

# TriageLink - Week 1 Deliverable

## Data Source Summary & Schema Map

**Project:** AI-Driven ER Wait-Time Prediction and Dashboard  
**Team Member:** Zuhair Abbas  
**Role:** Machine Learning & Data Integration (60 hours)  
**Week:** 1 of 6  
**Date:** November 17-23, 2025  
**Hours Logged:** 10 hours

## Executive Summary

This document provides a comprehensive analysis of three primary data sources for TriageLink's predictive ER wait-time model:

- 1. **ER Watch** - Ontario hospital ER wait times (real-time)
- 2. **HowLongWillIWait** - Multi-hospital wait time aggregator (real-time)
- 3. **Pediatric Triage Dataset** - CTAS-based triage conditions (historical)

All three datasets have been reviewed, their schemas documented, and integration pathways identified for Week 2 data cleaning and merging.

## Data Source 1: ER Watch

### Overview

- **Source:** <https://www.er-watch.ca/>
- **Update Frequency:** ~15 minutes
- **Coverage:** Ontario hospitals with public ER wait time reporting
- **Data Format:** JSON (via web scraping) or HTML parsing

### Key Features

- ☒ Near real-time ER wait times
- ☒ Hospital-specific data
- ☒ Publicly available (no authentication required)
- ☒ Updates every 15 minutes (average)

### Data Schema (Inferred)

Field	Type	Description	Example	Notes
hospital_name	String	Full hospital name	"Toronto General (University Health Network)"	Primary identifier

Field	Type	Description	Example	Notes
wait_time	String/Int	Current ER wait time	"2 hr 12 min" or "Not available"	Requires parsing
timestamp	DateTime	Data collection time	"2025-11-09 23:52:26"	System-generated
hospital_system	String	Parent health system	"University Health Network"	Extracted from name
data_available	Boolean	Whether data is reported	True/False	Derived field

Data Quality Issues

- **Missing Data:** Some hospitals show "Not available"
- **Format Inconsistency:** Wait times have multiple formats:
  - "2 hr 12 min"
  - "1 hr or less to 1 hr 9 min" (ranges)
  - "Not available"
- **Hospital Naming:** Inconsistent naming conventions across sources

Sample Data (from testing)

```
{
  "Toronto General (University Health Network)": "2 hr 22 min",
  "Sunnybrook": "8 hr 25 min",
  "CHEO (Childrens Hospital of Eastern Ontario)": "12 hr 35 min",
  "Windsor Regional Hospital - Oullette Campus": "Not available"
}
```

Collection Method

- Automated scraping script developed** (already completed for interview assignment):
- Python-based collector using `requests` library
  - Collects data every 30 minutes
  - Auto-saves to Excel with timestamps
  - Comprehensive error handling for network issues

**Collection Rate:** 48 data points/day per hospital × 50+ hospitals = 2,400+ daily records

Data Source 2: HowLongWillIWait

Overview

- **Source:** <https://howlongwilliwait.com/sample.json>
- **Update Frequency:** Real-time (API-based)
- **Coverage:** 53 Ontario hospitals
- **Data Format:** JSON (REST API)

Key Features

- ☒ REST API endpoint (easier than scraping)
- ☒ JSON format (machine-readable)
- ☒ Same hospital coverage as ER Watch
- ☒ Already tested and validated (interview assignment)

Data Schema

Field	Type	Description	Example	Notes
hospital_name	String	Hospital identifier	"Toronto General (University Health Network)"	Key field
wait_time	String	Current wait time	"2 hr 12 min"	Consistent format with ER Watch
region	String	Geographic region	"GTA", "Eastern Ontario"	To be derived
hospital_tier	String	Hospital classification	"Tier1_Pediatric_Centre"	To be merged from pediatric dataset

Sample API Response

```
{
  "Toronto General (University Health Network)": "2 hr 12 min",
  "Toronto Western (University Health Network)": "3 hr 7 min",
  "Sunnybrook": "8 hr 25 min",
  "Markham Stouffville (Oak Valley Health)": "2 hr 30 min"
}
```

Collection Status

- ☒ **Already implemented** - Data collection script operational
- ☒ **Tested successfully** - Collected 106 records (53 hospitals × 2 timestamps)
- ☒ **Output validated** - Excel export working correctly

