4 | 26 | 8 | 2 | 26 | 10 | 26 | 136 | 238 | 57.1 | 4.9 |
| cTotal-i | 761 | 591 | 1 | 97 | 287 | 3088 | 1628 | 73 | 1647 | 384 | 551 | 254 | Box: 8010 out of 9644 cases |  |  |
| cAgree-i (%) |  |  | 100 | 80.4 | 6.6 | 67.5 | 41.8 | 17.8 | 66.7 | 41.7 | 69.3 | 53.5 | 4651 / 8010 = 58.1 ± 0.9% |  |  |
| cStd dev |  |  | 100 | 9.1 | 1.5 | 1.5 | 1.6 | 4.9 | 2.0 | 3.3 | 3.5 | 4.6 | 4651 cases on the diagonal |  |  |

|  | SEER Individual T Stage T-i Categories |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | cBlank | cTX | cT0 | cTis | cT1 | cT1a | cT1b | cT2 | cT2a | cT2b | cT3 | cT4 | pTotal-i | pAgree-i (%) | p Std dev(%) |
| pBlank |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| pTX |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| pT0 |  |  | 1 | 0 | 0 | 2 | 2 | 0 | 3 | 0 | 0 | 2 | 10 | 10 | 10 |
| pTis |  |  | 0 | 78 | 1 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 86 | 90.7 | 10.3 |
| pT1 |  |  | 0 | 0 | 19 | 17 | 3 | 0 | 2 | 0 | 0 | 0 |  |  |  |
| pT1a |  |  | 0 | 9 | 135 | 2083 | 295 | 7 | 50 | 7 | 38 | 21 |  | 3613 / 3976 |  |
| pT1b |  |  | 0 | 4 | 59 | 322 | 680 | 5 | 172 | 8 | 24 | 16 |  | 90.9 ± 1.5% |  |
| pT2 |  |  | 0 | 1 | 4 | 38 | 22 | 13 | 9 | 1 | 0 | 1 |  |  |  |
| pT2a |  |  | 0 | 5 | 51 | 473 | 538 | 26 | 1099 | 99 | 56 | 33 |  | 1578 / 2854 |  |
| pT2b |  |  | 0 | 0 | 3 | 0 | 16 | 8 | 163 | 160 | 25 | 10 |  | 55.3 ± 1.4% |  |
| pT3 |  |  | 0 | 0 | 11 | 121 | 64 | 12 | 123 | 98 | 382 | 35 | 846 | 45.2 | 2.3 |
| pT4 |  |  | 0 | 0 | 4 | 26 | 8 | 2 | 26 | 10 | 26 | 136 | 238 | 57.1 | 4.9 |
| cTotal-i |  |  | 1 | 97 |  | 3613 / 5003 |  |  | 1578 / 2104 |  | 551 | 254 | Box: 8010 out of 9644 cases |  |  |
| cAgree-i (%) |  |  | 100 | 80.4 |  | 72.2 ± 1.2% |  |  | 75.0 ± 1.9% |  | 69.3 | 53.5 | 5788 / 8010 = 72.2 ± 0.9% |  |  |
| cStd dev |  |  | 100 | 9.1 |  |  |  |  |  |  | 3.5 | 4.6 | 5788 cases on the diagonal |  |  |

**Fig. 1.** A) Comparison of the clinical vs pathological T-stage, under TNM7. Excluding "Blank" (no data) and TX categories, there remain 8,010 cases. The highlighted diagonal cases are when the clinical and pathological T-stage category classification agree, cT-*i* = pT-i, where "i" is a generic index referring to each stage category. The pTotal-i and cTotal-i sum number of cases in each row or column, respectively, do not include the "Blank" cases. The Agree-i% is the percent-concordance for each category i. The "Std dev" is the statistical error. B) The T1 subcategory cases are aggregated and T2 subcategory cases are aggregated, and each aggregate is treated as one category, as highlighted. See text.

corresponding errors are all expressed as percentages in this article. For instance, the 1.6% statistical error associated with the 72.7% results from square-root(2083) / (3088+2645)/2 × 100 (see the numbers in Fig. 1A).

Results are more meaningful if they also include an estimated systematic error along with the statistical error. For example, rigorous results in physics report a combined, or separately, statistical and systematic errors because both types of errors constitute a measure of whether the results are significant [15]. We summarized sources of potential clinical biases and staging categorization biases in Reference [7], but there also exist pathological biases (Table 1) that may systematically increase or decrease the final T-stage category assigned to the cancer. For instance, formalin fixation tends to shrink some types of tumor tissue so that the pathologically-measured tumor length would be systematically smaller on average and the corresponding T-stage may be lower [16, 17]. To capture the collective effect of all systematic biases in causing clinical-pathological staging discordance, we define an aggregated systematic error as the difference between the clinical and pathological percent-concordances (Agree-i%) for each category i that otherwise theoretically or ideally would have agreed. As an example from the values in Fig. 1A, the aggregated agreement difference, pT1a – cT1a is 78.8 – 67.5% = 11.3%, where we arbitrarily used the pathological value first in the difference. The invariant concordance may now be expressed with both a statistical and systematic error as 72.7 ± 1.6 + 11.3%. The plus sign in front of 11.3 compactly indicates the pathological Agree-i% value is larger than the clinical Agree-i% value for category i, whereas a minus sign means the pathological Agree-i% is smaller. A large systematic error means a large discrepancy between cT-i and pT-i. This quantifies discrepancies for each stage category. Previously, we evaluated clinical-pathological staging concordance using the linear-weighted Cohen-kappa index [7,18-22], and we continue to report the k results herein. However, there is not an easy way to express statistical and systematic errors separately for Kappa index so that it is unclear as to meaning of the Kappa index results.

To correlate staging discordance with cancer characteristics, we performed ordinal-output multivariate logistic regression and likelihood ratio tests for categorical and linear characteristics in Reference [7]. Now, we instead use the three ordinal-outputs to detect trends versus staging concordance, when pathology T-stage is higher than, equal to or lower than the corresponding clinical T-stage. Hospitals can tabulate the percentage of occurrence for each of the cancer characteristics versus each ordinal-output in a spreadsheet-template (e.g. Table 3). This is a simple way to detect correlations with concordance by tabulating a number or percentage of occurrences in spreadsheet templates.

