

# Structured EHR Data Underestimates Prevalence And Misses Large Proportions Of Patients With Pulmonary Hypertension

R. Chen<sup>1</sup>, J. Pfeifer<sup>2</sup>, D. Leibowitz<sup>1</sup>, D. M. Vidmar<sup>1</sup>, W. Thompson<sup>1</sup>, J. Leader<sup>2</sup>, K. Morland<sup>2</sup>, A. Nelsen<sup>2</sup>, B. Fornwalt<sup>1</sup>

<sup>1</sup>Tempus AI, Inc., Chicago, IL, United States, <sup>2</sup>United Therapeutics, Research Triangle Park, NC, United States

## INTRODUCTION

- Pulmonary hypertension (PH) is a progressive, debilitating disease which is difficult to accurately identify in real-world datasets
- Such challenges limit opportunities to optimize medical therapy and improve clinical research
- Few estimates exist of population-wide prevalence
- We sought to determine:
  - 1) whether patients with PH could be reliably identified using structured electronic health record (EHR) data and
  - 2) estimate population-wide prevalence of PH

## METHODS

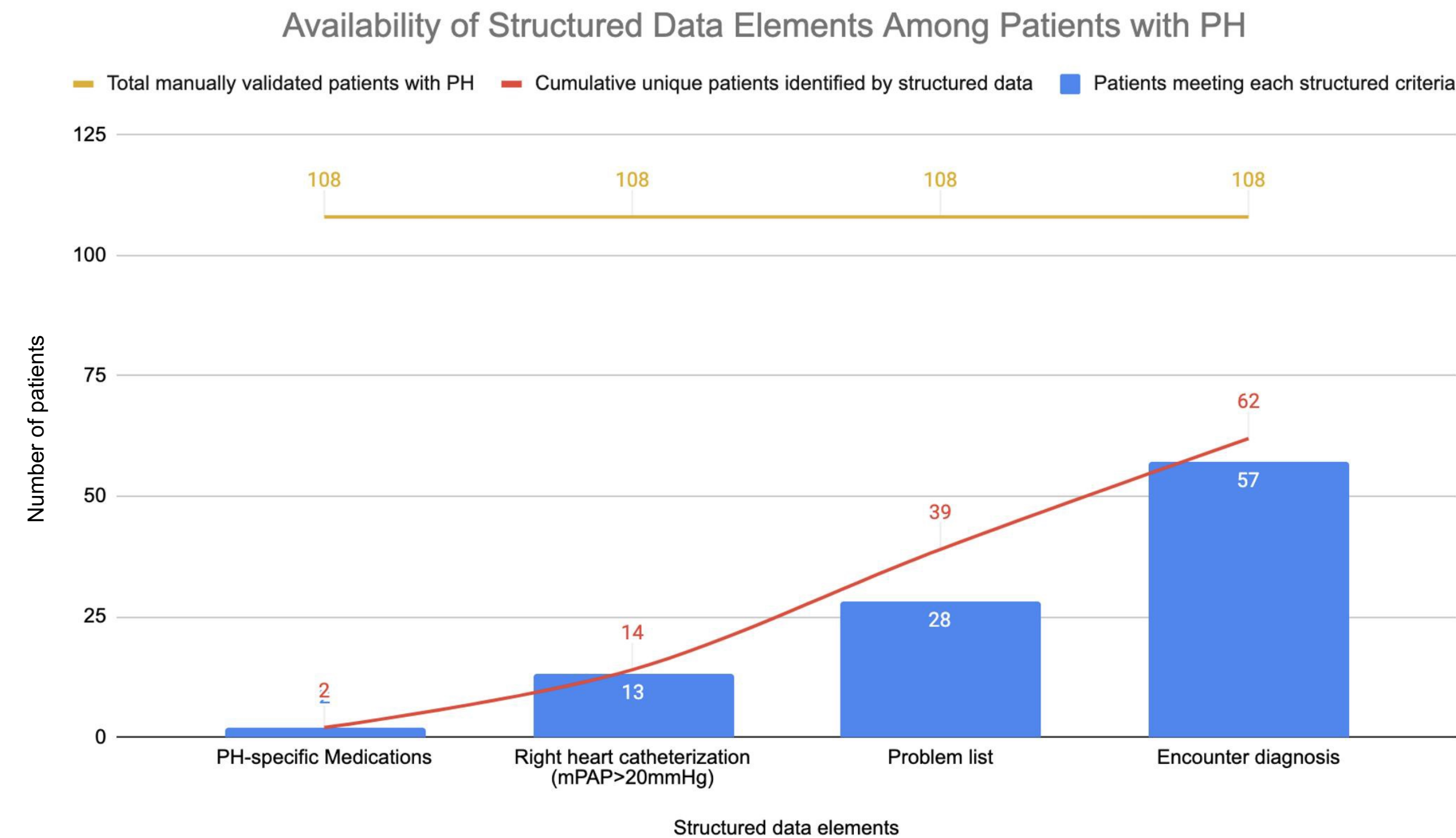
- We used a longitudinal de-identified, multi-modal EHR dataset of 2+ million patients
- We first use a filtering method to extract all patients with at least one piece of evidence of possible PH anywhere in the EHR (such as clinical notes or structured data)
  - These methods enable random sampling for a reliable, unbiased estimate of sensitivity within a feasible number of chart reviews
- From the remaining cohort (filtered population), we randomly sampled 300 patients for blinded, two-physician review
  - We identified patients with clinically diagnosed PH; defined as a clear diagnosis or mention in unstructured note text by a clinician
- We then characterized the availability of structured data elements for these patients with PH for the following:
  - 1) any occurrence of a PH-related diagnosis code on the problem list,
  - 2) any occurrence of a PH-related diagnosis code in any encounter,
  - 3) any order or medication reconciliation of PH-specific vasodilator therapies, or
  - 4) any right heart catheterization (RHC) with mean pulmonary artery pressure >20mmHg
- Additionally, we estimated population-wide prevalence by extrapolating the proportion of PH+ patients within the random sample of the filtered population to the overall EHR population