Integration Notes

- **Overlap with ER Watch:** ~90% hospital overlap (good for validation)
- **Data freshness:** Both sources update at similar intervals
- **Primary vs Secondary:** Will use HowLongWillIWait as primary source (more reliable API)

## Data Source 3: Pediatric Triage Dataset

### Overview

- **Source:** Internal dataset (provided by Langyue/Tony)
- **Format:** Excel (.xls)
- **Purpose:** CTAS-based pediatric triage conditions for model training
- **Records:** 15 pediatric conditions

### Dataset Structure

**Shape:** 15 conditions × 9 attributes

### Data Schema

Column	Type	Description	Example	Unique Values
condition	String	Pediatric medical condition	"Fever under 3 months", "Asthma exacerbation"	15
age_range	String	Applicable age range	"<1 year", "2–12 years"	12
ctas_level	Integer	Canadian Triage Acuity Scale (1-5)	2, 3, 4	4 levels
red_flag_criteria	String	Critical warning signs	"Temp > 38.5°C or lethargy"	15
recommended_destination	String	Hospital tier recommendation	"Tier1_Pediatric_Centre"	2 tiers
hospital_tier	String	Specific hospital recommendation	"SickKids", "CHEO", "McMaster Children's Hospital"	7 hospitals
vital_sign_thresholds	String	Clinical thresholds	"Temp > 38.5°C", "O <sub>2</sub> < 92%"	11 thresholds
symptom_keywords	String	Search keywords	"fever, lethargy, infant"	15 keyword sets
notes	String	Clinical notes	"Urgent referral for possible sepsis"	15

### CTAS Level Distribution

Level 2 (Emergent):	4 conditions (27%)
Level 3 (Urgent):	7 conditions (47%)

Level 4 (Less Urgent):	3 conditions (20%)
Level 5 (Non-urgent):	1 condition (7%)

Hospital Tier Breakdown

Tier 1 - Pediatric Centres:

- SickKids (5 conditions)
- CHEO (3 conditions)
- McMaster Children's Hospital (3 conditions)

Tier 2 - Pediatric Units:

- Humber River Hospital (2 conditions)
- Unity Health Toronto (2 conditions)
- William Osler Health System (1 condition)

Sample Records

Example 1: High Acuity (CTAS Level 2)

Condition: Fever under 3 months
Age Range: <1 year
CTAS Level: 2 (Emergent)
Red Flags: Temp > 38.5°C or lethargy
Destination: SickKids (Tier1_Pediatric_Centre)
Thresholds: Temp > 38.5°C
Keywords: fever, lethargy, infant
Notes: Urgent referral for possible sepsis

Example 2: Medium Acuity (CTAS Level 3)

Condition: Asthma exacerbation
Age Range: 2-12 years
CTAS Level: 3 (Urgent)
Red Flags: RR > 40 or O <sub>2</sub> < 92%
Destination: Humber River Hospital (Tier2_Pediatric_Unit)
Thresholds: O <sub>2</sub> < 92%
Keywords: wheeze, shortness of breath
Notes: Monitor response to bronchodilator

Data Quality Assessment

- ☒ **Complete data** - No missing values in critical fields
- ☒ **Consistent format** - Standardized structure across conditions

- ☒ **Clinical validity** - Aligns with Pre-CTAS Manual v1.5
- ☐ **Limited scope** - Only 15 conditions (Langyue will expand)

# Integration Plan: Merging Three Data Sources

## Phase 1: Hospital Master Table (Week 2)

**Goal:** Create unified hospital reference table

Source	Provides	Join Key
ER Watch	Real-time wait times	hospital_name
HowLongWillIWait	Real-time wait times (primary)	hospital_name
Pediatric Dataset	Hospital tier, pediatric capability	hospital_tier → hospital_name

**Output:** triage\_hospital\_master.csv

## Phase 2: Feature Engineering (Week 3)

**Derived Variables:**

### 1. Temporal Features

- hour\_of\_day (0-23)
- day\_of\_week (Mon-Sun)
- is\_weekend (Boolean)
- is\_night\_shift (18:00-06:00)

