Hyperdimensional Change Detection for Novel Data Set Exploration and Continuous Unsupervised Monitoring

Sean Coffinger

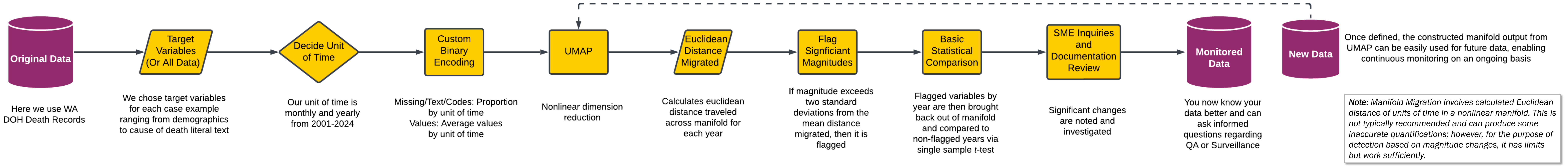
Washington State Department of Health, The Center for Health Statistics

THE PROBLEM

Practitioners who interact with data often utilize unfamiliar longitudinal data sets and are required to detect historical abnormalities and continuously monitor for new and unexpected changes. These data can vary across time in quality, values, coding, etc. The lack of comprehensive documentation of historical variable changes can exacerbate these challenges. To gain insight, a data user can investigate these changes independently to pursue clarification of key changes from responsible subject matter experts. In addition to retroactive analyses of data, the practice of continuous monitoring for quality assurance, and/or informing robust and targeted surveillance, is warranted. Here we suggest a global, unsupervised, and untargeted approach to identify significant changes in individual and sets of variables. The goal is to inform practitioners during the 'exploratory' phase of a project of significant changes. Furthermore, the aim is not to quantify all abnormalities, but rather monitor *when* and *where* interesting changes have occurred with the benefit of generalizability, flexibility, control, speed and automation.

OUR SOLUTION

Change detection via manifold migration can be employed to identify points of interest (Time x Variable). This can be applied to any time scale and number/type of variables.


Original Data
Here we use WA DOH Death Records
Target Variables (Or All Data)
We chose target variables for each case example ranging from demographics to cause of death literal text
Decide Unit of Time
Our unit of time is monthly and yearly from 2001-2024
Custom Binary Encoding
Missing/Text/Codes: Proportion by unit of time Values: Average values by unit of time
UMAP
Nonlinear dimension reduction
Euclidean Distance Migrated
Calculates euclidean distance traveled across manifold for each year
Flag Signficiant Magnitudes
If magnitude exceeds two standard deviations from the mean distance migrated, then it is flagged
Basic Statistical Comparison
Flagged variables by year are then brought back out of manifold and compared to non-flagged years via single sample *t*-test
SME Inquiries and Documentation Review
Significant changes are noted and investigated
Monitored Data
You now know your data better and can ask informed questions regarding QA or Surveillance
New Data

Once defined, the constructed manifold output from UMAP can be easily used for future data, enabling continuous monitoring on an ongoing basis

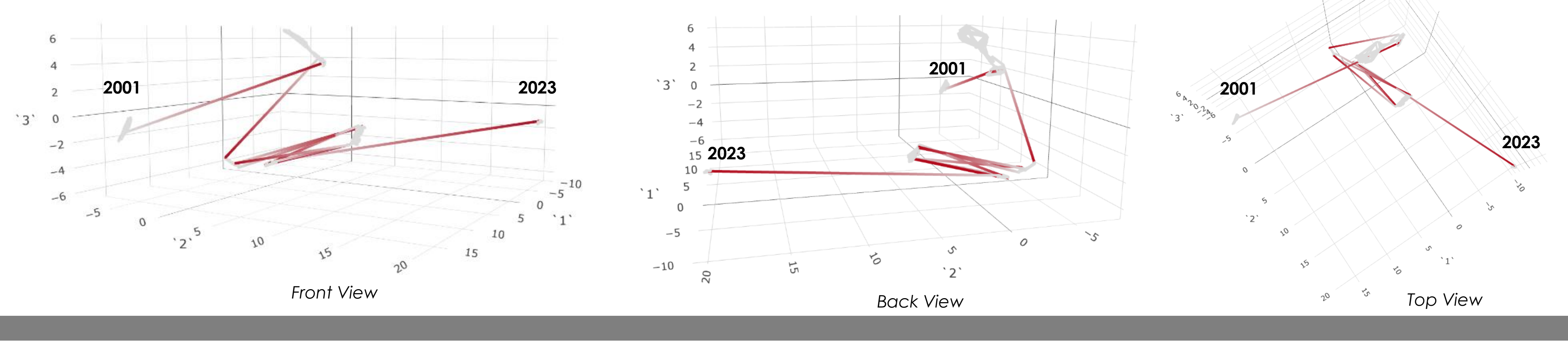
Note: Manifold Migration involves calculated Euclidean distance of units of time in a nonlinear manifold. This is not typically recommended and can produce some inaccurate quantifications; however, for the purpose of detection based on magnitude changes, it has limits but work sufficiently.