Another useful comparison is the concordance between the pathological and clinical tumor dimension measurements to help crosscheck the T-staging results. In Reference [7], we computed a few ways to check the concordance of tumor dimension as measured pathologically after

**Table 1**
Potential pathological issues that may cause the measured and "true" greatest tumor dimension to disagree [12, 16, 17, 25-32]. Clinical measurement biases are summarized in [7].

| Measuring tumor length | Tumor may appear shorter than actual | Tumor may appear longer than actual | Size may appear shorter or longer than actual |
|---|---|---|---|
| **Pathological** | Not including spiculation or invasion | Gross tumor examination slicing interval, 2–3 or 5 mm | Physician or assistant variability, or rotation of residents |
| | Biopsy removed some of the tissue; more lymph nodes examined | Lobectomy, pneumonectomy or large amount of resected material | Problem delineating boundary of tumor: multi-nodules, multi-foci cases, positive resection margins |
| | Formalin fixation | Large tumors | Rounding or estimating |
| | Extracorporeal tumor specimen | Irregularly-shaped tumor | Vision bias: focusing on bold reticles on the ruler |
| | Solid portion only | Inflammation, edema | Positive resection margin or sub-solids |
| | Not finding true longest axis of the tumor, breadloaf slicing orientation not perpendicular to longest axis | Measuring tumor with a flexible ordinary ruler | Stiff ordinary ruler to measure tumor size |
| | | Higher N-stage<br>Higher tumor grade | Epithelial tumors<br>Periodic length spikes at every 5 mm. |

**Table 2**
Summary of concordance from Figs. 1, S2, S3. The first column is the percentage of clinical cases that agree with pathological staging. The second column is the percentage of pathological cases that agree with clinical staging. The last column is the invariant concordance expressed as a percentage.

**A. SEER 2016 cTNM vs pTNM agreement, with respect to different denominators, under TNM7**

| Group TNM-stage i | With respect to the number of clinical cases in category i (columns of Fig. S2) | With respect to the number of pathology cases in category i (rows of Fig. S2) | **Invariant agreement** with respect to the average of the number of clinical and pathology cases for category i |
|---|---|---|---|
| **0** | 82.3 ± 9.3 – 8.5% | 90.8 ± 10.2 + 8.5% | 86.3 ± 9.7 + 8.5% |
| **IA** | 67.6 ± 1.2 – 21.9% | 89.5 ± 1.7 + 21.9% | 77.0 ± 1.4 + 21.9% |
| **IB** | 55.3 ± 2.0 + 12.3% | 43.0 ± 1.6 – 12.3% | 48.4 ± 1.8 – 12.3% |
| **IIA** | 41.1 ± 2.6 + 11.3% | 30.1 ± 1.9 – 11.3% | 34.8 ± 2.2 – 11.3% |
| **IIB** | 57.2 ± 3.7 + 18.4% | 38.8 ± 2.5 – 18.4% | 46.3 ± 3.0 – 18.4% |
| **IIIA** | 49.4 ± 3.1 + 17.0% | 32.4 ± 2.0 – 17.0% | 39.2 ± 2.5 – 17.0% |
| **IIIB** | 24.1 ± 6.7 + 11.0% | 35.1 ± 9.7 – 11.0% | 28.6 ± 7.9 – 11.0% |
| **IV** | 91.8 ± 5.4 + 4.4% | 87.4 ± 5.1 – 4.4% | 89.6 ± 5.2 – 4.4% |

**B. SEER 2016 cT vs pT agreement, with respect to different denominators, under TNM7**

| T-stage i | With respect to clinical cases (columns of Fig. 1A) | With respect to pathology cases (rows of Fig. 1A) | **Invariant agreement** with respect to the average of clinical and pathology cases |
|---|---|---|---|
| **TX** | 2.3 ± 0.6% (exclude data blank cases) | 14.6 ± 4.0% (exclude blanks) | |
| **T0** | 100 ± 100 + 90% | 10 ± 10 – 90% | 18.2 ± 18.2 – 90% |
| **Tis** | 80.4 ± 9.1 – 10.3% | 90.7 ± 10.3 + 10.3% | 85.2 ± 9.7 + 10.3% |
| **T1** | 6.6 ± 1.5 | 46.3 ± 10.6 | n/a |
| **T1a** | 67.5 ± 1.5 – 11.3% | 78.8 ± 1.7 + 11.3% | 72.7 ± 1.6 + 11.3% |
| **T1b** | 41.8 ± 1.6 – 10.9% | 52.7 ± 2.0 + 10.9% | 46.6 ± 1.8 + 10.9% |
| **T2** | 17.8 ± 4.9 | 14.6 ± 4.1 | n/a |
| **T2a** | 66.7 ± 2.0 + 20.5% | 46.2 ± 1.4 – 20.5% | 54.6 ± 1.6 – 20.5% |
| **T2b** | 41.7 ± 3.3 + 0.1% | 41.6 ± 3.3 – 0.1% | 41.6 ± 3.3 – 0.1% |
| **T3** | 69.3 ± 3.5 + 24.1% | 45.2 ± 2.3 – 24.1% | 54.7 ± 2.8 – 24.1% |
| **T4** | 53.5 ± 4.6 – 2.6% | 57.1 ± 4.9 + 2.6% | 55.2 ± 4.7 + 2.6% |

**C. SEER 2016 cN vs pN agreement, with respect to different denominators, under TNM7 or TNM8**

| N-stage i | With respect to clinical cases (columns of Fig. S3) | With respect to pathology cases (rows of Fig. S3) | **Invariant agreement** with respect to the average of clinical and pathology cases |
|---|---|---|---|
| **NX** | 4.6 ± 2.7% (exclude blanks) | 7.2 ± 1.2% (exclude blanks) | |
| **N0** | 86.8 ± 1.1 – 8.9% | 95.7 ± 1.2 + 8.9% | 91.0 ± 1.2 + 8.9% |
| **N1** | 54.3 ± 3.5 + 28.3% | 26.0 ± 1.7 – 28.3% | 35.1 ± 2.3 – 28.3% |
| **N2** | 46.3 ± 3.6 + 16.5% | 29.5 ± 2.3 – 16.5% | 36.0 ± 2.8 – 16.5% |
| **N3** | 12.5 ± 6.3 – 37.5% | 50 ± 25 + 37.5% | 20 ± 10 + 37.5% |

resection versus radiologically prior to resection. For example, for each T-stage size range ("bin"), we compared the pathological value with the clinical value, and tabulated the frequency that the clinical value is above or below the range assuming that the pathological value is the "correct" value. This is another way to assess whether cT-stage agrees with pT-stage, with respect to the tumor size component of T-stage.