### 2. Geographic Features

- region (GTA, Eastern, Western, Northern, Central)
- distance\_to\_pediatric\_centre (km)

### 3. Hospital Attributes

- has\_pediatric\_er (Boolean)
- is\_trauma\_centre (Boolean)
- hospital\_tier (1 or 2)
- 24hour\_coverage (Boolean)

### 4. Severity Features

- ctas\_level (1-5)
- has\_red\_flags (Boolean)
- severity\_score (calculated)

## Phase 3: Training Dataset Structure (Week 3-4)

**Target Variable:** wait\_time\_minutes (converted from "2 hr 12 min" → 132)

Feature Set (20+ features):

Category	Features	Count
Temporal	hour, day_of_week, is_weekend, is_night	4
Geographic	region, latitude, longitude	3
Hospital	tier, pediatric_capable, trauma_centre, bed_count	4
Triage	ctas_level, has_red_flags, condition_category	3
Historical	avg_wait_30min_ago, avg_wait_1hr_ago, rolling_avg	3
Weather (future)	temperature, precipitation	2

**Rows:** ~50 hospitals × 48 datapoints/day × 90 days = **216,000 training records**

Data Dictionary

Master Data Dictionary

Field Name	Data Type	Source	Description	Possible Values	Required	Notes
record_id	Integer	Generated	Unique record identifier	1, 2, 3...	Yes	Auto-increment
timestamp	DateTime	Collector	Data collection time	ISO 8601 format	Yes	UTC timezone
hospital_id	Integer	Generated	Unique hospital ID	1-53	Yes	Foreign key
hospital_name	String	ER Watch / HLWIW	Full hospital name	Various	Yes	Primary identifier
hospital_system	String	Derived	Parent health network	UHN, Unity Health, etc.	No	Extracted from name
region	String	Derived	Ontario health region	GTA, Eastern, Western, etc.	Yes	Geographic grouping
hospital_tier	Integer	Pediatric Dataset	Hospital classification	1 (Pediatric Centre), 2 (Pediatric Unit)	No	Null if not pediatric

Field Name	Data Type	Source	Description	Possible Values	Required	Notes
has_pediatric_er	Boolean	Pediatric Dataset	Pediatric ER capability	True/False	Yes	Derived from tier
wait_time_raw	String	ER Watch / HLWIW	Original wait time string	"2 hr 12 min"	Yes	As-received format
wait_time_minutes	Integer	Derived	Wait time in minutes	0-999	Yes	Target variable
data_available	Boolean	Derived	Data reporting status	True/False	Yes	False if "Not available"
hour_of_day	Integer	Derived	Hour (0-23)	0-23	Yes	For temporal patterns
day_of_week	Integer	Derived	Day (0=Mon, 6=Sun)	0-6	Yes	For temporal patterns
is_weekend	Boolean	Derived	Weekend flag	True/False	Yes	Sat/Sun = True
ctas_level	Integer	Pediatric Dataset	Triage acuity (1-5)	1 (Resuscitation) to 5 (Non-urgent)	No	For specific conditions
condition	String	Pediatric Dataset	Medical condition	"Fever under 3 months", etc.	No	For pediatric cases
red_flag_criteria	String	Pediatric Dataset	Critical warning signs	Various	No	Clinical indicators

Data Quality Summary

ER Watch / HowLongWillIWait

Metric	Value	Status
Total Hospitals	53	<input checked="" type="checkbox"/> Good coverage
Data Availability Rate	~75%	<input type="checkbox"/> 25% show "Not available"
Update Frequency	15 minutes	<input checked="" type="checkbox"/> Suitable for real-time

Metric	Value	Status
Historical Data	None (need to collect)	⦿ Requires 90-day collection
Format Consistency	Moderate	⚠ Requires parsing

Pediatric Triage Dataset

Metric	Value	Status
Total Conditions	15	⚠ Limited (Langyue expanding to 20+)
CTAS Coverage	Levels 2-5	☑ Good distribution
Hospital Mapping	7 hospitals	⚠ Need full Ontario mapping
Completeness	100%	☑ No missing values
Clinical Validity	High	☑ Based on Pre-CTAS Manual

Next Steps (Week 2)

Data Cleaning Tasks (10 hours)