CASE #1 – MISSINGNESS EXAMPLE

To investigate global system changes or large shifts in variable usage, we used this strategy to look at global missingness. Here we ingested all descriptive demographic variables used in a typical linkage project and evaluation from the death table and calculated proportion missing by month.

Date	Var_1	Var_2	…	Var_n
01/2001	.01	.12	…	.99
02/2001	.02	.15	…	.99
…	…	…	…	…
03/2024	.99	.09	…	.00

UMAP

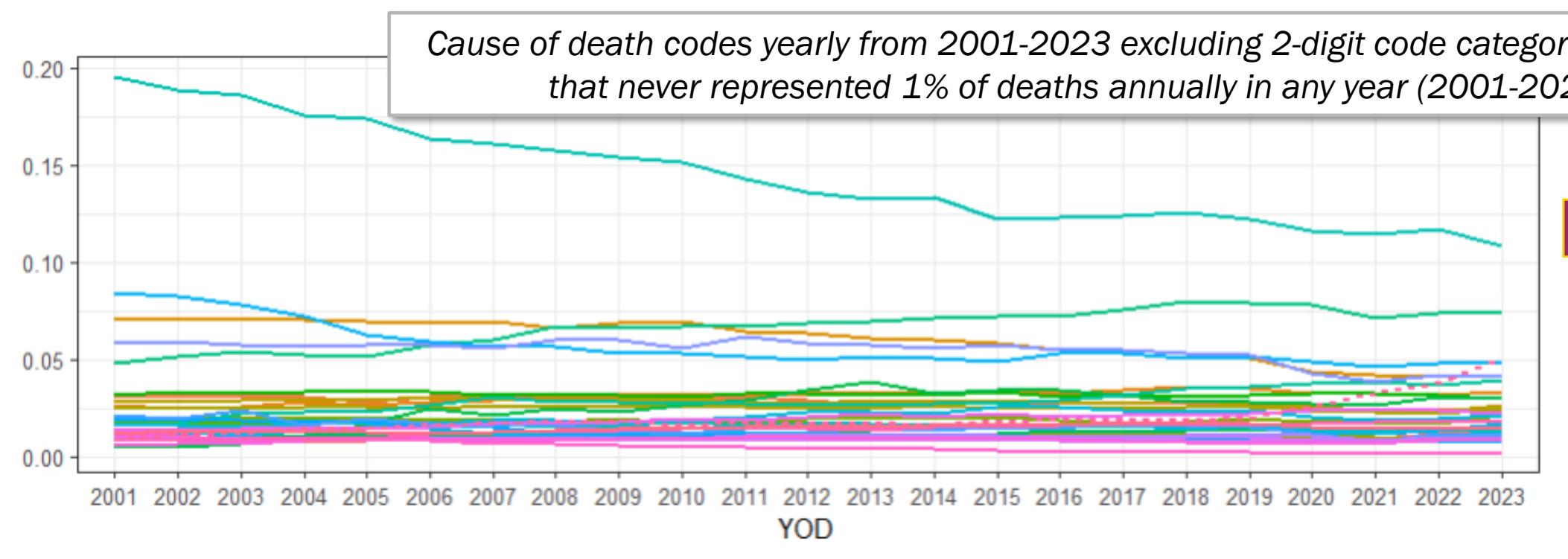


Significant Changes Identified Post-SME Inquiry	Count
Large scale system changes/migrations	2
Mass variable discontinuations	2
Variable value changes	1
Variable category addition	1
Potentially unidentified change	2
TOTAL	8