## Results

Fig. 1A shows a T-stage cT vs pT concordance of 58.1 ± 0.9% ($k = 0.57$, 95% confidence interval CI: 0.54–0.60) over categories T0–T4. Approximately 29% of the cases are pT > cT, and 12% pT < cT. One possible source of ambiguity in T-staging is due to not subcategorizing cases within the T1 and T2 categories. Since the introduction of TNM7, lung cancer staging guidelines instruct that the subcategories be used in order to refine prognosis and treatment so that ideally, there would be no cases categorized as merely "T1" or "T2". To estimate the effect of not subcategorizing, we aggregated all T1 subcategories and aggregated all T2 sub subcategories (Fig. 1B) which yields a concordance of 72.2 ± 0.9%, which is much better than the 58.1%. The 72.2% can also be used to compare with results before 2010, before TNM7, when there were no subcategories.

Table 2B summarizes the percent-concordances for each T-stage category i, cT-i vs pT-i, with respect to the total number of clinical cases, pathological cases, and their average. The invariant concordance is very good for low T-stages, starting at ~85% but drops to ~40–55% for higher T-stages. For completeness, we summarize the updated SEER numerical results for TNM group stage and N-stage in Table 2 and the Appendix. The group staging, cTNM vs pTNM, concordance over all categories 0–IV, is only 62.4 ± 0.9% ($k = 0.58$, CI: 0.55–0.61). Approximately 30% of the cases are pTNM-stage higher than cTNM-stage, and 10% are lower. The cN vs pN concordance is very good 83.4 ± 1.0% over categories N0–N3 ($k = 0.37$, CI: 0.32–0.42). Approximately 13% of the cases are pN > cN, and 4% pN < cN.

Fig. 1A contains outlier cases occurring away from the diagonal elements and appearing as "spotty" unexpected increases instead of gradual decreases from the peak numbers in the diagonal elements. For instance, the cT1a, cTX, cBlank, cT3 columns in Fig. 1A contain spotty increases that may motivate a hospital to investigate the outliers using their patient records (EMR).

Tables 3-4 tabulate the percentage of cases of certain cancer characteristics for each T-stage concordance category (pT > cT, pT = cT, pT < cT). There is more discordance for larger resections (lobectomy, pneumonectomy), higher pN and cN except N3, higher tumor grade, and certain SEER registry regions (see Appendix). For different SEER regional registries, T-stage concordance ranges from 51.0 ± 7.1% to

**Table 3**
Percentage of cases for different types of surgical resection versus concordance between pathological and clinical T-stage, (pT stage < cT stage), (pT stage = cT stage), (pT stage > cT stage). Other tabulated cancer characteristics are in Appendix Tables S1-S5. The “Std dev” is the statistical error, one standard deviation.

| Surgery Primary Site 8553 cases | pT < cT% | Std dev | pT = cT% | Std dev | pT > cT% | Std dev | No. cases |
|---|---|---|---|---|---|---|---|
| 21 - wedge resection | 10.9 | 0.9 | 63.0 | 2.0 | 26.1 | 1.3 | 1498 |
| 22 - segmentectomy | 8.6 | 1.4 | 62.9 | 3.8 | 28.5 | 2.6 | 428 |
| 30 - lobectomy | 12.4 | 1.4 | 55.5 | 2.9 | 32.0 | 2.2 | 668 |
| 33 - lobectomy mediastinal lymph node dissection | 12.0 | 0.5 | 56.6 | 1.0 | 31.4 | 0.8 | 5519 |
| 45 - lobectomy NOS | 12.8 | 2.8 | 55.5 | 5.8 | 31.7 | 4.4 | 164 |
| 46 - lobectomy chest wall | 8.5 | 4.3 | 44.7 | 9.8 | 46.8 | 10.0 | 47 |
| 55 - pneumonectomy | 10.7 | 6.2 | 39.3 | 11.8 | 50.0 | 13.4 | 28 |
| 56 radical pneumonectomy | 16.4 | 2.9 | 51.7 | 5.1 | 31.8 | 4.0 | 201 |
| Total | | | | | | | 8553 |

**Table 4**
Percentage of cases for different histologies. The percentage is with respect to the total number of cases of a particular histology, which is based on the SEER lung histology groups, https://seer.cancer.gov/tools/mphrules/2007/lung/terms_defs.pdf and NAACCR ICD-O-3 codes. The two largest cancer histologies, adenocarcinoma and squamous cell carcinoma have statistically same results.

| Histology ICD-O-3 SEER 8614 cases | pT < cT% | Std dev | pT = cT% | Std dev | pT > cT% | Std dev | No. cases |
|---|---|---|---|---|---|---|---|
| Malignant neoplasm NOS, 0.4% of 8614 | 5.6 | 3.9 | 58.3 | 12.7 | 36.1 | 10.0 | 36 |
| Adenocarcinoma, 65.6% of 8614 | 11.9 | 0.5 | 57.4 | 1.0 | 30.8 | 0.7 | 5647 |
| Squamous cell, 21.2% | 10.4 | 0.8 | 57.7 | 1.8 | 31.9 | 1.3 | 1829 |
| Adenosquamous, 2% | 12.2 | 2.6 | 49.5 | 5.1 | 38.3 | 4.5 | 188 |
| Large cell carc., 2% | 9.7 | 2.5 | 52.3 | 5.8 | 38.1 | 5.0 | 155 |
| Sarcomatoid, 0.7% | 12.3 | 4.6 | 43.9 | 8.8 | 43.9 | 8.8 | 57 |
| Neuroendocrine Carcinoid tumors NOS, 6.4% of 8614 | 14.4 | 1.6 | 66.5 | 3.5 | 19.1 | 1.9 | 550 |
| General NSCLC, 0.9% | 9.5 | 3.6 | 63.5 | 9.3 | 27.0 | 6.0 | 74 |
| Neuroendocrine NOS, 0.7% of 8614 | 19.6 | 5.9 | 62.5 | 10.6 | 17.9 | 5.6 | 56 |
| Salivary gland-type tumors, 0.3% of 8614 | 31.8 | 12.0 | 45.5 | 14.4 | 22.7 | 10.2 | 22 |
| Total number of cases | | | | | | | 8614 |