- 1. **Normalize Hospital Names** (2 hours)
  - Create hospital ID mapping table
  - Handle naming inconsistencies
  - Map to standard identifiers
- 2. **Merge Hospital Attributes** (3 hours)
  - Add pediatric ER flag
  - Add trauma centre flag
  - Add 24-hour coverage flag
  - Add bed count (if available)
- 3. **Handle Missing Values** (3 hours)
  - Strategy for "Not available" wait times
  - Imputation vs exclusion decision
  - Document missing data patterns
- 4. **Encode Categorical Features** (2 hours)
  - One-hot encoding for regions
  - Ordinal encoding for CTAS levels
  - Binary encoding for flags

Week 2 Deliverable: `triage_hospital_master.csv`

Technical Notes

## Collection Infrastructure

### Already Implemented:

- Python-based data collector ☒
- Automated 30-minute intervals ☒
- Excel export functionality ☒
- Error handling and retry logic ☒

### To Be Developed:

- Database storage (PostgreSQL recommended)
- Automated data validation
- Data quality monitoring dashboard

## Tools & Libraries

### Current Stack:

- Python 3.12
- pandas 2.1.0
- requests 2.31.0
- openpyxl 3.1.2

### Week 2+ Requirements:

- scikit-learn (model training)
- XGBoost (advanced models)
- matplotlib/seaborn (visualization)
- SQLAlchemy (database integration)



---

## Collaboration Checkpoints

### Completed

- ☒ Week 1 kickoff meeting with Tony + Langyue (Nov 4, 2025)
- ☒ Data source review and schema documentation
- ☒ Initial pediatric dataset analysis

### Upcoming

-  Week 3 sync with Langyue - verify CTAS field integration
-  Week 5-6 end-to-end testing with Langyue's dashboard

---

## Appendix: Data Samples

### A. HowLongWillIWait API Sample (Nov 9, 2025 23:52)

--

```
Timestamp,Hospital Name,Wait Time
2025-11-09 23:52:26,Toronto General (University Health Network),2 hr 22 min
2025-11-09 23:52:26,Toronto Western (University Health Network),1 hr 20 min
2025-11-09 23:52:26,Sunnybrook,8 hr 25 min
2025-11-09 23:52:26,CHEO (Childrens Hospital of Eastern Ontario),12 hr 35 min
2025-11-09 23:52:26,Markham Stouffville (Oak Valley Health),2 hr 30 min
```

B. Pediatric Triage Dataset Sample

```
condition,age_range,ctas_level,red_flag_criteria,recommended_destination,hospital_
tier
Fever under 3 months,<1 year,2,Temp > 38.5°C or
lethargy,Tier1_Pediatric_Centre,SickKids
Asthma exacerbation,2-12 years,3,RR > 40 or O2 < 92%,Tier2_Pediatric_Unit,Humber
River Hospital
Seizure,1-12 years,2,Active seizure or postictal state,Tier1_Pediatric_Centre,CHEO
```

## Summary & Conclusions

Key Findings

- 1. **Data Availability:** ER Watch and HowLongWillIWait provide comprehensive real-time coverage of 53 Ontario hospitals
- 2. **Data Quality:** ~75% data availability rate; 25% missing data requires handling strategy
- 3. **Integration Feasibility:** Three data sources can be merged via hospital name matching with manual mapping support
- 4. **Feature Richness:** Combining wait times + CTAS data + hospital attributes creates robust feature set for ML
- 5. **Collection Status:** Data collection infrastructure already operational from interview assignment

Risks & Mitigation

Risk	Impact	Mitigation
Missing historical data	High	Begin immediate 90-day collection
Hospital name mismatches	Medium	Create manual mapping table Week 2
Limited pediatric conditions	Medium	Langyue expanding dataset Week 2
Data availability gaps	Low	Use imputation + exclusion strategy

Readiness Assessment

- ☒ **Ready for Week 2:** Data schemas documented, integration plan clear
  - ☒ **Collection operational:** Can begin historical data accumulation
- 

**Document Prepared By:** Zuhair Abbas

**Date:** November 17-23, 2025

**Hours Logged:** 10 hours (Week 1 complete)

**Next Deliverable:** `triage_hospital_master.csv` (Week 2)