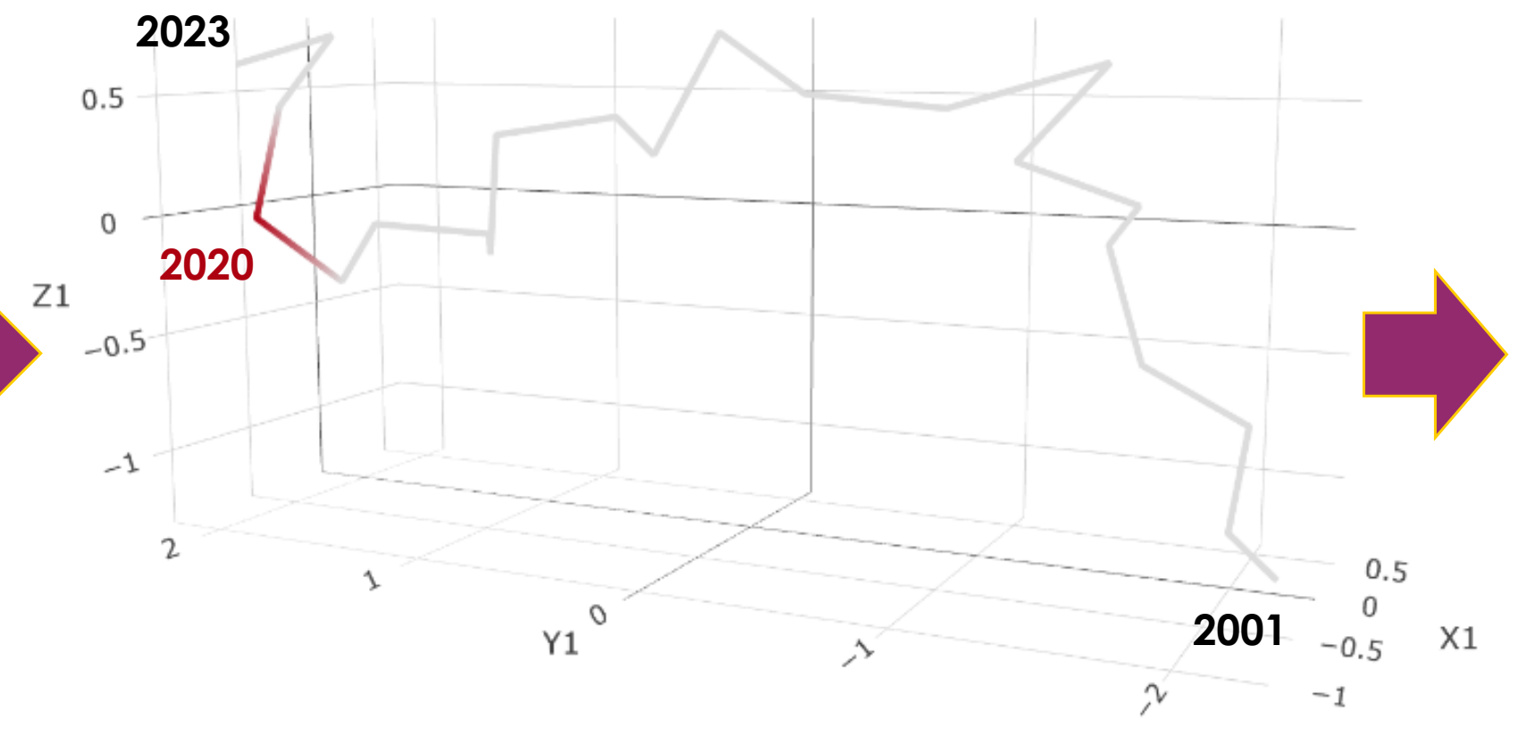
CASE #2.1 – VALUE CHANGE EXAMPLE

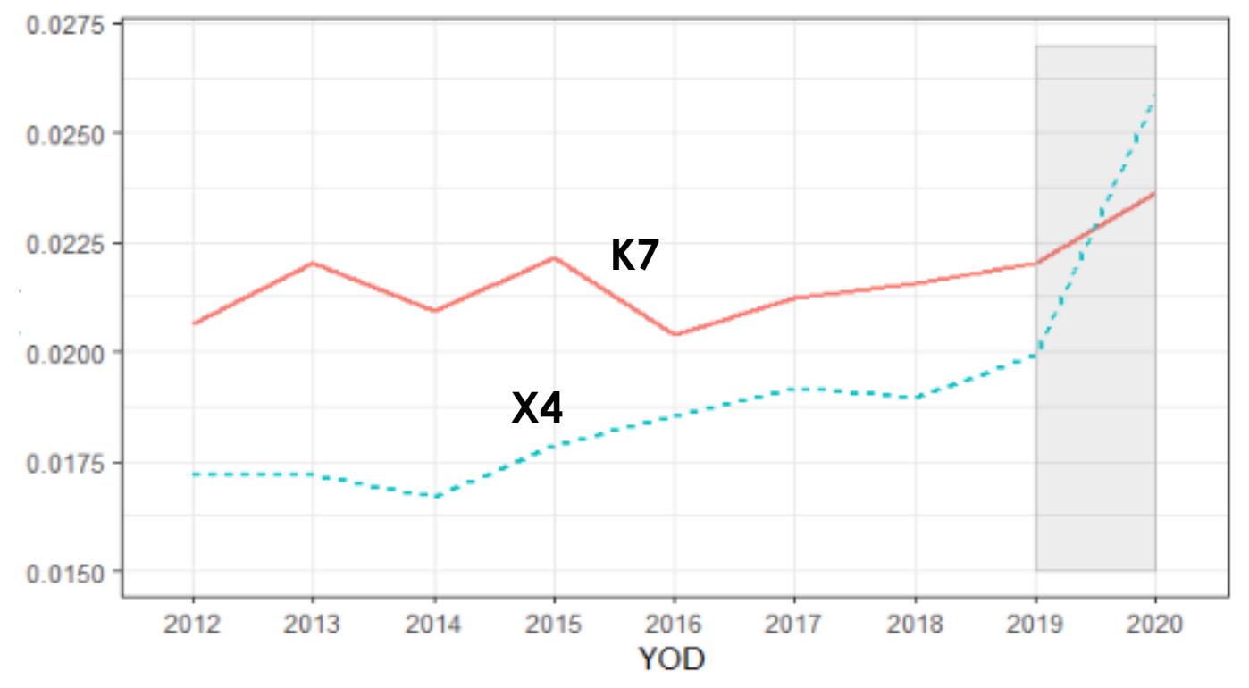
To demonstrate how we can track value changes in variable(s) we used this strategy to monitor significant changes in coded trends. Here we chose the non-quantitative value of Cause of Death (COD). Note this is even easier with quantitative values.

Of course we can already do this 2- dimensionally: set thresholds and detect abnormal changes. However, doing this untargeted can be tedious and inefficient for large variable change detection.


Cause of death codes yearly from 2001-2023 excluding 2-digit code categories that never represented 1% of deaths annually in any year (2001-2023)