71.8 ± 7.1% (Table S5). The concordance is better away from the major T-stage tumor size thresholds for the averaged clinical-pathology tumor length values; pT is more-often upstage of cT right near the T-stage size thresholds. The two largest types of NSCLC, adenocarcinoma and squamous cell carcinoma, exhibit the statistically-same staging concordance (Table 4).

Regarding tumor size, Reference [7] showed, for example, that for pathology tumor length under TNM8, ≤ 1 cm, 51.8 ± 2.8% of the corresponding clinically measured length is over 1 cm, meaning that 51.8% of the clinical cases would be in a higher T-stage category than the pathological T-stage. Between 1 cm – 2 cm, 9.6 ± 0.6% of the clinical size measurements are lower, and 18.1 ± 0.8% are higher. Between 2 cm – 3 cm, 28.4 ± 1.2% of the clinical size measurements are lower, and 15.2 ± 0.9% are higher. These interval ranges coincide with the TNM8 size cutoff values for stage T1. Under TNM7 cutoff values, for pathology tumor length ≤ 2 cm, 15.9 ± 0.7% of the corresponding clinically measured size is over 2 cm, meaning that 15.9% of the clinical cases would be in a higher T-stage category than the pathological T-stage. Between 2 cm – 3 cm, the results are the same as for TNM8. Between 3 cm – 5 cm, 29.8 ± 1.4% of the clinical measurements are lower, and 8.2 ± 0.7% are higher. Between 5 cm – 7 cm, 50.4 ± 3.2% of the clinical measurements are lower, and 6.0 ± 1.1% are higher.

## Discussion

Clinical and pathological T-staging agrees only moderately, and has not improved over pre-TNM7 results, despite improved imaging resolution [1-4, 23, 24]. Table 2B shows large discordances between the clinical and pathological individual T-stage categories, T1a through T2b, which determine the type of resection. At the T1–T2 threshold, Fig. 1B shows >20% discordance, and resection decisions may require additional patient exams. Over all categories, the ~29% pT > cT patients may not have received adequate pre-operative care and monitoring. T-staging discordance also contributes to approximately 78% of the clinical versus pathological discordance in the group TNM-stage results.

Stage I, NSCLC studies point out that clinical staging is understaged due to possible medical or practice issues [1, 13, 25]. Mathematically, there is an additional non-clinical explanation. This phenomenon tends to occur for finitely-categorized measurements due to a bounded probability distribution. For low TNM and T stages, clinical staging will be “understaged” on average; for high stages, clinical staging will be “overstaged” on average when subsequently-performed stage classification disagrees with the initial classification. For low stages, subsequent differing classification can *only* increase because the categories are bounded below (i.e. no more categories). Likewise, for high stages, subsequent differing classification can *only* decrease because it is bounded above. Examining the row and column directions of the matrices reflects the phenomenon. Surgical resection is a treatment modality mainly for lower-stage, non-metastatic NSCLC cases so that resection datasets would probabilistically have more clinical understaging.

The discordance between the number of pathological versus clinical cases for different ranges of tumor length also generally suffers from the bounded probability effect. At first glance, it is seemingly not true for the lowest bin, <1 cm tumor size; however, this is due to the fact that we binned the cases based on the pathology size rather than the clinical size. Had we selected cases binned by clinical size <1 cm, instead, approximately 57.7% of the tumors were subsequently measured longer pathologically, a result which exhibits the bounding effect. Nevertheless, the pathological tumor size is considered the reference standard, and for such small tumors the measurement is believed reasonably accurate because the specimens are within the ~2 cm field of view of the microscope, and the methods of pathological measurements were more standardized than radiological measurements during the 2016 time period when the SEER data was collected. So if we take such tumors as having a “true” size of <1 cm, then ~58.1% are over measured clinically. The large 58.1% discrepancy for clinically-measured tumor sizes for the <1 cm bin has consequences when categorizing tumors under the TNM8 staging system. Patient monitoring is affected when tumors clinically “measure” over 8 mm and patients receive additional tests and scans, which increase costs. There are also treatment consequences. For peripheral tumors measured clinically, below ~1 cm, a non-anatomic lung resection (wedge resection) by video-assisted thoracoscopic surgery may be performed. However, above ~1 cm, more invasive anatomic segmentectomies are performed for deep tumors smaller than ~2 cm, and lobectomies for deep tumors larger than ~2 cm. Alternatively, for early stage T1 or T2 NSCLC, patients may instead be treated by radiotherapy instead of resection with dose-fractions based on tumor size and location. Studies are also underway to investigate new radiotherapies such as stereotactic ablative radiotherapy with immunotherapy. Patients with progressively higher cT-stage NSCLC generally need treatment intensification, including neoadjuvant and adjuvant therapy. For example, for patients not undergoing surgical resection and have no pathological stage assigned to them, those with tumors greater than clinically measured 4 cm will generally receive adjuvant chemotherapy.

The clinical size measurements and corresponding cT-stage category are not only critical in the choice of treatment, but also in patient prognosis. The lung cancer staging committee for TNM8 proposed finer T-stage cutoffs because they found that there can be tumors less than 2 cm with significantly different prognosis [36]. They also pointed out that the smaller tumors could constitute a particular group worthy of further studies regarding growth rate, tumor density, PET scan results, type of resection, alternative non-surgical therapies, molecular characterization, and genetic signatures. However, if the assigned tumor size range and corresponding cT-stage are incorrect, then comparison of patient cohorts is incorrect, which hinders fine-tuning of the best possible management of and treatments for different stratifications of NSCLC and their prognosis.