## SUMMARY

- Structured data elements in EHR datasets fail to identify a large proportion of patients with clinically diagnosed PH
- Leveraging unstructured data from clinical notes, such as in natural language processing-based phenotyping approaches, may be necessary for more complete identification of patients with PH in EHR data

## RESULTS

Figure 1. Limited availability of EHR structured data in patients with known PH



Within the random sample of the filtered population, we identified 108 patients with clinically diagnosed PH

Only 62 (57%) of these 108 patients had structured EHR data suggestive of PH

Population-wide prevalence estimates:

- 36% (95% CI: 31%-42%) of patients in random sample found to have PH => extrapolated to entire post-filter cohort = 59K
- Presuming all positives are in the post-filtered set (a conservative lower bound estimate), 59K/2.23M patients= a prevalence of 2.6%
- Population-wide prevalence of PH estimated to be 2.6% (much higher than existing literature estimate of ~1%)<sup>1</sup>

Figure 2. Conservative estimate of PH prevalence is 2.6%

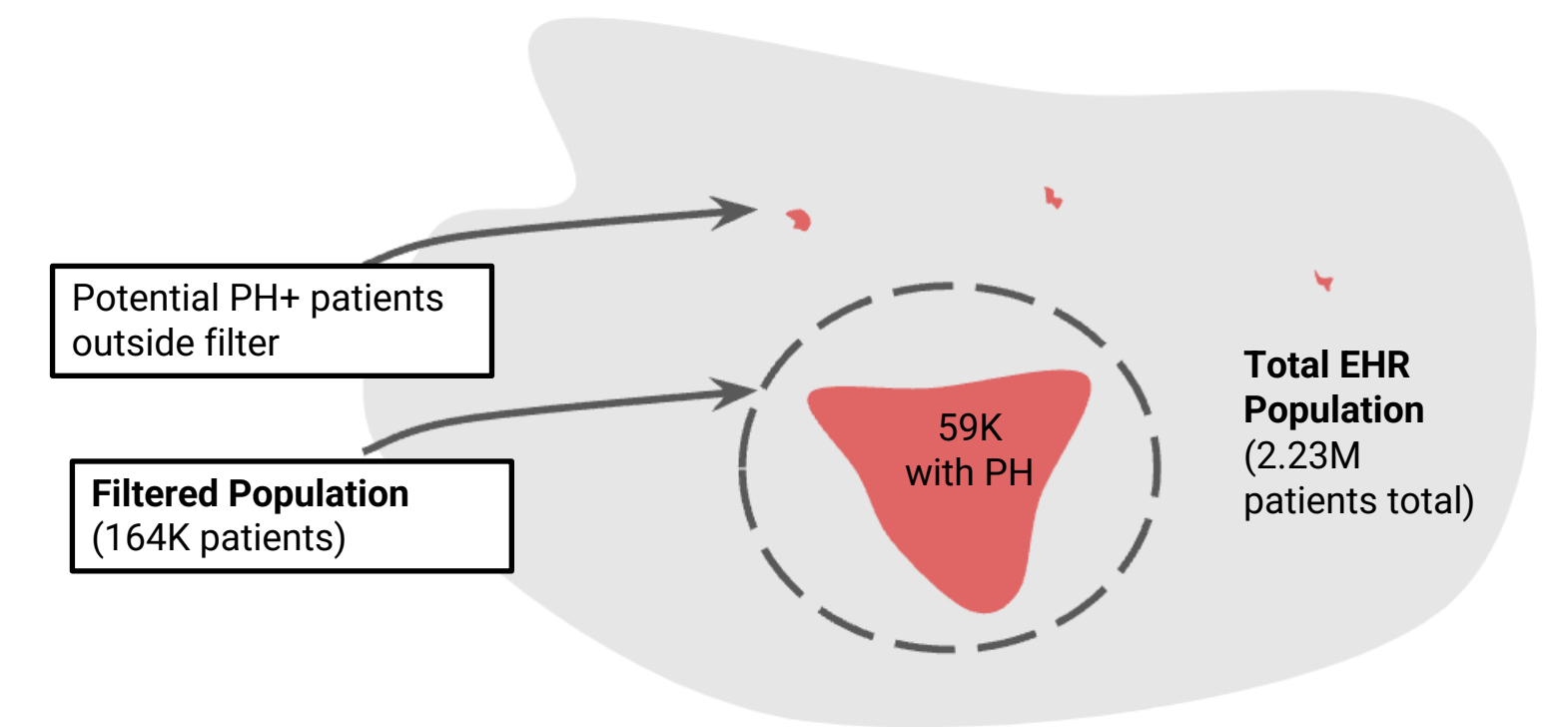


Figure 2. Presuming no additional patients outside the filter have PH, 59K/2.23M patients total=2.6% prevalence. If there are additional missed PH patients outside the filtered cohort, the estimated prevalence would be even higher

Figure 3. PH diagnosis codes are neither sensitive nor specific for PH

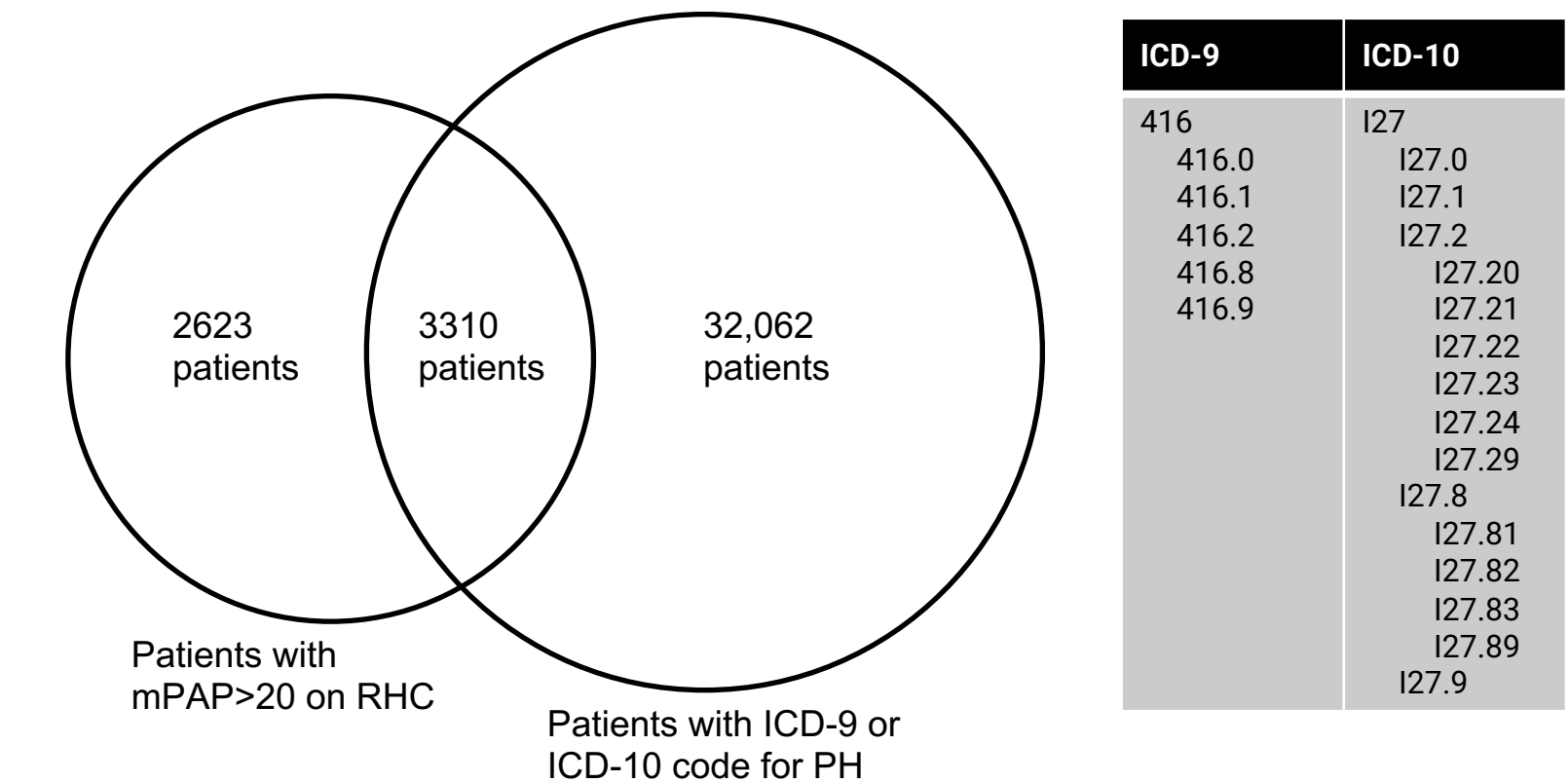


Figure 3. 44% (2623) of patients with mPAP>20 do not have any PH diagnosis codes (ICD-9 or 10) from any encounter in their HER history

1. Hoepfer MM, Humbert M, Souza R, et al. A global view of pulmonary hypertension. *Lancet Respir Med.* 2016;4(4):306-322.