UMAP





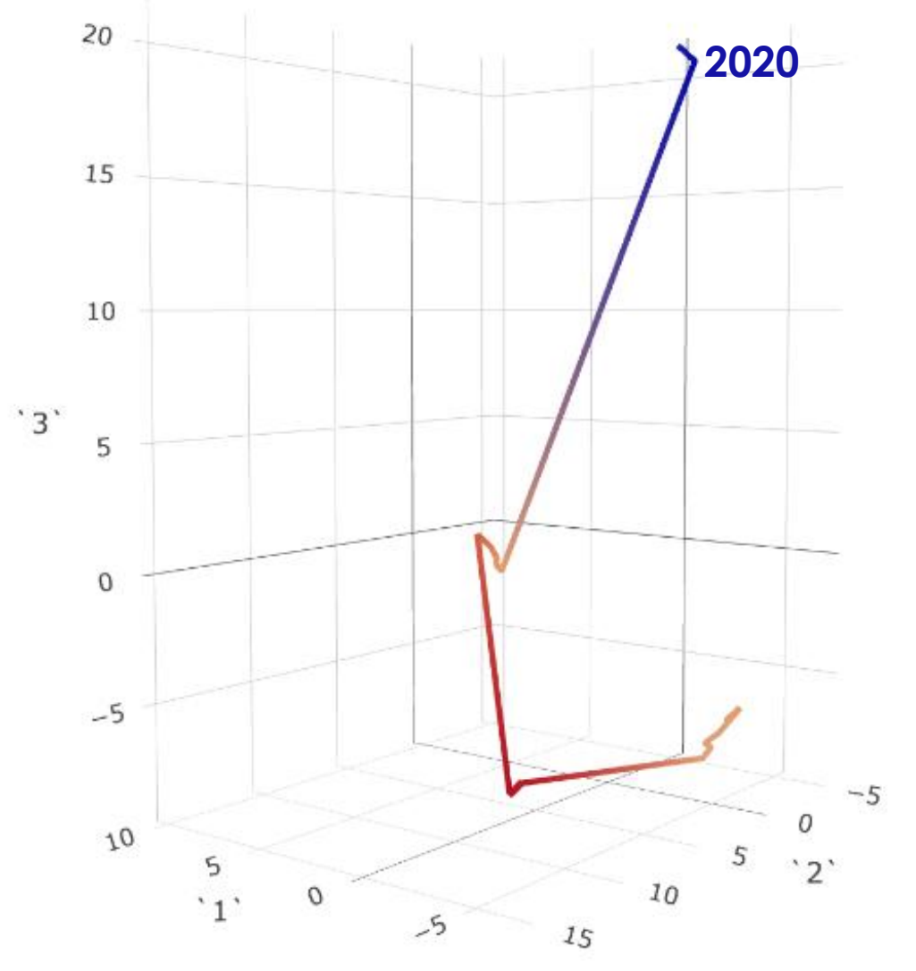
2020 was the only 'flagged' year based on migration distance. We then compared the year-to-year differences (Δ) of proportions coded for each cause of death. A *t*-Test (α=.001) between 2019 to 2020 Δ values and all previous Δs confirmed that several codes experienced a significant change in 2020. Two of these 2-digit ICD-10 mortality codes were detected to have increases: X4 (Accidental poisoning by an exposure to noxious substances) and K7 (Diseases of Liver).

CASE #2.2 – COD LITERALS EXAMPLE

In Case 2.1 we discovered that a year flagged via migration magnitude identified two CODs that significantly increased their trajectory in coded proportion. To demonstrate the ability to utilize these methods and strategy on literals, qualitative and text fields, we employ similar methods to analyze individual words in COD literal text fields that are changing in terms of Δ frequency.

For both K7 and X4 COD records, literals were queried then tokenized into individual words. A medical coding "common garbage word" list was used to filter out common words (i.e., and, or, additional, etc.). Additionally, infrequent words and numeric values were excluded (<1% max annual proportion per code). Annual proportion of incidence for each word was calculated then fed into 4-dimensional UMAP manifold generation.