When the SEER TNM8 data becomes available, the T-stage concordance is expected to worsen under the TNM8 staging system as indicated by the worsening tumor length concordance in the SEER data when the size interval range is reduced (e.g. 2cm-wide interval versus the 1cm-wide intervals). In addition, the concordance is worse right at some of the T-stage tumor size thresholds, which has been attributed to measurement issues such as rounding [26], which is also likely to worsen T-stage concordance under TNM8 because there are now more T-stage categories and thresholds. The accuracy of patient prognosis is worsened by measurements that artificially congregate right around the thresholds. Prognosis is further worsened if the patient monitoring and treatment plans are suboptimal and based on the clinical cancer stage assigned to a patient when borderline-threshold measurements can more easily cause a patient to fall into one stage subcategory as opposed to another one when the bin size range is reduced. TNM7 divided the T1 stage into T1a ($\leq$ 2 cm) and T1b ($>$ 2 cm and $\leq$ 3 cm), whereas TNM8 divides the T1 stage into three one-centimeter intervals, T1a, T1b and T1c. T2 stage under TNM8 has a size cutoff of 5 cm rather than 7 cm, involvement of the mainstem bronchus is now T2 rather than T3, and all of atelectasis/pneumonitis is now T2. TNM8 T3 stage now includes tumors between 5 cm and 7 cm, whereas these tumors were stage T2 under TNM7. Tumors longer than 7 cm are now category T4a under TNM8 instead of T3 under TNM7, and diaphragm invasion is now T4 instead of T3. In summary, TNM8 primarily has more T-stage lung cancer categories that have finer size bin ranges than TNM8.

As a concrete example effect under TNM8, we consider Rami-Porta et al. plots of the 5-year NSCLC survival rate of the patients for clinical staging and pathological staging, separately [36]. Under the TNM8, pT1a stage has a predicted survival rate of 91%. But in the SEER data, 58.1% of such patients would instead have been assigned the cT1b stage, with a lower predicted survival rate of 83%. There is a continuous trend of survival discrepancy between pathological and clinical stage, although both predict a reduction in the survival rate for increasing-sized tumors [36]. Aside from receiving different patient management and care that may affect survival and/or finances, there may also be a psychological negative effect on some patients if they think they have less chance of survival based on their initial diagnosis of clinical staging.

### *How to implement the tabulation framework*

This section covers the steps to implement the framework. We address how to standardize quantifying the concordance, and magnitude of systematic and statistical errors so that hospitals can periodically benchmark whether they have really progressed. Hospitals can use these methods for quality control, to track down the outlier cases that may occur in their own matrices (e.g. Fig. 1) or trends in Tables 3, S1, etc., to identify situations where staging may potentially be improved. In addition, biases in the tumor-size measurements such as those listed in Table 1 [12, 16, 17, 25-32] suggest it may be possible to standardize the measurements of pathological tumor size, which could help improve T-staging. Likewise, some potential clinical measurement and staging improvements were discussed in [7].

Steps:

1 For each patient, save cancer data in discrete fields format in a database or in a spreadsheet. Many cancer hospitals have a cancer registry or departmental database that already performs this function. The data fields needed are: the tumor size measured clinically and pathologically, the pathological and clinical assigned cancer stages, and laterality. Also save information related to treatment in discrete fields (e.g. neoadjuvant therapy), dates of diagnosis and treatment. This is to eliminate cases where the measured tumor size may have changed much due to growth or shrinkage. Also save as discrete fields variables that may affect tumor size or stage discrepancies, e.g., histology, measurement modality (e.g. rounding, 2-D vs 3-D images). Narrative patient reports containing varying terminology are not as efficient as databases with standardized discrete data for analysis purposes.

2 Create a spreadsheet matrix as in Fig. 1 and a table similar to Table 3. We used Excel because it is widely available and provides arithmetic functions: only summation, subtraction, division, multiplication, and square roots are needed. The Excel spreadsheets can be saved as templates with the needed equations to compute agreement and statistical errors. Templates can also include a list of definitions to avoid ambiguity.

3 For patients who have both a pathological and a clinical stage and no neoadjuvant therapy, populate the matrix with the number of cases for each stage category such as pT-stage vs cT-stage in Fig. 1 or pN-stage vs cN-stage in Fig. S3. If the equations are stored in the spreadsheet cells, the matrix should self calculate all the other derived numbers: the Total number of cases in each row or column; the percent-agreement (number of cases with the same assigned pathology and clinical stage (the diagonal elements), divided by the Total number of cases in that row or column, respectively); then multiply by 100 to express the ratio as a percentage. The standard deviation statistical error should self calculate as well (square-root of the number of cases with same assigned pathology and clinical stage, divided by the Total number of cases in that row or column); then multiply by 100 to express the standard deviation as a percentage.

4 To calculate the invariant agreement for each stage category, compute the average of the Total number of cases in each row and respective column, and use that averaged number as the denominator and then follow step 3 to compute percent-agreement.

5 The total "systematic error" for each stage category is the pathological agreement minus the clinical agreement. Users can also treat this value as a linear measure of the biases and measurement issues that systematically cause the pathological and clinical measurements to differ. Issues may include those listed in Table 1 and Reference [7].

6 To glean correlations between the concordance of pathological and clinical stage values with different cancer and treatment parameters, populate tables like Table 3 and 4 with the number of cases in each of the categories. The percentage of cases and standard deviation should self calculate if the Excel table was set up with equations in the cells to compute the ratio and standard deviation (see step 3).

7 For each patient having both clinical and pathological tumor size data, also plot histograms of the number of cases of clinical and pathological tumor size like in Reference [7]. Check whether the distribution is monotonically decreasing or if there are sudden spikes or dips in the data. Also plot the difference between the pathological size and clinical size. Excel can also compute the Pearson correlation if desired.

8 If there are unexpected spikes or scarcity of cases in the data, look at the patient EMR to glean potential causes. For example, if there are too many cTx cases, perhaps check whether these are due to transfer patients. As another example, if there is too much discordance in the T1 stage as opposed to T2 stage, perhaps check the CT images or pathology specimens (or images of specimens) for systematic problems in the clinical measurement of smaller tumors. And so on.