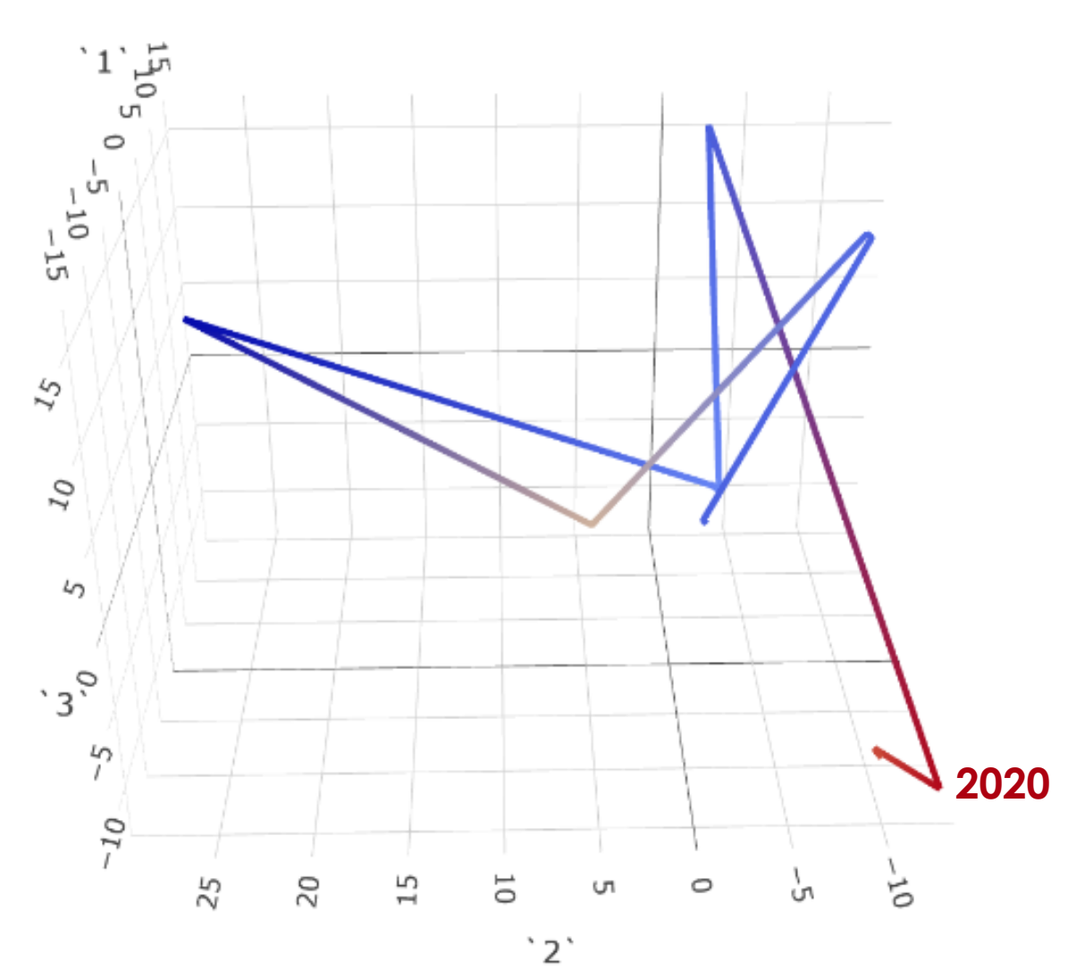
K7



2020 Interesting Words	% Points *Above* Upper 95% CI	
Alcoholic	7.82	*(0.433, CI 0.288-0.355)*
Kidney	4.61	*(0.122, CI 0.031-0.076)*
Cirrhosis	3.88	*(0.729, CI 0.649-0.729)*
Nonalcoholic	3.25	*(0.071, CI 0.015-0.038)*

2020 Interesting Words	% Points *Below* Lower 95% CI	
Alcoholism	4.41	*(0.093, CI 0.137-0.199)*
Renal	1.99	*(0.130, CI 0.149-0.167)*
GI	2.28	*(0.001, CI 0.023-0.056)*
Hepatitis C	0.84	*(0.120, CI 0.129-0.150)*

X4



2020 Interesting Words	% Points *Above* Upper 95% CI	
Methamphetamine	12.00	*(0.441, CI 0.171-0.321)*
Fentanyl	11.39	*(0.383, CI 0.039-0.269)*
Heroin	3.13	*(0.241, CI 0.110-0.210)*
Obesity	0.15	*(0.036, CI 0.024-0.343)*

2020 Interesting Words	% Points *Below* Lower 95% CI	
Oxycodone	3.44	*(0.052, CI 0.087-0.130)*
Diazepam	2.20	*(0.010, CI 0.032-0.055)*
Morphine	1.93	*(0.041, CI 0.060-0.088)*
Hydrocodone	1.51	*(0.022, CI 0.037-0.611)*

SUMMARY

Case #1 successfully demonstrated how we can use manifold migration monitoring to successfully investigate and QA a novel historical dataset. Eight significant changes were detected in demographic variables, six were confirmed by SMEs and documentation. Case #2 demonstrated the generalizability of hyperdimensional change detection by looking at yearly death trends to discover two notable increases in 2020. Then, a separate manifold was constructed to delve further into our initial finding to gain insight on possible contributing factors. In all, manifold migration has limitations compared to targeted analyses, but can provide fast, generalized, untargeted information to any data user.

CONTACT

Sean Coffinger, MA
Health Statistics Manager
Center for Health Statistics
Sean.coffinger@doh.wa.gov

Publication # DOH 422-284
To request this document in another format, call 1-800-525-0127. Deaf or hard of hearing customers, please call 711 (Washington Relay) or email doh.information@doh.wa.gov