Cancer hospitals, and also registries, can use these simplified techniques to monitor staging and tumor length measurement quality. They can compare their results with SEER's, first to eliminate sources of the more-obvious biases or practice problems discerned from their matrices and tables and EMR, then iteratively repeat the Steps to uncover more subtle problems or correlations with other variables. These same techniques can also be applied to other cancers by revising the names of the stage categories, histology, etc., in the templates and entering the relevant numerical case data. Because some SEER regions have better results than others, this indicates there is room for improvement. With these tools, the possibility of making improvements to cancer staging is beneficial to patients.

### *Advantages of the tabulation framework*

Firstly, the framework is easily implemented to quantify staging quality, using spreadsheets and basic arithmetic. Matrix patterns readily reveal off-diagonal outlier cases that worsen clinical-pathological concordance: clinical T1 cases that have not been properly subcategorized as T1a or T1b (Fig. 1A), six-fold more cTX than pTX cases, which categories are more error-prone, etc. Individual hospitals can access their patient EMR to investigate the outliers. By contrast, present publications often use Kappa indices, multivariate logistic regressions, etc., that are essentially black-box calculations requiring statisticians, who are either scarce or non-existent at some hospitals.

Compared to percent-concordances, the Cohen-kappa concordance index is difficult to interpret, obscures patterns of problems, and is intended for comparing two raters or single-rater repeatability of a single object [22]. However, for cancer staging, different objects are instead being rated by two varying groups of unknown number of raters among the clinicians and pathologists. Additionally, there is a known mathematical problem that Cohen-kappa can be lacking even when the percent-concordance is fine, which is what occurred with our N-stage overall concordance results in Fig. S3 [33, 34]. One way to verify this problem is noting that the order of the N-stage categories is arbitrary. If we re-order the categories in the N-stage matrix, the Cohen-kappa value changes, but the percent-concordance does not. The magnitude of the index also depends on setup parameters such as the weighting scheme on the entries in the matrix. Another drawback of the Kappa index is that the error on the index, is also hard to interpret, whether it is purely statistical in nature or not.

Secondly, the matrices yield clinical-pathological concordances for individual stage categories (T1a, T2b, etc.) using two different denominators. Using the pathological values as the denominator is actually the accepted mathematical method when they are a gold standard or reference value [3, 14, 38]. However, other studies instead referenced the concordance by the total clinical cases (cTotal-i). One problem with this is that the number of clinical T-stage cases is itself a varying value, having wide variance over the years [1, 35], resulting in large discrepancies among studies. Nevertheless, both values are useful, and the difference between the pathological and clinical results provides an easy way to quantify an estimated overall systematic error.

One advantage of the third choice of denominator, averaging (pTotal-i, cTotal-i), is that the concordance is invariant because it is *unique*, allowing comparison with future studies unambiguously and compatibly. This choice is consistent with the calculation of overall T-stage concordance, where the denominator is the total number of cases in the entire matrix, effectively averaging over both pathological and clinical cases. This averaged denominator is akin to Bland-Altman plots that average values when the true value is unknown. Averaging is consistent with TNM8 treating pathological and clinical measurements equally valid [9, 36]. Staging guidelines point out that the pathological tumor length is not so "golden" a reference standard because it too displays biases like in Table 1. For NSCLC that is diagnosed by ever-improving radiologic images, the accuracy of clinical staging is improving, perhaps warranting being treated equally with pathological staging, especially now that prognosis has also been calibrated with respect to clinical tumor length [36].

Thirdly, when analyzing characteristics that cause cT vs pT discordance, the multivariate logistic regressions are hard to interpret and implement, requiring a statistics package, aid of a biostatistician, iterative processing, and manipulating or omnibussing input predictor-variables that have many species (e.g. histology). The results may vary depending on which input variables or tool settings are included. Univariate logistic regressions are easier to perform and understand but may yield biased results due to under-specification errors [37]. Even the multivariate analysis may be under-specified due to missing predictor-variables (i.e. not including all relevant pieces of a whole pie). By contrast, simple concordance vs cancer characteristics tabulations are easily set up in a spreadsheet to obtain insights like with Tables 3 and 4.

Table 3 tabulates the percentage of SEER cases cancer characteristics versus three ordinal-output T-stage concordance categories. Instead of merely numerically associating a characteristic with clinical-pathological concordance by a p-value, which is what multivariate regressions determine, tabulations instead make it easier to detect trends. One could also readily plot a histogram from the tabulations and visually see the trends. For example, Table 3 shows that as resected specimen size increases, the T-stage concordance declines, and pT becomes upstage of cT. To detect trends with lower data statistics such as <1000 cases, the three ordinal-outputs categories may need to be reduced to two values: agree and disagree, depending on the number of species of a characteristic (e.g. number of types of surgery) [13].

### *Limitations and recommendations*

In this article we address how to quantify clinical and pathological stage discordance and look for patterns of issues (e.g. which stage subcategories or tumor size ranges exhibit the largest discordance, whether the upstaging jumps by one or more than one subcategory, which variables affect stage discordance, etc.). We also address how to standardize the reporting of the arithmetic magnitude of discordance and its systematic and statistical errors so that different studies can be compared apples-to-apples. However, we do not address methods to correct or reduce the discordance such as by standardizing certain practices, or methods to reduce the biases that contribute to and increase the magnitude of systematic errors. Rather, in Reference [7], we addressed how to reduce some of the clinical staging biases and standardize some of the methods of measuring clinical tumor length and reporting a clinical NSCLC stage. In an upcoming publication on breast cancer in the American Journal of Clinical Pathology, we address how to overcome some challenges in gross tumor measurements that determine the pathological cancer stage. Lung cancer gross tumor measurements are performed in a manner similar to breast tumors measurements, and thus there may be benefits in considering the breast cancer article. Nevertheless, more research and guidance are needed in standardizing NSCLC staging to make clinical NSCLC staging more accurate.

Registrars can preserve only a finite amount of data. However, to better assess and quantify staging discordance, it would be useful to have additional data fields such as the source of the stage values assigned to a patient (e.g. CT with or without contrast, EBUS/EUS procedures). Nevertheless, at least the SEER patient survival data will become available in a few years to perform correlations with cancer stage and the magnitude of pathological-clinical discordance.

Our tabulation framework methods require sufficient data. A typical cancer hospital has much fewer cases than SEER. Consequently, we applied the methods to only 450 in-house NSCLC cases to check how well they work with fewer cases. We obtained 61.1 ± 3.7% concordance for TNM-stage, 79.6 ± 4.2% for N-stage, and 58.8 ± 3.5% for T-stage, which are statistically identical to the SEER results. Tis, T1a, T1b, T2a, T2b, T3, T4 invariant concordances are 90 ± 30%; 74.4 ± 7.5 + 5.6%; 53.6 ± 8.8 + 19.2%; 53.6 ± 6.5 – 31.5%; 59.3 ± 14.8 – 18.0%; 61.7 ± 11.5 – 18.8%; 75 ± 19 + 7.6%, respectively. Except for T2b, these values are statistically identical to the SEER results. It is not necessary to have even 450 cases to ascertain patterns in the data from the matrices (e.g. Fig. 1). Too many outliers off-diagonal would be readily apparent and indicate problems may exist.

Our definition of an aggregated systematic error for each category helps to further quantify the magnitude of discordance, but it is not the traditional root sum square (RSS) systematic error that is difficult to calculate due to the existence of many biases (e.g. Table 1). We state an overall concordance of 58.1 ± 0.9%, which includes a statistical error but lacks information about the systematic error. The total RSS error on the overall concordance would include the error contribution from each bias. To quantify the contribution from any particular individual bias entails tabulating a new matrix of results after including the bias, and then calculate the difference between corresponding percent-concordances between the original and new matrices. For example, the T-stage concordance for delay-to-resection of <120 days versus longer periods, is 58.1% versus 57.7%, respectively. The difference of 0.4% is an estimate of the individual systematic error due to delay, during which time the tumor may have progressed. The cT vs pT concordance accounting for systematic error due only to delay can be reported as 58.1 ± 0.9–0.4%, where the minus sign signifies delays reduce concordance.

## Conclusion

Clinical NSCLC staging accuracy can be improved, and a first step in the overall process towards improvement is to standardize and unify the method of quantifying accuracy and discordance and identifying where issues may exist. This tabulation-template framework simplifies comparing staging results compatibly and detecting problems. Cancer hospitals and registries can implement the intuitive framework to quantify, monitor and improve staging accuracy at their sites with an overall goal of improving cancer patient management, treatment and prognosis.

## Credit author statement

**Dolly Y. Wu:** Conceptualization, Resources, Methodology, Data curation, Software, Validation, Formal analysis, Investigation, Visualization, Project administration, Funding acquisition, Writing - original draft, Writing - review & edit

**Ann E. Spangler:** Conceptualization, Resources, Writing - review & edit, Investigation

**Dat T. Vo:** Conceptualization, Resources, Writing - review & edit, Methodology, Investigation, Funding acquisition

**Alberto de Hoyos:** Conceptualization, Resources, Writing - review & edit

**Stephen J. Seiler:** Conceptualization, Resources, Writing - review & edit, Project administration, Funding acquisition

## Conflict of interest and support

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The retrospective patient review study protocol was approved by the UTSW Institutional Review Board (STU042018–003, with waived informed consent).

## Acknowledgements

We especially thank the U.S. National Cancer Institute's SEER Program; UT Southwestern Medical Center Cancer Registry; UTSW Thoracic Surgery Database (Phillip Escarsega); Ana Barragan Montero (PhD), Yasmeen Butt (MD), H. Michael Crowson (PhD), Fernando Kay (MD), Mia Lv (PhD), Ayobami Odu (MBBS), Kenneth Westover (MD,PhD).

## Appendix and Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ctarc.2020.100253. Here is a link to the TNM7 lung cancer staging categories. http://cancerstaging.org/references-tools/quickreferences/documents/lungmedium.pdf.

## References

[1] D.J. Heineman, M.G. Ten Berge, J.M. Daniels, et al., Clinical staging of stage I non-small cell lung cancer in the Netherlands-need for improvement in an era with expanding nonsurgical treatment options: data from the Dutch lung surgery audit, Ann. Thorac. Surg. 102 (2016) 1615–1621.

[2] D.J. Heineman, M.G. Ten Berge, J.M. Daniels, et al., The quality of staging non-small cell lung cancer in the Netherlands: data from the Dutch lung surgery audit, Ann. Thorac. Surg. 102 (2016) 1622–1629.

[3] A. Lopez-Encuentra, R. Garcia-Lujan, J.J. Rivas, et al., Comparison between clinical and pathologic staging in 2,994 cases of lung cancer, Ann. Thorac. Surg. 79 (2005) 974–979.discussion 979

[4] I. Macia, J. Moya, I. Escobar, et al., Quality study of a lung cancer committee: study of agreement between preoperative and pathological staging, Eur. J. Cardiothorac. Surg. 37 (2010) 540–545.

[5] B.H. Heidinger, K.R. Anderson, E.M. Moriarty, et al., Size measurement and T-staging of lung adenocarcinomas manifesting as solid nodules <=30mm on CT: radiology-pathology correlation, Acad. Radiol. 24 (2017) 851–859.

[6] J. D'Cunha, J.E. Herndon, Herzan DL 2nd, et al., Poor correspondence between clinical and pathologic staging in stage I non-small cell lung cancer: results from CALGB 9761, a prospective trial, Lung Cancer 48 (2005) 241–246.

[7] D.Y. Wu, A. de Hoyos, D.T. Vo, et al., Clinical non-small cell lung cancer staging and tumor length measurement results from U.S. cancer hospitals, Acad. Radiol. (2020), https://doi.org/10.1016/j.acra.2020.04.007.

[8] S.B. Edge, The AJCC Staging Manual, 7th Edition, Springer, 2009.

[9] M.B Amin, AJCC Cancer Staging Manual 8th Edition (and Update to Breast Chapter), Springer, 2017.

[10] L.H. Sobin, M.K. Gospodarowicz, C. Wittekind, International union against Cancer. TNM Classification of Malignant Tumours, Wiley-Blackwell, Hoboken, NJ, 2010. Chichester, West Sussex, UK.

[11] Brierley J., Gospodarowicz M.K., Wittekind C. TNM classification of malignant tumours. Chichester, West Sussex, UK; Hoboken, NJ: John Wiley & Sons, Inc., 2017.

[12] C.I. Henschke, D.F. Yankelevitz, R. Yip, et al., Lung cancers diagnosed at annual CT screening: volume doubling times, Radiology 263 (2012) 578–583.

[13] B.M. Stiles, E.L. Servais, P.C. Lee, et al., Point: clinical stage IA non-small cell lung cancer determined by computed tomography and positron emission tomography is frequently not pathologic IA non-small cell lung cancer: the problem of understaging, J. Thorac. Cardiovasc. Surg. 137 (2009) 13–19.

[14] C.M. Washington, Principles and Practice of Radiation Therapy - E-Book, Elsevier, 2015, p. 265. Also see http://www.phy.ilstu.edu/slh/percent%20difference%20error.pdf or http://science.clemson.edu/physics/labs/tutorials/error/index.html.

[15] D.Y. Wu, K. Hayes, M.L. Perl, et al., Radiative tau-production and decay, Phys. Rev. D 41 (1990) 2339–2342.

[16] P.K. Hsu, H.C. Huang, C.C. Hsieh, et al., Effect of formalin fixation on tumor size determination in stage I non-small cell lung cancer, Ann. Thorac. Surg. 84 (2007) 1825–1829.

[17] H.S. Park, S. Lee, S. Haam, G.D. Lee, Effect of formalin fixation and tumour size in small-sized non-small-cell lung cancer: a prospective, single-centre study, Histopathology 71 (2017) 437–445.

[18] Cardillo G. Cohen's Kappa: compute the cohen's kappa ratio. . In. http://www.mathworks.com/matlabcentral/fileexchange/15365: 2007.

[19] J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychol. Bull. 70 (1968) 213–220.

[20] N. Wongpakaran, T. Wongpakaran, D. Wedding, K.L Gwet, A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples, BMC Med. Res. Methodol. 13 (2013) 61.

[21] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.

[22] M.C. Banerjee, , Michelle, Laura McSweeney, Michelle, Beyond kappa: a review of interrater agreement measures, Canad. J. Stat. 27 (2008) 20.

[23] N. Navani, D.J. Fisher, J.F. Tierney, et al., The accuracy of clinical staging of stage I-IIIa non-small cell lung cancer: an analysis based on individual participant data, Chest 155 (2019) 502–509.

[24] E. Cetinkaya, A. Turna, P. Yildiz, et al., Comparison of clinical and surgical-pathologic staging of the patients with non-small cell lung carcinoma, Eur. J. Cardiothorac. Surg. 22 (2002) 1000–1005.

[25] M.J. Bott, A.P. Patel, T.D. Crabtree, et al., Pathologic upstaging in patients undergoing resection for stage I non-small cell lung cancer: are there modifiable predictors? Ann. Thorac. Surg. 100 (2015) 2048–2053.

[26] K.R. Anderson, B.H. Heidinger, Y. Chen, et al., Measurement bias of gross pathologic compared with radiologic tumor size of resected lung adenocarcinomas: implications for the T-stage revisions in the eighth edition of the american joint committee on cancer staging manual, Am. J. Clin. Pathol. 147 (2017) 641–648.

[27] C.H. Park, T.H. Kim, S. Lee, et al., Correlation between maximal tumor diameter of fresh pathology specimens and computed tomography images in lung adenocarcinoma, PLoS ONE 14 (2019), e0211141.

[28] N.V. Adsay, O. Basturk, B. Saka, Pathologic staging of tumors: pitfalls and opportunities for improvements, Semin. Diagn. Pathol. 29 (2012) 103–108.

[29] K. Lampen-Sachar, B. Zhao, J. Zheng, et al., Correlation between tumor measurement on Computed Tomography and resected specimen size in lung adenocarcinomas, Lung Cancer 75 (2012) 332–335.

[30] G.E. Orchard, M. Shams, T. Nwokie, et al., Development of new and accurate measurement devices (TruSlice and TruSlice Digital) for use in histological dissection: an attempt to improve specimen dissection precision, Br. J. Biomed. Sci. 72 (2015) 140–145.

[31] T. Radonic, C. Dickhoff, M. Mino-Kenudson, et al., Gross handling of pulmonary resection specimen: maintaining the 3-dimensional orientation, J. Thorac. Dis. 11 (2019) S37–S44.

[32] G.L. Sica, A.A. Gal, Lung cancer staging: pathology issues, Semin. Diagn. Pathol. 29 (2012) 116–126.

[33] M.L McHugh, Interrater reliability: the kappa statistic, Biochem. Med. (Zagreb) 22 (2012) 276–282.

[34] S. Xu, M.F. Lorber, Interrater agreement statistics with skewed data: evaluation of alternatives to Cohen's kappa, J. Consult. Clin. Psychol. 82 (2014) 1219–1227.

[35] A. Gdeedo, P. Van Schil, B. Corthouts, et al., Comparison of imaging TNM [(i)TNM] and pathological TNM [pTNM] in staging of bronchogenic carcinoma, Eur. J. Cardiothorac. Surg. 12 (1997) 224–227.

[36] R. Rami-Porta, V. Bolejack, J. Crowley, et al., The IASLC lung cancer staging project: proposals for the revisions of the T descriptors in the forthcoming eighth edition of the TNM classification for lung cancer, J. Thorac. Oncol. 10 (2015) 990–1003.

[37] H. Wang, J. Peng, B. Wang, et al., Inconsistency between univariate and multiple logistic regressions, Shanghai Arch. Psychiatry 29 (2017) 124–128.

[38] H.C. Fernando, P. Goldstraw, The accuracy of clinical evaluative intrathoracic staging in lung cancer as assessed by postsurgical pathologic staging, Cancer 65 (1990) 2503–2506.